Recognition of disease-associated alleles in the reference sequence is critical for accurate risk assessment through genome sequencing

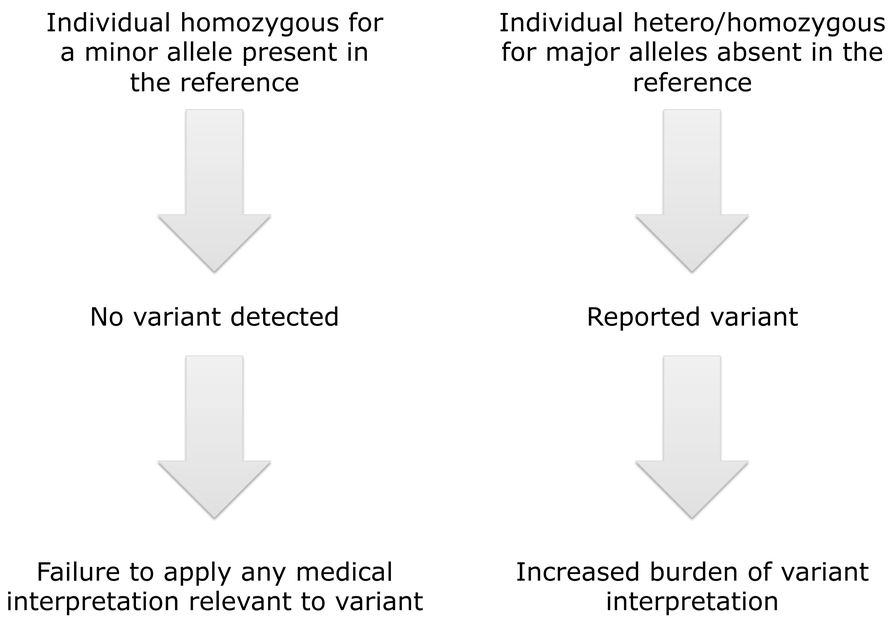
Contact:
gemma.chandratillake@personalis.com

Personalis, Inc., Menlo Park, CA

Gemma Chandratillake, Anil Patwardhan, Sarah Garcia, Michael James Clark, Stephen Chervitz, Daniel Newburger, Hugo Lam, John West, Richard Chen

Inaccuracies with the Public Reference

The public reference genome sequence (GRCh37) contains minor alleles at >1 million positions. The presence of minor alleles in the reference impacts the detection of variants in an individual and in many cases negatively affects the genetic and medical interpretation of their results.



An Enhanced Reference Sequence

Leveraging both public and proprietary data sources, we identified medically relevant minor allele positions in GRCh37, where the reference allele is the minor allele by frequency (via 1000 Genomes) in four different HapMap populations. These positions include those variants previously associated with Mendelian disease, pharmacogenomic response, and complex disease.

Among the variants identified were:

38 variants in GRCh37 previously reported to be involved in Mendelian disease (HGMD designation DM/DM?)

- e.g. rs4784677 in *BBS2* associated with Bardet-Biedl Syndrome
- e.g. rs1529927 in *SLC12A3* associated with Gitelman Syndrome

217 variants designated disease-associated polymorphisms with functional evidence (DFP)

- e.g. rs6025, the Factor V Leiden allele.

4 presumed-deleterious alleles (nonsense, frameshift, splice-site) in HGMD genes

- e.g. rs276936 in *DSC3*
- e.g. rs9959632 in *PIGN*.

77 variants with pharmacogenetic associations listed in PharmGKB

- e.g. rs1954787 involved in citalopram response

985 variants associated with complex disease in the Personalis Disease Variant Database.

While it is unlikely that all of these variants are pathogenic, they warrant in-depth review. Such variants would be completely missed in homozygous individuals and likely filtered out due to population frequency in heterozygous individuals due to recognition of the major allele as variant.

Extending previous work[1], we revised 1.1 million positions in GRCh37 where the reference-allele differed from the major allele. This enhanced reference sequence is used for alignment and variant calling both in our research discovery work and in our clinical diagnostics.

