Technologies for Sequencing and Interpreting Personal Genomes

A dissertation presented

by

Abraham Meir Rosenbaum

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Genetics

Harvard University

Cambridge, Massachusetts

January, 2010

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Dissertation Advisor: Professor George M. Church Author: Abraham M. Rosenbaum

Technologies for Sequencing and Interpreting Personal Genomes

Abstract

This dissertation is focused on the development of technologies to increase our understanding of the correlation between genomes and phenomes. Particularly, it describes advances in sequencing technology and targeted capture of genomic regions of interest to increase the efficiency of collecting genomic information. Additionally, it discusses improvements in interpreting these genomic regions using existing databases. The introduction starts by describing the nature of our genetic individuality and our understanding of it, beginning with the mapping of genes using extensive pedigrees, followed by the mapping of common diseases to common variants using large populations, and finally the sequencing of genetic material from large populations to correlate phenotypes with rare variants. It then provides a brief review of Second Generation Sequencing (SGS), genomic targeting methods and the clinical applicability of disease-related variants found in healthy individuals. Chapter 2 describes a web-based software tool and its performance in annotating all genomic variants identified in 25 genomes using several general mutation databases, together with an algorithm for identifying those variants with the potential for clinical utility. Additionally, arguments for clinical geneticists prioritizing various sets of genes are presented and the error rates of different sequencing platforms are discussed. Chapter 3 describes molecular inversion probes (MIPs) and their use in targeting exons for sequencing from the first ten participants from the Personal Genome Project. Additionally, it describes various tools

developed to analyze this data. Chapter 4 describes different biases of MIPs and presents design criteria for future experiments targeting genomic regions with MIPs. Appendices A, B and C describe technical improvements to the open source Polonator SGS platform, including the development of a microfluidics flow-cell (Appendix A), the replacement of emulsion PCR amplified beads with rolling circle amplified colonies ("rolonies", Appendix B) and the use of ordered arrays to increase the density of sequenced features (Appendix C). Taken together, these technical improvements represent a >400x decrease in sequencing cost. Appendix D describes the first open-source SGS platform, Appendix E improvements to MIP targeting, Appendix F the MIPTAG Pro algorithm for designing MIPs and Appendix G the analysis of a genome sequenced for a consumables cost of less than $4,400 through the use of rolonies and ordered arrays.

For my family,

Without their support this could not have happened

Acknowledgements

לְדָוִד מִזְמוֹר: לה' הָאָרֶץ וּמְלוֹאָהּ תֵּבֵל וְיֹשְׁבֵי בָהּ (תהלים פרק כ"ד)

I would like to express my gratitude to George for his patience, guidance and inspiration. While his encyclopedic knowledge, creativity and imagination are well known, it is not often that someone with these talents makes them available in such a generous way. Despite his hectic schedule he is extremely adept at making time for his students' questions, even when they are the naïve ones of a young graduate student. Looking back through my notebooks, I am amazed at how futile and poorly planned many of my experiments were, yet George never dissuaded me and encouraged me to pursue my goals at my own pace. His exhortation to "think like a molecule" served me well through all my work with surface chemistry, single molecule analysis and molecular biology, and his encouragement through the slow analysis of hundreds of variants identified in various genomes was invaluable. In terms of patience and easy-going nature, I do not think that I will have a boss like George, even should I be self-employed.

Part of the lure of the Church lab is the amount of experience and expertise present among the lab members, and the camaraderie and sense of uninhibited collaboration fostered by George. Having spent almost six years there, I had the unique privilege to learn from many of my colleagues – I cannot name all of them, but I would like to acknowledge a few. I am thankful to Jay Shendure and Greg Porreca, not only for putting up with my use of their pipette-men when I first joined the lab, but for their patience and genuine desire to help me understand the intricacies of multiplex polony sequencing. Other members of the lab that I am grateful for include Nick Reppas, one of the best molecular biologists that I have ever met; John Aach for his ability to explain and present complicated analyses in an understandable manner; Mark Umbarger a fellow graduate student with whom I could both share triumphs and commiserate over failed experiments; Kyriakos Leptos whose range of interests outside of biology always led to

interesting conversation; Gautam and Morten for adding some spice to the lab; and Sara Vassallo for making the lab a cheerier place, among many others who have positively contributed to my development in the Church Lab.

After getting my feet wet, there were four individuals with whom I closely collaborated and I am particularly indebted to. Brian Chow, an excellent surface chemist with an ever broadening knowledge of biological processes, was a great partner to have in the development of rolonies and nanogrids. Craig Forest, a methodical engineer with whom I was privileged to work with for only a year before he was off to Georgia Tech as a PI, helped me to develop both short term and long term outlooks that proved very useful to my work. Mike Chou and I worked together on the development of MIPTAG Pro; his desire to learn from everyone is truly inspirational. Finally, I would like to thank Sasha Wait Zaranek for his "big picture" thinking and ability to abstract away the minutiae, and for helping me tie together the work done over my graduate training. I would also like to thank him for introducing me to Scalable Computing Experts, where Tom, Ward and Miron were great to work with in the development of bioinformatics tools and the making of these tools user-friendly.

I am also appreciative of all the guidance I received from the Dissertation Advisory Committee over the years as they heard of my triumphs and frustrations. Thank you to Matt Meyerson, Jack Szostak and Joe Jacobson. My dissertation examiners were also very helpful in providing a thorough editing job of this manuscript and an insightful critique of the ideas presented. Thank you to Matt Meyerson, Jim Gusella, David Walt and Matt Warman.

From a young age I was interested in biology and the natural processes; my interest in pursuing a Ph.D. was piqued through freshman microbiology taught by Rabbi Dr. Moshe D. Tendler at Yeshiva College. While the subject matter was interesting, it was really the anecdotes culled from his long career that made his class so fascinating. My summer work with Sara

Shanske and Ali Naini at Columbia University introduced me to DNA diagnostics and mutation analysis and eventually led to my current work. Finally, I would like to express gratitude to all my Rabbis and teachers for constantly encouraging me to accomplish my best.

My parents and grandparents (Martin ob"m and Rose Romerovski and Edmund ob"m and Rosa ob"m Rosenbaum) have been without peer in their support of my learning and education. My parents as well as all four of my grandparents encouraged me, together with my siblings, to take full advantage of the unique opportunities afforded by the American education system, and took great pride in our achievements. Thank you.

My in-laws, Paul and Chavi Jacobs, have also been extremely supportive, not only allowing me to keep their only daughter and half of their grandchildren over 500 miles away for the past six and a half years, but for also encouraging and aiding my research by constantly scanning the popular press for related discoveries. Additional eyes are always welcome!

Finally, I would like to thank my lovely wife Elyan and our children: Maidy, Racheli, Mordechai and Tzvi. People are often surprised at my ability to raise four children while in graduate school – if they knew my wife they would no longer be surprised. She is the epitome of the "Woman of Valor" exemplified in Proverbs 31. We were married 9 days before the start of graduate school and she has been my most ardent supporter since then. Thank you.

הוא היה אומר האב זוכה לבן בנוי ובכח ובעושר ובחכמה ובשנים

(עדויות פרק ב משנה ט)

(Rabbi Akiva) would say: A father merits his son beauty, strength, wealth, wisdom and longevity.

(Mishnah)

ואמר רבי יצחק אין הברכה מצויה אלא בדבר הסמוי מן העין

(תלמוד בבלי מסכת תענית דף ח עמוד ב)

Rabbi Issac said: good fortune is only found in items that are hidden from the eye.

(Babylonian Talmud)

Table of Contents

Chapter 1	Introduction: Personal Genomics	1
Chapter 2	Clinical Analysis of Individual Genomes	26
Chapter 3	Initial Data Release from the Personal Genome Project	115
Chapter 4	Analysis of MIP Targeted Sequencing Biases and Recommendations for Future Design	155
	Future Directions	175
Appendix A	DNA Sequencing by Ligation on Surface-Bound Beads in a Microchannel Environment	186
Appendix B	Multiplex Rolony Sequencing	192
Appendix C	Nanogrids for Creating Self-Ordered Arrays of Library Molecules for Second Generation Sequencing	214
Appendix D	Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome	225
Appendix E	Multiplex padlock targeted sequencing reveals human hypermutable CpG variations	231
Appendix F	Computational design of molecular inversion probes for targeted genomic sequencing using MIPTAG Pro	242
Appendix G	Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays	276

Figures and Tables

Figure 2-1	Approach and Results	32
Table 2-1	Variants with Sufficient Evidence to Warrant Sequence Confirmation and Clinical Follow-up	36
Figure 2-2	*MYL2* Ala13Thr Pedigrees	39
Figure 2-3	Agreement between Substitution Calls for a Single Individual across Different Studies	41
Figure S2-1	Correlation between HapMap and 1000 Genome Project Frequencies for Identified Alleles	49
Figure S2-2	Previously dbSNP Incorrectly Calculated Amino Acid Changes for Codons Split by Splice Junctions	50
Figure S2-3	Comparison of Genome Statistics	76
Table S2-1	Source Data	79
Table S2-2	Results of Processing Genomic Data by Trait-o-matic	80
Table S2-3	Systematic Manual Processing of HGMD/OMIM Results	81
Table S2-4	Variants with Potential Phenotypic Expression	82
Table S2-5	Variants that according to their HGMD citation should have phenotypic expression, but later research (not cited in HGMD) raise doubts as to this conclusion	85
Table S2-6	Common Variants of Clinical Interest in our Genomes	88
Table S2-7	High Odds Ratio GWAS Variants	89
Table S2-8	Drug Dosage Variants	90
Table S2-9	ApoE Status for Nine Full Genome Sequences	93

Table S2-10	Variants Implicated in Disease with Unreported Frequencies, but Appearing Frequently in YRI Genomes	94
Table S2-11	Consensus between Illumina (Bentley et al.) and SOLiD (McKernan et al.) Sequencing of NA18507	95
Table S2-12	Variation Frequency in OMIM/Genetests Genes	96
Figure 3-1	Overview of the MIP Capture Process	119
Figure 3-2	Reproducibility of Capture and Downstream Steps	122
Figure 3-3	MIP Capture Efficiency	124
Figure 3-4	EcoP15i Digestion	126
Figure 3-5	Initial Genomerator Concordance	129
Figure 3-6	Independent Read Analysis	132
Figure 3-7	Comparison of Quality/Coverage Thresholds from Two Libraries	135
Figure 3-8	SNP Composition and Comparison with Other Sequencing Efforts	137
Table 3-1	Thresholds, Estimated Target Size and SNPs for Each Library	138
Table 3-2	Analysis of Variants Prioritized by Trait-o-matic	141
Figure 4-1	Polymerase Effect on Capture Efficiency	160
Figure 4-2	Targets Analyzed by Length	162

Figure 4-3	Targets Analyzed by GC Content	164
Figure 4-4	Targets Analyzed by Extension Arm Tm	166
Figure 4-5	Targets Analyzed by Ligation Arm Tm	167
Figure 4-6	Targets Analyzed by Probe Secondary Structure	169
Figure 4-7	Targets Analyzed by Target Secondary Structure	170
Figure 4-8	Pair-wise Analysis of Target Capture Variables	172
Figure 5-1	Mask Design and Fabrication Process	189
Figure 5-2	Photograph of Channel with C_4F_8 Surface Passivation	189
Table 5-1	Bead Binding Selectivity on Microchannel Walls	189
Figure 5-3	Assembled Vacuum Chuck and Evaluation of Temperature Controls	190
Figure 5-4	DNA Sequencing by Ligation Demonstrated in a 50 μm Wide Channel	190
Figure 6-1	Surface Bound Primer Extension	197
Figure 6-2	Secondary Extension off Surface Bound Primers	197
Figure 6-3	Binding Concatemers Post-Amplification	199
Figure 6-4	DNA Fiber Stretching	200
Figure 6-5	Fluorescence Intensity as a Function of Polymerization Time	201

Figure 6-6	Rolony Sequencing by Ligation	203
Figure 7-1	Surface Modifications for Rolony Self Assembly	217
Figure 7-2	Versatility and Limitations of Pattern Size and Treatment	217
Figure 7-3	Passivation Difficulties	218
Figure 7-4	Rolonies Bound to E-beam Grids	219
Figure 7-5	Clonality Analysis of Concatemer Hybridization	220
Figure 7-6	Ordered Bead Arrays	222
Figure 7-7	Residual Resist after Descumming	222
Figure 7-8	Rolonies on UV and 2-beam Interference Lithography Generated Nanogrids	223

Chapter 1

Introduction – Personal Genomics

Sir Archibald Garrod introduced the concept of chemical individuality in 1902[1], and later wrote in *Inborn Errors of Metabolism*:

> The delicate ultra-chemical methods which the researches of recent years have brought to light … teach the lesson that the members of each individual species are built up of their own specific proteins, which resemble each other the more closely the more nearly the species are related…. The existence of chemical individuality follows of necessity from that of chemical specificity, but we should expect the differences between individuals to be still more subtle and difficult of detection (page 2)[2].

Less than a decade later, Morgan described chromosomes as the basis for the heritability of many aspects of our individuality[3], and Sturtevant began mapping phenotypes to these chromosomes[4].

The inability to port many of the tools used for non-human genetics to human genetics, however, significantly slowed this endeavor in humans. It was not until 1968 that the first phenotype linked to a non-sex chromosome was described using a cytogenic approach[5]. Furthermore, the molecular basis of human phenotypes was primarily connected with the causative mutation in the protein[6], not the causative chromosomal change. The elucidation of the structure of DNA[7] greatly increased our understanding of the phenotype-genotype connection, and this process was aided by mapping phenotypes to chromosomal regions utilizing restriction fragment length polymorphisms[8-9]. The technologies enabling practical DNA sequencing[10-11], and the concept of using DNA polymorphisms to map genes in humans[12] led to the identification of the first nucleotide variant, a trinucleotide repeat expansion, associated with a disease in 1983[13].

While these techniques allowed for mapping of monogenic highly penetrant mutations in families, the mapping of quantitative trait loci with less-pronounced

phenotypic effects remained elusive. In 1996-1997 a number of groups proposed that causes of common diseases can be traced to common variants (the CV-CD hypothesis)[14-16]. Completion of the HapMap project[17-18] which cataloged common single nucleotide polymorphisms (SNPs) together with the realization that linkage disequilibrium allowed relatively few SNPs to identify the probable genotype of all 3.1 million SNPs enabled the analysis of millions of positions in an affordable fashion and the analysis of entire related populations instead of families. The cost effectiveness of the process allowed for studies exceeding tens of thousands of individuals[19] in genome wide association studies (GWAS). While technically any study of the human genome for chromosomal regions associated with disease is a "GWAS," this term is usually specifically used to refer to analyzing common variants for their associations with common phenotypes, typically quantitative trait loci. For a comprehensive review of the history and capabilities of genetic mapping, see Altschuler, et al.[20].

The history of human genotype-phenotype correlation from its initial family-based studies searching for chromosomal aberrations and chromosomal markers, and then gradually to molecular markers and population based large-scale screens for common variant associations, has yielded the molecular basis for a large number of phenotypes. The CV-CD hypothesis, however, has not yielded the amount of anticipated correlations and our understanding of genetic causes for our individuality remains woefully inadequate.

The current database of simple nucleotide polymorphisms (SNPs in dbSNP 130) maintained by The National Center for Biotechnology Information (NCBI) reports almost 18,000,000 variants in the human genome, and for variants comprising duplications and

insertions of at least 100bp, the database for copy number variations (CNVs, http://projects.tcag.ca/variation/[21]) contains almost 50,000 entries as of August, 2009. Additionally, each of the initial ten whole genome sequences reports another approximately 300,000 single nucleotide changes not on this list[22-30], as well as numerous CNVs and small insertions/deletions. While all these data contribute to a person's individuality, perhaps the most interesting variants are those associated with phenotypes.

As of 2009, the Online Mendelian Inheritance in Man (OMIM), a NCBI-supported repository for genetic variation lists 2,239 genes correlating with phenotypes[31]. While the majority are monogenic, the database also includes oligogenic, quantitative trait loci and epigenetic dependent variants[32]. Shortly after the publication of the rough draft of the Human Genome Project[33-34] in 2000, OMIM contained 1000 genes with associated phenotypes[35]. Many assumed that the rough draft would usher in a new era for the discovery of disease genes[36]. While undoubtedly it has helped, the list of phenotype-associated genes has little more than doubled to 2,239 in the nine years since then[31], while the cost of DNA sequencing has dropped by over six orders of magnitude[28].

This dramatic reduction in sequencing cost has prompted a number of researchers to take a new direction in genetic mapping. Instead of using microarrays to sample only common variants in large populations, targeted sequencing and whole genome sequencing can enable associations of even rare variants. These proposed projects involve sequencing large amounts of genetic material from large numbers of individuals to elucidate both the full extent of the heritability of our individualities and the phenotypic effects of the remaining 80% of genes. While some groups have opted to sequence large cohorts without collecting phenotype data[37], others collect this data but only for clinically

relevant traits[38]. One group has estimated that to achieve a 40% chance of discovering genes associated with an effect of 0.25σ, (e.g., a difference of 0.5 inch from mean height) in the European population, one would need to phenotype 100,000 individuals and sequence all genes from 5,000 individuals from each of the highest and lowest quintiles[39]. For an effect of 0.5σ (e.g., a difference of 1 inch from mean height) the power would increase to 77%. Following along these lines, the Personal Genome Project (PGP) was formed to develop "highly integrated and comprehensive human genome and phenome datasets[40]," and it currently has IRB approval to accept 100,000 participants. The comprehensiveness of the phenotype is crucial for this project, as errors caused by incomplete phenotypic information have plagued a number of GWAS[41-43]. This is only possible through maintaining a connection with the participants and obtaining cell lines for additional *in vitro* studies.

This dissertation explores a number of areas related to the PGP and the study of personal genomes in general. Particularly, it focuses on the transition from using microarrays to map common variants to technologies allowing cost-effective targeted sequencing of large amounts of genetic material and ultimately entire genomes from large numbers of individuals. These include: (1) improvements to second generation sequencing (multiplex, cyclical enzymatic-based sequencing methods) to decrease cost by more than one order of magnitude, (2) improvements to targeted sequencing to enable cost effective sequencing of particular subsets of genes. Additionally, since thousands of genomic variants will be generated for each participant, undoubtedly some will be already identified as being clinically important and should be prioritized when data is made available to the volunteers. This dissertation also focuses on (3) defining these

variants and creating a user-friendly method for presenting these variants, together with all annotated variants, to the individual and the community. This repository is designed as an interactive database that can be updated in a wiki-like fashion to reflect new discoveries. We envision the addition of phased data, DNA and RNA editing data, epigenetic data as well as tissue specific and environmentally specific information to our understanding of phenotypes, and eventually this database, as our ability to assess these aspects of our individuality in a high throughput manner is developed.

Second Generation DNA Sequencing Technologies

A prime development enabling the transition from associating common variants with phenotype to associating rare ones was the advent of cyclical multiplex sequencing technologies in 2005[44-45]. These technologies were originally referred to as Next Generation Sequencing platforms (NGS) to describe their replacing Sanger sequencing, the "First Generation." As even newer technologies are being developed, the preferred term has come to be Second Generation Sequencing (SGS). The definition of SGS is still debated; I prefer using it to describe cyclical enzymatic sequencing reactions on prepared library molecules with fairly short read lengths. Pure third generation would refer to methods with none of these aspects (i.e. processive, non-enzymatic sequencing of DNA requiring no preparation with extremely long read lengths). With the ongoing development of second generation sequencing technologies, the reagent cost for DNA sequencing has decreased by an average of one order of magnitude every year since then, to the current cost of $0.5/megabase. Currently, there are six commercially available platforms, reviewed extensively by Shendure and Ji[46] who cover the basic technologies

and capabilities of all but the most recent platform. The most relevant statistics for the platforms reviewed are compiled in their Table 1.

The first platform, 454 FLX, relies upon emulsion PCR amplified polystyrene beads[47-48] and pyrosequencing in fiber-optic microarray microwells[24,44]. In this platform library molecules are amplified onto 28-micron beads so that each bead has many thousand copies of the same molecule. The beads are then deposited into picoliter-scale wells. In a cyclical fashion, a single nucleotide species together with luciferin and adenosine 5'-phosphosulfate is deposited in each well. With incorporation of the correct nucleotide (or nucleotides in the case of a homopolymer run) the level of light released is detected and registered. This is the only platform capable of contiguous read-lengths in the 400-800bp range, but it has the highest consumable cost of $47 / megabase.

The Illumina Genome Analyzer II relies upon a modified version of bridge-PCR[49-51] to amplify library molecules using surface bound primers[23]. Four-color reversible terminators are incorporated into the complementary strand, and the fluorescent signal is registered via a CCD camera through a total internal reflection (TIRF) microscopy setup. The device is capable of sequencing a contiguous read length of ~100 bp and has an estimated consumable cost of <$0.5 / megabase.

The Polonator and SOLiD platforms utilize emulsion PCR to amplify the library onto one-micron beads which are then covalently attached to a modified glass surface. The nucleotide identity is detected through ligation[45] utilizing either nonamers and single-base encoding for Polonator[45] or octamers with two-base encoding for SOLiD[29]. SOLiD is capable of contiguous 50bp reads and has a consumable cost similar to that of

Illumina while Polonator, while maintaining a similar consumable cost is only capable of contiguous 13bp reads.

Complete Genomics (CGI) only offers a sequencing service for whole human genomes using their platform. Library molecules are circularized and amplified via rolling circle amplification[52-57]. These molecules are then electrostatically attached to a slide patterned with sub-micron sized chemically-activated spots, and sequence is detected with combinatorial probe anchor ligation[28]. 20bp of contiguous sequence can be read, at a published consumable cost of $0.0124 / megabase.

The Heliscope platform is a quasi-third generation sequencing platforms in that single molecules are sequenced, but the process is still cyclical, enzymes are required and library preparation is still necessary. Single molecules are A-tailed with terminal transferase and hybridized to surface-bound primers consisting of thymine homopolymers. Single-base four-color extensions are performed and the signal is detected with a customized microscope setup. The read-length average is 33bp and the published consumable cost is $0.6 / megabase[30,58].

Other than consumable cost, four other factors of note in platform comparison are accuracy, throughput, machine cost and library preparation. In terms of accuracy, the reported false positive rates for raw reads range from 3.5% for Heliscope to 0.06% for SOLiD due to their built in error correction with two-base encoding. Since the current trend is to rely upon read depths of 30-40X to minimize heterozygous dropouts, the processed false positive rates for each platform should be negligible. Regarding throughput and machine cost, each of the platforms are roughly within a 2-fold range. In terms of library preparation, Helicos requires the fewest enzymatic steps while Complete

Genomics requires the most. The cost to generate a library for CGI, however, is only a small fraction of the total cost of consumables.

While each of these platforms is constantly being updated to decrease cost and improve read-length and accuracy – the advantages presented by the Complete Genomics platform are worth noting. Their cost-savings is primarily due to two factors: (1) the cost of amplifying the library and placing it in the flowcell, and (2) the amount of reagent required per library molecule. SOLiD and Polonator have a high cost of generating the clonally amplified beads, but a low cost to covalently attach them to the surface, while Solexa has a low cost to amplify via bridge-PCR, but a high cost to create the flow-cell within which they are amplified due to the complicated manufacturing procedure and surface chemistry involved. 454 has a high cost for both the emulsion PCR and fiber-optic microarray generation. While both Heliscope and Complete Genomics have very low costs for library amplification and library attachment to the sequencing surface, the cost of manufacturing the flow-cell for Heliscope is significantly higher than that of CGI.

The second contributing factor is influenced by the density of library features, the height of the flow-cell and the need to use enough reagents to drive the reaction to completion. Published results show that the CGI ordered array allows the packing over 500,000 features / mm^2, with a readily achievable density over 2,000,000 features / mm^2. The Polonator has demonstrated a maximum feature density of 600,000 features / mm^2, Illumina GA2 250,000 features / mm^2, SOLiD3 160,000 features / mm^2, Heliscope 70,000 features / mm^2 and 454 FLX 480 features / mm^2. In terms of flow cell height, Heliscope obtains the lowest volume with a 5 micron chamber height, the CGI platform utilizes 50 micron beads to form a gasket, and the Polonator relies upon an 80 micron

gasket. SOLiD3 and Illumina GA2 both have flow-cell heights exceeding 100 microns and 454 FLX has a height of 300 microns. Since with increasing feature density the savings scale to the squared, and with decreasing flow-cell volume the savings scale in a linear fashion, the advantages of CGI are demonstrable. Finally, the ligation-based platform utilized with Polonator and Complete Genomics is unchained, so that the second cycle of ligation does not rely upon the efficiency of the first. With the exception of Heliscope, the nature of the cyclical chained reaction requires that enough reagent to drive the cycle to completion so that the molecules in each cluster, or polony, do not fall out of phase. With an unchained reaction each cycle is an independent event there is no need to drive the reaction to completion and a lower concentration of sequencing reagents can be used. Additionally, the CGI platform uses capillary action instead of a pump to exchange reagents, ensuring that no reagents are lost to large lengths of tubing.

In this dissertation technological advances designed for the open-source Polonator platform will be presented. They include (1) a modified flow-cell with a height of 20 microns and potential height of 8.5 microns (Appendix A), (2) the preparation of a library amplified through rolling circle amplification and sequenced with both synthesis based and ligation based sequencing (Appendix B), and (3) the development of an ordered array to enable a higher density of either rolling circle amplified library molecules or emulsion PCR amplified beads (Appendix C. Taken together, these modifications, when fully integrated into the Polonator platform, have the potential to drive down the cost of sequencing to less than $0.003/megabase and enable the collection of large number of genome sequences for phenome-genome correlations.

Targeted Sequencing

Ideally, to obtain the genetic component of individuality the entire genome of each person should be sequenced and phased. Unfortunately, even with the cost reductions enabled by second generation sequencing, this is often not affordable and targeted sequencing of a genomic subset is done instead. The first methods to sequence many genomic regions relied upon automating millions of PCR's[59-60]; this was achieved but at a very high cost and low throughput. To better match the speed of second generation sequencing, a number of different targeting methods have been developed. Generally, they can be categorized as solid-support hybridization enrichment, solution-based hybridization enrichment and PCR-based selection.

Three publications introducing microarray-based hybridization enrichment appeared in 2007. The first two studies created whole genome libraries and then, through hybridization to a single Nimblegen microarray of oligos complementary to the region of interest, enriched for 304kbp[61] or ~5mbp[62]. A third study utilized seven custom microarrays to target all exons[63]. Each of these methods used ~20 micrograms of genomic DNA, and led to the commercialization of the Nimblegen Sequence Capture 2.1M Human Exome array targeting 180,000 exons and 551 miRNA exons. Targeted capture has also been accomplished using less genomic DNA by Febit Technologies in a microfluidic apparatus with 185kbp of target and tiled probes[64] and using a typical microarray setup targeting 4mbp[65]. Recently, Illumina has also developed a microarray targeting all exons which has been successfully applied in uncovering an autosomal dominant and autosomal recessive disease-associated gene[66-67]. An alternative method based upon the Southern blot relies upon an initial series of PCRs to create the probes,

but through this process an immortal probe set is generated. These probes are then attached to a filter which was successfully used to enrich for 115kbp[68] from a whole genome library.

An alternative, solution based method was introduced in 2009. This method relies upon chip-synthesized long oligomers from Agilent that are cleaved off the microarray, processed and bound to one micron ferromagnetic beads. A library of genomic DNA is mixed with these beads, and the targets are enriched for through hybridization to the complementary sequence. This technique allows for targeting of much smaller amounts of library DNA due to the greater efficiency of solution based hybridization. This method was used to target 2.5mbp[69] and 3.9mbp[70] from a genomic DNA library and cDNA from 467 genes from a cDNA library[71]. This technique is commercialized as the Agilent SureSelect Target Enrichment System.

Two different PCR-based methods have been used to successfully enrich for targets of interest: molecular inversion probes (MIPs) and microdroplet PCR. Multiplex PCR is very inefficient due to the creation of primer-dimers and other spurious material[72]. Careful design of primers has allowed for the detection of thousands of SNPs[73], but there have been no published technique successfully targeting longer regions. The first approach, based upon MIPs[74-76] relies upon Agilent probes (produced as above) where each oligomer consists of three parts (from the 5' end): a "ligation arm" complementary to a region upstream from the target, a universal sequence, and an "extension arm" complementary to a region downstream from the target. For the capture reaction polymerase is used to extend the 3' end of the probe, copying the target region, and ligase is used to seal the polymerized strand to the synthetic one. The first attempt at targeting

6.7mbp, captured only 22% of targets[77], but later attempts showed better results for 485 targets[78], the original 6.7mbp and 1.7mbp[79], and 500kbp of 10bp regions around CpG sites[80]. The second method, microdroplet PCR, utilizes the RDT1000 (Raindance Technologies) to simultaneously perform almost 4,000 PCR reactions in isolated microdroplets, targeting 1.49mbp[81].

Accurate comparison of these methods is very difficult due to the different targets chosen by each group and the different goals for each publication[81-82]. Each of the platforms achieves a high capture rate with a high level of accuracy, and each is under constant improvement. Currently, there are a number of differences between hybridization and PCR based targeted capture. PCR methods fair better insofar as (1) all hybridization based methods require the initial creation of a shotgun library, while the PCR methods are "library-free." (2) Hybridization methods cannot differentiate efficiently between homologous genes or any sequence with strong similarity to a target, while PCR based methods can more easily rely upon a few SNPs in designing the primers. (3) Sequence adjacent to the probe sequence will be captured also ("near target capture") with the hybridization methods depending upon the size of fragment lengths of the shotgun library. It should be noted, however, that this third aspect has proved beneficial in cDNA targeting where unexpected gene fusions were uncovered due to near target capture[71]. The benefits of hybridization capture are (1) greater uniformity of captured species when compared with MIPs, (2) simpler probe design algorithms and (3) greater flexibility in allowing SNPs in the target region and (4) lower cost of entry when compared with the RDT1000 microfluidics device. It should also be noted that the high reproducibility of MIP capture has led to the segregation of probes into a number of

subsets that can be normalized and combined for sequencing. This allows MIPs to achieve a level of uniformity exceeding that of single-microarray whole exome capture.

Improvements to MIP based targeted sequencing are presented in this dissertation in the context of sequencing large numbers of exons for each of the first ten participants in the PGP (Chapter 3). Furthermore, some capture biases are described and overcome while others are noted so as to circumvent them with new algorithms for designing the molecular inversion probes (Chapter 4 and Appendix F).

Sequence Analysis

The new ability to efficiently scan an entire genome (or with the above targeting techniques, a whole exome) for rare variants have already yielded success in the analysis of individuals with specific phenotypes. Specifically, novel rare variants have been associated with Freeman-Sheldon syndrome[66], Miller syndrome[67], Bartter syndrome[83] and acute myeloid leukemia[84-85]. In these reports, the algorithms utilized by each group were custom tailored for the predicted mode of disease-inheritance. Various genome sequencing methods are being developed to aid in the discovery of rare variants connected with specific phenotypes, such as cancer[86]. These are in addition to more general analysis algorithms such as SIFT[87] and PolyPhen[88] which are designed to predict the effect of nonsynonymous variants.

Large studies searching for disease variants, however, will undoubtedly uncover variants of importance (incidental findings) not associated with the disease studied, especially if the study is looking for variants associated with non-disease phenotypes. Whether and how to report such data has been a matter of discussion, both in theory with

the recommendation in 2008 for a two year study to assess different possibilities[89], and in practice with whole genome scanning for CNVs to diagnose autism spectrum disorder in a young child revealing a deletion in BRCA1inherited from the probands asymptomatic 24yo mother[90]. This latter case is a variant of unknown phenotypic effect, and while they may be quite prevalent in these types of studies[91], there is little consensus as to what should be reported even when the literature identifies the consequences of the particular mutation. Corriel has developed an Informed Cohort Oversight Board (http://cpmc.coriell.org/Sections/About/SAB.aspx?PgId=51) to individually assess each genetic variant reported in their Personalized Medicine Collaborative. The criteria for inclusion are "(1) actionability of disease based on available medical or lifestyle interventions, and (2) the association between the disease and the genetic variant." The complexity in these decisions is noted by Khoury and Wagener, who at the beginning of the Human Genome Project expounded upon the possibilities of using probabilistic data associated with common variants for disease prevention[92]. They report that a variant with a sevenfold increased risk for preeclampsia may contribute to 50-80% of cases in a given population, but if the variant prevalence is >65%, and preeclampsia only occurs in 5% of cases, it clearly has little positive predictive value (PPV). This is echoed by more recent opinions explaining why individual based genetics will not be replaced by variant based genomics[93]. While Khoury does profess optimism for certain genetic variants improving individuals' health (although not most GWAS-based ones[94]), others are not as optimistic[95]. Khoury, however, does make an argument for variants with relative risks exceeding 50x, a few of which have already been identified[96-97].

While it may be enticing to specifically report variants already used for diagnostic sequencing (NCBI maintains a web-based list of diagnostic labs and disease genes covered for health care providers[98]), the variant level information utilized by diagnostic companies is rarely available. This has led efforts such as Phenopedia and Genopedia[99] to report only gene level information. Clinically relevant moderately penetrant variants, such as those causative for hemochromatosis (HFE C282Y), have been touted as a good example of the utility of genetic information for primary care physicians[100]. A study by Quest Diagnostics, however, showed that this is actually more complex[101], as "a homozygous patient for C282Y who also has increased ferritin saturations and arthritis is almost certainly affected with (hemochromatosis). However, the same genotype in an asymptomatic patient tested because of a family history of (hemochromatosis) indicates that the patient is at <1% risk of developing (hemochromatosis) in his or her lifetime[102-103]." Others have also recommended that variant associations can be applied most appropriately only when more phenotypic data is available[104]. For variants associated with penetrant recessive diseases for which the participant is heterozygous, the information would primarily be useful for family planning (e.g. variants detected by Dor Yeshorim[105] for the Ashkenazi Jewish population) and not the participant's personal health, which raises yet another set of ethical, legal and social implications (ELSI).

Additionally, the actionability of diagnostically used variants in the absence of any symptoms remains questionable. A recent review[106] notes the successes of testing for phenylketonuria[107] and the so-far unique case of BRCA testing warranting drastic measures[108]. The authors note, however, that for most variants "the constellation of behaviors that constitute a healthful lifestyle—regular exercise, a diet low in fats and high

in fruits and vegetables, and avoidance of smoking and obesity," are the advised course of action irrespective as to whether the variants is present. While some have argued that the knowledge of one's genetic predispositions, however inapplicable they are on an individual basis, may influence the adoption of a healthier lifestyle[109-110], there are those that question these results[111].

An exception to these studies claiming the irrelevance of such data for clinical utility, appears to be variants associated with pharmacogenetic affects[112], where early access to this information may have a high impact on clinical treatment. PharmGKB[113], a database of such variants currently lists 14 drugs with pharmacogenetic recommendations, but this is likely to increase as GWAS begin pursuing more pharmacogenetic variants[114].

To mitigate some of these complex questions as to the clinical relevance of variants, the PGP has adopted a comprehensive entrance exam to ensure that all participants are aware of the predictability of phenotypes based upon our current knowledge. Additionally, they consent to release all their phenotypic and genotypic data to the community of researchers and the general public. The questions above, therefore, are applicable only to the prioritization of variants, i.e. which ones we should bring to a participant's attention and recommend clinical follow-up for. These issues are more fully discussed in Chapter 2 and Future Directions, along with the potential utility of various general mutation databases for defining variants of clinical interest. We create a tool called Trait-o-matic to prioritize variants based upon existing genotype-phenotype association databases, yet we conclude that there is no perfect database for this purpose, and have begun the creation of a new database called genomes-environments-traits

(GET). Initial population of this database will be with both rare and common variants, with those that are applicable to individual clinical utility clearly marked. Associating the individual phenomes with these variants will help elucidate the penetrance of these variants and, possibly the existence of modifying variants. Recently, high throughput technologies enabling genome-wide analysis of epigenetic effects[115], DNA and RNA editing[116-117], allele specific and tissue specific expression[118-120] and chromosomal structure[121], have been developed. As the usage of these and other genome-wide assays increase and correlations with phenotypes are discovered we anticipate adding these to the database. The name of this database appropriately frames the goals of sequencing personal genomes – to define the chemical individuality of each person.

References

1 Garrod, A. E. The incidence of alkaptonuria: a study in chemical individuality. 1902 [classical article]. *Yale J Biol Med* **75**, 221-231 (2002).

2 Garrod, A. E. *Inborn Errors of Metabolism*. On-line Fascimile Edition: Electronic Scholarly Publishing edn, 2 (Henry Frowde and Hodder & Stoughton, 1923).

3 Morgan, T. H. Chromosomes and Heredity. *The American Naturalist* **44**, 47 (1910). <http://www.esp.org/foundations/genetics/classical/holdings/m/thm-10b.pdf>.

4 Sturtevant, A. H. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *Journal of Experimental Zoology* **14**, 43-59 (1913).

5 Donahue, R. P., Bias, W. B., Renwick, J. H. & McKusick, V. A. Probable assignment of the Duffy blood group locus to chromosome 1 in man. *Proc Natl Acad Sci U S A* **61**, 949-955 (1968).

6 Baglioni, C. The fusion of two peptide chains in hemoglobin Lepore and its interpretation as a genetic deletion. *Proc Natl Acad Sci U S A* **48**, 1880-1886 (1962).

7 Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953).

8 Grodzicker, T., Williams, J., Sharp, P. & Sambrook, J. Physical mapping of temperature-sensitive mutations of adenoviruses. *Cold Spring Harb Symp Quant Biol* **39 Pt 1**, 439-446 (1975).

9 Sambrook, J., Williams, J., Sharp, P. A. & Grodzicker, T. Physical mapping of temperature-sensitive mutations of adenoviruses. *J Mol Biol* **97**, 369-390 (1975).

10 Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74**, 560-564 (1977).

11 Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-5467 (1977).

12 Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* **32**, 314-331 (1980).

13 Gusella, J. F. *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234-238 (1983).

14 Lander, E. S. The new genomics: global views of biology. *Science* **274**, 536-539 (1996).

15 Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516-1517 (1996).

16 Collins, F. S., Guyer, M. S. & Charkravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580-1581 (1997).

17 Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861, doi:nature06258 [pii]
10.1038/nature06258 (2007).

18 A haplotype map of the human genome. *Nature* **437**, 1299-1320, doi:10.1038/nature04226 (2005).

19 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678, doi:nature05911 [pii]
10.1038/nature05911 (2007).

20 Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881-888, doi:322/5903/881 [pii]
10.1126/science.1156409 (2008).

21 Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nat Genet* **36**, 949-951, doi:10.1038/ng1416
ng1416 [pii] (2004).

22 Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol* **5**, e254, doi:07-PLBI-RA-1258 [pii]
10.1371/journal.pbio.0050254 (2007).
23 Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59, doi:nature07517 [pii]
10.1038/nature07517 (2008).
24 Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872-876, doi:nature06884 [pii]
10.1038/nature06884 (2008).
25 Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60-65, doi:nature07484 [pii]
10.1038/nature07484 (2008).
26 Ahn, S. M. *et al.* The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group. *Genome Res*, doi:gr.092197.109 [pii]
10.1101/gr.092197.109 (2009).
27 Kim, J. I. *et al.* A highly annotated whole-genome sequence of a Korean individual. *Nature*, doi:nature08211 [pii]
10.1038/nature08211 (2009).
28 Drmanac, R. *et al.* Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science*, doi:1181498 [pii]
10.1126/science.1181498 (2009).
29 McKernan, K. J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two base encoding. *Genome Res*, doi:gr.091868.109 [pii]
10.1101/gr.091868.109 (2009).
30 Pushkarev, D., Neff, N. F. & Quake, S. R. Single-molecule sequencing of an individual human genome. *Nat Biotechnol*, doi:nbt.1561 [pii]
10.1038/nbt.1561 (2009).
31 Amberger, J., Bocchini, C. A., Scott, A. F. & Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res* **37**, D793-796, doi:gkn665 [pii]
10.1093/nar/gkn665 (2009).
32 McKusick, V. A. Mendelian Inheritance in Man and its online version, OMIM. *Am J Hum Genet* **80**, 588-604, doi:S0002-9297(07)61121-5 [pii]
10.1086/514346 (2007).
33 Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351, doi:10.1126/science.1058040
291/5507/1304 [pii] (2001).
34 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).
35 Antonarakis, S. E. & McKusick, V. A. OMIM passes the 1,000-disease-gene mark. *Nat Genet* **25**, 11, doi:10.1038/75497 (2000).
36 Peltonen, L. & McKusick, V. A. Genomics and medicine. Dissecting human disease in the postgenomic era. *Science* **291**, 1224-1229 (2001).
37 Siva, N. 1000 Genomes project. *Nat Biotechnol* **26**, 256, doi:nbt0308-256b [pii]
10.1038/nbt0308-256b (2008).

38 Biesecker, L. G. *et al.* The ClinSeq Project: Piloting large-scale genome sequencing for research in genomic medicine. *Genome Res*, doi:gr.092841.109 [pii]
10.1101/gr.092841.109 (2009).
39 Kryukov, G. V., Shpunt, A., Stamatoyannopoulos, J. A. & Sunyaev, S. R. Power of deep, all-exon resequencing for discovery of human trait genes. *Proc Natl Acad Sci U S A* **106**, 3871-3876, doi:0812824106 [pii]
10.1073/pnas.0812824106 (2009).
40 Church, G. M. The personal genome project. *Mol Syst Biol* **1**, 2005 0030, doi:msb4100040 [pii]
10.1038/msb4100040 (2005).
41 Manolio, T. A., Bailey-Wilson, J. E. & Collins, F. S. Genes, environment and the value of prospective cohort studies. *Nat Rev Genet* **7**, 812-820, doi:nrg1919 [pii]
10.1038/nrg1919 (2006).
42 Beutler, E., Felitti, V. J., Koziol, J. A., Ho, N. J. & Gelbart, T. Penetrance of 845G--> A (C282Y) HFE hereditary haemochromatosis mutation in the USA. *Lancet* **359**, 211-218, doi:S0140-6736(02)07447-0 [pii]
10.1016/S0140-6736(02)07447-0 (2002).
43 Snyder, M., Weissman, S. & Gerstein, M. Personal phenotypes to go with personal genomes. *Mol Syst Biol* **5**, 273, doi:msb200932 [pii]
10.1038/msb.2009.32 (2009).
44 Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380, doi:nature03959 [pii]
10.1038/nature03959 (2005).
45 Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728-1732, doi:1117389 [pii]
10.1126/science.1117389 (2005).
46 Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat Biotechnol* **26**, 1135-1145, doi:nbt1486 [pii]
10.1038/nbt1486 (2008).
47 Diehl, F. *et al.* BEAMing: single-molecule PCR on microparticles in water-in-oil emulsions. *Nat Methods* **3**, 551-559, doi:nmeth898 [pii]
10.1038/nmeth898 (2006).
48 Dressman, D., Yan, H., Traverso, G., Kinzler, K. W. & Vogelstein, B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci U S A* **100**, 8817-8822, doi:10.1073/pnas.1133470100
1133470100 [pii] (2003).
49 Adessi, C. *et al.* Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res* **28**, E87 (2000).
50 Shapero, M. H., Leuther, K. K., Nguyen, A., Scott, M. & Jones, K. W. SNP genotyping by multiplexed solid-phase amplification and fluorescent minisequencing. *Genome Res* **11**, 1926-1934, doi:10.1101/gr.205001 (2001).
51 Fedurco, M., Romieu, A., Williams, S., Lawrence, I. & Turcatti, G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res* **34**, e22, doi:34/3/e22 [pii]
10.1093/nar/gnj023 (2006).
52 Lizardi, P. M. *et al.* Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nat Genet* **19**, 225-232, doi:10.1038/898 (1998).

53 Blanco, L. *et al.* Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *J Biol Chem* **264**, 8935-8940 (1989).
54 Baner, J., Nilsson, M., Mendel-Hartvig, M. & Landegren, U. Signal amplification of padlock probes by rolling circle replication. *Nucleic Acids Res* **26**, 5073-5078, doi:gkb813 [pii] (1998).
55 Goransson, J. *et al.* A single molecule array for digital targeted molecular analyses. *Nucleic Acids Res* **37**, e7, doi:gkn921 [pii]
10.1093/nar/gkn921 (2009).
56 Larsson, C. *et al.* In situ genotyping individual DNA molecules by target-primed rolling-circle amplification of padlock probes. *Nat Methods* **1**, 227-232, doi:nmeth723 [pii]
10.1038/nmeth723 (2004).
57 Melin, J. *et al.* Thermoplastic microfluidic platform for single-molecule detection, cell culture, and actuation. *Anal Chem* **77**, 7122-7130, doi:10.1021/ac050916u (2005).
58 Harris, T. D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320**, 106-109, doi:320/5872/106 [pii]
10.1126/science.1150427 (2008).
59 Sjoblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268-274, doi:1133427 [pii]
10.1126/science.1133427 (2006).
60 Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108-1113, doi:1145720 [pii]
10.1126/science.1145720 (2007).
61 Okou, D. T. *et al.* Microarray-based genomic selection for high-throughput resequencing. *Nat Methods* **4**, 907-909, doi:nmeth1109 [pii]
10.1038/nmeth1109 (2007).
62 Albert, T. J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nat Methods* **4**, 903-905, doi:nmeth1111 [pii]
10.1038/nmeth1111 (2007).
63 Hodges, E. *et al.* Genome-wide in situ exon capture for selective resequencing. *Nat Genet* **39**, 1522-1527, doi:ng.2007.42 [pii]
10.1038/ng.2007.42 (2007).
64 Bau, S. *et al.* Targeted next-generation sequencing by specific capture of multiple genomic loci using low-volume microfluidic DNA arrays. *Anal Bioanal Chem* **393**, 171-175, doi:10.1007/s00216-008-2460-7 (2009).
65 Hodges, E. *et al.* Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat Protoc* **4**, 960-974, doi:nprot.2009.68 [pii]
10.1038/nprot.2009.68 (2009).
66 Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, doi:nature08250 [pii]
10.1038/nature08250 (2009).
67 Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet*, doi:ng.499 [pii]
10.1038/ng.499 (2009).
68 Herman, D. S. *et al.* Filter-based hybridization capture of subgenomes enables resequencing and copy-number detection. *Nat Methods* **6**, 507-510, doi:nmeth.1343 [pii]
10.1038/nmeth.1343 (2009).

69 Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* **27**, 182-189, doi:nbt.1523 [pii]
10.1038/nbt.1523 (2009).
70 Tewhey, R. *et al.* Enrichment of sequencing targets from the human genome by solution hybridization. *Genome Biol* **10**, R116, doi:gb-2009-10-10-r116 [pii]
10.1186/gb-2009-10-10-r116 (2009).
71 Levin, J. Z. *et al.* Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol* **10**, R115, doi:gb-2009-10-10-r115 [pii]
10.1186/gb-2009-10-10-r115 (2009).
72 Markoulatos, P., Siafakas, N. & Moncany, M. Multiplex polymerase chain reaction: a practical approach. *J Clin Lab Anal* **16**, 47-51, doi:10.1002/jcla.2058 [pii] (2002).
73 Wang, H. Y. *et al.* A genotyping system capable of simultaneously analyzing >1000 single nucleotide polymorphisms in a haploid genome. *Genome Res* **15**, 276-283, doi:15/2/276 [pii]
10.1101/gr.2885205 (2005).
74 Hardenbol, P. *et al.* Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat Biotechnol* **21**, 673-678, doi:10.1038/nbt821
nbt821 [pii] (2003).
75 Hardenbol, P. *et al.* Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay. *Genome Res* **15**, 269-275, doi:15/2/269 [pii]
10.1101/gr.3185605 (2005).
76 Wang, Y. *et al.* Analysis of molecular inversion probe performance for allele copy number determination. *Genome Biol* **8**, R246, doi:gb-2007-8-11-r246 [pii]
10.1186/gb-2007-8-11-r246 (2007).
77 Porreca, G. J. *et al.* Multiplex amplification of large sets of human exons. *Nat Methods* **4**, 931-936, doi:nmeth1110 [pii]
10.1038/nmeth1110 (2007).
78 Krishnakumar, S. *et al.* A comprehensive assay for targeted multiplex amplification of human DNA sequences. *Proc Natl Acad Sci U S A* **105**, 9296-9301, doi:0803240105 [pii]
10.1073/pnas.0803240105 (2008).
79 Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat Methods* **6**, 315-316, doi:nmeth.f.248 [pii]
10.1038/nmeth.f.248 (2009).
80 Li, J. B. *et al.* Multiplex padlock targeted sequencing reveals human hypermutable CpG variations. *Genome Res* **19**, 1606-1615, doi:gr.092213.109 [pii]
10.1101/gr.092213.109 (2009).
81 Tewhey, R. *et al.* Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nat Biotechnol* **27**, 1025-1031, doi:nbt.1583 [pii]
10.1038/nbt.1583 (2009).
82 Kirkness, E. F. Targeted sequencing with microfluidics. *Nat Biotechnol* **27**, 998-999, doi:nbt1109-998 [pii]
10.1038/nbt1109-998 (2009).
83 Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* **106**, 19096-19101, doi:0910672106 [pii]
10.1073/pnas.0910672106 (2009).

84 Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66-72, doi:nature07485 [pii]
10.1038/nature07485 (2008).
85 Mardis, E. R. *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med* **361**, 1058-1066, doi:NEJMoa0903840 [pii]
10.1056/NEJMoa0903840 (2009).
86 Carter, H. *et al.* Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res* **69**, 6660-6667, doi:0008-5472.CAN-09-1133 [pii]
10.1158/0008-5472.CAN-09-1133 (2009).
87 Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-3814 (2003).
88 Sunyaev, S. *et al.* Prediction of deleterious human alleles. *Hum Mol Genet* **10**, 591-597 (2001).
89 Wolf, S. M. *et al.* Managing incidental findings in human subjects research: analysis and recommendations. *J Law Med Ethics* **36**, 219-248, 211, doi:JLME266 [pii]
10.1111/j.1748-720X.2008.00266.x (2008).
90 Ali-Khan, S. E., Daar, A. S., Shuman, C., Ray, P. N. & Scherer, S. W. Whole genome scanning: resolving clinical diagnosis and management amidst complex data. *Pediatr Res* **66**, 7, doi:10.1203/PDR.0b013e3181b0cbd8 (2009).
91 Kohane, I. S., Masys, D. R. & Altman, R. B. The incidentalome: a threat to genomic medicine. *JAMA* **296**, 212-215, doi:296/2/212 [pii]
10.1001/jama.296.2.212 (2006).
92 Khoury, M. J. & Wagener, D. K. Epidemiological evaluation of the use of genetics to improve the predictive value of disease risk factors. *Am J Hum Genet* **56**, 835-844 (1995).
93 Gusella, J. & MacDonald, M. No post-genetics era in human disease research. *Nat Rev Genet* **3**, 72-79, doi:10.1038/nrg706
nrg706 [pii] (2002).
94 Khoury, M. J., Little, J., Higgins, J., Ioannidis, J. P. & Gwinn, M. Reporting of systematic reviews: the challenge of genetic association studies. *PLoS Med* **4**, e211, doi:07-PLME-C-0409 [pii]
10.1371/journal.pmed.0040211 (2007).
95 Rockhill, B., Kawachi, I. & Colditz, G. A. Individual risk prediction and population-wide disease prevention. *Epidemiol Rev* **22**, 176-180 (2000).
96 Schaumberg, D. A., Hankinson, S. E., Guo, Q., Rimm, E. & Hunter, D. J. A prospective study of 2 major age-related macular degeneration susceptibility alleles and interactions with modifiable risk factors. *Arch Ophthalmol* **125**, 55-62, doi:125/1/55 [pii]
10.1001/archopht.125.1.55 (2007).
97 Maller, J. *et al.* Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration. *Nat Genet* **38**, 1055-1059, doi:ng1873 [pii]
10.1038/ng1873 (2006).
98 Pagon, R. A. GeneTests: an online genetic information resource for health care providers. *J Med Libr Assoc* **94**, 343-348 (2006).
99 Yu, W., Clyne, M., Khoury, M. J. & Gwinn, M. Phenopedia and Genopedia: Disease-centered and Gene-centered Views of the Evolving Knowledge of Human Genetic Associations. *Bioinformatics*, doi:btp618 [pii]

10.1093/bioinformatics/btp618 (2009).
100 Burke, W. & Emery, J. Genetics education for primary-care providers. *Nat Rev Genet* **3**, 561-566, doi:10.1038/nrg845
nrg845 [pii] (2002).
101 Strom, C. M. Mutation detection, interpretation, and applications in the clinical laboratory setting. *Mutat Res* **573**, 160-167, doi:S0027-5107(05)00034-5 [pii]
10.1016/j.mrfmmm.2004.09.017 (2005).
102 Steinberg, K. K. *et al.* Prevalence of C282Y and H63D mutations in the hemochromatosis (HFE) gene in the United States. *JAMA* **285**, 2216-2222, doi:joc02264 [pii] (2001).
103 Burt, M. J. *et al.* The significance of haemochromatosis gene mutations in the general population: implications for screening. *Gut* **43**, 830-836 (1998).
104 Guttormsen, B. N. *et al.* Rationale for targeted rather than population based screening with C-reactive protein using the National Health and Nutrition Examination Survey (1999 to 2002). *Am J Cardiol* **100**, 1130-1133, doi:S0002-9149(07)01258-1 [pii]
10.1016/j.amjcard.2007.05.037 (2007).
105 Ekstein, J. & Katzenstein, H. The Dor Yeshorim story: community-based carrier screening for Tay-Sachs disease. *Adv Genet* **44**, 297-310 (2001).
106 Burke, W. & Psaty, B. M. Personalized medicine in the era of genomics. *JAMA* **298**, 1682-1684, doi:298/14/1682 [pii]
10.1001/jama.298.14.1682 (2007).
107 National Institutes of Health Consensus Development Conference Statement: phenylketonuria: screening and management, October 16-18, 2000. *Pediatrics* **108**, 972-982 (2001).
108 Genetic risk assessment and BRCA mutation testing for breast and ovarian cancer susceptibility: recommendation statement. *Ann Intern Med* **143**, 355-361, doi:143/5/355 [pii] (2005).
109 Marteau, T. M. & Weinman, J. Self-regulation and the behavioural response to DNA risk information: a theoretical analysis and framework for future research. *Soc Sci Med* **62**, 1360-1368, doi:S0277-9536(05)00422-3 [pii]
10.1016/j.socscimed.2005.08.005 (2006).
110 Green, R. C. *et al.* Disclosure of APOE genotype for risk of Alzheimer's disease. *N Engl J Med* **361**, 245-254, doi:361/3/245 [pii]
10.1056/NEJMoa0809578 (2009).
111 Kane, R. A. & Kane, R. L. Effect of genetic testing for risk of Alzheimer's disease. *N Engl J Med* **361**, 298-299, doi:361/3/298 [pii]
10.1056/NEJMe0903449 (2009).
112 Robertson, J. A. The $1000 genome: ethical and legal issues in whole genome sequencing of individuals. *Am J Bioeth* **3**, W-IF1, doi:10.1162/152651603322874762 (2003).
113 Klein, T. E. *et al.* Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. *Pharmacogenomics J* **1**, 167-170 (2001).
114 Guessous, I., Gwinn, M. & Khoury, M. J. Genome-wide association studies in pharmacogenomics: untapped potential for translation. *Genome Med* **1**, 46, doi:gm46 [pii]
10.1186/gm46 (2009).
115 Ball, M. P. *et al.* Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotechnol* **27**, 361-368, doi:nbt.1533 [pii]
10.1038/nbt.1533 (2009).

116 Li, J. B. *et al.* Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324**, 1210-1213, doi:324/5931/1210 [pii]
10.1126/science.1170995 (2009).
117 Gottlieb, B. *et al.* BAK1 gene variation and abdominal aortic aneurysms. *Hum Mutat* **30**, 1043-1047, doi:10.1002/humu.21046 (2009).
118 Lee, J. H. *et al.* A robust approach to identifying tissue-specific gene expression regulatory variants using personalized human induced pluripotent stem cells. *PLoS Genet* **5**, e1000718, doi:10.1371/journal.pgen.1000718 (2009).
119 Zhang, K. *et al.* Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat Methods* **6**, 613-618, doi:nmeth.1357 [pii]
10.1038/nmeth.1357 (2009).
120 Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868-874, doi:nature08625 [pii]
10.1038/nature08625 (2009).
121 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:326/5950/289 [pii]
10.1126/science.1181369 (2009).

Chapter 2

Clinical Analysis of Individual Genomes

This chapter was submitted to Nature for publication and is currently under revision.

Abraham M. Rosenbaum[1]*, Joseph Thakuria[1,2]*, Xiaodi Wu[1]*, Alexander Wait Zaranek[1]*, Gerard Berry[3], Marsha F. Browning[2], Matthew J. Callow[4], Michael F. Chou[1], Wendy K. Chung[5], Gerald Cox[6], Shawn Douglas[1], Peter Hulick[7], Jong-Il Kim[8], Jin Billy Li[1], Michael Murray[9], Geoffrey B. Nilsen[4], Hugh Y. Rienhoff[10], John Aach[1], Radoje Drmanac[4], Stephen R. Quake[11], Christine E. Seidman[12], Jeong-Sun Seo[8], Kun Zhang[13], Heidi L. Rehm[14], George M. Church[1]

[1]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA; [2]Division of Clinical and Biochemical Genetics, Massachusetts General Hospital, Boston, MA 02114, USA; [3]Division of Clinical and Biochemical Genetics, Childrens Hospital Boston, Boston, MA 02115, USA; [4]Complete Genomics, Inc., Mountain View, CA 94043, USA; [5]Departments of Pediatrics and Medicine, Columbia University, New York, NY, 10032, USA; [6]Genzyme Corporation, Cambridge, MA 02142, USA; [7]Division of Genetics, North Shore University Health System, Evanston, IL 60201, USA; [8]Genomic Medicine Institute, Seoul National University, College of Medicine, Seoul, 110-799, Korea; [9]Division of Genetics, Brigham and Women's Hospital, Boston, MA 02115, USA; [10]Laguna Honda Hospital; San Francisco, CA 94116, USA; [11]Dept of Bioengineering, Stanford University and Howard Hughes Medical Institute, Stanford CA 94305, USA; [12]Department of Genetics, Harvard Medical School and Howard Hughes Medical Institute, Boston, MA 02115, USA; [13]Department of Bioengineering, University of California, San Diego, CA 92093, USA; and [14]Department of Pathology, Harvard Medical School, Boston, MA, 02115, USA
*These authors contributed equally

Author Contributions X.W., J.T. and G.M.C. conceived and developed Trait-o-matic with help from S.D. and A.W.Z.; A.M.R., A.W.Z., J.T. and G.M.C. conceived and developed the systematic manual analysis with help from G.B., M.F.B., G.C., P.H., M.M. and H.Y.R.; A.M.R. performed all analyses with help from A.W.Z., J.T., and M.F.C; A.M.R., K.Z. and A.W.Z. produced the PGP9.3 data release with help from J.B.L., and J.A; Ala13Thr workup was performed by J.T. with help from C.E.S., H.L.R. and W.K.C.; C.E.S. performed medical examination of PGP6; NA07022 genome and advice was provided by R.D., G.B.N. and M.J.C.; AK genome and advice was provided by J.S.S. and J.I.K; P0 genome and advice was provided by S.R.Q.; A.M.R., J.T., A.W.Z., X.W. and M.F.C. wrote the manuscript; G.M.C. supervised all aspects of the study.

Acknowledgements We are grateful for the help and advice provided by all the member of the Church Laboratory, Jason Bobe, Jeantine Lunshof and other members of the Personal Genome Project Community, Ting Wu and other members of PGEd, and the computational support and help provided by Scalable Computing Experts. We thank NHGRI, NHLBI and Personalgenomes.org for funding support.

Abstract

Whole human genome sequencing will help enable comprehensive clinical genetic diagnosis and intervention as we increase our understanding of human biology in both normal and pathogenic states. As the field rapidly advances, we see a need for a community-updated algorithm for prioritization of findings in genome sequence data. We focus on the potentially most predictable and actionable traits using an automated component called Trait-o-matic and curation to further refine output. Applying this approach prioritizes 11 variants associated with medically-relevant gross phenotypes for confirmation and possible clinical action from three million variants in each of nine presumed-healthy individuals with publicly available, nearly complete human genome sequences and 60,000 variants across 16 partial genomes. We confirm and provide detailed clinical workup of a cardiac mutation in a Personal Genome Project participant who is currently asymptomatic. Additionally, we report four common pathogenic recessive variants, 98 rare recessive variants presumed pathogenic, 16 quantitative trait loci with large risks for disease, and 32 pharmacogenetic-related recommendations, one of which is recommended by the Food and Drug Administration (FDA). Furthermore, we find 24 variants associated with disease in existing databases, but either reclassified as benign due to more recent studies or likely benign due to their high frequencies in these genomes. Finally, we find significantly fewer deleterious variants (indels, frameshifts, and rare missense/nonsense mutations) in genes listed in Online Mendelian Inheritance in Man than all other genes sequenced. This supports the prioritization of such genes in searching for novel variants of clinical utility.

Introduction

With the publication of the whole genome sequence of the first Personal Genome Project (PGP[1]) volunteer in this letter, there are now nine whole genomes in the public domain, and it is appropriate to begin to address the similarities and differences that can contribute to clinical utility for normal individuals. Early whole genome sequencing efforts have taken different approaches to making medically-relevant phenotypic inferences from the variants uncovered. Some opt for no discussion[2-4]; others make note of common, functionally neutral, or heterozygous recessive variants found in databases such as Online Mendelian Inheritance in Man[5] (OMIM) and/or the Human Gene Mutation Database[6] (HGMD)[7-10]; and still others generate their own OMIM-derived database for analysis.[11] Several projects have been announced to sequence large numbers of individuals using next-generation sequencing technologies[12-14] with different charters for choosing variants for confirmatory sequencing and phenotypic analysis. A growing movement exists to deliver serious, predictable and actionable findings to research subjects[15], and another trend has emerged to deliver even alleles with limited impact via "direct-to-consumer" companies. Very few physicians are trained in adult genetics, and even the most trained find it challenging to integrate familial disease segregation, incomplete penetrance, variable expressivity and multigenic and multi-allelic, interactions.

Here we present a systematic method for analyzing human genome variants called Trait-o-matic which synthesizes available data from major databases. Khoury et al.[16] describe the potential for disease prevention inherent in personal genomics; our approach helps prioritize variants towards this goal while complementing genomic browsers such

as HuGE Navigator[17]. We have used our tool to analyze variants of clinical relevance in over 25 genomes (one of these analyses was recently published[18]), and here we present analysis from the eight publicly available genomes, the PGP 9.3 data release comprising exonic material from ten PGP volunteers[19], and eight public HapMap partial genomes from Ng et al.[20]. Additionally, we present here the first full genome sequence from the PGP.

Trait-o-matic has a web-based interface and an extensible database that, in part, utilizes data parsed from the OMIM, HGMD, SNPedia[21], and PharmGKB[22] databases. Since ~90% of well-defined, highly penetrant genetic diseases are the result of either point mutations or deletions occurring in coding regions (HGMD 2009.2), we have initially focused on point mutations and a few relatively common small insertions and deletions already used in clinical testing. We expected to find few, if any, variants for highly-penetrant genetics diseases due to the presumed good health of these individuals. Indeed, we identify only eleven variants expected to cause clinically-relevant gross phenotypic changes in our 25 individuals that would require confirmation. Of these, only one was reported late-onset and had both sufficient reports of it segregating with disease within families and functional studies to be a cause for concern, and we describe our clinical followup for this variant.

Methods Summary

We generated lists of variants for the entire genome of PGP1 and for the partial genomes of the first ten PGP participants. In addition, we downloaded variants from publicly available genomes (see table S2-1); we focused our comprehensive analysis on

substitutions because these were available for each genome. To identify those with clinical utility, we first matched these lists to over 1,500 features extracted from each of SNPedia (June 2009) and PharmGKB (April 2009). We then converted every non-synonymous substitution to its amino acid change incorporating known splice variants; these were matched to our database comprising over 11,000 nonsense and missense substitutions from OMIM (June 2009) extracted through custom scripts, and 44,776 from HGMD Professional 7.1 (March 2007). We then computed the trait-associated allele frequency (TAF) for every match using HapMap Project data for each population and prioritized rare ones for further analysis. All code for processing these data was consolidated into a publicly available repository (see the "Trait-o-matic" website). Finally, using the GeneTests database, we noted whether clinical sequencing was available for the gene in which the substitution was identified.[23] Trait-o-matic–generated descriptors were used to manually filter for both disease-related phenotypes and likely expression, and Trait-o-matic generated literature citations were used to manually generate a final list based upon clinical evidence and actionability. Additionally, to enable novel phenotype–genotype correlations, we used Trait-o-matic to identify and rank by predicted deleterious effect every non-synonymous substitution found in an OMIM Morbid Map gene (the cytogenic location corresponding to the disease listed in OMIM, see supplemental methods for additional detail).

Results

To reveal actionable clinical data, we would ideally have case and control statistics for all clinically relevant variations. This database would also quantify the

effects of allelic combinations,[24] list disease modifiers where appropriate, and provide hypotheses tailored to each individual's phenotype. As genome sequence data becomes more complete, this database would include all types of variants, including phase information, indels, mitochondrial variants, structural variants, methylation status and other types of genetic information not routinely assayed in a high-throughput manner. Because such a database does not yet exist, we rely upon four existing databases and adapt our analysis to each one to take advantage of its strengths in prioritizing clinical variants found within it. Our approach uses HGMD and OMIM to identify monogenic, highly penetrant mutations, SNPedia to identify clinically important quantitative-trait loci and poorly penetrant monogenic mutations discovered through genome-wide association studies (table S2-7), and PharmGKB for pharmacogenetically relevant variants (table S2-8). By using our approach, we cover drug dosage, clinical assessment and discovery—the major areas relevant to an ideal, individualized genome interpretation, summarized in Figure 2-1a.

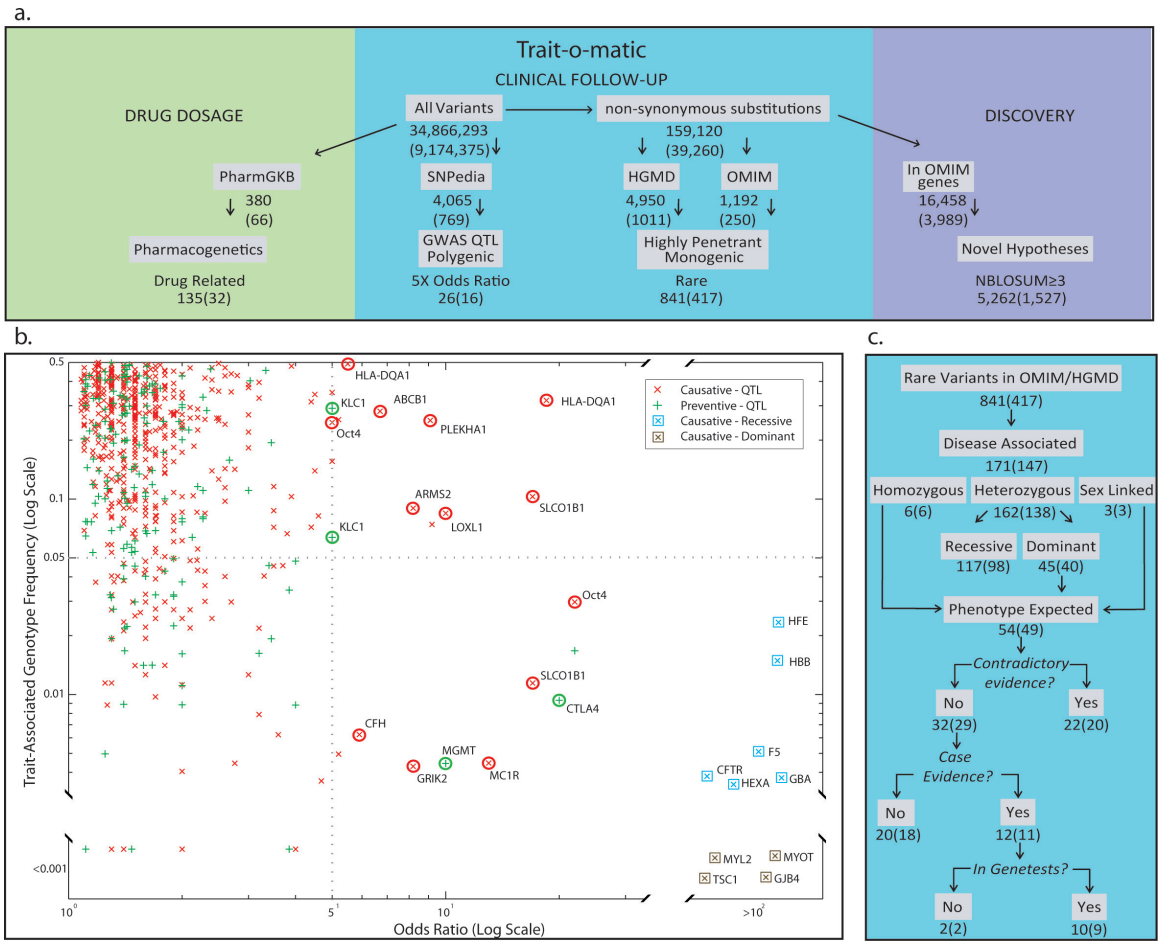


Figure 2-1. Approach and Results (a) Automated Analysis by Trait-o-matic. Variants are automatically filtered for relevance to pharmacogenetics, high odds-ratio quantitative trait loci (QTL), highly penetrant monogenic disease, and novel hypotheses generation. The number of variants analyzed across all 25 genomes are shown, along with the number of unique variants in parentheses. **(b) Relationship between trait-associated genotype frequency and odds ratio for the associated phenotype.** The frequency for each QTL is the weighted average frequency of CEU, JPT, CHB and YRI from the HapMap populations. The frequencies for the affected genotypes caused by recessive and dominant variants are calculated based upon the population with the highest frequency for that variant. All QTL variants are from the SNPedia database, and recessive variants are chosen from those that are routinely used for clinical sequencing with a causative allele frequency ≥0.05 (table S2-6). The dominant variants are the ones uncovered in this study with an OR≥100. The vertical line demarcates an odds ratio of 5x; all variants greater are circled or boxed and labeled with the gene name. As research in this field advances, this cutoff will be replaced with more complex functions of multiple alleles, environmental risk factors, availability of other clinical phenotypes or hypotheses and the actionability of the particular variant as well as consequences of action with false positives. **(c) Manual Analysis.** Evaluations of rare monogenic substitutions are made as shown to identify clinically actionable mutations.

The HGMD and OMIM databases report variants according to amino acid position and change; to match against these we first convert each of our substitutions to its corresponding amino acid change, accounting for potential splice variants. Our algorithm for converting chromosomal coordinates to amino acid positions has identified 19% more amino acid changes than previously reported in the Asian diploid genome[10] (see supplementary methods). Furthermore, our parsing of the OMIM database compares favorably with that of McKernan et al.[11]; we found 13 variants not revealed with their algorithm and successfully re-identified 45 of 67 variants on their list. 13 of the remaining 22 variants are, however, found on our HGMD-generated list. Overall, we identified an average of 266 HGMD and 64 OMIM non-synonymous substitutions per genome (table S2-2).

These variants contain a large number of non-clinical and benign polymorphisms, and the descriptors often do not contain sufficient information for prioritization. To test the efficiency of HapMap-based frequency filtering in enriching for variants with high correlation with disease, we sampled >700 variants from the SNPedia database with HapMap genotype frequencies and defined odds-ratios (OR) for disease. We found that a 5% frequency filter retained 16.2% of all variants and ~39% of high disease correlation variants (OR≥5x), for a >2.4x enrichment (p=0.0062) (fig. 2-1b). We expect this enrichment to be even larger in HGMD and OMIM where the expected ORs are higher and the variants typically rarer (over 25% of variants from these databases identified in our genomes were either not sampled by HapMap or were reported with a frequency lower than 5% in the Caucasian (CEU) population). To enrich for deleterious variants, we used an allele frequency filter to retain only those variants occurring at a frequency less

than 5% in the HapMap ethnic population that is most closely matched to that of each individual. On average, this step removed 68% of the variants. Since this proxy will miss relevant recessive variants with disease frequency greater than 1/400, we created a clinical utility database of the few relatively common diseases whose causative variants (both indels and substitutions) are more common than 5%, and unconditionally retain matches to this list to all common variants (supplemental table S2-6a, S2-6b). While we have utilized a strict OR cutoff for this study, we envision customizing this threshold based upon individual phenotype, as well as cost–benefit and risk–reward ratios for clinical actionability.

To further curate the 417 rare variants in OMIM and HGMD identified by Trait-o-matic, we systematically analyze them for clinical relevance (figure 2-1c). Remarkably, only 147 were actually disease-related and highly penetrant, with the rest being poorly penetrant variants (identified by the terms "susceptibility""association" and "polymorphism"), rare olfactory receptors, blood types, pigment variations and other non-clinical variants. We surveyed the primary OMIM citation or sole HGMD citation for each of these variants: 49 would appear to predict phenotypes, either through homozygosity, sex linkage, compound heterozygosity, or a dominant form of expression. Subsequent published reports ("contradictory evidence"), however, reclassified 20 of these variants as likely benign (table S2-5). Absent medical hypotheses indicating candidate genes, where functional variants would also be considered, the remaining 29 variants (table S2-4) were analyzed for sufficient case-based evidence to warrant follow-up: we required that the variants be reported in either (a) two unrelated symptomatic individuals, or (b) three symptomatic individuals in a single family. Four such variants,

however, were found in two or more individuals of Yoruba descent, leading us to conjecture that they may not be disease-causing (table S2-10a). Using our systematic approach, Table 2-1 lists eleven remaining variants, which, if sequence verified to be present in the individual, should produce a gross-phenotypic effect in the individual.. Table S2-3 shows the effect of applying the above filters to each genome analyzed.. Other lists of potential clinical interest (the final list of 29 potential clinical variants (S2-4), 98 recessive variants (supplemental), APOE status (S2-9) and common disease-causing variants (S2-6)) are included in the supplemental data. The age and health of these individuals make it likely that the variants listed in Table 2-1 are either sequencing errors, poorly penetrant or benign polymorphisms mistakenly associated with disease. Of the three late-onset diseases in identified individuals, only two of them were viewed as being extremely critical (hypertrophic cardiomyopathy due to mutations in *MYL2* and *GLA*). Lack of access to DNA from P0 prevented us from confirming this variant, but we informed him of this finding. Our close collaboration with PGP6 allowed us to perform a more thorough follow-up for this variant.

Table 1. Variants with Sufficient Evidence to Warrant Sequence Confirmation and Clinical Follow-up. After analyzing 25 genomes with Trait-o-matic we obtained 417 rare monogenic variants; subsequent systematic prioritization, summarized in figure 1c, reveals eleven variants for clinical followup. Genes available for clinical diagnostics ("In Genetests") are identified; these variants are more likely to be actionable. Further detail, including frequency data and all sources are provided in Table S2-4.

Genome/ Gender	Disease/ In Genetests?	Variant/ Alteration/ rsID	State/ Inheritance Case/Control	Recommendation
PGP6 Male	Hypertrophic cardiomyopathy Yes	Chr12:109841347 *MYL2*, A13T	Het Dominant 3; 0/339	Found in three Caucasian individuals (in two families) in a late-onset form; recommend follow-up by cardiologist.
JCV/ NA07022 Male Male	Thin membrane basement disease Yes	Chr2:227624091 *COL4A4*, G999E rs13027659	Het Dominant 4; 0/50	Be advised of microscopic hematuria, renal function should be followed by primary care physician. While this condition is benign, it is important when considering Alport Syndrome.
JW Male	Myotilinopathy Yes	Chr5: 137234459 *MYOT*, Q74K rs6890689	Het Dominant 2; 0/100 1 individual had second mutation	Seen in one symptomatic heterozygote and one compound heterozygote. Late-onset progressive distal muscle weakness and peripheral neuropathy with hyporeflexia; physical therapy and assistive devices may help.
JW Male	Idiopathic epilepsy No	Chr16:1196127 *CACNA1H*, A876T	Het Dominant 4; 0/100 Segregates across 3 generations	Various phenotypes have been reported in autosomal dominant epilepsy, but data for this gene is weak and this gene is unavailable for clinical sequencing.
P0 Male	Neuro-hypophyseal diabetes insipidus Yes	chr20:3011659 *AVP*, G96C	Het Dominant 11; 0/50 2 kindred (6 and 5). Segregates across 3 generations	Found segregating with disease in one Polish family. Consider biochemical confirmation and follow-up with endocrinologist.
P0 Male	Fabry disease Yes	chrX:100545469 *GLA*, Q119*	X-Linked 2 families, 1 individual and no controls	Found in one British individual and two families with classical phenotype. Consider biochemical confirmation; there are mutations that correlate with left ventricular hypertrophy in the 6^{th}-8^{th} decade.
YH Male	Tuberous sclerosis complex Yes	Chr9: 134770826 *TSC1*, Q654E	Het Dominant 5; 0/100 4 familial 1 sporadic	Known to have variable expressivity and should be evaluated for features associated with TSC; found in one Korean, one Japanese with TSC and one newborn with cardiac rhabdomyoma. There is complete penetrance for distinctive skin rash; recommend follow-up with primary care physician.
SJK Male	Nonsyndromic deafness Yes	Chr1:34999551 *GJB4*, V37M	Het Dominant 2; 0/120	Reported in two individuals with hearing loss. Clinical sequencing for this gene is available only for its more frequent association with erythrokeratodermia variabilis.
NA18507 Male	Polycystic kidney disease Yes	Chr16: 2092866 *PKD1*, E2966D rs13337123	Het Dominant 3; no controls Found in 3 unrelated patients from a cohort of 90	This variant was seen in 3 unrelated individuals. In AD PKD1, 50% progress to end-stage renal failure; early signs include hypertension. Recommend followup with primary care physician.
NA18517 Female	Leber congenital amaurosis IV Yes	Chr17:6269533 *AIPL1,* P376S rs61757484 Chr17:6272486 *AIPL1,* T114I rs8069375	Compound Het Recessive 2; no controls	Found in two individuals; one African American with severe maculopathy. Recommend followup with ophthalmologist.
NA12878 Female	Emery Dreifuss muscular dystrophy No	Chr14:63746504 *SYNE2*, T6211M rs36215895	Het Dominant 5; 384 alleles Found in father + two children and mother + son	This variant presented a variety of phenotypes including general weakness and cardiac hypertrophy. Recommend followup with primary care physician.

We confirmed the sequence of the clinically important *MYL2* Ala13Thr in the sixth PGP volunteer (PGP6) and performed a more comprehensive clinical workup. A literature search and examination of the CardioGenomics Database[25] reveal five reports including two probands[26-28] and functional studies evaluating the Ala13Thr variant[28-30]. Poetter et al[26] present a single case harboring the Ala13Thr variant with no family history data available and report its absence in 189 controls; the phenotype of that affected patient was hypertrophic cardiomyopathy (HCM) that predominantly involved the midventricular region (MVH). Andersen et al[27] and Hougs et al[28] describe a single Danish HCM kindred (one deceased without available genotype). The variant was found to segregate in 2 out of 3 affected living individuals. The individual without the variant was initially thought to be a phenocopy due to concurrent obesity and hypertension and then later suspected to have disease due to another variant identified in the *MYH7* gene, Asn1327Lys. In the Anderson study, the variant was absent from 150 controls and 197 other probands. Functional studies performed by Szczesna et al. using purified protein extracts of MYL2 suggested that the Ala13Thr mutation impacted Ca^{2+} binding but not ATPase activity[28-29]. In summary, the pathogenicity of the Ala13Thr variant is possible but not certain, based upon published literature alone.

Because this gene is found in GeneTests,[23] we contacted all four laboratories in the United States (Harvard-Partners Laboratory for Molecular Medicine (LMM), Correlagen, GeneDx, PGxHealth) offering Clinical Laboratory Improvement Amendments (CLIA)–approved diagnostic sequencing of *MYL2* for cardiomyopathy to check for relevant unpublished data. Only two had observed this variant. Correlagen reported finding the variant in one patient with HCM, though no clinical or family history

data was available. The LMM studied a family with two siblings and their father with HCM; the father was not tested and is deceased from cardiomyopathy treated with heart transplant, one affected sibling has the *MYL2* Ala13Thr variant, but the other affected sibling does not have the Ala13Thr variant and instead harbors a *MYBPC3* Glu619Lys variant. Although it is possible that the two separate mutations identified in this family are individually responsible for HCM, the 79-year-old mother's echocardiogram is normal, suggesting there may be an as yet unidentified third mutation that is primarily responsible for disease in this family. It should also be noted that more recent data from the LMM suggest that the Asn1327Lys variant, identified as the second variant in the Anderson/Hougs family, is "likely benign" based upon frequency and lack of segregation in an additional family (unreported results, H. Rehm). Therefore, it is also possible that a third, as yet unidentified, variant may also be primarily responsible for disease in the Anderson/Hougs family. However, at this time, it is not possible to rule out a primary deleterious or contributory role of the Ala13Thr variant.

The earliest age of onset in both families appears to be in the early-forties, while the latest age of onset (presumably in the mother (II:3) of the proband reported by Anderson et al.) is not clear, but the sister (III:5) in the LMM family is unaffected at 56 years of age. PGP6, who is heterozygous for this mutation, is asymptomatic at 54 years of age and has no family history of cardiac disease. His ethnicity is Ashkenazi Jewish (AJ), as are the Anderson/Hougs family (P. Anderson, personal communication) and the LMM family we report here. To test whether this variant is a polymorphism in the probands' populations we screened an AJ DNA panel and did not detect it in 116 controls. Confirmatory testing was performed at the LMM, where he was also confirmed negative

for the Asn1327Lys (*MYH7)* and Glu619Lys (*MYBPC3)* mutations that were also identified in these pedigrees. We informed individual PGP6 of these findings, reviewed the literature with him, and recommended cardiac follow up for a non-invasive, non-urgent, baseline echocardiogram. The echocardiogram was normal, and while we cannot rule out pathogenicity it is likely a rare polymorphism in the AJ population. Nevertheless, PGP6 will be re-evaluated periodically as part of ongoing research. Figure 2-2 shows pedigrees of all known reports of *MYL2* Ala13Thr.

