Shiny-SoSV: A web-based performance calculator for somatic structural variant detection

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

[1]Garvan Institute of Medical Research, Darlinghurst, New South Wales, Australia

[2]Central Clinical School, University of Sydney, Camperdown, New South Wales, Australia

[3]St Vincent's Clinical School, University of New South Wales, Randwick, New South Wales, Australia

* Corresponding author

Email: eva.chan@health.nsw.gov.au (EC)

Abstract

Somatic structural variants are an important contributor to cancer development and evolution. Accurate detection of these complex variants from whole genome sequencing data is influenced by a multitude of parameters. However, there are currently no tools for guiding study design nor are there applications that could predict the performance of somatic structural variant detection. To address this gap, we developed Shiny-SoSV, a user-friendly web-based calculator for determining the impact of common variables on the sensitivity and precision of somatic structural variant detection, including choice of variant detection tool, sequencing depth of coverage, variant allele fraction, and variant breakpoint resolution. Using simulation studies, we determined singular and combinatoric effects of these variables, modelled the results using a generalised additive model, allowing structural variant detection performance to be predicted for any combination of predictors. Shiny-SoSV provides an interactive and visual platform for users to easily compare individual and combined impact of different parameters. It predicts the performance of a proposed study design, on somatic structural variant detection, prior to the commencement of benchwork. Shiny-SoSV is freely available at https://hcpcg.shinyapps.io/Shiny-SoSV with accompanying user's guide and example use-cases.

Introduction

Somatic structural variants (SVs) are large (> 50 bp) genomic rearrangements that arise in tumours and could directly contribute to cancer development and progression [1-3]. 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 or closer apart than expected are suggestive of a deletion or insertion event

respectively, while read pairs aligned with inconsistent orientation are indicative of an inversion or duplication event. Attempts to improve SV detection have thus spurred many computational developments, resulting in the publication and availability of numerous SV detection tools (e.g. [4-8]). These tools are based on different, sometimes overlapping algorithms and multiple evaluation studies have shown that these tools differ substantially in their sensitivity and breakpoint precision [9-11].

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 captured as harbouring the variant at a given genomic locus). Although histopathology can provide an overall estimate of tumour purity, it typically only reflects an approximate estimate of VAF, due to further complexity created through tumour sub-clonality and variability between histopathology and sequencing samples. Although deeper sequencing may increase sensitivity [10], numerous studies have shown that the benefit of increasing coverage does saturate (e.g. [12]). 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 easily made, especially by non-bioinformaticians in clinical settings. To address this, we have developed Shiny-SoSV, a web-based interactive application to help estimate the performance (sensitivity and precision) of somatic SV detection based on four user-modifiable parameters: VAF, sequencing depth of tumour and matched-normal samples, and SV breakpoint precision.

Materials and methods

Simulation of structural variants

To evaluate the sensitivity and accuracy of SV detection, a simulation study was devised. Two sets of SVs each including 200 of each of six SV types (deletion (DEL), duplication (DUP), inversion (INV), domestic insertion (DINS), foreign insertion (FINS) and translocation (TRA)) were simulated with SVEngine [13], as previously described [14](**S1 Table**). For each SV type, variable lengths were included ranging from 50 bp to 1,000,000 bp. Here, SV length is the length of the rearranged sequence; for example, the SV length of a TRA is the length of the translocated piece of DNA fragment.

SVs were randomly distributed along the genome (GRCh38) without overlap, masking N-gap, centromeric and telomeric regions. N-gap and centromeric regions were based on gap [last updated 24/12/2013] and centromeres [last update 17/08/2014] tables from the University of California at Santa Cruz (UCSC) Genome Browser, while telomeric regions are defined as the 10 Mbp of both ends of each chromosome of GRCh38.

Simulation of short paired-end reads

NGS data was simulated to reflect the typical matched tumour-normal whole genome sequencing (WGS) approach for cancer genomics, namely using paired-end short-read sequencing, where genomic DNA is fragmented to a typical size range of 500 bp (insert size) and the two ends of each fragment sequenced inwards to up to 150 bp (read length).

The first 1,200 simulated SVs were treated as germline and spiked into both the normal and tumour genomes (FASTA format), while the second simulation of 1,200 SVs were treated as cancer and added to the tumour genome. Paired short-reads were sampled using SVEngine [13] from the altered FASTA to 120x depth of coverage. Insert size of read-pairs was simulated with a normal distribution with mean of 500 bp and standard deviation of 100 bp, while read length was set to 150 bp. Other parameters were kept as SVEngine default, including random embedding of sequencing error and small variants (base error rate = 0.02, rate of mutation = 0.001, fraction of indels = 0.15, probability an indel is extended = 0.3). Sequencing reads (FASTQ format) were aligned to human genome reference GRCh38 using BWA-MEM v0.7.17-r1194 [15], generating alignment files (BAM format). Different depths of sequencing coverage were obtained by subsampling from the two 120x datasets using Picard DownsampleSam (http://picard.sourceforge.net). Different tumour purity levels (VAF) were emulated by merging varying ratios of normal to tumour aligned reads, using Picard MergeSamFiles, to create the final tumour files (BAM). Here, tumour purity and VAF are used interchangeably as SVs are assumed independent of each other.

