Helix Copy Number Variations White Paper

September 20, 2018

The Helix CNV Caller delivers clinically validated copy number variations (CNVs) across the Helix Exome+, an assay that queries the full human exome as well as hundreds of thousands of non-coding targets in Helix's CLIA-certified and CAP-accredited production laboratory. These results, delivered through a separate API endpoint from our small variants endpoint, offer an expansion on the type of genetic variation that can be detected, studied, and reported upon in the Helix model (*Figure 1*).

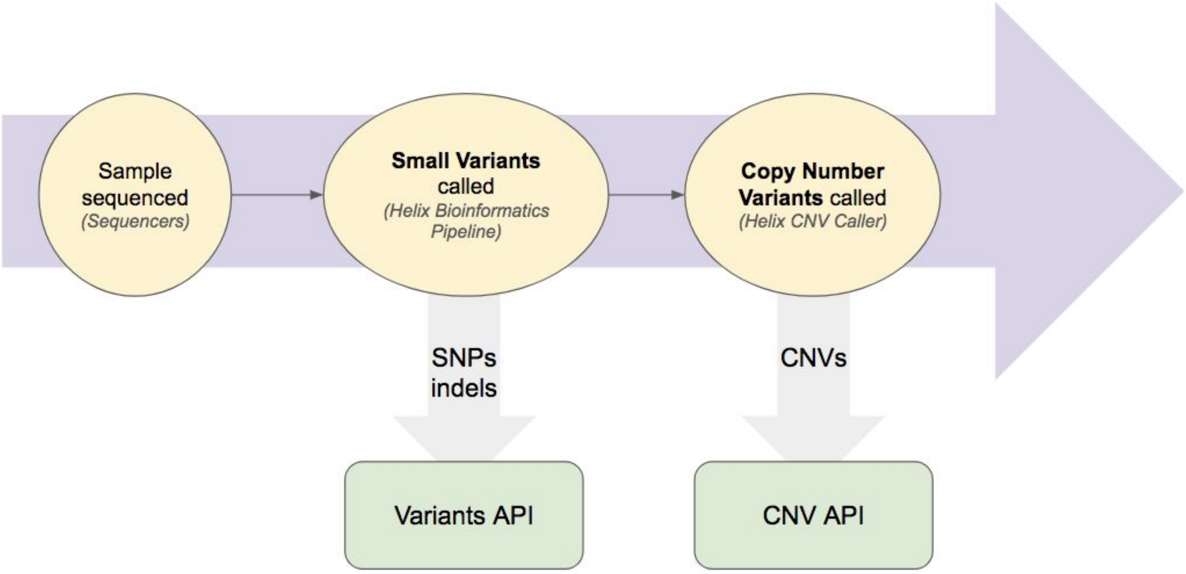


Figure 1. *Small Variants and CNVs are offered via independent APIs.*

Detection of Exome-wide CNVs

Establishing CNV Targets

Prior to identifying CNVs, the Exome+ is segmented into hundreds of thousands of CNV Targets, which are predefined genomic regions representing the smallest unit for which copy number is determined. The majority of CNV Targets represent a single exon, but due to proximity in space, a subset of CNV Targets are merged so that they extend past an exon to include nearby non-coding SNPs or to represent multiple adjacent exons (*Figure 2*).

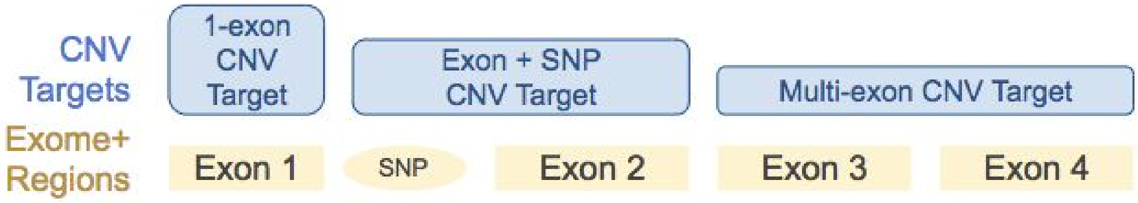


Figure 2. *CNV Targets represent the smallest unit for which copy number is determined. These usually span a single exon but they may also span larger regions within the Helix Exome+. Because CNV Targets are pre-defined, it is possible to determine which regions have merged targets and which regions have targets spanning a single exon for any gene of interest. Refer to the Appendix for specific examples of CNV Targets. Spacing in this figure not shown to scale.*

CNV Targets were created based on the distribution of Exome+ regions and then assessed to determine if they qualify for inclusion in CNV analysis. CNV Targets that did not meet the quality requirements were blacklisted, resulting in their exclusion prior to CNV analysis. These quality checks included: (1) a check to remove CNV Targets with a high no-call rate; (2) a check to remove CNV Targets that cover common CNVs; and (3) a check to remove CNV Targets where the underlying sequence is not unique within the genome.

