Prospective evaluation of genome sequencing versus standard-of-care as a first molecular diagnostic test

Brockman, Austin-Tse et al.
Supplementary Information

Table of Contents

Supplementary Methods

Small Sequence Variant Calling and Analysis

Reads were aligned to the human reference sequence (GRCh37) using Burrows-Wheeler Aligner (BWA), and variant calls were made using the Genome Analysis ToolKit (GATK). Our sequence variant filtration methods are described in Figure S1. In patients who elected to receive such information, we additionally screened for previously reported and novel variants in 59 genes of medical significance[1] that may be unrelated to the patient phenotype (secondary findings). Furthermore, when patients were enrolled as trios, an additional analysis was performed to identify any variants that were identified in the proband but absent from both parents (*de novo* variation).

SV Calling and Analysis

SV calling was conducted on the cGS data using a single sample version of GATK-SV2 v.0.7 (https://github.com/broadinstitute/gatk-sv-single-sample) that we developed for these studies to be interoperable on the Terra cloud platform (https://terra.bio). GATK-SV is an ensemble SV pipeline that maximizes SV discovery by combining SV algorithms that capture orthogonal evidence for SV detection into a single callset and adjudicates the aggregate SV set with evidence directly from the aligned BAM files to improve specificity. For this study four SV algorithms were processed from each individual, including two paired-end/split-read algorithms (Manta v.1.5.0, WHAM-GRAPHENING v.1.7.0)3,4 and two read-depth algorithms (cnMops v.1.12.0 and GATK-gCNV)5,6. In the single sample mode SVs are genotyped against an existing panel of control genomes which is used in downstream filtering. All SVs are annotated for predicted genic impact as previously described2, and resulting variant calls were filtered to identify variants that were predicted to cause loss-of-function, which included deletions that span coding sequence, any SVs that directly disrupted a canonical transcript (e.g. inversion, insertion, translocation, intragenic exonic duplication, or complex SV with one or more breakpoints that disrupted a canonical transcript), or whole gene copy gain. These SVs were then further filtered to rare variants with an allele frequency <5% within our cohort. Partial gene duplications without data to suggest disruption of the primary copy of the gene were excluded given their uncertain functional impact7. Additional post hoc quality control filters were applied to normalize samples that harbored an unusually large number of SVs due to abnormal dosage profiles of read counts that contribute to spurious read depth-based SV detection2. Among the ten samples with an excess of SVs discovered via read depth, we restricted variants discovered by read depth alone to large CNVs >25 kb for seven cases (65CGS, 80CGS, 152CGS, 169CGS, 120CGS, 147CGS, 148CGS), while three extreme outlier cases were omitted from read depth-based analyses (21CGS, 176CGS, 183CGS).

Of note, the SV calling and analysis methods described above have not been clinically validated. While the cGS-derived CNVs reported in this paper represent high confidence calls, orthogonal confirmation of these variants was ongoing at the time of publication of this manuscript, with the exception of the homozygous STRC/CATSPER2 deletion identified in participant 170CGS, which was confirmed via droplet digital PCR as previously described8.

Variant Assessment and Reporting

The evidence for gene-disease validity and relevance to the patient phenotype was evaluated for each variant resulting from the filtering strategies above and variants were classified based on ACMG/AMP criteria with ClinGen rule specifications (http://www.clinicalgenome.org/working-groups/sequence-variant-interpretation)[7,9,10].

Variants were included on the cGS report if they met one of the following criteria: (1) VUS/LP/P in a dominant gene related to the patient's phenotype, (2) VUS/LP/P biallelic variants in a recessive gene related to the patient's phenotype, (3) monoallelic VUS – Favor Pathogenic/LP/P in a recessive gene related to the patient's phenotype, (4) LP/P variants in a gene related to a documented family history of disease, or (5) LP/P variants in secondary findings genes. All clinically reported sequence variants were confirmed via Sanger sequencing.

