Lions and tigers and bears, oh my!
The spectrum of technically challenging pathogenic variants in patients presents new requirements for genetic test development and validation

Stephen Lincoln[1], Rebecca Truty[1], Justin Zook[2], Birgit Funke[3,4], Catherine Huang[5], Jessica Dickens[5], Matthew Ghent[1], Shan Yang[1], Adam Rosendorff[1], Swaroop Aradhya[1], Marc Salit[2,6], and Robert Nussbaum[1,7]

[1]Invitae, San Francisco, CA; [2]National Institute for Standards and Technology, Gaithersburg, MD; [3]Laboratory for Molecular Medicine, Cambridge, MA; [4]Harvard Medical School, Boston, MA; [5]Seracare Life Sciences, Gaithersburg, MD; [6]Stanford University, Palo Alto, CA; [7]University of California, San Francisco

Summary

Clinical genetic tests are increasingly used but can be complex to implement: Many medically relevant genes are located in technically challenging regions of the genome. For example, 23% of potentially medically relevant genes have exons with high homology to another genomic region (e.g., a pseudogene, paralog, or other segmental duplication).[1] Furthermore, medically important but technically challenging classes of variation are also now well documented.[2]

The impacts of this on diagnostic yield and technology choices for clinical tests have not yet not been thoroughly described. In this study, we examine the prevalence and distribution of pathogenic (disease-causing) variants in commonly clinically tested genes as a function of the technological hurdles that must be overcome in order to accurately detect them. We find that "hard" variants are prevalent (~11%), and that simplistic laboratory approaches would thus likely have a significant false-negative rate.

Validating tests for these events can be challenging, in part because it is difficult to obtain positive control specimens. We explore one possible solution: creating synthetic standards by spiking variants into a known genomic background.

Methods

We examined 30,000 individuals clinically tested for physician-specified gene panels underlying a hereditary cancer, cardiovascular, neurological, or pediatric condition. A diverse set of targeted techniques was used to detect and to orthogonally confirm the presence of variants in the ordered genes:

- Illumina 2x150 NGS (Next-Generation Sequencing)
- Array CGH (Comparative Genomic Hybridization)
- Sanger sequencing of standard PCR and long-range PCR products
- PacBio (Pacific Biosciences) long-read sequencing
- MLPA (Multiplex Ligation-Dependent Probe Amplification)

The NGS methods used multi-technology hybridization-based target enrichment and five variant-calling algorithms, an update to methods described and validated previously.[2] Exons and flanking regions of 974 clinically relevant genes were targeted, as were intronic regions known to harbor pathogenic variants or breakpoints of inherited structural variants. These methods were shown to have high sensitivity through a painstaking process of sourcing specimens with rare hard variants from biobanks, clinics, and other clinical laboratories. Notably, these samples often have little DNA, which cannot be replenished. The classification of variants as pathogenic (including likely pathogenic) followed the ACMG 2015 guidelines.[3] We also compared the results of these diagnostic-grade tests to exome sequences from a variety of sources and to the Genome in a Bottle (GIAB) data sets.[4]

Genome in a Bottle

Until recently, the GIAB project provided reference sequence for a single widely used reference sample, NA12878. Though a tremendous asset, it had limitations:

- 23% of the NA12878 genome did not have high confidence calls, including all or part of the coding exons of 30% of commonly clinically tested genes.
- This 23% was biased, representing the technically challenging regions of the genome in which errors are most likely and validation is thus most critical.
- In clinical genes, NA12878 contains only relatively easy variants: SNVs and (a few) small indels. Some genes (15%) have no exonic variants of any type.

The GIAB v3.3 is an important advance, but hard variants remain lacking.

	Previous	Version 3.3
Samples with reference data	1[a]	5[a,b,c]
High-confidence regions (genome)	77%	83%[a], 76–79%[b,c]
High-confidence (clinical genes)	70%	86%[a], 75–84%[b,c]
CNV/SV data	No	In process[d]

a. NA12878; b. NA24385, NA24149, NA24143 (new trio); c. NA24631 (new individual)
d. These individuals are likely to be CNV negative in most or all clinically relevant genes.

Genome in a Bottle Consortium

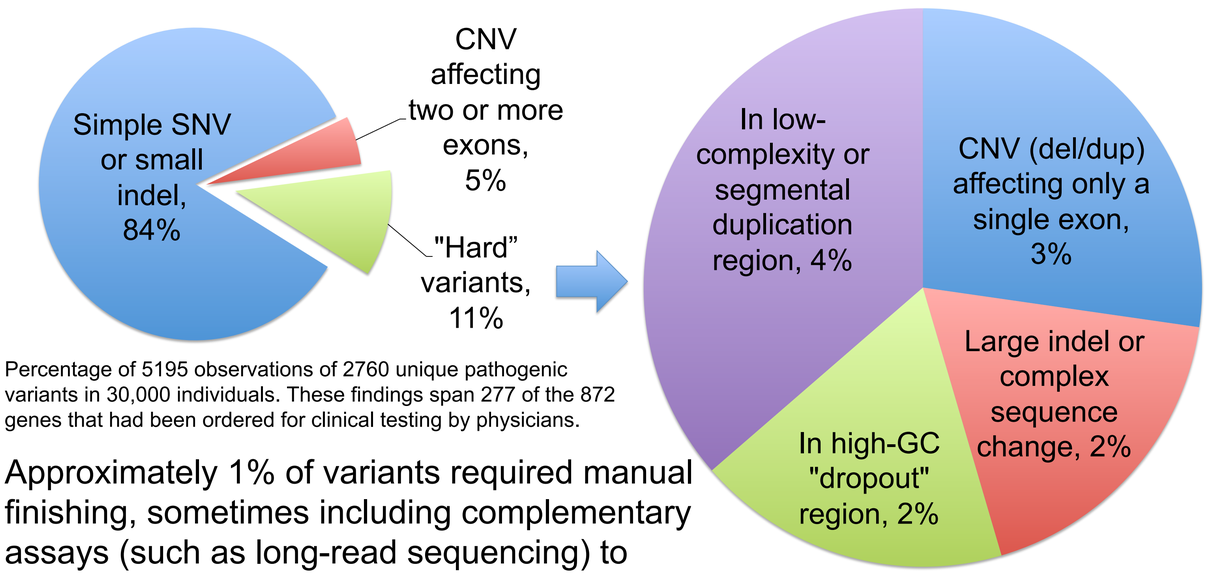
New!

