Published in final edited form as:
Hum Mutat. 2012 April ; 33(4): 609–613. doi:10.1002/humu.22033.

Detecting false positive signals in exome sequencing

Karin V Fuentes Fajardo[1], **David Adams**[1,2], **NISC Comparative Sequencing Program**[3], **Christopher E Mason**[4,5], **Murat Sincan**[2], **Cynthia Tifft**[1,2], **Camilo Toro**[1], **Cornelius F Boerkoel**[1], **William Gahl**[1,2,6], and **Thomas Markello**[6]

[1]NIH Undiagnosed Diseases Program, NIH Office of Rare Diseases Research and NHGRI, Bethesda, MD, USA

[2]Medical Genetics Branch, NHGRI, NIH, Bethesda, MD, USA

[3]NIH Intramural Sequencing Center, NIH, Bethesda MD, USA

[5]Department of Physiology and Biophysics, Weill Medical College, Cornell University, New York, NY 10065, USA.

[5]HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Medical College, Cornell University, New York, NY 10065, USA.

[6]Office of the Clinical Director, NHGRI, NIH, Bethesda, MD, USA

Abstract

Disease gene discovery has been transformed by affordable sequencing of exomes and genomes. Identification of disease-causing mutations requires sifting through a large number of sequence variants. A subset of the variants are unlikely to be good candidates for disease causation based on one or more of the following criteria: (1) being located in genomic regions known to be highly polymorphic, (2) having characteristics suggesting assembly misalignment, and/or (3) being labeled as variants based on misleading reference genome information. We analyzed exome sequence data from 118 individuals in 29 families seen in the NIH Undiagnosed Diseases Program (UDP) to create lists of variants and genes with these characteristics. Specifically, we identified several groups of genes that are candidates for provisional exclusion during exome analysis; 23,389 positions with excess heterozygosity suggestive of alignment errors; and 1,009 positions in which the hg18 human genome reference sequence appeared to contain a minor allele. Exclusion of such variants, which we provide in supplemental lists, will likely enhance identification of disease-causing mutations using exome sequence data.

Keywords

Exome sequencing; Inherited disease; False positives; Next Generation Sequencing; Genomics; Illumina; Sequencing errors; Alignment errors; WES; Sureselect Human All exon

INTRODUCTION

Identification of disease-causing genes among the variants generated by exome sequencing (ES) requires the separation of candidates with high pathogenic potential from variants that have a low-probability for disease causation. Numerous well-described mechanisms can

Thomas C Markello, Medical Genetics Branch, NIH/NHGRI, 10 Center Drive, Building 10/10C103, Bethesda MD 20892, USA., markellot@mail.nih.gov.

Supporting Information for this preprint is available from the *Human Mutation* editorial office upon request (humu@wiley.com)

generate low-interest variants. Biological sources of low-interest variants include both common and rare population variation. Certain regions of the genome are unusually variable and the study of exome sequencing data from even a few individuals reveals genes that vary from the reference sequence in most, if not all, sequenced individuals.

High throughput sequencing techniques also generates low-interest variants in the form of genotype false-positives. Errors can arise from biases in the library construction (Aird, et al., 2011; Bentley, et al., 2008; Koboldt, et al., 2010; Teer, et al., 2010), errant polymerase reactions (Aird, et al., 2011), difficulty making genotype calls at the end of short reads, loss of synchrony among DNA sequencing reactions within a cluster (Ledergerber and Dessimoz, 2011) or manufacturer/platform-specific mechanistic problems such as overlap in absorption spectra for guanine and thymine in the Illumina system (Dohm, et al., 2008; Meacham, et al., 2011). Misalignment of sequencing reads to a reference sequence (RefSeq) and inaccuracies or biases of the RefSeq compared to a specific local population are other sources of false positive genotype calls in Next Generation Sequencing (NGS) data (Church, et al., 2011). Misalignments of short-length sequencing reads to a reference sequence are influenced by the choice of seed-based strategies or algorithms for complete alignment permutations (Homer and Nelson, 2010; Li and Durbin, 2009; Schatz, et al., 2010). These problems often arise in regions with low complexity (Landan and Graur, 2007) or result from misalignment of multiple copies of genes, paralogues or pseudogenes (Blankenberg, et al., 2010).