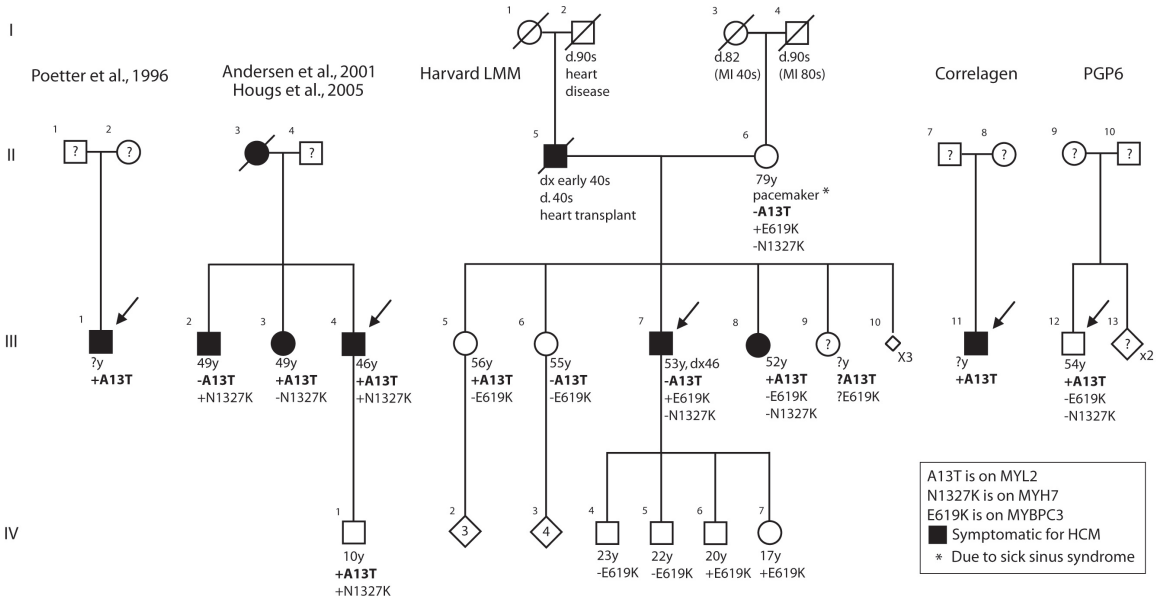


Figure 2-2. MYL2 Ala13Thr Pedigrees. The Poetter, Hougs and Andersen pedigrees are derived from their respective publications. In the LMM pedigree, II.5 is assumed to have harbored the Ala13Thr mutation. III.5 and III.12 (PGP6) are the only asymptomatic carriers for this mutation over the age of fifty.

Although we have focused our analysis on previously identified, clinically relevant variants, we have also begun prioritizing additional variants for discovery. We have found that while OMIM and GeneTests genes comprise 12.6% and 9.3% of exonic base-pairs, respectively, they contain only 10.3% and 7% of potentially deleterious

variants (p=1.36e-104 and p=1.70e-99, respectively; see table S2-12). The relative scarcity of mutations suggests the evolutionary importance of these genes. To further prioritize variants, we have utilized scoring (NBLOSUM) based on the negative of a BLOSUM100 value, which is more computationally efficient than PolyPhen and SIFT with some loss of predictive value (see supplemental data), and have found that each genome contains an average of 50 novel variants NBLOSUM≥3 in OMIM genes.
The utility of this filter can be shown using the variants from the genome of an individual with acute myeloid leukemia[31] where we find two relevant mutations. The eight novel variants described in Ley, et al.[31] are not in OMIM genes, but we do find one germline heterozygous variant, presumed to be deleterious (NBLOSUM=4), in a gene previously implicated in acute promyelocytic leukemia (*NUMA1* E1334G) and one heterozygous (NBLOSUM=5) in a gene implicated in T-cell acute lymphocytic leukemia (*RAP1GDS1* D119V). Other approaches tailored to cancer genes could further refine the algorithm[32].

Our approach prioritizes a manageable number of variants for confirmatory sequencing. False positive rates in existing studies, therefore, do not constitute a significant obstacle for clinical follow-up. Several groups have sequenced the genomes of NA18507, and we take advantage of this by only prioritizing those variants identified by two out of three groups. For non-synonymous substitutions, this consensus is 84%, as shown in Figure 2-3. If we consider only novel non-synonymous substitutions, that number drops to 57% (figure 2-3b and table S2-11). As databases improve, we anticipate an increase in the number of prioritized variants, while we also expect a concurrent decrease in the fraction of those present due to sequencing error. Improvements can also help with high false negative rates. Simply providing a coverage map would clarify the

unexpected absence of non-reference variations, e.g. for ApoE status in table S2-9. Similarly, it would enable the prioritization of clinically significant genotypes that match the reference, such as Factor V Leiden (table S2-6). Overall, despite continual refinement, we expect our approach to continue to produce manageable numbers of variants.

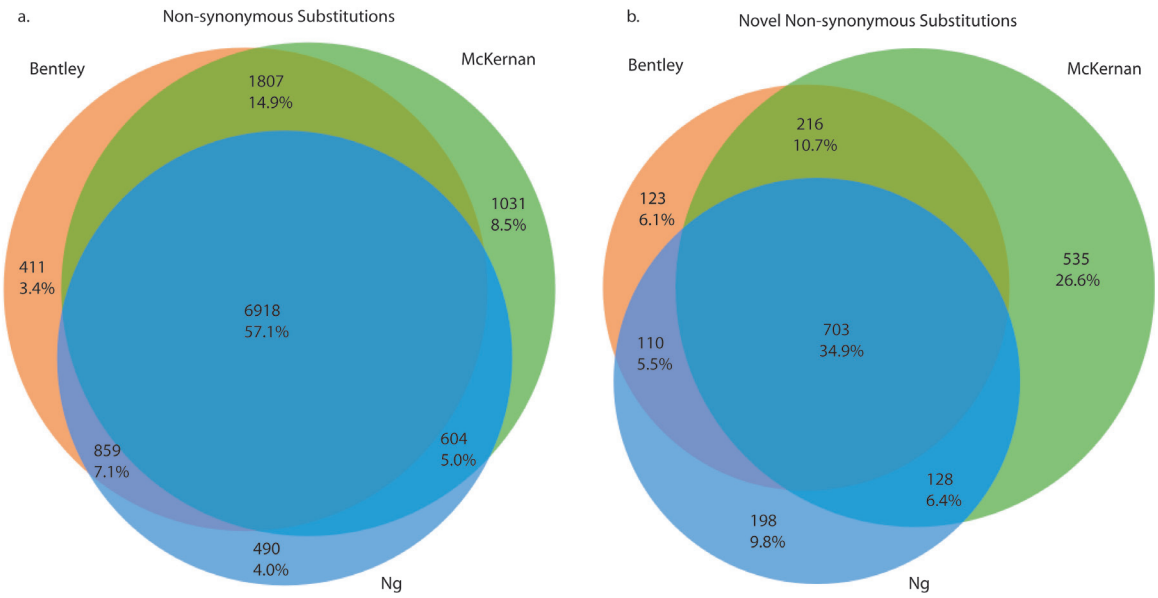


Figure 2-3. Agreement between Substitution Calls for a Single Individual (NA18507) Across Different Studies. (a) shows the agreement between all non-synonymous substitutions, with 84% of calls seen by two of three groups. **(b)** depicts all novel non-synonymous substitutions where only 57% of calls are seen by two of three groups. This suggests, as expected, an increased error rate among rare variants and the need for confirmation of variants identified.

The rapid and accelerating growth of DNA sequencing is enabling the creation of vast personal genome datasets. As we have indicated, there is a great opportunity for improvement and standardization of data and application interfaces. We hope that, to aid this, comprehensive phenotypes—and the release of cell lines to enable functional studies—will become widespread. Here, we provide a tool for scientists, clinicians and

members of the general public to come together to advance medicine via future changes to the Trait-o-matic application on the World Wide Web. Building on the successes and experience of modern molecular genetics, we have examined more than 25 personal genome datasets and demonstrated an approach that can prioritize variants for further clinical follow-up, drug dosage and medical discovery.

Trait-o-matic—and all output used to generate this paper—is available at http://snp.med.harvard.edu

HGMD is a proprietary database available via password protection. Access can be granted via http://www.hgmd.cf.ac.uk/

PharmGKB and SNPedia are available through a non-commercial license.

Various components may be subject to local and/or HIPAA regulation.

No data should be used to draw clinical conclusions without consulting a qualified physician.

Funding support for the AML Sequencing Project was provided by a gift from Alvin J. Siteman, and from grants from the NCI (CA101937), NHGRI (U54HG003079), and the Barnes-Jewish Foundation. The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap through dbGaP accession number phs000159.v1.p1.

References

1 Church, G. M. The personal genome project. *Mol Syst Biol* **1**, 2005 0030, doi:msb4100040 [pii]
10.1038/msb4100040 (2005).

2 Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59, doi:nature07517 [pii]
10.1038/nature07517 (2008).

3 Ahn, S. M. *et al.* The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group. *Genome Res*, doi:gr.092197.109 [pii]
10.1101/gr.092197.109 (2009).

4 Pushkarev, D., Neff, N. F. & Quake, S. R. Single-molecule sequencing of an individual human genome. *Nat Biotechnol*, doi:nbt.1561 [pii]
10.1038/nbt.1561 (2009).

5 Amberger, J., Bocchini, C. A., Scott, A. F. & Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res* **37**, D793-796, doi:gkn665 [pii]
10.1093/nar/gkn665 (2009).

6 Stenson, P. D. *et al.* The Human Gene Mutation Database: 2008 update. *Genome Med* **1**, 13, doi:gm13 [pii]
10.1186/gm13 (2009).

7 Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol* **5**, e254, doi:07-PLBI-RA-1258 [pii]
10.1371/journal.pbio.0050254 (2007).

8 Ng, P. C. *et al.* Genetic variation in an individual human exome. *PLoS Genet* **4**, e1000160, doi:10.1371/journal.pgen.1000160 (2008).

9 Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872-876, doi:nature06884 [pii]
10.1038/nature06884 (2008).

10 Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60-65, doi:nature07484 [pii]
10.1038/nature07484 (2008).

11 McKernan, K. J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two base encoding. *Genome Res*, doi:gr.091868.109 [pii]
10.1101/gr.091868.109 (2009).

12 McGuire, A. 1000 genomes: on the road to personalized medicine. *Personalized Medicine* **5**, 195-197 (2008).

13 Blow, N. Genomics: the personal side of genomics. *Nature* **449**, 627-630, doi:449627a [pii]
10.1038/449627a (2007).

14 Biesecker, L. G. *et al.* The ClinSeq Project: Piloting large-scale genome sequencing for research in genomic medicine. *Genome Res*, doi:gr.092841.109 [pii]
10.1101/gr.092841.109 (2009).

15 Greely, H. T. The uneasy ethical and legal underpinnings of large-scale genomic biobanks. *Annu Rev Genomics Hum Genet* **8**, 343-364, doi:10.1146/annurev.genom.7.080505.115721 (2007).

16 Khoury, M. J. *et al.* The Scientific Foundation for personal genomics: recommendations from a National Institutes of Health-Centers for Disease Control and Prevention multidisciplinary workshop. *Genet Med* **11**, 559-567, doi:10.1097/GIM.0b013e3181b13a6c (2009).

17 Yu, W., Gwinn, M., Clyne, M., Yesupriya, A. & Khoury, M. J. A navigator for human genome epidemiology. *Nat Genet* **40**, 124-125, doi:ng0208-124 [pii]
10.1038/ng0208-124 (2008).

18 Kim, J. I. *et al.* A highly annotated whole-genome sequence of a Korean individual. *Nature*, doi:nature08211 [pii]
10.1038/nature08211 (2009).

19 PGP. Personal Genome Project Data Release 9.3. (2009). <http://www.personalgenomes.org/data/PGP9.3/>.

20 Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, doi:nature08250 [pii]
10.1038/nature08250 (2009).

21 Cariaso, M., Lennon, G. *http://www.SNPedia.com/*, <http://www.SNPedia.com/>

22 Hernandez-Boussard, T. *et al.* The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic Acids Res* **36**, D913-918, doi:gkm1009 [pii]
10.1093/nar/gkm1009 (2008).

23 Pagon, R. A. GeneTests: an online genetic information resource for health care providers. *J Med Libr Assoc* **94**, 343-348 (2006).

24 Kathiresan, S. *et al.* Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet* **41**, 334-341, doi:ng.327 [pii]
10.1038/ng.327 (2009).

25 Genomics of Cardiovascular Development, Adaptation, and Remodeling. NHLBI Program for Genomic Applications, Harvard Medical School. (2009). <http://www.cardiogenomics.org>.

26 Poetter, K. *et al.* Mutations in either the essential or regulatory light chains of myosin are associated with a rare myopathy in human heart and skeletal muscle. *Nat Genet* **13**, 63-69, doi:10.1038/ng0596-63 (1996).

27 Andersen, P. S. *et al.* Myosin light chain mutations in familial hypertrophic cardiomyopathy: phenotypic presentation and frequency in Danish and South African populations. *J Med Genet* **38**, E43 (2001).

28 Hougs, L. *et al.* One third of Danish hypertrophic cardiomyopathy patients with MYH7 mutations have mutations [corrected] in MYH7 rod region. *Eur J Hum Genet* **13**, 161-165, doi:5201310 [pii]
10.1038/sj.ejhg.5201310 (2005).

29 Szczesna, D. *et al.* Familial hypertrophic cardiomyopathy mutations in the regulatory light chains of myosin affect their structure, Ca2+ binding, and phosphorylation. *J Biol Chem* **276**, 7086-7092, doi:10.1074/jbc.M009823200
M009823200 [pii] (2001).

30 Szczesna-Cordary, D., Guzman, G., Ng, S. S. & Zhao, J. Familial hypertrophic cardiomyopathy-linked alterations in Ca2+ binding of human cardiac myosin regulatory light chain affect cardiac muscle contraction. *J Biol Chem* **279**, 3535-3542, doi:10.1074/jbc.M307092200
M307092200 [pii] (2004).

31 Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66-72, doi:nature07485 [pii]
10.1038/nature07485 (2008).

32 Carter, H. *et al.* Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res* **69**, 6660-6667, doi:0008-5472.CAN-09-1133 [pii]
10.1158/0008-5472.CAN-09-1133 (2009).

Supplemental Data 2-1: Supplemental Methods

Construction of the Trait-o-matic Database

To extract a list of OMIM phenotype-allele correlations that is more comprehensive than the 3744 rs IDs mapping to 955 OMIM entries contained in the dbSNP table "OmimVarLocusIdSNP," a Python script capable of extracting non-synonymous single amino acid changes from the OMIM full text was run. This led to the creation of a database table with over 11,000 entries, representing a much larger set of OMIM allelic variants. These data are not strictly a superset of OmimVarLocusIdSNP, however, as certain variants have either incorrect information in the text or are compound mutations.

A second script made use of dbSNP data retrieved from UCSC (snp129) to extract the reference allele for each nsSNP appearing in dbSNP.

Another script was implemented to extract data from SNPedia via the MediaWiki API. Structured markup was parsed for information about genotypes and their corresponding effects, and links to PubMed literature references were isolated from the free text and recorded as accompanying references. Where descriptions of effect specified degree but not condition (i.e. "increased risk" instead of "increased risk for [disease]"), free text was parsed for links preceded by the text "associated with" or "association with"; where found, the accompanying text was appended to the effect description. Where descriptions of effect suggested that a genotype was associated only with an "average," "common," or "normal" phenotype, that particular genotype– phenotype pair was discarded. Output data were then manually edited for spelling and consistency. An accompanying script was used to format the edited data (~1500 entries) for insertion into

a database table, with an optional flag to output only those entries with genotypes homozygous for the reference allele.

Subsequently, a script was implemented to insert HapMap allele frequency data into a database table, summing allele counts for populations of related geographic origin. For example, allele frequencies for East Asians were aggregated from HapMap data for two Chinese populations (CHB, CHD) and one Japanese population (JPT). Aggregation was intended to maximize the HapMap frequency data available for each aggregated region.

Trait-o-matic queries several databases to retrieve information for each variant; this functionality was implemented in a set of scripts written in Python.

A fundamental set of functionality was first constructed as a "utils" library. Portions of this library drew from pre-existing open source code; most notably, a selection of code written in C by W. James Kent, then partially ported to Python as part of the Pennsylvania State University "Galaxy" project, implements functionality to read and write efficiently from a compressed sequence format known as "2bit," by which the entire reference genome can be represented in approximately 700 MB on disk.

Above this foundation, Python scripts were created that query the necessary databases and perform scoring and filtering on the data provided to round out the utility's "core" functionality. These scripts can be invoked via the command line, and are also exposed by a Python script that implements an XML-RPC server, which responds to XML-formatted requests transmitted over HTTP. In addition to core functionality, we implemented a web interface to permit users to view sample data, upload data to the server, and retrieve results in sortable tabular format. This interface passes data to the

core via XML-RPC calls; separation of functionality allows the outward-facing web components to be located separately from the core utility, thus making any future necessity to run many interface or core instances in parallel much easier to implement.

Frequency Filter

The frequency filter bins variants into four frequency categories based upon either the causative allele (when known) or the minor allele (when both variants are potentially "causative," as in hypothetical data, pharmacogenetic and susceptibility data). These four categories are: rare ($f < 0.05$), minor ($0.05 \leq f < 0.5$), major ($f \geq 0.5$), and unknown based upon frequency data from the HapMap Project (release 27) aggregated from populations of related geographic origin to maximize frequency data available (e.g. East Asian is a combination of CHB, CHD and JPT, while Caucasian is a combination of CEU and TSI; see *Construction of the Trait-o-matic Database*). We considered merging the initial 1000 Genome Project data (April 2009) with HapMap project data for this filter, but did not after finding that for the first ~2,500 variants little additional information was provided by the preliminary 1000 Genome Project release data (figure S1). We did, however, supplement the HapMap frequency data with that of the 1000 Genome Project as part of the systematic manual processing.

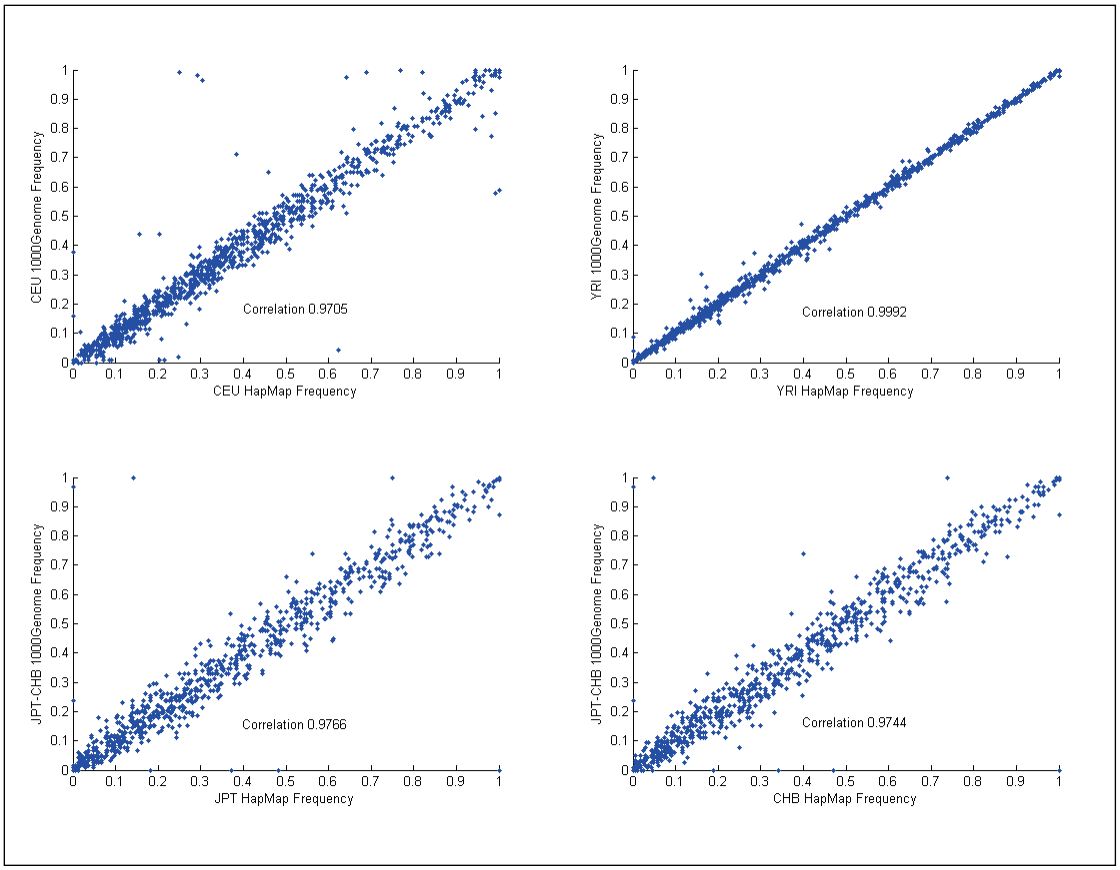


Figure S2-1. Correlation between HapMap and 1000 Genome Project Frequencies for Identified Alleles The Pearson Correlation generated from the trimmed list of variant frequencies is shown for each plot.

Non-synonymous Substitution Calling

As OMIM and HGMD data are stored in gene and amino acid coordinates, only non-synonymous single nucleotide substitutions and not all substitutions are queried against these datasets. Amino acids corresponding to a given nucleotide change are calculated by a script that translates DNA sequences in the appropriate frame after consulting a database table that describes coding sequence locations (UCSC table "refFlat")[1]. Comparison of the Trait-o-matic function inference algorithm against dbSNP revealed an error in the dbSNP algorithm whereby amino acids encoded across a splice junction may be incorrectly reckoned. In rs9370096, for example, a C→T mutation in coding strand of the polycystic kidney and hepatic disease 1 (*PKHD1*) gene produces a

cysteine residue. While the third letter of the corresponding codon is located at the beginning of exon 22, the dbSNP algorithm apparently ignores the splice junction and proceeds incorrectly into the intron, reading [C/T]GG instead of [C/T]GC. Genetic code redundancy shields the effect of this error in the case of the C allele but not the T allele, with the result that dbSNP gives the amino acid change as R760W instead of R760C (figure S2-2). Recently, this error in dbSNP seems to have been corrected.

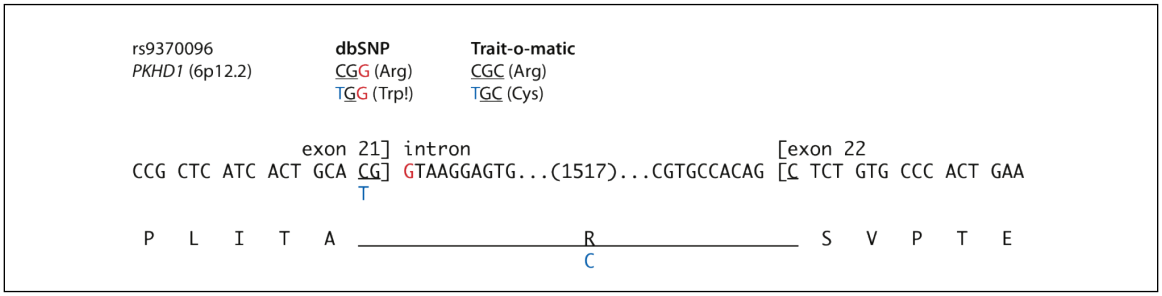


Figure S2-2. Previously dbSNP incorrectly calculated amino acid changes for codons split by splice junctions. Differences arise between dbSNP and Trait-o-matic function inferences due to erroneous use of intron sequences by dbSNP. *Top*, dbSNP and Trait-o-matic function inferences for rs9370096; *middle*, excerpt of coding strand nucleotide sequence for *PKHD1* near rs9370096 (blue); *bottom*, corresponding amino acid sequence.

Source Data and Explanation of Non-synonymous Substitution Discrepancies with Source Data

Data for each genome were obtained from the source listed in table S2-1. Also noted are the number of SNP variants reported in each primary publication (where available), and the number that we obtained. Summary statistics for all genomes is presented in figure S2-3. Trait-o-matic retrieved 9060 non-synonymous substitutions from YH data, 28% more than the published claim of 7062. Subsequent examination revealed that Wang *et al.*[2], authors of the original data, had already labeled 8166 substitutions as nonsynonymous in the retrieved file, 16% more than their own published claim. 8128 of these 8166 (99.5%) were among the 9060 designated as such by Trait-o-matic. Manual follow-up on a random subset of the remaining 38 shows a mixture of

substitutions where: (a) the coding sequence claimed by Wang *et al.* has been permanently suppressed in NCBI databases due to insufficient evidence for it being part of a gene; (b) dbSNP and Trait-o-matic agree that the substitution is synonymous; or (c) the claimed exon is absent in the table of coding sequences ("refFlat") consulted by Trait-o-matic, presumably because it belongs to a suppressed isoform. Conversely, 11 of the 9060 non-synonymous substitutions designated by Trait-o-matic were erroneously labeled as synonymous by Wang *et al.* because these substitutions are silent in at least one coding sequence but nonsynonymous in at least another (data not shown). The remaining 894 non-synonymous substitutions designated by Trait-o-matic but not Wang *et al.* were annotated as non-coding by the authors, likely reflecting the use of an older dataset for coding sequence locations.

PGP Data Release 9.3

We have previously reported the targeted capture of 55,000 exons[3] and, utilizing protocol improvements by Li et al.[4], we captured exons for the first ten volunteers in the Personal Genome Project (PGP). The hybridization arms were the same as previously reported[3], while the universal sequence was updated to reflect the changes introduced by Li et al. and to include an EcoP15i recognition site 35bp from either terminus. With this improved protocol we nevertheless found a 20-fold decrease in capture efficiency (compared to the mean) when the 5' end of the ligation arm was thymidine, and a 2.5-fold decrease when it was adenosine. Upon switching to Pfu DNA polymerase, we found a remarkable improvement, with the bias reduced to 2-fold and 1.1-fold, respectively. An

added benefit of this polymerase was its lack of sensitivity to higher concentrations of dNTP, allowing the dNTP concentration to better match the K_m of the polymerase.

After PCR amplification of the captured exons, we digested with EcoP15i to remove the universal sequence and approximately half of each anchor arm. While most restriction endonucleases require their recognition site to be present no more than 6 nucleotides from the 5' or 3' end for efficient cleavage, we found that EcoP15i exhibited no cleavage even with the recognition site 10nt from the terminus (data not shown). We PCR amplified with 5' biotinylated primers containing the EcoP15i recognition site 20nt from the 5' end, digested and removed the cleaved sequence with Streptavidin coated beads. This permitted us to avoid a size selection and served to retain the ~12% of exonic sequence that contained the CAGCAG recognition sequence, and where therefore of a similar size to the primers. The fragments were then polished, concatenated via T4 DNA ligase and sheared to produce a shotgun library for the Illumina GAII. Upon average, only 10% of material was recovered after the EcoP15i digestion and streptavidin-bead enrichment despite a 2x over-digestion. The enrichment, however, was successful with ~36% of positions within the retained material seen at least 3 times, while only ~3% of the removed material was seen at least 3 times. Remarkably, the presence of an additional EcoP15i site in a capture region increased mean capture 3.7-fold. While it is possible that the additional chance for the enzyme to cleave would yield this result, it is also likely that the 20bp distance is insufficient for efficient cleavage and positioning the recognition site even further from the terminus would aid EcoP15i digestion.

Utilizing the Illumina GAII, a total of 2 gigabases of raw sequence data was generated from these libraries with 59% of it aligning uniquely to the reference—without

insertions or deletions and up to two mismatches using MAQ 0.6.7. Of the 1.188 gigabases of placed reads, 873 Mbp placed against the target regions when placement was allowed against the entire genome. Of the 6.7Mbp target region, the average sequenced region was 5.67Mbp (4.31-7.08), implying an average coverage of 15.4x. For each participant we optimized the baseline coverage threshold and consensus quality score so that the individual's concordance with the ~1000 target sites overlapping each individual's Affymetrix 500K microarray would be 99%. This threshold was chosen taking into account the self-concordance for the Affymetrix DM calling algorithm (http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf). We manually calculated the results using both a MAQ consensus quality/coverage threshold and a unique sequence/coverage threshold to filter PCR duplicates and found that the former filter consistently outperformed the latter; by applying both together we were able to further increase accuracy. All positions in which any non-reference call was covered by at least three definitely independent reads were compared to the matching Affymetrix data. The 99% concordance was based upon an average of 225 positions, ranging from 86 to 418. The accuracy for heterozygous calls was 356/361 (98.6%), 1453/1471 (98.8%) for reference calls and 352/355 (99.2%) for non-reference bases. After application of this filter and extrapolating the call rate of these 1000 positions to the entire 6.7Mbp target region, we calculate an average call rate of 1.61Mbp/participant, with a range of 0.6-3.0Mbp. We used this top 28% of the data for all further calculations. The reason for the large spread in coverage was partly due to the raw reads generated by each sequencing lane, but mostly due to the limited complexity of a few of the libraries.

In the exon data we focused on the 7,412 substitutions that were called with an accuracy of 99%. 658 of these were the result of chimeric sequence (a library construction artifact) or possible synthesis errors in the capture arms, and 68 were off-target calls, potentially an artifact of the placement algorithm. The on-target capture rate after application of these filters was 96.9%, with a range from 91.9 – 99.8%. For the remaining 6,686 substitutions a bed file was generated with the calls from all 10 genomes for those positions, for a total of 16,267bp. These substitutions represent 3,110 unique positions of which 40% are non-synonymous and 827 are novel.

PGP1 WGS

The whole genome sequence of the first Personal Genome Project volunteer (PGP1) was performed by Complete Genomics, Inc (Mountain View, CA). Although CGI is developing technology for delivery of phased diploid genome sequences, the PGP1 genome is not phased. Of autosomal bases called in the reference genome, 90.7% were called in this genome. In addition to explicitly specified variations, CGI describes 134.1Mb (~4.7%) of the genome incompletely as merely "consistent" or "inconsistent" with the reference sequence. Comparisons with 500K Affy SNP Chip data on PGP1 indicates > 99.3% concordance in all regions fully called by CGI, and > 93.5% if uncalled or incompletely described regions are counted as false negatives.

Supplemental Discussion 2-1 - Recessive Variants

Caucasian Genomes

For NA07022

1. chr17:37949759 NAGLU, R737G: Also found in **JW and PGP1** genome, this variants is found by one group in a patient with Sanfilippo Syndrome B, together with a known Sanfilippo variant.[5] TAF is 0.049.
2. chr9:135291831 ADAMTS13, P457L: This variant was found as part of a compound heterozygote in Thrombotic thrombocytopenic purpura patient.[6]
3. chr11:66050228 BBS1, M390R: When homozygous this variant is causative for Bardet-Beidl syndrome in an oligogenic fashion.[7] (OMIM*209901.0001)
4. chr2:201782343 CASP10, V410I: This variant was implicated in autosomal recessive Autoimmune lymphoproliferative syndrome type II.[8]
5. chr1:97754009 DPYD, S534N: It is debated in the literature as to whether the heterozygote causes lower levels of DPD expression. Gross et al.,[9] note a severe phenotype in two instances of compound heterzygosity.
6. chr15:78259581 FAH, R341W: This variant is noted in OMIM (OMIM+276700.0006) as being a pseudodeficiency allele fumarylacetoacetase levels connected to hereditary tyrosinemia type 1; it has been found to be a factor in compound heterozygotes.[10]
7. chr16:3244464 MEFV, R202Q: Ritis et al find that this variant (TAF of 0.125 in Greek population (?)) is possibly an autosomal recessive causative variant for Familial Mediterranean fever.[11]

8. chr22:16946288 PEX26, L153V: This variant, also seen in **NA12878**, is found as part of a compound heterozygote (together with a frameshift/nonsense mutation) causative for infantile Refsum disorder.[12]
9. chr19:46550716 TGFB1, R25P: This variant is causes a variation in TGFβ1 levels. It has been associated with the stage of hepatic fibrosis in patients with chronic hepatitis C in an association study.[13]
10. chr9:107406320 FKTN, G125S: This variant was implicated in autosomal recessive Walker-Warburg syndrome, after being found in one Spanish patient as a compound heterozygote with 473-bp deletion in the same gene[14] (OMIM*607440.0012). TAF is unknown.

For JCV

1. chr12:1823893 CACNA2D4, Y802*: this variant is autosomal recessive for retinal cone dystrophy through truncation of one-third of the ORF (OMIM *608171.0001).[15] TAF is unknown.
2. chr1:5862830 NPHP4, R848W: This variant has been found together with R682X as a compound heterozygote in three nephronophthisis and retinitis pigmentosa, (Senior-Løken syndrome) patients from one family.[16-17] TAF is 0.036.
3. chr16:55493820 SLC12A3, R919C: This variant, also **seen in NA18555**, is implicated in Gitelman's syndrome, having been seen in two patients: one compound heterozygous for this, Gly741Arg and Arg968Ter and another for this and Arg968Ter. It was also seen in 1/50 controls.[17] TAF is 0.009.

4. chr2:227625327 COL4A4, S969*: This variant has been implicated in two families with autosomal recessive Alport syndrome as both a compound heterozygote and homozygote.[18] TAF is unknown.
5. chr3:15660878 BTD, A171T: This variant, together with D444H (also in JCV) makes up the most common cause of biotinidase deficiency in newborn screenings in one early study (14/31).[19] Later research finds that the biotin level in homozygous individuals is not less than that of individuals heterozygous for more penetrant variants.[20] TAF is unknown.
6. chr4:100740828 MTTP, D384A: Found homozygously, this variant is implicated in familial hypobetalipoproteinemia as a compound heterozygote with G661A and with a background of homozygosity for ε2 allele for apolipoprotein E.[21] TAF is 0.033.
7. chrX:100543309 GLA, C172G: This variant was implicated in Fabry disease where it was found together with D313Y in one patient.[22] TAF is unknown.

For JW

1. chr16:55106002 BBS2, N70S: This variant was found as part of compound homozygote with a previously reported causative mutation in the MKKS gene[23] in a patient with Bardet Biedl Syndrome 2 (OMIM*606151.0013). TAF is 0.00925.
2. chr5:74017026 HEXB, L62S: This variant was reported in a patient with infantile Sandhoff disease as part of a compound heterozygote together with a 25kb deletion spanning part of the gene[24] (OMIM*606873.0012). TAF is 0.042.

3. chr1:156891152 SPTA1, A970D: It is debated in the literature as to whether this variant is presumed causative for autosomal recessive spherocytosis type 3[25] (OMIM*182860.0009). TAF is 0.0087.
4. chr16:49303691 NOD2, R790W: This variant was found as a compound heterozygote with p.Leu1007fs in a patient with severe Crohn's disease.[26] TAF is unknown.
5. chr16:55462088 SLC12A3, G264A: This variant, **also found in PGP3**, was implicated in Gitelman's syndrome in a patient compound heterozygous for this variant and P643L.[27] TAF is 0.044.
6. chr16:88342653 FANCA, S1088F: Also found in **PGP4, PGP9**, **and YH,** this variant has been implicated as causative for autosomal recessive Fanconi anemia, and was not found in 100 normal chromosomes.[28] TAF is 0.05.
7. chr18:32039091 MOCOS, M358V: This variant was found as a compound heterozygote (with R419*) in a patient with classical xanthinuria type II.[29] TAF is 0.038.
8. chr3:33030725 GLB1, C521R: This variant (also referred to as C491R) is implicated in autosomal recessive G_{M1}-Gangliosidosis, with gene expression being 25% of normal.[30] TAF is 0.0089.
9. chr4:3444964 DOK7, S45L: This variant is implicated in congenital myasthenic syndromes when in a compound heterozygote with P376P fsX30 or P469H (one patient each).[31] TAF is unknown.
10. chr4:619702 PDE6B, Y219H: This variant is implicated in autosomal recessive retinitis pigmentosa as a compound heterozygote.[32] TAF is 0.0087.

11. chr5:149341070 SLC26A2, T574I: This variant is classified as uncertain pathogenic in autosomal recessive diastrophic dysplasia. It is found as a heterozygote in one putative diastrophic dysplasia patient where a second mutation could not be found, and in one compound heterozygote with two additional mutations.[33] TAF is 0.0089.
12. chr6:49688206 RHAG, V270I: This variant is implicated in autosomal recessive Rh_{null} disorder. It was found as a compound heterozygote with G280R.[34] TAF is 0.018.
13. chr17:37949759 NAGLU, R737G: See NA07022.
14. chr1:16243654 *CLCNKB, L27R:* This variant, **also found in P0 and NA19129,** has a TAF of 0.25 in the "general population" and is considered a benign polymorphism[35] that may be a Bartter syndrome 3 modifier in a homozygous state (found in one Japanese patient).[36] TAF is unknown.

For PGP1

1. chr11:133634133 ACAD8, S171C: This variant (reported in the literature as S149C) is implicated in autosomal recessive isobutyryl-CoA dehydrogenase deficiency in one patient. Natural history is not well defined,[37] so there it is likely of low penetrance. TAF is unknown.
2. chr4:6355034 WFS1, V871M: This variant was identified in one individual with autosomal recessive Wolfram syndrome in a heterozygous state and in two deaf sisters with heterozygosis, the sisters' father, however carried the variant and was asymptomatic.[38] TAF is unknown.
3. chr16:55493820 SLC12A3, R919C: See JCV.
4. chr17:37949759 NAGLU, R737G: See NA07022.

For P0

1. chr1:16243654 *CLCNKB, L27R:* See JW.
2. chr17:7956211 ALOXE3 L237M: This variant is implicated in nonbullous congenital ichthyosiform erythroderma, having been seen in one homozygous patient[39] (OMIM*607206.0004). TAF is unknown.
3. chr1:25762834 *LDLRAP1, R238W*: This variant was found homozygous together with a homozygous frame-shift in two British siblings with autosomal recessive Hypercholesterolaemia.[40] TAF is unknown.
4. chr1:94289842 *ABCA4, G863A:* This variant was found as part of a compound heterozygote in three Stargardt patients.[41] TAF is 0.018.
5. chr1:94301301 *ABCA4, R572Q*: This variant was found as part of a compound heterozygote in two separate families with autosomal recessive Stargardt disease.[42] TAF is unknown.
6. chr1:155115542 *NTRK1, H598Y*: See below. TAF is 0.045.
7. chr1:155115570 *NTRK1, G607V:* These two variants are likely in cis as they were implicated autosomal recessive congenital pain insenstivity when both were homozygous in the offspring of a consanguineous union.[43] TAF is unknown for this variant.
8. chr2:44393296 *SLC3A1, M467T:* This variant represents 40% of chromosomes from cystinuria patients and was found in a homozygous state in four patients.[44] TAF is unknown.
9. chr3:15661697 *BTD, D444H*: See JCV variant list.

10. chr5:149340823 *SLC26A2, R492W*: Found in "numerous" autosomal recessive diastrophic dysplasia patients.[33] TAF is unknown.
11. chr8:24828217 *NEFM, G336S*: Found in one **autosomal dominant** early-onset Parkinson disease patient with three asymptomatic siblings harboring the variant; penetrance is therefore 25% at age 44.[45] TAF is unknown.
12. chr12:32940857 *PKP2, D26N*: This homozygous variant was implicated in autosomal dominant Arrhythmogenic right ventricular dysplasia, but was "unclassified pathogenicity" due to (1) it not being in a conserved region and (2) predicted not very deleterious.[46] TAF is unknown.
13. chr17:7925204 *ALOX12B, P127S*: This variant, also seen in **NA12878**, was implicated in autosomal recessive congenital ichthyosis, in two Turkish siblings, but the second mutation was not found. The phenotype involved self-healing abdominal scaling.[39] TAF is unknown.
14. chr17:71338086 *UNC13D, R928C*: This variant was implicated in autosomal recessive familial haemophagocytic lymphohistiocytosis. It was found in 2 unrelated Italian patients as a compound heterozygote; both required chemotherapy and received unmatched bone marrow donations.[47] TAF is unknown.

For PGP3

1. chr16:55462088 SLC12A3, G264A: See JW.

For PGP4

1. chr15:25902148 OCA2, A481T: This variant was implicated in autosomal recessive ocular albinism,[48-49] and shows a reduction in expression to 70% of wild-type.[50] This variant is quite common in East Asia,[51] but very rare in the European population.[52] TAF is unknown.
2. chr16:88342653 FANCA, S1088F: See JW.

For PGP9

1. chr1:46428232 POMGNT1, D556N: This variant was implicated in autosomal recessive late onset limb-girdle muscular dystrophy without mental retardation after it was seen in one patient.[53] While it has an altered kinetic profile, it appears to be of low penetrance, having been seen with MAF of 0.037 in French controls and in one unaffected homozygous sibling of a proband.[54] TAF is unknown.
2. chr16:88342653 FANCA, S1088F: See JW.

For NA12156

1. chr7:138068331 *ATP6V0A4, M580T* A/G *G*: See **SJK**. CEU TAF 0.033.
2. chr8:55700113 *RP1, T373I* C/T T: This variant, implicated in autosomal recessive Retinitis Pigmentosa, was found in 2 consanguineous Pakistani families[55] (OMIM 603937.0006).
3. chr1:94337071 *ABCA4, R212H* C/T *T*: This variant was found in one heterozygote in a screen of 112 Chinese AMD patients and was not seen in 94 controls.[56] Baum et

al.[57] find it in controls, and state that it may have some effect on AMD. CEU TAF 0.049

4. chr1:94341274 *ABCA4, R152Q* C/T *T* found in one Spanish homozygous for autosomal recessive stargardt disease.[58] TAF is unknown.
5. chr1:214605014 *USH2A, V230M* C/T *T* found as compound heterozygote in one Caucasian autosomal recessive Usher syndrome 2a patient. It was also seen in 1/100 control chromosomes.[59] TAF is unknown.
6. chr2:179034062 *DFNB59, G292R* G/A *A* This variant was found in one homozygous Iranian and implicated in autosomal recessive non-syndromic hearing loss; it was also seen in 1/100 control chromosomes.[60] CEU TAF 0.04386
7. chr2:208816546 *IDH1, Y183C* T/C *C* This variant was found segregating with early onset osteoarthritis in one Caucasian family, but was also found with a 0.02 frequency in subjects without significant enrichment for disease. The disease often has complex inheritance.[61] TAF is unknown.

NA12878

1. chr5:179978517 *FLT4, P954S* G/A *A* This variant was found in 1/15 infantile hemangioma specimens.[62] (OMIM 136352.0005) TAF is unknown.
2. chr12:101758382 *PAH, Y414C* T/C *C* This variant is the most common cause mild recessive non-PKU Hyperphenylalaninemia in N. Europe, accounting for 5% of cases[63] (OMIM 612349.0017). TAF is unknown.
3. chr17:7925204 *ALOX12B, P127S* G/A *A* See P0.

African Genomes

For NA18507

1. chr10:55261259 PCDH15, Q1342K: This variant, found only in Bentley's and Ng's version, is implicated in Usher syndrome type I when present as a compound heterozygote with T1867del in a patient of European descent.[64] TAF is unknown.
2. chr2:220148070 INHA, G227R: This variant, found in Bentley's and Ng et al.'s version and seen in **NA18517**, is implicated in low penetrance pediatric adrenocortical tumors as a compound heterzygote with TP53 R337H in 3 Brazilian patients.[65] TAF is unknown.
3. chr4:178468213 NEIL3, R38C: This variant, found in Bentley's and Ng's version, was found as a compound heterozygote (together with R15 (a common polymorphism)) in three patients with multiple colorectal adenomas.[66] TAF is unknown.
4. chr11:95209096 MTMR2, N545S: This variant, present in all versions, was found in two patients with autosomal recessive Charcot-Marie-Tooth type 4B, but the compounding variant was not found.[67] TAF is 0.056.
5. chr15:26000537 OCA2, G27R: This variant, present in all versions, was found in an Ashkenazi Jewish individual with autosomal recessive oculocutaneous albinism.[68] TAF is 0.017.
6. chr15:25909324 OCA2, I370T: This variant, present in all versions, was implicated in the autosomal recessive oculocutaneous albinism ("the most common recessively inherited disorder among Southern African Blacks") in two compound heterozygous

individuals.[69] TAF is unknown (in the SNP500 dataset in dbSNP it is 0.021 in a sample of 24 individuals).

7. chr1:94285190 ABCA4, V931M: This variant, present in all versions, is implicated in autosomal recessive Stargardt macular dystrophy in a Saudi Arabian family.[41] TAF is unknown.
8. chr3:10064689 FANCD2, L456R: This variant, present in all versions, is reported pathogenic in the Ashkenazi Jewish population in autosomal recessive Fanconi anemia with severe phenotype.[70] TAF is unknown.
9. chr15:40467295 CAPN3, T184M: Found in all versions; see PGP10.

For PGP10

1. chr15:40467295 CAPN3, T184M: This variant, also **found in NA18507,** is implicated in patients expressing "clinical criteria of a classic LGMD phenotype—namely, autosomal recessive inheritance, progression of muscle weakness in a limb-girdle distribution."[71] TAF is unknown. (dbSNP reports a TAF of 0.013 in a mixed population of African Americans and Caucasians)

For 18517

1. chr12:51047224 *KRT85, R78H C/T T* This variant was found segregating with recessive ectodermal dysplasia in a consanguineous Pakistani family. Not found in 200 control chrom from Pakistan[72] (OMIM 602767.0001). TAF is unknown.

2. chrX:84449850 *POF1B, R329Q C/T T* This variant was found in sex-linked recessive premature ovarian failure 2b in Lebanese family[73] (OMIM 300603.0001). TAF is unknown.
3. chr2:220148070 *INHA, G227R* G/A *A* See NA18507.
4. chr8:24831447 *NEFM, P725Q* C/A *A* Found in 2 Parkinson's disease patients (Caucasian and Arab) but not in 236 controls, and the authors argue for this being a susceptibility factor.[74] YRI TAF 0.032
5. chr12:5998328 *VWF, S1506L* G/A *A* This variant is implicated in autosomal recessive Von Willebrand disease.[75] TAF is unknown.
6. chr12:101784513 *PAH, N167S T/C C* This variant was found in one recessive phenylketonuria newborn in Texas newborn screening.[76] TAF is unknown.
7. chr17:7856637 *GUCY2D, P701S* C/T *T* See SJK. YRI TAF is 0.026.

NA19129

1. chr1:16243654 *CLCNKB, L27R* G/T *G* See JW. TAF is unknown.
2. chr1:53440651 *CPT2, A101V* C/T *T* Found in one patient with carnitine palmitoyltransferase 2 deficiency, authors conjecture dominant expression of this variant. No controls were reported.[77] YRI TAF 0.045.
3. chr6:162542229 *PARK2, P153R* G/C *C* Found heterozygously in one patient with recessive Parkinson disease; the second mutation is not reported. Controls were not reported.[78] This variant was also reported (without the second mutation) in an individual from Yoruba with sporadic PD and was seen in 1/51 Nigerian controls.[79] YRI TAF 0.018

4. chr16:8849089 *PMM2, N216S A/G G* This variant was found in one Scandinavian patient with recessive congenital disorder of glycosylation 1a and not seen 100 control chromosomes.[80] TAF is unknown.
5. chr19:43656115 *RYR1, S1342G* G/G *G* found in study of N. African and Caucasians with dominant malignant hyperthermia (central core disease).[81] Later reports note the frequency in control populations and the need for more research to understand the susceptibility to disease.[82] TAF is unknown.

NA19240

1. chr1:55284810 *PCSK9, Y142** C/G *G* This variant is quantitative trait locus for lower low density lipoprotein cholesterol level[83] (OMIM 607786.0004). TAF is unknown.
2. chr1:20847608 *PINK1, A383T* G/A *A* This variant is found in a screen or Parkinson Disease variants in the UK population. Found in one patient in a heterozygous patient with a brother with PD. The brother was not available to be checked. Not seen in 1,576 control chromosomes.[84] Ibanez et al. view it as benign when found in one Moroccan patient,[85] and Ishihara-Paul et al. lend credence to this view by finding it in a number of older controls.[86] TAF is unknown.
3. chr12:5965032 *VWF, R2287W* G/A *A* In a European study for Von Willebrand disease 1 factors this variant was found as a compound heterozygote in one individual.[87] TAF is unknown.
4. chr13:31812839 *BRCA2, H2116R* A/G *G* This variant was seen in one African American family with various cancers; it was labeled an "unclassified variant."[88] TAF is unknown.

5. chrX:153236201 *FLNA, A1764T* C/T *T* This variant found in one female patient with periventricular heterotopias. She has 5 sons, two of whom are mentally retarded, but it cannot be confirmed if they carry this variant or if this is the cause.[89] TAF is unknown.

Asian Genomes

For YH

1. chr12:101773223 PAH, R176Q: This variant is implicated in autosomal recessive mild hyperphenylalaninemia in 8.6% of 93 patients as a compound heterozygote in Spain.[90] TAF is unknown.
2. chr16:3244627 MEFV, E148Q: This variant, also **found in AK and SJK and NA18555,** is implicated in autosomal recessive Familial Mediterranean Fever with penetrance of ~55%. It has been conjectured to be a polymorphism[91-93] (OMIM*608107.0005) and may be so in the Asian population. TAF is unknown.
3. chr16:87411936 GALNS, V488M: This variant found in autosomal recessive mucopolysaccharidosis IVA patients of Japanese origin, is assumed to be a benign polymorphism based upon cell line analysis.[94] TAF is unknown.
4. chr16:88342653 FANCA, S1088F: See JW.
5. chr1:155115542 *NTRK1, H598Y*: See P0.
6. chr1:155115570 *NTRK1, G607V:* See P0.

For AK1

1. chr1:214438429 USH2A, S1311*: This variant is predicted disease causing for autosomal recessive Usher syndrome type IIa, found in a single patient of Swedish descent.[95] TAF is unknown.
2. chr19:5782672 FUT6, R303G: This variant is autosomal recessive for deficiency in plasma α3-fucosyltransferase activity in two Indonesian individuals.[96] TAF is unknown.
3. chr2:38155078 CYP1B1, V320L: This variant is implicated in autosomal recessive Primary congenital glaucoma, and was found in two patients of Japanese descent as a compound heterozygote.[97] TAF is unknown.
4. chr16:3244627 MEFV, E148Q: See YH.

For SJK

1. chr10:13380242 PHYH, P29S: This fairly common variant in the CEU population (TAF 0.18, also found in P0 and NA12156) has been implicated in autosomal recessive Refsum's disease in two patients of unreported ethnicity.[98] TAF for Asian is unknown.
2. chr10:96530400 CYP2C19, W212*: This variant is implicated in autosomal recessive poor metabolism of (S)-mephenytoin (anticonvulsant drug) in 34 patients.[99] This has also been implicated in autosomal recessive poor proguanil metabolism (an anti-malarial drug) [100] (OMIM*124020.0003). TAF is 0.045 (and 0.033 in a a NCBI study).

3. chr17:7856637 GUCY2D, P701S: This variant, also seen in **NA18517 and NA18555**, is implicated in autosomal recessive Leber congenital amaurosis in several patients from diverse ethnicities.[101] TAF is unknown. (TAF for YRI is 0.026 and for CEU is 0.0087)
4. chr20:4628521 PRNP, E219K: The Glu variant has been found homozygous in ~85 cases of CJD in Japan[102] (OMIM*176640.0019). TAF is unknown (other sources estimate it to be 0.06 in Japan[103] (OMIM*176640.0019).
5. chr7:138069331 ATP6V0A4, M580T: This variant, also seen in **NA12156**, was implicated in autosomal recessive distal renal tubular acidosis when it was found in 4yo Turkish patient[104] (OMIM*605239.0005).
6. chr16:3244627 MEFV, E148Q: See YH.

NA18555

1. chr7:44071364 *PGAM2, G97D* C/T *T* This variant was found in 2 Japanese individuals in a heterozygous manner with glycogen storage disease X and exercise intolerance[105] (OMIM 6120931.0004). TAF is unknown.
2. chr16:3244627 *MEFV, E148Q* C/G *G* See YH.
3. chr17:7847244 *GUCY2D, A52S* G/T *T* This variant segregated with autosomal recessive Leber congenital amaurosis type 1, but also found in 2/100 controls[106] (OMIM 600179.0004). TAF is unknown.
4. chr13:19661305 *GJB2, S139N* C/T *T* This variant was found in 1 family with autosomal recessive congenital deafness and 0/116 controls.[107] TAF is unknown.
5. chr16:55493820 *SLC12A3, R919C* C/T *T* See JCV. JPTCHB TAF 0.020

6. chr17:7856637 *GUCY2D, P701S* C/T *T* See SJK.
7. chr17:7847151 *GUCY2D, W21R* T/C *C* This and the previous variant are claimed to a benign/weak polymorphism in at least some populations in Leber congenital amaurosis.[108]
8. chrX:107752652 *COL4A5, G953V* G/T *T* found in one family with x-linked recessive Alport syndrome segregating with disease.[109] TAF is unknown.

NA18956

1. chr21:45755537 *COL18A1, D1437N* G/A *A* This variant was found as part of a compound het with a 1bp del in autosomal recessive knobloch syndrome type 1.[110] (OMIM 120328.0004) JPTCHB TAF 0.020
2. chr1:94268627 *ABCA4, V1433I* C/T *T*: This variant was found in one family in a screen of 150 families with autosomal recessive Stargardt disease, most of whom were Caucasian.[42] JPTCHB TAF 0.0085

Supplemental Discussion 2-2 - Comparison of NBLOSUM and SIFT/POLYPHEN

To score non-synonymous substitutions not found in database sources, Trait-o-matic uses a simple scoring system based on a substitution matrix to minimize computational demands, inferring that a serious mutation—a non-conservative or nonsense mutation—in a gene associated with a particular disease is more likely to be deleterious. Of the two commonly used sets of substitution matrices, we have chosen to use a BLOSUM (block substitution matrix) over a PAM (point accepted mutation) matrix because the latter is asymmetric. The presence of symmetry obviates consideration of which allele at a polymorphic locus is ancestral; this is not difficult for novel mutations (the reference allele can reasonably be taken as ancestral) but requires additional database queries for common polymorphisms. Of the block substitution matrices, we have used the highest threshold of sequence similarity (BLOSUM100), since we are only examining variants within the same species. We have also taken the negative of the matrix (NBLOSUM) in answer to the intuitive notion that less conservative amino acid changes ought to have a higher score, since these scores are used as a measure of potential deleterious effect. Though this matrix-based procedure assigns integer values normalized so that stop codons have a score of 10, we make no assumption about any linear progression of phenotypic effect between integer increments, and hence use non-parametric methods to assess the effectiveness of this scoring method in discriminating benign and deleterious mutations.

As expected, pairwise Mann-Whitney U tests show that score distributions do not differ significantly for non-synonymous substitutions in the JCV, JW, YH and NA18507 (Bentley) genomes. Two pairwise tests showed $p < 0.05$ (between the Venter and YH

genomes ($U = 4.75 \times 107$, $n_1 = 10{,}690$, $n_2 = 8742$, $p = 0.0437$ with continuity correction, two-tailed) and between the Venter and Yoruba genomes ($U = 5.26 \times 10^7$, $n_1 = 10{,}690$, $n_2 = 9650$, $P = 0.0118$ with continuity correction, two-tailed)), but these results are not significant after Bonferroni correction for multiple hypotheses. However, the score distribution of any complete genome does differ significantly with that of aggregated 9.3 PGP exome data (the least significant p-value arising between the Venter genome and the PGP, $U = 3.62 \times 10^7$, $n_1 = 10{,}690$, $n_2 = 7189$, $p = 3.36 \times 10^{-11}$ with continuity correction, two-tailed). Also as expected, the OMIM dataset is shifted significantly towards higher matrix-based scores as compared with any genome; a representative test between OMIM and Watson's genome gives $U = 2.10 \times 10^7$, $P < 2.2 \times 10^{-16}$ with continuity correction, one-tailed.

Since nearly all OMIM alleles are deleterious and ~80% of non-synonymous substitutions in an individual's genome are estimated to be benign, we use OMIM alleles as positive controls and Watson's alleles as negative controls to estimate true and false positive rates for the use of matrix-based scores in predicting deleterious alleles. Watson's advanced age and relative health imply that his genome contains relatively few deleterious alleles. In this case, plotting these data yields an area under the receiver operating characteristic (ROC) curve of 0.695, and setting a decision boundary at $\geq +3$ yields a maximal difference of 29.1% between true and false positive rates. As deleterious alleles are correlated with higher scores, one expects that a true set of negative controls would show lower classification error.

To compare the effectiveness of NBLOSUM-based scoring with that of SIFT[111], we retrieved precomputed SIFT predictions for Watson and OMIM non-synonymous

substitutions. Because the most efficient method of accessing these predictions is through submission of dbSNP rs IDs via a web interface, we limited positive controls to OMIM alleles mapped directly to dbSNP rs IDs (in OmimVarLocusIdSNP) and negative controls to Watson alleles found in dbSNP. For additional comparison, ROC curves were plotted for SIFT predictions, NBLOSUM-based scores, minor-allele-frequency (MAF)-based scores, and summed NBLOSUM/MAF-based scores based on this set of positive and negative controls. To generate MAF-based scores, we consider the lowest MAF across aggregated HapMap populations and apply a logarithmic function f(MAF) such that $f(0.5) = -10$, $f(0.05) = 0$, $f(0.005) = 10$, to a maximum score of 15. We presume that polymorphisms at loci with no frequency data are rare and assign them a score of 15. This function was somewhat arbitrarily chosen so that we could assign approximately equal weight to the NBLOSUM-based score and MAF-based score by addition, and was devised without the use of a training set for parameter optimization. Examination of ROC curves shows that SIFT (area under curve = 0.848) had the highest predictive accuracy and NBLOSUM-based scores had the lowest (area = 0.659), but summed NBLOSUM/MAF-based scoring dramatically increased accuracy (area = 0.787) and was more effective than MAF-based scoring alone.

PolyPhen[112] claims still higher accuracy than SIFT, but because PolyPhen results are qualitative only (benign, possibly damaging, probably damaging, unknown), an ROC curve cannot be generated. Still, if we accept PolyPhen's claimed true positive rate of 82% and false positive rate of 8%, the NBLOSUM-based scoring method clearly provides inferior results for predicting deleterious alleles in exchange for significant increases in computational speed. It remains to be seen if adjusting minor-allele

frequency-based scoring or its combination with NBLOSUM-based scoring using training sets can push the accuracy of this simple algorithm beyond that of SIFT and PolyPhen.

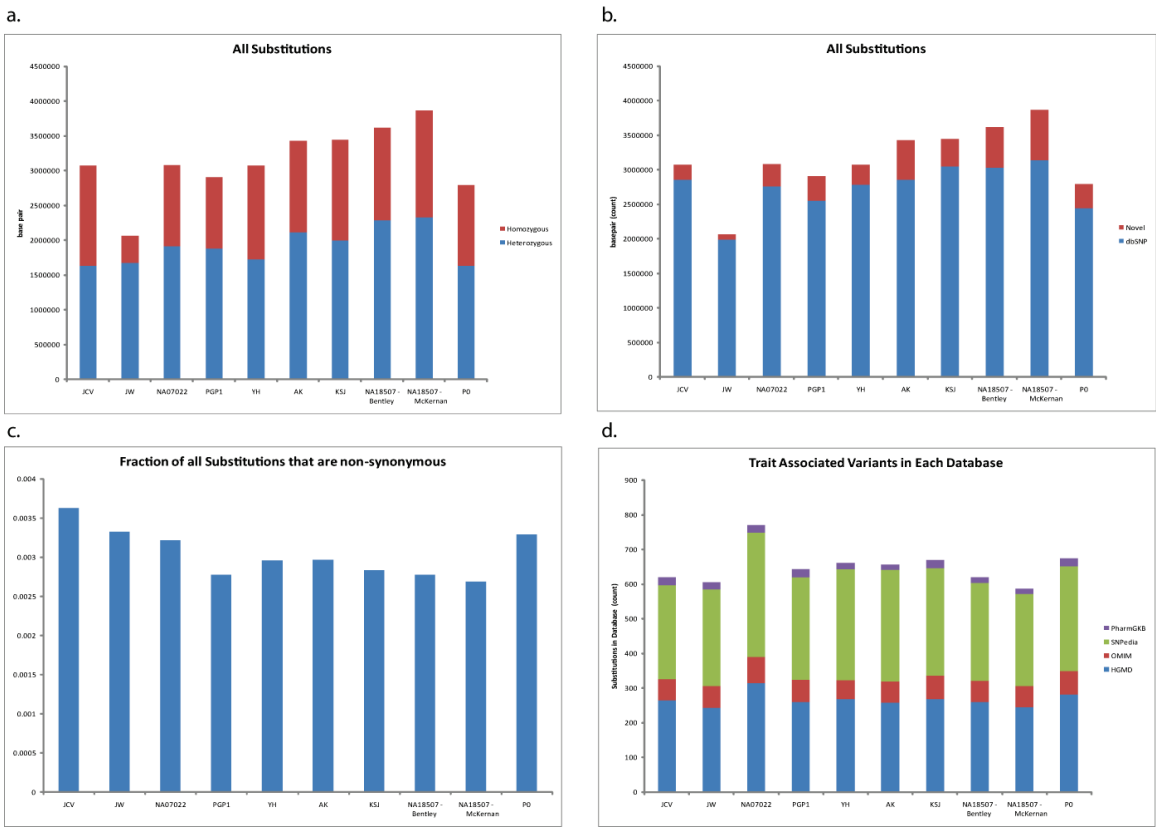


Figure S2-3.1. Comparison of Genome Statistics. (a) Heterozygous/Homozygous Variant Calls for Each Whole Genome Sequence. (b) dbSNP/Novel Variant Calls for Each Whole Genome Sequence. (c) Fraction of Variant Calls that are Non-Synonymous. (d) Variants in Each Database. It is difficult to gauge sequencing quality based upon these data alone since individual variability is still not well understood.

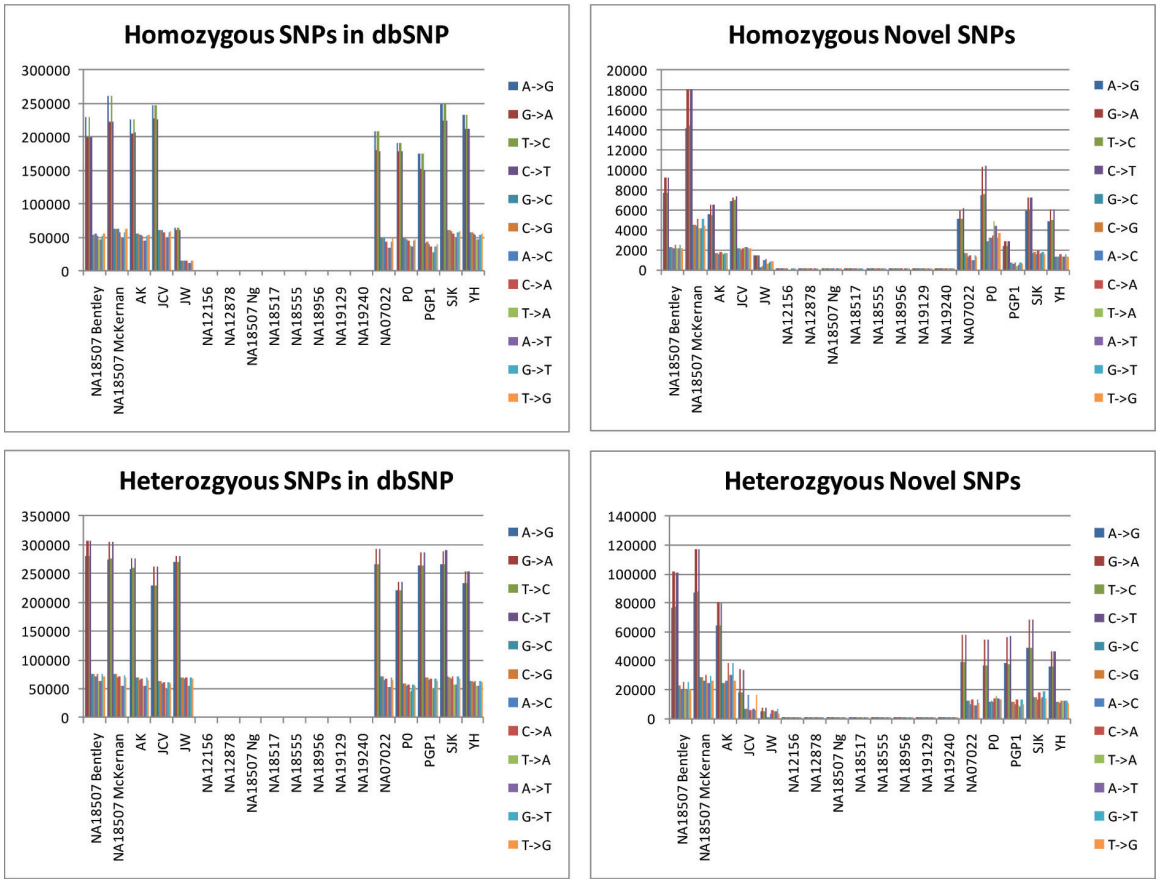


Figure S2-3.2 Detailed Comparison of Single Nucleotide Substitutions between Sequencing Projects
Each chart is as labeled, with the number of times each particular substitution is seen plotted for each genomes analyzed. The consistency between the different genomes and different platforms is remarkable. Especially striking is the predominance of the ancestral allele in homozygous transitions from the reference while in heterozygous variants transitions are typically away from the ancestral allele.

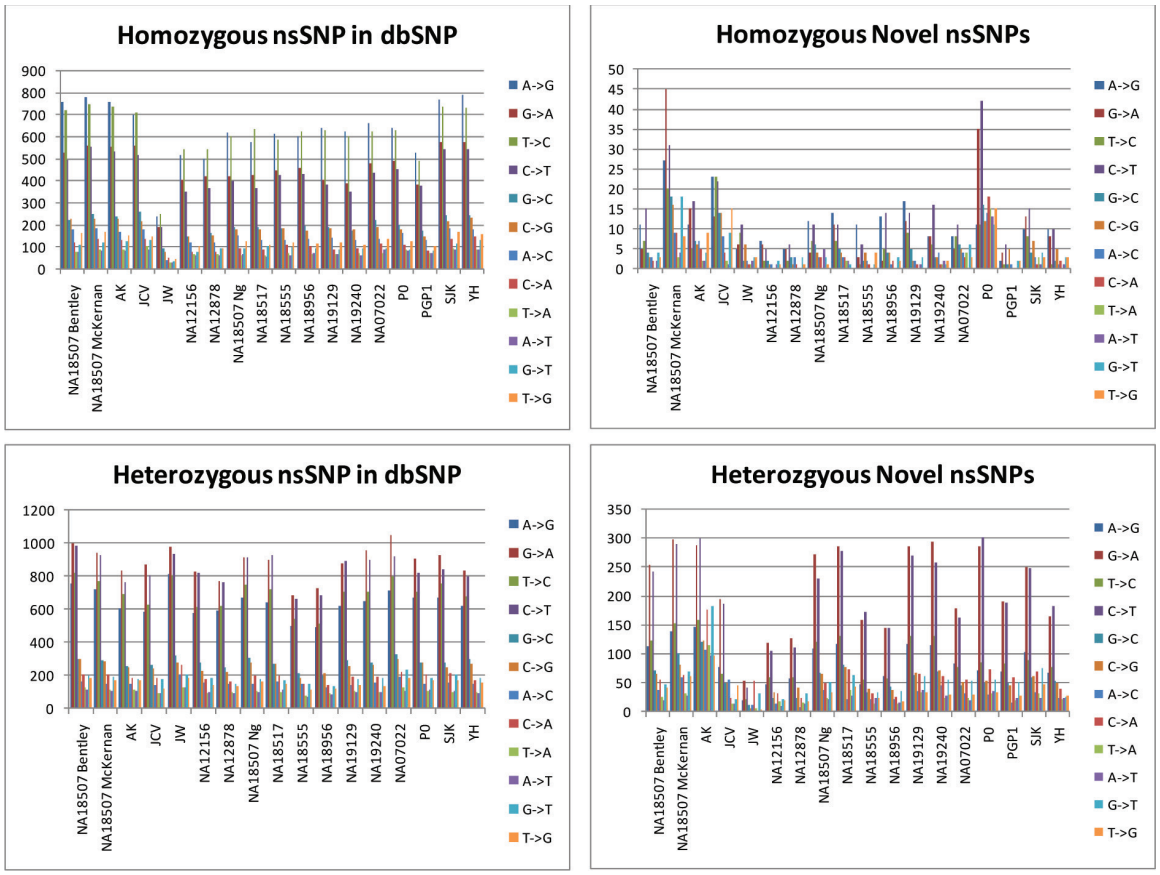


Figure S2-3.3 Detailed Comparison of non-synonymous Single Nucleotide Substitutions between Sequencing Projects. Each chart is as labeled. The consistency between the different genomes and different platforms is remarkable. As seen with the analysis of all single nucleotide substitutions, especially striking is the predominance of the ancestral allele in homozygous transitions from the reference while in heterozygous variants transitions are typically away from the ancestral allele.

Table S2-1. Source Data. Differences between the reported SNPs and obtained SNPs reflect redactions or data released from a different analysis pipeline than the one used for the initial publication.

Genome	Publication	Source	Reported SNPs	Obtained SNPs
JCV	Levy, et al.[113]	http://huref.jcvi.org/	3213401	3074686
JW	Wheeler, et al.[114]	http://jimwatsonsequence.cshl.edu/	3322093	2060544
NA07022	Unpublished	via correspondence	N/A	3077756
PGP1	This paper	N/A	N/A	2966187
YH	Wang, et al.[2]	http://yh.genomics.org.cn/	3074097	3074097
AK	Kim, et al.[115]	via correspondence	3453653	3424779
SJK	Ahn, et al.[116]	ftp://ftp.kobic.kr/pub/KOBIC-KoreanGenome/KOREF_20090224/genetic_variations/	3439107	3439107
NA18507-Bentley	Bentley, et al.[117]	via correspondence	3612498	3612498
NA18507-McKernan	McKernan, et al.[118]	ftp://solidanon:solidanon1mmoorex@ftp1.solidsoftwaretools.com/anonymous/yoruban/Yoruban_var.tar.bz2	3866085	3866085
P0	Pushkarev, et al.[119]	via correspondence	2805471	2794408
NA18507-Ng	Ng, et al.[120]	http://krishna.gs.washington.edu/12_exomes/	28575	28575
NA12156	Ng, et al.[120]	http://krishna.gs.washington.edu/12_exomes/	23457	23457
NA12878	Ng, et al.[120]	http://krishna.gs.washington.edu/12_exomes/	23203	23203
NA18517	Ng, et al.[120]	http://krishna.gs.washington.edu/12_exomes/	28581	28581
NA18555	Ng, et al.[120]	http://krishna.gs.washington.edu/12_exomes/	23258	23258
NA18956	Ng, et al.[120]	http://krishna.gs.washington.edu/12_exomes/	23284	23284
NA19129	Ng, et al.[120]	http://krishna.gs.washington.edu/12_exomes/	28519	28519
NA19240	Ng, et al.[120]	http://krishna.gs.washington.edu/12_exomes/	28247	28247
9.3-PGP1	This paper	N/A	N/A	520
9.3-PGP2	This paper	N/A	N/A	415
9.3-PGP3	This paper	N/A	N/A	1045
9.3-PGP4	This paper	N/A	N/A	1569
9.3-PGP5	This paper	N/A	N/A	401
9.3-PGP6	This paper	N/A	N/A	1262
9.3-PGP7	This paper	N/A	N/A	449
9.3-PGP8	This paper	N/A	N/A	238
9.3-PGP9	This paper	N/A	N/A	1360
9.3-PGP10	This paper	N/A	N/A	551

Table S2-2. Results of processing genomic data by Trait-o-matic

Genome	Substitutions	Non-synonymous	HGMD	SNPedia	OMIM	PharmGKB
JCV	3074686	8895	265	271	60	23
JW	2060544	6831	243	278	63	20
NA07022	3077496	9310	314	359	75	21
PGP1	2966187	8217	260	296	64	22
P0	2794408	9200	280	302	69	22
YH	3074097	9060	267	321	55	18
AK	3424779	10119	257	321	62	16
SJK	3439107	9714	268	310	68	22
NA18507-Bentley	3612498	9995	259	281	62	17
NA18507-McKernan	3866085	10360	245	265	61	16
NA18507-Bentley intersect McKernan	3079069	8725	225	256	56	15
NA18507-Ng	28575	8871	243	89	61	18
NA12156	23457	7338	267	107	71	26
NA12878	23203	7241	269	111	69	32
NA18517	28581	8859	236	94	42	13
NA18555	23258	7215	245	93	50	18
NA18956	23284	7278	235	99	56	18
NA19129	28519	8771	225	88	51	11
NA19240	28247	8739	232	91	59	17
9.3-PGP1	520	183	8	3	2	2
9.3-PGP2	415	191	5	1	2	2
9.3-PGP3	1045	465	11	3	3	1
9.3-PGP4	1569	538	26	10	8	2
9.3-PGP5	401	178	1	0	0	0
9.3-PGP6	1262	499	25	5	11	3
9.3-PGP7	449	208	5	2	0	1
9.3-PGP8	238	89	3	0	0	1
9.3-PGP9	1360	493	25	7	9	3
9.3-PGP10	551	263	6	2	3	0
Union	9174375	39260	1011	769	250	66

Table S2-3. Systematic Manual Processing of HGMD/OMIM Results. The column headers are defined and explained in figure 2-1c and the text accompanying that figure.

Genomes	Rare Variants	Disease Associated	Phenotype Expected	Literature Confirmation	Case Evidence
JCV	44	16	9	5	1
JW	47	20	6	3	2
NA07022	62	12	2	2	1
PGP1	43	10	6	2	0
P0	58	17	3	3	2
YH	46	8	2	1	1
AK	42	9	5	1	0
SJK	38	10	4	2	1
NA18507-Bentley	43	12	3	2	0
NA18507-McKernan	50	11	5	3	1
NA18507-Bentley Intersect McKernan	36	8	2	1	0
NA18507-Ng	49	14	5	3	1
NA12156	53	8	1	0	0
NA12878	37	5	2	1	1
NA18517	49	8	1	1	1
NA18555	35	8	0	0	0
NA18956	43	4	2	1	0
NA19129	47	8	3	2	0
NA19240	37	5	0	0	0
9.3-PGP1	0	0	0	0	0
9.3-PGP2	1	0	0	0	0
9.3-PGP3	1	1	0	0	0
9.3-PGP4	6	2	0	0	0
9.3-PGP5	0	0	0	0	0
9.3-PGP6	5	1	1	1	1
9.3-PGP7	0	0	0	0	0
9.3-PGP8	0	0	0	0	0
9.3-PGP9	3	2	0	0	0
9.3-PGP10	2	1	1	1	0
Union	417	147	49	29	11

Table S2-4. Variants with Potential Phenotypic Expression (29 variants)

Genome	State	Chr	Location/ Gene, Alteration	TAF	Phenotype	Notes on Variants
NA 07022, JCV	Het AD	2	227624091 *COL4A4, G999E*	Unk	Thin membrane basement disease	Found in 3 symptomatic and one possibly symptomatic individual(s)[121-122].
NA 07022, PGP1	Het AD	12	55711185 *MYO1A, S797F*	Unk	Early-onset moderate sensorineural hearing loss	Found in one family affecting father and son.[123]
JCV	Hom AR	3	15661697 *BTD, D444H*	0.018	Biotinidase deficiency	Hom. infant produced trace amounts of enzyme.[19]
JCV	Het AR	3	15660878 *BTD, A171T*	Unk	Biotinidase deficiency	Together with previous mutation this is a common cause for disease.[19] Two adults with similar genotype were found asymptomatic.[124-125]
JCV	Het AD	1	206025069 *CD46, A309V*	Unk	Non-Shiga toxin-associated hemolytic uremic syndrome	This variant is also called A304V in lit. Found in 7 patients with related phenotypes.[126-127] Penetrance is 54%, and expression usually occurs before adulthood.
JW	Het AD	5	137234459 MYOT, Q74K	0.00	Myotilinopathy	Found in two patients, one of which involving weakness of the pectoralis muscles.[128-129]
JW	Het AD	16	1196127 CACNA1H A876T	Unk	Idiopathic epilepsy	Found segregating with disease in 1 family (4 ind) with a variety of phenotypes.[130]
JW PGP1	Het AD	12	119921765 *HNF1A, G574S*	Unk	Hepatic Adenoma	Implicated in one case of hepatic adenoma in an individual with familial diabetes.[131] (OMIM*142410.0013)
PGP1	Het AD	1	221351823 TLR5, R392X	0.035	Impaired ability to generate immune response to flagellated bacteria	Found due to susceptibility to Legionnaires' Disease.[132] It does not prevent immune response to typhus.[133]
PGP1	Het AD	4	6353700 WFS1, C426Y	Unk	Major familial depression	Found in one Caucasian.[134]
NA 18507 – McKernan	X	X	69172053 EDA, A349T	Unk	X-linked hypohidrotic ectodermal dysplasia	Found in two independent patients.[135]
NA 18507–	Het	16	2092866	Unk	Polycystic disease	Found in three unrelated

McKernan Ng	AD		PKD1, E2966D			patients.[136]
PGP10	Het AD	17	38451240 BRCA1, I1858T	Unk	Prostate cancer	Found in early-onset prostate cancer in one family (2 siblings).[137]
NA 18507 – All	Het AD	10	72030654 PRF1, R4H	Unk	Acquired aplastic anemia	Found in one African Individual[138] and OMIM*170280.0013.
NA 18507– Bentley Ng	Het AD	21	34664672 KCNE2, Q9E	0.015	Long QT Syndrome, SIDS	Confers susceptibility to LQTS (OMIM) and was found in a screen for SIDS genes.[139] "Its relatively high frequency may confer arrhythmia susceptibility, particularly during exposure to antibiotics like clarithromycin".[140]
SJK	X	X	153246807 FLNA, V528M	Unk	Bilateral periventricular nodular heterotopias	Found in heterozygous Japanese female, and authors conjecture lethality in males.[141]
SJK	Het AD	1	34999551 GJB4, V37M	Unk	Nonsyndromic deafness	This variant has been implicated in nonsyndromic deafness in 2 unrelated individuals in Taiwan.[142]
AK	Het AD	11	73394991 UCP3, R70W	Unk	Severe obesity and diabetes	Found in two Chinese Individuals [143].
YH	Het AD	9	134770826 TSC1, Q654E	Unk	Tuberous sclerosis complex and cardiac rhabdomyoma	Found in one Japanese and one Korean patient in two studies, but there is a wide phenotypic range.[144-145] Also found in one fetus later born live with cardiac rhabdomyoma [146].
PGP6	Het AD	12	109841347 MYL2, A13T	Unk	Hypertrophic cardiomyopathy	Found in one patient in Poetter et al.,[147] and in two siblings (one a compound het) in Andersen et al. and Hougs et al.[148-149] Three papers report on the molecular basis for the disease.[150-152]
JCV	Het AD	6	7526746 DSP, R1775I	Unk	Arrhythmogenic right ventricular cardiomyopathy	Found in one family one symptomatic and one 52yo asymptomatic[153] female being monitored, and the authors consider it probably pathogenic.[154]
P0	Het AD	15	87671182 POLG, G517V	Unk	Progressive external ophthalmoplegia	Found in 3 generations of one Central European family with varying phenotypes (epilepsy, neuropathy and myopathy,

						mild neuropathy). TAF 0.0015 in German controls.[155]
P0	Het AD	20	3011659 AVP, G96C	Unk	Diabetes insipidus, neurohypophyseal	Found in two publications in at least three families.[156-157]
P0	X-linked	X	100545469 GLA, Q119*	Unk	Fabry disease	First identified in one British individual in Davies et al., and also seen in two families in Ashley et al.[158-159] The mutation occurs in a mutation hotspot identified by Eng et al.[160]
NA12878	Het AD	14	63746504 *SYNE2, T6211M*	Unk	Emery Dreifuss muscular dystrophy	Described as T89M in the literature, found in five patients[161] (OMIM 608442.0001)
NA18517	Com. Het	17	6269533 *AIPL1, P376S;* 6272486 *AIPL1, T114I*	0.20 0.054	Leber congenital amaurosis IV	Initially reported in one African-American child[162-163], later found in in an additional patient.[163]
NA18956	Het AD	9	134762748 *TSC1, T899S*	Unk	Tuberous sclerosis	Initially found in 1 Japanese sporadic female autosomal dominant tuberous sclerosis patient without mental retardation in a screen for TSC1/TSC2 mutations. It was not seen in 100 controls.[144] The authors find another patient[164] but raise doubts due to the rarity of causative missense mutations in TSC1.
NA19129	Hom AD	2	237909660 *COL6A3, A2941V*	Unk	Bethlem myopathy	This variant was found in a screen of autosomal dominant mild bethlem myopathy this was seen in 1 individual heterozygous, and not seen in 156 controls.[165]
NA19129	Hom AD	19	43656115 *RYR1, S1342G*	Unk	Malignant Hyperthermia	This variant was found in study of N. African and Caucasians with dominant malignant hyperthermia (central core disease).[81] Later reports note the frequency in control populations and the need for more research to understand the susceptibility to disease.[82]

Table S2-5. Variants that according to their HGMD citation should have phenotypic expression, but later research (not cited in HGMD) raise doubts as to this conclusion (20 variants).

