Implementing an Augmented Clinical Exome and Reference Improvements to Enhance Diagnostic Yield and Discovery

Richard Chen, Deanna M. Church, Mark Pratt, Gabor Bartha, Jason Harris, Shujun Luo, Ming Li, Nan Leng, Anil Patwardhan, Steve Chervitz, Sarah Garcia, John West

Personalis, Inc. | 1350 Willow Road, Suite 202 | Menlo Park, CA 94025

Contact: richard.chen@personalis.com

Introduction

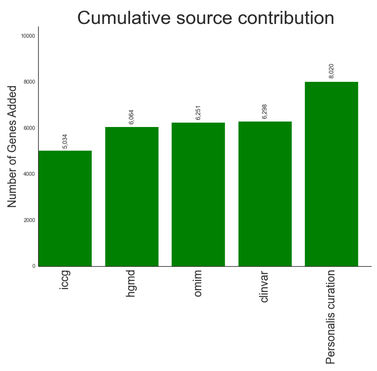
Clinical exome sequencing is increasingly used for solving diagnostic odyssey cases in children with suspected genetic syndromes and cancer. The complex process of going from DNA sample to clinical report involves multiple, technologically, scientifically, and medically complex steps; despite early success, significant improvements can be made to increase the overall diagnostic yield of exome sequencing tests.

To improve diagnostic yield, we have developed an augmented exome approach to increase gene finishing as well as sensitivity and specificity of exome sequencing to achieve clinical grade performance. Furthermore, we explore methods of enhancing the informatics pipeline performance through human reference improvements.

Methods

Augmented Exome Approach

Even when sequenced at high average coverage, exomes (and whole genomes) have poor actual coverage in many important regions, including those areas linked to Mendelian disease, complex disease, and pharmacogenomics. To address these issues, we have developed our third generation augmented exome approach that boosts accuracy and coverage in over 7000 genes that are medically relevant to Mendelian disease, cancer, and pharmacogenomics as referenced by multiple databases and literature sources. We developed ACE™ sample preparation, enrichment and sequencing protocols to address coverage shortfalls found in these content regions with standard exome sequencing protocols.



The FIGURE TO THE LEFT shows the content sources used to define the clinically important genes that in turn become the targets for augmentation in our augmented exome where we target over 7000 genes. In addition to augmenting exons in those genes. We also augment UTR regions for those genes, thousands of intronic variants, and 20bp into the introns in order to capture splice sites.

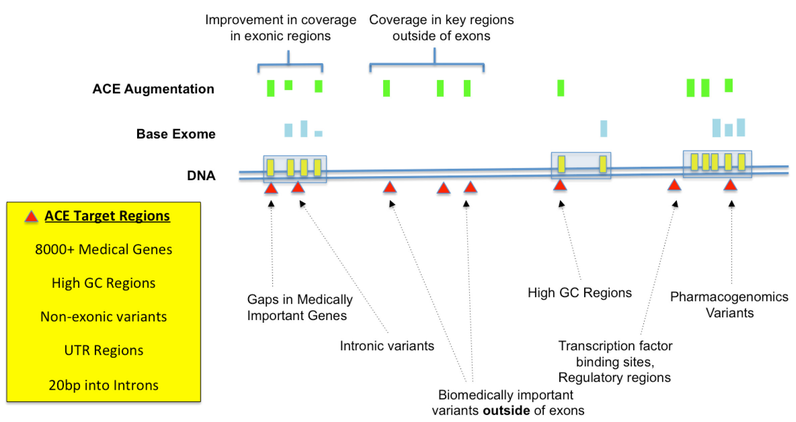
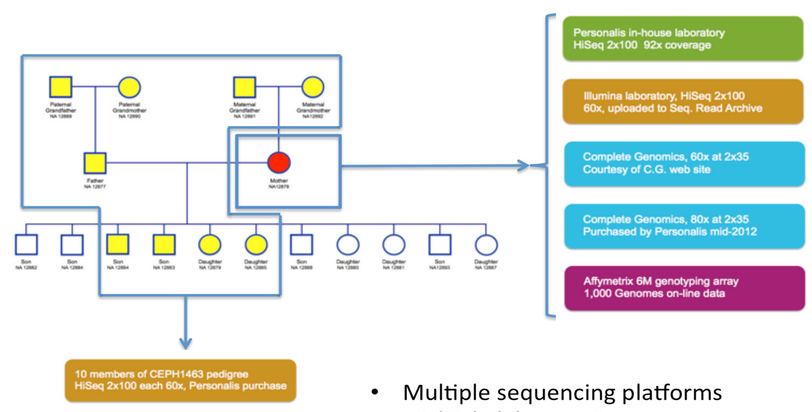


FIGURE ABOVE: Overall high GC regions are specifically targeted with optimized sample prep and specifically placed probes to fill in systematic coverage gaps.