The reference sequence itself may be an additional source of variants. For some base positions, the reference sequence specifies a minor allele in most large human populations. Such biases occur because of the limited number of individuals on which the original reference sequence was based, plus sequencing and alignment errors (Lander, et al., 2001). As a result, the NCBI human genome reference sequence includes minor variants, unique variants and, possibly, disease-causing mutations (Snyder, et al., 2010) (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/).

Some of these variants can be identified and provisionally excluded during a search for disease-causing variants. Herein we provide example exclusion lists based upon our accumulated hg18 ES data. In addition, it is important for researchers to generate similar exclusion lists from their own datasets to take into account errors, that may be specific to the sequencing and analysis methodology or the human genome reference version used.

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 of the National Human Genome Research Institute and gave written, informed consent. The patients were members of 29 different families and had unique and widely divergent phenotypes. These 29 families contained 55 founders and additional affected and unaffected siblings of the probands, summing to a total of 118 individuals. Of the 29 families, 5 have been diagnosed (see accompanying manuscripts; Gahl, et al., 2011; Gahl and Tifft, 2011; Pierson, et al., 2011) and several others have strong candidate gene leads identified by ES.

An additional anonymized dataset of 401 exome sequences derived from the ClinSeq™ study (Biesecker, et al., 2009) was used as a crosscheck on the population characteristics of specific variants discovered in the UDP data set.

DNA extraction

DNA was extracted from 10 mL of peripheral whole blood using the Puregene kit (Qiagen, Inc., Valencia, CA) according to the manufacturer's protocol.

Exome sequencing, sequence alignment and variant annotation

Initially the Agilent (Santa Clara, CA) Human 38Mb all exome capture method was used for enrichment and, as the design improved to capture additional exons and previously unannotated genes, the 50Mb capture method was substituted. (Coffey, et al., 2011; Gnirke, et al., 2009). 150 base pair (bp) insert libraries were used for capture without indexing. The Illumina GAIIx platform was used to obtain paired end 76 bp and 101 bp sequencing reads (Bentley, et al., 2008) as technology progressed. Potential duplicate reads arising from PCR duplication were retained since the National Institutes of Health Intramural Sequencing Center (NISC) has observed that their PCR duplicate levels are consistently <10% of reads and that their genotypes have >99.9% concordance with genotypes from SNP arrays.

Alignment to the human genome reference sequence (UCSC assembly hg18, NCBI build 36) was carried out using the Efficient Large-Scale Alignment of Nucleotide Databases (ELAND) package (Illumina, Inc., San Diego, CA). ELAND was used in such a way that paired-end reads were aligned independently, and those that aligned uniquely were grouped into genomic sequence intervals of about 100 kb. Those that failed to align were binned with their paired-end mates, thus making use of paired-end information not utilized by ELAND. Reads that mapped equally well in more than one location were discarded. *Cross_Match* (P. Green, http://www.phrap.org), a Smith-Waterman based local alignment algorithm, was used to align binned reads to their respective 100 kb genomic sequence, using the parameters –minscore 21 and –masklevel 0. Cross_Match alignments were converted to the SamTools BAM format.

Because of the large number of exome sequences already aligned to hg18, NISC has continued to use this as the reference for exome sequence alignment. In order to compare our exome sequences to the ClinSeq exome sequences, which we use as an internally controlled allele frequency filter, we did not realign to the hg19 reference. We also elected to align to hg18 because even though hg18 has technically been superseded by hg19, hg18 still has more UCSC annotation. Consequently, all positions within the main text and supporting information refer to hg18 genome coordinates.