Genome	State	Chr	Location/ Gene, Alteration	TAF	Phenotype	Notes on Variants
JCV	Hom	1	43576927 MPL, V114M	0.04	Congenital amegakaryocytic thrombocytopenia	Found in one com. hom. with R90X with onset <1yo,[166] but Fox et al.,[167] report it as a benign polymorphism
JW NA12878	Hom	2	219993120 DES, A213V	Unk	Restrictive myopathy and progressive skeletal myopathy	Due to a MAF of 0.02 in controls, Goldfarb et al., postulates that it is low penetrance variant.[168] Kostareva cites new evidence to claim a benign polymorphism.[169]
JW	Het AD	1	214563555 USH2A, E478D	Unk	Usher syndrome type II	Found in >6 patients, but of uncertain pathogenicity,[170] later reports find it in all three forms of the syndrome but also with MAF of 0.017 in Dutch controls.[171]
JW	Het AD	13	31827387 BRCA2 A2466V	0.009	Ovarian Cancer	Initially reported in one ovarian cancer patient[172], it was later reported to be a polymorphism with no cancer risk in YRI population.[173]
JCV	Het AD	11	20579513 SLC6A5, A89E	Unk	Hyperekplexia	Initially reported as such causative,[174] but later reports find control TAF of 0.02 and *in vitro* expression same as wildtype.[175-176]
JCV	Het AD	16	8814415 PMM2, E197A	Unk	Congenital disorders of glycosylation type Ia	Found in Matthijs et al.,[177] but later reports claim that it is benign.[178-179]
JCV	Het AD	5	110455984 WDR36, D33E	Unk	Primary open-angle glaucoma	Found in 8/399 patients but in 1/376 controls it was initially labeled as incompletely

						penetrant,[180] but Fottz et al. emphasize that it is susceptibility factor.[181]
PGP1	Het AD	21	46228730 COL6A1, S116N	Unk	Bethlem myopathy	Found in 2 moderate and one mild case,[165] but later studies find TAF of 0.01 in controls.[182]
PGP1	Het AD	16	67413467 CDH1, A592T	0.005 cancer500	Prostate/papillary thyroid/colorectal cancer	Found in 3 individuals and one family in a screen for cancer-related variants.[183] A later review points out its lack of functional significance and polymorphic frequency[184].
NA12156	Het AD	5	149720925 *TCOF1, A41V*	Unk	Treacher-Collins syndrome	This variant was found in one Treacher-Collins syndrome patient[185], but later found to not segregate with disease in that family.[186]
NA18507-Bentley Ng	Het AD	16	2101118 PKD1, R1351W	Unk	Polycystic disease	Found in 1/82 patients and no controls,[187] but later reports declare it benign.[188]
NA19129	Het AD	19	60359419 TNNI3, P82S	Unk	Elderly-onset hypertrophic cardiomyopathy	Found in 2 patients with onset 52.5 ± 3.6,[189] later reports find TAF of 0.03 in Afro-Caribbean controls.[190]
NA18507-All	Het AD	3	37028554 MLH1, V213M	0.023	Non-polyposis colorectal cancer	Found in a Portuguese family,[191] but three studies cited in Raevaara[192] label it non-pathogenic.
YH	Het AD	13	31812277 BRCA2, I1929V	Unk	Sporadic breast cancer	Found in 1/97 Chinese patients and labeled "unclassified pathogenicity"[193], later found in 4 Korean cases, 1 control and 1 benign breast cancer case[194] and one 51yo Korean male with

						stomach cancer, but still labeled of "unknown significance."[195]
AK	Het AD	17	36993050 KRT14, A413T	Unk	Epidermolysis bullosa simplex	Found in one patient from Taiwan with onset 4-5yo,[196] but later found with control TAF of 0.11.[197]
SJK	Het AD	4	89148477 PKD2, A190T	Unk	Polycystic kidney disease	Found in Zhang et al.,[198] but later reports claim it is benign.[199]
AK	Het AD	7	50709706 GRB10, P95S	Unk	Russel-Silver syndrome	Found in Yoshihashi et al., [200] but doubts have been raised.[201]
AK NA18956	Het AD	10	13218772 OPTN, R545Q	Unk	Open angle glaucoma	Initially found in three individuals,[202] further studies have shown it to be a polymorphism.[203]
AK	Het AD	2	166807404 SCN9A, R1150W	0.049	Primary erythermalgia	Originally reported in one sporadic case,[204] it was later determined to be a polymorphism.[205]
SJK	Hom AR	8	143953695 CYP11B1, A386V	Unk	Congenital adrenal hyperplasia	This variant is listed as causative[206] and was also found in a 9yo Chinese male as a compound het.[207] Merke et al declare it a polymorphism after it is found in 1 control.[208]

Table S2-6a. Diseases with Contributing Variants with Trait-Associated Frequencies > 0.05.

Disease	Variant	Highest Pop-Specific Frequency
Sickle cell disease	Chr11:5204808; HBB, Glu7Val (rs334)	0.125 (OMIM+141900.0243)
Cystic fibrosis	Chr7:116986883:116986885; CFTR, Phe508del (rs332)	0.067 (OMIM*602421.0001)
Hemochromatosis	Chr6:26199158; HFE, His63Asp (rs1799945); HFE Cys282Tyr	0.157 (OMIM+235200)
Factor V Leiden	Chr1:167785673; F5, Arg506Gln (rs6025)	0.07 (OMIM#612309)
Gaucher's disease	Chr1:153472257; GBA, Asn370Ser	0.067 (OMIM#231000)
Tay Sach's disease	Chr15:70425975:70425975; HEXA, 1277insTATC	0.08 (OMIM#272800)
Alzheimer's disease (ApoE4)	Chr19:50103781; APOE, Cys130Arg (rs429358) & Chr19:50103919 APOE, Arg176Cys (rs7412)	0.41 (OMIM+107741)
Age-related macular degeneration	Chr1:194908856; CFH, Ile62Val (rs800292) & Chr1:194925860; CFH Tyr402His (rs1061170)	0.192, 0.182 (HapMap)

Table S2-6b. Common Disease Variants in our Genomes

Genome	State	Chr	Location/ Gene, Alteration	TAF	Phenotype
PGP1 P0	Het AR	6	26199158 HFE, His63Asp	0.157	HFE-associated hemochromatosis has low penetrance even with homozygosity for C282Y, the common variant. H63D associated disease has even lower penetrance, with the majority of disease only being present when compound het with C282Y
NA12156 NA12878 AK JW P0	Het AR	1	194908856 CFH, V62I	0.192	
NA18507 -All NA18517 NA18555 NA18956 NA19129 NA19240	Hom AR	1	194908856 CFH, V62I	0.192	
YH JCV NA07022 JW	Het AR	1	194925860 CFH, Y402H	0.182	The heterozygote has a ~30% risk of developing ARMD, while the homozygote has a 53% risk

JW	Het AR	1	167785673; F5, Arg506Gln	0.07	Note: A is the causative allele and the reference allele, so homozygous causative genotypes will not be noted in variant lists
NA07022	Het AR	7	116986881:116986884 del CTT	unk	This novel deletion overlaps the CFTR deletion by 2bp and introduces a fs in the protein.
NA19240	Het AR	11	5204808 HBB E7V	0.125	Carrier for Sickle Cell Anemia Trait

Table S2-7. **High Odds Ratio GWAS variants**.

Genome	Position	Geno -type	SNPedia Link	Phenotype
NA18507 JCV PGP1	chr1:118766777	A/C	(http://www.snpedia.com/index.php/Rs2145418)	5.0x increased thyroid cancer risk
NA18507	chr7:87002922	G/G	(http://www.snpedia.com/index.php/Rs10248420)	7x more likely to respond to certain antidepressants
SJK	chr7:87002922	A/G	(http://www.snpedia.com/index.php/Rs10248420)	7x more likely to respond to certain antidepressants
NA18507	chr7:87037500	A/C	(http://www.snpedia.com/index.php/Rs2235015)	7x more likely to respond to certain antidepressants
NA18507	chr8:130015336	A/A	(http://www.snpedia.com/index.php/Rs987525)	6x increased risk for cleft lip
YH, P0, SJK	chr10:124145371	T/T	(http://www.snpedia.com/index.php/Rs4146894)	9.1x risk for AMD
YH, SJK	chr10:124204438	T/T	(http://www.snpedia.com/index.php/Rs10490924)	8.2x risk for AMD
YH, SJK	chr10:124210534	A/A	(http://www.snpedia.com/index.php/Rs11200638)	~10x risk for wet AMD
NA07022	chr6:31248026	A/G	(http://www.snpedia.com/index.php/Rs1265159)	~5x risk for psoriasis
NA07022	chr6:31263764	C/G	(http://www.snpedia.com/index.php/Rs1265181)	~5 risk for psoriasis among Chinese
NA07022	chr6:32712350	A/A	(http://www.snpedia.com/index.php/Rs9272346)	18.5x risk for type-1 diabetes
P0, JW, SJK	chr6:32712350	A/G	(http://www.snpedia.com/index.php/Rs9272346)	5.5x risk for type-1 diabetes
JCV	chr6:31248026	A/A	(http://www.snpedia.com/index.php/Rs1265159)	22x risk for psoriasis
JCV	chr6:31263764	C/C	(http://www.snpedia.com/index.php/Rs1265181)	22x risk for psoriasis among Chinese
NA18555 NA18956	chr7:86998554	A/A	http://www.snpedia.com/index.php/Rs2032582	6.7x risk for Crohn's disease
NA19129 NA19240	chr1: 194925860	C/C	http://www.snpedia.com/index.php/Rs1061170	5.9x risk for AMD; higher mortality among nonagenarians

Table S2-8 Drug Dosage Variants see http://www.pharmgkb.org/ for sources and details

Genome	Position	Genotype	Drug	Phenotype
NA18507, NA07022, NA12156, NA12878, NA18517, NA18956, NA19240	chr1:12175542 TNFRSF1B, M196R	G/T	infliximab	low response
NA18507, JW, AK, PGP1, NA19129	chr1:98121473 DPYD, R29C	A/G	fluoruracil	increased incidence of nauseau
NA18507, NA07022, P0, JW, NA12156, NA12878, NA18517, NA18555	chr11:67109265 GSTP1, I105V	A/G	fluoruracil	recessive for hematological toxicity
NA18507, YH, AK, SJK, PGP1, NA18555, NA19240	chr11:74561225 SLCO2B1, R168Q	A/G	montelukast	poor response
NA18507, P0, JCV, AK, NA18517, NA18555, NA19129, NA19240	chr12:21221005 SLCO1B1, N130D	G/G	pravastatin	low plasma AUC, but no change in *in-vitro* transport
SJK, PGP1, NA12156, NA12878	chr12:21221005 SLCO1B1, N130D	A/G	pravastatin	low plasma AUC, but no change in *in-vitro* transport
NA18507	chr19:46210061 CYP2B6, I328T	C/T	bupropion efavirenz nevirapine	elevated plasma levels of efavirenz
YH, JCV, AK, NA12878	chr1:97753983 DPYD, I543V	C/T	fluoruracil	increased incidence of nauseau
SJK, NA12156	chr1:97753983 DPYD, I543V	C/C	fluoruracil	increased incidence of nauseau
YH, SJK, NA12156,	chr2:169719231 LRP2, K4094E	C/T	Cisplatin	associated with hearing loss

NA18555, NA18956				
YH, AK, SJK, NA18555, NA18956	chr4:89271347 ABCG2, Q141K	G/T	diflomotecan; rosuvastatin	Signicantly altered kinetics; increased plasma AUC
YH, JW, NA18555, NA18956	chr6:154402490 OPRM1, N40D	A/G	ethanol; naltrexone	better clinical outcome
YH, P0, JCV, SJK, NA12878, NA18555, NA18956	chr10:115795046 ADRB1, G389R	C/C	atenolol; verapamil	better outcome from treatment with atenolol vs. verapamil
JW, NA12156, NA19240	chr10:115795046 ADRB1, G389R	C/G	atenolol; verapamil	better outcome from treatment with atenolol vs. verapamil
NA07022, NA12156, NA12878, NA19129, NA19240	chr1:159781166 FCGR3A, F176V	A/C	cetuximab; rituximab	better response
NA07022, P0, JW, PGP1	chr2:21117405 APOB, T98I	A/G	atenolol; irbestartan	irbestartan more effective than atenolol in lowering blood pressure
NA07022, P0, JCV, JW, AK, SJK, NA12156, NA12878, NA19240	chr5:176452849 FGFR4, G388R	A/G	cisplatin; cyclo-phosphamide; fluorouracil; methotrexate; tamoxifen	possible association with sensitivity to cisplatin; poor outcome
NA07022, JW	chr6:31886251 HSPA1L, T493M	A/G	carbamazepine	protection from hypersensitivity
NA07022, NA18956	chr10:115794026 ADRB1, S49G	A/G	atenolol; verapamil	better outcome from treatment with atenolol vs. verapamil
NA07022, P0, JCV, JW	chr17:35133114 ERBB2, I655V	A/G	trastuzumab	associated with cardiac toxicity
NA07022, JW, NA19129	chr19:46204681 CYP2B6, Q172H	G/T	efavirenz	part of *6 haplotype with low mean plasma levels
NA07022, JCV,	chr19:50103781 APOE, C130R	C/T	acitretin	APOE e4 not associated with

NA12156, NA18517, NA19129				response
NA07022, JCV	chr19:50103919 APOE, R176C C/T	C/T	Acitretin	APOE e4 not associated with response
NA07022, JCV, NA12878	chr19:50546759 ERCC2, K751Q	G/T	fluoruracil; leucovorin	increase risk of early relapse in Asian patients
P0, NA12156	chr10:101553805 ABCC2, V417I	A/G	Talinolol	increased clearance
P0, NA12878	chr11:67110155 GSTP1, A114V	C/T	Thiotepa	homozygotes may have increased exposure
P0, NA12156, NA12878	chr15:76669980 CHRNA5, D398N	A/G	Nicotine	associated with early-onset addiction
AK, NA12878	chr12:21222816 SLCO1B1, V174A	C/T	repaglinide	increased plasma AUC
SJK	chr10:96530400 CYP2C19, W212*	A/G	copidogrel	associated with lower levels and higher rate of adverse cardiovascular events (**FDA Approved**)
SJK, PGP1, PGP2, PGP6, PGP9, NA12156, NA18517	chr19:15851431 CYP4F2, V433M	C/T	Warfarin	recessive for requiring 1mg/day more
PGP4	chr19:15851431 CYP4F2, V433M	T/T	Warfarin	require 1mg/day more
PGP1, NA12878	chr10:96692037 CYP2C9, R144C	C/T	fluvastatin; glipizide; phenytoin; tolbutamide; warfarin	require lower dosage of warfarin; affects clearance of other drugs
PGP1, NA12878	chr10:96788739 CYP2C8, K399R	C/T	paclitaxel; repaglinide; rosiglitazone	impaired metabolism *in vitro*; lower plasma concentration of repaglinide and rosiglitazone
PGP1, NA12878	chr10:96817020 CYP2C8, R139K	C/T	paclitaxel; repaglinide; rosiglitazone	impaired metabolism *in vitro*; lower plasma concentration of repaglinide and rosiglitazone

NA12156	chr12: 123914216 *SCARB1, G2S*	C/T	fenofibrate	Higher responsiveness
NA19240	chr10:96808096 *CYP2C8, I269F*	A/T	paclitaxel	Recessive for allele with 2x lower clearance

Table S2-9. **ApoE status for the 9 full genomes**. Zygosity not known (ZNK) depicts no variant called at that position, but we cannot be certain that the call is reference without re-placing the reads from the short read archive. The "reference" frequency for each of these two positions is ~10%.[209] The binomial probability of 22/28 alleles sampled matching reference is p= 2.00e-17. PGP1 has released a file depicting every sequenced position, and this sequence was not covered, and JW has requested that this data be redacted.

Genome	rs429358	rs7412
NA07022	C/T	C/T
JCV	ZNK	C/T
PGP1	Not covered	Not covered
JW	Redacted	Redacted
AK	ZNK	ZNK
SJK	ZNK	ZNK
YH	ZNK	ZNK
NA18507	ZNK	ZNK
P0	ZNK	ZNK
NA12156	ZNK	C/T
NA12878	ZNK	ZNK
NA18517	ZNK	C/T
NA18555	ZNK	ZNK
NA18956	ZNK	ZNK
NA19129	ZNK	C/T
NA19240	ZNK	ZNK

Table S2-10a. Variants Implicated in Disease with Unreported Frequencies, but Appearing Frequently in YRI Genomes

Genome	State	Location/ Gene, Alteration	TAF	Phenotype	Case; controls Notes
NA18507-all NA19129 NA19240	Het AD	Chr4: 88752564 DSPP, Arg68Trp	Unk	Dentinogenesis imperfecta type II	14; 0/42 Found in a Swedish family segregating with disease;[210] reviewed by Kim et al., who reports additional cases.[211]
NA18507-all NA19129 NA19240	Hom AD	chr19:15152576 NOTCH3, Ala1020Pro	Unk	Cerebral arteriopathy with subcortical infarcts and leukoencephalopathy	4; 0/100 Found in four patients of unknown ethnicity, one of whom diagnosed at 77yo.[212]
NA18507-all NA18517	Het AD	Chr6: 134252293 TCF21, G22V	Unk	Dilated cardiomyopathy	Found in 12yo female, with mother symptomatic for DCM and grandmother with sensorineural hearing loss.[213]
NA18507-all NA18517 NA19240	Het AD	Chr4: 5806425 EVC, R443Q	Unk	Ellis-van Creveld syndrome	Although this syndrome is usually inherited recessively, this was found dominant in an Amish family (father-daughter).[214]

Table S2-10b. NA18507 Variants Found by Only One Group

Genome	State	Location/ Gene, Alteration	TAF	Phenotype	Notes
NA18507 McKernan	Hom	ChrX: 69172053 EDA, Ala349Thr Yes	Unk	X-linked hypohidrotic ectodermal dysplasia	Abnormal development of hair, teeth and eccrine sweat glands; if dental development normal assumed to be a sequencing error[135].
NA18507 McKernan	Het AD	chr19: 60359419 TNNI3, P82S	Unk	Elderly-onset hypertrophic cardiomyopathy	Found in 2 patients with onset 52.5 ± 3.6[189], later reports find TAF of 0.03 in Afro-Caribbean controls[190].
NA18507 Ng	Het	chr10:115795046 ADRB1, G389R		Pharmacogenetic	PharmGKB: Better outcome from treatment with atenolol vs. verapamil
NA18507 Ng	Het	chrX:152661677 ABCD1, G608D		Adreno-leukodystrophy	Seen in one European boy, not seen in 100 control chromosomes[215].
NA18507 Ng	Het	chr1:55296443 PCSK9, A443T		Hyper-cholesterolemia	Found in one individual, but the variant was asymptomatic in his daughter; not found in 340 controls - likely a benign but rare variant[216].

Table S2-11: Consensus between Illumina (Bentley) and SOLiD (McKernan) sequencing

Variant	Consensus	Illumina Only	SOLiD only	% Consensus
All Substitutions	3079069	533429	787016	0.699866
Transcription	1490065	248452	356757	0.711155
Exon + intron	1291271	214043	304876	0.713335
Exons	28620	4628	6283	0.723989
Non-synonymous Substitutions	8725	1270	1635	0.750215
In OMIM/HGMD Database	234	35	20	0.809688
TAF < 0.05	36	7	14	0.631579
Phenotype Expected	6	1	4	0.545455
Case Evidence	1	0	2	0.333333

Table S2-12a: Variation Frequency in OMIM Genes. Significance is calculated using the Fisher Exact Two-Tailed test.

Variant Type	% in OMIM	Expected	Odds Ratio	Significance
Non-synonymous Substitutions	10.73%	12.60%	0.85	2.87E-50
Novel non-synonymous Substitutions	9.87%	12.60%	0.78	2.31E-50
Nonsense Substitutions	8.47%	12.60%	0.67	2.23E-03
Insertions/Deletions	8.95%	12.60%	0.71	3.02E-04
Frameshifts	5.46%	12.60%	0.43	1.55E-05
All	10.32%	12.60%	0.82	1.36E-104

Table S2-12b: Variation Frequency in Genetests Genes. Significance is calculated using the Fisher Exact Two-Tailed test.

Variant Type	% in Genetests	Expected	Odds Ratio	Significance
Non-synonymous Substitutions	7.26%	9.29%	0.78	3.13E-50
Novel non-synonymous Substitutions	6.72%	9.29%	0.72	5.12E-46
Nonsense Substitutions	5.94%	9.29%	0.64	8.18E-03
Insertions/Deletions	6.22%	9.29%	0.67	1.64E-03
Frameshifts	4.04%	9.29%	0.43	3.89E-04
All	7.01%	9.29%	0.75	1.70E-99

References

1 Kuhn, R. M. *et al.* The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res* **37**, D755-761, doi:gkn875 [pii]
10.1093/nar/gkn875 (2009).

2 Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60-65, doi:nature07484 [pii]
10.1038/nature07484 (2008).

3 Porreca, G. J. *et al.* Multiplex amplification of large sets of human exons. *Nat Methods* **4**, 931-936, doi:nmeth1110 [pii]
10.1038/nmeth1110 (2007).

4 Li, J. B. *et al.* Multiplex padlock targeted sequencing reveal human hypermutable CpG variations. *Genome Res*, doi:gr.092213.109 [pii]
10.1101/gr.092213.109 (2009).

5 Villani, G. R., Pontarelli, G., Vitale, D. & Di Natale, P. Gene symbol: NAGLU. Disease: Sanfillipo syndrome B. *Hum Genet* **115**, 173 (2004).

6 Assink, K. *et al.* Mutation analysis and clinical implications of von Willebrand factor-cleaving protease deficiency. *Kidney Int* **63**, 1995-1999, doi:kid24 [pii]
10.1046/j.1523-1755.63.6s.1.x (2003).

7 Mykytyn, K. *et al.* Identification of the gene (BBS1) most commonly involved in Bardet-Biedl syndrome, a complex human obesity syndrome. *Nat Genet* **31**, 435-438, doi:10.1038/ng935
ng935 [pii] (2002).

8 Wang, J. *et al.* Inherited human Caspase 10 mutations underlie defective lymphocyte and dendritic cell apoptosis in autoimmune lymphoproliferative syndrome type II. *Cell* **98**, 47-58, doi:S0092-8674(00)80605-4 [pii]
10.1016/S0092-8674(00)80605-4 (1999).

9 Gross, E. *et al.* Detailed analysis of five mutations in dihydropyrimidine dehydrogenase detected in cancer patients with 5-fluorouracil-related side effects. *Hum Mutat* **22**, 498, doi:10.1002/humu.9201 (2003).

10 Rootwelt, H., Brodtkorb, E. & Kvittingen, E. A. Identification of a frequent pseudodeficiency mutation in the fumarylacetoacetase gene, with implications for diagnosis of tyrosinemia type I. *Am J Hum Genet* **55**, 1122-1127 (1994).

11 Ritis, K. *et al.* Non-isotopic RNase cleavage assay for mutation detection in MEFV, the gene responsible for familial Mediterranean fever, in a cohort of Greek patients. *Ann Rheum Dis* **63**, 438-443 (2004).

12 Furuki, S. *et al.* Mutations in the peroxin Pex26p responsible for peroxisome biogenesis disorders of complementation group 8 impair its stability, peroxisomal localization, and interaction with the Pex1p x Pex6p complex. *J Biol Chem* **281**, 1317-1323, doi:M510044200 [pii]

10.1074/jbc.M510044200 (2006).

13 Tag, C. G. *et al.* Analysis of the transforming growth factor-beta1 (TGF-beta1) codon 25 gene polymorphism by LightCycler-analysis in patients with chronic hepatitis C infection. *Cytokine* **24**, 173-181, doi:S1043466603003077 [pii] (2003).

14 Cotarelo, R. P. *et al.* Two new patients bearing mutations in the fukutin gene confirm the relevance of this gene in Walker-Warburg syndrome. *Clin Genet* **73**, 139-145, doi:CGE936 [pii]
10.1111/j.1399-0004.2007.00936.x (2008).

15 Wycisk, K. A. *et al.* Mutation in the auxiliary calcium-channel subunit CACNA2D4 causes autosomal recessive cone dystrophy. *Am J Hum Genet* **79**, 973-977, doi:S0002-9297(07)60841-6 [pii]
10.1086/508944 (2006).

16 Otto, E. *et al.* A gene mutated in nephronophthisis and retinitis pigmentosa encodes a novel protein, nephroretinin, conserved in evolution. *Am J Hum Genet* **71**, 1161-1167, doi:S0002-9297(07)60408-X [pii]
10.1086/344395 (2002).

17 Lemmink, H. H. *et al.* Novel mutations in the thiazide-sensitive NaCl cotransporter gene in patients with Gitelman syndrome with predominant localization to the C-terminal domain. *Kidney Int* **54**, 720-730, doi:10.1046/j.1523-1755.1998.00070.x (1998).

18 Dagher, H., Yan Wang, Y., Fassett, R. & Savige, J. Three novel COL4A4 mutations resulting in stop codons and their clinical effects in autosomal recessive Alport syndrome. *Hum Mutat* **20**, 321-322, doi:10.1002/humu.9065 (2002).

19 Norrgard, K. J. *et al.* Double mutation (A171T and D444H) is a common cause of profound biotinidase deficiency in children ascertained by newborn screening the the United States. Mutations in brief no. 128. Online. *Hum Mutat* **11**, 410, doi:10.1002/(SICI)1098-1004(1998)11:5<410::AID-HUMU10>3.0.CO;2-8 [pii]
10.1002/(SICI)1098-1004(1998)11:5<410::AID-HUMU9>3.0.CO;2-Q (1998).

20 Ng, P. C. *et al.* Genetic variation in an individual human exome. *PLoS Genet* **4**, e1000160, doi:10.1371/journal.pgen.1000160 (2008).

21 Di Leo, E. *et al.* Mutations in MTP gene in abeta- and hypobeta-lipoproteinemia. *Atherosclerosis* **180**, 311-318, doi:S0021-9150(04)00634-3 [pii]
10.1016/j.atherosclerosis.2004.12.004 (2005).

22 Yasuda, M. *et al.* Fabry disease: characterization of alpha-galactosidase A double mutations and the D313Y plasma enzyme pseudodeficiency allele. *Hum Mutat* **22**, 486-492, doi:10.1002/humu.10275 (2003).

23 Katsanis, N. *et al.* Triallelic inheritance in Bardet-Biedl syndrome, a Mendelian recessive disorder. *Science* **293**, 2256-2259, doi:10.1126/science.1063525
293/5538/2256 [pii] (2001).

24 Zhang, Z. X. *et al.* A second, large deletion in the HEXB gene in a patient with infantile Sandhoff disease. *Hum Mol Genet* **4**, 777-780 (1995).

25 Boivin, P. *et al.* Spectrin alpha IIa variant in dominant and non-dominant spherocytosis. *Hum Genet* **92**, 153-156 (1993).

26 Schnitzler, F. *et al.* Eight novel CARD15 variants detected by DNA sequence analysis of the CARD15 gene in 111 patients with inflammatory bowel disease. *Immunogenetics* **58**, 99-106, doi:10.1007/s00251-005-0073-2 (2006).

27 Pantanetti, P., Arnaldi, G., Balercia, G., Mantero, F. & Giacchetti, G. Severe hypomagnesaemia-induced hypocalcaemia in a patient with Gitelman's syndrome. *Clin Endocrinol (Oxf)* **56**, 413-418, doi:1223 [pii] (2002).

28 Wijker, M. *et al.* Heterogeneous spectrum of mutations in the Fanconi anaemia group A gene. *Eur J Hum Genet* **7**, 52-59, doi:10.1038/sj.ejhg.5200248 (1999).

29 Ichida, K., Matsumura, T., Sakuma, R., Hosoya, T. & Nishino, T. Mutation of human molybdenum cofactor sulfurase gene is responsible for classical xanthinuria type II. *Biochem Biophys Res Commun* **282**, 1194-1200, doi:10.1006/bbrc.2001.4719
S0006-291X(01)94719-9 [pii] (2001).

30 Caciotti, A. *et al.* Role of beta-galactosidase and elastin binding protein in lysosomal and nonlysosomal complexes of patients with GM1-gangliosidosis. *Hum Mutat* **25**, 285-292, doi:10.1002/humu.20147 (2005).

31 Muller, J. S. *et al.* Phenotypical spectrum of DOK7 mutations in congenital myasthenic syndromes. *Brain* **130**, 1497-1506, doi:awm068 [pii]
10.1093/brain/awm068 (2007).

32 McLaughlin, M. E., Ehrhart, T. L., Berson, E. L. & Dryja, T. P. Mutation spectrum of the gene encoding the beta subunit of rod phosphodiesterase among patients with autosomal recessive retinitis pigmentosa. *Proc Natl Acad Sci U S A* **92**, 3249-3253 (1995).

33 Rossi, A. & Superti-Furga, A. Mutations in the diastrophic dysplasia sulfate transporter (DTDST) gene (SLC26A2): 22 novel mutations, mutation review, associated skeletal phenotypes, and diagnostic relevance. *Hum Mutat* **17**, 159-171, doi:10.1002/humu.1 [pii]
10.1002/humu.1 (2001).

34 Huang, C. H. *et al.* Molecular basis for Rh(null) syndrome: identification of three new missense mutations in the Rh50 glycoprotein gene. *Am J Hematol* **62**, 25-32, doi:10.1002/(SICI)1096-8652(199909)62:1<25::AID-AJH5>3.0.CO;2-K [pii] (1999).

35 Konrad, M. *et al.* Mutations in the chloride channel gene CLCNKB as a cause of classic Bartter syndrome. *J Am Soc Nephrol* **11**, 1449-1459 (2000).

36 Fukuyama, S., Hiramatsu, M., Akagi, M., Higa, M. & Ohta, T. Novel mutations of the chloride channel Kb gene in two Japanese patients clinically diagnosed as Bartter syndrome with hypocalciuria. *J Clin Endocrinol Metab* **89**, 5847-5850, doi:89/11/5847 [pii]

10.1210/jc.2004-0775 (2004).

37 Oglesbee, D. *et al.* Development of a newborn screening follow-up algorithm for the diagnosis of isobutyryl-CoA dehydrogenase deficiency. *Genet Med* **9**, 108-116, doi:10.1097/GIM.0b013e31802f78d6

00125817-200702000-00008 [pii] (2007).

38 Domenech, E., Gomez-Zaera, M. & Nunes, V. WFS1 mutations in Spanish patients with diabetes mellitus and deafness. *Eur J Hum Genet* **10**, 421-426, doi:10.1038/sj.ejhg.5200823 (2002).

39 Eckl, K. M. *et al.* Mutation spectrum and functional analysis of epidermis-type lipoxygenases in patients with autosomal recessive congenital ichthyosis. *Hum Mutat* **26**, 351-361, doi:10.1002/humu.20236 (2005).

40 Eden, E. R. *et al.* Restoration of LDL receptor function in cells from patients with autosomal recessive hypercholesterolemia by retroviral expression of ARH1. *J Clin Invest* **110**, 1695-1702, doi:10.1172/JCI16445 (2002).

41 Allikmets, R. *et al.* A photoreceptor cell-specific ATP-binding transporter gene (ABCR) is mutated in recessive Stargardt macular dystrophy. *Nat Genet* **15**, 236-246, doi:10.1038/ng0397-236 (1997).

42 Lewis, R. A. *et al.* Genotype/Phenotype analysis of a photoreceptor-specific ATP-binding cassette transporter gene, ABCR, in Stargardt disease. *Am J Hum Genet* **64**, 422-434, doi:S0002-9297(07)61748-0 [pii]

10.1086/302251 (1999).

43 Mardy, S. *et al.* Congenital insensitivity to pain with anhidrosis: novel mutations in the TRKA (NTRK1) gene encoding a high-affinity receptor for nerve growth factor. *Am J Hum Genet* **64**, 1570-1579, doi:S0002-9297(07)63659-3 [pii]

10.1086/302422 (1999).

44 Calonge, M. J. *et al.* Cystinuria caused by mutations in rBAT, a gene involved in the transport of cystine. *Nat Genet* **6**, 420-425, doi:10.1038/ng0494-420 (1994).

45 Lavedan, C., Buchholtz, S., Nussbaum, R. L., Albin, R. L. & Polymeropoulos, M. H. A mutation in the human neurofilament M gene in Parkinson's disease that suggests a role for the cytoskeleton in neuronal degeneration. *Neurosci Lett* **322**, 57-61, doi:S0304394001025137 [pii] (2002).

46 van Tintelen, J. P. *et al.* Plakophilin-2 mutations are the major determinant of familial arrhythmogenic right ventricular dysplasia/cardiomyopathy. *Circulation* **113**, 1650-1658, doi:CIRCULATIONAHA.105.609719 [pii]

10.1161/CIRCULATIONAHA.105.609719 (2006).

47 Santoro, A. *et al.* Novel Munc13-4 mutations in children and young adult patients with haemophagocytic lymphohistiocytosis. *J Med Genet* **43**, 953-960, doi:jmg.2006.041863 [pii]

10.1136/jmg.2006.041863 (2006).

48 Lee, S. T. *et al.* Mutations of the P gene in oculocutaneous albinism, ocular albinism, and Prader-Willi syndrome plus albinism. *N Engl J Med* **330**, 529-534 (1994).

49 Kawai, M. *et al.* A patient with subclinical oculocutaneous albinism type 2 diagnosed on getting severely sunburned. *Dermatology* **210**, 322-323, doi:DRM2005210004322 [pii]
10.1159/000084758 (2005).

50 Sviderskaya, E. V. *et al.* Complementation of hypopigmentation in p-mutant (pink-eyed dilution) mouse melanocytes by normal human P cDNA, and defective complementation by OCA2 mutant sequences. *J Invest Dermatol* **108**, 30-34, doi:S0022202X9781932X [pii] (1997).

51 Yuasa, I. *et al.* OCA2 481Thr, a hypofunctional allele in pigmentation, is characteristic of northeastern Asian populations. *J Hum Genet* **52**, 690-693, doi:10.1007/s10038-007-0167-9 (2007).

52 Duffy, D. L. *et al.* A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. *Am J Hum Genet* **80**, 241-252, doi:S0002-9297(07)62682-2 [pii]
10.1086/510885 (2007).

53 Godfrey, C. *et al.* Refining genotype phenotype correlations in muscular dystrophies with defective glycosylation of dystroglycan. *Brain* **130**, 2725-2735, doi:awm212 [pii]
10.1093/brain/awm212 (2007).

54 Jimenez-Mallebrera, C. *et al.* A Comparative Study of alpha-Dystroglycan Glycosylation in Dystroglycanopathies Suggests that the Hypoglycosylation of alpha-Dystroglycan Does Not Consistently Correlate with Clinical Severity. *Brain Pathol*, doi:BPA198 [pii]
10.1111/j.1750-3639.2008.00198.x (2008).

55 Khaliq, S. *et al.* Novel association of RP1 gene mutations with autosomal recessive retinitis pigmentosa. *J Med Genet* **42**, 436-438, doi:42/5/436 [pii]
10.1136/jmg.2004.024281 (2005).

56 Pang, C. P. & Lam, D. S. Differential occurrence of mutations causative of eye diseases in the Chinese population. *Hum Mutat* **19**, 189-208, doi:10.1002/humu.10053 [pii]
10.1002/humu.10053 (2002).

57 Baum, L. *et al.* ABCA4 sequence variants in Chinese patients with age-related macular degeneration or Stargardt's disease. *Ophthalmologica* **217**, 111-114, doi:10.1159/000068553 OPH17111 [pii] (2003).

58 Paloma, E., Martinez-Mir, A., Vilageliu, L., Gonzalez-Duarte, R. & Balcells, S. Spectrum of ABCA4 (ABCR) gene mutations in Spanish patients with autosomal recessive macular dystrophies. *Hum Mutat* **17**, 504-510, doi:10.1002/humu.1133 [pii]
10.1002/humu.1133 (2001).

59 Dreyer, B. *et al.* Identification of novel USH2A mutations: implications for the structure of USH2A protein. *Eur J Hum Genet* **8**, 500-506, doi:10.1038/sj.ejhg.5200491 (2000).

60 Hashemzadeh Chaleshtori, M. *et al.* Novel mutations in the pejvakin gene are associated with autosomal recessive non-syndromic hearing loss in Iranian families. *Clin Genet* **72**, 261-263, doi:CGE852 [pii]
10.1111/j.1399-0004.2007.00852.x (2007).

61 Min, J. L., Meulenbelt, I., Kloppenburg, M., van Duijn, C. M. & Slagboom, P. E. Mutation analysis of candidate genes within the 2q33.3 linkage area for familial early-onset generalised osteoarthritis. *Eur J Hum Genet* **15**, 791-799, doi:5201829 [pii]
10.1038/sj.ejhg.5201829 (2007).

62 Walter, J. W. *et al.* Somatic mutation of vascular endothelial growth factor receptors in juvenile hemangioma. *Genes Chromosomes Cancer* **33**, 295-303, doi:10.1002/gcc.10028 [pii] (2002).

63 Nowacki, P. M., Byck, S., Prevost, L. & Scriver, C. R. PAH Mutation Analysis Consortium Database: 1997. Prototype for relational locus-specific mutation databases. *Nucleic Acids Res* **26**, 220-225, doi:gkb038 [pii] (1998).

64 Ouyang, X. M. *et al.* Characterization of Usher syndrome type I gene mutations in an Usher syndrome patient population. *Hum Genet* **116**, 292-299, doi:10.1007/s00439-004-1227-2 (2005).

65 Longui, C. A. *et al.* Inhibin alpha-subunit (INHA) gene and locus changes in paediatric adrenocortical tumours from TP53 R337H mutation heterozygote carriers. *J Med Genet* **41**, 354-359 (2004).

66 Colley, J. *et al.* Rapid recognition of aberrant dHPLC elution profiles using the Transgenomic Navigator software. *Hum Mutat* **26**, 165, doi:10.1002/humu.9354 (2005).

67 Bolino, A. *et al.* Denaturing high-performance liquid chromatography of the myotubularin-related 2 gene (MTMR2) in unrelated patients with Charcot-Marie-Tooth disease suggests a low frequency of mutation in inherited neuropathy. *Neurogenetics* **3**, 107-109 (2001).

68 Spritz, R. A. *et al.* Novel mutations of the P gene in type II oculocutaneous albinism (OCA2). *Hum Mutat* **10**, 175-177, doi:10.1002/(SICI)1098-1004(1997)10:2<175::AID-HUMU12>3.0.CO;2-X [pii]
10.1002/(SICI)1098-1004(1997)10:2<175::AID-HUMU12>3.0.CO;2-X (1997).

69 Kerr, R. *et al.* Identification of P gene mutations in individuals with oculocutaneous albinism in sub-Saharan Africa. *Hum Mutat* **15**, 166-172, doi:10.1002/(SICI)1098-1004(200002)15:2<166::AID-HUMU5>3.0.CO;2-Z [pii]
10.1002/(SICI)1098-1004(200002)15:2<166::AID-HUMU5>3.0.CO;2-Z (2000).

70 Kalb, R. *et al.* Hypomorphic mutations in the gene encoding a key Fanconi anemia protein, FANCD2, sustain a significant group of FA-D2 patients with severe phenotype. *Am J Hum Genet* **80**, 895-910, doi:S0002-9297(07)60945-8 [pii]
10.1086/517616 (2007).

71 Richard, I. *et al.* Calpainopathy-a survey of mutations and polymorphisms. *Am J Hum Genet* **64**, 1524-1540, doi:S0002-9297(07)63655-6 [pii]

10.1086/302426 (1999).

72 Naeem, M., Wajid, M., Lee, K., Leal, S. M. & Ahmad, W. A mutation in the hair matrix and cuticle keratin KRTHB5 gene causes ectodermal dysplasia of hair and nail type. *J Med Genet* **43**, 274-279, doi:43/3/274 [pii]

10.1136/jmg.2005.033381 (2006).

73 Lacombe, A. *et al.* Disruption of POF1B binding to nonmuscle actin filaments is associated with premature ovarian failure. *Am J Hum Genet* **79**, 113-119, doi:S0002-9297(07)60012-3 [pii]

10.1086/505406 (2006).

74 Kruger, R. *et al.* Mutation analysis of the neurofilament M gene in Parkinson's disease. *Neurosci Lett* **351**, 125-129, doi:S0304394003009030 [pii] (2003).

75 Hassenpflug, W. A. *et al.* Impact of mutations in the von Willebrand factor A2 domain on ADAMTS13-dependent proteolysis. *Blood* **107**, 2339-2345, doi:2005-04-1758 [pii]

10.1182/blood-2005-04-1758 (2006).

76 Yang, Y., Drummond-Borg, M. & Garcia-Heras, J. Molecular analysis of phenylketonuria (PKU) in newborns from Texas. *Hum Mutat* **17**, 523, doi:10.1002/humu.1140 [pii]

10.1002/humu.1141 (2001).

77 Isackson, P. J., Bennett, M. J. & Vladutiu, G. D. Identification of 16 new disease-causing mutations in the CPT2 gene resulting in carnitine palmitoyltransferase II deficiency. *Mol Genet Metab* **89**, 323-331, doi:S1096-7192(06)00284-8 [pii]

10.1016/j.ymgme.2006.08.004 (2006).

78 Clark, L. N. *et al.* Construction and validation of a Parkinson's disease mutation genotyping array for the Parkin gene. *Mov Disord* **22**, 932-937, doi:10.1002/mds.21419 (2007).

79 Okubadejo, N. *et al.* Analysis of Nigerians with apparently sporadic Parkinson disease for mutations in LRRK2, PRKN and ATXN3. *PLoS ONE* **3**, e3421, doi:10.1371/journal.pone.0003421 (2008).

80 Bjursell, C. *et al.* PMM2 mutation spectrum, including 10 novel mutations, in a large CDG type 1A family material with a focus on Scandinavian families. *Hum Mutat* **16**, 395-400, doi:10.1002/1098-1004(200011)16:5<395::AID-HUMU3>3.0.CO;2-T [pii]

10.1002/1098-1004(200011)16:5<395::AID-HUMU3>3.0.CO;2-T (2000).

81 Robinson, R., Carpenter, D., Shaw, M. A., Halsall, J. & Hopkins, P. Mutations in RYR1 in malignant hyperthermia and central core disease. *Hum Mutat* **27**, 977-989, doi:10.1002/humu.20356 (2006).

82 Bayarsikhan, M., Cappacchione, M., Brandom, B., Muldoon, S. & Sambuughin, N. in *American Society of Anesthesiologists* (Annual Meeting Abstracts, 2008).

83 Cohen, J. C., Boerwinkle, E., Mosley, T. H., Jr. & Hobbs, H. H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N Engl J Med* **354**, 1264-1272, doi:354/12/1264 [pii]

10.1056/NEJMoa054013 (2006).

84 Abou-Sleiman, P. M. *et al.* A heterozygous effect for PINK1 mutations in Parkinson's disease? *Ann Neurol* **60**, 414-419, doi:10.1002/ana.20960 (2006).

85 Ibanez, P. *et al.* Mutational analysis of the PINK1 gene in early-onset parkinsonism in Europe and North Africa. *Brain* **129**, 686-694, doi:awl005 [pii]

10.1093/brain/awl005 (2006).

86 Ishihara-Paul, L. *et al.* PINK1 mutations and parkinsonism. *Neurology* **71**, 896-902, doi:01.wnl.0000323812.40708.1f [pii]

10.1212/01.wnl.0000323812.40708.1f (2008).

87 Goodeve, A. *et al.* Phenotype and genotype of a cohort of families historically diagnosed with type 1 von Willebrand disease in the European study, Molecular and Clinical Markers for the Diagnosis and Management of Type 1 von Willebrand Disease (MCMDM-1VWD). *Blood* **109**, 112-121, doi:blood-2006-05-020784 [pii]

10.1182/blood-2006-05-020784 (2007).

88 Carvalho, M. A. *et al.* Determination of cancer risk associated with germ line BRCA1 missense variants by functional analysis. *Cancer Res* **67**, 1494-1501, doi:67/4/1494 [pii]

10.1158/0008-5472.CAN-06-3297 (2007).

89 Sheen, V. L. *et al.* Mutations in the X-linked filamin 1 gene cause periventricular nodular heterotopia in males as well as in females. *Hum Mol Genet* **10**, 1775-1783 (2001).

90 Desviat, L. R. *et al.* Genetic and phenotypic aspects of phenylalanine hydroxylase deficiency in Spain: molecular survey by regions. *Eur J Hum Genet* **7**, 386-392, doi:10.1038/sj.ejhg.5200312 (1999).

91 Bernot, A. *et al.* Non-founder mutations in the MEFV gene establish this gene as the cause of familial Mediterranean fever (FMF). *Hum Mol Genet* **7**, 1317-1325, doi:ddb164 [pii] (1998).

92 Aksentijevich, I. *et al.* Mutation and haplotype studies of familial Mediterranean fever reveal new ancestral relationships and evidence for a high carrier frequency with reduced penetrance in the Ashkenazi Jewish population. *Am J Hum Genet* **64**, 949-962, doi:AJHG980834 [pii] (1999).

93 Medlej-Hashim, M. *et al.* Familial Mediterranean fever: the potential for misdiagnosis of E148V using the E148Q usual RFLP detection method. *Clin Genet* **61**, 71-73, doi:010114 [pii] (2002).

94 Ogawa, T. *et al.* Mucopolysaccharidosis IVA: screening and identification of mutations of the N-acetylgalactosamine-6-sulfate sulfatase gene. *Hum Mol Genet* **4**, 341-349 (1995).

95 Dreyer, B. *et al.* Spectrum of USH2A mutations in Scandinavian patients with Usher syndrome type II. *Hum Mutat* **29**, 451, doi:10.1002/humu.9524 (2008).

96 Elmgren, A. *et al.* Identification of two functionally deficient plasma alpha 3-fucosyltransferase (FUT6) alleles. *Hum Mutat* **16**, 473-481, doi:10.1002/1098-1004(200012)16:6<473::AID-HUMU4>3.0.CO;2-T [pii]

10.1002/1098-1004(200012)16:6<473::AID-HUMU4>3.0.CO;2-T (2000).

97 Mashima, Y. *et al.* Novel cytochrome P4501B1 (CYP1B1) gene mutations in Japanese patients with primary congenital glaucoma. *Invest Ophthalmol Vis Sci* **42**, 2211-2216 (2001).

98 Jansen, G. A. *et al.* Human phytanoyl-CoA hydroxylase: resolution of the gene structure and the molecular basis of Refsum's disease. *Hum Mol Genet* **9**, 1195-1200, doi:ddd134 [pii] (2000).

99 De Morais, S. M. *et al.* Identification of a new genetic defect responsible for the polymorphism of (S)-mephenytoin metabolism in Japanese. *Mol Pharmacol* **46**, 594-598 (1994).

100 Kaneko, A., Kaneko, O., Taleo, G., Bjorkman, A. & Kobayakawa, T. High frequencies of CYP2C19 mutations and poor metabolism of proguanil in Vanuatu. *Lancet* **349**, 921-922, doi:S0140-6736(05)62696-7 [pii]

10.1016/S0140-6736(05)62696-7 (1997).

101 Dharmaraj, S. R. *et al.* Mutational analysis and clinical correlation in Leber congenital amaurosis. *Ophthalmic Genet* **21**, 135-150 (2000).

102 Shibuya, S., Higuchi, J., Shin, R. W., Tateishi, J. & Kitamoto, T. Protective prion protein polymorphisms against sporadic Creutzfeldt-Jakob disease. *Lancet* **351**, 419, doi:S0140-6736(05)78358-6 [pii]

10.1016/S0140-6736(05)78358-6 (1998).

103 Nishida, Y. *et al.* Creutzfeldt-Jakob disease with a novel insertion and codon 219 Lys/Lys polymorphism in PRNP. *Neurology* **63**, 1978-1979, doi:63/10/1978 [pii] (2004).

104 Smith, A. N. *et al.* Mutations in ATP6N1B, encoding a new kidney vacuolar proton pump 116-kD subunit, cause recessive distal renal tubular acidosis with preserved hearing. *Nat Genet* **26**, 71-75, doi:10.1038/79208 (2000).

105 Hadjigeorgiou, G. M. *et al.* Manifesting heterozygotes in a Japanese family with a novel mutation in the muscle-specific phosphoglycerate mutase (PGAM-M) gene. *Neuromuscul Disord* **9**, 399-402 (1999).

106 Perrault, I. *et al.* Retinal-specific guanylate cyclase gene mutations in Leber's congenital amaurosis. *Nat Genet* **14**, 461-464, doi:10.1038/ng1296-461 (1996).

107 Marlin, S. *et al.* Connexin 26 gene mutations in congenitally deaf children: pitfalls for genetic counseling. *Arch Otolaryngol Head Neck Surg* **127**, 927-933, doi:ooa00164 [pii] (2001).

108 Henderson, R. H. *et al.* An assessment of the apex microarray technology in genotyping patients with Leber congenital amaurosis and early-onset severe retinal dystrophy. *Invest Ophthalmol Vis Sci* **48**, 5684-5689, doi:48/12/5684 [pii]

10.1167/iovs.07-0207 (2007).

109 Knebelmann, B. *et al.* Spectrum of mutations in the COL4A5 collagen gene in X-linked Alport syndrome. *Am J Hum Genet* **59**, 1221-1232 (1996).

110 Menzel, O. *et al.* Knobloch syndrome: novel mutations in COL18A1, evidence for genetic heterogeneity, and a functionally impaired polymorphism in endostatin. *Hum Mutat* **23**, 77-84, doi:10.1002/humu.10284 (2004).

111 Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-3814 (2003).

112 Sunyaev, S. *et al.* Prediction of deleterious human alleles. *Hum Mol Genet* **10**, 591-597 (2001).

113 Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol* **5**, e254, doi:07-PLBI-RA-1258 [pii]

10.1371/journal.pbio.0050254 (2007).

114 Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872-876, doi:nature06884 [pii]

10.1038/nature06884 (2008).

115 Kim, J. I. *et al.* A highly annotated whole-genome sequence of a Korean individual. *Nature*, doi:nature08211 [pii]

10.1038/nature08211 (2009).

116 Ahn, S. M. *et al.* The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group. *Genome Res*, doi:gr.092197.109 [pii]

10.1101/gr.092197.109 (2009).

117 Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59, doi:nature07517 [pii]

10.1038/nature07517 (2008).

118 McKernan, K. J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two base encoding. *Genome Res*, doi:gr.091868.109 [pii]

10.1101/gr.091868.109 (2009).

119 Pushkarev, D., Neff, N. F. & Quake, S. R. Single-molecule sequencing of an individual human genome. *Nat Biotechnol*, doi:nbt.1561 [pii]

10.1038/nbt.1561 (2009).

120 Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, doi:nature08250 [pii]

10.1038/nature08250 (2009).

121 Buzza, M. *et al.* Mutations in the COL4A4 gene in thin basement membrane disease. *Kidney Int* **63**, 447-453, doi:kid780 [pii]

10.1046/j.1523-1755.2003.00780.x (2003).

122 Wang, Y. Y. *et al.* COL4A3 mutations and their clinical consequences in thin basement membrane nephropathy (TBMN). *Kidney Int* **65**, 786-790, doi:10.1111/j.1523-1755.2004.00453.x (2004).

123 Donaudy, F. *et al.* Multiple mutations of MYO1A, a cochlear-expressed gene, in sensorineural hearing loss. *Am J Hum Genet* **72**, 1571-1577, doi:S0002-9297(07)60457-1 [pii]

10.1086/375654 (2003).

124 Wolf, B. *et al.* Profound biotinidase deficiency in two asymptomatic adults. *Am J Med Genet* **73**, 5-9, doi:10.1002/(SICI)1096-8628(19971128)73:1<5::AID-AJMG2>3.0.CO;2-U [pii] (1997).

125 Baykal, T. *et al.* Asymptomatic adults and older siblings with biotinidase deficiency ascertained by family studies of index cases. *J Inherit Metab Dis* **28**, 903-912, doi:10.1007/s10545-005-0161-3 (2005).

126 Caprioli, J. *et al.* Genetics of HUS: the impact of MCP, CFH, and IF mutations on clinical presentation, response to treatment, and outcome. *Blood* **108**, 1267-1279, doi:blood-2005-10-007252 [pii]

10.1182/blood-2005-10-007252 (2006).

127 Fang, C. J. *et al.* Membrane cofactor protein mutations in atypical hemolytic uremic syndrome (aHUS), fatal Stx-HUS, C3 glomerulonephritis, and the HELLP syndrome. *Blood* **111**, 624-632, doi:blood-2007-04-084533 [pii]

10.1182/blood-2007-04-084533 (2008).

128 Olive, M., Goldfarb, L. G., Shatunov, A., Fischer, D. & Ferrer, I. Myotilinopathy: refining the clinical and myopathological phenotype. *Brain* **128**, 2315-2326, doi:awh576 [pii]

10.1093/brain/awh576 (2005).

129 Bar, H. *et al.* Conspicuous involvement of desmin tail mutations in diverse cardiac and skeletal myopathies. *Hum Mutat* **28**, 374-386, doi:10.1002/humu.20459 (2007).

130 Heron, S. E. *et al.* Extended spectrum of idiopathic generalized epilepsies associated with CACNA1H functional variants. *Ann Neurol* **62**, 560-568, doi:10.1002/ana.21169 (2007).

131 Bluteau, O. *et al.* Bi-allelic inactivation of TCF1 in hepatic adenomas. *Nat Genet* **32**, 312-315, doi:10.1038/ng1001

ng1001 [pii] (2002).

132 Hawn, T. R. *et al.* A common dominant TLR5 stop codon polymorphism abolishes flagellin signaling and is associated with susceptibility to legionnaires' disease. *J Exp Med* **198**, 1563-1572, doi:10.1084/jem.20031220

jem.20031220 [pii] (2003).

133 Misch, E. A. & Hawn, T. R. Toll-like receptor polymorphisms and susceptibility to human disease. *Clin Sci (Lond)* **114**, 347-360, doi:CS20070214 [pii]

10.1042/CS20070214 (2008).

134 Torres, R. *et al.* Mutation screening of the Wolfram syndrome gene in psychiatric patients. *Mol Psychiatry* **6**, 39-43 (2001).

135 Monreal, A. W., Zonana, J. & Ferguson, B. Identification of a new splice form of the EDA1 gene permits detection of nearly all X-linked hypohidrotic ectodermal dysplasia mutations. *Am J Hum Genet* **63**, 380-389, doi:S0002-9297(07)61481-5 [pii]

10.1086/301984 (1998).

136 Afzal, A. R. *et al.* Novel mutations in the duplicated region of the polycystic kidney disease 1 (PKD1) gene provides supporting evidence for gene conversion. *Genet Test* **4**, 365-370, doi:10.1089/109065700750065108 (2000).

137 Zuhlke, K. A. *et al.* Truncating BRCA1 mutations are uncommon in a cohort of hereditary prostate cancer families with evidence of linkage to 17q markers. *Clin Cancer Res* **10**, 5975-5980, doi:10.1158/1078-0432.CCR-04-0554

10/18/5975 [pii] (2004).

138 Solomou, E. E. *et al.* Perforin gene mutations in patients with acquired aplastic anemia. *Blood* **109**, 5234-5237, doi:blood-2006-12-063495 [pii]

10.1182/blood-2006-12-063495 (2007).

139 Arnestad, M. *et al.* Prevalence of long-QT syndrome gene variants in sudden infant death syndrome. *Circulation* **115**, 361-367, doi:CIRCULATIONAHA.106.658021 [pii]

10.1161/CIRCULATIONAHA.106.658021 (2007).

140 Ackerman, M. J. *et al.* Ethnic differences in cardiac potassium channel variants: implications for genetic susceptibility to sudden cardiac death and genetic testing for congenital long QT syndrome. *Mayo Clin Proc* **78**, 1479-1487 (2003).

141 Kakita, A. *et al.* Bilateral periventricular nodular heterotopia due to filamin 1 gene mutation: widespread glomeruloid microvascular anomaly and dysplastic cytoarchitecture in the cerebral cortex. *Acta Neuropathol* **104**, 649-657, doi:10.1007/s00401-002-0594-9 (2002).

142 Yang, J. J. *et al.* Identification of mutations in members of the connexin gene family as a cause of nonsyndromic deafness in Taiwan. *Audiol Neurootol* **12**, 198-208, doi:000099024 [pii]

10.1159/000099024 (2007).

143 Brown, A. M., Dolan, J. W., Willi, S. M., Garvey, W. T. & Argyropoulos, G. Endogenous mutations in human uncoupling protein 3 alter its functional properties. *FEBS Lett* **464**, 189-193, doi:S0014-5793(99)01708-1 [pii] (1999).

144 Zhang, H. *et al.* Mutational analysis of TSC1 and TSC2 genes in Japanese patients with tuberous sclerosis complex. *J Hum Genet* **44**, 391-396, doi:10.1007/s100380050185 (1999).

145 Choi, J. E., Chae, J. H., Hwang, Y. S. & Kim, K. J. Mutational analysis of TSC1 and TSC2 in Korean patients with tuberous sclerosis complex. *Brain Dev* **28**, 440-446, doi:S0387-7604(06)00023-4 [pii]

10.1016/j.braindev.2006.01.006 (2006).

146 Milunsky, A., Ito, M., Maher, T. A., Flynn, M. & Milunsky, J. M. Prenatal molecular diagnosis of tuberous sclerosis complex. *Am J Obstet Gynecol* **200**, 321 e321-326, doi:S0002-9378(08)02204-7 [pii]

10.1016/j.ajog.2008.11.004 (2009).

147 Poetter, K. *et al.* Mutations in either the essential or regulatory light chains of myosin are associated with a rare myopathy in human heart and skeletal muscle. *Nat Genet* **13**, 63-69, doi:10.1038/ng0596-63 (1996).

148 Andersen, P. S. *et al.* Myosin light chain mutations in familial hypertrophic cardiomyopathy: phenotypic presentation and frequency in Danish and South African populations. *J Med Genet* **38**, E43 (2001).

149 Hougs, L. *et al.* One third of Danish hypertrophic cardiomyopathy patients with MYH7 mutations have mutations [corrected] in MYH7 rod region. *Eur J Hum Genet* **13**, 161-165, doi:5201310 [pii]
10.1038/sj.ejhg.5201310 (2005).

150 Szczesna, D. *et al.* Familial hypertrophic cardiomyopathy mutations in the regulatory light chains of myosin affect their structure, Ca2+ binding, and phosphorylation. *J Biol Chem* **276**, 7086-7092, doi:10.1074/jbc.M009823200
M009823200 [pii] (2001).

151 Szczesna-Cordary, D., Guzman, G., Ng, S. S. & Zhao, J. Familial hypertrophic cardiomyopathy-linked alterations in Ca2+ binding of human cardiac myosin regulatory light chain affect cardiac muscle contraction. *J Biol Chem* **279**, 3535-3542, doi:10.1074/jbc.M307092200
M307092200 [pii] (2004).

152 Roopnarine, O. Mechanical defects of muscle fibers with myosin light chain mutants that cause cardiomyopathy. *Biophys J* **84**, 2440-2449, doi:S0006-3495(03)75048-6 [pii]
10.1016/S0006-3495(03)75048-6 (2003).

153 Bauce, B. *et al.* Clinical profile of four families with arrhythmogenic right ventricular cardiomyopathy caused by dominant desmoplakin mutations. *Eur Heart J* **26**, 1666-1675, doi:ehi341 [pii]
10.1093/eurheartj/ehi341 (2005).

154 Rampazzo, A. in *European Society of Cardiology*.

155 Horvath, R. *et al.* Phenotypic spectrum associated with mutations of the mitochondrial polymerase gamma gene. *Brain* **129**, 1674-1684, doi:awl088 [pii]
10.1093/brain/awl088 (2006).

156 Hedrich, C. M. *et al.* Autosomal dominant neurohypophyseal diabetes insipidus in two families. Molecular analysis of the vasopressin-neurophysin II gene and functional studies of three missense mutations. *Horm Res* **71**, 111-119, doi:000183900 [pii]
10.1159/000183900 (2009).

157 Rittig, S. *et al.* Identification of 13 new mutations in the vasopressin-neurophysin II gene in 17 kindreds with familial autosomal dominant neurohypophyseal diabetes insipidus. *Am J Hum Genet* **58**, 107-117 (1996).

158 Davies, J. P. *et al.* Fabry disease: fourteen alpha-galactosidase A mutations in unrelated families from the United Kingdom and other European countries. *Eur J Hum Genet* **4**, 219-224 (1996).

159 Ashley, G. A., Shabbeer, J., Yasuda, M., Eng, C. M. & Desnick, R. J. Fabry disease: twenty novel alpha-galactosidase A mutations causing the classical phenotype. *J Hum Genet* **46**, 192-196, doi:10.1007/s100380170088 (2001).

160 Eng, C. M. *et al.* Fabry disease: twenty-three mutations including sense and antisense CpG alterations and identification of a deletional hot-spot in the alpha-galactosidase A gene. *Hum Mol Genet* **3**, 1795-1799 (1994).

161 Zhang, Q. *et al.* Nesprin-1 and -2 are involved in the pathogenesis of Emery Dreifuss muscular dystrophy and are critical for nuclear envelope integrity. *Hum Mol Genet* **16**, 2816-2833, doi:ddm238 [pii]
10.1093/hmg/ddm238 (2007).

162 Sohocki, M. M. *et al.* Prevalence of AIPL1 mutations in inherited retinal degenerative disease. *Mol Genet Metab* **70**, 142-150, doi:10.1006/mgme.2000.3001
S1096-7192(00)93001-4 [pii] (2000).

163 Dharmaraj, S. *et al.* The phenotype of Leber congenital amaurosis in patients with AIPL1 mutations. *Arch Ophthalmol* **122**, 1029-1037, doi:10.1001/archopht.122.7.1029
122/7/1029 [pii] (2004).

164 Yamamoto, T. *et al.* Novel TSC1 and TSC2 mutations in Japanese patients with tuberous sclerosis complex. *Brain Dev* **24**, 227-230, doi:S0387760402000177 [pii] (2002).

165 Lampe, A. K. *et al.* Automated genomic sequence analysis of the three collagen VI genes: applications to Ullrich congenital muscular dystrophy and Bethlem myopathy. *J Med Genet* **42**, 108-120, doi:42/2/108 [pii]
10.1136/jmg.2004.023754 (2005).

166 Ballmaier, M. *et al.* c-mpl mutations are the cause of congenital amegakaryocytic thrombocytopenia. *Blood* **97**, 139-146 (2001).

167 Fox, N. E. *et al.* Compound heterozygous c-Mpl mutations in a child with congenital amegakaryocytic thrombocytopenia: functional characterization and a review of the literature. *Exp Hematol* **37**, 495-503, doi:S0301-472X(09)00005-8 [pii]
10.1016/j.exphem.2009.01.001 (2009).

168 Goldfarb, L. G., Vicart, P., Goebel, H. H. & Dalakas, M. C. Desmin myopathy. *Brain* **127**, 723-734, doi:10.1093/brain/awh033
awh033 [pii] (2004).

169 Kostareva, A. *Genetic and Pathophysiological Study of Desmin Derangements in Cardiomyopathies* Ph.D. thesis, Karolinska institutet, (2007).

170 Seyedahmadi, B. J., Rivolta, C., Keene, J. A., Berson, E. L. & Dryja, T. P. Comprehensive screening of the USH2A gene in Usher syndrome type II and non-syndromic recessive retinitis pigmentosa. *Exp Eye Res* **79**, 167-173, doi:10.1016/j.exer.2004.03.005
S0014483504000892 [pii] (2004).

171 Cremers, F. P. *et al.* Development of a genotyping microarray for Usher syndrome. *J Med Genet* **44**, 153-160, doi:jmg.2006.044784 [pii]
10.1136/jmg.2006.044784 (2007).
172 Takahashi, H. *et al.* Mutations of the BRCA2 gene in ovarian carcinomas. *Cancer Res* **56**, 2738-2741 (1996).
173 Freedman, M. L. *et al.* Common variation in BRCA2 and breast cancer risk: a haplotype-based analysis in the Multiethnic Cohort. *Hum Mol Genet* **13**, 2431-2441, doi:10.1093/hmg/ddh270
ddh270 [pii] (2004).
174 Eulenburg, V. *et al.* Mutations within the human GLYT2 (SLC6A5) gene associated with hyperekplexia. *Biochem Biophys Res Commun* **348**, 400-405, doi:S0006-291X(06)01576-2 [pii]
10.1016/j.bbrc.2006.07.080 (2006).
175 Rees, M. I. *et al.* Mutations in the gene encoding GlyT2 (SLC6A5) define a presynaptic component of human startle disease. *Nat Genet* **38**, 801-806, doi:ng1814 [pii]
10.1038/ng1814 (2006).
176 Harvey, R. J., Topf, M., Harvey, K. & Rees, M. I. The genetics of hyperekplexia: more than startle! *Trends Genet* **24**, 439-447, doi:S0168-9525(08)00200-X [pii]
10.1016/j.tig.2008.06.005 (2008).
177 Matthijs, G. *et al.* Mutations in PMM2 that cause congenital disorders of glycosylation, type Ia (CDG-Ia). *Hum Mutat* **16**, 386-394, doi:10.1002/1098-1004(200011)16:5<386::AID-HUMU2>3.0.CO;2-Y [pii]
10.1002/1098-1004(200011)16:5<386::AID-HUMU2>3.0.CO;2-Y (2000).
178 Grubenmann, C. E. *et al.* Deficiency of the first mannosylation step in the N-glycosylation pathway causes congenital disorder of glycosylation type Ik. *Hum Mol Genet* **13**, 535-542, doi:10.1093/hmg/ddh050
ddh050 [pii] (2004).
179 Le Bizec, C. *et al.* A new insight into PMM2 mutations in the French population. *Hum Mutat* **25**, 504-505, doi:10.1002/humu.9336 (2005).
180 Pasutto, F. *et al.* Profiling of WDR36 missense variants in German patients with glaucoma. *Invest Ophthalmol Vis Sci* **49**, 270-274, doi:49/1/270 [pii]
10.1167/iovs.07-0500 (2008).
181 Footz, T. K. *et al.* Glaucoma-associated WDR36 variants encode functional defects in a yeast model system. *Hum Mol Genet* **18**, 1276-1287, doi:ddp027 [pii]
10.1093/hmg/ddp027 (2009).
182 Martoni, E. *et al.* C.P.2.05 Molecular analysis of COL6 genes in patients with Bethlem myopathy and Ullrich congenital muscular dystrophy. *Neuromuscular Disorders* **17**, 844-845 (2007).

183 Jonsson, B. A., Bergh, A., Stattin, P., Emmanuelsson, M. & Gronberg, H. Germline mutations in E-cadherin do not explain association of hereditary prostate cancer, gastric cancer and breast cancer. *Int J Cancer* **98**, 838-843, doi:10.1002/ijc.10258 [pii] (2002).

184 Suriano, G., Seixas, S., Rocha, J. & Seruca, R. A model to infer the pathogenic significance of CDH1 germline missense variants. *J Mol Med* **84**, 1023-1031, doi:10.1007/s00109-006-0091-z (2006).

185 Ellis, P. E., Dawson, M. & Dixon, M. J. Mutation testing in Treacher Collins Syndrome. *J Orthod* **29**, 293-297; discussion 278 (2002).

186 Macaya, D. *et al.* A synonymous mutation in TCOF1 causes Treacher Collins syndrome due to mis-splicing of a constitutive exon. *Am J Med Genet A* **149A**, 1624-1627, doi:10.1002/ajmg.a.32834 (2009).

187 Garcia-Gonzalez, M. A. *et al.* Evaluating the clinical utility of a molecular genetic test for polycystic kidney disease. *Mol Genet Metab* **92**, 160-167, doi:S1096-7192(07)00161-8 [pii]
10.1016/j.ymgme.2007.05.004 (2007).

188 Rossetti, S. *et al.* Comprehensive molecular diagnostics in autosomal dominant polycystic kidney disease. *J Am Soc Nephrol* **18**, 2143-2160, doi:ASN.2006121387 [pii]
10.1681/ASN.2006121387 (2007).

189 Niimura, H. *et al.* Sarcomere protein gene mutations in hypertrophic cardiomyopathy of the elderly. *Circulation* **105**, 446-451 (2002).

190 Mogensen, J. *et al.* Frequency and clinical expression of cardiac troponin I mutations in 748 consecutive families with hypertrophic cardiomyopathy. *J Am Coll Cardiol* **44**, 2315-2325, doi:S0735-1097(04)01851-0 [pii]
10.1016/j.jacc.2004.05.088 (2004).

191 Fidalgo, P. *et al.* Detection of mutations in mismatch repair genes in Portuguese families with hereditary non-polyposis colorectal cancer (HNPCC) by a multi-method approach. *Eur J Hum Genet* **8**, 49-53, doi:10.1038/sj.ejhg.5200393 (2000).

192 Raevaara, T. *Functional Significance of Minor MLH1 Germline Alterations Found in Colon Cancer Patients* Ph.D. thesis, University of Helsinki, (2005).

193 Seo, J. H. *et al.* BRCA1 and BRCA2 germline mutations in Korean patients with sporadic breast cancer. *Hum Mutat* **24**, 350, doi:10.1002/humu.9275 (2004).

194 Suter, N. M. *et al.* BRCA1 and BRCA2 mutations in women from Shanghai China. *Cancer Epidemiol Biomarkers Prev* **13**, 181-189 (2004).

195 Lim, M. C. *et al.* BRCA1 and BRCA2 germline mutations in Korean ovarian cancer patients. *J Cancer Res Clin Oncol*, doi:10.1007/s00432-009-0607-3 (2009).

196 Chao, S. C., Yang, M. H. & Lee, S. F. Novel KRT14 mutation in a Taiwanese patient with epidermolysis bullosa simplex (Kobner type). *J Formos Med Assoc* **101**, 287-290 (2002).

197 Hattori, N. *et al.* A case of epidermolysis bullosa simplex with a newly found missense mutation and polymorphism in the highly conserved helix termination motif among type I keratins, which was previously reported as a pathogenic missense mutation. *Br J Dermatol* **155**, 1062-1063, doi:BJD7425 [pii]
10.1111/j.1365-2133.2006.07425.x (2006).

198 Zhang, D. Y. *et al.* [Mutation detection of PKD2 gene in Chinese by denaturing high-performance liquid chromatograph]. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi* **21**, 211-214, doi:940621050 [pii] (2004).

199 Chung, W. *et al.* PKD2 gene mutation analysis in Korean autosomal dominant polycystic kidney disease patients using two-dimensional gene scanning. *Clin Genet* **70**, 502-508, doi:CGE721 [pii]
10.1111/j.1399-0004.2006.00721.x (2006).

200 Yoshihashi, H. *et al.* Imprinting of human GRB10 and its mutations in two patients with Russell-Silver syndrome. *Am J Hum Genet* **67**, 476-482, doi:S0002-9297(07)62656-1 [pii] (2000).

201 Mergenthaler, S. *et al.* Conflicting reports of imprinting status of human GRB10 in developing brain: how reliable are somatic cell hybrids for predicting allelic origin of expression? *Am J Hum Genet* **68**, 543-545, doi:S0002-9297(07)64109-3 [pii]
10.1086/318192 (2001).

202 Rezaie, T. *et al.* Adult-onset primary open-angle glaucoma caused by mutations in optineurin. *Science* **295**, 1077-1079, doi:10.1126/science.1066901
295/5557/1077 [pii] (2002).

203 Ayala-Lugo, R. M. *et al.* Variation in optineurin (OPTN) allele frequencies between and within populations. *Mol Vis* **13**, 151-163, doi:v13/a18 [pii] (2007).

204 Drenth, J. P. *et al.* SCN9A mutations define primary erythermalgia as a neuropathic disorder of voltage gated sodium channels. *J Invest Dermatol* **124**, 1333-1338, doi:JID23737 [pii]
10.1111/j.0022-202X.2005.23737.x (2005).

205 Waxman, S. G. & Dib-Hajj, S. Erythermalgia: molecular basis for an inherited pain syndrome. *Trends Mol Med* **11**, 555-562, doi:S1471-4914(05)00231-5 [pii]
10.1016/j.molmed.2005.10.004 (2005).

206 Amin, H. K., Hoeppner, W., Shaarawy, M. & Barakat, M. Gene Symbol: CYP11B1 Disease: Congenital Adrenal Hyperplasia. *Hum Genet* **110**, 2, doi:10.1007/s00439-002-0703-9 (2002).

207 Lee, H. H., Won, G. S., Chao, H. T., Lee, Y. J. & Chung, B. C. Novel missense mutations, GCC [Ala306]- > GTC [Val] and ACG [Thr318]- > CCG [Pro], in the CYP11B1 gene cause steroid 11beta-hydroxylase deficiency in the Chinese. *Clin Endocrinol (Oxf)* **62**, 418-422, doi:CEN2234 [pii]
10.1111/j.1365-2265.2005.02234.x (2005).

208 Merke, D. P. *et al.* Novel CYP11B1 mutations in congenital adrenal hyperplasia due to steroid 11 beta-hydroxylase deficiency. *J Clin Endocrinol Metab* **83**, 270-273 (1998).

209 Hill, J. M., Bhattacharjee, P. S. & Neumann, D. M. Apolipoprotein E alleles can contribute to the pathogenesis of numerous clinical conditions including HSV-1 corneal disease. *Exp Eye Res* **84**, 801-811, doi:S0014-4835(06)00338-1 [pii]
10.1016/j.exer.2006.08.001 (2007).

210 Malmgren, B., Lindskog, S., Elgadi, A. & Norgren, S. Clinical, histopathologic, and genetic investigation in two large families with dentinogenesis imperfecta type II. *Hum Genet* **114**, 491-498, doi:10.1007/s00439-004-1084-z (2004).

211 Kim, J. W. & Simmer, J. P. Hereditary dentin defects. *J Dent Res* **86**, 392-399, doi:86/5/392 [pii] (2007).

212 Scheid, R. *et al.* Cysteine-sparing notch3 mutations: cadasil or cadasil variants? *Neurology* **71**, 774-776, doi:71/10/774 [pii]
10.1212/01.wnl.0000324928.44694.f7 (2008).

213 Arbustini, E. A. *et al.* Gene symbol: CMD1A. Disease: Dilated cardiomyopathy associated with conduction system disease. *Hum Genet* **117**, 295 (2005).

214 Ruiz-Perez, V. L. *et al.* Mutations in a new gene in Ellis-van Creveld syndrome and Weyers acrodental dysostosis. *Nat Genet* **24**, 283-286, doi:10.1038/73508 (2000).

215 Dvorakova, L. *et al.* Eight novel ABCD1 gene mutations and three polymorphisms in patients with X-linked adrenoleukodystrophy: The first polymorphism causing an amino acid exchange. *Hum Mutat* **18**, 52-60, doi:10.1002/humu.1149 [pii]
10.1002/humu.1149 (2001).

216 Allard, D. *et al.* Novel mutations of the PCSK9 gene cause variable phenotype of autosomal dominant hypercholesterolemia. *Hum Mutat* **26**, 497, doi:10.1002/humu.9383 (2005).

Chapter 3

Initial Data Release from the Personal Genome Project

Abraham M. Rosenbaum[1]*, Alexander Wait-Zaranek[1]*, Xiaodi Wu[1], Joseph Thakuria[1,2], Gregory J. Porreca[1], Jin Billy Li[1], Michael F. Chou[1], Kun Zhang[3], John Aach[1], Emily LeProust[4], George M. Church[1]

[1]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA; [2]Division of Clinical and Biochemical Genetics, Massachusetts General Hospital, Boston, MA 02114, USA; [3]Department of Bioengineering, University of California, San Diego, CA 92093, USA; [4]Genomics Solution Unit, Agilent Technologies, 5301 Stevens Creek Boulevard, Santa Clara, CA 95051, USA

*These authors contributed equally

Author Contributions A.M.R. performed the MIP capture reaction with help from J.B.L., K.Z., G.J.P. and E.L.; A.M.R. created the libraries with help from J.B.L. and G.J.P.; A.M.R. mapped the reads with help from A.W.Z. and J.B.L.; K.Z. performed the microarray analysis; A.W.Z. created the genomerator software tool; A.M.R. and A.W.Z. analyzed the capture efficiency with the help of M.F.C. and J.A.; A.M.R. performed the variant analysis with help from A.W.Z., J.T. and X.W.; and G.M.C. supervised all aspects of the study.

Acknowledgements We are grateful for the help and advice provided by all the member of the Church Laboratory, specifically Madeleine Price-Ball and Graham Rockwell for computational support; Jason Bobe, Jeantine Lunshof and other members of the Personal Genome Project Community; Ting Wu and other members of PGEd; and the computational support and help provided by Scalable Computing Experts. We thank NHGRI, NHLBI and Personalgenomes.org for funding support.

Abstract

The Personal Genome Project is a community-oriented endeavor that seeks to engage researchers, volunteers and clinicians in furthering our knowledge of the human genome and all its variations. In this paper we report on the first semiannual data release from the project. We discuss improvements in targeted sequencing that enabled the sequencing of large exonic regions from each of the first ten volunteers, and the computational pipeline that we set up to analyze the data. Approximately 20% of exons were targeted and over 100 Mbp of mappable exonic sequence was acquired for each individual. Application of strict thresholds, however, filtered out most of this information and we released to the community 3-5% of each individual's exonic sequence with high confidence. This data was then processed for variants of potential clinical interest, and we present a list of seventeen variants with more detailed explanations as to their importance to our volunteers.

Introduction

A number of projects have recently been announced to further our knowledge of human genetic diversity in both the pathological and normal state. Both the 1,000 Genome Project[1] and the ClinSeq effort[2] propose a mixture of publicly available and secured data. They will release aggregate data that cannot be traced back to individual participants, and also deposit all data, including information that may lead to identification, in the NIH Database of Genotypes and Phenotypes (dbGaP)[3]. Access to dbGaP data is limited to NIH approved research, and prior to gaining access one must agree to not attempt to identify any individual present in the database. The Personal Genome Project[4] (PGP), however, focuses on making its information fully available. This data sharing policy is predicated upon two premises: (1) complete genome-phenome databases will lead to the most thorough understanding of individual variation[5], and (2) there is no way to guarantee the security and anonymity of such data[6-7]. In the PGP the participants are subject to a thorough open-consent policy[8-9], that both clarifies the impossibility of guaranteeing anonymity and confirms that the participants know the positive and potentially negative consequences of participating. Allowing this data to be accessible to everyone will certainly increase the number of hypotheses being tested and our overall knowledge of human genome diversity. As has been reported in recent surveys[10], many research volunteers desire continual interaction with researchers analyzing their genetic information, and the PGP has already attracted over 1,000 volunteers.

After recruiting the first ten participants (PGP10), the decision was made to defer the sequencing of their entire genomes until it was more affordable, and instead concentrate on a subset of each genome. The large cost of whole genome sequencing has

led to a number of high-throughput targeting methods for enriching regions of interest. Surface-based microarray hybridization methods[11-12] first create a whole-genome small-insert library and then enrich for desired regions through hybridization to long surface bound oligomers. These methods typically require large amounts of library input and long incubation times, yielding libraries enriched for sequences present in a Poisson distribution around the targeted regions. An alternative approach, based upon Molecular Inversion Probes (MIP)[13], relies upon the specificity of polymerase chain reaction (PCR) in targeting the regions of interest. In this process, two probes are designed to flank each region of interest, and oligomers are designed so that each set of these probes are attached to either end of a universal sequence (the "backbone"). After hybridizing the probes to their respective regions, the upstream probe (the "extension arm") is extended by polymerase copying the genomic target. This synthesized region is then ligated to the downstream arm (the "ligation arm") creating a circle that is used in downstream processes. See Figure 3-1 for an overview of the process. While PCR cannot be multiplexed very efficiently[14], the MIP capture process has proven effective in targeting over 10,000 single nucleotide positions[15], and in Porreca et al., in targeting over 50,000 exons (approximately 20% of all exons)[16]. In this initial publication, MIP capture was limited by capture efficiency and uniformity. A later publication[17] attempted to improve efficiency by increasing the size of the backbone; concurrent research in our lab[18] (see Appendix E), however, demonstrated a 435x improvement in capture efficiency achievable through a combination of increased concentration of probes and enzymes, longer incubation times, and better probe design. We utilized this method for the target enrichment reported here.

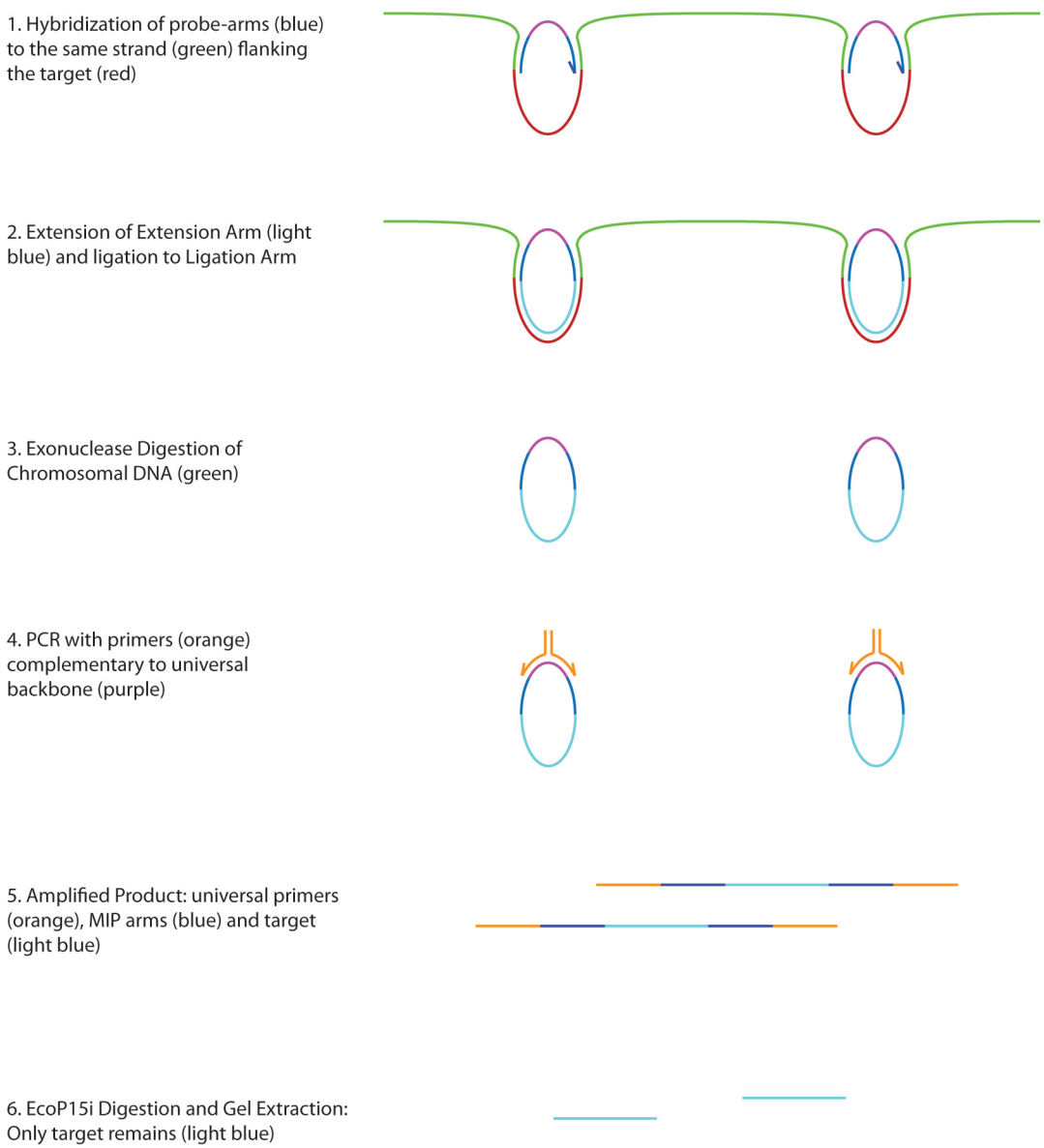


Figure 3-1. Overview of the MIP capture Process. Three libraries were created from this capture reaction. In the end-sequencing library the product in step 5 was sequenced by appending Illumina-GA2 clustering adapters to the universal primers (orange) and the custom sequencing primers complementary to the universal primers were used. For the shotgun libraries either the product from step 5 was circularized and made high molecular weight through rolling circle amplification before creation of a Polonator library (see Appendix D) or the product from step 6 was concatenated into high molecular weight concatemers for an Illumina GAII library.

