Genetics inMedicine | ORIGINAL RESEARCH ARTICLE

Open

Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios

Xiaolin Zhu, MD[1], Slavé Petrovski, PhD[1,2], Pingxing Xie, PhD[1,13], Elizabeth K. Ruzzo, PhD[1], Yi-Fan Lu, BS[1], K. Melodi McSweeney, BS[1], Bruria Ben-Zeev, MD[3,4], Andreea Nissenkorn, MD[3,4], Yair Anikster, MD, PhD[3,4], Danit Oz-Levi, MS[5], Ryan S. Dhindsa[1], Yuki Hitomi, PhD[1,14], Kelly Schoch, MS, CGC[6], Rebecca C. Spillmann, MS, CGC[1], Gali Heimer, MD, PhD[3,7], Dina Marek-Yagel, PhD[8], Michal Tzadok, MD[3,4], Yujun Han, PhD[1], Gordon Worley, MD[6], Jennifer Goldstein, PhD, CGC[6], Yong-Hui Jiang, MD, PhD[6,9], Doron Lancet, PhD[5], Elon Pras, MD[3,10], Vandana Shashi, MD[6], Duncan McHale, MBBS, PhD[11], Anna C. Need, PhD[1,12] and David B. Goldstein, PhD[1]

Purpose: Despite the recognized clinical value of exome-based diagnostics, methods for comprehensive genomic interpretation remain immature. Diagnoses are based on known or presumed pathogenic variants in genes already associated with a similar phenotype. Here, we extend this paradigm by evaluating novel bioinformatics approaches to aid identification of new gene–disease associations.

Methods: We analyzed 119 trios to identify both diagnostic genotypes in known genes and candidate genotypes in novel genes. We considered qualifying genotypes based on their population frequency and in silico predicted effects we also characterized the patterns of genotypes enriched among this collection of patients.

Results: We obtained a genetic diagnosis for 29 (24%) of our patients. We showed that patients carried an excess of damaging de novo mutations in intolerant genes, particularly those shown to be essential in mice ($P = 3.4 \times 10^{-8}$). This enrichment is only partially explained by mutations found in known disease-causing genes.

Conclusion: This work indicates that the application of appropriate bioinformatics analyses to clinical sequence data can also help implicate novel disease genes and suggest expanded phenotypes for known disease genes. These analyses further suggest that some cases resolved by whole-exome sequencing will have direct therapeutic implications.

Genet Med advance online publication 15 January 2015

Key Words: diagnosis; genic intolerance; *HNRNPU*; rare disease; whole-exome sequencing

INTRODUCTION

Whole-exome sequencing (WES) has emerged as a successful diagnostic tool in the study of genetic disease and has proven to be particularly effective in identifying disease-associated genes that are refractory to linkage analysis. Furthermore, WES has shown diagnostic utility in routine clinical settings. One recent exome sequencing study reported a diagnostic rate of 25% in a heterogeneous population of 250 patients,[1] whereas other studies have reported variable, sometimes higher, rates.[2–4] However, a limitation of most current studies, and indeed clinical analyses, is that a diagnosis is only possible if the gene has been previously implicated in a similar condition. The immense value that WES offers in identifying novel disease genes often remains unexplored.

In this study, we applied trio WES (i.e., sequencing the patient and both unaffected biological parents) to a cohort of 119 patients who had been referred to medical geneticists for a variety of conditions. They comprise 113 trios reported for the first time and 6 previously unresolved trios[2] that are reinterpreted. Importantly, this cohort of patients reflects a heterogeneous collection of clinical presentations, making the cohort representative of a typical genetics clinic. Through trio sequencing,

The last two authors contributed equally to this work.
The first two authors contributed equally to this work.
[1]Center for Human Genome Variation, Duke University School of Medicine, Durham, North Carolina, USA; [2]Department of Medicine, University of Melbourne, Austin Health and Royal Melbourne Hospital, Melbourne, Australia; [3]Edmond and Lily Safra Children's Hospital, Sheba Medical Center, Ramat Gan, Israel; [4]Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel; [5]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel; [6]Department of Pediatrics, Duke University School of Medicine, Durham, North Carolina, USA; [7]Pinchas Borenstein Talpiot Medical Leadership Program, Pediatric Neurology Unit, Chaim Sheba Medical Center, Tel HaShomer, Israel; [8]Metabolic Disease Unit, Edmond and Lily Children's Hospital, Sheba Medical Center, Ramat Gan, Israel; [9]Department of Neurobiology, Duke University, Durham, North Carolina, USA; [10]Danek Gertner Institute of Human Genetics, Sheba Medical Center, Ramat Gan, Israel; [11]UCB NewMedicines, Slough, UK; [12]Division of Brain Sciences, Department of Medicine, Imperial College London, London, UK; [13]Present address: Department of Human Genetics, McGill University, Montreal, Quebec, Canada; [14]Present address: Department of Human Genetics, Graduate School of Medicine, University of Tokyo, Tokyo, Japan. Correspondence: David B. Goldstein (dbgoldstein@columbia.edu) or Anna C. Need (a.need@imperial.ac.uk) or Slavé Petrovski (slave.petrovski@duke.edu)

Submitted 27 August 2014; accepted 19 November 2014; advance online publication 15 January 2015. doi:10.1038/gim.2014.191

we identify qualifying genotypes, focusing on genotypes seen only in the patient and not in unaffected parents or controls.[2] This approach not only provides diagnoses based on mutations found in already known genes but also can provide pointers toward novel disease genes, such as the identification of *NGLY1* in our previous work,[2] which was subsequently confirmed in seven additional rare cases with similar clinical presentations.[5] We expect such examples will emerge more frequently as careful bioinformatics analysis of candidates becomes a central part of diagnosing genetic disease.[6]