Genotypes were called using bam2mpg2 (http://research.nhgri.nih.gov/software/bam2mpg) for all positions with high-quality sequence (Phred-like Q20 or greater) using a Bayesian algorithm (Most Probable Genotype (MPG))(Teer, et al., 2010). Genotypes with an MPG score ≥10 had a > 99.89% concordance to genotypes from SNP array data. Similar to the method for false positive reduction in the GATK software (Depristo, et al., 2011), an optional data quality filter MPG/coverage ratio ≥0.5 was also applied to reduce false positives due to alignment errors (Ajay, et al., 2011; Wei, et al., 2011). Missense variants were then assigned a delta score depending on the predicted degree of severity for functional disruption using the Conserved Domain-based Prediction (CDPred) algorithm (Bell, et al., 2011; Johnston, et al., 2010; McLaughlin, et al., 2010; Prickett, et al., 2011) (http://research.nhgri.nih.gov/software/CDPred/). Variants with a CDPred delta score between -1 and -30 are classified as "predicted deleterious". CDPred scores are based on well-annotated and manually curated protein domains when the variant can be aligned to an entry in the NCBI Conserved Domain Database (CDD). When alignment to a CDD entry is not possible, CDPred defaults to the BLOSUM 62 substitution matrix. The positive or negative magnitude of output scores is more restricted when the substitution matrix is used reflecting the paucity of data in those regions. CDPred was chosen over other programs such

as SIFT (Ng and Henikoff, 2003) and Polyphen (Adzhubei, et al., 2010) because it was easy to incorporate into an automated pipeline, provided suitable output characteristics for our analyses and performed similarly to SIFT and Polyphen (unpublished data).

Filtering and statistical analysis of variants

The variant lists provided by NISC were sorted and filtered using the VarSifter software (http://research.nhgri.nih.gov/software/VarSifter) (Teer, et al., 2011) and then exported to Excel (Microsoft Corp., Renton, WA) for further analysis. Boolean logic filtering was performed using the JavaSDK package implemented in VarSifter. Conditional exact Hardy Weinberg equilibrium (HWE) one-sided testing was performed on all the available data for the UDP variants using the conditional HWExact module with the "greater" option selected. The R language package was developed by Jan Graffelmann (Engels, 2009; Graffelman, 2010; Wigginton, et al., 2005) (http://www.r-project.org, http://www-eio.upc.edu/~jan).

Varsifter formatted Bedfiles mentioned in this manuscript (Supp. Files S1–S4) were generated using the online software analysis suite Galaxy (http://main.g2.bx.psu.edu/root) (Blankenberg, et al., 2010; Goecks, et al., 2010). These bedfiles are provided as supporting information in the online version of this manuscript. Since the primary purpose of these files is to exclude suspected false positive variants from the data being queried within Varsifter, the genome wide complement data option in Galaxy was used to meet the program's formatting requirements of only an "include bedfile positions" option. Therefore, although these files can be viewed using the UCSC genome browser (http://genome.ucsc.edu/), the display of these custom tracks shows variant positions in white and all other regions of the genome as a solid black line.

Comparison of identified variants with other platforms and alignment methods

The subset of putative confounding variants identified in the UDP exomes and meeting the criteria of MPG ≥10 and MPG/Coverage ≥0.5 were compared to four whole genome sequencing datasets generated with the Illumina HiSeq 2000 sequencer and paired end reads (100bp). The sequence data were aligned using the Burrows Wheeler Algorithm (Li and Durbin, 2009). We also compared ES variants to 69 human genomes publically released from Complete Genomics, Inc. (Drmanac, et al., 2010) using BEDtools (v2.12) and in-house Perl scripts.

Gene exclusion list generation

Developing the provisional gene name exclusion list began by grouping all variants for the 29 families by locus name. The total number of predicted deleterious variants per family at each locus was recorded and the loci were sorted by number of occurrences. Validated NCBI pseudogenes were identified in the latest version of Gene (http://www.ncbi.nlm.nih.gov/gene) and added to an alternative gene exclusion list.

