10 Common Misconceptions in Clinical NGS Sequencing

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

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

Introduction

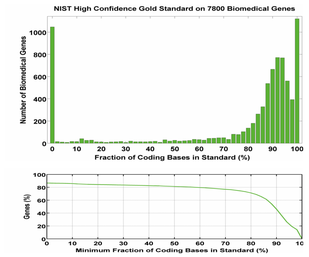
Whole exome and genome sequencing are increasingly used for clinical diagnosis of rare genetic syndromes yet clinicians often overlook critical issues with next-generation sequencing (NGS) that can impact diagnostic accuracy. Here we seek to raise, and dispel some of the important misconceptions regarding accuracy and the clinical application of NGS. These common misconceptions include: (1) NGS gold standards are gold (2) Exomes cover the whole exome (3) Standard exomes are sufficient for clinical use (4) Average depth of sequencing is an adequate predictor of quality and accuracy (5) Whole genomes are better than exomes for clinical applications (6) Coverage gaps in genes can be fixed by simply sequencing to higher depth (7) 10X coverage is enough to call variants (8) The human reference used has minimal clinical diagnostic impact (9) The sequencing, alignment/variant calling, and interpretation steps of the clinical NGS workflow can be optimized independently of each other (10) There is sufficient data in public databases for clinical variant interpretation.

Misconception #1: NGS Gold Standards are Gold

The NIST v2.18 call set on NA12878 is the recognized genomic-scale accuracy standard.

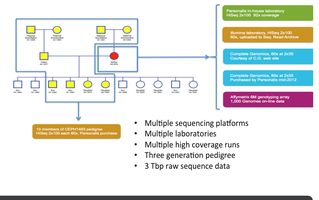
However there are some known limitations:

- 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.
- Half of genes of biomedical interest have 10% or more of their coding bases outside the high confidence gold standard.
- Structural variants are not currently included in the set.



Solutions:

- 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 below).
- Personalis is collaborating with NIST to develop a structural variant gold set as well.



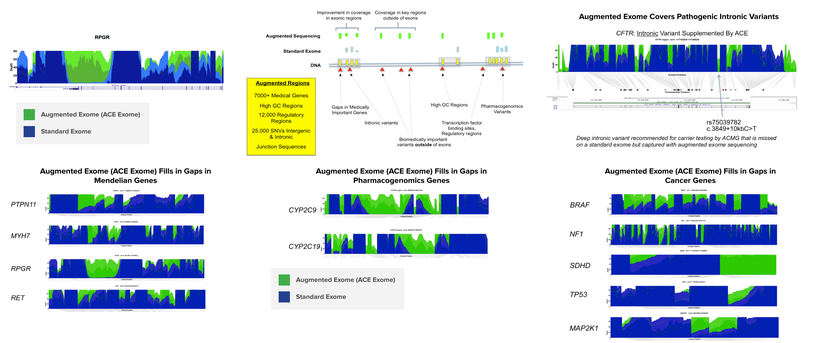
- Multiple sequencing platforms
- Multiple laboratories
- Multiple high coverage runs
- Three generation pedigree
- 3 Tbp raw sequence data

Misconception #2: Standard Exomes Cover the Whole Exome

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. Furthermore there is variation that can occur from run to run. The result is that even if a base is well covered on average, for any given sample, the coverage may be subpar.

Solutions:

Augmented Exomes can Improve Coverage Significantly. We have developed an Augmented Exome (ACE) that is a custom targeted pullout augmenting coverage in over 7000 medically important genes in Mendelian disease, cancer, and pharmacogenomics. It is also augmented to cover pathogenic intronic variants, UTR regions, and to improve SV detection.



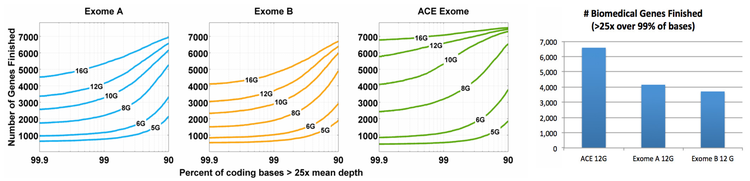
Misconception #3: Standard Exomes are Sufficient for Clinical Use

Given the systematic gaps that can occur in standard exomes as described previously, we measured the ability of standard exomes to "finish" the 7000+ medically important genes (Mendelian, cancer, and pharmacogenomics genes). The results of that analysis are shown below. In summary, at 12G of sequencing both standard exome platforms finish less than 50% of the medical exome genes.

