Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two base encoding

Kevin Judd McKernan, Heather E. Peckham, Gina Costa, et al.

Genome Res. published online June 22, 2009
Access the most recent version at doi:10.1101/gr.091868.109

SEQUENCE AND STRUCTURAL VARIATION IN A HUMAN GENOME UNCOVERED BY SHORT-READ, MASSIVELY PARALLEL LIGATION SEQUENCING USING TWO BASE ENCODING

McKernan, K.J.*[++1], Peckham, H.E.*[1], Costa, G.L.*[1]., McLaughlin, S.F.[1], Fu, Y.[1], Tsung, E.F.[1], Clouser, C.R.[1], Duncan, C[1]., Ichikawa, J.K.[1], Lee, C.C.[1], Zhang, Z.[2], Ranade, S.S.[2], Dimalanta, E.T.[1], Hyland, F.C.[2], Sokolsky, T.D.[1], Zhang, L.[1], Sheridan J.A.[1], Fu, H.[2], Hendrickson, C.L.[7], Li, B.[2], Kotler, L.[1], Stuart, J.R.[1], Malek, J.A.[3], Manning, J.M.[1], Antipova, A.A.[1], Perez, D.S.[1], Moore, M.P.[1], Hayashibara, K.C.[2], Lyons, M.R.[1], Beaudoin, R.E.[1], Coleman, B.E.[1], Laptewicz, M.W.[1], Sannicandro, A.E.[1], Rhodes, M.D.[2], Gottimukkala, R.K.[2], Yang, S.[2], Bafna, V.[5], Bashir, A.[5], MacBride, A.[4], Alkan, C.[6], Kidd, J.M.[6], Eichler, E.E.[6], Reese, M.G. [4], De La Vega, F.M.[2], Blanchard A.P*[1].

*These authors contributed equally to the project

1. Life Technologies, 500 Cummings Center, Suite 2400, Beverly, MA 01915
2. Life Technologies, 850 Lincoln Center Drive, Foster City, CA 94404
3. Weill Cornell Medical College in Qatar, Doha, Qatar
4.Omicia Inc, 2200 Powell Street, Suite 525, Emeryville, CA 94608
5.University of California, San Diego, 9500 Gillman Drive, La Jolla, CA 92093-0404
6.Department of Genome Sciences, School of Medicine, University of Washington, Seattle, WA 98195
7. New England Biolabs, 240 County Road, Ipswich, MA 01938-2723

[++]Corresponding Author: Kevin McKernan, Life Technologies, 500 Cummings Center, Suite 2400, Beverly, MA 01915; Kevin.McKernan@appliedbiosystems.com

Running Title: SOLiD™ Sequencing of a Yoruba Genome

Keywords: Massively Parallel Sequencing, Ligation, Two Base Encoding, Color Space, Human Genome, Structural Variation, Haplotype Phasing, SNPs, Personal Genomes

ABSTRACT

We describe the genome sequencing of an anonymous individual of African origin using a novel ligation based sequencing assay that enables a unique form of error correction that improves the raw accuracy of the aligned reads to >99.9% allowing us to accurately call SNPs with as few as 2 reads per allele. We collected several billion mate-paired reads yielding ~18× haploid coverage of aligned sequence and close to 300× clone coverage. Over 98% of the reference genome is covered with at least one uniquely placed read and 99.65% is spanned by at least one uniquely placed mate-paired clone. We identify over 3.8 million SNPs, 19% of which are novel. Mate-paired data is used to physically resolve haplotype phases of nearly 2/3 of the genotypes obtained and produce phased segments of up to 215 kb. We detect 226,529 intra-read indels, 5,590 indels between mate-paired reads, 91 inversions and 4 gene fusions. We use a novel approach for detecting indels between mate-paired reads that are smaller than the standard deviation of the insert size of the library and discover deletions in common with those detected with our intra-read approach. Dozens of mutations previously described in OMIM and hundreds of non-synonymous single-nucleotide and structural variants in genes previously implicated in disease are identified in this individual. There is more genetic variation in the human genome still to be uncovered and we provide guidance for future surveys in populations and cancer biopsies.

The data has been submitted to NCBI. The accession number at the Short Read Archive is indicated in the manuscript. Supplementary material is also provided.

INTRODUCTION

The sequencing of the human genome provided the springboard to understand the role of genetic variation on disease and human evolution (Lander et al. 2001; Venter et al. 2001). More recently the HapMap project surveyed the frequencies of approximately 4 million highly ascertained single nucleotide variants on a sample of four major human populations (International Hapmap Consortium 2005; Frazer et al. 2007), providing useful background information to design and analyze genetic association studies for common variants. However, a hypothesis free view of the full complement of genetic variants on individual genomes has been explored in only a few instances (Levy et al. 2007; Bentley et al. 2008; Ley et al. 2008; Wang et al. 2008; Wheeler et al. 2008), mostly due to cost and time constraints. Yet, these five studies, together with other recent population studies (Redon et al. 2006; Korbel et al. 2007; Kidd et al. 2008), have revealed a much more dynamic picture of genomic variation in the human species, suggesting that the prevalence of structural variation has been largely underestimated. Therefore, the picture of human genomic variation is far from complete and the introduction of second generation, high throughput and inexpensive sequencing opens the possibility of sequencing hundreds to thousands of human genomes to gain a better understanding of the full extent of human variation (Kaiser 2008). Here, we introduce the use of massively parallel short-read sequencing by ligation of mate-paired and fragment libraries to uncover single nucleotide, insertion and deletion (indel) and inversion variants of the genome from an anonymous donor from Yoruba in Ibadan, Nigeria, which was part of the International HapMap Project (Frazer et al. 2007). We also explore the benefits of a novel error correction system that aids in SNP discovery.

Several techniques for massively parallel DNA sequencing have recently been described (Ronaghi et al. 1999; Brenner et al. 2000; Braslavsky et al. 2003; Margulies et al. 2005; Shendure et al. 2005; Ju et al. 2006; Gibbs et al. 2007; Bentley et al. 2008; Eid et al. 2009). They broadly fall into two assay categories (polymerase or ligase based) and two detection categories ("asynchronous single molecule" and "synchronous multi-molecule" or often referred to as "ensemble" read outs). SOLiD™ (Sequencing by Oligo Ligation Detection) sequencing, the method used here, is a DNA ligase based synchronous ensemble detection method utilized to read 500 million to over 1 billion reads per instrument run (Cloonan et al. 2008; Valouev et al. 2008).

All of these techniques are theoretically compatible with mate-paired sequencing but they differ in how they generate the mate-paired reads. Campbell utilizes an approach that generates short pairs from cluster PCR colonies (Campbell et al. 2008) often referred to as "paired-ends". These paired-end reads have limited insert sizes due to the efficiency and representation of PCR amplification of long amplicons via cluster PCR. Consequently very few paired-end reads are generated that are longer than a Sanger capillary electrophoresis read (<10× clone coverage in pairs >1.0 kb). Korbel et al. and Bentley et al. utilize DNA circularization and random shearing circumventing the need to PCR the entire pairing distance at the cost of more input DNA (Korbel et al. 2007; Bentley et al. 2008). These pairs each differ substantially in their tag length due to the random shearing step. These asymmetrical tag lengths reduce the pairing efficiency and often contaminate the library prep with a high number of 200 bp inserts, thus no more

than 100× clone coverage is obtained and many tags are sequenced that are not paired or are paired in the wrong distance or orientation. Furthermore, these techniques may result in many inverted molecules that complicate the detection of inversions.

The preferred pairing method would provide both high sequence coverage and high clone or "physical" coverage with flexible insert sizes such that SNPs, small indels, larger structural variations and copy number polymorphisms (CNVs) could be surveyed in one method. Here we utilize two pairing methods that retain less variable tag lengths while enabling both high sequence coverage (14.9×) and high clone coverage (297×) of the human genome to enable the broadest survey of variation possible.

Use of ligases for massively parallel short-read DNA sequencing of human genomes offers several unique attributes next to polymerases. Most notable is the use of an error correcting probe labeling scheme (2-base encoding or 2BE) which provides error correction concurrent with the color called alignment of the data (i.e., without having to re-sequence the reads). This correction property has specific utility in bisulfite sequencing, de novo assembly, indel detection and SNP detection, but in this manuscript we will focus only on the SNP detection attributes.

Here, we demonstrate that SOLiD sequencing is capable of efficiently surveying single nucleotide polymorphisms and many forms of structural variation concurrently at relatively modest coverage levels. The unprecedented clone coverage allows us to uncover a significantly larger number of structural variants in a size range not efficiently

explored in previous studies, helping to complete the picture of functional variants in this genome. The observed pattern of putatively functional genetic variation in the Yoruba genome recapitulates population signatures of natural selection and suggests a higher than expected individual load of potentially deleterious variants in the human genome.

RESULTS

Sequencing performance with short reads

We generated 76.53 Gb of mate-paired reads that align to the reference human genome (hg18) of which 49.07 Gb have both tags mapping to the genome as a uniquely placed pair. 46.55 Gb of these pairs align at the expected distance, order and orientation and of these 42.56 Gb are comprised of mate pairs in which at least one tag has a unique start point and are therefore derived from unique and unamplified molecules (Table S1). In addition, we generated "fragment" reads (45-50bp) of which approximately 10.5 Gb and 8.6 Gb align and align uniquely to hg18, respectively.

Overall we cover on average 17.9-fold the haploid genome length. While the genomic regions covered by uniquely placed mate-paired and fragment reads overlap substantially (80.35% of the covered regions), the mate-paired reads are mapped to an additional 19.59% of the genome not accessible by the fragments, whereas there is only 0.06% of the genome covered exclusively with fragments.

Over 98.57% and 96.58% of the autosomal genome is covered by at least 1 and 3 reads, respectively, while over 99.6% and 95% is spanned by a single clone and 124 clones, respectively, when we require that a mate pair be placed uniquely in the genome (Figure 1). When we remove this criteria of uniqueness, over 99.8% and 99.0% of the autosomal genome is covered by at least 1 and 3 reads, respectively, while 100% and 95% of the genome is spanned by a single clone and 212 clones, respectively. Ns in the reference

assembly are excluded. Further investigation of the increased coverage gained from mate-paired libraries demonstrates that a more comprehensive sampling of the human genome is achieved with uniquely placed mate pairs than with the unique placement of each of the independent tags (Figure 2). The coverage per chromosome is lower on the sex chromosomes as expected. While theoretically mapping all possible 25-mers allowing up to 2 mismatches uniquely back to the autosomal reference leaves 75.38% of the reference covered, or "mappable", we cover 77.1% of the genome with 25-bp reads allowing up to 2 mismatches. This performance is likely due to the presence of sequence that is missing from the reference. Color space alignments yield greater specificity than base space in which the genome is 70.86% uniquely mappable with 25-mers allowing up to 2 mismatches. The mate pair coverage nearly matches what is expected based on the estimated mappability of mate-paired tags (using a constant insert size and sampling every 25 bases).

The uncovered genomic regions tend to be of low GC content and enriched in repetitive sequences, in particular long repeats and segmental duplications (Table S2). This is expected since it is difficult to uniquely map short reads to segmental duplications and other repeats – when the requirement of uniqueness is removed these features are much less prominent or the correlation reverses. These uncovered regions also tend to be enriched in dbSNP entries, which suggests many of these are likely to be pseudo-SNPs generated by assembly or mapping artifacts as previously reported (Tuzun et al. 2005). Longer repeats are particularly enriched in areas not covered by uniquely placed mate-paired reads and this is certainly due to an inability to cross over the repeat at the insert

sizes used to generate the libraries in this work (the longest being 3.5 kb) and suggests that using longer inserts would help to overcome this limitation.

To assess the sampling of sequence tags across the genome we plotted the average sequence coverage across the genome according to GC-content (see Methods) along with the 10^{th} and 90^{th} centiles and compared these to what is expected according to the Poisson distribution. While the 2×25 mate pairs and fragments lack the most in low GC-content and the 2×50 mate pairs lack the most in high GC-content, the combination of all 3 library types compensates for each other and has good coverage in all but the most extreme GC-content (Figure S4).

