Journal of Genomes and Exomes

ORIGINAL RESEARCH

Systematic Identification and Definition of Consistently Well-Characterized Protein-Coding Exons Using Next Generation Sequencing Technology

Praveen F. Cherukuri[1,2], Murat Sincan[3], John P. Accardi[3], Karin Fuentes Fajardo[1,2], Thomas C. Markello[1,2], Cornelius F. Boerkoel[1,2], Cynthia J. Tifft[1,2], William A. Gahl[1–3] and David Adams[1,3]

[1]NIH Undiagnosed Diseases Program and NHGRI, National Institutes of Health, Bethesda, Maryland, USA. [2]Office of the Clinical Director, National Human Genome Research Insitute, NIH, Bethesda, Maryland, USA. [3]Medical Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, Maryland, USA.
Corresponding author email: cherukur@mail.nih.gov

Abstract: The ability to target and capture known exons in the human genome, and characterize them by massively parallel sequencing, has led to the identification of the genetic causes of many Mendelian disorders. Several factors suggest that exome sequencing will be the preferred clinical next generation technology for some time to come. Advantages of high sequencing depth include the low cost/coverage compared with genome sequencing, and the fact that non-coding-sequence interpretation is still in the early stages of development. In this study of data from the NIH Undiagnosed Diseases Program (UDP), we investigated a novel approach to quantify the quality of exome sequencing data. We systematically and thoroughly evaluated the genotypable fraction across well-characterized protein-coding exons and found that >88% are genotyped to completion and, on average, >93% of all coding bases were genotyped (with target sequencing efficiency of 96%). We also demonstrate a methodology for robust identification of consistently genotyped exons using a new statistical metric, the index of dispersion. This methodology allowed us to define the overall genotypeability of all 167,717 autosomal exons and 95.5% of these had a reproducible pattern of sequencing. Finally, we developed a computational application to take advantage of the reproducible and predictable pattern to confidently detect homozygous deletion events of protein-coding exons. We exploited the sequence pattern information towards reduction of search complexity to detect homozygous deletion events. Of our 11 predictions of homozygous exon-deletion events, we studied 3, performing wet lab experiments that confirmed and validated each of them. We conclude that our systematic approach to analyzing exome sequence data across our patient cohort provides a powerful computational methodology to evaluate, assess, interpret and predict patterns that are relevant to the pathophysiology of the sequenced individuals.

Keywords: Next generation sequencing reproducibility, index of dispersion, homozygous exon-deletion detection

Journal of Genomes and Exomes 2013:2 1–18

doi: 10.4137/JGE.S10089

This article is available from http://www.la-press.com.

Introduction

Current methods to identify genetic causes of Mendelian disorders rely upon exon capture and massively parallel sequencing.[1–5] This methodology is becoming increasingly popular as the cost of sequencing falls. Typically, between 20,000 and 50,000 variants are identified per sequenced exome, of which between 8,000 and 10,000 are non-synonymous variants[6,7] and are readily interpretable. As a result of gene-agnositc sequencing approach, studies continue to reveal unexpected genes as the primary cause of new diseases.[8–11]

In considering exome sequencing as a clinical and diagnostic tool, clinicians and basic researchers alike face several key challenges, such as preferential selection of certain genomic regions,[12,13] variable definitions of exons by different gene annotation databases,[6] variable and non-uniform genotypeability of a given genomic locus by exome sequencing across members of a given family, and complete lack of sequence data for exons of known disease-causing genes.[14] In order to routinely employ exome sequencing as a clinical tool,[9] it is imperative systematically to address these hurdles.

We employed a rigorous, systematic analytical approach to evaluate the reproducibility of exome sequencing data from patients enrolled in the NIH Undiagnosed Diseases Program (UDP).[15] We investigated whether statistical analysis of the quality of genotypeability within a set of individuals, across well-annotated bases in the human exome, could help identify candidate variants. Using genotypeability of these genomic regions as a metric, we present a novel statistical approach for defining reproducibility of sequencing. These results show that the majority of exons in our data are highly reproducible across the individuals that we evaluated. In addition, we show that using quantitative metrics for defining well-sequenced exons permits the use of these regions to detect gross chromosomal abnormalities, such as large exon-size homozygous deletion events.

Methods

Patients

Patients accepted into the NIH Undiagnosed Diseases Program (UDP) were enrolled in clinical protocol 76-HG-0238, approved by the Institutional Review Board (IRB) of the National Human Genome Research Institute (NHGRI), and gave written, informed consent. An additional anonymized dataset of 801 exome sequences derived from the ClinSeq™ study[16] was used for validation and filtering data.

DNA extraction

DNA was extracted from 10 mL of peripheral whole blood from each individual in study using the Puregene kit (Qiagen, Inc, Valencia, CA) according to the manufacturer's protocol as previously described.[17]

Exome sequencing, alignment and genotype-calling

Exome sequencing was performed using genomic DNA extracted from peripheral blood. In-solution exome capture was performed according to the manufacturer's protocol using SureSelect Human All Exon Kits (Agilent Technologies, Santa Clara, CA) or the TruSeq Exome Enrichment Kit (Illumina, San Diego, CA). The Agilent 38 Mb kit was used for 79 UDP samples, the Agilent 50 Mb kit was used for 47 UDP samples, and the Illumina TruSeq kit was used for 45 UDP samples. Flow-cell preparation and 76 to 100-bp paired-end (PE) read sequencing were performed per the protocol for the Illumina Genome Analyzer II_x and Illumina HiSeq 2000 (Illumina, San Diego, CA). For all subsequent analyses, sequence data derived from libraries constructed using each kit were grouped to give three data sets. The 45 TruSeq samples were further grouped into cohorts of 11 (test-set), 15 (founders) and 19 (probands), respectively.

For the 171 exome data, reads were aligned to the human genome reference sequence (UCSC assembly hg18, NCBI build 36) using `eland` (Illumina, Inc, San Diego, CA). The program `eland` was used in such a way that paired-end (PE) reads were aligned independently, and those that aligned uniquely were grouped into genomic sequence intervals of about 100 kb. Reads that failed to align were binned with their PE mates without eland using the PE information. Reads that mapped equally well in more than one location were discarded. The program `Cross_Match`, a Smith-Waterman based local-alignment algorithm, was used to align binned reads to their respective 100 kb genomic reference sequence, using the parameters, `–minscore 21` and `–masklevel 0`

(http://www.phrap.org).[18] Genotypes in the next-generation data were called using a Bayesian genotype-caller, Most Probable Genotype (MPG).[19] The MPG calling algorithm calculates a Bayesian posterior probability of all possible genotypes at a given position, and reports a score, which is computed as the difference in probabilities between the most probable and the second most probable genotypes.

Sanger sequencing

Sanger sequencing was performed for 40 randomly selected exons that were poorly genotyped during NGS to confirm sequence rescue according to standard protocols. We performed PCR amplification using HotStar Taq (Qiagen, Valencia, CA). PCR products were sequenced by dideoxynucleotide chain-termination sequencing (Macrogen, Seoul, Korea). Sequences were aligned and analyzed using Sequencher software program (v.4.10.1, Gene Codes, Ann Arbor, Michigan).

Exon genotypeability and reproducibility

Model

Given N independent random variables $x_1, x_2, \ldots, x_N$ each following the same probability law, the quantity $D = (1/N)\sum_{i=1}^{N}(x_i - \bar{x})^2 \Big/ (1/N)\sum_{i=1}^{N} x_i$ is known as the index of dispersion or variance to mean ratio (VMR). The value of D allows for testing the distribution of observed data, compared to the null hypothesis of random observations. If the variable *xi* follows a Poisson model, then the variance of the distribution is equal to its mean, and the value of D is equal to 1. The index of dispersion (D) therefore allows testing whether observations are uniformly dispersed ($D = 0$), underdispersed ($0 \leq D < 1$) or overdispersed ($D > 1$). The uniform and underdispersed data correspond to having more variables being equal to or closer to the mean than in the Poisson distribution. In contrast, overdispersed data indicate that there are clustered random variables compared to the Poisson distribution. Statistically significant deviations in either direction will lead to the rejection of the hypothesis of randomness (null hypothesis in Poisson goodness of fit testing).