The simulation was conducted in three replicates (Set 1-3) to include a large combination of depths of coverage of the normal samples, tumour samples and VAF (**Table 1**). To examine the effect of stochastic noise, each of the 216 *in silico* experiments in Set 1 were replicated another two times using the same simulated SV set but different sampling sequencing reads. In all, this resulted in a total of 1080 tumour/normal pairs of BAM files.

Table 1. Simulated and evaluated variables

Tumour coverage			Normal coverage			VAF			*T* (bp)[1]		
Set1	Set2	Set3	Set1	Set2	Set3	Set1	Set2	Set3	Set1	Set2	Set3
20x	25x	20x	15x	25x	15x	0.05	0.12	0.05	2	8	2
30x	38x	35x	30x	38x	35x	0.1	0.25	0.15	5	15	10
45x	50x	40x	45x	50x	40x	0.2	0.4	0.3	50	30	60
60x	70x	60x	60x	70x	60x	0.5	0.6	0.5	100	90	100
75x	85x	80x	75x	85x	80x	0.8	0.75	0.7	150	120	130
90x	95x	100x	90x	95x	100x	1	0.95	0.9	200	180	190

[1]*T* is breakpoint precision threshold. This is the maximum allowed difference between called and simulated breakpoint positions to be defined as true positive.

Somatic SV detection

Somatic SVs were detected using five SV callers, namely Manta [4], Lumpy [5], GRIDSS [6], Delly [7] and SvABA [8] for each tumour/normal pair and high-confidence calls were post-filtered as described previously [14]. Each SV caller was executed using default parameters.

Defining true positive calls and concordant callsets

Candidate SV callsets were compared against "true" simulated SV sets as previously described [14]. In brief, a true positive (TP) SV call must meet two criteria: i) reported SV type must match the simulated type and ii) detected breakpoints must be within *T* bp from the simulated breakpoints. To pass the second criterion, all breakpoint positions corresponding to a SV event must be within a prescribed breakpoint precision threshold (*T*). We have applied varying values of *T* for each replicate dataset (**Table 1**).

The total number of true positives is calculated as the number of calls in a callset satisfying both TP criteria. The number of false positives (FP) is the number of SV calls in a callset not satisfying either TP criteria. Thus, following the confusion matrix, we define: sensitivity = TP / (TP + FN) and precision = TP / (TP + FP).

The union and intersection callsets for each pair of SV callers were also derived, resulting in another 40 callsets for each simulated dataset. Two SV calls were considered the same if they have matching SV type [14] and their reported breakpoint positions are within 5 bp of each other. The union callset of two SV callers is SV calls detected by either Caller_1 or Caller_2, while the intersection callset are calls reported by both Caller_1 and Caller_2. In both union and intersection cases, Caller_1 is the "dominant caller" such that any overlapping calls in the final callset is taken from the output from Caller_1 (including coordinates and SV types).

Predictive model selection

A generalised additive model (GAM) was used to assess and predict the relationships between predictor variables (VAF, tumour coverage, normal coverage and breakpoint precision threshold *T*) and response variables (sensitivity and precision) for each SV caller and all pairs of callers. The choice of GAM was motivated by non-linear relationships observed between some response and predictor variables.

In particular, we observed VAF to have a non-linear effect on sensitivity for all SV callers (**Fig 1a and S1a Fig**) and precision on some SV callers (**Fig 2a**). However, the non-linear effect of VAF on precision is less obvious at lower breakpoint precision resolution (higher *T*) and for some SV callers, such as GRIDSS (**S1b Fig**). Tumour coverage has a greater non-linear impact on sensitivity and precision on samples at low VAF (**S2 Fig**). Different SV callers have different breakpoint detection resolution, as previously described [14] and SvABA was found to have the lowest breakpoint resolution among the five SV callers examined (**Fig 3)**. Consequently, breakpoint precision threshold (*T*) has greater effect for SvABA and less so for other callers (**S3 Fig**). Additionally, most SV callers have lower breakpoint resolution at low tumour coverage and low VAF (e.g. 20x and 0.2 respectively in **Fig 3)**. The impact of *T* on sensitivity, therefore, appears to be variably dependent on tumour coverage and VAF. For example, the sensitivity of Manta increases with increasing *T* at low VAF (e.g. 0.2) and is only obvious at tumour coverage < 60x (**S3a Fig**). Similarly, precision is predominantly influenced by *T* at low tumour coverage and VAF (**S3b Fig**). Normal coverage has no notable impact,

except for Manta, which was observed to have only a minor effect on sensitivity and only at low normal coverage (< 30x) (**S4 Fig**).