Identification of SNPs

The NA18507 genome has not been Sanger sequenced to more than 0.5× sequence coverage (Kidd et al. 2008). However, it has been extensively genotyped as part of the HapMap project (Frazer et al. 2007) and some regions shotgun sequenced to higher depth as part of the ENCODE project (Birney et al. 2007). As a result, false negatives can only be approximated by estimating how much of the hg18 genome is left uncovered (Figure 1) but false positives can be assessed with several techniques. Once reads are uniquely mapped to the genome, sequence variants can be discovered by comparing read sequence to the reference genome, and taking into account the redundancy obtained to distinguish natural sequence variation from sporadic sequencing errors. Using 2 base encoding "2BE", only adjacent sequencing or color differences to the reference can be candidates for single SNPs. Thus the majority (>92%) of our sequencing errors can be eliminated as

putative SNPs. Adjacent color differences that can encode a base change are termed ”valid adjacent” and are explained in more detail in the Supplementary Information on color space.

Across the entire genome, we call 3,866,085 SNPs of which 81% are in dbSNP (release 129). Due to the error correcting qualities of the color space reads, we only need to see each allele twice to call a variant in contrast to other methods that require 3 and 4 alleles (Bentley et al 2008; Wang et al 2008). This allows us to call a comparable number of SNPs in the human genome at lower coverage levels without sacrificing accuracy as suggested by the high dbSNP concordance. To assess the false discovery rate in an unknown genome we performed an analysis similar to Wheeler et al. in which we counted the number of times our sample contains an allele that is not annotated at a bi-allelic loci in dbSNP 129. Using this as a proxy for false discovery, we observe 99.88% of dbSNP loci in our sample are called as one of the two annotated alleles and a non-annotated third allele is only called 0.12% of the time.

Comparing new sequencing technologies to array based genotyping assays that have selected SNPs with high assay conversion rates can be misleading. This does not address our goal of understanding our ability to detect all polymorphisms in a human genome. As a result we choose to focus our laboratory validation on SNPs that are novel as indicated by an absence in dbSNP. To confirm our findings we randomly selected 333 of our 734,662 novel SNP calls (those not found in dbSNP 129) for validation with SNPlex™ genotyping assays (Tobler et. al., 2005). There were 34 assay failures that left us with

299 successful assays. 86% of our novel SNP calls are heterozygous and 94% of the successful assays (280 of the 299) are heterozygous. The SNPlex genotypes are in >95% agreement with the SOLiD detected genotypes. There are 14 cases in which SNPlex disagrees with the SOLiD detected genotype. 2 calls are SOLiD-detected homozygous SNPs which SNPlex calls as heterozygous indicating that SOLiD undercalled the heterozygous state in favor of the variant allele (in both cases SOLiD detects the reference allele but the variant allele is present in greater numbers than the reference allele and the SNP was called a homozygous variant). These calls are expected due to the lower likelihood of sampling both alleles proportionately at fairly low coverage. The remaining disagreements are SOLiD-detected homozygous (1) and heterozygous (11) SNPs in which SNPlex detects the genotypes as homozygous for the reference allele. 12/299 (4.01%) of our novel SNP calls detect one allele that is unsubstantiated by SNPlex and could be considered to be true false-positives. Since our novel SNP calls are made up of approximately 19% of all of our SNP calls, it can be inferred that 0.76% of all of our SNP calls are FPs. 8 of the 11 heterozygous FPs calls are low coverage (5× to 8×) in which the variant allele is on only one strand. If we discount these from the analysis then 1.34% of our novel SNP calls are indicated as FP and it can be inferred that 0.25% of all of our SNPs are FPs. Previous validation of SNP detection with the fragment, 2×25 mate-paired and a different set of 2×50 mate-paired data with manually reviewed Sanger CE sequencing confirmed 111 out of 112 assayed novel heterozygous SNPs that were not in dbSNP.

Our goal is not only to detect SNPs (alternative alleles to the reference), but also to infer whether the sample is heterozygous or homozygous at a given position, the most challenging being the detection of a heterozygous state given the sampling introduced by the shotgun process and the bias induced by mapping to a reference sequence. However, this task is facilitated by the error detection and correction scheme of the SOLiD 2BE sequencing chemistry, which reduces the average sequencing error rate to less than 0.1% (Table S1).

To further assess our ability to call genotypes correctly by sequencing we compared our data with those from the 3.9M SNP calls for this individual in the HapMap r26 data release. Out of the 3,026,465 HapMap homozygous genotypes called, (there are 3,054,399 in HapMap genotyped for NA18507 but some are not called by us due to low coverage or noise; we make calls for >99%) we had an overall 99.16% agreement with HapMap. Removing SNPs with a MAF <5%, where genotypes are less confident in the HapMap due to genotyping assay artifacts (Welch et al. 2008), the agreement climbs to 99.68%.

Higher coverage is required to adequately sample two rather than one allele and thus call a heterozygote rather than a homozygote SNP. Reads which contain variant alleles utilize 2 color mismatches to cover the variant and therefore these reads are allowed less sequencing errors than reads with no variants. As a result more reads match the reference alleles than the variant alleles. Approximately 57% of all reads which map to either allele

at HapMap r26 heterozygous loci map to the reference allele. We are assessing mapping valid adjacent mismatches as a single mismatch to compensate for this.

Figure 3 shows the dependency of homozygous and heterozygous calling with coverage and suggests that both types of calling are highly concordant with HapMap even at modest coverage levels and despite a tendency for reads to overmap to the reference. For the cases where we do not detect a heterozygous SNP at known HapMap r26 loci despite having adequate coverage, we have demonstrated that diBayes, an alternative algorithm, can identify them in most cases.

We call 60.3% of the SNPs we identify as heterozygous and amongst these 67.3% are transitions and 32.7% are transversions - a 2.05 ratio which is very close to expectation for the human genome (Lander et al. 2001; Venter et al. 2001). 27.2% of the heterozygotes we call are novel and amongst these 65.3% are transitions and 34.7% are transversions. SNPs are more densely represented on autosomes than on the X and Y chromosomes.

Single base changes have the potential to be disruptive to gene structure when they are within exons. Amongst the SNPs that we identify in NA18507, 68,624 of them are in a known exon and 16.1% of these SNPs within exons are novel. 12.5% of exons contain at least one SNP and one exon contains 49 SNPs. The distribution of the number of SNPs identified per exon is shown in Figure S6. From these, we identify 9,902 SNPs that produce non-synonymous changes in coding sequences (ns-SNPs). While exons comprise

2.6% of the total genome, only 1.78% of all SNPs we identified are in exons indicating that there is an under-representation of SNPs in exons due to their deleterious potential (Lander et al. 2001; Venter et al. 2001).

Resolving haplotype phases with mate-paired data

Mate-paired data with accurate reads can be used to resolve haplotype phases when two reads in a pair cover alleles at different loci (Kidd et al. 2008). We combined our 3,759,673 genome-wide autosomal reference-variant SNP calls with autosomal HapMap r26 genotypes (which includes NA18507 genotypes for homozygous reference-allele loci) to obtain 6,184,461 distinct potential genotypes. We evaluated the number of mate-paired reads which cover two or more of these loci to establish the upper limit for potentially resolving phases with the 14.89× mate-paired data. We observed 4,027,548 potential genotypes (65.12%) covered by 6,767,943 pairs (1.1% of all mate pairs) with 24.03% of the mate pairs calling SOLiD-detected heterozygous/heterozygous, 35.90% heterozygous/homozygous and 40.07% homozygous/homozygous pairing events. 3,734,035 pairs (55.17%) have both a Forward (F3) and a Reverse (R3) tag covering 3,580,440 distinct genotype loci and the remaining 44.83% have only a F3 or R3 tag covering 1,205,779 distinct genotype loci. Nearly 2/3 of the genotypes for this individual are covered by at least 1 mate-paired read that is in phase with another genotyped location and 43% that we detect as heterozygous are in phase with another that we also detect as heterozygous.

We evaluated all mate pairs which cover HapMap r22 phased heterozygous genotypes and determined that our phases agree with 98.95% of the HapMap phases and we cover 21.74% of the HapMap-phased heterozygous genotypes with at least one mate pair. We also looked for mate pairs in which both a novel heterozygous locus and a HapMap-phased heterozygous locus are covered to determine if the phases of the novel loci are in agreement with the HapMap loci. 76,300 of our novel heterozygotes fall into this category and for the 15,946 pairs where both alleles of the novel heterozygotes are covered the alleles are in opposite HapMap phases compared to each other, as expected, 99.52% of the time. Mate pairs can be collapsed into longer haplotype "blocks" of sizes up to 215Kb (Figure S5). In principle, these physically phased blocks can be used together with HapMap genotype data and statistical phasing algorithms to produce more accurate and complete haplotype phases for this individual.

Intra-read or "split-read" insertions and deletions

The most prevalent class of small insertions and deletions are less than 5bp in length (Mills et al. 2006; Levy et al. 2007) with 57% of the events being single base indels which were undetectable with the approach taken by Wheeler et al. Since SOLiD is a terminating chemistry it can accurately call single and multiple base indels under the sequence read by mapping the indel read to the reference genome. We find 226,529 indels, including 89,679 insertions of up to 3 bases, 124,024 deletions of up to 11 bases, and 12,826 larger indels (Figure 4).

Approximately 67% of the small indels found (insertions up to 3 bp and deletions up to 11 bp) are present in dbSNP and 49% are seen in data from the nine individuals in Kidd et al. that had 2.78× total sequence coverage. The concordance with Kidd et al. drops to 37% and 22% when considering only Yoruba samples (1.46× sequence coverage) and only NA18507 (0.50× sequence coverage), respectively. This low concordance is expected as dbSNP under represents multiple nucleotide variants as seen in Kidd et al., in which 75% of the discovered indels <100bp are novel and the low sequence coverage of Kidd et al. prevents much of the genome from being accessible to small indel detection. Additionally, we find 10,525 insertions of length 4 to 14, with dbSNP and Kidd concordances of 74% and 64%, respectively, and 233 insertions of length 15 to 19, with concordances of 52% and 51%, respectively. We also detect 2,068 deletions from size 12 to 498, with dbSNP and Kidd concordances of 41% and 38%, respectively.

In regards to the impact of small indels (insertions up to 3 bp and deletions up to 11 bp) on gene structure, we find that 2,788 exons contain at least 1 indel while 392 exons have more than 1 indel and 2,241 genes contain an indel within an exon (Figure S6). We observe an over abundance of indels in first and last exons as observed with SNPs. In total 76% of 2,241 genes containing indels have them in the first and last exon (362 in the first, 1,492 in the last, and 159 in their only exon). There is a noticeable preference for even sized indels across the genome due to the existence of dinucleotide and tetranucleotide repeats as reported by Mills and Levy (Figure 4). However, this preference is skewed in favor of 3 and 6 base (3n) indels in coding exons (Figure S7) in which 34.3% of the indels within and 12.7% of the indels outside of coding exons are 3

or 6 bps in size. The observation that indels in coding exons show a size distribution that favors 3 and 6 base codon skipping mutations when the rest of the genome exhibits even sized indels is expected due to purifying selection against frameshifts in coding regions.

Large inter-read insertions and deletions

We have identified 1,515 insertions and 4,075 deletions by clustering of mate-paired reads with discordant distances when mapped to the human reference assembly hg18 (Methods). Deletions range in size from 86 to 96,957 bp and insertions from 30 to 1,287 bp. Figure 5 illustrates the size distributions of the insertions and deletions and highlights the abundance of variations in the size range of Alu and LINE elements. The insertions and deletions have clones with discordant distances spanning them that are deviated by at least 6 standard deviations from expectation at the given level of clone coverage. Figure S8 illustrates the size limit of detection of inter-read insertions and deletions at various levels of clone coverage at 6 standard deviations of significance given our library insert size of 1400 bp with a standard deviation of 199 bp.

Comparing the large deletions to the Venter, Watson and YH genomes, we find that 40% of the deletions >= 200 bp and 30% of the deletions >= 1000 bp have been previously identified in these genomes. Common to all 4 of the genomes are 353 deletions >= 200 bp and 49 deletions >= 1000 bp. These deletions in common in all 4 genomes from 3 ethnically distinct populations indicate the possibility of the presence of a minor allele in the hg18 reference sequence. A pair-wise comparison of the number of deletions identified in each of the 4 genomes is presented in Table S3. While we detect more

deletions in NA18507 than have been detected in the other genomes, in each pair-wise comparison over 20% of the deletions >= 200 bp detected in NA18507 have been detected in the other genome. We also detect over 38% of the deletions >= 200 bp that have been detected in each of the other genomes, significantly more than when considering most of the other pairs of genomes with respect to each other.

