nature biotechnology

Performance comparison of exome DNA sequencing technologies

Michael J Clark[1,4], Rui Chen[1,4], Hugo Y K Lam[1], Konrad J Karczewski[1], Rong Chen[2], Ghia Euskirchen[1,3], Atul J Butte[2] & Michael Snyder[1,3]

Whole exome sequencing by high-throughput sequencing of target-enriched genomic DNA (exome-seq) has become common in basic and translational research as a means of interrogating the interpretable part of the human genome at relatively low cost. We present a comparison of three major commercial exome sequencing platforms from Agilent, Illumina and Nimblegen applied to the same human blood sample. Our results suggest that the Nimblegen platform, which is the only one to use high-density overlapping baits, covers fewer genomic regions than the other platforms but requires the least amount of sequencing to sensitively detect small variants. Agilent and Illumina are able to detect a greater total number of variants with additional sequencing. Illumina captures untranslated regions, which are not targeted by the Nimblegen and Agilent platforms. We also compare exome sequencing and whole genome sequencing (WGS) of the same sample, demonstrating that exome sequencing can detect additional small variants missed by WGS.

It is now possible to analyze the genomic DNA of individuals using whole genome sequencing and exome sequencing[1–3], and these strategies have become popular for basic[4,5] and translational[6–11] research. Exome sequencing involves the capture of RNA coding regions by hybridizing genomic DNA to oligonucleotide probes (baits) that collectively cover the human exome regions. These enriched genomic regions are then sequenced using high-throughput DNA sequencing technology[12]. Although WGS is more comprehensive, exome sequencing has become more common because it captures the highly interpretable coding region of the genome and is more affordable, thereby allowing large numbers of samples to be analyzed. Exome sequencing has been used for analyses of disease loci that segregate in families[13,14], large disease cohorts (National Heart, Lung, and Blood Institute) and validation in WGS studies (such as The 1000 Genomes Project[15]).

There are currently three major exome enrichment platforms: Agilent's SureSelect Human All Exon 50Mb, Roche/Nimblegen's SeqCap EZ Exome Library v2.0 and Illumina's TruSeq Exome Enrichment. Each platform uses biotinylated oligonucleotide baits complementary to the exome targets to hybridize sequencing libraries prepared from fragmented genomic DNA. These bound libraries are enriched for targeted regions by pull-down with magnetic streptavidin beads and then sequenced. The technologies differ in their target choice, bait lengths, bait density and molecule used for capture (DNA for Nimblegen and Illumina, and RNA for Agilent). The performance of each technology was systematically analyzed and compared, thereby revealing how design differences and experimental parameters (e.g., sequencing depth) affect variant discovery.

RESULTS

Platform design differences

There are substantial differences in the density of oligonucleotide baits between the three platforms (**Fig. 1a**). Nimblegen contains overlapping baits that cover the bases it targets multiple times, making it the highest density platform of the three. Agilent baits reside immediately adjacent to one another across the target exon intervals. Illumina relies on paired-end reads to extend outside the bait sequences and fill in the gaps.

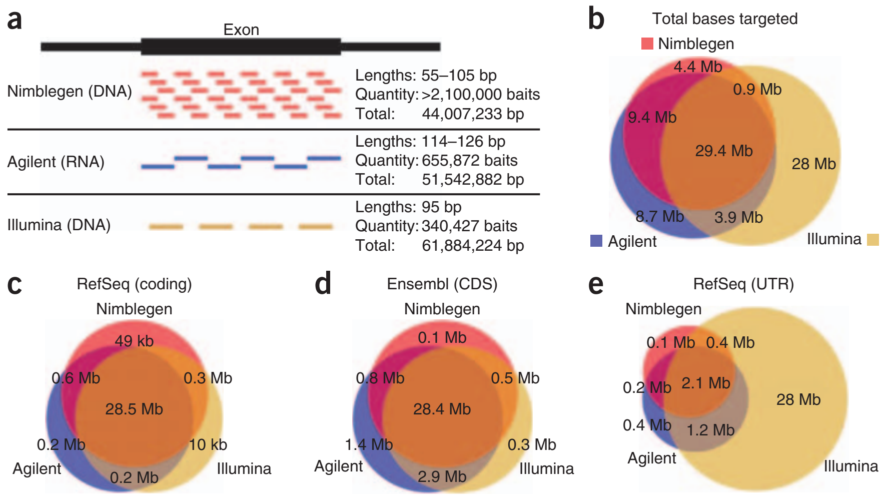
The exome enrichment platforms also have different target regions. The exome consists of all the exons of a genome that are transcribed into mature RNA. Numerous databases of mRNA coding sequences exist (including RefSeq[16], UCSC KnownGenes[17] and Ensembl[18]). They contain different numbers of noncoding RNA genes, and the start and end positions of some transcripts differ between them. Each commercial platform targets particular exomic segments based on combinations of the available databases. We compared the exact regions of the genome covered by each platform (based on individual design documents obtained from the company websites or through correspondence) (**Fig. 1b**). A large number of bases (29.45 Mb) are targeted by all three platforms. The Nimblegen and Agilent platforms share more with each other (38,830,789 bp) than either does with the Illumina platform (30,304,987 bp and 33,299,208 bp, respectively) and each platform possesses 4.4–28 Mb of unique target regions.

We first examined coverage of major RNA databases—RefSeq (coding and untranslated region (UTR)), Ensembl (total and coding sequence (CDS)) and the microRNA (miRNA) database miRBase[19] (**Supplementary Table 1**). Coverage of mRNA coding exons in both RefSeq (**Fig. 1c**) and Ensembl (**Fig. 1d**) was strikingly similar between all platforms. The shared bases in mRNA coding exons account for nearly all of the 29.45 Mb common to the three platforms. Nonetheless, each platform does target specific regions. The majority of the Illumina-specific 27.73 Mb targets UTR regions (**Fig. 1e**). Nimblegen covers a greater portion of miRNAs, and Agilent better covers Ensembl genes.

[1]Department of Genetics, Stanford University School of Medicine, Stanford, California, USA. [2]Division of Systems Medicine, Department of Pediatrics, Stanford University School of Medicine, Stanford, California, USA. [3]Center for Genomics and Personalized Medicine, Stanford University, Stanford, California, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to M.S. (mpsnyder@stanford.edu).