Supplementary Figures

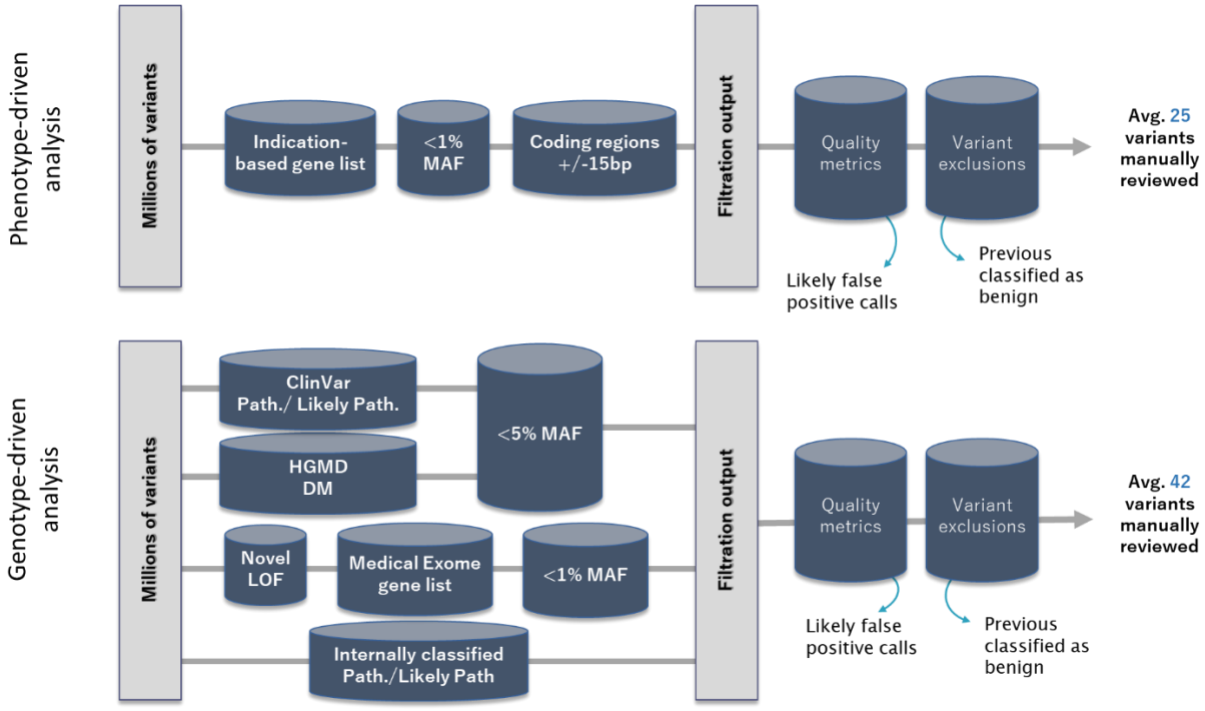


Figure S1. cGS analysis strategy. Each case underwent a two-tiered filtration strategy composed of both genotype-driven and phenotype-driven analyses. Top: Phenotype-driven analyses were designed to capture all rare variants in genes relevant to the patient phenotype. Relevant gene lists were manually curated for each indication from literature and database searches, and varied in size depending on the phenotype. Bottom: Genotype-driven analyses were designed to capture all highly suspicious variation in a patient's genome, including previously published disease-causing variants found in HGMD[11] and ClinVar[12], loss-of-function variants in Medical Exome genes (a custom-generated list of ~5000 genes, which was designed to capture all genes that have been reported in association with human disease), and all internally classified pathogenic and likely pathogenic variants. All variants returned by these filtration criteria were reviewed for disease causality and relevance to the patient phenotype. Abbreviations: DM – Disease-causing mutation; HGMD – Human Gene Mutation Database; LOF – loss of function; MAF – maximum minor allele frequency in gnomAD: Path. – Pathogenic; Likely Path. – Likely pathogenic

