Published in final edited form as:

Ann Hum Genet. 2009 September ; 73(Pt 5): 502–513. doi:10.1111/j.1469-1809.2009.00530.x.

Combining Microarray-based Genomic Selection (MGS) with the Illumina Genome Analyzer Platform to Sequence Diploid Target Regions

D. T. OKOU¹, A. E. LOCKE^{1,3}, K. M. STEINBERG^{1,2}, K. HAGEN¹, P. ATHRI¹, A. C. SHETTY¹, V. PATEL¹, and M. E. ZWICK^{1,2,3}

¹ Department of Human Genetics, Emory University School of Medicine, Atlanta, GA

² Graduate Program in Population Biology, Ecology and Evolution, Emory University, Atlanta, GA

³ Graduate Program in Genetics and Molecular Biology, Emory University, Atlanta, GA

SUMMARY

Novel methods of targeted sequencing of unique regions from complex eukaryotic have generated a great deal of excitement, but critical demonstrations of these methods efficacy with respect to diploid genotype calling and experimental variation are lacking. To address this issue, we optimized microarray-based genomic selection (MGS) for use with the Illumina Genome Analyzer (IGA). A set of 202 fragments (304 kb total) contained within a 1.7-Mb genomic region on human chromosome X were MGS/IGA sequenced in ten female HapMap samples generating a total of 2.4 GB of DNA sequence. At a minimum coverage threshold of 5X, 93.9% of all bases and 94.9% of segregating sites were called, while 57.7% of bases (57.4% of segregating sites) were called at a 50x threshold. Data accuracy at known segregating sites was 98.9% at 5X coverage, rising to 99.6% at 50X coverage. Accuracy at homozygous sites was 98.7% at 5X sequence coverage and 99.5% at 50X coverage. Although accuracy at heterozygous sites was modestly lower, it was still over 92% at 5X coverage and increased to nearly 97% at 50X coverage. These data provide the first demonstration that MGS/IGA sequencing can generate the very high quality sequence data necessary for human genetics research.

All sequence generated in this study have been deposited in NCBI Short Read Archive (<http://www.ncbi.nlm.nih.gov/Traces/sra>, Accession # SRA007913).

Keywords

Personal Genomes; Direct Selection; Microarray-based Genomic Selection; Illumina Genome Analyzer; Targeted Sequencing; Human Genetics

INTRODUCTION

The application of genomics technologies to identify the causative variants underlying phenotypic traits is one of the central challenges of genetics today. As we approach an era of personal genomes, the apparently complex genomic architecture underlying many human traits poses significant technical challenges to both basic research and medical genomics (Zwick *et al.*, 2000). On the one hand, we have those variants most easily identified as causative: the rare, single genotypic changes that result in major phenotypic differences.

Even in such cases, however, obtaining the genome sequence of the multiple loci or, in some cases, very large genes, can slow the development of efficient genetic testing assays. At the other extreme, the technological development and use of genome-wide association (GWA) studies in a case-control framework to identify common single nucleotide polymorphisms (SNP) that either cause disease, or are in linkage disequilibrium with causative variants, has enabled the genetic dissection of a wide variety human complex disease traits (Consortium, 2005, Frazer *et al.*, 2007, Consortium, 2007, McCarroll *et al.*, 2008, Raychaudhuri *et al.*, 2008). Yet despite the many successes of GWA studies, a substantial genetic contribution to these disorders remains to be discovered. One possible explanation is that such diseases are caused by rare variants that would not be easily detected by whole genome association (Zwick *et al.*, 2000, Pritchard, 2001, Pritchard & Cox, 2002). If this is the case, the direct sequencing of genomic regions and personal genomes to identify causative variants should become of increasing utility for exploring the role of rare variation in human disease.

A number of second generation sequencing technologies are beginning to give investigators enormous raw sequencing power at a dramatically lower cost per sequenced base (Cutler *et al.*, 2001, Margulies *et al.*, 2005, Shendure *et al.*, 2005, Bentley *et al.*, 2008, Shendure *et al.*, 2004), and applying these technologies for the targeted resequencing of large genomic regions could yield many new research and clinical applications. Yet, major challenges remain, among them isolating target DNA, sequencing to the appropriate depth for data completeness and accuracy, and developing bioinformatics tools for data analysis.

Large genomic regions, ranging in size from hundreds to thousands of kilobases, are hard to isolate as target DNA for sequencing using direct PCR of targeted fragments. To ease the isolation of target DNA, direct genomic selection was developed, but because it was paired with the more expensive traditional Sanger sequencing, it did not enjoy wide use (Bashiardes *et al.*, 2005). More recent efforts to overcome this technical challenge include a number of solid and liquid phase genomic selection methods paired with second generation sequencing (Okou *et al.*, 2007, Albert *et al.*, 2007, Porreca *et al.*, 2007, Hodges *et al.*, 2007, Krishnakumar *et al.*, 2008, Bau *et al.*, 2008, Gnirke *et al.*, 2009).

