nature genetics

Factors influencing success of clinical genome sequencing across a broad spectrum of disorders

Jenny C Taylor[1,2,43], Hilary C Martin[2,43], Stefano Lise[2], John Broxholme[2], Jean-Baptiste Cazier[3], Andy Rimmer[2], Alexander Kanapin[2], Gerton Lunter[2], Simon Fiddy[2], Chris Allan[2], A Radu Aricescu[2], Moustafa Attar[2], Christian Babbs[4], Jennifer Becq[5], David Beeson[6], Celeste Bento[7], Patricia Bignell[8], Edward Blair[9], Veronica J Buckle[4], Katherine Bull[2,10], Ondrej Cais[11], Holger Cario[12], Helen Chapel[13], Richard R Copley[1,2], Richard Cornall[10], Jude Craft[1,2], Karin Dahan[14,15], Emma E Davenport[2], Calliope Dendrou[16], Olivier Devuyst[17], Aimée L Fenwick[18], Jonathan Flint[2], Lars Fugger[16], Rodney D Gilbert[19], Anne Goriely[18], Angie Green[2], Ingo H Greger[11], Russell Grocock[5], Anja V Gruszczyk[18], Robert Hastings[20], Edouard Hatton[2], Doug Higgs[4], Adrian Hill[2,21], Chris Holmes[2,22], Malcolm Howard[1,2], Linda Hughes[2], Peter Humburg[2], David Johnson[23], Fredrik Karpe[24], Zoya Kingsbury[5], Usha Kini[9], Julian C Knight[2], Jonathan Krohn[2], Sarah Lamble[2], Craig Langman[25], Lorne Lonie[2], Joshua Luck[18], Davis McCarthy[2], Simon J McGowan[18], Mary Frances McMullin[26], Kerry A Miller[18], Lisa Murray[5], Andrea H Németh[27], M Andrew Nesbit[28], David Nutt[29], Elizabeth Ormondroyd[20], Annette Bang Oturai[30], Alistair Pagnamenta[1,2], Smita Y Patel[13], Melanie Percy[31], Nayia Petousi[32], Paolo Piazza[2], Sian E Piret[28], Guadalupe Polanco-Echeverry[2], Niko Popitsch[1,2], Fiona Powrie[33], Chris Pugh[32], Lynn Quek[4], Peter A Robbins[34], Kathryn Robson[4], Alexandra Russo[35], Natasha Sahgal[2], Pauline A van Schouwenburg[13], Anna Schuh[1,36], Earl Silverman[37], Alison Simmons[16,33], Per Soelberg Sørensen[30], Elizabeth Sweeney[38], John Taylor[1,39], Rajesh V Thakker[28], Ian Tomlinson[1,2], Amy Trebes[2], Stephen R F Twigg[18], Holm H Uhlig[32], Paresh Vyas[4], Tim Vyse[40], Steven A Wall[23], Hugh Watkins[20], Michael P Whyte[41], Lorna Witty[2], Ben Wright[2], Chris Yau[2], David Buck[2], Sean Humphray[5], Peter J Ratcliffe[32], John I Bell[42], Andrew O M Wilkie[18], David Bentley[5], Peter Donnelly[2,22,44] & Gilean McVean[2,44]

To assess factors influencing the success of whole-genome sequencing for mainstream clinical diagnosis, we sequenced 217 individuals from 156 independent cases or families across a broad spectrum of disorders in whom previous screening had identified no pathogenic variants. We quantified the number of candidate variants identified using different strategies for variant calling, filtering, annotation and prioritization. We found that jointly calling variants across samples, filtering against both local and external databases, deploying multiple annotation tools and using familial transmission above biological plausibility contributed to accuracy. Overall, we identified disease-causing variants in 21% of cases, with the proportion increasing to 34% (23/68) for mendelian disorders and 57% (8/14) in family trios. We also discovered 32 potentially clinically actionable variants in 18 genes unrelated to the referral disorder, although only 4 were ultimately considered reportable. Our results demonstrate the value of genome sequencing for routine clinical diagnosis but also highlight many outstanding challenges.

The mainstream application of whole-genome sequencing in clinical diagnosis holds much promise. In contrast to existing genetic tools, such as targeted gene sequencing, array comparative genomic hybridization (aCGH) and exome sequencing[1–5], only whole-genome sequencing can characterize all types of genetic variation in all parts of the genome. Such completeness, coupled with efforts to chart the distribution of genetic variation in populations[6,7], will enable the identification of pathogenic variants and hence influence diagnosis, genetic counseling and treatment.

Nevertheless, clinical adoption of whole-genome sequencing faces many challenges, including cost, speed of delivery, sensitivity, specificity and heterogeneity in variant detection, ambiguities and errors in variant annotation, a substantial informatics burden and the difficulties posed by incidental findings[8,9]. Consequently, although technological improvements to enhance speed in critical situations, such as neonatal intensive care[10], and detailed evaluations of whole-genome and whole-exome sequencing data in specific disorders[11] are providing the opportunity for the wider use of this approach, its reach

A full list of affiliations appears at the end of the paper.

Received 8 July 2014; accepted 22 April 2015; published online 18 May 2015; doi:10.1038/ng.3304

into the clinic is thus far limited[12]. For whole-genome sequencing to be adopted as a routine clinical platform, it would be necessary to demonstrate its diagnostic yield for patients with likely genetic disorders identified by clinicians across a broad range of medical specialties, within a hospital setting. Furthermore, the challenges of reliably identifying and validating potential pathogenic variants at scale across such a disease spectrum would need to be met.

To address these challenges, we undertook the WGS500 program to sequence 500 patient genomes from diverse genetic disorders referred by a range of medical specialists. For all disorders, study leaders had access to additional samples and/or could follow up with functional studies for validation. Some results from this study have already been published[13–19]. Here we report an overview of the results from the mendelian and immunological disorders, representing 156 independent individual cases or families, selected because a strong genetic component was suspected (on the basis of family history, early onset or disorder severity) but previous genetic screening had failed to identify any pathogenic variants (**Fig. 1**). The disorders varied substantially in the number of independent cases recruited, the availability of additional family members and the likely disease model. Here we identify and quantify the effect on success of factors relating to the genetic architecture of a disease, experimental design and analytical strategy.

RESULTS

Variant calling, filtering and annotation

Individuals were sequenced to an average of 31.8× (range of 22.7–60.8×) such that, on average, 82.7% of the genome (88.2% of the exome) was covered by at least 20× (**Supplementary Fig. 1**). We found no significant correlation between sequencing coverage and diagnostic success (Pearson $r = -0.1$; $P = 0.13$), indicating that, at this depth, fluctuations in coverage for whole-genome sequencing have a minor role in determining success for germline disorders. For the few samples with low levels of contaminating DNA (**Supplementary Fig. 1**), we took additional care in the interpretation of candidate pathogenic variants rather than returning to the patient for additional material; one individual with substantial contamination was excluded.

We processed all samples with the same pipelines for sequencing, variant calling and annotation (Online Methods). The concordance between the whole-genome sequencing data and genotypes from SNP arrays was over 99.9% at heterozygous sites (**Supplementary Fig. 2** and **Supplementary Tables 1** and **2**). Our pipeline included two key steps. First, we used a two-stage variant calling procedure with an initial round of independent calling followed by a second round that revisited the evidence in each individual for any variant called across all samples. This approach improves genotype accuracy by, for example, using strong evidence for a variant in a child to enhance support for the same variant in a parent (and vice versa). Joint calling substantially increased the accuracy of *de novo* mutation detection in families. For example, the number of candidate coding *de novo* mutations was reduced from a mean of 32.1 after independent variant calling (filtering against the 1000 Genomes Project and National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP) databases) to 2.6 after joint calling of the parents and proband (**Supplementary Table 3**).

The second key step was that, when identifying likely pathogenic variants, in addition to filtering against external data sources, we also filtered against other WGS500 samples. For example, when filtering against external data sources only, individuals had an average of 80.8 rare or new (frequency < 0.5%) homozygous coding variants but had only 1.5 if variants present above this frequency in other WGS500 samples were also excluded (**Supplementary Table 4**). Using control samples sequenced with the same technology and processed through the same pipeline reduced the impact of systematic differences between our studies and others with respect to coverage, sequencing technology, the experimental protocol and data processing (**Supplementary Fig. 3** and **Supplementary Table 3**).

Finally, we found that the choice of transcript set and annotation software can affect variant annotations[20]. Comparison of annotations using the RefSeq and Ensembl transcript sets found only 44% agreement for putative loss-of-function variants. Similarly, we found agreement of only 66% for loss-of-function annotations and 87% for all exonic annotations between the ANNOVAR[21] and Ensembl Variant Effect Predictor (VEP)[22] tools. In both comparisons, the greatest discrepancy was for splicing annotations (agreement of 25% between transcript sets and 57% between software tools). Such heterogeneity in how variants are annotated can substantially reduce the efficacy of whole-genome sequencing in clinical analysis. We therefore used multiple annotation approaches to identify candidate variants.

Evaluating the biological candidacy of variants

To identify candidate disease-causing variants, we used a combination of predicted functional impact, frequency in the population, transmission within a family (where appropriate) and, when multiple independent cases were available, statistical evidence for association (Online Methods). Because most genes harbor large numbers of rare variants[6,23]—many of which are absent from existing databases and affect the protein produced, but only a fraction of which may influence disease risk—care has to be taken in interpreting newly identified variants in known disease-associated genes. To assess the burden of such 'variants of unknown significance' across a range of disorders, we defined candidate genes for early-onset epilepsy (EOE), X-linked mental retardation (XLMR) and craniosynostosis (CRS). For EOE, we used a semiautomated approach based on a three-tiered system taking into account medical genetics and biological information (Online Methods and **Supplementary Table 5**). Tier 1 comprised the set of known genes for the disorder (from the Human Gene Mutation Database, HGMD[24]), tier 2 added genes known for related disorders (from HGMD) or whose products interact directly with those for tier 1 genes (from the Mammalian Protein-Protein Interaction Database, MIPS[25]), and tier 3 added genes in relevant biological pathways (from HGMD and the Gene Ontology (GO) database). For XLMR, we only examined tier 1 genes. For CRS, we used lists generated by expert curation. For individuals with the disorder, additional family members were used to identify the most likely pathogenic variants (**Supplementary Table 6**).

