Genome analysis

SVPV: a structural variant prediction viewer for paired-end sequencing datasets

Jacob E. Munro¹, Sally L. Dunwoodie¹,²,³ and Eleni Giannoulatou¹,²,*

¹Victor Chang Cardiac Research Institute, Sydney, NSW 2010, Australia, ²St Vincent's Clinical School and ³School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW 2052, Australia

*To whom correspondence should be addressed.
Associate Editor: Bonnie Berger

Received on September 27, 2016; revised on February 21, 2017; editorial decision on February 22, 2017; accepted on February 23, 2017

Abstract

Motivation: A wide range of algorithms exist for the prediction of structural variants (SVs) from paired-end whole genome sequencing (WGS) alignments. It is essential for the purpose of quality control to be able to visualize, compare and contrast the data underlying the predictions across multiple different algorithms.

Results: We provide the structural variant prediction viewer, a tool which presents a visual summary of the most relevant features for SV prediction from WGS data. SV calls from multiple prediction algorithms may be visualized together, along with annotation of population allele frequencies from reference SV datasets. Gene annotations may also be included. The application is capable of running in a Graphical User Interface (GUI) mode for visualizing SVs one by one, or in batch mode for processing many SVs serially.

Availability and Implementation: SVPV is available at GitHub (https://github.com/VCCRI/SVPV/).

Contact: e.giannoulatou@victorchang.edu.au

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Structural variants (SVs) are typically classified as sequence variants longer than 50 bp, and may be balanced (inversions and translocations) or unbalanced (insertions, deletions, duplications and copy number variations). Due to their large size compared to single nucleotide variants (SNVs) and insertion/deletions (indels), SVs comprise the majority of variable bases in the human genome. Evidence from the 1000 Genomes Project suggests that humans have a median of 8.9 Mb of SVs versus 3.6 Mb of SNVs (Sudmant et al., 2015). Additionally, there is mounting evidence for the contribution of SVs to common and rare human diseases (Weischenfeldt et al., 2013).

SVs may be predicted from paired-end whole genome sequencing (WGS) based on several features of the alignments. SV breakpoints give rise to split read alignments (clipping). Deletions are characterized by decreased read depth and increased mapping distance for paired reads spanning the breakpoint (Supplementary Fig. S1). Tandem duplications show increased read depth and inverted paired read orientation (Supplementary Fig. S2). Inversions result in paired reads mapping to the same strand (Supplementary Fig. S3). Insertions will show clipping and possibly a decrease in mapping distance or orphaned reads (Supplementary Fig. S4). Translocations will show a high proportion of mates mapping to a different molecule (Supplementary Fig. S5). There are many tools available for SV prediction, including CNVnator (Abyzov et al., 2011), Delly2 (Rausch et al., 2012), LUMPY (Layer et al., 2014) and GenomeSTRiP (Handsaker et al., 2013). These tools use some combination of the alignment features to predict SVs.

Whilst a large range of applications exist for viewing WGS alignments (e.g. Integrative Genomics Viewer (IGV); Robinson et al., 2011), these are typically geared towards inspecting SNVs and indels, and operate by loading each individual aligned read in a given genomic locus. This approach is not ideal for SVs as they cover considerably larger genomic regions, and as such loading individual reads may exhaust a computer’s memory. Additionally, summary statistics on how reads are aligned at a given position are not given. We present structural variant prediction viewer (SVPV), an application designed to present visual summaries of WGS alignment data relevant to SV prediction. This will aid in the validation of individual SV calls, allow manual calling of SVs by inspection of the figures, and enable researchers to compare and contrast different SV prediction methods.
2 Input and example dataset

Alignments and VCFs: Alignment files in sorted, indexed BAM/CRAM format are parsed by SAMtools (Li et al., 2009) and piped into SVPV. Similarly, SV calls from VCF/BCF files are parsed by BCFtools and piped into the program. SVPV runs with one primary SV call set along with an arbitrary number of alternative call sets. A reference population SV call set may also be provided. SV types supported by SVPV are deletions, duplications, copy number variants (CNVs), insertions, translocations and inversions.

Genes: SVPV supports annotation of RefSeq genes. These must be in ‘refgene’ formatted tables, and are currently available for a variety of reference genome assemblies from the UCSC table browser (Karolchik et al., 2004; Pruitt et al., 2014).

Example Dataset: SVPV was tested on the publicly available sequence data of eight members of the CEPH pedigree 1463, whole-genome sequenced to a mean coverage of ~50x. Paired reads were aligned to the hg38 reference using BWA mem (H.Li, manuscript in preparation). SVs were predicted on the resulting alignments using Delly2, LUMPY and CNVnator. As CNVnator does not natively outputs results as VCF, the results had to be coerced into this format (see Supplementary Material for full details). SV VCFs from the 1000 Genomes Project were used to provide a reference population frequency. RefSeq genes for assembly hg38 were downloaded from the refgene format from the UCSC table browser and used for annotation.

3 Functionality and output

SVPV parses the provided alignment files within a window around a given SV call, and compiles binned summary statistics on the aligned reads. These statistics are used to generate a figure consisting of a set of tracks for each sample, as well as a set of annotation tracks (Fig. 1). Representative examples of the supported SV types are given in the Supplementary Material. If run in GUI mode, SVPV allows visualization of individual calls on some subset of the input samples (Supplementary Fig. S4). Calls can be filtered based on length, SV type, allele frequency, sample genotypes and intersection with genes of interest. In batch mode SVPV will produce pdf figures for each SV in the input VCF matching the provided filters. Sample tracks consist of SV calls (from both primary and alternate VCFs), read depth, forward and reverse mapping distances, and the proportion of reads mapping that are clipped, mapped to a different molecule, orphaned (mate does not map), inverted (forward and reverse reads oriented outwards) or same-strand (forward and reverse reads mapped to the same strand). Annotation tracks provide the size, type and allele frequency of SVs within the current window from the primary, alternate and reference VCFs. RefSeq genes with exons are also shown if they fall within the window. GC content is presented if a fasta sequence of the reference genome is supplied.

4 Conclusions

SVPV provides a convenient way for researchers to interrogate SV call sets and determine that predictions appear correct before proceeding to further analysis. By comparing multiple call sets, the strengths and weaknesses of different algorithms may become apparent, aid researchers in the choice of which algorithm to use for their dataset, and ultimately lead to refined prediction algorithms.

Funding

This work has been supported by the Chain Reaction (The Ultimate Corporate Bike Challenge) and the Office of Health and Medical Research, NSW State Government. S.L.D. has a NHMRC Senior Research Fellowship (1042002).