MATERIALS AND METHODS

Subjects

A total of 113 patients with suspected genetic disorders, and their unaffected biological parents, are reported here for the first time. In general, the clinical presentations of most patients were considered severe (**Supplementary Table S1** online), and the patients were either judged to have an *undiagnosed* genetic disorder or suspected to have a specific genetic disorder that was genetically *unresolved* based on known diagnostic panels. Sixty-five trios were recruited at the Genome Sequencing Clinic at Duke University Medical Center, and 48 were recruited at the pediatric clinic of the Sheba Medical Center in Tel HaShomer, Israel. Clinical phenotypes are shown in **Supplementary Table S1** online. We previously published an analysis of 12 pilot trios recruited in a similar way at Duke.[2] Here, we reinterpret six trios that were unresolved in those analyses.[2] A total of 119 trios are reported.

All 119 patients underwent a clinical genetics evaluation, with traditional genetic diagnostics performed whenever clinically indicated. All patients reported here remained undiagnosed or unresolved after tests including candidate mutation and/or gene testing, karyotyping, chromosomal microarray analysis, and gene panels. The appropriate institutional review boards approved this research protocol. Written informed consent was received from all participants or their guardians.

To estimate the population frequency of both variants and genotypes, we used two independent sources of population control data, totaling up to 9,530 individuals. The internal control cohort comprised subjects enrolled in the Center for Human Genome Variation through Duke institutional review board–approved protocols (n = 3,027). Among these controls, 83.5% were Caucasian, 6.7% were Middle Eastern, 3.5% were African, and the remaining 6.3% were of Asian or other ancestries. Although these internal controls were not ascertained for rare disorders, their individual phenotypes were not analyzed and some could have rare pathogenic variants. The external control cohort comprises subjects enrolled in the National Heart, Lung, and Blood Institute Grand Opportunity Exome Sequencing Project (n = 6,503, http://evs.gs.washington.edu/EVS/).

Sequencing and bioinformatics pipeline

DNA was extracted from a peripheral blood sample. To capture the coding regions, we used the 65-Mb Illumina TruSeq Exome Enrichment Kit (Illumina, San Diego, CA), the 64-Mb Roche NimbleGen SeqCap EZ Exome Library Kit (Roche NimbleGen, Madison, WI), or the 50-Mb Agilent SureSelect Human All Exon Kit (Agilent, Santa Clara, CA). The capture kit type was consistent within a given trio. All sequencing was performed on the Illumina HiSeq 2000 platform (Illumina) at the Genomic Analysis Facility in the Center for Human Genome Variation. The overall coverage statistics for each individual are shown in **Supplementary Table S2** online.

Using Burrows-Wheeler Aligner (BWA-0.5.10),[7] sequencing reads were mapped to a Genome Reference Consortium Human Genome Build 37 (GRCh37)-derived alignment set including decoy sequences; the same reference genome is used in the 1000 Genomes Project (http://www.1000genomes.org/). Polymerase chain reaction duplicates were removed using picard-tools-1.59 (http://picard.sourceforge.net). Single-nucleotide variants and small insertions/deletions (indels) were called using the UnifiedGenotyper of the Genome Analysis Toolkit (GATK-1.6–11)[8] and annotated using SnpEff-3.3 (Ensembl 73 database).[9] The six revisited trios were also subjected to this bioinformatics pipeline; previously, they were aligned to GRCh36 and variants were called using SAMtools.[10]

Identifying qualifying genotypes

For each patient, we hypothesized that completely penetrant genotypes would explain the major clinical manifestations. Using the trio WES data along with in silico and population site frequency data of genetic variants, we generated a list of "qualifying genotypes" for each trio. Each variant had to meet specific quality control thresholds (**Supplementary Method 1.1** online). In terms of functional annotation, we included only protein-altering variants, including truncating variants (stop gain/loss, start loss, or frameshift), missense variants, canonical splice-site variants, inframe indels affecting protein-coding regions, and variants within the intron–exon boundary (eight bases flanking the exonic boundaries). We primarily focused on genotypes absent in control data sets. We systematically considered four different genetic models, using stratified European and African Americans in the Exon Variant Server (EVS) for minor allele frequency estimations: (i) germ-line de novo mutations, also absent in the available control populations (extended to include mitochondrial DNA sequence by requiring mother to have only the reference allele while the patient has only the mutant allele); (ii) recessive homozygous genotypes, which were heterozygous in both parents, never homozygous in controls, with a control allele frequency <1%; (iii) hemizygous X chromosome variants inherited from an unaffected heterozygous mother, with a control allele frequency <1% and never observed in male controls or homozygous in female controls; and (iv) compound heterozygous genotypes in the patient (one variant inherited from each heterozygous parent, with the two variants occurring at different genomic positions within the same gene), for which neither variant was ever homozygous in controls, and each had a control allele frequency <1%. For the compound

heterozygous genotypes, we further required that regardless of phasing, the two variants never co-occurred in the Center for Human Genome Variation controls. Genotypes meeting these criteria were referred to as "qualifying genotypes," with the genes harboring qualifying genotypes referred to as "qualifying genes."

