

Supplement

Unique Features of Genome Sequencing for Rare Disease Diagnosis

Contents:

Methods.....	2
Sequencing.....	2
Analysis.....	3
Variant classification.....	5
References.....	5
Table S1 Standard searches used by the Broad team for ES/GS analysis.....	9
Figure S1 Classification of variants.....	10
Table S2. Diagnoses.....	11
Table S3. Candidates.....	12
Table S4. Diagnostic yield by imputed ancestry.....	15

Sequencing Methods

Genome sequencing (GS) was performed by the Genomics Platform at the Broad Institute of MIT and Harvard. PCR-free preparation of sample DNA (350 ng input at >2 ng/ul) is accomplished using Illumina HiSeq X Ten v2 chemistry. Libraries are sequenced to a mean target coverage of 30x. GS data was processed through a pipeline based on Picard, using base quality score recalibration and local realignment at known indels. The BWA aligner was used for mapping reads to the human genome build 38 (GRCh38). Single nucleotide variants and insertions/deletions (indels) were jointly called across all samples using Genome Analysis Toolkit (GATK) HaplotypeCaller package version 4.0. Default filters were applied to SNV and indel calls using the GATK Variant Quality Score Recalibration (VQSR) approach. Annotation was performed using Variant Effect Predictor (VEP). GATK-SV¹ was used to detect structural variants (SVs), which were annotated with the GATK SVAnnotate tool. Mitochondrial DNA (mtDNA) single nucleotide and small indel variants were called from GS data using the gnomAD-mitochondria pipeline² and large mtDNA deletions were called by MitoSAlt³. ExpansionHunter v5 was used to genotype known disease-associated tandem repeat expansions (TREs).⁴ Lastly, the variant call set was uploaded to *seqr* for collaborative analysis between the CMG and investigator or for analysis by the RGP team.⁵

ES was performed prior to GS for many of these cases through a variety of clinical diagnostic laboratories or by the Genomics Platform at the Broad Institute. For these cases, libraries from DNA samples (>250 ng of DNA, at >2 ng/ul) were created with an Illumina Nextera exome capture (38 Mb target) and sequenced (150 bp paired reads) to cover >80% of targets at 20x and a mean target coverage of 80x until January 2019 and thereafter using a Twist exome capture (~38 Mb target) and sequenced (150 bp reads) to cover > 90% of targets at 20x and a mean target

coverage of 80x. Sample identity quality assurance checks were performed on each sample. The ES data was de-multiplexed and each sample's sequence data were aggregated into a single Picard BAM file. ES data was subsequently processed similar to GS data as previously described.

Analysis

The Broad CMG analysis team has developed four standard searches that are applied for each family (Table S1). We prioritize variants for further study that have high pathogenicity scores using common *in silico* predictors (e.g. REVEL⁶, CADD⁷, SIFT⁸, PolyPhen-2⁹, MutationTaster¹⁰, MPC¹¹), occur at highly conserved residues as determined by manual review on the UCSC genome browser and evaluating the Genomic Evolutionary Rate Profiling (GERP) score¹². We visually inspect the read data using the Integrated Genomics Viewer (IGV) for our candidate variants to ensure they are not sequencing artifacts. Top candidate variants are typically confirmed by an orthogonal method such as Sanger sequencing.

To detect SVs, we initially utilized multiple SV-calling tools including Manta¹³, DELLY¹⁴, and Smoove (<https://github.com/brentp/smoove>), and more recently applied GATK-SV: an ensemble SV detection tool that discovers, genotypes, and resolves the diverse classes of SVs that can be captured from GS data, including balanced and unbalanced CNVs, inversions, insertions, translocations, and a spectrum of complex SVs. Briefly, GATK-SV maximizes sensitivity by harmonizing five algorithms, then adjudicating and re-genotyping SVs from raw read evidence¹. GATK-SV considers all SV evidence available from GS, including discordant paired-end (PE) or split reads (SR) crossing a breakpoint, and normalized read-depth (RD) or B-allele frequencies. Each CRAM file is processed with five algorithms, which currently include two PE/SR algorithms (Manta¹³, Wham¹⁵), two RD algorithms (cnMOPS¹⁶ and GATK-gCNV¹⁷), and a

mobile element algorithm, MELT¹⁸. GATK-SV is publicly available on GitHub (<https://github.com/broadinstitute/gatk-sv>). For CNV analysis, we also apply germline Copy Number Variant caller (gCNV), a coverage-based CNV detection method that normalizes coverage across the exome by adjusting for systematic bias and uses a probabilistic framework to infer copy number from the normalized coverage. We manually evaluate the CNV data, filtering out low-quality calls and inherited variants (based on family history) and focus our analysis on CNVs overlapping protein-coding genes. For cases with a strong phenotype pointing to a particular gene or genes as the likely candidate, we may also manually search for SVs by visually-inspecting the reads across the gene in question (using the Integrated Genomics Viewer¹⁹).

