Helix

Variants Pipeline Performance White Paper

Last Updated May 15, 2019

Executive Summary

Helix is a personal genomics platform company with a simple but powerful mission: to empower every person to improve their life through DNA. Our platform includes saliva sample collection, next-generation DNA sequencing (NGS), secure data storage, and secure APIs. Our partners can integrate with our APIs to deliver insights into ancestry, entertainment, family, fitness, health, and nutrition.

Sequence Once, Query Often

For each of its users, Helix sequences a full Exome+ dataset and securely stores this DNA information. With its "*sequence once, query often*" approach, Helix enables its partners to digitally query a predefined *in silico* panel derived from this data. As changes to their panel require only software updates, our partners can easily offer and update products. Helix partners do not need to build or staff their own next-generation sequencing facilities to incorporate DNA insights into their products.

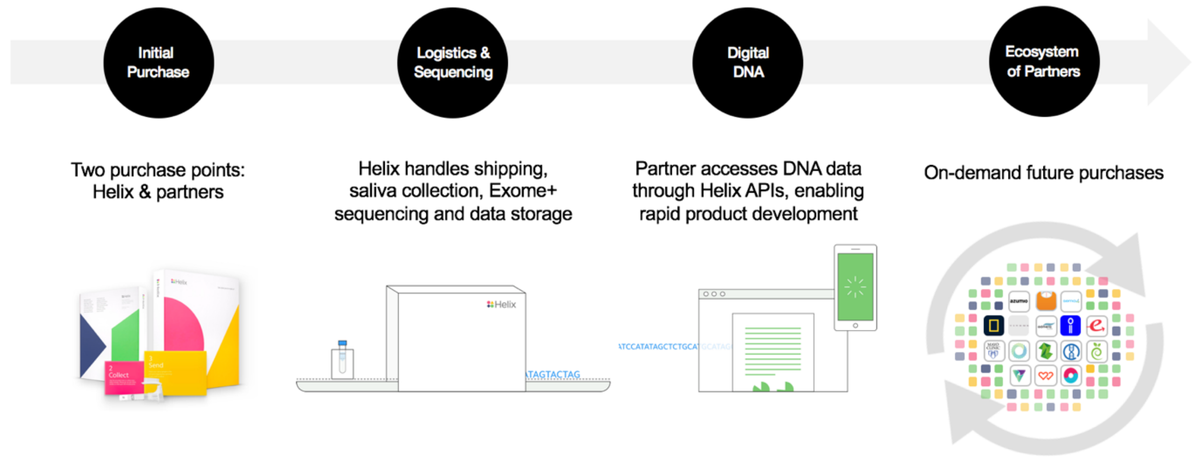

Initial Purchase
Logistics & Sequencing
Digital DNA
Ecosystem of Partners
Two purchase points: Helix & partners
Helix handles shipping, saliva collection, Exome+ sequencing and data storage
Partner accesses DNA data through Helix APIs, enabling rapid product development
On-demand future purchases
Helix
ATAGTACTAG
ATCCATATAGCTCTGCATCATAG

Figure 1: Helix Platform Services: Helix enables partners to access *in silico* panels of genetic information generated from Helix's Exome+, using only an APIs.

Helix's Genomic APIs include the Variants endpoint, delivering single nucleotide variants (SNVs) and indels across the Exome+; the CNV endpoint, delivering copy number variants across the Exome+; ancestry endpoints delivering summarized ancestry estimates; and more.