Determination of genetic diagnosis

For each trio, the list of qualifying genes was checked against OMIM. Specifically, we required that (i) the OMIM disease phenotype overlapped the patient's clinical features; (ii) the qualifying genotype was consistent with the reported OMIM inheritance pattern (e.g., dominant or recessive); and (iii) the qualifying mutation itself was reported in a similarly affected patient *or* the qualifying mutation was of the same functional class (e.g., loss-of-function, missense) as those reported in a similarly affected patient. The genetics team then communicated directly with the treating clinicians to discuss whether a relevant qualifying genotype could explain the clinical presentation of the patient. When both the genetics team and the treating clinicians agreed that the qualifying genotype was the final diagnosis, the qualifying genotype was considered to be a "genetic diagnosis" (**Supplementary Table S6** online). We referred to cases assigned a diagnosis in this way as "resolved."

Each of the variants leading to a diagnosis was visually inspected using Integrative Genomics Viewer[11] followed by Sanger validation. These Sanger validations were performed at the Center for Human Genome Variation for trios recruited from Sheba and by a CLIA-certified laboratory for trios recruited from Duke.

For trios without a genetic diagnosis determined (unresolved), and with a sufficiently small number of qualifying genotypes, we performed a broader literature inspection to highlight potentially interesting candidates to be followed up in future studies (**Supplementary Method 2** online).

Bioinformatic signatures of causal variants

Among the 103 patients who did not have a genetic diagnosis determined by an inherited genotype, we used a previously described gene-level and variant-level prioritization framework to interpret the properties of their de novo mutations in comparison with control trio de novo mutations.[12] For the gene-level score, we used the Residual Variation Intolerance Score scoring system introduced by Petrovski et al.[12] For the variant-level score, we took the Ensembl PolyPhen-2 HumVar scores[13] for missense de novo mutations and assigned nonsense and canonical splice-site mutations a score of 1. Synonymous mutations were assigned a score of 0. We prioritized de novo mutations that reside in the "hot zone" previously defined by Petrovski et al.[12] by a Residual Variation Intolerance Score percentile score (y-axis) ≤0.25 and a PolyPhen-2 score (x-axis) ≥0.95. We used data from 728 published control trio subjects;[14–19] 337 controls had at least one assessable de novo mutation to estimate the empirical distribution of the gene-level and variant-level scores in the 2D space, particularly the expected proportion of de novo mutations within the hot zone (**Supplementary Method 1.2** online).

We subsequently incorporated new information about essential mouse genes as extracted from the Mouse Genome Database by Georgi et al..[20] A correlation had been previously established between essential mouse genes and genes that are intolerant to functional variation in the human population (Residual Variation Intolerance Score; $P = 1.3 \times 10^{-114}$) (Table 1 of Petrovski et al.[12]). Despite this correlation, we wanted to assess whether integrating intolerant scores with the essential gene list would create a stronger signature of pathogenic mutations. We therefore assessed the value of taking the intersection between de novo hot zone mutations and 2,472 essential genes—human orthologs of genes that result in lethality when either or both copies are disrupted in mice.[20,21]

RESULTS

Genetic diagnosis based on known gene–disease associations

On average, 94.2% of the exome-wide consensus coding sequence (CCDS) sequence (release 14) was covered with at least 10-fold coverage (**Supplementary Table S2** online). We identified an average of 12 qualifying genotypes for each trio, averaging one de novo, one newly hemizygous, three newly homozygous, and seven compound heterozygous genotypes (**Supplementary Table S3** online). Compared with Duke trios (n = 71), Sheba trios (n = 48) had, on average, more newly homozygous qualifying genotypes (**Supplementary Table S3** online); this is consistent with the higher percentage of consanguinity among Sheba trios (18.8 vs. 1.4% among Duke trios; **Supplementary Table S1** online). Sheba trios also had more qualifying genotypes than Duke trios (**Supplementary Table S3** online), consistent with the fact that our control cohorts (comprising primarily Caucasians) are more ethnically matched with Duke trios than with Sheba trios (Middle Eastern origin).

Discussion with the treating physicians followed by Sanger validation established the diagnoses for a total of 29 patients (**Supplementary Table S6** online). Thirteen (45%) were due to a de novo mutation, seven (24%) were due to a newly homozygous genotype, five (17%) were due to a newly hemizygous genotype, and four (14%) were due to a compound heterozygous genotype. The percentage of cases diagnosed with recessive conditions is higher than that reported in previous studies,[1] presumably due to the increased level of consanguinity in the Israeli cohort as compared with the populations in most other published WES diagnostic studies.

We have four examples in which the genetic diagnosis led to an immediate change in management. For two of the four patients, the genetic diagnoses informed specific pharmacotherapies. These patients are the patient of trio 10, who has a de novo missense mutation in *KCNQ2* (**Supplementary Table S6** online) and has been prescribed retigabine,[22] and the patient of trio 53, who has a de novo missense mutation in *KCNT1* (**Supplementary Table S6** online) and has been treated with

quinidine.[23] Retigabine treatment has reduced seizure frequency in the first patient (trio 10) despite a lack of an observable positive effect on development. In two additional patients (trio 24 and trio 26), the genetic diagnoses led to specific diet interventions that significantly improved the patients' metabolic conditions.

Identifying novel disease genes: bioinformatic signatures can point toward novel genes

Prioritizing based on genic intolerance[12] highlights the clear presence of hot zone de novo mutations among our clinically heterogeneous collection of 103 patients unresolved by a recessive genotype. Seventy patients had at least one assessable de novo mutation, and 29 (41.4%) of these resided in the hot zone (**Supplementary Table S7** online). This is compared with the frequency of hot zone de novo mutations among control trios with 13.1% (44/337) of controls with at least one assessable de novo mutation ($P = 2.3 \times 10^{-7}$, Fisher's exact test; **Figure 1**). Similar to the previous publications of this approach in specific neuropsychiatric ascertainments,[12,24] this indicates that presence of a hot zone de novo mutation is a strong candidate even among our heterogeneous collection of rare disorders. Considering the 103 patients unresolved by a recessive disorder, 28.2% of the patients had a hot zone de novo mutation as compared with 6.0% among the 728 sequenced control trios ($P = 3.0 \times 10^{-10}$; **Figure 1**).