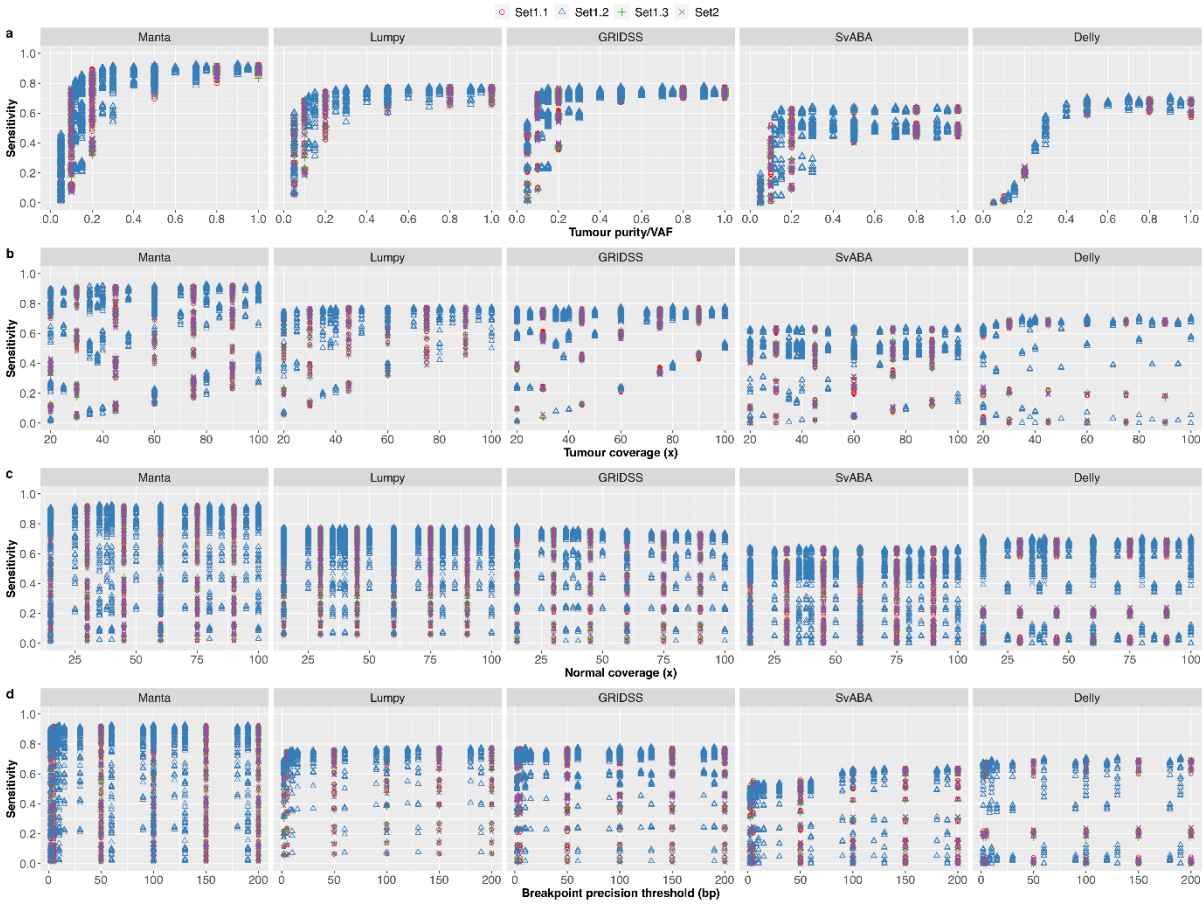

Fig 1. Relationship between sensitivity and predictor variables. Shown are the effects of (a) tumour purity/VAF, (b) tumour coverage, (c) normal coverage and (d) breakpoint precision threshold (d) on somatic SV detection sensitivity for five SV callers (Manta, Lumpy, GRIDSS, SvABA, Delly).

