Performance of the Helix Exome+® Assay

Helix's proprietary Exome+® assay is a panel-grade clinical exome enhanced by ~300,000 informative non-coding regions. Due to its custom design and proprietary bioinformatics solutions, it enables both clinical return of results and supports research applications with:

- Comprehensive and highly uniform coverage (≥ 99.5% call rate at ≥ 20x for clinically relevant regions)
- Clinically-validated intragenic and multigenic CNVs (100% sensitivity for ≥ 2 exons)
- Clinically-validated star allele calls for pharmacogenetic regions (accurate detection of > 100 CYP2D6 star alleles)
- Array-equivalent genome-wide imputation of tens of millions of high-confidence SNPs for discovery and polygenic risk scores
- Inclusion of the full mitochondrial genome

Clinicians and researchers are able to get the benefits of a targeted panel, the breadth of a microarray, and the completeness of an exome— all from one sample and one assay. The discovery and analysis of rare and novel variants, genome-wide imputation, polygenic risk score calculation, ancestry inference, replication and stratification of GWAS findings, and more are all delivered by the Exome+.

The Helix Exome+ assay is run exclusively at the Helix's CLIA and CAP accredited laboratory facility in San Diego, CA (CLIA #05D2117342, CAP #9382893). The Helix Laboratory is a highly automated facility with the ability to process millions of Exome+ assays annually. Our assay validation process adheres to guidelines from the College of American Pathologists (CAP)[1], and the Nex-StoCT workgroup for Standardization of Clinical Testing by NGS[2]. Our validation studies include DNA from saliva samples and well-characterized cell lines.

This paper details the performance characteristics of all variant types reported from the Helix Exome+ assay (small variants, copy number variants, CYP2D6 star alleles, and imputation).

Small Variants

Small variants are generally defined as single nucleotide changes and small insertions/deletions (indels) < 20bps. We evaluated the performance of our assay against public reference materials from the Platinum Genomes[3] and the National Institute of Standards and Technology (NIST) Genome in a Bottle (GIAB)[4] datasets. Exome+ replicates were generated for CEPH samples NA12877 and NA12878, an Ashkenazi Jewish trio (NA24385, NA24149, NA24143), and a Han Chinese sample (NA24631)[5] obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research. Exome+ data were compared with high confidence calls from the Platinum Genomes and GIAB datasets. Ten samples with previously identified variants were also included. These data allowed the evaluation of variants within the Helix analytical and reportable range, including SNVs, insertions ≤ 20 bp, deletions ≤ 20 bp, multiple nucleotide variants (MNV), substitutions, and complex variants (Table 1).

Table 1: Summary performance metrics for small variants on the Helix Exome+ assay.

Variant Type	Sensitivity	Specificity
SNV	99.9%	99.9999%
Deletion	99.9%	
Insertion	99.8%	
MNV	99.7%	
Complex	99.2%	
Substitution	99.6%	

Sensitivity: (True Positives / (Total Positives + False Negatives)).
Specificity: (True Negatives / (True Negatives + False Positives)).

The Exome+ delivers high-confidence variant calls for ≥ 99.5% bases across the ~ 600 genes most relevant to the proactive genetic testing.

Comparison to Sanger Sequencing

Sanger sequencing has long served as a gold standard method to confirm small nucleotide variants (SNVs) and indels. We conducted validation studies to compare Exome+ assay performance against saliva samples with Sanger sequencing and against cell lines with documented events in Coriell[6]. Analysis included 1,236 samples (1,141 sourced from saliva and 95 from cell lines) with 1,251 variants and 172,711 reference sites, and demonstrated > 99.9% concordance between Exome+ and Sanger or documented truth[6] (see Table 2).

Table 2: Concordance of NGS and Sanger results for salivas and documented truth for cell lines, broken down by variant category.