The performance of this augmented exome was assessed using three methods: coverage over all exonic bases in our ACE exome assay compared to other standard exomes, comparison against the NIST standard genome, and examination of the structure and coverage of the 7000 medical genes in the latest well annotated reference assembly (GRCh37). The NIST v2.18 call set on NA12878 is the recognized genomic-scale accuracy standard. However there are some known limitations:

- The high accuracy call set is restricted to highly confident regions and excludes segmental duplications, CNVs, simple repeats and other challenging regions
- This call set covers 71% of the genome.
- Importantly, the high accuracy standard excludes large fractions of coding bases on genes of biomedical interest.



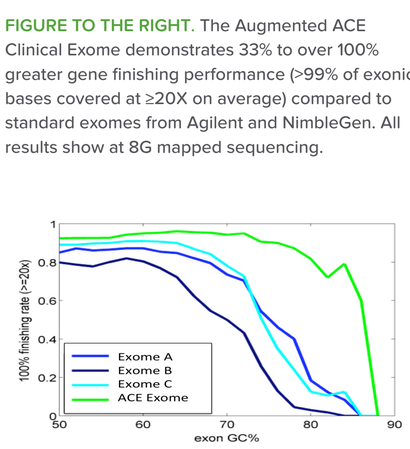
- Multiple sequencing platforms
- Multiple laboratories
- Multiple high coverage runs
- Three generation pedigree
- 3 Tbp raw sequence data

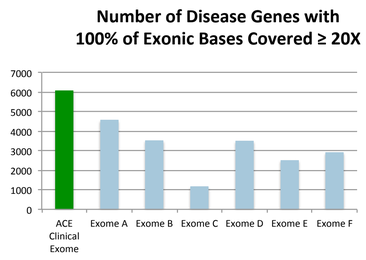
To address some of these issues, Personalis has created an internal gold standard by sequencing members of the CEPH pedigree to high depth in multiple labs, on different platforms, including difficult to sequence regions that are not included in the high confidence NIST standard (FIGURE ABOVE).

Results

ACE Augmented Clinical Exome Approach Achieves Greater Sensitivity and Clinical Grade Performance

FIGURE TO THE RIGHT. The Augmented ACE Clinical Exome demonstrates 33% to over 100% greater gene finishing performance (>99% of exonic bases covered at ≥20X on average) compared to standard exomes from Agilent and NimbleGen. All results show at 8G mapped sequencing.





The FIGURE TO THE LEFT shows improved coverage in GC rich regions due to augmentation in the ACE Clinical Exome compared to other standard exomes (Comparison at 8G Mapped).

NIST GiB High-Conf. Std. | Personalis Gold Std.

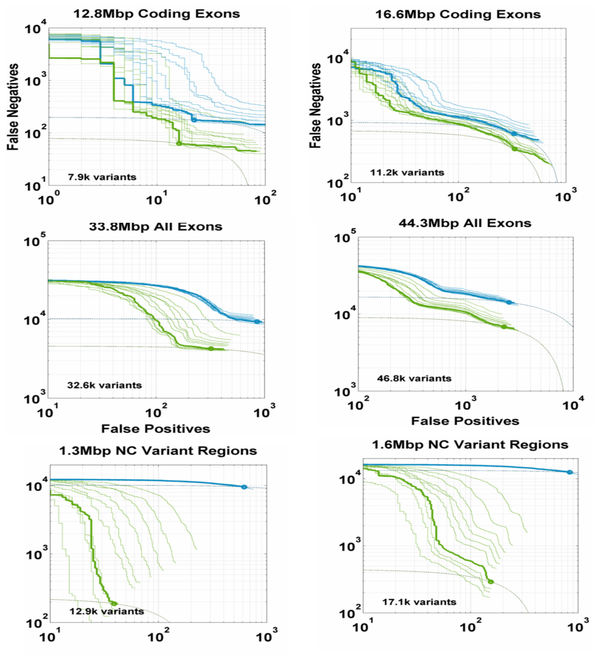


FIGURE TO THE LEFT: Exome accuracy on content regions. Shown are FP-FN error curves for SNVs and indels for two different protocols (Blue – standard exome, Green – Personalis ACE™ exome) for a range of total sequence ranging from 3-20Gb. The top panels measure accuracy against the NIST GIAB call set while the lower panels measure against our larger gold set. Genotyping and mis-characterization errors are included as False Positives.