While these approaches hold great promise, whether targeted sequencing on second-generation sequencing platforms can achieve the level of accuracy and data completeness necessary for many medical and research applications remains to be seen (Olson, 2007). For instance, there have been two solid phase selection studies published that did not report raw sequence accuracy, making it difficult to assess the utility of the approach (Albert *et al.*, 2007, Hodges *et al.*, 2007). Furthermore, although other studies have shown that variable homozygous sites can be identified with great accuracy (Okou *et al.*, 2007, Porreca *et al.*, 2007), detecting both alleles of known heterozygous genotypes is reportedly accurate at only ~31% of variable sites in a single sample (Porreca *et al.*, 2007). A recently published liquid phase hybrid selection method reported improved results, sequencing 64% of targeted exons and obtaining highly accurate SNP calls at 67% of targeted SNPs located within 2.5 Mb of targeted exonic sequences in three HapMap samples (Gnirke *et al.*, 2009). To date, there have been no published studies that have demonstrated that solid phase selection and sequencing is capable of making highly accurate genotype calls in multiple diploid samples at the vast majority of targeted sites.

Here we provide the first demonstration of a solid phase selection and sequencing protocol capable of making highly accurate genotype calls at a majority of targeted sites. We have seamlessly integrated microarray-based genomic selection (MGS) with sequencing on the Illumina Genome Analyzer (IGA) platform. In order to focus on data quality and completeness, we used MGS to directly select and sequence 304 kb from a targeted 1.7 Mb-sized region on the X chromosome in 10 HapMap females. Our data provide the first

demonstration that MGS/IGA sequencing is a robust method capable of making highly accurate genotype calls at more than 90% of known segregating sites in the ten samples that we sequenced. Furthermore, we report changes in the MGS protocol that significantly improves the obtained level of enrichment. Finally, we find no evidence of allelic bias in the capture of both alleles at heterozygous sites. Our data show that MGS/IGA sequencing is a sufficiently repeatable and accurate methodology that will surely contribute to the identification and interpretation of human genomic variation that will be revealed by the targeted sequencing of personal genomes.

RESULTS

Figure 1 shows the MGS/IGA protocol outlined in schematic form with specific details of its implementation contained within the Materials and Methods and our latest complete protocol in Supplemental Data 1. We have integrated the standard Illumina Genome Analyzer adaptors directly into the MGS/IGA protocol. To validate our approach, we used a 385,000-probe custom microarray (Roche NimbleGen, Inc.) targeted toward 202 non-overlapping genomic fragments located on the human X chromosome. In total, these fragments consisted of 304 kb of unique sequence surrounding and including three protein-coding genes (*FMR1*, *FMR1NB*, and *AFF2*) from a larger 1.7-Mb genomic region (Figure 2a, 2b). Our sample population consisted of ten females from the HapMap: five of European descent (NA07000, NA07055, NA11993, NA12057, and NA12145) and five of African (Yoruban) descent (NA18502, NA18505, NA18508, NA18517, and NA18523).

Using ten IGA lanes for sequencing after selection by MGS, we generated 2.14 gigabases (Gb) of total sequence. We obtained the highest levels of enrichment for samples NA18505 and NA18523 where we used 1X COT, hybridized at 55C and sequenced on the GAI platform (Supplemental Data 2). The median coverage across the 202-targeted genomic regions ranged from 9.5 to 270.5, and the mean coverage ranged from 13.2 to 356.1 (Table 1). Across all ten samples sequenced, approximately 2% of the 2020 fragments sequenced had a median coverage of less than 5 (Figure 3). Most of the low coverage fragments were found in a single sample (NA18505), which had the lowest IGA sequence output. We repeated the sample NA18505 two additional times and obtained poor coverage, suggesting that the cause of the relatively poor MGS/IGA sequencing performance was a property of that specific DNA sample (data not shown). Our coverage data imply that there was no systematic failure of any of the 202-targeted fragments across the different samples sequenced.

The proportion of reads mapping to the targeted genomic region varied approximately fourfold across all samples (Table 1). Estimated enrichment among all IGA sequence reads that map uniquely to the human genome range between 956 and 6465 (mean 2786). Using a slightly more conservative criterion that estimates enrichment relative to the total IGA sequence obtained from each lane resulted in a similar observed level of enrichment (Table 1). The fold enrichment obtained is correlated with the total number of IGA reads, suggesting that at least a portion of the variation among samples we observed arises from IGA sequencing of targets ($r^2=0.03$, $p=0.048$). The cause of this correlation probably arises as a consequence of imprecise DNA quantitation prior to IGA cluster generation. The median coverage at the 2020 fragments showed a slight negative relationship with fragment size, although this association was not statistically significant ($r^2=0.001$, $p=0.057$, Figure 4a). In contrast, we found that the median coverage at the 2020 fragments exhibited a weak positive correlation with GC content that was statistically significant ($r^2=0.03$, $p=2.11 \times 10^{-15}$, Figure 4b). Notably, this modest correlation is in the opposite direction of that reported in human whole genome sequencing studies using the IGA platform (Bentley *et al.*, 2008; Wang *et al.*, 2008).

We evaluated the data completeness of our MGS/IGA sequence at both variant and invariant sites among the 2020 fragments we resequenced (Figure 5). The regions we targeted contain 329 (CEPH) and 331 (YRU) SNPs that had already been genotyped by the HapMap project (Frazer *et al.*, 2007). At a minimum coverage threshold of 5X, 93.9% of all bases are called, and 94.9% of segregating sites are called. These percentages decrease linearly as we increase the threshold, with 57.7% of bases (57.4% of segregating sites) called at a 50x threshold. These data suggest no apparent bias in the basecalling rates between invariant and segregating sites, since their data completeness is similar at all coverage levels. We note, however, that our estimate of theta at 20X (0.001) and 50X coverage (0.0008) is approximately 1.6 to 2-fold higher than we expected (0.0005) for this region on the X chromosome. The cause of this observation remains unknown, although we believe it highly likely that improved methods of assembly and genotype calling would likely reduce this discrepancy.