RESULTS

ES variants shared among probands with dissimilar phenotypes

Patients enrolled in the NIH Undiagnosed Diseases Program (UDP) exhibit heterogeneous, striking, and unusual phenotypes that have eluded diagnosis and may reflect new diseases. To determine the genetic bases of these disorders, we performed exome sequencing on a subset of participants. Using the 38 Mb Agilent Human SureSelect Kit, which targets the NCBI Consensus Coding Sequence, and two GAIIX flow cell lanes gave an average of 2.8 Gb of aligned sequence bases per sample, and on average 88.9% of baited nucleotides had an MPG score ≥10. For the 50 Mb Agilent Human SureSelect Kit, we obtained an average of 3.9 Gb of aligned sequence bases per sample, and on average 88.8% of baited nucleotides

had an MPG score ≥10. Further details on coverage statistics are provided in the accompanying manuscript of (Dias, et al., 2012).

The total set of sequenced exomes comprised 118 individuals, including a founder subset of 55 individuals. Using a quality cutoff of MPG ≥10, a subset of 698,248 unique variants was detected in the total set, and subset of 549,242 in the founder set (Table 1). Many variants were recurrent, despite highly divergent phenotypes among the probands. The diseases represented by the UDP cohort are likely rare and highly penetrant. We reasoned that any ES variant shared by multiple families with different proband phenotypes is unlikely to be disease-causing even if it is predicted to be deleterious by algorithms such as SIFT (Ng and Henikoff, 2003), Polyphen (Adzhubei, et al., 2010) or CDPred (Johnston, et al., 2010). This provided a component of the rationale for creating lists of low-interest variants.

Highly variable genes

We hypothesized that genes frequently containing numerous pathogenic variants had a low probability of being the source of disease-causing candidates for most exome/genome projects. Therefore, we sought to exclude genes that had frequent variations from the RefSeq that were plausibly pathogenic (missense, nonsense, frame-shifting, canonical-splice-site modifying) and rare enough to remain after filtering out common polymorphisms. For some exclusion lists, we applied a software prediction of variant pathogenicity using CDPred. The genes we identified are enumerated in the lists Supp. Tables S1–S7 and are listed along with construction notes in Supp. Table S8. For our exome projects, we applied gene exclusion lists as a provisional filtration step, adding back subsets of the excluded genes if no convincing disease-causing variants were found.

Deviations from Hardy-Weinberg Equilibrium: Excess Heterozygosity

The presence of excess heterozygosity in a cohort of exome sequence data is suggestive of sequence-read alignment errors, wherein two similar sequences, each homozygous for a different nucleotide at one or more positions, are aligned. We investigated whether such patterns existed in our data and found 392 variants with an MPG ≥10 that were heterozygous in all 118 UDP exomes (Supp. File S1 and Supp. Table S9 (tab heterozygous_nonref_annotations.xls)).

Previous publications concerning SNP (Doron and Shweiki, 2011) and exome data have proposed that the genotypes of misaligned sequences will be in Hardy-Weinberg disequilibrium (Engels, 2009; Graffelman, 2010; Wigginton, et al., 2005). The *a priori* probability of only heterozygous genotypes, based upon equal allele frequencies, is $p \leq (½)^{-55}$ for the 55 independent founder genotypes and is $p \leq (½)^{-118}$ when including the entire cohort. Applying a Bonferonni correction of 7.0×10^{-5} to the 549,242 ES variants identified in the founders, we concluded that a conditional, single tailed, HWE exact p value of less than 7.0×10^{-8} would be significant for inclusion into a false positive list at $p < 0.05$. Using this criterion, we identified 23,389 positions with excess heterozygosity (Supp. File S2 and Supp. Table S9 (tab heterozygous_nonref_annotations_2.xls)); each variant had an MPG ≥10 in at least one exome.

We reasoned that, if these heterozygous variants arose from a compression block, a region where highly similar sequences are inadvertently compressed computationally (Roach, et al., 2010), the two nearly identical component sequences that were misaligned might show up as copy number variations detected by other means. Confirming our suspicion, we found 15,140 variants in CNV regions listed in the Database of Genomic Variants (DGV) and as identified by RepeatMasker, 2,104 variants within repeat regions and 593 variants within

tandem repeats using SeattleSeq variant annotation (http://gvs.gs.washington.edu/SeattleSeqAnnotation/).