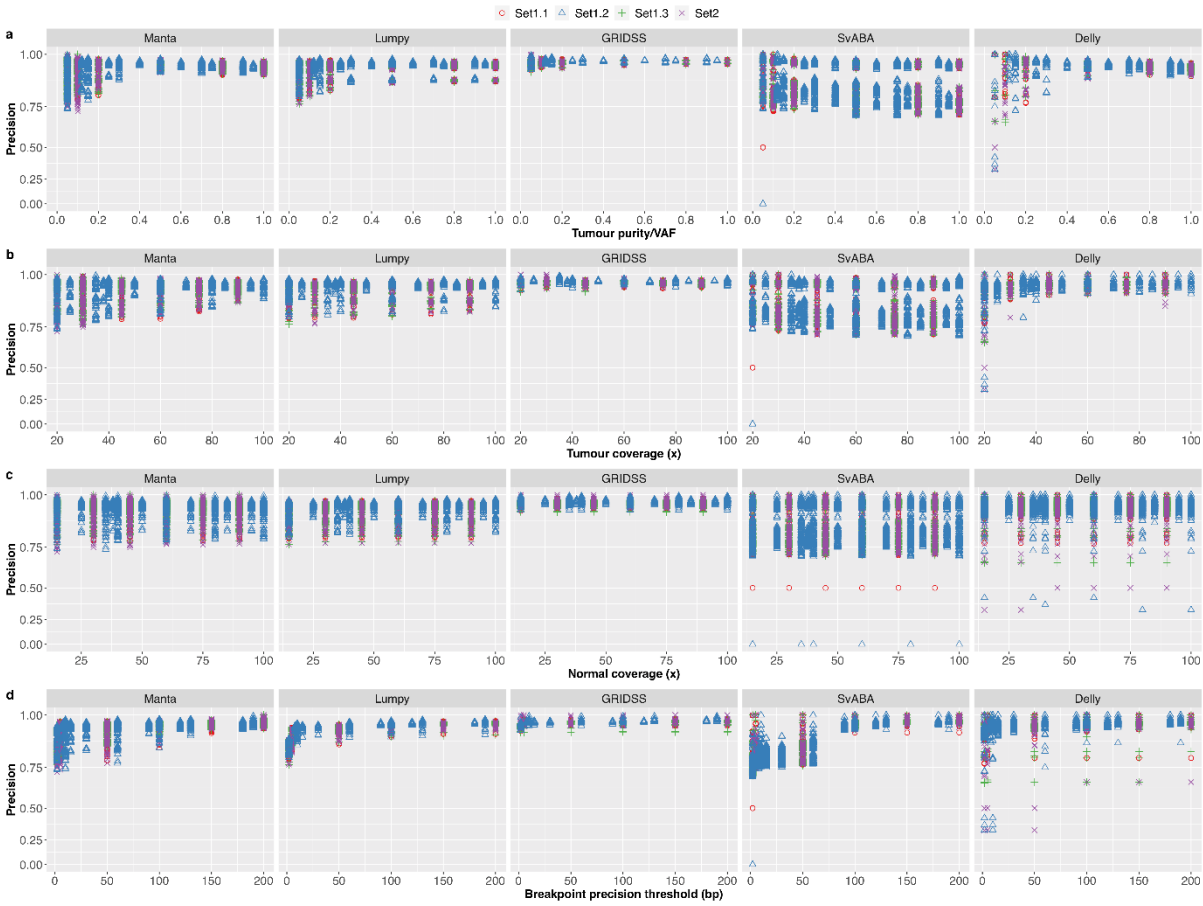

Fig 2. Relationship between precision and predictor variables. Shown are the effects of (a) tumour purity/VAF, (b) tumour coverage, (c) normal coverage and (d) breakpoint precision threshold (d) on somatic SV detection precision, represented in log-scale, for five SV callers (Manta, Lumpy, GRIDSS, SvABA, Delly).