To assess the accuracy of our MGS/IGA sequence data, we compared our genotype calls at 3300 known SNPs with genotype data publicly available from the HapMap project (Frazer *et al.*, 2007). The overall accuracy at variable sites was 98.9% at 5X coverage and increased to 99.6% at 50X coverage (Figure 5). We saw that accuracy at homozygous sites was 98.7% at 5X sequence coverage and 99.5% at 50X coverage (Figure 5). Although accuracy at heterozygous sites was modestly lower, it was still over 92% at 5X coverage and increased to nearly 97% at 50X coverage.

In our initial analysis of the MGS/IGA sequence data, we observed 63 discrepant genotype calls at 10X coverage. Using Sanger sequencing to independently verify these genotypes revealed 16 cases (25.3%) where the HapMap genotyping was incorrect while the MGS/IGA sequencing call was correct. Another 28 discrepant SNPs (44.4%) had at least three or more IGA reads of one or two alleles consistent with the HapMap genotype (2 homozygous, 26 heterozygous), but in each case MAQ failed to correctly call the correct diploid genotype. Nineteen of these discrepancies had over 100X total coverage at the variant site, with greater than 20X coverage of both alleles. Our data suggests that improved methods of calling diploid genotypes can be expected to increase data accuracy at these types of sites. Combined, 66.7% of the discrepant bases either show strong evidence for or were unambiguously confirmed as being correctly sequenced by MGS/IGA (Figure 5). The remaining 19 MGS/IGA sequencing errors occurred at heterozygous sites and showed a small, but not statistically significant bias toward calling the reference allele (12 matched reference allele, 7 matched other allele, sign test, $p = 0.36$).

DISCUSSION

The targeted sequencing of unique genomic regions from complex eukaryotic genomes will enable a host of potential new applications. In human genetics, these methods can be expected to enable more detailed studies of human genome variation while at the same time, speeding the discovery of causative alleles underlying human Mendelian disorders and common multifactorial diseases. We have shown that MGS/IGA sequencing can be combined successfully to generate the kind of very high quality sequence data necessary for both research and medical genomics applications. With an overall accuracy rate of 98.9% at targeted variable sites, this combination represents a significant step forward, with accuracy on par with the HapMap (Frazer *et al.*, 2007). Of particular note, we report dramatic improvements at heterozygous sites and data completeness over previously published data (Porreca *et al.*, 2007, Gnirke *et al.*, 2009).

The improved accuracy at segregating sites observed in this study is likely a function of the almost five-fold greater enrichment achieved with our current protocol as compared to our

previous work (Okou *et al.*, 2007). MGS/IGA sequencing did not show a decreased level of coverage at smaller fragments, which had been a common finding in earlier studies. We believe this may arise as a consequence of both our protocol modifications and the high density of capture probes for each targeted region. The fact that we can obtain the very high level of enrichment necessary to obtain nearly complete high quality sequence coverage among the 202 fragments (304 kb) in the 1.7 Mb-sized genomic region implies that larger genomic regions might also be sequenced nearly completely to generate highly accurate data with MGS/IGA sequencing. We are currently exploring reducing the number of probes and optimizing probe selection in order to expand the size of regions that can be resequenced, while maintaining high data completeness and accuracy. On the other hand, producing arrays with even greater densities of capture probes would be expected to improve the performance of the MGS/IGA sequencing assay.

While our data demonstrate that MGS/IGA sequencing is robust, the variation in enrichment we observe among samples reveals that some sources of significant experimental variation remain to be understood and provide opportunities for future improvement. Prior to our work, a careful presentation of the extent of variation among different samples that a user might expect to observe has been lacking. One potential cause of this variation lies in the amount of sequence generated per IGA lane, which clearly influenced our ability to successfully detect variant sites successfully. Increasing the amount of sequence coverage can be expected to further improve detection of both alleles in heterozygotes. Furthermore, our analysis revealed that existing genotype calling software might fail to detect variable sites, even when sequence coverage is very high. Other potential sources of variation lie in the MGS protocol itself, and we are working to identify and minimize their effects. Finally, all methods of targeted sequencing will be most successfully applied to unique sequence regions in complex eukaryotic genomes. Because repetitive sequences, that include simple repeats, transposable elements, and gene families, may not be able to be uniquely enriched, we do not expect that they will be able to be reliably sequenced. Thus detecting genetic variation in repetitive regions will likely have to be pursued with alternative approaches.

All of these results lead us to the conclusion that genomic selection technologies, though still in their infancy, are not only capable of enriching for target sequences, but when teamed with high-throughput sequencing technologies, they are capable of meeting the stringent standards of completeness and accuracy necessary for studies in the genomic era of biomedical research. Adapting the MGS/IGA protocol for use with paired-end sequencing is straightforward and can be expected as a next step to improve sequence coverage as to enable the detection of insertion and deletion variation. Future improvements in MGS array design also seem likely to improve overall performance. The ability to quickly redesign a MGS array is a particular strength of this technology, especially for medical genomic applications where one may want to offer a personalized genetic test. Nevertheless, we believe that both solid and liquid phase enrichment protocols will prove useful for a wide variety of applications as their reliability, data completeness and sequence coverage continue to improve.

Although we have stressed the importance of the MGS/IGA sequencing for medical genomics, it is clear that this technique can be adapted easily to many research applications, not only for humans, but other model systems. From selecting and sequencing an association or linkage peak, the comprehensive sequence analysis of a candidate pathway, rapid mapping of induced mutations in model systems, or clinical applications in human genetics, continued improvement of methods like MGS/IGA sequencing will prove their worth as a viable and convenient alternative to generate target DNA for novel DNA sequencing platforms.