Received 16 May; accepted 18 August; published online 25 September 2011; doi:10.1038/nbt.1975

Figure 1 Exome enrichment designs include different biochemical methods, bait lengths, quantity and overlap of baits and number of bases targeted. (**a**) Bait design details for each commercial platform. (**b**) Venn diagram showing the overlap of targeted genome regions for all three platforms. (**c**) Venn diagram showing coverage of RefSeq coding exons and overlap between platforms. (**d**,**e**) Same as **c**, but for Ensembl CDS exons and RefSeq UTR exons respectively.



Various metrics from the physical protocols for each platform were compared (**Supplementary Table 2**). Input genomic DNA ranged from 1 μg (Illumina) to 3 μg (Nimblegen and Agilent). The total procedure time before sequencing ranged from 3.5 d (Agilent, Illumina) to 7 d (Nimblegen). Pre- and post-hybridization PCR cycles varied across platforms. Agilent uses RNA in its selection rather than DNA. All three platforms can be automated. Although the list price for each platform varies, the per-reaction prices are highly negotiable with the vendors and currently range from <$400 to >$1,000.

Target enrichment efficiency

To assess the enrichment efficiency of each platform, we sequenced libraries generated from genomic DNA derived from peripheral blood mononuclear cells (PBMCs) of a healthy volunteer of European descent. Exome DNA was enriched with each platform according to the manufacturers' recommendation. For each exome library, 112–184 million (M) 101-bp paired-end reads were generated using one lane of an Illumina HiSeq 2000 and mapped using the Burrows-Wheeler Alignment tool (BWA). BWA mapped 99% of reads to human DNA (with 88–95% to unique regions of the genome), and 10–15% of those reads were duplicates (PCR artifacts) that were removed during post-processing (**Supplementary Table 3**). For comparison at constant read depth, 80M mapped reads were randomly drawn from each data set.

Overall targeting efficiency was assessed by measuring base coverage over all targeted bases for each platform at 80M reads. With Nimblegen enrichment, 98.6% of the targeted bases were covered at least once, and 96.8% at ≥10×; with Illumina, 97.1% of bases were covered at least once, and 90.0% at ≥10×; with Agilent, 96.6% of bases were covered at least once, and 89.6% at ≥10× (**Fig. 2a**).

To assess targeting and enrichment efficiency as a function of sequencing depth, we randomly chose aligned reads from the 80M read pool in 10M read increments from 20–80M reads. The percent of targeted bases covered at depths of at least 10×, 20× and 30× were assessed (**Fig. 2b–d**). At all read counts and depth cut-offs, the Nimblegen platform enriched a higher percentage of its targeted bases than the other two platforms. Illumina and Agilent enriched a higher total number of bases at higher read counts (**Fig. 2e–g**). The efficient baits became saturated by 40M (Nimblegen), 50M (Agilent) and 60M (Illumina) reads, with <2% increase in bases covered at ≥10×. These findings indicate that design differences dramatically affect the balance between targeting efficiency and total number of bases targeted. A higher density design, targeting a smaller genomic interval, results in higher efficiency. Lower density designs can capture a greater total number of bases but require substantially larger amounts of sequencing.

Off-target enrichment

Off-target regions may be enriched if there is high sequence similarity between those regions and bait regions. The number of reads that unambiguously mapped to regions outside the targeted bait intervals for each platform (±500 bp) was quantified in the normalized 80M read data sets to assess off-target enrichment; 9.3% of Nimblegen, 12.8% of Agilent and 35.6% of Illumina reads uniquely mapped to off-target regions (**Fig. 3a**). The percent of off-target enrichment correlated strongly with the enrichment trends mentioned previously, suggesting that off-target enrichments have a dramatic effect on targeting efficiency. Off-target reads were cross-referenced with RepeatMasker and segmental duplications, genomic structures known to confound targeted assays. For all three platforms, a higher fraction of off-target enrichments mapped to repeat elements (**Fig. 3b**) and segmental duplications (**Fig. 3c**) than did on-target regions.

Enrichment bias owing to GC content

Another source of potential inefficiency may come from targeting regions with high or low GC content. Lower coverage in sequencing regions with high GC or high AT content has long been observed[20]. GC bias in sequencing studies is in large part due to early PCR steps during library generation[21] where high and low GC content cause reduced amplification and therefore lower sequencing coverage. GC content has also been shown to affect the efficiency of hybridization to oligonucleotides[22,23], and therefore may also influence target enrichment by oligonucleotide baits. To investigate the effect of GC content on efficiency, we plotted GC content against mean read depth across target regions using the normalized 80M–read data sets (**Fig. 3d–f**). The density plots show that each platform demonstrates a marked reduction in read depth over high and low GC targets. All three platforms showed a sharp drop in read depth as GC content increased from 60% to 80%. As the GC content dropped from 40% to 20%, the performance of both Nimblegen and Illumina diminished with lower read depth over those targets (**Fig. 3e**,**f**). The Agilent platform displayed only a slight reduction in read depth across low GC targets (**Fig. 3d**), possibly because of its lower number of PCR cycles, longer baits and/or the use of RNA probes.

Detection of single-nucleotide variations

Although enrichment efficiency is a function of read depth, it does not necessarily correlate with ability to identify variants. Single-nucleotide variations (SNVs) represent the most numerous sequence variations in the human exome[15] and their accurate and comprehensive identification is a major goal of exome sequencing. To evaluate

