CNVThresher: Combining multiple lines of evidence to construct high-quality CNV call sets

Jason Harris, Deanna M. Church, Stephen Chervitz, Richard Chen; Personalis, Inc.

Abstract

We present **CNVThresher**, a tool that annotates CNV detections in WGS data with metrics designed to reflect the level of evidence present in the aligned reads, in support of each CNV call. The tool examines the pileup of aligned reads in the vicinity of the call, and looks for several features that are expected for real CNVs:

(1) **Morphological features of the read-depth profile**: the read depth should transition sharply at the breakpoint to a level reflecting the CNV's copy number, and this level should be sustained throughout the feature's width.

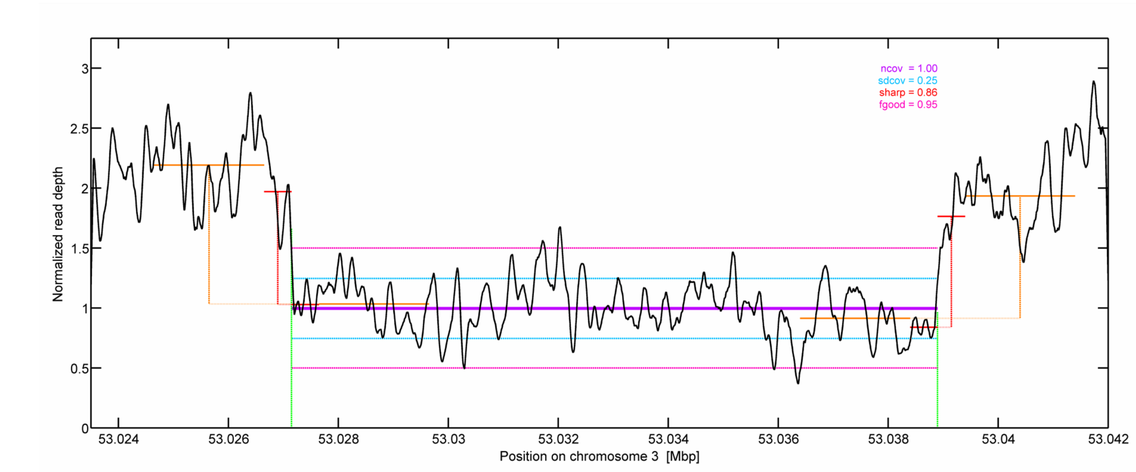
(2) **Read mapping anomalies**: including read pairs with anomalous insert size, and aligned soft-clip edges at the breakpoints.

(3) **Allele distribution features**: we expect a lack of biallelic positions inside heterozygous deletions, and we expect to see positions with unbalanced allele ratios inside duplications.

(4) **Quality control metrics**: to flag calls that may be more likely false positives, we also measure the fraction of reads that had a mapping quality of zero (both inside the feature, and in the flanking regions), and the density of non-Reference bases in the flanking regions.

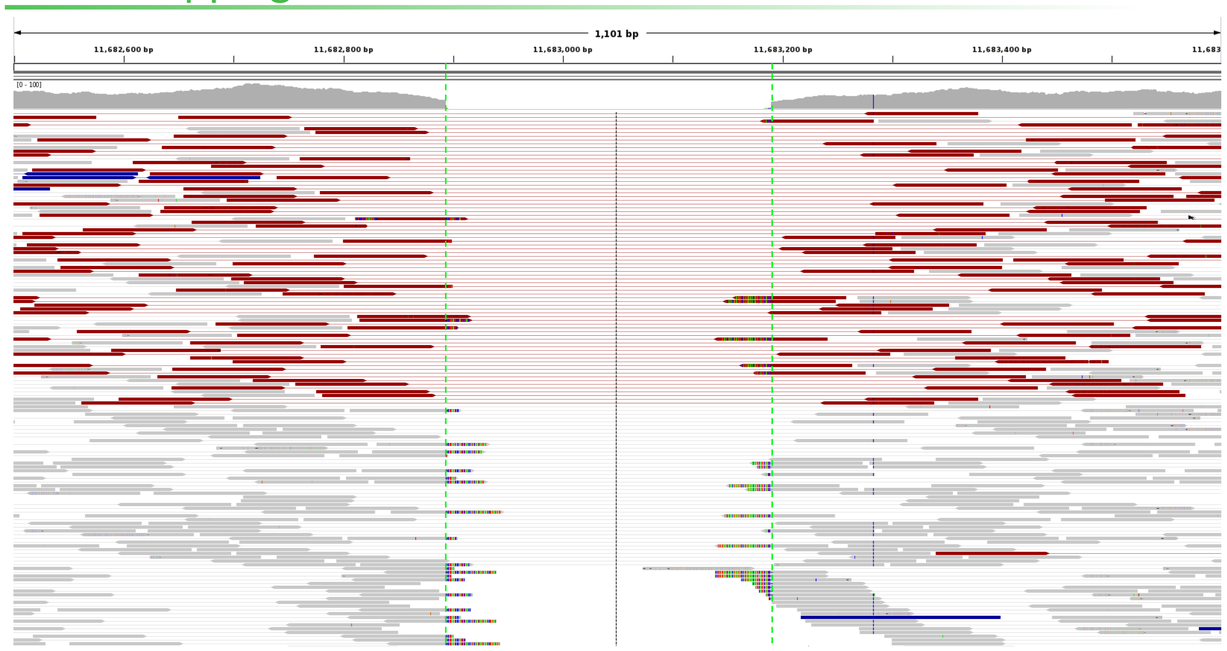
Together, the CNVThresher annotations provide a robust method for evaluating the level of evidence present in the aligned reads for any set of CNV calls.