Comparing the 392 positions where heterozygosity was the genotype in every exome to the Agilent SureSelect baited regions, we found that fully half of the heterozygous variation in these positions arose from incidental capture of nontargeted regions. In addition, two baits had targeted regions that are now annotated as pseudogenes.

Deviations from Hardy-Weinberg Equilibrium: Excess Homozygosity

The presence of excess non-reference homozygosity for a given base pair in an exome cohort suggests that the reference sequence contains a minor-allele nucleotide designation—one that does not represent the major allele in the population from which the exome cohort was derived. We identified 1,009 positions in which every exome was homozygous for a non-reference genotype with an MPG ≥10. Using the UCSC genome browser to compare a subset of 187 randomly selected variants to cDNA sequences aligned to the hg18 human genome reference sequence, we found that in all cases where a cDNA sequence was available, the reference cDNA sequence agreed with the non-reference genotype call in our exome data. The 1009 nonreference homozygous variants are provided as a varsifter BED formatted file Supp File S3, and additional data about the variants are included in Supp. Table S9 (tab homozygous_nonref_annotations.xls).

Presence of Variants in dbSNP

The mechanisms discussed above may also produce DNA variant genotype calls with other types of genotyping technology. We searched dbSNPv130 to see if our variants had been previously reported. Of all the homozygous non-reference variants of high quality (MPG≥10), 96.8% were in dbSNP. For the heterozygous variants identified by HWE testing, 68% were in dbSNP. The SNP reference numbers for all variants are provided in Supp. Table S9.

DISCUSSION

To filter false positive variants from exome sequence data and thereby aid discovery of disease-causing mutations, we present a set of three tools, i.e., lists of highly polymorphic genes, positions of recurrent suspected misalignments to the human genome reference sequence, andpositions at which the hg18 human genome reference sequence contains a minor allele. These lists, complementing previously published exome variant analysis tools (Bilguvar, et al., 2010; Choi, et al., 2009; Hoischen, et al., 2010; Ng, et al., 2010a; Ng, et al., 2010b), were derived from analyses of exome data generated by the NIH UDP exome set, not from analysis of public databases. The tools are likely to be platform or alignment specific to some extent, particularly the list of heterozygous sites. However, they may be applicable in general to exome data collected with the widely used Agilent-Illumina sequencing technologies.

Our list of highly “polymorphic” genes starts with those identified empirically as having a large number of variants. We arbitrarily defined “large” as at least 10 predicted deleterious variants by the CDPred algorithm (CDPred ≤ -1). One method to select genes that look to be enriched for false positive variants, is to examine a diverse human population for genes that consistently generate many variants. This could be due to sequence alignment artifacts or true polymorphic nature of the genes; the ES samples obtained from the NIH UDP cohort constitute an ideal cohort of such individuals with disparate conditions (Gahl and Tifft, 2011). To this list, we added genes of lower relevance for our UDP patients, such as members of the olfactory and taste receptor gene families. Although these genes could cause

disease by duplication and divergence to a moonlighting function or by exerting a dominant negative effect, we considered these possibilities unlikely; also, adding them to our list of excluded genes on a first pass analysis does not preclude the option of revisiting them later. After thorough characterization of genes and regions where such variants are found, future refinements could involve excluding only the highly polymorphic regions of these genes rather than those entire gene loci.

Another common problem in analyzing ES data involves confounding variants arising from misalignment of sequences to the human genome reference sequence. These false positives can be identified using family data. Any ES data set, if produced in a consistent way, can be analyzed for HWE deviations. Given the large number of variant positions generated by ES, a conservative Bonferonni correction for multiple sampling was used to avoid spurious exclusion of variants that by chance appeared out of HWE. Furthermore, lists of false positive variants derived by looking at HWE should be generated using a single alignment algorithm, as was done in this case. However, even when looking across platforms and algorithms at the 37% (146/393) of the ES variants where there are only heterozygous genotypes, we found concordance in more than 50 of the 69 publically available Complete Genomics genomes.