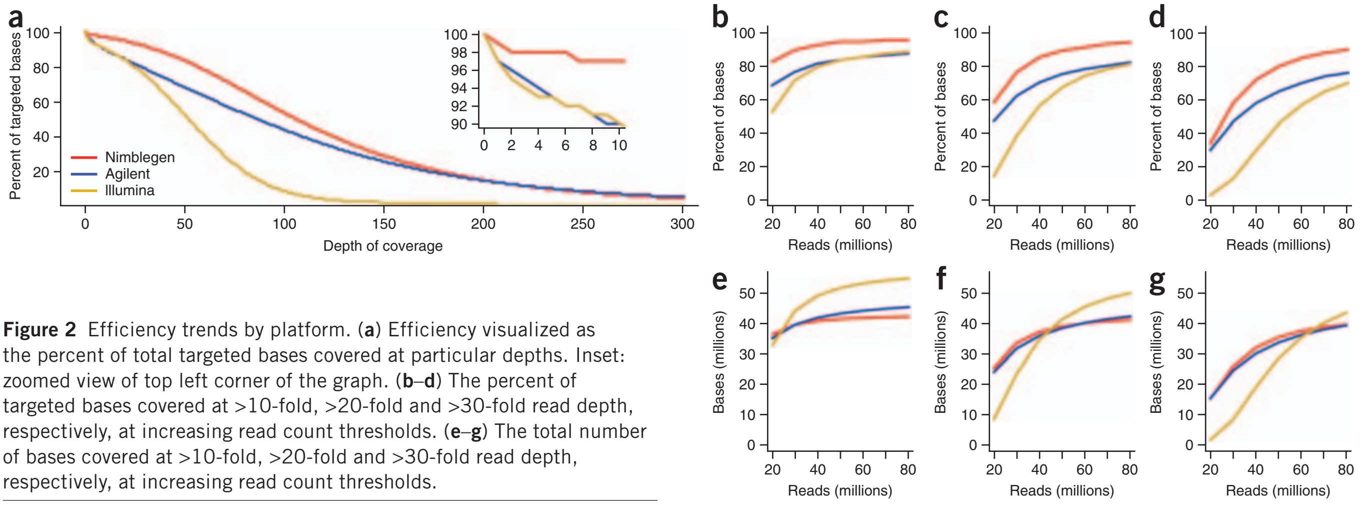


Figure 2 Efficiency trends by platform. (**a**) Efficiency visualized as the percent of total targeted bases covered at particular depths. Inset: zoomed view of top left corner of the graph. (**b–d**) The percent of targeted bases covered at >10-fold, >20-fold and >30-fold read depth, respectively, at increasing read count thresholds. (**e–g**) The total number of bases covered at >10-fold, >20-fold and >30-fold read depth, respectively, at increasing read count thresholds.

SNV detection performance, we called variants in each normalized data set using the Genome Analysis Toolkit (GATK)[24]. From the normalized 80M read data sets, a total of 46,960 (Nimblegen), 50,634 (Agilent) and 52,859 (Illumina) SNVs were detected (**Supplementary Data 1** and **Supplementary Table 4**).

Single-nucleotide polymorphism (SNP) calls were validated by analyzing the same sample with Illumina Human 1M-Duo SNP Chip. Data were restricted to bases within the targeted regions that received a Phred-based quality score ≥30 by GATK. Heterozygous positions in the SNP Chip were compared to the genotype calls in the normalized 80M read exome sequencing data. Concordance rates were 99.3% for Agilent, 99.5% for Nimblegen and 99.2% for Illumina. For each platform, all nonconcordant genotype calls were calls of homozygous reference. Reference bias is a phenomenon often observed in sequencing studies[25]. Allelic balance (AB) was calculated by determining the ratio of reference base calls over the total number of calls at every SNV with a quality score ≥30. For Agilent, AB = 0.55, for Nimblegen and Illumina, AB = 0.53. These biases were not strong, but explain a fraction of the discordance with the SNP Chip data. SNP Chips also have their own error rates that may account for some of the discordance.

Trends in SNV calls

In general, although the oligonucleotides, bait length and type (DNA and RNA) differ, no biases toward or against specific nucleotide substitutions were observed among the three platforms. There was a slight increase in G→A/C→T transitions and slight decrease in non-G→C/C→G transversions in the Nimblegen data because a larger percent of its target bases are in coding regions, which have a higher GC content and therefore different nucleotide substitution rates from the rest of the genome[26]. The transition/transversion (ts/tv) ratio of total variants ranged from 2.53 to 2.67 and was slightly lower than estimates of ~2.8 from the exome based on 1000 Genomes data[15]. As expected, the platform with the most target sequence outside coding exons (Illumina) had the lowest ts/tv, whereas the platform with the least (Nimblegen) had the highest. No significant difference in the ratio of heterozygous to homozygous variants between platforms was observed.

We next examined whether the differences in efficiency at lower read counts affected overall SNV detection. For all platforms, the total number of SNVs detected increased as read count increased (**Fig. 4a**). There was a correlation between the total bases targeted by the platform and the number of SNVs detected, particularly at higher read counts. This increase was not linear, and for more than 30M reads, fewer than 5% additional SNVs were found by increasing read depth. SNV detection across targeted bases shared by all three platforms was assessed (**Fig. 4b**). In shared regions, Nimblegen consistently captured the most SNVs and became saturated with the lowest number of reads, followed by Agilent and then Illumina, indicating a correlation between bait density and sensitivity to SNV detection.

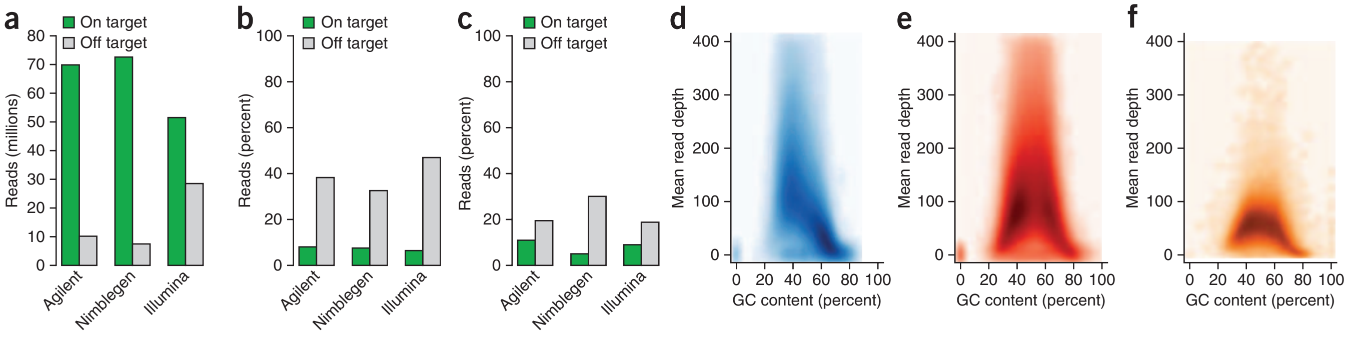


Figure 3 Off-target enrichment and GC bias. (**a**) Off-target enrichment by platform is represented by total number of on-target (green) and off-target (gray) post-alignment reads from data sets normalized to 80M reads total. (**b**,**c**) The percent of on-target and off-target reads that overlap RepeatMasker entries (**b**) and known segmental duplications (**c**). (**d–f**) Density plot shows the correlation between mean read depth across targeted regions and GC content in the Agilent exome sequencing data (**d**), Nimblegen (**e**) and Illumina (**f**). GC content across every target region was determined by dividing the number of G and C bases by the total number of bases in the target region. Mean read depth was determined across each target region independently. These plots were generated with smoothScatter from the Bioconductor package "geneplotter" (http://www.bioconductor.org/).

