Supporting Information

Belkadi et al. 10.1073/pnas.1418631112

SI Text

Analysis of High-Throughput Sequencing Data. We used the Genome Analysis Software Kit (GATK) best-practice pipeline to analyze our WES and WGS data (1). Reads were aligned to the human reference genome (hg19) using the Maximum Exact Matches algorithm in Burrows–Wheeler Aligner (BWA) (2). Local realignment around indels was performed by the GATK (3). PCR duplicates were removed using Picard tools (broadinstitute.github.io/picard). The GATK base quality score recalibrator was applied to correct sequencing artifacts. We called our six WES simultaneously together with 24 other WES using Unified Genotyper (UG) (3), as recommended by the software to increase the chance that the UG calls variants that are not well-supported in individual samples rather than dismiss them as errors. All variants with a Phred-scaled SNP quality ≤ 30 were filtered out. The UG calling process in WGS was similar to that used for WES; we called our six WGS together with 20 other WGS. In both WES and WGS, the calling process targeted only regions covered by the WES 71 Mb kit + 50 bp flanking each exon (4). When we expanded the WES regions with 100 and 200 bp flanking each exon, as performed in some previous studies (5–7), we observed a higher genotype mismatch in variants called by WES and WGS, with a much lower quality of the WES variants located in those additional regions.

Matched and mismatched genotype statistics, analyses of variant coverage depth (CD) (i.e., the number of reads passing quality control used to calculate the genotype at a specific site in a specific sample), genotype quality (GQ) (i.e., a Phred-scaled value representing the confidence that the called genotype is the true genotype), and minor-read ratio (MRR) [i.e., the ratio of reads for the less covered allele (reference or variant allele) over the total number of reads covering the position where the variant was called] were performed using a homemade R software script (8).

We then filtered out variants with a CD of <8 or a GQ of <20 or an MRR of <20% as suggested in ref. 9 using a homemade script .We used the Annovar tool (10) to annotate high-quality (HQ) variants that were detected exclusively by one method. We checked manually some HQ coding variants detected exclusively by WES or WGS using the Integrative Genomics Viewer (IGV) (11), and we observed that some HQ-coding WES-exclusive variants were also present in WGS but miscalled by the UG tool. To recall the UG miscalled SNVs, we used the GATK haplotype caller (HC) tool (3). Indels and SNVs were called simultaneously using HC on six WES and six WGS. We then split SNVs and indels into two combined variant call format (vcf) files. The same DP, GQ, and MRR filters were applied for both SNV and indels, and we used Annovar to annotate the HQ resulting variants.

CNVs were detected on WES data from our 6 samples together with 24 other samples originating from Europe using XHMM (12) and Conifer (13). For XHMM, we first calculated the depth of coverage in the 789,124 WES targets using GATK. XHMM was then run using default parameters to infer CNVs from read depths as previously described (12). For Conifer, the SVD-ZRPKM thresholds algorithm was used with the default parameters to find CNV breakpoints (13). For WGS data, we ran Genome STRiP (14) with the default parameters to detect large deletions on our 6 WGS together with 24 other WGS European samples from the 1000 Genomes database (15). Genome STRiP looks for signatures of large deletions indicated by unusual spacing or orientation read pairs. We then kept only deletions that overlapped with at least one WES-targeted region. We looked whether the CNVs identified were present in the DGV database in February 2015 (16).

All scripts are available on https://github.com/HGID/WES_vs_WGS.

Sanger-Sequencing Methods.

Selection of variants. We randomly selected variants detected exclusively by WES or WGS to test them by Sanger sequencing. We only sequenced once exclusive variants that were identified in multiple samples. We chose fewer variants in sample S1 because we had few genomic DNA (gDNA) available for this sample, and we could not test any of the variants in S2 because of the absence of remaining gDNA. No other criteria (position, gene, CADD score, frequency, size of indel, etc.) were used for deciding which variants to Sanger sequence. For SNVs, we chose more variants in the two categories of WES fully-exclusive and WGS fully-exclusive as we first hypothesized (wrongly) that most, if not all, partly-exclusive variants would be real.