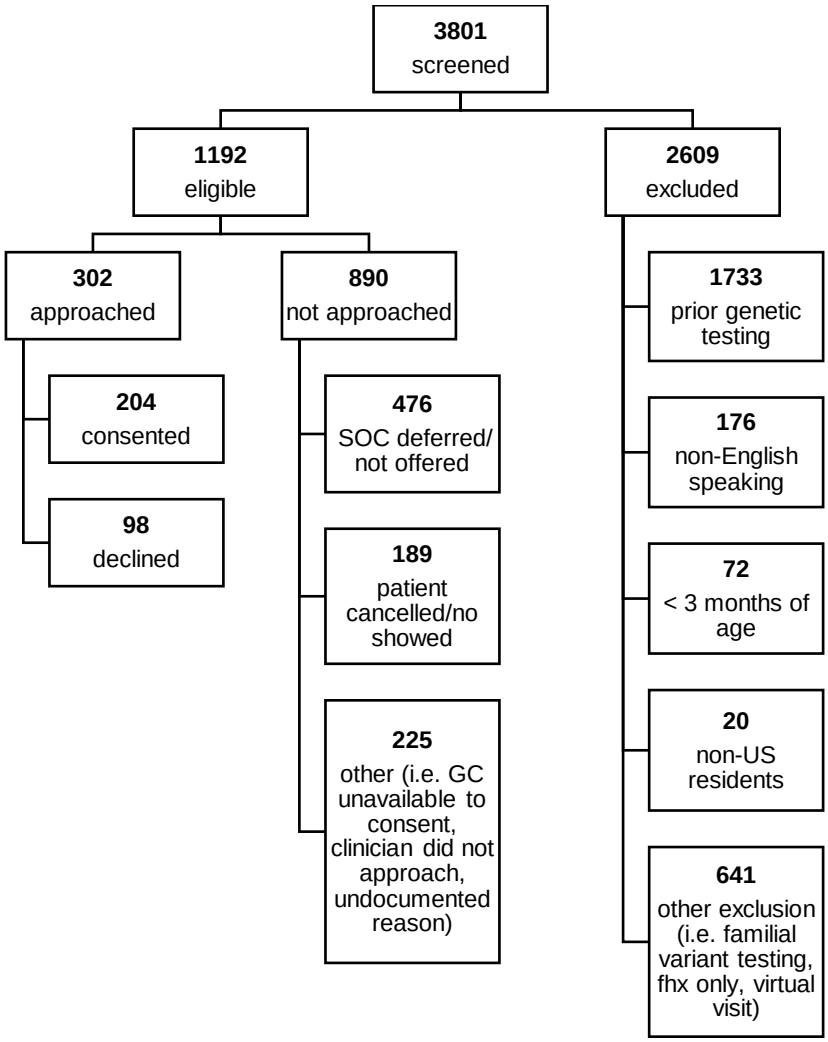
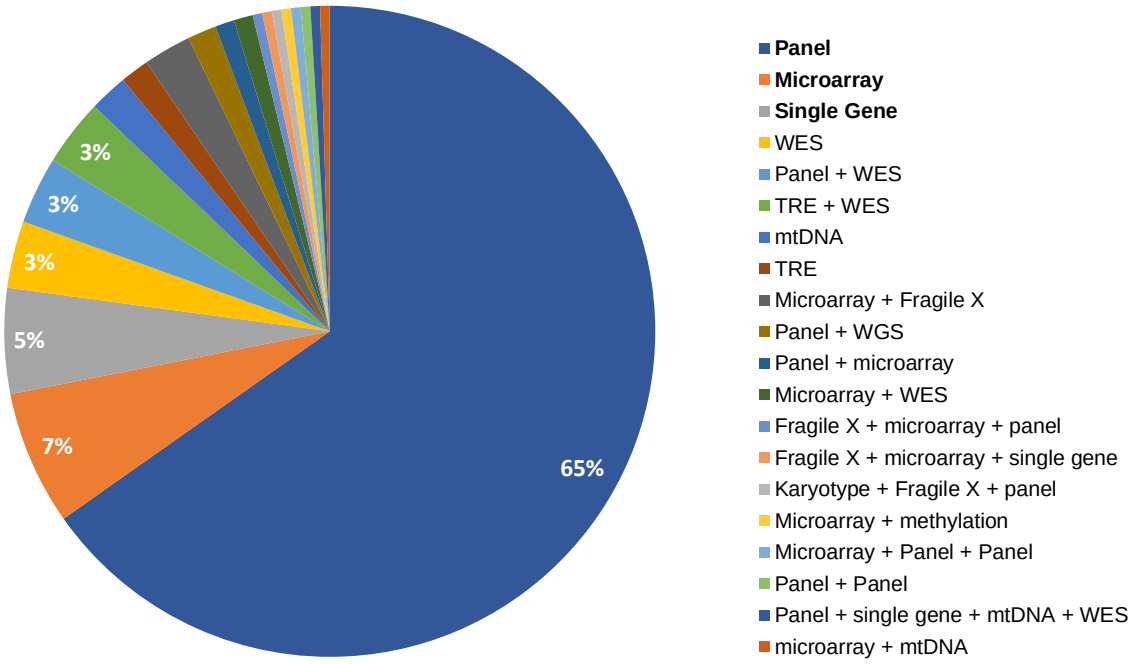
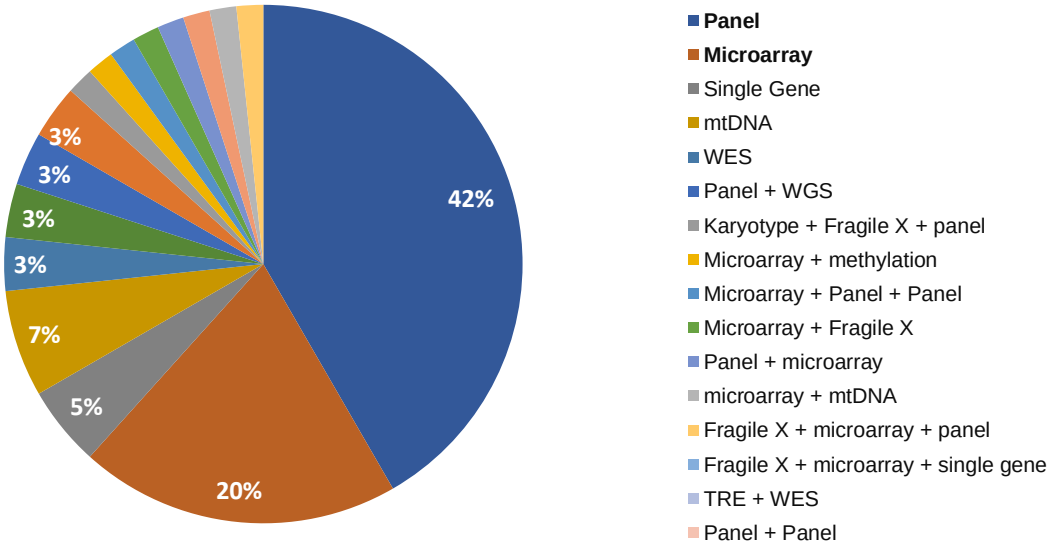