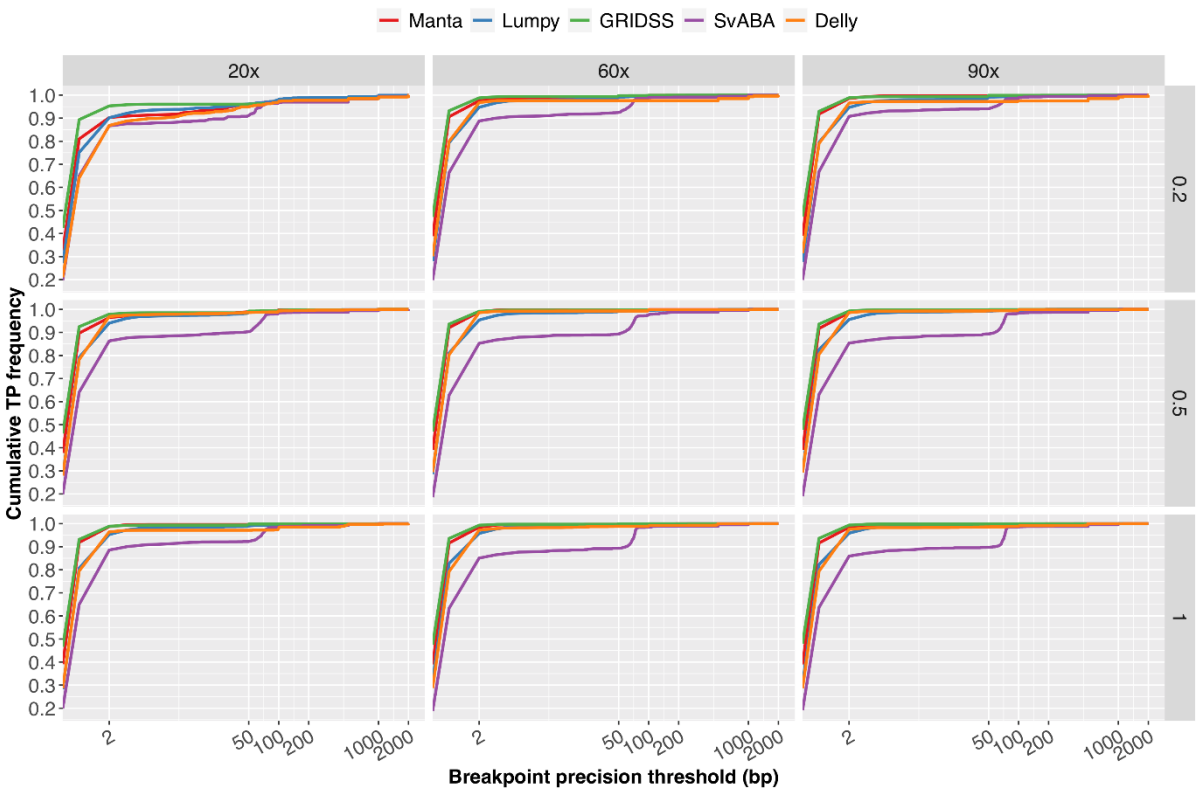

Fig 3. Breakpoint resolution of structural variant calling. Results are based on simulation SV Set 1 with tumour coverage of 20x, 60x and 90x, matched normal coverage of 60x and variant allele frequency of 0.2, 0.5 and 1. This figure shows the cumulated frequency of the number true positive structural variants detected by each caller for breakpoint precision thresholds less than 2,000 bp.

In order to select the model with the most appropriate set of parameters and that best explain the relationship between the predictor and response variables, we examined on five sensitivity models (Model (1) - (5)) and six precision models (Model (6) - (11)) independently for each SV caller. To account for non-linear joint relationships, smooth functions (f) were applied to the tensor product of selected predictor variables. Smoothing parameters were estimated using restricted maximum likelihood (REML), rather than the default generalised cross validation (GCV), to avoid local minima [16]. Further, as the response variables are proportions with values between 0 and 1, the beta regression (betar) with logistic link function was used.

$$\begin{aligned} Sensitivity \approx\ & f_1(VAF \otimes Tumour\ coverage) + f_2(VAF \otimes T) \\ & + f_3(VAF \otimes Normal\ coverage) + f_4(Tumour\ coverage \otimes T) \end{aligned} \tag{1}$$

$$\begin{aligned} Sensitivity \approx\ & f_1(VAF \otimes Tumour\ coverage) + f_2(VAF \otimes T) \\ & + f_3(VAF \otimes Normal\ coverage) \end{aligned} \tag{2}$$

$$\begin{aligned} Sensitivity \approx\ & f_1(VAF \otimes Tumour\ coverage) + f_2(VAF \otimes T) \\ & + f_3(Tumour\ coverage \otimes T) + Normal\ coverage \end{aligned} \tag{3}$$

$$\begin{aligned} Sensitivity \approx\ & f_1(VAF \otimes Tumour\ coverage) + f_2(VAF \otimes T) \\ & + Normal\ coverage \end{aligned} \tag{4}$$

$$Sensitivity \approx f_1(VAF \otimes Tumour\ coverage) + T + Normal\ coverage \tag{5}$$

$$Precision \approx f_1(VAF \otimes Tumour\ coverage) + f_2(VAF \otimes T) + f_3(Tumour\ coverage \otimes T) + Normal\ coverage \quad (6)$$

$$Precision \approx f_1(VAF \otimes Tumour\ coverage) + f_2(VAF \otimes T) + Normal\ coverage \quad (7)$$

$$Precision \approx f_1(VAF \otimes Tumour\ coverage) + f_2(Tumour\ coverage \otimes T) + Normal\ coverage \quad (8)$$

$$Precision \approx f_1(VAF \otimes Tumour\ coverage) + T + Normal\ coverage \quad (9)$$

$$Precision \approx f_1(VAF \otimes T) + Tumour\ coverage + Normal\ coverage \quad (10)$$

$$Precision \approx VAF + T + Tumour\ coverage + Normal\ coverage \quad (11)$$

Replicate Sets 1 and 2 (**Table 1**) were used for model selection and Set 3 was only used for model validation. To select the models for sensitivity and precision, we used cross-validation for accuracy calculation of each model for each SV caller. Each time, one simulated value of a variable (e.g. all data with 20x tumour coverage) was left out for model fitting and error estimation. The average root mean squared error (RMSE) and mean absolute error (MAE) for each candidate model and five SV callers are reported in **Table 2**. For sensitivity, Model 1 was selected for Manta and SvABA and Model 5 for Lumpy and GRIDSS, as they displayed lowest RMSE and MAE. For Delly, four models achieved similar RMSE and MAE and so the simplest model (Model 5) was selected in order to avoid overfitting. In addition, we found that matched normal coverage has no significant effect on Lumpy, SvABA and Delly sensitivity, therefore, the term was excluded in the models for these SV callers. For precision, Model 6 was selected for Manta and Model 7 for Lumpy, GRIDSS and Delly. For precision of SvABA, Model 9 was selected based on lowest RMSE. As with sensitivity, matched normal coverage was not observed to have any significant effect on GRIDSS, and thus was excluded from its final prediction models.

