Towards Clinical-grade Detection, Annotation, and Interpretation *of* Structural Variants from Next-Gen Sequencing Data

Poster 569

Stephen Chervitz, Elena Helman, Jason Harris, Anil Patwardhan, Sarah Garcia, Jeanie Tirch, Michael J. Clark, Deanna M. Church, John West, Richard Chen

Contact: schervitz@personalis.com

[1]Personalis, Inc. | 1350 Willow Road, Suite 202 | Menlo Park, CA 94025

Introduction

Structural variations (SVs), including copy-number variations (CNVs - deletions, duplications) and copy-neutral variations (inversions, balanced translocations), collectively account for more differences between individual genomes than do single-nucleotide variations and small indels. Spanning 50 to millions of base pairs, SVs occur due to a variety of mutational mechanisms. The rapid drop in next-generation sequencing (NGS) costs has lead to an explosion of new SV detection tools & datasets, yet hurdles remain for reliable SV detection. We highlight the current technological and methodological challenges in SV detection, annotation, and interpretation that impede their clinical utilization, drawing on our experience with numerous whole exome & whole genome datasets from trios, proband-only Mendelian samples, as well as matched tumor/normal and tumor-only cancer samples.

SV Detection Challenges

Accurately detecting SVs from NGS data remains a significant challenge. Sequence read and insert size relative to the variation events lead to biases in the types and sizes of events that can readily be detected. Furthermore, the genomic architecture at some chromosomal locations leaves these regions inaccessible to interrogation by short read technology.

The GOOD

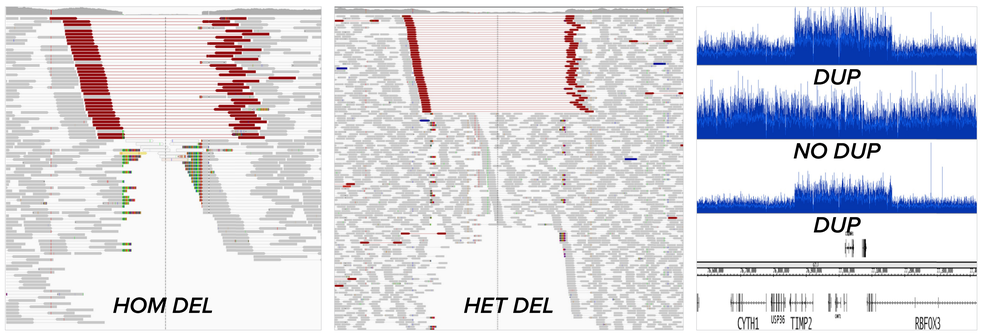

FIGURE 1: Good CNV detections possess multiple confirmatory attributes when examining the aligned NGS read pairs in a genome browser (IGV or IGB shown here): sustained increase or decrease in the read depth profile, soft-clip edges at breakpoints, split read pairs with anomalous insert size spanning CNV location, lack of heterozygous SNVs in the location of heterozygous deletions.

The BAD

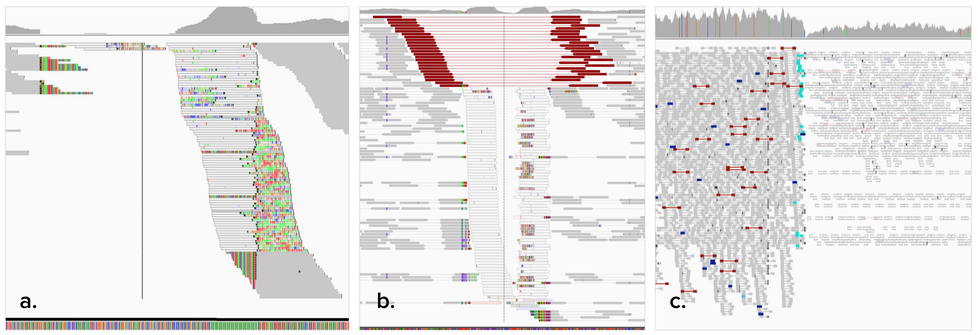

FIGURE 2: Low-quality or spurious CNV detections lack one or more lines of evidence characteristic of real CNVs and may show other anomalies. **(a)** Long poly(A) homopolymer near an apparent homozygous deletion causes large pile-up of reads. **(b)** Clustering of poorly mapped reads within a set of split reads. **(c)** Reference genome artifact causes an apparent heterozygous deletion due to a rare duplication present in reference genome.