Augmented Exome (ACE Exome) Fills in Gaps in Cancer Genes

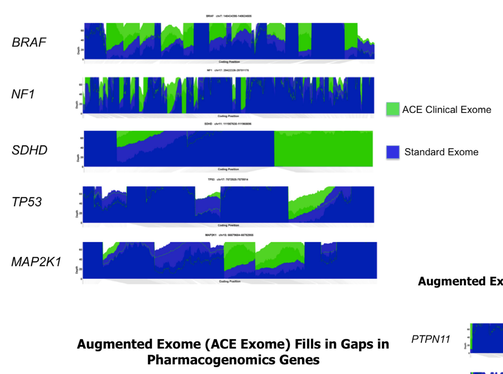
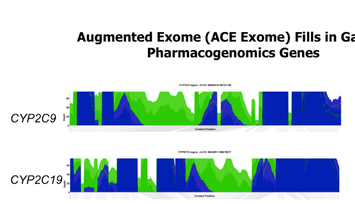
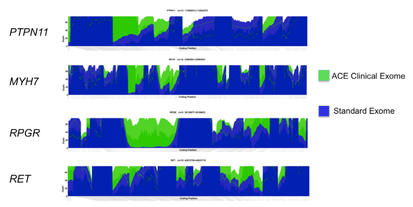


FIGURE TO THE LEFT AND BELOW show examples of how the augmented exome fills in gaps in critical cancer, Mendelian, and pharmacogenomics genes.

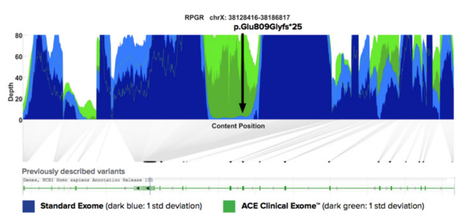
Augmented Exome (ACE Exome) Fills in Gaps in Pharmacogenomics Genes



Augmented Exome (ACE Exome) Fills in Gaps in Mendelian Genes

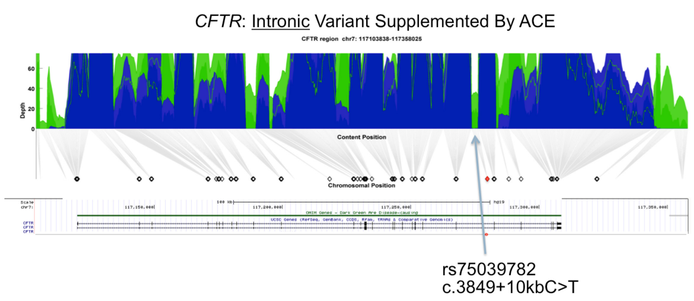


Impacts on Downstream Interpretation



EXAMPLE ON THE LEFT shows just one example of a pathogenic variant we detected in ***RPGR*** in one of our cases that would have been missed with a standard exome approach.

Augmented Exome Covers Pathogenic Intronic Variants



Deep intronic variant recommended for carrier testing by ACMG that is missed on a standard exome but captured with augmented exome sequencing.

Using Information from GRCh38 to Improve Exome Analysis

While we develop gold standard variant sets for GRCh38 and evaluate tools that allow us to take advantage of the entire assembly structure, we can use information in GRCh38 to improve our exome analysis.

Using Fix Patches to Improve Detection

The GRC releases quarterly patches after each major assembly release. There were 13 such patch releases for GRCh37. The fix patches in these releases provide corrections to problems in the main portions of GRCh37, and act as a preview of sequence that will be in GRCh38. Using the data we can recover exons in medically important genes that are not available in GRCh37.

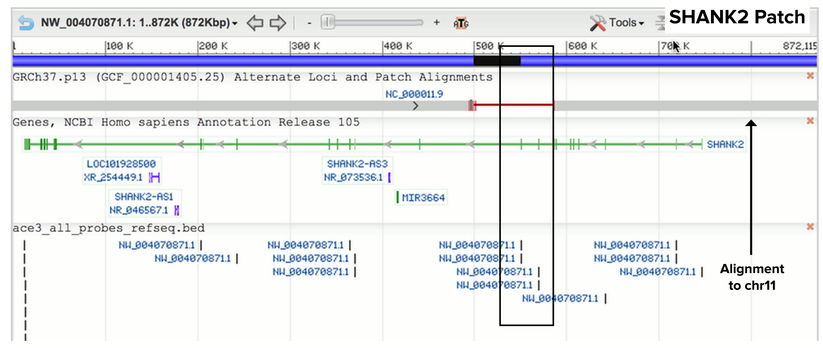


FIGURE ABOVE: A fix patch for GRCh37 containing the SHANK2 gene, an autism susceptibility was released. This sequences contains two exons (highlighted in gray). ACEv3 contains probes to these exons, allows us to interrogate these exons using our fix patch version of the reference assembly.