Table 2. Sensitivity and precision model comparison and selection

	Error Estimate	Models	Manta	Lumpy	GRIDSS	SvABA	Delly
Sensitivity	RMSE	(1)	0.037	0.038	0.047	0.050	0.017
		(2)	**0.037**	0.038	0.047	0.051	0.016
		(3)	0.040	0.038	0.047	**0.049**	0.016
		(4)	0.040	0.038	0.047	0.050	0.016
		(5)	0.039	**0.037**	**0.045**	0.052	**0.016**
	MAE	(1)	0.028	0.030	0.036	0.038	0.013
		(2)	**0.028**	0.030	0.036	0.038	0.013
		(3)	0.030	0.030	0.035	**0.037**	0.013
		(4)	0.030	0.030	0.035	0.037	0.013
		(5)	0.030	**0.030**	**0.035**	0.040	**0.013**
Precision	RMSE	(6)	**0.012**	0.016	0.008	0.039	0.027
		(7)	0.013	**0.016**	**0.008**	0.039	**0.027**
		(8)	0.017	0.016	0.009	0.039	0.032

		(9)	0.023	0.020	0.009	**0.038**	0.035
		(10)	0.021	0.017	0.011	0.061	0.040
		(11)	0.030	0.023	0.011	0.071	0.046
	MAE	(6)	**0.009**	0.012	0.006	0.028	0.016
		(7)	0.010	**0.012**	**0.006**	0.028	**0.016**
		(8)	0.013	0.013	0.006	**0.027**	0.019
		(9)	0.018	0.016	0.007	0.028	0.021
		(10)	0.015	0.013	0.009	0.045	0.024
		(11)	0.023	0.018	0.008	0.053	0.030

The R package *mgcv* default of thin plate regression spline (tp) was used as the smoothing basis. As tp tend to result in lower mean squared error [16], it was implemented for our purpose. Basis dimension was set as the default for all smoothing terms.

The models for pairwise union and intersection of SV callers were chosen to be the same as the dominant caller. GAM models were then fitted for each SV caller and each pair of SV callers.

Web-based application design

Shiny-SoSV is a web application developed using R package shiny v1.3.2, and is hosted on shinyapps.io, built to provide a visual platform for exploring the behaviour of sensitivity and precision with various predictor variables through two main interactive plots. R objects of class "gam" from the GAM models make up the data underlying Shiny-SoSV. Manipulation of response variables by users via the web app are registered by R/shiny, which then signals a call to the function predict(), returning predicted sensitivity and precision values along with estimated standard errors. Plots of predicted sensitivity and precision with confidence intervals are generated with R package ggplot2 version 3.1.1 [17]. Comparison and interaction effects of two or more predictors are possible by selecting additional variables, including selection of multiple SV callers via checkboxes and one or more numeric parameters via slider bars or checkboxes (**Fig 4)**.