Here we report on the MIP-targeted capture, sequencing and analysis of large exonic regions from the PGP10. In particular, we demonstrate improvements in capture efficiency, specificity and uniformity when compared with Porreca et al[16]. We then

describe the creation of two different shotgun libraries, specifically detailing our experience with the restriction enzyme EcoP15i. Finally, we discuss the bioinformatics analysis of these libraries both in terms of minimizing false positives and negatives and in terms of prioritizing variants for greater phenotypic workup.

Results

55,000 MIPs targeting 55,000 exons were synthesized. The sequences of the hybridization and ligation arms, as well as that of the universal backbone were the same as Porreca et al.[16]. The universal primers used to amplify the raw probes, however, were changed to reflect improvements by Li et al.[18] (see methods and Appendix E). In brief, 55,000 oligomers are synthesized on a solid substrate and cleaved. These oligomers are PCR-amplified using universal primers. The universal primers are then enzymatically removed, yielding the final MIPs. The final probes consist of two 20bp unique hybridization arms separated by a universal 30bp backbone. The arms recognize regions flanking one of 55,000 exons (the target set). In the capture reaction, the probes are mixed with genomic DNA and polymerase and ligase are used to replicate the target exons and circularize the MIPs, generating the MIP-capture product. The MIP-capture product consists of a universal backbone and two unique reference-sequence 20bp arms flanking a replica of the genomic target. Sequencing of the region between the two flanking arms will provide information of any single nucleotide variants and small indels. Additionally, variations in the number of times a target is seen can imply copy number variations. In this report, the captured targets were analyzed via three separate libraries: an end-sequencing library in which the reference-sequence arms and 11bp of the target

replica region is sequenced, and two shotgun libraries, one for the Illumina GAII sequencer and one for the Polonator Sequencer[19].

To assess the feasibility of MIP capture for targeted sequencing across a number of individuals, we first analyzed the reproducibility of the process. Porreca et al. reported a correlation of 0.558 between the end-sequencing of two different capture reactions from the same individual. With our improvements, end sequencing of two of our libraries from two different individuals had a Pearson correlation of 0.7674. Furthermore, we suspect that the true correlation is even higher, because, while the same sample (PGP9) run simultaneously on two separate lanes of the GAII had a correlation of 0.9965, the correlation between separate runs was only 0.8516, primarily due to differences in loading densities (Figure 3-2A). The pair-wise correlation for each of the ten shotgun libraries (Figure 3-2B) show some level of inconsistency, but the highest concordance achieved (0.93) show the potential for very high reproducibility. Furthermore, since the fraction of targets seen more than 10x, which included a large number of PCR duplicates, showed a lower correlation than the entire target set, better library construction methods will likely improve these numbers even more.

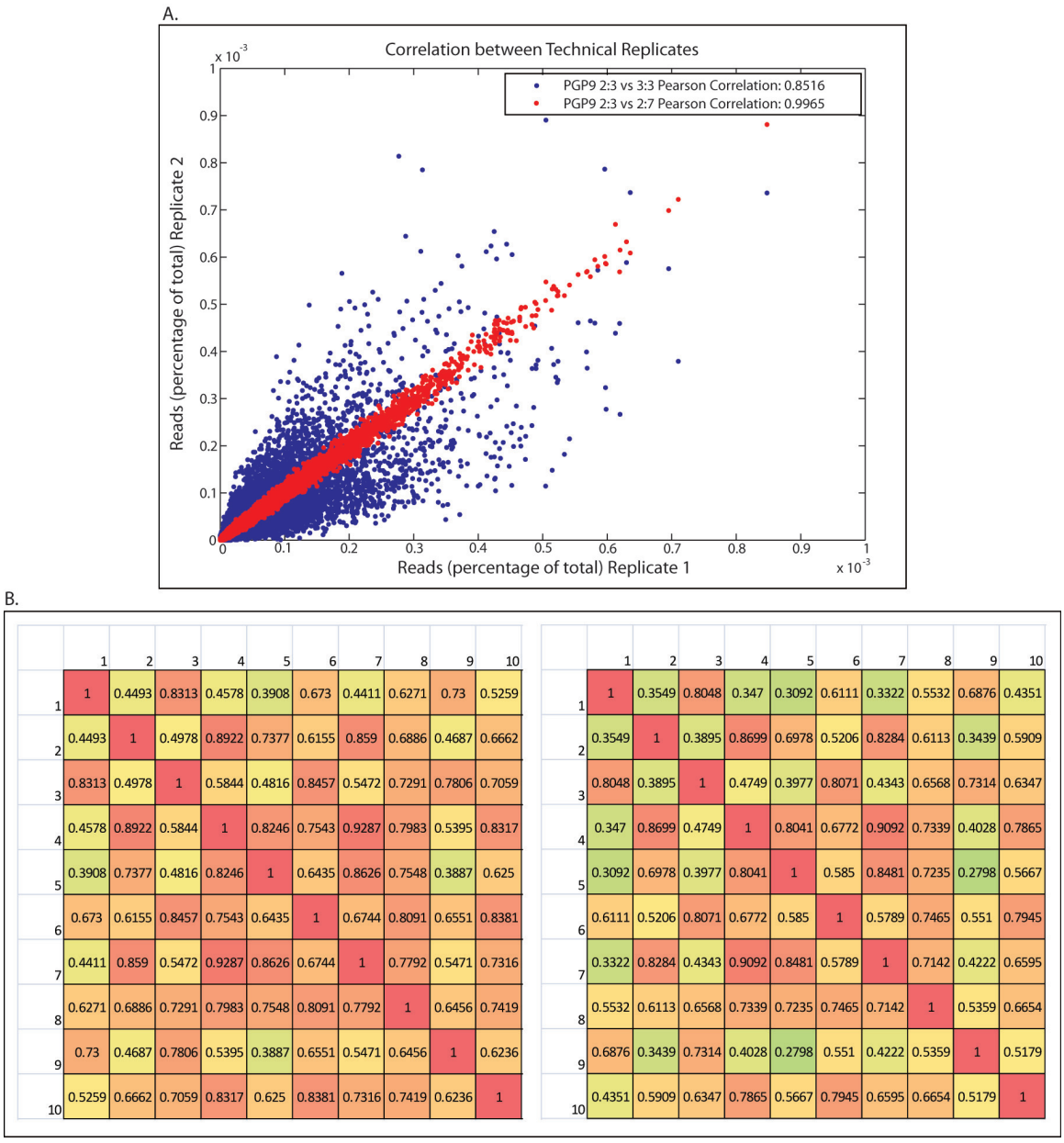


Figure 3-2. Reproducibility of Capture and Downstream Steps A. Technical Replicates. Red depicts the correlation between two sequencing lanes loaded with the same PGP9 shotgun library and run simultaneously, and the blue depicts the correlation between two runs of the same PGP9 shotgun library run on different days on the same sequencing machine. **B. Correlation Between Shotgun Libraries.** The left grid depicts the pairwise correlation between different PGP samples sampling the normalized number of times each target is seen. The right grid only samples targets seen more than 10x. There is significant variation in capture between different libraries, but the existence of very highly correlated pairs may implicate library construction as the source of the variability. This is supported by the decrease in correlation between high-capture targets in which there is likely to be a greater number of PCR duplicates, a by-product of library construction.

To compare our results with the MIP capture benchmarks, we also analyzed the uniformity within each library, the accuracy of capture in determining zygosity, and the sensitivity towards copy number variants. The uniformity of capture within each sample shows a vast improvement over Porreca et al.; this reflects both changes in the capture process and in the initial probe synthesis. End sequencing showed the successful capture 87% of targets, with ~80% and ~50% found in a 100-fold and 10-fold range, respectively (Figure 3-3A). While there still remained a range of almost four orders of magnitude between the most and least abundant targets, this range would still permit sequencing of greater than 60% of the targets with more than 10x coverage utilizing one lane of the Illumina GAII with 36bp reads. Similarly, the benchmark for the ability to discern between heterozygous and homozygous calls was vastly improved over previous efforts; while in Porreca et al. even 500x coverage was generally insufficient for determining a position heterozygous or homozygous, 50x coverage was sufficient for this dataset. This result is comparable to that achieved by Turner et al. when they utilize only the best 30% of 55,000 probes[20] (Figure 3-3C). To assess the feasibility of correctly identifying copy number variations we used the normalized read count from all targets for the shotgun libraries, and compared the fraction of reads originating from the X-chromosome for males vs. females. On average the read depth for the X-chromosome for female participants was similar to the autosomal read depth. For male participants, however, read depth was 20-30% lower on average (Figure 3B). The capture quality as assessed by these results encouraged us to use this method to capture and sequence exons from each of these participants.

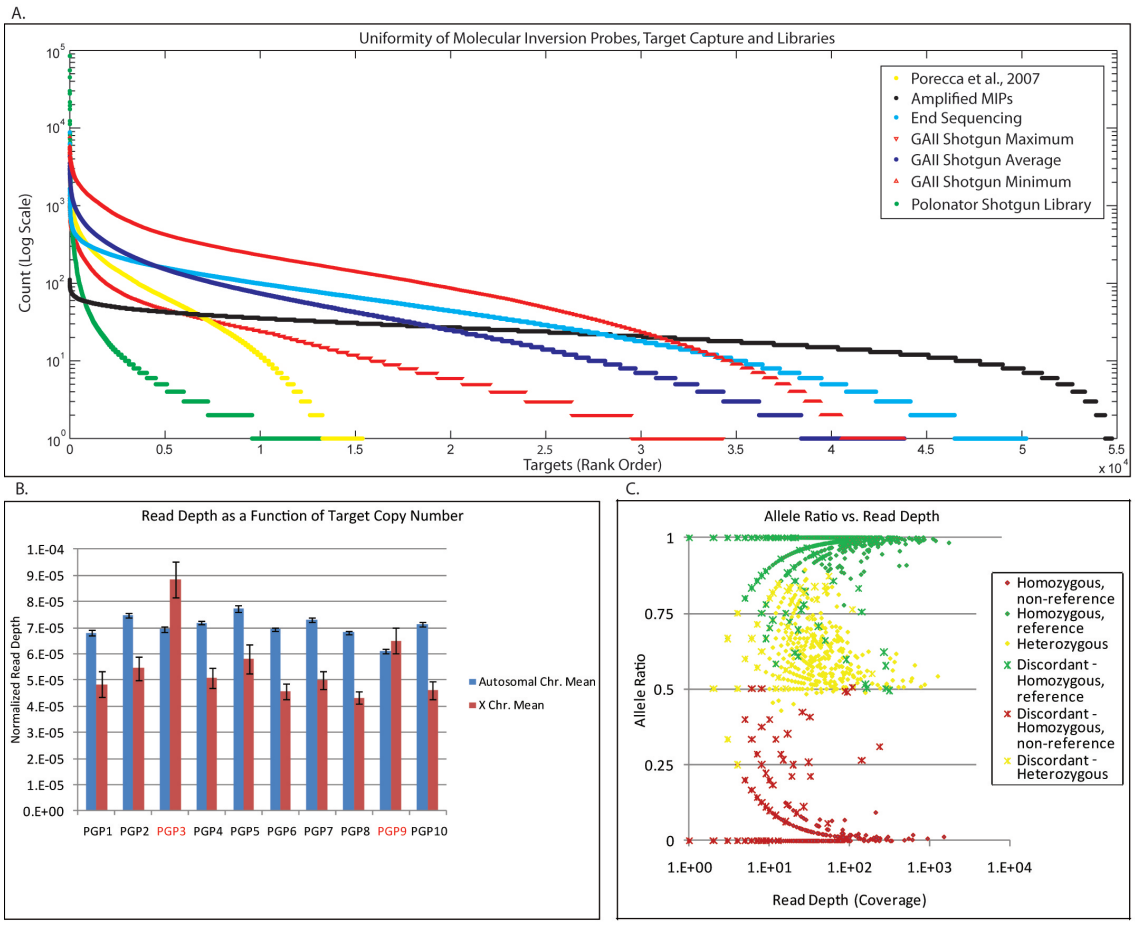


Figure 3-3. MIP Probe Capture Efficiency. A. Uniformity of Molecular Inversion Probes, Target Capture and Libraries. All sequences (with the exception of Porecca et al.) were generated with the Illumina GAII DNA Sequencer. Amplified MIPs and the Polonator Shotgun Library are from a single GAII lane, End Sequencing is the average of two different libraries and GAII Shotgun Average is the average of ten libraries. Placement of the shotgun reads was performed with MAQ, and end sequencing with BLAST. When applicable, all counts are rounded up to the nearest integer. The preparation of the MIPs introduces relatively little bias in their uniformity, and the end sequencing is greatly improved from Porecca et al. The GAII shotgun shows a slight loss in uniformity, and the Polonator Shotgun Library shows limited complexity, presumably due to a bottleneck in library production. **B. Copy Number Variation.** To estimate the sensitivity of MIP capture to copy number variation we compared the X-chromosome capture between male and female participants. Capture data for each participant was normalized, and the average read depth of ~11,000 autosomal targets and ~200 X Chromosome targets with at least 10 raw reads are shown. The error bars show the standard error for each data point. A comparison between the female (red) and male (black) samples show a marked decrease in read depth when only one chromosome is present. **C. Allele Frequency vs. Read Depth.** MIP capture has also proven successful at discriminating between heterozygous and homozygous calls, acheiving an even ratio of both reference and non-reference calls for heterozygous positions with sufficient coverage. Positions in which the zygosity is confirmed by independently derived microarray-based genotyping are marked with diamonds and when they are in disagreement the calls are portrayed with X-es.

Analysis of EcoP15i Digestion

To minimize the amount of non-captured material in the libraries (see Figure 3-1 step 5) we removed most of the 20bp synthetic arms and universal primers through enzymatic digestion with EcoP15i. This enzyme cleaves 27bp downstream from its recognition site, and its activity is improved when two recognition sites are present in a head-to-head orientation[21]. Additionally, there have been contradictory reports as to whether sinefungin improves digestion[22-23]. Since our universal backbone did not contain the recognition site for EcoP15i, tailed-PCR primers were used to append this sequence to either end of the library molecules. Although no restriction enzyme is reported to require that the recognition site be more than six nucleotides from the terminus, we have found that even 10nt was insufficient for cleavage by EcoP15i; we positioned the recognition site 20nt from the terminus. We initially considered gel extraction of the expected size after the digestion, but an analysis of our targets revealed that 12% of the targets contained the enzyme recognition site. These molecules would potentially be cleaved more than once and be lost in the gel extraction due to their shorter length. We decided, therefore, to not use gel extraction; instead we used biotinylated primers and streptavidin beads to remove the cut ends. This would also enrich for doubly-cut molecules as any molecule with a single uncut side would be removed. Remarkably, analysis comparing targets containing the additional recognition site to all other targets showed that the presence of an additional EcoP15i site in a capture region increased average capture 3.7-fold. While it is possible that the additional chance for the enzyme to cleave would yield this result, we hypothesize that the 20bp distance is insufficient for efficient cleavage and positioning the recognition site even further from the terminus

would aid EcoP15i digestion. For an overview of the process and the successful removal of universal sequence and half of the synthetic arms, see Figure 3-4. We generated ten shotgun libraries through concatenation of EcoP15i cleaved fragments and four paired end libraries via a protocol modified from Porreca et al[19]. These latter libraries produced for the Polonator proved to be of insufficient complexity to generate much information.

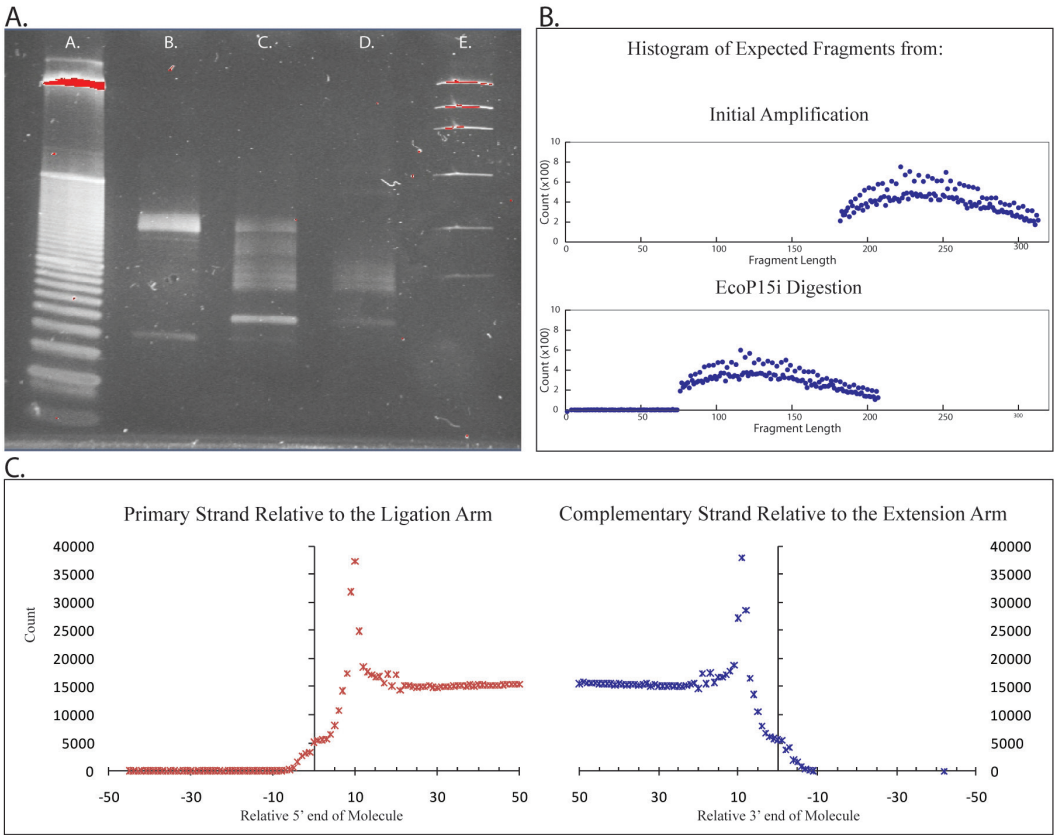


Figure 3-4. A. Polyacrylamide gel of target capture and EcoP15i digestion. Lane A: 10bp ladder; Lane B: 20ng amplified MIP capture product with the bulk of material at 180-320bp and primers present at 40bp; Lane C: 20ng EcoP15i digested product; the uncut material is still present at the original size; Lane D: 20ng streptavidin bead extracted product - the uncut product has been removed with the streptavidin coated beads, along with most of the primers; Lane E: Invitrogen Low Mass Ladder. **B. Expected size range of EcoP15i digestion.** The expected size range of the amplified targets is 180bp - 320bp. EcoP15i digestion is expected to produce primarily a range from 80bp - 210bp. Targets with additional enzyme recognition sites will generate a long tail ranging from 1-80bp. **C. Targeting specificity and EcoP15i digestion.** Each point represents the 5' end (red) or 3' end (blue) of a library molecule; the relative position 0 represents the 5' end (left) and 3' end (right) of each target (including the 20bp MIP arms). The relative dearth of molecules beginning in the first 10bp of each end is due to the successful removal of these regions. Molecules that do place in these regions are likely an artifact of chimeric sequence and short read lengths. The large number of molecules 10bp from the termini are due to the sequencing primers ligating to unconcatenated library molecules (see Methods for more detail).

Reference Placement and Quality Analysis

We acquired one lane of Illumina GAII sequencing (two lanes for PGP8 and PGP9) for each of the shotgun libraries and developed a web-based tool which we call Genomerator (http://genomerator.freelogy.org), for generating consensus sequence and high quality variant lists. This application consists of three pipelines. The first takes as input Next Generation Sequencing reads along with their quality scores, a reference sequence, and, optionally, three thresholds: minimum variant consensus quality, minimum read depth and minimum independent reads. For this project the program used MAQ as the alignment algorithm. In addition to creating the standard MAQ-generated files, it creates a general feature format (http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml) file of all variants conforming to the desired thresholds. Optionally, this gff file can be tailored to contain the genotypes at specific positions (irrespective as to whether they are variants), or to contain variants from particular genomic regions.

The second pipeline is designed for microarray data. This pipeline takes as input raw microarray data, a reference genome, probe margins, and, optionally, a list of genomic regions to limit the output to. Currently, this pipeline is optimized for Affymetrix data, and is processed using the dynamic model (DM) algorithm[24]. The output consists of a BED file with the genotype calls.

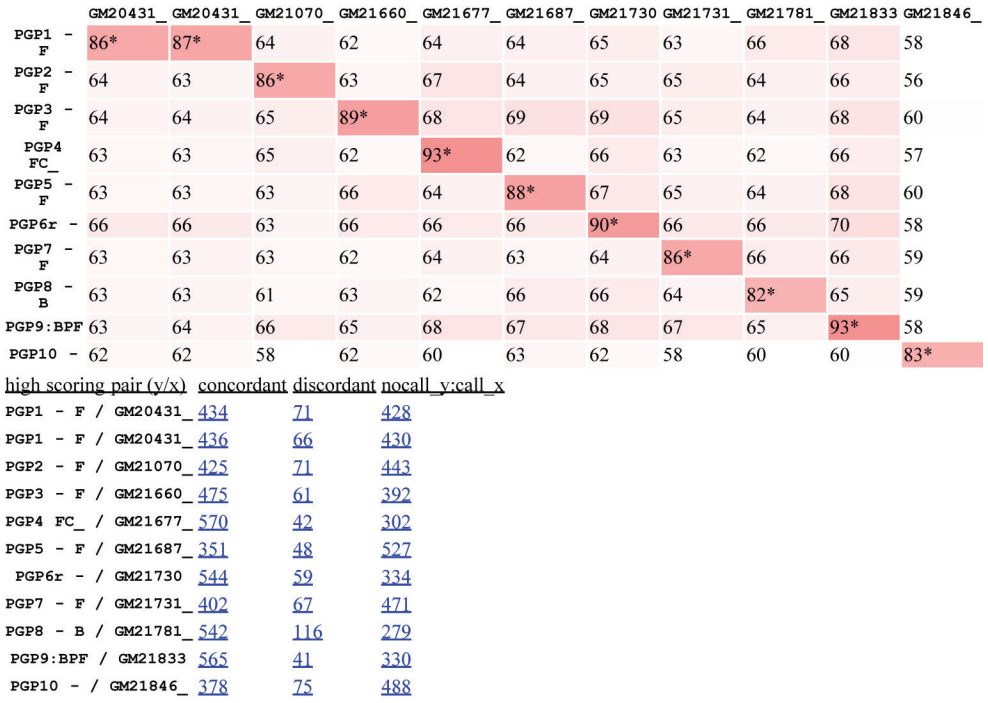
The third pipeline compares the results from two runs. This can be a comparison of two different sequencing runs, two different microarray sets or a microarray set and sequencing run. The output for this pipeline consists of three files for each input pair: a file of concordant calls, a file of discordant calls and a no-call file where one input file

contained a call but the second did not. Additionally, if presented with numerous input files of one type, it can display the concordance ratio for each of them and display further details for the pair(s) with the highest concordance.

Sequencing of the ten libraries generated two Gbp of raw data. 59% of reads aligned uniquely to the HG18 reference — without insertions or deletions and up to two mismatches using MAQ 0.6.7. Of the 1.188 gigabases of placed reads, 873 Mbp placed against the target regions when placement was allowed against the entire genome. Despite the 6.7Mbp target region, the average sequenced region was 5.67Mbp (4.31-7.08), implying an average coverage of 154x. While the non-uniformity of capture efficiency highlighted in Figure 3-3A complicates this claim, nevertheless on average, greater than 25,000 targets (3Mbp) were covered by at least 10 36bp reads for a minimum average coverage of 3x, and over 18,000 targets (2.2Mbp) were covered by at least 30 36bp reads for a minimum average coverage of 9x. To evaluate the frequency of misplacement caused by the MAQ algorithm with short reads, we processed a complete *in-silico* shotgun library of our target set through this pipeline. We found that only 3.7% of targets had less than 90% of their *in-silico* reads correctly mapped, and 93.6% of targets had greater than 99% of their reads correctly identified. Due to the high placement accuracy of this algorithm with 36bp reads from our target set, we did not perform any additional corrections for potential misplacement of reads.

Having previously obtained Affymetrix 500K microarray data for each of the participants, we compared the sequencing generated list with the list of all microarray-based genotypes in the target region. To create the sequencing-based list we used the optional SNP list feature in Genomerator to generate the sequence-based genotype for all

positions sampled on the microarray (irrespective as to whether they were variants) and used the third pipeline to compare these results with the microarray data (Figure 3-5). Based upon this data, without applying any thresholds, the average percentage of microarray-genotyped positions that were also sequenced is 57%, which is a more accurate assessment of the target coverage. The overall raw accuracy, however, was only 87.7%. Even with this poor raw accuracy, it is remarkable that there exists a clear distinction between the PGP participants of European descent (PGP1-PGP9) and the participant of African descent (PGP10). Additionally, one is able to re-identify each of the participants, which further reinforces concerns of privacy and data protection.



	GM20431_	GM20431_	GM21070_	GM21660_	GM21677_	GM21687_	GM21730	GM21731_	GM21781_	GM21833	GM21846_
PGP1 - F	86*	87*	64	62	64	64	65	63	66	68	58
PGP2 - F	64	63	86*	63	67	64	65	65	64	66	56
PGP3 - F	64	64	65	89*	68	69	69	65	64	68	60
PGP4 FC_	63	63	65	62	93*	62	66	63	62	66	57
PGP5 - F	63	63	63	66	64	88*	67	65	64	68	60
PGP6r -	66	66	63	66	66	66	90*	66	66	70	58
PGP7 - F	63	63	63	62	64	63	64	86*	66	66	59
PGP8 - B	63	63	61	63	62	66	66	64	82*	65	59
PGP9:BPF	63	64	66	65	68	67	68	67	65	93*	58
PGP10 -	62	62	58	62	60	63	62	58	60	60	83*

high scoring pair (y/x)	concordant	discordant	nocall_y:call_x
PGP1 - F / GM20431_	434	71	428
PGP1 - F / GM20431_	436	66	430
PGP2 - F / GM21070_	425	71	443
PGP3 - F / GM21660_	475	61	392
PGP4 FC_ / GM21677_	570	42	302
PGP5 - F / GM21687_	351	48	527
PGP6r - / GM21730	544	59	334
PGP7 - F / GM21731_	402	67	471
PGP8 - B / GM21781_	542	116	279
PGP9:BPF / GM21833	565	41	330
PGP10 - / GM21846_	378	75	488

Figure 3-5. Initial Genomerator Concordance. This screenshot from the Genomerator concordance pipeline shows the results from comparison of genotypes called by the Affymetrix 500K Microarray with those called by MIP targeted sequencing. Sequence data is presented on the y-axis and microarray data on the x-axis. The Affymetrix data is filtered to sample only the approximately 1000 calls appearing in the target region. Two microarrays were processed for PGP1 and the concordance is displayed for both. All sequence-based genotypes were used for each participant with the exception of PGP6 who redacted some genotypes associated with Alzheimer's risk. Of note is both the relatively high discordant rate, which nevertheless does not impede the re-identification of the participants. Also of interest is the higher similarity between the individuals of European descent when compared to PGP10 who is of African-American descent.

Analysis of these placed reads, however, showed that a large fraction (50%-90%) of sequencing reads start from the same position as another read (Figure 3-6A). The effect of PCR duplicates has been noted by other groups, with reports as high as 60% of reads being from duplications after a bottleneck[25]. The preponderance of possible PCR duplicates gives an additional reason why the calculated 154x mean coverage is inaccurate. It also complicates the strict utilization of coverage cutoffs in demonstrating confidence that a variant is indeed correct. To choose thresholds permissive enough to retain much of the variant information, but sufficiently restrictive to remove most false positives, we processed different permutations of the three thresholds enabled by Genomerator. These variant lists were then compared to microarray data for each individual with the goal of achieving 99% concordance, the maximum self-concordance of the Affymetrix DM algorithm (http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf). To first confirm the accuracy of the Affymetrix DM algorithm, we compared the results of the two microarrays performed on the PGP1 DNA. We found that for the ~1000 genotyped positions contained in the target region 911/913 positions were concordant (99.8%) with 39 positions (4.1%) successfully genotyped in only one of the two runs. This closely matched the overall concordance for all positions, where 475,214/476,925 (99.64%) positions were in agreement, and 19,673 (3.96%) of genotypes appeared in only one of the two runs. While these two runs point to accuracy higher than 99%, we retained the documented accuracy for further analysis.

Comparing our filters to the respective microarray data, we found that different combinations of coverage and minimum unique start sites (to minimize the effect of

PCR-duplicates) were generally insufficient to achieve 99% concordance even with very high thresholds. Specifically, we applied the minimum unique start site filter to all variant calls, and the coverage filter to all calls. Sampling combinations of up to ten unique start sites and 50x coverage we found that on average, the highest concordance, 98%, was achieved with a minimum threshold of six start sites and 30x coverage. This combination of filters retained ~17% of all variants called (Figure 3-6C). A detailed look, however, showed that disparities between the individual libraries would complicate the analysis. The most highly concordant library, PGP4, achieved 99% concordance with 4x coverage and six independent reads while retaining 73% of its variants. The least concordant (PGP8), however, could not obtain a concordance above 91% with six independent reads even with 30x coverage, and even then only 20% of variants were retained (Figure 3-6B). While this is partially due to the relaxation of the independent read threshold for non-variant calls, the primary reason is a few reads with a large number of low quality consensus calls in this library (Figures 3-7A and 3-7B).

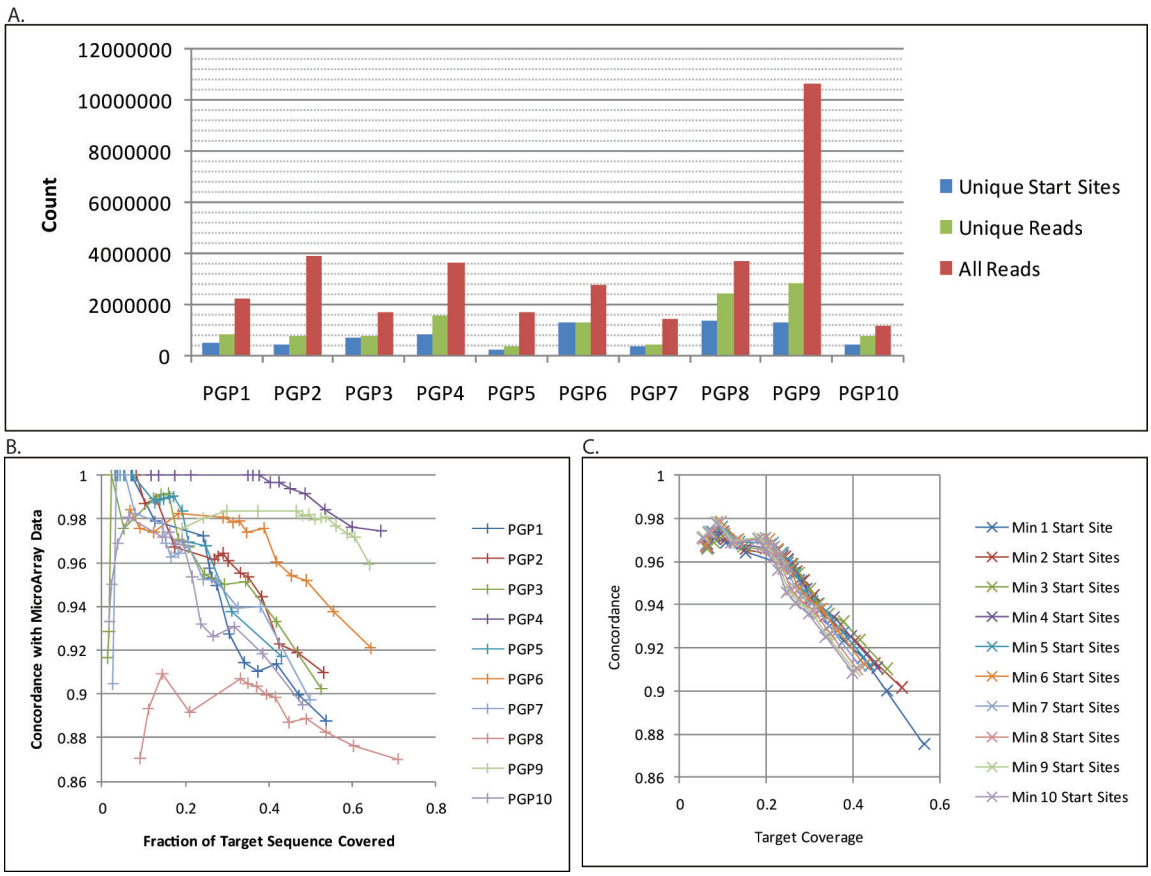


Figure 3-6. Independent Read Analysis. A. Shotgun Read Composition. Each library represents one lane of Illumina GAII sequencing, with the exception of PGP8 and PGP9 for which we have combined data from two lanes. Bottlenecks in library construction led to the preponderance of PCR duplicates. With a target space of 6.7 Mbp, we would expect that a large fraction of our reads would be obtained from molecules initiating from different positions. Instead we find 4e5-1.2e6 unique start positions for each library sequenced, or 10%-50% of all reads for a given library. While it is not certain that that reads sharing the same start site are indeed PCR duplicates, it is likely that a large fraction of them are. **B. Microarray Concordance Using Minimum Start Site and Coverage Thresholds.** Each of the PGP10 shotgun libraries were analyzed with different thresholds and the results were compared with microarray data. In this plot all variants are based upon at least six independent library molecules (judged by the number of start sites), while calls matching the reference are not subject to this threshold. The right-most marker in each line shows a coverage of 1 and successive markers depict greater coverage (one marker each for 1-10, followed by 20, 30, 40 and 50x coverage). Both reference and non-reference calls are subject to this threshold. While PGP4 obtains 100% concordance with 8x coverage and almost 40% of its targets sequenced, PGP8 never achieves greater than 91% concordance. **C. Effect of Independent Read Threshold on Concordance.** To estimate the threshold for independent reads that would provide the largest increase in concordance with the smallest decrease in allowed variants we sampled the effects of allowing variants with a minimum of 1-10 independent reads. For each independent read threshold we generated a curve marking read depths of 1-10, 20, 30, 40 and 50x. The greatest increase in concordance is gained through requiring three independent reads.

Transitioning to a flexible threshold based on both MAQ consensus quality and coverage enabled a concordance of 99% with the microarray data. Additionally, the fraction of variants retained by these thresholds was typically comparable to that retained

by the unique start site/coverage thresholds highest concordance level. With these cutoffs, both PGP4 and PGP8 were able to achieve 99% concordance while retaining 67% and 21% of raw variants. Further analysis of the fraction of novel non-synonymous SNPs in these variants, however, proved to raise suspicion as to the number of false positives permitted by this filter. In dbSNP build 129 there are 34,560 single-nucleotide substitutions in the target region, or 0.52% of all base-pairs. Initial whole-genome and whole exome sequences have been very consistent in identifying 10% of the variants in each genome as novel (see Figure S2-3), and, for these genomes, the target region does not deviate from this trend. The thresholds above, however, produce variant lists with an average of 42% substitutions not in dbSNP (Figure 3-7C), and barring PGP8, every variant list had more than 29% of their substitutions not appearing in dbSNP. Furthermore, when the variant list was analyzed with Trait-o-matic to prioritize potentially deleterious rare variants, 13/28 variants reported had an allele with only one or two independent reads supporting it, raising suspicion as to their being false positives. To determine the number of independent reads that would provide the biggest increase in concordance with minimal variant loss, we compared the effects of requiring 1-10 independent reads to confirm each variant, and found that three independent reads provided the greatest increase in concordance with the smallest loss of total variants called (Figure 3-6C). We proceeded to develop a new threshold based upon all three criteria. We calculated new consensus quality / read depth thresholds using only variants based upon at least three independent reads, where independence is defined as having a unique start position for the molecule. The fraction of novel nsSNPs was half of that reported when no unique start site threshold was required; while this was still

significantly higher than that reported by other sequencing efforts, we found that the expected number of false positives was sufficiently low to not hinder downstream analysis (Figure 3-7C). Although the overall data looks promising, we were unable to account for the large variability between different libraries (Figure 3-7D), and this process will require greater fine-tuning in future efforts.

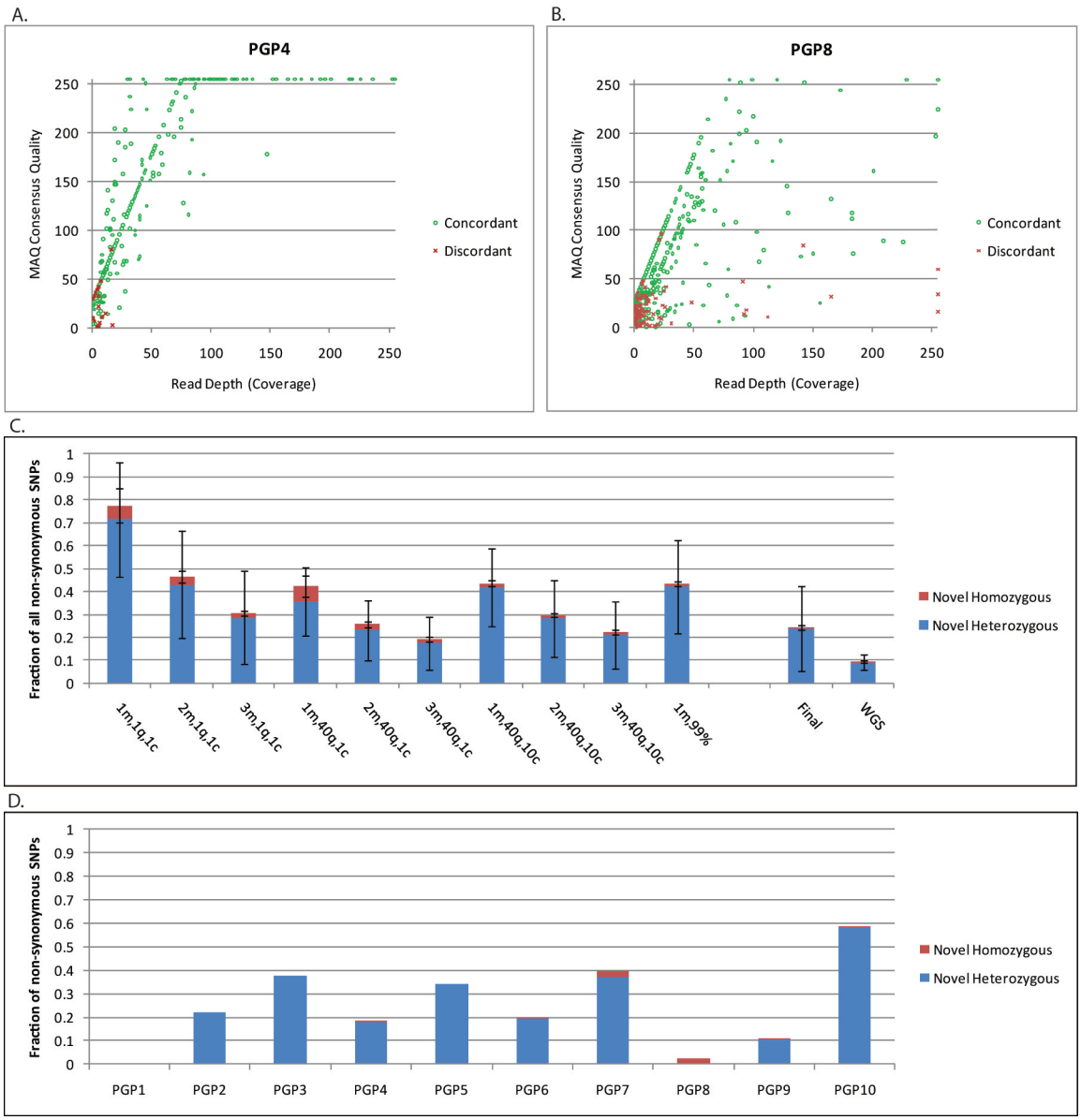


Figure 3-7. A-B. Comparison of Quality/Coverage Thresholds from Two Libraries. The shotgun libraries from PGP4 and PGP8 highlight the utility of the MAQ consensus quality threshold. Both samples have the bulk of discordant calls with a quality score below 50, while the coverage of such calls vary greatly between the two libraries. The default MAQ consensus quality threshold is 40. **C. Stringency Cutoffs**. Sequencing efforts have adopted different cutoffs for quality control. We compare the effects of different filters entailing minimum start sites (m) to control for PCR duplicates, together with MAQ consensus quality (q) and read depth (c). Additionally, we compare the 99% concordance with microarray data when no threshold is established for minimum independent reads. The average of 18 whole genome and whole exome sequencing efforts have been remarkably consistent in reporting 10% of nsSNPs in the target region as being novel heterozygous variants, and 0.5% novel homozygous. With our libraries we have found that requiring numerous start sites was the most important feature in minimizing false positives. The final cutoffs required three start sites for every variant called, along with a read depth and consensus quality cutoff that fulfilled the requirement for a 99% concordance with independently derived genomic data. While this data still appeared to contain a large number of false positives, our prioritization scheme for deleterious variants allowed us to manage the data effectively. Error bars represent one standard deviation. **D. Novel Heterozygous/Homozygous Ratios for Individual Libraries**. The threshold levels for each library was optimized to reach 99% concordance with independently derived genotype information. Nevertheless, large deviations from the expected ratios exist in a non-uniform manner across all libraries. This is likely the result of inconsistent library production.

As an additional estimate of false positive rates we compared the frequencies of different nucleotide substitutions causing the non-synonymous SNPs. Corroborating previous research, we found significantly more transitions than transversions[26-27]. While fully random errors would find a ratio of 1:3 for transitions to transversions, the relatively frequent mutations of cytosine to thymine through depurination accounts for the abundance of transitions. This was found to be the case for the variants identified in our libraries when utilizing this three-part threshold (Figure 3-8A). Remarkably, we find that for the whole genome sequences (and for two of our libraries) it is more likely for the reference allele to be the non-ancestral allele (Figure 3-8B and 3-8C, see also Figures S2-3.2 and S2-3.3). Additionally, we find that mutations between adenine and thymine are the least likely to occur.

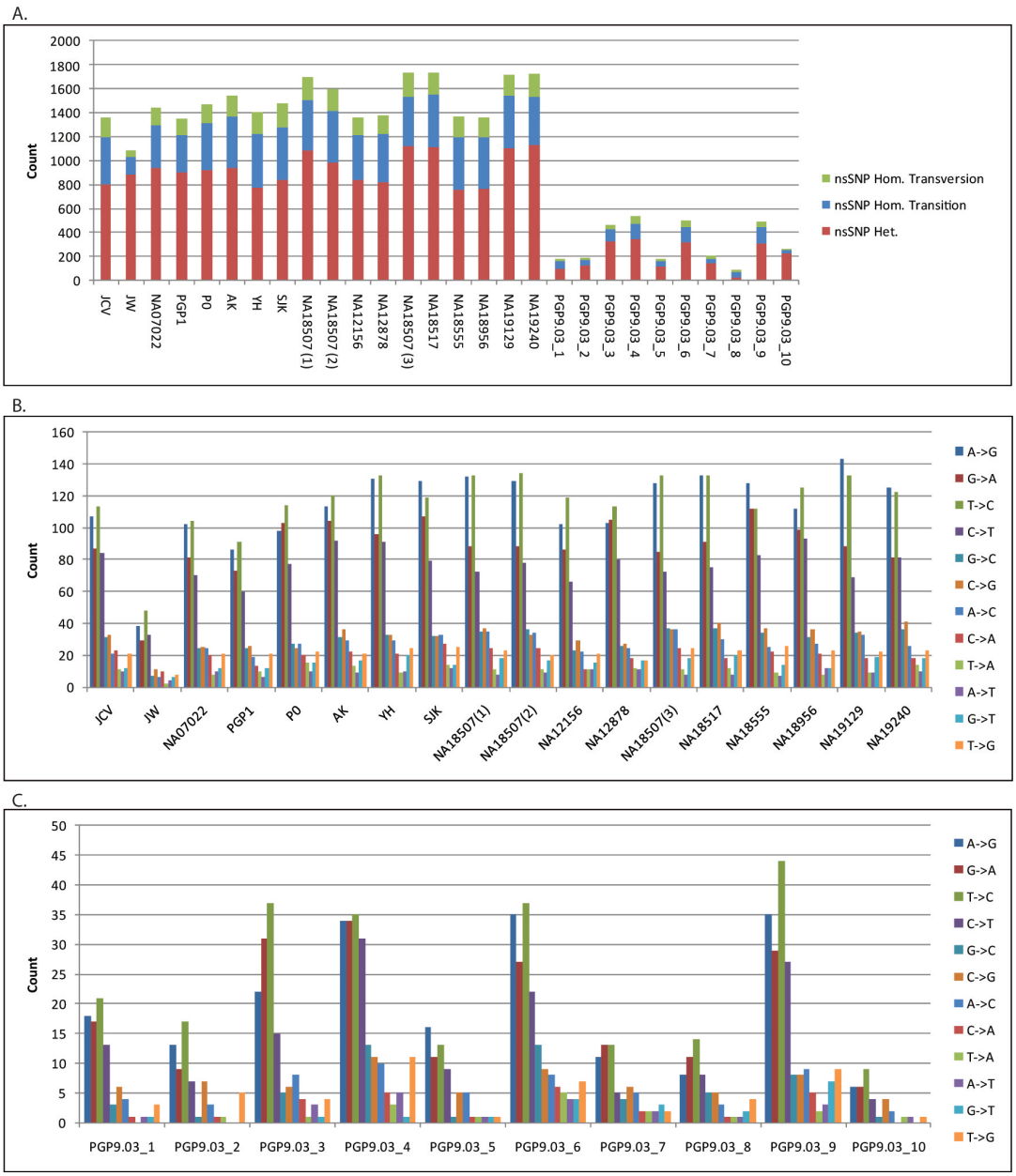


Figure 3-8. SNP Composition and Comparison with Other Sequencing Efforts. A. nsSNPs Reported. Despite collecting 2Gbp of raw sequence data, quality cutoffs and control for PCR duplicates limited the number of variants reported with high confidence. The makeup of these non-synonymous SNPs conform to the heterozygous/homozygous and transition/transversion makeup of this region as reported by published genome sequencing efforts utilizing Sanger, 454, Illumina, SOLiD, Complete Genomics and Agilent Capture Arrays. NA18507 (1) is from Bentley et al., NA18507 (2) is from McKernan et al., and NA18507 (3) is from Ng et al. These data were downloaded from their respective sources as described in Chapter 2. Overall, we report an average of 20% of the expected SNPs for this region for each individual when compared with other sequencing efforts. Since this region represents ~17% of all exons, the sequence space of the ten libraries represents the equivalent of 35% of one full exome sequence. **B. Single Base Substitution Details for Other Sequencing Efforts.** All ethnicities show an enrichment for transitions over transversions. Remarkably, these data suggest that the mutation is often towards the ancestral allele with the reference genome showing the mutation caused by the depurination of cytosine. **C. Single Base Substitution Details.** For the ten libraries presented here the transition/transversion ratios similarity to those of other efforst suggests high accuracy.

High Confidence Data

The final 99% concordance with microarray data was based upon an average of 225 positions called in both the microarray and sequence data (range: 86 – 418). The overall accuracy for heterozygous calls was 356/361 (98.6%), and for homozygous calls it was 1453/1471 (98.8%) for reference and 352/355 (99.2%) for non-reference calls. Extrapolation of the call rate of these 1000 positions to the entire 6.7Mbp target region leads to an estimate sequencing of 1.61Mbp/participant (range: 0.6-3.0Mbp) (Table 3-1). This fraction represents the top 28% of our raw data, and the data from all ten libraries would be the equivalent of one half-exome. Comparison to the amount of non-synonymous SNPs expected from the target region, however, point to a slightly lower overall sequence space of 35% of one exome. The reason for the large differences in coverage between libraries was partly due to the loading density of each sequencing lane, but mostly due to the limited complexity of a few of the libraries.

Table 3-1. Thresholds, Estimated Target Size and SNPs for each Library. The first two columns show the thresholds applied for library. The Fraction of Microarray Positions Called shows the fraction of independently derived genotypes from the sequence space that was detected through sequencing and the Estimated Callable bp is the extrapolation of the microarray data to estimate the fraction of the target sequenced. The final column lists the number of SNPs identified for each individual.

Sample	Minimum Consensus Quality	Minimum Coverage	Fraction of Microarry Positions Called	Estimated Callable bp	SNPs
PGP1	20	11	0.22	1.47	520
PGP2	49	35	0.12	0.77	415
PGP3	15	6	0.29	1.95	1045
PGP4	37	1	0.45	3.03	1569
PGP5	30	7	0.15	1.01	401
PGP6	9	6	0.38	2.55	1262
PGP7	10	9	0.15	0.99	449
PGP8	61	22	0.15	0.99	238
PGP9	55	1	0.46	3.09	1360
PGP10	19	14	0.10	0.63	551
Average	**30.5**	**11.2**	**0.24**	**1.65**	**781.00**

Overall, 7,412 SNPs were called; 658 of these, however, were in the MIP arm regions and do not accurately reflect the genotype of the individual. These 658 SNPs are present either as a result of synthesis error or chimeric sequence resulting from library

construction. While 68 of the remaining 6,754 SNPs were off-target calls (1%), potentially an artifact of the placement algorithm, the average on-target capture was 96.9% for each of the ten libraries (range: 91.9 – 99.8%). For the remaining 6,686 substitutions a bed file was generated with the calls from all ten individuals for those positions, for a total of 16,267bp. These substitutions represent 3,110 unique positions of which 40% are non-synonymous and 827 novel.

Trait-o-matic Analysis

We analyzed these variant lists with the Trait-o-matic program to prioritize variants with possible clinical utility. This tool converts each variant to its corresponding amino acid change and matches to variants listed in OMIM, HGMD, SNPedia and PharmGKB, prioritizing clinically important variants. For details of this process see Chapter 2. 18 variants were flagged as being rare, potentially clinically important variants in these ten individuals. Unexpectedly, despite the reported rarity of these variants, four of them had been found in previous whole genome sequences (*PIGR* A580V, *APOA5* S19W, *C3* L314P, and *CAPN3* T184M) and one (*MSH5* P803S) was seen in two PGP individuals. Sanger sequencing was performed on these loci, and one (*BRCA1* H1564P) proved to be a sequencing error. The literature was searched for more information on each of these variants, and these data and recommendations for follow-up are listed in Table 3-2. A description of the follow-up of the *MYL2* A13T variant is described in Chapter 2, as well as further information for the *POMGNT1* and *CAPN3* mutations where they are reclassified as likely benign (Supplemental Discussion 2-1). In addition to these variants, it was interesting to find two variants in *OCA2* in PGP4. Both of these variants are reported as very rare in the European population, but there has been no specific study

of the Ashkenazi Jewish population, the ethnic group to which PGP4 belongs. We look forward to analyzing more genetic material from this individual to see whether his genome contains more SNPs usually associated with Asian ethnicity. To understand the source of our one false positive, we compared its raw reads to that of *MYL2* A13T, and found no apparent way to differentiate between it and a true variant.

Table 3-2. Analysis of Variants Prioritized by Trait-o-matic. We analyzed the filtered variants with Trait-o-matic to prioritize those that were potentially of clinical interest. 18 rare variants were prioritized, and 17 were confirmed by Sanger sequencing, with one being a sequencing error. The literature was searched for more information on these 17 variants, and the conclusions are present in this table. HapMap Frequency reflects the frequencies calculated by the Trait-o-matic tool, found at http://snp.med.harvard.edu. CAF Ethnic Groups shows the causative allele frequency for the ethnic group the individual associates with using a weighted average of HapMap, 1000 Genomes and literature data. (low-high) shows the lowest and highest frequencies for this variant in all ethnicities.

PGP	Gene dbSNP ID	Geno-type	Position	Literature PMID	HapMap Frequency	Present in Genetests?	CAF Ethnic Group (low-high)	Phenotype	Conclusion
2	*PON1 L55M* *rs854560*	A/T	Chr7:94784020	11335891 9011577 10669651 15028278 10811591 10856521	Unknown	No	0.38 (0.022-0.4)	PON1 activity is important is reducing oxidation of LDLs and HDLs and the prevention of atherosclerosis, and this variant (with LD) accounts for ~15% of the variance in expression	Expression of the variant is highly variable and it is very common in the individual's ethnic background
3	*NPPA V32M* *rs5063*	C/T	Chr1:11830235	16368448 15028278 10525492	Rare	No	0.048 (0.019-0.126)	With LD this variant has OR of 2.0x for stroke in Caucasians and is associated with protection against essential hypertension and pharmacogenetic effects with Irbesartan in Chinese	Association studies with moderate correlation with increased stroke and heart disease
4	*ADCY6 A674S* *rs3730071*	C/A	Chr12:47455065	15903125 17916776	Rare	No	0.045 (0-0.045)	The variant shows a 2x increase in ADCY6 activity in Rat vascular muscle	Unknown human phenotype
4	*PIGR A580V* *rs291102*	A/A	Chr1:205173101	12740691	Rare	No	0.021 (0.019-0.846)	Association with Immunogobulin A Nephropathy with OR of 2.71x	Association study with a not-identified mechanism
4, 6	*MSH5 P803S* *rs1802127*	T/T	Chr6:31837904	16574953 17409188	Rare	No	0.014 (0-0.156)	OR of 0.62x with chronic lymphocytic leukemia. Also, significantly associated with IgAD and CVID, but with incomplete penetrance	Variable penetrance, in dissociation with CLE and association with IgAD and CVID.
4	*N/A* *rs7951*	A/G	Chr19:6632991	18174230	Rare	No	0.045 (0.045-0.16)	OR of 1.4x for systemic lupus erythematosus and statistically lower C3 serum levels in Japanese population	Association with SLE in Japanese population
4	*OCA2 A481T*	C/T	Chr15:25902148	8302318 8980282 17568986 17236130 15942220	Unknown	Yes	~0 (0-0.24)	This variant, while very rare in the individual's population is very frequent in NE Asia. It reduces gene expression to 70% of wild type, and homozygus individuals may be at a risk for subclinical ocular albinism in the Asian population	Slight decrease in gene expression, but no strong association with disease
4	*OCA2 R305W* *rs1800401*	G/A	Chr15:25933648	17236130 12163334	Unknown	Yes	0.046 (0.021-0.068)	Dissociation with blue/grey eyes and brown black hair	Eye/hair color association
6	*APOA5 S19W* *rs3135506*	G/C	Chr11:116167617	12417524	Unknown	No	0.058 (0-0.15)	This variant is associated with plasma triglyceride concentrations above the 90th percentile independent of dietary regimen	Association with plasma triglyceride concentrations above the 90th percentile
6	*BCL10 A5S* *rs12037217*	C/A	Chr1:85514611	16229939	Unknown	No	0.068 (Japanese)	OR of 6.25 in progressing from stage I to stage II/III in Japanese testicular cancer patients	Strong association with testicular cancer progression in Japanese population
6	*C3 L314P* *rs56326450*	G/A	Chr19:6664262	18325906	Unknown	Yes	0.22 (Caucasian)	This variant is in LD with a variant providing a population attributable risk of 0.17 to age-related macular degeneration	In tight LD with a minor cause for AMD
6	*MYL2 A13T*	C/T	Chr12:109841347	8673105 11748309 15483641 14594949 11102452 12668451	Unknown	Yes	0.0057 (0-0.0057) (mixed pop.)	Two different case reports view this as a causative mutaiton for HCM, but one of those studies did see compound heterozygosity. Three different *in vitro* studies confirm the abberant behavior of the mutant protein	Likely pathogenic for HCM, this variant should be followed up
9	*KEL T193M* *rs8176058*	G/A	Chr7:142365130	7849312 13705100	Rare	Yes	0.023 (0-0.045)	Immune Response to K_2 genotype in homozygotes, 5% of blood transfusions to them generate anti-K antibodies. There is also the risk of hemolytic disease of the newborn	Kell K_2 heterozygote
9	*PCSK9 R237W*	C/T	Chr1:55290962	16465619 15358785	Unknown	Yes	0.24 (0.028-0.24)	Not significantly associated with high LDL (>95%)	No statistically significant correlation with disease
9	*POMGNT1 D556N*	C/T	Chr1:46428232	17878207 18691338	Unknown	Yes	0.037 (French)	Case report associates this variant with Limb Girdle Muscular Dystrophy without Mental Retardation, but the report also found a homozygous sibling of an affected heterozygote that was not affected	No proven phenotype, possibly incomplete penetrance for LGMD without mental retardation
10	*CAPN3 T184M* *rs35889956*	C/T	Chr15:40467295	10330340	Unknown	Yes	0.013 (mixed pop.)	Case report associates this variant with classical, recessive Limb Girdle Muscular Dystrophy in one family from Réunion Island	Potentially recessive variant in the Caucasian population, of unknown effect in the African population

We present here the first data release from the Personal Genome Project and a discussion of some of the tools being developed to aid the analysis and presentation of the data from the PGP. We envision releasing new data every six months and we hope that the tools developed here will help both the scientific and PGP communities further our understanding of the correlations between genomes and phenomes.

Methods

Illumina GA Sequencing of Probes

Generation of padlock probes. Using a programmable microarray (Agilent Technologies), 55,000 oligos (110-mers) were synthesized, cleaved off, and collected in a single eppendorf tube. Each of the oligo species is approximately 0.2 fmol, totaling ~10 pmol of oligos synthesized on one array. The sequence of the 110-mer oligo is ATCAAGCCGAAGACAGTGTT[ligation_arm]CTTCAGCTTCCCGATATCCGAC-GGTAGTGT[extension_arm]GATCCAGGAAATTCGCGCTA. To amplify the probe set from Agilent, a 100 μL reaction was assembled containing 2.5 units of Platinum Taq Polymerase (Invitrogen), 1x Platinum Taq buffer, 200 nmol of each dNTP, 1.5mM $MgCl_2$, 0.5x SybrGreen (Invitrogen), 1nM Template (From Agilent), and 500nM each forward and reverse primers (Integrated DNA Technologies). The forward primer was AP1 (A*T*C*AAGCCGAAGACAGTGT/3deoxyU/ *:phosphorothioate bond) and reverse was AP2 (/5Phos/TAGCGCGAATTTCCTGGATC). Thermocycling was carried out on the MJ Research Opticon 2 Real Time DNA Engine and involved a 5m denaturation at 95°C, followed by 9 cycles of 95°C (30s), 58°C (1m) and 72°C (1m), and a final extension of 72°C (5m). The product was purified over a single Qiaquick column (Qiagen)., and the quantity of product was assessed using Nanodrop-100 spectrophotometer. One ng of the product was amplified via the above PCR protocol for 10 cycles using the following adapters: AP1-SLXA (AATGATACGGCGACCACCGAATCAAGCCGAAGACAGTGT, IDT) and AP2-SLXA (CAAGCAGAAGAGGCATACGATAGCGCGAATTTCCTGGATC, IDT). The product was cleaned over a Qiaquick column (Qiagen) and then gel-excised from a 6%

TBE polyacrylamide gel. Sequencing was performed on one lane of Illumina-GAI using the custom primer: TAGCGCGAATTTCCTGGATC. Custom scripts were used to match these sequences to the reference sequence.

MIP Probe Preparation

To create large amounts of capture probes from the unfinished amplified probes, 3 96-well plates of 100 µL PCR reactions were assembled. Each PCR contained 1x Platinum Taq Supermix (Invitrogen), 500nM each AP1 and AP2 primer, 0.5x SybrGreen (Invitrogen), and 20 pM template (from the amplification in the previous section). 12 cycles of PCR were performed using the above conditions on a thermocycler (MJ Research DNA Engine). The reaction volume was concentrated using ethanol precipitation, followed by purification with 12 Qiaquick Columns (Qiagen) and elution in a total of 600 µL of buffer EB (Qiagen). Quantification of DNA (Nanodrop-100 spectrophotometer) showed a recovery of 875 pmol. 66 µL of Lambda Exonuclease buffer (NEB) and 11U of enzyme (NEB) were added and the reaction was incubated 45m at 37°C, followed by 10m at 75°C to heat-inactivate the enzyme. The sample was purified over 7 Qiaquick columns (Qiagen), and eluted in 350 µL of buffer EB (Qiagen). Recovery was estimated at 390 pmol. 17.5U of USER enzyme (NEB) was added and incubated for 30m at 30°C. DpnII buffer (NEB) was then added to 1x final concentration along with 4nmol of Guide_DpnII (GCGCGAATTTCCTGGATC, IDT), and the sample was heat denatured 5m at 95°C, ramped at a rate of -0.1°C/s to 60°C where it was held for 3m. 112U of DpnII (NEB) was added and the sample was incubated for 15m at 37°C. An additional 10U of USER was then added for an additional 30m digestion at 37°C. The correct band (70bp) was gel excised from a 6% TB-Urea PAGE

gel. The gel slabs were crushed and suspended for 2h at 55°C in buffer EB with periodic mixing, and then filtered through Nanosep 0.2 μm columns (Pall). The material was then ethanol-precipitated and resuspended in 100 μL of buffer EB. Total recovery was estimated to be 75 pmol of prepared probes.

Paired Tag Library for Polonator Sequencing

Target Capture

For target capture, a 20 μL mixture containing 2.6pmol of probes and 3 μg of gDNA in 1x AmpLigase buffer (Epicentre) was assembled. The solution was denatured for 5m at 95°C and then allowed to hybridize for 40 hours at 60°C. No probe controls were prepared for each sample, and a single no-gDNA control was assembled. After 40h, 20 pmol each dNTP, 2U Taq Stoffel Fragment (NEB) and 2.5U AmpLigase (Epicentre) were added to each reaction. The solution was maintained for 10 hours at 60°C, and then cycled 10 times as follows: 5m at 95°C, 1 hr at 60°C, with a final extension of 60°C for 10h. After the incubation, non-circularized DNA was digested through the addition of 6.0 μL of a solution containing 3x ExoI Buffer (NEB), 40U of ExoI (NEB) and 200U of ExoIII (NEB). Digestion was performed for 1hr at 37°C followed by 15m at 80°C to heat inactivate the enzymes.

Exon Amplification

The following 50 μL reaction was assembled to amplify the captured exons (each sample in triplicate): 300fmol of each primer CP-2-FA (GCACGATCCGACGGTAGTGT, IDT) and CP-2-R CCGTAATCGGGAAGCTGAAG, IDT), 200nM each dNTP (Invitrogen), 0.3x SybrGreen (Invitrogen), 1U Accuprime Pfx (Invitrogen), and 2.0 μL of captured exons (from above). Thermocycling was performed with the same conditions as above

while being monitored in real-time (MJ Research Opticon II). Samples were removed from the thermocycler while still in the exponential amplification phase (typically 7-17 cycles). The three samples for each genomic DNA sample were pooled and then run on a 6% TBE-PAGE gel from which the expected band was excised. The DNA was eluted from the crushed acrylamide by incubating overnight at 68°C. The samples were filtered through Spin-X column (GE), ethanol precipitated and resuspended in 10 μL of TE pH 7.0. The DNA was then phosphorylated and circularized in a 20 μL reaction containing 5pmol each CP_2_CIRC_A (ACACTACCGTCGGATCGTGCACCGTAATCGGGAAGCTGAAG, IDT) and CP_2_CIRC_NO_A (ACACTACCGTCGGATCGTGCCCGTAATCGGGAAGCTGAAG, IDT), 20nmol dATP (Invitrogen), 20nmol DTT (Invitrogen), 5U AmpLigase (Epicentre), 5U T4 PNK (NEB) in 1x AmpLigase buffer (Epicentre) and 4.5 μL from the above reaction. The mixture was incubated for 60m at 37°C, and then cycled 10x: 30s at 95°C and 15m at 60°C. 40U of Exonuclease I (NEB) and 200U of Exonuclease III (NEB) were then added to each sample and incubated 60m at 37°C followed by heat inactivation for 15m at 80°C. For each sample two 50-μL hyper-branched rolling circle amplificatio reactions were assembled: 1nM each dNTP (Invitrogen), 0.4x SybrGreen (Invitrogen), 1.25nM random hexamer primer (6N IDT), 0.5U/μL Phi29 DNA Polymerase (Epicentre), and 10 μL from the above reaction. Amplification was run overnight on a real-time PCR (MJ Research Opticon II) at 30°C.

The product was diluted 1:20 in dH2O and 600 μL were shorn in a modified hydroshear. The product was then cleaned with a Qiaquick column and eluted in 34 μL of

EB.

The ends of the fragments were blunted using the End-It Kit (Epicentre), following the manufacturer's protocol, and the sample was phenol-chloroform extracted, ethanol precipitated and resuspended in 20 μL E30-UP-L (/5Phos/CTGCTGGATCTAC, IDT) and E30-DN-L (ACGAGTAGATCCAGCAG, IDT) as well as E30-UP-R (TCGTCATGTAGCAGCAG, IDT) and E30-DN-R (/5Phos/CTGCTGCTACATG, IDT) were combined in separate tubes at a concentration of 10 μM, heat denatured for 5m at 95°C and then allowed to slowly cool to room temperature. A ligation mixture of 50 μL was assembled containing 7.5pmol of each adapter-pair, 20 μL of DNA from the previous step, 100U/μL T4 DNA Ligase (NEB) in 1x Quick Ligase buffer (NEB), and incubated at room temperature for 90m. The reaction was stopped through phenol-chloroform extraction and ethanol precipitation. The samples were run on a 6% PAGE non-denaturing gel, the fragment excised, crushed and DNA eluted for 60m at 55°C. The DNA was filtered through a Spin-X column (GE) and ethanol precipitated. The DNA was resuspended in 40 μL of dH2O. The 5' ends were then phosphorylated with 0.2U/μL T4 PNK (NEB) in 1X T4 DNA Ligase buffer (NEB) for 30m at 37°C. The samples were cleaned through a Qiaquick Column (Qiagen) and the amount of DNA quantified. The DNA was diluted to 0.9ng/μL, and circularized with 100U/μL QuickLigase (NEB) in 1x QuickLigase Buffer for 2hr at 16°C. The samples were phenol-chloroform extracted, ethanol precipitated and resuspended in 30 μL of dH2O. The samples were amplified via hRCA in three 100 μL reactions containing 1mM each dNTP (Invitrogen), 50 μM 6N (IDT), 0.3x SybrGreen (Invitrogen), 5U/μL Phi29 DNA Polymerase (Epicentre) and 10 μL DNA from the previous step in 1x Phi29 reaction buffer (Epicentre). The samples

were incubated overnight at 30°C. The samples were then phenol-chloroform extracted, ethanol precipitated, and resuspended in 10 µL of dH2O. The DNA was digested with EcoP15i with the following 100 µL reaction: 1mM ATP (NEB), 100 µg/ml BSA (NEB), 20 µM Sinefungin (Sigma-Aldrich), 0.5U/µL EcoP15i (NEB), and the DNA from the previous step. The reaction proceeded for 90m at 37°C, after which an additional 1nmol of ATP was added together with 0.4U/µL EcoP15i and the reaction was incubated for an additional 90m at 37°C. The product was cleaned over a Qiaquick column and the correct size product was excised from a 6% non-denaturing TBE PAGE gel. Total recovery was estimated to be 1.5-3pmol/sample. Sequencing adapters FDV-B (ATCACCGACTGCCCATAGAGAGGAAAGCGGAGGCGTAGTGGTT, IDT) / FDV-T (AACCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGAT, IDT) and RDV-T (AACTGCCCCGGGTTCCTCATTCTCT, IDT) / RDV-B (AGAGAATGAGGAACCCGGGGCAGTT, IDT) were annealed as above, and were ligated on in a reaction volume of 20 µL containing 20:1 adapter:template and 2000U of T4 DNA Ligase (NEB) in 1x T4 DNA Ligase buffer (NEB). The reaction was incubated overnight at 16C. The sample was phenol-chloroform extracted, ethanol precipitated, and resuspended in 10 µL of TE pH 8.0. Nick translation was used to seal the nicks in a 25 µL reaction containing 500nM each dNTP (Invitrogen), 0.4U/µL DNA Polymerase I (NEB), 1x NEB buffer 2 and the DNA from the previous step. After 30m at 16°C the reaction was cleaned with a phenol-chloroform extraction and ethanol precipitation and the resuspended product was gel purified over a 6% TBE-PAGE gel. The correct product was then enriched for with 10 cycles of PCR run containing: 200nM each dNTP, 1.5mM MgCl2, 200nM each FDV-T and RDV-T primers, 0.3x SybrGreen (Invitrogen),

0.02U/μL Platinum Tag (Invitrogen) in 1x PCR Buffer (Invitrogen), with 1 μL of DNA from the previous step. The sample was then cleaned over a Qiaquick column and gel quantified.

For Illumina GA sequencing of this library, the following adapters were used to amplify a dilution of the final product: RDV-T-SLXA (AACTGCCCCGGGTTCCTCATTCTCTCAAGCAGAAGAGGCATACGA, IDT) and FDV-T-SLXA (AATGATACGGCGACCACCGAAACCACTACGCCTCCGCTTTCC, IDT). Sequencing was performed with custom primer: AACTGCCCCGGGTTCCTCATTCTCT.

End-Sequencing of Captured Targets for Illumina GA

The captured exons from above were amplified with CP2-FA-SLXA (CAAGCAGAAGACGGCATACGAGCACGATCCGACGGTAGTGT, IDT) and CP2-RA-SLXA (AATGATACGGCGACCACCGAGATCTCCGTAATCGGGAAGCTGAAG, IDT). Sequencing was performed with the custom primer: CCGTAATCGGGAAGCTGAAG.

Shotgun Library for Illumina GA

Exon capture reactions were assembled containing 3ng of prepared probes, 1μg of gDNA in a total of 20 μL of 1x AmpLigase Buffer (Epicentre). After incubation for 40hr at 60°C 2U of Pfu DNA Polymerase (Stratagene), 2.5U of AmpLigase (Epicentre), and 0.5nmol of each dNTP (Invitrogen) were added to each reaction. After an additional 80m at 60°C 1.5 μL of each Exonuclease I (NEB) and Exonuclease III (NEB) were added as well as 1.5 μL of 10x ExoI buffer. The samples were incubated for 60m at 37°C before heat inactivation for 15m at 80°C. Three 100-μL PCR reactions were assembled for each

sample 1.25U/100 µL Pfu (Stratagene) with the following primers: CP-2-FA-Biotin3 (/5Biosg/CTCATTACCCTCTCCCTCATCAGCAGATCCGACGGTAGTGT, IDT) and CP-2-RA-Biotin3 (/5Biosg/CCCCTAAATCCCAACCTCAACAGCAGATCGGGAAGCTGAAG, IDT). The PCR reactions were pooled, cleaned over a Qiaquick column and eluted in 50 µL of TE, pH 7.0. To remove the universal sequence and anchor arms, the samples were digested with EcoP15i using one unit of enzyme per 1.5pmol of material in a reaction containing 1mM ATP (NEB), BSA (NEB) and 20 µM Sinefungin (Sigma-Aldrich). 25 µL of C1 beads (Invitrogen) were resuspended in 100 µL of Bind and Wash Buffer and used to remove the biotinylated probes. After incubating with the beads for 20m rotating at room temperature, the supernatant was removed and cleaned over a Qiaquick column. The total recovery of DNA was 5-30%. The samples were then blunted and phosphorylated using the End-It kit (Epicentre) following the manufacturer's protocol, and the material cleaned over a Qiaquick column (Qiagen). The DNA was concatenated in a 20 µL reaction using 100U/µL T4 DNA Ligase (NEB) in 1x T4 DNA Ligase buffer (NEB) overnight at16°C. dH2O was added to increase the volume of each sample to 200 µL and the enzyme was heat inactivated by heating for 10m at 65°C. The samples were then sheared using ultrasonic waves (BioRuptor, Diagenode). The device was set at high, alternating on and off every 30s for 60m. Every 15m the DNA was collected at the bottom of the tube and the device was refilled with ice. The ends were then polished using the End-It kit, and the samples cleaned over a Qiaquick column. 45pmol of Solexa sequencing primers were then ligated on for 30m at 16°C and cleaned over a Qiaquick column. Nick translation was assembled as above and the reaction proceeded for 30m at

16°C. The DNA was run on a 6% PAGE and a band between 150-200bp was excised for each of the samples, as above. The correct product was PCR enriched and quantified.

References

1 Siva, N. 1000 Genomes project. *Nat Biotechnol* **26**, 256, doi:nbt0308-256b [pii]
10.1038/nbt0308-256b (2008).

2 Biesecker, L. G. *et al.* The ClinSeq Project: Piloting large-scale genome sequencing for research in genomic medicine. *Genome Res*, doi:gr.092841.109 [pii]
10.1101/gr.092841.109 (2009).

3 Mailman, M. D. *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* **39**, 1181-1186, doi:ng1007-1181 [pii]
10.1038/ng1007-1181 (2007).

4 Church, G. M. The personal genome project. *Mol Syst Biol* **1**, 2005 0030, doi:msb4100040 [pii]
10.1038/msb4100040 (2005).

5 Snyder, M., Weissman, S. & Gerstein, M. Personal phenotypes to go with personal genomes. *Mol Syst Biol* **5**, 273, doi:msb200932 [pii]
10.1038/msb.2009.32 (2009).

6 Homer, N. *et al.* Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* **4**, e1000167, doi:10.1371/journal.pgen.1000167 (2008).

7 Zerhouni, E. A. & Nabel, E. G. Protecting aggregate genomic data. *Science* **322**, 44, doi:1165490 [pii]
10.1126/science.1165490 (2008).

8 Lunshof, J. E., Chadwick, R., Vorhaus, D. B. & Church, G. M. From genetic privacy to open consent. *Nat Rev Genet* **9**, 406-411, doi:nrg2360 [pii]
10.1038/nrg2360 (2008).

9 Church, G. *et al.* Public access to genome-wide data: five views on balancing research with privacy and protection. *PLoS Genet* **5**, e1000665, doi:10.1371/journal.pgen.1000665 (2009).

10 Murphy, J. *et al.* Public expectations for return of results from large-cohort genetic research. *Am J Bioeth* **8**, 36-43, doi:906418390 [pii]
10.1080/15265160802513093 (2008).

11 Okou, D. T. *et al.* Microarray-based genomic selection for high-throughput resequencing. *Nat Methods* **4**, 907-909, doi:nmeth1109 [pii]
10.1038/nmeth1109 (2007).

12 Albert, T. J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nat Methods* **4**, 903-905, doi:nmeth1111 [pii]
10.1038/nmeth1111 (2007).

13 Hardenbol, P. *et al.* Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat Biotechnol* **21**, 673-678, doi:10.1038/nbt821
nbt821 [pii] (2003).

14 Markoulatos, P., Siafakas, N. & Moncany, M. Multiplex polymerase chain reaction: a practical approach. *J Clin Lab Anal* **16**, 47-51, doi:10.1002/jcla.2058 [pii] (2002).

15 Hardenbol, P. *et al.* Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay. *Genome Res* **15**, 269-275, doi:15/2/269 [pii]
10.1101/gr.3185605 (2005).

16 Porreca, G. J. *et al.* Multiplex amplification of large sets of human exons. *Nat Methods* **4**, 931-936, doi:nmeth1110 [pii]
10.1038/nmeth1110 (2007).

17 Krishnakumar, S. *et al.* A comprehensive assay for targeted multiplex amplification of human DNA sequences. *Proc Natl Acad Sci U S A* **105**, 9296-9301, doi:0803240105 [pii]
10.1073/pnas.0803240105 (2008).

18 Li, J. B. *et al.* Multiplex padlock targeted sequencing reveals human hypermutable CpG variations. *Genome Res* **19**, 1606-1615, doi:gr.092213.109 [pii]
10.1101/gr.092213.109 (2009).

19 Porreca, G. J., Shendure, J. & Church, G. M. Polony DNA sequencing. *Curr Protoc Mol Biol* **Chapter 7**, Unit 7 8, doi:10.1002/0471142727.mb0708s76 (2006).

20 Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat Methods* **6**, 315-316, doi:nmeth.f.248 [pii]
10.1038/nmeth.f.248 (2009).

21 Mucke, M., Reich, S., Moncke-Buchner, E., Reuter, M. & Kruger, D. H. DNA cleavage by type III restriction-modification enzyme EcoP15I is independent of spacer distance between two head to head oriented recognition sites. *J Mol Biol* **312**, 687-698, doi:10.1006/jmbi.2001.4998
S0022-2836(01)94998-8 [pii] (2001).

22 Moncke-Buchner, E. *et al.* Functional characterization and modulation of the DNA cleavage efficiency of type III restriction endonuclease EcoP15I in its interaction with two sites in the DNA target. *J Mol Biol* **387**, 1309-1319, doi:S0022-2836(09)00218-6 [pii]
10.1016/j.jmb.2009.02.047 (2009).

23 Raghavendra, N. K. & Rao, D. N. Exogenous AdoMet and its analogue sinefungin differentially influence DNA cleavage by R.EcoP15I--usefulness in SAGE. *Biochem Biophys Res Commun* **334**, 803-811, doi:S0006-291X(05)01409-9 [pii]
10.1016/j.bbrc.2005.06.171 (2005).

24 Di, X. *et al.* Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays. *Bioinformatics* **21**, 1958-1963, doi:bti275 [pii]
10.1093/bioinformatics/bti275 (2005).

25 Quail, M. A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nat Methods* **5**, 1005-1010, doi:nmeth.1270 [pii]
10.1038/nmeth.1270 (2008).

26 Zhang, Z. & Gerstein, M. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res* **31**, 5338-5348 (2003).

27 Zhao, Z. & Boerwinkle, E. Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Res* **12**, 1679-1686, doi:10.1101/gr.287302 (2002).

Chapter 4

Analysis of MIP Targeted Sequencing Biases and Recommendations for Future Design

Abraham M. Rosenbaum[1], Michael F. Chou[1], Jin Billy Li[1], Madeleine Ball[1], John Aach[1], George M. Church[1]

[1]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

Author Contributions A.M.R. performed the analysis with computational help from M.F.C. and M.B., statistical and presentation advice from J.A. and M.F.C., and helpful discussions with M.F.C., J.B.L., M.B., J.A. and G.M.C. G.M.C. supervised all aspects of the study.

Acknowledgements We are grateful for the help and advice provided by all the member of the Church Laboratory. We thank NHGRI and NHLBI for funding support.

Abstract

Targeted capture of genomic regions based upon molecular inversion probes (MIPs) has been successfully used in a number of recent publications. In each of these studies, however, there was significant non-uniformity in sequencing depth, with the most frequently sequenced targets appearing four orders of magnitude more often than the least. We explore bias introduced by GC content, length and secondary structure of the targets, as well as secondary structure of the probes and melting temperature of the hybridization arms of the probes. While some of these biases (particularly GC content) are complicated by downstream steps of PCR and ligation, we tend to view the entire process as one and not differentiate between the capture and necessary downstream steps. In addition to calculating the overall capture efficiency for each criterion, we also calculate the pair-wise efficiency for these factors. Furthermore, we bin each of the factors and calculate the variance for all species in each bin to enable separate pooling of species followed by calibration and mixing prior to sequencing. This data is compiled and used as the basis for the MIPTAG-Pro algorithm described in Appendix F.

Introduction

The development of second generation sequencing has created a need for new high-throughput DNA targeting methods as a front-end for these processes. A number of these methods are discussed in Chapter 1. Each of these methods has distinct advantages and capture biases. Broadly categorized, the hybridization-based methods, whether surface based[1-5] or solution based[6-7] typically capture in a Poisson distribution around the target, have difficulty differentiating between homologues, and are broadly influenced by GC content[6-7]. It is not clear whether the loss of AT and GC rich regions is caused by the

limitations of Second Generation Sequencing[8-9], biases in synthesis of the capture DNA, or PCR biases (a five-fold decrease in PCR efficiency with extreme nucleotide compositions has been reported[6]). PCR-based targeting[10-15], on the other hand, does have the ability to capture homologues through careful design of the primers, but the different processes do have their own biases. Typically, PCR can rely upon algorithms such as Primer3[16] to optimize the primers, and, additionally, temperature, salt concentration and PCR additives can be optimized to maximize the desired product. When designing thousands of reactions, however, it becomes exceedingly difficult to optimize the conditions for each pair and impossible if they are all done in the same reaction tube. The added problem of creating primers as close to the desired target as possible while maintaining optimized design criteria is also more acutely felt in such a multiplex reaction. Finally, in the case of Molecular Inversion Probes, the additional ligation step has two consequences in the design of the probe. Firstly, the ligation step necessitates that the ligation arm be neither degraded nor displaced by the polymerase. Secondly, while typical PCR and hybridization techniques can rely upon RNA probes to prevent cross-hybridization between species, ligation necessitates that at least part of the MIP be constructed of DNA. Here we present an analysis of the biases present in the MIP capture reactions described in the previous chapter, and the elucidation of some design rules. These design rules have been consolidated into a software program "MIPTAG Pro" that is described in Appendix F.

Results

To find possible sources of bias, we analyzed our end-sequencing results for correlation between capture efficiency and (1) target length, (2) extension and (3) ligation

arm melting temperature (Tm), (4) target GC-content, secondary structure of both the (5) probe and (6) target region, and (7) nucleotide composition of the ends of the hybridization and extension arms.

The most striking correlation with poor capture in early experiments was the identity of the terminal nucleotide on the ligation arm; this was in stark contrast with the extension arm where there was little correlation. We found that both base pairing and base stacking of this nucleotide influenced the percentage of targets captured and the number of sequenced reads coming from the targets that were captured. Both of these elements provide information as to how well the molecular inversion probe performs. The nucleotide with the poorest pairing and stacking ability, thymine, showed an overall capture efficiency of 50% (that is, only 50% of the targets were seen at least once) and a read depth almost 17.1x lower than expected. While adenine showed a similar target capture rate to that of cytosine (83% and 85%, respectively), the read depth showed a 2.23x drop for adenine, while the decrease for cytosine was only 1.25x. The MIPs with the ligation arm terminus being guanine, the nucleotide with the strongest base pairing and base stacking interactions, captured 93% of targets and showed an overrepresentation of reads of 1.5x. We hypothesized that Stoffel Large Fragment, *Taq* Polymerase mutated so as to abolish 5 to 3' exonuclease activity[17], was displacing the 5' nucleotide of the ligation arm, thereby preventing its ligation to the nascent strand due to overlapping nucleotides. Replacing this polymerase with Pfu polymerase generated a more balanced representation of targets in the shotgun libraries. The thymine and adenine under-representations dropped to 1.98x and 1.16x respectively, cytosine was slightly over-represented at 1.09x and guanine was less over-represented at 1.15x (Figures 4-1A and 4-

1B). Additionally, while previous work has shown a decrease in capture efficiency when the concentration of dNTP exceeded 100x the minimally necessary amount[13], Pfu polymerase showed no such decrease even when the concentration was 12,000x (Figures 4-1C and 4-1D) We used Pfu polymerase for targeted capture from each of the ten Personal Genome Project samples, and this capture reaction was used to generate the shotgun libraries described in the previous chapter. For each of the other six variables noted, we binned the variables into evenly sized groups ranging from 32 to 132 bins for each variable. For each one of these bins we calculated the fraction of MIPs that successfully captured their target at least once, the fraction of all sequencing reads that come from that bin, and the read depth distribution. Finally, we compared the results of the end-sequencing and shotgun libraries, to account for potential complications due to the polymerase change and additional enzymatic steps required for the shotgun library production.

