Shiny-SoSV: A web app for interactive evaluation of somatic structural variant calls

Tingting Gong[1,2], Vanessa M Hayes[1,2,3], Eva KF Chan[1,3*]

[1]Garvan Institute of Medical Research, Darlinghurst, NSW 2010, Australia,
[2]Central Clinical School, University of Sydney, Camperdown, NSW 2006, Australia,
[3]St Vincent's Clinical School, University of New South Wales, Randwick, NSW 2052, Australia.

*To whom correspondence should be addressed

Abstract

Somatic structural variants play a significant role in cancer development and evolution. Accurate detection of these complex variants from whole genome sequencing data is influenced by many variables, the effects of which are not always linear. With increasing demand for the application of whole genome sequencing in clinical settings, there is an unmet need for clinician scientists to easily make technical decisions for every unique patient and sample. To address this, we have developed Shiny-SoSV, an interactive web application for evaluating the effects of five common variables on the sensitivity and precision of somatic structural variant calls, thereby enabling users to quickly make informed sequencing and bioinformatics decisions early on in their study design.

Availability: Shiny-SoSV is freely available at https://hcpcg.shinyapps.io/Shiny-SoSV and source code available at https://github.com/tgong1/Shiny-SoSV

Introduction

Somatic structural variations (SVs) are large (> 50 bp) genomic rearrangements that arise in tumours and significantly contribute to cancer development and progression (Tubio, 2015; Northcott et al., 2014; Baca et al., 2013). The advent of next generation sequencing (NGS) has facilitated an increase in the efficiency and accuracy of detecting somatic variants in cancer genomes. However, due to limitations of short-read NGS, large SVs can only be inferred through alignment signatures. For example, read pairs that map farther apart than expected are suggestive of a deletion event. Attempts to improve SV detection have thus spurred many computational developments, resulting in the publication and availability of numerous SV detection tools (Chen et al., 2009; Ye et al., 2009; Rausch et al., 2012; Layer

et al., 2014; Chen et al., 2016; Cameron et al., 2017). These tools are based on different, sometimes overlapping, algorithms and therefore can differ substantially in their sensitivity and breakpoint precision (Liu et al., 2015; Tattini et al., 2015).

SV detection is further complicated in cancer samples due to variable tumour purity (proportion of cancer to non-cancerous cells in a sample) and variant allele frequency (VAF, proportion of the sequencing reads harbouring the variant at a given genomic locus). Although histopathology can provide an overall estimate of tumour purity, it typically only reflects the upper bound of observed VAF, due to the possibility of sub-clonality. Moreover, although deeper sequencing may increase sensitivity, numerous studies have shown that the benefit of increasing coverage does saturate (Chaisson et al., 2019; Griffith et al., 2015; Sims et al., 2014). Thus, knowing how much sequencing depth to increase is not obvious with decisions typically based on experience or "gut instinct".

Taken together, there are many alterable variables that can affect somatic SV detection in NGS analyses. Yet, informed decisions on these variables are rarely easy, especially by non-bioinformaticians in clinical settings. To address this, we have developed Shiny-SoSV, a web-based interactive application to help evaluate and guide decisions on modifiable parameters impacting SV detection in cancer genomes.

Methods

Simulation data generation and Somatic SV calling

To evaluate the sensitivity and accuracy of SV detection, a simulation study was devised. In short, two sets of SVs including six SV types (deletion, duplication, inversion, domestic insertion, foreign insertion and translocation) were simulated using SVEngine (Xia et al., 2018). SVs were randomly distributed along the genome (GRCh38) without overlap, masking gap, centromeric and telomeric regions as recommended (Dixon et al., 2018; Xia et al., 2018). The first set of 1,200 SVs were used as germline SVs and spiked into both the germline and tumour genomes (fasta). The second set of 1,200 SVs were used as tumour SVs and further added to the tumour genome. Paired short-reads were sampled using SVEngine (Xia et al., 2018) from the altered fasta (detail in Supplementary Methods), then aligned to GRCh38 using BWA-MEM (Li and Durbin, 2009), generating alignment bam files. Tumour purity (VAF) was emulated by merging different ratios of normal to tumour aligned reads to create the final

tumour bam files. Here, tumour purity and VAF are used interchangeably as SVs are assumed independent of each other. In all, 216 pairs of bam files, imitating varying depths of coverage of the normal samples (15x, 30x, 45x, 60x, 75x, 90x), tumour samples (20x, 30x, 45x, 60x, 75x, 90x) and VAF (5%, 10%, 20%, 50%, 80%, 100%) were generated, each then replicated three times for a total of 648 datasets.