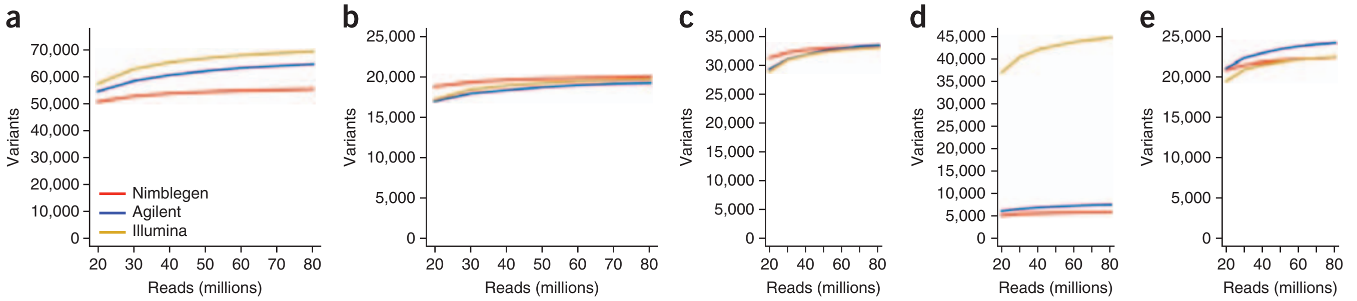


Figure 4 SNV trends by platform. Sensitivity toward SNVs is compared between each platform at increasing read counts. (**a**) Total number of SNVs detected at increasing read count thresholds. Sensitivity increases at higher read counts, particularly for the lower efficiency platforms. (**b**) SNVs detected in bases targeted by all three platforms. Nimblegen detects the most SNVs at all read counts because it is the most efficient. There is <2% increase in total variants detected for all platforms past 50M reads. (**c**) SNVs detected in RefSeq coding exons. These curves match the shared interval curves very closely because the genomic region shared by all three platforms is made up almost entirely by the RefSeq coding exons. (**d**) SNVs detected in RefSeq UTRs. UTRs are generally only targeted by the Illumina platform, so it detects far more in the UTRs at all read counts. (**e**) SNVs detected in Ensembl CDS. The Nimblegen and Illumina curves are very similar to their RefSeq coding curves in **c**. The Agilent curve is shifted upwards compared to its RefSeq coding curve because Agilent targets a large segment (1.4 Mb) of Ensembl CDS missed by the other two platforms.

SNV detection in mRNA coding regions

SNV detection in regions covered by particular exome databases was examined. Nimblegen-enriched libraries consistently enabled detection of the greatest number of RefSeq coding region variants at every read count (**Fig. 4c**). Illumina enrichment detected many more mutations in the UTR than either Agilent or Nimblegen (**Fig. 4d**). By 30M reads, Agilent's additional coverage allowed it to identify the most Ensembl CDS SNVs (**Fig. 4e**).

Variants specifically detected by one platform were examined in more detail. Platform-specific SNVs were typically discovered because of higher coverage in their targeted regions. The higher efficiency of Nimblegen's dense baits led to higher relative coverage of low complexity, hard-to-target regions and therefore detection of more SNVs in these regions (**Supplementary Fig. 1a**). Agilent detected unique SNVs most often in introns because Agilent baits sometimes extend farther outside the exon targets than the baits of other platforms (**Supplementary Fig. 1b**). Most of the SNVs detected uniquely by Illumina lie in UTRs (**Supplementary Fig. 1c**).

Small insertion and deletion detection

Small insertions and deletions (indels), ranging in size from −84 to +18 bases, were detected at a frequency of 12.5–14.5% that of SNVs (**Supplementary Data 2** and **Supplementary Table 4**), similar to the percentage reported by others[15,26]. As with SNVs, the total number of indels detected correlated with read count (**Fig. 5a**). Notably, at low read counts, more indels were detected after Agilent enrichment than after Illumina enrichment. At 50M reads, Illumina surpassed Agilent.

Coverage of regions containing indels largely matched coverage over the targeted regions. In shared and RefSeq regions, Nimblegen had the highest sensitivity for detecting indels because of higher average read depth. Agilent surpassed Illumina in indel detection at low read counts (**Fig. 5b**,**c**). Many more indels in UTRs were detected after Illumina enrichment (**Fig. 5d**). Agilent enrichment led to the largest number of detected indels at every read count in Ensembl CDS exons (**Fig. 5e**).

Most indels were 1 base in size (**Supplementary Fig. 2a**). Notably, there were slight enrichments at indel sizes of 4 and 8 bases in the total captured DNA data, consistent with findings in comparisons between human and primate genomes[27]. As expected[28], the frequency of indels present in the protein coding segments was much lower than in the total covered regions, which contain introns, UTRs and intergenic sequences (**Supplementary Fig. 2b**). There was a strong bias toward indels of a size equal to multiples of three bases in coding regions. This pattern was presumably due to selective pressure against deleterious frameshift mutations in the coding regions.

Comparison with WGS

WGS requires a much greater amount of sequencing to achieve equivalent coverage as exome sequencing, but its performance relative to exome sequencing has not been well described. To do such a comparison, we performed WGS to high read depth on an Illumina HiSeq 2000 on a blood sample from the same individual analyzed for the exome sequencing comparison. A subset of those reads (seven lanes) was extracted, mapped and had duplicates removed. This yielded 1,194,622,756 unambiguously mapped, nonduplicate 10-bp paired-end reads and a mean 35× genome-wide coverage. To compare this level of coverage to what can be obtained using exome sequencing, we normalized our exome sequencing data to 50M reads for each platform because this level allows multiplexing at least 3 and up to 6 exomes per lane. The resulting coverage for each platform was 30× mean target coverage for Illumina, 60× for Agilent and 68× for Nimblegen. Thus, using <5% of the number of unambiguously mapped reads, exome sequencing achieved coverage over targets that was in one case nearly equal to (Illumina) and in the two others almost two times as high as (Agilent and Nimblegen) that of WGS.