To evaluate for tandem repeat expansions (TREs), we run ExpansionHunter v5²⁰ on GS samples to genotype 60 known disease-associated repeat loci. The locus specifications we use are publicly available on github (<https://github.com/broadinstitute/str-analysis>) and represent the same list of loci for which population frequencies are available in the gnomAD browser (https://gnomad.broadinstitute.org/short-tandem-repeats?dataset=gnomad_r3). We also run REViewer²¹ to generate read visualizations. Then, to identify candidate pathogenic expansions, we evaluate individuals with the most-expanded genotypes for each locus, comparing them to the pathogenic threshold and population frequencies for this locus in gnomAD. We also evaluate genotype qualities based on reviewing read visualizations.

To evaluate mtDNA SNVs and indels, we run the mitochondria mode of Mutect2 followed by the gnomAD-mitochondria pipeline². Then, to identify candidate variants, we search for “confirmed” variants listed in MITOMAP²² and P/LP variants listed in ClinVar. We finally review all mtDNA variants of uncertain significance (VUS) reported in ClinVar and/or with

“reported” status in MITOMAP, in addition to unreported variants with pathogenic *in silico* prediction based on mtDNA-specifications of the ACMG/AMP guidelines (APOGEE ≥ 0.5 for missense variants; MitoTIP > 12.66 plus HmtVar ≥ 0.35 for tRNA variants), that are absent at high heteroplasmy level ($\geq 80\%$) or homoplasmy ($\geq 95\%$) in reference databases (gnomAD v3 and HelixMTdb).

Variant classification

In order to systematically assess the pathogenicity of the structural variants that we identified, the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen) standards for classification and reporting of constitutional copy-number variants were applied²³. Variants in novel gene-disease relationships are classified as VUS until the gene-disease relationship has at least moderate evidence supporting it. CNV associated with disorders that follow an autosomal recessive or X-linked mode of inheritance are not addressed in these standards and required additional consideration; the classification criteria were modified to optimally capture evidence for pathogenicity for the range of variants that we identified.

Relative proportions of VUS to pathogenic/likely pathogenic variants are presented in Figure S1.

References

1. Collins RL, Brand H, Karczewski KJ, et al. A structural variation reference for medical and population genetics. *Nature* 2020;581(7809):444-451.
2. Laricchia KM, Lake NJ, Watts NA, et al. Mitochondrial DNA variation across 56,434 individuals in gnomAD. *Genome Res* 2022;32(3):569-582.
3. Basu S, Xie X, Uhler JP, et al. Accurate mapping of mitochondrial DNA deletions and duplications using deep sequencing. *PLoS Genet* 2020;16(12):e1009242.

4. Dolzhenko E, Deshpande V, Schlesinger F, et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics* 2019;35(22):4754-4756.
5. Pais LS, Snow H, Weisburd B, et al. seqr: A web-based analysis and collaboration tool for rare disease genomics. *Hum Mutat* 2022;43(6):698-707.
6. Ioannidis NM, Rothstein JH, Pejaver V, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* 2016;99(4):877-885.
7. 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(3):310-5.
8. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acid Res* 2012;40:452.
9. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;7(4):248-9.
10. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nature methods* 2014;11(4):361-362.
11. Kaitlin E. Samocha, Jack A. Kosmicki, Konrad J. Karczewski, Anne H. O’Donnell-Luria, Emma Pierce-Hoffman, Daniel G. MacArthur, Benjamin M. Neale, Mark J. Daly. bioRxiv 148353; doi: <https://doi.org/10.1101/148353>.
12. Huber CD, Kim BY, Lohmueller KE. Population genetic models of GERP scores suggest pervasive turnover of constrained sites across mammalian evolution. *PLoS Genet* 2020;16(5):e1008827.

13. Chen X, Schulz-Trieglaff O, Shaw R, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 2016;32(8):1220-2.
14. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 2012;28(18):i333-i339. .
15. Kronenberg ZN, Osborne EJ, Cone KR, et al. Wham: Identifying Structural Variants of Biological Consequence. *PLoS Comput Biol* 2015;11(12):e1004572.
16. Klambauer G, Schwarzbauer K, Mayr A, et al. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res* 2012;40(9):e69.
17. Fu JM, Satterstrom FK, Peng M, et al. Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. *Nat Genet* 2022;54(9):1320-1331.
18. Gardner EJ, Lam VK, Harris DN, et al. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res* 2017;27(11):1916-1929.
19. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013;14(2):178-92.
20. Dolzhenko E, van Vugt J, Shaw RJ, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res* 2017;27(11):1895-1903.
21. Dolzhenko E, Weisburd B, Ibanez K, et al. REViewer: haplotype-resolved visualization of read alignments in and around tandem repeats. *Genome Med* 2022;14(1):84.