Design of the primers. The first step was to create a bed file with each row representing a region of 400 bp centered on the variants chosen for Sanger sequencing. The bed file was then uploaded in the University of California, Santa Cruz (UCSC) genome browser using the "add custom tracks" tab. The reference genome assembly used was GRCh37/hg19 (https://genome.ucsc.edu/cgi-bin/hgGateway). Fasta files with the sequence for each region were then downloaded from the UCSC website and uploaded to BatchPrimer3 v1.0 (batchprimer3.bioinformatics.ucdavis.edu/cgi-bin/batchprimer3/batchprimer3.cgi) (17). We noticed that BatchPrimer3 worked better if the fasta files were copied and pasted rather than uploaded using a link. We then requested for Sequencing primers using the following parameters: nb of return = 1 (1 toward 3′, and 1 toward 5′); sequencing start = −1; primer size: Min = 18, Opt = 22, Max = 25; primer Tm: Min = 55, Opt = 58, Max = 62; Max self complementarity = 8; Max 3′ self complementarity = 3. Lastly, variants for which one of the two primers was closer to 60 bp to the variant were excluded from further sequencing and analysis. M13F or M13R sequences were added at the 5′ end of the forward or reverse primers. The full list of primers ordered is available in Dataset S1.

Sequencing of the variants. Amplification of the variants was performed by using, per reaction, the following: H_2O = 11.5 μL, 40% (vol/vol) glycerol = 4.5 μL, 10× buffer (Denville without $MgCl_2$) = 2.25 μL, $MgCL_2$ (25 mM) = 0.9 μL, dNTP (10 mM) = 0.225 μL, primers (10 μM) = 0.5 μL each, Taq Polymerase (CB4050-2; Denville) = 0.5 μL, DNA = 50–100 ng. DNA was substituted by H_2O in negative controls. Thirty-eight cycles of 94 °C (30 s), 60 °C (30 s), 72 °C (1 min) were performed on a Veriti Thermal Cycler (Life Technologies). Sequencing PCR was done using the Big Dye 1.1 (Life Technologies) protocol with 1 μL of amplification PCR product and either the M13F or the M13R primer on a Veriti Thermal Cycler (Life Technologies). Lastly, the samples were sequenced on an ABI 3730 XL sequencer (Life Technologies). Sanger sequencing was attempted only once for each variant.

Analysis of the Sanger sequences. For SNVs, the analysis of the Sanger sequences was done using the DNASTAR SeqMan Pro software (v11.2.1) using the default settings. To facilitate the localization of the potential variants, we assembled the sequences obtained by Sanger with a 20-bp fasta sequence centered on each variant. This sequence was obtained by creating a bed file of the region in the same way as described for the primer design. Variants where either the forward or reverse sequence did not work were excluded from the analysis and assigned an NA on the Sanger sequencing results (Dataset S1). For indels, the analysis of the Sanger sequences was

much more difficult, and it was not possible to use the DNASTAR SeqMan Pro software. Instead we used the software ApE (A plasmid Editor) to visualize every peak as clearly as possible. We then reconstructed the two alleles manually for each variant tested. For several indels, the analysis or results seemed intermediate. We considered that a variant was a false positive if (*i*) there was no insertion or deletion at the place identified, or (*ii*) the size or sequence of the indel was incorrect, or (*iii*) the height of the peaks corresponding to the mutant allele were higher than the background noise usually observed; in practice, we validated indels that had sequencing peaks with a height that was >20% of the height of WT peaks. Lastly, we encountered several indels that were a combination of a deletion and an insertion. For example, the WT sequence would be AAAAAAAAA, and the mutated sequence would be AAACAAA. Analysis of WES and WGS did not integrate these calls into one. We considered the results of WES or WGS true if WES or WGS called both the deletion of AAA and an insertion of a C in this example.