METHODS

Array Design

We used the UCSC Table Browser function with repeats masked on the latest human genome build (March 2006) to identify the unique sequences within a selected genomic region (Thomas *et al.*, 2003). The CGG repeat sequence of FMR1 from the human genome reference sequence was included in the design. Since genetic variants in regulatory elements away from the coding sequences may influence gene expression (Kleinjan & van Heyningen, 2005), unique sequence upstream and downstream of the target genes were also included. We then selected among the unique sequence to obtain 304 kb of unique sequence. We excluded unique sequences of 100 bp or less and in some cases, included short (<100 bp) stretches of previously masked sequence, to avoid breaking up large genomic regions into smaller fragments.

The sequences, in FASTA format, were then provided to chip design engineers at Roche-NimbleGen to select oligonucleotides for the Microarray-based Genomic Selection (MGS) chip. Standard bioinformatics filters that check for genomic uniqueness against an indexed human genome (15mers) were used to select capture oligonucleotides (oligos). The oligos were between 50 and 93 basepairs long and were designed to achieve optimal isothermal hybridization across the microarray. The MGS microarrays used contain ~385,000 capture probes. For the 202 fragments (304 kb), there were two pairs of probes for every 3 bases.

Sample Selection

DNA samples were purchased from the Coriell Cell Repository (<http://ccr.coriell.org>) and included 10 females representing two different populations: one of European descent (n=5) selected from the Centre d'Etude du Polymorphisme Humain (CEPH) panel with Coriell Cell Repository numbers NA07000, NA07055, NA11993, NA12057, and NA12145; and a second population of African descent (n=5) selected from the HapMap Project with Coriell Cell Repository numbers: NA18502, NA18505, NA18508, NA18517 and NA18523.

Adaptor and Primers Oligonucleotides

The adaptor oligos used in this project were ordered from Invitrogen Corp. and represented the genomic DNA adaptor sequences indicated by Illumina. Each adaptor oligo (forward and reverse) was diluted in water to 400 μ M. The adaptors for repaired-end ligation were prepared by mixing equal volumes of forward and reverse oligonucleotide to generate a double stranded molecule as would be supplied by Illumina. The mixture was heated at 95°C for ten minutes in a heating block. The heating block was then lowered 65°C to allow the oligos to slowly cool and anneal for two hours. The PCR and sequencing primers used were either ordered from Invitrogen or purchased directly from Illumina. When obtained from Invitrogen, the PCR and sequencing primers were prepared in water at 25 μ M and 100 μ M respectively.

Genomic DNA Preparation and Target Library Construction

Whole genome amplification was performed on 250 ng of genomic DNA using the RepliG Kit (Qiagen Inc.). Following amplification, the unpurified samples were diluted to 250 μ l with water. They were sonicated (Misonix sonicator S-4000) in Eppendorf tubes with a microtip probe using the following parameters: six pulses of 30 second each, with two minutes of rest at a power output level of 20%. After fragmentation, samples were purified with Promega Wizard® SV Gel and PCR Clean-Up System (Promega). Each sample required two purification columns to prevent saturation and maximize recovery. The samples were quantified using a spectrophotometer (NanoDrop ND1000) and approximately 250 ng of each sample was run on a 1.5% TAE agarose gel against 300 ng of a 1 Kb plus

ladder (Invitrogen) to verify that fragments averaged 250 bp in size. 20 to 25 µg of each sample was aliquoted into a sterile Eppendorf tube and the samples were then dried down in a SpeedVac at medium heat (75°C) to 47 µl.

Repairing Ends of the DNA Library

To 55 µl of fragmented DNA we added 10 µl of dNTPs (2.5 mM, TaKaRa), 10 µl of 10X T4 DNA Polymerase Buffer (New England Biolab (NEB)), 1 µl of 100X BSA (NEB) and 15 µl of T4 DNA Polymerase (3U/µl, NEB). The mixture was incubated at 12°C for 20 minutes followed by 75°C for 20 minutes. After incubation, the fragments were given A tails by adding 3 µl of 100mM dATP (Sigma), 3 µl of 50mM MgCl₂, and 5 µl of Taq DNA Polymerase (5U/µl, NEB) directly to the mixture. This was followed with incubation in a thermocycler at 72°C for 35 minutes. The sample was then purified with the Promega Wizard® SV Gel and PCR Clean-Up System following the manufacturer recommendation. Each column was eluted with 50 µl of water. After quantification, the volume was adjusted to 40 µl for phosphorylation. To the A tailed fragments, we added 5 µl 10X T4 DNA ligase Buffer (NEB), 1 µl 100mM ATP, and 4 µl of T4 Polynucleotide Kinase (10U/µl, NEB). The mixture was incubated at 37°C for 30 minutes followed by purification as described above. Samples were eluted with 70 µl of water and adjusted to 65 µl after Nanodrop quantification.

Ligation of Adaptors

In a PCR tube containing 65 µl (63 µl for samples NA18508 and NA18523) of the above repaired product, 10 µl of 10X T4 DNA Ligase Buffer (NEB), 20 µl of Adaptors (22 µl for NA18508 and NA18523) and 5 µl of T4 DNA Ligase (2000U/µl, NEB) were added. The mixture was incubated at 25°C for two hours. The ratio of adaptor ends to repaired DNA fragment ends was at least 12:1. The ligation product was purified using PureLink PCR purification kit and Binding Buffer HC (Invitrogen). Two columns were used for each sample and eluates were combined by sample after each column was eluted with 100 µl of water.