Somatic SVs were called using three SV callers (Manta, Lumpy and GRIDSS) for each tumour/normal pair. The union and intersection callsets for each pair of SV callers were also generated (see Supplementary Methods), resulting in a total of 15 callsets for each simulated dataset. Sensitivity and precision of SV calls were determined by comparing against the set of simulated tumour SVs, requiring matching SV type for consideration as true positives, and evaluating precision of breakpoint positions. Further details of the evaluation criteria for true and false positive calls are in Supplementary Methods.

Prediction model fitting

Predictions of sensitivity and precision were based on a generalised additive model (GAM), fitting on SV caller, VAF, depth of coverage of tumour and normal samples and breakpoint precision threshold as predictors. The choice of GAM was motivated by non-linear relationships observed between some response and predictor variables (Figure S3). In particular, we found that VAF has a non-linear effect on sensitivity, and both VAF and breakpoint precision threshold have non-linear impact on precision. Therefore, smooth functions were applied to these predictors using restricted maximum likelihood (Wood, 2011). As the response variables are proportions with values between 0 and 1, the beta regression (betar) with logistic link function was used. See Supplementary Methods for full model specification with implementation using the R package *mgcv* (Wood, 2011).

Web-application design

The web application was developed using R *shiny* v1.3.2 to provide a visual platform for evaluating the behaviour of sensitivity and precision with various predictor variables through two main interactive plots. Comparison and interaction effects of two or more predictors are possible by selecting additional variable, including selection of multiple SV callers via checkboxes and one or more numeric parameters via slider bars or checkboxes. Parameter values are easy to change and evaluation plots can be interpreted quickly. To enhance usability, Shiny-SoSV includes a user guide and example use cases.

Results

To illustrate the utility of Shiny-SoSV, we present a common use case.

Suppose, based on histopathology, a user is aware their cohort of cancer samples have tumour purities between 20% and 60%, and they want to know to what depth of coverage they should sequence their tumour samples, assuming all matched normal samples will be sequenced to the default of 30x.

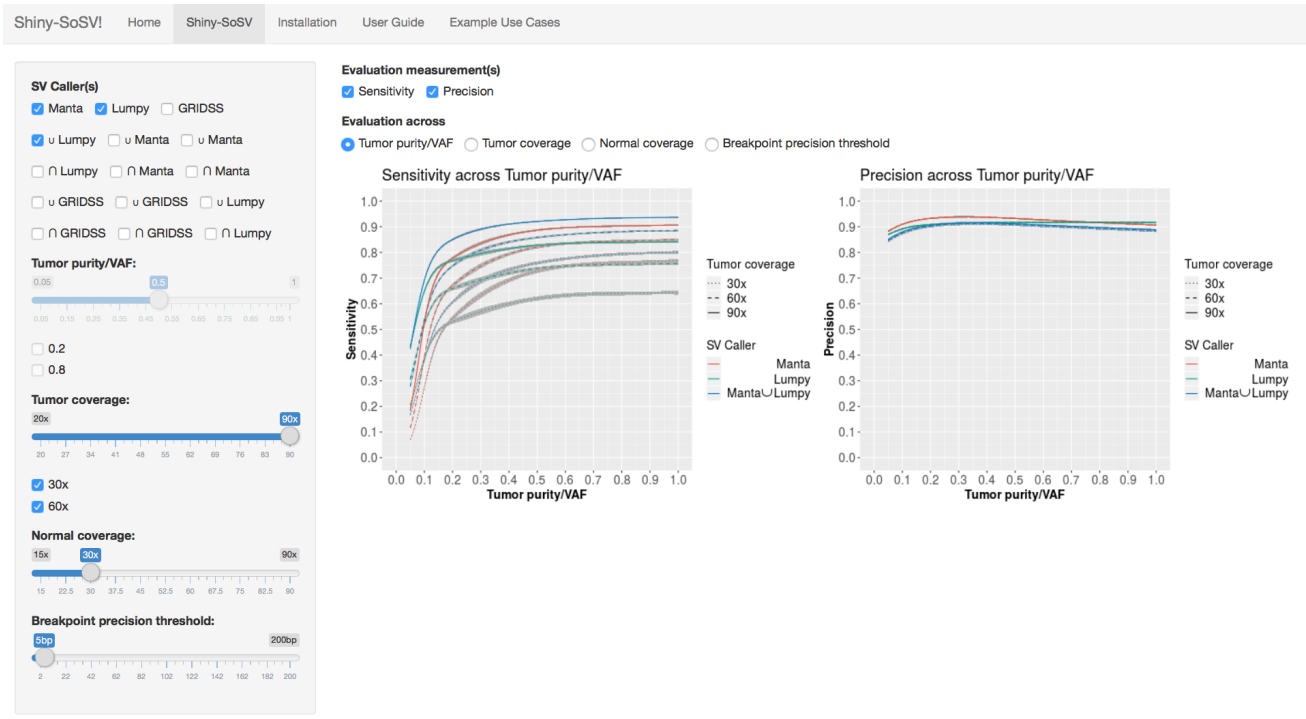

Fig. 1. Shiny-SoSV interface. Shown are sensitivity and precision estimates evaluated across variant allele frequency. In this example, the performances of two SV callers and their union callsets (distinguished by line colours) are displayed for three different tumour depths of coverage settings (line types). Confidence intervals of all estimates are shown as grey ribbons.