Deletions in NA18507 identified by both intra- and inter-read approaches

The long tag mate-paired data allows us to identify some of the same deletions by both the intra- and inter-read approaches. Amongst the 2,068 deletions ranging in size from 12 to 498 bases detected within reads, 193 of them are in common with the deletions identified by discordant read-pair clusters (Figure 6). Amongst these deletions, 60 of them have also been identified in the Venter, Watson and YH genomes. Figure 6 also illustrates a 328 bp deletion that has been identified by both the intra- and inter-read approaches in NA18507 by SOLiD and in each of the other 3 genomes. The ability to identify deletions with both inter- and intra-read approaches bridges the gap between these two methods, and not only allows the corroboration of the identified events but also permits the detection of the full range of sizes of deletions in the human genome.

Inversions

While inversion detection is sensitive to both sequence coverage and clone coverage, the clone coverage is especially important. If a read is in the center of an inversion it is more likely that its mate is flipped with respect to itself the farther away its pair is located and therefore able to detect an inversion breakpoint. We analyzed all mate pairs with both

ends mapped but with sequences on opposite strands (SOLiD mate-paired libraries create mate-paired tags in which both sequence reads are normally on the same strand). We look for multiple mate pairs to exhibit the same inverted orientation at the same coordinates in the genome and across multiple libraries. The supporting evidence of all inversions can be visually inspected using the SOLiD Alignment Browser (Figure S9).

We observe 91 inversions of which 22 are amongst the 90 inversions observed by Levy et al. and 37 are amongst the 72 inversions observed by Korbel et al. in NA18505 (Korbel et al. 2007). Some of the inversions will be missed by short reads due to an enrichment of repeats at the edges of the variations. A complex inversion and deletion is discovered on chromosome 4 (Figure S9) in which a homozygous deletion is nested within a heterozygous inversion indicating a hotspot for variation in which the deletion occurred prior to the inversion. This genomic event is at the same location in which others have observed copy number changes (Redon et al. 2006; McCarroll et al. 2008) and where we have observed a high discordance in our SNP calling compared with the HapMap. The observed homozygous deletion near 190.845Mb of chromosome 4 corresponds to two annotated segmental duplications (chr4:190,844,046-190,845,646 and chr4:190,845,653-190,847,295).

There is potential with mate-paired reads to resolve the phase of these inversions by linking them with SNPs that are identified within the sequenced tags. Among the 68 heterozygous large (>400bp) non-overlapping inversions, 55 (81%) are detected with mate-paired reads that contain SNPs compared to the reference sequence and 49 (72%) of

them contain a heterozygous SNP. 41 of the 49 inversions that are in phase with a heterozygous SNP have more than 1 set of mate-paired reads confirming the phasing. This preliminary data suggests the potential for resolving the phase of all types of structural variations by linking them to heterozygous SNPs.

Copy Number Variations

Through the analysis of changes in depth of coverage in windows, we observe 565 CNVs in the size range of 2 kb to 937 kb in the autosomes. The distribution of size ranges is shown in Figure 7A. We identify 116 out of the 179 CNV events detected with the Affymetrix genotyping array 6.0 (McCarroll and Altshuler 2007). These CNVs are plotted in Figure 7B with the SOLiD-called copy numbers on the X-axis and CGH-called copy numbers in various colors. We demonstrate that there is good agreement between the copy numbers called by the different technologies. When the CNV regions are compared to the deletions detected using our mate-pair clusters, 102 out of 409 1-copy regions are supported by heterozygous deletions, 51 out of 78 0-copy regions are supported by homozygous deletions and there are no 1-copy regions overlapping with homozygous deletions or 0-copy regions overlapping with heterozygous deletions.

Evaluation of sequencing strategies for genetic variation discovery

Whole-genome human sequencing with massively parallel platforms is poised to become the workhorse of population and medical genetic studies. However, little information exists on how to better design these studies and balance cost versus benefit of different sequencing strategies. We thus investigated the fraction of genetic variation discovery at

relatively low levels of average sequence coverage: 2×, 4×, 8× and 10×. Figure 8 shows the percent of the genome that meets the coverage requirements for SNPs and intra-read indels at various levels of average sequence coverage (panel A) as well as the actual number of these variants that are detected (panel B). Significant homozygous SNP discovery can be achieved at 2× with a significant increase at 4×, but only a modest increase at 8×. Determining heterozygote status requires more coverage but at 8× 84.3% of the genome is accessible to heterozygous SNP detection. The identification of small indels under the sequence read is fairly incomplete at these low coverage depths due to more stringent mapping requirements.

To understand the benefits of deeper sequencing in identifying structural variants we assessed the indels detected between mate-paired reads from sets of 1, 2, 3 and 4 slides of long mate-paired data that amount to 2.2×, 4.0×, 5.6× and 8.4× average sequence coverage, respectively. Figure 8c illustrates that the number of insertions and deletions rises steadily from 2.2× to 8.4× average coverage and indicates that further sequencing will enable more variants to be detected. While the full data set in this study is from 8.4× average sequence coverage of long mate-paired data, 37%, 61% and 72% of the deletions >= 200 bp in the 8.4× data set are detected at 2.2×, 4.0× and 5.6× average sequence coverage, respectively. The advantage of deeper sequencing is even larger for insertions in which 16%, 49% and 68% of the insertions >= 200 bp in the 8.4× data set are detected at 2.2×, 4.0× and 5.6× average sequence coverage, respectively.

In regards to the combination of mate-pair libraries and the choice of insert size, we investigated the contributions of multiple insert sizes to the power to detect and resolve large indels by simulation and comparison with our empirical data. Figure S10 shows the theoretical and observed probability of detecting an insertion or a deletion with two different insert sized libraries independently (green and blue lines) and combined (red line). Our results imply that while a larger insert size increases detection, the probability of resolving a breakpoint increases with a combined library approach such that the probability of detecting a break point with a 600 bp and a 2,841 bp library is higher than using either library exclusively. Higher resolution break point mapping is anticipated to greatly facilitate any PCR based validation of next generation results.

Diversity amongst human genomes

We compared the SNPs and structural variations identified in NA18507 to those found in the Venter (Levy et al. 2007), Watson (Wheeler et al. 2008) and YH (Wang et al. 2008) genomes. Figure S11 demonstrates that over 20% of the SNPs found in NA18507 are in each of the other genomes while 20-40% of the SNPs identified in each genome are unique to that genome. Less insertions, deletions and inversions are in common amongst the four genomes and a higher proportion of them are unique to the genome in which they are identified. While it is noteworthy to compare this data it must be understood in the context of what is still yet to be uncovered in each genome. The percent of the NA18507 variants that are in common with the other genomes is a lower bound while the number of variants that are unique to each genome is an upper bound. These values will certainly shift as more variation is uncovered in each of the genomes. Since it is likely

that we have identified most of the SNPs that are present in NA18507 while structural variations are typically more difficult to identify than SNPs with current sequencing technologies and have not yet all been identified, it will be exciting to discover whether the tendency for structural variations to be more distinct to single genomes than SNPs will hold as more of the variations in each genome are revealed or whether this is an artifact of the current state of detection.

Functional consequences of genetic variation in the Yoruba individual genome

To assess disease relevant variations present in the NA18507 sequence as described previously in the literature, we used the disease variants as described in the Online Mendelian Inheritance in Man (OMIM) database (Hamosh et al. 2005; McKusick 2007), a database of gene-disease relationships. Here we investigate only the amino acid allele variants from the OMIM database – a list of 9,239 variants of amino acid and terminator mutations that we can position uniquely and with confidence onto the genomic sequence as described in Methods. We then compare the NA18507 sequence variants to this list encompassing 2,161 human genes found in OMIM as of August 2008. Within these 9,239 mutations, Table S4 shows the list of OMIM variants for which this individual is a carrier of the disease–related allele in the homozygous or heterozygous states. Based on the annotations in OMIM and after reviewing the corresponding literature, we further filtered the homozygous and heterozygous alleles into high, medium and low reliability for annotation (Table S5). In total, NA18507 carries five disease relevant OMIM alleles with stronger evidence in the homozygous form. For all five of these the sequence quality is sufficient to trust the zygosity assignment. We found an additional ten for

which the annotations in OMIM were inconclusive ("medium" reliability). The remaining three overlaps had contradictory or disease irrelevant annotations ("low" reliability). Furthermore, NA18507 carries 49 OMIM alleles in the heterozygous state.

In reviewing the disease associations for the homozygous alleles, each association is with a common, multi-factorial disease including susceptibility to obesity, drug addiction, atopy, thrombocytosis, bladder cancer as well as the trait of slow acetylation. As expected, from this initial analysis it seems none of the many Mendelian disease-causing variants listed in OMIM are found in the NA18507 genome sequence in the homozygous state. In addition, in Table S6 we list all variations in NA18507 that generate an in-frame stop codon in homozygous and heterozygous states and overlap a known SNP in dbSNP.

We studied the impact of large insertions and deletions (detected by the inter-read approach) on gene integrity, by looking for breakpoint regions that overlap with gene regions, defined as the transcription start site to the end of the mRNA transcript including all exons and introns of the gene. We detected 2,477 potential gene disruption events (44.3% of the 5,590 large indels) that fall within gene boundaries in the NA18507 genome sequence, disrupting a total number of 2,015 unique human genes including some that are disrupted by multiple events. Amongst these 2,015 disrupted genes, 303 (15.0%) are contained in a curated collection of 3,600 human disease genes (~15% of all human genes) assembled using a previously published collection of 923 human disease genes (Jimenez-Sanchez et al. 2001), all genes listed in OMIM and in the Human

Genome Mutation Database (HGMD v7.1) (Stenson et al. 2003) as of August 2008 and further extended by comprehensive and rigorous literature review (Table S7). The number of observed events should be treated as an upper bound of the total number of gene overlaps since the boundaries for the large indels are wider than the actual events so the structural variation is either occurring within the gene or in close proximity to it. In summary, we can see a trend for disruption events to cluster around genes, but no clear preference to cluster around disease genes. Further analysis of these disruption events along with an evaluation of whether an exon is disrupted is warranted.

We identify a number of gene pairs that appear to be fused by a structural variation. Based on the empirical distribution of clone lengths, and the positioning of the discordant mate-paired reads, we compute a confidence estimate on the gene fusion event (Bashir et al. 2008). Table 1 reports five events, with the probability of genomic fusion equal to 1, which are supported by at least 3 distinct pairs of reads. Interestingly, all but one of the predicted fusions involve tandem duplicated genes. The distance between the pairs of duplicated fused genes ranges from 16 bp to 368 kb. Two of the gene pairs have been previously observed as chimeras in the literature: *APOBEC3A-APOBEC3B* (Kidd et al. 2007) and *EMR2-CD97* (Chiu et al. 2008). The gene pair *CTRB1-CTRB2* is caused by an inversion in which six of the mate-paired reads that detect the inversion also contain a SNP compared to the reference sequence and of which three of them contain a heterozygous SNP providing the potential to resolve the phase of this fusion.

Signature of purifying selection in the pattern of non-synonymous mutations

We discovered 6,131 non-synonymous SNPs that are pre-annotated using PolyPhen for the expected degree of damage they will cause to the protein (Ramensky, Bork, and Sunyaev 2002). 4,912 of these non-synonymous SNPs are annotated as 'benign' (79.6%); 765 are 'possibly damaging' (12.4%); 454 are 'probably damaging' (7.4%). This compares to the proportions of these categories among all 76,434 non-synonymous SNPs in the PolyPhen database. Of annotated SNPs, 66.2% are 'benign'; 18.9% are 'possibly damaging' and 14.9% are 'probably damaging'. The nonsynonymous SNPs in this Yoruba sample are significantly less damaging than the full collection of dbSNP nonsynonymous SNPs ($p<10^{-5}$). The homozygote state is significantly underrepresented for 'probably damaging' and 'possibly damaging' alleles as compared to 'benign' variants in this individual genome (Table S8).

We investigated whether damaging SNPs were over or under represented in certain protein classes. There are 986 proteins with annotated function in the Panther protein classification database (Thomas et al. 2003) (http://www.pantherdb.org) containing possibly or probably damaging nsSNPs. When comparing with the distribution of proteins in Panther categories of the human proteome, we identify protein families significantly under-represented for damaging SNPs (binomial test, $P<0.05$ with Bonferroni correction), including nucleic acid binding proteins ($p=0.00012$), ligases ($p=0.0053$), transferases ($p=0.0063$), transcription factors ($p=0.0086$) and kinases ($p=0.084$). Categories overrepresented for damaging SNPs include receptors ($p=0.0013$) (especially G_protein coupled receptors, $p=10^{-9}$), extracellular matrix glycoproteins

(p=0.009), cell adhesion molecules (p=0.03), and cytoskeletal protein (p=0.07); as well as biological functions of genes including sensory perception ($p=10^{-11}$) specifically olfaction ($p=10^{-17}$), G-protein mediated signaling (p=0.00057) and cell adhesion-mediated signaling (p=0.054).