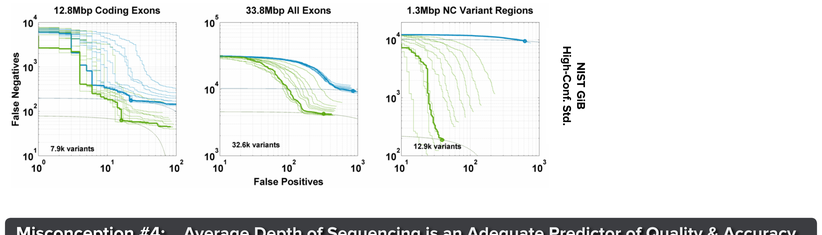
Solutions:

An augmented exome (ACE Exome) approach can significantly improve gene finishing. Compared to standard exomes, the ACE augmented exome finishes 50-80% more genes at comparable sequencing depth resulting in significantly improved analytical validity.

Shown below is the relationship between stringency of finishing criteria, total amount of sequence, protocol and number of genes finished. For example, at 12Gb the ACE protocol achieves 25x local mean depth on 99% of bases in 6,500 of the 7,800 genes versus 4,100 and 3,700 for the comparison protocols at the same 12G sequencing level.

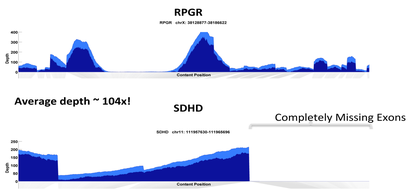


Shown are FP-FN error curves for SNVs and indels for two different protocols (Blue – standard exome, Green – Personalis ACE Exome™) for a range of total sequence ranging from 3-20Gb. We measure accuracy against the NIST GiB call set. Genotyping and mis-characterization errors are included as False Positives. Our standard clinical protocol of 12Gb is shown a bold line with the minimum error contour marked.



Misconception #4: Average Depth of Sequencing is an Adequate Predictor of Quality & Accuracy

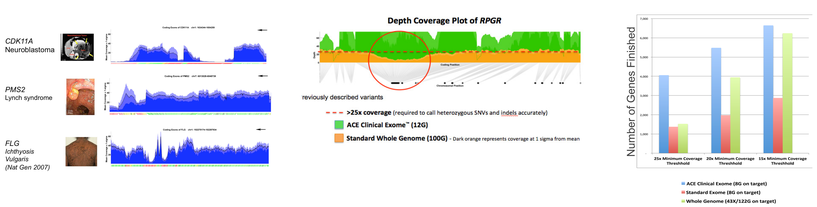
There is a common misconception that simply sequencing to high average depth (>100x) on standard exomes fixes sequencing gaps issues. However the problem is that coverage is uneven and can result in peaks with excess coverage and "coverage deserts" due to systematic errors in sequencing. In the two examples shown here, the coverage desert in RPGR is due to high GC. The gaps in SDHD are due to pseudogenes and changing reference transcripts. These are gaps that do not get filled with any amount of sequencing.



Misconception #5: Whole Genomes are Better Than Exomes for Clinical Applications

Whole genomes can suffer from the same systematic biases that cause gaps in exomes such as regions of high GC, etc. Also because sequencing to high depth on whole genomes can be costly, traditionally 30-60x exomes have been considered "clinical" grade. At this level of sequencing, whole genomes do not perform better at "finishing" the medical genes than a high depth augmented exomes such as ACE. See figure to the right. Furthermore, since augmented exomes focus on achieving high coverage of all clinically interpretable genes, it is unclear what additional data whole genomes provide that is clinically reportable at this time.

Whole genomes traditionally are seen as performing better than standard exomes in detecting structural variants. However augmented exomes such as ACE add additional targeted capture to detect structural variants at high sensitivity.

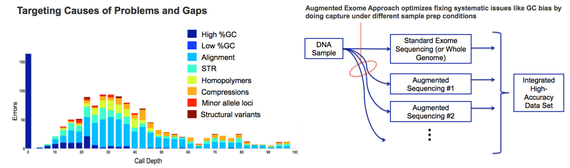


Misconception #6: Coverage Gaps in Genes Can be Fixed by Simply Sequencing to Higher Depth

Even at higher coverage, exomes and genomes can still have significant gaps. Many of these areas cannot be fixed by simply sequencing to higher depth because they are due to inherent limitations of the sequencing platform in regions of high GC content or repetitive sequence.