Sensitivity for detecting SNVs was compared between WGS and the exome sequencing experiments. Variants were called from the WGS data using GATK with the same cut-offs and filters as exome sequencing. The WGS data had 98.5% concordance with SNP Chip at heterozygous positions. The WGS data were restricted to the regions targeted by each platform for comparison. The majority of SNVs were detected by both exome sequencing and WGS across all three platforms, but there were both exome sequencing–specific and WGS-specific SNVs (**Supplementary Table 5**).

The average Phred-based quality scores for SNVs from exome sequencing were much higher than those of SNVs from WGS for Nimblegen (573 in exome versus 320 in WGS) and Agilent (428 versus 192), and very close for Illumina (341 versus 380). The exome sequencing– and WGS-specific SNVs had lower average quality than those SNVs detected by both. A greater proportion of WGS-specific SNVs were of a low quality compared to exome sequencing–specific SNVs for Agilent and Nimblegen (**Supplementary Fig. 3a**,**b**). For Illumina, exome sequencing– and WGS-specific SNV qualities were very similar (**Supplementary Fig. 3c**). This was because 50M reads only generate 30× coverage by Illumina exome sequencing.

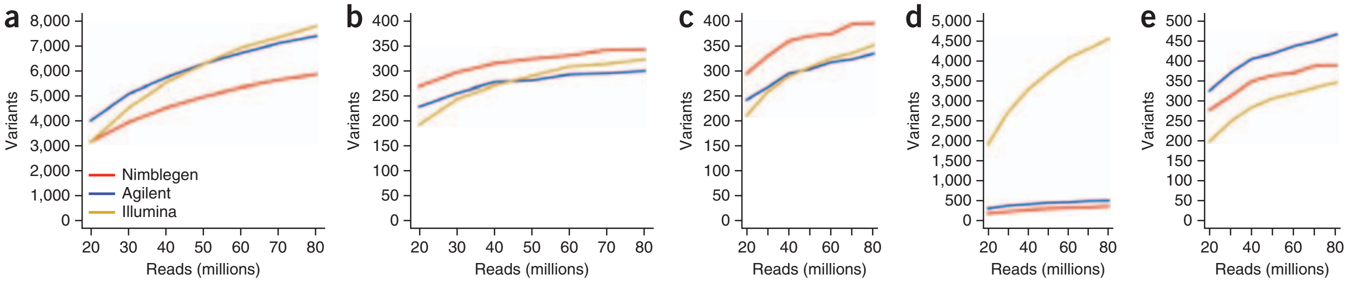


Figure 5 Sensitivity toward indels compared between each platform at increasing read counts. Indel sensitivity may be more intimately tied to factors such as bait length and density compared with SNV sensitivity. (**a**) Total number of indels detected at increasing read count thresholds. As with SNVs, sensitivity increases at higher read counts. Agilent detects the highest quantity at lower read counts because its baits appear more robust toward indels than Illumina's. (**b**) Indels detected in bases targeted by all three platforms. Nimblegen detects the most indels at all read counts because it is the most efficient. Very few indels are detected in the shared interval because it is mostly made up of coding exons, which have a strong bias against indels. (**c**) Indels detected in RefSeq coding exons. These curves match the shared interval curves from **b** closely, much like for SNVs. (**d**) Indels detected in RefSeq UTRs. Again, Illumina detects far more of these because it is the only platform that specifically targets UTRs. (**e**) Indels detected in Ensembl CDS. Agilent detects the most indels in Ensembl CDS due to a combination of the additional 1.4 Mb of targeted Ensembl CDS bases and its high sensitivity toward indels.

Therefore, Illumina was compared again at 60M, 70M and 80M reads yielding coverages of 36×, 42× and 48×, respectively. The quality of all variants increased accordingly, as did the quality of exome sequencing–specific variants.

As variant quality scores are closely tied to read depth over the variant positions, we compared coverage between exome sequencing and WGS over variant positions. Variants detected uniquely by exome sequencing or WGS did indeed have greater average coverage in their detection platform as compared with the total variants detected by both (**Fig. 6a–c**). The WGS-specific SNVs often had zero coverage by exome sequencing (169 Nimblegen SNVs with zero coverage out of 1,235; 615 Agilent SNVs out of 3,362; 2,275 Illumina SNVs out of 6,126), suggesting enrichment failure in these regions. In contrast, very few of the exome sequencing–specific SNVs had zero coverage in WGS (14 Nimblegen-specific SNVs out of 2,291; 13 Agilent-specific SNVs out of 5,199; 24 Illumina-specific SNVs out of 4,385). However, these exome sequencing–specific SNVs tended to have lower than average base coverage in WGS. This even held true for the Illumina exome sequencing at 30× mean coverage relative to WGS. These results indicate that enrichment is capable of bolstering coverage at specific positions that are missed by WGS and leads to more sensitive variant calling in those regions.

To determine as precisely as possible the number of true exome sequencing–specific and WGS-specific SNVs, we estimated the quantity of false-positive SNVs for each experiment. Based on the assumption that SNVs detected by both exome sequencing and WGS (hereafter called shared SNVs) are highly robust, novel variant rates were estimated by comparing shared SNVs with known common SNPs (>1% allele frequency in the population according to dbSNP132). Considering shared SNVs, 13.3% (4,704/35,448) of Agilent's, 11.2% (3,385/30,097) of Nimblegen's and 12.1% (5,151/42,633) of Illumina's were novel by this definition. The novel variant rates were significantly ($P < 10^{-16}$) higher in the exome sequencing–specific SNVs (60.3% of Agilent's, 39.3% of Nimblegen's, 48.9% of Illumina's) and WGS-specific SNVs (56.2% of Agilent's, 59.8% of Nimblegen's, 34.6% of Illumina's), suggesting large quantities of false-positive SNVs in these sets. False positives were estimated by calculating the expected number of novel SNVs in the experiment-specific sets based on the