Hybridization of Sample to MGS Array

To 8 µg of the ligated sample, a 5-fold amount (in µg) of human Cot-1 DNA (Invitrogen) (equal amount of repaired DNA and cot-1 DNA for samples NA18508 and NA18523) was added. The samples were dried down to the pellet in a Speed-Vac at medium heat (75°C). To each pellet, 16.2 µl of VWR water, 20 µl of 2X Hybe Buffer and 3.5 µl Hybe Component A (Roche NimbleGen) were added. The sample pellets were gently resuspended and denatured at 95°C for ten minutes. The samples were quickly spun down and placed in a 50°C MAUI heat block (Biomicro) (55°C for samples NA18508 and NA18523) until ready to use. Each sample was loaded onto a custom MGS chip prefitted with SL lid (Biomicro) and hybridized at 50°C (55°C for samples NA18508 and NA18523) for 60 hours with mixing.

Elution of Target Fragments

After hybridization, the MGS arrays were quickly rinsed in warm (42°C) Wash Buffer 1 (Roche NimbleGen), followed by two five minute stringent washes at 55°C (60°C for samples NA18508 and NA18523) with a Stringent Buffer (Roche NimbleGen). The arrays were then rinsed at room temperature with Wash Buffer 1, Wash Buffer 2 and Wash Buffer 3 (Roche NimbleGen) for two minutes, one minute, and 30 seconds respectively. The MGS chips were then transferred to a custom-made heating block and the selected fragments for each sample were eluted at 95°C with three aliquots of VWR water (400 µl each), the first two following a five minute incubation and the third after a quick rinse. Each sample eluate was dried to a pellet in a Speed-Vac at 75°C. The pellets were rehydrated in 33 µl of VWR

water and samples quantified with a Nanodrop (single strand measurement) to determine their concentration.

Amplification of MGS Eluted Fragments by PCR

The entire reconstituted MGS eluate was amplified using high fidelity polymerase. The forward primer was designed to insert the sequencing primer binding site into the adaptor during the amplification process. Each PCR reaction included 5 µl of 10X LA PCR buffer (TaKaRa), 5 µl of 2.5 mM dNTPs mix (TaKaRa), 2 µl of 20 µM FWD LMPCR primer, 2 µl of 20 µM REV LMPCR primer, and 2 µl of LA Taq (5U/µl, TaKaRa), and VWR water to 50 µl volume. The reactions were incubated in a thermocycler at (1) 98°C for 30 sec, (2) 98°C for 10 seconds, (3) 65°C for 30 seconds, (4) 72°C for 30 seconds, (5) Repeat step 2 18 times (18 cycles), then at 72°C for 5 minutes and a final hold at 4°C. Each PCR reaction was transferred into a 1.5 ml tube and purified with the Promega Wizard® SV Gel and PCR Clean-Up. Each column was eluted with 100 µl of water and the sample concentration was determined with the picogreen method.

Cluster Generation of MGS Selected Target DNA

From the picogreen quantification, 0.025 picomoles (in 19 µl of EB buffer) of amplified MGS-selected target DNA template was denatured with 1 µl of 2N NaOH at room temperature for five minutes. The denatured template was then diluted with pre-chilled hybridization buffer, to a final concentration of 4 pM (40 pM for samples NA18508 and NA18523). On the Illumina Cluster Station, 120 µl of each template sample corresponding to 0.47 ng of DNA (1.9 ng for samples NA18508 and NA18523) was loaded onto each lane of a flow cell pre grafted with oligos complementary to the adaptors. Each fragment will hybridize to the grafted oligos and generate a unique cluster through isothermal bridge amplification of a single molecule. Lane five of the flow cell was always used for bacterial Phi-X as a control. After amplification, the bridged cluster was linearized, blocked and denatured. A sequencing primer was then attached to the binding site inserted during amplification.

Single End Resequencing of MGS Selected Target DNA

The flow cell, with MGS targets amplified and primed for sequencing, was transferred onto the Illumina Genome Analyzer (IGA). A 36 cycle step-wise sequencing-by-synthesis process using four-color labeled nucleotides was performed, according to the manufacturer's instructions. Each run generated 300 tile images per lane per cycle (200 tile images for samples NA18508 and NA18523). Each tile contained an average of 19,000 clusters for IGA version 1 (IGA_I) and 74,000 clusters for IGA version 2 (IGA_II).

IGA Image Processing

The data analysis pipeline for the Illumina 1G analyzer was used, without the ELAND option for sequence alignment. This portion consisted of two different modules. The first module (Firecrest) performed analysis of images captured by the instrument by remapping cluster positions. The second module (Bustard) called bases from the image files. Analysis parameters were chosen to extract all sequences without quality filter (QF_PARAMS '(1==1)'), in a format that includes the quality score (fastq format) of each base and is meant to be exportable into other alignment programs. The output of this pipeline consisted of text files containing sequence fragments up to 36 bases.