22. Kogelnik AM, Lott MT, Brown MD, Navathe SB, Wallace DC. MITOMAP: a human mitochondrial genome database--1998 update. *Nucleic Acids Res* 1998;26(1):112-5.
23. Riggs ER, Andersen EF, Cherry AM, et al. Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genet Med* 2020;22(2):245-257.
24. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;581(7809):434-443.
25. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, et al. Predicting Splicing from Primary Sequence with Deep Learning. *Cell* 2019;176(3):535-548.e24.

Table S1: Standard searches used by the Broad team for ES/GS analysis

Search	Variant annotations	Variant frequency (Broad callset, gnomAD ²⁴ , gnomAD SV ¹)	Variant call quality	Annotation overrides (SpliceAI score ²⁵)
Dominant/ <i>de novo</i> restrictive	Coding variants, essential and extended splice site, LOF SVs	0.001 gnomAD 0.01 callset	Pass VQSR GQ 40 AB 20	ClinVar LP/P SpliceAI >0.2
Recessive restrictive	Coding variants, essential and extended splice site, LOF SVs	0.01 gnomAD 0.03 callset	Pass VQSR GQ 40 AB 20	ClinVar LP/P SpliceAI >0.2
Dominant/ <i>de novo</i> permissive	Coding variants, synonymous, splice, 5/3'UTR, Non-coding exons, TFBS, regulatory region, LOF/intronic/UTR/promoter SVs	0.001 gnomAD 0.01 callset	GQ 40 AB 10	ClinVar LP/P/VUS SpliceAI >0.1
Recessive permissive	Coding variants, synonymous, splice, 5/3'UTR, Non-coding exons, TFBS, regulatory region, LOF/intronic/UTR/promoter SVs	0.01 gnomAD 0.03 callset	GQ 40 AB 10	ClinVar LP/P/VUS SpliceAI >0.1

Legend: LOF: Loss of function; VQSR: Variant Quality Score Recalibration; GQ: Genotype quality; AB: allele balance; LP/P: Likely pathogenic/pathogenic; UTR: Untranslated region; TFBS: transcription factor binding site; VUS: Variant of uncertain significance

Figure S1. Classification of variants. Pathogenicity of 284 variants in 218 families solved via GS, classified as per the ACMG/AMP/ClinGen standards.

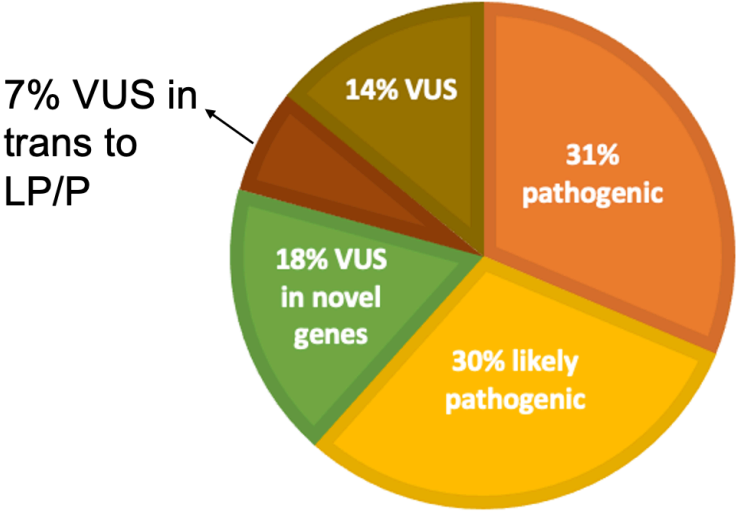


Table S2. Diagnoses

See separate spreadsheet

Table S3. Candidates. Candidate novel disease genes identified in this cohort.