Variant Type	Count of Genotypes	True Positives	True Negatives	False Positive[1]	False Negative[2]	Percent Agreement
SNV	978	977	--	--	1	99.897751
Deletion	218	218	--	--	0	100.00000
Insertion	55	55	--	--	0	100.00000
Reference	172,711	--	172,710	1	--	99.999421

[1] There was one False Positive where Sanger suggests a homozygous reference result captured from sequence 278 bases upstream of the variant of interest. The Exome+ results suggest a heterozygous call at chr5: 177992754A>C or G (rs759632048) in the PROP1 gene, consistent with a known SNP at this location.

[2] There was one False Negative, where Sanger sequencing identified a variant in the gene HGD to be heterozygous (C/T) whereas the Exome+ assigned it as homozygous reference (C/C).

Copy Number Variants (CNVs)

Although CNVs are less common than SNVs and indels, they can similarly impact predisposition to disease [7]. Pathogenic CNVs might affect entire genes or might span only parts of genes. Exon-level copy number is reported across the Exome+ genes, with performance as described in Table 3.

Table 3: Summary performance metrics for CNV on the Helix exome+ assay.

	Result
CNV Sensitivity	98.5%
CNV Sensitivity, ≥ 2 exons	100%
CNV Specificity	100%
CNV Call Rate	99.9%

Sensitivity was evaluated across 44 samples carrying 45 documented CNVs of varied size, ranging from single-exon CNVs to multi-genic CNVs. Samples were replicated across different runs, resulting in 108 total data points. Of the CNVs that were missed, all spanned only one exon, though the majority of single-exon events were in fact detected (19 of 27 single-exon events measured were accurately identified). Specificity calculations were based on 45 samples across 116 replicates. No unexpected CNVs were identified in this sample set. CNV Call Rate was determined based on 351 samples. In order to assign copy number, we require a quality score *callQuality* ≥ 20. Where callQuality < 20, the CNV Target is no-called, reducing the CNV Call Rate.

To understand the frequency of CNVs identified across a relevant section of the exome, we counted the number of CNV Events identified across 59 medically actionable genes[9], across eleven CDC Tier 1[10] genes, and across four Familial Hypercholesterolemia (FH) genes using 27,513 production samples that pass CNV QC on the Exome+ (Table 4). This frequency of CNVs found in the general population is consistent with what has been reported in the past[11].

Table 4. Characteristics of CNV Events in clinically relevant genes. Description of the frequency of CNV events identified across 27,513 samples overlapping 59 medically-actionable genes, CDC Tier 1, or FH.

	Medically Actionable (59 Genes)	**CDC Tier 1 (11 genes)**	**FH (4 genes)**
% of samples carrying a CNV	0.90%	0.12%	0.03%

In addition to coding regions, some non-coding events are captured as part of the CNV output, including promoters for APC, BMPR1A, LDLR, PTEN, GREM1.

CYP2D6 Star Allele Typing for Pharmacogenomics

The defining variants used to infer the star alleles for most pharmacogenomics (PGx) genes are accessible as SNPs and indels from the Helix Bioinformatics Pipeline, with performance metrics described in Table 1. As an exception, CYP2D6 is outside of the reportable range of the Helix Bioinformatics Pipeline due to complications resulting from common recombination with the adjacent pseudogene CYP2D7[12]. Instead, CYP2D6 star alleles are delivered by a specialized analytical tool, the proprietary Helix PGx Pipeline.

To assess performance of CYP2D6 typing by the PGx Pipeline, 153 samples with documented CYP2D6 genotypes were run with a subset replicated both within and across runs. For this sample set, all star alleles were classified as either simple alleles or complex alleles, where simple alleles were defined as those determined by SNPs, indels, or whole gene deletions and either present in only a single copy or as two copies but in the absence of any other simple alleles. Complex alleles, in contrast, represented CYP2D6/CYP2D7 gene hybrids, star alleles that were duplicated in the presence of other simple star alleles, or star alleles with three or more copies. Performance characteristics were evaluated independently for simple alleles and complex alleles due to the inherent increased difficulty in identifying complex alleles. Results of these analyses are described in Table 5.