The EXOME

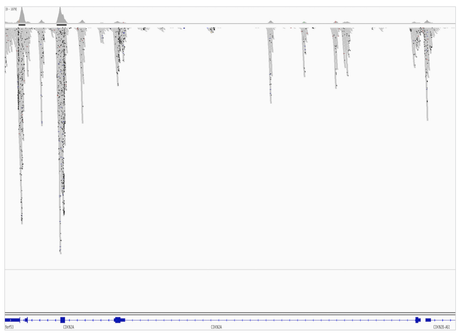

FIGURE 3: CNV detection in whole exome sequencing data is complicated by exon-specific targeting, leading to non-uniform coverage across the genome. Careful normalization is required to factor in %GC and other parameters.

Somatic (Tumor) SV Challenges

Somatic SV detection in cancer samples poses unique challenges, due to small sample amounts, low cellularity, and tumor heterogeneity, as well as wildly aberrant genomic events. Read-depth and minor-allele frequencies may be used to assess relative CNV levels in whole-exome data, but sophisticated corrections for purity and clonal mixtures must be applied to elucidate absolute copy number. Additionally, a matched normal and/or a set of normals is necessary to rule out germline events and establish baseline normal variation to which to compare tumor data.

Absolute Copy Number Profiles

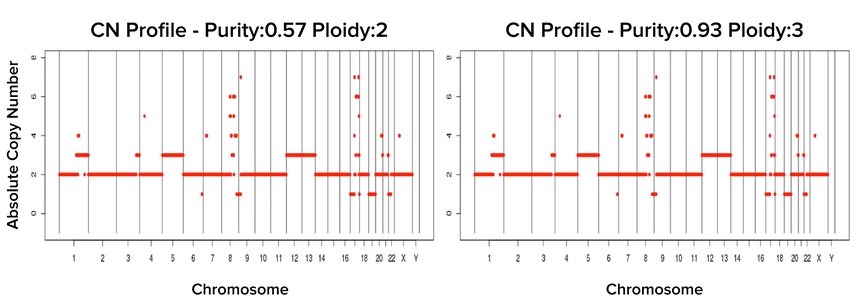

FIGURE 4: Assigning absolute copy-number values to each genomic segment requires accounting for tumor purity, heterogeneity and ploidy. On the left is a mostly diploid case with 0.57 tumor content; on the right is a mostly triploid case with 0.93 tumor content.

Allelic Skew and LOH

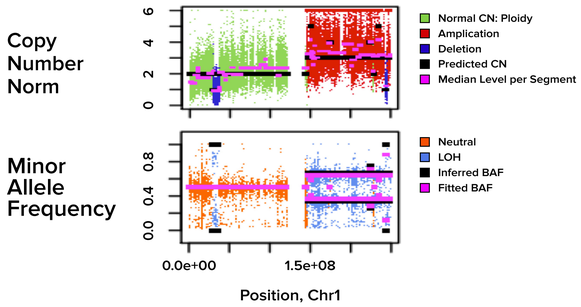

FIGURE 5: Minor allele frequency information can be used along with read-depth to elucidate allelic skew and loss of heterozygosity, common events in tumor biology.

