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.

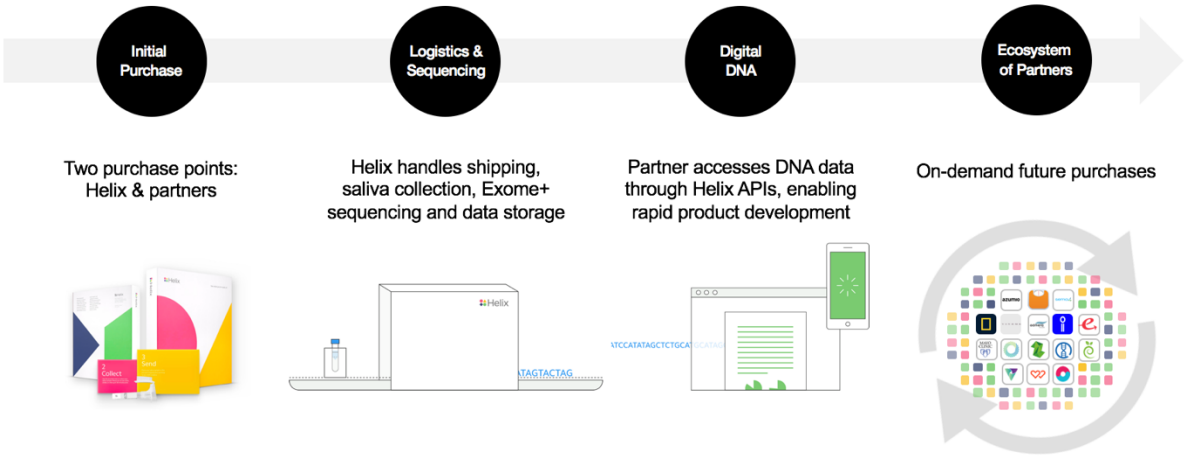

Figure 1: Helix Platform Services: Helix enables partners to access *in silico* panels of genetic information generated from Helix's Exome+, using only an API.

Helix Exome+ Data Completeness and Quality

Coverage Performance on Hypothetical Partner Target Regions

Helix provides detailed data describing genome coverage of desired target regions to partners to enable 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 hypothetical partner panels that could be available on Helix. In the below examples, we show the *Fraction of bases covered ≥20x in 95% of samples* metric, which provides the percentage of the target bases in which at least 95% of samples have at least 20x coverage computed across the 4,000 sample reference dataset.

Disease		Gene	Median Coverage	Fraction of bases covered ≥20x in 95% of Samples
Familial Hyper-cholesteremia		APOB	88	100.0%
		LDLR	96	99.6%
		PCSK9	100	100.0%
MODY		GCK	99	100.0%
		HNF1A	102	100.0%
		HNF1B	101	100.0%
		HNF4A	90	100.0%
		PDX1	75	100.0%
Carrier Screening	Cystic Fibrosis	CFTR	93	100.0%
	Tay Sachs	HEXA	101	100.0%
	Canavan	ASPA	91	100.0%
	Bloom	BLM	90	99.7%
	Wilson's	ATP7B	97	100.0%
	PKU	PAH	99	100.0%

Table 1: Coverage for a hypothetical panel.
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; for example, in CFTR in "95% of Samples", 100% of the target bases are ≥20x coverage.

1) *Familial Hypercholesterolemia (Table 1)*: Most people develop high cholesterol due to a combination of environmental, lifestyle, and genetic factors that add up to cause the disease. But about 1 in 250 people, over 800,000 adults in the US, have an inherited condition called familial hypercholesterolemia (FH) in which a change in a single gene can result in high cholesterol and lead to early onset cardiovascular disease.

2) *Maturity Onset of Diabetes of the Young (MODY, Table 1)*: Some people with diabetes have a rare, inherited form that is often misdiagnosed, but can impact how the diabetes is treated. People with MODY often respond best to therapies that are different from those that people with Type 1 and Type 2 diabetes receive.

3) *Carrier Screening (Table 1)*: Carrier screening is a genetic test that 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.

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. Helix's Exome+ assay provides ≥20x coverage for approximately 83% of all variants in the current NHGRI-EBI GWAS Catalog[2]. 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
rs3184504	134	100.00%	rs2047009	52	99.97%	rs6725887	47	99.43%
rs2259816	102	100.00%	rs17114036	50	99.97%	rs1878406	39	98.93%
rs3798220	96	100.00%	rs9319428	50	99.97%	rs11984041	39	98.65%
rs46522	90	100.00%	rs1561198	50	99.92%	rs10947789	37	98.65%
rs6544713	83	100.00%	rs9515203	46	99.92%	rs12190287	38	97.65%
rs273909	82	100.00%	rs11556924	52	99.90%	rs2246833	35	97.40%
rs3825807	69	100.00%	rs11206510	44	99.87%	rs646776	36	96.60%
rs515135	66	100.00%	rs17465637	56	99.85%	rs2023938	34	96.13%
rs4773144	66	100.00%	rs7173743	45	99.85%	rs4977574	33	95.17%
rs17609940	63	100.00%	rs12526453	46	99.82%	rs501120	37	93.07%
rs17514846	60	100.00%	rs2505083	44	99.80%	rs12936587	29	88.00%
rs2252641	59	100.00%	rs4299376	43	99.78%	rs7692387	28	87.88%
rs1412444	58	100.00%	rs4252120	45	99.73%	rs974819	31	86.80%
rs1122608	58	100.00%	rs9818870	44	99.73%	rs3217992	29	85.25%
rs216172	56	100.00%	rs964184	44	99.73%	rs2954029	27	84.83%
rs2048327	53	100.00%	rs10953541	44	99.70%	rs10455872	28	84.68%
rs579459	50	100.00%	rs1746048	43	99.70%	rs599839	28	81.53%
rs12413409	56	99.97%	rs4845625	40	99.50%	rs2895811	24	70.20%

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. A total of 1,152 Exome+ datasets were generated for these experiments.