Another source of confounding variants arises when the hg18 human genome reference sequence contains a minor allele, rare disease-causing variant, or a simple sequencing error. Although some of our variants might occur due to systemic errors in NGS compared to the Sanger method used to generate the human genome reference sequence (Balasubramanian, et al., 2011; van der Maarel, et al., 2011), comparing NGS of exomes with that of Illumina genomes confirmed that the vast majority of the variant genotypes were correctly called. For 85% (863/1009) of the ES homozygous non-RefSeq genotypes, we also found concordance in more than 50 of the 69 publically available Complete Genomics genomes. This suggests that these variants are not always platform-, chemistry- or alignment-specific. Fortunately, these errors will become less common as non-disease causing variations in the human genome are identified and annotated (Church, et al., 2011). In fact, the accelerating accumulation of sequencing data continues to contribute to the accuracy of a variety of data sets including dbSNP and the human pan-genome (Li, et al., 2010).

Many of the variants detected by HWE and exome-dataset analysis also occur in dbSNP. For the homozygous non-reference variants, this is not surprising, since minor alleles in the reference sequence will differ from sequence data obtained using any technology. For heterozygous variants, the percentage of variants in dbSNP is smaller, and may represent similar sequencing specificity issues as those that arise using NGS. Filtration using unselected dbSNP records introduces a well-described hazard of excluding important and possibly disease-causing variants. Identification of variants using the methods we describe allows for the construction of ES variant filters with known and rationally formulated characteristics.

In conclusion, incremental improvements in the analysis of genome data will occur with improved sequencing chemistry, better alignments, longer read lengths, deeper coverage, and advanced technologies that address inadequacies in long-range sequencing and gap filling (Homer and Nelson, 2010; Schatz, et al., 2010). For now, lists of problematic genes or variant locations (e.g., heterozygous genotypes with HWE inconsistencies or all homozygous nonhuman genome reference alleles) help to identify false positive signals (Supp. File S4 and Supporting Information text). Such lists assist in the winnowing of ES variants and are essential for disease-causing gene discovery.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank our patients and their families, who are partners in the pursuits of the NIH UDP. We appreciate the excellent technical skills of Roxanne Fischer and Richard Hess. We also thank Dr. Ajay Shankar Subramaniam and Dr. Elliott H Margulies, who helped with the correlation of suspected false positive variants in exome sequencing against Illumina genome sequences. We value the help from Taylor Davis, who assisted in the manual investigation of the “all homozygous” variants in our dataset using the UCSC browser and Dr. Praveen Cherukuri’s advise regarding the use of CdPred for our dataset.

Grant Sponsor

This work was supported by NHGRI intramural funding and by the UDP program through the office of the Director NIH.

REFERENCES

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nature methods. 2010; 7:248–249. [PubMed: 20354512]

Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome biology. 2011; 12:R18. [PubMed: 21338519]

Ajay SS, Parker SCJ, Abaan HO, Fuentes Fajardo KV, Margulies EH. Accurate and comprehensive sequencing of personal genomes. Genome Research. 2011 In press.

Balasubramanian S, Habegger L, Frankish A, MacArthur DG, Harte R, Tyler-Smith C, Harrow J, Gerstein M. Gene inactivation and its implications for annotation in the era of personal genomics. Genes & development. 2011; 25:1–10. [PubMed: 21205862]

Bell DW, Sikdar N, Lee KY, Price JC, Chatterjee R, Park HD, Fox J, Ishiai M, Rudd ML, Pollock LM, et al. Predisposition to cancer caused by genetic and functional defects of mammalian Atad5. PLoS Genet. 2011; 7:e1002245. [PubMed: 21901109]

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008; 456:53–59. [PubMed: 18987734]

Biesecker LG, Mullikin JC, Facio FM, Turner C, Cherukuri PF, Blakesley RW, Bouffard GG, Chines PS, Cruz P, Hansen NF, et al. The ClinSeq Project: piloting large-scale genome sequencing for research in genomic medicine. Genome research. 2009; 19:1665–1674. [PubMed: 19602640]