DISCUSSION

Benefits of 2BE for error correction and SNP discovery

For human sequencing it is advantageous to have a substitution error rate which is substantially lower than the anticipated substitution polymorphism rate (10^{-3}) such that any single read can be trusted for homozygous SNP detection (Venter et al. 2001). In absence of this level of accuracy one must build confidence based on the overlap of aligned reads. Due to Poisson sampling limitations of shotgun sequencing this overlap usually requires at minimum 3 fold more reads per allele (Lander and Waterman 1988) and thus there is distinct value in higher accuracy read generation to maximize polymorphism detection per Gb of sequence generated.

The 2BE method introduced here provides in theory a 37.5 fold gain in sensitivity for detecting real SNPs next to raw measurement noise (see Supplement "Error Correction" for details and further discussion). We evaluated the raw sequencing error and the remaining error after 2BE correction for all of the runs in this study and we indeed observe a >99.9% average accuracy (Table S1). This gain in accuracy is reflected in the ability to call SNPs and in particular heterozygotes at relatively modest levels of average coverage (see Figures 3 and 9).

One possible draw back to color space is that it requires 2 mismatches per SNP in a given read length. As a result a 50 bp read with 5 SNPs is difficult to align to the human reference as it consumes 10 color mismatches. Unless the alignment tools treat pairs of

valid adjacent mismatches as a single mismatch, such highly polymorphic sequences can exhibit a reference bias with strict reference matching algorithms. Therefore mate pairs are preferred when aligning to polymorphic regions such as the MHC of the human genome since one read can act as an anchor and more relaxed or gapped alignments can be performed with the polymorphic tag. Alternatively, four-frame dynamic programming alignment implemented in tools such as SHRIMP (Lee et al. 2008) can be utilized. A second traversable concern is de novo assembly with color space discussed in the Supplement "Error Correction".

On the sequencing of 100s-1000s of human genomes with short reads

Variation discovery is clearly coverage dependent, asymptotic and in the case of SNPs very sensitive to read error rate. An alternative for projects considering the sequencing of hundreds of genomes where the cost is still a driving factor is to sequence these at a low sequence coverage (e.g. 4-6×) at the expense of partial genetic variation discovery. For structural variation discovery by discordant mate-paired clones, the relevant parameter is clearly clone coverage and this is easier to achieve with larger inserts at equivalent sequencing cost. We have demonstrated the number of homozygous and heterozygous SNPs as well as intra- and inter-read indels that are detected at a variety of coverage levels and how these compare to what is detected with our full data set of ~18× sequence coverage. We show that it is possible to physically phase clusters of alleles interconnected by mate-paired reads, providing a significant amount of information that should improve complete haplotype resolution by statistical methods. This information can be used as a guide when planning experiments to understand how much of the

genome will be accessible to each type of variant given the average level of sequence coverage.

We believe these data support the recent priorities in building a more comprehensive human reference sequence which better captures these forms of variation which whole genome sequence assemblers may have condensed or under surveyed in regions of the reference which were haploid derived (BACs and PACs) (Gresham and Kruglyak 2008). Algorithms can improve variant detection substantially if they are informed of regions of common CNVs or structural variations.

Comparison with other personal genomes

It is tempting to compare the individual African genome reported here and the two previously reported genomes of European decent in terms of the patterns of genetic variation. Our expectation is to find more genetic variants in a heterozygous state in the African vs. European genomes (Li et al. 2008). However, there are several factors that complicate this comparison. First, the levels of coverage between these studies varies significantly, in part due to the cost, but also due to length of the reads underlying sequencing technologies, 500-1000 bp in Venter, 200 bp in Wheeler, and 25-50 bp here. Clearly, shorter reads have more difficulty spanning different types of repeats in the human genome and could be more susceptible to chemistry biases due to local sequence composition (e.g., GC content). Some of these limitations can be overcome by deeper coverage (Venter and Wheeler 7×, Bentley 30×, our work 18×), which at the same time renders the comparison challenging. In addition, the different sequencing platforms used

in these studies have different error patterns and rates (0.01% for Sanger (Shendure and Ji 2008), 3% for pyrosequencing (Quinlan and Marth 2007), 0.1% for SOLiD, 1-2% for Illumina/Solexa (Hillier et al. 2008), error rates for all platforms continue to improve and these numbers represent an ephemeral comparison), which together with the shotgun sampling statistics and coverage biases results in different demands for calling SNPs and genotypes.

Furthermore, the power to detect other types of genetic variation varies widely between studies due to the different sequencing strategies used: in Venter indels were detected by read overlap or comparisons to the reference genome which restricted discovery to the 1-100 bp range; in Wheeler reads were aligned to the reference genome and thus restricted discovery of indels to much shorter than the read length; and in our case we utilized read overlap with the reference in addition to mate-paired mapping analysis to infer larger inter-read indels, which is coverage dependent and thus gave us power on the 100-100,000 bp range. Therefore, differences between these studies are more likely reflective of the power of the different methods involved rather than population of origin differences and simply reflect the ranges of unexplored variation by former studies.

On the signatures of natural selection and demography in a single genome

A question of considerable interest that could be addressed with data from individual genomes is the number of deleterious mutations per human genome. Based on population genetic theory, it has been suggested in the past that the number of lethal mutations per individual should be extraordinarily low (Morton et al. 1956). Lohmueller et al. recently

addressed this question based on a sequencing survey of the exons of over 20,000 genes in 20 European and 19 African American individuals (Lohmueller et al. 2008). Their results suggest that the number of potentially damaging mutations per individual (as predicted by PolyPhen) is much higher than expected, on the order of several hundreds, and that the population of European decent has more probably damaging mutations than that of African decent, presumably due to the population bottleneck suffered by the former. Our results are consistent with these findings.

We demonstrate that in the genome of this Yoruba individual, over 50 mutations previously implicated with disease, as well as over 1,500 putatively deleterious mutations (as predicted by PolyPhen) and 2,000 possible gene disruptions events (300 in genes previously implicated with disease), are present. Potentially damaging SNPs are under-represented in exons of genes with molecular functions that are essential for cell survival, but over-represented in exons of genes undergoing rapid evolution in human populations. Many of these biological functions, most notably olfaction and immunity, have previously been demonstrated to be over-represented among gene variants that differ in frequency among human populations, suggesting at minimum a relaxation of purifying selection and perhaps directional selection and/or an increase in mutation rate. Our observation that both SNPs and indels tend to reside outside of exons, and that both disease susceptibility alleles and putatively deleterious mutations are present in a homozygous state less than expected, is consistent with the operation of purifying selection.

Conclusions

With the rapid improvements in next generation sequencing, at the time of this writing, the data for this study could be generated in just one to two 30-50 Gb runs from a SOLiD instrument at an estimated reagent cost of under $30,000. The time to analyze such large data sets is not keeping pace with these increases in data generation and we anticipate much pioneering work ahead on whole genome sequence analysis. We have placed many of the provisional analysis tools as open source tools on the web and encourage critique and improvement to this software (http://solidsoftwaretools.com).

Previous studies have considered mostly SNPs as sources of deleterious mutations; however, it is becoming clear that structural variation can have functional implication in gene integrity and function (Kidd et al. 2008). Here we are able to complete the landscape of potentially deleterious variation by considering insertion and deletion events that damage genes and/or elicit potential gene fusions. These results suggest that it is important to consider structural variation in determining the potential disease alleles in a genome and population studies.

It seems that even in a single human genome, the signatures of natural selection and human demography are ever present, and that the exploration of personal genomes has significant potential to be of importance in healthcare and personalized medicine. Our studies provide guidance for future exploration of human genetic variation with ultra-high throughput short-read sequencing technologies such as SOLiD and confirm that accuracy is an important factor that interplays with throughput in determining the cost-

effectiveness of the new sequencing methods in whole human re-sequencing. As with the initial sequencing of the human genome, it appears longer-range mate pairs continue to provide structure and phasing information of significant value to understand personal genomes.

METHODS

DNA sample. We sequenced the genome of the Yoruba sample NA18507, obtained from the Coriell Institute. This sample has been previously consented for genomic research by the International HapMap Project (Frazer et al. 2007).

Sequencing. We sequenced the genomic DNA of the sample using a combination of mate-paired libraries and fragment libraries with the Applied Biosystems SOLiD™ System analyzer according to the manufacturers' instructions. Two different methods for generating mate-paired libraries were utilized; TypeIII generated libraries and Nick Translation libraries.

TypeIII Libraries

Briefly, using the method described by Smith et al. (Smith et al. 2006) we generated mate-paired libraries with the TypeIII restriction endonuclease EcoP15I (Applied Biosystems SOLiD Library Preparation Guide).

Nick Translation Libraries

To generate paired 50 base tags, the EcoP15I cap adaptors were left dephosphorylated so that circularization of the target DNA left a nick on the 3' ends of the internal adaptor. These nicks were bi-directionally extended into the insert DNA using a timed nick translation reaction. Tags were liberated with S1 nuclease, end repaired with the Epicentre Endit kit and varied in size from 50-75bp per tag. All libraries were primer

ligated with T4 DNA ligase (Ambion) and utilized identical adaptors (P1 and P2, Applied Biosystems SOLiD Library Oligos #4392456) for emulsion PCR (Applied Biosystems Long Mate Pair Library Protocol).

Fragment Libraries

Additionally, we generated sheared "fragment" libraries that were sequenced as unidirectional reads. Briefly, fragment libraries were generated by shearing genomic DNA to a 60-90 bp range using various shearing methods (DNaseI, Nebulization, and adaptive focused acoustic bombardment with a Covaris S2) and end repairing the DNA.

Emulsion PCR was performed according to Dressman et al. (Dressman et al. 2003) with a few minor modifications (Supplement Emulsion Methods). Since limited dilution of DNA is utilized to produce clonal bead amplification, 70-80% of the beads in any given emulsion are un-amplified beads. An enrichment step is performed to select for the templated beads and provide a higher number of sequence generating features per run. Enrichment of amplified beads was performed as previously described (Shendure et al. 2005) with a few modifications (Supplement Enrichment). Once emulsions are broken the beads are enriched, end modified and deposited on a microscope slide ready for SOLiD sequencing (Supplement End Modification and Deposition).

Ligation sequencing is performed in five different frames of sequencing. As a result five different 5'phosphorylated primers that are each offset by 1 base with respect to each other are used. The detection probes have a cleavable phosphorothiolate linkage fixed

between the 5th and 6th base such that sequencing with 1 primer generates partial dinucleotide information in 5 base increments. Primer 1 will survey dinucleotides 1,2 and 6,7 and 11,12 and so on to bases 46 and 47. Primer 2 will survey dinucleotides 0,1 and 5,6 and 10,11….45,46. Primers 3, 4 and 5 will be nested more than 2 bases into the known adaptor sequence and thus do not require their 1st ligation cycle to imaged. (Supplementary Figure S1).

SOLiD sequence alignment. The AB SOLiD alignment tool, mapreads, translates the reference sequence to dibase encoding (color space) and aligns the reads in color space. The program guarantees finding all alignments between a read and the reference sequence with up to M mismatches (a user specified parameter). It uses multiple spaced seeds (discontinuous word patterns) to achieve a rapid running time (Zhang et al 2009, in preparation). Reads which align in only one location in the color space reference with up to the given number of mismatches are referred to as uniquely aligned.

Representation of reads in terms of GC-content. The %GC of the human genome is calculated for each 250 bp window. The average coverage depths of the windows are grouped by associated GC content and ranked within the groups. The mean of each group defines a Poisson distribution and the 10th and 90th centiles are compared with those of actual data.

SNP identification. Single nucleotide polymorphisms are initially identified using the SOLiD Consensus Calling algorithm. SNPs are called by a consensus of valid adjacent

2-base encoded mismatches. The confidence of each base call is determined by the type of call in color space, the position in the read and the 6-mer base space context in which the base call occurs and this confidence is used to weight the contribution of each set of adjacent base calls to the consensus call. The SNPs are further filtered to eliminate all variants with coverage > 3 times the mean, variants amongst 3 SNPs called in a 10 bp window and variants within 15 bp of an intra-read indel that we identify in this analysis.