We also found that de novo mutations occurring in the essential genes were enriched among patients as compared with controls (31.1% (32/103) vs. 11.5% (84/728); $P = 1.1 \times 10^{-6}$). Moreover, we found that de novo hot zone mutations in essential genes occurred in 16 (15.5%) of the 103 case trios and in only 14 of the 728 control trios (1.9%) ($P = 3.4 \times 10^{-8}$, Fisher's exact test). This translates to ~88% (14/16) of the patients having been directly ascertained for such hot zone de novo mutations among essential genes (**Figure 1** and **Supplementary Table S7** online), and it strongly suggests that in the majority of the patients with such a mutation, the mutation is causing or contributing to the disorder. Interestingly, 5 (31%) of the 16 hot zone mutations in essential genes are already considered to be causal (*KCNQ2, NOTCH2, GNAO1,* and two individuals with a *DYNC1H1* de novo mutation; **Supplementary Table S6** online). An additional seven are among genes with an existing OMIM or PubMed disease association with a less consistent clinical phenotypic overlap (*DCTN1, BIN1, GRIN1, GRIN2B, COL4A1, MYO5A*, and *HUWE1*). Finally, four mutations occur in genes that are strong genetic candidates with no existing (OMIM or PubMed) literature support for germ-line disease association (*SLC9A1*, *HNRNPU*, *CELSR3*, and *EWSR1*) (**Supplementary Table S7** online). In the case of *HNRNPU*, recent studies identified de novo mutations in patients with epileptic encephalopathy,[25,26] showing a partially overlapping phenotype with our patient.

DISCUSSION

In this study of 119 trios, we diagnosed 29 patients through WES, which is in accordance with other recent studies (**Supplementary Table S6** online).[1–4] In another 21 (17.6%) patients, we identified strong candidate genes based on comparing the properties of de novo qualifying mutations seen in cases to those of a collection of control trios (**Supplementary Table S7** online). Among the unresolved trios, our inspection further identified candidate genes. Some trios had a qualifying genotype in an OMIM disease gene (**Supplementary Table S4** online), but they did not meet all the criteria described in "determination of genetic diagnosis." Four trios had qualifying genes of interest that are not currently OMIM disease-associated genes (**Supplementary Table S5** online).

Importantly, we successfully identified the genetic diagnoses for two trios that were negative in our pilot study[2] (see Materials and Methods). The first patient (trio N8) had an inherited hemizygous missense mutation in *ATRX*. This variant was not identified in the original analysis, in which alignment was to an older version of the human reference genome (GRCh36) and variant calling was performed using SAMtools. In the other patient (trio N12), we found a de novo stop gain in *SRCAP*; however, literature supporting a genetic diagnosis of Floating-Harbor syndrome (OMIM 136140) emerged after our original paper was submitted.[2,27] These two examples emphasize the value of reinterpreting clinical exomes on a regular basis to leverage the latest advances in medical literature and bioinformatics.

A solid diagnosis often has value for patients and families even when no new treatments result from the diagnosis. In a minority of cases, however, the correct diagnosis can improve patient management. The potential effectiveness of such genotype-driven treatments is often deducible from the assayed biological consequence of the disease-causing mutation and known mechanisms of action of candidate drugs, and is further testable from the behavior of mutant proteins in in vitro assays in the presence or absence of the candidate drugs.[22,23] It is also important to emphasize that unlike retigabine, a recently marketed anticonvulsant for which the potential utility in treating *KCNQ2*-associated epileptic encephalopathy might be better appreciated, quinidine was not considered as a possible therapy for refractory seizures and would never have been tried in a patient without knowledge of the *KCNT1* genotype. Similarly, the genetic diagnoses for patients affected with metabolic disorders, with extensive yet uninformative metabolic workups, guided active clinical management (e.g., diet interventions for trio 24 and trio 26) that clearly benefitted the patients.

We found that 15 (13%) of the 119 interpreted trios achieved a genetic diagnosis based on having the same variant as those previously reported in patients with similar phenotypes. This could indicate low allelic heterogeneity among some of the disease-causing genes. As an example, for each of these 15 genes we considered the "opportunity" to have identified an overlapping pathogenic variant based on the collection of OMIM (http://www.omim.org/), ClinVar,[28] and HGMD[29] reported pathogenic variants. We found the range to be 5–235 unique "pathogenic" variants per gene, or 0.02–2.79% of the possible single-nucleotide substitution events after accounting for gene size. Among the 15 genes for which we identified the same variant in our