For each disorder, we found multiple unaffected individuals in WGS500 with variants in the candidate genes for XLMR, CRS and EOE that would be interpreted as potentially pathogenic had those individuals presented with the disorder in question. Within the 10 known genes for EOE (tier 1), we found that 3 of 216 individuals (1.3% of the sample) carried a new heterozygous candidate variant and 1 (0.5%) carried a rare homozygous candidate variant (**Fig. 2a**); none of these individuals had epilepsy. As the strength of gene candidacy decreased, the number of putative pathogenic variants increased; 36% of individuals carried at least one heterozygous candidate variant among the tier 1 genes or the additional 82 genes implicated in milder forms of epilepsy (tier 2), and 96% of individuals carried one such variant in a tier 1 or 2 gene or in one of the 771 genes involved in brain development or function (tier 3). The proportions for homozygous candidate variants were 3% and 17% for tier 1 and 2 genes and tier

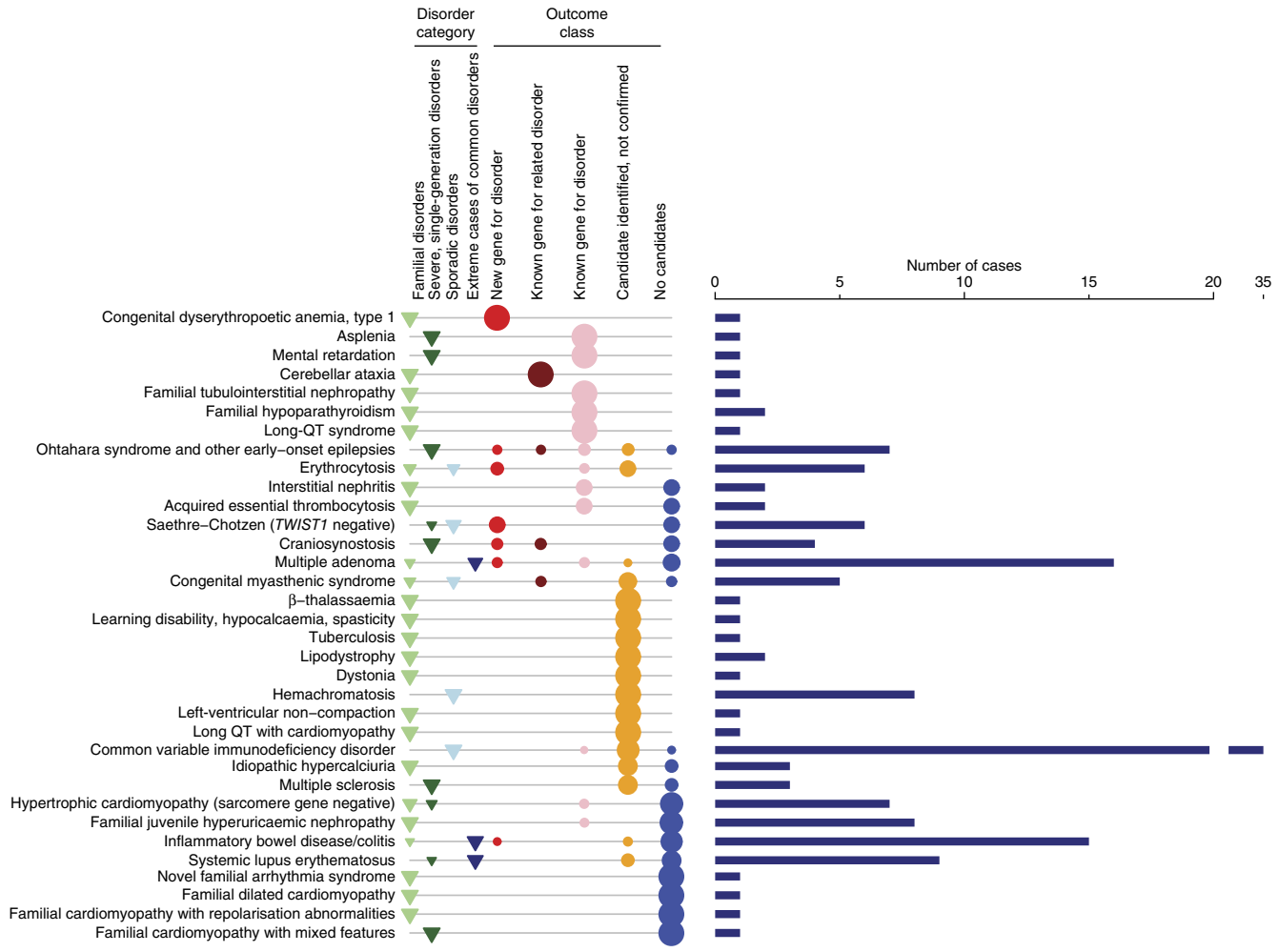

Figure 1 Overview of projects and results. For each disorder, the number of independent cases studied (bars) is shown alongside information about the nature of the disorder (triangles): disorders were classified as familial disorders (category 1; light green), severe single-generation disorders suspected to be caused by *de novo* or recessive mutations (category 2; dark green), unrelated sporadic disorders (category 3; light blue) and extreme cases of common complex diseases (category 4; dark blue). The proportion of cases with each outcome class (A–E) is also shown, by circles, with proportion indicated by point size (Online Methods): outcomes were classified by the identification of a pathogenic variant in a new gene for the disorder (class A; red), a pathogenic variant in a gene for a related disorder (class B; brown), a pathogenic variant in a known gene for the disorder (class C; pink), a candidate pathogenic variant with validation studies underway (class D; orange), and no single candidate variant or negative results for validation of top candidate(s) (class E; blue). Disorders are ranked by the fraction of cases with confirmed pathogenic variants (classes A–C).

1–3 genes, respectively. We found no enrichment for either heterozygous or homozygous candidate variants in tier 1 and 2 genes among the six patients with EOE (**Supplementary Table 7**), and only two of ten tier 1 and 2 variants found in EOE samples were thought to be pathogenic on the basis of family information; for homozygous variants in tier 1–3 genes, the corresponding figure was one of three (**Supplementary Table 6**)[15].

We found similar results for genes in other disorders. For CRS, 57 of 216 (26%) samples carried at least one new heterozygous coding variant in the 38 expert-curated known causative genes, although no sample had any rare homozygous coding variants (**Supplementary Fig. 4**). Five CRS samples carried tier 1 or 2 variants, but none were thought to be pathogenic as they were not of *de novo* origin. For XLMR, the effect was striking: 30 of 109 (28%) male samples carried at least one previously unreported missense variant at a conserved residue within the 83 known genes for XLMR (**Fig. 2b**). In only two of these cases (two brothers with mental retardation) was the variant thought to be pathogenic.

We also investigated the burden of potentially pathogenic regulatory variants, focusing on conserved positions in regulatory regions defined by the Ensembl Regulatory Build that are less than 50 kb away from candidate genes (**Supplementary Fig. 5**). The mean number of new heterozygous variants per individual was 203 (s.d. = 102; range = 102–614), more than twice as many as the equivalent number of new coding variants (mean = 75; **Supplementary Fig. 3**), although we note that this number is inflated because there are fewer control individuals in publicly available data sets for regulatory variants than in those for exonic variants. Many individuals had previously unreported or rare variants at conserved sites in regulatory regions close to candidate genes for EOE and CRS (**Supplementary Fig. 6**). Moreover, in samples from patients with the disorder, there were typically multiple potential regulatory variants that were consistent with a plausible inheritance model, although none of these were considered likely to be pathogenic because stronger candidate variants were present (**Supplementary Table 6**).

These results demonstrate that the combined use of gene candidacy, predicted functional consequence, variant frequency and

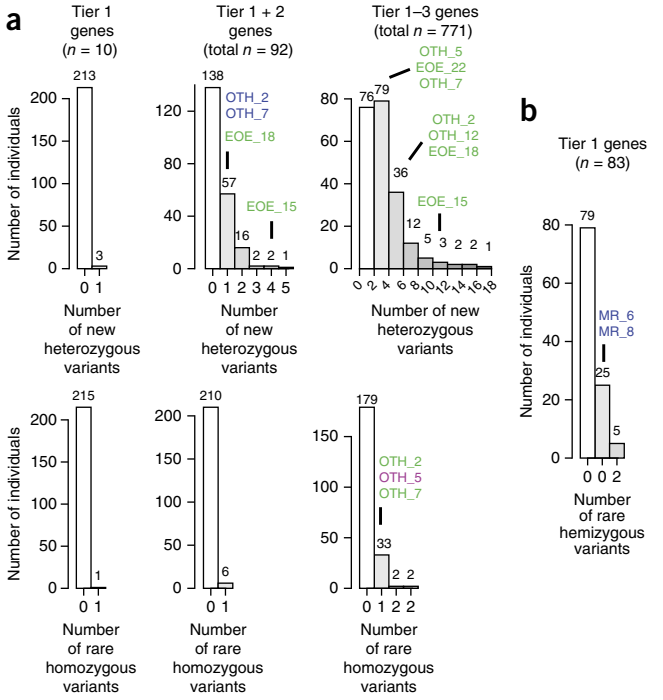

Figure 2 The burden of variants of unknown significance. (**a**) Histograms of the number of previously unreported coding variants at conserved positions in different sets of candidate genes (tiers 1, 1 + 2, and 1–3) for EOE, under different inheritance models, across 216 WGS500 samples. (**b**) Histogram of the number of previously unreported coding variants at conserved positions in known XLMR-associated genes for the 99 male WGS500 samples. The candidate genes were chosen by high-throughput searches (Online Methods). Sample identifiers indicate individuals with the disorder in question. Sample names in green text indicate that the identified variant is not likely to be pathogenic (as it does not fit a plausible inheritance model or is less functionally compelling than another candidate); sample names in blue text indicate that the identified variant is thought to be causal (**Supplementary Table 6**). OTH, Ohtahara syndrome; EOE, nonsyndromic early-onset epilepsy; MR, mental retardation. See **Supplementary Figure 4** for the analysis of CRS.

evolutionary conservation, although these are widely used filters within pipelines for identifying pathogenic candidate variants, will not by themselves differentiate between pathogenic and non-pathogenic variants. Naive application of such rules will lead to a high rate of false positive diagnosis, even in rare disorders with mutations occurring in limited numbers of known genes. Moreover, focusing only on candidate genes will lead to a high false negative rate; in the eight families with EOE, CRS or XLMR for which a strong candidate (class A–C) pathogenic variant has been identified (**Supplementary Table 6**), only four of these variants were in candidate genes found using automated database searches (tiers 1–3). In this study, as in others, additional evidence, such as functional data, familial transmission, *de novo* status and screening of other patients, was needed to establish pathogenicity.

Overview of the findings

In 33 of the 156 cases (21%), we identified at least one variant with a high level of evidence of pathogenicity (classes A–C as described in the Online Methods; **Fig. 1**, **Table 1** and **Supplementary Tables 8** and **9**). These comprised 5 nonsense variants, 15 missense variants, 3 noncoding variants, 2 frameshift variants, 1 in-frame indel, 5 variants that disrupted splicing and 2 compound heterozygotes, each with 1 missense and 1 either nonsense or splicing variant (and, additionally, 1 variant that was reported independently of WGS500). Altogether, we identified 12 cases with variants in new genes for which we found additional compelling genetic and/or functional evidence of pathogenicity (class A), 4 cases with variants in genes known for other phenotypes but not for the disorder in question, supported by additional genetic and/or functional evidence (class B), and 7 cases with variants in genes already known for that phenotype (class C). This rate of success is comparable to those in recent exome sequencing studies for various disorders[3,5,8,26]. Below, we describe the range of the findings and some of the outstanding challenges identified.

Variants missed by previous genetic testing

We identified four cases where a candidate variant lay within a gene that had previously been screened by a clinical or research genetics laboratory (UK or overseas) but had been missed. These variants were in *UMOD* in familial juvenile hyperuricemic nephropathy (FJHN; **Supplementary Fig. 7**), in *KCNQ1* in long-QT syndrome (LTS), and in *APC* and *MSH6* in multiple adenoma. The rate of false negative results from Sanger sequencing is likely to vary considerably between genes and types of variant. Nevertheless, across the samples studied here, a relatively small fraction (2.5%) of cases resulted from variants in genes with false negative test results in standard clinical genetics testing.