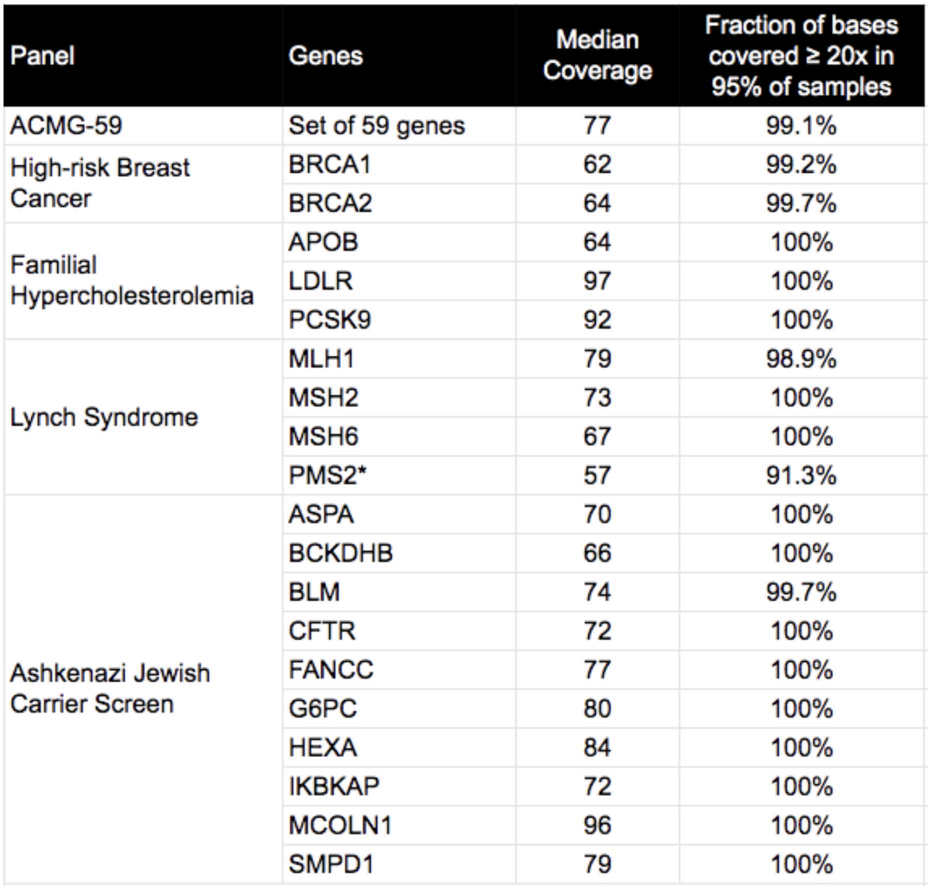
This White Paper describes performance of the SNVs and indels identified by the Variants Pipeline through the Variants API. Results are based on Exome+ v2 data, where results on Exome+ v1 data are described at https://cdn.shopify.com/s/files/1/0004/5416/4542/files/Helix-Performance-White-Paper_v2.pdf?3324109126506576509.

Helix Exome+ Data Completeness and Quality

Coverage Performance for Hypothetical Partner Panels

Helix provides detailed data describing genome coverage for target regions of interest to partners, enabling their design of robust and reliable panels. Coverage calculations are based on 4,000 Exome+ results approved by the Helix Laboratory Director and delivered from our production laboratory (2,000 males and 2,000 females). These coverage datasets include full base-pair level coverage histograms, which can be aggregated into detailed variant, exon, gene, or panel level statistics. This Exome+ assay performance data allows our partners to leverage the “sequence once, query often” approach to define and modify panels using only software.

For illustrative purposes, we provide coverage performance for different types of partner panels in Tables 1 and 2.



Panel	Genes	Median Coverage	Fraction of bases covered ≥ 20x in 95% of samples
ACMG-59	Set of 59 genes	77	99.1%
High-risk Breast Cancer	BRCA1	62	99.2%
	BRCA2	64	99.7%
Familial Hypercholesterolemia	APOB	64	100%
	LDLR	97	100%
	PCSK9	92	100%
Lynch Syndrome	MLH1	79	98.9%
	MSH2	73	100%
	MSH6	67	100%
	PMS2*	57	91.3%
Ashkenazi Jewish Carrier Screen	ASPA	70	100%
	BCKDHB	66	100%
	BLM	74	99.7%
	CFTR	72	100%
	FANCC	77	100%
	G6PC	80	100%
	HEXA	84	100%
	IKBKAP	72	100%
	MCOLN1	96	100%
	SMPD1	79	100%