[1]*Phased whole-genome genetic risk in a family quartet using a major allele reference sequence, Dewey et.al., PLoS Genet. 2011 Sep; 7(9)*

Other Personalis Posters

Clark et. al., **2127W**, Weds., Oct. 23rd 10:30am
Li et. al., **1642T,** Thurs., Oct 24th 11:30am
Garcia et. al., **1550F**, Fri., Oct 25th 11:30am
Pratt et. al., **2608F**, Fri., Oct 25th 11:30am

Improved Accuracy- An Example

A sample homozygous for a variant in human F5 (the factor V Leiden allele), associated with an inherited disorder of blood clotting, was purchased from the Coriell Cell Repositories (http://ccr.coriell.org) and submitted for exome sequencing using the proprietary pipeline analysis at Personalis. Analysis was run against both the GRCh37 reference and our enhanced reference sequence.

Below are snapshots of Integrative Genomics Viewer (IGV) plots showing failure to detect the known Factor V variant (rs6025) when calling against the public reference genome, and its detection when using our enhanced reference sequence.

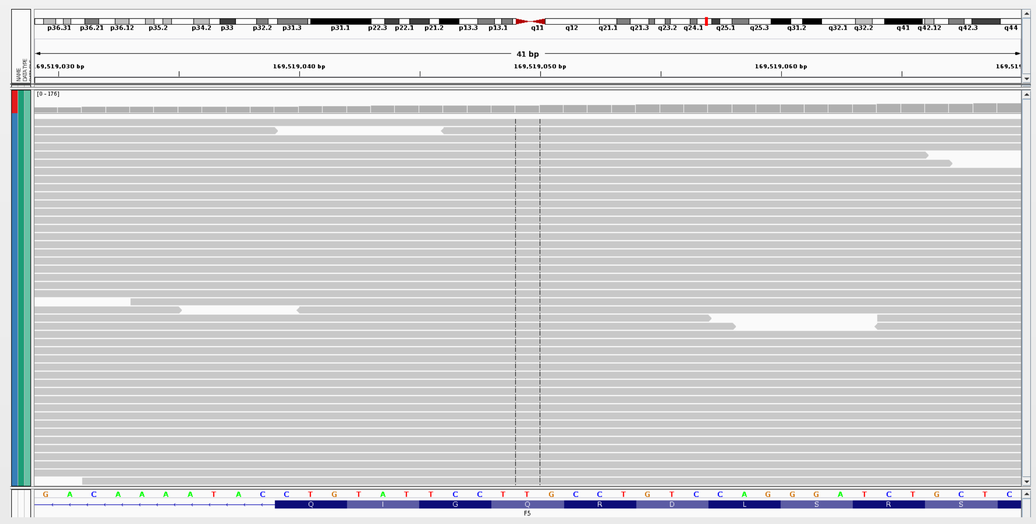


Figure 1: Using the standard GRCh37 reference, the known factor V allele is not identified as variant, due to the presence of the minor allele in the reference

Figure 2: In contrast, the use of the enhanced reference sequence, with a revised reference allele (T>C) based on pan-ethnic major allele frequencies, correctly resolves the variant (G>A on coding strand) in this homozygous individual.

Accurate and complete variant detection, combined with additional downstream annotation content, allows us to provide medically relevant interpretation for this individual. The Personalis pipeline provides structural and functional information for every variant identified through sequencing, with updates provided on a continual basis. Among the comprehensive list of content available, are curated variant-disease relationships, pharmacogenomics and pathways.

Position	Ref	Alt	Gene	dbSNP ID	HGMD	ClinVar	OMIMgene	Drugbank
Chr1:169514006	A	.	F5	..	..	..	..	..
Chr1:169519049	C	T	F5	rs6025	Thrombosis, increased risk; Altered FV1/FV2 ratio...Coronary artery disease...Venous thromboembolism...	Ischemic stroke...Thrombophilia due to activated protein c resistance; Factor V deficiency	...Thrombophilia due to activated protein C resistance...	Drotrecogin alfa; Phenylmercury
Chr1:169521553	G	.	F5	..	..	..	..	..

Table 1: The Personalis pipeline combines increased accuracy in variant calling with relevant structural and functional information drawn from over 40 curated public and proprietary data sources. For the rs6025 variant, an excerpt of the variant-disease and pharmacogenetic information is shown. Additional information (not shown) includes: structural and functional content for genes, non-coding RNA, and regulatory elements; multi-ethnic population frequencies; putative functional effects; proprietary pharmacogenomic information; and variant call-quality information.

Conclusion

The use of an enhanced human reference sequence is critical for accurate and complete variant detection in exome and whole-genome sequencing studies. Its use improves accuracy in both discovery studies and in an individual's disease-, carrier-, and pharmacogenetic-risk assessment.