Challenges in establishing pathogenicity

For several disorders, likely pathogenic variants were identified in genes not previously reported for the corresponding conditions or related phenotypes. When additional variants of major coding consequence were found in screens of other cases (and not controls), the evidence for pathogenicity was considered strong, including for *POLE* and *POLD1* in multiple adenoma and colorectal cancer[19], *TCF12* in Saethre-Chotzen–like syndrome[16], *ALG2* for congenital myasthenic syndrome[17] and *C15orf41* in congenital dyserythropoetic anemia, type 1 (ref. 14). In some cases, mouse models provided supportive evidence (for example, confirming the role of *SPTBN2* in cerebellar ataxia[18]) and/or functional work demonstrated that the variant affected protein function (for example, a *de novo* mutation in *KCNT1* found in a patient with Ohtahara syndrome was shown to affect potassium channel activity[15]).

In six cases, likely pathogenic variants were identified in genes where variants cause disorders with related phenotypes. For example, a *de novo* mutation that disrupts *CBL* splicing (NM_005188; exon 9; c.1228–1G>A) was identified in a patient with severe epilepsy, microcephaly and developmental delay[15]. Cbl is a ubiquitin ligase that regulates the Ras-MAPK (mitogen-activated protein kinase) pathway[27], and heterozygous missense variants in *CBL* cause facial, cutaneous and cardiac abnormalities, hypotonia and developmental delay[28,29], as well as microcephaly and a predisposition to juvenile myelomonocytic leukemia[30,31]. However, although our patient had unusual cutaneous and cardiac features, these were not typical of neuro-cardio-facial-cutaneous (NCFC) syndrome, and review by clinicians did not alter the original diagnosis. *CBL* variants have previously been noted for their variable phenotypes and incomplete penetrance[31]. Thus, the *CBL* variant is a strong candidate, but no other likely pathogenic variants in *CBL* were identified in a panel of over 500 other patients with epilepsy[15].

The difficulties in establishing pathogenicity are also illustrated by a *de novo* missense mutation (NM_031407; c.329G>A; p.Arg110Gln) in *HUWE1* identified in a girl with CRS and learning difficulties (**Fig. 3a** and **Supplementary Note**). Mutations in *HUWE1* have been reported to cause XLMR and macrocephaly[32–34], although not previously CRS. The

Table 1 Summary of conditions for which pathogenic variants (class A–C) were identified

Disorder	Project category	Outcome class	Gene[a]	Coding consequence	Inheritance (zygosity)[b]	Variant
Acquired essential thrombocytosis	1.1	C	*THPO*	Splicing	D (het)	NM_001177598: c.13+1G>C
Asplenia	2.2	C	*RPSA*[c]	Splicing	D (het)	NM_002295.4: c.–34+5G>C
Cerebellar ataxia	1.2	B	*SPTBN2*	Nonsense	AR (hom)	NM_006946: c.1881G>A; p.Cys627*
Common variable immunodeficiency disorder	3	C	[d]	Missense	[d]	[d]
Congenital dyserythropoietic anemia, type 1	1.2	A	*C15ORF41*	Missense	AR (hom)	NM_001130010: c.533T>A; p.Leu178Gln
Congenital myasthenic syndrome	3	B	*ALG2*	Missense	AR (hom)	NM_033087: c.203T>G; p.Val68Gly
Craniosynostosis	2.1	A	*ZIC1*	Nonsense	DN (het)	NM_003412.3: c.1163C>A; p.Ser388*
	2.1	B	*HUWE1*	Missense	DN (het)	NM_031407.6: c.329G>A; p.Arg110Gln
Erythrocytosis	1.1	A	*EPO*	Noncoding	D (het)	NM_000799.2: c.–136G>A
	1.1	A	*EPO*	Noncoding	D (het)	NM_000799.2: c.–136G>A
	3	C	*BPGM*	Missense	D (het)	NM_001724: c.269G>A; p.Arg90His
Familial hypoparathyroidism	1.3	C	*SOX3*	Noncoding	XL (hemi)	Deletion of chr. X: 139,502,946–139,504,327, 1.5 kb downstream of *SOX3*
	1.1	C	*CASR*	Missense	D (het)	NM_000388: c.2299G>C; p.Glu767Gln
Familial juvenile hyperuricemic nephropathy	1.4	C	*UMOD*	Missense	D (het)	NM_001008389: c.410G>A; p.Cys137Tyr
Familial tubulointerstitial nephropathy	1.1	C	*UMOD*	Missense (in-frame insertion/ deletion)	D (het)	NM_001008389: c.279_289del; p.93_97del NM_001008389: c.278_279insCCGCCTCC; p.Val93fs
Hypertrophic cardiomyopathy (sarcomere gene-negative)	1.1	C	*MYBPC3*	Nonsense	[e]	NM_000256: c.1303C>T; p.Gln435*
Inflammatory bowel syndrome/colitis	4	A	[d]	Missense	[d]	[d]
Interstitial nephritis	1.4	C	*MUC1*[c]	–	D (het)	[d]
Long-QT syndrome	1.1	C	*KCNQ1*	Missense	D (het)	NM_000218: c.1195_1196insC; p.Ala399fs
Mental retardation	2.1	C	*GRIA3*	Missense	XL (hemi)	[d]
Ohtahara syndrome and other early-onset epilepsies	2.1	A	*PIGQ*	Splicing	SR (hom)	NM_004204: c.690-2A>G
	2.1	B	*KCNT1*	Missense	UPID (hom)	NM_020822: c.2896G>A; p.Ala966Thr
	2.1	C	*KCNQ2*	Missense	DN (het)	NM_004518: c.827C>T; p.Thr276Ile
	2.1	C	*SCN2A*	Missense	DN (het)	NM_001040143: c.5558A>G; p.His1853Arg
Multiple adenoma	1.1	A	*POLD1*	Missense	D (het)	NM_002691: c.1433G>A; p.Ser478Asn
	1.1	A	*POLD1*	Missense	D (het)	NM_002691: c.1433G>A; p.Ser478Asn
	4	A	*POLE*	Missense	D (het)	NM_006231: c.1270C>G; p.Leu424Val
	4	C	*MSH6*	Missense and nonsense	CR (het; het)	NM_000179: c.2315G>A; p.Arg772Gln
	4	C	*BMPR1A*	Frameshift	AR (hom)	NM_004329.2: c.142_143insT; p.Thr49Asnfs*22
	4	C	*APC*	Splicing	D (het)	NM_001127511: c.251–2A>G
Saethre-Chotzen syndrome (*TWIST1* negative)	3	A	*TCF12*	Nonsense	DN (het)	NM_207037.1: c.1283T>G; p.Leu428*
	3	A	*TCF12*	Splicing	DN (het)	NM_207037.1: c.1035+3G>C
	2.1	A	*CDC45*	Synonymous (splicing) and missense	CR (het; het)	NM_001178010.2: c.318C>T; p.Val106 = NM_001178010.2: c.773A>G; p.Asp258Gly

[a]Each row represents a separate case or family; if the same gene is reported in two rows, this signifies that the gene is thought to be pathogenic in both cases. Some genes have two mutations in the same affected individual, likely representing compound heterozygous inheritance, which is indicated in the "Inheritance" column. [b]D (het), dominant inheritance with affected individual(s) heterozygous; AR (hom), autosomal recessive inheritance with affected individual(s) homozygous; DN (het), *de novo* variant with affected individual(s) heterozygous; XL (hemi), X-linked recessive inheritance with affected male(s) hemizygous and affected female(s) homozygous; UPD (hom), uniparental isodisomy inheritance with affected individual(s) homozygous; CR (het; het), compound recessive inheritance with affected individual(s) heterozygous for two different variants in the same gene. [c]Causal variant discovered independently of WGS500. [d]Details will be reported in an independent publication. [e]Form of inheritance not clear. See **Supplementary Table 8**.

mutation affects a highly conserved residue in a domain of unknown function (DUF908; **Supplementary Fig. 8**). The gene spans 154,641 bp and comprises 84 exons; because of the extensive heterogeneity in CRS, the contribution of this gene to disease is likely to be low. Thus, it was not surprising that no other *HUWE1* mutations were found in a cohort of 47 unrelated cases with complex CRS. The mutation originated on the paternal X chromosome (**Fig. 3b** and **Supplementary Fig. 8**). Unexpectedly, cells from the patient showed preferential inactivation of the maternally inherited, wild-type X chromosome (**Fig. 3c**); consistent with these two observations, only the mutant allele was expressed in the tissues available for analysis (fibroblasts and transformed lymphoblasts) (**Fig. 3d**). Seven other X-linked *de novo* point mutations (three in genes: a 5′ UTR change in *CCDC160* and intronic changes in *FRMPD4* and *IGSF1*) were identified in the same individual (**Fig. 3e**), although none were considered pathogenic. The finding of a substitution at a highly conserved residue in a known XLMR-related gene, in combination with exclusive expression of the mutant allele, suggested that this mutation contributed at least to the learning difficulties in this child but also suggested that this is a highly unusual case, and it was hence challenging to establish true pathogenicity. Recently, however, we identified, using whole-exome sequencing, a different *de novo* hemizygous mutation altering the same amino acid of HUWE1 (c.328C>T; p.Arg110Trp) in a boy presenting with metopic CRS, moderate to severe learning disability and other dysmorphic features, supporting the evidence for causality.

Candidates for pathogenic regulatory variants

Strong candidate pathogenic variants were detected outside the coding fraction of the genome in two conditions. The same variant at a