Index of dispersion of exons

Consider an exon k of length L and let y_{kj} be an indicator variable of genotypeability (genotype-score (MPG $\geq$ 10)) at position j. We define the genotypeability (x_k) of a given exon k as the percentage of the sum of indicator variables over the length of a given exon k as described below.

$$x_k = \left(\frac{1}{\mathrm{L}}\sum_{1}^{\mathrm{L}} y_{kj}\right) * 100$$

The mean of genotypeability of the exon k in a cohort of N individuals is defined as $\bar{x}_k$ (average genotypeability) and its variance is defined as σ_k^2. Given these variables, we calculated the index of dispersion D for a given exon k across N individuals as follows:

$$D_k = \frac{\sigma_k^2}{\bar{x}_k}$$

We computed the value of D for all well annotated exons, given a set of individuals sequenced using the the same chemistry and exome capture kit. The code to compute values of D for each cohort of individuals and all exons was written in PERL programming language (version 5.12). Scatterplots and matrices of D were generated using the statistical computing software R (version 2.12.2; 32-bit build).

Quantitative PCR (qPCR) amplification

qPCR amplification was performed using 200 ng of UDP2179 or UDP2473 genomic DNA and Bio-Rad SSO Fast EvaGreen supermix (Hercules, CA). The amplification was carried out with an initial denaturation at 95 °C for 30 s, followed by 30 cycles of denaturation at 95 °C for 5 s, annealing and extension at 65 °C for 1 s. Unaffected human genomic DNA and no template control (NTC) samples were run for each set of primers as positive and negative controls, respectively.

Statistical methods

Pearson's Chi-squared test of independence between variables D and other sequence related features (GC-content and low-mappability) were performed in the statistical computing software R (version 2.12.2) using *chisq.test*. The *chisq.test* was applied to the contingency tables for each of these different variables. Two-sample Wilcoxon Test was performed

in R using *wilcox.test* to assess whether two samples of independent observations were from distinct populations.

Results

Exome sequencing provides genotypeability of >93% and target sequencing efficiency of >96% of well annotated coding bases

The interpretation of massively parallel sequencing success of human exons and especially protein-coding bases is highly dependent on the choice of gene annotation database used during analysis.[6] To ensure independence in interpretation of sequencing results, we standardized the annotation of coding bases as defined by the three major gene annotation databases—RefSeq,[20] UCSC,[21] and Ensembl.[22] We defined the union of coding bases from the above three databases to represent the 'all coding bases' dataset (ACB: 37,008,680 bases) and the intersection to represent the 'common coding bases' dataset (CCB: 32,271,709 bases) (Fig. 1). After defining these two datasets, we interrogated overlap of the annotation databases, capture kits and target sequencing efficiency of these bases in the UDP patient cohort. Target sequencing efficiency was defined as the percentage of bases observed to be successfully sequenced of all bases expected to be sequenced in target regions designed to be captured by commercial capture kits (Supplementary Table S1). Overall, the CCB constituted 87% of the ACB. The Ensembl and UCSC coding bases exclusively annotated 0.7 Mb and 1.9 Mb of the all coding bases respectively (Supplementary Fig. S1). Coverage of all Ensembl coding bases across all three capture kits—Agilent 38 Mb, Agilent 50 Mb, Illumina TruSeq kits, was strikingly similar (78%), whereas coverage of UCSC known gene and RefSeq coding bases was 93%–95% for the Agilent 50 Mb and the Illumina TruSeq capture kits and 81% for the Agilent 38 Mb capture kit, consistent with previous reports.[6] The Agilent 38 Mb, Agilent 50 Mb and TruSeq capture kits, respectively, targeted 85%, 97% and 95% of the CCB and 75%, 94% and 88% of the ACB (See

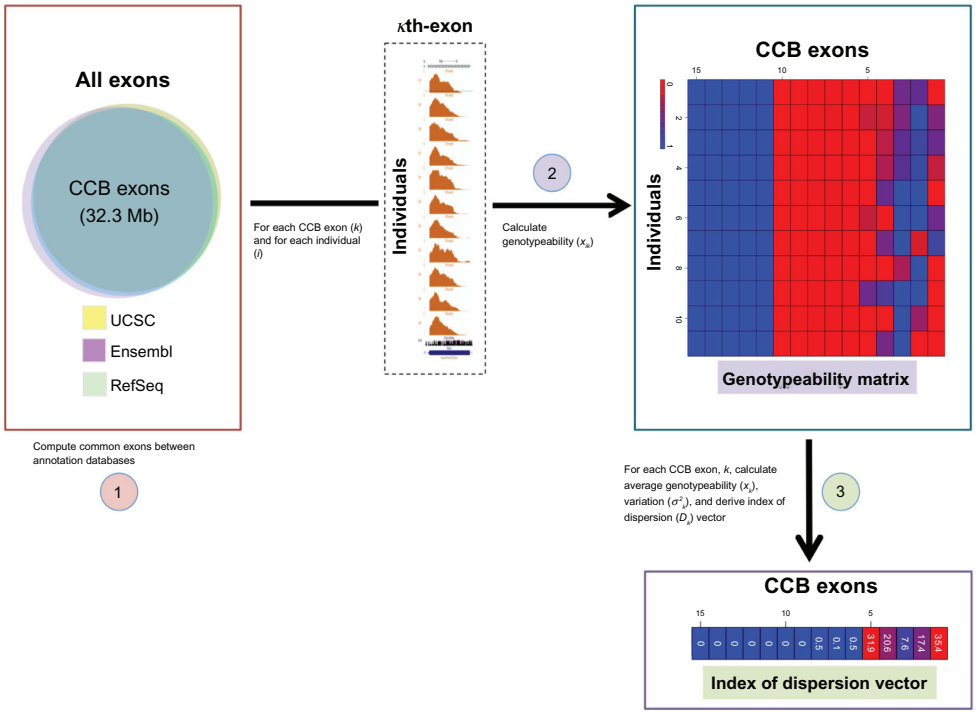

Figure 1. A schematic diagram demonstrating the process of computing the index of dispersion for each exon in our dataset across individuals with exome sequencing data.
Notes: The process of exome sequence data involves three main steps: *Step* (1) Determination of exonic bases to consider for the analysis, *Step* (2) Calculation of genotypeability of exons, and finally, *Step* (3) Calculation of index of dispersion vector. The color *blue* in *genotypeability matrix* indicates complete genotypeability, and *red* indicates lack of genotypeability. The color *blue* in *index of dispersion* vector depicts exons with values less than ten, and *red* and corresponding shades greater than 10.

Supplementary Fig. S2). The Illumina and Agilent platforms target specific mutually exclusive regions; the majority of the 25.0 Mb of unique regions targeted by the TruSeq kit were the untranslated portions of the exons (UTRs). Focusing on the CCB, we observed that the TruSeq kit had better target-sequencing efficiency (96.5%) compared to the Agilent 50 Mbkit (91%) (Supplementary Fig. S3), even though the latter targeted 31,344,815 bases (97.1% of CCB) compared to TruSeq's 30,768,285 bases (95.3% of CCB). Genotypes at each position sequenced were called by Most Probably Genotype (MPG) calling algorithm, which reports a score that is computed as the difference in Bayesian posterior probabilities between the most probable and the second most probable genotypes.[19] We required bases to have qualities of Q20 or higher and a genotype call to have a MPG score of 10 or greater. A MPG score of 10 or greater signifies that the theoretical probability of the call being incorrect is e^{-10} or 1 in 22,026 genotypes called. On average, TruSeq provided over 93% genotypeability (MPG $\geq$ 10) across the common coding bases (Table 1). Given these results, we restricted further analyses to the sequence data derived from libraries generated using the TruSeq capture kit.