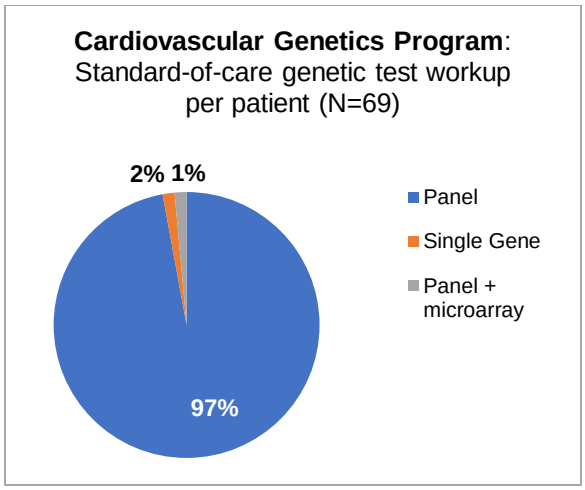
Figure S2. Study exclusion reasons

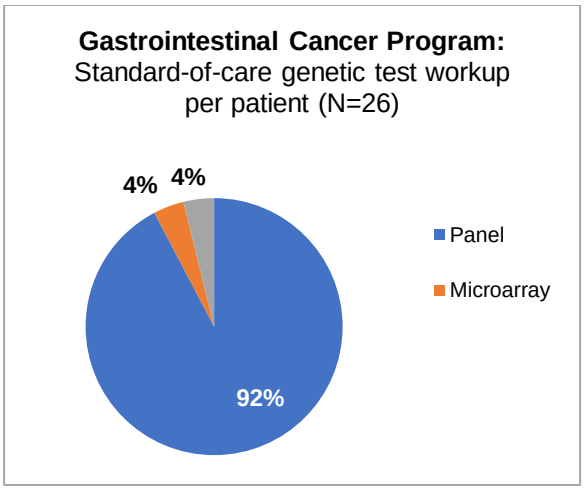
All participants:
Standard-of-care genetic test workup per patient (N=204)

