Disease Diagnosis by Exome Sequencing: Issues Hampering Causative Variant Discovery

Personalis, Inc., Menlo Park, CA

Gemma Chandratillake, Sarah Garcia, Mark Pratt, Gabor Bartha, Shujun Luo, Richard Chen, John West

Contact
gemmac@personalis.com

Introduction

Exome sequencing is increasingly utilized in clinical genetics practice to diagnose cases where other genetic testing has proven futile or cost-inefficient. However, the limitations of next-generation sequencing technologies, particularly with respect to commonly utilized "off-the-shelf" exome enrichments, result in poor coverage of certain disease-causing mutations. Several issues also exist in variant identification and annotation.

Exome sequencing misses exons

Despite prevalent use of the term "whole exome sequencing", exome enrichment kits do not provide coverage over all exonic content. Several genes are completely absent from exome sequences, and many more genes with clinical relevance are only partially covered. Reasons for poor coverage include lack of targeting by exome platforms, and high guanine/cytosine content, which is particularly prevalent in first exons.

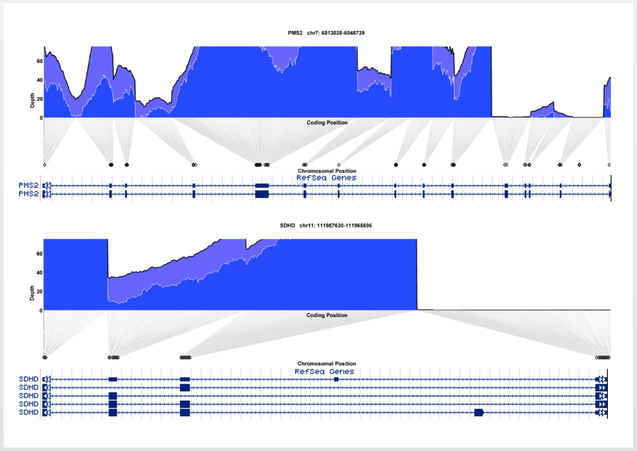


Figure 1: Incomplete coverage of genes with standard exome platforms.

Plots show incomplete coverage of exonic sequence of two genes from the set recommended by ACMG for reporting of incidental findings:
Top panel: *PMS2*
Lower panel: *SDHD*
Black outline: mean coverage
White outline: one standard deviation below mean coverage.

Diamonds indicate biomedically annotated variants from the Personalis Disease Variant Database

Exome sequencing misses non-coding variants

By design, exome enrichments primarily target coding content and so disease-causing variants located in UTRs, intronic, promoter, and intergenic regulatory regions are missed. One example is the intronic variant of *CFTR*, 3849+10kbC>T, recommended by ACMG for cystic fibrosis carrier screening. This variant, rs75039782, is located at chr7:117280015, indicated by the dotted line in Figure 2. The deep intronic position of this variant means that it is not covered by standard exome sequencing platforms, illustrated in this figure by a lack of corresponding reads at this position.

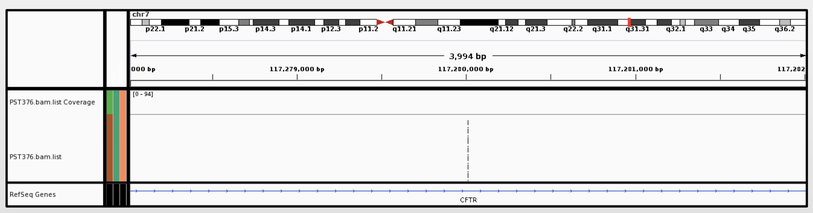


Figure 2: Lack of coverage of CFTR variant

Exome sequencing misses structural variants

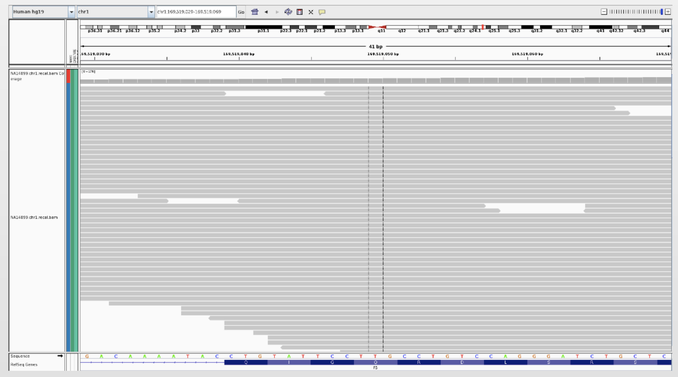
Tools for detecting structural variants have largely been developed for genome sequences and cannot be readily applied to exomes. Therefore, most exome sequencing services limit variant identification to SNVs and small indels thereby incompletely interrogating genes for mutations, leading to false negative results.