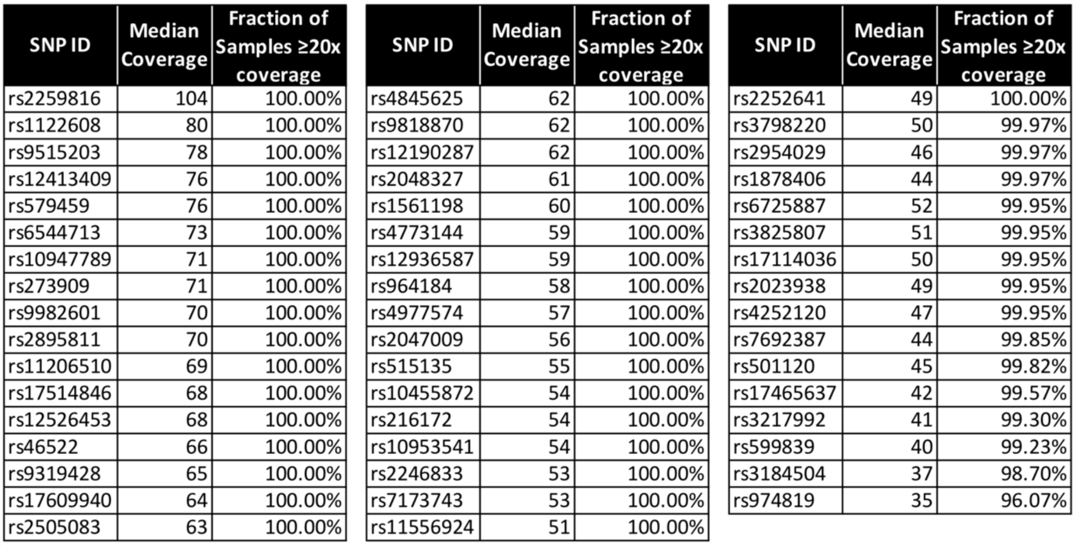
Table 1: Coverage for potential partner panels.
Median Coverage. Number of reads covering the 50th percentile of bases in the 50th percentile of samples.
Fraction of bases covered ≥20x in 95% of Samples. The percentage of bases that have at least 95% of samples with at least 20x coverage over the target region.
* Excludes PMS2 exons 2 and 12-15.

1) *ACMG-59 (Table 1)*: The American College of Medical Genetics and Genomics recommends reporting pathogenic and likely pathogenic findings in a set of 59 genes considered relevant to preventive adult-onset disease[1]. The Helix Exome+ delivers consistent coverage across these genes. Results shown exclude PMS2 exons 2 and 12-15.

2) *Hereditary Breast and Ovarian Cancer, Familial Hypercholesterolemia, Lynch Syndrome (Table 1)*: These diseases are defined by CDC's Office of Public Health Genomics (PHG) as those having significant opportunity for improving public health as a result of early genetic testing[12].

3) *Carrier Screening (Table 1)*: Carrier screening evaluates if a person carries a faulty copy of a recessive allele that results in a serious inherited disorder if an offspring receives two faulty copies of the recessive allele (one from each parent). Up to 24% of the general US population are carriers of at least one disease-causing recessive allele[3]. Another study suggests that expanded carrier screening may expose an even larger fraction of carriers[4]. While our assay cannot detect all known carrier conditions, we can provide partners coverage data to inform product planning for expanded lists of carrier conditions. Results shown represent a typical carrier screen for individuals of Ashkenazi Jewish background.

4) *50-SNP risk panel for coronary artery disease (Table 2):* Helix's Exome+ assay targets non-coding regions outside of the exome that are informative for a variety of products. An example of an application for these SNPs comes from a 2016 New England Journal of Medicine report describing the use of a 50-SNP polygenic risk score for coronary artery disease risk[5].



SNP ID	Median Coverage	Fraction of Samples ≥20x coverage	SNP ID	Median Coverage	Fraction of Samples ≥20x coverage	SNP ID	Median Coverage	Fraction of Samples ≥20x coverage
rs2259816	104	100.00%	rs4845625	62	100.00%	rs2252641	49	100.00%
rs1122608	80	100.00%	rs9818870	62	100.00%	rs3798220	50	99.97%
rs9515203	78	100.00%	rs12190287	62	100.00%	rs2954029	46	99.97%
rs12413409	76	100.00%	rs2048327	61	100.00%	rs1878406	44	99.97%
rs579459	76	100.00%	rs1561198	60	100.00%	rs6725887	52	99.95%
rs6544713	73	100.00%	rs4773144	59	100.00%	rs3825807	51	99.95%
rs10947789	71	100.00%	rs12936587	59	100.00%	rs17114036	50	99.95%
rs273909	71	100.00%	rs964184	58	100.00%	rs2023938	49	99.95%
rs9982601	70	100.00%	rs4977574	57	100.00%	rs4252120	47	99.95%
rs2895811	70	100.00%	rs2047009	56	100.00%	rs7692387	44	99.85%
rs11206510	69	100.00%	rs515135	55	100.00%	rs501120	45	99.82%
rs17514846	68	100.00%	rs10455872	54	100.00%	rs17465637	42	99.57%
rs12526453	68	100.00%	rs216172	54	100.00%	rs3217992	41	99.30%
rs46522	66	100.00%	rs10953541	54	100.00%	rs599839	40	99.23%
rs9319428	65	100.00%	rs2246833	53	100.00%	rs3184504	37	98.70%
rs17609940	64	100.00%	rs7173743	53	100.00%	rs974819	35	96.07%
rs2505083	63	100.00%	rs11556924	51	100.00%			