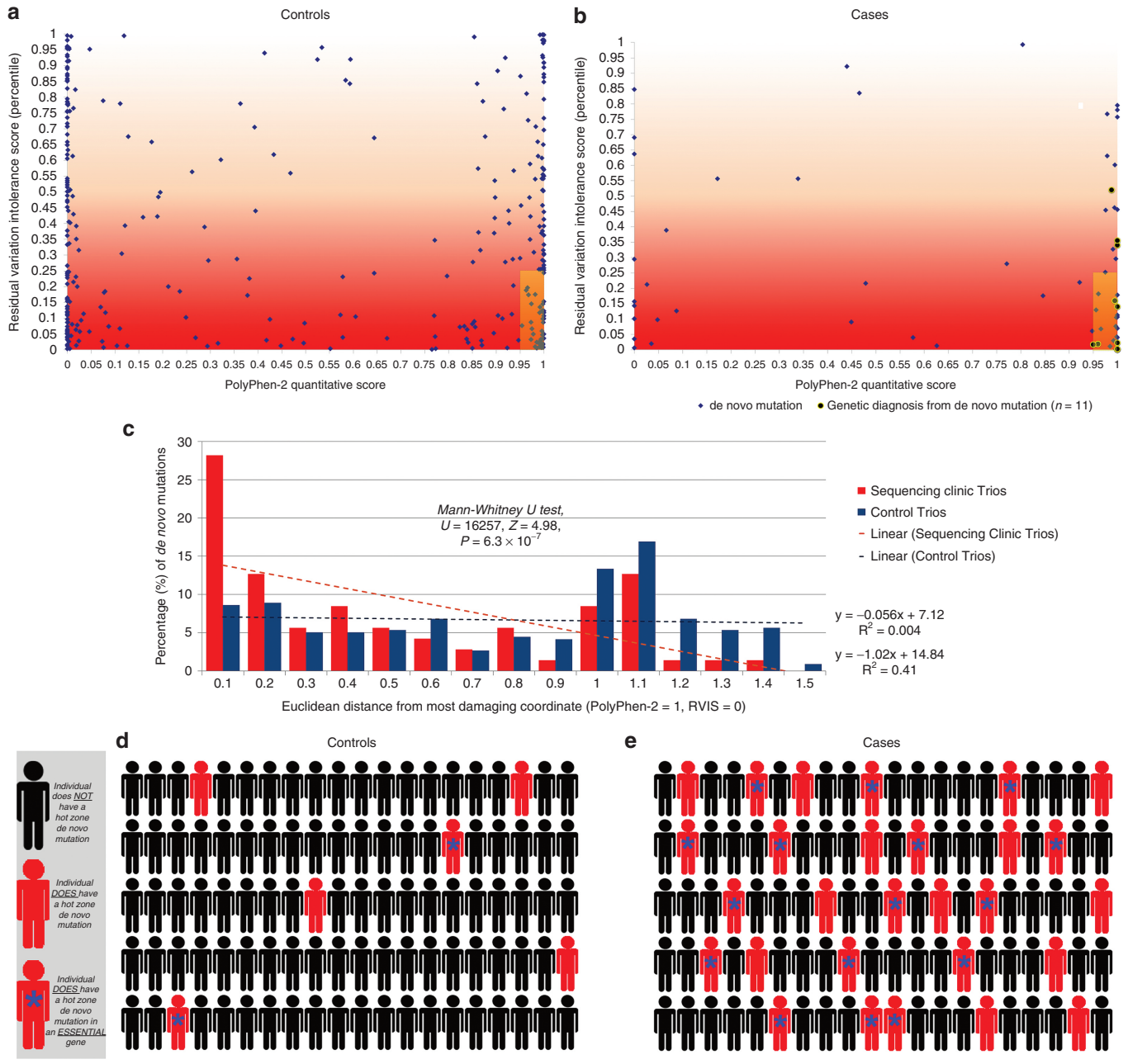

Figure 1 Hot zone bioinformatic signatures of de novo mutations. Among (**a**) controls and (**b**) cases who had at least one protein-coding de novo mutation identified, the gene-level (Residual Variation Intolerance Score (RVIS) genic intolerance percentile) and variant-level (PolyPhen-2) scores are plotted in two-dimensional space. Black circles in (**b**) are used for trios achieving a genetic diagnosis via the plotted de novo mutation. Blue diamonds represent de novo mutations among trios that have at least one de novo mutation but are currently unresolved by a de novo or inherited mutation. In (**a**) and (**b**) the hot zone is the shaded region corresponding to PolyPhen-2 (x-axis) ≥0.95 and RVIS (y-axis) ≤0.25. Of the 337 control de novo mutations, 44 (13.1%) occur in the hot zone as compared with 29 (41.4%) of 70 de novo mutations observed among cases (Fisher's exact test, $P = 2.3 \times 10^{-7}$). This indicates that among the cases there is an excess of 20 (69%) de novo mutations in the hot zone. To further illustrate the difference between the two populations (cases = red; controls = blue), (**c**) a histogram shows the distribution of Euclidean distances to the most damaging coordinate (PolyPhen-2 = 1 and RVIS = 0) for de novo mutations plotted in (**a**) and (**b**). It is strikingly clear that (**b**) de novo mutations identified among patients ascertained for severe undiagnosed genetic conditions are drawn from a distribution that is significantly closer in Euclidean distance to the most damaging coordinate than are (**a**) de novo mutations ascertained from a control population (Mann–Whitney *U*-test, $P = 6.3 \times 10^{-7}$). Linear regression lines were generated for both populations. Population-level representations (**d**) and (**e**) of hot zone de novo mutation incidence among the two groups. Red silhouettes represent carriers of de novo hot zone mutations. For (**e**) cases, 29 of 103 (28.2%) patients ascertained for an undiagnosed genetic condition, without an inherited genetic diagnosis, had a hot zone de novo mutation as compared with (**d**) controls, for which only 44 of 728 (6.0%) sequenced control trios had a hot zone de novo mutation (Fisher's exact test, $P = 3.0 \times 10^{-10}$; 79% excess observations among cases). Moreover, adding in the layer of information regarding essential gene status (red silhouette with a blue asterisk) further pinpointed toward putative pathogenic mutations because among the control population only 1.9% had a hot zone de novo mutation occurring in an essential gene. This is compared with the 15.5% of cases (Fisher's exact test, $P = 3.4 \times 10^{-8}$; 88% excess observations among cases).