Presence of disease-causing alleles in the reference sequence

The public reference genome sequence (GRCh37) contains minor alleles at >1 million positions. The presence of minor alleles in the reference negatively impacts both sequence read alignment and variant calling. For example, an individual who is homozygous for a minor allele present in the reference will not be reported as variant at that locus, resulting in failure to apply any medical interpretation relevant to that variant to the individual. Examples of such disease-associated minor alleles present in the reference sequence are: rs4784677 in *BBS2* associated with Bardet-Biedl Syndrome, rs1529927 in *SLC12A3* associated with Gitelman Syndrome and hypertension, and the factor V Leiden allele, rs6025. Figure 3 illustrates one such example.

Figure 3: Effect of presence of F5 Leiden allele in the reference sequence

IGV plot of aligned sequencing reads from an individual homozygous for the *F5* Leiden allele. The deleterious C>T allele (G>A on coding strand) is not recognized as a variant since it is consistent with the reference sequence.



Inconsistent variant nomenclature

The identification of variants relevant to diseases and traits in a genome is dependent on the ability to annotate variants against a database of known variant-phenotype relationships. This requires the unambiguous mapping of variants described in the primary literature onto a modern, systematized coordinate system.

The primary literature, however, does not always provide sufficient information to make unambiguous mapping of some previously published variants possible. Despite the development of a standardized coordinate system, historical or legacy names for variants often persist in practice, making queries of databases and primary literature difficult. Figure 4 illustrates one such example.

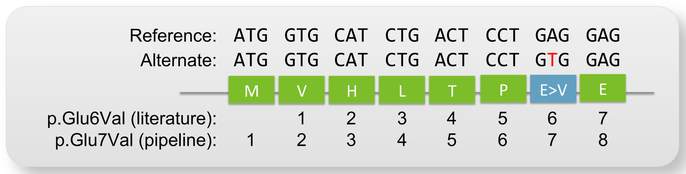


Figure 4: Differences in the Nomenclature for the Sickle-Cell Mutation in HBB

Indels are particularly difficult variants to map; modern variant-calling pipelines generally "left justify" indel calls where multiple variant coordinates are equivalent. However, this convention has not been adopted by the literature, or by the commonly used public variant databases. Figure 5 illustrates one such example.

Lack of reporting of "no-calls"

Generally positions not confidently called are treated as non-variant, i.e. no distinction is made in reporting between positions that are the same as the reference sequence, and sites for which the base was not able to be determined. This leads to false negative results.

Conclusion

Clinicians should take account of the limitations of exome sequencing when considering testing strategies for their patients and be proactive in seeking such information from laboratories offering these tests.

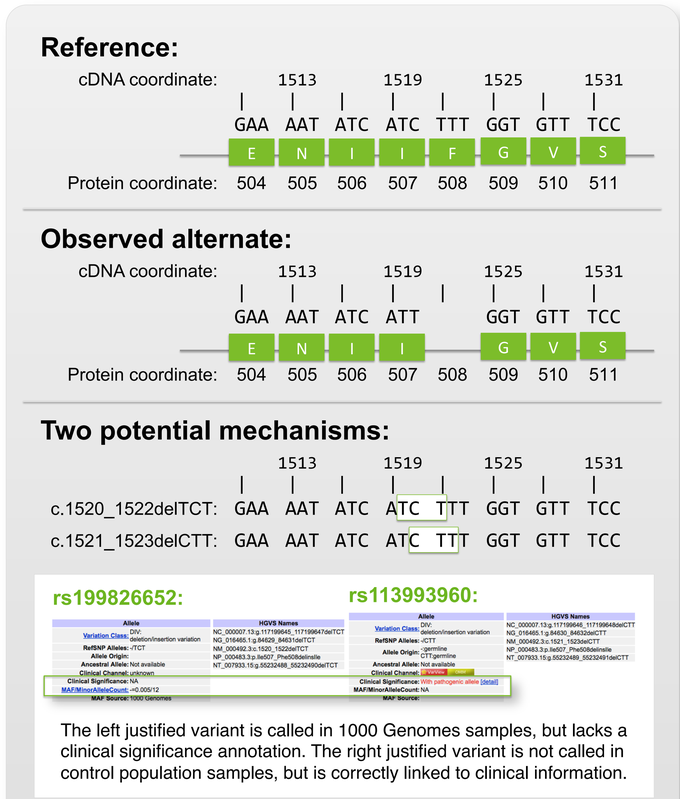


Figure 5: Ambiguity in Mapping & Annotation of CFTR ΔF508

Further Information

- Sarah Garcia et al. **"Finding the Clinical Answer in Genomic Sequence: Narrowing the Search Space for Disease-Causing Mutations"**. Poster #94.
- Gemma Chandratillake & Sarah Garcia: **"Issues Hampering Exome/ Genome Interpretation in Diagnostic and Predictive Medicine"**. Personalized Medicine SIG meeting, Thurs 12.30pm 208-B.
- **Personalis Booth #124**