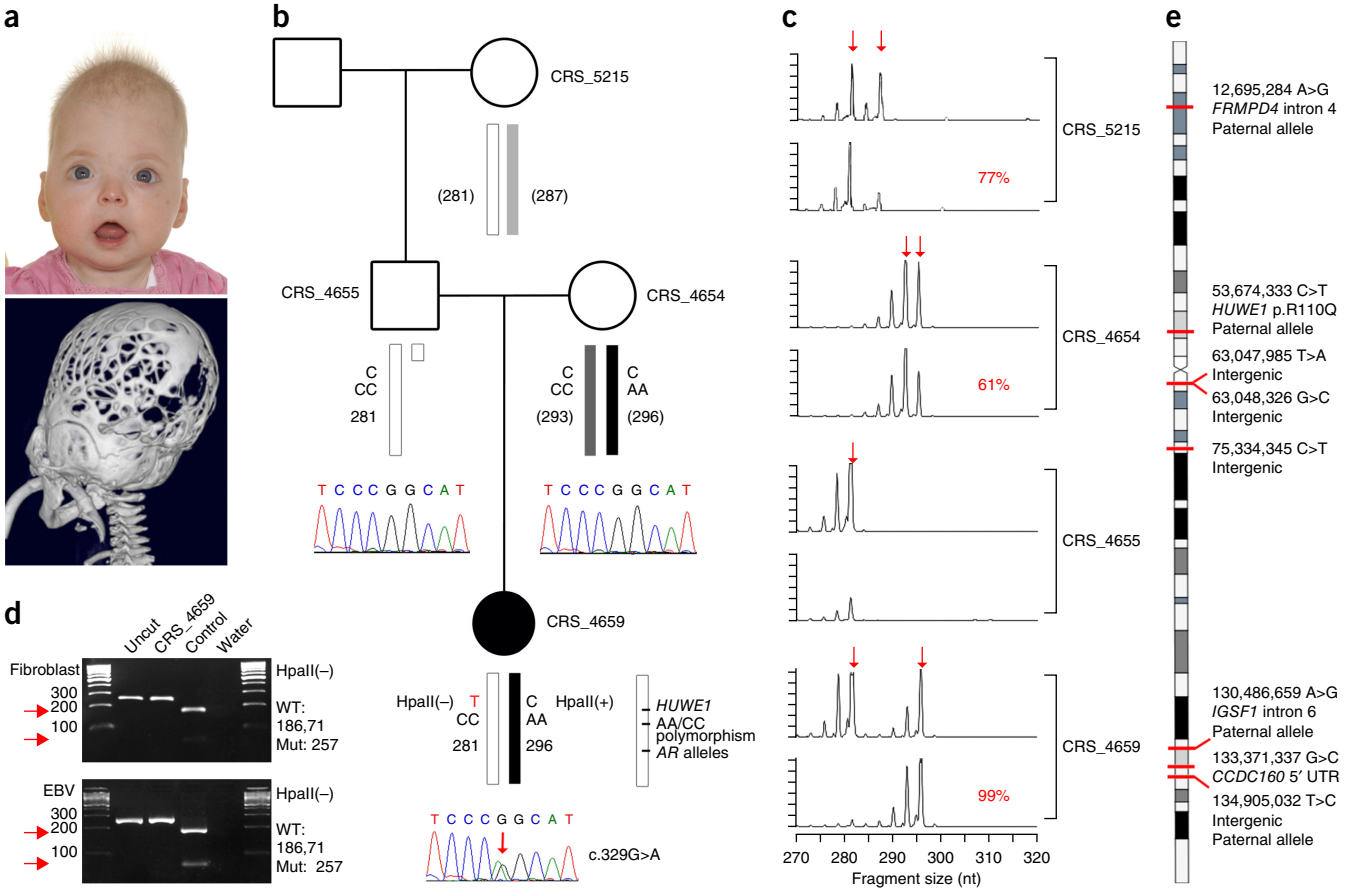

Figure 3 Identification of a *de novo* mutation in *HUWE1* associated with severe CRS. (**a**) Top, the proband (CRS_4659; female, aged 6 months) presented with an abnormal skull shape; informed consent to publish this photograph was obtained from the family. Bottom, the three-dimensional computed tomography (CT) scan at age 5 months shows multisuture synostosis with multiple craniolacunae. (**b**) Family pedigree showing dideoxy sequencing chromatograms with the *de novo* G>A mutation of the X-linked *HUWE1* gene in the proband (red arrow). Schematic X chromosomes are annotated, from top to bottom, with the *HUWE1* alleles, the haplotype for AA/CC polymorphisms located 1.15 kb away from the mutation and used to deduce paternal origin, and androgen receptor (*AR*) trinucleotide-repeat allele size (allele sizes in CRS_4654 and CRS_5215 are in parentheses to emphasize that the phase is unknown relative to other parts of the two X chromosomes). Note that the *HUWE1* mutation abolishes an HpaII restriction site. (**c**) Analysis of X-chromosome inactivation in whole-blood samples at the *AR* locus. For each individual, *AR* alleles are indicated by arrows in the upper panel, and the lower panel shows the proportions of methylated alleles and percentage representation of the more highly inactivated X chromosome. (**d**) Exclusive expression of cDNA from the mutant *HUWE1* allele in both fibroblast and Epstein-Barr virus (EBV)-transformed lymphoblastoid cells from the proband. Arrows highlight the absence of expression of the normal allele in both cell types. Product sizes (bp) from different alleles are shown on the right. WT, wild type; Mut, mutant. (**e**) X-chromosome ideogram showing eight *de novo* mutations identified. Where known, the parental allele on which the variant arose is indicated.

highly conserved base (chr. 7; g.100318468G>A; **Fig. 4a**) within the 5′ UTR of the *EPO* gene was identified in two independent families with erythrocytosis and cosegregated with the disease (**Fig. 4b** and **Supplementary Note**). Moreover, this was the only rare exonic variant found in an 8-Mb region that was identical by descent in the affected individuals in these two unrelated families (the only such region), suggesting that it had a single and probably recent mutational origin. *EPO* is a strong candidate gene for this disease, as the encoded erythropoietin is essential for red blood cell production and increased levels cause increased red blood cell mass, the hallmark of erythrocytosis[35,36]. However, genetic variation in *EPO* has not previously been linked with erythrocytosis, and further functional data would be necessary to definitively prove causality.

In another case, we discovered a complex event leading to deletion of 1.4 kb of the X chromosome and insertion of 50 kb from chromosome 2p (**Fig. 4c** and **Supplementary Fig. 9**) in a patient with X-linked hypoparathyroidism (**Supplementary Note**). This variant lay 81.5 kb downstream of *SOX3*, segregated with the disease (**Fig. 4d**) and is similar to an event reported previously in an independent kindred[37]. *SOX3* is a strong candidate gene for this disease as it influences the development of the parathyroid gland[38]. Although the pathogenicity for these variants is not proven, that such candidates can be identified using whole-genome sequencing demonstrates the value of screening the noncoding genome.

Incidental findings

The identification of variants unrelated to the referred condition but which have potential clinical and actionable significance is a major challenge for screening by whole-genome sequencing. To evaluate the burden of such incidental findings, we followed the recommendations of the American College of Medical Genetics and Genomics[39] and used HGMD[24] assignments of pathogenic status to identify 32 variants in 18 genes of possible clinical relevance (4 nonsense, 3 splice-site and 25 nonsynonymous variants). After detailed and lengthy review of the literature and curated variant databases, 26 could be eliminated (**Supplementary Table 10**), leaving 6 variants in

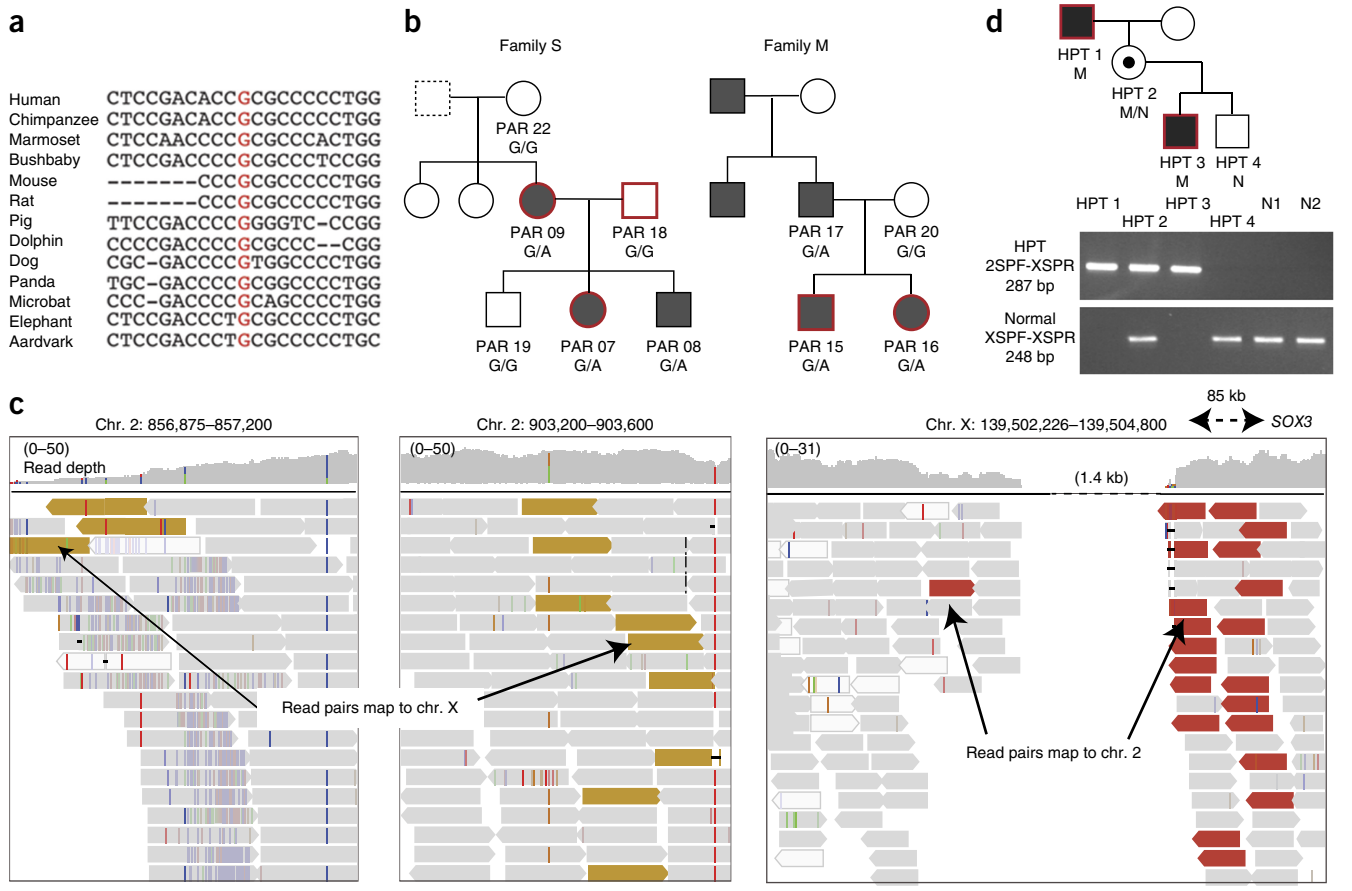

Figure 4 Candidate pathogenic noncoding variants. (**a**) Multiple-species alignment of a region of the 5′ UTR of *EPO* in which a variant was identified at a conserved position (red text) in two families with erythrocytosis. (**b**) Pedigrees for the families with Erythrocytosis studied, showing affected individuals (shaded gray), those sequenced (red borders) and the genotypes of all individuals for whom we had DNA. We had no information about the father of PAR09 (dashed box). (**c**) Summary of read-mapping in an individual with hypoparathyroidism showing evidence for an interstitial insertion-deletion event in which a ~50-kb region of chromosome 2p25.3 (left) has been duplicated and inserted into the X chromosome, resulting in a 1.4-kb deletion 81.5 kb downstream of *SOX3* (right). Yellow reads, mate mapping to the X chromosome; red reads, mate mapping to chromosome 2; gray reads, read and mate mapping to the same chromosome; white reads, read with mapping quality of 0. (**d**) Pedigree showing segregation of the complex variant, with PCR validation below. M, mutation; N, normal. Primers 2SPF and XSPR flank the distal breakpoint of the deletion-insertion (sequences in **Supplementary Fig. 9**). Primers XSPF and XSPR detect the normal allele. The mutation was not seen in 150 alleles from 100 unrelated normocalcemic individuals (50 males and 50 females, including N1 and N2, who are shown).

4 genes, each present in a single case (**Table 2**). Although the majority of these variants have been published in association with a relevant disease, major doubts remain about their clinical interpretation owing to incomplete information on (i) variant frequencies in large populations of healthy people, (ii) phenotypes when variants segregate within families and (iii) corroborative functional studies[40]. When a variant occurs at an appreciable frequency in public databases (for example, *RYR1* variants), the rarity of associated case reports suggests that penetrance is low or indeed zero. Even in the most apparently clear-cut case of a nonsense variant in *BRCA2*, the actual disease risk is likely to be reduced in the absence of a documented family history[41].