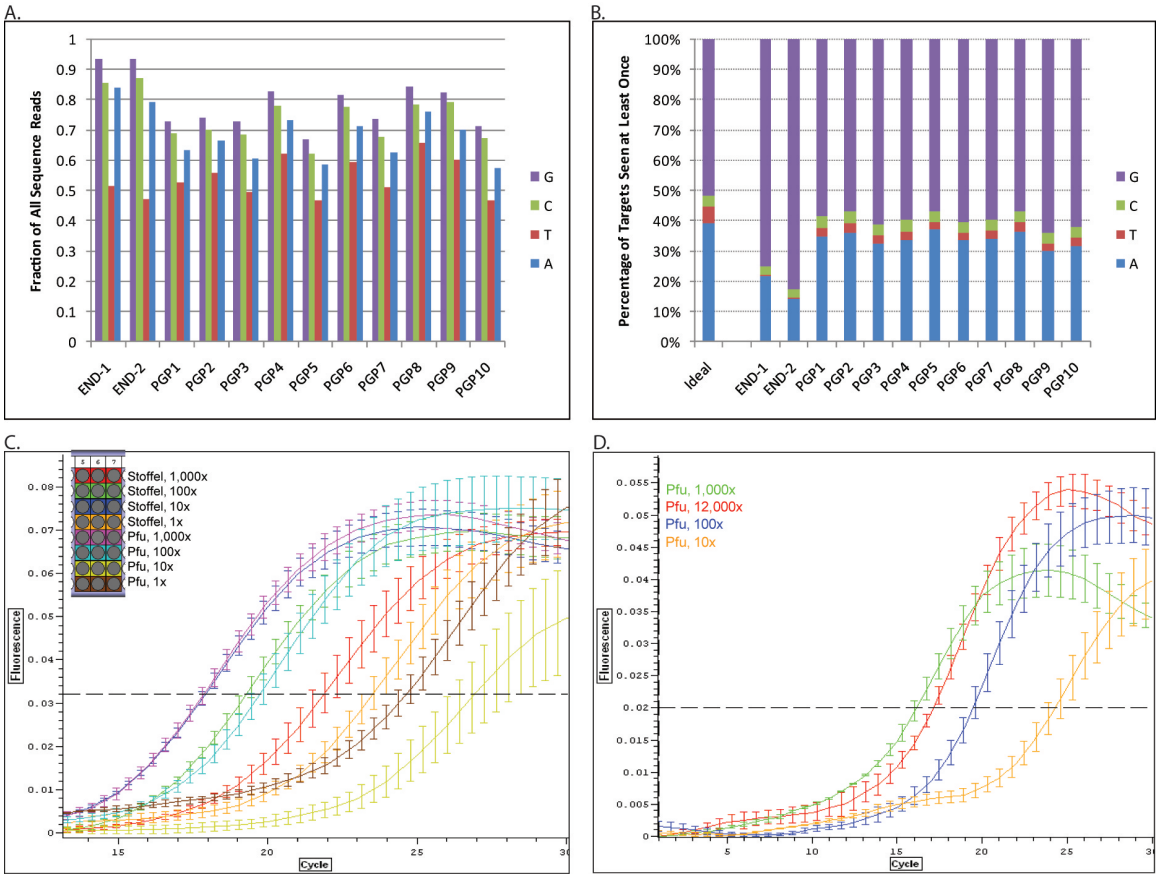


Figure 4-1. Polymerase Effect on Capture Efficiency. A. Capture Efficiency as a Function of Ligation Arm Terminus. Plot of the percentage of targets seen at least once grouped by the nucleotide identity of the 5' terminus of the ligation arm. Shown are data from the two end-sequencing libraries (END 1 and END 2) where Stoffel Large Fragment was the polymerase used for MIP capture, and the ten shotgun libraries (PGP1-PGP10) where the polymerase was Pfu. While the shotgun libraries overall efficiency was lower, the uniformity of the capture was improved. **B. Relative Capture Efficiency as a Function of Ligation Arm Terminus.** The makeup of all sequenced reads is expected to reflect the makeup of the targets. The left-most plot shows the fraction of all targets with the noted terminal nucleotide. END 1 and END 2 depict the fraction of all reads originating from the specified bins from the end-sequencing library, and PGP1-PGP10 show the same for all reads from the ten shotgun libraries. **C-D. Effect dNTP Concentration on Capture Efficiency.** MIP capture reactions were assembled with different concentrations of dNTPs and different polymerases (Stoffel or Pfu). 1x dNTP is calculated as the minimum amount of dNTPs necessary to fill in every gap assuming each nucleotide is equally represented in the gap region. After the capture reactions an aliquot of each sample was amplified in triplicate via real-time PCR to estimate the number of captured targets. While Stoffel is very sensitive to dNTP concentration, Pfu shows no such limitation.

The MIP targets ranged in size from 60nt to 191nt, and each size target was analyzed separately. While the targets were fairly evenly distributed throughout this range (80% of the data lay between 75nt and 170nt, with slight tapering towards the extremes), there was a marked three base-pair fluctuation, reflecting the nature of three

base pair codons. We checked for additional periodicity reflecting the 10.7bp helical pitch of DNA, which had been seen in previous work in our lab involving circularization of DNA[18] (see Appendix D), but our data proved too noisy. The fraction of targets captured decreased with increasing length, and this decrease was exacerbated in the shotgun library. This resulted in 80% of the data lying within a range of 73nt and 71nt for the End-sequencing and Shotgun Libraries (lengths of 60nt-131nt and 60nt-133nt), respectively. While the fraction of all reads originating from each target length reflected the slight increase in targets for mid-range sizes in the end-sequencing library, this was masked by the much larger downward trend in the shotgun library. The primary cause for this downward slope appears to be an increase in the highest capturing short targets, but not all short targets, resulting in decreased uniformity for shorter targets. Remarkably, although both libraries show a similar fraction of all reads originating from long targets, the shotgun library shows increased uniformity in this range (Figure 4-2). While it is unlikely that the switch to Pfu polymerase would have such a profound effect on length, we considered the hypothesis that the shorter molecule length increased the likelihood of successful EcoP15i digestion, or increased the likelihood of concatenation due to improved kinetics (see the previous chapter for library construction methods). Either of these would lead to an over-representation of these targets in the final shotgun library. We tested this latter hypothesis by comparing the number of reads originating from shotgun sequence with those originating from the ligation of a sequencing adapter directly to the end of the amplified MIP. The lower the ratio, the less efficient the concatenation was. We found a consistent ratio of 1:30 regardless of target length.

Recently, solutions to the general inefficiency of concatenation have been suggested[19], and we look forward to seeing whether this will improve our results.

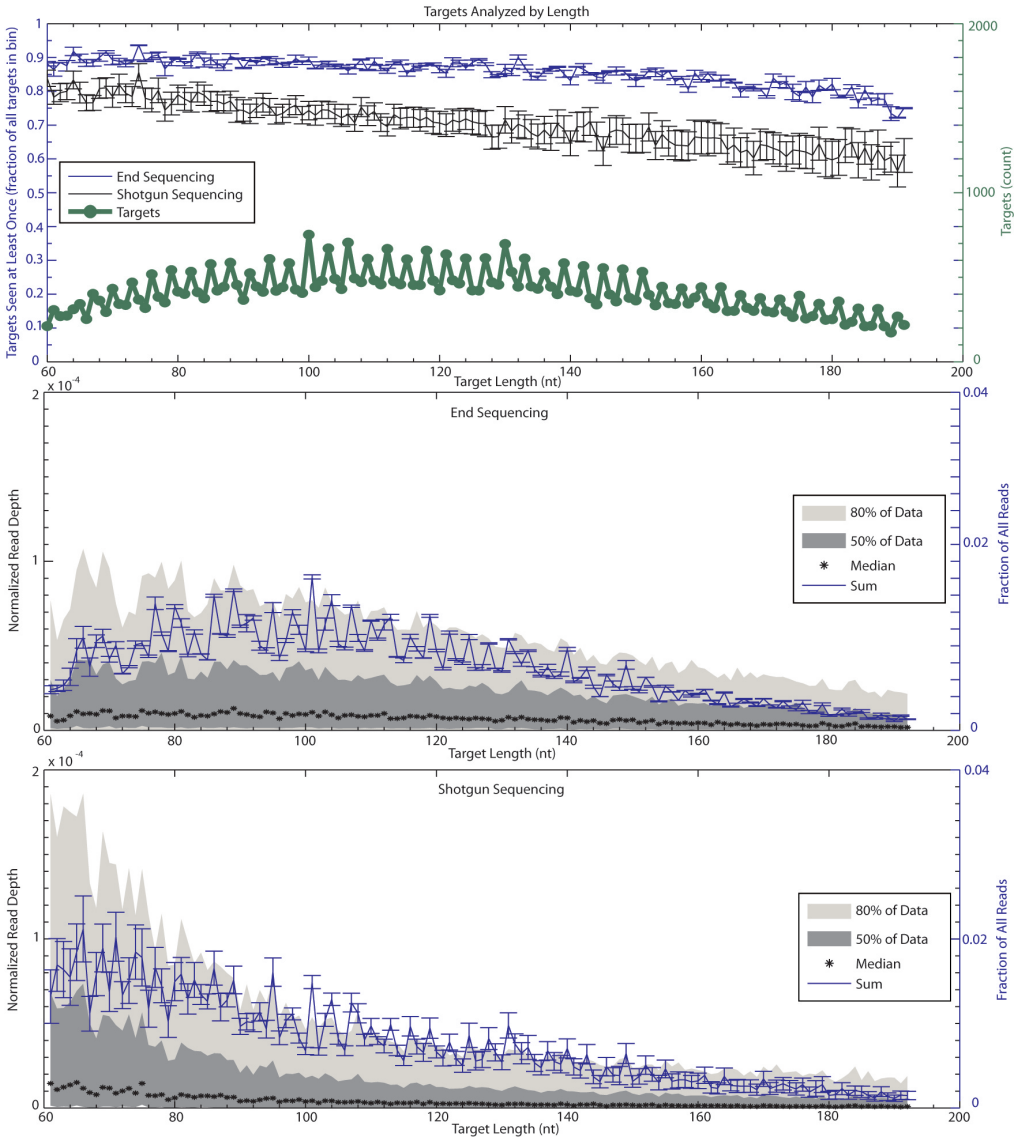


Figure 4-2. Top Plot: Absolute Target Capture. The number of targets of each nucleotide length are marked with green circles against the right hand axis. The blue and black lines show the fraction of targets of each length that were seen at least once for the end-sequencing and shotgun-sequencing libraries, respectively. These are plotted against the left-hand axis. The end-sequencing reads were placed against a database of expected targets using WU-BLAST to account for sequencing errors, and the shotgun-sequencing libraries were placed with MAQ using the HG18 reference genome. The relative inefficient capture of longer targets exists in both datasets, with a more dramatic difference in the shotgun library. **Middle Plot: End-Sequencing Capture Efficiency.** The blue line represents the fraction of all data that comes from targets of the given length. The black stars show the normalized median capture of each target length, the dark grey represents 50% of the normalized data (25% - 75%) and the light grey represents 80% of the normalized data (10% - 90%). **Bottom Plot: Shotgun Sequencing Capture Efficiency.** The plot is arranged as above. In the shotgun library the increased efficiency of short-target capture is mostly due to an increase in the most abundant species, while the median remains fairly constant. Conversely, the 80% range of capture becomes more uniform as the targets increase in length. All error bars represent one standard deviation.

The target GC content ranged from 0.24 to 0.86, with 80% of the targets lying between 0.38 and 0.6. The bins with the highest fraction of targets captured were those with GC content between 0.5 and 0.6, with the shotgun library showing a greater loss of low-GC content targets than the end-sequencing library. For both libraries, however, 80% of the data lay between 0.4 and 0.58. Uniformity of capture generally followed this trend, with greater capture correlating with lesser uniformity. Finally, the shotgun library contained very little high capture material outside of a very tight band with GC content from 0.5 to 0.6, in contrast with the end-sequencing library (Figure 4-3). Many groups have found loss of GC-low regions a problem with library preparation for the Illumina GAII, particularly when gel-extraction is involved [9], and this may account for some of the loss. Alternatively, this difference may be alleviated by further optimization of the polymerase.

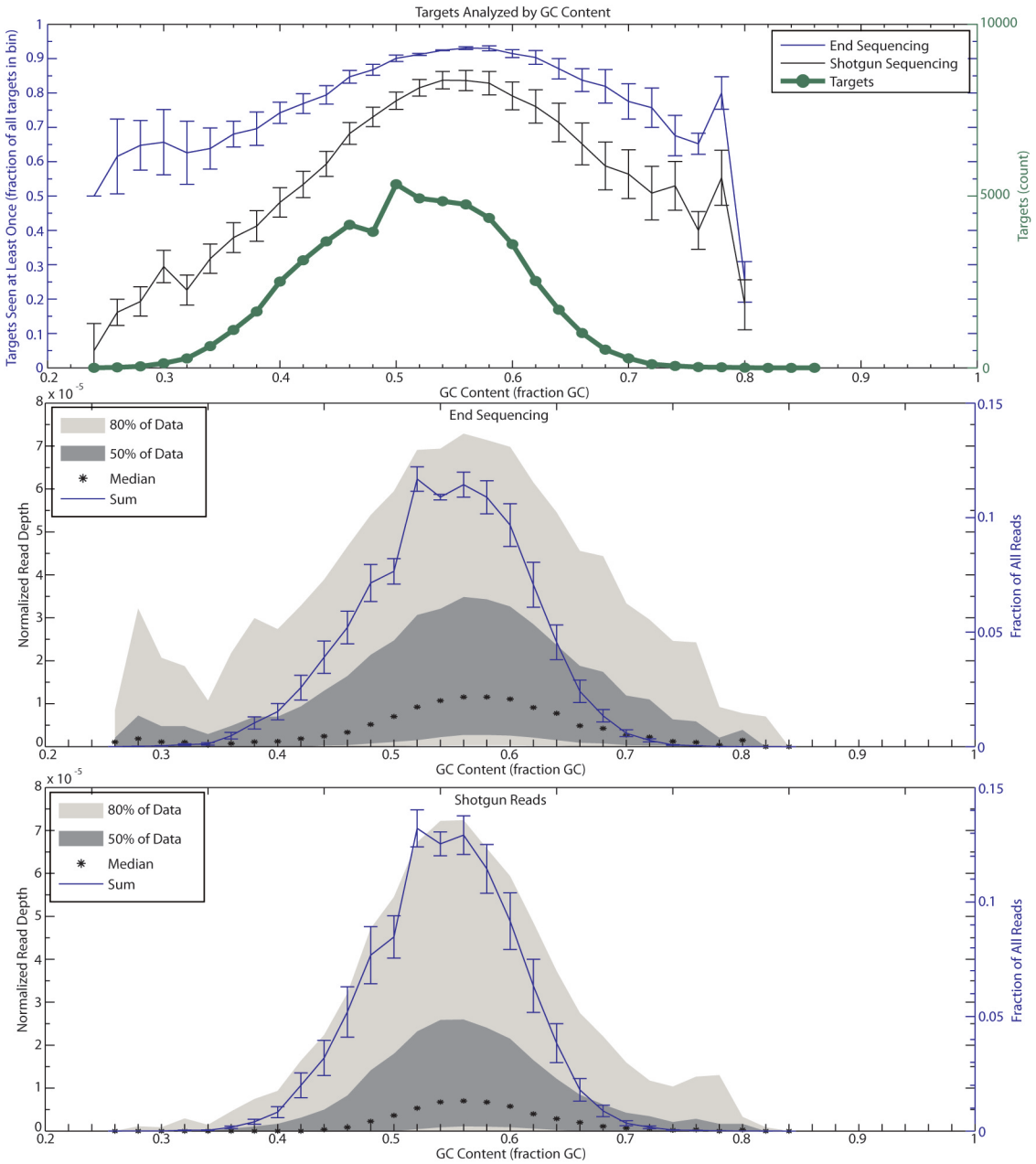


Figure 4-3. Top Plot: Absolute Target Capture. The number of targets of each nucleotide length are marked with green circles against the right hand axis. The blue and black lines show the fraction of targets of each length that were seen at least once for the end-sequencing and shotgun-sequencing libraries, respectively. These are plotted against the left-hand axis. The end-sequencing reads were placed against a database of expected targets using WU-BLAST to account for sequencing errors, and the shotgun-sequencing libraries were placed with MAQ using the HG18 reference genome. The inefficient capture of low and high GC targets is compounded by the difficulty the Illumina GAII has with this type of sequence, but the shotgun library production shows an even greater loss of these targets. **Middle Plot: End-Sequencing Capture Efficiency.** The blue line represents the fraction of all data that comes from targets of the given length. The black stars show the normalized median capture of each target length, the dark grey represents 50% of the normalized data (25% - 75%) and the light grey represents 80% of the normalized data (10% - 90%). **Bottom Plot: Shotgun Sequencing Capture Efficiency.** The plot is arranged as above. In the shotgun library the capture distribution is even narrower, lacking the few frequently appearing targets with low-GC content. All error bars represent one standard deviation.

The extension and ligation arm Tm analysis show similar trends (calculated with http://arep.med.harvard.edu/kzhang/cgi-bin/myOligoTm.cgi). Both have a Tm range from 37.5°C to 64°C, with 80% of targets falling between 43°C and 57°C. For the extension arm 80% of the data lay between 44.5°C/45.5°C and 57°C for the End-Sequencing and Shotgun libraries, respectively. For the ligation arm, while the upper bound increased by 0.5°C, the lower bound increased by 2°C for each of the libraries. Regarding uniformity of reads both the End-sequencing and Shotgun libraries were very similar: they both had fairly consistent level of uniformity for Tm-s between 50°C and 64°C despite a diminishing fraction of total reads over 55°C. The shift of the lower bound of 80% of the data 1.5°C-2.5°C for the extension (Figure 4-4) and 3.5°C-4.5°C for the ligation arms (Figure 4-5) reflect the sensitivity of MIP capture to these metrics. The greater permissiveness of the extension arm likely reflects the ability of the polymerase to quickly extend the relatively short primer, while the Tm of the ligation arm is more limited due to the necessity that the end of the arm be hybridized to its complement in order to prevent the polymerase from over-extending the synthesized strand into the anchor region. The difference between the End-sequencing and shotgun libraries may be due to the difference in processivity between Stoffel and Pfu, but it is more likely due to the GC-effect insofar as the correlation coefficient between GC content and Tm is 0.543 and 0.547 for the extension and ligation arms, respectively.

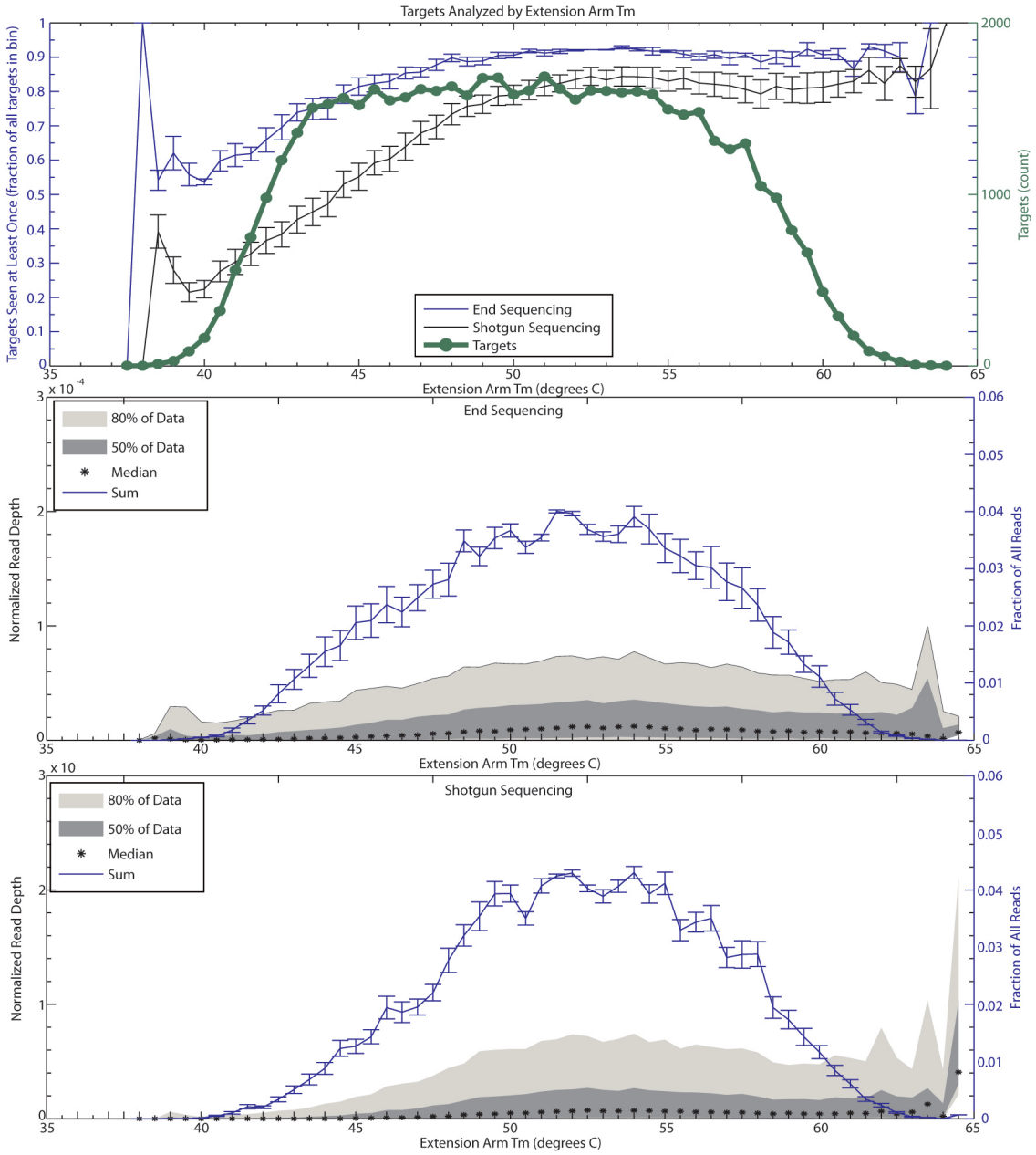


Figure 4-4. Top Plot: Absolute Target Capture. The number of targets of each nucleotide length are marked with green circles against the right hand axis. The blue and black lines show the fraction of targets of each length that were seen at least once for the end-sequencing and shotgun-sequencing libraries, respectively. These are plotted against the left-hand axis. The end-sequencing reads were placed against a database of expected targets using WU-BLAST to account for sequencing errors, and the shotgun-sequencing libraries were placed with MAQ using the HG18 reference genome. The inefficient capture of targets with low extension arm Tm is consistent, but the greater sensitivity for the shotgun-library may be due to this metric's correlation with GC content. **Middle Plot: End-Sequencing Capture Efficiency.** The blue line represents the fraction of all data that comes from targets of the given length. The black stars show the normalized median capture of each target length, the dark grey represents 50% of the normalized data (25% - 75%) and the light grey represents 80% of the normalized data (10% - 90%). **Bottom Plot: Shotgun Sequencing Capture Efficiency.** The plot is arranged as above. In the shotgun library the capture distribution is even narrower, possibly reflecting the correlation between extension arm Tm and GC content. Error bars represent one standard deviation.

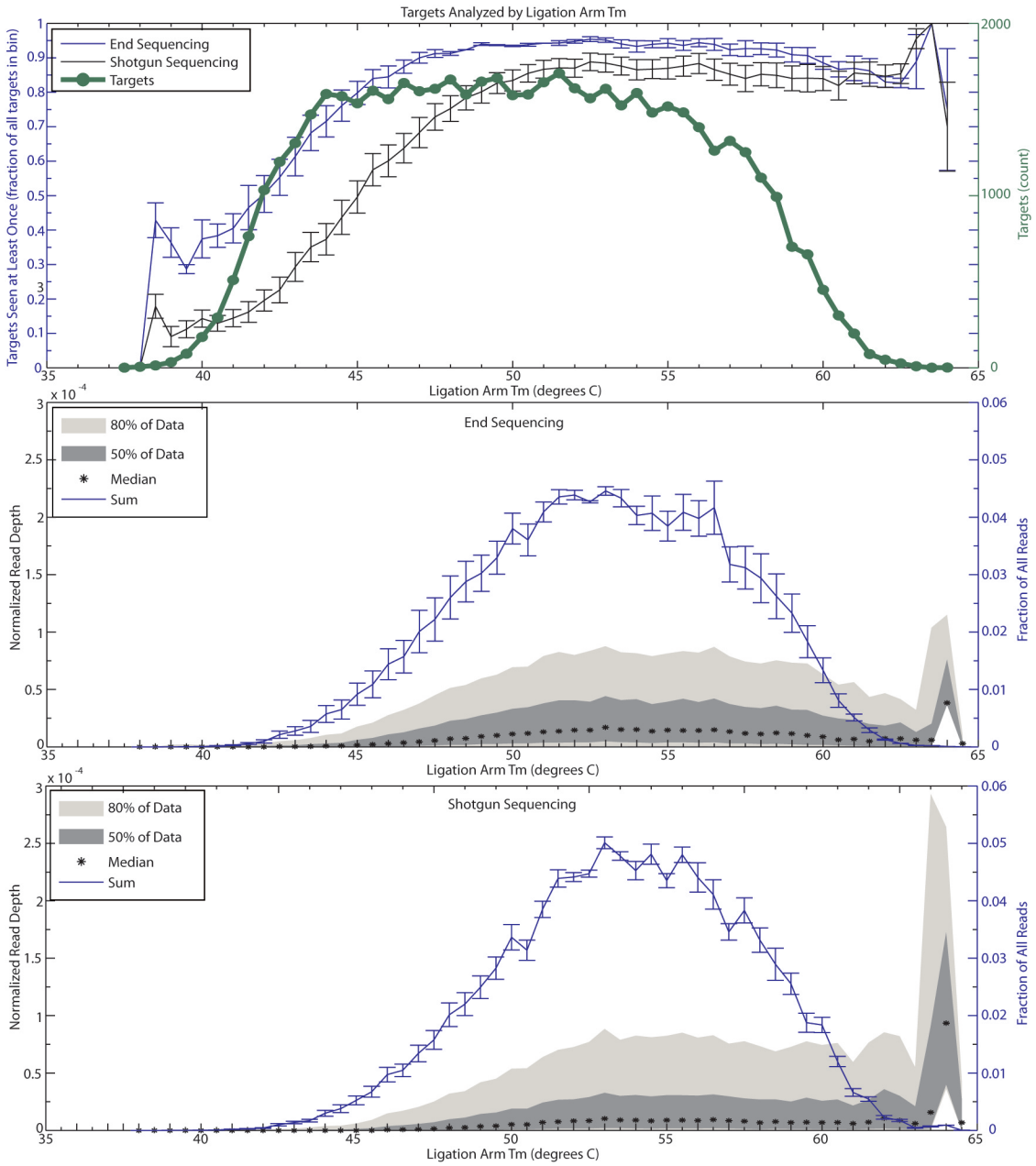


Figure 4-5. Top Plot: Absolute Target Capture. The number of targets of each nucleotide length are marked with green circles against the right hand axis. The blue and black lines show the fraction of targets of each length that were seen at least once for the end-sequencing and shotgun-sequencing libraries, respectively. These are plotted against the left-hand axis. The end-sequencing reads were placed against a database of expected targets using WU-BLAST to account for sequencing errors, and the shotgun-sequencing libraries were placed with MAQ using the HG18 reference genome. The inefficient capture of targets with low ligation arm Tm is consistent, but the greater sensitivity for the shotgun-library may be due to this metric's correlation with GC content. **Middle Plot: End-Sequencing Capture Efficiency.** The blue line represents the fraction of all data that comes from targets of the given length. The black stars show the normalized median capture of each target length, the dark grey represents 50% of the normalized data (25% - 75%) and the light grey represents 80% of the normalized data (10% - 90%). **Bottom Plot: Shotgun Sequencing Capture Efficiency.** The plot is arranged as above. In the shotgun library the capture distribution is even narrower, possibly reflecting the correlation between ligation arm Tm and GC content. Error bars represent one standard deviation.

Finally, we analyzed the Gibbs Free Energy of the probes and targets to see whether secondary structure influences MIP targeted capture. To calculate the free energy we ran a local instance of mfold (http://mfold.bioinfo.rpi.edu/cgi-bin/dna-form1.cgi)[20-22] using the default settings with the exception that the ion concentrations and temperature were set to match the experimental conditions. Free energy of the probes showed a very narrow range (which is expected due to 30 of the 70 nucleotides being shared by all probes) and 24% of the probes had a ΔG of 0.4-0.5, with 80% lying between -1.3 and 0.5 (Figure 4-6). Free energy for the targets had a larger distribution, with 80% of the targets between -7 and -0.5 (Figure 4-7). Both the libraries showed a distribution of reads very similar to the expected distribution, reflecting the fact that secondary structure of the probes and immediate target region seem to have little effect on target capture.

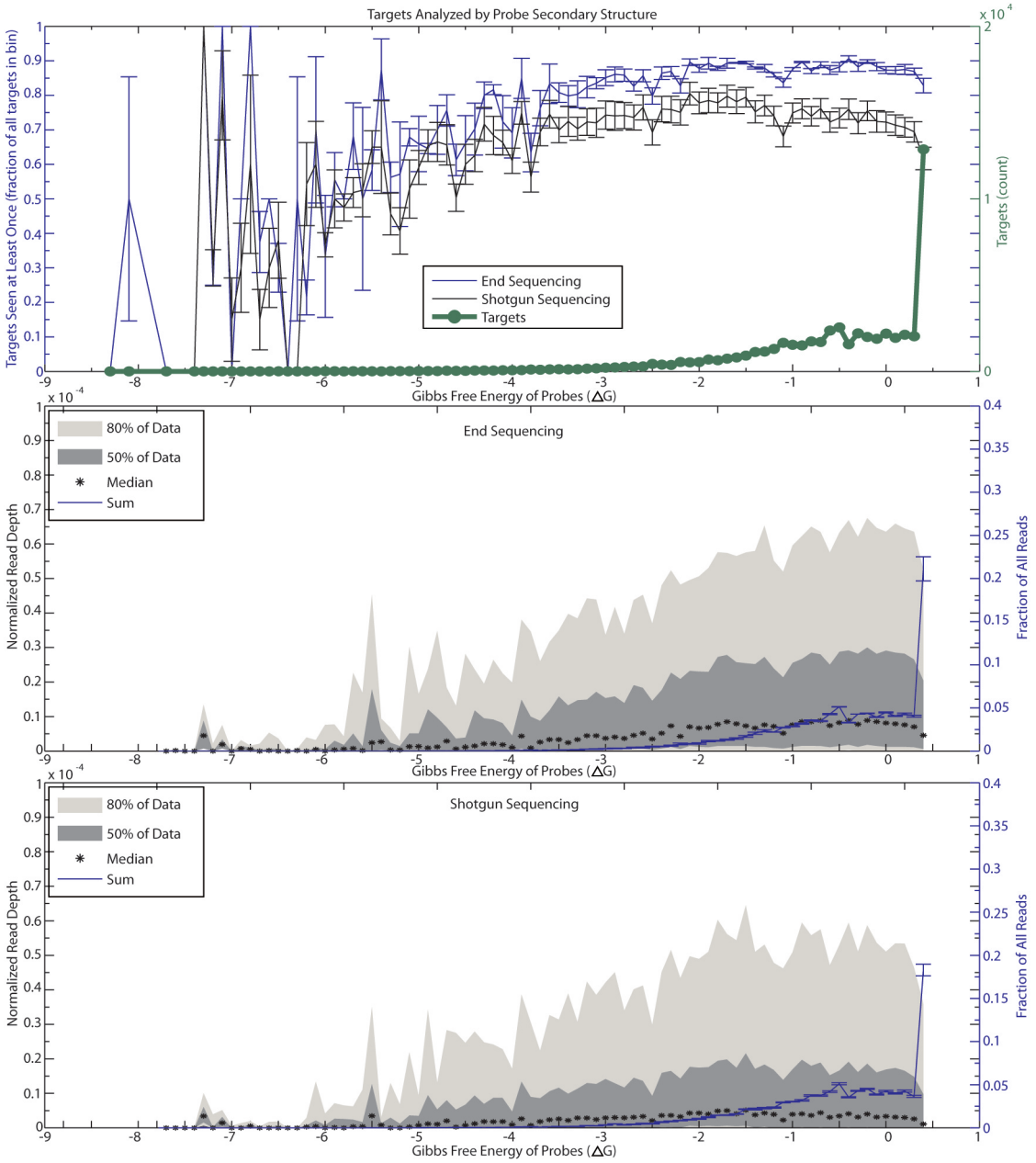


Figure 4-6. Top Plot: Absolute Target Capture. The number of targets of each nucleotide length are marked with green circles against the right hand axis. The blue and black lines show the fraction of targets of each length that were seen at least once for the end-sequencing and shotgun-sequencing libraries, respectively. These are plotted against the left-hand axis. The end-sequencing reads were placed against a database of expected targets using WU-BLAST to account for sequencing errors, and the shotgun-sequencing libraries were placed with MAQ using the HG18 reference genome. The relatively short length of the probes and the >40% of sequence shared among all probes result in their very limited distribution. **Middle Plot: End-Sequencing Capture Efficiency.** The blue line represents the fraction of all data that comes from targets of the given length. The black stars show the normalized median capture of each target length, the dark grey represents 50% of the normalized data (25% - 75%) and the light grey represents 80% of the normalized data (10% - 90%). **Bottom Plot: Shotgun Sequencing Capture Efficiency.** The plot is arranged as above. Error bars represent one standard deviation.

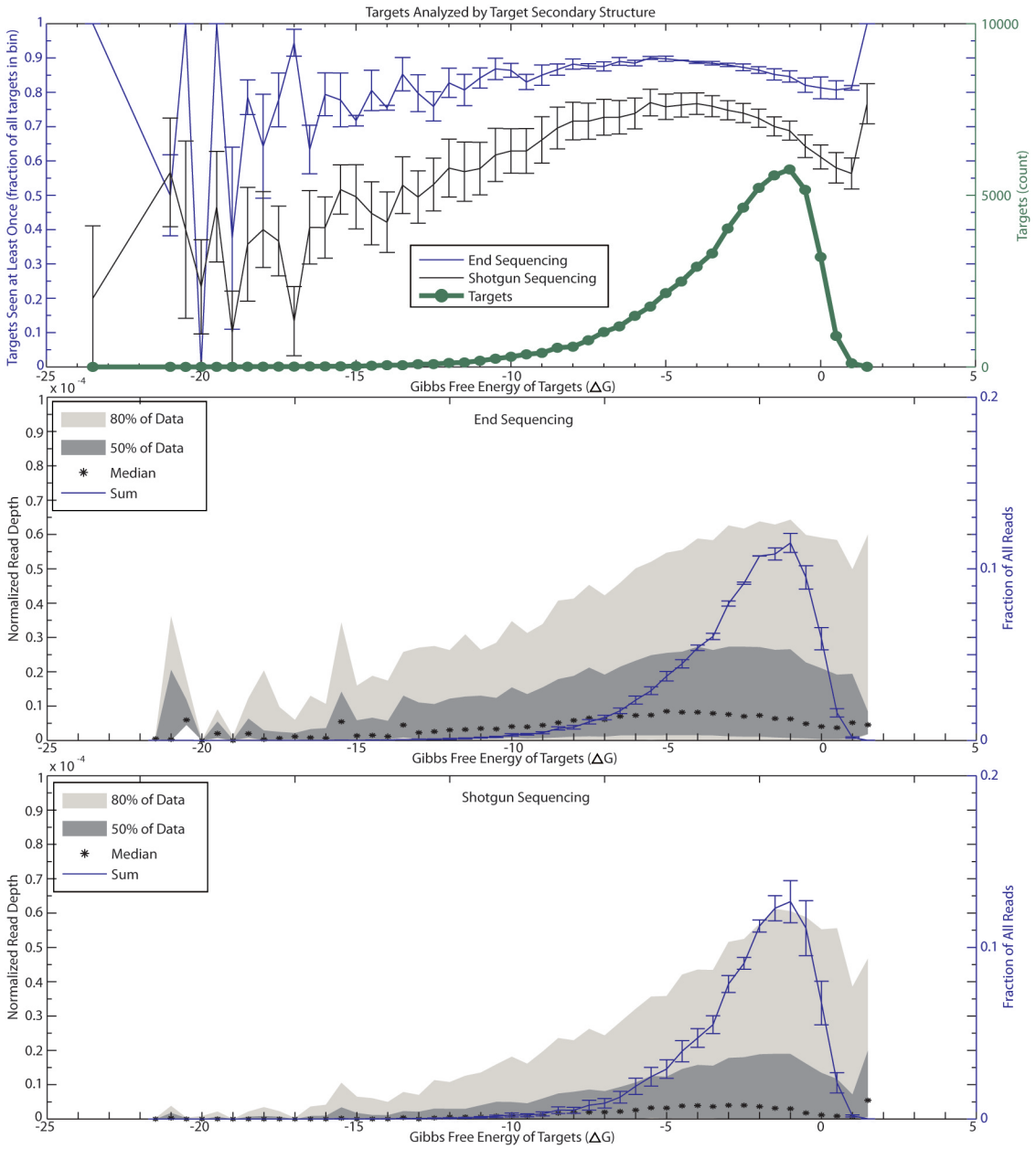


Figure 4-7. Top Plot: Absolute Target Capture. The number of targets of each nucleotide length are marked with green circles against the right hand axis. The blue and black lines show the fraction of targets of each length that were seen at least once for the end-sequencing and shotgun-sequencing libraries, respectively. These are plotted against the left-hand axis. The end-sequencing reads were placed against a database of expected targets using WU-BLAST to account for sequencing errors, and the shotgun-sequencing libraries were placed with MAQ using the HG18 reference genome. Remarkably, it appears that with increasing free energy of the target, there is decreasing capture efficiency. **Middle Plot: End-Sequencing Capture Efficiency.** The blue line represents the fraction of all data that comes from targets of the given length. The black stars show the normalized median capture of each target length, the dark grey represents 50% of the normalized data (25% - 75%) and the light grey represents 80% of the normalized data (10% - 90%). **Bottom Plot: Shotgun Sequencing Capture Efficiency.** The plot is arranged as above. Error bars represent one standard deviation.

To address whether comparing the criteria in a pair-wise fashion would decrease the variability, we sampled pairs of the four most striking effect: ligation arm Tm, extension

arm Tm, GC content and length. The results do not reveal a perfect set of capture parameters (Figure 4-8), but they do reinforce the need to avoid certain regions when designing probes as the capture efficiency is predicted to be very poor. Overall, the variance increases with capture efficiency, and we do not identify any regions exhibiting low variance. In general, the higher the melting temperature of the ligation and extension arms, the more successful the capture will be. The probes with shorter lengths appear to have a higher capture efficiency, but the variance may increase in a disproportional manner (see Figure 4-2) potentially making them an undesirable target. Designing targets to have a GC range close to 0.5 appears to improve the results, but it is not clear whether this will improve with better target design or with more optimized sequencing platforms. It remains to be seen, however, if combining shorter probes with un-optimized GC content or melting temperatures will normalize the results and improve targeted sequencing results.

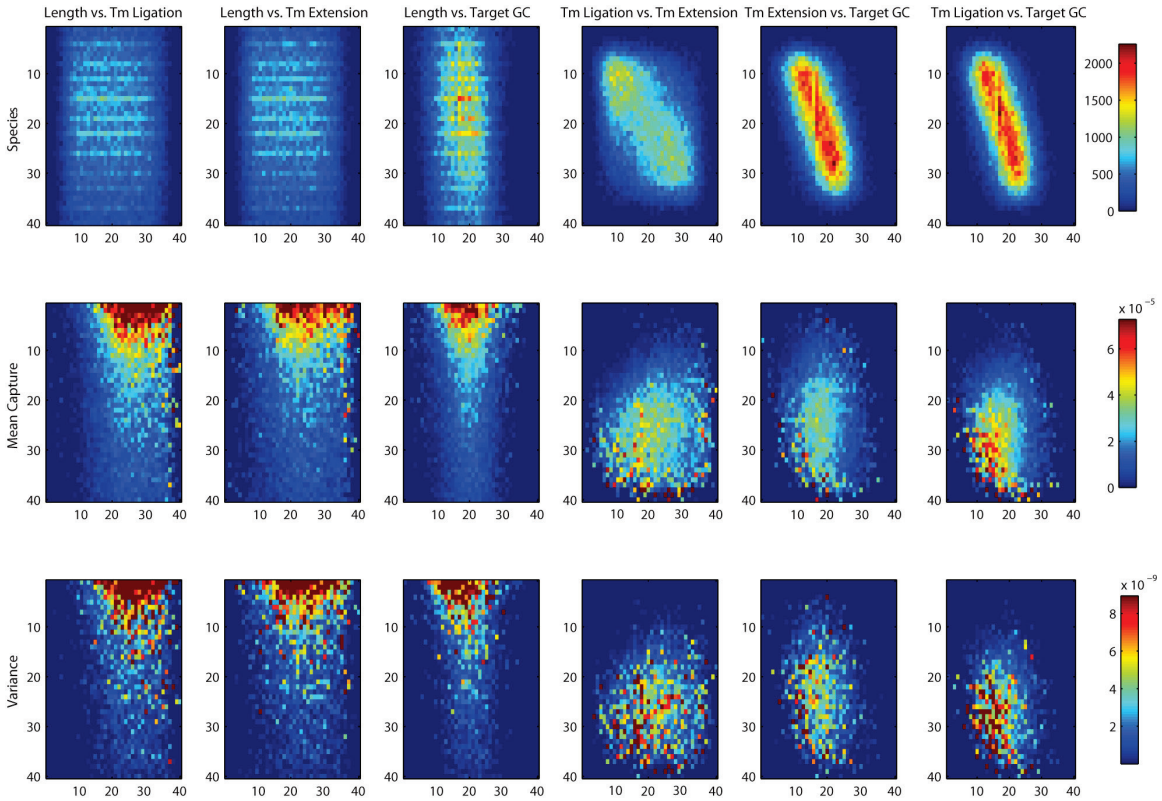


Figure 4-8. Pair-wise Analysis of Target Capture Variables. The full range of values for ligation arm Tm, extension arm Tm, length and target GC content was divided into 40 bins. The values were paired and a heat map was generated detailing the amount of targets designed for each bin, the amount of sequence generated from each bin and the variation of values in the bin. Generally, variation tracked with number of sequence reads. Overall, this data corroborates the earlier figures, but more clearly defines the poor capture for probes with low arm melting temperatures and targets with extreme GC content.

Recent work has attempted to alleviate the biases of molecular inversion probe targeted sequencing through empirically dividing the targets into separate reactions and then normalizing and mixing them before sequencing[23]. While our work here may still need to rely upon such empirical measurements, these guidelines can be utilized to design a more uniform first pass in a targeted sequencing experiment.

References

1 Okou, D. T. *et al.* Microarray-based genomic selection for high-throughput resequencing. *Nat Methods* **4**, 907-909, doi:nmeth1109 [pii]
10.1038/nmeth1109 (2007).

2 Albert, T. J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nat Methods* **4**, 903-905, doi:nmeth1111 [pii]
10.1038/nmeth1111 (2007).

3 Herman, D. S. *et al.* Filter-based hybridization capture of subgenomes enables resequencing and copy-number detection. *Nat Methods* **6**, 507-510, doi:nmeth.1343 [pii]
10.1038/nmeth.1343 (2009).

4 Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, doi:nature08250 [pii]
10.1038/nature08250 (2009).

5 Hodges, E. *et al.* Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat Protoc* **4**, 960-974, doi:nprot.2009.68 [pii]
10.1038/nprot.2009.68 (2009).

6 Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* **27**, 182-189, doi:nbt.1523 [pii]
10.1038/nbt.1523 (2009).

7 Tewhey, R. *et al.* Enrichment of sequencing targets from the human genome by solution hybridization. *Genome Biol* **10**, R116, doi:gb-2009-10-10-r116 [pii]
10.1186/gb-2009-10-10-r116 (2009).

8 Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36**, e105, doi:gkn425 [pii]
10.1093/nar/gkn425 (2008).

9 Quail, M. A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nat Methods* **5**, 1005-1010, doi:nmeth.1270 [pii]
10.1038/nmeth.1270 (2008).

10 Porreca, G. J. *et al.* Multiplex amplification of large sets of human exons. *Nat Methods* **4**, 931-936, doi:nmeth1110 [pii]
10.1038/nmeth1110 (2007).

11 Tewhey, R. *et al.* Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nat Biotechnol* **27**, 1025-1031, doi:nbt.1583 [pii]
10.1038/nbt.1583 (2009).

12 Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat Methods* **6**, 315-316, doi:nmeth.f.248 [pii]

10.1038/nmeth.f.248 (2009).

13 Li, J. B. *et al.* Multiplex padlock targeted sequencing reveals human hypermutable CpG variations. *Genome Res* **19**, 1606-1615, doi:gr.092213.109 [pii]

10.1101/gr.092213.109 (2009).

14 Krishnakumar, S. *et al.* A comprehensive assay for targeted multiplex amplification of human DNA sequences. *Proc Natl Acad Sci U S A* **105**, 9296-9301, doi:0803240105 [pii]

10.1073/pnas.0803240105 (2008).

15 Markoulatos, P., Siafakas, N. & Moncany, M. Multiplex polymerase chain reaction: a practical approach. *J Clin Lab Anal* **16**, 47-51, doi:10.1002/jcla.2058 [pii] (2002).

16 Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**, 365-386 (2000).

17 Lawyer, F. C. *et al.* High-level expression, purification, and enzymatic characterization of full-length Thermus aquaticus DNA polymerase and a truncated form deficient in 5' to 3' exonuclease activity. *PCR Methods Appl* **2**, 275-287 (1993).

18 Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728-1732, doi:1117389 [pii]

10.1126/science.1117389 (2005).

19 Harismendy, O. & Frazer, K. Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis technology. *Biotechniques* **46**, 229-231, doi:000113082 [pii]

10.2144/000113082 (2009).

20 Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**, 3406-3415 (2003).

21 SantaLucia, J., Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci U S A* **95**, 1460-1465 (1998).

22 Peyret, N. *Prediction of Nucleic Acid Hybridization: Parameters and Algorithms* Ph.D. thesis, Wayne State University, (2000).

23 Deng, J. *et al.* Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat Biotechnol* **27**, 353-360, doi:nbt.1530 [pii]

10.1038/nbt.1530 (2009).

Future Directions

The speed with which the cost of DNA sequencing is decreasing is perhaps without parallel in the history of technology development. The initiation of Second Generation Sequencing with the publication of two different methods in 2005 (Appendix D) culminating in the most recent effort (Appendix G) has led to a drop in the consumables cost of DNA sequencing from $38.7/megabase to $0.0125/megabase. Along with this dramatic increase in sequencing capabilities comes the task of analyzing and presenting the data so that it can be of the greatest use. This dissertation discusses work performed to both increase our sequencing capabilities and make this information accessible to individuals and their health care clinicians. To further this goal there are many future directions that should be pursued. I will first outline directions for each set of experiments, and then for the broader issues facing this research.

Chapter 2 discusses the development of Trait-o-matic and its utilization in analyzing 25 public whole or partial genomes. This tool currently calculates and then annotates all non-synonymous single nucleotide substitutions with descriptors found in four different databases. Additionally, it calculates HapMap-based frequencies for each variant, and predicts the deleterious nature of every non-synonymous substitution. Currently, only single nucleotide substitutions are analyzed by the tool – insertions deletions and splice site variants need to be incorporated into the algorithm. While only a small number of diseases have been associated with these types of variants, it is partially due to our lack of tools allowing genomic scanning for them. New technologies have increasingly brought copy number variants to the forefront in potential causes for neurological disorders, and they need to be incorporated into Trait-o-matic.

While the deleterious ranking and frequency features are useful in bringing certain potentially clinically relevant variants to the individual's attention, it cannot yet be considered a useable clinical tool. The current annotation and filtering tools still leave too many variants in the "high priority bin" that the clinician would need to sort through for clinical utility. Many diagnostic laboratories utilize a system of ranking to include five or six different levels of causation, ranging from pathogenic to benign[1]. We would ideally have a database of all previously described variants together with their clinical causation-ranking. Furthermore, clinical decisions are rarely based solely upon the DNA genotype of the individual. As described in regards to HFE in Chapter 1, many variants take on clinical relevance only when associated with specific phenotypes. Currently, our tool does not collect phenotypic data nor annotate variants with the phenotypic data that should be utilized to amend its clinical ranking.

Furthermore, regarding the creation of new correlations between variants and phenotypes, there are many discovery tools that should be implemented into the genome analysis utility. As discussed in Chapter 1, a number of groups have used the power of whole genome or whole exome sequencing to uncover the molecular basis for Mendelian disorders. According to OMIM, only 41% of the 6,500 Mendelian or possibly Mendelian diseases listed have a causative gene identified. To uncover genes for monogenic disorders, the ability to contrast the variants found in controls and cases needs to be implemented, taking into account the inheritance patterns of the disease. For polygenic diseases and quantitative trait loci, complex algorithms need to be developed to sample the probabilities of different genetic interactions causing the disease. With these

functionalities in place, Trait-o-matic is a serious contender for the preliminary analysis of whole genome sequences.

In the analysis of genomic variants, one question that still faces Trait-o-matic is that the clinical nature of these variants is open to interpretation, and with each additional genome (currently) requiring close to one hour of analysis, it is not feasible for large studies. To overcome these issues, the "Trait-o-matic team" has begun the development of the GET evidence database. Its name is an acronym for Genome + Environment = Traits provided sufficient evidence, which we hope encapsulates the future of genome analysis. While this project is still in its infancy, we have begun this new database by documenting clinically relevant variants. This will entail identifying the variants, relevant literature, associated phenotypes and clinical prognosis. The database has been established as a wiki to encourage community participation.

Furthermore, in this new database, the variants will be connected with any genomes that they are found in, enabling detailed analysis of the particular genome, and potentially the participant's phenotype when an unexpected symptomatic or asymptomatic phenotype is reported. The wiki nature of the databases also provides a way for the individual to communicate any clinically relevant outcomes related to the variants in discussion.

Chapter 3 discusses the Personal Genome Project and the sequencing and analysis underlying the project's first data release. Targeted sequencing was used to capture exonic regions, and a web-based utility was used to generate thresholds for high-confidence data. A bottleneck in the library construction led to difficulty in using strict coverage criterion for assessing the accuracy of called variants, since many of the reads

were actually PCR duplicates. While the problem of PCR duplicates has been discussed by a number of papers (see Quail et al[2]., quoted in Chapter 3), we have added another tool to ameliorate its affects through a user-defined threshold necessitating a specific number of independent reads. Additionally, we have developed an interface allowing user defined coverage and quality thresholds so as to maximize correlation with independently derived data.

The accuracy-threshold algorithm currently compares microarray data with sequencing data, and the user can optimize three thresholds to maximize the concordance of the subset of variants, and presumably, the accuracy of all variants called. A natural extension of this tool would be to automate it so that the thresholds tailored to the library at hand are automatically calculated. Furthermore, many different algorithms have been developed to map NGS reads and calculate the accuracy of the reads. While this tool currently uses MAQ, incorporating in different algorithms and giving the user the choice of which to run is a logical extension. Finally, a number of groups have begun comparing the accuracy of the different sequencing platforms by trying to re-identify the same variants when sequencing the same sample with each device[3-4] (see Figure 2-3). A desirable algorithm would automatically compare data from any sequencing or microarray platform and, in addition to providing higher accuracy scores to concordant variants, it would incorporate the known strengths and weaknesses of each platform in terms of predicted error types and variant detection sensitivity.

The fourth chapter discusses targeted sequencing using molecular inversion probes (MIPs) and elaborates upon the biases uncovered through a set of probes targeting approximately 20% of all exons. It also sets the stage for Appendix F where this data is

incorporated into an algorithm, MIPTAG Pro, which automates the design of MIPs for a user-defined set of targets. This algorithm has been used to design and order MIPs for two collaborators' projects and the data is still being analyzed. This analysis will allow us to judge the effectiveness of the algorithm and further refine it. Concurrently, our lab is designing MIPs targeting all exons using an independent algorithm and empirically binning the probes into different subsets and subsequently normalizing and combining the libraries. This data will also prove invaluable in discovering the nature of these targeting and sequencing biases.

The initial action taken to improve uniformity of capture – transitioning from Stoffel Large Fragment to Pfu Polymerase, proved very effective. It remains to be seen whether different polymerases or ligases would similarly increase the uniformity. Additionally, PCR additives may reduce some biases, such as the addition of betaine for GC-content normalization.

Appendices A, B and C represent ongoing work on further decreasing the cost of DNA sequencing. They discuss methods for decreasing reagent consumption through the use of microfluidics technology, replacing the emulsion PCR with much more cost-effective rolonies, and increasing the density of library molecules through arraying them in an orderly fashion with nanogrids. Taken together these improvement represent a >400x increase in sequencing capacity on the Polonator, decreasing the cost of a 30x genome on the Polonator from $97,000 to less than $250 (these calculations include consumables and machine amortization). While prototypes for each of these improvements have been created, more work is still needed before they can be implemented on the Polonator.

For the microfluidics device, channels were etched into a silicon wafer which was then anodically bonded to a coverglass. While silicon etching provides the greatest flexibility in terms of design and channel depth, the process of anodic bonding is not amenable to high-throughput production. Typically, the process is used for silicon wafers which are then diced into dozens of individual pieces. Due to the large size of the flowcell, our initial design allowed one flowcell per wafer; further optimization, however, fit three onto each wafer. With this format vendors were still unable to provide us with cost estimates less than a few hundred dollars per flowcell after an initial large payment for further prototyping. To avoid the anodic bonding and the cost of silicon wafers we also pursued a glass on glass bonding technology offered by Micronit Microfluidics. While the cost savings promised by such a technology would justify the cost per flowcell, it was difficult for a research lab to justify the large upfront cost for prototyping.

The task of creating a microfluidic device began when the gasket for the Polonator was 160 microns in height. Since then this gasket has been reduced to 80 microns, which is basically the limitation for this technology. To create an even lower chamber height, Complete Genomics creates a mixture of 50 micron beads and adhesive that is patterned to create channels (Appendix G). It remains to be seen whether this method can be used with increasingly smaller beads of 28 or 15 microns.

In Appendix B, research is presented showing the sequencing of rolonies with both ligation and synthesis based methods. Circularization of library molecules and amplification with Phi29 both proved to be less reproducible than desired, and we were never able to track down the source of these inconsistencies. Different buffers and reducing agents were sampled and it appears that none of these actually inhibit the

polymerase from working. Rather, it seems that either residual exonucleases or repeated freeze-thaw cycles are very deleterious to the single-stranded circular DNA. It also appears that a fresh stock of Phi29 polymerase is important.

To attach rolonies to a glass coverslip numerous chemistries were used both to tether the rolony at only its terminus or throughout the entire molecule. Intuitively, there appears a tradeoff between these two methods; with terminal attachment more of the molecule is accessible to the sequencing reagents, but the free-floating end can rotate through a fairly large radius and a single break in the single-stranded rolony will leave it free to float away. Imaging such end-tethered rolonies on our epifluorescence microscope showed a number of such "lift-offs" in a relatively short timeframe. This may have been exacerbated by the photons used to generate the fluorescence, but it is still a matter of concern when one wants to sample molecules over a span of 3-4 days.

Ironically, the best surface attachment did not involve any chemical bond, but rather charging the glass surface with positive amine groups to which the negatively charged DNA will bind. It can be assumed that ionic charges in the buffer will affect the static binding but this hypothesis has yet to be tested.

Complete Genomics (Appendix G) utilizes a 14bp inverted repeat (7 A's followed by 7 T's) to make their molecule fold up on itself. We have not seen a difference between rolonies containing a hairpin and rolonies without this sequence. One major issue plaguing rolonies is their relative lack of signal intensity when compared with beads. Theoretically, a loose sphere created by a concatemer should rival the signal of beads given a long enough concatemer, especially since half the signal from every bead is blocked by the bead itself. This has not been our experience, however, and the rolonies

typically need exposures much longer than those required by beads. An increase in rolony signal would aid much of the research presented here.

Appendix C addresses nanogrids and ordered arrays of both beads and rolonies. Initial experiments were done with e-beam lithography in collaboration with Brian Chow from Joe Jacobson's group at MIT. While this route was chosen due to the flexibility of grid specifications enabled by e-beam technology, the small patterned areas and difficulties maintaining cleanliness with exceedingly small spots created difficulties. The e-beam wafers were comprised of 25X25 spots every 5mm along the entire surface of the wafer, and the silicon was coated with chemicals to make it repel the DNA (passivated) forcing the rolonies to the chemically active spots. We hypothesize that the large hydrophobic passivated silicon surface repelled the large DNA molecules so effectively that the rolonies rarely encountered the activated spots, making the scheme appear to not work very well. We later discovered that while bare silicon will bind rolonies indiscriminately, the rolonies will preferentially bind to chemically activated spots, allowing for ordered arrays even without the additional passivation. This only occurs, however, when the spots are relatively close, so a large patterned surface is needed.

Deep-UV patterning of <400nm features requires very expensive equipment, and there are very few suppliers available for prototyping. As the lithography continues to develop we expect that access to these capabilities will increase. In the meantime we have demonstrated the feasibility of patterning features as small as 300nm with 2-beam interference lithography. The capabilities of this method are limited to perfect grids, but for our process that is entirely acceptable. Additionally, patterning can be done on any

substrate, not just silicon. As we refine our grids we expect this technology to be fully integrated into the Polonator in a relatively short while.

In broader terms, while I am sure that the implementation of some of the technologies outlined in Appendices A-C will lead to further cost-savings, I am also confident that the community will be able to maintain the pace of cost reduction. The potential for third generation sequencing is truly exciting. While there is no clear definition for the third generation of DNA sequencing, the proposed methods include non-cyclical processing of DNA with dramatically longer read lengths, not requiring enzymes or library construction. While there may be no platform capable of combining all these elements in the near future, any of them can potentially provide a dramatic cost reduction.

In the same dramatic way that SGS required a viable replacement for the upfront methods of cloning and PCR currently in use for Sanger sequencing, third generation sequencing will also require us to develop new ways to think about targeting genomic regions. Read lengths in the megabases will require new technologies to select chromosomal regions instead of exons or genes. Promising technologies for this includes sorting methods based upon flow cytometry or selection methods based upon laser capture microdissection technologies. Efficient high throughput methodologies based on these capture methods will have to be developed to match the potential speeds of these new sequencing methods.

The last 5 years have seen a decrease in DNA sequencing cost of more than one order of magnitude per year. With this speed of development the $100 genome is well within our reach and personal genomes will soon be a reality. Some technologies for

sequencing and tools for analysis are discussed in this dissertation, but an analysis of the ethical and moral implications of such information is beyond the scope of this research. Will we better off knowing the susceptibilities inherent in our genomes? A recent paper on Alzheimer's disease susceptibility finds that individuals react favorably to this information[5], but the opposite appears to be the case for Huntington's disease, where even those found to not be carriers were more prone to depression, especially when there were pre-existing psychological issues[6]. How can we classify diseases and personality types to maximize clinical benefit and avoid depression? While it may seem proper to not safeguard and not disclose this data, how can this be accomplished and under what circumstances should the data be queried by the individual or a health professional? These questions need to be addressed by the community, but undoubtedly a greater understanding of human individuality will eventually increase the health and well-being of all mankind.

References

1 Plon, S. E. *et al.* Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum Mutat* **29**, 1282-1291, doi:10.1002/humu.20880 (2008).

2 Quail, M. A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nat Methods* **5**, 1005-1010, doi:nmeth.1270 [pii]
10.1038/nmeth.1270 (2008).

3 Shen, Y., Sarin, S., Liu, Y., Hobert, O. & Pe'er, I. Comparing platforms for C. elegans mutant identification using high-throughput whole-genome sequencing. *PLoS ONE* **3**, e4012, doi:10.1371/journal.pone.0004012 (2008).

4 Harismendy, O. *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* **10**, R32, doi:gb-2009-10-3-r32 [pii]
10.1186/gb-2009-10-3-r32 (2009).

5 Green, R. C. *et al.* Disclosure of APOE genotype for risk of Alzheimer's disease. *N Engl J Med* **361**, 245-254, doi:361/3/245 [pii]
10.1056/NEJMoa0809578 (2009).

6 Gargiulo, M. *et al.* Long-term outcome of presymptomatic testing in Huntington disease. *Eur J Hum Genet* **17**, 165-171, doi:ejhg2008146 [pii]
10.1038/ejhg.2008.146 (2009).

Appendix A

DNA Sequencing by Ligation on Surface-Bound Beads in a Microchannel Environment

This paper was presented at μTAS 2008 as a Refereed Conference Publication.

C.R. Forest[1†]*, A.M. Rosenbaum[1]*, and G.M. Church[1], DNA sequencing by ligation on surface-bound beads in a microchannel environment, Proceedings of the 12*th* *InternationalConference on Miniaturized Chemical and Biochemical Analysis Systems* (μTAS), San Diego, CA, October 12-16, 2008

[1]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA; [†]Current address: The George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

*These authors contributed equally

Author Contributions C.R.F. and A.M.R.designed and performed all experiments and G.M.C. supervised all aspects of the study.

Acknowledgements We are indebted to Greg Porreca, Rich Terry, Erik Garrison Mirko Palla and the MIT-Microsystems Technology Laboratories for their assitance. This work was sponsored by the NIH-NHGRI Centers for Excellence in Genomics Grant.

Abstract

Advancements in DNA sequencing can potentially enable genome-wide studies of associations between mutations and traits in large population pools. In this work we have developed and implemented a microchannel chip to reduce biochemical reagent volumes for sequencing by ligation on surface-bound beads by 12×. Further, bead binding selectivity in 8.5 μm deep channels is enhanced with C_4F_8 surface passivation. This chip integrates with a thermally controlled vacuum chuck and an automated instrument, the Polonator, to perform cyclic biochemical reactions and imaging with reagent volume and cost reduction over preceeding technologies.

Keywords: DNA sequencing, ligation, microchannels, surface passivation

Introduction

DNA sequencing advancements in cost and throughput are undergoing intensive development to enable widespread discovery of genomics information. Some groups have focused on conventional dideoxy sequencing, making improvements that include performing separations in microchannels[1], while others have shifted to using cyclic array sequencing technologies[2]. Our group has previously published a cyclic array technique, sequencing by ligation, using DNA templates tethered to immobilized, 1 μm diameter, ferromagnetic beads[3]. In this work we report on the extension of this technique to a microchannel chip for the open-source Polonator instrument, with resulting 12× reduction in reagent usage-a dominant cost.

Design and Manufacture

The microchannel chip contains an array of 16 or 32 addressable channels etched in silicon bonded to borofloat glass. The channels are addressable through individual ports to permit multiplexing of bead arrays. Subsequently sealing these ports with poly-imide tape allows a single inlet and outlet to deliver common reagents for sequencing. Beads are bound in monolayers on the channel's glass surface by selectively passivating and silanizing the silicon and glass, respectively. Passivation is achieved through C_4F_8 fluoropolymer deposition prior to anodic bonding, and silanization is performed with aminopropyltriethoxysilane that enables bead binding through NHS-ester crosslinking. When mounted on a vacuum chuck with peltier thermal control, the Polonator's epi-fluorescence microscope and reagent handling system allows sequencing from an area containing almost 4×10^9 DNA-loaded beads.

Results and Discussion

The channel layout and fabrication process are shown respectively in Figures 5-1*a* and *b*. Each channel is 2 mm wide with 160 mm^2 active, silanized area that is typically arrayed with 60×10^6 beads. Thru-hole channel ports, either plasma-etched or drilled, are located at either end and the center of each channel for bead loading.

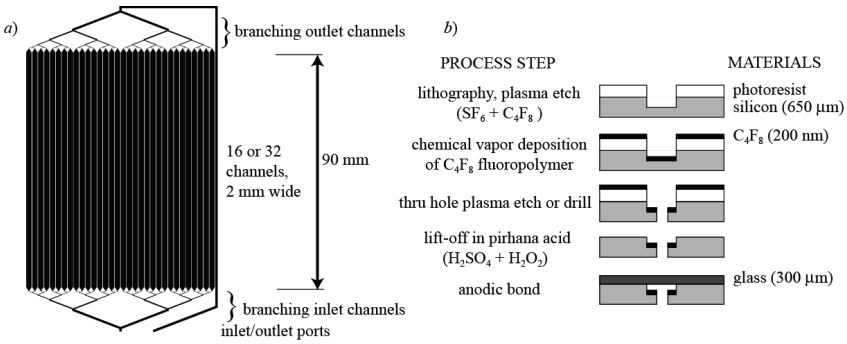


Figure 5-1. a) Mask design with 16-32 channels and b) fabrication process for microchannel chip, as patterned onto 150 mm diameter silicon wafers.

Figure 5-2 shows a portion of a channel surface as fabricated using this process and design. The silicon channel is 8.5 µm deep with a 200 nm C_4F_8 passivation treatment that inhibits subsequent silanization. This passivation layer has been measured to be 5% more autofluorescent than bare silicon at 550 nm (for Cy3 fluorescent dye), which is tolerable; this surface modification mitigates bead binding by more than 18× relative to bare SiO_2 and 2× relative to bare silicon as indicated in Table 5-1.

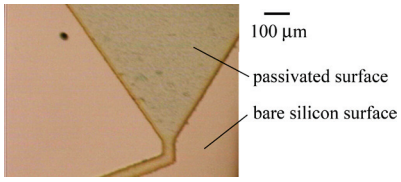


Figure 5-2. Photograph of channel with C_4F_8 surface passivation. The 200 nm thick passivation layer enables selective silanization for bead binding in a monolayer.

Table 5-1. Bead binding selectivity on microchannel walls. Gravity predisposes the beads to bind to the SiO_2 wall. Selectivity is enhanced with C_4F_8 passivation.

	Polymer (C_4F_8)	**Silicon**	**Glass (SiO_2)**
Beads bound material wall	33	66	594
Beads bound to SiO_2 wall	19,700	17,600	17,800
% spurious binding	0.17%	0.38%	3.34%

The microchannel chip and its thermally controlled vacuum chuck are shown in Figure 5-3. Kinematic registration features on the chuck accurately and repeatably located the

chip to 25 µm laterally. Thermal mapping of the chuck was performed at biochemistry temperatures to assess accuracy and uniformity as shown in Figure 5-3*c*. At the temperature extremes (5 °C and 56 °C), the total surface temperature range is 4.9 and 3.7 °C respectively, which is acceptable for selective DNA hybridization and refrigeration after sequencing.

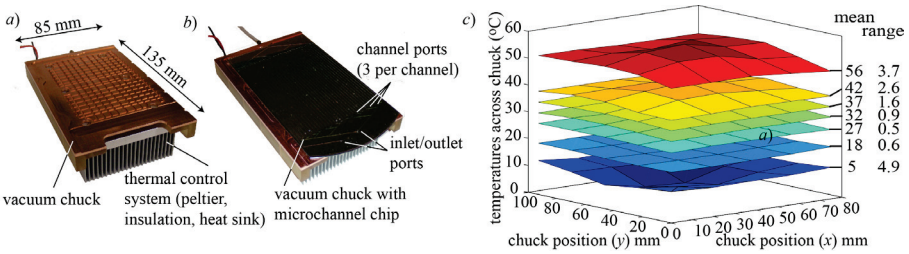


Figure 5-3. a) Vacuum chuck b) with microchannel chip. The assembly constrains the chip accurately and repeatably for fluorescence microscopy and c) controls its temperature between 5-56 °C sufficiently uniformly.

DNA sequencing results on surface-bound beads in the microchannel environment using sub-µL reagent volumes are shown in Figure 5-4. This false-colored image is a composite of separate fluorescent images corresponding to each nucleotide base. Bead confinement in the 50 µm wide channel is evident.

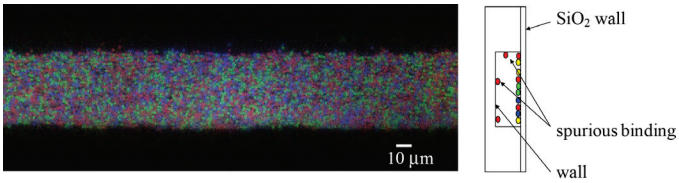


Figure 5-4. DNA sequencing by ligation demonstrated in a 50 µm wide channel. The surface bound beads have unique tethered DNA templates that can be queried using biochemical protocols previously published[3].

Conclusions

This flexible, addressable array of surface-modified microchannels offers substantial reagent reduction for bead-based cyclic array sequencing, bringing us one step closer to realizing the vision for a $1,000 sequenced human genome.

References

1 Emrich, C. A., Tian, H., Medintz, I. L. & Mathies, R. A. Microfabricated 384-lane capillary array electrophoresis bioanalyzer for ultrahigh-throughput genetic analysis. *Anal Chem* **74**, 5076-5083 (2002).

2 Shendure, J. A., Porreca, G. J. & Church, G. M. Overview of DNA sequencing strategies. *Curr Protoc Mol Biol* **Chapter 7**, Unit 7 1, doi:10.1002/0471142727.mb0701s81 (2008).

3 Kim, J. B. *et al.* Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science* **316**, 1481-1484, doi:316/5830/1481 [pii]
10.1126/science.1137325 (2007).

Appendix B

Multiplex Rolony Sequencing

Abraham M. Rosenbaum[1]*, Brian Y. Chow[2]*, Gregory J. Porreca[1], Jay A. Shendure[1], Peter Lee[1], Jun Zhu[1], Kun Zhang[1], John Aach[1], Joseph M. Jacobson[2], George M. Church[1]

[1]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA; [2]MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

*These authors contributed equally

Author Contributions A.M.R. performed all molecular biology with help from G.J.P., J.A.S., J.Z., and K.Z.; Surface chemistry was designed by B.Y.C. and performed with help from A.M.R. and P.L.; G.M.C. and J.M.J. supervised all aspects of the study.

Acknowledgements We are grateful for the help and advice provided by all the members of the Church Laboratory and the Molecular Machines Group, as well as helpful discussions with Jenny Göransson (Nilsson Laboratory, Upsalla University), Rade Drmanac (Complete Genomics) and Xiaohua Huang (UCSD). We thank NHGRI for funding support.

Abstract

The cost of DNA sequencing has dropped by over four orders of magnitude through the continuing development of Second Generation Sequencing (SGS) techniques. Polonator, the only open source SGS platform, currently costs less than $1/megabase, and 50% of that cost comes from the emulsion amplification of one micron beads. Concatemers generated through rolling circle amplification of circular library molecules have been shown to collapse upon themselves in a clonal fashion to form one micron spheres detectable through hybridization of fluorescent probes. Here we present the attachment of these molecules to a chemically modified slide surface, and the identification of the different synthetic molecules through both sequencing by synthesis and sequencing by ligation.

Introduction

In the past five years second generation sequencing (SGS) methods have decreased the cost of DNA sequencing by over four orders of magnitude. The SGS process consists of a number of processes: library molecule preparation, library molecule deposition in a flow cell, cyclical enzymatic based analysis of nucleotide composition, and detection of a signal generated by the enzymatic process. For a general overview of the approach each of the five platforms have for each step, see Chapter 1. Both the open-source Polonator and the related SOLiD platform prepare the library through emulsion PCR[1-3] to clonally amplify the library to one micron ferromagnetic beads[4-6] which are then covalently linked to the surface of a flow-cell. Creating bead-based PCR colonies (polonies) as a separate step from library binding allows for maximal packing of library molecules, as contrasted with bridge PCR used by Illumina GAII where the molecules

must be diluted to prevent overlapping polonies, resulting in a lower density. For the Polonator, however, this comes at a very high cost – emulsion PCR comprises almost 50% of the total per megabase cost.

A faster and more cost efficient method for creating PCR-like colonies involves the circularization of the library molecule and its amplification with rolling circle amplification (RCA). Two papers in 1998 first report the use of Phi29 polymerase based RCA and its ability to synthesize a contiguous concatemer many kilobases in length in 12h[7], and the ability to initiate the polymerization from surface based primers allowing detection of several thousand unique molecules on a 25X75mm slide[8]. Later it was demonstrated that in solution these RCA colonies ("rolonies") reach a diameter of 1 micron after about an hour of polymerization and they remain discrete concatemers, not clumps of numerous molecules[9].

Different methods have been used to attach large DNA molecules to the surface of a slide. These include initially attaching the terminus of the DNA molecule to a surface bound vinyl group, followed by stretching the DNA molecule and drying it onto the surface[10]. Various methods exist for attaching just the terminus to the surface while the rest of the molecule remains free in solution. These include biotin-streptavidin conjugation[11-12] and covalent attachment through NHS-ester chemistry[13]. These methods, however, only tether one end of a long molecule to the surface which creates packing and imaging problems as the free end of the molecule is free in solution and subject to Brownian motion. The random-walk of the free end of end-tethered Lambda DNA (48,502bp) exhibits a circumference of 2.7μm[14]; limiting the maximum density to 135,000 features/mm^2. Furthermore, the ability of single-stranded DNA (ssDNA)

molecules over 20nt long to interact statically with the modified slide surface can be a complicating factor[15].

Here, we perform both sequencing by synthesis and sequencing by ligation of rolonies averaging one micron in size. Additionally, we demonstrate the clonal attachment of rolonies to surface modified slides at densities exceeding 35,000 features/mm^2 and their detection through sequencing by ligation. Improvements in the process are likely to increase this number, as clonal rolonies have been demonstrated at densities exceeding 40,000,000/mm^2 detectable with gold nanospheres and electron microscopy[16], and in a study published since this work at densities of 300,000 features/mm^2 on a poly-lysine activated surface detected through hybridization of fluorescent probes[17].

Results

To introduce rolony technology to Polonator sequencing we first assessed some of the methods reported by other groups. One of the earliest approaches[8] entailed covalently attaching primers to a slide surface, hybridizing circular templates to these primers and initiating rolling circle amplification. In this paper, detection was performed through hybridization of fluorescent probes. We successfully replicated this protocol, and demonstrated specificity through comparison of different ratios of input probes (Figure 6-1A and 6-1B). Additionally, when utilizing amino-NHS chemistry to couple amine-functionalized primers to a PDC-treated silanized surface, we found that separating the functional group from the primer with a 12-carbon linker allowed denser packing than when it was separated with a 6-carbon linker. Due to the relative inefficiency of the surface-primed reaction as assessed by the total number of fluorescent probes hybridizing

to the rolonies, we developed two similar approaches that yielded greater overall fluorescence. In the first, instead of hybridizing circular template to surface-bound primers, we pre-amplified the circular templates and bound the concatemers to the surface bound primer. In this manner, many surface bound primers would hybridize to each concatemer, and the strand-displacing polymerase would extend each primer to the end of the concatemer. In this instance, to distinguish between different species we utilized sequencing by extension (SBE) and the results are shown in Figure 6-1C. Additionally, we found that this signal can be increased even more through a second round of amplification. To perform this we hybridized a primer complementary to the surface-bound strand and repeated the strand-displacement amplification. To confirm the increase in surface-bound molecular repeats we used a stringent denaturation protocol to remove all material not directly attached to the surface, and then hybridized a fluorescently labeled primer. Comparison of a region with only one round of amplification to one with two rounds show a marked increase in signal (Figure 6-2). While this experiment was performed using Bst Polymerase, similar results were found when using either Vent Polymerase or thermophilic Helicase-Dependent Amplification (tHDA kit, Biohelix).

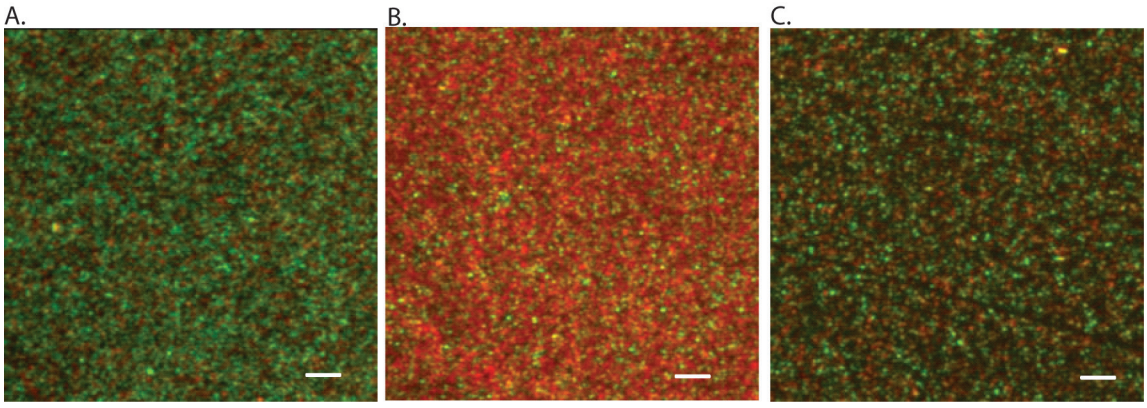


Figure 6-1. Surface-Bound Primer Extension. A. Extension of a Mixture of Two Circular Templates in a Ratio of 1:4. A mixture of two species of circles with a region complementary to the surface-bound primers was hybridized to the primers. RCA was performed to extend the primers and the identity of the surface-bound concatemers was queried through hybridization of fluorescently labeled probes. The first concatemer hybridized to a Cy5-labeled primer and is false-colored red in the image, and the second concatemer hybridized to a Cy3 labeled primer and is false colored green in this image. **B. Extension of a Mixture of Two Circular Templates in a Ratio of 4:1.** This image was generated in the same manner as B, with the exception that the templates were present in the opposite ratio. The size-bar shows a length of 10 micron and all images are taken with a 20x objective (NA 0.75). **C. Extension from Rolling Circle Amplification-generated Concatemers Hybridization.** Two concatemers with a region complementary to the surface-bound primers were generated with RCA and then hybridized to the primers. The primers were then extended with a strand displacing polymerase. Since numerous primers could bind to a single concatemer, the signal from each original molecule was greater than in B and C. To detect the concatemers sequencing-by-extension was performed as described in the methods.

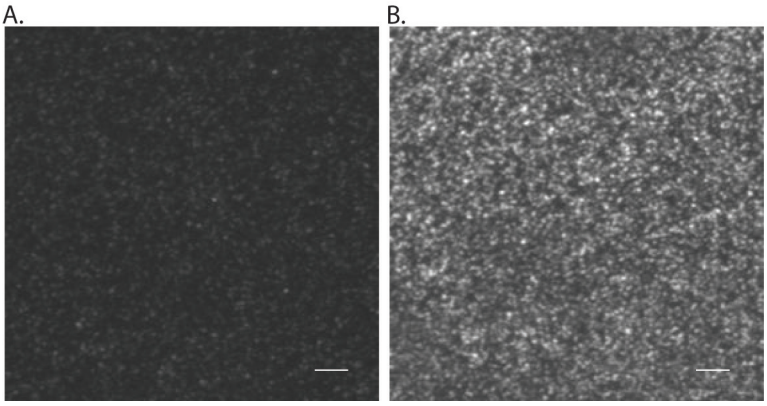


Figure 6-2. Secondary Extension off Surface Bound Primers. A. Extension from RCA generated Concatemer Hybridization. The slide was prepared as in figure 1A. After extension a Cy3 labeled primer complementary to the concatemer was hybridized to visualize the extended molecules. **B. Secondary Amplification.** All DNA not directly bound to the surface was removed through stringent denaturation, and an amplification mixture containing strand-displacing polymerase, complementary primers and dNTPs was added. After this reaction was stopped fluorescent probes were hybridized to measure the increase in molecule size. The same exposure and objective was used for both images; the measure bar depicts 10 microns.

Dilutions of these latter experiments demonstrated the difficulty in confining the rolonies to small locations. Upon hybridization to the surface bound primers, the concatemers, even in high salt concentrations, would unravel as they found

complementary sequence, thereby leaving a very large footprint. To increase the signal without a secondary amplification through surface bound primers we used hyper-branched RCA (hRCA), where a much larger input molecule is created. While this technique proved useful when each synthetic template was present in a separate reaction mixture (Figure 6-3), a mixture of different library species in a single reaction would most likely lead to a large proportion of chimeric molecules. Additionally, binding the biotin-containing hRCA products to a streptavidin coated surface generated rolonies that were not static and were subject to Brownian motion. This is consistent with a large molecule in which only the bottom-most portion is bound to the surface. While this larger sphere allows for increased signal intensity, it also complicates image analysis by necessitating additional algorithms to analyze a moving target. The smaller linear RCA molecule, however, remained bound to even an unmodified surface, presumably due to electrostatic interactions. We found that binding rolonies made in solution to the surface achieved a sufficiently high S/N ratio to enable sequencing even with a 250ms exposure, and remained sufficiently compact to enable analysis of sufficient numbers to rival the current sequencing throughput.

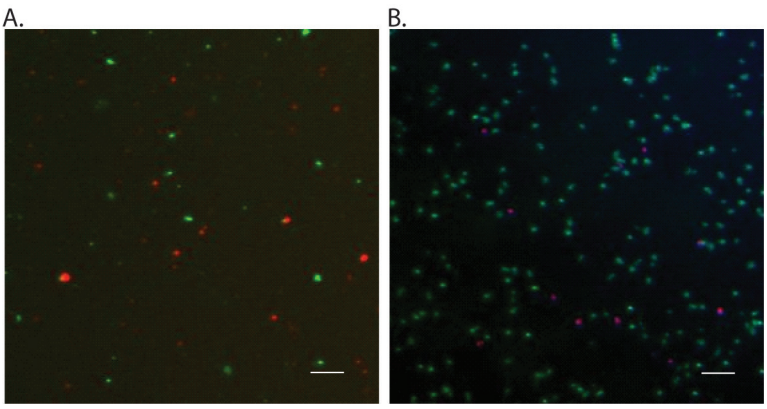


Figure 6-3. Binding Concatemers Post-Amplification. A. Surface Bound Biotinylated hRCA Concatemers. In two separate reactions containing either cT4' or cT5' templates, biotinylated forward primers and fluorescently labeled reverse primers were used to amplify the template which was later bound to a strepatividin coated surface. These molecules exhibitied Brownian motion, demonstrating that a significant part of their mass was not attached to the surface. **B. Linear RCA Concatemers Amplified Clonally in a Single Reaction.** The above templates were then mixed in the presence of only one universal primer for a linear RCA reaction. After the reaction was stopped, fluorescent probes were added to demonstrate the clonality of the reaction. The molecules are deposited on the surface through electrostatic interactions. Cy3 is false colored green and Cy5 red in both images and the measure bar is 10 micron in length.