The majority of CCB Exons (>88%) are confidently genotyped to 100% in exome sequencing data

To determine the proportion of CCB exons that are confidently genotyped over all the bases in their respective exons, we analyzed the proportion of bases in a given exon that scored above the given genotype score threshold (MPG $\geq$ 10). For this, 188,881 unique CCB exons were evaluated for genotypeability above a confidence threshold of MPG $\geq$10 in 11 individuals (testset) captured by the TruSeq capture kit and sequenced on the Illumina HiSeq 2000 (Fig. 1, steps 1 and 2). Of 188,881 exons, 165,803 (88%) exons were confidently genotyped to completion over all the bases in a given exon. Of the total exons, 7,678 (4%) did not have any bases with confident genotype sequencing (0% genotypeability). Of these 7,678 exons with complete lack of interpretable genotype data, 487 (6%) were from the sex-chromosomes. In autosomes, chromosome 6 (2,767 exons; 36%), followed by chromosome 1 (578 exons; 7%), had the highest number of exonic regions with no interpre-

Table 1. Target coverage and sequencing efficiency of three exome capture kits in 137 individuals sequenced in the NIH undiagnosed diseases program (UDP), relative to common coding bases in the human genome.

Exome capture kit	Sequence data	On-target (CCB)	Off-target (CCB)	Total (CCB)	Theoretical		Observed			
					On-target (CCB%)	Off-target (CCB%)	CCB specific sequencing (%)	CCB non-specific sequencing (%)	CCB sequenced (%)	Target efficiency*
Agilent 38 Mb	Expected	27,381,137	4,890,572	32,271,709	84.8	15.2	–	–	–	–
	Observed ($N = 79$)	24,721,316	997,856	25,719,172	–	–	76.6	3.09	79.7	90.3
Agilent 50 Mb	Expected	31,344,815	926,894	32,271,709	**97.1**	2.9	–	–	–	–
	Observed ($N = 47$)	28,663,767	192,294	28,856,061	–	–	88.8	0.60	**89.4**	**91.4**
TruSeq	Expected	30,768,285	1,503,424	32,271,709	**95.3**	4.7	–	–	–	–
	Observed ($N = 11$)	29,686,420	361,998	30,048,418	–	–	92.0	1.12	**93.1**	**96.5**

Notes: *Target efficiency is defined as the percentage of observed bases sequenced of all bases expected to be sequenced in a given target region. Target regions are defined as regions targeted by commercial exome capture kits.
Abbreviation: CCB, common coding bases.

table genotype data (genotypes over the threshold of MPG $\geq$ 10) over the entire length of a given exon (See Supplementary Table S2). Subsequent analyses through rest of the experiments were restricted to autosomal chromosomes.

Autosomal CCB exons with low average genotypeability are significantly enriched for extremes of GC content and/or low sequence read mappability

To determine if genotypeability of an exon could demostrate a relationship with sequence composition, we analyzed the CCB exons using GC content and short-read mappability as defined by the "Broad alignability track" from the UCSC genome browser.[21] For this analysis, we removed chromosomes annotated as "hap" and "random". This left 179,114 autosomal exons. We also removed very short exons (<10 bases in length). The final set for GC analysis and low-mappabilty analysis contained 178,105 autosomal exons. GC content for each of these exons was calculated as a percent of the length of the exon. The exons were binned into three categories : Low-content (<33%), Medium-GC content (33%–66%) and High-GC content (>66%). We also divided all exons (178,105) into four categories based on their average genotypeablity (number of individuals sequenced, N = 11) : No genotypeability (0%), Low-medium genotypeability (>0 to ≤50%), High-medium genotypeability (>50% to <100%) and Fully genotyped (100%). We then tested if the GC content of an exon was related to its observed average genotypeability. We observed a significant relationship between the GC content and the average exon genotypeability, with the best genotypeability values observed in medium GC content exons (χ^2 = 31744.94, df = 6, $P < 2 \times 10^{-16}$). Both high and low GC content were correlated with low genotypeability (Table 2). We investigated whether exon length had any effect on this relationship. We repeated the same analyses on short (<50 bases), medium (50–300 bases) and long exons (>300 bases) and observed that the strong correlation between the GC content and genotypeability was not affected by exon length (See Supplementary Tables S3–S5).

To investigate further if the lack of genotypeable data was due to mappability of short-reads, we looked at the relationship between short-read mappability of an exon and its observed average genotypeability. It is well established that sequencability of a given exon obtained from mapping short reads to a reference sequence is a direct function of the success in correctly placing the short reads on their original locations on the reference genome. For this purpose, we used the "Broad alignability track" from the UCSC genome browser. The Broad alignability track displays whether a region is made up of mostly unique or mostly non-unique sequence. To generate the track, every 36-mer in the genome was marked as "unique" if the most similar 36-mer elsewhere in the genome had at most 2 mismatches. Position 'x' in the alignable track is marked by 1 if >50% of the bases in [x–200,x+200] are "unique" and by 0 otherwise.[21] We found a significant relationship between the mappability of an exon and the observed average genotypeability of that exon (χ^2 = 13165.31, df = 6, N = 178105, $P < 2 \times 10^{-16}$). (Supplementary Table S6). The exons with high-mappability values were genotyped to near completion with most having full genotypeability (100%). Finally, we analyzed the relationship between the repeat regions of the genome as defined by the repeat masker track of the UCSC genome browser and exon genotypeability (Supplementary Table S7). We found

Table 2. Distribution of GC content of CCB exons (178,105) grouped by their average genotypeability in 11 UDP individuals sequenced using the TruSeq exome capture kit on Illumina HiSeq 2000.

GC content (%) categories	Average genotypeability categories			
	No genotypeability (0%)	Low-medium genotypeability (>0% to ≤50%)	High-medium genotypeability (>50% to <100%)	Fully genotyped (100%)
Low-GC content (<33%)	95	15	294	3,553
Medium-GC content (33%–66%)	3,152	1,249	15,321	140,830
High-GC content (>66%)	1,139	1,885	6,106	4,466

a significant relationship between the repeat masked regions and low genotypeability (χ^2 = 26282.69, df = 3, $P < 2 \times 10^{-16}$).

From the set of exons that either had high GC content or low mappability, we randomly selected 40 exons with high GC (25) and low mappability (15) to test if these exons can be rescued by Sanger sequencing. Of the 40 randomly selected exons, 36 (90%) were successfully sequenced and confirmed by Sanger sequencing under various optimizing conditions (See Supplementary Table S8). Of these 36 confirmed exons, 22 were from extreme GC and 14 from low-mappability regions. Based on these findings, we investigated if exons with a consistent genotypeability signature can differentiate regions that derive from noisy data.