Any decision on clinical action must balance multiple potential harms (invasion of personal autonomy, the severity of proposed preventive intervention, associated healthcare costs) against the anticipated benefits to health. We propose that only four variants are clinically reportable (**Table 2**), and a further two variants are of uncertain relevance and warrant further investigation (**Table 2**). For example, the p.Arg397Trp-encoding variant in *KCNQ1* is potentially associated with long-QT syndrome (LQTS) and sudden death. The frequency of this single variant in the Exome Variant Server (EVS) exceeds that of LQTS overall, suggesting very low absolute risk; nevertheless, it is probably reasonable to recommend avoidance of certain classes of medication (even if the subject does not have any obvious electrocardiogram (ECG) abnormality), as this intervention can very occasionally be lifesaving. By contrast, we do not believe that intensive electrophysiological investigation or clinical cascade screening of the extended family are indicated. These observations highlight the urgent need for unbiased data from large biobanks to support clinical decision-making.

DISCUSSION

The goal of the WGS500 study was to evaluate the potential value of whole-genome sequencing in mainstream genetic diagnosis. In routine clinical settings, the opportunities for time-consuming investigation of multiple variants emerging from whole-genome sequencing are limited. We identified multiple strategies in analysis (joint variant calling, filtering of variants against local databases and the use of multiple annotation algorithms) that improve the reliability of the variants called and improve sensitivity and specificity in detecting candidate disease-causing variants.

Table 2 Incidental findings with potentially actionable consequences

Incidental finding condition	Gene	Amino acid change	UK10K	EVS_EA	Comments
Reportable incidental findings					
Arrhythmogenic right-ventricular cardiomyopathy (ARVC)	*DSG2*	NM_001943: c.2397T>G; p.Tyr799*	Absent	Absent	Stop-gain mutation, not previously reported, but mutation class considered pathogenic[44].
	DSG2	NM_001943: c.2554G>T; p.Glu852*	Absent	Absent	Stop-gain mutation, not previously reported, but mutation class considered pathogenic[44].
Breast and ovarian cancers	*BRCA2*	NM_000059: c.7558C>T; p.Arg2520*	Absent	0.0001	Stop-gain mutation; five independent p.Arg2520* alterations in affected patients[45–49]. Four reports of variant being pathogenic in ClinVar[50] (submitted by independent clinical laboratories) and seven records rated as causal in the UMD-BRCA2 database[51]. Mutations of this class described in Brohet *et al.*[52].
Long-QT syndrome	*KCNQ1*	NM_000218: c.877C>T, p.Arg293Cys	Absent	Absent	Two independent reports in the literature: (i) 4/2,500 independent cases from the FAMILION cohort referred for long-QT genetic testing[53] and (ii) 1 case of 388 consecutive unrelated patients with swimming-triggered arrhythmia syndromes[54], as compound heterozygote with p.Gly269Asp. The location of the substitution in the pore is suggestive of pathogenicity.
Incidental findings of uncertain significance					
Long-QT syndrome	*KCNQ1*	NM_000218: c.1189C>T; p.Arg397Trp	Absent	0.0006	Three independent reports in the literature, including 3/2,500 independent cases referred for long-QT testing[55], 5/600 cases in the LQT registry[53] and 1/91 independent cases of intrauterine fetal death[56]. Functional data from heterologous expression of mutation in HEK293 cells, which show markedly reduced current on whole-cell patch clamp as compared with wild type[56], and in inside-out membrane patches from *Xenopus laevis* oocytes, which showed markedly reduced ATP binding[57]. Taken together, this suggests that the mutation should not be disregarded clinically, as it may be weakly pathogenic, albeit with low absolute risk.
Malignant hyperthermia	*RYR1*	NM_000540: c.5036G>A; p.Arg1679His	0.0006	0.0014	Variant observed in single subject with complication and positive functional testing[58] but no independent replication.

Variants deemed to be reportable and clinically actionable are listed in the top part of the table. Those for which the evidence was not considered sufficient to be clinically actionable are reported or that are of uncertain relevance are listed in the bottom part. UK10K, frequency in the UK10K twin cohort; EVS, Exome Variant Server; EVS_EA, frequency in European Americans in the EVS; HGMD, Human Gene Mutation Database; UMD, Universal Mutation Database (see URLs).

With these innovations, whole-genome sequencing proved to be effective for the molecular diagnosis of severe disorders for which a strong genetic component was suspected but where screening of known associated genes had previously failed to identify candidate variants. Overall, whole-genome sequencing identified a pathogenic variant in 33 of 156 cases (21%), including 23 of 68 (33.8%) mendelian cases (class A, B or C in category 1 or 2), with the proportion increasing to 57% (8/14) in cases where *de novo* or recessive models of inheritance were suspected and both parents were sequenced (category 2.1) (**Supplementary Table 8**). The majority of these variants lie within genes and are hence typically accessible through whole-exome sequencing. However, in an independent study of 141 exomes, 3 of 33 sites reported in **Table 1** lay outside the targeted exome, and a further 6 lay within the targeted region but had low coverage (median <20×; **Supplementary Fig. 10**). If a minimum of six reads, three of which support the variant, are required for detecting a new heterozygous variant, we estimate that 15% of the causal variants identified in this whole-genome sequencing study (including coding and noncoding changes) would likely have been missed by whole-exome sequencing (0.5% in WGS500). Conversely, using 20 variant sites identified as causal from an independent exome sequencing project, we estimate that whole-genome sequencing at this coverage has 99.6% power to identify a new heterozygous variant (as compared with 96.1% power in whole-exome sequencing; **Supplementary Fig. 10**).

Moreover, whole-genome sequencing has additional benefits. For example, in the CRS case discussed above, whole-genome sequencing was important for (i) identifying the *HUWE1* mutation, (ii) identifying nearby variants to establish parent of origin, so that we could subsequently show that only the mutant chromosome was being expressed, and (iii) assessing other *de novo* mutations on the X chromosome that might affect X-chromosome inactivation. The latter two points could not have been addressed with whole-exome sequencing data. Moreover, whole-genome sequencing identified, in other cases, two likely pathogenic noncoding variants and unusual chromosomal features including large deletions (for example, a 30-Mb deletion on the X chromosome of a patient with congenital myasthenia, although this deletion is not thought to be relevant to the disorder), distant consanguinity (**Supplementary Fig. 11**) and uniparental disomy (as in the case of a child with Ohtahara syndrome[15]).

In other types of disorders, whole-genome sequencing proved less successful. The number of candidate variants in families with dominantly inherited disorders makes functional validation time-consuming, and many such cases remain in active follow-up. Furthermore, our hypothesis that extreme forms of complex disorders (involving early onset or severe disease) would be enriched for monogenic forms was not confirmed. In only 2 cases out of 49 (1 case of common variable immunodeficiency disorder (CVID) and 1 case of inflammatory bowel disease (IBD)) did whole-genome sequencing on unrelated individuals with extreme immune-related disorders identify strong candidates for pathogenic variants, despite substantial sample sizes ($n = 34$ for CVID and 15 for IBD). Several other candidates have been identified, but pathogenicity has not been confirmed. This low success rate likely reflects the influence of multiple genetic factors, even in extreme cases. Only very large patient cohorts are well powered for the identification of new genes with a modest contribution to the phenotype[42,43]; in any specific case, it will be difficult to assign pathogenicity to a particular variant.

Our results also highlight the outstanding challenges of whole-genome sequence interpretation. Every individual carries multiple rare variants that could potentially be assessed as pathogenic for a

given disorder on the basis of biological information about the gene, the coding consequence of the variant and its frequency within the population. Such variants may be benign or have variable penetrance, making their clinical interpretation challenging without additional information (such as *de novo* status or cosegregation with the disease within a family). Conversely, rigid application of biological candidacy filters will lead to false negatives. Ultimately, whole-genome sequencing will only be able to reliably assess the diagnostic and predictive value of any specific variant if this variant, or another variant in the same gene, is identified in other individuals with the same disorder for whom detailed phenotypic and clinical data are available.

Finally, the identification of pathogenic variants, the exclusion of potential candidate variants and the identification of incidental findings relied on close collaboration between analysts, scientists knowledgeable about the disease and genes, and clinicians with expertise in the specific disorders. The availability of resources and expertise for functional validation studies was critical to the assignment of causality. Provision of this network may be challenging to establish in a clinical setting, but it will be an important aspect of successful translation of whole-genome sequencing into healthcare.

URLs. Stampy read mapper, http://www.well.ox.ac.uk/project-stampy; Platypus variant caller, https://github.com/andyrimmer/Platypus; Picard, http://broadinstitute.github.io/picard/; Ensembl regulatory build, http://www.ensembl.org/info/genome/funcgen/regulatory_build.html; NHLBI Exome Variant Server (EVS), http://evs.gs.washington.edu/EVS/; Copenhagen disease gene association list, http://diseases.jensenlab.org/Search; Eurexpress database, http://discovery.lifemapsc.com/gene-expression-signals/high-throughput/ish-large-scale-dataset-eurexpress; Universal Mutation Database for *BRCA2*, http://www.umd.be/BRCA2/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The majority of samples studied in WGS500 were consented for clinical investigation only. A small number of samples were collected for general research, and whole-genome sequencing data for these samples are available from the European Nucleotide Archive (ENA) under accession PRJEB9151.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the patients and their families who consented to these studies and the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics for the generation of the sequencing data. Additionally, we are grateful to F. Harrington, C. Mignion, V. Sharma, I. Taylor and I. Westbury for assistance with molecular genetic analysis and the staff of the Oxford University Hospitals Regional Genetics and Immunology Laboratories for the DNA preparation for some of the samples.

This work was funded by a Wellcome Trust Core Award (090532/Z/09/Z) and a Medical Research Council Hub grant (G0900747 91070) to P.D., the NIHR Biomedical Research Centre Oxford, the UK Department of Health's NIHR Biomedical Research Centres funding scheme and Illumina. Additional support is acknowledged from the Biotechnology and Biological Science Research Council (BBSRC) (BB/I02593X/1) to G.L. and G.M.; Wellcome Trust grants 093329, 091182 and 102731 to A.O.M.W. and 100308 to L.F.; the Newlife Foundation for Disabled Children (10-11/04) to A.O.M.W.; AtaxiaUK to A.H.N.; the Haemochromatosis Society to K.R.; European Research Council (FP7/2007-2013) grant agreements 281824 to J.C.K. and 305608 to O.D.; the Jeffrey Modell Foundation NYC and Baxter Healthcare to S.Y.P. and H. Chapel; Action de Recherche Concertée (ARC10/15-029, Communauté Française de Belgique) to O.D.; Fonds de la Recherche Scientifique (FNRS), Fonds de la Recherche Scientifique Médicale (FRSM) and Inter-University Attraction Pole (IUAP; Belgium federal government) to O.D.; the Swiss National Centre of Competence in Research Kidney Control of Homeostasis Program to O.D.; the Gebert Rüf Stiftung (project GRS-038/12) to O.D.; Swiss National Science Foundation grant 310030-146490 to O.D.; the Shriners Hospitals for Children (grant 15958) to M.P.W.; and UK Medical Research Council grants G9825289 and G1000467 to R.V.T., L009609 to A.R.A., G1000801 to D.H. and MC_UC_12010/3 to L.F. The views expressed in this publication are those of the authors and not necessarily those of the UK Department of Health.