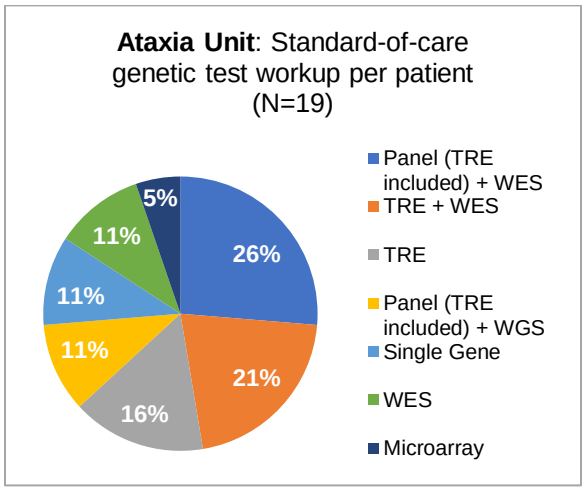


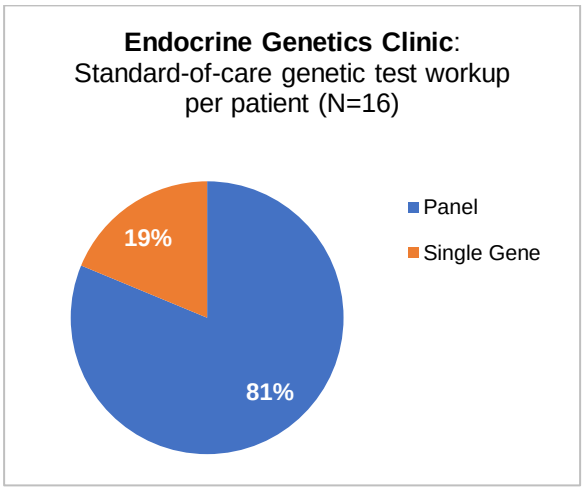
Medical Genetics and Metabolism Program:
Standard-of-care genetic test workup per patient (N=60)











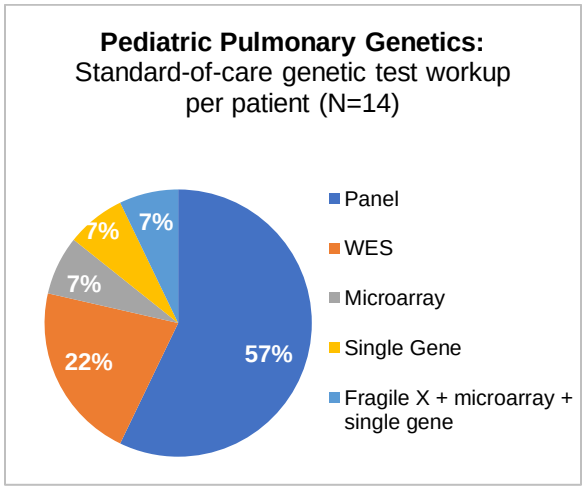


Figure S3. SOC genetic tests ordered for all patients enrolled

Supplementary Tables

Table S1. Eligibility criteria

Inclusion Criteria	Exclusion Criteria
Pediatric and adult (≥ 3 months of age)	Non-English speaking
Genetics evaluation and genetic testing ordered at MGH	Prior genetic testing for current referral indication
Have a *suspected genetic disorder* in which the genetic cause is unknown	
< 18 years of age [a]	

[a] *additional eligibility criterion to be enrolled as a trio*

Table S2. HPO terms by clinic. Across different clinics including both study arms, there were significantly different mean numbers of total HPO terms, primary HPO terms and body systems (P<0.001), but the mean number of non-primary HPO terms was not significantly different (P=0.2). This analysis was done using ANOVA.

	Mean Number of HPO Terms				
Clinic	**Total**	**Primary - phenotype**	**Non-primary - phenotype**	**No. Body Systems**	**No. Genes**
ATX	9.40	6.50	2.90	4.50	228.30
CGP	4.56	1.79	2.76	2.56	214.97
ETG	6.63	2.63	4.00	3.38	154.75
GIC	2.92	1.58	1.50	2.25	401.64
MGP	7.50	4.83	2.93	4.37	219.04
PUL	7.14	6.71	0.43	4.14	331.29

Abbreviations: Cardiovascular Genetics Program (CGP), Medical Genetics and Metabolism Program (MGP), Ataxia Genetics Unit- Neurology (ATX), Gastrointestinal Cancer Program (GIC), Endocrine Genetics (END), and Pulmonary Genetics Clinic (PUL)

References

1. Kalia, S. S. *et al.* Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.* **19**, 249–255 (2017).
2. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
3. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–2 (2016).
4. Kronenberg, Z. N. *et al.* Wham: Identifying Structural Variants of Biological Consequence. *PLoS Comput. Biol.* **11**, e1004572 (2015).
5. Klambauer, G. *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.* **40**, e69 (2012).
6. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–8 (2011).
7. Riggs, E. R. *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.* **22**, 245–257 (2020).
8. Mandelker, D. *et al.* Comprehensive diagnostic testing for stereocilin: an approach for analyzing medically important genes with high homology. *J. Mol. Diagn.* **16**, 639–47 (2014).
9. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–24 (2015).
10. Strande, N. T. *et al.* Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. *Am. J. Hum. Genet.* **100**, 895–906 (2017).
11. Stenson, P. D. *et al.* The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. *Hum. Genet.* **139**, 1197–1207 (2020).
12. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980-5 (2014).