1. DePristo MA, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5):491–498.
2. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26(5):589–595.
3. McKenna A, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9): 1297–1303.
4. Meynert AM, Ansari M, FitzPatrick DR, Taylor MS (2014) Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics* 15:247.
5. Sulonen A-M, et al. (2011) Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol* 12(9):R94.
6. Wang K, et al. (2011) Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. *Nat Genet* 43(12):1219–1223.
7. Szpiech ZA, et al. (2013) Long runs of homozygosity are enriched for deleterious variation. *Am J Hum Genet* 93(1):90–102.
8. R Development Core Team (2013) R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria).
9. Carson AR, et al. (2014) Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Bioinformatics* 15:125.
10. Wang K, Li M, Hakonarson H (2010) ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38(16):e164.
11. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief Bioinform* 14(2): 178–192.
12. Fromer M, et al. (2012) Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet* 91(4):597–607.
13. Krumm N, et al.; NHLBI Exome Sequencing Project (2012) Copy number variation detection and genotyping from exome sequence data. *Genome Res* 22(8):1525–1532.
14. Handsaker RE, Korn JM, Nemesh J, McCarroll SA (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* 43(3):269–276.
15. 1000 Genomes Project Consortium, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56–65.
16. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW (2014) The Database of Genomic Variants: A curated collection of structural variation in the human genome. *Nucleic Acids Res* 42(Database issue):D986–D992.
17. You FM, et al. (2008) BatchPrimer3: A high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics* 9:253.

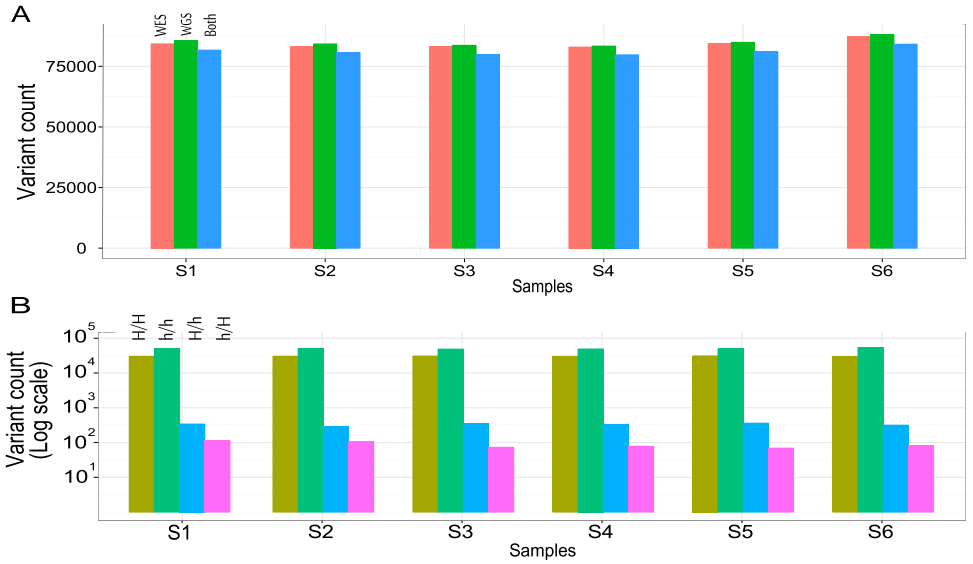

Fig. S1. Number and general characteristics of single-nucleotide variants (SNVs) called by WES and WGS. (*A*) Total number of SNVs called by WES alone, WGS alone, and both platforms. (*B*) Characteristics of the SNVs called by both WES and WGS for each sample, with four columns indicating the number of SNVs called homozygous by both methods (H/H, light green), called heterozygous by both methods (h/h, dark green), called homozygous by WES and heterozygous by WGS (H/h, blue), and called heterozygous by WES and homozygous by WGS (h/H, purple).