Morphological features



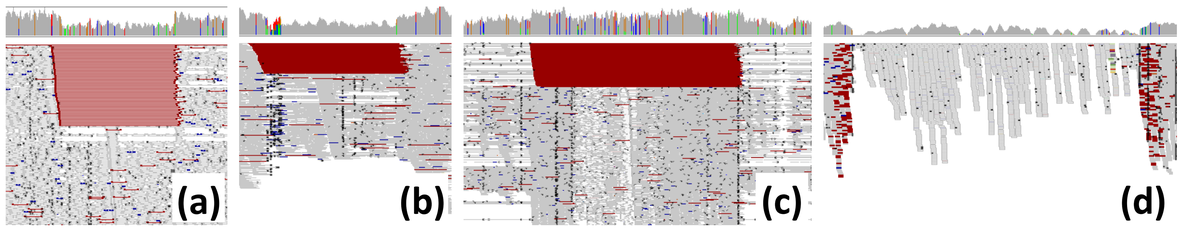
Morphological metrics measured by CNVThresher are illustrated in this view of the normalized read-depth profile in the vicinity of a 12 kbp heterozygous deletion. The **normalized coverage data** is shown in black, and the **measured CNV breakpoints** are indicated. Metrics include the **mean normalized coverage in the feature (ncov)**, the **standard deviation of the normalized coverage (sdcov)** in the feature, the **fraction of loci in the feature with coverage consistent with the CNV type (fgood)**. We also measure the edge sharpness (**sharp**) as the ratio of the **delta-coverage immediately surrounding the breakpoints** to the **delta-coverage measured at a distance from the breakpoints**.