AUTHOR CONTRIBUTIONS

P.D. and G.M. jointly supervised and oversaw the WGS500 project. C. Babbs, D. Beeson, P.B., E.B., H. Chapel, R.C., J.F., L.F., D.H., A.H., F.K., U.K., J.C.K., A.H.N., S.Y.P., C.P., F.P., P.J.R., P.A.R., K.R., A. Schuh, A. Simmons, R.V.T., I.T., H.H.U., P.V., H.W. and A.O.M.W. were principal investigators on individual projects. V.J.B., K.B., C.D., O.D., R.D.G., J.K., C.L., M.A.N., N. Petousi, S.E.P., S.R.F.T., T.V. and M.P.W. were lead investigators on individual projects. H. Cario, M.F.M., C. Bento, K.D., O.D., R.D.G., D.J., C.L., D.N., E.O., A.B.O., M.P., A. Russo, E. Silverman, P.S.S., E. Sweeney, S.A.W. and M.P.W. contributed clinical samples and clinical data. C.A., M.A., A. Green, S.H., Z.K., S. Lamble, L.L., P.P., G.P.-E., A.T. and L.W. prepared libraries and generated whole-genome sequences, led by D. Buck (High-Throughput Genomics Group, Oxford) and D. Bentley (Illumina Cambridge). J. Becq, J. Broxholme, S.F., R.G., E.H., C.H., L.H., P.H., A.K., S. Lise, G.L., D.M., L.M., A. Rimmer, N.S., B.W., C.Y. and N. Popitsch performed study-wide bioinformatic analysis of whole-genome sequence data, led by J.-B.C. and R.R.C. J.T. performed the whole-exome sequence analysis presented in **Supplementary Figure 10**. E.E.D., A.V.G., M.H., J.L., H.C.M., S.J.M., K.A.M., A.P., L.Q. and P.A.v.S. performed project-specific bioinformatic analysis of whole-genome sequence data. A.R.A., O.C., A.L.F., A. Goriely, I.H.G., A.V.G., R.H., J.L., K.A.M. and A.P. performed project-specific genetic and functional validation studies. G.M. wrote the manuscript with help from H.C.M., J.C.T. and A.O.M.W. and further contributions from S. Lise, D.M., A.P., R.V.T. and S.E.P. J.C. collated information for the paper. P.D. chaired the Steering Committee and the Operations Committee. J.I.B., D. Bentley, G.M., P.J.R., J.C.T. and A.O.M.W. were members of the Steering Committee. J. Broxholme, D. Buck, J.-B.C., R.C., J.C.K., G.L., G.M., J.C.T., I.T., A.O.M.W. and L.W. were members of the Operations Committee.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Need, A.C. *et al.* Clinical application of exome sequencing in undiagnosed genetic conditions. *J. Med. Genet.* **49**, 353–361 (2012).
2. Bamshad, M.J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12**, 745–755 (2011).
3. Yang, Y. *et al.* Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.* **369**, 1502–1511 (2013).
4. Gonzaga-Jauregui, C., Lupski, J.R. & Gibbs, R.A. Human genome sequencing in health and disease. *Annu. Rev. Med.* **63**, 35–61 (2012).
5. Dixon-Salazar, T.J. *et al.* Exome sequencing can improve diagnosis and alter patient management. *Sci. Transl. Med.* **4**, 138ra78 (2012).
6. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
7. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
8. Beaulieu, C.L. *et al.* FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project. *Am. J. Hum. Genet.* **94**, 809–817 (2014).
9. Biesecker, L.G. & Green, R.C. Diagnostic clinical genome and exome sequencing. *N. Engl. J. Med.* **370**, 2418–2425 (2014).
10. Saunders, C.J. *et al.* Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Sci. Transl. Med.* **4**, 154ra135 (2012).
11. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344–347 (2014).
12. Jacob, H.J. *et al.* Genomics in clinical practice: lessons from the front lines. *Sci Transl. Med.* **5**, 194cm5 (2013).
13. Cazier, J.B. *et al.* Whole-genome sequencing of bladder cancers reveals somatic *CDKN1A* mutations and clinicopathological associations with mutation burden. *Nat. Commun.* **5**, 3756 (2014).
14. Babbs, C. *et al.* Homozygous mutations in a predicted endonuclease are a novel cause of congenital dyserythropoietic anemia type I. *Haematologica* **98**, 1383–1387 (2013).
15. Martin, H.C. *et al.* Clinical whole-genome sequencing in severe early-onset epilepsy reveals new genes and improves molecular diagnosis. *Hum. Mol. Genet.* **23**, 3200–3211 (2014).

16. Sharma, V.P. *et al.* Mutations in *TCF12*, encoding a basic helix-loop-helix partner of TWIST1, are a frequent cause of coronal craniosynostosis. *Nat. Genet.* **45**, 304–307 (2013).
17. Cossins, J. *et al.* Congenital myasthenic syndromes due to mutations in *ALG2* and *ALG14*. *Brain* **136**, 944–956 (2013).
18. Lise, S. *et al.* Recessive mutations in *SPTBN2* implicate β-III spectrin in both cognitive and motor development. *PLoS Genet.* **8**, e1003074 (2012).
19. Palles, C. *et al.* Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat. Genet.* **45**, 136–144 (2013).
20. McCarthy, D.J. *et al.* Choice of transcripts and software has a large effect on variant annotation. *Genome Med.* **6**, 26 (2014).
21. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
22. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).
23. Nelson, M.R. *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**, 100–104 (2012).
24. Stenson, P.D. *et al.* The Human Gene Mutation Database: 2008 update. *Genome Med.* **1**, 13 (2009).
25. Pagel, P. *et al.* The MIPS mammalian protein-protein interaction database. *Bioinformatics* **21**, 832–834 (2005).
26. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367**, 1921–1929 (2012).
27. Swaminathan, G. & Tsygankov, A.Y. The Cbl family proteins: ring leaders in regulation of cell signaling. *J. Cell. Physiol.* **209**, 21–43 (2006).
28. Denayer, E. & Legius, E. What's new in the neuro-cardio-facial-cutaneous syndromes? *Eur. J. Pediatr.* **166**, 1091–1098 (2007).
29. Martinelli, S. *et al.* Heterozygous germline mutations in the *CBL* tumor-suppressor gene cause a Noonan syndrome-like phenotype. *Am. J. Hum. Genet.* **87**, 250–257 (2010).
30. Niemeyer, C.M. *et al.* Germline *CBL* mutations cause developmental abnormalities and predispose to juvenile myelomonocytic leukemia. *Nat. Genet.* **42**, 794–800 (2010).
31. Pérez, B. *et al.* Germline mutations of the *CBL* gene define a new genetic syndrome with predisposition to juvenile myelomonocytic leukaemia. *J. Med. Genet.* **47**, 686–691 (2010).
32. Nava, C. *et al.* Analysis of the chromosome X exome in patients with autism spectrum disorders identified novel candidate genes, including *TMLHE*. *Transl. Psychiatry* **2**, e179 (2012).
33. Isrie, M. *et al. HUWE1* mutation explains phenotypic severity in a case of familial idiopathic intellectual disability. *Eur. J. Med. Genet.* **56**, 379–382 (2013).
34. Froyen, G. *et al.* Submicroscopic duplications of the hydroxysteroid dehydrogenase *HSD17B10* and the E3 ubiquitin ligase *HUWE1* are associated with mental retardation. *Am. J. Hum. Genet.* **82**, 432–443 (2008).
35. McMullin, M.F. The classification and diagnosis of erythrocytosis. *Int. J. Lab. Hematol.* **30**, 447–459 (2008).
36. Jelkmann, W. Regulation of erythropoietin production. *J. Physiol.* (*Lond.*) **589**, 1251–1258 (2011).
37. Bowl, M.R. *et al.* An interstitial deletion-insertion involving chromosomes 2p25.3 and Xq27.1, near *SOX3*, causes X-linked recessive hypoparathyroidism. *J. Clin. Invest.* **115**, 2822–2831 (2005).
38. Zajac, J.D. & Danks, J.A. The development of the parathyroid gland: from fish to human. *Curr. Opin. Nephrol. Hypertens.* **17**, 353–356 (2008).
39. Green, R.C. *et al.* ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* **15**, 565–574 (2013).
40. MacArthur, D.G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014).
41. Metcalfe, K. *et al.* Family history of cancer and cancer risks in women with *BRCA1* or *BRCA2* mutations. *J. Natl. Cancer Inst.* **102**, 1874–1878 (2010).
42. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc. Natl. Acad. Sci. USA* **111**, E455–E464 (2014).
43. Moutsianas, L. *et al.* The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS Genet.* **11**, e1005165 (2015).
44. Kapplinger, J.D. *et al.* Distinguishing arrhythmogenic right ventricular cardiomyopathy/dysplasia-associated mutations from background genetic noise. *J. Am. Coll. Cardiol.* **57**, 2317–2327 (2011).
45. Castéra, L. *et al.* Next-generation sequencing for the diagnosis of hereditary breast and ovarian cancer using genomic capture targeting multiple candidate genes. *Eur. J. Hum. Genet.* **22**, 1305–1313 (2014).
46. Chong, H.K. *et al.* The validation and clinical implementation of BRCAplus: a comprehensive high-risk breast cancer diagnostic assay. *PLoS ONE* **9**, e97408 (2014).
47. Borg, A. *et al.* Characterization of *BRCA1* and *BRCA2* deleterious mutations and variants of unknown clinical significance in unilateral and bilateral breast cancer: the WECARE study. *Hum. Mutat.* **31**, E1200–E1240 (2010).
48. Rebbeck, T.R. *et al.* Bilateral prophylactic mastectomy reduces breast cancer risk in *BRCA1* and *BRCA2* mutation carriers: the PROSE Study Group. *J. Clin. Oncol.* **22**, 1055–1062 (2004).
49. Håkansson, S. *et al.* Moderate frequency of *BRCA1* and *BRCA2* germ-line mutations in Scandinavian familial breast cancer. *Am. J. Hum. Genet.* **60**, 1068–1078 (1997).
50. Landrum, M.J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2014).
51. Caputo, S. *et al.* Description and analysis of genetic variants in French hereditary breast and ovarian cancer families recorded in the UMD-BRCA1/BRCA2 databases. *Nucleic Acids Res.* **40**, D992–D1002 (2012).
52. Brohet, R.M. *et al.* Breast and ovarian cancer risks in a large series of clinically ascertained families with a high proportion of *BRCA1* and *BRCA2* Dutch founder mutations. *J. Med. Genet.* **51**, 98–107 (2014).
53. Moss, A.J. *et al.* Clinical aspects of type-1 long-QT syndrome by location, coding type, and biophysical function of mutations involving the *KCNQ1* gene. *Circulation* **115**, 2481–2489 (2007).
54. Choi, G. *et al.* Spectrum and frequency of cardiac channel defects in swimming-triggered arrhythmia syndromes. *Circulation* **110**, 2119–2124 (2004).
55. Kapplinger, J.D. *et al.* Spectrum and prevalence of mutations from the first 2,500 consecutive unrelated patients referred for the FAMILION long QT syndrome genetic test. *Heart Rhythm* **6**, 1297–1303 (2009).
56. Crotti, L. *et al.* Long QT syndrome–associated mutations in intrauterine fetal death. *J. Am. Med. Assoc.* **309**, 1473–1482 (2013).
57. Li, Y. *et al.* Intracellular ATP binding is required to activate the slowly activating K^+ channel I_{Ks}. *Proc. Natl. Acad. Sci. USA* **110**, 18922–18927 (2013).
58. Vukcevic, M. *et al.* Functional properties of *RYR1* mutations identified in Swedish patients with malignant hyperthermia and central core disease. *Anesth. Analg.* **111**, 185–190 (2010).