Assembly and Analysis of IGA Sequences

The open-source software MAQ (<http://maq.sourceforge.net>) was used to map the IGA short reads to a reference sequence (Li *et al.*, 2008). To increase the efficiency of mapping at the

beginning and end of a reference sequence, each segment in the reference genome was padded 75 bases at each end. The padding applied was not used in computing final statistics. The mapping algorithm used in MAQ has been described elsewhere (Li *et al.*, 2008). In brief, MAQ first indexes sequence reads by building multiple hash tables (one table per read) and then scanning the reference genome sequence against the tables. This allows the identification of read positions (hits) that are subsequently scored. By default, the indexing is done on the first 28 bases of each read (assumed to be the most accurate portion). Also, alignments with up to two mismatches of the 28 bases are detected with certainty. For alignment, MAQ scans the reference three times against six hash tables (templates). The use of six templates ensures that only hits of sequence with up to two mismatches are recorded. Finally, MAQ assigns each individual aligned read a mapping quality that represents the *phred*-scaled probability (Ewing and Green 1998) that a read alignment could be wrong. For mapping, assembly, and SNP analysis of this single-end sequencing, the following MAQ parameters were chosen. A maximum mismatch (-n) of two, a minimum mapping quality (-q) of 30, a minimum read depth (-d) of five and a fraction of heterozygotes among all sites (-r) of 0.001. More than 52.3 millions reads were obtained after quality filtering, yielding over two gigabase (Gb) of DNA sequence.

Mapping of IGA Reads

In order to estimate the enrichment obtained with MGS/IGA sequencing, we mapped all of the reads of a given IGS run using the following approach. First, the file containing the IGA reads was first split into smaller files to decrease the time requirement of the analysis. We then use a local version of BLAT (<http://www.soe.ucsc.edu/~kent>) and a '.2bit' file of the human genome to compute the score, percent identity, and the number of mismatches for each IGA sequence read. The results were filtered keeping those that had less than 5 mismatches and the top hits for each read was obtained. Based on these hits, the reads were then separated into three groups, namely those that lie entirely in the ROI, those that lie elsewhere in the genome and those that do not match to the genome. The reads that did not map to the genome or had more than 5 mismatches were then tested if they mapped completely to the Illumina Genomic Analyzer (IGA) adaptors or sequencing primers (both forward and reverse). Thus all of the reads were categorized into 5 groups, namely reads that entirely mapped to the ROI, reads that mapped to other regions of the genome, reads that entirely mapped to the IGA adaptors, reads that entirely mapped to the IGA primers and reads that fell off due to the stringent filtering. The results of this analysis are contained in Supplemental Data 2.

Derivation of Statistics From the Pileup File

A pileup file was generated by MAQ version 0.6.6 for each sample using default parameters. The pileup file header contains the following fields: chromosome, position, reference base, depth, and the bases of the read that cover the position (<http://maq.sourceforge.net>). The mean, median, variance, and standard deviation of the depth of coverage and melting temperature (T_m) of each segment were computed. Padding applied to the reference sequence was accounted for but not used in computing segment statistics. The T_m for each segment was estimated by using a sliding window of 50 bases and a shift of five bases. The mean, variance, and standard deviation of T_m were calculated from the T_m value of each window. T_m was calculated using the model and parameters for oligonucleotides bound to a surface as described in (Vainrub & Pettitt, 2004). Windows that contained 'N' calls were not used in computing T_m. Some segments have a higher standard deviation value than the mean because of a non-normal distribution.

Calculation of Fold Enrichment

Fold enrichment was calculated using the following method. Consider the following variables:

p = proportion of reads that map to a targeted region of interest

g = size of genome (in this case human genome, 3×10^9)

t = size of target region (in this case, 304,000)

x = degree of enrichment

From these variables, we can write:

$$(x * t) / ((x * t) + (g - t)) = p \quad (1)$$

as an expression showing the proportion of reads that map to a genomic region as a function of the degree of enrichment, size of the target region and size of the genome. With some algebra, this can be rearranged to solve for x :

$$x = (p * (g - t)) / ((1 - p) * t) \quad (2)$$

Comparison and IGA and HapMap Data

For each HapMap sample, we compared our IGA basecalls with genotype calls generated by the HapMap project (www.hapmap.org) using a program developed in house. Only positions for which bases were called in both HapMap and IGA were used to calculate the basecalling rate (completeness score for HapMap and IGA) and identify discrepancies (mismatches between the two technologies). The accuracy of the IGA sequence was determined assuming that the HapMap data was 100% accurate. Our program also reported homozygous and heterozygous sites called by both HapMap and IGA or by either one separately. The final results were updated with the data from the validation of discrepancies.

Validation Sequencing

Discrepancies between IGA and HapMap data were evaluated using by traditional Sanger method of sequencing in the forward and reverse direction (Agencourt Biosciences). PCR primers were chosen using in-house primer picking software (unpublished). PCR reactions were composed of 400 ng of sample DNA mixed with 8 μ l of dNTP mix (TaKaRa), 5 μ l of 10X LA Taq buffer (TaKaRa), 1.5 μ l LA Taq (TaKaRa), 0.8 μ l of each forward and reverse primer and VWR water to 50 μ l total volume. DNA was amplified using the following parameters: 94°C for 4 min, 30 cycles of 94°C for 20 sec, 58°C for 1 min, and 72°C followed by 72°C for 5 minutes. The primers that amplified the SNP discrepancies are listed in Supplementary Table 1. PCR products were run on a 1% TAE agarose gel, excised from the gel, purified using the Promega Wizard® SV Gel and PCR Clean-Up System, and sent to Agencourt. Each chromatogram was interrogated manually for confirmation of the SNPs in question.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institutes of Health/National Institute of Mental Health and Gift Fund grant MH076439 (MEZ), the Simons Foundation Autism Research Initiative (MEZ), and in part by PHS Grant (UL1 RR025008, KL2 RR025009 or TL1 RR025010) from the Clinical and Translational Science Award program, National Institutes of Health, National Center for Research Resources.