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. 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.

Variant Type	Count	Accuracy / SD		Sensitivity / SD	
		Platinum Genomes	NIST GIAB	Platinum Genomes	NIST GIAB
SNV	137,456	99.70 / 0.01	99.45 / 0.05	99.74 / 0.01	99.92 / 0.01
INSERTION ≤ 20bp	1,988	98.73 / 0.27	97.12 / 0.17	98.91 / 0.17	98.53 / 0.15
DELETION ≤ 20bp	2,117	99.01 / 0.16	98.74 / 0.33	99.19 / 0.18	99.44 / 0.12
MNV	662	99.36 / 0.51	98.35 / 0.43	100.00 / 0.00	100.00 / 0.00
SUBSTITUTION	100	98.80 / 0.03	93.61 / 2.97	100.00 / 0.00	99.65 / 0.82
COMPLEX	81	98.62 / 0.25	97.23 / 1.15	98.89 / 0.94	98.30 / 1.05

Table 3: Reference validation.
Count. Mean count of PASS variants (TP+FP+FN) across all samples.
Accuracy. The average accuracy across samples of the same type. Sample accuracy is calculated as TP/(TP+FP+FN).
SD. Standard deviation.
Sensitivity. The average sensitivity across samples of the same type. Sample sensitivity is calculated as TP/(TP+FN).

Selected Positive Control Evaluation

The 65 gene panel in the Mount Sinai Carrier Check product was further validated with positive control samples that had been previously collected and tested in the clinic. They were re-sequenced using Helix's Exome+ assay and evaluated using the Helix's Bioinformatics Pipeline v2.6.1. This dataset consisted of 91 samples with a total of 126 known alleles, of which all 126 were called correctly and reportable.

Variant Type	Detected	Not Detected	Accuracy
SNV	67	0	100.0%
Insertion ≤ 20bp	10	0	100.0%
Deletion ≤ 20bp	44	0	100.0%
Substitution	0	0	100.0%
All	**126**	**0**	**100.0%**

Table 4: Positive control variant analysis.

Reproducibility and Repeatability

Robustness of the Helix Laboratory Platform demonstrates high technical precision for all variant types. Intra-assay repeatability was evaluated using 16 saliva and 16 cell line triplicates from the same run. Inter-assay reproducibility was evaluated using 47 saliva and 48 cell line samples processed by two different operators and sequenced on separate runs using the same instrument.

Variant Type	Count	Non-Reference Repeatability / SD	Non-Reference Reproducibility / SD
SNV	137,456	99.85 / 0.03	99.90 / 0.03
INSERTION ≤ 20bp	1,988	98.13 / 0.43	98.57 / 0.41
DELETION ≤ 20bp	2,117	98.71 / 0.35	99.04 / 0.39
MNV	662	98.00 / 0.81	98.87 / 0.52
SUBSTITUTION	100	95.62 / 2.49	96.38 / 2.63
COMPLEX	81	97.37 / 2.28	98.02 / 1.89

Table 5: Reproducibility, Repeatability
Repeatability. Measured as concordance between sample triplicates from the same run.
Reproducibility. Measured as concordance between sample triplicates from different runs.
SD. Standard deviation.
Count. Mean count of variants (TP+FP+FN) across all samples.

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. Based on these data, Helix demonstrated a precision of ~99% with sites with MAF >5% in the 1000 Genomes Phase 3 site list. For the site list, each individual sample had a per-sample recall rate of approximately 85%. 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+ assay is a targeted DNA sequencing assay that targets ~22,000 genes and known non-coding SNPs that occur outside of the Exome. The assay has been optimized to improve coverage depth and uniformity of the whole exome and mitochondria. In addition, Helix has added priority coverage to ~6,000 medically informative genes as well as select regulatory and intergenic regions. Finally, several hundred thousand additional positions are covered outside of the exome that include known GWAS findings, ancestry informative markers, and common SNPs to 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 Bioinformatics Pipeline performs imputation by pre-phasing samples and then imputing. Pre-phasing is performed 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. We then perform quality filtering to provide only high precision imputed variant calls. Imputed variant calls are distinguished from observed variant calls in the Helix Genomics API.

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 bases.
3. Deletion is the removal of bases.
4. MNV are phased, or linked, SNVs. Also includes an insertion and a deletion at the same location with the same length.
5. Substitutions have an insertion and a deletion at the same location with different lengths.
6. Complex variants are two different variant types at the same location (different on each 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. 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. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2013).

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/