Table 5: Summary performance metrics for CYP2D6 genotyping on the Helix Exome+ assay.

	Result
Simple Allele Sensitivity	100%
Complex Allele Sensitivity	98.9%
CYP2D6 Specificity	100%
Repeatability	100%
Reproducibility	100%

Sensitivity was calculated per allele, with all simple alleles identified correctly and only one complex allele mis-identified. In the case of the incorrect complex allele, the Helix PGx Pipeline called (*2, *41, *41), whereas the documented call was (*2, *41, *41, *41), suggesting inaccuracy in exact copy number of the *41 duplication.

Specificity calculations were calculated by treating *1 as the reference allele. Repeatability and reproducibility assessed concordance of intra-run and inter-run replicates, respectively.

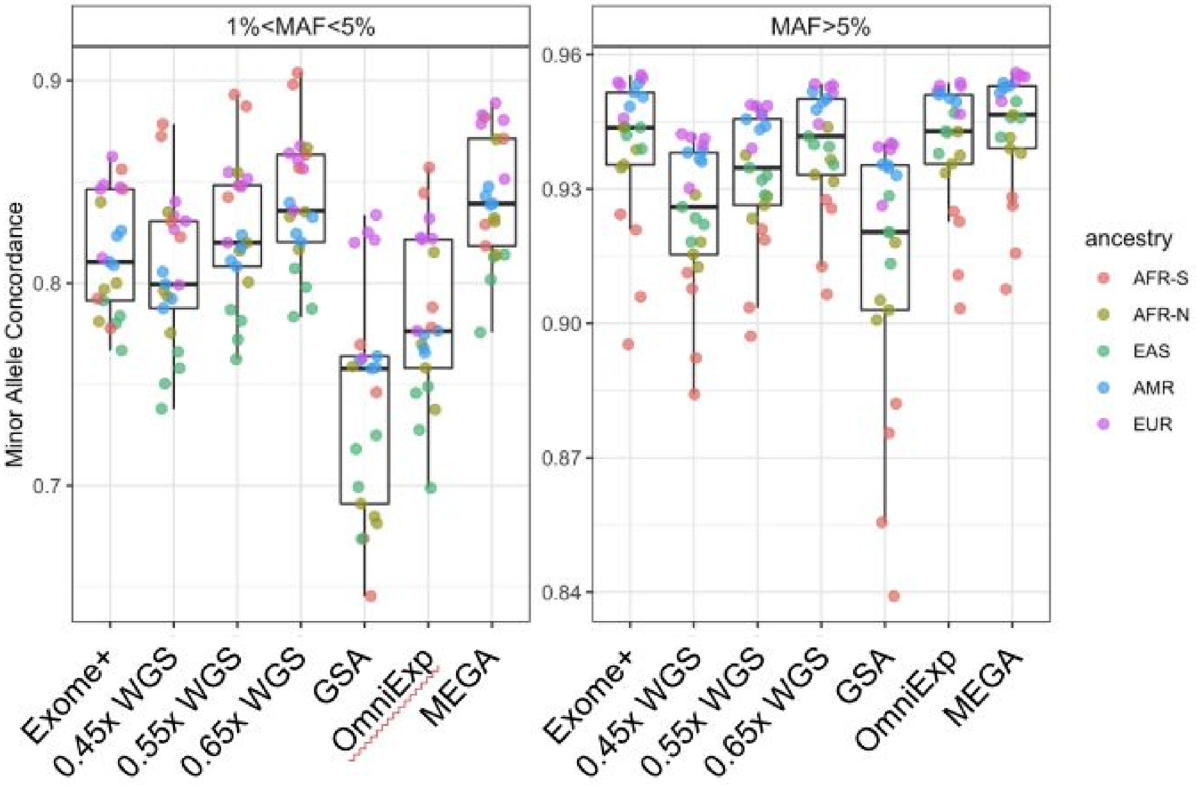
The PGx Pipeline delivers 106 (of 107) CYP2D6 star alleles described in PharmVar version 3.4[13].