Bilguvar K, Ozturk AK, Louvi A, Kwan KY, Choi M, Tatli B, Yalnizoglu D, Tuysuz B, Caglayan AO, Gokben S, et al. Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. Nature. 2010; 467:207–210. [PubMed: 20729831]

Blankenberg, D.; Von Kuster, G.; Coraor, N.; Ananda, G.; Lazarus, R.; Mangan, M.; Nekrutenko, A.; Taylor, J. Current protocols in molecular biology / edited by Frederick M. Ausubel …[et al.] Chapter 19: Unit 19. 2010. Galaxy: a web-based genome analysis tool for experimentalists; p. 1-21.

Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloglu A, Ozen S, Sanjad S, 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; 106:19096–19101. [PubMed: 19861545]

Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen HC, Agarwala R, McLaren WM, Ritchie GR, et al. Modernizing reference genome assemblies. PLoS Biol. 2011; 9:e1001091. [PubMed: 21750661]

Coffey AJ, Kokocinski F, Calafato MS, Scott CE, Palta P, Drury E, Joyce CJ, Leproust EM, Harrow J, Hunt S, et al. The GENCODE exome: sequencing the complete human exome. European journal of human genetics : EJHG. 2011

Depristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas MA, Hanna M, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics. 2011; 43:491–498. [PubMed: 21478889]

Dias C, Sincan M, Rupps R, Briemberg H, Selby K, Mullikin J, Markello T, Adams D, Gahl WA, Boerkoel CF. Exome sequencing: diagnosis of genetically heterogeneous neuromuscular disorders. Hum Mutat. 2011; 33

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. [PubMed: 18660515]

Doron S, Shweiki D. SNP uniqueness problem: a proof-of-principle in HapMap SNPs. Human mutation. 2011; 32:355–357. [PubMed: 21412948]

Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science. 2010; 327:78–81. [PubMed: 19892942]

Engels WR. Exact tests for Hardy-Weinberg proportions. Genetics. 2009; 183:1431–1441. [PubMed: 19797043]

Gahl WA, Markello TC, Toro C, Fajardo KF, Sincan M, Gill F, Carlson-Donohoe H, Gropman A, Pierson TM, Golas G, et al. The National Institutes of Health Undiagnosed Diseases Program: Insights into rare diseases. Genet Med. 2011 PMID: 21952431 (Published online, ahead of print).

Gahl WA, Tifft CJ. The NIH Undiagnosed Diseases Program: lessons learned. JAMA. 2011; 305:1904–1905. [PubMed: 21558523]

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature biotechnology. 2009; 27:182–189.

Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome biology. 2010; 11:R86. [PubMed: 20738864]

Graffelman J. The number of markers in the HapMap project: some notes on chi-square and exact tests for Hardy-Weinberg equilibrium. American journal of human genetics. 2010; 86:813–818. author reply 818-9. [PubMed: 20466092]

Hoischen A, van Bon BW, Gilissen C, Arts P, van Lier B, Steehouwer M, de Vries P, de Reuver R, Wieskamp N, Mortier G, et al. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nature genetics. 2010; 42:483–485. [PubMed: 20436468]

Homer N, Nelson SF. Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA. Genome biology. 2010; 11:R99. [PubMed: 20932289]

Johnston JJ, Teer JK, Cherukuri PF, Hansen NF, Loftus SK, Chong K, Mullikin JC, Biesecker LG. Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate. American journal of human genetics. 2010; 86:743–748. [PubMed: 20451169]

Koboldt DC, Ding L, Mardis ER, Wilson RK. Challenges of sequencing human genomes. Briefings in bioinformatics. 2010; 11:484–498. [PubMed: 20519329]

Landan G, Graur D. Heads or tails: a simple reliability check for multiple sequence alignments. Molecular biology and evolution. 2007; 24:1380–1383. [PubMed: 17387100]

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. Initial sequencing and analysis of the human genome. Nature. 2001; 409:860–921. [PubMed: 11237011]

Ledergerber C, Dessimoz C. Base-calling for next-generation sequencing platforms. Briefings in bioinformatics. 2011