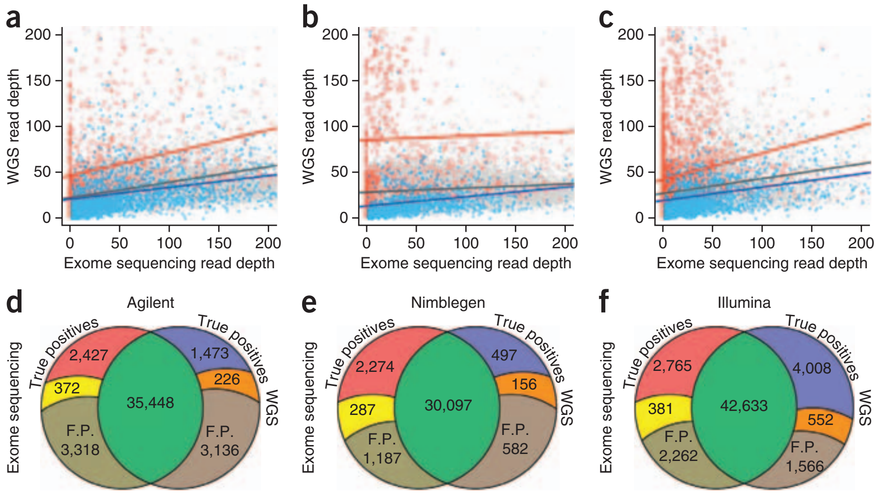


Figure 6 SNVs detected uniquely by exome sequencing or WGS, but not both. A standard WGS experiment at 35× mean genomic coverage was compared to exome sequencing experiments on each platform at 50M reads yielding exome target coverage of 30× for Illumina, 60× for Agilent and 68× for Nimblegen. SNVs were called in the WGS and then restricted to the regions targeted by each platform for comparison. (**a**) SNVs called in Agilent target regions by exome sequencing and WGS plotted as a function of coverage in exome sequencing versus coverage in WGS. Gray dots represent SNVs detected by both exome sequencing and WGS. Light blue dots represent SNVs uniquely called by exome sequencing. Red dots represent SNVs uniquely called by WGS. Lines represent the linear regression of the corresponding points. (**b**,**c**) The same plot as for **a**, but for Nimblegen and Illumina, respectively. For all three exome sequencing platforms, SNVs detected uniquely by exome sequencing had lower than average coverage in WGS. SNVs detected uniquely by WGS were often in targets with zero or very low coverage by exome sequencing. (**d**) Venn diagram of SNVs detected by Agilent exome sequencing and WGS across Agilent targets. SNVs detected by both are in the green section. True-positive exome sequencing–specific SNVs are divided into novel (yellow) and known (red) slices. True-positive WGS-specific SNVs are divided into novel (orange) and known (blue) slices. False positives are in brown. (**e**,**f**) Same as **d**, but for Nimblegen (**e**) and Illumina (**f**), respectively. F.P., false positives.

number of known SNPs. We detected 2,799 Agilent, 2,561 Nimblegen and 3,146 Illumina exome sequencing–specific SNVs and 1,699 Agilent, 653 Nimblegen and 4,560 WGS-specific SNVs (**Fig. 6d**–**f**). The false-positive SNV sets generally have low quality scores, so using a higher quality score threshold on these variants recovers most of the true-positive novel SNVs.

These data demonstrate that there are some regions (and therefore variants) that are missed by a typical WGS but observed by exome-sequencing enrichment because of the higher coverage attainable with target-enriched sequencing over specific regions. Similarly, there are some targeted regions and variants missed by each exome sequencing platform but detected by WGS. Comparison with a large database of disease-related variants[29,30] classified 456 Agilent, 369 Nimblegen and 467 Illumina exome sequencing–specific SNVs as associated with human diseases. Three hundred and one of these were common between all three platforms, suggesting that some regions missed by WGS but captured by exome sequencing have clinical relevance.

DISCUSSION

A comparison study such as this one is intended to inform the research community of the pros and cons of each platform and to help investigators make an informed decision about which platform is best for their research. In this case, all three exome enrichment platforms demonstrate a very high level of targeting efficiency and cover a very large portion of the overall exome. The question of which enrichment platform is best must be answered with respect to specific parameters. We have observed that the densely packed, overlapping baits of the Nimblegen SeqCap EZ Exome demonstrate the highest efficiency target enrichment, able to adequately cover the largest proportion of its targeted bases with the least amount of sequencing. Therefore, the Nimblegen platform is superior to the other two platforms for research restricted to the regions that it covers.

However, greater genomic coverage is desirable to many researchers. The per-base cost of sequencing is plummeting[31], and as a result the optimal balance between efficiency and coverage is changing. As sequencing becomes cheaper, efficiency often becomes less valuable relative to coverage. We have detailed the regions of the genome uniquely covered by each platform because particular regions targeted by one platform may be of interest to specific researchers. Although the Illumina platform demonstrated less targeting efficiency than the others, it is the only platform that is designed to enrich UTRs, which are almost completely untargeted by the other two platforms, and is therefore the natural choice for researchers interested in those regions.

Many researchers performing exome sequencing are most interested in coding regions. Although coding regions can be difficult to define because they differ depending on the database used, the main goal of exome sequencing is discovering variations associated with particular phenotypes. Exomes are particularly powerful for research on Mendelian disorders, as these disorders are often caused by small mutations in gene coding regions. Our results suggest that with regard to the RefSeq exome, Nimblegen has a slight edge in sensitivity for SNPs and small indels. However, with regard to the Ensembl CDS regions, the Agilent SureSelect Human All Exon kit can detect the most SNPs and small indels given slightly more sequencing. All of these platforms can detect disease-associated variants, of which a small proportion are unique to each platform.

Our findings with exome sequencing can be extended to general enrichment principles and custom enrichment assays. We demonstrated multiple levels of bait density and genome coverage that can be used as a guide when designing custom enrichment bait sets. Although it is evident that overlapping baits improved sensitivity, the number of overlapping baits that are necessary remains unclear. What is clear is that an overlapping design is superior to an immediately adjacent or spaced design with regard to enrichment efficiency. Moreover, we observed that the relatively long baits and/or RNA methodology of the Agilent SureSelect allowed for increased sensitivity toward indels. Therefore, longer baits of this type are more desirable in custom assay designs.