Imputation for Polygenic Risk Scores

While standard Exome assays do not support comprehensive genome-wide imputation of common polymorphisms due to a lack of coverage in intergenic regions of the genome, the Exome+ assay includes ~300,000 non-coding regions selected for their relevance to GWAS findings, ancestry, and to power imputation. As a result, Helix is able to offer robust genome-wide imputation with tens of millions of high-confidence SNPs imputed with technical equivalence to ~0.6x WGS, Infinium OmniExpress, and Infinium MEGA for common alleles (Figure 1). For rare alleles, the Exome+ was approximately equivalent to 0.5x WGS, with improved performance over GSA and OmniExp (Figure 1).

This strong performance is attained because imputation is not limited to the hundreds of thousands of SNPs targeted in the Exome+ assay, but instead leverages all directly sequenced data, including the flanking regions of those SNPs and the full exome.

Figure 1: Demonstration of minor allele concordance between imputed variants and directly sequenced variants across five ancestral populations. Results are presented for typical Exome+ runs, WGS at different coverage levels (0.45x, 0.55x, 0.65x), and three microarrays (GSA, OmniExp, and MEGA). AFR-S: Sub-saharan African. AFR-N: North African. EAS: East Asian. AMR: Indigenous American. EUR: European.



Materials and Methods

Laboratory

The Helix Laboratory Platform is a highly automated laboratory process for generating robust and accurate sequencing results. The clinical laboratory at Helix is CLIA Certified #05D2117342 and CAP Accredited #9382893. Helix utilizes a Quality Management System that employs in-process monitoring and Six Sigma methodologies to ensure robust processes around DNA isolation, library preparation, enrichment, sequencing, and bioinformatics. This allows us to generate repeatable, accurate, and high-quality sequencing data.

Assay

The Exome+ v2 assay is a targeted DNA sequencing assay that targets ~19,000 genes, ~300,000 non-coding SNPs, and the mitochondrial genome. The assay has been optimized to provide consistent coverage across the whole exome and mitochondria with increased coverage of medically informative genes and select regulatory and intergenic regions. Additionally, hundreds of thousands of non-coding positions, including known GWAS markers, ancestry informative markers, and common SNPs, are used to support high-confidence genome-wide imputation results.

Small Variants

The Bioinformatics Pipeline uses well-established algorithms for alignment and quality control metrics. Helix utilizes a customized version of Sentieon's optimized variant calling software, which provides superior computational and analytical performance when compared to GATK[14].

For benchmarking purposes, only variants that pass our analytical standards are included and all variants belong to one of six variant type categories:

- SNV is a single base changed to a different base.
- Insertion is the addition of 1 to 20 bases.
- Deletion is the removal of 1 to 20 bases.
- MNV are phased, or linked, SNVs. This includes adjacent SNPs representing an overlapping insertion and deletion with the same length.
- Substitutions are variants resulting from an insertion and a deletion with different lengths sharing the same location and strand.
- Complex variants are two different variant types sharing the same location but each mapping to a different allele.

Variants that are two or more different types are binned in descending order: complex, MNVs, substitution, deletion, insertion, SNV.

The Helix Variants Pipeline performs imputation by pre-phasing samples and then imputing. Pre-phasing is done using reference databases which include the 1000 Genomes Phase 3 data. This is followed by genotype imputation for all 1000 Genomes Phase 3 sites that have genotype quality (GQ) values less than 20. Imputation results are then filtered for quality so that only high precision imputed variant calls are reported. Imputed variant calls are distinguished from observed variant calls in the Helix Genomics API by use of filter flags.