We also present SNPs called with another SOLiD SNP detection algorithm called diBayes. diBayes is a Bayesian algorithm that includes color space error detection (Hyland, F.C.L. et al., in preparation), an error model that uses probe and positional errors as well as color quality values, and the prior probability of population heterozygosity, in a framework similar to that of PolyBayes (Marth et al. 1999).

SNP validation. To confirm our findings we randomly selected 333 of our novel autosomal SNP calls (those not found in dbSNP 129) for validation using the SNPlex Genotyping System (Tobler et al. 2005). There were 34 assay failures that left a total of 299 successful assays in eight multiplex pools. The genotyping was performed as described (Tobler et al. 2005) on a panel of 46 Coriell DNA samples (22 African American, and 24 Caucasian samples) to populate genotype clusters and the Yoruban DNA sample NA10859.

Identification of small insertions and deletions under the sequencing tag. Indels can be detected within the actual sequencing reads (intra-read indels) or by observing the

expected clone sizes to stretch or compress (inter-read indels). Intra-read indels have the benefit of single base resolution of the variation while inter-read indels are less precise on the coordinate of the event. We surveyed indels with one end anchored (OEA) mate pairs (Kidd et al. 2008) in order to increase the accuracy since the search range for the unanchored tag is drastically reduced compared to searching the entire human genome.

Small insertions (up to 3 bp) and deletions (up to 11 bp)

Using mate-paired libraries, we realign our OEA pairs using the anchored pair as a seed and perform a more aggressive alignment with the other tag in a several kb window (depending on the insert size of the library) from the anchored mate. Using the unanchored tag we align both ends of the read until the maximum number of two mismatches for 2×25-mers or five mismatches for 2×50-mers occur. Disallowing for indels 1 or 2 bases of either end of the read (not including the first base of the read), we identify if we are able to piece together both ends only allowing for a single gap of up to 3 base pairs inserted (present in read but not in reference) or up to 11 base pairs deleted. Furthermore, we identify a gapped alignment if the above joining can be done with the fewest (up to a maximum of two for 2×25 and five for 2×50) number of mismatches and identify the location of it by where this joining occurs. The algorithm considers only unanchored tags in which only this gapped alignment can be found and condenses 2 or more non-redundant alignments allowing up to 4 candidates to be within 5 bases and unlimited candidates within 2 bases of another. It also requires that ¾ of the supporting reads have a consistent indel size and that, for cases with only two supporting non-redundant reads, the indels are on average greater than 9.1 bases from the end of the read.

Ambiguity of this location is common but is reduced by checking the color space compatibility of the sequences that the gap traverses.

We use fragment libraries to provide additional evidence to identify indel candidates. We use the same algorithm with the fragment data except that in this case the first and last 20 bases of the read are matched against the genome allowing for one mismatch and the window locations are 40 bases upstream to 80 bases downstream from the match location. Only fragment reads that do not align ungapped and align uniquely to the genome with a gap are considered for indel detection. Using the gapped alignment procedure described above with OEA pairs, we find gapped alignments by disallowing evidences that are within 9 bases for insertions and 12 bases for deletions of the ends of the reads and allowing a maximum of 3 mismatches in the read.

Medium insertions (up to 19 bp) and deletions (up to 500 bp)

We are able to take advantage of the longer reads in the 2×50 mate pairs to search for intra-read insertions up to 19 bp and deletions up to 500 bp. With these longer tag lengths, we disallow evidences that have gapped deletions within 19 bases of either end of the read and align both ends of the read until the maximum number of 5 mismatches occurs. For gapped insertions, we disallow 13 bases, but allow up to 5 mismatches for sizes 4 to 14, and 3 mismatches, for sizes 15 to 19. For medium insertions, we condense these results to call candidates in the same manner as small indels. For medium deletions, however, we do not require that the indel size is the same in ¾ of the supporting molecules.

Large structural variation inference. To search for larger insertions or deletions (inter-read indels) from 100 bp to 100 kb we evaluate the average pairing distance across the genome and look for pairs that significantly deviate from the expected insert size (Tuzun et al. 2005). We harness the power of high clone coverage to enable us to detect smaller indels (< 1 kb) with a high level of statistical significance which have been previously undetectable with mate-paired distances. A look-up table is created in which the amount that the clones must be deviated to achieve one standard deviation of significance is the standard error at each level of clone coverage. $SE = \frac{SD}{\sqrt{n}}$ in which SD is the standard deviation of all of the normal clones in the library and n is the number of clones. This produces an asymptotic curve in which the minimum size of detectable indels at a given level of significance drops rapidly as the clone coverage increases. We use the look-up table to determine the significance of the deviation in average insert size at each position in the genome. Regions of the genome that are significantly deviated are selected as candidate indels and hierarchical clustering is used to segregate the clones into groups in which the difference in the sizes of all clones in a group is less than the range of normal insert sizes of the given libraries. All clusters with less than two clones are removed and the candidates are assessed to determine if there is a homozygous or heterozygous population of deviated insert sizes. Any candidate indels with more than two populations with at least two clones in each are removed from consideration. All clones deviated by ≥ 100 kb are discarded.

Clones from various libraries with various insert sizes contribute to a single indel call by combining the probabilities associated with the clones from each library (Supplement Inter-read Insertions and Deletions). We limited the study of these structural variations to the longer tag mate-paired libraries. The clones used to detect the large indels are limited to those in which both tags place uniquely against the reference sequence allowing up to 5 mismatches per 50 bp tag and in which the number of mismatches in both tags sum to 5 or less.

Inversions. An inversion is defined by its two breakpoints. The number of mate pairs with one flipped end supporting the occurrence of the starting or ending inversion breakpoint is counted for each base pair. The genomic ranges corresponding to local peaks of these counts, if above a score threshold, are called as candidate breakpoint ranges. To define an inversion, its starting and ending breakpoints are paired up only if they are the reciprocal nearest neighbor of each other in the correct order. The score for such an inversion is the harmonic mean of its two breakpoints. Finally, each breakpoint range is scanned for coverage of normal mate pairs to identify a sub-range with the lowest normal mate-paired coverage as the most probable location of a breakpoint and to differentiate homozygous inversions from heterozygous ones.

Copy Number Variations. Copy Number Variations are analyzed using a hidden Markov model on variable-length windows. The mappability of the genome is calculated for a given tag length and number of mismatches by matching a uniformly fragmented genome to itself. Subsequently, the per-base count of sequence coverage is summed by

sliding windows (the size of which are varied to keep the sum of the corresponding mappable coverage constant). GC normalization is calculated using the empirical distribution of coverage as a function of GC content, normalizing observed coverage to adjust for GC content. A hidden Markov model is used for segmentation and a final set of filters is used to remove CNVs at segments with low mappability or small size.

SNP annotation analysis. Amino acid variants, found in the allele variant list in each OMIM entry, were transferred onto the reference genome using a battery of computational methods described elsewhere (Hendrix, MacBride, Salas and Reese, in preparation). In short, OMIM alleles were aligned to a translated genomic sequence, reverse-translated and transcribed, and each mutation was located as a single base change in the genomic sequence using Golden Path coordinates (Kent et al. 2002), creating a whole-genome map of OMIM in genomic coordinates. The results of this pipeline are 9,239 variants of amino acid and terminator mutations within OMIM that we can position uniquely and with confidence onto the genomic sequence.

PolyPhen is a tool to predict the possible impact of an amino acid substitution on the structure and function of a human protein, using sequence, phylogenetic and structural information characterizing the substitution (Ramensky et al. 2002). PolyPhen mapped dbSNP SNPs to the protein identifier from the SWALL database (SwissProt). We compared the non-synonymous SNPs we detected against the predicted impact of 76,434 non-synonymous SNPs in dbSNP build 126. We found 2,892 non-synonymous SNPs in this Yoruba sample that have a PolyPhen annotation.

Prediction and validation of gene fusions. The gene fusion prediction was performed as described in Bashir et al. utilizing mate-paired reads that match the human genome with zero mismatches and in only one location. The mate-paired data used in this analysis consists of all of the 2×25 data sets (see Table S1) as well as a small sample of 2×50 mate-paired data that is available at the NCBI Short Read Archive via the accession SRA000272 under the slide name Florence_20080201_1. The constraints imposed by multiple, spanning reads were used to reduce uncertainty in breakpoint location and the probability of any particular breakpoint was evaluated by additionally considering the insert size distribution of each mapped read. Six of the predicted breakpoints (four of which resulted from inversions, one from a deletion, and one from an undefined rearrangement) have a breakpoint probability equal to 1 and are supported by at least 3 pairs of reads. Five of these breakpoints (the four inversions and one deletion) were selected for further validation.

The regions predicted to contain the fusion genes were PCR-amplified from genomic DNA and then sequenced by conventional Sanger sequencing (Agencourt Bioscience) (Supplement Sanger Confirmation of Gene Fusions). Assembly of these sequences and alignment to the human genome (hg18) either identified the precise fusion point, or, in cases of high sequence homology between the fused genes, localized the breakpoint to within 100 bp. We Sanger sequenced both fused and reference alleles to validate possible heterozygous rearrangements. Four of the five predicted gene fusions have been confirmed by Sanger sequencing.

All analysis algorithms are freely available in open source format at http://solidsoftwaretools.com. The variants identified in this study are available at http://solidsoftwaretools.com/gf/project/yoruban. All data has been submitted to the Short Read Archive at NCBI and is available via the accession SRA000272.

ACKNOWLEDGEMENTS

We thank Jason Warner, Lynne Apone, Ali Aslam, Deyra Rodriguez and Chunlin Xiao for preliminary feasibility studies for the project, Ryan Koehler for early algorithmic development for CNV analysis, Brian Tuch for a ns-SNP annotation script, Charles Scafe for data mining and Aaron Kitzmiller for making the variant lists publicly available. We thank Eugene Spier and Michael Wenz for valuable advice and Roger Canales for assistance in bioinformatics tool release. This work was supported in part by NIH grants HG004120 to E.E.E., HG002993 to M.G.R. and NIH R01 HG004962-01 to V. Bafna and A. Bashir.

FIGURE LEGENDS

Figure 1. Cumulative plot of sequence and clone coverage from uniquely placed fragments and uniquely placed mate pairs. The sequence coverage is derived from the fragment, 2×25 mate-paired and 2×50 mate-paired libraries while the clone coverage is from only the mate-paired libraries (2×25 and 2×50).

Figure 2. Uniquely placed mate pairs provide a more comprehensive sampling of the human genome than the unique placement of each of the tags independently. The coverage is separated by mate-paired data treated as single tags before pairing (Mate Pairs, Unpaired) and mate-paired data treated as mate pairs (Mate Pairs, Paired).

Figure 3. Dependence of genotype calling on depth of sequence coverage. The NA18507 genotypes called by SOLiD at all HapMap loci are compared to the HapMap genotypes by SOLiD coverage per genome position (average 18× coverage). Coverage includes alleles representing the reference or a valid base change, i.e., alleles with single or invalid adjacent mismatches are not included. No prior information about SNP presence or SNP alleles was used in making SOLiD gentoype calls. The number of HapMap loci with a given level of SOLiD coverage ("Count") are shown and the percentage of these loci for which SOLiD gives the same genotype as HapMap for homozygotes and heterozygotes is represented by the colored lines (graphed using the left-hand y-axis and referred to as “% Concordance”) using two genotyping algorithms: Consensus Caller and diBayes. diBayes is more sensitive at heterozygous SNP detection and yields a lower FN-rate than Consensus Caller but we did not attempt to estimate the FP-rate of diBayes with validation data. SOLiD genotypes that differ from HapMap gentoypes are nearly always heterozygous undercalls (i.e., the position is called homozogyous for one of the two alleles) or called as N (insufficient evidence to make a confident genotype call).

Figure 4. Length distributions of small and medium insertions and deletions under sequencing reads with respective concordances. Deletions are detected up to 500 bp and insertions up to 20 bp. A high prevalence of small indels, even sized indels and Alu sized deletions (300-350 bp) are found in this genome. Larger indels (deletions 12 bp and higher and insertions 4 bp and higher) are called with more restrictive settings (see Methods) than smaller ones.

Figure 5. Length distributions of large insertions and deletions identified between mate-paired tags. There is an abundance of insertions and deletions in the size range of Alus as well as a spike in the number of deletions in the size range of LINEs (6000 bp).