References

- Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, Weinstock GM, Gibbs RA. Direct selection of human genomic loci by microarray hybridization. *Nat Methods*. 2007; 4:903–5. [PubMed: 17934467]
- Bashiardes S, Veile R, Helms C, Mardis ER, Bowcock AM, Lovett M. Direct genomic selection. *Nat Methods*. 2005; 2:63–9. [PubMed: 16152676]
- Bau, Schracke N.; Kranzle, M.; Wu, H.; Stahler, PF. Targeted next-generation sequencing by specific capture of multiple genomic loci using low-volume *Analytical and Bioanalytical Chemistry*. 2008
- Bentley D, Balasubramanian S, Swerdlow H, Smith G, Milton J, Brown C, Hall K, Evers D, Barnes C, Bignell H, Boutell J, Bryant J, Carter R, Keira Cheetham R, Cox A, Ellis D, Flatbush M, Gormley N, Humphray S, Irving L, Karbelashvili M, Kirk S, Li H, Liu X, Maisinger K, Murray L, Obradovic B, Ost T, Parkinson M, Pratt M, Rasolonjatovo I, Reed M, Rigatti R, Rodighiero C, Ross M, Sabot A, Sankar S, Scally A, Schroth G, Smith M, Smith V, Spiridou A, Torrance P, Tzonev S, Vermaas E, Walter K, Wu X, Zhang L, Alam M, Anastasi C, Aniebo I, Bailey D, Bancarz I, Banerjee S, Barbour S, Baybayan P, Benoit V, Benson K, Bevis C, Black P, Boodhun A, Brennan J, Bridgham J, Brown R, Brown A, Buermann D, Bundu A, Burrows J, Carter N, Castillo N, Chiara E, Catenazzi M, Chang S, Neil Cooley R, Crake N, Dada O, Diakoumakos K, Dominguez-Fernandez B, Earnshaw D, Egbujor U, Elmore D, Etchin S, Ewan M, Fedurco M, Fraser L, Fuentes Fajardo K, Scott Furey W, George D, Gietzen K, Goddard C, Golda G, Granieri P, Green D, Gustafson D, Hansen N, Harnish K, Haudenschild C, Heyer N, Hims M, Ho J, Horgan A, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008; 456:53–59. [PubMed: 18987734]
- Consortium IH. A haplotype map of the human genome. *Nature*. 2005; 437:1299–320. [PubMed: 16255080]
- Consortium IH. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007; 447:661–78. [PubMed: 17554300]
- Cutler DJ, Zwick ME, Carrasquillo MM, Yohn CT. High-Throughput Variation Detection and Genotyping Using Microarrays. *Genome Research*. 2001
- Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Sun W, Wang H, Wang Y, Xiong X, Xu L, Wayne MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 2007; 449:851–61. [PubMed: 17943122]
- Gnirke A, Melnikov A, Maguire J, Rogov P, Leproust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol*. 2009; 27:182–9. [PubMed: 19182786]

- Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, McCombie WR. Genome-wide in situ exon capture for selective resequencing. *Nat Genet.* 2007; 39:1522–7. [PubMed: 17982454]
- Kleinjan DA, Van Heyningen V. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am J Hum Genet.* 2005; 76:8–32. [PubMed: 15549674]
- Krishnakumar S, Zheng J, Wilhelmy J, Faham M, Mindrinos M, Davis R. A comprehensive assay for targeted multiplex amplification of human DNA sequences. *Proceedings of the National Academy of Sciences of the United States of America.* 2008; 105:9296–301. [PubMed: 18599465]
- Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research.* 2008; 18:1851–1858. [PubMed: 18714091]
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. Genome sequencing in microfabricated high-density picolitre reactors. *Nature.* 2005; 437:376–80. [PubMed: 16056220]
- Mccarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shaperro MH, De Bakker PI, Maller JB, Kirby A, Elliott AL, Parkin M, Hubbell E, Webster T, Mei R, Veitch J, Collins PJ, Handsaker R, Lincoln S, Nizzari M, Blume J, Jones KW, Rava R, Daly MJ, Gabriel SB, Altshuler D. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet.* 2008; 40:1166–74. [PubMed: 18776908]
- Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME. Microarray-based genomic selection for high-throughput resequencing. *Nat Methods.* 2007; 4:907–9. [PubMed: 17934469]
- Olson MDLD. Enrichment of super-sized resequencing targets from the human genome. *CTYP- 3. Nat Methods.* 2007; 4:891–2. [PubMed: 17971778]
- Porreca G, Zhang K, Li J, Xie B, Austin D, Vassallo S, Leproust E, Peck B, Emig C, Dahl F, Gao Y, Church G, Shendure J. Multiplex amplification of large sets of human exons. *Nature Methods.* 2007; 4:931–936. [PubMed: 17934468]
- Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet.* 2001; 69:124–37. [PubMed: 11404818]
- Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet.* 2002; 11:2417–23. [PubMed: 12351577]
- Raychaudhuri S, Remmers EF, Lee AT, Hackett R, Guiducci C, Burt NP, Gianniny L, Korman BD, Padyukov L, Kurreeman FA, Chang M, Catanese JJ, Ding B, Wong S, Van Der Helm-Van Mil AH, Neale BM, Coby J, Cui J, Tak PP, Wolbink GJ, Crusius JB, Van Der Horst-Bruinsma IE, Criswell LA, Amos CI, Seldin MF, Kastner DL, Ardlie KG, Alfredsson L, Costenbader KH, Altshuler D, Huizinga TW, Shadick NA, Weinblatt ME, De Vries N, Worthington J, Seielstad M, Toes RE, Karlson EW, Begovich AB, Klareskog L, Gregersen PK, Daly MJ, Plenge RM. Common variants at CD40 and other loci confer risk of rheumatoid arthritis. *Nat Genet.* 2008; 40:1216–23. [PubMed: 18794853]
- Shendure J, Mitra RD, Varma C, Church GM. Advanced sequencing technologies: methods and goals. *Nat Rev Genet.* 2004; 5:335–44. [PubMed: 15143316]
- Shendure J, Porreca GJ, Reppas NB, Lin X, Mccutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science.* 2005; 309:1728–32. [PubMed: 16081699]
- Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC, Maskeri B, Hansen NF, Schwartz MS, Weber RJ, Kent WJ, Karolchik D, Bruen TC, Bevan R, Cutler DJ, Schwartz S, Elnitski L, Idol JR, Prasad AB, Lee-Lin SQ, Maduro VV, Summers TJ, Portnoy ME, Dietrich NL, Akhter N, Ayele K, Benjamin B, Cariaga K, Brinkley CP, Brooks SY, Granite S, Guan X, Gupta J, Haghighi P, Ho SL, Huang MC, Karlins E, Laric PL, Legaspi R, Lim MJ, Maduro QL, Masiello CA, Mastrian SD, Mccloskey JC, Pearson R, Stantripop S, Tiongson EE, Tran JT, Tsurgeon C, Vogt JL, Walker MA, Wetherby KD, Wiggins LS, Young AC, Zhang LH, Osoegawa K, Zhu B, Zhao B, Shu CL, De