To estimate the number of repeats in reach rolony, we stretched the rolonies using a single attachment point for each rolony and a slowly receding meniscus to fully extend each molecule. While existing protocols called for creating this meniscus through the angling and slow removal of a coverslip, we found that with this method it was difficult to fully extend the rolonies and obtain reproducible results (Figure 6-4). Much better results were obtained by incubating biotin-capped rolonies on an activated streptavidin slide in a 55°C humidity chamber. Over the period of 30 minutes, the slow drying of the spot led to the extension of all rolonies along the periphery. While the length of a single base pair of double-stranded DNA is well established as 0.34nm, the length for single-stranded DNA depends upon the base, but lies somewhere between 0.32 and 0.51nm[18-19]. With a 60m RCA reaction and Phi29 polymerase processing 2,240bp/minute, the maximum length of the expected 134kb DNA strand is between 40 and 80 microns, depending upon the makeup of the ssDNA strand. With the fully stretched DNA we find

that the most common strands are less than 20 microns, which would correspond to 25-50kb, but it is likely that these strands are not fully extended. The longest strands found are between 50-65 microns, approximately the expected length. Since, however, ssDNA can be stretched an additional 0.6nm/bp[20] when force is applied, it is possible that these numbers grossly overestimate the length of the rolonies.

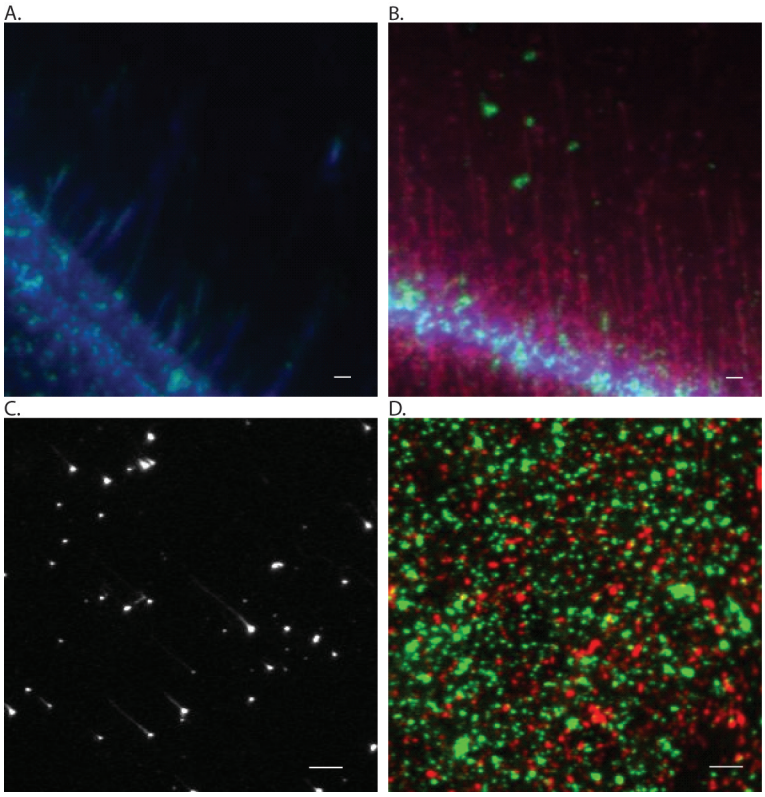


Figure 6-4. DNA Fiber Stretching A-B. Unintentional Stretching. A mixture of concatemers generated by RCA were spotted onto a silanized slide and incubated in a humidity chamber. Drying towards the edges of the spot enabled estimates as to the length of the concatemers. Assuming the molecules are fully extended, the length of ssDNA would be 0.58nm/base. The mode length of the molecules was 6 microns, or 13,000 bp, while the longest molecules were 52 micron, or 110kb. **C. Methodical Stretching.** Using a protocol for fiber stretching, this experiment was repeated with concatemers containing a single biotin attached to a streptavidin activated slide, and, while it was difficult to achieve full stretching, the above lengths were corroborated. **D. Unstretched Molecules.** Nevertheless, even when following protocols for DNA stretching many of the molecules appeared to bind the surface in numerous places, preventing their extension with the receding meniscus. All flourophores are false-colored as above and the measure bars are 10 microns in length.

The wide range of rolony sizes is further supported by data analyzing the fluorescence intensity of the rolonies. Aliquots were removed from a RCA reaction at eleven different intervals over 180 minutes, bound to a slide and then probed with

fluorescently modified primers. To analyze the signal, custom software developed by our lab was used to generate a binary mask to define >1000 rolonies in a single image[21], and the average and standard deviation of all the rolonies in a single image was calculated (Figure 6-5). Ignoring two outliers, the fluorescence increased in a linear fashion through 90 minutes. Between 90 and 180 minutes, however, the mean fluorescent intensity remained the same while the standard deviation markedly increased. This demonstrates the cessation of elongation experienced by a large number of rolonies, while others kept growing. Since a single Phi29 polymerase is capable of replicating the Phi29 genome, demonstrating a processivity of at least 19,285bp, and *in vitro* reactions have demonstrated processivity of 70kb[22], it stands to reason that this is partially due to some molecules having the enzyme fall off and others possible experiencing a re-initiation of elongation. While the amount of ssDNA around the elongation site will make re-initiation difficult, the former is perhaps the more plausible explanation.

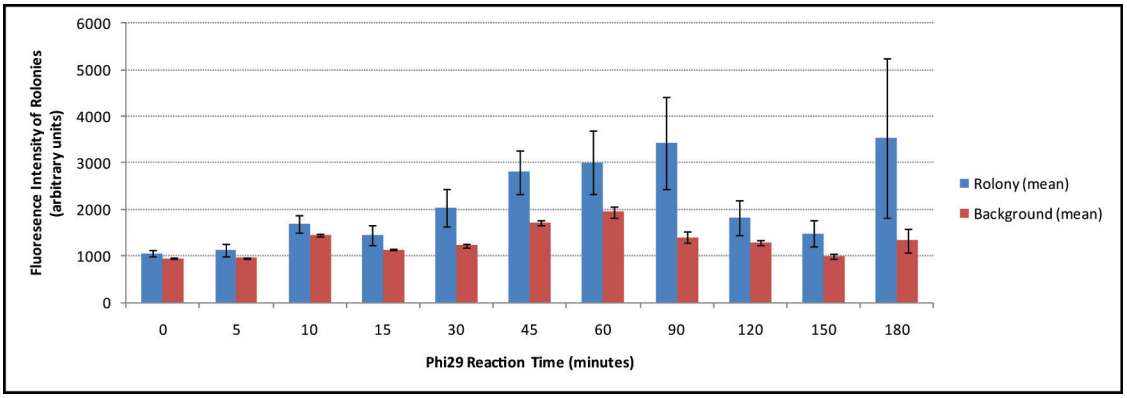


Figure 6-5. Fluorescence Intensity as a Function of Polymerization Time. Aliquots from a RCA reaction were removed at the noted timepoints and the enzyme was heat inactivated. The concatemers were then bound with a fluorescently labeled complementary primer, and bound to the surface of a slide. The intensity of rolonies was calculated by creating a binary mask marking each rolony and then calculating the average pixel intensity. Since the rolonies comprised a very small percentage of all pixels, the background was estimated as the average of all pixels. As the reaction proceeds, the growing error bars point to some concatemers increasing in size in a linear fashion, with others remaining short; this is likely explained by the kinetics of Phi29 polymerase.

To test whether the sequencing by ligation (SBL), the current enzymatic sequencing technique utilized by the Polonator, was amenable to rolonies, we created a

circular synthetic template with the universal sequencing primers and two variable regions: a single base that varied between A/T and an eight base fully degenerate region. Assuming complete circularization of the template, the rolonies were amplified via RCA with amino-modified primers at a concentration of 15E12 molecules/ml. 1% of this mixture was then bound to an epoxysilane modified slide, and SBL was performed for the first minus position with A/T being the possible bases and the fully degenerate minus 5 position. Figure 6-6A and 6-6B show the success of this reaction, where all rolonies are shown in the first image, and only half the rolonies represented by A/T are shown in the second. While the morphology of the rolonies made creation of a mask and automated analysis difficult, rolonies created by the Nilsson lab[17] utilizing poly-lysine chemistry yielded rolonies with a more compact morphology. While regions from some of our slides did exhibit a compact morphology and densities exceeding 35,000 features/mm^2, we had difficulty achieving this in a consistent manner. The mask generated by PISA is shown in Figure 6-6D and SBL was performed for the first minus position. The tetrahedron generated from the analyzed calls demonstrates the possibility of using rolonies for next-generation sequencing, and we anticipate a significant cost reduction with its application to Polonator SGS.

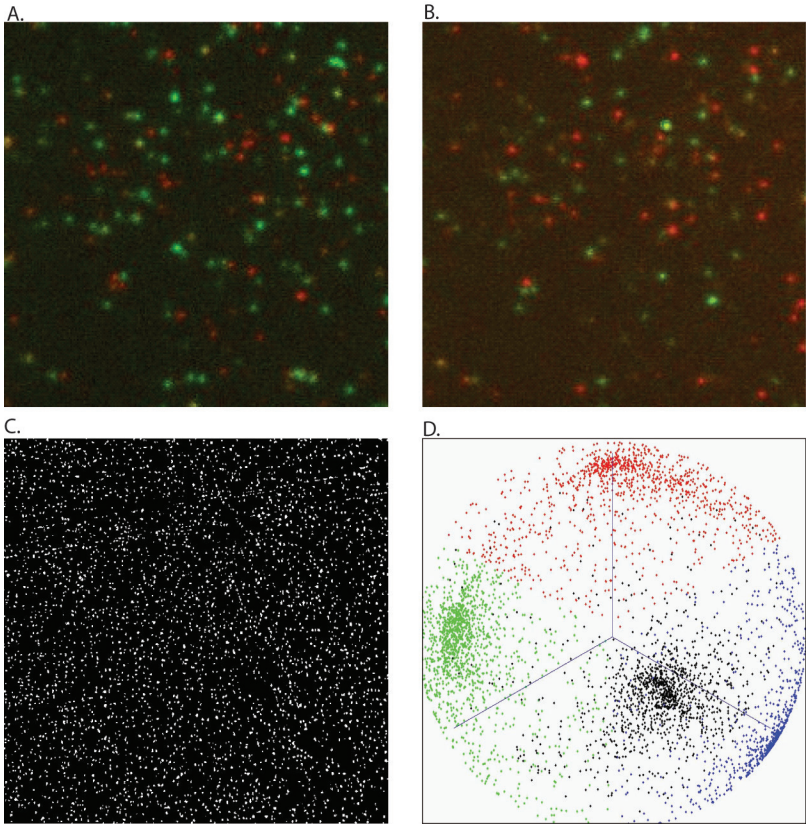


Figure 6-6. Rolony Sequencing by Ligation A. SBL Position Minus 1 B. SBL Position Minus 5. Rolonies from synthetic templates were generated and surface bound as described. The minus one position had two possible bases (A/T), while the fifth position was fully degenerate. Shown for both positions are A and T, detected by Cy3 (false-colored green) and Cy5 (false-colored red), respectively. **C. PISA Mask for Rolonies.** The rolonies in the above image proved too irregularly shaped for the PISA algorithm to define a binary mask (see methods). To demonstrate the feasibility of this algorithm in ideal situations, we applied it to rolony slides generated by collaborator's in the Nilsson lab featuring over 15,000 features/mm^2. **D. Tetrahedron for Position Minus 1.** With the mask generated in C we analyzed the output of SBL minus 1 from a separate experiment done in collaboration with the Nilsson Laboratory. All four nucleotides were present and we found that the rolonies separated nicely into four groups when analyzed with the Polonator software.

The successful prototyping of RCA based colonies (rolonies) demonstrate their feasibility as a replacement for emulsion PCR amplified beads. Additional experiments are still required to increase the signal intensity and the density of rolonies to make them a part of the Polonator platform and dramatically reduce the cost of DNA sequencing.

Methods

Generation of Circular Templates

To phosphorylate the five synthetic oligos (T1, T2, T3, T4, and T5) five reactions were assembled with 10 pmol of each oligo in 10 µL of 1x T4 DNA Ligase Buffer (NEB) with 10 U of T4 Polynucleotide Kinase (NEB). The reactions were incubated for 30m at 37°C and the enzyme was heat inactivated by heating for 20m at 65°C. To circularize the oligos, 40 µL of a mixture containing 20 pmol T-Guiding oligo, 1.25x AmpLigase Buffer (Epicentre) and 5U of AmpLigase (Epicentre) was added to each reaction. The reactions were then cycled (2m at 94°C, 10m at 42°C) three times on a MJ Research DNA Engine PTC-200. 20U of Exonuclease I (NEB) and 100U of Exonuclease III (NEB) were then added to each reaction to degrade all non-circularized material; the reaction was incubated 30m at 37°C, followed by 20m at 80°C to heat inactivate the enzymes. The five oligos were then phenol-chloroform extracted according to standard protocols, and the circlular DNA templates were resuspended in 50 µL of TE, each. These circles are referred to as cT#, and the stock concentration was calculated as 200nM.

Circle to Circle Amplification was then used to generate the complement of these circles. In brief, for each circlular template a 25 µL reaction containing 10 µL of each circle, 2pmol of T-Guiding Oligo, 20U of Phi29 Polymerase (Epicentre), 25nM each dNTP (Epicentre) and 1x Phi29 Buffer was assembled. The reactions were incubated for 70h at 30°C. The concatemers were digested in a 50 µL reaction containing 20 µL from the previous step, 2nmol of RO+ primer, 0.8x NEB2 buffer (NEB) and 100U MspI (NEB). The reaction was incubated for 2h at 37°C, and the enzyme was heat inactivated for 20m at 65°C. To form the circle complementary to the above template, a 100 µL

reaction was assembled containing 45 µL from the previous step and 10U AmpLigase (Epicentre) in 1x AmpLigase Buffer (Epicentre). The reaction was cycled as above for 10 cycles followed by an overnight incubation at 55°C. The samples were phenol-chloroform extracted and ethanol precipitated, and then resuspended in 50 µL of TE. Uncircularized DNA was digested with 40U Exonuclease I (NEB) and 50U of Exonuclease III (NEB) through a 3h incubation at 37°C, followed by heat inactivation of the enzyme. The samples were phenol-chloroform extracted and ethanol precipitated, and then resuspended in 50 µL of TE. These circles are referred to as cT#'. The stock concentration was calculated at 1µM.

Phi29-mediated Rolling Circle Amplification

A 50 µL reaction was assembled containing 2pmol RO+ primer for cT#' or RO- primer for cT#, 1pmol template, 0.5mM dNTP (Epicentre) and 20U Phi29 in 1x RepliPhi Buffer (Epicentre). The reaction was incubated 30m at 37°C and then heat inactivated for 20m at 65°C.

Slide Surface Activation

The protocol for activating the slides is adapted from Bulyk et al.[13]

Glass slides (Gold Seal) were cleaned for 0.5 to 2 hrs in 2 N nitric acid. After rinsing in distilled water, the slides were soaked in distilled water for 5 to 15 minutes, and then washed once with acetone. The slides were silanized by immersing them for 15 minutes in a solution of 1% 3 (aminopropyl)-triethoxysilane (Sigma-Aldrich) dissolved in 95% acetone. After washing the slides twice in acetone, the slides were baked for >30 minutes at 70°C. The surface of the slides was then activated by placing the slides in a solution of 0.5% 1,4-diphenylene-diisothiocyanate (PDC) (Fluka) dissolved in a solution consisting of 40 ml pyridine and 360 ml anhydrous N,N-dimethylformamide for 2 to 4

hrs. The slides were then washed twice with methanol, twice with acetone, and stored in a dessicator until use. The rest of the PDC surface was inactivated by a 10 min incubation in 1% ammonium hydroxide/0.1% SDS/200 mM NaCl. After washing in 4xSSC, the slides were neutralized in 6xSSPE/0.01% Triton X-100, washed twice in 4xSSC, then washed in 2xSSC and spun dry in a clinical centrifuge. Slides were stored in a closed box at room temperature until use.

While the original protocol uses a 60m incubation of 500pmol amino-labeled primers in 300mM K_2HPO_4, pH 9.0 at 37°C, we have found that 15m of coupling followed by 3m of blocking is sufficient.

The following protocol is adapted from X. Huang.

A 50 µL solution was assembled consisting of 500pmol Amino Fisseq-F in 300mM K_2HPO_4 was incubated for 15m on an activated slide, and the slide was inactivated as above. A 50 µL solution containing 2.5 µL of a Phi29 RCA reaction utilizing cT1and cT2 primed by RO- (as above), 1.2x SSPE and 0.04% Triton X-100 was added to the slide and spread by placing a coverslip on top of the solution. The slide was incubated for 60m at 55°C in a humidity chamber and then washed. The slide was coated with a 50 µL solution containing 0.8x Thermopol Buffer (NEB), 0.5 µM dNTP, 1x BSA (NEB), 3.85 µg SSB, and 40U Bst Polymerase (NEB). The slide was incubated for 30m at 55°C. It was then washed with 2x SSC/0.05% Triton X-100 then 2x SSC/0.1%SDS followed by a wash with 4x SSC and 2x SSC. 1 pmol of Fisseq-R FITC was then hybridized to the slide in 6x SSPE/0.1% Triton X-100 for 5m at 55°C, and the slide was washed as above. A single-base extension (SBE) was performed using Cy3 labeled dATP and Cy5 labeled dCTP. In brief, a 50 µL solution containing 25U of Klenow (NEB) and

25pmol of each labeled dinucleotide was placed and spread onto the slide in 1x Klenow Buffer. The reaction proceeded for 2m at room temperature after which the slide was washed, dried with compressed air and 20 μL of SlowFade (Invitrogen) was used to cover the slide. As an alternative to the above protocol, cT4' and cT5' were hybridized directly to the surface bound primers; switching to a 12-carbon linker (Amino-Fisseq-F Long) also improved the binding density.

Biotin-SE

The surface treatment part of this method was adapted from Yildiz et al[11].

The streptavidin-coated coverslip was made as follows: the coverslip was first cleaned by sonicating in 1M KOH solution and dried under nitrogen gas. 50 μL of 1 mg/ml BSA-biotin (A-8549 Sigma) in T50 buffer (10 mM Tris pH 8.0, 50 mM NaCl) was added to the coverslip via a laboratory-built flow-chamber, allowed to sit for ten minutes, and then washed with 100 μL of T50 buffer. 50 μL of 0.2 mg/ml Streptavidin (S-888 Molecular Probes) in T50 buffer was then added, allowed to sit for ten minutes, and then washed with 100 μL T50 buffer.

To generate hyper-branched biotin-containing RCA products, a 25 μL reaction containing 1 pmol each of cT4' and cT5', 1 pmol of **bio-Fisseq-F, 1mM dNTP (Epicentre) and 50U Phi29 Polymerase (Epicentre) in 1x Repliphi Buffer (Epicentre) was assembled. The reaction was incubated for 55m at 30°C. 800pmol of Fisseq-R, 800pmol T-filler 1, 2400 pmol of RO+ primer were then added together with Cy3-T4 long/T4 Filler or Cy5 T5 Long/T5 Filler. The reaction temperature was maintained for an additional 55m after which the polymerase was heat inactivated for 20m at 65°C. 10 μL of this solution was diluted in 190 μL of 1mg/ml BSA, and 100 μL was flowed over a

modified flowcell. The image was obtained with a 20x objective and 5s exposure, and the CY3 false-colored green and Cy5 false-colored red.

Free-floating linear rolonies bound statically to the surface

A 50 µL reaction was assembled containing: 5 fmol of each cT4' and cT5', 8 fmol Amino-Fisseq-F, 0.25mM dNTP, 50U Phi29 in 1x Repliphi Buffer. The reaction was incubated for 105m at 30°C and heat inactivated for 10m at 65°C. 5 µL of this reaction was in 45 µL of 6x SSPE/0.1% Triton containing 1pmol each Cy3-T4, Cy5-T5 and Fisseq-FITC-R. 5 µL of this solution was imaged on a coverslip.

Fiber Stretching

This was performed using the protocol outlined by Kraus et al [23], but we found it easier to slowly dry end-tethered rolonies in a heated humidity chamber. The experiment was repeated with 105m RCA product and this slide was used for the images in the text.

To replicate the experiments from Kraus et al, the DNA was coupled for 5 minutes to activated glass covered with plastic, and the glass was lifted while generating a meniscus. The slide was then heat dried and incubated at 37°C with 3% BSA in 2x SSC for 30 minutes. The slide was washed with 2x SSC and 70% and 100% ethanol, air-dried and hybridized 0.03 pmol of fluores per slide. The exposure time for all images was 250ms.

Synthetic Template for Sequencing

To generate a new template enabling sequencing by ligation (SBL) I ordered the following four oligomers from IDT: T' Bridge 1, T6' Bridge 2, PT' Part 1, T8' Part 2. These oligomers, when combined form the following synthetic oligo:

CCA||CTACGCCTCCGC||TTTCCTCTCTATGGGCAGT|||CGGTGATAGAGTGGTGG

A||[A/T]NNNNNNNNG||ATGGCAGAGAATGAGGAAC|||CCGGGGCAG The || marks the junction between the two sequences and the ||| marks the junction between the two bridging oligomers.

A 25 µL solution containing 25pmol of each of these oligos was made in 1x AmpLigase Buffer, and 2.5U of AmpLigase was added. The solution was incubated overnight at 55°C. 10U of Exonuclease I and 50U of Exonuclease III were added to digest uncircularized material and then the enzymes were heat-inactivated as above.

This template was used in a Phi29-RCA reaction (with amine-labeled primer) using 2.5pmol of template in the 50m reaction. Five µL of material was added to a mixture with a final concentration of 0.8M NaCl in Potassium Bicarbonate Buffer with 0.01% Triton X-100 and 4 µL of this solution was spotted onto an epoxy-activated slide.

Epoxy-activation of the slide is achieved through incubation of the cleaned slide with 2% 3-glycidoxytrimethoxysilane in acetone (v:v) for at least 30 minutes at room temperature followed by a wash with acetone.

SBL Protocol

Any hybridized probes were denatured with 50 µL 0.1 M NaOH wash for ten minutes at RT. The slides were then washed 2x in Wash 1E (10 mM Tris (pH 7.5), 50 mM KCl, 2 mM EDTA (pH 8.0), 0.01% Triton X-100). To hybridize the anchor primer, 50 µL 6X SSPE with 0.01% Triton-X-100 containing 0.5 µL 1 mM Anchor Primer was used. For minus positions, the LigFix DD anchor primer was used. Hybridization was performed for five minutes at 56°C followed by two minutes at 42°C. Subsequently, the slides were washed 2x in Wash 1E. A 50 µL ligation reaction was assembled containing 40pmol appropriate nonamer mix, and 100U High Concentration DNA Ligase in 1x T4

DNA Ligase Buffer (NEB). The sample was incubated for 30m at 35°C and then wash 2x in Wash 1E prior to imaging.

Oligos (Synthesized by Integrated DNA Technologies (IDT))

T1: CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGATAGAG TGGTGGAGTGTGTGTGTGTGTGAGAGAATGAGGAACCCGGGGCAG

T2: CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGATAGAGT GGTGGACACACACACACACACAGAGAATGAGGAACCCGGGGCAG

T3: CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGATAGAGT GGTGGATCACGTGTGTGAGCACTAGAGAATGAGGAACCCGGGGCAG

T4: CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGATAGAGTG GTGGATCGGTCGTTCGGCTGAGAGAATGAGGAACCCGGGGCAG

T5: CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGATAGAGTG GTGGACGACAGCTCTCACATAGAGAATGAGGAACCCGGGGCAG

T-Guiding oligo: GGAGGCGTAGTGGCTGCCCCGGGTTC

RO+ primer: ATGAGGAACCCGGGGC*A*G

RO- primer: CTGCCCCGGGTTCCTC*A*T

Amino Fisseq-F: /5AmMC6/AACCACTACGCCTCCGCTTTCC*T*C

Fisseq-R FITC: /56Org488XN/CTGCCCCGGGTTCCTCATTCTCT

Amino Fisseq-F Long: /5AmMC12/CCACTACGCCTCCGCTTTCCTCTCTATGGG

****bio-Fisseq-F**: /5Bio/CCACTACGCCTCCGCTTTCCTCTCTAT GGGCAGTCGGTG*A*T

Fisseq-R: CTGCCCCGGGTTCCTCATTCTCT

T-filler 1: CTCTATCACCGACTG

Cy3-T4 long: /5Cy3/CAGCCGAACGACCGATCCACCA

T4 Filler: CAGCCGAACGACCGATCCACCA

Cy5 T5 Long: /5Cy5/ATGTGAGAGCTGTCGTCCACCA

T5 Filler: ATGTGAGAGCTGTCGTCCACCA

T' Bridge 1: TTTCCTCTCTATGGGCAGTCGGTGATAGAGTGGTGGA

T6' Bridge 2: ATGGCAGAGAATGAGGAACCCGGGGCAGCCA

PT' Part 1: /5Phos/ACTGCCCATAGAGAGGAAAGCGGAGGCGTAGTGG

CTGCCCCGG

T8' Part 2: /5Phos/GTTCCTCATTCTCTGCCATCNNNNNNNNWTCCACCACT

CTATCACCG

LigFix DD: /5Phos/ATCACCGACTGCCCA

References

1 Diehl, F. *et al.* BEAMing: single-molecule PCR on microparticles in water-in-oil emulsions. *Nat Methods* **3**, 551-559, doi:nmeth898 [pii]
10.1038/nmeth898 (2006).

2 Dressman, D., Yan, H., Traverso, G., Kinzler, K. W. & Vogelstein, B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci U S A* **100**, 8817-8822, doi:10.1073/pnas.1133470100
1133470100 [pii] (2003).

3 Li, M., Diehl, F., Dressman, D., Vogelstein, B. & Kinzler, K. W. BEAMing up for detection and quantification of rare sequence variants. *Nat Methods* **3**, 95-97, doi:nmeth850 [pii]
10.1038/nmeth850 (2006).

4 Kim, J. B. *et al.* Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science* **316**, 1481-1484, doi:316/5830/1481 [pii]
10.1126/science.1137325 (2007).

5 Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728-1732, doi:1117389 [pii]
10.1126/science.1117389 (2005).

6 McKernan, K. J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two base encoding. *Genome Res*, doi:gr.091868.109 [pii]
10.1101/gr.091868.109 (2009).

7 Baner, J., Nilsson, M., Mendel-Hartvig, M. & Landegren, U. Signal amplification of padlock probes by rolling circle replication. *Nucleic Acids Res* **26**, 5073-5078, doi:gkb813 [pii] (1998).

8 Lizardi, P. M. *et al.* Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nat Genet* **19**, 225-232, doi:10.1038/898 (1998).

9 Melin, J. *et al.* Thermoplastic microfluidic platform for single-molecule detection, cell culture, and actuation. *Anal Chem* **77**, 7122-7130, doi:10.1021/ac050916u (2005).

10 Bensimon, A. *et al.* Alignment and sensitive detection of DNA by a moving interface. *Science* **265**, 2096-2098 (1994).

11 Yildiz, A. *et al.* Myosin V walks hand-over-hand: single fluorophore imaging with 1.5-nm localization. *Science* **300**, 2061-2065, doi:10.1126/science.1084398
1084398 [pii] (2003).

12 Ha, T. *et al.* Initiation and re-initiation of DNA unwinding by the Escherichia coli Rep helicase. *Nature* **419**, 638-641, doi:10.1038/nature01083
nature01083 [pii] (2002).

13 Bulyk, M. L., Huang, X., Choo, Y. & Church, G. M. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc Natl Acad Sci U S A* **98**, 7158-7163, doi:10.1073/pnas.111163698
111163698 [pii] (2001).

14 Dinu, C. Z. *et al.* Parallel manipulation of bifunctional DNA molecules on structured surfaces using kinesin-driven microtubules. *Small* **2**, 1090-1098, doi:10.1002/smll.200600112 (2006).

15 Steel, A. B., Levicky, R. L., Herne, T. M. & Tarlov, M. J. Immobilization of nucleic acids at solid surfaces: effect of oligonucleotide length on layer assembly. *Biophys J* **79**, 975-981, doi:S0006-3495(00)76351-X [pii]
10.1016/S0006-3495(00)76351-X (2000).

16 Nie, B., Shortreed, M. R. & Smith, L. M. Quantitative detection of individual cleaved DNA molecules on surfaces using gold nanoparticles and scanning electron microscope imaging. *Anal Chem* **78**, 1528-1534, doi:10.1021/ac052067g (2006).

17 Goransson, J. *et al.* A single molecule array for digital targeted molecular analyses. *Nucleic Acids Res* **37**, e7, doi:gkn921 [pii]
10.1093/nar/gkn921 (2009).

18 Adamcik, J., Klinov, D. V., Witz, G., Sekatskii, S. K. & Dietler, G. Observation of single-stranded DNA on mica and highly oriented pyrolytic graphite by atomic force microscopy. *FEBS Lett* **580**, 5671-5675, doi:S0014-5793(06)01101-X [pii]
10.1016/j.febslet.2006.09.017 (2006).

19 Mills, J. B., Vacano, E. & Hagerman, P. J. Flexibility of single-stranded DNA: use of gapped duplex helices to determine the persistence lengths of poly(dT) and poly(dA). *J Mol Biol* **285**, 245-257, doi:10.1006/jmbi.1998.2287
S0022-2836(98)92287-2 [pii] (1999).

20 Pant, K., Karpel, R. L., Rouzina, I. & Williams, M. C. Mechanical measurement of single-molecule binding rates: kinetics of DNA helix-destabilization by T4 gene 32 protein. *J Mol Biol* **336**, 851-870 (2004).

21 Zhang, K. *et al.* Long-range polony haplotyping of individual human chromosome molecules. *Nat Genet* **38**, 382-387, doi:ng1741 [pii]
10.1038/ng1741 (2006).

22 Blanco, L. *et al.* Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *J Biol Chem* **264**, 8935-8940 (1989).

23 Kraus, J. *et al.* High-resolution comparative hybridization to combed DNA fibers. *Hum Genet* **99**, 374-380 (1997).

Appendix C

Nanogrids for Creating Self-Ordered Arrays of Library Molecules for Second Generation Sequencing

Abraham M. Rosenbaum[1]*, Brian Y. Chow[2]*, Jaebum Joo[2], Gregory J. Porreca[1], Craig R. Forest[1], Dae H. Kim[1], Francois Vigneault[1], Rich Terry[1], Joseph M. Jacobson[2], George M. Church[1]

*These authors contributed equally

Author Contributions Surface patterns were designed by G.M.C., A.M.R. and B.Y.C. with help from C.R.F., J.J. and R.T.; Surface chemistry was designed by B.Y.C. and performed with help from A.M.R., D.H.K., and R.T.; Experiments on e-beam substrates were performed by A.M.R. with help from B.Y.C. and G.J.P.; Experiments on Lincoln Lab substrates were performed by A.M.R. and C.R.F. with help from B.Y.C. and R.T., and experiments on 2-beam interference substrates were performed by A.M.R. and D.H.K. with help from B.Y.C., R.T. and F.V.; G.M.C. and J.M.J. supervised all aspects of the study.

Acknowledgements We are grateful for the help and advice provided by all the members of the Church Laboratory and the Molecular Machines Group, as well as helpful discussions with Rade Drmanac (Complete Genomics). We are also grateful for the manufacturing help and advice received from Craig Keast, Jeffrey Knecht and Bruce Wheeler (Lincoln Laboratories, MIT) and Michael Skvarla (Cornell Nanofabrication Facility). We thank NHGRI for funding support.

Abstract

A number of second generation sequencing (SGS) platforms perform sequencing reactions on library molecules randomly arrayed on the surface of a flowcell. Ordered arrays have the potential for increasing the density of library molecules, thereby increasing the leverage of sequencing reagents and decreasing cost. Additionally, when detecting fluorescent signal with a microscope mounted CCD, having fewer pixels dedicated to each molecule decreases the scan time per feature thereby increasing the throughput. Here we report on advances towards ordered arrays of both emulsion amplified beads and RCA generated concatemer colonies ("rolonies") to be used in conjunction with the open-source Polonator platform. We demonstrate the potential of the process using grids patterned with a range of lithography techniques and chemistries and discuss potential pitfalls that should be avoided. We then provide direction for the successful application of this process in SGS.

Introduction

The first published Second Generation Sequencing (SGS) method utilized emulsion PCR (ePCR) amplified 28 micron beads in a fiber-optic microarray consisting of 44 micron wells[1]. The second such platform opted instead to use one micron emulsion PCR beads in an unordered array[2]. Fiber-optic arrays[3-4] and nanofabricated DNA binding racks for single molecule studies[5-7] both exhibit the potential density and spot size to enable an array significantly smaller than currently in use for SGS. Additionally, this technology has the ability to optimize the utilization of the CCD camera, as when each pixel or 2X2 pixel array is used to detect an additional molecule the sequencing can be viewed as being fully optimized from a detection vantage point. While a fiber-optic array

is amenable to bead deposition in high densities, the limitations of bead-polonies generated by ePCR led us to pursue nanofabricated grids (nanogrids) of DNA binding spots that are capable to binding both ePCR beads and RCA-colonies (rolonies).

Results

To enable self-assembly of ordered arrays of rolonies to both maximize their packing and facilitate their identification in downstream image analysis, we made arrays of activated spots to bind the rolonies and passivated the inter-spot regions. This passivation was needed to prevent the rolonies from indiscriminately binding to the non-activated regions. Different dimensions, binding chemistries and passivation chemistries were sampled to maximize the binding only one rolony per spot and prevent their binding outside the activated spots. We call these patterned substrates nanogrids. We made prototypes with e-beam and UV photolithography, and 2-beam interference lithography (Figure 7-1). While the e-beam allowed us greater flexibility in terms of one-off prototypes, enabling spots as small as 100nm (Figure 7-2), it cannot be used to generate large numbers of grids at an affordable price. We were, however, able to prototype many of the combinations of spot size and pitch and the different substrates that would be available upon the transition to a more high-throughput method. Furthermore, we were able to reject certain combinations either due to small pitch (Figure 7-2D), likelihood of resist lifting off (Figure 7-2E), or conversely, the likelihood for the resist to become irreversibly bound to the substrate (Figure 7-2F). For the UV photolithography we utilized both a deep-UV stepper with 248nm output to generate 200 and 350nm spots on a 0.7 and 1.1 micron pitch and a UV stepper with 365nm output for 700nm spots on a 2 micron pitch. For the two-beam lithography we used a custom setup.

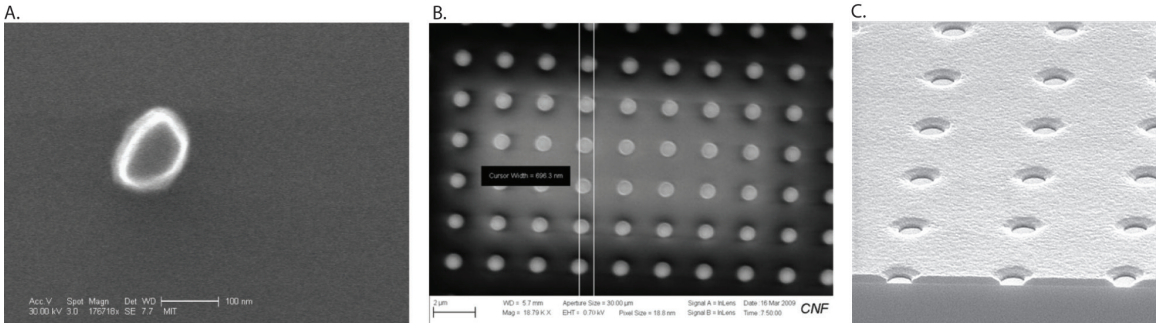


Figure 7-1. Surface Modifications for Rolony Self-Assembly. A. E-Beam. A 100nm spot created with an e-beam using a negative resist. **B. UV Photolithography of a Nanogrid.** Grids of DNA binding spots were made with 365nm UV Photolithography. The spot size was approximately 600nm with a 2 micron pitch. **C. 2-Beam Laser Interference Lithography.** Spots are 300nm with 1.5 micron spacing.

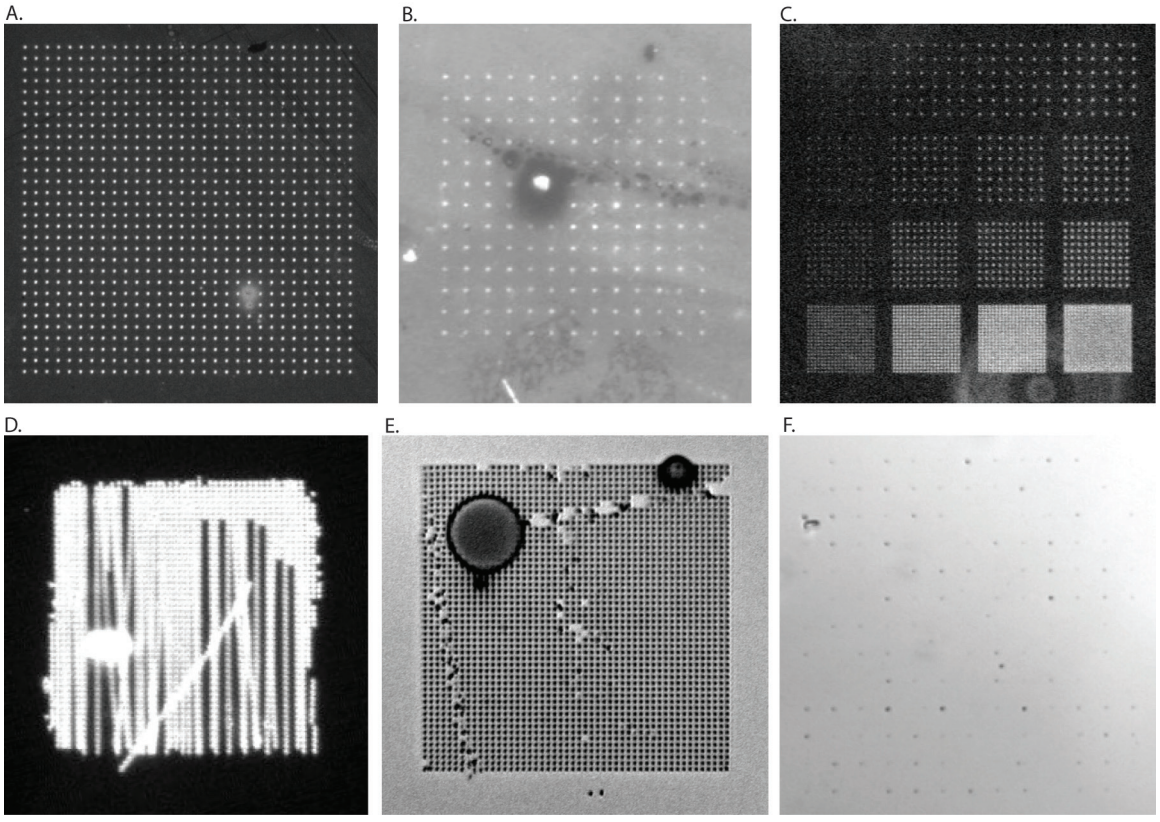


Figure 7-2. Versatility and Limitations of Pattern Size and Treatment. A. 1 Micron Spots with ~3.6 Micron Pitch on Si. B. 100nm Spots with 2.4 Micron Pitch on ITO. C. 200, 300, 400, 500 nm Spots with 1.2, 2.4, 3.6 and 4.8 Micron Pitch. D. 500nm Spots with 1.2 Micron Pitch. This combination proved difficult to pattern, and many of the spots fused together, generating lines. **E. 100nm Spots with 1.2 Micron Pitch.** In this early run many of the spots lifted, revealing the underlying Si. **F. 200nm Spots with 7.2 Micron Pitch.** The resist on this substrate was irreversibly fused to the Si. Acetone and utrasonic cleaning were insufficient to remove the resist. All images were taken with a 40x objective (NA 0.9) on an inverted epifluorescence microscope. In images A, B, and C the spots were silanized, activated and coupled to amino-labeled fluorescently labeled probes with NHS-chemistry. Figures D, E, and F were taken with a phase-contrast objective.

To activate the spots we used both NHS-ester based chemistries (PDC, glycidoxyl and BS3) and biotin-streptavidin interactions. For surfaces that were passivated, we used either perfluorination or PEG-ylation. While the passivation process was very effective for repelling short molecules and creating differences in hydrophobicity, in our hands it appeared to relatively ineffective in directing rolonies towards activated spots (Figure 7-3).

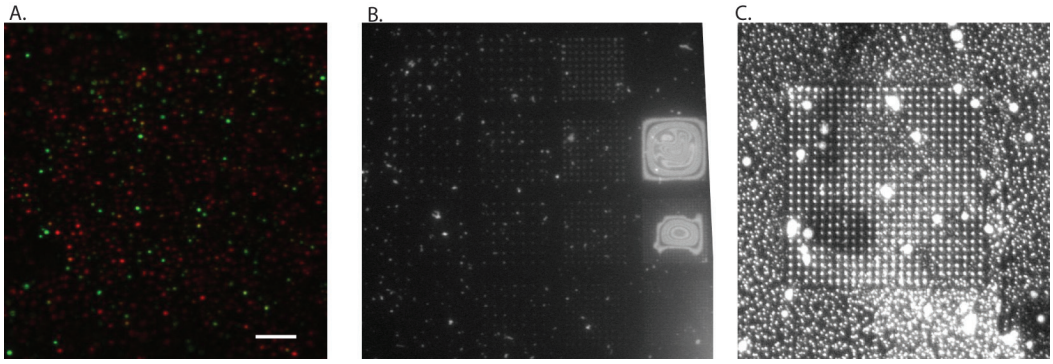


Figure 7-3. Passivation Difficulties. A. Rolonies Densely Bound to Passivated Region of 200X1200nm grid. B. Rolonies Bound to non-patterned Portion of Combinatorial Grid. Here the passivation is clearly affecting the hydrophobicity of the substrate, as apparent by the liquid being confined to the gridded area. Nevertheless, the rolonies are able to bind irrespective of the activated portions. **C. Indiscriminate Binding of Fluorescent Probes Due to Expired Resist.** In this image, the photoresist used to pattern the grids had expired which adversely affected the passivation and silanization chemistries. Alternatively, it is possible that blocking with a 5% BSA solution was ineffective.

Nevertheless, we were able to generate patterned grids with each spot containing a single rolony with a number of different chemistries. Similar to the results on unpatterned surfaces (see Appendix B), extending primers to replicated hybridized concatemers created an inseparable mixture of molecules, and was not amenable to sequencing. Coupling of hyper-branch RCA generated biotinylated concatemers to streptavidin containing surfaces mimicked the results from unpatterned substrates and led to distinct binding of rolonies to individual spots. Additionally, we found that hybridization of concatemers to the surface bound primers without extension from those primers led to distinct and definable molecules on an unpatterened surface. Upon

replicating this procedure on a patterned surface, the same result was demonstrated. This process, however, is not amenable to sequencing because of the denaturation steps (Figure 7-4). Further analysis of this image demonstrates the clonal capture of concatemers by each spot (Figure 7-5).

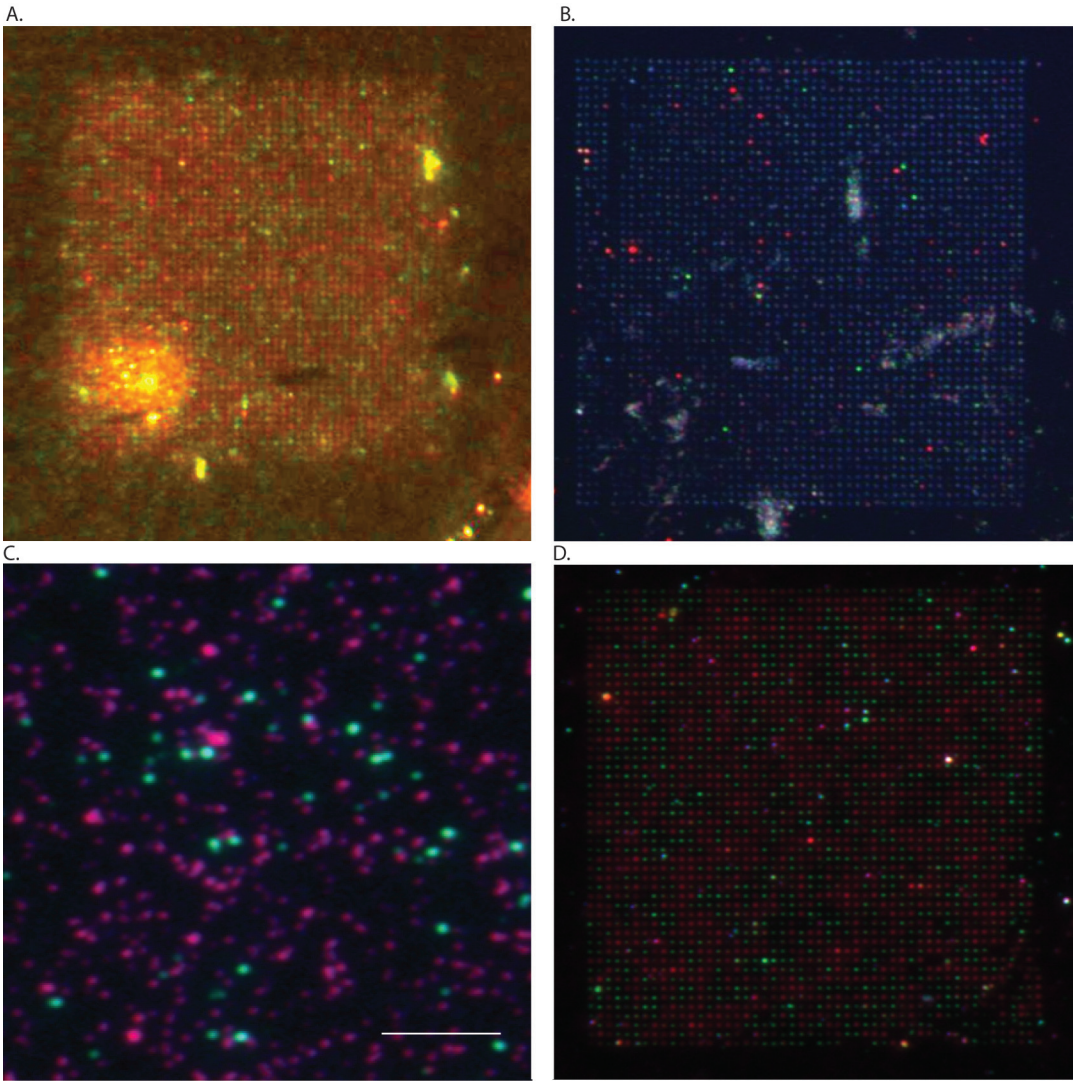


Figure 7-4. Rolonies Bound to E-beam Grids. A. Surface-bound Primers Extended off Hybridized Concatemers on 100X1200nm Grid. With this protocol very few grids contained only one library molecule. **B. Hyper-Branched Biotinylated RCA Molecules Bound to 200X1200nm Streptavidin Activated Spots.** While the molecules did preferentially bind the grids, it is not clear what role the residual resist (fluorescing in the FITC channel, false-colored blue) played in this coupling reaction. **C. Concatemers Hybridized to Surface-bound Primers.** Unexpectedly, while the extension of surface primers resulted in overlapping molecules, the hybridization without extension yielded distinct molecules when the primers were on an unpatterned surface. The measure-bar is 10 microns in length. **D. Concatemers Hybridized to Surface-bound Primers on a 200X1200nm PDC-activated Grid.** Utilizing the same protocol as in C yielded a clonal array of library molecules.

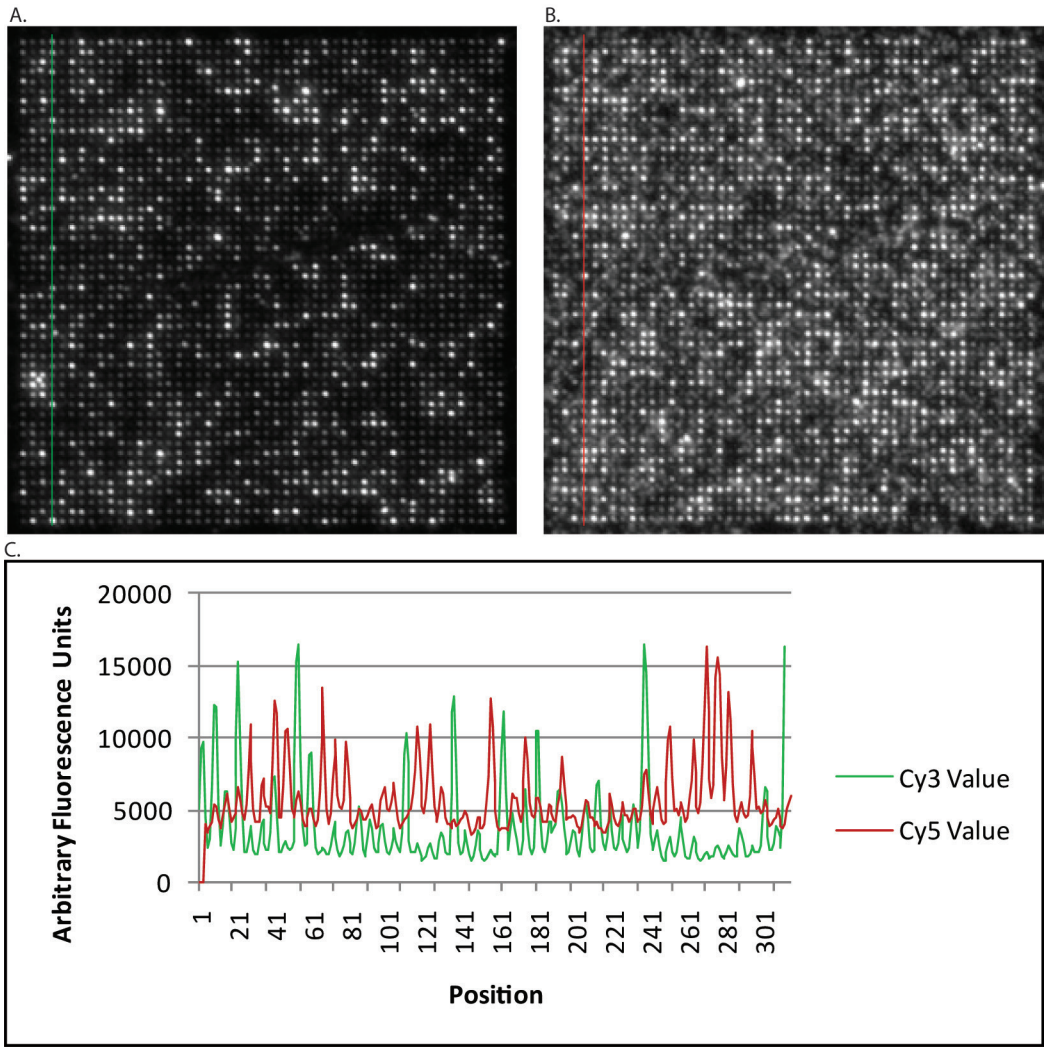


Figure 7-5. Clonality Analysis of Concatemer Hybridization. A. Image of Cy3 Signal. B. Image of Cy5 Signal. C. Line-Width Analysis of Figure 7-4D. The fourth column of spots was analyzed, as marked in A and B. The background in B is approximately 4x that of A due to the weaker Cy5 signal in this image. Clonality is clearly demonstrated insofar as the signal from both channels is never increased in simultaneously in a given spot.

After these successes we transitioned to photolithography to enable larger patterned surfaces. We processed one batch of deep-UV patterned wafers, made with 248nm light containing four groups of patterns mixing 200 and 350nm spots and 0.7 and 1.2 micron spacing. To enable a second set of experiments on these wafers, we had some etched to a depth of 300nm using reactive ion etching (RIE). This would allow us to deposit beads in a regular array[8-9]. While this did enable self assembly of both 1 micron and 510nm beads when loaded through capillary action (Figure 7-6), it had the additional

benefit of providing superior activation by thoroughly descumming the resist from the bare silicon. This process, however, also made the resist irremovable to solvents benign enough to not destroy the silanization. With our second batch of UV lithography we were restricted to a 365nm wavelength lithography and opted for 600nm spots and 2 micron spacing. We included in our process a 30s RIE to descum the patterned areas. While this is standard procedure and is effective in many other scenarios, we found that the descumming was insufficient for our wafers (Figure 7-7). The initial attempt to bind rolonies directly to the activated grids with NHS-ester chemistry appeared to show clonality despite the large amount of debris, later analysis on more dilute samples demonstrated that each rolony was potentially stretching over many activated spots depending upon the flow of the rolony solution and whether the solution was allowed to dry. When this was controlled for with the 600nm size grids generated by both UV photolithography and 2-beam intereference lithography, the rolonies maintained their compactness and remained isolated to only one spot. The size of these spots, however, allowed for more than one rolony to occupy each spot (Figure 7-8).

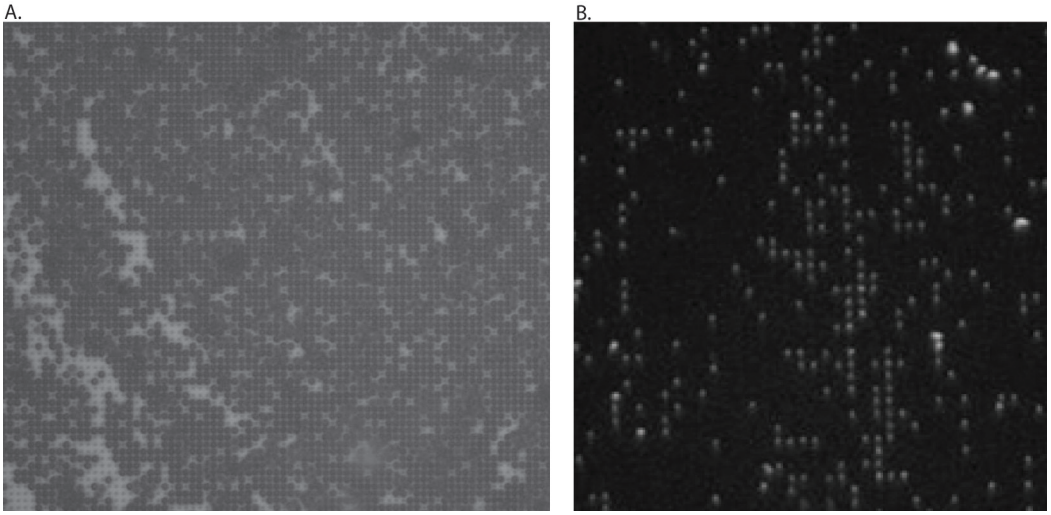


Figure 7-6. Ordered Bead Arrays. A. Ordered Arrays of 1 Micron Beads on a 500X1,100nm Grid. B. Ordered Array of 510nm Beads on 500X1,100nm Grid. A coverslip was placed on the grids and capillary action was used to fill the area, depositing beads in the holes. The 1 micron beads were imaged with white light, and the 510nm beads were bound to fluorescent primers and imaged using the epifluorescent settings.

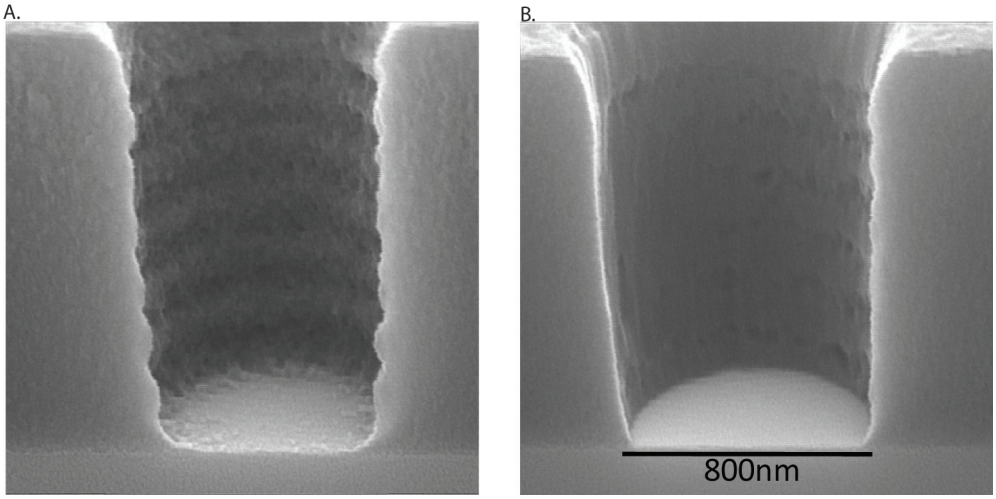


Figure 7-7. Residual Resist after Descumming. A. After Descumming. B. After Additional Deep RIE. EM analysis of the grids from CNF showed that despite descumming, significant residue remained in the spots. The thickness of the resist is also apparent as a barrier to rolony binding. An additional round of RIE served to remove most of the residue, but also increased the size of the spots from 600nm to 800nm.

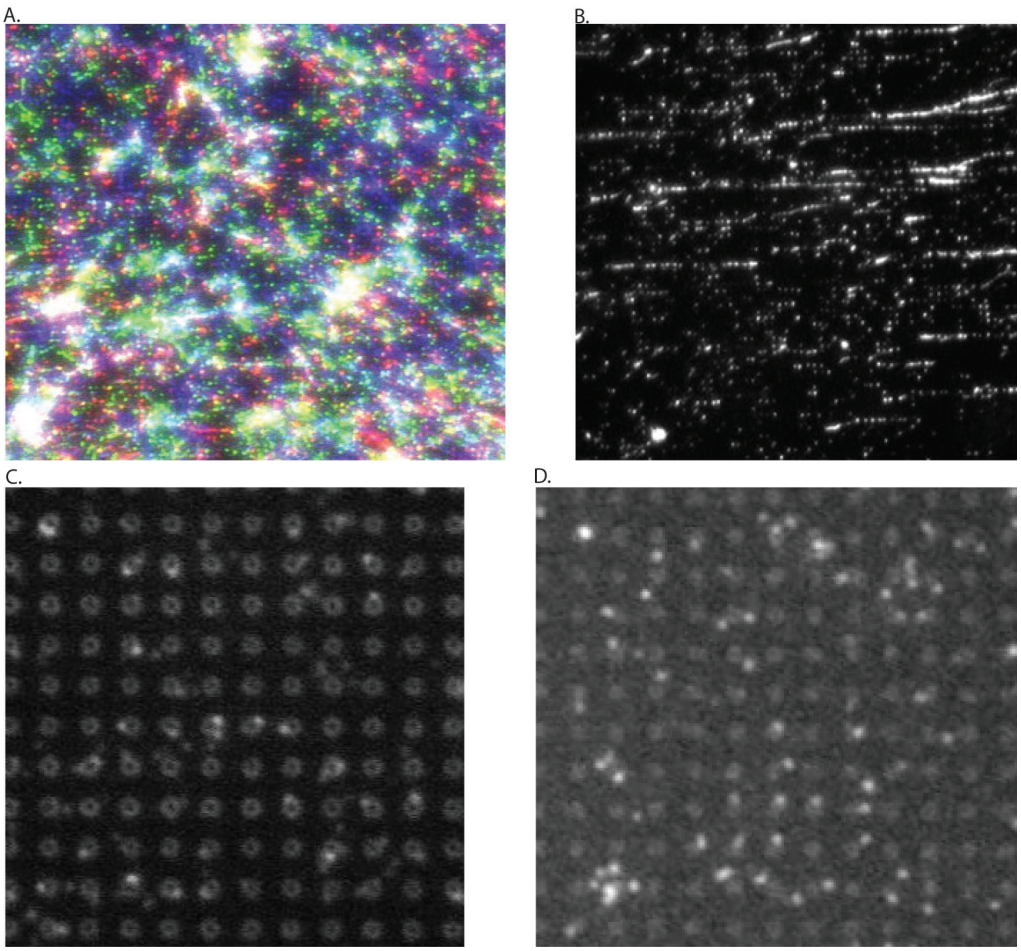


Figure 7-8. Rolonies on UV and 2-beam Interference Lithography Generated Nanogrids. A. Lincoln Laboratory Grids. Significant binding was partially obscured by a large buildup of concatemers not bound to the surface. **B. 200nmX1.1 micron unetched spots from Lincoln Labs.** The rolonies extended, but the comparative hydrophilicity of the activated spots caused the rolonies to "relax" in the spots, thereby allowing greater hybridization to fluorescent probes. **C. CNF 600nmX2 micron grids. D. 600nmX2 micron 2-beam Interference grids.** In both C and D the activated spots are clearly visible as well as the rolonies hybridized with fluorescently labeled probes. The binding of the rolonies to the activated spots and the presence of 2 or more rolonies in each of these spots is apparent in a number of places.

Further development of this technology using 2-beam lithography with 300nm spot sizes should lead to a maximum of one rolony per spot and a higher percentage of occupied spots. When combined with other SGS advances on the Polonator platform, this will enable an additional order of magnitude decrease in the cost of DNA sequencing.

Methods

Surface Chemistries Adapted for Nanogrids

All chemistries were performed as described in Appendix B. When passivation was required, the perfluorination or PEG-ylation was performed by our collaborator, Brian Y. Chow, at MIT. To silanize the substrates, the above protocol was followed using water as the solvent. Following the silanization, the resist was stripped with acetone, optionally in an ultrasonic bath, for 30m. All downstream processes were performed as in Appendix B.

References

1 Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380, doi:nature03959 [pii]
10.1038/nature03959 (2005).

2 Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728-1732, doi:1117389 [pii]
10.1126/science.1117389 (2005).

3 Walt, D. R. Fibre Optic Microarrays. *Chemical Society Reviews*, doi:10.1039/b809339n (2010).

4 Tam, J. M., Song, L. & Walt, D. R. DNA detection on ultrahigh-density optical fiber-based nanoarrays. *Biosens Bioelectron* **24**, 2488-2493, doi:S0956-5663(08)00691-X [pii]
10.1016/j.bios.2008.12.034 (2009).

5 Gorman, J., Fazio, T., Wang, F., Wind, S. & Greene, E. C. Nanofabricated Racks of Aligned and Anchored DNA Substrates for Single-Molecule Imaging. *Langmuir*, doi:10.1021/la902443e (2009).

6 Visnapuu, M. L., Fazio, T., Wind, S. & Greene, E. C. Parallel arrays of geometric nanowells for assembling curtains of DNA with controlled lateral dispersion. *Langmuir* **24**, 11293-11299, doi:10.1021/la8017634 (2008).

7 Dimalanta, E. T. *et al.* A microfluidic system for large DNA molecule arrays. *Anal Chem* **76**, 5293-5301, doi:10.1021/ac0496401 (2004).

8 Kraus, T. *et al.* Closing the Gap Between Self-Assembly and Microsystems Using Self-Assembly, Transfer, and Integration of Particles. *Advanced Materials* **17**, 2438-2442 (2005).

9 Xia, Y., Yin, Y., Lu, Y. & McLellan, J. Template-Assisted Self-Assembly of Spherical Colloids into Complex and Controllable Structures. *Advanced Functional Materials* **13**, 907-918 (2003).

Appendix D

Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome

This work was originally published as:

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D. & Church, G. M. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science,* 309**,** 1728-32.

Supplementary Online Material can be found at:
http://www.sciencemag.org/cgi/data/1117389/DC1/1

Author Contributions: Except for specific contributions listed, J.A.S. and G.J.P. were equally responsible for all of the sequencing biochemistry, software, hardware, molecular biology protocol development, and integration described in the chapter.

N.B.R. and X.L. engineered and evolved the *Escherichia coli* strain we resequenced, and N.B.R. performed some of the confirmatory Sanger sequencing of the mutations predicted. J.P.M. and R.D.M. contributed Supplementary Figures 1 and 6. N.B.R. and M.D.W. contributed significantly to the development of the *in vitro* shotgun genomic library protocol. A.M.R. contributed to the optimization of emulsion PCR conditions.

3. S. Georgi, www.batteriesdigest.com/id380.htm (accessed June 2005).
4. R. M. Alexander, *J. Exp. Biol.* **160**, 55 (1991).
5. G. A. Cavagna, N. C. Heglund, C. R. Taylor, *Am. J. Physiol.* **233**, R243 (1977).
6. J. Drake, *Wired* **9**, 90 (2001).
7. S. Stanford, R. Pelrine, R. Kornbluh, Q. Pei, in *Proceedings of the 13th International Symposium on Unmanned Untethered Submersible Technology* (Autonomous Undersea Systems Institute, Lee, NH, 2003).
8. T. Starner, J. Paradiso, in *Low Power Electronics Design* (CRC Press, Boca Raton, FL, 2004), p. 45–1.
9. G. A. Cavagna, M. Kaneko, *J. Physiol.* **268**, 647 (1977).
10. S. A. Gard, S. C. Miff, A. D. Kuo, *Hum. Mov. Sci.* **22**, 597 (2004).
11. Supporting material is available on *Science* Online.
12. Because it is a prototype, there has been no attempt to reduce the weight of the backpack—indeed, it is substantially "overdesigned." Further, the 5.6 kg includes the weight of six load cells and one 25-cm-long transducer, each with accompanying brackets and cables, as well as other components that will not be present on a typical pack. In future prototypes, we estimate that the weight will exceed that of a normal backpack by no more than 1 to 1.5 kg.
13. Under high-power conditions (5.6 km hour^{-1} with 20- and 29-kg loads and 4.8 km hour^{-1} with a 38-kg load), power generation on the incline was the same as on the flat. Under low-power conditions (4.8 km hour^{-1} with 20- and 28-kg loads), electricity generation on the incline was actually substantially greater than that on the flat (table S1).
14. R. Margaria, *Biomechanics and Energetics of Muscular Exercise* (Clarendon, Oxford, 1976).
15. R. A. Ferguson *et al.*, *J. Physiol.* **536**, 261 (2001).
16. G. A. Cavagna, P. A. Willems, M. A. Legramandi, N. C. Heglund, *J. Exp. Biol.* **205**, 3413 (2002).
17. A. Grabowski, C. T. Farley, R. Kram, *J. Appl. Physiol.* **98**, 579 (2005).
18. J. M. Donelan, R. Kram, A. D. Kuo, *J. Exp. Biol.* **205**, 3717 (2002).
19. J. M. Donelan, R. Kram, A. D. Kuo, *J. Biomech.* **35**, 117 (2002).
20. J. S. Gottschall, R. Kram, *J. Appl. Physiol.* **94**, 1766 (2003).
21. Because this savings in metabolic energy represents only 6% of the net energetic cost of walking with the backpack (492 W) (table S3) (*17*, *18*), accurate determinations of the position and movements of the center of mass, as well as the direction and magnitude of the ground reaction forces, are essential to discern the mechanism. This will require twin–force-platform single-leg measurements, as well as a complete kinematics and mechanical energy analysis (*19*, *20*). The energy analysis is made more complex because the position of the load with respect to the backpack frame and the amount of energy stored in the backpack springs vary during the gait cycle. Finally, electromyogram measurements are also important to test whether a change in effective muscle moment arms may have caused a change in the volume of activated muscle and hence a change in metabolic cost (*20*, *27*, *28*).
22. K. Schmidt-Nielsen, *Animal Physiology: Adaptation and Environment* (Cambridge Univ. Press, Cambridge, ed. 3, 1988).
23. This assumes that electronic devices are being powered in real time. If there were a power loss of 50% associated with storage (such as in batteries) and recovery of electrical energy, then these factors would be halved.
24. When not walking, the rack can be disengaged and the generator cranked by hand or by foot. Electrical powers of ~3 W are achievable by hand, and higher wattage can be achieved by using the leg to power it.
25. R. Kram, *J. Appl. Physiol.* **71**, 1119 (1991).
26. A. E. Minetti, *J. Exp. Biol.* **207**, 1265 (2004).
27. A. A. Biewener, C. T. Farley, T. J. Roberts, M. Temaner, *J. Appl. Physiol.* **97**, 2266 (2004).
28. T. M. Griffin, T. J. Roberts, R. Kram, *J. Appl. Physiol.* **95**, 172 (2003).

29. This work was supported by NIH grants AR46125 and AR38404. Some aspects of the project were supported by Office of Naval Research grant N000140310568 and a grant from the University of Pennsylvania Research Foundation. The authors thank Q. Zhang, H. Hofmann, W. Megill, and A. Dunham for helpful discussions; R. Sprague, E. Maxwell, R. Essner, L. Gazit, M. Yuhas, and J. Milligan for helping with the experimentation; and F. Letterio for machining the backpacks.

Supporting Online Material

www.sciencemag.org/cgi/content/full/309/5741/1725/DC1
Materials and Methods
SOM Text
Figs. S1 and S2
Tables S1 to S4
References

14 February 2005; accepted 25 July 2005
10.1126/science.1111063

Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome

Jay Shendure,[1]*† Gregory J. Porreca,[1]*† Nikos B. Reppas,[1] Xiaoxia Lin,[1] John P. McCutcheon,[2,3] Abraham M. Rosenbaum,[1] Michael D. Wang,[1] Kun Zhang,[1] Robi D. Mitra,[2] George M. Church[1]

We describe a DNA sequencing technology in which a commonly available, inexpensive epifluorescence microscope is converted to rapid nonelectrophoretic DNA sequencing automation. We apply this technology to resequence an evolved strain of *Escherichia coli* at less than one error per million consensus bases. A cell-free, mate-paired library provided single DNA molecules that were amplified in parallel to 1-micrometer beads by emulsion polymerase chain reaction. Millions of beads were immobilized in a polyacrylamide gel and subjected to automated cycles of sequencing by ligation and four-color imaging. Cost per base was roughly one-ninth as much as that of conventional sequencing. Our protocols were implemented with off-the-shelf instrumentation and reagents.

The ubiquity and longevity of Sanger sequencing (*1*) are remarkable. Analogous to semiconductors, measures of cost and production have followed exponential trends (*2*). High-throughput centers generate data at a speed of 20 raw bases per instrument-second and a cost of $1.00 per raw kilobase. Nonetheless, optimizations of electrophoretic methods may be reaching their limits. Meeting the challenge of the $1000 human genome requires a paradigm shift in our underlying approach to the DNA polymer (*3*).

Cyclic array methods, an attractive class of alternative technologies, are "multiplex" in that they leverage a single reagent volume to enzymatically manipulate thousands to millions of immobilized DNA features in parallel. Reads are built up over successive cycles of imaging-based data acquisition. Beyond this common thread, these technologies diversify in a panoply of ways: single-molecule versus multimolecule features, ordered versus disordered arrays, sequencing biochemistry, scale of miniaturization, etc. (*3*). Innovative proof-of-concept experiments have been reported, but are generally limited in terms of throughput, feature density, and library complexity (*4*–*9*). A range of practical and technical hurdles separate these test systems from competing with conventional sequencing on genomic-scale applications.

Our approach to developing a more mature alternative was guided by several considerations. (i) An integrated sequencing pipeline includes library construction, template amplification, and DNA sequencing. We therefore sought compatible protocols that multiplexed each step to an equivalent order of magnitude. (ii) As more genomes are sequenced de novo, demand will likely shift toward genomic resequencing; e.g., to look at variation between individuals. For resequencing, consensus accuracy increases in importance relative to read length because a read need only be long enough to correctly position it on a reference genome. However, a consensus accuracy of 99.99%, i.e., the Bermuda standard, would still result in hundreds of errors in a microbial genome and hundreds of thousands of errors in a mammalian genome. To avoid unacceptable numbers of false-positives, a consensus error rate of 1×10^{-6} is a more reasonable standard for which to aim. (iii) We sought to develop sequencing chemistries compatible with conventional epifluorescence imaging. Diffraction-limited optics with charge-coupled device detection achieves an excellent balance because it not only provides submicrometer resolution and high sensitivity for rapid data acquisition, but is also inexpensive and easily implemented.

[1]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. [2]Department of Genetics, [3]Howard Hughes Medical Institute, Washington University, St. Louis, MO 63110, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: shendure@alumni.princeton.edu (J.S.), gregory_porreca@student.hms.harvard.edu (G.J.P.)

Conventional shotgun libraries are constructed by cloning fragmented genomic DNA of a defined size range into an *Escherichia coli* vector. Sequencing reads derived from opposite ends of each fragment are termed "mate-pairs." To avoid bottlenecks imposed by *E. coli* transformation, we developed a multiplexed, cell-free library construction protocol. Our strategy (Fig. 1A) uses a type IIs restriction endonuclease to bring sequences separated on the genome by ~1 kb into proximity. Each ~135–base pair (bp) library molecule contains two mate-paired 17- to 18-bp tags of unique genomic sequence, flanked and separated by universal sequences that are complementary to amplification or sequencing primers used in subsequent steps. The in vitro protocol (Note S1) results in a library with a complexity of ~1 million unique, mate-paired species.

Conventionally, template amplification has been performed by bacterial colonies that must be individually picked. Polymerase colony, or polony, technologies perform multiplex amplification while maintaining spatial clustering of identical amplicons (*10*). These include in situ polonies (*11*), in situ rolling circle amplification (RCA) (*12*), bridge polymerase chain reaction (PCR) (*13*), picotiter PCR (*9*), and emulsion PCR (*14*). In emulsion PCR (ePCR), a water-in-oil emulsion permits millions of noninteracting amplifications within a milliliter-scale volume (*15*–*17*). Amplification products of individual compartments are captured via inclusion of 1-μm paramagnetic beads bearing one of the PCR primers (*14*). Any single bead bears thousands of single-stranded copies of the same PCR product, whereas different beads bear the products of different compartmentalized PCR reactions (Fig. 1B). The beads generated by ePCR have highly desirable characteristics: high signal density, geometric uniformity, strong feature separation, and a size that is small but still resolvable by inexpensive optics.

Provided that the template molecules are sufficiently short (fig. S1), an optimized version of the ePCR protocol described by Dressman *et al*. (*14*) robustly and reproducibly amplifies our complex libraries (Note S2). In practice, ePCR yields empty, clonal, and nonclonal beads, which arise from emulsion compartments that initially have zero, one, or multiple template molecules, respectively. Increasing template concentration in an ePCR reaction boosts the fraction of amplified beads at the cost of greater nonclonality (*14*). To generate populations in which a high fraction of beads was both amplified and clonal, we developed a hybridization-based in vitro enrichment method (Fig. 1C). The protocol is capable of a fivefold enrichment of amplified beads (Note S3).

Iterative interrogation of ePCR beads (Fig. 1D) requires immobilization in a format compatible with enzymatic manipulation and epifluorescence imaging. We found that a simple acrylamide-based gel system developed for in situ polonies (*6*) was easily applied to ePCR beads, resulting in a ~1.5-cm^2 array of disordered, monolayered, immobilized beads (Note S4, Fig. 2A).

With few exceptions (*18*), sequencing biochemistries rely on the discriminatory capacities of polymerases and ligases (*1*, *6*, *8*, *19*–*22*). We evaluated a variety of sequencing protocols in our system. A four-color sequencing by ligation scheme ("degenerate ligation") yielded the most promising results (Fig. 2, B and C). A detailed graphical description of this method is shown in fig. S7. We begin by hybridizing an "anchor primer" to one of four positions (immediately 5′ or 3′ to one of the two tags). We then perform an enzymatic ligation reaction of the anchor primer to a population of degenerate nonamers that are labeled with fluorescent dyes. At any given cycle, the population of nonamers that is used is structured such that the identity of one of its positions is correlated with the identity of the fluorophore attached to that nonamer. To the extent that the ligase discriminates for complementarity at that queried position, the fluorescent signal allows us to infer the identity of that base (Fig. 2, B and C). After performing the ligation and four-color imaging, the anchor primer:nonamer complexes are stripped and a new cycle is begun. With T4 DNA ligase, we can obtain accurate sequence when the query position is as far as six bases from the ligation junction while ligating in the 5′→3′ direction, and seven bases from the ligation junction in the 3′→5′ direction. This allows us to access 13 bp per tag (a hexamer and heptamer separated by a 4- to 5-bp gap) and 26 bp per amplicon (2 tags × 13 bp) (fig. S7).

Although the sequencing method presented here can be performed manually, we benefited from fully automating the procedure (fig. S3). Our integrated liquid-handling and microscopy setup can be replicated with off-the-shelf components at a cost of about $140,000. A detailed description of instrumentation and software is provided in Notes S5 and S7.

As a genomic-scale challenge, we sought a microbial genome that was expected, relative to a reference sequence, to contain a modest number of both expected and unexpected differences.

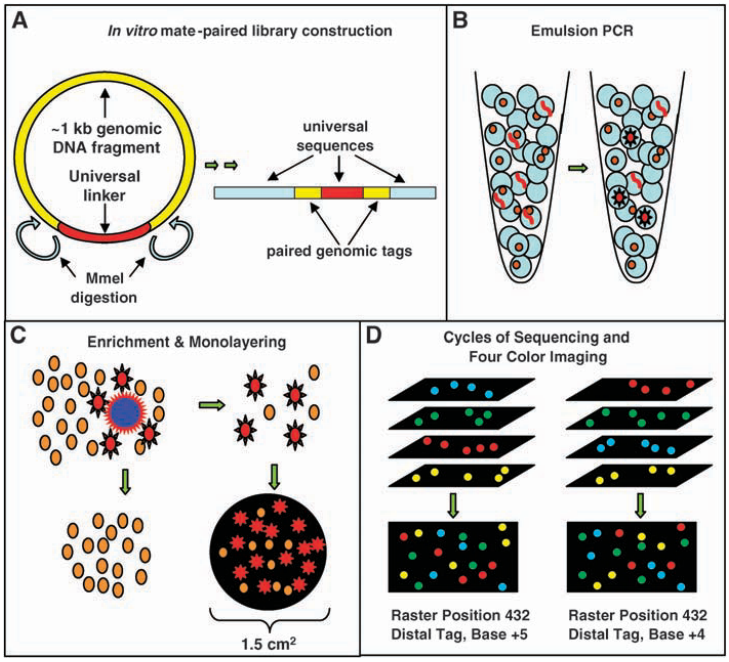


Fig. 1. A multiplex approach to genome sequencing. (**A**) Sheared, size-selected genomic fragments (yellow) are circularized with a linker (red) bearing Mme I recognition sites (Note S1). Subsequent steps, which include a rolling circle amplification, yield the 134- to 136-bp mate-paired library molecules shown at right. (**B**) ePCR (*14*) yields clonal template amplification on 1-μm beads (Note S2). (**C**) Hybridization to nonmagnetic, low-density "capture beads" (dark blue) permits enrichment of the amplified fraction (red) of magnetic ePCR beads by centrifugation (Note S3). Beads are immobilized and mounted in a flowcell for automated sequencing (Note S4). (**D**) At each sequencing cycle, four-color imaging is performed across several hundred raster positions to determine the sequence of each amplified bead at a specific position in one of the tags. The structure of each sequencing cycle is discussed in the text, Note S6, and fig. S7.

We selected a derivative of *E. coli* MG1655, engineered for deficiencies in tryptophan biosynthesis and evolved for ~200 generations under conditions of syntrophic symbiosis via coculture with a tyrosine biosynthesis–deficient strain (*23*). Specific phenotypes emerged during the laboratory evolution, leading to the expectation of genetic changes in addition to intentionally engineered differences.

An in vitro mate-paired library was constructed from genomic DNA derived from a single clone of the evolved Trp^{-}_{v} strain. To sequence this library, we performed successive instrument runs with progressively higher bead densities. In an experiment ultimately yielding 30.1 Mb of sequence, 26 cycles of sequencing were performed on an array containing amplified, enriched ePCR beads. At each cycle, data were acquired for four wavelengths at 20× optical magnification by rastering across each of 516 fields of view on the array (Fig. 1D). A detailed description of the structure of each sequencing cycle is provided in Note S6. In total, 54,696 images (14 bit, 1000 ×1000) were collected. Cycle times averaged 135 min per base (~90 min for reactions and ~45 min for imaging), for a total of ~60 hours per instrument run.

Image processing and base calling algorithms are detailed in Note S7. In brief, all images taken at a given raster position were aligned. Two additional image sets were acquired: brightfield images to robustly identify bead locations (Fig. 2A) and fluorescent primer images to identify amplified beads. Our algorithms detected 14 million objects within the set of brightfield images. On the basis of size, fluorescence, and overall signal coherence over the course of the sequencing run, we determined 1.6 million to be well-amplified, clonal beads (~11%). For each cycle, mean intensities for amplified beads were extracted and normalized to a 4D unit vector (Fig. 2, B and C). The Euclidean distance of the unit vector for a given raw base call to the median centroid of the nearest cluster serves as a natural metric of the quality of that call.

The reference genome consisted of the *E. coli* MG1655 genome (GenBank accession code U00096.2) appended with sequences corresponding to the *cat* gene and the lambda Red prophage, which had been engineered into the sequenced strain to replace the *trp* and *bio* operons, respectively. To systematically assess our power to detect single-base substitutions, we introduced a set of 100 random single-nucleotide changes into the reference sequence at randomly selected positions ("mock SNCs") (Table 1).

An algorithm was developed to place the discontinuous reads onto the reference sequence (Note S7). The matching criteria required the paired tags to be appropriately oriented and located within 700 to 1200 bp of one another, allowing for substitutions if exact matches were not found. Of the 1.6 million reads, we were able to confidently place ~1.16 million (~72%) to specific locations on the reference genome, resulting in ~30.1 million bases of resequencing data at a median raw accuracy of 99.7%. At this stage of the analysis, the data were combined with reads from a previous instrument run that contributed an additional ~18.1 million bases of equivalent quality (Fig. 2D). In this latter experiment, ~1.8 million reads were generated from ~7.6 million objects (~24%), of which ~0.8 million were confidently placed (~40%).