Figure 6. The distribution of the 193 deletions identified in NA18507 with SOLiD by both the intra-read and inter-read approaches. The inset illustrates a 328 bp deletion detected using both the inter- and intra-read approaches. Four non-redundant molecules identify the deletion with the intra-read approach while 81 clones identify the deletion with the inter-read approach. This deletion has also been found in the Venter, Watson and YH genomes.

Figure 7: Copy Number Variations detected with SOLiD mate-paired reads in NA18507. (a) The size distribution of CNVs detected with SOLiD mate-paired reads. (b) Overlap of copy numbers computed from normalized SOLiD coverage and from Affymetrix array CGH (McCarroll et al). Colors indicate CNV calls from aCGH. On the top of the figure are the numbers of SOLiD CNV calls that overlap with aCGH data at each copy number.

Figure 8. Theoretical and actual detection of SNPs and indels at various levels of average sequence coverage. (a) The upper bound on the number of SNPs and intra-read indels that can be detected at various levels of coverage. This is calculated by assessing how much of the genome meets the coverage requirements for each type of variant, 2× coverage for homozygous SNPs, 4× coverage for heterozygous SNPs and 6× coverage without considering the 3 bp on each end of the reads for intra-read indels. For small indels, two split reads are required to make a call, but due to the more restrictive manner of these calls, only approximately 1 in 3 reads (as found in simulations) can be used for this. (b) The actual number of SNPs and intra-read indels detected at various levels of average sequence coverage. (c) The number of insertions and deletions >= 200 bp detected between mate-paired reads at various average levels of sequence coverage.