[1]National Institute for Health Research (NIHR) Comprehensive Biomedical Research Centre, Oxford, UK. [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [3]Centre for Computational Biology, University of Birmingham, Edgbaston, UK. [4]Medical Research Council (MRC) Molecular Haematology Unit, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. [5]Illumina Cambridge, Ltd., Saffron Walden, UK. [6]Neurosciences Group, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. [7]Hematology Department, Centro Hospitalar e Universitário de Coimbra, Coimbra, Portugal. [8]Molecular Haematology Department, Oxford University Hospitals National Health Service (NHS) Trust, Oxford, UK. [9]Department of Clinical Genetics, Oxford University Hospitals NHS Trust, Oxford, UK. [10]Centre for Cellular and Molecular Physiology, University of Oxford, Oxford, UK. [11]Neurobiology Division, MRC Laboratory of Molecular Biology, Cambridge, UK. [12]Department of Pediatrics and Adolescent Medicine, University Medical Center, Ulm, Germany. [13]Primary Immunodeficiency Unit, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [14]Centre de Génétique Humaine, Institut de Génétique et de Pathologie, Gosselies, Belgium. [15]Cliniques Universitaires Saint-Luc, Université Catholique de Louvain, Brussels, Belgium. [16]MRC Human Immunology Unit, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. [17]Institute of Physiology, Zurich Center for Integrative Human Physiology, University of Zurich, Zurich, Switzerland. [18]Clinical Genetics Group, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. [19]University Hospital Southampton NHS Foundation Trust, University of Southampton, Southampton, UK. [20]Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [21]Jenner Institute, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [22]Department of Statistics, University of Oxford, Oxford, UK. [23]Craniofacial Unit, Department of Plastic and Reconstructive Surgery, Oxford University Hospitals NHS Trust, Oxford, UK. [24]Oxford Laboratory for Integrative Physiology, Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Oxford, UK. [25]Kidney Diseases, Feinberg School of Medicine, Northwestern University and the Ann and Robert H. Lurie Children's Hospital of Chicago, Chicago, Illinois, USA. [26]Centre for Cancer Research and Cell Biology, Queen's University, Belfast, UK. [27]Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK. [28]Academic Endocrine Unit, Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Oxford, UK. [29]Centre for Neuropsychopharmacology, Division of Brain Sciences, Imperial College, London, UK. [30]Danish Multiple Sclerosis Center, Department of Neurology, Copenhagen University Hospital, Copenhagen, Denmark. [31]Department of Haematology, Belfast City Hospital, Belfast, UK. [32]Nuffield Department of Medicine, University of Oxford, Oxford, UK. [33]Translational Gastroenterology Unit, University of Oxford, Oxford, UK. [34]Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, UK. [35]Department of Pediatrics, University Hospital, Mainz, Germany. [36]Department of Oncology, University of Oxford, Oxford, UK. [37]Division of Rheumatology, The Hospital for Sick Children, Toronto, Ontario, Canada. [38]Department of Clinical Genetics, Liverpool Women's NHS Foundation Trust, Liverpool, UK. [39]Oxford NHS Regional Molecular Genetics Laboratory, Oxford University Hospitals NHS Trust, Oxford, UK. [40]Division of Genetics, King's College London, Guy's Hospital, London, UK. [41]Center for Metabolic Bone Disease and Molecular Research, Shriners Hospital for Children, St. Louis, Missouri, USA. [42]Office of the Regius Professor of Medicine, University of Oxford, Oxford, UK. [43]These authors contributed equally to this work. [44]These authors jointly supervised this work. Correspondence should be addressed to G.M. (mcvean@well.ox.ac.uk).

ONLINE METHODS

Overview of the WGS500 Project Consortium. Whole-genome sequencing was carried out as part of a collaboration between the Wellcome Trust Centre for Human Genetics at the University of Oxford, the Oxford NIHR Biomedical Research Centre and Illumina. We sought to sequence 500 whole genomes from patients for whom findings could be of immediate clinical use in terms of diagnosis, prognosis, treatment selection, or genetic counseling and reproductive choices.

Process and criteria for sample inclusion. Proposals were invited from clinicians and researchers in Oxford, commencing in December 2010, and were reviewed by the scientific Steering Committee. Variants in known candidate genes and large chromosomal copy number changes had to have been excluded for the patient to be included in the study, to maximize the likelihood of identifying variants in new disease-related genes. Projects were categorized as follows:

1. Families with suspected mendelian conditions with affected individuals across multiple generations (dominant or recessive). In these cases, we usually sequenced one or a few family members (chosen to maximize power for exclusion analysis) and obtained SNP array data on all available other members, to identify regions identical by descent between affected individuals.
    1.1. Dominant model suspected.
    1.2. Recessive model suspected (often due to consanguinity).
    1.3. X-linked model suspected.
    1.4. Multiple unrelated families with linkage to the same region(s).
2. Families with suspected mendelian conditions with one or more affected individuals in a single generation. For these cases, we hypothesized that they were due to *de novo* or recessive mutations.
    2.1. Affected offspring and both parents sequenced.
    2.2. Only affected offspring sequenced, not the parents.
3. Cohort of unrelated sporadic cases with no known family history.
4. Individuals with extreme forms of common disorders (early-onset or severe forms).

Ethics. Individual researchers had explicit research consent to undertake genetic investigation into the cause of the relevant disease and/or samples were obtained with clinical consent as part of efforts to identify the cause of the patient's disease. Ethics committee reference numbers for every individual research project have been provided to the journal editors.

Sequencing library preparation. We obtained 3–5 µg of DNA from each individual, usually from blood or otherwise from saliva or immortalized cell lines. Samples were diluted to 80 ng/µl in 10 mM Tris-HCl, pH 8.5, and then quantified using the High-Sensitivity Qubit system (Invitrogen). Sample integrity was assessed using 1% E-Gel EX (Invitrogen). Focused ultrasonication was carried out to fragment 2 µg of DNA, using the Covaris S2 system with the following settings: duty cycle = 10%, intensity = 5%, cycles/burst = 200 and time = 60 s. Libraries were constructed using the NEBNext DNA Sample Prep Master Mix Set 1 kit (New England BioLabs), with minor modifications to the protocol provided by the manufacturer. Ligation of adaptors was performed using 6 µl of Illumina adaptors (Multiplexing Sample Preparation Oligonucleotide kit). Ligated libraries were selected for size using 2% E-Gel EX (Invitrogen), and the distribution of sizes for the fragments in the purified fraction was determined using the Tapestation 1DK system (Agilent/Lab901). Each library was enriched by PCR with 25 µM of each of the following custom primers: Multiplex PCR primer 1.0, 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′); Index primer, 5′-CAAGCAGAAGACGGCATACGAGAT[INDEX]CAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3′. Index sequences were 8 bp long and formed part of an indexing system developed in house[59]. Four independent PCRs were prepared per sample using 25% of the volume of the pre-PCR library for each reaction. After eight cycles of PCR (cycling conditions were as recommended by Illumina), the four reactions were pooled and purified with AMPure XP beads (Beckman Coulter). The final size distribution was determined using the Tapestation 1DK system. The concentration of each library was determined by RT-PCR using the Agilent qPCR Library Quantification kit and an MX3005P instrument (Agilent Technologies).

Whole-genome sequencing and quality control. Whole-genome sequencing was performed on either the Illumina HiSeq 2000 or HiSeq 2500 instrument run in standard mode, either by the Oxford Genomics Centre at the Wellcome Trust Centre for Human Genetics or Illumina Cambridge. We generated 100-bp reads and used v2.5 or v3 clustering and sequencing chemistry. A PhiX control was spiked into the libraries. We aimed for a mean coverage of 30× and obtained a minimum coverage of 22.7×. The modal number of lanes required to reach the desired coverage was 2.33.

We used the recommended quality metrics in the Illumina Sequence Analysis Viewer in analyzing each lane. Additionally, we generated our own quality metrics for each lane (or, in the case of multiplexes, each part of a lane) and required the following criteria to be met: <2% duplicate pairs; most frequent *k*-mer <2%; >99% mapped; <2.5% read pairs mapping to different chromosomes; mean insert size between 340 and 440 bp, with a median absolute deviation of <30 bp; approximately uniform genomic coverage by GC content; ~1% exonic coverage; <2% N bases at any cycle; and an approximately equal number of reads per tag (three samples multiplexed per lane), standard deviation <25%.

Read mapping. Sequence reads were generated using the Illumina offline basecaller (OLB v1.9.3) and mapped to the GRCh37d5 human reference sequence. This reference genome was obtained from the 1000 Genomes Project and is based on hg19 but contains a 35.48-Mb decoy chromosome that reduces the misalignment of repetitive sequence and improves the accuracy of SNP discovery. Mapping was performed using the Burrows-Wheeler Aligner (BWA v0.5.6)[60] and Stampy (versions 1.0.12 –1.0.22; see URLs)[61], and merging and deduplication were performed using Picard (v1.67; see URLs).

SNP and indel calling and genotyping. Variant calling was performed with Platypus[62] (version 0.1.9; see URLs) using the default settings. This algorithm can detect SNPs and short indels (<50 bp) and is sensitive to somatic mosaic mutations at low allele frequencies[63,64].

The variant calling included two stages. First, we used Platypus to identify SNPs and short indels in all samples individually (raw calling). We then ran it a second time to genotype the union of all variants in all samples. We performed the raw calling on groups of related samples together ('joint calling'), so that the same sites were interrogated for all samples in the family. It was thus possible to distinguish homozygous reference from a missing call, and the observation of a variant in one of the individuals in the family reduced the required threshold for calling it in the other family members[62].

We retained variants with a "PASS" flag that had a posterior probability (Phred scale) of >20 that the variant segregates. Variants with a "clustered" flag (within 25 bp of another variant) were manually checked in the Integrated Genomics Viewer (IGV)[65] but not discounted.

To check for sample contamination, we plotted the distribution of the ratios of the number of reads containing the alternate allele to the total number of reads (ALT:TOTAL) for known SNPs (from dbSNP) and new SNPs for each sample (**Supplementary Fig. 1d**). To check for duplicates and cryptically related individuals, we ran principal-components analysis on the WGS500 data. We included three populations (CEU, YRI and JPT+CHB) from the HapMap Project, which allowed us to identify a few individuals with a particularly high number of variants as ancestry outliers.