Copy Number Variants

The CNV Caller uses CNV Targets as the smallest unit for copy number assessment, such that the majority of CNV Targets equate to single exons or short non-coding regions. In some cases, exons that are a short distance apart may be merged into a single CNV Target. Read depth for each CNV Target is normalized using similar data from samples run through the laboratory at the same time. CNV events are then determined using a Hidden Markov Model (HMM).

CYP2D6 Genotyping

The PGx Pipeline uses a probabilistic approach to calculate the likelihood of a given star allele solution based on the observed data. The input includes the allele depths at 96 defining variants as well as exon-level copy number across both CYP2D6 and CYP2D7.

Limitations

Helix is excited to offer its partners the ability to query data from the Exome+ assay for each of its users. There are several caveats to its assay. The Exome+ assay does not sequence the whole genome. While we provide deep and broad coverage of the exonic regions of the genome, as well as several hundred thousand non-coding regions, this is still only ~ 2% of the entire genome.

Helix's assay does not perform equally across all regions of the exome. Regions that are hard to sequence, such as extremes of GC content, low complexity regions and segmentally duplicated regions may not have robust coverage. Further, indels greater than 20 bp are excluded from the analytical range, as are variants in or adjacent to homopolymer runs of > 10 bp, dinucleotide repeats of > 12 bp, or trinucleotide repeats of > 21 bp. Multinucleotide variants, substitutions, and complex variants are also excluded from short tandem repeat regions and homopolymer runs > 7 bp. Detection of heteroplasmic variants on the mitochondrial chromosome is not supported.

For CNVs, non-unique regions such as PMS2, exons 12-15, are outside of the reportable range. Events smaller than a CNV Target are likely to be missed, else they are reported as if they represent the full CNV Target. In the case of whole chromosome aneuploidy or large but partial chromosome aneuploidy, the entire chromosome is excluded from analysis. An exception is that CNVs will continue to be called in the presence of Trisomy 21. Mosaic events and structural variations such as inversions and translocations are outside of the Helix CNV Analytical Range. CNV results must be confirmed by a diagnostic laboratory prior to making any medical decisions or taking any medical actions.

CNV detection has been optimized for detection of rare CNVs across clinically-relevant genes. During the application of exome-wide CNV in research, it should be known that: (1) CNV detection is tuned toward rare events, and there will be reduced sensitivity to common CNVs, and (2) some regions may be more prone to false positives.

The PGx Pipeline is believed to have decreased sensitivity to CYP2D6*13 (representing a CYP2D7/CYP2D6 hybrid gene structure).

References

1. Aziz, N. et al. College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. Arch. Pathol. Lab. Med. **139**, 481–493 (2015).
2. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Research. **42**, 1001–1006 (2014).
3. Eberle, M. A. et al. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. Genome Res. **27**, 157–164 (2017).
4. Zook, J. M. et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat. Biotechnol. **32**, 246–251 (2014).
5. Zook, J. M. et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Scientific Data **3**, 160025 (2016).
6. https://www.coriell.org/
7. Conrad, et al., Origins and functional impact of copy number variation in the human genome. Nature. 464(7289):704-12 (2010).
8. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, et al. (2011) Mapping copy number variation by population-scale genome sequencing. Nature 470: 59–65.
9. Kalia et al., Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genet Med. 2017 Feb; 19(2): 249-255.
10. https://www.cdc.gov/genomics/implementation/toolkit/tier1.htm
11. Truty, R et al., Prevalance and properties of intragenic copy-number variation in Mendelian disease genes. Genetics in Medicine. **21** (1) 2019.
12. Structural Variation CYP2D6 (PharmVar). https://www.pharmvar.org/gene-support/Variation_CYP2D6.pdf
13. https://www.pharmvar.org/gene/CYP2D6, with v3.4 star alleles found in the excel spreadsheet dated 'Nov 5, 2018', downloaded here: https://api.pharmgkb.org/v1/download/file/attachment/CYP2D6_allele_definition_table.xlsx
14. Sentieon.com, DNAseq, for consistent and confident germline variant detection. (2017). https://www.sentieon.com/products/