Index of dispersion robustly distinguishes consistently genotyped exons from inconsistently genotyped exons

To determine if given exons can be classified into categories that reflect their overally reproducibility of genotypeability across a cohort of patients, we began with a set of 179,114 autosomal exons sequenced using the same sequencing chemistry (HiSeq 2000) and capture kit (TruSeq) in 11 UDP individuals. Exonic regions from chromosomes annotated as 'hap' and 'random' were excluded from this set. We also excluded short exons (<50 bases in length) from further analysis. The final data set consisted of 167,717 distinct exons. The index of dispersion (D) for each of these regions was calculated as described in the Methods section (Fig. 1). Based on the characteristics of D (see Methods), the exons were classified into three main categories : (a) Constant/underdispersed exons; $D < 1$, (b) Random / low-overdispersed exons; $1 \leq D \leq 10$, and (c) High-overdispersed exons; $D > 10$. These three bins had 160,115 (95.5%), 4,447 (2.7%) and 3,155 (4.1%) exons respectively (Table 3). The category of constant/underdispersed exons consists of a subset of all exons whose genotypeability results are highly similar between all individuals sequenced, while on the other extreme, highly overdispersed exons category describes the subset of regions for which the genotypeability sequencing results differ widely. To assess degree of completeness of a given exon's genotypeability, all exons were binned based on their average genotypeability value (total number of individuals analyzed : $N = 11$) into 5 major groups: No genotypeability (0%), Low genotypeability exons (0 to ≤30%), Medium genotypeability exons (>30 to ≤70%), High genotypeability exons (>71 to ≤100%), and Complete genotypeability exons (100%). Given the categories above, we investigated the distribution of exons as a function of their overall genotypeability. The majority of exons (140,325; 83.7%) were genotyped to completion over their full length (Table 3).

Next, we tested if underdispersion of genotypeability of a given exon ($0 < D < 1$) can be used as a metric to reliably select exons with consistent genotypeability coverage patterns and distinguish

Table 3. Distribution and relationship between genotypeability of exons and their corresponding index of dispersion.

Average genotypeability categories	Index of dispersion (D) categories			
	Constant/ underdispersed exons ($0 \geq D > 1$)	Random/low-overdispersed exons ($1 \geq D \geq 10$)	High overdispersed exons ($D > 10$)	Total (%)
No genotypeability exons (0%)	3,690	–	–	3,690 (2.2%)
Low genotypeability exons (0 to ≤30%)	68	482	1,253	1,803 (1.1%)
Medium genotypeability exons (>30 to ≤70%)	244	758	1,440	2,442 (1.5%)
High genotypeability exons (>71 to ≤100%)	15,788	3,207	462	19,457 (11.6%)
Complete genotypeability exons (100%)	140,325	–	–	140,325 (83.7%)
Total	160,115 (95.5%)	4,447 (2.7%)	3,155 (1.9%)	167,717

Note: The values of D were calculated from a cohort of 11 individuals sequenced with the TruSeq exome sequencing kit.

them from the rest of the exons with inconsistent patterns ($D > 1$). In order to test this hypothesis, 5 exons were randomly and independently selected from each of the 15 classes of exons described in Table 3. Of the 15 possible classes of grouped exons, only 11 sub-groups had data (see Table 3). This yielded a total of 55 randomly selected exons (Fig. 2A). For statistical robustness of the two-sample Wilcox test (in a 3 × 3 format), no genotypeability and complete genotypeability exons were not included to fit the test format. We tested the 45 exons (9 data points × 5 exons) to evaluate if the index of dispersion could differentiate between constant/underdispersed exons and overdispersed exons. We found a significant difference between constant/underdispersed and

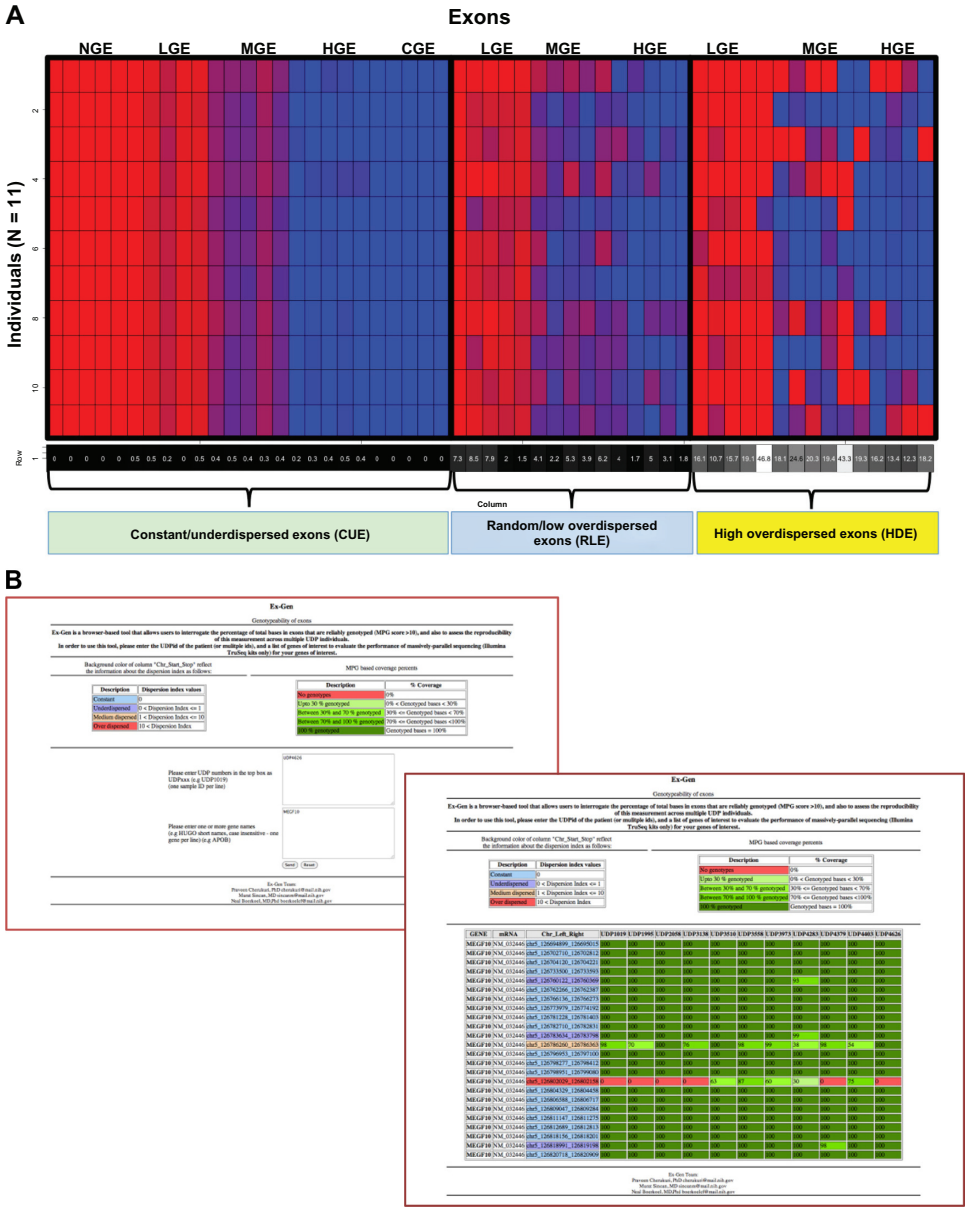

Figure 2. Classification of exons into three distinct groups based on their index of dispersion of genotypeability values. (**A**) The three main categories based on the index of dispersion of genotypeability are:Constant/Underdispersed Exons (CUE; *light green*), Random/Low overdispersed Exons (RLE; *light blue*) and High overdispersed Exons (HDE; *yellow*). In total, 55 exons were randomly selected (5 from each genotypeability bin) that were sequenced in 11 individuals. Each cell in the genotypeability matrix is color coded in a range of shades from *red* (no genotypeability) to *blue* (complete genotype). The index of dispersion of genotypeability vector is shown below the matrix, and is shaded from *black* (low dispersion) to *white* (high dispersion). (**B**) Ex-Gen: An internal web-tool for dynamically analyzing gene-specific genotypeability of whole-exome sequence data in UDP patients. This tools allows users to evaluate and interpret whole-exome sequencing results in an individual, gene, and exon-specific manner.

overdispersed exons (two-sample Wilcoxon test; $P = 3 \times 10^{-5}$). Based on these results, we labeled a total of 156,425 exons as being consistently and repeatably genotyped in multiple samples (3,690 exons had no genotypeability across the entire length of the exon). Addtionally, we cataloged a total 6,845 poorly performing exons that either had no genotypeability (3,690) or exons that had a very high index of dispersion (3,155) See Table 3. Finally, for the purpose of internal querying of performance of exome sequence data, we developed a web-portal tool called ExGen for dynamic display and interrogation. Screen-shots of web-portal are shown in Figure 2B.