Solutions:

Augmented exome approaches can optimize sample prep and targeted sequencing designed to specifically fix systematic biases such as high GC, repeats, etc.

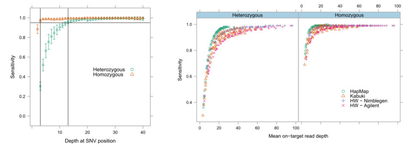


Misconception #7: 10X Coverage is Enough to Call Variants Clinically

Several papers have shown that heterozygous variants are not consistently and accurately called in regions with a local read depth below 13x to 20x coverage (Mynert *et al*, 2013). Meeting this threshold consistently can be a problem for standard exomes at 50x average coverage where one can expect over 20% of exome regions to fall below the 20x coverage threshold. This can also be an issue for whole genomes at the 30x "clinical" standard.

Solutions:

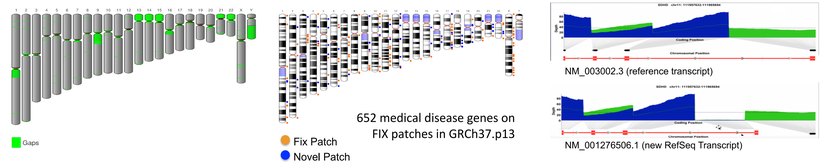
Establish clinical standard for NGS platforms requiring >15-20 minimum coverage across all exonic bases.



Source: Meynert *et al.* BMC Bioinformatics 2013

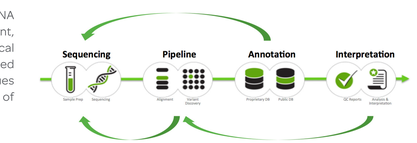
Misconception #8: The Human Reference Used has Little Clinical Impact

NGS solutions such as NGS panels, exomes, and whole genomes are sensitive to the human reference used for alignment. There are many gaps in the reference that are in various stages of "patching" in GRCh37 and GRCh38. For examples 652 medical disease genes are present on fix patches for GRCh37 alone. Migration to the newest reference GRCh38 will require significant changes to the informatics pipelines and annotation engines that are currently used with GRCh37.



Misconception #9: The Sequencing, Alignment/Variant Calling, and Interpretation Steps of the Clinical NGS Workflow Can be Optimized Independently of Each Other

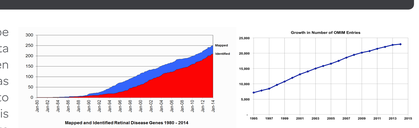
There is tremendous technical and scientific complexity in going from DNA to interpretation. This includes sample prep, NGS sequencing, alignment, variant calling, annotation, and clinical interpretation. Quality of final clinical interpretation can suffer when the components of the workflow are optimized independently from each other. Solutions to difficult NGS accuracy issues often require simultaneous and coordinated changes in multiple parts of the workflow to achieve an optimal solution.



For example, improving structural variant calling in augmented exomes involves simultaneous changes in the targeted capture method and the informatics pipeline that calls the variants. Downstream content for annotation can feedback to guide targeted capture for new disease causing genes.

Misconception #10: There is Sufficient Data in Traditional Public Databases for Clinical Variant Interpretation

Traditional curated sources such as OMIM and HGMD continue to be useful, but are unable to keep pace with astounding amount of data being generated in academia and laboratories. Furthermore, it has been estimated that in some of these resources the error rate can be as high as 20%. Data sharing among labs will be critical as well as continued efforts to structure data from the literature. Up-to-dateness of these databases is critical for improving diagnostic yield for patients. For example, shown are graphs that show examples of how quickly genes are being discovered in retinal disease and growth of OMIM.



Source: RetNet

Further Information from Personalis

Wednesday, June 11 @ 3:30 - 5:00 PM

Panels vs. Exomes: An Interactive Panel Discussion

Garcia *et al.*, The Clinical Exome: Personalis' Experience Using an Enhanced Exome and Genome- wide Structural Variant Detection for the Diagnosis of Diseases of Unknown Genetic Etiology

Church *et al.*, Impacts of Updating the Genome Assembly on Genome Interpretation

Contact:
richard.chen@personalis.com

Table 5

POS-CMRC-1.0 June 2014

www.personalis.com | info@personalis.com | +1 855-436-6634 | +1 650-752-1300

Personalis®
Pioneering Genome-Guided Medicine