To address this using Shiny-SoSV, as demonstrated in **Figure 1**, the user could select to evaluate the effect of "Tumour purity/VAF" ("Evaluation across") on either or both "Sensitivity" and "Precision" ("evaluation measurements"). On the side bar, they could select individual or combination(s) of SV callers (e.g. Manta, Lumpy, and their union set) for comparison and select up to three tumour coverage settings (e.g. 30x, 60x and 90x) simultaneously, while keeping all other parameters unchanged. From this, it is immediately obvious that VAF has a great impact on sensitivity, while little impact on precision. Sensitivity of all SV callers

increases rapidly from VAF of 5% to 30%. At VAF > 30%, improvements in sensitivity notably slows for all SV callers, with Manta showing relatively larger improvements until it reaches the limit with this combination of parameters. Depending on the user's budget and objective (i.e. acceptable sensitivity and precision level), they may elect to sequence at 30x using the SV caller(s) chosen (e.g. union set of Manta and Lumpy) for all tumour samples. Alternatively, they may decide to forgo samples with tumour purity < 30%, which may free up resources allowing remaining tumour samples to be sequenced to 60x or even 90x depth of coverage.

Conclusion

Shiny-SoSV provides an easy to use and visually interactive platform for evaluating the effects of multiple variables impacting somatic SV detection. The current release allows evaluation of five common effectors of SV calls and three popular SV callers. Inclusion of addition SV callers can easily be incorporated with existing simulation datasets, while assessment of additional variables (such as mapping quality, insert sizes and nucleotide complexity) can be achieved with further simulation datasets. In sum, we believe Shiny-SoSV will enable bioinformatician, as well as non-bioinformatician, the ability to optimally design whole genome sequencing experiments for detecting SVs in cancer genomes.

Acknowledgements

We acknowledge the high-performance computing resources generously provided by the National Computational Infrastructure (Raijin), the Garvan Institute of Medical Research (Wolfpack) and the University of Sydney (Artemis).

Funding

TG is supported by an Australian Government Research Training Program Scholarship. VMH is supported by the University of Sydney Foundation and Petre Foundation, Australia. This work was supported in part by a Movember Revolutionary Team Award funded by Movember Australia and Prostate Cancer Foundation Australia.

References

Baca, S.C., et al. (2013) Punctuated evolution of prostate cancer genomes. Cell, 153, 666-677.

Cameron, D.L., et al. (2017) GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res, 27, 2050-2060.

Chaisson, M.J.P., et al. (2019) Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nature Communications, 10, 1784.

Chen, K., et al. (2009) BreakDancer: An algorithm for high resolution mapping of genomic structural variation. Nature methods, 6, 677-681.

Chen, X., et al. (2016) Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics, 32, 1220-1222.

Dixon, J.R., et al. (2018) Integrative detection and analysis of structural variation in cancer genomes. Nature genetics, 50, 1388-1398.

Griffith, M., et al. (2015) Optimizing cancer genome sequencing and analysis. Cell Syst, 1, 210-223.

Layer, R.M., et al. (2014) LUMPY: a probabilistic framework for structural variant discovery. Genome Biol, 15, R84.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754-1760.

Liu, B., et al. (2015) Structural variation discovery in the cancer genome using next generation sequencing: computational solutions and perspectives. Oncotarget, 6, 5477-5489.

Northcott, P.A., et al. (2014) Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. Nature, 511, 428-434.

Rausch, T., et al. (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics (Oxford, England), 28, i333-i339.

Sims, D., et al. (2014) Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics, 15, 121.

Tattini, L., et al. (2015) Detection of Genomic Structural Variants from Next-Generation Sequencing Data. Frontiers in bioengineering and biotechnology, 3, 92-92.

Tubio, J.M.C. (2015) Somatic structural variation and cancer. Briefings in Functional Genomics, 14, 339-351.

Wood, S.N. 2011 Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society (B), 73, 3-36.

Xia, L.C., et al. (2018) SVEngine: an efficient and versatile simulator of genome structural variations with features of cancer clonal evolution. Gigascience, 7.

Ye, K., et al. (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics, 25, 2865-2871.