Application of index of dispersion to detect exon-scale homozygous deletion events in affected patients

Since the index of dispersion efficiently distinguished consistently genotyped exons from noisy regions in the exome data, we hypothesized that differences in the dispersion signatures of founder data and proband data would rapidly enable us to detect homozygous deleted regions in probands. To test this, we built two datasets of 15 founder individuals (controls) and 19 proband individuals (affected cases). A total of 167,717 CCB exons were processsed and the average genotypeability and indices of dispersion were derived for each dataset. The indices of dispersion of founders and probands were compared (Fig. 3A). Control exons from the founder dataset consisted of exons from constant/underdispersed category ($0 \leq D \leq 1$) that had greater than 70% average genotypeability. The control dataset consisted of 153,654 distinct exons. The case exons from the proband dataset consisted of 7,263 random/low-overdispersed exons ($1 \leq D \leq 10$). Exons in proband dataset with potential exon size homozygous deletion events ($D \geq 4$) were tested if in founders (controls) these exons were grouped as constant/underdispersed exons

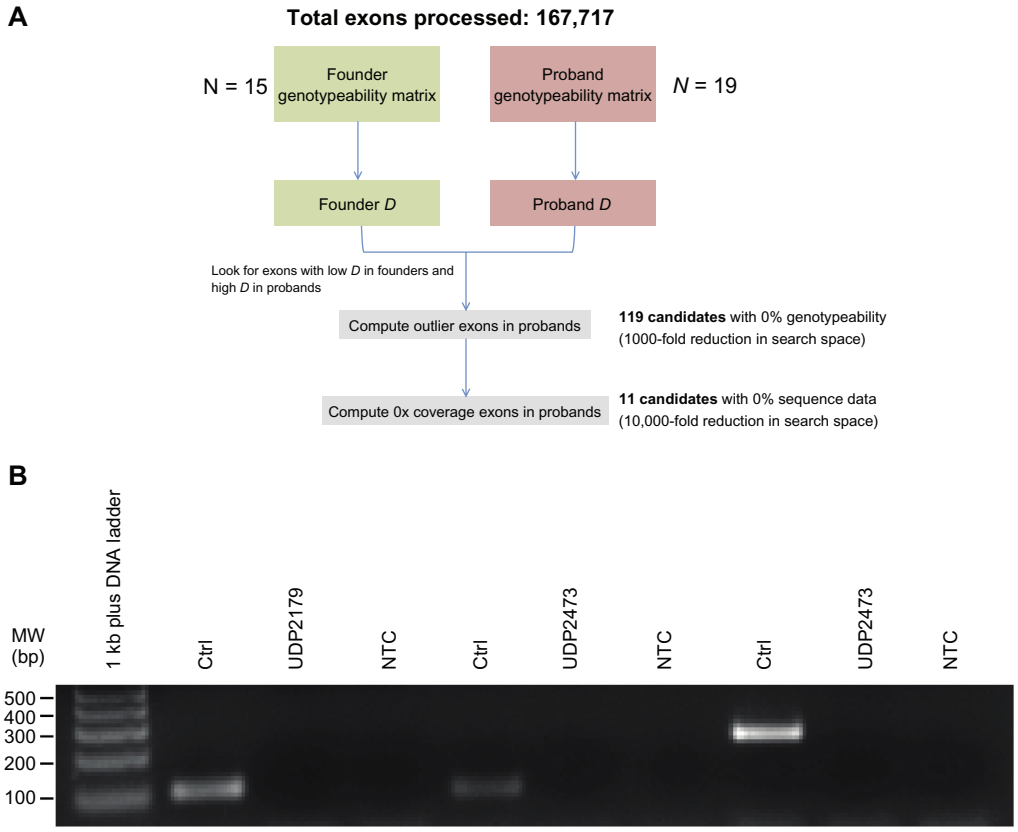

Figure 3. Prediction and validation of homozygous deletions in patient whole-exome sequence data. (**A**) Schematic diagram showing the procedure to rapidly detect homozygously deleted exon events by deriving dispersion index vectors from massively-parallel exome sequence data of founders and probands. (**B**) PCR validation of deletion event in three genes—*UGT2B17*, *KRT77* and *MEGF10* in two affected individuals—UDP2179 and UDP2473. *UGT2B17* (lanes 2–4), *KRT77* (lanes 5–7) and *MEGF10* (lanes 8–10) were tested with unaffected (Ctrl), affected (UDP2179 and UDP2473) and negative control (NTC).

($D \leq 0.6$). This rapidly found 702 candidate proband exons for potential homozygous deletion events. These 702 candidate exons with high dispersion index values were further investigated for complete lack of sequence genotypeability for each individual sequenced, as high index of dispersion for a given region in a given test group indicates an outlier in genotypeability due to one or a small subset of individuals. For each of these 702 candidate exons, we checked for complete lack of genotypeability across the entire length of the exon. This step narrowed the total number of candidates from 702 to 119 independent exonic regions from in a set of 14 proband individuals. This result exactly mapped exonic regions with lack of genotype data to their respective individual with potential homozygous deletion event. We finally tested these 119 candidate homozygous deletion exons for complete lack of sequence data (sequence coverage = 0X) in 14 individuals to eliminate 108 exons with low-sequence coverage in these probands, but not enough for confident genotypes to be called over the full length of these exons. After this final step, there were 11 exons in 3 proband individuals (UDP2179, UDP2473, and UDP865), in 6 distinct genes – *UGT2B17, PRB1, KRT77, MEGF10, CHRFAM7A,* and *UGT2B28* (Table 4) with complete lack of sequence data (0X) suggesting a potential true homozygous deletion events. We were able to successfully test and validate three (of 11) exons in three genes (*UGT2B17, KRT77* and *MEGF10*) by quantitative polymerase chain-reaction (qPCR) (Supplementary Fig. S4) based on based on the following criteria: (i) the sequence similarity of test region was less than 95% compared to the rest of the human genome; (ii) the overall GC-content of the PCR product was <60% and did not contain runs of poly-G or poly-C. The homozygous deletion events were confirmed and validated for all three predicted homozygous deletion events (Fig. 3B and Supplementary Table S9).

Discussion

In this study we evaluated a systematic and thorough computational approach to define and characterize the genotypeability of protein-coding genes sequenced using massively-parallel sequencing technology. First, we demonstrated the value of a systematic approach to evaluate protein-coding bases that is agnostic to any one particular gene annotation database, the advantage of not having to report database-specific results,[6] and the performance of the latest exome-capture technologies compared to previous versions. Using this consensus based approach, we systematically demonstrated that over 88% of protein-coding exons in the human genome are confidently genotyped to completion in all individuals sequenced using latest exome capture sequencing technology. As the research community progresses towards implementation of genome and exome sequencing in clinical diagnostic and translational settings,[23,24] our findings highlight and pin-point protein coding regions in the human genome that are highly complete and reproducible, and amenable to statistically reliable interpretation.