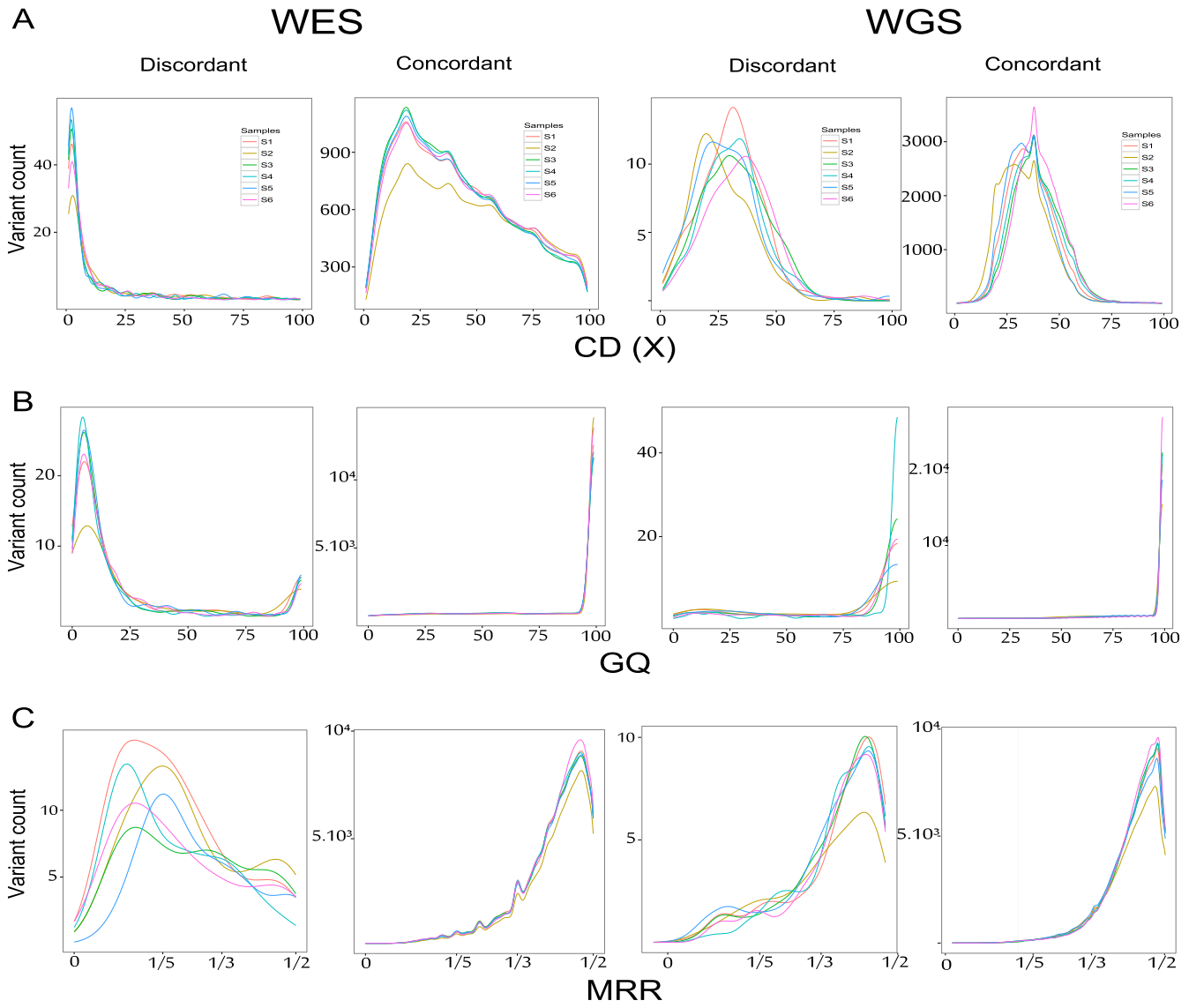

Fig. S2. Distribution of the three main quality parameters for the SNVs with genotypes discordant between WES and WGS. (*A*) Coverage depth (CD), (*B*) genotype quality (GQ) score, and (*C*) minor-read ratio (MRR). For each of the three parameters, four panels are shown: the two panels on the *Left* show the characteristics of discordant and concordant SNVs in WES samples; the two panels on the *Right* show the characteristics of discordant and concordant SNVs in WGS samples.