High-confidence consensus calls were determined for 70.5% of the *E. coli* genome for which sufficient and consistent coverage was available (3,289,465 bp; generally positions with ~4× or greater coverage). There were six positions within this set that did not agree with the reference sequence, and thus were targeted for confirmation by Sanger sequencing. All six were correct, although in one case we detected the edge of an 8-bp deletion rather than a substitution (Table 2). Three of these six mutations represent heterogeneities in lambda Red or MG1655, or errors in the

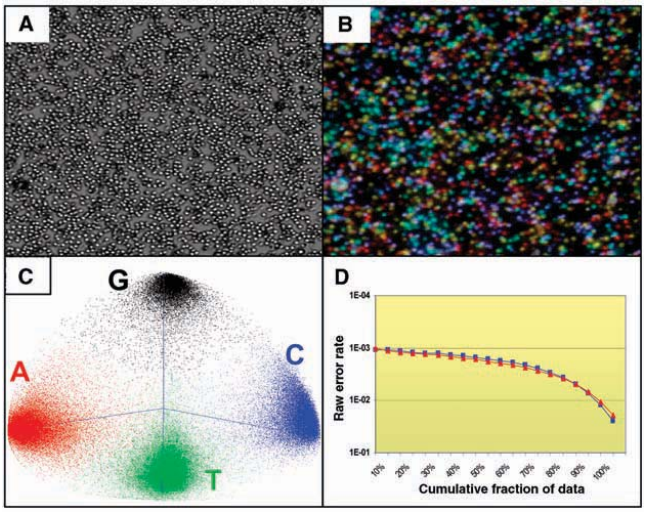


Fig. 2. Raw data acquisition and base calling. (**A**) Brightfield images (area shown corresponds to 0.01% of the total gel area) facilitate object segmentation by simple thresholding, allowing resolution even when multiple 1-μm beads are in contact. (**B**) False-color depiction of four fluorescence images acquired at this location from a single ligation cycle. A, gold; G, red; C, light blue; T, purple. (**C**) Four-color data from each cycle can be visualized in tetrahedral space, where each point represents a single bead, and the four clusters correspond to the four possible base calls. Shown is the sequencing data from position (−1) of the proximal tag of a complex *E. coli*–derived library. (**D**) Cumulative distribution of raw error as a function of rank-ordered quality for two independent experiments (red triangles, 18.1-Mb run; blue squares, 30.1-Mb run). The *x* axis indicates percentile bins of beads, sorted on the basis of a confidence metric. The *y* axis (logarithmic scale) indicates the raw base-calling accuracy of each cumulative bin. Equivalent Phred scores are Q20 = 1×10^{-2}, Q30 = 1×10^{-3} {Phred score = $-10[\log_{10}(\text{raw per-base error})]$}. Cumulative distribution of raw error with sequencing by ligation cycles considered independently is shown in fig. S8.

Table 1. Genome Coverage and SNC prediction. Bases with consistent consensus coverage were used to make mutation predictions. To assess power, the outcome of consensus calling for the mock SNC positions with various levels of coverage was determined. Data from two independent sets of mock SNCs are shown. "86 of 87," for example, means that 87 of the 100 mock SNCs were present in the sequence that was covered with 1× or more reads, and 86 of these were called correctly.

Coverage	Percent of genome	Correctly called mock substitutions
1× or greater	91.4%	86 of 87 88 of 90
2× or greater	83.3%	78 of 78 75 of 76
3× or greater	74.9%	67 of 67 68 of 68
4× or greater	66.9%	58 of 58 62 of 62

reference sequence; three were only present in the evolved variant (Table 2). Of the 100 mock SNCs, 53 were at positions called with high confidence. All of these were correctly called as substitutions of the expected nucleotide (59 of 59 on a second set of mock SNCs). The absence of substitution errors in ~3.3 Mb of reference sequence positions called with high confidence suggests that we are achieving consensus accuracies sufficient for resequencing applications. Percentage of the genome covered and mock SNC discovery at various levels of coverage are shown in Table 1.

Despite 10× coverage in terms of raw base pairs, only ~91.4% of the genome had at least 1× coverage (fig. S4). Substantial fluctuations in coverage were observed owing to the stochasticity of the RCA step of library construction. We are currently generating libraries that are more complex and more evenly distributed.

A Gaussoid distribution of distances between mate-paired tags was observed, consistent with the size selection during library construction (Fig. 3, A and B). Notably, the helical pitch of DNA (~10.6 bp per turn) is evident in the local statistics of ~1 million circularization events (Fig. 3B). As a function of the number of bases sequenced, we generated over an order of magnitude more mate-pairing data points than an equivalent amount of conventional sequencing. To detect genomic rearrangements, we mined the unplaced mate-pairs for consistent links between genomic regions that did not fall within the expected distance constraints. In addition to detecting the expected replacements of the *trp* and *bio* operons with *cat* and lambda Red prophage (Fig. 3D), we detected and confirmed the absence of a 776-bp IS1 transposon (Fig. 3C), a previously described heterogeneity in MG1655 strains (*24*). We also detected and confirmed a ~1.8-kb region that was heterogeneously inverted in the genomic DNA used to construct the library (Fig. 3E), owing to activity of *pin* on the invertible P region (*25*).

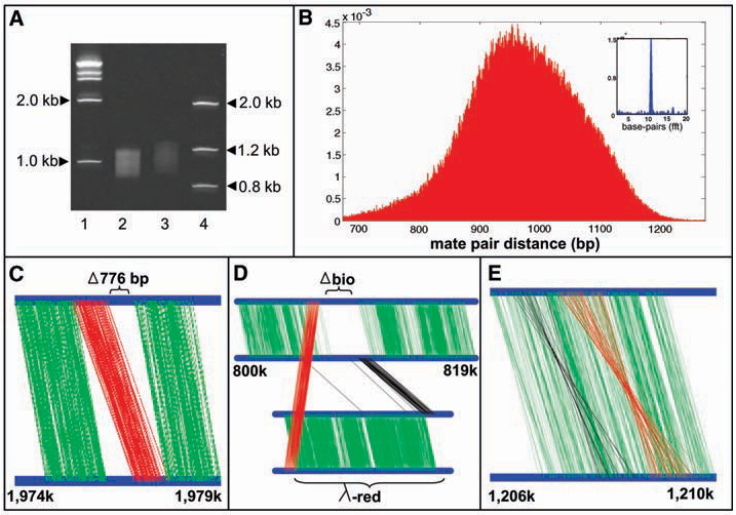


Fig. 3. Mate-paired tags and rearrangement discovery. (**A**) Diagnostic 6% polyacrylamide gel of the sheared, size-selected genomic DNA from which the library was constructed. Lanes 1 and 4 are molecular size markers. Lane 2 represents the material used in the library sequenced to generate the paired-tag mappings in (B), and lane 3 represents genomic DNA for a different library. (**B**) Histogram of distances between ~1 million mapped mate-pair sequences. The probability of circularization favors integrals of the helical pitch of DNA, such that the Fourier transform of the distribution (**inset**) yields a peak at 10.6 bp (*27*) (**C** to **E**). Consistent, aberrant mapping of unplaced mate-pairs to distal sequences revealed information about underlying rearrangements. Top and bottom blue bars indicate genomic positions for proximal and distal tags, respectively. Green connections indicate mate-pairings that fall within expected distance constraints, whereas red and black connections indicate aberrant connections (red indicates connections between the same strand, and black, connections between opposite strands). (C) Detection of a 776-bp deletion in the *flhD* promoter (*24*). (D) Detection of the replacement of the *bio* locus with the lambda red construct. (E) Detection of the P-region inversion (*25*). Detection of the inversion on a background of normally mate-paired reads indicates that the inversion is heterogeneously present.

Table 2. Polymorphism discovery. Predictions for mutated positions were tested and verified as correct by Sanger sequencing. We found three mutations unique to the evolved strain—two in *ompF*, a porin, and one in *lrp*, a global regulator.

Position	Type	Gene	Context	Confirmation	Comments
986,328	T → G	*ompF*	−10 region	Yes	Evolved strain only
931,955	8-bp deletion	*lrp*	Frameshift	Yes	Evolved strain only
985,791	T → G	*ompF*	Glu → Ala	Yes	Evolved strain only
1,976,527−1,977,302	776-bp deletion	*flhD*	Promoter	Yes	MG1655 heterogeneity
3,957,960	C → T	*ppiC*	5′ UTR	Yes	MG1655 heterogeneity
λ-red, 3274	T → C	*ORF61*	Lys → Gly	Yes	λ-red heterogeneity
λ-red, 9846	T → C	*cI*	Glu → Glu	Yes	λ-red heterogeneity

We observe error rates of ~0.001 for the better half of our raw base calls (Fig. 2D). Although high consensus accuracies are still achieved with relatively low coverage, our best raw accuracies are notably one to two orders of magnitude less accurate than most raw bases in a conventional Sanger sequencing trace. The PCR amplifications before sequencing are potentially introducing errors at a rate that imposes a ceiling on the accuracies achievable by the sequencing method itself. One potential solution is to create a library directly from the genomic material to be sequenced, such that the library molecules are linear RCA amplicons. Such concatemers, where each copy is independently derived from the original template, would theoretically provide a form of error correction during ePCR.

Our algorithms were focused on detection of point substitutions and rearrangements. Increasing read lengths, currently totaling only 26 bp per amplicon, will be critical to detecting a wider spectrum of mutation. A higher fidelity ligase (*20*) or sequential nonamer ligations (*20*, *21*) may enable completion of each 17- to 18-bp tag. Eco P15 I, which generates ~27-bp tags, would allow even longer read lengths while retaining the same mate-pairing scheme (*26*).

We estimate a cost of $0.11 per raw kilobase of sequence generated (Note S8), roughly one-ninth as much as the best costs for electrophoretic sequencing. Raw data in all sequencing methods are generally combined to form a consensus. Even though costs are generally defined in terms of raw bases, the critical metric to compare technologies is consensus accuracy for a given cost. There is thus a need to devise appropriate cost metrics for specific levels of consensus accuracy.

If library construction costs are not included, the estimated cost drops to $0.08 per raw kilobase. Higher densities of amplified beads are expected to boost the number of bases sequenced per experiment. While imaging, data were collected at a rate of ~400 bp/s. Although enzymatic steps slowed our overall throughput to ~140 bp/s, a dual flowcell instrument (such that the microscope is always imaging) will allow us to achieve continuous data acquisition. Enzymatic reagents, which dominate our cost equation, can be produced in-house at a fraction of the commercial price.

We demonstrate low costs of sequencing, mate-paired reads, high multiplicities, and high consensus accuracies. These enable applications including BAC (bacterial artificial chromosome) and bacterial genome resequencing, as well as SAGE (serial analysis of gene expression) tag and barcode sequencing. Simulations suggest that the current mate-paired libraries are compatible with human genome resequencing, provided that the read length can be increased to cover the full 17- to 18-bp tag (fig. S5).

What are the limits of this approach? As many as 1 billion 1-μm beads can potentially be fit in the area of a standard microscope slide (fig. S6). We achieve raw data acquisition rates of ~400 bp/s, more than an order of magnitude faster than conventional sequencing. From another point of view, we collected ~786 gigabits of image data from which we gleaned only ~60 megabits of sequence. This sparsity—one useful bit of information per 10,000 bits collected—is a ripe avenue for improvement. The natural limit of this direction is single-pixel sequencing, in which the commonplace analogy between bytes and bases will be at its most manifest.

References and Notes

1. F. Sanger *et al.*, *Nature* **265**, 687 (1977).
2. F. S. Collins, M. Morgan, A. Patrinos, *Science* **300**, 286 (2003).
3. J. Shendure *et al.*, *Nat. Rev. Genet.* **5**, 335 (2004).
4. I. Braslavsky, B. Hebert, E. Kartalov, S. R. Quake, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 3960 (2003).
5. T. S. Seo *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 5926 (2005).
6. R. D. Mitra, J. Shendure, J. Olejnik, O. Edyta Krzymanska, G. M. Church, *Anal. Biochem.* **320**, 55 (2003).
7. M. J. Levene *et al.*, *Science* **299**, 682 (2003).
8. M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlen, P. Nyren, *Anal. Biochem.* **242**, 84 (1996).
9. J. H. Leamon *et al.*, *Electrophoresis* **24**, 3769 (2003).
10. http://arep.med.harvard.edu/Polonator/Plone.htm
11. R. D. Mitra, G. M. Church, *Nucleic Acids Res.* **27**, e34 (1999).
12. P. M. Lizardi *et al.*, *Nat. Genet.* **19**, 225 (1998).
13. C. P. Adams, S. J. Kron, U.S. Patent 5,641,658 (1997).
14. D. Dressman, H. Yan, G. Traverso, K. W. Kinzler, B. Vogelstein, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 8817 (2003).
15. D. S. Tawfik, A. D. Griffiths, *Nat. Biotechnol.* **16**, 652 (1998).
16. F. J. Ghadessy, J. L. Ong, P. Holliger, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 4552 (2001).
17. M. Nakano *et al.*, *J. Biotechnol.* **102**, 117 (2003).
18. A. M. Maxam, W. Gilbert, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 560 (1977).
19. F. Barany, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 189 (1991).
20. J. N. Housby, E. M. Southern, *Nucleic Acids Res.* **26**, 4259 (1998).
21. S. C. Macevicz, U.S. Patent 5,750,341 (1998).
22. S. Brenner *et al.*, *Nat. Biotechnol.* **18**, 630 (2000).
23. N. B. Reppas, X. Lin, in preparation.
24. C. S. Barker, B. M. Pruss, P. Matsumura, *J. Bacteriol.* **186**, 7529 (2004).
25. R. H. Plastерk, P. van de Putte, *EMBO J.* **4**, 237 (1985).
26. M. Mucke, S. Reich, E. Moncke-Buchner, M. Reuter, D. H. Kruger, *J. Mol. Biol.* **312**, 687 (2001).
27. D. Shore, R. L. Baldwin, *J. Mol. Biol.* **170**, 957 (1983).

28. For advice, encouragement, and technical assistance, we are deeply indebted to J. Zhu, S. Douglas, J. Chou, J. Aach, M. Nikku, A. Lee, N. Novikov, and M. Wright (Church Lab); A. Blanchard, G. Costa, H. Ebling, J. Ichikawa, J. Malek, P. McEwan, K. McKernan, A. Sheridan, and D. Smith (Agencourt); S. Skiena (SUNY–Stony Brook) C. Felts (RPI); R. Fincher (Alcott); D. Focht (Bioptechs); and M. Hotfelder and J. Feng (Washington University). We thank B. Vogelstein, J. Edwards, and their groups for assistance with emulsion PCR. This work was supported by the National Human Genome Research Institute–Centers of Excellence in Genomic Science and U.S. Department of Energy–Genomes to Life grants.

Supporting Online Material
www.sciencemag.org/cgi/content/full/1117389/DC1
SOM Text
Figs. S1 to S8

14 July 2005; accepted 27 July 2005
Published online 4 August 2005;
10.1126/science.1117389
Include this information when citing this paper.

PUMA Couples the Nuclear and Cytoplasmic Proapoptotic Function of p53

Jerry E. Chipuk,[1]* Lisa Bouchier-Hayes,[1] Tomomi Kuwana,[1,2] Donald D. Newmeyer,[1] Douglas R. Green[1]*†

The *Trp53* tumor suppressor gene product (p53) functions in the nucleus to regulate proapoptotic genes, whereas cytoplasmic p53 directly activates proapoptotic Bcl-2 proteins to permeabilize mitochondria and initiate apoptosis. Here, we demonstrate that a tripartite nexus between Bcl-xL, cytoplasmic p53, and PUMA coordinates these distinct p53 functions. After genotoxic stress, Bcl-xL sequestered cytoplasmic p53. Nuclear p53 caused expression of *PUMA*, which then displaced p53 from Bcl-xL, allowing p53 to induce mitochondrial permeabilization. Mutant Bcl-xL that bound p53, but not PUMA, rendered cells resistant to p53-induced apoptosis irrespective of *PUMA* expression. Thus, PUMA couples the nuclear and cytoplasmic proapoptotic functions of p53.

The antineoplastic function of p53 occurs primarily through the induction of apoptosis (*1*). p53 undergoes posttranslational modification in response to oncogene-activated signaling pathways or to genotoxic stress; this allows stabilization of p53, which accumulates in the nucleus and regulates target gene expression. Numerous genes are regulated by p53, such as those encoding death receptors [for example, *FAS* (*CD95*)] and proapoptotic Bcl-2 proteins (for example, *BAX*, *BID*, *Noxa*, and *PUMA*) (*2*–*7*). In parallel, p53 also accumulates in the cytoplasm, where it directly activates the proapoptotic protein BAX to promote mitochondrial outer-membrane permeabilization (MOMP) (*8*–*10*). Once MOMP occurs, proapoptogenic factors (for example, cytochrome c) are released from mitochondria, caspases are activated, and apoptosis rapidly ensues (*11*). Thus, p53 possesses a proapoptotic function that is independent of its transcriptional activity (*12*–*15*).

If p53 directly engages MOMP in cooperation with BAX, no further requirement for p53-dependent transcriptional regulation of additional proapoptotic Bcl-2 proteins would be expected. Nevertheless, PUMA (p53–upregulated modifier of apoptosis), a proapoptotic BH3-only protein, is a direct transcriptional target of p53. Furthermore, mice deficient in *Puma* are resistant to p53-dependent, DNA damage–induced apoptosis even though p53 is stabilized and accumulates in the cytoplasm (*6*, *16*–*18*). A better understanding of the distinct nuclear and cytoplasmic proapoptotic functions of p53 may reveal strategies for the prevention and treatment of cancer.

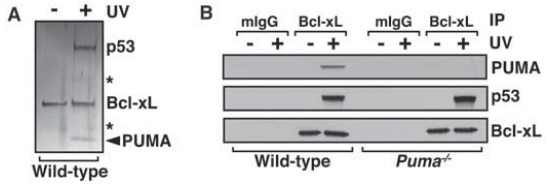


Fig. 1. DNA damage–induced p53·Bcl-xL and PUMA·Bcl-xL complexes. (**A**) Proteins from cytosolic extracts prepared from wild-type or *Puma*$^{-/-}$ MEFs treated with 5 mJ/cm^2 UV were immunoprecipitated with an agarose-conjugated antibody to Bcl-xL, eluted, subjected to SDS-PAGE, and visualized by silver staining. Bands were excised and subjected to tryptic digestion and mass spectrometry. The asterisk (*) indicates a fragment of Bcl-xL or p53. (**B**) Cytosolic extracts were treated as in (A), but protein complexes were analyzed by Western blot. mIgG (mouse immunoglobulin G) is a control antibody.

[1]Division of Cellular Immunology, La Jolla Institute for Allergy and Immunology, 10355 Science Center Drive, San Diego, CA 92121, USA. [2]University of Iowa, Carver College of Medicine, Department of Pathology, Iowa City, IA 52242, USA.

*Present address: Department of Immunology, St. Jude Children's Research Hospital, 332 North Lauderdale Street, Memphis, TN 38105, USA.
†To whom correspondence should be addressed. E-mail: dgreen5240@aol.com

Appendix E

Multiplex padlock targeted sequencing reveals human hypermutable CpG variations

This work was originally published as:

Li, J. B., Gao, Y., Aach, J., Zhang, K., Kryukov, G. V., Xie, B., Ahlford, A., Yoon, J. K., Rosenbaum, A. M., Zaranek, A. W., LeProust, E., Sunyaev, S. R. & Church, G. M. 2009. Multiplex padlock targeted sequencing reveals human hypermutable CpG variations. *Genome Res,* 19**,** 1606-15.

Supplementary Online Material can be found at:
http://genome.cshlp.org/content/suppl/2009/07/14/gr.092213.109.DC1/Supp.pdf

Author Contributions: A.M.R. provided helpful discussion regarding MIP capture optimization and library construction.

Methods

Multiplex padlock targeted sequencing reveals human hypermutable CpG variations

Jin Billy Li,[1,6,9] Yuan Gao,[2,6] John Aach,[1,6] Kun Zhang,[3,6] Gregory V. Kryukov,[4,6] Bin Xie,[2] Annika Ahlford,[1,7] Jung-Ki Yoon,[1,8] Abraham M. Rosenbaum,[1] Alexander Wait Zaranek,[1] Emily LeProust,[5] Shamil R. Sunyaev,[4] and George M. Church[1,9]

[1]*Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA;* [2]*Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, Virginia 23284, USA;* [3]*Department of Bioengineering, University of California, San Diego, California 92093, USA;* [4]*Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA;* [5]*Genomics Solution Unit, Agilent Technologies Inc., Santa Clara, California 95051, USA*

Utilizing the full power of next-generation sequencing often requires the ability to perform large-scale multiplex enrichment of many specific genomic loci in multiple samples. Several technologies have been recently developed but await substantial improvements. We report the 10,000-fold improvement of a previously developed padlock-based approach, and apply the assay to identifying genetic variations in hypermutable CpG regions across human chromosome 21. From ~3 million reads derived from a single Illumina Genome Analyzer lane, ~94% (~50,500) target sites can be observed with at least one read. The uniformity of coverage was also greatly improved; up to 93% and 57% of all targets fell within a 100- and 10-fold coverage range, respectively. Alleles at >400,000 target base positions were determined across six subjects and examined for single nucleotide polymorphisms (SNPs), and the concordance with independently obtained genotypes was 98.4%–100%. We detected >500 SNPs not currently in dbSNP, 362 of which were in targeted CpG locations. Transitions in CpG sites were at least 13.7 times more abundant than non-CpG transitions. Fractions of polymorphic CpG sites are lower in CpG-rich regions and show higher correlation with human–chimpanzee divergence within CpG versus non-CpG sites. This is consistent with the hypothesis that methylation rate heterogeneity along chromosomes contributes to mutation rate variation in humans. Our success suggests that targeted CpG resequencing is an efficient way to identify common and rare genetic variations. In addition, the significantly improved padlock capture technology can be readily applied to other projects that require multiplex sample preparation.

[Supplemental material is available online at http://www.genome.org. The sequence data from this study have been submitted to the NCBI Short Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi) under accession no. SRA007914.]

For decades, DNA sequencing has been pivotal in understanding biology, yielding over 900 whole-genome sequences, and identifying genetic variations and somatic mutations that underlie human diseases (Frazer et al. 2007; Stenson et al. 2008). Recent sequencing-based studies suggest that a large panel of genes is mutated in various cancers (Sjoblom et al. 2006; Jones et al. 2008; Parsons et al. 2008). Individually rare but cumulatively frequent variations contribute to the inheritance of common multifactorial diseases (Cohen et al. 2004, 2006; Bodmer and Bonilla 2008; Ji et al. 2008). Recently, "deep sequencing" has been enabled by "next-generation" technologies that reduce sequencing costs by several orders of magnitude (Shendure and Ji 2008). However, it is still prohibitively expensive to sequence whole human genomes, particularly when sample sizes are large. Thus, multiplexed targeted amplification of many genomic regions of interest is crucial for rapid and cost-effective sequencing-based research projects.

[6]**These authors contributed equally to this work.**
Present addresses: [7]Department of Medical Sciences, Uppsala University, S-751 85 Uppsala, Sweden; [8]College of Medicine, Seoul National University, Seoul 110-799, Korea.
[9]**Corresponding authors.**
E-mail http://arep.med.harvard.edu/gmc/email.html; fax (617) 432-6513.
E-mail jli@genetics.med.harvard.edu; fax (617) 432-6513.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.092213.109.

Parallel targeted amplification of selected genome regions is a challenging task (Garber 2008). Two different categories of methods have been developed to enrich or capture desired genomic regions, such as exons. One category employs hybridization of sheared genomic DNA to probes complementary to targeted regions. The probes can be oligonucleotides on a microarray surface (Albert et al. 2007; Hodges et al. 2007; Okou et al. 2007) or in solution (Gnirke et al. 2009). Although most of the desired regions are captured, the specificity of enriched genomic DNA tends to be limited due to "off target" and "near target" capture. In addition, due to the low efficiency of hybridization on the surface of a microarray, large amounts of genomic DNA are needed. The other category of methods requires hybridization in regions flanking both sides of the target and subsequent circularization of the targets. One way is to use "selector" oligonucleotides to guide the directed circularization of the target sequences digested with restriction enzymes (Dahl et al. 2005, 2007). Another method, which is independent of the presence of flanking restriction enzyme sites and thus is more flexible, applies padlock (molecular inversion) probes that anchor targeted regions and are circularized after polymerization and ligation (Hardenbol et al. 2003; Porreca et al. 2007). In our initial study, we targeted 55,000 exons but observed only ~10,000 unique sites in over 2 million end-sequencing reads, and most heterozygous loci were called incorrectly (Porreca et al.

19:1606–1615 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

2007), indicating the need for substantial improvement in capturing efficiency.

In the human genome, CpG dinucleotides are about fivefold less abundant than expected by chance (Sved and Bird 1990). This is due to the widespread methylation of cytosine in CpG and the deamination of 5-methylcytosine to thymidine (Wang et al. 1982); CpG is thus frequently mutated to TpG (or CpA on the complementary DNA strand). Overall, CpG elevates the mutation rate for transitions by 14- to 15-fold and for transversions by three- to fourfold (Kondrashov 2003; Hwang and Green 2004; Schmidt et al. 2008).

Mutations within the CpG context are a predominant cause of human diseases. At least one-third of mutations implicated in Mendelian diseases originated within CpG contexts (Cooper and Youssoufian 1988; Cooper and Krawczak 1993). Although CpG sites are depleted in the bulk of noncoding human DNA, they are selectively maintained in protein-coding genes and other functional genomic regions despite the elevated mutation rate (Subramanian and Kumar 2003; Kondrashov et al. 2006). Thus, the prevalence of CpG-induced mutations among disease mutations is much greater than among all mutations.

CpG context plays an important role in somatic mutations involved in human cancer. In the *TP53* gene, which is mutated in >50% of all human tumors, ~30% of all mutations occur at CpG dinucleotides, and all five major mutation hotspots are found at CpGs (Olivier et al. 2002). Recently, sequencing of nearly all protein-coding regions in cancer genomes revealed that 17%, 38%, 43%, and 48% of point mutations occur at CpGs in breast, pancreatic, brain, and colorectal cancers, respectively (Sjoblom et al. 2006; Jones et al. 2008; Parsons et al. 2008).

In this work, we describe improvements to our padlock capturing technology that yield an estimated 10,000-fold increase in capture efficiency over previous reports and significant improvements in sensitivity and uniformity. We designed 53,777 padlock probes to cover ~24% of all CpGs on human chromosome 21, where each probe captured a 40-bp region containing at least one CpG, and applied them to discover genetic variants in the genomic DNA from one HapMap CEPH individual (NA10835) and five volunteers from the Personal Genome Project (PGP; http://www.personalgenomes.org). We report the improved performance and high reproducibility of our optimized methods, and demonstrate the utility of the data for identification of known and novel SNPs and unbiased analysis of CpG variation rates.

Results

Site selection, probe design, synthesis, and processing

We designed 53,777 padlock probes to capture CpGs on human chromosome 21. The probes were designed to maximize the number of CpGs in a nonoverlapping manner. Each probe flanks a 40-bp gapped target containing at least one CpG, with the entire probe set covering 90,158 CpGs (24%) of 383,729 total on chromosome 21. The gap size of 40 bp was chosen to simplify sequencing library construction, as this size is comparable to the read length of most next-generation technologies. The two anchoring sequences adjacent to the 40-bp targets are termed the extension and ligation arms. The extension arm is the anchoring sequence from which polymerization initiates, and the ligation arm is the sequence to which the polymerized DNA is ligated. In the gap, one CpG (the *targeted* CpG) was located right next to the ligation junction. Successful circularization of padlock probes depends on the fidelity of both polymerase and ligase. The melting temperature (Tm) of extension arms is generally 5°C lower than the Tm of ligation arms (Supplemental Fig. 1). This design may stabilize the ligation arm and minimizes polymerase displacement of the ligation arm at the ligation junction (Akhras et al. 2007). We also took measures to ensure the uniqueness of the arms by avoiding repetitive regions (see Methods for details).

The padlock probes, flanked by amplifiable primers, were synthesized as 150-nucleotide (nt) oligonucleotides on and then released from a programmable microarray solid support. The padlock probe precursor oligonucleotides appeared as a single band of the desired size on a denaturing gel (Fig. 1A). To minimize PCR amplification bias while producing ample amounts of products, we performed two rounds of PCR with about 10 cycles each (see Methods). We sequenced the precursors before and after two rounds of amplification on an Illumina Genome Analyzer and found the number of reads per site was not very different from the Poisson distribution (Fig. 1B), suggesting that probe precursors were roughly evenly distributed both before and even after two rounds of PCR amplification. This result indicates that a small initial quantity of padlock precursors can be faithfully amplified to generate very large quantities of padlock probes, which greatly reduces the probe cost for analyzing large numbers of samples. In contrast to our previous method of using nicking enzymes whose recognition sites cannot be present on the capture arms (Porreca et al. 2007), we developed a novel approach that allows more flexibility in arm selection to generate single-stranded padlock probes (see Methods).

Improvement of padlock capturing

We previously reported an attempt to capture 55,000 exons in the human genome using padlock probes (Porreca et al. 2007). However, the capturing efficiency appeared to be very low—as a consequence, only ~20% of the targeted sites could be detected, and most of the heterozygous SNPs were incorrectly called as homozygous. After experimentation aimed at increasing the efficiency, we identified three factors that, together with better probe design and synthesis, yielded substantial improvements: (1) The amount of time allowed for hybridization reactions between genomic DNA and padlock probes must be extended; (2) increased amounts of reactants are required to ensure adequate generation of circles—especially, increased amounts of padlock probes are required for reactions that use only low amounts of genomic DNA (0.5–1 μg); and (3) dNTP concentrations must be carefully adjusted to possibly minimize the strand displacement activity of the polymerase.

We used a systematic approach to optimize padlock capture. To quantitate the capturing efficiency, we developed a simple SYBR Green-based real-time PCR assay to measure the number of circularized padlock probes. We define 100% capturing efficiency as the condition in which one padlock probe is circularized at each genomic copy of a target locus. When we hybridized 1 μg of human genomic DNA (0.5 amol or ~300,000 copies of haploid genomes) and 10× excess of probes at each target site at 60°C for up to 1 h, the capturing efficiency is ~0.0025%. This translates to an average of 7.5 circles formed per site (0.0025% × 300,000 = 7.5). When we used the previous 55k exon set (Porreca et al. 2007), the efficiency was only 0.0001% (an average of 0.3 circles per site). The 25-fold improvement in this work, even under the same reaction conditions, was clearly due to better probe design and synthesis that led to higher and more uniform hybridization efficiency. An

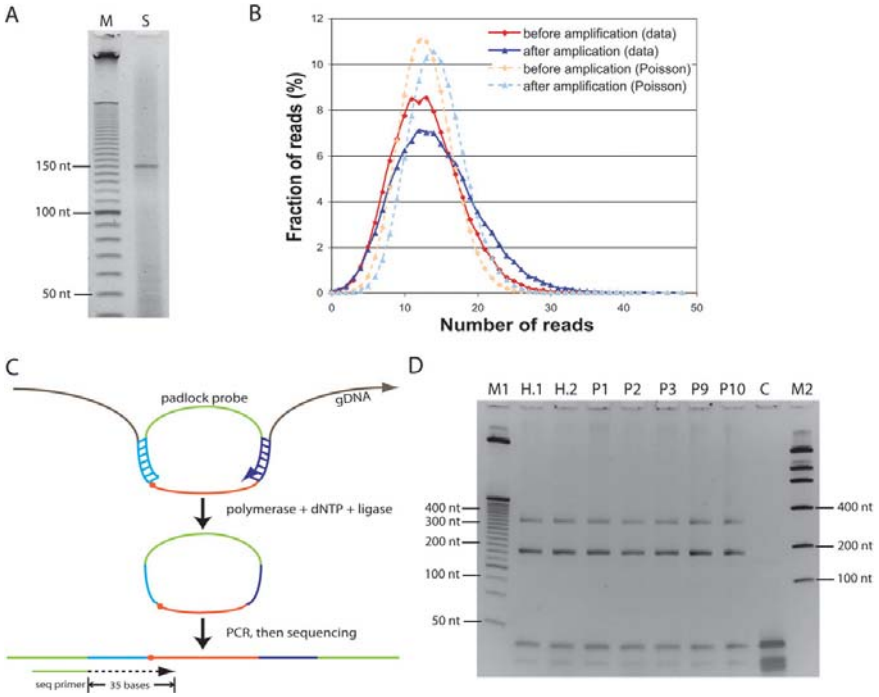


Figure 1. Padlock probe capture of 53,777 CpG sites. (*A*) The raw probe precursor (150-mer) sample from Agilent (*S*) was loaded along with a 10-bp ladder (M) on a 6% denaturing PAGE gel. (*B*) The probe precursors before and after two rounds of PCR amplification were end-sequenced by Illumina Genome Analyzer. (*C*) The padlock probes were hybridized to the targeted genomic CpG sites with a uniform 40-nt size. To simplify library construction, a target CpG (dot) was located immediately next to the ligation arm of the probe. Enzymatic filling and ligation of the gap (brown) allowed a copy of the target site to form a circle with the padlock probe. The circles were then PCR amplified using the backbone sequences (green) as primers. The common backbone sequence immediately upstream of the ligation arm served as a sequencing primer. (*D*) Amplification of circles derived from padlock probes. PCR products were loaded on a 6% PAGE gel. The two *upper* DNA bands had the expected amplicon sizes: 184 bp (subject to gel purification and Illumina sequencing) and 334 bp (if polymerization extended around the circle twice); the *lower* bands below 50 nt were derived from PCR primers. (Lane *M1*) 25-bp DNA ladder (Invitrogen); (lanes *H.1*,*H.2*) technical replicates of HapMap sample NA10835; (lanes *P1*,*P2*,*P3*,*P9*,*P10*) Personal Genomes 1, 2, 3, 9, and 10, respectively; (lane C) no genomic DNA control; (lane *M2*) low mass DNA ladder (Invitrogen).

additional ~10-fold increase was obtained by extending the reaction time from 1 h to at least 24 h (Fig. 2, left panel). We then explored further conditions with larger amounts of padlock probes. When we increased the amount of probes to 50×, 100×, and 250× excess, the capturing efficiency climbed to as high as 0.6% (an average of 1800 circles per site) and did not yet seem to saturate (Fig. 2, middle panel).

In addition to the complex hybridization reaction of the human genome and thousands of probes in a single tube, another challenge is to assure efficient circularization through polymerization and ligation. After the polymerase finishes filling the gap, it has to dissociate from the DNA to enable the ligase to close the gap. We used the Stoffel fragment of the AmpliTaq DNA polymerase (Applied Biosystems) that lacks 5′→3′ exonuclease activity. Fewer circles were formed when 10× more or 10× less amount of Stoffel was used (data not shown). We further tested the effect of dNTP concentration on the polymerase strand displacement ability, and found that there seemed to be a narrow range of dNTP concentration that gave rise to the highest circularization efficiency (Fig. 2, right panel). With the optimal dNTP concentration along with at least 24 h hybridization and a probe to genome molar ratio of 100:1, we were able to form padlock circles from >1% of genomic DNA copies (i.e., >3000 circles) at each target site on average, which should virtually guarantee detection of each allele of a diploid DNA locus. This is a combined 435-fold improvement compared with the least optimal reaction condition using the optimized probe set (Fig. 2). When the additional 25-fold increase due to refined probe design and synthesis is factored in, we achieved a total of >10,000-fold improvement after the optimization. This suggests that even smaller amounts of genomic DNA can be analyzed successfully in cases where input sample DNA is limiting, e.g., when DNA is from tumor samples or microdissected tissue: 10 ng of genomic DNA would generate an average of more than 30 circles per site, which is sufficient for seeing at least three copies of each allele with a false-negative rate of 10^{-5}.

Performance of improved padlock capturing

Single-end sequencing of the ligation arm (25 bp) and 10 bases of the adjacent polymerized extension region (35 bp total) were

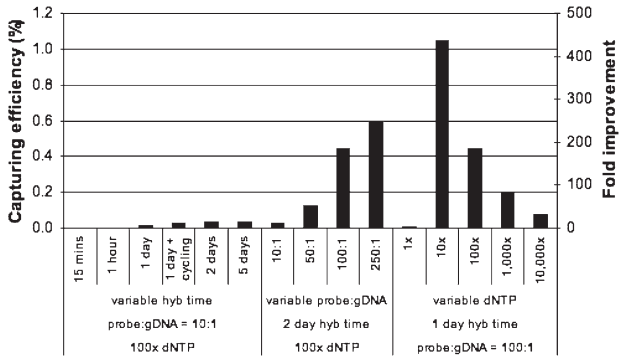


Figure 2. Improvement of padlock capturing efficiency with longer hybridization time (*left*), more probes (*middle*), and appropriate dNTP concentration (*right*). The ratios (10:1, 50:1, 100:1, and 250:1) are molar ratios between each of the padlock probes and genomes. dNTP (1×) is defined as the minimum amount of dNTP needed to capture all genomic copies at each target region. The fold improvement (vertical axis at *right*) is relative to the reaction with 10:1 probe ratio, hybridized for 15 min at 60°C, and with 100× dNTP. Similar results were observed in independent experiments.

obtained with an Illumina Genome Analyzer 1 (Fig. 1C,D) ("run 1"); subsequently, for reasons described below (and in the Supplemental Text), we resequenced these libraries on an updated Illumina Genome Analyzer 2 ("run 2"). The 10 bases of the polymerized extension region (which we called the "Target10" region) are the only parts of each sequence read that are copied from the subject genomes versus synthesized as parts of the padlock probes. The Target10 region thus comprises read positions 26–35, with each probe's target CpG occupying positions 26–27. Although there is an option to trim the ligation arm with the Type IIS restriction enzyme EcoP15I built into the probe design and then ligate with sequencing adapters, we chose the simple end-sequencing approach (Fig. 1C), which focuses on the CpGs at or near the ligation junction by design and that fall within the sequencing read length. Filtered, mappable read counts for each subject library ranged from 1.5 M to 3.8 M in run 1, and from 2.3 M to 4.8 M in run 2 (Supplemental Table 1).

We evaluated our improved padlock method with four metrics: sensitivity, uniformity, reproducibility, and the accuracy of the genotypes.

1. Sensitivity: The sensitivity of our multiplexed capture, which is related to multiplexity, can be assessed by the fraction of loci that are covered by at least one mapped read. With 2–3 million mapped reads, the sensitivity ranged from 90.8% to 94.0% (Fig. 3A; Supplemental Fig. 3). Each library offered the potential to resequence 537,770 Target10 base positions for each subject, and yielded an average of 346,749 (64.5%) Target10 base position allele determinations per library (replicates uncombined, Supplemental Table 1) due to residual probe circularization limitations, sequencing limitations, and filters used to ensure genotyping accuracy.
2. Uniformity: For both sequencing runs over all samples, 54.4%–56.5% of all captured targets had coverage levels within a 10-fold range, and 87.2%–92.7% had coverages within a 100-fold range (Fig. 3A; Supplemental Fig. 4).
3. Reproducibility: We conducted a technical replicate of our capturing experiment for HapMap sample NA10835 using our improved protocols. Read counts between the NA10835 replicates were correlated at >0.98 within sequencing runs (Fig. 3B; Supplemental Fig. 6), and at >0.95 across sequencing runs (see Supplemental Text). High reproducibility across the genotypes determined from the NA10835 replicates was also observed within the individual sequencing runs (>99.8% agreement in genotypes and >0.92 Spearman's rank correlation between genotype scores; see Supplemental Text).
4. Accuracy: Details are described in depth in following sections. Furthermore, to identify characteristics that may lead to further improvement of capture efficiency, we analyzed the relationship between coverage (i.e., the number of reads per site) and sequence features inherent in the probe design (Supplemental Fig. 7).

Genotype determination and spurious sequencing artifact

We determined genotypes for Target10 sites using an algorithm that makes use of Illumina base call quality scores, base-specific base call error rates, and Bayesian prior probabilities for the likelihood of a locus being a site of variation (see Methods and Supplemental Text). The algorithm delivers not only a genotype call for each site but also a genotype score indicating the confidence of the call. We generally used a genotype score of 5 as a cutoff, which represents a 1×10^{-5} chance that the genotype was called in error given the algorithm's error model. We optimized genotype calling accuracy using four measures, including the heterozygosity of Target10 loci by read position, CpG polymorphic allele fractions by read position, genotype call concordance with independent SNP data for the subjects, and profiles of dbSNP and non-dbSNP sites of variation (findings below and in Supplemental Text). Using this latter measure in our initial work with run 1, we found a consistently and apparently anomalously high degree of sharing of some heterozygous genotypes across subjects. Using conventional Sanger sequencing, we sequenced five dbSNP and 10 non-dbSNP highly shared heterozygous sites in all six subjects. The highly shared dbSNP sites were all validated, but none of the non-dbSNP sites were correctly called in any of the six subjects. Ultimately, we resequenced the same libraries used for run 1 but found that the anomaly persisted with run 2. However, the anomalously shared loci were distinct in the two runs, and the anomaly disappeared when we considered only those loci that were called with the same genotype in both runs (the "intersection" of the runs) (See Supplemental Text for extensive discussion and Supplemental Fig. 15 for examples). Since the libraries and computational processing were identical in all other respects, these observations suggest that incorrect reads for small subsets of probes may be generated within the sequencing process itself by an as yet unidentified mechanism. On this basis, we performed all analysis of genotypes based on the intersection, which comprises 442,937 Target10 loci genotyped in at least one subject sample.

Genotype validation

Of all dbSNP loci annotated to be in our Target10 regions (5059), we verified Target10 location matches of 5008 by aligning Target10 reference sequences against dbSNP sequences with stringent

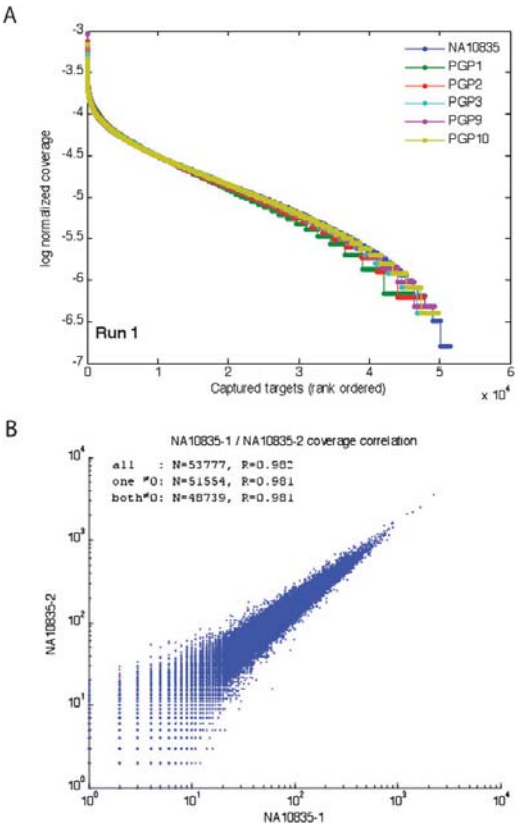


Figure 3. Improved performance of padlock technology. (*A*) Uniformity of target sites. For each sample, log-normalized coverage levels from sequencing of padlock probe reaction products were computed for each captured target as the $\log_{10}$ of the number of target-mapped, filtered reads divided by the total number of mapped, filtered reads from the reaction. Targets were then ranked for each sample from highest to lowest numbers of mapped, filtered reads and plotted. Except at the extremes, curves exhibit a gradually decreasing slope, indicating that a large number of targets have coverage levels within two orders of magnitude. The plot above depicts sequencing run 1; sequencing run 2 is very similar (Supplemental Fig. 4). For both sequencing runs, overall samples, 54.4%–56.5% of all captured targets had coverage levels within a 10-fold range, and 87.2%–92.7% had coverage within a 100-fold range. (*B*) Reproducibility of padlock capture. Scatter plot of read coverage of the technical replicate libraries sequenced for NA10835. Pearson correlation coefficients (R) between read counts are provided for all 53,777 target sites (all), all target sites for which one of the replicates has nonzero coverage (one $\neq 0$), and all for which both replicates have nonzero coverage (both $\neq 0$). All Pearson correlation coefficients are >98.1%. The scatter plot is presented on a log–log scale and therefore only contains points corresponding to targets in the "both $\neq 0$" set. The plot above depicts sequencing run 1; sequencing run 2 is very similar (Supplemental Fig. 6). For details on sequencing runs and read mapping and filtering, see text and Supplemental Text.

parameters (Supplemental Text; Supplemental Table 2). The number of these dbSNP loci that could be genotyped in an individual subject depended strongly on coverage ($R^2 = 0.885$; Supplemental Fig. 8). When we compared the bases actually found at these sites with previously observed alleles annotated for them in dbSNP, we found consistency rates >99.5% for all subjects, with statistically significantly greater consistency for dbSNP loci annotated as having been validated compared with dbSNP loci annotated with "unknown" validation (Supplemental Fig. 9). Additionally, among these sites, 2025 had been genotyped by the HapMap project for NA10835 (Frazer et al. 2007) and 217–245 had been genotyped independently on the Affymetrix SNP 500K platform for the other five Personal Genome Project subjects. Between 98.5% and 100% of these independently assessed genotypes matched our Target10 genotypes (Supplemental Table 3). For those genotypes assessed as heterozygous by the independent data source, the fraction of Target10 genotypes that did not match the independent genotypes was 0%–3%. These results confirm the accuracy of our padlock capture library sequencing and computational methods for genotyping.

Candidate novel SNPs

We identified 489 loci within the Target10 regions that were heterozygous in at least one subject and were not among our 5008 location-verified dbSNP loci. There were also 13 loci that were homozygous in each subject but for which the homozygous genotypes differed. These 502 loci represent candidate novel SNPs (Supplemental Table 4). Three hundred fifty-four of the heterozygous sites (71%) and eight of the 13 homozygous but differing loci (62%) are at the target CpG positions (26 and 27). Subject PGP10 is heterozygous at 215 non-dbSNP positions, a much higher number than the other subjects, which range from 65 to 87, depending on coverage (Supplemental Fig. 8). Taking the coverage relationship into account, PGP10 exhibits about three times the number of non-dbSNP heterozygous loci than the other subjects would have at PGP10's coverage level. PGP10 also exhibits less sharing of heterozygous genotypes compared with the other subjects (Supplemental Text; Supplemental Fig. 11). This and other findings below may relate to PGP10's African-American (AA) ancestry versus the European-American (EA) ancestry of the other five subjects, as the higher degree of genetic variation in AA populations may present SNPs uncommon in EA populations.

Biases in target capture

Target capture by padlock probes involves different enzyme behavior at the ligation junction than at other gap positions, while Illumina sequence quality can vary with read position (Supplemental Fig. 5). Both phenomena can potentially introduce biases into sequence data by read position. To assess for the presence of such effects, we analyzed heterozygosity (nucleotide diversity) by Target10 read position, and also compared the frequencies of CpG polymorphisms measured on an allele basis between target and nontarget positions. We found evidence of biases toward increased polymorphism rates in target CpG positions, especially in position 27 (see Discussion and Supplemental material). However, the size of the bias is small compared with the magnitude of the CpG polymorphism overall. For instance, the variation in heterozygosity between positions 26 and 27 is only ~1.25-fold compared with a ninefold change between these positions and 28–35 (Supplemental Table 5), while the difference between CpG polymorphic allele fractions across all positions is ~1.31-fold compared with a 12.5-fold difference between CpG and non-CpG allele fractions (see Supplemental material). Heterozygosity is expected to be

higher in positions 26–27 than 28–35 because of the high mutability of CpG dinucleotides, while positions 28–35 should be close to overall chromosome 21 averages. In fact, the average heterozygosity in positions 28–35 is 0.00067, slightly higher than the 0.00052 value found for chromosome 21 by the HapMap project (Sachidanandam et al. 2001). A higher value for positions 28–35 is expected because Target10 regions are often in CpG islands, so that these positions have a high frequency of CpGs compared with chromosome 21 as a whole, even though they are not CpG target positions (see Supplemental material). Meanwhile, the overall 12.5-fold difference between CpG polymorphic allele fractions and non-CpG allele fractions is close to estimates of CpG versus non-CpG mutation rates (see below).

CpG variation rates and forces influencing CpG and non-CpG variation

To estimate the impact of CpG context on mutation rate, we determined the ancestral state of each SNP using the chimpanzee genome. Comparison of densities of CpG and non-CpG SNPs in our padlock probes suggests that the rate of transitions originated at CpG sites is 13.7 times higher than the rate of non-CpG transitions. This estimate is in good agreement with previous studies (Kondrashov 2003; Hwang and Green 2004; Schmidt et al. 2008). This is a slight underestimate of the impact of CpGs on mutation rate because mutations in the chimpanzee lineage reduce the estimate of the number of CpG dinucleotides in the ancestral sequence.

Our estimates are unlikely to be biased toward various genomic features as CpG sites surveyed were chosen to be a large representative subset of all chromosome 21 CpG sites irrespective of known annotations. Correspondingly, <5% of CpG sites surveyed were located in known protein-coding regions, and only 20 SNPs there were detected.

In comparison with intergenic regions, CpG transition density per site was 15% lower in intronic regions and 3.2 times lower in coding regions. As expected within coding regions, SNP density per site was higher at fourfold degenerate sites than at nondegenerate sites. We observed one nonsense CpG transition (arginine to stop codon). It corresponded to the known rs4148974 SNP present in a facultative exon of the NADH dehydrogenase (ubiquinone) flavoprotein 3 gene (*NDUFV3*).

We also applied these data to ongoing analyses into the mechanisms of mutation rate and the basis of mutation rate heterogeneity. Mutation rate in mammals is known to be heterogeneous at a megabase scale (Wolfe et al. 1989; Smith and Lercher 2002; Gaffney and Keightley 2005). Various factors may contribute to mutation rate heterogeneity. Local CpG methylation level would affect the rate of CpG mutations and would leave the rate of non-CpG mutations unaffected. Fidelity of DNA replication and efficiency of the mismatch repair (MMR) system would primarily impact non-CpG mutation rate because mutations arising from deamination of methylcytosines escape the MMR system. Other factors, such as DNA exposure to damage and efficiency of base excision repair (BER) would impact both CpG and non-CpG mutation rate. Thus, comparison of CpG and non-CpG SNP densities along the chromosome may be informative about relative contributions of forces influencing mutation rate.

To analyze changes in the variation rate along the chromosome 21, we pooled CpG transitions and all non-CpG variants into 3-Mb windows. We observed a negative correlation between CpG content (fraction of CpG dinucleotides in a window) and the CpG polymorphism density per CpG site (Fig. 4A). Divergence with chimpanzee shows a similar but weaker effect (Fig. 4B). This suggests that the CpG mutation rate is heterogeneous and CpG dinucleotides are preserved in regions of lower mutation rate. An alternative (although less likely) explanation may be provided by natural selection maintaining hypermutable contexts.

The observed heterogeneity of CpG variation rate (and consequently the heterogeneity of the CpG content) along the chromosome could parallel the overall pattern of variation, or it could be CpG-specific. We find that CpG and non-CpG SNP densities are highly correlated (Fig. 4C). However, this correlation is expected to be strong because many factors unrelated to mutation rates impact both CpG and non-CpG variation. These factors include variation in coalescent times influenced by historic changes in human population size and population structure (Marth et al. 2003), background selection (Charlesworth et al. 1993), hitchhiking effect (Smith and Haigh 1974), and biased gene conversion (Webster and Smith 2004). To exclude the effect of these population factors, we analyzed the dependency of the CpG polymorphism rate on species divergence in CpG and non-CpG sites. SNP densities in both CpG and non-CpG sites do not display a statistically significant correlation with non-CpG divergence in the chimpanzee lineage in our data set (Fig. 4D,E). This correlation would reflect the influence of mutation rate components common to CpG and non-CpG mutations, such as exposure of DNA to damage and the efficiency of the BER system (Stamatoyannopoulos et al. 2009). It may also reflect the effect of natural selection. Contributions of these forces to the megabase-scale heterogeneity of the human polymorphism rate appear limited, at least in our sparse data set. The density of SNPs originated in CpG sites shows a stronger significant correlation with CpG divergence in the chimpanzee lineage (Fig. 4F). This suggests that local CpG mutation rate is primarily influenced by factors specific to CpG dinucleotides. We hypothesize that heterogeneity of the methylation rate along the chromosome may underlie the heterogeneity of the CpG mutation rate.

A more precise analysis would be possible with larger targeted CpG data sets than these relatively sparse chromosome 21-based data. However, the results above were confirmed using species divergence and population variation data in the genomic regions included in phase I of the ENCODE project (data not shown). In these regions, correlation of SNP densities in both CpG and non-CpG sites with non-CpG divergence appears statistically significant, but still remains much weaker than correlation between diversity and divergence in CpG sites.

Discussion

The cost of DNA sequencing has been continuously dropping in the past several years toward the goal of sequencing a human genome at $1000. However, regardless of the ever-lowering cost of DNA sequencing, it is always more cost-effective to selectively sequence regions of interest, thus allowing more samples to be analyzed. Until the cost of the target capture overwhelms the sequencing cost, targeted sequencing remains a viable and highly demanded approach in genetic studies and diagnosis.

In this work, we improved the padlock-based capturing method significantly by refining the probe design algorithm and probe synthesis protocol, extending the hybridization time, increasing the amount of the probes, and tuning the dNTP concentration. We applied this improved technology to amplify 53,777 regions of 40 bp containing CpGs across human chromosome 21.

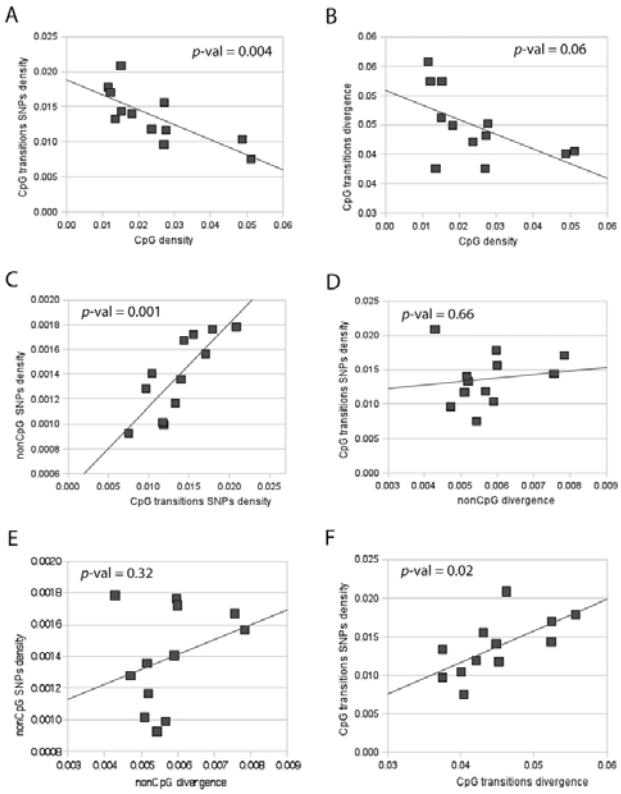


Figure 4. Correlations between polymorphism, interspecies divergence, and CpG content. We analyzed divergence in the chimpanzee lineage after divergence from human using orangutan as an outgroup in order to compensate for bias due to padlock probe design based on the presence of CpGs in the human sequence. SNP densities were calculated as normalized densities per site of a specific type. To calculate the density of CpG SNPs, we divided the total number of the observed CpG polymorphisms in the region by the combined length of all surveyed CpG nucleotides in the region. Correspondingly, to calculate the density of non-CpG SNPs, we divided the total number of observed non-CpG polymorphisms in the region by the combined length of all surveyed non-CpG nucleotides in the region. (*A*) Correlation between densities of SNPs originated as transitions in CpG sites and fraction of CpGs in the region. (*B*) Correlation between substitutions due to CpG transitions in the chimpanzee lineage after divergence with humans and fraction of CpGs in the region in the human genome. (*C*) Correlation between densities of SNPs originated as transitions in CpG sites and non-CpG SNP density. (*D*) Correlation between densities of SNPs originated as transitions in CpG sites with non-CpG divergence in the chimpanzee lineage after split with humans. (*E*) Correlation between non-CpG SNP density with non-CpG divergence in the chimpanzee lineage after split with humans. (*F*) Correlation between densities of SNPs originated as transitions in CpG sites with divergence in the chimpanzee lineage due to CpG transitions.

A uniform gap size was chosen for the simplicity of sequencing library construction; new protocols developed herein have also successfully applied to probe sets with various target sizes, such as the exon set (JB Li, K Zhang, and GM Church, unpubl.).

The capture efficiency has increased >10,000-fold compared with our previous report (Porreca et al. 2007), where <20% of 55,000 exons were amplified and most of the heterozygous SNPs were erroneously called as homozygous. We are now able to observe over 50,000 sites of 53,777 desired targets with ~3 million reads from a single lane of the Illumina Genome Analyzer. The uniformity of different targets was also significantly improved; ~90% and ~55% of all targets were within a 100- and 10-fold range of each other, respectively. The accuracy of genotyping calls was estimated to be 98.5% for NA10835 in comparison with HapMap data, and 99.2%–100% accurate for the other subjects with PGP Affymetrix 500K SNP data. The modified experimental protocol mainly accounted for the significant improvement. In addition, these metrics were achieved with a series of constraints in the probe design. Lastly, steady improvement of our oligonucleotide synthesis technology (Agilent) enabled us to start with roughly equal amounts of probe precursors (Fig. 1B).

Compared with the microarray-based "on surface" hybridization approach to enrich targets (Albert et al. 2007; Hodges et al. 2007; Okou et al. 2007), our significantly improved padlock-based "in solution" technology has led to (1) efficient hybridization as >1% of genomic copies successfully form circles at each locus, (2) a requirement of ~20-fold less input genomic DNA, and (3) better scalability of the reactions, particularly when the sample size is large. Although a recently developed method based on "in solution" hybridization has solved these problems (Gnirke et al. 2009), the close to 100% specificity of the padlock approach is unmatched since two linked probing sequences, rather than one, need to be precisely hybridized to form circles (Porreca et al. 2007). However, padlock-based target capture technology is subject to biases in circle formation due to the enzymatic requirements of the reaction. With the new protocols developed in this work, the uniformity of distribution among different target regions has been significantly improved. For example, the fraction of sites within 100-fold in abundance has increased from 16% to ~90%.

The uniformity could be further improved based on what we learned in this work. First, selection of extension and ligation arms could be more flexible in a wider window flanking the target. This would make it possible to design probes for difficult regions or probes that better satisfy the design criteria. Additionally, we observed that the first base on the ligation arm of the padlock probe (proximal end to the target gap) contributed to the success of circle formation, and the G+C content of targets led to amplification bias (Supplemental Fig. 7). These pitfalls could be avoided or alleviated by the flexibility in probe design. In addition, longer padlock probes may further improve the uniformity as demonstrated in a recent smaller scale study (Krishnakumar et al. 2008). Lastly, due to the fact that the abundance of reads per site spans a three-log magnitude range and this difference in abundance is systematic

rather than random (Fig. 3), the probes could be divided into three or more subsets so that the sites would fall into a 10-fold range in each subset (Deng et al. 2009).

Our improved padlock-based capture technology can be extended to a variety of applications. For example, one can efficiently capture genomic regions, such as SNPs, exons, and contiguous regions containing susceptible mutations from gene mapping studies. Such tools are greatly demanded in projects including personal genomes, cancer genomes, and genome-wide association follow-up studies. In addition, the cDNA derived from RNA can also be targeted to quantitatively measure allele frequencies in gene expression (Zhang et al. 2009) and to identify RNA editing events (Li et al. 2009). Because of the superior specificity of the padlocks, we recently applied them to bisulfite converted human genomes to profile cytosine methylation (Ball et al. 2009; Deng et al. 2009).

We showcased the utility of data obtained using our padlock probes by accurately estimating the CpG mutation rate as at least 13.7 times the non-CpG mutation rate in humans, using the chimpanzee genome to identify ancestral CpGs. A simple quantitative analysis of the data obtained using padlock probes suggests that CpG polymorphism density is highly variable along the chromosome. Comparison of human polymorphism and species divergence in CpG versus non-CpG sites demonstrates that this heterogeneity is primarily due to factors specific to CpG dinucleotides. This suggests that variation in methylation rate may be an important determinant of local mutation rate in humans. This analysis does not exclude the alternative hypothesis that hypermutable CpG contexts are preserved by purifying selection.

We believe that the inexpensive and rapid analysis of CpG mutations has a potential for a number of applications in human genetics. First, genetic diagnostics of autosomal dominant genetic disorders and discovery of genes involved in developmental defects requires detection of de novo mutation events. Highly accurate sequencing of the complete genome or even "exome" is likely to remain prohibitively expensive in coming years. At the same time, a much less expensive screening of CpG mutations will have more than one-third of a chance to find the relevant de novo sequence change at a small fraction of the cost (Cooper and Youssoufian 1988; Cooper and Krawczak 1993). Therefore, CpG screening can be used as an efficient first pass for detecting de novo mutations. Second, rare alleles involved in human complex phenotypes can be detected by systematic resequencing of phenotyped populations. CpG screening may provide a more efficient study design because it will detect a large fraction of rare alleles at a small fraction of the cost. We expect to carry this study design forward in larger phenotyped populations such as those from Genome-Wide Association Studies and the Personal Genome Project.

Methods

Selection of 53,777 CpG loci across human chr21

Human genomic sequences (hg18) were downloaded from the UCSC Genome Browser (http://genome.ucsc.edu). The total number of CpGs on human chr21 is 371,971. We started with 223,058 nonoverlapping probes that could possibly cover the maximum number of CpGs on human chromosome 21, and 101,822 of them were nonrepetitive, which made probe design possible. We designed 53,777 probes that can be accommodated by the capacity of oligonucleotide synthesis on a single microarray, with the following stringent considerations: (1) the CGs did not fall into the repetitive regions; (2) the Tm of the extension arm ranges 50°C–58°C, with 53°C being optimal; the Tm of the ligation arm ranges 53°C–61°C, with 58°C being optimal; (3) the length of extension and ligation arms ranges from 17 to 25 nt; (4) the arms were compared against a pre-computed occurrence table of all 12-mers in the human genome. The sum of the 12-mer occurrences in the arms was normalized by the arm length. Only arms with a normalized 12-mer count less than 2000 were retained; (5) the pair of extension and ligation arms separated by a gap was BLASTed against the human genome. When both arms matched a second location in the genome, the probe was discarded; and (6) the GC content of both arms was 30%–70%.

Improved padlock capturing protocol

Generation of padlock probes

Using a programmable microarray (Agilent Technologies), we synthesized 53,777 oligos (150-mers), cleaved them off the microarray, and collected them in a single Eppendorf tube. Each of the oligo species is ~0.2 fmol, totaling ~10 pmol of oligos synthesized on one array. The sequence of the 150-mer oligo is ATCAAGC CGAAGACAGTGTT[ligation_arm][random_ligation]TCTCTGCT GCTTCAGCTTCCCAGTCGTGGTACATACGAGCGATATCCGAC GGTAGTGTAC[random_extension][extension_arm]GATCCAGG AAATTCGCGCTA. Random sequences may be added to extend ligation and extension arms to 25 bases for uniform probe size.

A 100-μL PCR reaction was assembled with 90 μL of Platinum Taq supermix (Invitrogen), 50 pmol each of forward primer, A*T*C*AAGCCGAAGACAGTGT/deoxyUridine/ (* denotes phosphothiorate bond), and reverse primer, /5phos/TAGCGCGAATTT CCTGGATC (both from IDT), 0.5× SYBR Green (Invitrogen), and 10–100 fmol of template. The quantitative PCR program was 5 min at 95°C; nine to 15 cycles of 30 sec at 95°C, 1 min at 58°C, and 1 min at 72°C; and 5 min at 72°C. We first used 1% (~100 fmol) of oligos provided by Agilent in a single 100-μL reaction with nine cycles. Approximately 10 fmol of the purified PCR product was used as template in each of the 5 × 96 100-μL reactions with 12–15 cycles in the second round of PCR. The PCR product was purified with a QIAquick PCR purification kit (Qiagen).

Sixteen tubes of 100-μL reactions with 1× λ exonuclease buffer, 5.6 μg of PCR products, and 25 units of λ exonuclease were incubated for 1 h at 37°C, then 15 min at 75°C. The reaction was purified with a QIAquick PCR purification kit, eluted with 400 μL of dH_2O, and quantified with a Nanodrop (Thermo Scientific) at 45 ng/μL. Eight tubes of 60-μL reactions with 2.25 μg of λ exonuclease-treated single-stranded DNA, 1× DpnII reaction buffer (NEB), and 200 pmol of Guide_DpnII (GCGCGAATTTCCTG GATCNN) were denatured for 5 min at 95°C, ramped to 60°C at 0.1°C/sec, and incubated for 10 min at 60°C and 1 min at 37°C. For each of the 60-μL reactions, we then added 50 units of DpnII (NEB) and five units of USER (NEB), and incubated the reaction for 3 h at 37°C. The post-processing 110-mer padlock probes were size selected on 6% denaturing acrylmide gels (Invitrogen) and eluted in 100 μL of dH_2O. The total concentration of the 110-mer padlock probes was estimated to be 16 ng/μL (441 nM) by a denaturing acrylmide gel.

Hybridization

The genomic DNAs of the HapMap sample, NA10835, and five Personal Genome Project (PGP) samples (NA20431, NA21070, NA21660, NA21781, and NA21833) were obtained from Coriell. In a 15-μL reaction, 1× Ampligase buffer (Epicentre), 500 ng (0.25 amol) of genomic DNA, and 48 ng (1.32 pmol) of probes were mixed (each probe to gDNA molar ratio = 100:1; numbers change

accordingly for other ratios), denatured for 10 min at 95°C, ramped at 0.1°C/sec to 60°C, and then hybridized for 24 h at 60°C. We then added 2 μL of gap filling and sealing mix (5.4 μM dNTPs [100×, numbers change accordingly for 1×, 10×, 1000×, and 10,000×], two units of Taq Stoffel fragment [Applied Biosystems], and 2.5 units of Ampligase [Epicentre] in Ampligase storage buffer [Epicentre]), and incubated the reaction for 15 min, 1 h, 1 d, 2 d, or 5 d at 60°C. We also tried cycling the reaction: after 1 d at 60°C, we applied 10 cycles of 2 min at 95°C followed by 2 h at 60°C. To remove the linear DNA, we lowered the incubation temperature to 37°C, immediately added 2 μL of Exonuclease I (20 units/μL) and 2 μL of Exonuclease III (200 units/μL) (both from USB), and incubated the reaction for 2 h at 37°C followed by 5 min at 94°C.

Amplification of captured circles

The circles were amplified by two 100-μL PCR reactions with 50 μL of 2× iQ SYBR Green supermix (Bio-Rad), 10 μL of circle template (from above), and 40 pmol each of forward (CAAGCAGAAGACG GCATACGAGCGATATCCGACGGTAGTGTAC) and reverse (AATG ATACGGCGACCACCGACACTAACACGACTGGGAAGCTGAAGC AGCAG) primers (IDT). The PCR program was 3 min at 96°C; three cycles of 30 sec at 95°C, 30 sec at 60°C, and 30 sec at 72°C; and 10 cycles of 30 sec at 95°C, 1 min at 72°C, and 5 min at 72°C. The desired PCR products were gel purified and quantified. For each sample, 10–20 fmol of DNA was sequenced by both Illumina Genome Analyzer version 1 and updated version 2 with the custom primer (CACTAACACGACTGGGAAGCTGAAGCAGCAGAGA).

Illumina sequence processing, genotyping, and SNP identification

Illumina reads were mapped to target sequences by aligning them along their full 35-bp lengths to the 35 bp of human genome reference sequence corresponding to each of the 53,777 targets using in-house built dynamic programming software. Only reads with at most one mismatch or gap compared with a target, and no alignment better than two mismatches away from distinct target sequences, were accepted.

Genotypes were computed using an in-house-developed algorithm that takes into account base-specific base calling error rates determined from mapped reads, base call quality scores, and prior probabilities for genotypes. The algorithm computes a probability P(genotype = xy | Illumina base calls and qualities) for all 10 possible genotypes xy, calls the genotype xy with the highest probability, and assigns the value $-\log_{10}(1 - P)$ as a confidence score. This score measures the probability of a miscall given the algorithm's error model. Additional details are provided in the Supplemental Text.

SNPs were identified as loci that were found to be heterozygous in at least one subject, or as homozygous in all subjects that could be genotyped for the locus with at least one subject having a different homozygous genotype than the others. Variants were associated with known dbSNP loci by downloading all dbSNP entries with moltype "single" and locType "exact" in the chromosome 21 ranges occupied by all 53,777 padlock probe Target10 regions by using the UCSC Genome Browser (http://genome.ucsc.edu), and then verifying that the location of the dbSNP SNP mapped precisely to the candidate SNP position (build 129). Additional details on SNP matching can be found in the Supplemental Text.

Illumina reads for both runs 1 and 2 are available from the NCBI Short Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi) under accession number SRA007914. All candidate novel SNPs identified in this study, and also all known SNP loci for which a heterozygous genotype or two disagreeing homozygous genotypes could be found among the six subjects, were submitted to dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) where they were assigned accession numbers ss120032891– ss120035117. dbSNP was also provided with individual subject genotypes for all of the known dbSNP and candidate novel SNP loci identified in this study. All dbSNP submissions used handle "CHURCH_CG54K."

Genotyping accuracy determination

Accuracy was determined primarily by comparing genotypes determined for all subjects with independently available SNP genotype data for subsets of loci that we could identify as assayed in both our and the independent studies. For NA10835, we used 2025 SNP loci genotyped by the HapMap project downloaded from http://ftp.hapmap.org/genotypes/latest/forward/non-redundant/. For the other five subjects, we used 217–246 SNPs that were assayed by Affymetrix SNP 500K arrays for the Personal Genome Project (http://personalgenomes.org). See Supplemental Text for additional details.

Heterozygosity and CpG polymorphic allele fraction determinations

Heterozygosity was computed as the number of heterozygous genotypes divided by the total number of genotypes. CpG polymorphic allele fractions were computed by counting the total number of alleles that could be determined in subject genotypes that correspond to CpGs in the hg18 human genome reference sequence in which Cs appeared as Ts (for CpG→TpG variations), or where Gs appeared as As (for CpG→CpA variations), and dividing by the total numbers of alleles of any sort determined for these CpG positions. Non-CpG polymorphic allele fractions were computed similarly. For additional details, see Supplemental Text.

Calculations of CpG and non-CpG divergence

We used orangutan as an outgroup to determine which nucleotide differences between the human and chimpanzee genomes occurred in the chimpanzee lineage. The human–orangutan alignment (hg18 vs. ponAbe2) was downloaded from the UCSC Genome Browser website (http://genome.ucsc.edu). CpG transition diversity was estimated by dividing the total number of CpG sites conserved between human and orangutan but harboring a transition in chimpanzee sequence by the total number of CpG sites conserved between human and orangutan and alignable with the chimpanzee genome. Non-CpG transitions and transversions were combined in a single class of non-CpG substitutions and their density was calculated in a similar way.

Calculation of CpG and non-CpG SNP densities

CpG transitions SNP density was calculated by dividing the number of reliably detected transitions in CpG nucleotides (with genotype scores of 5 or higher) by the total number of CpG sites among the first 10 probe's positions detected with a genotype score of 5. Non-CpG transitions and transversions were combined in a single class of non-CpG SNPs and their density was calculated in a similar way.

Statistical analysis of correlations

Statistical significance of the observed correlations was analyzed by three methods: (1) by construction of linear regression models with the subsequent application of an *F*-test, (2) by Spearman's

rank-order correlation test, and (3) by Kendall's rank correlation test. All three methods produced very similar *P*-values.

Acknowledgments

We thank Jay Shendure, Gregory Porreca, and Joseph Chou for input in the padlock technology development; Uri Laserson, Madeleine Price Ball, Michael Chou, Alon Keinan, and Heidi Rehm for discussion and/or critical reading of the manuscript; and the Harvard Research Institute Technology Group for computing resources. Funding came from NHGRI Center of Excellence in Genome Sciences and NHLBI targeted sequencing grants.

References

Akhras MS, Unemo M, Thiyagarajan S, Nyren P, Davis RW, Fire AZ, Pourmand N. 2007. Connector inversion probe technology: A powerful one-primer multiplex DNA amplification system for numerous scientific applications. *PLoS One* **2:** e915. doi: 10.1371/journal.pone.0000915.

Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, et al. 2007. Direct selection of human genomic loci by microarray hybridization. *Nat Methods* **4:** 903–905.

Ball MP, Li JB, Gao Y, Lee JH, LeProust EM, Park IH, Xie B, Daley GQ, Church GM. 2009. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotechnol* **27:** 361–368.

Bodmer W, Bonilla C. 2008. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* **40:** 695–701.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134:** 1289–1303.

Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. 2004. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305:** 869–872.

Cohen JC, Boerwinkle E, Mosley TH Jr, Hobbs HH. 2006. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N Engl J Med* **354:** 1264–1272.

Cooper DN, Krawczak M. 1993. *Human gene mutation*. BIOS Scientific, Oxford, UK.

Cooper DN, Youssoufian H. 1988. The CpG dinucleotide and human genetic disease. *Hum Genet* **78:** 151–155.

Dahl F, Gullberg M, Stenberg J, Landegren U, Nilsson M. 2005. Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic Acids Res* **33:** e71. doi: 10.1093/nar/gni070.

Dahl F, Stenberg J, Fredriksson S, Welch K, Zhang M, Nilsson M, Bicknell D, Bodmer WF, Davis RW, Ji H. 2007. Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proc Natl Acad Sci* **104:** 9387–9392.

Deng J, Shoemaker R, Xie B, Gore A, LeProust EM, Antosiewicz-Bourget J, Egli D, Maherali N, Park IH, Yu J, et al. 2009. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat Biotechnol* **27:** 353–360.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449:** 851–861.

Gaffney DJ, Keightley PD. 2005. The scale of mutational variation in the murid genome. *Genome Res* **15:** 1086–1094.

Garber K. 2008. Fixing the front end. *Nat Biotechnol* **26:** 1101–1104.

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, et al. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* **27:** 182–189.

Hardenbol P, Baner J, Jain M, Nilsson M, Namsaraev EA, Karlin-Neumann GA, Fakhrai-Rad H, Ronaghi M, Willis TD, Landegren U, et al. 2003. Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat Biotechnol* **21:** 673–678.

Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, et al. 2007. Genome-wide in situ exon capture for selective resequencing. *Nat Genet* **39:** 1522–1527.

Hwang DG, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci* **101:** 13994–14001.

Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, Simon DB, Newton-Cheh C, State MW, Levy D, Lifton RP. 2008. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet* **40:** 592–599.

Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, et al. 2008. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321:** 1801–1806.

Kondrashov AS. 2003. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum Mutat* **21:** 12–27.

Kondrashov FA, Ogurtsov AY, Kondrashov AS. 2006. Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites. *J Theor Biol* **240:** 616–626.

Krishnakumar S, Zheng J, Wilhelmy J, Faham M, Mindrinos M, Davis R. 2008. A comprehensive assay for targeted multiplex amplification of human DNA sequences. *Proc Natl Acad Sci* **105:** 9296–9301.

Li JB, Levanon EY, Yoon JK, Aach J, Xie B, LeProust E, Zhang K, Gao Y, Church GM. 2009. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324:** 1210–1213.

Marth G, Schuler G, Yeh R, Davenport R, Agarwala R, Church D, Wheelan S, Baker J, Ward M, Kholodov M, et al. 2003. Sequence variations in the public human genome data reflect a bottlenecked population history. *Proc Natl Acad Sci* **100:** 376–381.

Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME. 2007. Microarray-based genomic selection for high-throughput resequencing. *Nat Methods* **4:** 907–909.

Olivier M, Eeles R, Hollstein M, Khan MA, Harris CC, Hainaut P. 2002. The IARC TP53 database: New online mutation analysis and recommendations to users. *Hum Mutat* **19:** 607–614.

Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu IM, Gallia GL, et al. 2008. An integrated genomic analysis of human glioblastoma multiforme. *Science* **321:** 1807–1812.

Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F, et al. 2007. Multiplex amplification of large sets of human exons. *Nat Methods* **4:** 931–936.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928–933.

Schmidt S, Gerasimova A, Kondrashov FA, Adzuhbei IA, Kondrashov AS, Sunyaev S. 2008. Hypermutable non-synonymous sites are under stronger negative selection. *PLoS Genet* **4:** e1000281. doi: 10.1371/journal.pgen.1000281.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. *Nat Biotechnol* **26:** 1135–1145.

Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, et al. 2006. The consensus coding sequences of human breast and colorectal cancers. *Science* **314:** 268–274.

Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res* **23:** 23–35.

Smith NG, Lercher MJ. 2002. Regional similarities in polymorphism in the human genome extend over many megabases. *Trends Genet* **18:** 281–283.

Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. 2009. Human mutation rate associated with DNA replication timing. *Nat Genet* **41:** 393–395.

Stenson PD, Ball E, Howells K, Phillips A, Mort M, Cooper DN. 2008. Human Gene Mutation Database: Towards a comprehensive central mutation database. *J Med Genet* **45:** 124–126.

Subramanian S, Kumar S. 2003. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. *Genome Res* **13:** 838–844.

Sved J, Bird A. 1990. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc Natl Acad Sci* **87:** 4692–4696.

Wang RY, Kuo KC, Gehrke CW, Huang LH, Ehrlich M. 1982. Heat- and alkali-induced deamination of 5-methylcytosine and cytosine residues in DNA. *Biochim Biophys Acta* **697:** 371–377.

Webster MT, Smith NG. 2004. Fixation biases affecting human SNPs. *Trends Genet* **20:** 122–126.

Wolfe KH, Sharp PM, Li WH. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* **337:** 283–285.

Zhang K, Li JB, Gao Y, Egli D, Xie B, Deng J, Li Z, Lee JH, Aach J, LeProust E, et al. 2009. Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat Methods* **6:** 613–618.

Received February 11, 2009; accepted in revised form May 20, 2009.

Appendix F

Computational design of molecular inversion probes for targeted genomic sequencing using MIPTAG Pro

This work was originally published as Chapter 5 of a dissertation presented by M.F. Chou to Harvard Medical School in partial fulfillment for the degree of Ph.D. in the subject of Genetics in May, 2009.

Author Contributions: M.F.C. wrote and implemented the algorithm based upon input and advice provided by A.M.R. and J.B.L..

Chapter 5

Computational design of molecular inversion probes for targeted genomic sequencing using MIPTAG Pro

Michael F. Chou[1,2], Jin Billy Li[1], Abraham M. Rosenbaum[1] and George M. Church[1]

[1]Department of Genetics, Harvard Medical School, Boston, MA 02115

[2]Corresponding author:

Michael F. Chou
New Research Building
Harvard Medical School
77 Avenue Louis Pasteur
Boston, MA 02115
mchou@fas.harvard.edu

MFC did all of the algorithm design, development and programming. JBL and AMR provided useful performance data and discussions to guide the design.

ABSTRACT

Second generation high throughput sequencing platforms are revolutionizing genetic and genomic sequencing efforts. Using molecular inversion probes to capture genomic targets prior to sequencing allows these high throughput platforms to be used for very specific genes or genomic loci in many individuals. Until now, design of sequence capture probes for arbitrary targets in the genome has not been routinely available. Here we describe an automated software pipeline called MIPTAG Pro that takes a list of genes or genomic loci as input, and creates a list of molecular inversion probe sequences designed to capture those loci. This pipeline incorporates a number of rules learned in the use of earlier, less automated designs, and is currently being used for a number of targeted genetic and genomic projects.

INTRODUCTION

The creation of second generation high-throughput sequencing (HTS) technology developed in our lab and others[1-3] has dramatically decreased the cost of DNA sequencing. The availability of a number of commercial instruments has now allowed countless labs to approach large scale biological questions in a way that was almost unthinkable just a few years ago. Examples of studies enabled by this technology include: sequence-based expression analysis[4], Chip-Sequencing studies[5], small RNA studies[6], and individual genome sequencing of James Watson[7], J. Craig Venter[8] and an anonymous Asian individual[9, 10].

When HTS is combined with the ability to routinely capture and sequence specific subsets of genes or loci for genomic sequencing, the value of this technology is amplified even further because we can potentially use the same sequencing throughput to sequence a subset of genes in hundreds or thousands of individuals for the cost of sequencing just one individual completely. For example, by just sequencing the 2% of the human genome responsible for protein coding genes it is possible to sequence roughly 50 times as many individuals with the same sequencing capacity. This targeted approach is being taken with the first individuals in the Personal Genome Project (PGP) whose goal it is to sequence 100,000 individuals[11].

Methods of targeted capture of the exon coding subset of the genome have been published by at least two groups, but at this time, these methods are not routinely available to most of the scientific community. One of these approaches, hybridization with long 120 bp biotinylated capture probes developed at the Broad Institute, was described recently[12].

Our lab[13] and others[14] have taken the approach of targeted capture using molecular inversion probes (MIPs) that we first described as a pilot project to specifically capture a subset of 55,000 exons in the human genome (the 55K probe set). MIPs essentially allow for highly multiplexed yet also highly specific PCR reactions prior to sequencing. Probes in the 55K pilot project were designed to capture exons in the range of 60 - 191 nt in length[13]. Subsequent development of MIP protocols have led to an over 10,000 fold improvement in sequence capture and coverage efficiency making the method also very reliable for heterozygous genotype determination (Li et al, submitted).

The original 55K probe sets have now been used to determine preliminary sequence data from cell lines derived from the first 10 volunteers of the Personal Genome Project (the "PGP10"). These first generation MIP-based capture probes, while extremely specific, provide non-uniform sequence coverage with up to about 4 orders of magnitude systematic differences in final sequence coverage within the probe set (Rosenbaum et al, unpublished).

These 55K capture designs and others have provided empirical evidence of which probes result in high or low sequence coverage, and the following parameters are anti-correlated with high final sequence coverage: target length (particularly over 100 bp), high or low GC content within the target sequence, low melting temperature (Tm) of the probe capture arms, and the identity of the terminal base on the ligation end of the probe being thymidine (Li et al., unpublished).

Our experience with these 55K designs and the desire to utilize this technology on the complete human exome and other smaller targets motivated us to generalize and automate the entire probe design process and incorporate the ability to not only capture larger exons, but also a variety of other genomic features.

Here we describe a new MIP design strategy, which we call MIPTAG Pro (for Molecular Inversion Probes for Targeted Genomic sequence Producer) that can be used to target any genomic sequence – exonic or otherwise. We describe the computational approach that incorporates the lessons learned from the first generation probe design, and our algorithm for capturing long regions of target sequence that far exceed the exon sequence length restrictions of previous designs.

Probes generated using this new design methodology are now beginning to be used by our lab and collaborators to routinely sequence multiple genes and regions of the genome.