patient as in a previously reported case, *SRCAP*, *KCNT1*, and *NOTCH2* (with 6, 11, and 22 reported variants, respectively) occupy the (apparently) lower end of the allelic heterogeneity spectrum with less than 0.1% of possible single-nucleotide substitutions in those genes currently listed as "pathogenic" for a similar disorder. One hypothesis for this interesting pattern of restricted allelic heterogeneity is that some disease-causing mutations more frequently arise from specific mutagenic mechanisms. Alternatively, it is possible that disease-causing mutations are restrictively distributed in genes because they can only disrupt biological function in limited ways or would otherwise be nonviable.[30] With more diagnostic WES data being generated, we expect a better delineation and understanding of the allelic heterogeneity of disease-associated genes, with single-gene/disease resolution.

It is important to note that there is a difference between perfect controls—ethnically matched and screened for personal and family history of any relevant illness—and controls of convenience, such as the those provided by the EVS (http://evs.gs.washington.edu/EVS/). The number of candidate mutations in the Sheba cohort was higher than that in the Duke cohort (**Supplementary Table S3** online), and this reflects the benefits of having an ethnically matched control population for comparison. The EVS includes subjects selected for specific diseases such as early-onset heart disease and stroke, as well as for extreme phenotypes such as very high or low cholesterol. This must be borne in mind when using controls of convenience for screening out candidate variants. One illustrative example of this is the patient in trio 66 with a homozygous *DPYD* genotype (c.1905+1G>A) that has been reported to cause dihydropyrimidine dehydrogenase (DPD) deficiency (OMIM 274270), an autosomal-recessive disorder of pyrimidine metabolism.[31] Our patient had failure to thrive, global developmental delay, and high urine uracil levels, all consistent with DPD deficiency. However, we also found a homozygous genotype in 1 of 3,027 internal controls not known to have DPD deficiency. Examination of the literature shows that DPD deficiency is characterized by a highly variable phenotype, and some individuals with known pathogenic genotypes can be asymptomatic.[32] This fact, together with the observation that the mutation has already been reported among unrelated patients with DPD deficiency,[31] strongly supports the pathogenic nature of the genotype in our patient despite the occurrence of the genotype in a control.

Another example is the patient in trio 46, who was found to have a de novo nonsense mutation in *ASXL1* (p.Arg404Ter, also a ClinVar pathogenic variant).[28] This same mutation was previously reported in another patient with Bohring–Opitz syndrome.[33] Our (Caucasian) patient had overlapping clinical features with Bohring–Opitz syndrome, including growth failure, developmental delay, microcephaly, strabismus, hypotonia, and seizures; therefore, the de novo *ASXL1* nonsense mutation is highly likely to explain the clinical presentations in our patient. However, this same nonsense mutation (p.Arg404Ter) is observed in two (presumably unrelated) EVS subjects (both of African-American ancestry) who are very unlikely to be affected with Bohring–Opitz syndrome. Because Bohring–Opitz syndrome is an early-onset severe malformation syndrome and its causal mutations are presumably highly penetrant, the presence of p.Arg404Ter in two African-American EVS subjects leads us to consider this Sanger-validated de novo nonsense mutation as a very good candidate rather than a "genetic diagnosis determined." Further studies are required to elucidate the relevance of *ASXL1* loss-of-function variants observed in control population databases.

It is also important to bear in mind quality control differences when using controls sequenced elsewhere. For example, the patient of trio 19 (of Middle Eastern origin) is homozygous for a frameshift mutation in *TECPR2*, which causes autosomal-recessive spastic paraplegia-49 (SPG49; OMIM 615031).[34] The patient had overlapping manifestations with SPG49, including severe hypotonia, gastroesophageal reflux disease, areflexia, intellectual disability, and breathing abnormalities; however, this patient did not have any qualifying genotypes. Given the strong clinical evidence for SPG49, the patient's treating clinician requested that *TECPR2* be screened more liberally than the qualifying genotype criteria. As a result, we identified a homozygous genotype (p.Leu440ArgfsTer19) that has not been previously reported in SPG49 patients. Among our 9,530 controls, homozygosity of the same frameshift variant was found in one EVS subject (a European American). Because the EVS European-American genotypic distribution deviated from Hardy–Weinberg equilibrium (A1A1 = 1/A1R = 5/RR = 3,861, $P = 0.0027$),[35] we asked the Exome Sequencing Project directly about this EVS homozygous genotype and were informed that the homozygous indel genotype was likely to be heterozygous and mistakenly called as homozygous because of the low EVS sequencing coverage at this locus (Qian Yi, personal communication). We report this observation of misgenotyping to advocate for the careful evaluation of putatively pathogenic variants based on all lines of available evidence. We note that as part of this evaluation, we always find it useful to check Hardy–Weinberg equilibrium in control populations as a warning of the possibility of genotyping errors.

We were able to identify the *DPYD* and *ASXL1* genotypes (**Supplementary Table S6** online) by slightly relaxing the criteria to allow up to two control individuals to have the same genotype only when that exact genotype has been reported as pathogenic (**Supplementary Method 1.1** online). We also note that such relaxation of the rules to define "qualifying status" will become increasingly important as the sample size used as reference controls increases.