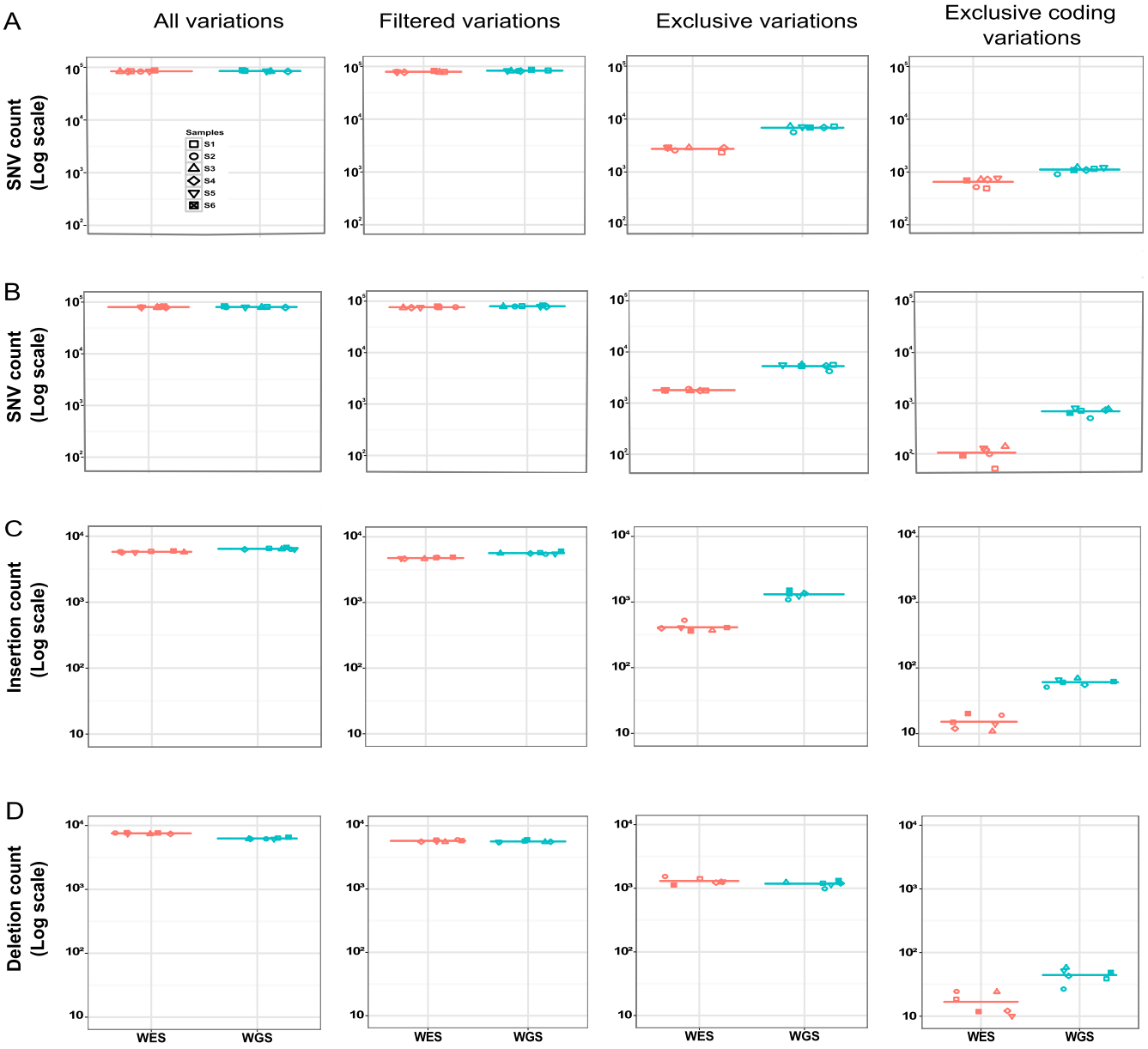

Fig. S3. Numbers of variations in each WES or WGS sample after the application of various filters: (*A*) SNVs called using Unified Genotyper, (*B*) SNVs called using the intersection of Unified Genotyper and Haplotype Caller, (*C*) insertions, and (*D*) Deletions. Insertions and deletions were called using Haplotype Caller. For each of the four panels, we show from *Left* to *Right*: total number of variations called by WES (red) or WGS (turquoise) for each sample; total number of high-quality variations satisfying the filtering criteria of a CD of ≥8×, a GQ of ≥20, and an MRR of ≥0.2 called by WES (red) or WGS (turquoise) for each sample; number of high-quality variations called by only one method, after filtering, high-quality exclusive WES variations (red) and high-quality exclusive WGS variations (turquoise); and number of exclusive WES (red) and exclusive WGS (turquoise) high-quality coding variations.