Using CNV Targets to Identify CNVs

Copy number is determined by assigning CNV Targets to CNV Events in which adjacent targets share the same ploidy. Ploidy is determined by comparing read depth per CNV Target for each individual against the expected profile of read pair counts as determined by a baseline of many individuals sequenced through the same processes at the same time. A tuned Hidden Markov Model (HMM) takes into account the expected dispersion for a CNV Target and the likelihood of transitions to new copy number states between CNV Targets to inform the start and end of each CNV Event as well as the copy number status for that CNV Event (output as “callPloidy”, see *Figure 3*).

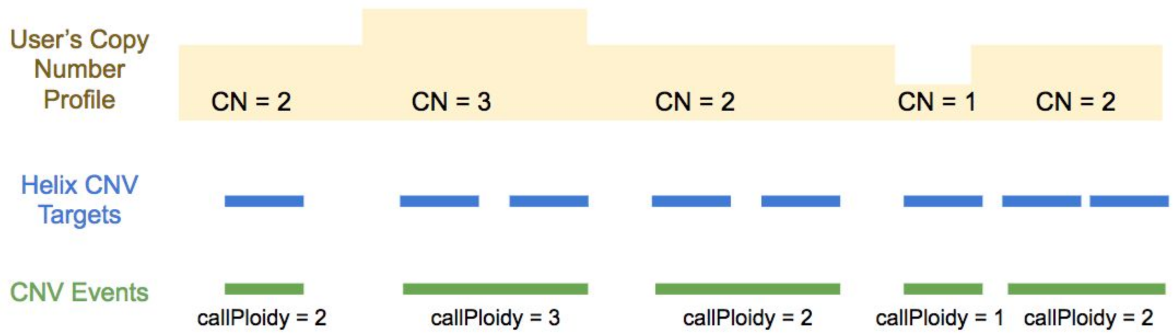


Figure 3. Coverage profiles for CNV Targets are used to distinguish regions with different copy number status, or CNV Events. Spacing in this figure not shown to scale.

Helix CNV Caller Performance Metrics

CNV Sensitivity

In order to determine the sensitivity of the Helix CNV Caller, a wide range of samples with documented CNVs were run through our standard production processes. Samples were processed as inter-batch replicates, meaning that independent library prep and enrichments were performed using the same DNA source. After removing samples that failed to meet sample-level QC metrics, there remained 220 results representing 85 documented CNVs ranging from single-exon events to cytogenetic CNVs.

All CNVs spanning two CNV Targets or more were detected successfully. In five cases, the documented CNV was missed, and in all of these cases the missed CNV spanned a single CNV Target (*Table 1*). This suggests high sensitivity to events covered by at least two CNV Targets.

True Positives	False Negatives*	Sensitivity	Sensitivity SE
215	5	97.7%	1.0%

Table 1. Sensitivity

Sensitivity: (TP/(TP+FN)) where TP = True Positives; FN = False Negatives. SE = Standard Error.
* All false negatives were CNV Events covered by a single CNV Target.

CNV Specificity

In the absence of established gold standards for CNV profiles across the exome, specificity estimates were measured using different approaches for different sets of samples (*Table 2*).

Relying on eight replicates of NA12878, specificity was determined by first collecting all documented deletions across the Exome+ based on a map of CNVs published in Mills et al.[1]. From there, all CNVs detected in the NA12878 replicates were compared to the known deletions and any novel CNVs were considered False Positives (FPs). Importantly, no effort was made to check if the presumed FPs may in fact be True Positives, just previously unknown. Yet, the previously documented CNVs only included deletions and not duplications. Therefore, it is assumed that a subset of the FPs are actually true and that specificity is under-estimated for this cohort. Among the NA12878 replicates, there were 2,168,046 true negative (TN) CNV Targets and 126 false positive (FP) CNV Targets (across 15 CNV Events), yielding a specificity estimate of > 99.994%.

To get a more informative specificity estimate for the regions of highest interest to Helix partners offering CNV within their product, specificity was measured by focusing on a set of 270 genes related to adult-onset disease, including the ACMG 59 genes. In this case, analysis was focused on a set of 26 ‘clinical samples’ (samples sent for clinical testing where a single established pathogenic CNV was detected). While the copy number profile of the individuals is unknown outside of the single documented pathogenic CNV, it was assumed that a second CNV within the set of 270 genes is highly unlikely. Therefore, for these samples, any CNV within the 270 gene panel that was not previously documented was considered a False Positive, though no effort was made to check if they may be True Positives and therefore that possibility cannot be ruled out. Two unexpected CNV Events were discovered, one deletion spanning a single CNV Target and one duplication spanning three CNV Targets (totalling a total of four FP CNV Targets).

Samples	True Negatives	False Positives	Specificity Estimate (TN/(TN+FP))	Specificity SE
8 NA12878 replicates	2,168,046	126	99.994%	0.0005%
26 clinical samples	82,465	4	99.995%	0.002%

Table 2. Specificity

TN = True Negatives; FP = False Positives; SE = Standard Error

CNV Target Call Rate

The Helix Exome+ is represented by hundreds of thousands of CNV Targets whose sensitivity and specificity are estimated above. In order to make a call on copy number, we require a quality score *callQuality* ≥ 20. Where callQuality < 20, the CNV Target is no-called. Across the full Exome+, we saw an average of 18 CNV Targets assigned a no-call across 526 replicates representing 186 samples, resulting in an estimate of a 99.99% Mean CNV Target Call Rate (*Table 3*), indicating that the Helix CNV Caller reliably delivers quality results with high sensitivity and high specificity across the entire Exome+.