FIGURES

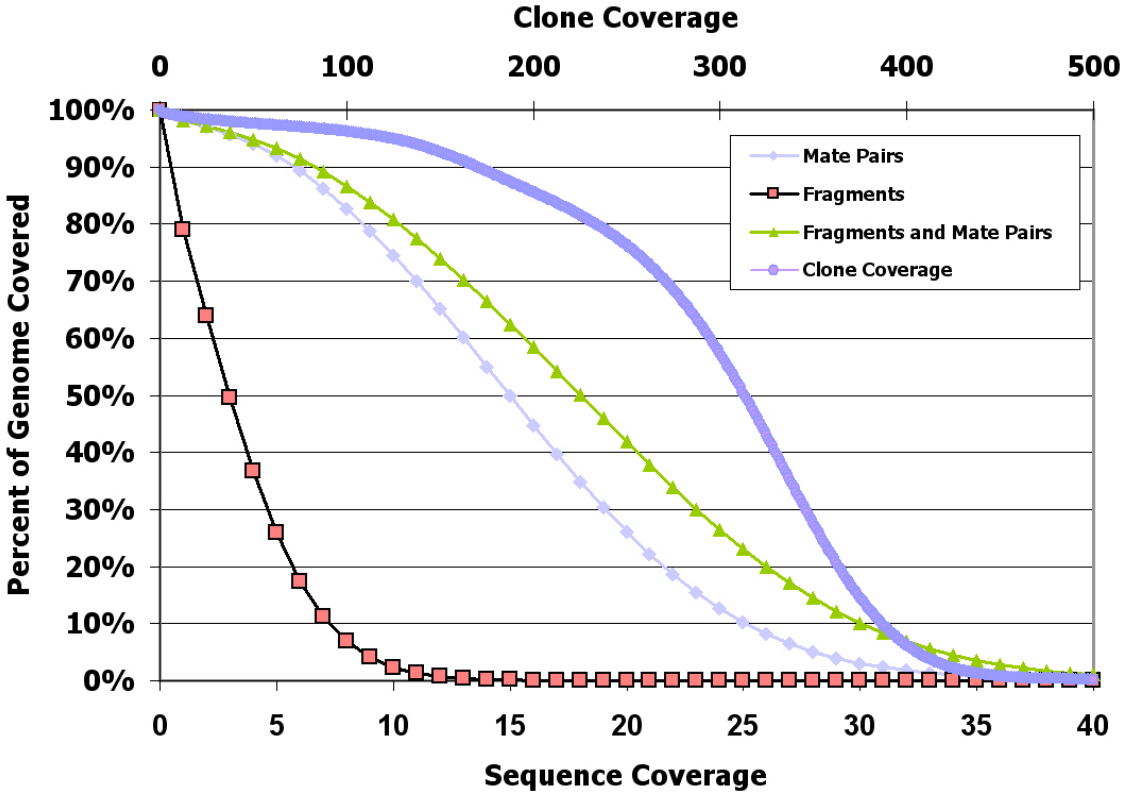


Figure 1

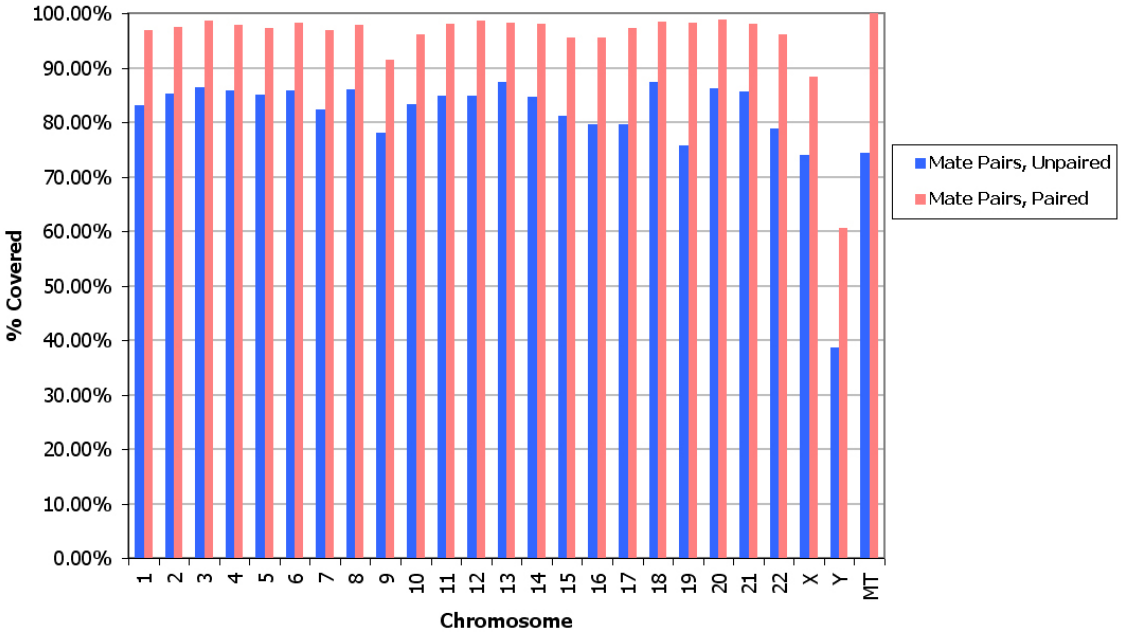


Figure 2

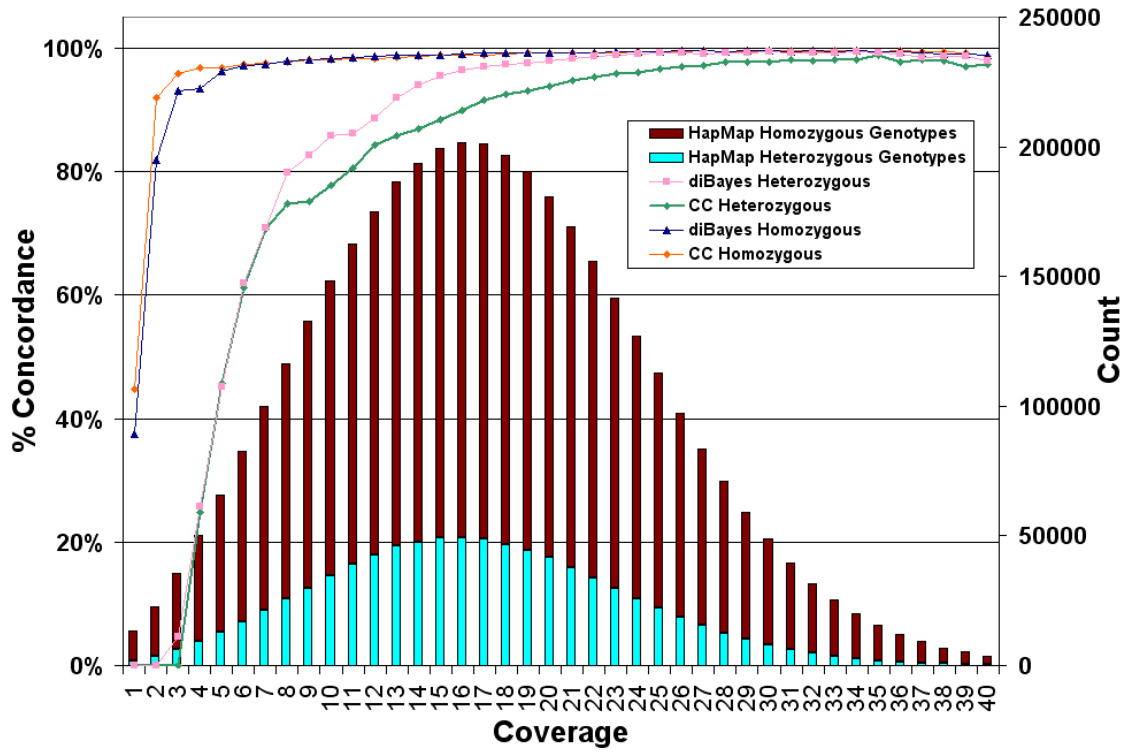


Figure 3

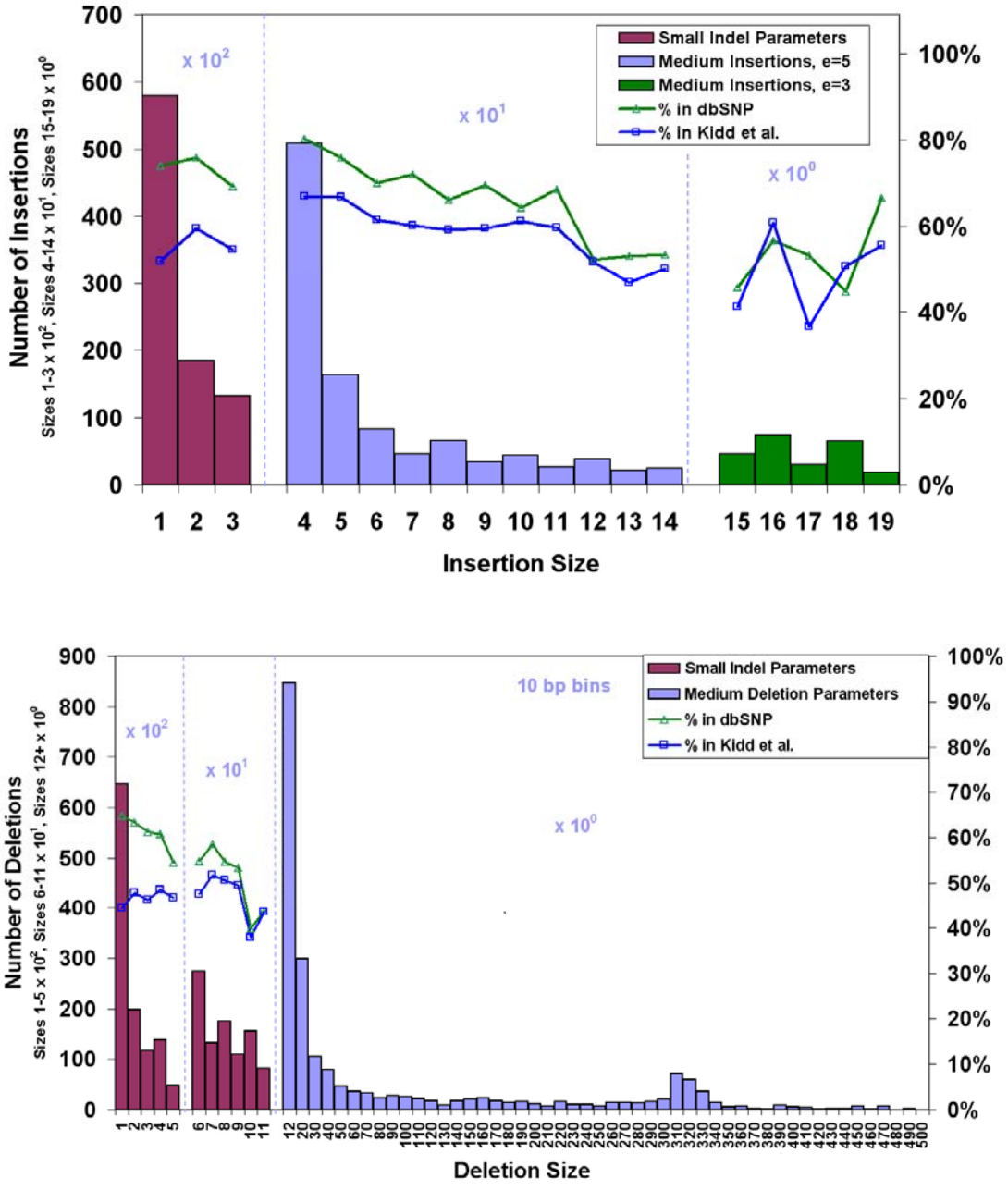


Figure 4

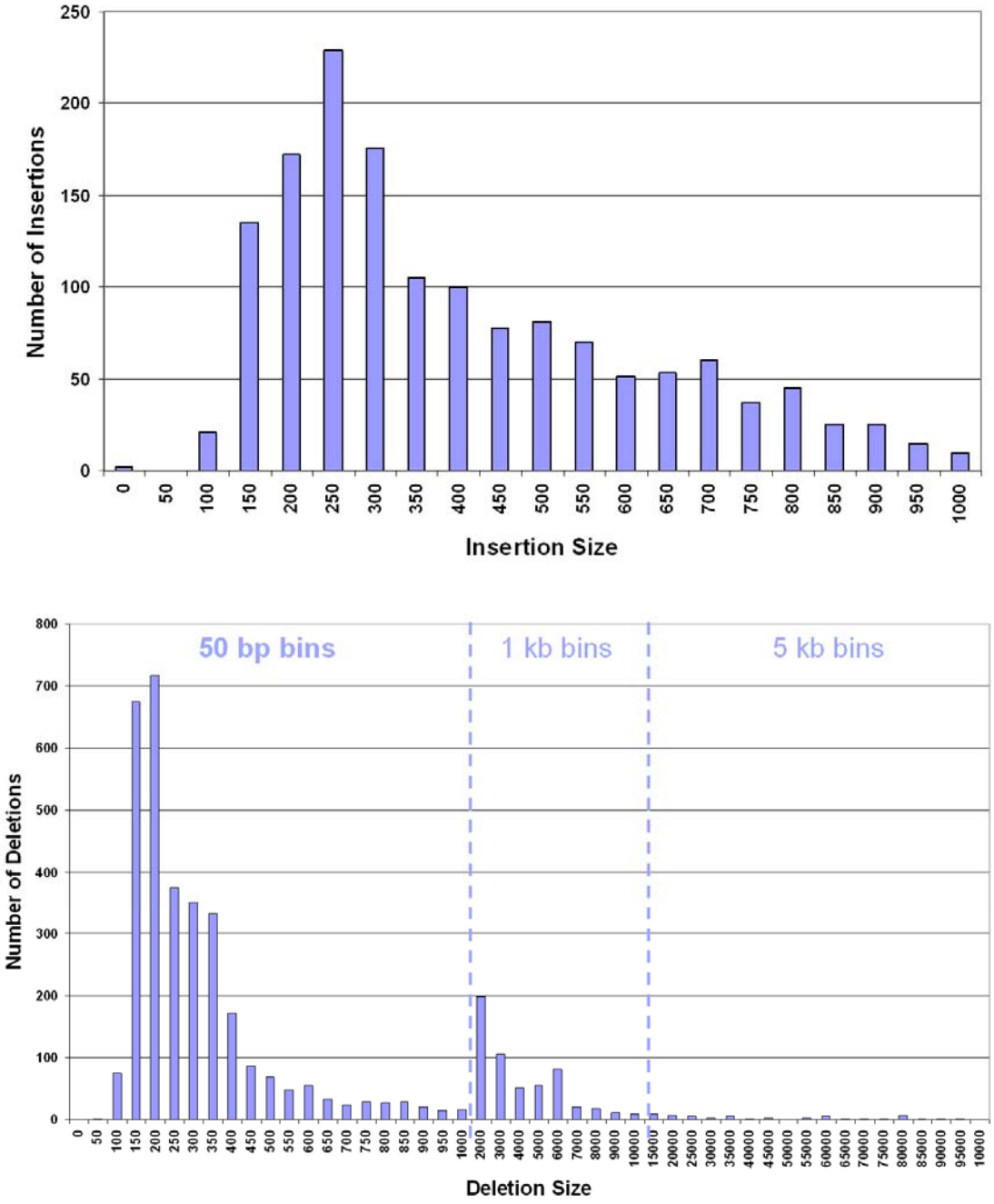


Figure 5

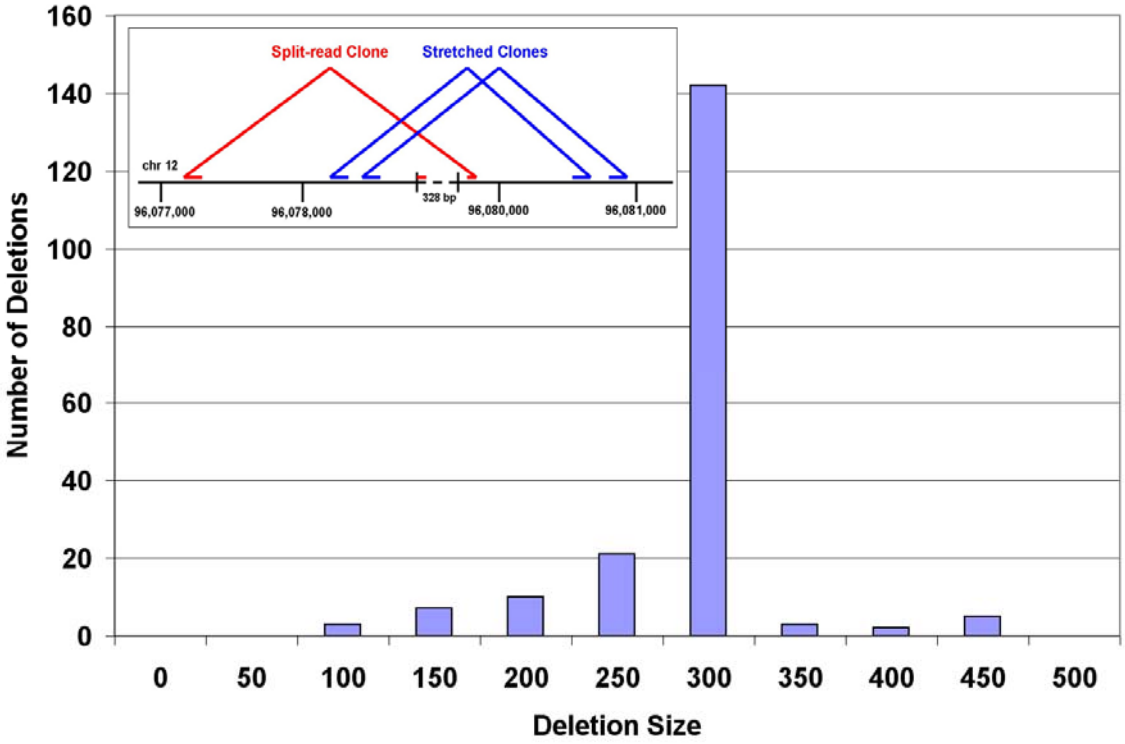
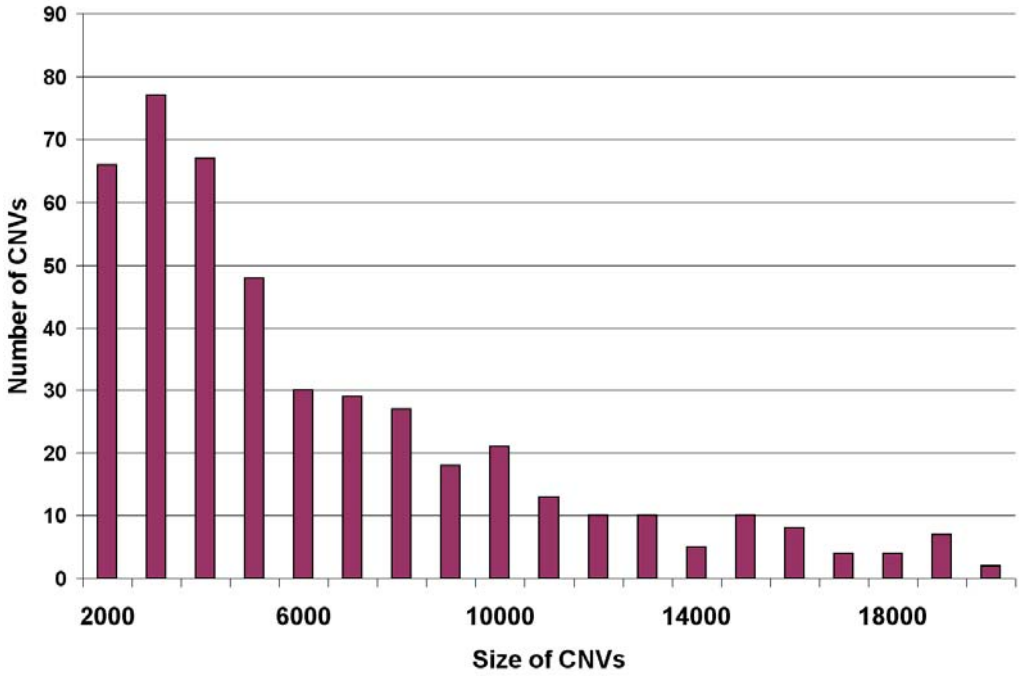
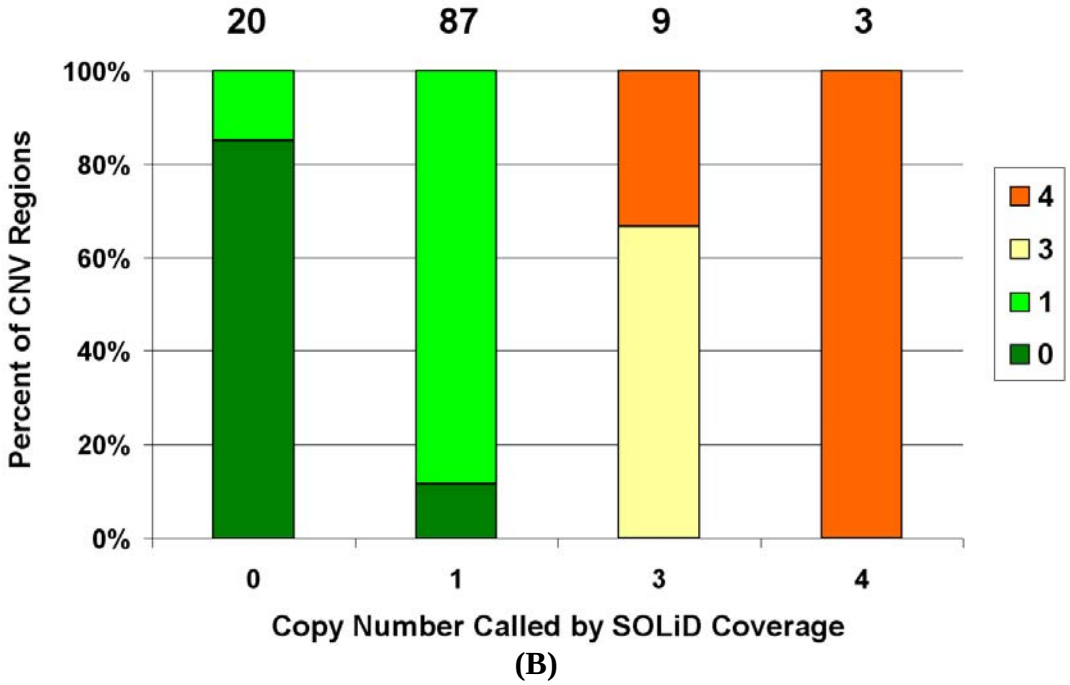