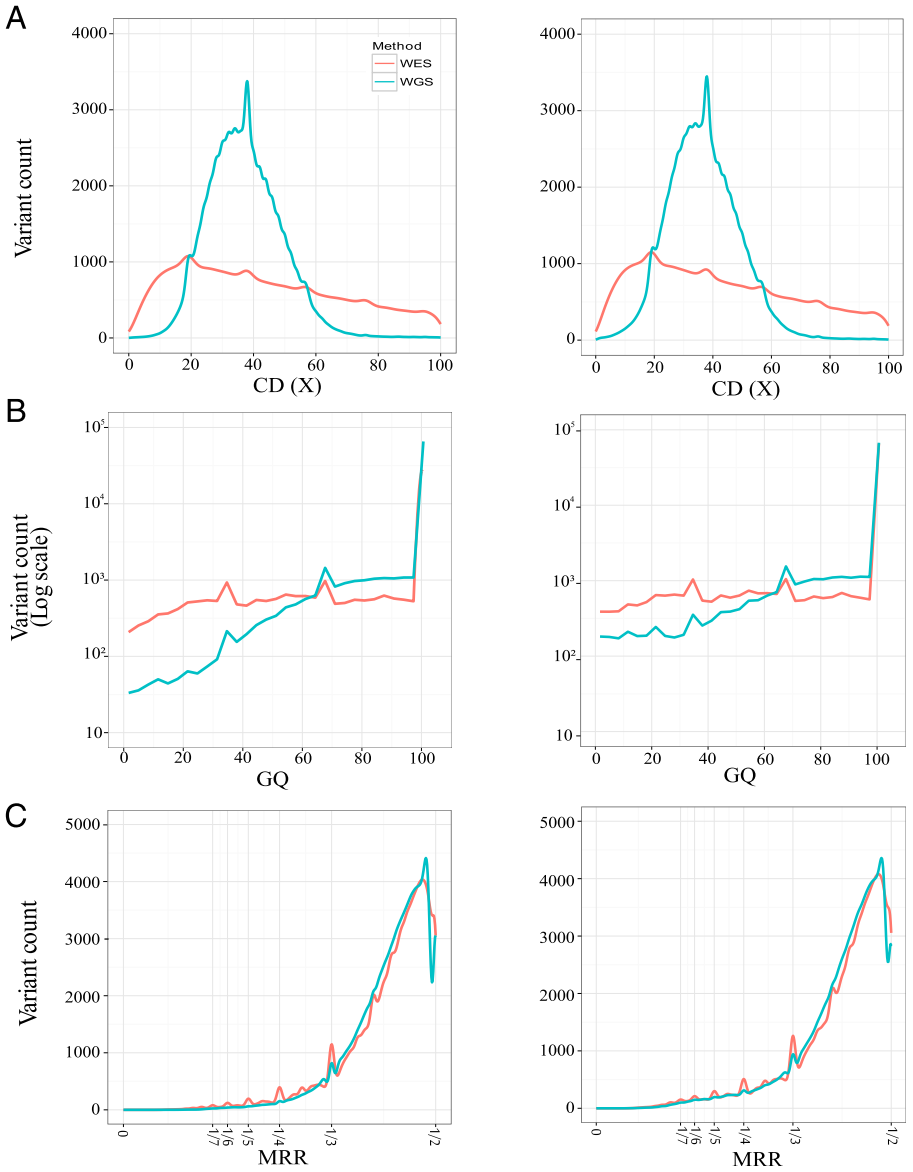

Fig. S4. Comparison of the distribution of the three main quality parameters for the SNVs detected by WES or WGS, with either the intersection of Unified Genotyper and Haplotype Caller, or with Unified Genotyper alone. (*A*) Coverage depth (CD), (*B*) genotype quality (GQ) score, and (*C*) minor-read ratio (MRR). For each of the three parameters we show the average over the six WES (red) and the six WGS (turquoise) samples for the intersection of callers (*Left*) and for Unified Genotyper alone (*Right*).