The use of standing human variation to evaluate genic intolerance is emerging as a critical tool in the interpretation of patient genomes.[12,36,37] We first introduced our "hot zone" bioinformatic signature in the work of Petrovski et al.,[12] in which we integrated gene- and variant-level predictions of pathogenicity to prioritize de novo mutations. Comparing the de novo hot zone mutations in cases versus controls represents a novel tool to understand the genetics of the rare diseases as a whole

and also on an individual basis. For example, we found clear evidence for a statistically significant enrichment of de novo hot zone mutations in rare disease exomes (69% excess; $P = 2.3 \times 10^{-7}$). Among the genes with a de novo hot zone mutation, some occur in diagnostic Mendelian disease genes and others are not disease-associated. This does not mean that all de novo hot zone mutations are pathogenic, but rather that it is clear that we see an enrichment of de novo hot zone mutations when ascertaining for severe rare diseases such as those studied here, and thus this collection of hot zone de novo mutations will harbor some real pathogenic variants. Using additional information, such as essential gene status, we note that among the general population an individual will rarely have a hot zone de novo mutation within an "essential" gene (1.9% of individuals, based on 728 control trios), yet we see the rate among our patients to be 15.5% (88% excess; $P = 3.4 \times 10^{-8}$). Taken together, these clear patterns from the entire collection of different rare diseases not only demonstrate the existence of risk factors but also provide valuable information for localizing pathogenic mutations in individuals. The genes carrying such variants can then be shared with clinical and research centers so that other similar patients with mutations in these genes might be identified and the pathogenicity of the gene might be confirmed.

Conclusion

Despite the increasing clinical use of WES, many of the prerequisites for a successful diagnosis are currently best realized in the research setting. This includes careful consideration of methods to evaluate the pathogenicity of variants. As we showed here, there will be exceptions that require careful interpretation. Furthermore, given the rapid pace of new gene discovery, it is essential to appreciate the need to dynamically reanalyze patient exomes. Finally, it is clear that deployment of bioinformatic prioritization tools provides important pointers toward apparent phenotype expansion of known genes (seven potential examples reported here), as well as pointers toward novel disease genes (four potential examples reported here). All these highlight the important role of research geneticists in implementing diagnostic WES and ultimately call for the collaborative efforts of scientists and clinicians to fully realize the discovery and translational potential of WES.

SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper at http://www.nature.com/gim

ACKNOWLEDGMENTS

We appreciate all the patients, their parents, the clinical research coordinators, and the referring physicians for participating in this study. We thank B. Krueger, J. Bridgers, J. Keebler, Z. Ren, and Q. Wang for bioinformatics support. We thank K. Cronin, S. Shaltz, A. Richards, and N. Walley for excellent technical support. The authors thank the National Heart, Lung, and Blood Institute Grand Opportunity (GO) Exome Sequencing Project and its ongoing studies that produced and provided exome variant calls for comparison: the Lung GO Sequencing Project (HL-102923); the WHI Sequencing Project (HL-102924); the Broad GO Sequencing Project (HL-102925); the Seattle GO Sequencing Project (HL-102926); and the Heart GO Sequencing Project (HL-103010). S.P. is a National Health and Medical Research Council CJ Martin Fellow. We acknowledge the following individuals for the contributions of control samples: W.B. Gallentine, E.L. Heinzen, A.M. Husain, K.N. Linney, M.A. Mikati, R.A. Radtke, and S.R. Sinha; J.P. McEvoy, J. Silver, and M. Silver; D.H. Murdock and The MURDOCK Study Community Registry and Biorepository; G. Cavalleri, N. Delanty, and C. Depondt; J. Burke, C. Hulette, and K. Welsh-Bohmer; J. Milner; J. Hoover-Fong, N.L. Sobreira, and D. Valle; E.J. Holtzman; W.L. Lowe; P. Lugar; S.M. Palmer; Z. Farfel, A. Poduri; M. Hauser; D. Marchuk; D. Koltai Attix, O. Chiba-Falek; E.T. Cirulli, V. Dixon, and J. McEvoy; K. Schmader, S. McDonald, H.K. White, M. Yanamadala, and the Carol Woods and Crosdaile Retirement Communities; R. Gbadegesin and M. Winn; D. Daskalakis; Q. Zhao; A. Holden, and E. Behr; R. Brown; and S. Kerns and H. Oster. The collection of control samples was funded in part by Bryan ADRC NIA P30 AG028377, the Ellison Medical Foundation New Scholar award AG-NS-0441-08, an award from SAIC-Frederick (M11-074), funding from Biogen Idec, NIMH awards RC2MH089915, R01MH097971, R01MH099216, and K01MH098126, the Epi4K Gene Discovery in Epilepsy study (NINDS U01-NS077303), the Epilepsy Genome/Phenome Project (EPGP - NINDS U01-NS053998), and the Center for HIV/AIDS Vaccine Immunology ("CHAVI") under a grant from the National Institute of Allergy and Infectious Diseases, National Institutes of Health (UO1AIO67854). E.K.R is funded by a predoctoral grant from the Epilepsy Foundation and Jo Rae Wright Fellowship for outstanding women in science (Duke University). Additional control samples were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through accession number phs000473. Samples used for data analysis were provided by the Swedish Cohort Collection supported by the NIMH grant R01MH077139, the Sylvan C. Herman Foundation, the Stanley Medical Research Institute, and the Swedish Research Council (grants 2009–4959 and 2011–4659). Support for the exome sequencing was provided by the NIMH Grand Opportunity grant RCMH089905, the Sylvan C. Herman Foundation, a grant from the Stanley Medical Research Institute, and multiple gifts to the Stanley Center for Psychiatric Research at the Broad Institute of MIT and Harvard.