Second, we thoroughly evaluated and examined the regions that were susceptible to deficiencies of

Table 4. List of exons in genes in UDP individuals predicted to be homozygously deleted based on exome sequencing data

UDP proband	Predicted homozygous deleted region	Gene	Length	Design success for confirmation
UDP2179	chr4:69085938-69086217	*UGT2B17*	280	–
UDP2179	chr4:69113885-69114033	*UGT2B17*	149	–
UDP2179	chr4:69116074-69116797	*UGT2B17*	724	Yes
UDP2473	chr12:11397851-11398203	*PRB1*	353	–
UDP2473	chr12:51372806-51372931	*KRT77*	126	Yes
UDP2473	chr5:126760123-126760369	*MEGF10*	247	Yes
UDP865	chr15:28446913-28447022	*CHRFAM7A*	110	–
UDP865	chr15:28452473-28452640	*CHRFAM7A*	168	–
UDP865	chr4:70182821-70182969	*UGT2B28*	149	–
UDP865	chr4:70187059-70187190	*UGT2B28*	132	–
UDP865	chr4:70194837-70195116	*UGT2B28*	280	–

massively-parallel sequence data, and systematically evaluated base composition to captured sequencing correlation. As the exomic regions significantly enriched for sequence features such GC content, low-sequence mappability were positively correlated with poor genotypeability, it is now possible to confidently identify these regions for further evaluation for improvement of exome capture kits. This data also provides *a priori* knowledge for using caution during interpretation of exome sequencing results which requires, for example, measuring frequency of a given 'potential disease' variant in general population. Further, this approach allowed us to develop a novel statistical method to classify regions that are consistently and reproducibly sequenced in our entire patient cohort. Our results confirm that dispersion index of average genotypeability is an robust metric to confidently classify regions of human genome based on sequencing performance.

Finally, we showed application of our statistical metric to rapidly achieve a dramatic reduction in search-space to confidently detect homozygous deletion events. The search computational complexity for finding homozygous deletions by a brute-force method would be on the order of *O(NxM)*, where '*N*' is defined as the number of individuals sequenced, and '*M*' is the total number of exons in the human genome (~180,000 exons). Our computational algorithm has a significant impact on the total search space considered for evaluation, which is on the order of *O(Nxm)*, where '*m*' is the subset of exons with high index of dispersion in test individuals, low index of dispersion in controls, no genotypeable and no sequence data (~11 exons in our study). Our findings confirm that true homozygous deletion events can be rapidly discovered (~10,000-fold to 15,000-fold faster) in panels of cases with appropriate control whole-exome sequence data using this algorithm. We successfully validated and confirmed candiate large-scale exon size deletions by PCR and qPCR on three genes (*UGT2B17, KRT77* and *MEGF10*) identified as homozygously deleted by our computational approach. The implications of the final step not only involves a rapid computational methodology to detect and annotate homozygous deletions, but also has the potential intrinsic value of providing information on potential disease pathogenesis using the patient data.

One limitation of our method is that care must be taken to ensure that the data under consideration are generated under similar conditions, namely, using the same sequence capture kit and sequencing platform. Dependence of final data on platform has been described before,[6] but we reiterate this condition. Future studies can be designed to evaluate similarities and correlations between different sequencing platforms and capture kits, which would enable researchers to analyze data for large-scale structural changes in a platform-independent manner.

In conclusion, we have described a systematic methodology to evaluate exome sequencing data for consistent genotypeability across protein coding exons in our patient population. In addition, we have developed a novel methodology that utilizes the advantages of the above principle to rapidly detect large homozygous deletions. Finally, we demonstrate that our methodology allows researchers and clinicians to interpret confidently the deletion events detected by exome sequencing based on the consistency of sequencing.

Author Contributions

PFC, CFB, and DA conceived and designed the experiments. PFC, CFB, MS and JPA analysed the data. PFC wrote the first draft of the manuscript. PFC, MS, CFB, DA and WAG contributed to the writing of the manuscript. PFC, MS, JPA, KFF, TCM, CFB, CJT, WAG, and DA agree with manuscript results and conclusions. PFC and CFB jointly developed the structure and arguments for the paper. WAG, CFB, and PFC made critical revisions and approved final version. All authors reviewed and approved of the final manuscript.

Funding

Author(s) disclose no funding sources.

Competing Interests

Author(s) disclose no potential conflicts of interest.

Acknowledgements

We thank Camilo Toro, MD., PhD, Peter Munson, PhD, and Thierry Vilboux, PhD for thoughtful input and discussion. We would also like to thank Mr. Philip Moors, Mr. Alvin Yun and Mr. Will Seabroook for support in computational infrastructure setup and implementation.

Disclosures and Ethics

As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References

1. Lieber DS, Vafai SB, Horton LC, et al. Atypical case of Wolfram syndrome revealed through targeted exome sequencing in a patient with suspected mitochondrial disease. *BMC medical genetics*. 2012;13:3.
2. Artuso R, Fallerini C, Dosa L, et al. Advances in Alport syndrome diagnosis using next-generation sequencing. European journal of human genetics: *EJHG*. 2012;20:50–7.
3. Ku CS, Naidoo N, Pawitan Y: Revisiting Mendelian disorders through exome sequencing. *Human Genetics*. 2011;129:351–70.
4. Ng SB, Buckingham KJ, Lee C, et al. Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics*. 2010;42:30–5.
5. Wang JL, Yang X, Xia K, et al. TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. *Brain: A Journal of Neurology*. 2010;133:3510–8.
6. Clark MJ, Chen R, Lam HY, et al. Performance comparison of exome DNA sequencing technologies. *Nature Biotechnology*. 2011;29:908–14.
7. Gilissen C, Hoischen A, Brunner HG, Veltman JA. Disease gene identification strategies for exome sequencing. *European Journal of Human Genetics: EJHG*. 2012;20:490–7.
8. Choi M, Scholl UI, Ji W, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America.* 2009;106:19096–101.
9. Worthey EA, Mayer AN, Syverson GD, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*. 2011;13:255–62.
10. Bonnefond A, Durand E, Sand O, et al. Molecular diagnosis of neonatal diabetes mellitus using next-generation sequencing of the whole exome. *PLoS ONE*. 2010;5:e13630.
11. Montenegro G, Powell E, Huang J, et al. Exome sequencing allows for rapid gene identification in a Charcot-Marie-Tooth family. *Annals of Neurology*. 2011;69:464–70.
12. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*. 2008;36:e105.
13. Quail MA, Kozarewa I, Smith F, et al. A large genome center's improvements to the Illumina sequencing system. *Nature Methods*. 2008;5:1005–10.
14. Bamshad MJ, Ng SB, Bigham AW, et al. Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*. 2011;12:745–55.
15. Gahl WA, Boerkoel CF, Boehm M: The NIH Undiagnosed Diseases Program: bonding scientists and clinicians. *Disease Models and Mechanisms*. 2012; 5:3–5.
16. Biesecker LG, Mullikin JC, Facio FM, et al. The ClinSeq Project: Piloting large-scale genome sequencing for research in genomic medicine. *Genome Research*. 2009;19:1665–74.
17. Dias C, Sincan M, Cherukuri PF, et al. An analysis of exome sequencing for diagnostic testing of the genes associated with muscle disease and spastic paraplegia. *Human Mutation*. 2012;33:614–26.
18. Wei X, Walia V, Lin JC, et al. Exome sequencing identifies GRIN2A as frequently mutated in melanoma. *Nat Genet*. 2011;43:442–6.
19. Teer JK, Bonnycastle LL, Chines PS, et al. Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. *Genome Research*. 2010;20:1420–31.
20. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Research*. 2012;40:D130–5.
21. Dreszer TR, Karolchik D, Zweig AS, et al. The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Research*. 2012;40: D918–23.
22. Flicek P, Amode MR, Barrell D, et al. Ensembl 2012. *Nucleic Acids Research*. 2012;40:D84–90.
23. Mestan KK, Ilkhanoff L, Mouli S, Lin S. Genomic sequencing in clinical trials. *Journal of Translational Medicine*. 2011;9:222.
24. Nelen M, Veltman JA. Genome and exome sequencing in the clinic: unbiased genomic approaches with a high diagnostic yield. *Pharmacogenomics*. 2012; 13:511–4.