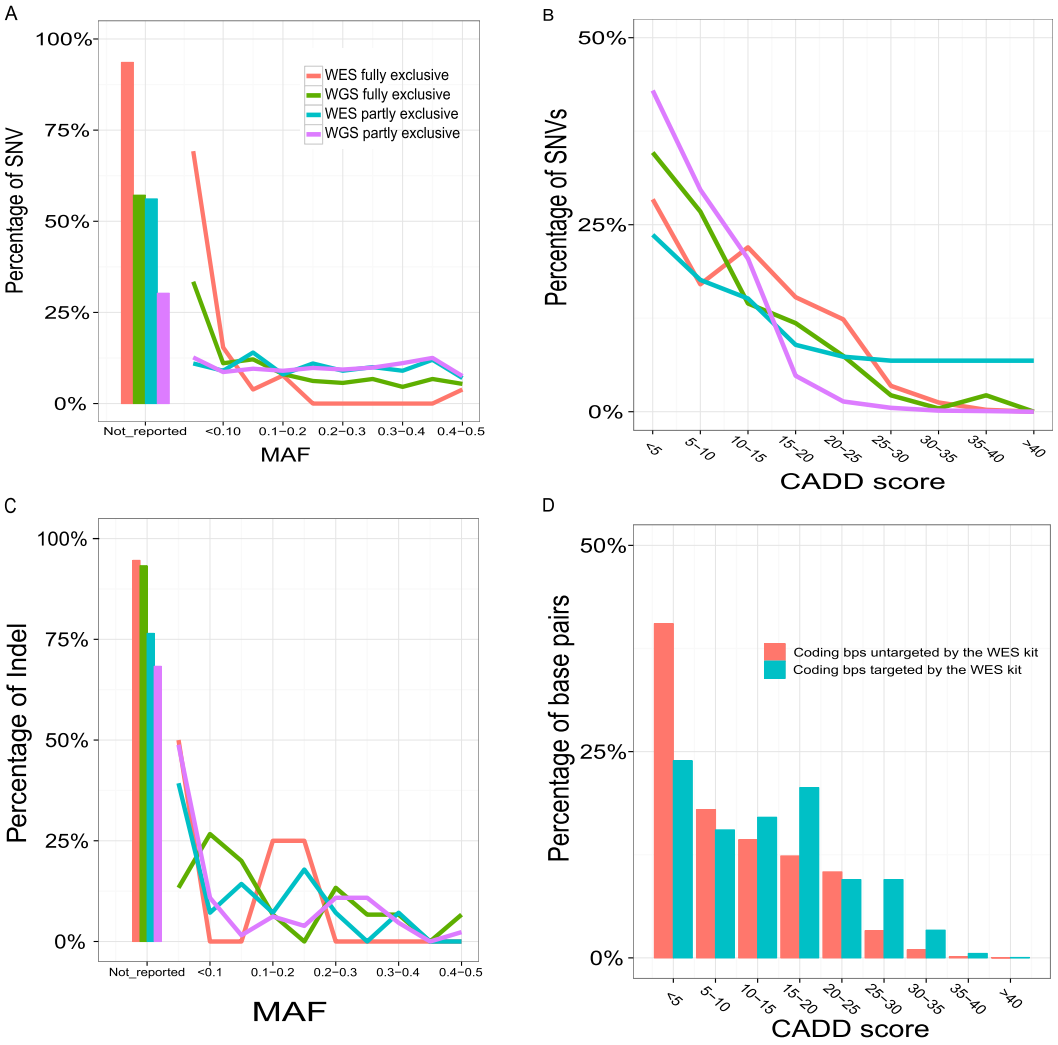

Fig. S5. Characteristics of variations missed by WES or WGS. (*A* and *C*) Distribution of high-quality coding SNVs (*A*) and indels (*C*) based on their presence and minor allele frequency (MAF) in the 1000 Genomes database. (*B* and *D*) Distribution of CADD (combined annotation-dependent depletion) scores (done on version 1.2) for high-quality coding SNVs identified exclusively by WES or by WGS (*B*) and for all base pairs included in the high-quality CCDS exons that were targeted (blue) or untargeted (red) with the 71 Mb ± 50 bp kit (*D*). For *A*, *B*, and *C*, red represents fully exclusive high-quality WES coding variation never identified by WGS; turquoise represents partly exclusive high-quality WES coding variations identified by WGS but filtered out due to their poor quality; green represents fully exclusive high-quality WGS coding variations never called by WES; and purple represents partly exclusive high-quality WGS coding variations identified by WES but filtered out due to their poor quality.