Mean CNV Target Call Rate	CNV Target Call Rate SE
99.99%	0.0003%

Table 3. CNV Target Call Rate
SE = Standard Error

Helix CNV Caller Limitations

The Helix CNV Caller has the following limitations:

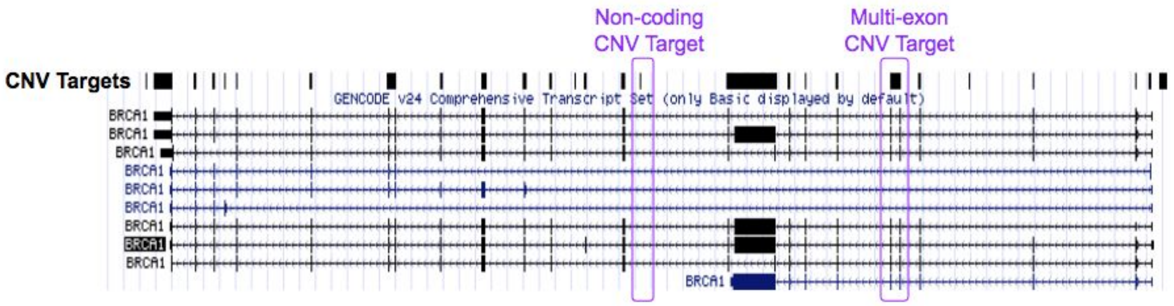
- Non-unique regions such as PMS2, exons 12-15, are excluded from analysis.
- Events smaller than a CNV Target are likely to be missed, else they are reported as if they represent the full CNV Target.
- In the case of whole chromosome aneuploidy or large but partial chromosome aneuploidy, the entire chromosome is excluded from analysis. An exception is that CNVs will continue to be called in the presence of Trisomy 21.
- If sex is not inferred by the Helix Bioinformatics Pipeline, which can occur in the presence of sex chromosome aneuploidy, then chrX is excluded from analysis.
- ChrY and chrM are excluded from analysis.
- Mosaic events and structural variations such as inversions and translocations are outside of the Helix CNV Analytical Range.
- CNV results should be confirmed by a diagnostic laboratory prior to making any medical decisions or taking any medical actions.
- Individuals who are symptomatic, pregnant (if performing carrier testing), or who have a family history should be directed toward comprehensive diagnostic testing in lieu of using a Helix-supported screen. Examples of such a family history include:
    - Onset of disease at an earlier age than population average, and/or
    - Family history of the same disease multiple times in multiple relatives (e.g., multiple relatives diagnosed with an arrhythmia), and/or

- Personal and/or family history suggestive of a syndrome (e.g., colon and uterine cancer in the same side of the family can indicate the family is at an increased odds to have Lynch syndrome), and/or
- Personal and/or family history of a diagnosis of a rare disease.

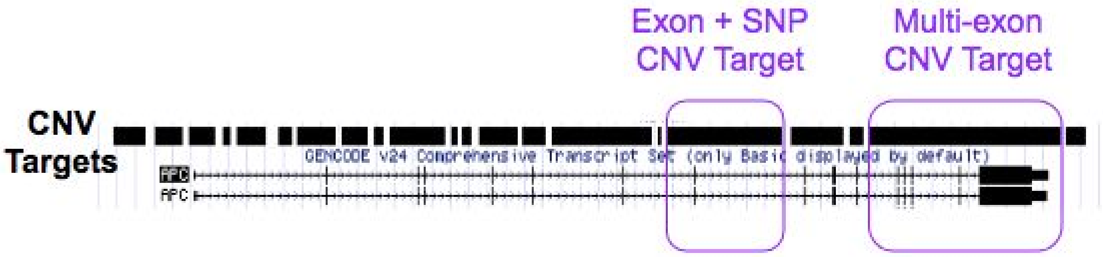
References

1. 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. https://www.nature.com/articles/nature09708
2. NCBI Genome Remapping Service. http://www.ncbi.nlm.nih.gov/genome/tools/remap

Appendix: Specific Examples of CNV Targets Across Health-related Genes



Appendix 1: *CNV Targets (the top row of black bars) are tiled across the BRCA1 gene. In most cases, a CNV Target represents a single exon. In a few cases, including those highlighted, the CNV Target represents an intronic, non-coding region or represents two adjacent exons within the same CNV Target because of their proximity.*


Exon + SNP
CNV Target
Multi-exon
CNV Target
CNV
Targets
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
APC
APC

Appendix 2. *CNV Targets (the top row of black bars) are tiled across the APC gene. In this case, the majority of CNV Targets span both an exon and non-coding intragenic space, and in multiple cases the CNV Targets also represent multiple exons. Though not highlighted in this picture, there are CNV Targets that cover the APC promoters.*