RESULTS

Given the potential to simultaneously target as many as half a million target sequences in a single complex multiplex capture reaction, we felt the need to satisfy two principle constraints. First, probes must be very unique to be highly specific; cross hybridization is undesirable even if it doesn't ultimately result in a mis-captured target because it competes with intended probe targets. Second, the melting temperature for each type of arm (extension or ligation) must be uniform for all of the probes. We describe our method to satisfy those two constraints (amongst others), and the results of the design pipeline with some example cases. (For details on the MIP capture method and terminology used throughout, see Porecca et al.[13])

The probe design pipeline

The process of probe design is broken up into phases, and a conceptual overview of the process for a hypothetical gene sequence is shown in Figure 5.1. (Details of different phases of the pipeline are discussed in more detail in Materials and Methods.)

Probe design input

Our probe design begins with a list of genomic loci and their coordinates, and a reference genomic sequence for the organism of interest. Coordinates of annotated genes,

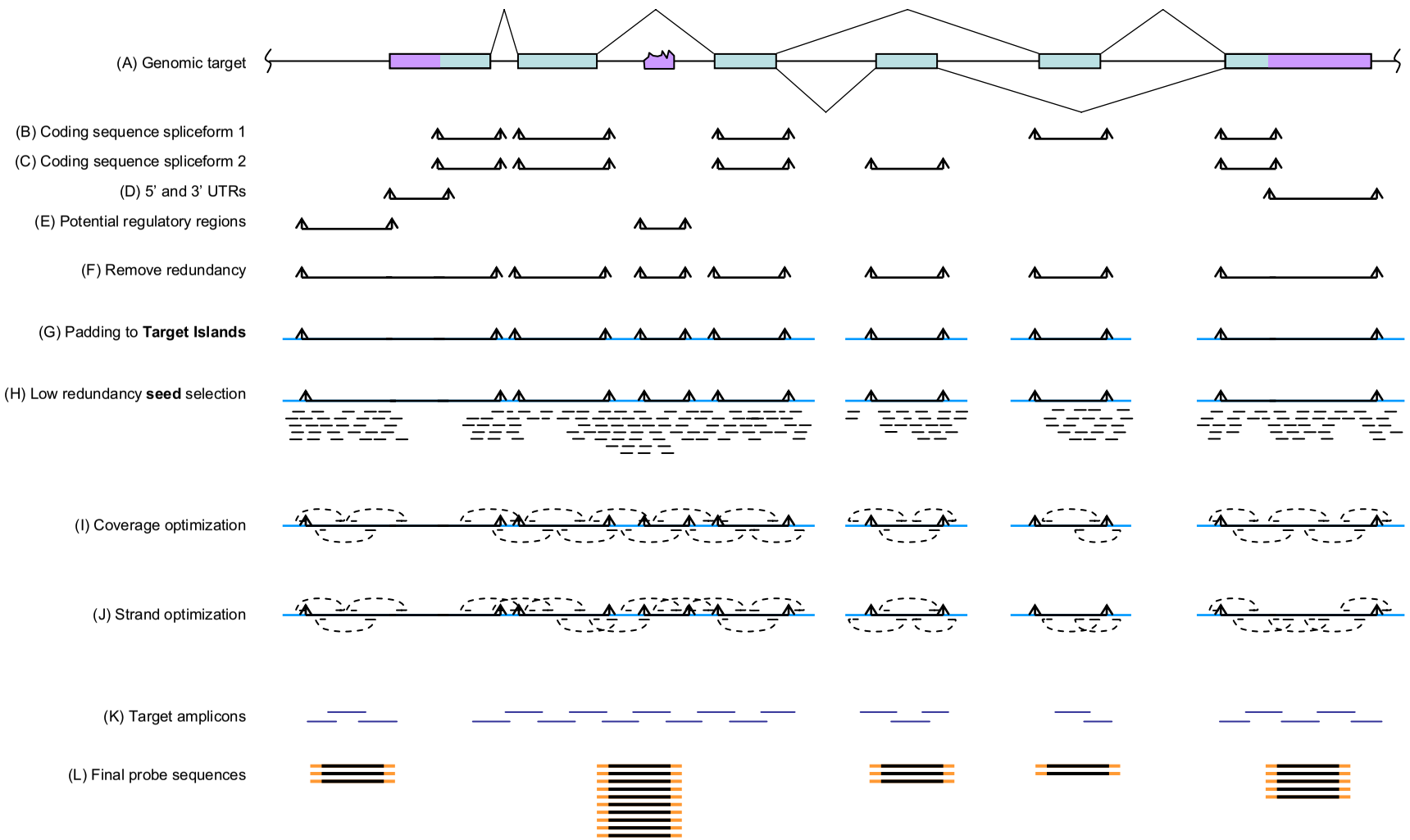


Figure 5.1: Probe design phases overview.

promoter regions, intergenic sequences, known conserved elements or micro-RNA encoding elements have all been used as input to the design pipeline.

Figures 5.1A-E illustrate some of these example inputs including target coding sequences (Figure 5.1B-C), UTR regions (Figure 5.1D), and regulatory or conserved regions (Figure 5.1E), but features may also be arbitrarily large regions such as a broad peak from a genetic linkage analysis.

Phase 1 – redundancy removal

Before actually designing probes for a given target region all target regions are consolidated and merged into contiguous regions (Figure 5.1F). In order to allow the design process more flexibility, and to allow probes to be designed starting from *outside* of the regions of interest, each target is padded usually with 200 bp. This is known as **non-target** DNA, and it is tracked separately from **target** DNA sequences throughout the entire pipeline. The addition of padding may again cause non-target sequences from different loci to overlap and be merged. Entire regions of overlapping targets and non-target sequences are then combined into contiguous sequences called **target islands** (Figure 5.1G).

Phase 2 – seed selection

Each target island (which includes target and non-target regions) is tiled with overlapping short 10 to 15 bp subsequences that we call **seeds** that occur less than 32 times throughout the entire genome of the organism. Each seed also includes the

coordinates of each of these occurrences throughout the genome (see Figure 5.1H). In subsequence steps, seeds are used as candidate starting points for probe arms.

Phase 3 – coverage optimization

Once all target islands have been seeded, each island can be processed independently of all others because they already contain global uniqueness information by the existence of the seeds themselves. Also, each island is far enough away from any other island so that any probes designed from one island will not overlap nor interfere with those of another island.

Coverage optimization incorporates all of the design rules based on previous experimental evidence that act as constraints in this process. Starting with an attempt to generate a single probe to cover the island, the goal is to test all pair-wise combinations of seeds as potential probe arms until a pair is determined that can maximally cover the **target** sequences within the island with a minimum of **non-target** sequence while also satisfying all of the design rule constraints. In practice, only the smallest islands with the fewest seeds can be exhaustively checked to find the best possible probe arm combinations, therefore, we use a greedy approach that can create probes satisfying all of the design constraints even for very large targets.

After a single probe is attempted, two are attempted, then three and so forth in a stereotypical interlocking pattern (see Supplementary Figure 5.1) until no further improvement in coverage is achieved with more probes. Even if a target island is fairly large, single probes are still attempted because there is no guarantee that an island will be

100% covered, so even a single probe solution may in theory be the best solution for such targets.

The probe design algorithm starts with the seeds tiled across a target island as potential probe arms because they already fulfill the basic criterion for uniqueness. Once seeds are chosen as potential probe arms, they can then be elongated (never shortened) to satisfy the Tm requirements of each arm. Therefore, any seed selected, and thus any probe arm arising from a seed, has been pre-screened for overall uniqueness.

Uniqueness is theoretically most important at the ends of the probe arms to avoid mis-priming. So, when additional bases from the reference sequence are added to seeds to allow them to grow to become probe arms with the correct Tm, the initial seed is always kept towards the 5' or 3' ends of the final probe (see Figure 5.2C). Elongating a seed to make it into a finished probe arm can only make it *more* unique than it already is. The end result is a set of probes that cover the target regions as optimally as possible given the starting seeds (see Figure 5.1I)

Phase 4 – target strand optimization

The final decision as to which strand is targeted by each probe is safely deferred until after the probes have been designed for optimal coverage in Phase 3. It is possible to just target one strand with all of the probes, and in some projects, that may be desirable. Another project may call for alternating strand targeting from one probe to the next and picking which of the two phases would be optimal across all of the probes in the set. The standard approach we have taken is to pick the optimal strand for each probe independently of all others. The decision is based on which target strand best satisfies the

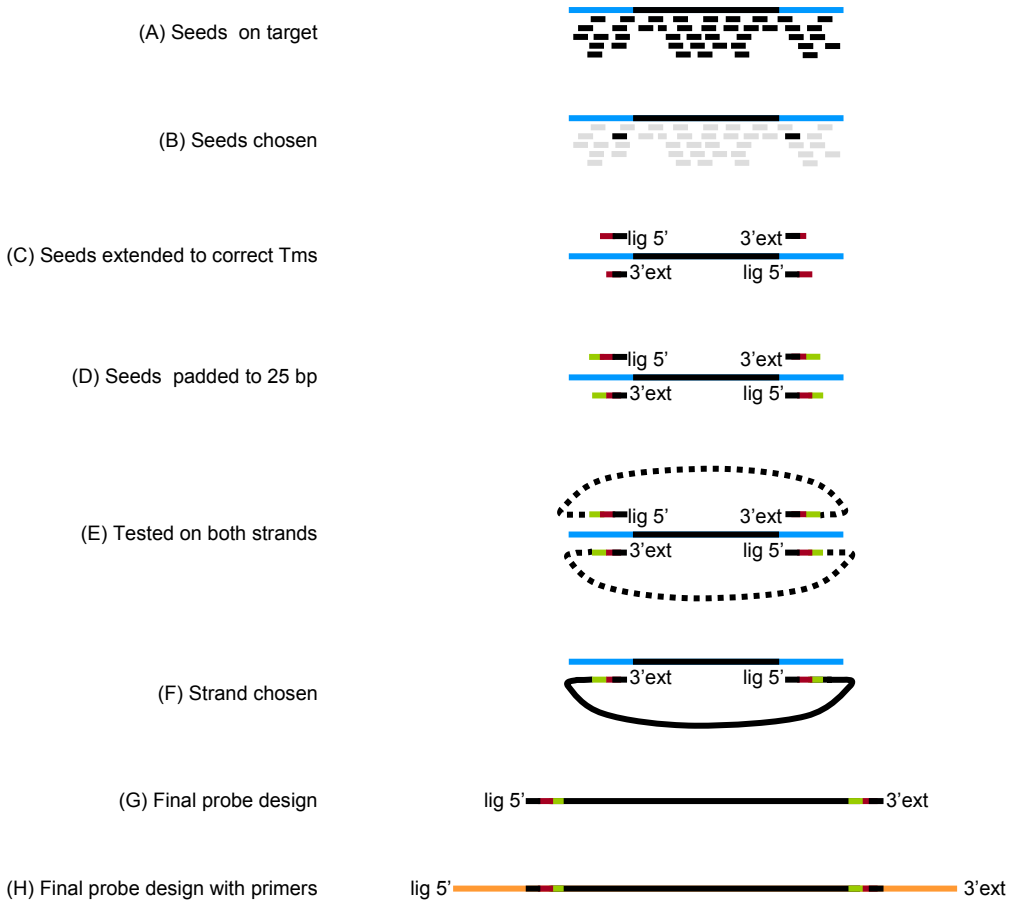


Figure 5.2: Probe design detail

optimal Tm of the arms unless only one strand is able to avoid a thymidine at the first position of the ligation arm in which case that strand is always chosen.

Flipping from the Watson to the Crick strand exchanges the role of each arm from that of extension to ligation and vice versa (see Figure 5.2C-E). Because the target Tm of the extension and ligation arms are usually different by design, flipping the target strand will also generally result in a different Tm for the arms and usually a different arm length for that Tm. Strand flipping is illustrated in Figures 5.1J and 5.2E-F.

We wanted to allow for separation of Phase 3 (coverage optimization) from Phase 4 (target strand optimization). So, prior to their use in Phase 3, each seed is pre-tested for its ability to become an extension or ligation arm without hitting a SNP. Because of this pre-test, the final strand determination can be deferred until the target strand optimization phase without invalidating the best coverage design generated in Phase 3.

Phase 5 – arm padding and backbone assembly

Although each probe arm is the result of seed elongation with reference sequence nucleotides, it is usually desirable for probe arms to be uniform in length. Therefore, probe arms are extended even further with non-complementary sequences in order to extend their length without increasing their melting temperature. The final probes include padding of each arm to 25 bp (which is user configurable) followed by insertion of the fixed MIP backbone between the two probe arms (see Figures 5.2F-G). The backbone, which can be designed per project, typically contains universal primer sequences for the inverted PCR amplification that occurs on successfully captured sequences as previously described[13]. Finally, outside primer sequences can be added to the ends of all of the

probes if amplification is desired or necessary before they are used as probes. In that case, users would follow protocols for MIP amplification including cleavage of these arms post-amplification and prior to use for capture (see Figures 5.1L and 5.2H and previously described protocols).

Probe design output

The final output from the pipeline is a file containing all probe IDs and sequences ready to be ordered for synthesis (Figure 5.1L).

Additional output files include BED format files of the probe target sequences for verification and visualization using the UCSC browser[14, 15] (See Figures 5.1K and Figure 5.3). Statistics on probe coverage and numbers are also part of the standard design output.

The role of uniqueness

When we first started the second generation probe design, we made the assumption that uniqueness of each probe arm would be an important factor in the design of a new probe set, and we proceeded with the assumption that probe arms would need to be very unique and also fit the melting temperature constraints. However, while we were developing these new algorithms, two different projects using MIPs within our lab demonstrated very little correlation between individual probe arm uniqueness and sequencing coverage (Li et al., unpublished). The reasons for this counterintuitive finding are not completely clear at this time.

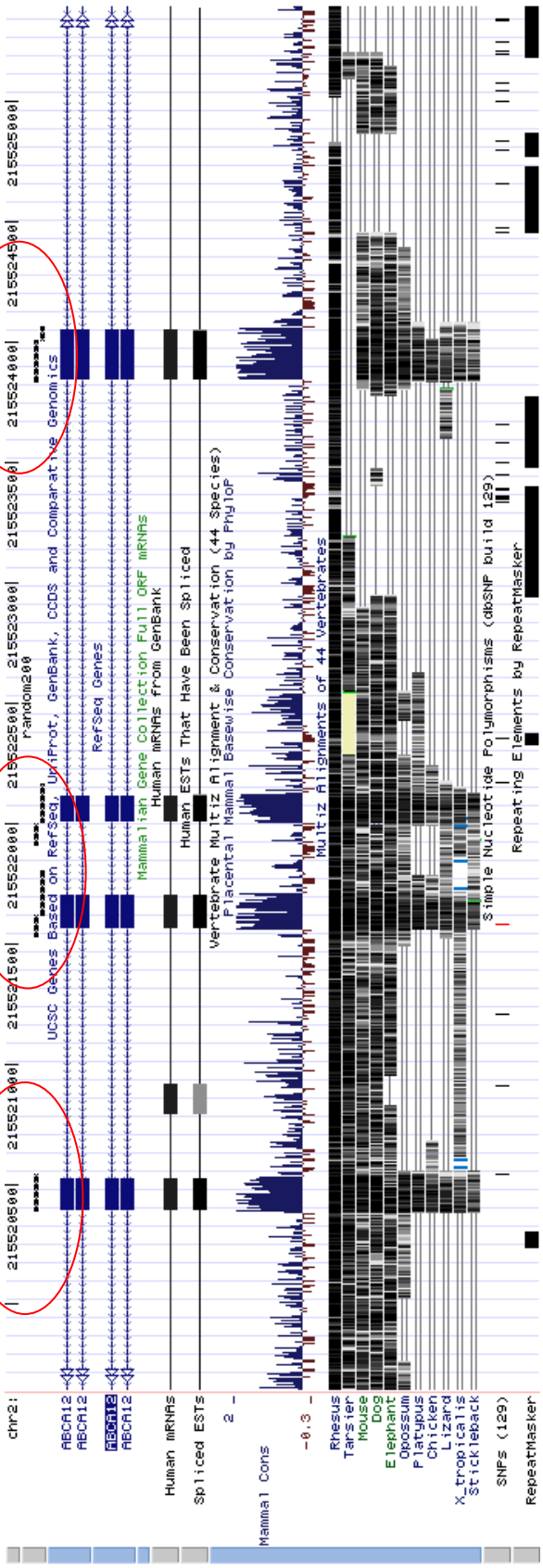


Figure 5.3: An example of BED file output used to visualize probe locations in the UCSC browser.

Despite not seeing the expected enrichment in sequence coverage resulting from highly unique probe arms, we also saw no systematic biochemical penalty for using unique probe arms. However, at the design stage, when we first used an absolute uniqueness criteria for the probe arms (i.e. exactly one occurrence allowed), we did pay a significant coverage penalty. Such a strict criterion often resulted in long stretches of over 200 bp (the maximum we chose for a target capture) with no seeds whatsoever. Thus we could achieve only approximately 50% target coverage.

Although we presently find no significant correlation between sequence capture efficiency and arm uniqueness, we still use seed uniqueness as a weak criterion for probe design by relaxing it enough to allow for seeds below 32 genome-wide occurrences. This filter allows all but extremely repetitive sequences through to Phase 3, and many targets can be completely tiled with probes under this cutoff level. Figure 5.4 shows the average uniqueness across the Human X chromosome. This example shows that allowing seeds occurring even less than 15 times in the genome allows high coverage. Allowing up to 32 repeats adds more flexibility in probe design, but we still ultimately enforce a pair-wise uniqueness constraint that only allows probe arm sequences to be chosen that are close to each other only at the intended target and not elsewhere in the genome.

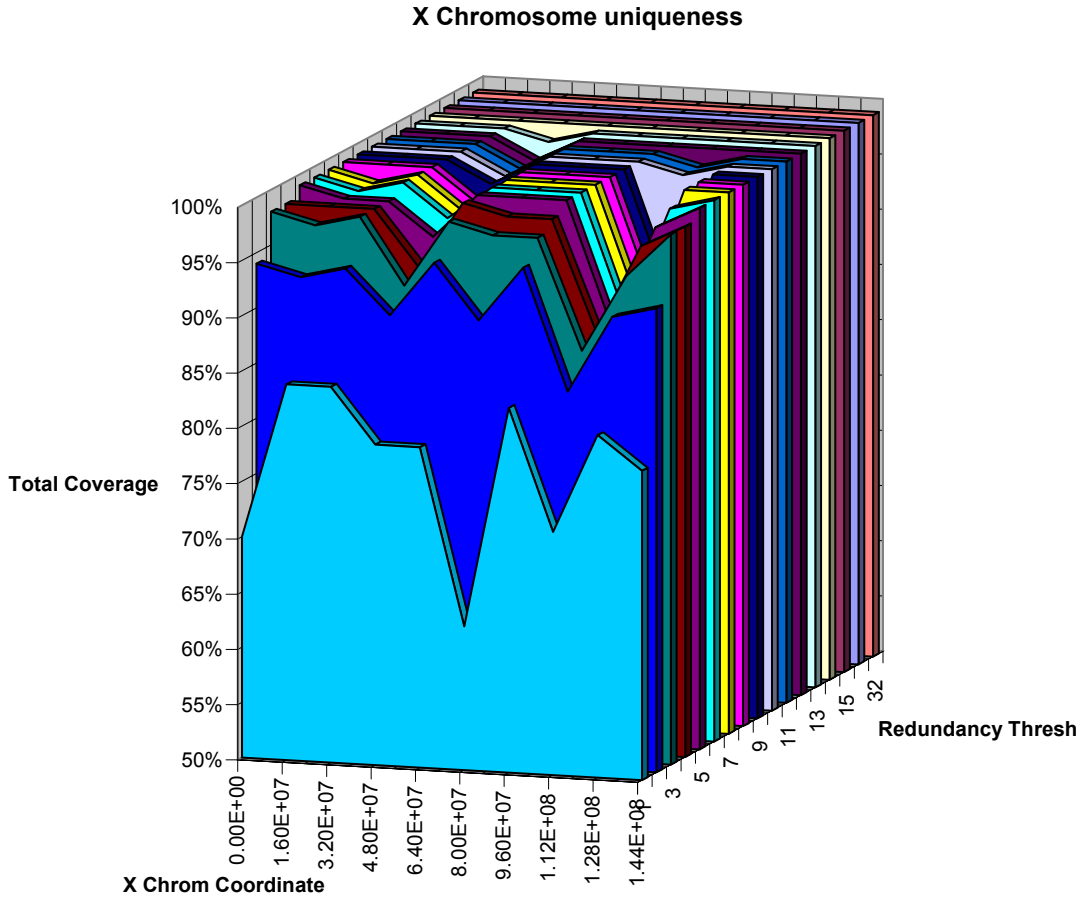


Figure 5.4: Uniqueness across the X Chromosome. The Y axis is how many sites have exactly N occurrences where N is from 1 to 32 on the Z axis. The X axis is chromosome X position. Note that at a redundancy of 11 or greater, almost all sequences are available to be seeded. Also note that the pseudo autosomal regions are at the beginning and end of the chromosome which explains the dips at the ends with an occurrence threshold of 1 because an exact copy exists in the Y chromosome reference sequence in these regions.

Per probe target capture size range

Per probe target capture size range is one of the parameters that may vary depending on the read length available on the intended sequencing platform or other protocol steps such as whether to perform random shotgun shearing or not. Generally speaking, the larger this target size range, the more flexibility the design pipeline will have, and the greater the total coverage will be.

Example projects

To illustrate the flexibility and performance of the design approach, we describe three different example projects along with their performance.

200 Random Genes. This scenario is similar to one in which one is interested in sequencing genes in a particular pathway or set of pathways with known or hypothesized relevance to a particular disease. Once a set of MIP probes is obtained, typically, sequencing would be performed on both cases and controls. To create this example, 30,600 RefSeq gene names were downloaded from the UCSC browser[14, 15], and 200 gene names were chosen at random. We then again used the browser to obtain all coding exons for these genes, and padded them by 5bp on either side in order to also capture any splicing signal variations.

The CFTR region. This example is similar to what one would encounter with a single large linkage peak and perhaps no candidate gene hypotheses available to narrow the search in the region. In the past it has been prohibitively expensive to simply sequence

the whole region. For this case, we chose the Encode ENm001 CFTR 1.9 Mb region as an example. Using a single pair of coordinates (chr7:115597756-117475182) as input to the pipeline, we were able to obtain probes that tile across the entire region. To demonstrate the ability of MIPTAG Pro to generate probes even in repetitive sequence regions, we ran it on two different instances of this genomic region. First we ran it on sequences containing SNPs that were filtered with RepeatMasker[16] which effectively removed all seeds from masked regions. Next we ran it on the same sequences containing SNPs that were not filtered with RepeatMasker.

All coding exons. As mentioned, the Personal Genome Project has begun sequencing the first subjects, and is initially targeting the protein coding exons in the entire genome. For this example, we downloaded all coding exons available from RefSeq[17], USCS known genes[14, 15] and the CCDS consortium (http://www.ncbi.nlm.nih.gov/projects/CCDS). After redundancy removal, this resulted in 196,616; 222,965 and 166,046 non-overlapping exons respectively. We padded each exon with 5bp on either side, and Phase 1 of the pipeline removed all redundant sequences to result in a final set of 225,831 padded exon targets.

Design results

We ran MIPTAG Pro on the example data sets with different constraints on allowable target capture sizes. For the 200 random gene set and the Encode region, we used 60-100 and 60-200 bp windows. For the All coding exons example, we used a 125-

165 bp window. The results and statistics for each example design set are summarized in Table 5.1.

Depending on the target, the 60-100 bp window runs performed slightly worse than those with larger windows, and the all exon example had the highest overall coverage at 93% (row G). However, this high coverage % is also the most inefficient (row H). This can be due to the fact that exon size is highly variable, and every time it falls below 125 bp, non-target sequences must be incorporated. This is reflected in row L which shows the average non-target sequence in each amplicon.

These results also show the effect of using repeat masked sequences on performance (compare Encode regions with and without RepeatMasker – note all other examples were run without RepeatMasker). Row F shows the best case coverage given the seeds before probe design takes place, and it is evident that the removal of seeds in repeat masked regions prevents the algorithm from covering the entire region. This also suggests that only about 57% of the region would be considered non-repetitive. However, repeat masking is likely too conservative for probe design as evidenced by the ability of MIPTAG Pro to achieve 89% coverage using 60-200 bp amplicons without repeat masking.

Finally, this table shows the diversity of input regions that MIPTAG Pro can use for design ranging from a single 1.9 Mb region up to 225,831 exons with an average size of 164 bp. While some examples demonstrate a significant amount of overhead (row H), this is mostly a function of the user constraints and selection of target regions because a single large region has 0% overhead, and thus is not a reflection of the algorithm as much as the inherent fragmentation given a fixed target window and targets of varying length.

Table 5.1: Target coverage results for a number of example projects under different target amplicons target constraint sizes, and the inclusion or exclusion of repeat sequences

input	Example Project:	200 random gene coding exons		Encode CFTR Region using RepeatMasker		Encode CFTR Region **not** using RepeatMasker		All coding exons
	Target Amplicon Range allowed:	60-100 bp	60-200 bp	60-100 bp	60-200 bp	60-100 bp	60-200 bp	125-165 bp
A	Total target bp	402,193	402,193	1,877,426	1,877,426	1,877,426	1,877,426	37,107,584
B	Number of input targets	2,232 exons	2,232 exons	1 region	1 region	1 region	1 region	225,831 exons
C	Average target bp (A/B)	180	180	1,877,426	1,877,426	1,877,426	1,877,426	164
D	Number of target islands	1,753	1,753	1	1	1	1	159,999
E	Average target bp per island (A/D)	229	229	1,877,426	1,877,426	1,877,426	1,877,426	232
F	Maximal theoretical coverage given the seeds	402,193 (100%)	402,193 (100%)	1,085,968 (57%)	1,166,774 (62%)	1,772,104 (94%)	1,826,621 (97%)	37,058,286 (99%)
G	Total target bp covered	354,057 (88%)	363,961 (90%)	862,573 (45%)	1,030,801 (54%)	1,422,708 (75%)	1,677,403 (89%)	34,707,171 (93%)
H	Total non-target bp covered	87,000 (22% overhead)	105,072 (26% overhead)	0 (0% overhead)	29 (0% overhead)	1 (0% overhead)	30 (0% overhead)	15,111,697 (41% overhead)
I	target islands 100% covered	78.9% (1,384 / 1,753)	88.7% (1,555 / 1,753)	0% (0 / 1)	0% (0 / 1)	0% (0 / 1)	0% (0 / 1)	89% (143,009 / 159,999)
J	target islands 1-99% covered	14.6% (256 / 1,753)	5.5% (97 / 1,753)	100% (1 / 1)	100% (1 / 1)	100% (1 / 1)	100% (1 / 1)	7% (11,512 / 159,999)
K	target islands 0% covered	6.4% (113 / 1,753)	5.7% (101 / 1,753)	0% (0 / 1)	0% (0 / 1)	0% (0 / 1)	0% (0 / 1)	3% (5,478 / 159,999)
L	Average Amplicon length (targ + non-targ)	87 (70 + 17)	146 (113 + 33)	91 (91 + 0)	158 (158 + 0)	91 (91 + 0)	163 (163 + 0)	145 (101+ 44)
M	Total MIPTAGS probes	5,178	3,259	9,987	6,947	16,562	11,079	348,951

Reasons for non-ideal probe efficiency

The human genome is filled with repetitive elements which plague all sequencing projects[18]. These problems are somewhat exacerbated by limitations of read length shared by most current HTS platforms. Generally, departures from an idealized design scenario are caused (a) by the unavailability of seeds where they are needed, and (b) by design rule constraints on allowable sequences. Both of these factors are ultimately a function of the particular target sequence and of the genomic background.

- The availability of seeds depends on the uniqueness of the region and the absence of SNPs at a potential seed location (illustrated by the empty seed region on the left side of Figure 5.1H). It is therefore possible that the best coverage is less than 100% simply because there are no available seeds flanking the desired target sequence. Alternatively, the final probe may need to capture non-target sequence along with target sequence because it may have been necessary to pick seeds displaced from the target sequence.
- Design rules evaluate sequence characteristics such as Tm for probe arms and GC content for target amplicons. Thus the probe design pipeline will sometimes create probes that capture non-target sequence around target sequences, and sometimes it will create probes that do not completely cover the target sequence.

MATERIALS AND METHODS

Software and hardware implementation. The entire probe design pipeline is written in a combination of Perl for rapid development and C for rapid execution of the core

algorithms. The pipeline is designed to take advantage of the distributed parallel architecture of the 176 node Linux-based computing cluster managed by Harvard Medical School Research Information Technology Group. Each node of the computing cluster has an Intel Xeon CPU with speeds ranging from 2.3 – 3.6 GHz and from 4 GB to 32 GB of RAM.

Software Architecture. Due to the computationally intensive nature of the probe design process, the process is divided amongst a suite of programs (see Supplementary Figure 5.2) which roughly correspond to the conceptual phases shown in Figures 5.1 and 5.2.

N-mer seed tallying pass. Each N-mer in the genome and its reverse complement is tallied using a sliding N-base window along the genome one base pair at a time. When a SNP is encountered in the reference sequence, alternative versions of each N-mer is tallied as a potential occurrence. For example, the 12-mer containing a SNP at its sixth position ACGTA**Y**GTACGT will be counted twice – once as ACGTA**C**GTACGT and once as ACGTA**T**GTACGT.

Normally, a single base would be tallied in 2N different N-mer windows (N on the Watson strand, and N on the Crick strand); however, each SNP has the potential to be counted at least two times (more for tri- or tetra-allelic SNPs) for each position in the window. So all of these numbers would be doubled or more depending on the number of alleles and how many SNPs occur per window. Although seeds will not ultimately be placed at locations containing a SNP, each SNP and its location will still count towards the total possible genomic background.

The seed picking algorithm requires the use of compute nodes with 32 GB of RAM. Seed lengths over 14 bp required too much memory to process in just one pass, so these longer length seeds are broken into multiple passes with 14 bp seeds requiring 4 passes, and 15 bp seeds requiring 16 passes each. (Overall, the seed picking process for the human genome takes approximately 24 CPU hours to run across several nodes almost independent of target size. The genomic frequencies could also be pre-computed once and stored but we have opted for increased computational overhead over increased storage space.)

Backbone sequence screening. To avoid any hybridization with other probes in the reaction, before use in Phase 3 (coverage optimization), seeds are completely eliminated if they are identical to any forward or reverse complement sequence within the intended probe backbone.

Seed placement. Seed placement occurs across an entire target island without regard to whether it covers a target or non-target sequence. Seeds are not placed if it would include a SNP or a repeat masked region. The final set of seeds is based solely upon their occurrence in the genome being less than 32. (This number was chosen for practical memory utilization reasons as well as empirical reasons because it covered most regions adequately as shown in Figure 5.4)

Probe design to optimize coverage is deployed across the compute cluster with each target island assigned to a single processor. This phase is computationally but not memory intensive, and uses heuristics to find a reasonably optimal solution.

Probe design constraints. Probe design begins with seeds and target island sequences (with indications of target and non-target sequence, repeat mask information (if used) and SNPs).

In order to obtain the best set of probes for an island, the probe design program can evaluate millions of alternative designs to select the set of probes with the highest **target** sequence coverage, the lowest **non-target** sequence coverage and the minimal total seed size in that order. For each island, a heuristic approach is used that is not guaranteed to be globally optimal, however, any design it generates is guaranteed to satisfy all of the following constraints. (These optimization criteria yield the most economical use of probe sequences by reducing redundancy; however, we can alternatively maximize *overlapping* **target** coverage.):

- minimum and maximum probe target size (The minimum is typically set to be 60 bp, and the maximum is typically set to be 100 to 200 bp depending on the ultimate sequencing platform and pre-sequencing protocol. Narrow ranges tend to decrease coverage.)
- minimum, maximum and optimal Tm of the ligation arm
- minimum, maximum and optimal Tm of the extension arm
- maximum probe arm length (typically 25 bp)
- minimum probe arm length (typically 12 bp)
- minimum and maximum GC optimal content for a probe target (typically 30-70%), and an optimal target probe length if these boundaries are exceeded

- minimum probe arm distance elsewhere in the genome. (This constraint assures pair-wise arm uniqueness in the genome so that other potential targets must not be smaller than this distance elsewhere in the genome. We know from previous experiments that the capture of very large target regions (over 500 bp) by MIPs is near zero. Setting this value to 1000 reduces non-specific capture of off target sequences to near zero.)
- a rule to attempt to avoid a thymidine nucleotide in the ligation arm directly adjacent to the target.

Many of these constraints such as the non-thymidine ligation base rule are the direct result of our experience with first generation probe designs. Variables indicated as *optimal* are not hard constraints but the program attempts to achieve these values. All of these variables are easily changed depending upon project goals.

Determination of the number of probes per target. The algorithm starts with one probe and tries designs with successively greater numbers of probes until the coverage plateaus with no further benefit achieved by using more probes. Probe design can fall into a number of scenarios that depend on the length and characteristics of the target sequence. For instance, if a target is only 100 bp, then the ideal scenario is usually one wherein a single probe (two arms) can be used to cover 100% of the entire target with 0% non-target sequence (see Supplementary Figure 5.1a). That is, the ends of the probe arms are placed at exactly 1 bp outside of the target sequence on either side, and the target is exactly filled in between the probe arms.

Long target sequences may require more than one probe, and the program generates probes in a stereotypical interlocking topology in order to cover a given region. For example, a 300 bp target cannot be covered by a single probe if the maximum target length is set to 200 bp. However, two probes (four arms) can cover this region in an interlocked manner as diagrammed in Supplementary Figure 5.1b. In order to capture all of the target sequence, one of the arms of each probe must be placed within the target region of the other.

There is an additional implicit constraint on the multi-probe design that no probe arm can directly overlap another probe arm. This is to reduce the chance of probe-probe hybridization which would compete with the intended target hybridization.

After trying many possible combinations of one probe (two arms), the program will try many (but not all) possible combinations of two probes (four arms) to achieve the most coverage with the least non-target coverage. The program will keep trying designs with more and more probes until it cannot do better than a design with fewer probes. Ultimately, it will keep the design with the fewest probes everything else being equal. A design requiring 3 or more probes has the stereotypical topology that is illustrated in Supplementary Figure 5.1c. While exons rarely require more than one probe, through this mechanism, the program can easily target large contiguous regions – even megabases in length.

Converting seeds into probe arms. All probe arms begin as **seeds** that are grown in the direction away from their capture sequence. This is the 5' direction if they become extension arms or the 3' direction if they are to become ligation arms. Before coverage

maximization, it is not necessarily clear which role a seed will ultimately play. In one potential design, a seed might end up being the start of an extension arm, but in an alternative design, that same seed might end up being the start of a ligation arm – even on the same strand. In order to calculate the correct length given a Tm, a seed is grown until it is as close to the ideal target Tm as possible and still within the minimum and maximum allowable Tms for that type of arm.

Seed pre-screening with Tm contraints. In order to reduce the complexity of the inner loop of the probe design algorithm, as a first pass, all seeds are pre-filtered to eliminate any seeds that would fail the Tm or length constrains of ligation or extension arms if they were to become part of a probe. In addition, should a seed (when grown into a probe arm) run into a SNP before achieving acceptable Tm, it is also eliminated. While a seed will ultimately only be used as either an extension or a ligation arm, requiring that it can potentially be either one at this stage simplifies the design process with a minimal impact on ultimate coverage.

Strand determination. Each probe is designed to optimize coverage while being flexible enough to flip between strands if required. This is because each arm has already been checked (as just described) for its ability to perform either as an extension arm (when it is on the strand that places it upstream from the target) or as a ligation arm (when it is on the strand that places it downstream from the target) (see Figures 5.2C-E). Generally speaking, strand determination is made by choosing the strand which allows the probe to have the best overall Tm for both arms (smallest overall departure from target Tms), and

flipping strands will, in general, change the probe arm lengths before padding (see Figure 5.2F). Strand determination can be performed independently for each probe, or collectively for a group of probes that are interlocking. Alternatively, strand determination can be forced to be one or the other. If a thymidine exists at the end of the ligation arm of only one strand's design, then the other strand is chosen regardless of which strand has a better optimal Tm.

Padding probe arms to uniform length. In order to make probes of uniform length for ease of synthesis and purification, padding may be added to the arms in a manner that will not alter their Tm. If the padded probe arm length is 25 bp, then for example, a probe arm that is 18 bp long at the right Tm, must have 7 bp added to the end not adjacent to the capture target to make it 25 bp long. Because random sequences can change from run to run, we systematically pad with the identical base from the opposite strand. Thus a G will be across from a G, an A across from an A, etc. SNPs *are* permitted in the padding region, and when possible, a base is chosen that will be non-complementary to either of the 2 possible nucleotides. When more than one nucleotide is possible, then a base is arbitrarily (but consistently) chosen that will be most antagonistic to the most alternative bases at that SNP. A perfect solution is not possible in all cases, but it will be deterministic.

Melting temperature calculation and conditions. The theoretical melting temperature of each probe arm is determined according to a C language version of a melting temperature calculator developed and kindly provided by Kun Zhang. The calculator incorporates the

formulas of SantaLucia J et al[19] which have been adjusted for MgCl2 and DMSO concentration based on von Ahsen et al.[20]. Based upon reaction conditions for our protocols, the default reagent concentrations used for these calculations is as follows: 100 pM primer concentration, 10 mM Magnesium, 25 mM monovalent cations (Na+ plus K+), 10 uM dNTPs, 0% DMSO, and a target of 50% complementary strand annealing.

Reference sequences and databases. The reference sequences are typically downloaded from NCBI or UCSC FTP sites. If available, versions of these genomes with annotated SNPs are used as a basis for probe design.

Probe synthesis. We have used Agilent as a source of array synthesized probes in our experiments to date, but other sources are possible including high density multi-well plates.

DISCUSSION

Genetics has a long history of using targeted sequencing of candidate genes and genomic loci to determine genetic causes of disease and phenotypes but it hasn't always been easy. We have developed a computational pipeline for the automatic design of molecular inversion probes for the capture of an arbitrary genomic target sequence or set of sequences. The ability to inexpensively capture and sequence targeted genomic regions in a large number of individuals eliminates a huge barrier to the rapid advancement of genetic and genomic experiments. With this technology, it is possible to not only sequence most exons of a genome efficiently; it is also possible to economically

sequence subsets of candidate genes and pathways for particular diseases in a large population.

Gene linkage analysis studies with inadequate numbers of families or inadequate family sizes have traditionally had to rely upon long manual sequencing efforts of candidate genes (or worse several megabases) under a broad linkage peak. It is now feasible to rapidly design and sequence – in one reaction – most genes or bases in such regions in a fraction of the time it would take to do so manually. We believe that using this approach, sequencing large or multiple regions of targeted genomic DNA will soon become routine.

The completely automated probe design pipeline takes as input a set of target sequences and we have given examples of a number of different project types and their performance in our pipeline. Currently, we are applying this design pipeline to all of these project types and more.

Although we have improved upon the first generation of probe designs by incorporating uniqueness, melting temperature, GC content and the ability to target large sequences, there are still many factors that affect capture and sequencing efficiency that are not well understood, causing MIP capture sensitivity to vary significantly by target.

The results presented here are without regard to other experimental variables introduced in sample handling and preparation. As we discover evidence of more factors that influence probe performance we plan to continue to incorporate these rules into our design methodology. It may be that some of the target regions will better lend themselves to other targeted sequencing approaches, but that is yet to be determined.

As sequencing continues to become even more economical, we believe that rapid targeted sequence capture will continue to enable an ever greater number of experiments and will be extremely useful for many years to come.

ACKNOWLEDGMENTS

The authors thank Eric Boyden for useful discussions regarding coverage requirements, and Kun Zhang for providing his melting temperature calculator written in the Perl programming language that we were able to port to the C programming language. M.F.C. and G.M.C. are funded by the Genomes To Life (GTL) Microbial Ecology, Proteogenomics and Computational Optima from the U.S. Department of Energy; J.B.L, A.M.R and G.M.C. are funded by the Center of Excellence in Genomic Science(CEGS) from the U.S. National Institute of Health-NHGRI; and A.M.R. is also funded by a Harvard Medical School, Genetics Department Seed Grant.

REFERENCES

1. Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376-80 (2005).

2. Shendure, J. et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309, 1728-32 (2005).

3. Bentley, D. R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53-9 (2008).

4. Kim, J. B. et al. Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. Science 316, 1481-4 (2007).

5. Chee, M. et al. Accessing genetic information with high-density DNA arrays. Science 274, 610-4 (1996).

6. Ruby, J. G. et al. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. Cell 127, 1193-207 (2006).

7. Wheeler, D. A. et al. The complete genome of an individual by massively parallel DNA sequencing. Nature 452, 872-6 (2008).

8. Levy, S. et al. The diploid genome sequence of an individual human. PLoS Biol 5, e254 (2007).

9. Wang, J. et al. The diploid genome sequence of an Asian individual. Nature 456, 60-5 (2008).

10. Li, G. et al. The YH database: the first Asian diploid genome database. Nucleic Acids Res 37, D1025-8 (2009).

11. Church, G. M. The personal genome project. Mol Syst Biol 1, 2005 0030 (2005).

12. Gnirke, A. et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 27, 182-9 (2009).

13. Porreca, G. J. et al. Multiplex amplification of large sets of human exons. Nat Methods 4, 931-6 (2007).

14. Kuhn, R. M. et al. The UCSC Genome Browser Database: update 2009. Nucleic Acids Res 37, D755-61 (2009).

15. Kent, W. J. et al. The human genome browser at UCSC. Genome Res 12, 996-1006 (2002).

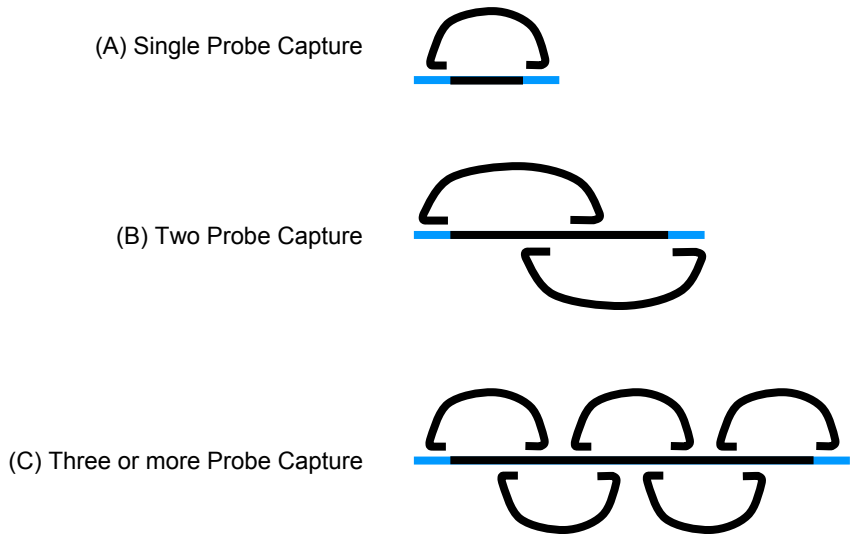
16. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker. unpublished data.

17. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35, D61-5 (2007).

18. Kent, W. J. & Haussler, D. Assembly of the working draft of the human genome with GigAssembler. Genome Res 11, 1541-8 (2001).

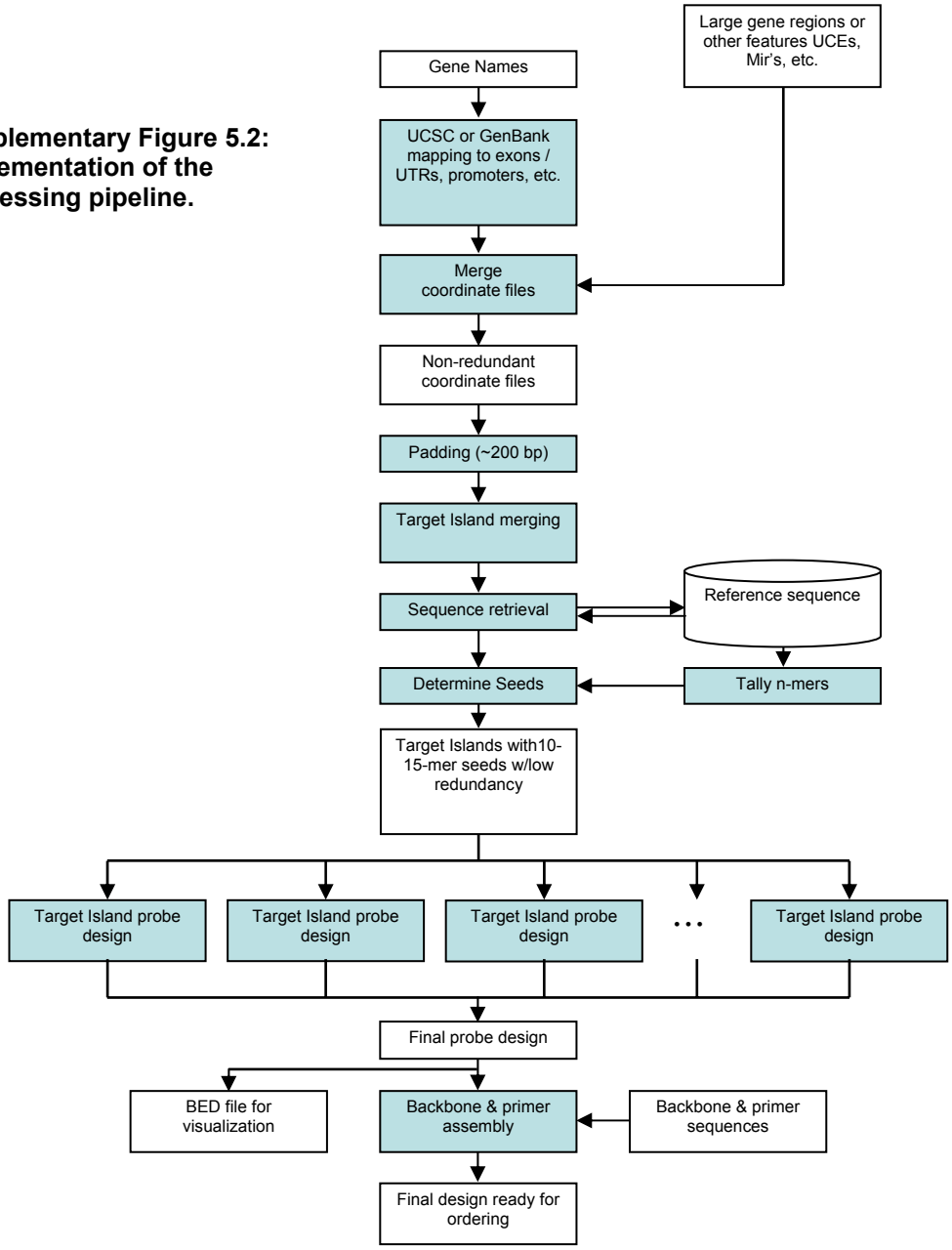
19. SantaLucia, J., Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc Natl Acad Sci U S A 95, 1460-5 (1998).

20. von Ahsen, N., Wittwer, C. T. & Schutz, E. Oligonucleotide melting temperatures under PCR conditions: nearest-neighbor corrections for Mg(2+), deoxynucleotide triphosphate, and dimethyl sulfoxide concentrations with comparison to alternative empirical formulas. Clin Chem 47, 1956-61 (2001).



Supplementary Figure 5.1: Stereotypical probe capture patterns for targets of various sizes. Shown here are interlocking probe patterns required to prevent gaps in sequence coverage. A single probe is shown in (a); two probes with only one arm landing in the middle of the sequence of the other is shown in (b); and three or more probes with two arms falling within other probes except at the ends is shown in (c). Note, the alternating strands are shown for illustrative clarity only, and is not meant to imply that probes must be on alternating strands.

Supplementary Figure 5.2: Implementation of the processing pipeline.



Appendix G

Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays

This work was originally published as:

Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., Carnevali, P., Nazarenko, I., Nilsen, G. B., Yeung, G., Dahl, F., Fernandez, A., Staker, B., Pant, K. P., Baccash, J., Borcherding, A. P., Brownley, A., Cedeno, R., Chen, L., Chernikoff, D., Cheung, A., Chirita, R., Curson, B., Ebert, J. C., Hacker, C. R., Hartlage, R., Hauser, B., Huang, S., Jiang, Y., Karpinchyk, V., Koenig, M., Kong, C., Landers, T., Le, C., Liu, J., McBride, C. E., Morenzoni, M., Morey, R. E., Mutch, K., Perazich, H., Perry, K., Peters, B. A., Peterson, J., Pethiyagoda, C. L., Pothuraju, K., Richter, C., Rosenbaum, A. M., Roy, S., Shafto, J., Sharanhovich, U., Shannon, K. W., Sheppy, C. G., Sun, M., Thakuria, J. V., Tran, A., Vu, D., Zaranek, A. W., Wu, X., Drmanac, S., Oliphant, A. R., Banyai, W. C., Martin, B., Ballinger, D. G., Church, G. M. & Reid, C. A. 2009. Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science*.

Pertinent excerpts from the supplementary material are included. The complete Supplementary Online Material can be found at: http://www.sciencemag.org/cgi/data/1181498/DC1/1

Author Contributions: A.M.R. prepared the published list of variants for analysis by the Trait-o-matic software program, and analyzed the results with A.W.Z., J.T., X.W. and G.M.C.. A.M.R. performed confirmatory Sanger sequencing on all variants prioritized using the algorithm outlined in Chapter 2.

Sciencexpress Report

Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays

Radoje Drmanac,[1]* Andrew B. Sparks,[1†] Matthew J. Callow,[1†] Aaron L. Halpern,[1†] Norman L. Burns,[1†] Bahram G. Kermani,[1†] Paolo Carnevali,[1†] Igor Nazarenko,[1†] Geoffrey B. Nilsen,[1†] George Yeung,[1†] Fredrik Dahl,[1†‡] Andres Fernandez,[1†] Bryan Staker,[1†] Krishna P. Pant,[1†] Jonathan Baccash,[1] Adam P. Borcherding,[1] Anushka Brownley,[1] Ryan Cedeno,[1] Linsu Chen,[1] Dan Chernikoff,[1] Alex Cheung,[1] Razvan Chirita,[1] Benjamin Curson,[1] Jessica C. Ebert,[1] Coleen R. Hacker,[1] Robert Hartlage,[1] Brian Hauser,[1] Steve Huang,[1] Yuan Jiang,[1] Vitali Karpinchyk,[1] Mark Koenig,[1] Calvin Kong,[1] Tom Landers,[1] Catherine Le,[1] Jia Liu,[1] Celeste E. McBride,[1] Matt Morenzoni,[1] Robert E. Morey,[1§] Karl Mutch,[1] Helena Perazich,[1] Kimberly Perry,[1] Brock A. Peters,[1] Joe Peterson,[1] Charit L. Pethiyagoda,[1] Kaliprasad Pothuraju,[1] Claudia Richter,[1] Abraham M. Rosenbaum,[2] Shaunak Roy,[1] Jay Shafto,[1] Uladzislau Sharanhovich,[1] Karen W. Shannon,[1‖] Conrad G. Sheppy,[1] Michel Sun,[1] Joseph V. Thakuria,[2] Anne Tran,[1] Dylan Vu,[1] Alexander Wait Zaranek,[2] Xiaodi Wu,[3] Snezana Drmanac,[1] Arnold R. Oliphant,[1] William C. Banyai,[1] Bruce Martin,[1] Dennis G. Ballinger,[1]* George M. Church,[2] Clifford A. Reid[1]

[1]Complete Genomics, Inc., 2071 Stierlin Court, Mountain View, CA 94043, USA. [2]Department of Genetics, Harvard Medical School, Cambridge, MA, USA. [3]School of Medicine, Washington University, St. Louis, St. Louis, MO, USA.

*To whom correspondence should be addressed. E-mail: rdrmanac@completegenomics.com (R.D.); dballinger@completegenomics.com (D.G.B.)

†These authors contributed equally to this work.

‡Present address: Ion Torrent Systems, San Francisco, CA, USA.

§Present address: San Diego State University, San Diego, CA, USA.

‖Present address: Life Technologies, Carlsbad, CA, USA.

Genome sequencing of large numbers of individuals promises to advance the understanding, treatment, and prevention of human diseases, among other applications. We describe a genome sequencing platform that achieves efficient imaging and low reagent consumption with combinatorial probe anchor ligation (cPAL) chemistry to independently assay each base from patterned nanoarrays of self-assembling DNA nanoballs (DNBs). We sequenced three human genomes with this platform, generating an average of 45- to 87-fold coverage per genome and identifying 3.2 to 4.5 million sequence variants per genome. Validation of one genome data set demonstrates a sequence accuracy of about 1 false variant per 100 kilobases. The high accuracy, affordable cost of $4,400 for sequencing consumables and scalability of this platform enable complete human genome sequencing for the detection of rare variants in large-scale genetic studies.

Genotyping technologies have enabled the routine assessment of common genetic variants at up to a million sites across the genome in thousands of individuals (*1*) and have increased our understanding of human genetic diversity and its biological and medical impact. Whole-genome sequencing costs have dropped from the >$100 M cost of the first human genomes (*2*, *3*) to the point where individual labs have generated genome sequences in a matter of months for material costs of as low as $48k (*4*–*12*) (table S5). Sequencing technologies, which use a variety of genomic microarray construction methodologies and sequencing chemistries (*13*–*32*), can determine human genetic diversity over an entire genome and identify common as well as rare SNPs, insertions and deletions. Despite these advances, improvements are still needed to enable the cost-effective characterization of the many hundreds of genomes required for complex disease genetic studies and for personalized disease prevention, prognosis and treatment.

We generated sequencing substrates (Fig. 1A and SOM) by means of genomic DNA fragmentation and recursive cutting with type IIS restriction enzymes and directional adaptor insertion (Fig. 1B and fig. S1). The resulting circles were then replicated with *Phi*29 polymerase (RCR) (*34*). Using a controlled, synchronized synthesis we obtained

hundreds of tandem copies of the sequencing substrate in palindrome-promoted coils of ssDNA, referred to as DNA nanoballs (DNBs) (Fig. 1C). DNBs were adsorbed onto photolithographically etched, surface modified (SOM) 25 x 75 mm silicon substrates with grid-patterned arrays of ~300nm spots for DNB binding (Fig. 1C). The use of patterned arrays increased DNA content per array and image information density relative to random genomic DNA arrays (*6*, *9*, *11*, *14*, 28). High-accuracy cPAL sequencing chemistry was then used to independently read up to 10 bases adjacent to each of eight anchor sites (Fig. 1D), resulting in a total of 31- to 35-base mate-paired reads (62 to 70 bases per DNB). cPAL is based on unchained hybridization and ligation technology (*15*, *27*, *28*, *31*), previously used to read 6-7 bases from each of four adaptor sites (26 total bases) (*28*), here extended using degenerate anchors to read up to 10 bases adjacent to each of the eight inserted adaptor sites (Fig. 1D, right) with similar accuracy at all read positions (fig. S3). This increased read length is essential for human genome sequencing.

Cell lines derived from two individuals previously characterized by the HapMap project (*33*), a Caucasian male of European decent (NA07022) and a Yoruban female (NA19240) were sequenced. NA12940 was selected to allow for a comparison of our sequence to the sequence of the same genome currently being assembled by 1000 genome project. In addition, lymphoblast DNA from a Personal Genome Project Caucasian male sample, PGP1 (NA20431) was sequenced because substantial data are available for biological comparisons (*35–37*). Automated cluster analysis of the four-dimensional intensity data produced raw base reads and associated raw base scores (SOM).

We mapped these sequence reads to the human genome reference assembly with a custom alignment algorithm that accommodates our read structure (fig. S4, SOM), resulting in between 124 and 241 Gb mapped and an overall genome coverage of 45- to 87-fold per genome.

To assess representational biases during circle construction we assayed genomic DNA and intermediate steps in the library construction process by quantitative PCR (QPCR) (fig. S2, SOM). This and mapped coverage showed a substantial deviation from Poisson expectation with excesses of both high and low coverage regions (fig. S5) but only a few percent of bases have coverage insufficient for assembly (Table 1). Much of this coverage bias is accounted for by local GC content in NA07022, a bias that was significantly reduced by improved adapter ligation and PCR conditions in NA19240 (fig. S5, SOM); the fraction of the genome with less than 15-fold coverage was accordingly reduced from 11% in NA07022 to 6.4% in NA19240 despite the latter having 25% less total coverage (Table 1).

Discordance with respect to the reference genome in uniquely mapping reads from NA07022 was 2.1% (range 1.4% – 3.3% per slide). However, considering only the highest scoring 85% of base calls reduced the raw read discordance to 0.47%, including about 0.1% of true variant positions.

Mapped reads were assembled into a best-fit, diploid sequence with a custom software suite employing both Bayesian and de Bruijn graph techniques (SOM). This process yielded diploid reference, variant or no-calls at each genomic location with associated variant quality scores. Confident diploid calls were made for 86 to 95% of the reference genome (Table 1), approaching the 98% that can be reconstructed in simulations. The 2% that is not reconstructed in simulations is composed of repeats that are longer than the ~400 base inserts used here and of high enough identity to prevent attribution of mappings to specific repeat copies. Longer mate-pair inserts minimize this limitation (*6*, *9*). Similar limitations affect other short read technologies.

We identified a range of 2.91 to 4.04 million SNPs with respect to the reference genome, 81 to 90% of which are reported in dbSNP, as well as short indels and block substitutions (Table 1 and table S6). Because of the use of local de novo assembly, indels were detected in sizes ranging up to 50 bp. As expected, indels in coding regions tend to occur in multiples of length 3, indicating the possible selection of minimally impacting variants in coding regions (fig. S6).

As an initial test of sequence accuracy, we compared our called SNPs with the HapMap phase I/II SNP genotypes reported for NA07022 (*1*). We fully called 94% of these positions with an overall concordance of 99.15% (Table 2) (the remaining 6% of positions were either half-called or not called). Furthermore; we fully called 96% of the Infinium (Illumina, San Diego, CA) subset of the HapMap SNPs with an overall concordance rate of 99.88%, reflecting the higher reported accuracy of these genotypes (*33*). Similar concordance rates with available SNP genotypes were observed in NA19240 (with a call rate of over 98%) and NA20431 (table S7).

We further characterized 134 of the 168 calls that were discordant with Infinium loci and Sanger sequencing of PCR products in NA07022, demonstrating that 55% of these discordances are errors in the reported HapMap genotypes (Table 2). The relationship between detection rate and read depth for about 1M Infinium HD SNPs that we subsequently genotyped in NA07022 shows that coverage of 25-fold at a position is sufficient to detect 91% of SNPs at heterozygous loci and 99% of SNPs at homozygous loci (fig. S5). Because the whole-genome false positive rate cannot be accurately estimated from known SNP loci, we tested a random subset of novel non-synonymous variants in

NA07022, a category that is enriched for errors (*10*). We extrapolated error rates from the targeted sequencing of 291 such loci, and estimated the false positive rate at about one variant per 100 kb, including <6.1 substitution-, <3.0 short deletion-, <3.9 short insertion- and <3.1 block-variants per Mb (Table 3 and table S8).

Aberrant mate-pair gaps may indicate the presence of length-altering structural variants and rearrangements with respect to the reference genome. A total of 2,126 clusters of such anomalous mate-pairs were identified in NA07022. We performed PCR-based confirmation of one such heterozygous 1,500-base deletion (fig. S7). More than half of the clusters are consistent in size with the addition or deletion of a single *Alu* repeat element.

Some applications of complete genome sequencing may benefit from maximal discovery rates, even at the cost of additional false-positives, while for others, a lower discovery rate and lower false-positive rate may be preferable. We used the variant quality score to tune call rate and accuracy (fig. S8). Additionally, novelty rate (relative to dbSNP) is also a function of variant quality score (fig. S9).

We processed the NA07022 data with Trait-o-Matic automated annotation software [as in (*12*)] yielding 1,159 annotated variants, 14 of which may have disease implications (table S10).

Because, the DNB sequencing substrates are produced by rolling-circle replication (*34*) in a uniform-temperature, solution-phase reaction with high template concentrations (> 20 billion per ml) this system avoids significant selection bottlenecks and non-clonal DNBs. This circumvents the stochastic inefficiencies of approaches that require precise titration of template concentrations for in situ clonal amplification in emulsion (*9*, *14*, *29*) or bridge PCR (*6*, *19*).

Our patterned arrays include high-occupancy and high-density nanoarrays self-assembled on photolithography-patterned, solid-phase substrates through electrostatic adsorption of solution-phase DNBs and yield a high proportion of informative pixels (site occupancies >95%) (fig. S12A)compared to random-position DNA arrays. This results in several hundred reaction sites in the compact (~300 nm diameter) DNB produce bright signals useful for rapid imaging of the sequences (SOM). Such small DNBs also allow for high density arrays. The data set reported herein was generated with arrays with ~350 million spots at a pitch of 1.29 µm. Such a spot density and higher ones achieved in proof of concept experiments (fig. S12B), result in high image efficiency and reduced reagent consumption that enable high sequencing throughput per instrument critical for high scale human genome sequencing for research and clinical applications.

Both sequencing by synthesis (SBS) and sequencing by ligation (SBL) use chained reads, wherein the substrate for cycle N+1 is dependent on the product of cycle N; consequently errors may accumulate over multiple cycles and data quality may be affected by errors (especially incomplete extensions) occurring in previous cycles. Thus, reactions need to be driven to near completion with high concentrations of expensive high purity labeled substrate molecules and enzymes. The independent, unchained nature of cPAL avoids error accumulation and tolerates low quality bases in otherwise high quality reads, thereby decreasing reagent costs. The average sequencing consumables cost for these three genomes was under $4,400 (table S5). The raw base and variant call accuracy achieved compares favorably with other reported human genome sequences (*2–12*).

As the sequencing substrates are produced by a DNA engineering process based on modified nick-translation for directional adaptor insertion (SOM), we obtained over 90% yield in adaptor ligation; and low chimeric rates of about 1% (SOM). DNA molecules with an inserted adaptor are further enriched with PCR (SOM). This recursive process can be implemented in batches of 96 samples and extended by inserting additional adaptors to read 120 bases or more per DNB (fig. S10). The current read length is comparable to other massively parallel sequencing technologies (*6–12*).

The sequence data reported here achieve sufficient quality and accuracy for complete genome association studies, the identification of potentially rare variants associated with disease or therapeutic treatments, and the identification of somatic mutations. The low cost of consumables and efficient imaging may enable studies of several hundreds of individuals. The higher accuracy and completeness required for clinical diagnostic applications provides incentive for continued improvement of this and other technologies.

References and Notes

1. T. A. Manolio, L. D. Brooks, F. S. Collins, *J. Clin. Invest.* **118(5)** 1590 (2008).
2. International Human Genome Sequencing Consortium, *Nature* **409**, 860 (2001).
3. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
4. S. Levy *et al.*, *PLoS Biol.* **5**, e254 (2007).
5. D. A. Wheeler *et al.*, *Nature* **452**, 872(2008).
6. D. R. Bentley *et al.*, *Nature* **456**, 53(2008).
7. J. Wang *et al.*, *Nature* **456**, 60 (2008).
8. S. M. Ahn *et al.*, *Genome Res.* **19**, 1622 (2009).
9. K. J. McKernan *et al.*, *Genome Res.* **19**, 1527 (2009).
10. T. J. Ley *et al.*, *Nature* **456**, 66 (2008).
11. D. Pushkarev, N. F. Neff, S. R. Quake, *Nat. Biotechnol.* **27**, 847 (2009).
12. J. I. Kim *et al*., *Nature*. **460**, 1011 (2009).
13. R. Drmanac *et al.*, *Genomics* **4**(2)**,** 114 (1989).
14. R. Drmanac, R. Crkvenjakov, *Scientia Yugoslavica*, **16** (1-2), 97 (1990).
15. R. Drmanac *et al.*, *Science* **260**, 1649 (1993).

16. P.C. Cheesman, US patent 5,302,509 (1994).
17. R. Drmanac, World Intellectual Property Organization WO/1995/009248 (1995).
18. M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlen, P. Nyren, *Anal. Biochem.* **242**, 84 (1996).
19. C. P. Adams, S. J. Kron, U.S. Patent 5,641,658 (1997).
20. P. M. Lizardi *et al.*, *Nat. Genet*. **19**, 225 (1998).
21. S. C. Macevicz, U.S. Patent 5,750,341 (1998).
22. S. Drmanac, D Kita, *et al.*, *Nat. Biotechnol*. **16,** 54 (1998).
23. R. D. Mitra, G. M. Church, *Nucleic Acids Res.* **27**, e34 (1999).
24. S. Brenner *et al.*, *Nat. Biotechnol.* **18**, 630 (2000).
25. I. Braslavsky, B. Hebert, E. Kartalov, S. R. Quake, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 3960 (2003).
26. R. D. Mitra, J. Shendure, J. Olejnik, O. E. Krzymanska, G. M. Church, *Anal. Biochem.* **320**, 55 (2003).
27. R. Drmanac *et al.*, World Intellectual Property Organization WO/2004/076683 (2004).
28. J. Shendure *et al.*, *Science* **309,** 1728 (2005).
29. M. Margulies *et al.*, *Nature* **437**, 376 (2005).
30. T.D. Harris *et al.*, *Science* **320**, 106 (2008).
31. A. Pihlak *et al.*, *Nat Biotechnol.* **26(6)**, 676 (2008).
32. J. Shendure, H. Ji, *Nat Biotechnol.* **26,** 1135 (2008).
33. The International HapMap Consortium, *Nature* **449**, 851 (2007).
34. L. Blanco *et al.*, *J Biol Chem.* **264**, 8935 (1989).
35. K Zhang *et al*., *Nature Methods* **6(8),** 613 (2009).
36. M.P. Ball, *et al*., *Nature Biotechnol.* **27,** 361 (2009).
37. J. B. Li *et al*., *Genome Research* Jul 13. PMID: 19525355 (2009).

38. We acknowledge and the ongoing contributions and support of all Complete Genomics employees and R. Mercado for manuscript preparation. Some of this work was supported from PersonalGenomes.org, and NHLBI. Data has been deposited at NCBI: reads in the Short Read Archive (SRA), accession SRA008092; and SNPs in dbSNP, accessions ss161884913 to ss175323894. Employees of Complete Genomics have stock options in the company and DGB has stock in Perlegen Sciences. Complete genomics has filed several patents on this work.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1181498/DC1
Materials and Methods
Figs. S1 to S12
Tables S1 to S9
References

3 September 2009; accepted 23 October 2009
Published online 5 November 2009;
10.1126/science.1181498
Include this information when citing this paper.

Fig. 1. Amplified DNA nanoarray platform. (**A**) Schematic flow diagram of the process used. (**B**) Library construction schematic (SOM, fig. S1). (**C**) DNB production (fig. S11, SOM) and nanoarray formation (SOM) schematics. (**D**) Schematic of combinatial probe anchor ligation (cPAL) products (SOM).

Table 1. Summary information from mapping and assembly of three genomes. All variations are with respect to the National Center for Biotechnology Information (NCBI) version 36 human genome reference assembly. Novel variations were ascertained by comparison to dbSNP (JDW, release 126; NA18507 (*6*), release 128; all other genomes, release 129). NA18507 and NA19240 are Yoruban HapMap samples, which may explain the number of SNPs and novelty rates. In partially called regions of the genome, one allele could be called confidently but not the other. The high call rate in NA19240 reflects reduced library bias (fig. S5).

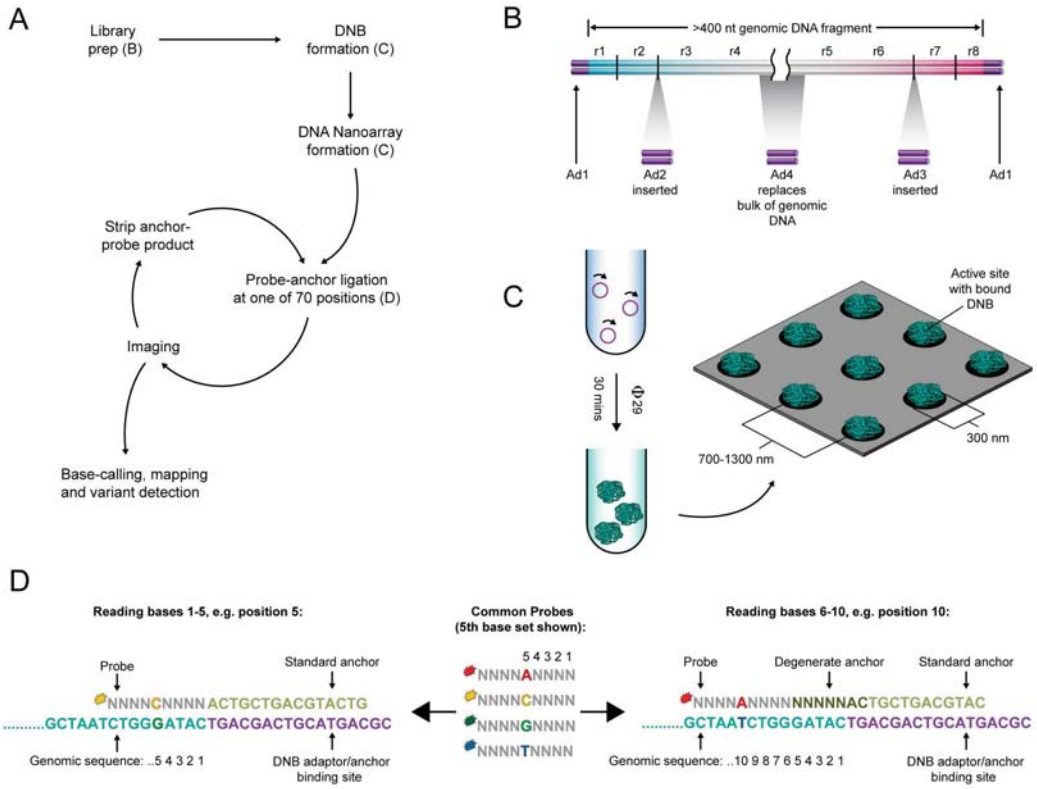
Sample	Mapped sequence (Gb)	Average Coverage depth (fold)	Percent of genome called		SNPs		Indels		Insertion:Deletion ratio
			Fully	Partially	Total	Novel %	Total	Novel %	
Genomes sequenced by Complete Genomics:									
NA07022 (*35*)	241	87	91%	2%	3,076,757	10%	337,604	37%	1.0
NA19240 (*36*)	178	63	95%	1%	4,042,801	19%	496,149	42%	0.96
NA20431 (*37*)	124	45	86%	3%	2,905,517	10%	269,794	37%	1.0
Genomes previously published:									
NA18507 (*6*)	135	47	–	–	4,139,196	26%	404,416	50%	0.77
NA18507 (*9*)	49	31	–	–	3,866,085	19%	226,529	33%	0.72
JCV (*3*)	21	7	–	–	3,213,401	15%	851,575	–	–
JDW (*4*)	21	7	–	–	3,322,093	18%	222,718	51%	0.4

Table 2. Concordance with genotypes for NA07022 generated by the HapMap Project (release 24) and the highest quality Infinium assay subset of those genotypes, as well as genotyping on Illumina Infinium 1M assay. Discordances with reported HapMap Infinium genotypes were verified by Sanger sequencing (SOM).

			Infinium 1M	HapMap phase I&II SNPs	HapMap Infinium subset	HapMap Infinium SNPs tested for accuracy by Sanger sequencing		
Published Concordance			–	99.03%	99.94%			
NA07022	# reported		1 M	3.9 M	143 K	These data correct	These data incorrect	% affirmed
	% called		95.98%	94. 39%	96.00%			
	% locus concordance		99.89%	99.15%	99.88%			
	HapMap genotype calls	Homozygous ref	99.96%	99.34%	99.96%	18	2	90%
		Heterozygous	99.78%	99.39%	99.80%	28	46	38%
		Homozygous alt	99.81%	98.14%	99.84%	28	12	70%

Table 3. False positive rates and FDRs were calculated for the entire set of variations called in NA07022 by extrapolating the heterozygous (Het) FDRs calculated from comparative Sanger sequencing of 291 selected novel variants (table S8) to all variants. This is a conservative approach (detailed in SOM text). The total number of all types of false positive variants is estimated at 7.5-16.1 per Mb.

Variation Type	Total detected	Novel	Het novel FDR (table S8)	Estimated false positives on genome	Estimated false positives / Mbp	Estimated FDR
SNP	3,076,869	310,690	2-6%	7k-17k	2.3-6.1	0.2-0.6%
Deletion	168,726	61,960	8-14%	5k-8k	1.8-3.0	3.0-5.0%
Insertion	168,909	61,933	11-18%	7k-11k	2.3-3.9	3.9-6.5%
Block substitution	62,783	30,445	11-29%	3k-9k	1.1-3.1	5.2-13.9%


A
Library prep (B)
DNB formation (C)
DNA Nanoarray formation (C)
Probe-anchor ligation at one of 70 positions (D)
Imaging
Strip anchor-probe product
Base-calling, mapping and variant detection
B
>400 nt genomic DNA fragment
r1
r2
r3
r4
r5
r6
r7
r8
Ad1
Ad2 inserted
Ad4 replaces bulk of genomic DNA
Ad3 inserted
Ad1
C
30 mins
Φ29
Active site with bound DNB
300 nm
700-1300 nm
D
Reading bases 1-5, e.g. position 5:
Probe
Standard anchor
NNNNCNNNN ACTGCTGACGTACTG
.....GCTAATCTGGGATACTGACGACTGCATGACGC
Genomic sequence: ..5 4 3 2 1
DNB adaptor/anchor binding site
Common Probes (5th base set shown):
5 4 3 2 1
NNNNANNNN
NNNNCNNNN
NNNNGNNNN
NNNNTNNNN
Reading bases 6-10, e.g. position 10:
Probe
Degenerate anchor
Standard anchor
NNNNANNNNNNNNNACTGCTGACGTAC
.....GCTAATCTGGGATACTGACGACTGCATGACGC
Genomic sequence: ..10 9 8 7 6 5 4 3 2 1
DNB adaptor/anchor binding site

www.sciencemag.org/cgi/content/full/1181498/DC1

Supporting Online Material for

Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays

Radoje Drmanac,* Andrew B. Sparks, Matthew J. Callow, Aaron L. Halpern, Norman L. Burns, Bahram G. Kermani, Paolo Carnevali, Igor Nazarenko, Geoffrey B. Nilsen, George Yeung, Fredrik Dahl, Andres Fernandez, Bryan Staker, Krishna P. Pant, Jonathan Baccash, Adam P. Borcherding, Anushka Brownley, Ryan Cedeno, Linsu Chen, Dan Chernikoff, Alex Cheung, Razvan Chirita, Benjamin Curson, Jessica C. Ebert, Coleen R. Hacker, Robert Hartlage, Brian Hauser, Steve Huang, Yuan Jiang, Vitali Karpinchyk, Mark Koenig, Calvin Kong, Tom Landers, Catherine Le, Jia Liu, Celeste E. McBride, Matt Morenzoni, Robert E. Morey, Karl Mutch, Helena Perazich, Kimberly Perry, Brock A. Peters, Joe Peterson, Charit L. Pethiyagoda, Kaliprasad Pothuraju, Claudia Richter, Abraham M. Rosenbaum, Shaunak Roy, Jay Shafto, Uladzislau Sharanhovich, Karen W. Shannon, Conrad G. Sheppy, Michel Sun, Joseph V. Thakuria, Anne Tran, Dylan Vu, Alexander Wait Zaranek, Xiaodi Wu, Snezana Drmanac, Arnold R. Oliphant, William C. Banyai, Bruce Martin, Dennis G. Ballinger,* George M. Church, Clifford A. Reid

*To whom correspondence should be addressed. E-mail: rdrmanac@completegenomics.com (R.D.); dballinger@completegenomics.com (D.G.B.)

Published 5 November 2009 on *Science* Express
DOI: 10.1126/science.1181498

This PDF file includes:

Materials and Methods
Figs. S1 to S12
Tables S1 to S9
References

the two allele sequences. A locus was determined to be confirmed if the corresponding traces aligned exactly to the expected read sequence at that variant position for at least one strand. Any strand contradiction or discrepancies due to background noise were resolved by visual inspection of the traces.

Section 10: Analysis of impact of coding SNPs

All SNP variants identified in NA07022 were analyzed with Trait-o-Matic software (as in *S25*). This software, run as a website, returns all non-synonymous SNP (nsSNP) variants found in HGMD, OMIM and SNPedia (cited SNPs), as well as all nsSNPs not specifically listed in the preceding databases, but that occur in genes listed in OMIM (uncited nsSNPs). Analysis of the NA07022 genome with Trait-o-Matic returned 1,141 variants, including 605 cited nsSNPs, and 536 uncited nsSNPs. Filtering of 320 variants with BLOSUM100 scores below 3 and 725 variants with a minor allele frequency (MAF) > 0.06 in the Caucasian/European (CEU) population (weighted average of HapMap and 1000 genomes frequency data) left 55 cited nsSNPs and 41 uncited SNPs. Forty one cited nsSNPs were removed either because their phenotypic evidence was based solely on association studies, or because they were not disease-associated (e.g. olfactory receptor, blood type, eye color), and 38 uncited nsSNPs were removed because they had non-obvious functional consequences. Table S9 lists the remaining 14 cited nsSNPs (12 heterozygous loci and one compound heterozygous locus), three uncited nsSNPs (two nonsense mutations and one homozygous mutation) as well as two common variants in APOE with potential phenotypic consequences.

Section 11. False Discovery rate (FDR) calculation for novel variations

Of the variations called in NA07022 that were novel with respect to dbSNP (build 129) and non-synonymous with respect to the NM_* set of NCBI Build 36.3 annotated transcripts, a random subset was assessed with Sanger sequencing (Table S8). For the purposes of this analysis, all indels that overlap the coding regions of transcripts were treated as non-synonymous changes irrespective of frame

State	Chr	Location	Gene	Alteration	Phenotype	Notes on Variants
Het	17	37949759	NAGLU	R737G	Sanfilippo Syndrome B	Identified in a patient with Sanfilippo Syndrome B, in association with a known Sanfilippo variant (*S8*). Also identified in Watson genome (*S9*) and NA20431.
Het	9	135291831	ADAMTS13	P426L	TTP	Identified as part of a compound heterozygote in Thrombotic Thrombocytopenic Purpura patient (*S10*).
Het	11	66050228	BBS1	M390R	Bardet-Beidl Syndrome	Homozygous variant reported as causative for Bardet-Beidl Syndrome in an oligogenic fashion (*S11*).
Het	19	6664262	C3	L314P	C3 structural variant	Codes for a structural variant of C3, of unknown clinical significance. Also identified in NA20431.
Het	2	201782343	CASP10	V410I	ALPS type II	Reported as recessive for ALPS type II (*S12*).
Het	2	227624091	COL4A4	G999E	TBMD	G->E mutations are often causative in TBMD; possibly pathogenic in a heterozygous form (*S13*). Also identified in Venter genome (*S5*).
Het	1	97754009	DPYD	S534N	DPYD deficiency	Heterozygote may reduce DPYD expression. Gross et al. (*S14*) note a severe phenotype in two compound heterozygotes.
Het	15	78259581	FAH	R341W	FAH deficiency	Is a pseudodeficiency allele for FAH and is observed in compound heterozygotes with FAH deficiency (*S15*).
Het	16	3244464	MEFV	R202Q	FMF	Possibly autosomal recessive causative variant for FMF (*S16*).
Het	12	55711185	MYO1A	S797F	early onset hearing loss	Reported as causative for dominant early onset moderate sensorineural hearing loss (*S17*). Also identified in NA20431.
Het	22	16946288	PEX26	L153V	Infantile Refsum Disorder	Reported as part of a compound heterozygote causative of Infantile Refsum Disorder (*S18*).
Het	19	46550716	TGFB1	R25P	hepatic fibrosis	Affects TGFβ1 levels. Associated with hepatic fibrosis in chronic HCV infections (*S19*).
Comp. Het	16	49303427/ 49314041	NOD2	R702W/ G908R	Crohn's disease	Compound heterozygote involving two variants (one with MAF of 0.03) associated with Crohn's disease (*S20*).
Het	18	19737949	LAMA3	K2069X	junctional epidermolysis bullosa	LAMA3 inactivation is implicated in autosomal recessive Epidermolysis Bullosa (*S21*). The most C-terminal mutation causative of disease is Q1368X.
Het	10	55296582	PCDH15	Y1181X	deafness	PCDH15 inactivation is implicated in autosomal recessive deafness (*S22*). The most C-terminal mutation causative of disease is S647X.
Hom	2	130996158	CFC1	W78R	Left-right axis abnormalities	BLOSUM score of 4. CFC1 has 4 OMIM-listed variants that exhibit a dominant expression for left-right axis abnormalities; two of these have incomplete penetrance (*S23*).
Comp. Het	19	50103781/ 50103919	APOE	C130R/ R176C	Alzheimer's Disease	These variants represent a ApoE4/ApoE2 heterozygote (*S24*)

Table S9: Summary of impact of coding variants in NA07022. See SOM text for details.

S1. G. A. Denisov, A. B. Arehart, and M. D. Curtin, US Patent 6681186 (2004).
S2. K. Li et al., BMC Bioinformatics 9, 1 (2008).
S3. P. Rice et al., TIG 16, 276 (2000).
S4. J.C. Venter, *et al. Science* **291**, 1304 (2001).
S5. S. Levy et al., PLoS Biol 5, e254 (2007).
S6. D.R. Bentley, *et al., Nature* **456**, 53(2008).
S7. D. Pushkarev, N.F. Neff, S.R. Quake, *Nat. Biotechnol.* **27**, 847 (2009).
S8. G.R. Villani, G. Pontarelli, D. Vitale, P. DiNatale, Hum Genet 115, 173 (2004).
S9. D. A. Wheeler et al., Nature 452, 872 (2008).
S10. K. Assink et al., Kidney Int 63, 1995 (2003).
S11. http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=209901
S12. J. Wang et al., Cell 98, 47 (1999).
S13. M. Buzza et al., Kidney Int 63, 447 (2003)
S14. E. Gross et al., Hum Mutat 22, 498 (2003).
S15. http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=276700
S16. K. Ritis et al., Ann Rheum Dis 63, 438 (2004).
S17. F. Donaudy F et al., Am J Hum Genet 72, 1571 (2003).
S18. S. Furuki et al., J Biol Chem 281, 1317 (2006).
S19. C. G. et al., Cytokine 24, 173 (2003).
S20. J. P. Hugot et al., Nature 411, 599 (2001).
S21. http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=600805
S22. http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=605514
S23. http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=605194
S24. http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=107741
S25. J.I. Kim, et al. *Nature*. **460,** 1011 (2009).