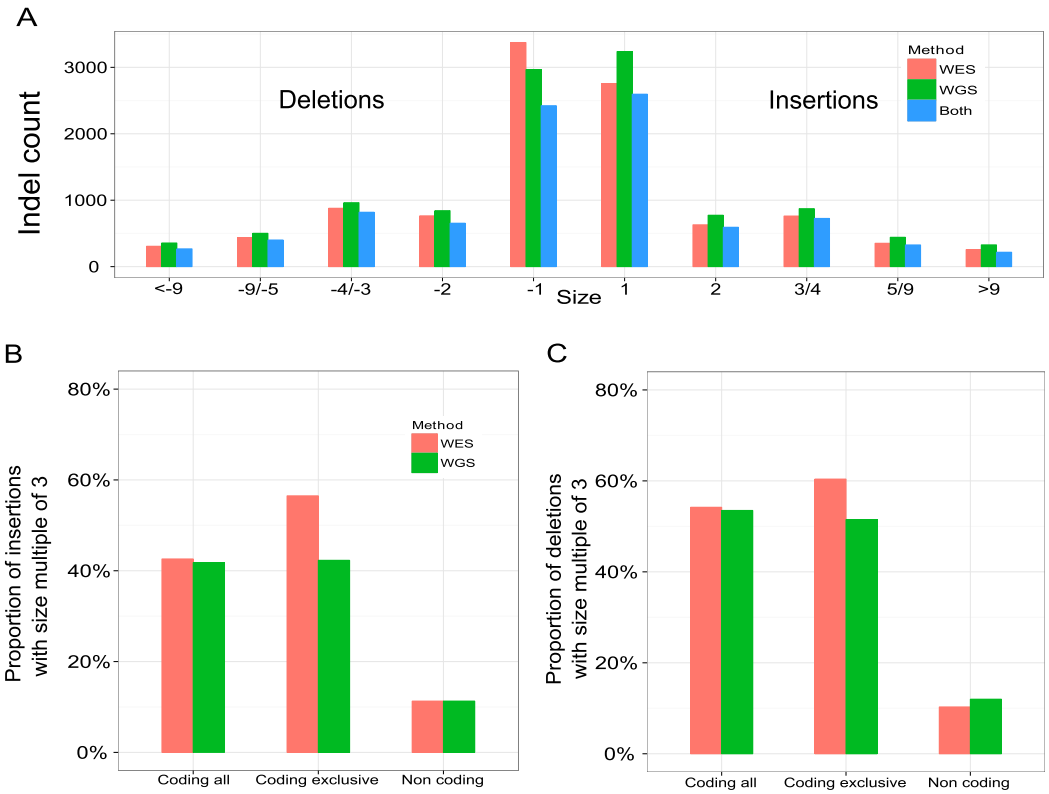

Fig. S6. Distribution of high-quality indel size. (*A*) Distribution of high-quality indels detected by WES (red), by WGS (green), and indicating those detected by both methods (blue) according to their size grouped in five categories: 1 bp, 2 bp, 3–4 bp, 5–9 bp, and ≥10 bp. (*B*) Proportion of high-quality insertions with size multiple of 3 in coding and noncoding regions detected by WES (red) and WGS (green). (*C*) Proportion of high-quality deletions with size multiple of 3 in coding and noncoding regions detected by WES (red) and WGS (green). For coding regions, we show both the total numbers of insertions/deletions and those that are WES- or WGS-exclusive.

Table S1. Reads and coverage statistics for each WES and each WGS

Sample	Total no. of WES reads	Total no. of WGS reads	No. of WES reads aligned in WES regions ± 50 bp	No. of WGS reads aligned in WES regions ± 50 bp	WES mean coverage in WES regions ± 50 bp	WSG mean coverage in WES regions ± 50 bp
S1	98,792,738	1,370,493,918	64,696,895	34,737,193	72.1	38.7
S2	124,483,242	1,303,868,290	80,970,674	31,743,245	90.3	35.3
S3	86,822,862	1,477,715,120	57,970,027	37,322,280	64.5	41.5
S4	89,521,104	1,438,287,290	59,084,117	36,600,011	65.9	40.7
S5	98,002,162	1,301,586,284	62,673,065	33,102,614	69.9	36.8
S6	100,056,600	1,445,702,068	68,002,983	37,619,386	75.8	41.9
Mean	99,613,118	1,389,608,828	65,566,294	35,187,455	73.1	39.2

Dataset S1. Sanger-sequencing results

Dataset S1

Dataset S2. List of poorly covered genes in WES data

Dataset S2