Read-mapping anomalies



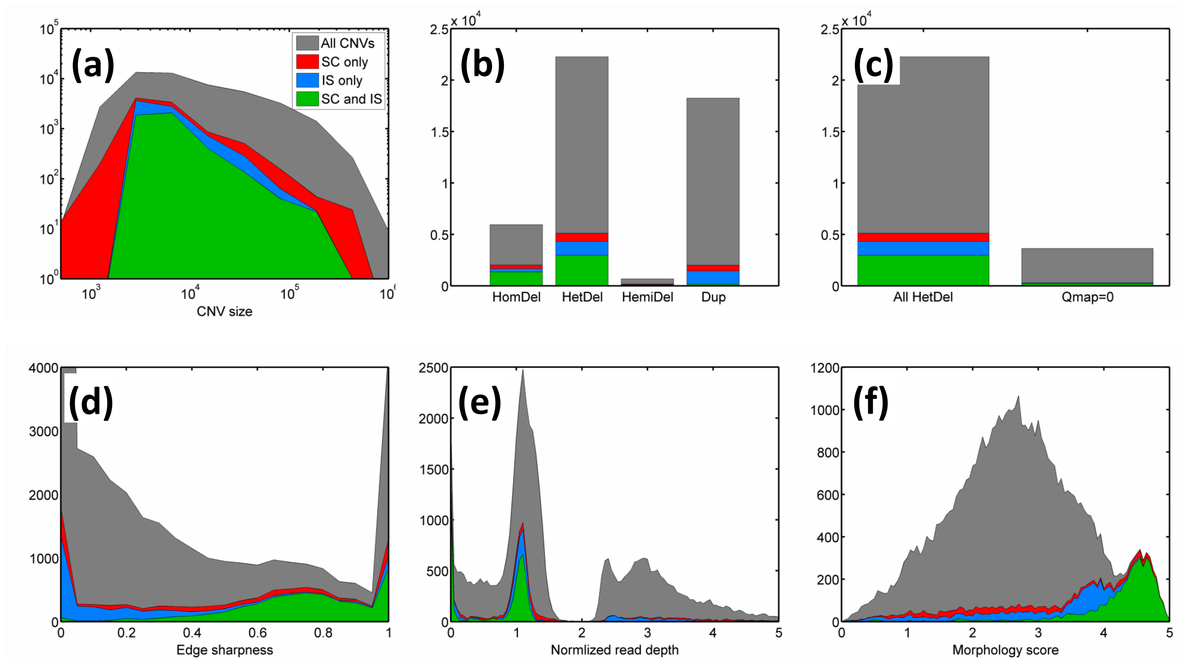
Read-mapping anomalies tracked by CNVThresher are illustrated in the IGV pileup view in the vicinity of a 300-bp homozygous deletion. Read pairs that bracket the breakpoints of a CNV will have **anomalous insert size**. Individual reads that cross a CNV breakpoint will either be split (for very small CNVs), or soft-clipped at the breakpoint (as shown here). IGV was configured to show clipped positions with color-coded nucleotides. Note that the soft-clipped sequence is coherent across reads (because in the sample, these reads map continuously, beyond the other breakpoint). CNVThresher requires this soft-clip coherence in order to flag a soft-clip breakpoint.

Allele distribution and Quality metrics



Other kinds of data extracted from the aligned BAMs are illustrated in these IGV pileup views of detected CNVs. In panel **(a)** we see a heterozygous deletion in which nearly all of the non-ref loci appear to be unanimous for one allele; this is expected for a het deletion since there is only one allele present. In panel **(b)** we see a different heterozygous deletion in which this expectation is violated: there is a cluster of loci near the upstream breakpoint which appear to be biallelic. CNVthresher provides a **fbiallelic** metric to help identify such cases. Similarly, in a duplication event, we expect to see unbalanced allele ratios when two alleles are present, as shown in panel **(c)**. CNVthresher provides a **funbal** metric that records the fraction of biallelic loci that appear to deviate from a 50/50 split. In regions of problematic mapping, we sometimes see heterozygous deletions that contain a high fraction of reads with mapping quality 0, suggesting that these reads may actually map elsewhere, and the CNV may actually be homozygous. This is illustrated in panel **(d)**. CNVthresher provides a **fQ0** metric that records the fraction of reads mapped in the feature with mapping quality 0.

Application to real data



We obtained 50x whole-genome sequencing data for a set of 28 samples. After alignment, we ran an internal CNV detection algorithm, obtaining a total of 47,000 CNV calls across the 28 samples. We then analyzed each sample's CNV call set with CNVthresher. In these six panels, we show some results from analyzing these CNV call sets with CNVthresher. In all panels, the colors show **All Detected CNVs**, those with **Soft-Clip Edge evidence (SC)**, those with **Insert-Size evidence (IS)**, and those with **Both SC and IS evidence**.

panel **(a)**: the size distribution of detected CNVs on a log/log scale.
panel **(b)**: the distribution of CNV type and zygosity.
panel **(c)**: among heterozygous deletions, the fraction that appear to be filled with Qmap=0 reads.
panel **(d)**: the edge-sharpness distribution (sharp).
panel **(e)**: the distribution of the normalized read depth (ncov).
panel **(f)**: the distribution of the "morphology score", a metric on a 5-point scale, composed of a linear combination of 5 morphological metrics.

Summary

CNVthresher is a tool for evaluating the level of evidence from the aligned BAM files, in support of each CNV call in a GFF. It traverses the aligned reads in the vicinity of each CNV call, and collects data on the morphology of the read-depth profile, read-mapping anomalies (such as large insert sizes or soft-clip edges), allele distributions, and read-mapping quality scores. Each metric is recorded in the GFF as an annotation tag. We have found this tool to be very useful for driving CNV detector development, and for constructing CNV Gold sets.

We are making CNVthresher available to the community. The github address is:
https://github.com/personalis/cnvthresher

Personalis®
Pioneering Genome-Guided Medicine

100-663-A