It may be argued that the importance of targeted sequencing is transient and will diminish as WGS becomes less expensive. However, we found that exome sequencing can identify variants that are not evident in WGS because of greater base coverage after enrichment. Even at equivalent coverage levels, specific regions had higher read depth in exome sequencing resulting in greater sensitivity in those regions. Target capture by exome sequencing unambiguously identified some of these difficult regions through preferential selection and observation at higher local read depth. These findings demonstrate a strong niche for target enrichment approaches even after WGS sequencing, where targeted sequencing is used to clarify results in regions where WGS yields low depth of coverage, to validate personal variations and to bolster discovery in the most interpretable part of the human genome.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Accession code. Exome sequencing and whole-genome sequencing data are available at the Sequence Read Archive under accession SRA040093.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank P. LaCroute for assistance with data processing and analysis. Thanks to A. Boyle and Y. Cheng for consulting with data analysis and display methods. We thank representatives from Agilent, Illumina and Nimblegen for their support and feedback as we performed these tests. We also thank the Hewlett Packard Foundation and Lucile Packard Foundation for Children's Health for support in creation of our disease/trait SNP database. This work was supported by the US National Institutes of Health grant no. HG002357.

AUTHOR CONTRIBUTIONS

M.S. and R.C. conceived and planned the study. R.C. performed the experiments. G.E. provided sequencing services. M.J.C. conducted the data analysis. R.C. and M.S. both contributed to the data analysis and discussion. H.Y.K.L. and M.J.C. analyzed the whole genome data. K.J.K., R.C. and A.J.B. created the disease/trait SNP database and analyzed our data against it. M.J.C., R.C. and M.S. prepared the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/nbt/index.html.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).
2. Hedges, D. *et al.* Exome sequencing of a multigenerational human pedigree. *PLoS ONE* **4**, e8232 (2009).
3. Lee, H. *et al.* Improving the efficiency of genomic loci capture using oligonucleotide arrays for high throughput resequencing. *BMC Genomics* **10**, 646 (2009).
4. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.* **11**, R119 (2010).
5. Bainbridge, M.N. *et al.* Whole exome capture in solution with 3 Gbp of data. *Genome Biol.* **11**, R62 (2010).
6. Ng, S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
7. Nazarian, R. *et al.* Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation. *Nature* **468**, 973–977 (2010).

8. Glazov, E.A. *et al.* Whole-exome re-sequencing in a family quartet identifies POP1 mutations as the cause of a novel skeletal dysplasia. *PLoS Genet.* **7**, e1002027 (2011).
9. Kalay, E. *et al.* CEP152 is a genome maintenance protein disrupted in Seckel syndrome. *Nat. Genet.* **43**, 23–26 (2011).
10. Shi, Y. *et al.* Exome sequencing identifies ZNF644 mutations in high myopia. *PLoS Genet.* **7**, e1002084 (2011).
11. Snape, K. *et al.* Mutations in *CEP57* cause mosaic variegated aneuploidy syndrome. *Nat. Genet.* **43**, 527–529 (2011).
12. Wheeler, D.A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
13. Ng, S.B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42**, 30–35 (2010).
14. Ng, S.B. *et al.* Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome. *Nat. Genet.* **42**, 790–793 (2010).
15. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
16. Pruitt, K.D., Tatusova, T., Klimke, W. & Maglott, D.R. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.* **37**, D32–D36 (2009).
17. Hsu, F. *et al.* The UCSC known genes. *Bioinformatics* **22**, 1036–1046 (2006).
18. Flicek, P. *et al.* Ensembl 2011. *Nucleic Acids Res.* **39**, D800–D806 (2011).
19. Griffiths-Jones, S., Saini, H.K., van Dongen, S. & Enright, A.J. miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36**, D154–D158 (2008).
20. Dohm, J.C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**, e105 (2008).
21. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18 (2011).
22. Kane, M.D. *et al.* Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.* **28**, 4552–4557 (2000).
23. Kucho, K., Yoneda, H., Harada, M. & Ishiura, M. Determinants of sensitivity and specificity in spotted DNA microarrays with unmodified oligonucleotides. *Genes Genet. Syst.* **79**, 189–197 (2004).
24. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
25. Degner, J.F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207–3212 (2009).
26. Zhang, Z. & Gerstein, M. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res.* **31**, 5338–5348 (2003).
27. Mills, R.E. *et al.* An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* **16**, 1182–1190 (2006).
28. Taylor, M.S., Ponting, C.P. & Copley, R.R. Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. *Genome Res.* **14**, 555–566 (2004).
29. Ashley, E.A. *et al.* Clinical assessment incorporating a personal genome. *Lancet* **375**, 1525–1535 (2010).
30. Chen, R., Davydov, E.V., Sirota, M. & Butte, A.J. Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. *PLoS ONE* **5**, e13574 (2010).
31. Wetterstrand, K.A. DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program. <http://www.genome.gov/sequencingcosts/> (accessed July 15, 2011).

ONLINE METHODS

Sample collection. Whole blood (100 ml) was drawn from a healthy, anonymous volunteer, from whom proper, informed consent was obtained, at the Stanford University Hospital. PBMCs were isolated with Ficoll gradient (Lymphocyte Separation Medium, MP Biomedicals) centrifugation according to the manufacturer's protocol. Genomic DNA was prepared from isolated PBMCs with the AllPrep DNA/RNA/Protein Mini Kit (QIAGEN) and treated with RNase A to remove remaining RNA. DNA concentration was quantified with Invitrogen's Qubit Fluorometer.

Exome enrichment with Agilent SureSelect Human All Exon kit. The kit was a gift from Agilent. Illumina sequencing libraries were prepared according to the manufacturer's instructions. Briefly, 3 µg of genomic DNA was sheared with the Covaris S2 system; the DNA fragments were end-repaired, extended with an 'A' base on the 3′ end, ligated with paired-end adaptors and amplified (four cycles). Exome-containing adaptor-ligated libraries were hybridized for 24 h with biotinylated oligo RNA baits, and enriched with streptavidin-conjugated magnetic beads. The final libraries were further amplified for 11 cycles with PCR, and subjected to Illumina sequencing on one lane of the HiSeq 2000 sequencer.

Exome enrichment with Roche/Nimblegen's SeqCap EZ Exome Library v2.0. The kit was a gift from Roche Nimblegen. Illumina sequencing libraries were made following Nimblegen's protocol with the following improvements: in Chapter 4 Steps 1–4 of the protocol two PCR reactions were set up for each sample with 15 µl of each unenriched sample library as template, and 2 µg of amplified sample library was used for each sample in the hybridization step described in Chapter 5 Step 2. In summary, 3 µg of genomic DNA was sheared with the Covaris S2 system, DNA fragments were concentrated with ethanol precipitation, end-repaired with the Epicentre End-It DNA End-Repair Kit, a deoxyadenosine was added at the 3′ end of the fragments with the Klenow 3′->5′ exo- enzyme (New England Biolabs), and ligated with Illumina's Paired-End Adaptor Oligo Mix. The ligated libraries were size selected for an average insert size of 250 bp (2 mm gel slice) by agarose gel excision and extraction, amplified for eight cycles by Pre-Capture LM-PCR, and hybridized for 72 h with biotinylated oligo DNA baits for exome-containing libraries. The hybridized libraries were enriched with streptavidin-conjugated magnetic beads and washed and amplified by PCR (18 cycles), and the quality of the libraries was checked by qPCR as described in the protocol. The final libraries were submitted for Illumina sequencing on one lane of the HiSeq 2000 sequencer.