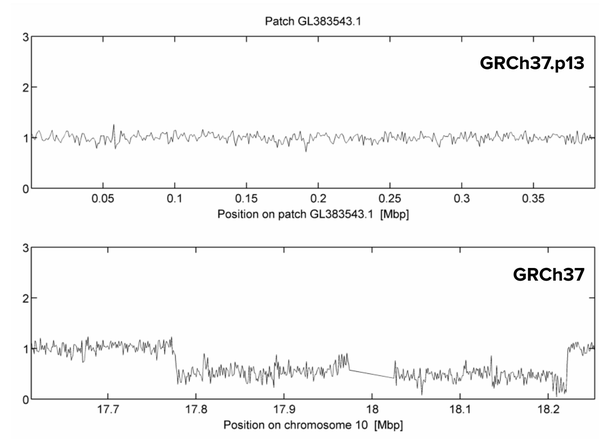


FIGURE TO THE LEFT: A whole genome sequencing analysis of NA12878 using our fix patch version of the reference assembly shows improved alignments in regions with fix patches, as shown in the normalized read depth plots above. The plots should be around 1, though we see depression of this in GRCh37, the alignments are improved in the fix patch version.

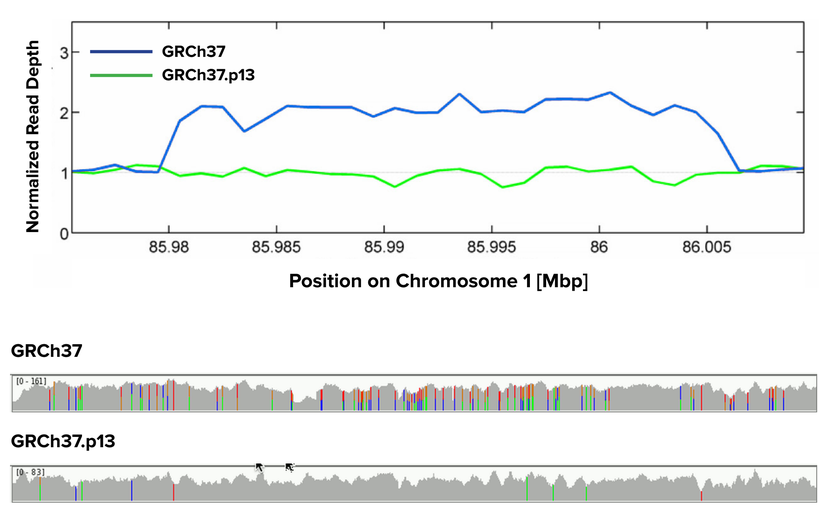


FIGURE ABOVE: Additionally, we can see improvements in regions outside of fix patch regions. The top graph shows the improvement in normalized read depth in a region of chr1. This translates into improved variant calling as seen in the IGV plots of this region in the lower panel.

Medically Relevant Gene Paralog Number Differences

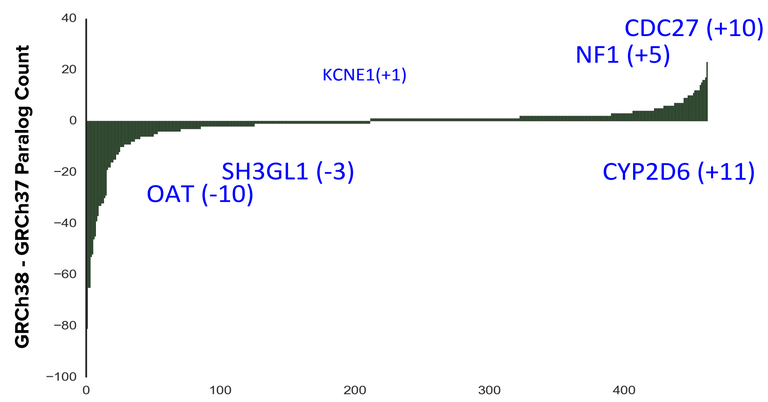


FIGURE ABOVE: Lastly, we can use GRCh38 annotation to improve interpretation on GRCh37 and its derivatives. The graph above shows the difference in the number of paralogs between GRCh38 and GRCh37 for almost 500 medically relevant genes. Genes that have gained paralogs are most concerning as they would be targets for off-target alignments in GRCh37, possibly leading to false positive variant calls. We can annotate high risk sites in GRCh37 to use as part of our genome interpretation.

Personalis®
Pioneering Genome-Guided Medicine