- Jong PJ, Lawrence CE, Smit AF, Chakravarti A, Haussler D, Green P, Miller W, Green ED. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*. 2003; 424:788–93. [PubMed: 12917688]
- Vainrub A, Pettitt BM. Theoretical aspects of genomic variation screening using DNA microarrays. *Biopolymers*. 2004; 73:614–20. [PubMed: 15048785]
- Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, Ma L, Li G, Yang Z, Zhang G, Yang B, Yu C, Liang F, Li W, Li S, Li D, Ni P, Ruan J, Li Q, Zhu H, Liu D, Lu Z, Li N, Guo G, Zhang J, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su Y, San A, Ping C, Yang S, Chen F, Li L, Zhou K, Zheng H, Ren Y, Yang L, Gao Y, Yang G, Li Z, Feng X, Kristiansen K, Wong G, Nielsen R, Durbin R, Bolund L, Zhang X, Li S, Yang H, Wang J. The diploid genome sequence of an Asian individual. *Nature*. 2008; 456:60–5. [PubMed: 18987735]
- Zwack ME, Cutler DJ, Chakravarti A. Patterns of Genetic Variation in Mendelian and Complex Traits. *Annu Rev Genomics Hum Genet*. 2000

Figure 1. Microarray-based Genomic Selection (MGS). Genomic DNA is fragmented, followed by adaptor ligation. Adaptors are identical to Illumina Genome Analyzer (IGA) adaptors. Ligated fragments are hybridized to a custom MGS array for 60 hours. Fragments that do not bind to the array are removed through a series of washes and the bound fragments are eluted in water. The eluted fragments are amplified with a single PCR reaction using IGA PCR primers. The amplified product is then processed for sequencing using Illumina's protocols.

Figure 2. Genomic Region and Fragment Size. a) Graphical display of 1.7-Mb genomic region on chromosome X with RefSeq genes (in dark blue) and the unique regions targeted on MGS array (in purple). b) Distribution of selected fragments by size. Fragments range from 149 bp to 7.29 kb with a mean of 1.48 kb and a median of 1.06 kb.

Figure 3. Median Coverage. a) Distribution of median sequencing coverage across all fragments. b) Distribution, by sample, of fragments with median coverage less than 5x.

Figure 4. Relationship of Median Coverage with Fragment Size and GC Content. Median coverage as a function of a) fragment size and b) GC content (regression lines in red).

Figure 5.

Data Completeness and Accuracy. The blue lines present data completeness as a function of the minimum depth of sequence coverage at all bases (square) and at segregating sites (circle). The red lines present genotype accuracy at all sites (circle), homozygous sites (diamond) and heterozygous sites (triangle) as a function of the minimum depth of sequence coverage.

MGS/IGA Sequencing Coverage and Fold Enrichment

Table 1

Sample ID	Mean Coverage	Median Coverage	Percent Reads Mapped to Region(IGA Mapped Reads)	Fold Enrichment (IGA Mapped Reads)	Percent Reads Mapped to Target Region (All IGA Reads)	Fold Enrichment (All IGA Reads)
NA07000	184.9	134.5	9.7%	1059	9.0%	973
NA07055	135.5	94.3	22.8%	2919	19.4%	2374
NA11993	28.0	19.0	8.8%	956	6.7%	708
NA12057	45.8	36.0	11.8%	1326	9.8%	1069
NA12145	39.2	32.0	21.9%	2767	17.4%	2073
NA18502	268.2	248.0	25.4%	3354	23.5%	3039
NA18505	13.2	9.5	9.9%	1085	7.4%	784
NA18508	356.1	270.5	39.6%	6465	36.9%	5775
NA18517	139.9	111.5	21.9%	2765	18.6%	2257
NA18523	269.5	240.0	34.4%	5164	31.3%	4502