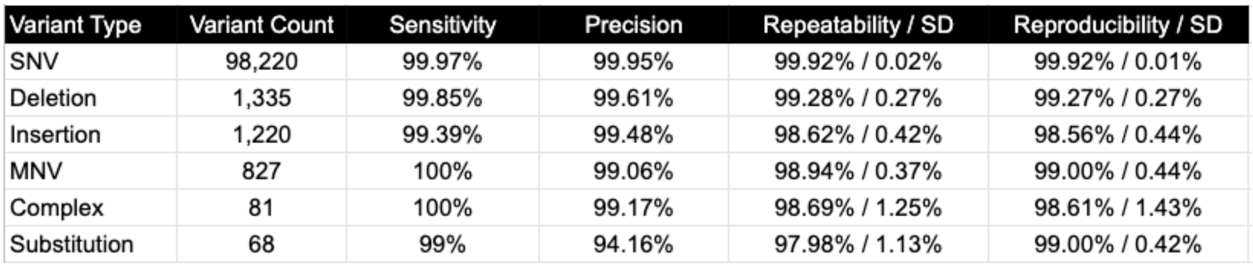
Table 2: Coverage of 50-SNP panel for coronary artery disease.

With Helix's Exome+ assay performance, we illustrate how Helix's "sequence once, query often" approach allows our partners to expand *in silico* panels as clinical recommendations evolve through software changes alone.

Validation using Reference Samples from NIST and GIAB

The Exome+ assay is performed in Helix's CLIA-certified and CAP-accredited laboratory. Our assay validation process adheres to guidelines from the College of American Pathologists (CAP)[6] and the Nex-StoCT workgroup for Standardization of Clinical Testing by NGS[7]. The validation study included DNA from saliva samples, well-characterized cell lines, and clinical positive control samples with known pathogenic variants. Results represent summary characteristics of variants that pass our analytical range.

We evaluated the performance of our assay against public reference materials from the Platinum Genomes[8] and the National Institute of Standards and Technology (NIST) Genome in a Bottle (GIAB)[9] datasets. Exome+ replicates were generated for cell lines for two individuals from CEPH pedigree 1463 (cell lines NA12877 and NA12878), an Ashkenazi Jewish trio (cell lines NA24385, NA24149, NA24143), and a Han Chinese sample (cell line NA24631)[10] obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research. These data were compared with high confidence calls from the Platinum Genomes and GIAB datasets. Four additional samples with known variant calls were also included. These data allowed the evaluation of variant calling accuracy for SNVs, insertions ≤ 20 bp, deletions ≤ 20 bp, multiple nucleotide variants (MNV), substitutions, and complex variants in sequence contexts that will be offered in products on the Helix platform (Table 3).



Variant Type	Variant Count	Sensitivity	Precision	Repeatability / SD	Reproducibility / SD
SNV	98,220	99.97%	99.95%	99.92% / 0.02%	99.92% / 0.01%
Deletion	1,335	99.85%	99.61%	99.28% / 0.27%	99.27% / 0.27%
Insertion	1,220	99.39%	99.48%	98.62% / 0.42%	98.56% / 0.44%
MNV	827	100%	99.06%	98.94% / 0.37%	99.00% / 0.44%
Complex	81	100%	99.17%	98.69% / 1.25%	98.61% / 1.43%
Substitution	68	99%	94.16%	97.98% / 1.13%	99.00% / 0.42%

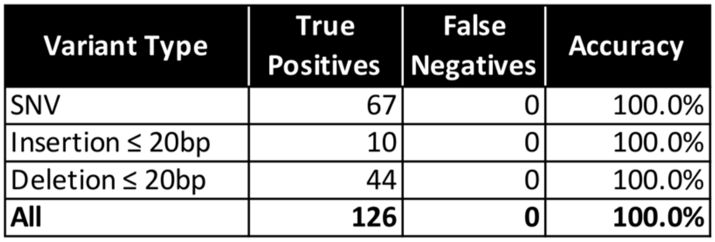
Table 3: Reference validation.
Variant Count: Average count of variant type per sample.
Sensitivity: Positive Percent Agreement.
Precision: Technical Positive Predictive Value.
Repeatability. Measured as concordance between sample triplicates from the same run.
Reproducibility. Measured as concordance between sample triplicates from different runs.
SD. Standard deviation.