Supplementary Data

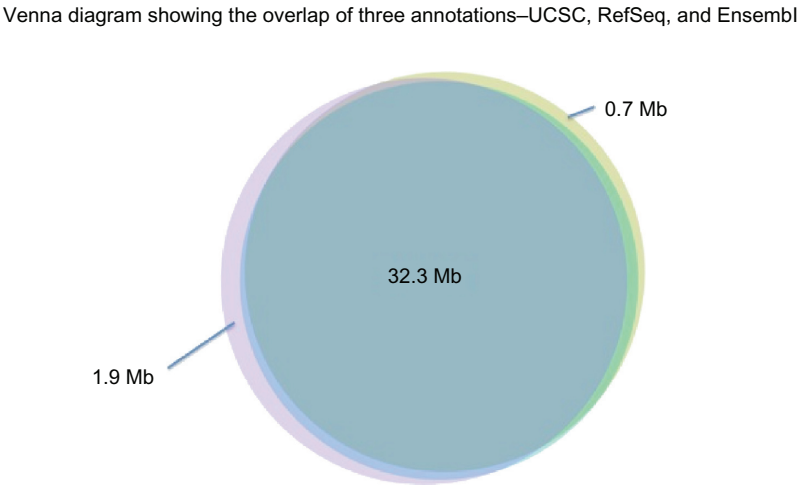

Figure S1. Area-proportional venn diagram showing overlap of coding bases as annotated by three major gene annotation databases—RefSeq (blue) Ensembl (purple) and UCSC (yellow).

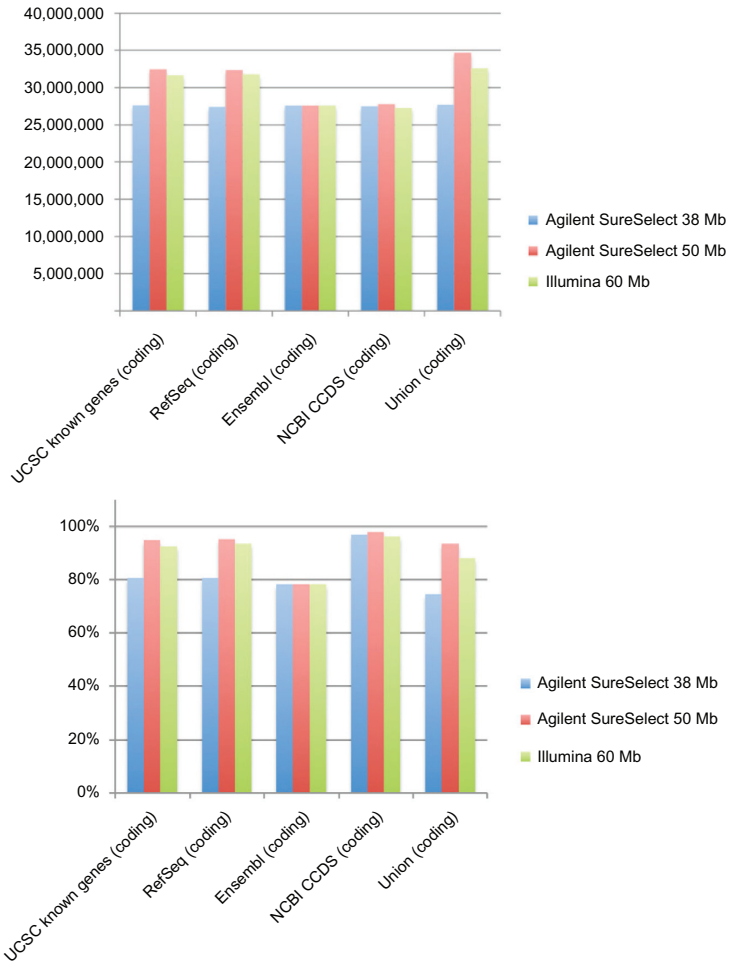

Figure S2. Comparision of three different exome capture kits (Agilent 38 Mb, Agilent 50 Mb, and Illumina TruSeq) and their relationship to four protein coding gene annotations (UCSC known genes, RefSeq, Ensembl, and NCBI CCDS).
Notes: Top panel shows the raw number of sequence bases targeted by each of the three kits and the bottom panel displays the percent of the annotations targeted by each of the three kits. The union of protein coding bases from the above three different annotation databases was generated to define and evaluate the "entire" coding space. We refer to this union bases as "all coding bases (ACB)" in the main manuscript.

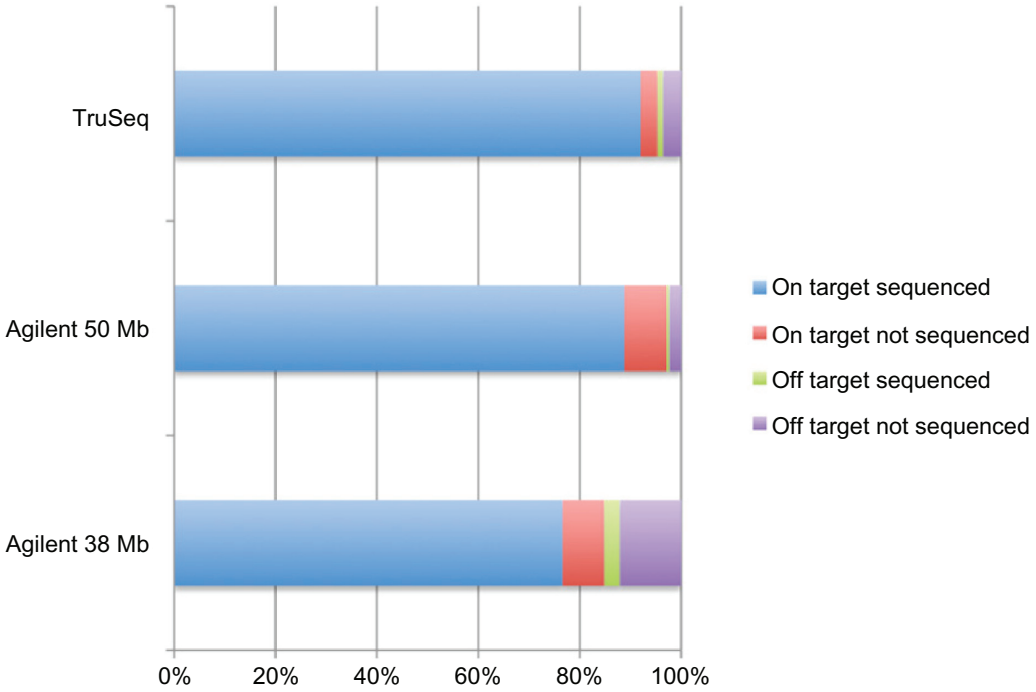

Figure S3. Targeting and sequencing efficiency of three different exome capture kits (Agilent SureSelect 38 Mb, Agilent SureSelect 50 Mb, and Illumina TruSeq) on the common coding bases (CCB, 32 Mb), which is an intersection of three independent human protein coding gene annotation databases.