In summary, we made two modifications to improve Nimblegen performance. With the original protocol, our experience was that sometimes we could not get enough amplified library for both the hybridization and the final qPCR validation, as quantified with Picogreen. Therefore we increased the cycles of PCR and split each reaction in two to ensure the sufficiency of PCR. In addition, we also increased the amount of the amplified libraries from 1 to 2 µg in the hybridization step to make the most use of the enrichment probes.

Illumina TruSeq Exome Enrichment. Illumina's TruSeq Exome Enrichment Kit was acquired as a free sample from Illumina. Pre-enrichment DNA libraries were constructed following Illumina's TruSeq DNA Sample Preparation Guide. A 300- to 400-bp band was gel selected for each library and exome enrichment was performed according to Illumina's TruSeq Exome Enrichment Guide. Two 20-h biotinylated bait-based hybridizations were performed with each followed with Streptavidin Magnetic Beads binding, a washing step and an elution step. A 10-cycle PCR enrichment was performed after the second elution and the enriched libraries were subjected to Illumina sequencing after quality check on one lane of HiSeq 2000.

Exome sequencing by Illumina HiSeq 2000. Libraries were denatured with sodium hydroxide and loaded onto an Illumina cBot for cluster generation according to the manufacturer's recommended protocols (TruSeq PE Cluster Kit v2). Lane 5 of each flow cell was reserved for a PhiX control. The primer-hybridized flow cells were then transferred to HiSeq 2000 sequencers and paired-end sequencing was done with TruSeq SBS kits (Illumina) in a 2 × 101b mode.

Libraries derived from exome samples from each of the three exome enrichment kits were run on one lane of the HiSeq each. Total read counts off the machine for each platform:

124,112,466 raw reads for Agilent SureSelect Human All Exon
184,983,780 raw reads for Nimblegen SeqCap EZ-Exome Library SR v2.0
112,885,944 raw reads for Illumina TruSeq Exome

Exome sequencing alignment. Raw reads in FASTQ format from each exome sequencing lane were aligned to the human reference genome (hg19) with BWA using default parameters with the -q 30 parameter to include soft clipping of low-quality bases. Total aligned read counts for each platform:

123,292,356 aligned reads for Agilent (99.3%)
183,502,451 aligned reads for Nimblegen (99.2%)
110,977,932 aligned reads for Illumina (98.3%)

Aligned reads were processed and sorted with SAMtools[32] and PCR duplicates were removed with Picard MarkDuplicates (http://picard.sourceforge.net). Final unambiguous, aligned read counts:

94,779,030 unambiguous, aligned reads for Agilent (76.4%)
154,270,343 unambiguous, aligned reads for Nimblegen (83.4%)
88,759,249 unambiguous, aligned reads for Illumina (78.6%)

Read-count normalization and thresholding. Total read count was normalized to 80M reads between all three exome sequencing experiments by randomly drawing 80M reads from each aligned and filtered read set. Further thresholding was done by randomly drawing 20M, 30M, 40M, 50M, 60M and 70M reads from the aligned and filtered read sets. Raw data were not realigned, but rather reads were directly taken from the whole aligned data sets.

Exome sequencing variant calling. Single-nucleotide variants were called with GATK[24] in a default mode as recommended by the GATK documentation for the normalized and read count thresholded data separately. The GATK variant calling pipeline was run on every data set independently. Briefly, reads around small variants and mapping artifacts were realigned, balanced based on covariates and assessed for genotyping. The UnifiedGenotyper was run with the *-baq CALCULATE_AS_NECESSARY* parameter and using a *stand_emit_conf* of 10.0 and *stand_call_conf* of 30.0. All variants with a Phred-based quality score <30.0 were called low quality and ignored.

Indels were called with the GATK UnifiedGenotyper using the Dindel model called by *-glm DINDEL*[33]. Indels were also called with a *stand_emit_conf* of 10.0 and *stand_call_conf* of 30.0. As with SNVs, all variants with a Phred-based quality score less than 30.0 were called low quality and ignored.

Exome sequencing variants for each platform are provided in **Supplementary Data 1** (SNVs) and **Supplementary Data 2** (indels). Variant counts and other metrics are presented in **Supplementary Table 4**.

Illumina whole genome sequencing library preparation. Paired-end 101b sequencing libraries were generated from the human PBMC whole genome DNA sample according to the Illumina HiSeq 2000 library generation protocol.

Illumina Whole Genome Sequencing. The whole genome library was handled exactly as the exome enrichment libraries were for sequencing on the HiSeq 2000 (see "Exome Sequencing by Illumina HiSeq 2000" section). The whole genome library was run on seven lanes of HiSeq (with lane 5 reserved for PhiX control). This yielded more than 1 billion total raw reads. Reads were aligned with BWA using the -q 30 parameter and duplicates were removed with Picard. After these filtering steps, 1,194,622,756 unambiguous, aligned reads were produced.

doi:10.1038/nbt.1975

WGS variant calling. Small variant calling on the WGS data was performed in the same manner as for the exome sequencing experiments. In short, the processed reads were run through GATK using the same parameters as were used with the exome sequencing experiments for both SNVs and indels. 3,773,305 raw SNVs and 616,355 raw indels were detected. After filtering out low-quality variants, 3,695,769 SNVs and 600,752 indels remained.

Whole genome variants will be hosted at the Sequence Read Archive as metadata with the sequence data.

SNP Chip. DNA derived from PBMCs was sent to Illumina to be run on the Illumina Human 1M-Duo SNP Chip. Illumina called SNP genotypes using their GenomeStudio program and returned the list of genotypes.

32. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
33. Albers, C.A. *et al.* Dindel: accurate indel calls from short-read data. *Genome Res.* **21**, 961–973 (2011).

doi:10.1038/nbt.1975