Figure 6

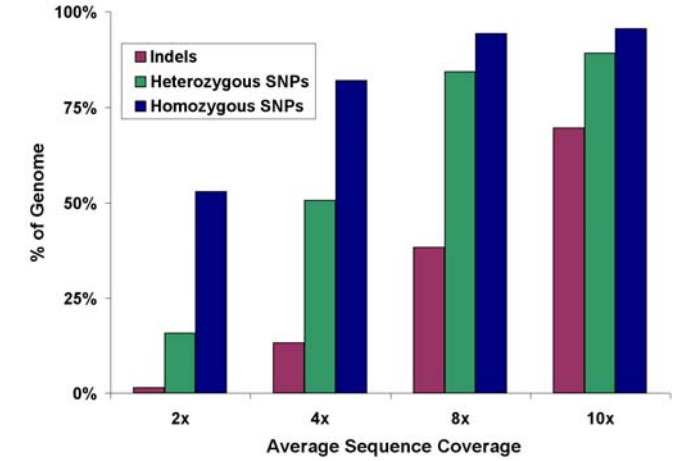


(A)

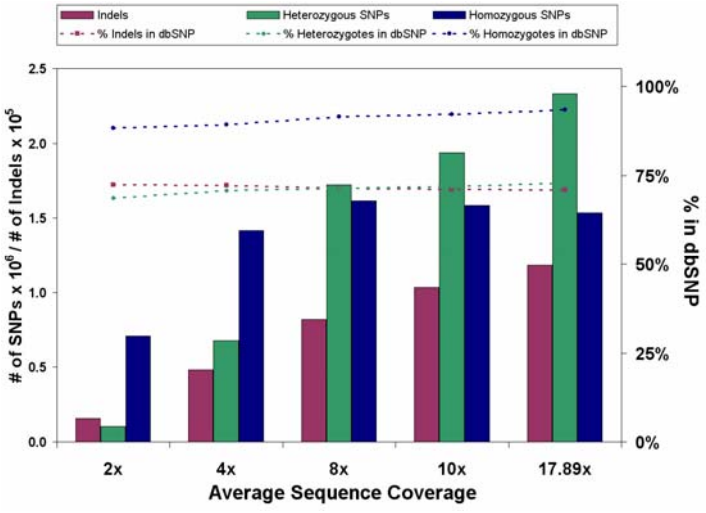


(B)

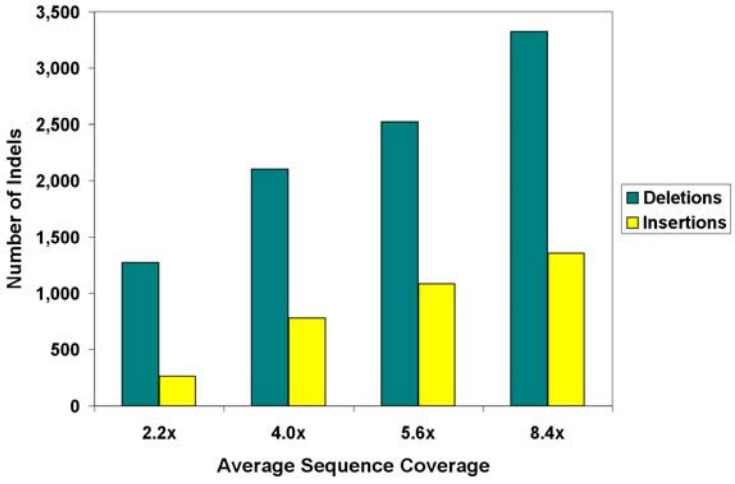
Figure 7



A



B



C

Figure 8

TABLES

Chromosome	Fusion	Rearrangement	Sanger Validation
2	*REG3G/REG3A*	Inversion	Yes
12	*CLEC1B/CLEC9A*	Inversion	No
16	*CTRB1/CTRB2*	Inversion	Yes
19	*EMR2/CD97*	Inversion	Yes
22	*APOBEC3A/APOBEC3B*	Deletion	Yes

Table 1. Predicted gene fusions created by structural variation events.

REFERENCES

2005. A haplotype map of the human genome. *Nature* **437**(7063): 1299-1320.

Bashir, A., Volik, S., Collins, C., Bafna, V., and Raphael, B.J. 2008. Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS Comput Biol* **4**(4): e1000051.

Bentley, D.R. Balasubramanian, S. Swerdlow, H.P. Smith, G.P. Milton, J. Brown, C.G. Hall, K.P. Evers, D.J. Barnes, C.L. Bignell, H.R. et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**(7218): 53-59.

Birney, E. Stamatoyannopoulos, J.A. Dutta, A. Guigo, R. Gingeras, T.R. Margulies, E.H. Weng, Z. Snyder, M. Dermitzakis, E.T. Thurman, R.E. et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**(7146): 799-816.

Braslavsky, I., Hebert, B., Kartalov, E., and Quake, S.R. 2003. Sequence information can be obtained from single DNA molecules. *Proc Natl Acad Sci U S A* **100**(7): 3960-3964.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M. et al. 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* **18**(6): 630-634.

Campbell, P.J., Stephens, P.J., Pleasance, E.D., O'Meara, S., Li, H., Santarius, T., Stebbings, L.A., Leroy, C., Edkins, S., Hardy, C. et al. 2008. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* **40**(6): 722-729.

Chiu, P.L., Ng, B.H., Chang, G.W., Gordon, S., and Lin, H.H. 2008. Putative alternative trans-splicing of leukocyte adhesion-GPCR pre-mRNAs generates functional chimeric receptors. *FEBS Lett* **582**(5): 792-798.

Cloonan, N., Forrest, A.R., Kolle, G., Gardiner, B.B., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., Bethel, G. et al. 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* **5**(7): 613-619.

Dressman, D., Yan, H., Traverso, G., Kinzler, K.W., and Vogelstein, B. 2003. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci U S A* **100**(15): 8817-8822.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B. et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* **323**(5910): 133-138.

Frazer, K.A. Ballinger, D.G. Cox, D.R. Hinds, D.A. Stuve, L.L. Gibbs, R.A. Belmont, J.W. Boudreau, A. Hardenbol, P. Leal, S.M. et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**(7164): 851-861.

Gibbs, R.A. Rogers, J. Katze, M.G. Bumgarner, R. Weinstock, G.M. Mardis, E.R. Remington, K.A. Strausberg, R.L. Venter, J.C. Wilson, R.K. et al. 2007.

Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**(5822): 222-234.

Gresham, D. and Kruglyak, L. 2008. Rise of the machines. *PLoS Genet* **4**(8): e1000134.

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research* **33**(Database issue): D514-517.

Hillier, L.W., Marth, G.T., Quinlan, A.R., Dooling, D., Fewell, G., Barnett, D., Fox, P., Glasscock, J.I., Hickenbotham, M., Huang, W. et al. 2008. Whole-genome sequencing and variant discovery in C. elegans. *Nat Methods* **5**(2): 183-188.

Jimenez-Sanchez, G., Childs, B., and Valle, D. 2001. Human disease genes. *Nature* **409**(6822): 853-855.

Ju, J., Kim, D.H., Bi, L., Meng, Q., Bai, X., Li, Z., Li, X., Marma, M.S., Shi, S., Wu, J. et al. 2006. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc Natl Acad Sci U S A* **103**(52): 19635-19640.

Kaiser, J. 2008. DNA sequencing. A plan to capture human diversity in 1000 genomes. *Science* **319**(5862): 395.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome research* **12**(6): 996--1006.

Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F. et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**(7191): 56-64.

Kidd, J.M., Newman, T.L., Tuzun, E., Kaul, R., and Eichler, E.E. 2007. Population stratification of a common APOBEC gene deletion polymorphism. *PLoS Genet* **3**(4): e63.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L. et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**(5849): 420-426.

Lander, E.S. Linton, L.M. Birren, B. Nusbaum, C. Zody, M.C. Baldwin, J. Devon, K. Dewar, K. Doyle, M. FitzHugh, W. et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.

Lander, E.S. and Waterman, M.S. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**(3): 231-239.

Lee, S., Cheran, E., and Brudno, M. 2008. A robust framework for detecting structural variations in a genome. *Bioinformatics* **24**(13): i59-67.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G. et al. 2007. The diploid genome sequence of an individual human. *PLoS Biol* **5**(10): e254.

Ley, T.J., Mardis, E.R., Ding, L., Fulton, B., McLellan, M.D., Chen, K., Dooling, D., Dunford-Shore, B.H., McGrath, S., Hickenbotham, M. et al. 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**(7218): 66-72.

Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L. et al. 2008.

Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**(5866): 1100-1104.

Lohmueller, K.E., Indap, A.R., Schmidt, S., Boyko, A.R., Hernandez, R.D., Hubisz, M.J., Sninsky, J.J., White, T.J., Sunyaev, S.R., Nielsen, R. et al. 2008. Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**(7181): 994-997.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z. et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**(7057): 376-380.

Marth, G.T., Korf, I., Yandell, M.D., Yeh, R.T., Gu, Z., Zakeri, H., Stitziel, N.O., Hillier, L., Kwok, P.Y., and Gish, W.R. 1999. A general approach to single-nucleotide polymorphism discovery. *Nat Genet* **23**(4): 452-456.

McCarroll, S.A. and Altshuler, D.M. 2007. Copy-number variation and association studies of human disease. *Nat Genet* **39**(7 Suppl): S37-42.

McCarroll, S.A., Huett, A., Kuballa, P., Chilewski, S.D., Landry, A., Goyette, P., Zody, M.C., Hall, J.L., Brant, S.R., Cho, J.H. et al. 2008. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat Genet* **40**(9): 1107-1112.

McKusick, V.A. 2007. Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins
University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda,
MD).

Mills, R.E., Luttig, C.T., Larkins, C.E., Beauchamp, A., Tsui, C., Pittard, W.S., and Devine, S.E. 2006. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res* **16**(9): 1182-1190.

Morton, N.E., Crow, J.F., and Muller, H.J. 1956. An Estimate of the Mutational Damage in Man from Data on Consanguineous Marriages. *Proc Natl Acad Sci U S A* **42**(11): 855-863.

Quinlan, A.R. and Marth, G.T. 2007. Primer-site SNPs mask mutations. *Nat Methods* **4**(3): 192.

Ramensky, V., Bork, P., and Sunyaev, S. 2002. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* **30**(17): 3894-3900.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W. et al. 2006. Global variation in copy number in the human genome. *Nature* **444**(7118): 444-454.

Ronaghi, M., Nygren, M., Lundeberg, J., and Nyren, P. 1999. Analyses of secondary structures in DNA by pyrosequencing. *Anal Biochem* **267**(1): 65-71.

Shendure, J. and Ji, H. 2008. Next-generation DNA sequencing. *Nat Biotechnol* **26**(10): 1135-1145.

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D., and Church, G.M. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**(5741): 1728-1732.

Smith, D., Malek, J., and Mckernan, K. 2006. METHODS FOR PRODUCING A PAIRED TAG FROM A NUCLEIC ACID SEQUENCE AND METHODS OF USE THEREOF. EP Patent 1,682,680.

Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S., Abeysinghe, S., Krawczak, M., and Cooper, D.N. 2003. Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* **21**(6): 577-581.

Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. 2003. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* **13**(9): 2129-2141.

Tobler, A.R., Short, S., Andersen, M.R., Paner, T.M., Briggs, J.C., Lambert, S.M., Wu, P.P., Wang, Y., Spoonde, A.Y., Koehler, R.T. et al. 2005. The SNPlex genotyping system: a flexible and scalable platform for SNP genotyping. *J Biomol Tech* **16**(4): 398-406.

Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D. et al. 2005. Fine-scale structural variation of the human genome. *Nat Genet* **37**(7): 727-732.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J.A., Costa, G., McKernan, K. et al. 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res* **18**(7): 1051-1063.

Venter, J.C. Adams, M.D. Myers, E.W. Li, P.W. Mural, R.J. Sutton, G.G. Smith, H.O. Yandell, M. Evans, C.A. Holt, R.A. et al. 2001. The sequence of the human genome. *Science* **291**(5507): 1304-1351.

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Guo, Y. et al. 2008. The diploid genome sequence of an Asian individual. *Nature* **456**(7218): 60-65.

Welch, R.A., Lazaruk, K., Haque, K.A., Hyland, F., Xiao, N., Wronka, L., Burdett, L., Chanock, S.J., Ingber, D., De La Vega, F.M. et al. 2008. Validation of the performance of a comprehensive genotyping assay panel of single nucleotide polymorphisms in drug metabolism enzyme genes. *Hum Mutat* **29**(5): 750-756.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T. et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**(7189): 872-876.