ID	Gene
UWA LAI963	<i>ABCD3</i>
HK115	<i>ACSL5</i>
SCO PED096	<i>ADGRE3</i>
BON B17-59	<i>AFAP1L1</i>
SOU FAM00008	<i>AGMAT</i>
HK103	<i>ANO1</i>
HK081	<i>ANO2</i>
RGP 696	<i>ARFGEF3</i>
CHU 05	<i>ARHGAP6</i>
RGP 245	<i>BAZ1B</i>
RGP 12	<i>BLOC1S1</i>
RGP 655	<i>BOD1</i>
CHU 23	<i>C10orf71</i>
RGP 572	<i>CACNA2D3</i>
RGP 658	<i>CAMK1D</i>
RGP 284	<i>CAMK4</i>
HK060	<i>CBX8</i>
HK017	<i>CDK11B</i>
RGP 735	<i>CDK16</i>
IK	<i>CDK5RAP3</i>
RGP 1374	<i>CEP192</i>
BEG 0761	<i>CFAP46</i>
RGP 868	<i>CFAP54</i>
CMG Laing_Ravencroft_WGS	<i>COL5A3</i>
RGP 1333	<i>DIPK2B</i>
RGP 45	<i>DNAH17</i>
RGP 1149	<i>EBF2</i>
49	<i>ELK1</i>
UWA LAI1646	<i>EP400</i>
RGP 726	<i>EPHA6</i>
RGP 54	<i>ERICH3</i>
RGP 1268	<i>ETV1</i>
827	<i>FAM193A</i>
RGP 1180	<i>FBXO42</i>
RGP 375	<i>FGF7</i>
HK085	<i>FLYWCH1</i>
RGP 1150	<i>FRG2</i>

RGP 86	<i>FRMPD3</i>
RGP 589	<i>FURIN</i>
RGP 1129	<i>GFPT1</i>
RGP 119	<i>GRM4</i>
HK028	<i>GTF2A1</i>
HK032	<i>HAPLN2</i>
HK044	<i>HEATR1</i>
HK075	<i>HELZ</i>
RGP 673	<i>HNRNPL</i>
VCGS FAM148	<i>HOXC8</i>
RGP 1016	<i>INTS6L</i>
CHU 04	<i>ISLR2</i>
RGP 1479	<i>KCNH8</i>
RGP 95	<i>KDM4A</i>
RGP 1392	<i>KDM8</i>
HK080	<i>KIAA0408</i>
RGP 1526	<i>KLHL13</i>
FAM39	<i>LATS2</i>
RGP 1498	<i>LBX1</i>
RGP 289	<i>MACO1</i>
RGP 105	<i>MARCH5</i>
RGP 20	<i>MAU2</i>
HK104	<i>MCRS1</i>
RGP 674	<i>MRPL54</i>
RGP 1175	<i>MYH1</i>
BON B18-54	<i>MYO7B</i>
RGP 526	<i>NCOR1</i>
RGP 1425	<i>NCOR2</i>
RGP 495	<i>NELL2</i>
RGP 314	<i>OSBPL9, SYNRG</i>
RGP 1138	<i>PACSIN3</i>
RGP 329	<i>PKP4</i>
RGP 1193	<i>PPP1R12C</i>
RGP 232	<i>PRICKLE3, GNA13</i>
RGP 1504	<i>PRPF4B</i>
VCGS FAM52	<i>PRPS2</i>
RGP 853	<i>PTPRG</i>
RGP 53	<i>RAB33A</i>
MAN 1601	<i>RCOR2</i>

RGP 230	<i>RHPN2</i>
RGP 469	<i>RIMS1</i>
235	<i>RPL37A</i>
GI	<i>SARNP</i>
RGP 402	<i>SCRIB</i>
RGP 522	<i>SCTR</i>
HK035	<i>SH3GL1</i>
1024	<i>SHCBP1</i>
HK072	<i>SMG6</i>
RGP 1099	<i>SMYD1</i>
FAM61	<i>SNED1</i>
RGP 682	<i>SRGAP2</i>
RGP 731	<i>SRRT</i>
RGP 5	<i>SSBP3</i>
RGP 135	<i>SSH1</i>
RGP 918	<i>SYNM</i>
CHU 01	<i>TBC1D22A</i>
RGP 431	<i>THAP12</i>
RGP 951	<i>THBS2</i>
RGP 677	<i>TLK1</i>
RGP 123	<i>TPPP</i>
FAM29	<i>TPR</i>
VCGS FAM147	<i>TRABD2B</i>
966	<i>TTC28</i>
RGP 452	<i>TXLNG</i>
RGP 1101	<i>UNC13B</i>
VCGS FAM2	<i>UNC5C</i>
RGP 1125	<i>VPS37D</i>
FAM1	<i>WWP1</i>
RGP 504	<i>ZBTB1</i>
RGP 763	<i>ZC3H11A</i>

Table S4. Diagnostic yield by imputed ancestry

Ancestry Category	Total (N, % of cohort)	Diagnosed (N, % of subgroup)
African/African American	22 (3.0%)	5 (22%)
Ashkenazi Jewish	33 (4.4%)	20 (61%)
East Asian	12 (1.6%)	2 (17%)
European	570 (76.6%)	170 (43%)
Latino/Admixed American	27 (3.6%)	4 (15%)
Middle Eastern	2 (0.3%)	1 (50%)
South Asian	13 (1.7%)	4 (31%)
Multiple/Unknown	65 (8.7%)	19 (29%)