Note: A further improved GIAB 3.3.1 will be announced at this meeting.

References: [1]Mandelker, *et al., Genet Med.* 2016. [2]Lincoln, *et al., J Mol Diag.* 2015. [3]Richards, *et al., Genet Med.* 2015. [4]Zook, *et al. Nature Biotech.* 2016. [5]Chong, *et al., PLOS One.* 2014. [6]Kang, *et al., PeerJ.* 2016. [7]Judkins, *et al., BMC Cancer.* 2015. [8]Lih, *et al. J Mol Diag.* 2015. [9]Sims, *et al., J Mol Diag.* 2016. [10]Kudalkar, *et al., J Mol Diag.* 2016.

Prevalence of Pathogenic Variants by Type

Of the 5195 pathogenic variants identified in the clinically tested individuals, 11% belong to a technically challenging category not trivially addressed by targeted short-read NGS or microarray methods.



Percentage of 5195 observations of 2760 unique pathogenic variants in 30,000 individuals. These findings span 277 of the 872 genes that had been ordered for clinical testing by physicians.

Approximately 1% of variants required manual finishing, sometimes including complementary assays (such as long-read sequencing) to resolve important details.

Off-the-shelf exome kits vary, but typically leave coverage gaps (exons <10x) in 5–10% of these genes, even at high average coverage (100–200x).

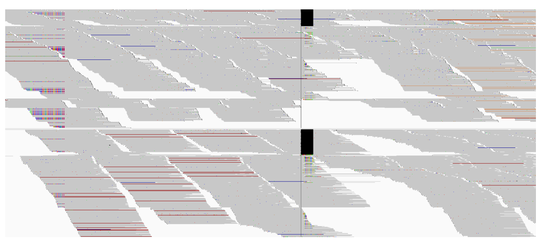
Validation of Genetic Tests

Most published validation studies unfortunately ignore this issue and include few, if any, examples of hard variants. For example, consider published validation studies of hereditary cancer tests from major U.S. laboratories:

	Chong *et al.*[5]	Kang *et al.*[6]	Judkins *et al.*[7]	Lincoln *et al.*[2]
Single nucleotide (SNVs)	3010[a]	536[a,b]	3884[a]	501[a]
Small indels	**11 (0.4%)**	**? [b]**	**39 (1.0%)**	**156 (22%)**
Multi-exon CNV (del/dup)	**2 (0.07%)**	**? [c]**	**41 (1.0%)**	**16 (2.3%)**
Single-exon CNV (del/dup)	**2 (0.07%)**	**? [b,c]**	**8 (0.2%)**	**13 (1.8%)**
Large indels, complex change	**0 (0%)**	**? [b]**	**0 (0%)**	**19 (2.7%)**

a. The vast majority (>95%) of SNVs in studies 5-7 were benign polymorphisms. Fewer (85%) were included in study 2.
b. Variant types are not separated clearly, but the number of "hardest" events is likely to be very small based on the study design.
c. Sixty altered exons summed across a much smaller number of events and individuals – the actual number is not stated.

Admittedly, sourcing positive controls for validation studies is challenging. We explored the use of synthetic controls created by spiking plasmids with large (>1000 bp) engineered inserts into GIAB sample NA24385. These inserts contained variants that appear heterozygous when sequenced. This approach was previously used to validate a somatic gene panel for the NCI-MPACT clinical trial[8,9] and a pilot control containing 10 cardiomyopathy variants was recently described by some of us.[10] Targeted NGS data for these synthetic variants mimic that of the endogenous events. Allele balances are similar, although coverage at these sites is artificially doubled and some new artifacts are present. These problems can be addressed by post-processing the BAM files. A similar sample containing 20 "hard" variants in cancer genes is currently undergoing testing.



Indel variants	Allele balance	Strand balance
NM_000256.3:c.2373_2374insG	0.49	0.51
NM_000363.4:c.532_534delAAG	0.50	0.61
NM_001001430.2:c.487_489delGAG	0.51	0.46
NM_000256.3:c.3628-41_3628-17del26	**0.28**	0.43

Allele balance is also skewed for the endogenous del26 variant (0.30–0.35), presumably due to targeting and mapping bias.

Conclusion

Technically challenging variants are rare, but they represent a disproportionate fraction of pathogenic and potentially medically actionable findings in genes that are commonly tested in clinical practice today. Test validation, particularly the measurement of sensitivity, for such variants is challenging and, unfortunately, sometimes ignored. Thus, these tests may have high clinical false-negative rates. The use of synthetic controls based on plasmid spike-ins appears to be a viable method with which to benchmark such tests. Other approaches, such as CRISPR editing of immortalized cell lines, may also be valuable, particularly for small copy number variants for which the spike-in approach is not applicable.