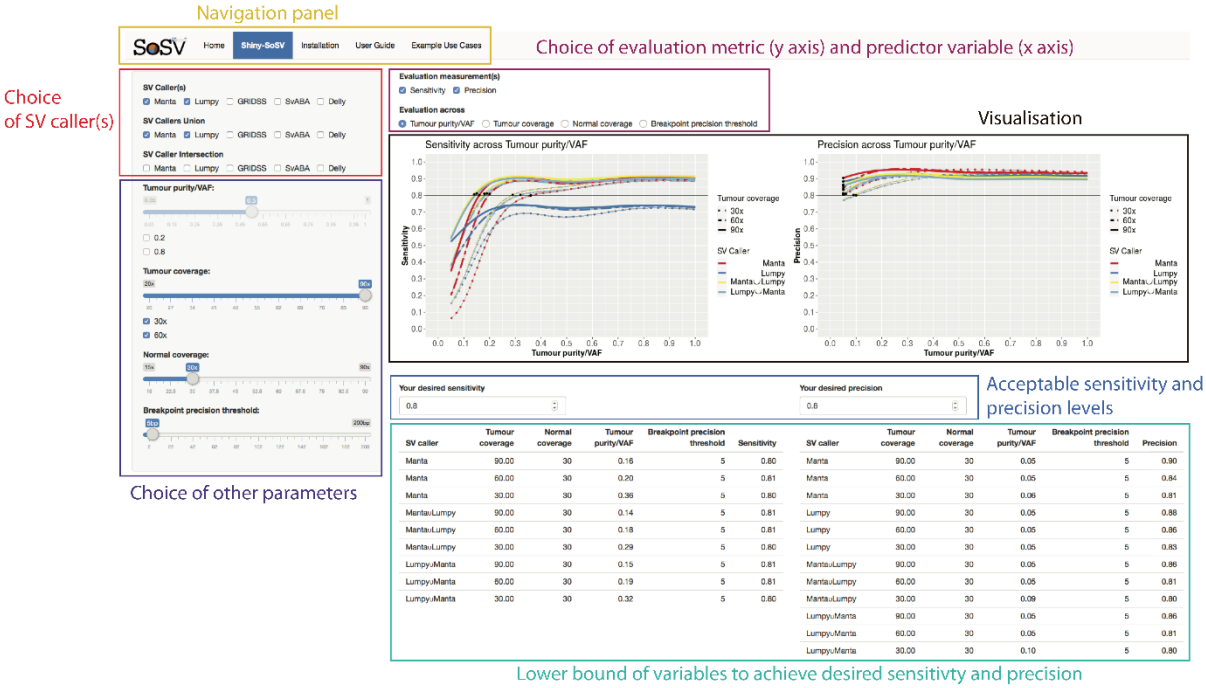

Fig 4. Shiny-SoSV user interface. Shown is a snapshot of Shiny-SoSV, displaying sensitivity and precision estimates evaluated across variant allele frequency (Tumour purity/VAF) (Black box labelled "Visualisation"). In this example, the performances of two SV callers (Manta and Lumpy) and their union callsets (distinguished by line colours) are displayed (controlled by red box labelled "Choice of SV caller(s)") for three different tumour depths of coverage settings (line types; 30x, 60x, 90x) (controlled by indigo coloured box labelled "Choice of other parameters"). The desired sensitivity and precision are set as 80% (Blue box labelled "Acceptable sensitivity and precision levels"). Confidence intervals of all estimates are shown as grey ribbons, which may be difficult to see due to tight intervals.

Results and Discussion

Prediction model validation

To test the robustness and predictive ability of the models, we used independent replicate Set3 shown in Table 1, generated using a separate simulated SV set and different combinations of parameter values. The results suggest that we can predict sensitivity and precision prediction within 8% error rate (**Table 3**).

Table 3. Sensitivity and precision model validation

Models	Error Estimates	Manta	Lumpy	GRIDSS	SvABA	Delly
Sensitivity	RMSE	0.027	0.031	0.036	0.045	0.025
	MAE	0.021	0.025	0.029	0.034	0.020
Precision	RMSE	0.014	0.016	0.013	0.079	0.047
	MAE	0.010	0.014	0.011	0.031	0.024

In addition, we validated our prediction model against the somatic SV evaluation results from the ICGC-TCGA DREAM Somatic Mutation Calling Challenge [10], which conducted a crowdsourced benchmarking of SV callers and reported results on three different simulated tumours from 15 teams. In this challenge, each tumour sample was simulated with SVs spiked into a sub-sampling of Illumina paired-end short-read sequencing reads from a cell line using BAMSurgeon [10] and each matched-normal sample was derived from a separate (non-overlapping) sub-sampling of the same original BAM file. Both of tumour and matched normal samples were sampled to around 30-40x coverage. Three synthetic datasets (denoted as in_silico 1 – 3) were generated, based on two different cell lines and three different SV sets containing different types and numbers of SVs (371, 655 and 2,886) spiked in at different frequencies (100%, 80% and mixture of subsets of SV at 20%, 33%, and 50% VAF). Detailed description of the challenge and datasets can be found at https://www.synapse.org/#!Synapse:syn312572/wiki/62018. SVs reported by each team were evaluated based on a single criterion that the called SV breakpoint is within 100 bp of a simulated SV, while correct SV type was not required. So, this TP criterion is more lenient than the TP criteria used in our study. Among the teams that participated, Teams 1 and 2 used Delly and Manta respectively, which allowed their sensitivity and precision to be predicted using our GAM models and compared to their reported evaluation results (**Fig 5**). Teams were allowed to use different versions and parameters on different datasets and make multiple submissions for each dataset, which resulted in multiple sensitivity and precision results for each team and dataset combination (e.g. sensitivity ranging from 53% to 74% of Team 1 for in_sillico 3). Overall, our models achieved low error rate (MAE = 3.3%) for precision but higher error rate (MAE = 12.2%) for sensitivity prediction, especially for in_silico 2 and 3. This is likely due to a combinations of differences in simulation and evaluation methods, and different versions and parameters of SV callers used by the DREAM Challenge participants [10].