Filtering variants in trios. There were 15 families in WGS500 from which both the parents and one or more affected children were sequenced: 6 trios and 1 quartet with EOE, 1 trio with hypertrophic cardiomyopathy, 1 trio with erythrocytosis (with mother and daughter affected), 4 trios with CRS, 1 trio with Saethre-Chotzen syndrome (a type of CRS) and 1 quartet with XLMR. If the parents were unaffected and only one child was affected, the initial hypothesis was that the causal mutation occurred *de novo*. To screen for these

variants, we searched for variants that were absent from all public databases (1000 Genomes Project, dbSNP and ESP) and from other WGS500 samples and that had been confidently called as heterozygous in the child and as homozygous reference in the parents (genotype log likelihood ratio, GLLR < -5). (This latter criterion for GLLR was not always applied when analyzing data for specific projects.) The value of applying these different filters is demonstrated in **Supplementary Table 3**.

We also investigated a simple recessive model in these families. Homozygous variants in the affected child (or children) had to have a frequency less than 0.5% in the 1000 Genomes Project and ESP databases (corresponding to an expected homozygous frequency of 1 in 40,000), and there had to be 0 homozygotes and ≤2 heterozygotes among the unrelated WGS500 samples. We required parents to be called as heterozygous. The results of these filters are shown in **Supplementary Table 4**. When filtering for compound recessive candidates (with the child having two rare heterozygous coding variants in the same gene, one inherited from each parent) and X-linked recessive candidates, we used the same frequency thresholds as for the simple recessive case.

Annotation of variants. The functional consequences of variants were predicted using several programs. We used ANNOVAR (February 2013 version)[66] to annotate variants with respect to RefSeq genes, adding information about segmental duplications, conservation (based on the UCSC alignment of 46 mammalian genomes), GERP, SIFT, MutationTaster, phyloP and PolyPhen-2 scores, dbSNP identifiers (version 1.35), and frequency in the ESP (see URLs) and Phase 1 of the 1000 Genomes Project[6]. We also annotated all variants using the Variant Effect Predictor from Ensembl (version 69) and the nonsynonymous SNPs using PolyPhen-2 (version 2.2.2r405).

Detection of copy number variants and extended homozygosity. We used several different methods to search for copy number variants (CNVs). First, we generated count profiles for each individual by dividing each chromosome into 10-kb bins and counting the number of reads in each bin. For each chromosome, we applied principal-components analysis to the log of the counts (training set; one per family) and then plotted the residuals of the predicted principal components along the chromosome. This procedure served to remove noise in the data due to biases in sequence composition. Candidate CNVs (down to about 10 kb) were identified visually as outliers. Second, we applied OncoSNP-SEQ[67] in germline mode to a subset of 300,000 reliable SNPs across the genome. Coverage and read counts at these locations were used as a proxy for the intensity and B-allelic frequency under a specific model of the hidden Markov model intended for next-generation sequencing data. Third, to identify exon-level CNVs, we used ExomeDepth[68].

To search for long regions of homozygosity, we calculated the fraction of heterozygous SNPs in 10-kb bins along the chromosomes for each individual, averaged these over 1-Mb regions and plotted them, ignoring centromeric and other repetitive regions. We classed as homozygous any segments with a heterozygous/homozygous ratio <0.2; this empirically chosen threshold avoided large homozygous regions being interrupted by genotyping errors in regions difficult to sequence but clearly distinguished them from the rest of the genome. Homozygous regions of up to 4 Mb in size are common in demonstrably outbred individuals[69]. Consanguinity had already been reported for many of the 39 individuals for whom regions of homozygosity larger than 4 Mb were identified (**Supplementary Fig. 11**); in cases for whom consanguinity had not already been reported, this finding prompted analysts to search for rare homozygous variants within these regions.

SNP array data. Illumina SNP arrays were run on some WGS500 samples and other relatives to check the genotyping accuracy of our sequencing pipeline, to refine linkage regions, to confirm familial relationships and, in two cases, to investigate whether large stretches of homozygosity were likely due to uniparental disomy or unreported consanguinity. We ran 200 ng of DNA on the Illumina Human CytoSNP-12 BeadChip or on the Human 1M-Duo BeadChip, following the manufacturer's guidelines. Concordance between the CytoSNP12 genotypes and the whole-genome sequencing data is shown in **Supplementary Tables 1** and **2**, and the dependence on coverage is shown in **Supplementary Figure 2**. In most cases, aCGH had already been performed before the submission of samples, but we also used QuantiSNP[70] to check for CNVs, as well as Nexus Copy Number version 7 (BioDiscovery). We used MERLIN[71] in familial studies to identify regions that were identical by descent.

Assessment of coverage in WGS500 and exome-sequence data. After removing duplicate reads, we used BEDTools[72] to measure coverage in all WGS500 samples and examined the cumulative distributions across the genome and exome (CCDS transcripts) (**Supplementary Fig. 1a**,**b**). To compare this coverage with that for a typical exome sequencing experiment, we took 141 whole-exome data sets for which capture was performed using the Roche NimbleGen SeqCap EZ v.2.0 kit at the Oxford Biomedical Research Centre, removed duplicate reads and measured coverage at each position in the CCDS transcripts. We compared the exome-wide coverage distributions between WGS500 and these exomes (**Supplementary Fig. 1**), as well as the coverage at specific variants thought to be causal (**Supplementary Fig. 10**).

Identifying variants of clinical relevance. After variant calling and annotation, cases from specific projects were investigated by different analysts, and variants were prioritized on the basis of mode of inheritance, functional consequence, population frequency, evolutionary consequence of the position and biological relevance. Where available, parental data, linkage information and repeated occurrence across multiple independent cases of a disorder were used to aid prioritization. Validation of biological consequence and/or screening of additional cohorts were used to confirm pathogenicity. Where a previously described variant or variant class (for example, loss of function or frameshift) was observed in a known gene for a disorder, it was assumed to be pathogenic and to have been missed by previous screening. All putatively causative variants were confirmed using Sanger sequencing. Information was returned to the clinicians responsible for managing individual patients, who decided whether and how information was reported to them.

Classification system for the results. We categorized the results for each independent case into five classes, as follows:

Class A:	Mutation found in a new gene for the phenotype, with additional genetic evidence (in unrelated cases) and/or functional data supporting causality.
Class B:	Mutation found in a gene known for a different phenotype, with additional genetic evidence and/or functional data supporting causality.
Class C:	Mutation found in a gene known for this phenotype.
Class D:	Mutation found in a new gene for the phenotype, with further genetic and functional validation studies in progress.
Class E:	No single candidate gene yet or negative results for validation of the original candidate gene(s).

The results for all projects are summarized in **Figure 1**. Note that one of the CVID cases recovered antibody production and was thus found to have been misdiagnosed.

Tiered strategy for the analysis of genes or variants. For 8 of the 13 families for which we sequenced the affected child (children) and healthy parents, we have identified the causal mutation with a reasonable level of confidence (class A, B or C in **Supplementary Table 6**). Thus, we used these families as test cases.

We compiled tiered lists of candidate genes for each of the three diseases: EOE, CRS and XLMR. For EOE (**Supplementary Table 5**), tier 1 contained genes that are recorded in HGMD[24] as causing Ohtahara syndrome or epileptic encephalopathy, tier 2 contained genes that interact with tier 1 genes (according to the MIPS database[25]) or that are listed in HGMD as causing more general epilepsy, and tier 3 contained genes that are components of biological pathways known to be involved in pathogenesis. For CRS, tier 1 comprised a manually curated list of genes mentioned in the literature as causing CRS in 2 or more cases, tier 2 comprised a list of additional genes associated with the term "craniosynostosis" in the Copenhagen disease gene association list (see URLs) and tier 3 comprised additional orthologs of 270 mouse genes that are expressed in the skull[73] (Eurexpress database; see URLs). For XLMR, we compiled the tier 1 list by searching HGMD for "mental retardation" and

"intellectual disability" and then restricting to chromosome X; we did not consider additional tiers because tier 1 already contained 83 genes.

We analyzed the mutational burden in these genes for all 216 samples, having excluded the contaminated sample, HCM_2361 (**Supplementary Fig. 1**). Specifically, we screened for coding variants in these genes that would appear to fit a *de novo* dominant, simple recessive or X-linked model in the absence of parental information, according to the following criteria:

- *De novo* model: new variant heterozygous in the proband and absent from the 1000 Genomes Project and ESP databases and other unrelated WGS500 samples.
- Simple recessive model: new or very rare variant homozygous in the proband with frequency <0.5% in the 1000 Genomes Project and ESP databases and no other homozygotes and ≤2 heterozygotes among the unrelated WGS500 samples.
- X-linked recessive model: new or very rare variant hemizygous in the male proband and with frequency <0.5% in the 1000 Genomes Project and ESP databases and no other homozygous females or hemizygous males and ≤2 heterozygotes among the unrelated WGS500 samples.

A variant was considered as coding if it was annotated by ANNOVAR[66] as missense, stop gain or stop loss, an indel or within a splice site, for one or more transcripts, and as conserved if it had a GERP or phyloP score greater than 2 or was in a constrained element as defined by the UCSC 46-way alignment. We also used VEP to add information about regulatory regions from the Ensembl v65 Regulatory Build (see URLs).

Actionable, pathogenic incidental findings. We followed the guidelines of the American College of Medical Genetics and Genomics[39] for reporting incidental findings. To identify potentially disease-causing mutations, we first took all the mutations in HGMD[24], aligned the indels to the left and removed variants with erroneous reference alleles. We retained variants classed as "DM" (disease mutation) that were within 10 bp of an exon of a gene in **Table 1** of Green *et al.*[39]. We then searched all WGS500 samples for these mutations and for nonsense or frameshift mutations in the same genes, which would be expected to be pathogenic. We removed variants that (i) were classified as being of unknown significance in curated, disease-specific mutation databases; (ii) had a high frequency (>1%) in the 1000 Genomes Project or EVS database; (iii) had data showing a lack of function; (iv) showed a lack of segregation in literature reports; or (v) were unlikely to be of relevance on the basis of their sequence context (for example, splicing variants without clear splicing signatures). These variants are listed in **Supplementary Table 10**. The remaining variants (**Table 2**) were then scrutinized by a panel of clinical experts, who decided which variants had enough data supporting pathogenicity to report.

59. Lamble, S. *et al.* Improved workflows for high throughput library preparation using the transposome-based Nextera system. *BMC Biotechnol.* **13**, 104 (2013).
60. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
61. Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21**, 936–939 (2011).
62. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).
63. Pagnamenta, A.T. *et al.* Exome sequencing can detect pathogenic mosaic mutations present at low allele frequencies. *J. Hum. Genet.* **57**, 70–72 (2012).
64. Ruark, E. *et al.* Mosaic *PPM1D* mutations are associated with predisposition to breast and ovarian cancer. *Nature* **493**, 406–410 (2013).
65. Thorvaldsdóttir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
66. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
67. Yau, C. OncoSNP-SEQ: a statistical approach for the identification of somatic copy number alterations from next-generation sequencing of cancer genomes. *Bioinformatics* **29**, 2482–2484 (2013).
68. Plagnol, V. *et al.* A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* **28**, 2747–2754 (2012).
69. McQuillan, R. *et al.* Runs of homozygosity in European populations. *Am. J. Hum. Genet.* **83**, 359–372 (2008).
70. Colella, S. *et al.* QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.* **35**, 2013–2025 (2007).
71. Abecasis, G.R., Cherny, S.S., Cookson, W.O. & Cardon, L.R. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30**, 97–101 (2002).
72. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
73. Diez-Roux, G. *et al.* A high-resolution anatomical atlas of the transcriptome in the mouse embryo. *PLoS Biol.* **9**, e1000582 (2011).