Matched Normal *vs.* Proxy

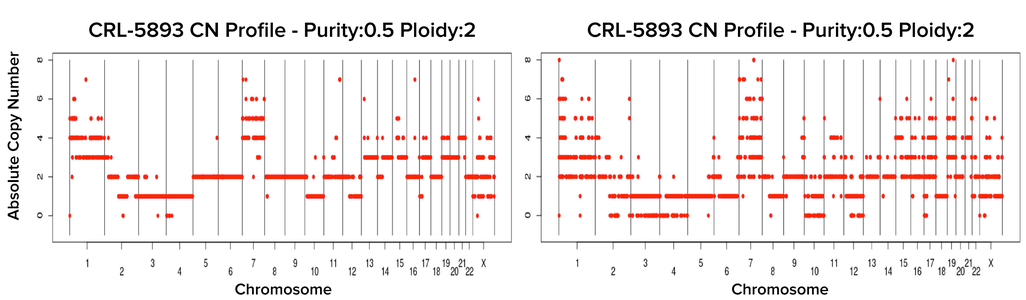

FIGURE 6: Individual copy number variation and coverage-based biases can be alleviated by normalizing to a matched normal sample (left) or, when a matched normal sample is unavailable, using a proxy normal sample from an unrelated individual (right).

Annotation and Interpretation Challenges

Annotation and interpretation are also a significant challenge due largely to the difficulty in determining if SV calls made using different technologies represent the same event. Importantly, integration of SV calls with smaller variant data is critical in order to accurately determine the clinical significance of variants present in a given individual or tumor.

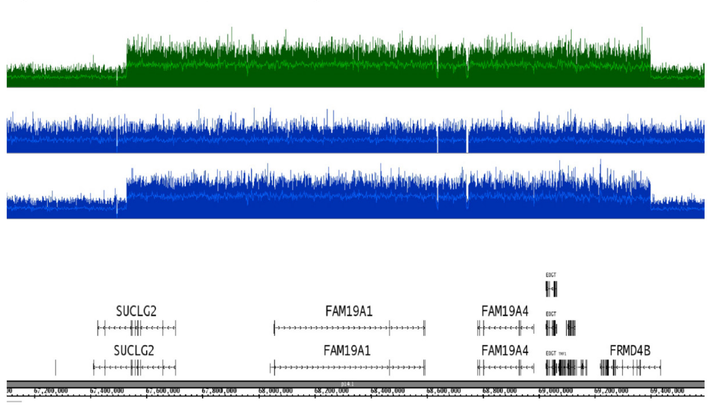

FIGURE 7: A large inherited duplication present in proband (top, green) and one parent (bottom) spans the location of two deletions inherited from the other parent (middle). Some SV detection algorithms reported the duplication as separate CNVs. Inheritance information from the trio facilitates elucidation of true CNV structures.

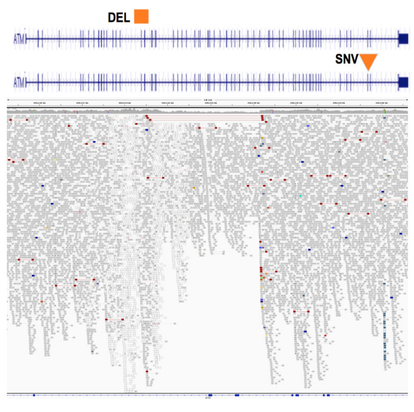

FIGURE 8: Two-exon ATM deletion (4.8kb) plus a compound heterozygous SNP in Coriell sample NA11254 was detected by Personalis integrated SV+SNV analysis.

FIGURE 9: Detected SVs are assessed using a comprehensive set of quality metrics to identify the best candidates.

Personalis SV Intelligence

Personalis is improving the clinical utility of NGS-detected SVs for both whole genome and whole exome analysis via:

- **Improved SV detection and annotation methodologies**
- **Integrated large (CNV) and small (SNV) variant analysis**
- **Comprehensive SV quality metrics**
- **Probands and trios –** *harnessing inheritance pattern information*
- **Reference genome expertise –** *accurate genome assembly knowledge*
- **Large and growing in-house sample repository –** *to identify common CNVs that complicate interpretation*

ACMG 2015 | Booth 1107

Personalis®
Pioneering Genome-Guided Medicine

www.personalis.com/clinical | clinical@personalis.com | +1 650-752-1349

POS-ACSC-1.0 MAR 2015