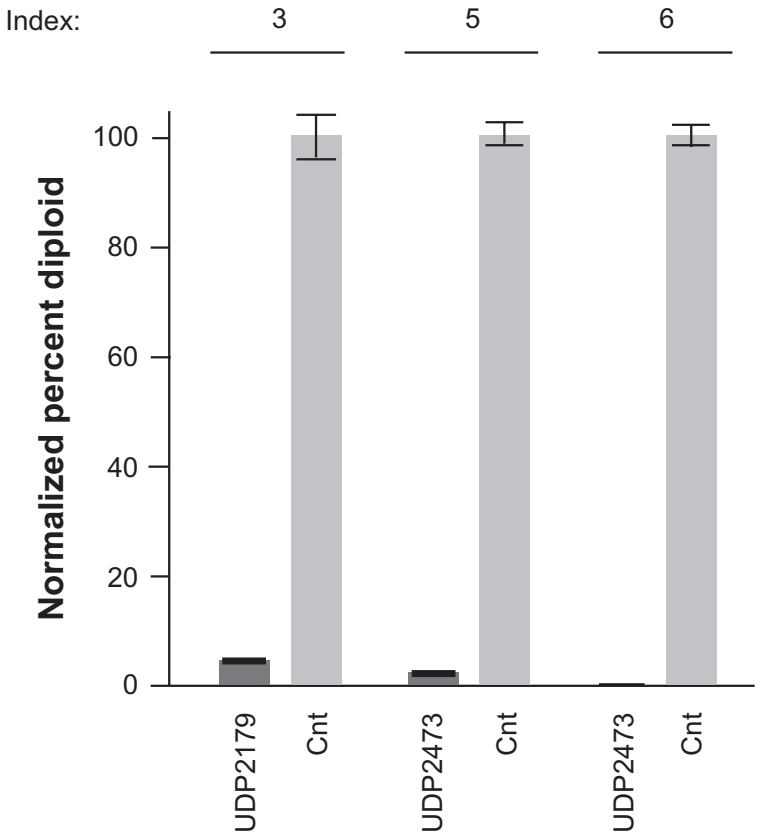

Figure S4. qPCR validation of homozygous deletion events detected by the automated alogrithm.
Notes: This analysis further confirmed the three homozygous deletions (*UGT2B17*—index 3, *KRT77*—index 5 and *MEGF10*—index 6) in the probands (UDP2173 and UDP2473). The quantification was normalized to percent diploid, and we have confirmed that the insignificant trace signal detected in index 3 and 5 were due to a trace quantity of cross-contaminant during qPCR handling.

Table S1. Bait design details for each of the three commercial platforms—Agilent SureSelect 38 Mb, Agilent SureSelect 50 Mb, and Illumina TruSeq.

Target kit intersections	Bases covered	Percent of union
SureSelect 38 only	16805	0.02
SureSelect 50 only	8,325,804	10.40
TruSeq only	28,563,968	35.67
SureSelect 38 and SureSelect 50	98,77,412	12.33
SureSelect 38 and TruSeq	1,377	0.00
SureSelect 50 and TruSeq	5,551,621	6.93
SureSelect 38 and SureSelect 50 and TruSeq	27,744,802	34.65
SureSelect 38 or SureSelect 50 or TruSeq	80,081,789	100.00

Table S2. Number and percentage of common coding base regions with no genotyping coverage in each chromosome in samples sequenced using TruSeq exome capture kit to show robustness of our methodolgy and its independence on number of individuals sequenced ($N = 11$, $N = 34$, $N = 184$).

Chromosome	Number of regions detected using different sample sizes			Average percentage
	$N = 11$	$N = 34$	$N = 184$	
chr1	578	611	588	7.7
chr10	325	325	312	4.2
chr11	190	211	207	2.6
chr12	153	164	155	2.0
chr13	54	60	61	0.8
chr14	80	85	86	1.1
chr15	212	219	215	2.8
chr16	276	294	289	3.7
chr17	399	333	319	4.6
chr18	37	41	43	0.5
chr19	253	275	266	3.4
chr2	395	386	380	5.0
chr20	98	114	112	1.4
chr21	49	56	54	0.7
chr22	122	126	127	1.6
chr3	195	199	191	2.5
chr4	130	142	142	1.8
chr5	220	222	221	2.9
chr6	2,767	2,772	2,762	36.0
chr7	236	251	245	3.2
chr8	205	200	195	2.6
chr9	217	239	221	2.9
chrX	322	298	251	3.8
chrY	165	160	160	2.1
Total	7,678	7,783	7,602	100

Table S3. GC-content of short exons (<50 nucleotides) and their average genotypeability values in UDP samples ($N = 11$).

GC content (%) of the exon	Average genotypeability of exons ($N = 10,388$) in 11 samples			
	None (0)	Low-medium (>0–≤50)	High-medium (>50–<100)	Full (100)
<33	28	3	37	560
33–66	518	193	668	7433
>66	150	80	187	531

Table S4. GC-content of medium exons (50–300 nucleotides) and their average genotypeability values in UDP samples ($N = 11$).

GC content (%) of the exon	Average genotypeability of exons ($N = 154,879$) in 11 samples			
	None (0)	Low-medium (>0–≤50)	High-medium (>50–<100)	Full (100)
<33	67	12	246	2955
33–66	2493	944	12220	125838
>66	865	1117	4470	3652

Table S5. GC-content of long exons (>300 nucleotides) and their average genotypeability values in UDP samples ($N = 11$).

GC content (%) of the exon	Average genotypeability of exons ($N = 12,322$) in 11 samples			
	None (0)	Low-medium (>0–≤50)	High-medium (>50–<100)	Full (100)
<33	0	0	8	16
33–66	130	107	2387	7165
>66	120	685	1438	266

Table S6. Mappability (Broad alignability track) and average exon genotypeability values in UDP samples ($N = 11$).

Mappability	Average coverage of exons ($N = 178,105$) in 11 samples			
	None (0)	Low-medium (>0–≤50)	High-medium (>50–<100)	Full (100)
<0.33	1617	496	1690	4210
0.33–0.66	31	17	123	625
>0.66	2738	2636	19908	144014

Table S7. Repeat masker annotations and average exon genotypeability values in UDP samples ($N = 11$).

Repeat masked region overlap of the exon	Average genotypeability of exons ($N = 178,105$) in 11 samples			
	None (0)	Low-medium (>0–≤50)	High-medium (>50–<100)	Full (100)
≥10%	1292	610	644	651
<10% (or none)	3094	2539	21077	148198

Table S8. Sanger sequencing confirmation of 40 randomly selected exons with either high GC content or low mappability.

Region	Sequence success	Comments
GC1	+	
GC2	+	
GC3	+	
GC4	+	
GC5	+	
GC6	+	
GC7	+	
GC8	+	
GC9	+	
GC10	+	
GC11	+	
GC12	+	
GC13	+	
GC14	+	
GC15	+	
GC16	+	
GC17	+	
GC18	+	
GC19	+	
GC20	+	
GC21	+	
GC22	+	Missing 3′ end
GC23	+	
GC24	–	No coverage
GC25	+	Missing 3′ end
LM1	+	
LM2	+	
LM3	+	
LM4	+	
LM5	+	
LM6	+	
LM7	+	
LM8	+	
LM9	+	
LM10	+	
LM11	+	
LM12	+	
LM13	+	
LM14	+	
LM15	–	No coverage

Abbreviations: GC, GC content; LM, Low mappability.

Table S9. Optimization conditions experimented with to rescue complete and valid Sanger sequence data in 4 regions of sequencing failure.

Region	Sequence status	Primer set	Conditions	PCR product
GC22	Missing 3′ end	#1 (5′ end)	H_2O/dmso/qbuffer	no/yes/no
		#2 (3′ end)	H_2O/betaine/dmso/qbuffer	no/no/no/no
		#3 (3′ end)	H_2O/dmso/qbuffer/betaine	22a—no/yes/no/no
				22b—no/no/no/no
GC24	No coverage	#2	H_2O/betainedmso/qbuffer	yes/no/no/no
		#3	H_2O/dmso/qbuffer/betaine	no/yes/no/no
GC25	Missing 3′ end	#1 (5′ end)	H_2O/dmso/qbuffer	no/yes/no
		#3 (3′ end)	H_2O/dmso/qbuffer/betaine	25a—no/yes/no/no
				25b—no/no/no/no
				25c—no/yes/no/no
LM15	No coverage	#1	H_2O/dmso/qbuffer/betaine	yes/no/no/yes
		#3	H_2O/dmso/qbuffer/betaine	yes/no/yes/yes