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]

Li R, Li Y, Zheng H, Luo R, Zhu H, Li Q, Qian W, Ren Y, Tian G, Li J, et al. Building the sequence map of the human pan-genome. Nat Biotechnol. 2010; 28:57–63. [PubMed: 19997067]

McLaughlin HM, Sakaguchi R, Liu C, Igarashi T, Pehlivan D, Chu K, Iyer R, Cruz P, Cherukuri PF, Hansen NF, et al. Compound heterozygosity for loss-of-function lysyl-tRNA synthetase mutations

in a patient with peripheral neuropathy. Am J Hum Genet. 2010; 87:560–566. [PubMed: 20920668]

Meacham F, Boffelli D, Dhahbi J, Martin DIK, Singer M, Pachter L. Identification and correction of systematic error in high-throughput sequence data. Nature Precedings. 2011

Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research. 2003; 31:3812–3814. [PubMed: 12824425]

Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nature genetics. 2010a; 42:790–793. [PubMed: 20711175]

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, et al. Exome sequencing identifies the cause of a mendelian disorder. Nature genetics. 2010b; 42:30–35. [PubMed: 19915526]

Pierson TM, Adams D, Bonn F, Martinelli P, Cherukuri PF, Teer JK, Hansen NF, Cruz P, Mullikin For The Nisc Comparative Sequencing Program JC. Blakesley RW, et al. Whole-Exome Sequencing Identifies Homozygous AFG3L2 Mutations in a Spastic Ataxia-Neuropathy Syndrome Linked to Mitochondrial m-AAA Proteases. PLoS Genet. 2011; 7:e1002325. [PubMed: 22022284]

Prickett TD, Wei X, Cardenas-Navia I, Teer JK, Lin JC, Walia V, Gartner J, Jiang J, Cherukuri PF, Molinolo A, et al. Exon capture analysis of G protein-coupled receptors identifies activating mutations in GRM3 in melanoma. Nat Genet. 2011; 43:1119–1126. [PubMed: 21946352]

Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science. 2010; 328:636–639. [PubMed: 20220176]

Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. Genome research. 2010; 20:1165–1173. [PubMed: 20508146]

Snyder M, Du J, Gerstein M. Personal genome sequencing: current approaches and challenges. Genes & development. 2010; 24:423–431. [PubMed: 20194435]

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat. 2003; 21:577–581. [PubMed: 12754702]

Teer JK, Bonnycastle LL, Chines PS, Hansen NF, Aoyama N, Swift AJ, Abaan HO, Albert TJ, Margulies EH, Green ED, et al. Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. Genome research. 2010; 20:1420–1431. [PubMed: 20810667]

Teer JK, Green ED, Mullikin JC, Biesecker LG. VarSifter: Visualizing and analyzing exome-scale sequence variation data on a desktop computer. Bioinformatics. 2011 [Epub ahead of print].

van der Maarel SM, Tawil R, Tapscott SJ. Facioscapulohumeral muscular dystrophy and DUX4: breaking the silence. Trends in molecular medicine. 2011; 17:252–258. [PubMed: 21288772]

Wei X, Walia V, Lin JC, Teer JK, Prickett TD, Gartner J, Davis S, Stemke-Hale K, Davies MA, Gershenwald JE, et al. Exome sequencing identifies GRIN2A as frequently mutated in melanoma. Nature genetics. 2011; 43:442–446. [PubMed: 21499247]

Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. American journal of human genetics. 2005; 76:887–893. [PubMed: 15789306]

Table 1

Analysis of 698264 sequence variants detected in ES data obtained from 29 UDP families

Data set	Number of variants	Number of genes
Variants arising in highly polymorphic genes		
≥10 variants in all families	N/A	17
≥10 variants in ≥3 families	N/A	166
Variants arising from misalignment		
Heterozygous in every exome	392	45
Excess heterozygosity	23,389	2,576
Variants arising from biases in the Hg18 human genome reference sequence		
Homozygous non-Hg18 human genome reference sequence in every exome	1,009	707