Robustness of the Helix Laboratory Platform demonstrates high technical precision for all variant types. Intra-assay repeatability was evaluated using triplicates from 32 samples processed in the same run.

Inter-assay reproducibility was evaluated using triplicates from 69 samples processed by different operators and sequenced on separate runs.

Selected Positive Control Evaluation

The 65 gene panel in the Sema4 Carrier Check product was further validated with positive control samples that had been previously collected and tested in the clinic. They were reevaluated using the Helix's Bioinformatics Pipeline v3.0.0. This dataset consisted of 91 samples with a total of 126 known alleles, of which all 126 were called correctly and reportable.



Variant Type	True Positives	False Negatives	Accuracy
SNV	67	0	100.0%
Insertion ≤ 20bp	10	0	100.0%
Deletion ≤ 20bp	44	0	100.0%
All	**126**	**0**	**100.0%**

Table 4: Positive control variant analysis.

Imputation

Imputation is a statistical technique for using population patterns of linkage disequilibrium to infer genotypes not directly observed. Standard Exome assays are not able to perform high quality imputation genome-wide due to a lack of coverage in intergenic regions of the genome. However, Helix's Exome+ assay includes several hundred thousand 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 services utilizing its Exome+ assay.

Helix evaluated the accuracy of its imputation by comparing Illumina Infinium genotype microarray results from ten individuals to a total of 1,060 Exome+ replicates of these individuals. While imputed genotypes have many useful applications, Helix does not allow the use of imputed results for physician ordered products.

Materials & 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 and known non-coding SNPs that occur outside of the exome. 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 are covered, including known GWAS markers, ancestry informative markers, and common SNPs that improve imputation accuracy.

Bioinformatics

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[11].

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.

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

1. SNV is a single base changed to a different base.
2. Insertion is the addition of 1 to 20 bases.
3. Deletion is the removal of 1 to 20 bases.
4. MNV are phased, or linked, SNVs. This includes adjacent SNPs representing an overlapping insertion and a deletion with the same length.
5. Substitutions are variants resulting from an insertion and a deletion with different lengths sharing the same location and strand.

6. 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.

Limitations

Helix is excited to offer its partners the ability to query data from the Exome+ for each of its users. There are several caveats to its assay. First the Exome+ does not assay 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. Copy number variants are not delivered via the Variants endpoint and instead are delivered through the CNV endpoint. We will work with our partners to understand the limitations of our assay for partner-specific products as we also work to reduce these limitations. We provide detailed coverage information across thousands of samples so that information on assay performance is transparent.

Conclusions

Helix's personal genome platforms offers its partners the ability to query highly robust and uniform Exome+ sequence data using a "sequence once, query often" model. This enables our partners to offer highly accurate interpretation services relying on software-only product development.

References

1. ACMG Board of Directors. ACMG policy statement: updated recommendations regarding analysis and reporting of secondary findings in clinical genome-scale sequencing. Genet Med. 17 (1) 68-69 (2015).
2. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
3. Nazareth, Shivani B. *et al.* Changing Trends in Carrier Screening for Genetic Disease in the United States. *Prenatal Diagnosis* **35**, 931–35 (2015).
4. Haque, I. S. *et al.* Modeled Fetal Risk of Genetic Diseases Identified by Expanded Carrier Screening. *JAMA* **316**, 734–742 (2016).

5. Khera, A. V. *et al.* Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. N. Engl. J. Med. 375, 2349–2358 (2016).
6. Aziz, N. *et al.* College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. *Arch. Pathol. Lab. Med.* **139**, 481–493 (2015).
7. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research.* **42**, 1001–1006 (2014).
8. 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).
9. 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).
10. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data* **3**, 160025 (2016).
11. Sentieon.com, DNAseq, for consistent and confident germline variant detection. (2017). https://www.sentieon.com/products/
12. https://www.cdc.gov/genomics/implementation/toolkit/tier1.htm