DISCLOSURE

Partial funding for this study was provided by UCB Celltech, including salary support for D.B.G., X.Z., P.X., E.K.R., K.S., R.C.S., Y.-H.J, and V.S. D.M. is an employee of UCB. The other authors declare no conflict of interest.

REFERENCES

1. Yang Y, Muzny DM, Reid JG, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med* 2013;369:1502–1511.
2. Need AC, Shashi V, Hitomi Y, et al. Clinical application of exome sequencing in undiagnosed genetic conditions. *J Med Genet* 2012;49:353–361.
3. Dixon-Salazar TJ, Silhavy JL, Udpa N, et al. Exome sequencing can improve diagnosis and alter patient management. *Sci Transl Med* 2012;4:138ra78.

4. de Ligt J, Willemsen MH, van Bon BW, et al. Diagnostic exome sequencing in persons with severe intellectual disability. *N Engl J Med* 2012;367:1921–1929.
5. Enns GM, Shashi V, Bainbridge M, et al.; FORGE Canada Consortium. Mutations in NGLY1 cause an inherited disorder of the endoplasmic reticulum-associated degradation pathway. *Genet Med* 2014;16:751–758.
6. Might M, Wilsey M. The shifting model in clinical diagnostics: how next-generation sequencing and families are altering the way rare diseases are discovered, studied, and treated. *Genet Med* 2014;16:736–737.
7. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–1760.
8. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;43:491–498.
9. Cingolani P, Platts A, Wang le L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 2012;6:80–92.
10. Li H, Handsaker B, Wysoker A, et al.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
11. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol* 2011;29:24–26.
12. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* 2013;9:e1003709.
13. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;7:248–249.
14. Sanders SJ, Murtha MT, Gupta AR, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 2012;485:237–241.
15. O'Roak BJ, Vives L, Girirajan S, et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 2012;485:246–250.
16. Iossifov I, Ronemus M, Levy D, et al. De novo gene disruptions in children on the autistic spectrum. *Neuron* 2012;74:285–299.
17. Gulsuner S, Walsh T, Watts AC, et al.; Consortium on the Genetics of Schizophrenia (COGS); PAARTNERS Study Group. Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* 2013;154:518–529.
18. Xu B, Ionita-Laza I, Roos JL, et al. De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat Genet* 2012;44:1365–1369.
19. Rauch A, Wieczorek D, Graf E, et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* 2012;380:1674–1682.
20. Georgi B, Voight BF, Bućan M. From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. *PLoS Genet* 2013;9:e1003484.
21. Blake JA, Bult CJ, Eppig JT, Kadin JA, Richardson JE; Mouse Genome Database Group. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res* 2014;42(Database issue):D810–D817.
22. Orhan G, Bock M, Schepers D, et al. Dominant-negative effects of KCNQ2 mutations are associated with epileptic encephalopathy. *Ann Neurol* 2014;75:382–394.
23. Milligan CJ, Li M, Gazina EV, et al. KCNT1 gain of function in 2 epilepsy phenotypes is reversed by quinidine. *Ann Neurol* 2014;75:581–590.
24. Zhu X, Need AC, Petrovski S, Goldstein DB. One gene, many neuropsychiatric disorders: lessons from Mendelian diseases. *Nat Neurosci* 2014;17:773–781.
25. Epi4K Consortium; Epilepsy Phenome/Genome Project; De novo mutations in epileptic encephalopathies. *Nature* 2013;501:217–221.
26. Du X, An Y, Yu L, et al. A genomic copy number variant analysis implicates the MBD5 and HNRNPU genes in Chinese children with infantile spasms and expands the clinical spectrum of 2q23.1 deletion. *BMC Med Genet* 2014;15:62.
27. Hood RL, Lines MA, Nikkel SM, et al.; FORGE Canada Consortium. Mutations in SRCAP, encoding SNF2-related CREBBP activator protein, cause Floating-Harbor syndrome. *Am J Hum Genet* 2012;90:308–313.
28. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 2014;42(Database issue):D980–D985.
29. Stenson PD, Ball EV, Mort M, et al. Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* 2003;21:577–581.
30. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. *Proc Natl Acad Sci USA* 2007;104:8685–8690.
31. Vreken P, Van Kuilenburg AB, Meinsma R, et al. A point mutation in an invariant splice donor site leads to exon skipping in two unrelated Dutch patients with dihydropyrimidine dehydrogenase deficiency. *J Inherit Metab Dis* 1996;19:645–654.
32. Van Kuilenburg AB, Vreken P, Abeling NG, et al. Genotype and phenotype in patients with dihydropyrimidine dehydrogenase deficiency. *Hum Genet* 1999;104:1–9.
33. Hoischen A, van Bon BW, Rodríguez-Santiago B, et al. De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. *Nat Genet* 2011;43:729–731.
34. Oz-Levi D, Ben-Zeev B, Ruzzo EK, et al. Mutation in TECPR2 reveals a role for autophagy in hereditary spastic paraparesis. *Am J Hum Genet* 2012;91:1065–1072.
35. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* 2005;76:887–893.
36. Pan S, Caleshu CA, Dunn KE, et al. Cardiac structural and sarcomere genes associated with cardiomyopathy exhibit marked intolerance of genetic variation. *Circ Cardiovasc Genet* 2012;5:602–610.
37. Samocha KE, Robinson EB, Sanders SJ, et al. A framework for the interpretation of de novo mutation in human disease. *Nat Genet* 2014;46:944–950.
38. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* 2010;6:e1001025.
39. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003;31:3812–3814.
40. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;46:310–315.