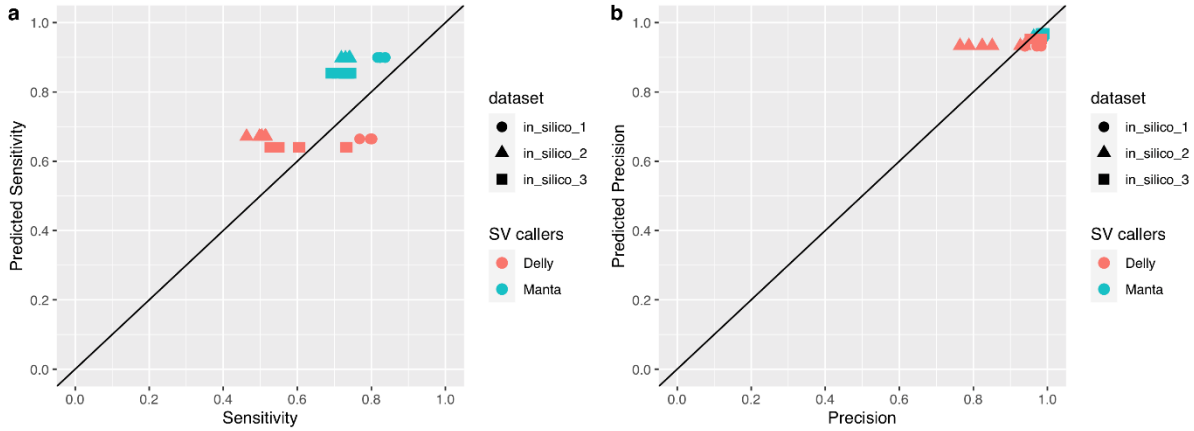

Fig 5 Comparison of GAM prediction and observed performance from the ICGC-TCGA DREAM Challenge. Shown are predicted sensitivity (a) and precision (b) by GAM model versus reported sensitivity and precision from Lee et al.,2018 [10]. Team 1 and Team 2 have used SV callers of Delly and Manta respectively. Sensitivity and precision are predicted based on predictor variables of tumour and normal coverage as 30x, breakpoint precision threshold as 100 bp and VAF as 100%, 80% and 50% for three datasets respectively.

Application features

The aim of Shiny-SoSV is to support decision-making on variables impacting one's ability to detect somatic SVs from WGS data. In particular, this app was designed to empower users in planning their WGS experiment for efficient somatic SV detection. To this end, five key features were implemented.

1. *Inclusion of important predictor variables.* We have included five variables in Shiny-SoSV. With the exception of VAF, all can be modified by the study investigator prior to the commencement of sequencing. This is important as it allows users to adjust experimental plans accordingly. Even though VAF is intrinsic to the biology of the sample, it is by far the most important predictor for somatic variant detection [10, 18]. Therefore, VAF is included mainly to determine minimum required tumour purity and support tuning of other modifiable variables.
2. *Predictor variables are easy to modify.* To make it user friendly, we have implemented tick boxes and slider bars for value selection across a range of variable choices. For example, users can choose any single or pairs of SV caller(s).
3. *Intuitive visualisation of the impact of predictor variables on responses.* To compare somatic SV detection performance with different choices of parameter values, GAM prediction models were implemented allowing prediction curves to be dynamically generated for easy visualisation. The prediction can be visualised via line plots with confidence intervals based on choices of different combinations of variables, in one figure.
4. *Ability to evaluate multiple variables simultanesously.* For example, users can examine the change of sensitivity (y-axis) across VAF (x-axis) with different choices of SV callers (line colour) and tumour coverages (line types) (**Fig 4**).
5. *Simple reporting of minimum requirements to achieve user's objective.* To further simplify decision-making, a summary table of minimum required value for each predictor variable, including SV callers, depth of coverage and tumour purity and breakpoint resolution, is reported for a user-defined detection sensitivity and precision level.

Application limitations

While we have endeavoured to create a performance calculator for real-world use, we recognise certain limitations and challenges. Firstly, the performance prediction was based on a series of simulated data, which we appreciate cannot fully encompass the complexity of real tumour sample and errors in real sequencing processes. However, by concurrently including both germline and somatic SVs, and including a range of SV types and SV sizes, we believe the simulation results provide a baseline estimate. Secondly, real unknown VAF can be lower than histopathological estimates of tumour purity, due to potential tumour sub-clonality and heterogeneity. We therefore present the prediction here to represent the upper limits of sensitivity and precision that can be achieved. Thirdly, we have included four predictor variables totalling 6,480 (3,888 unique) permutations of predictor values for prediction model fitting of each callset. However, other variables, such as average mapping quality, read length, insert sizes and nucleotide complexity, have been shown to impact variant detection, albeit

with lower effect size [9, 10]. Adding more variables and more data points could provide a better prediction model and more sophisticated calculator for SV detection performance prediction.

Application utility

The web-based app is hosted indefinitely at https://hcpcg.shinyapps.io/shiny-sosv by Shiny from RStudio. Both the input data to the GAM and R script for GAM fitting, as well as the resulting GAM R objects are provided with the Git release, available at https://github.com/tgong1/Shiny-SoSV. Additionally, the source code for launching the shiny app is available at this GitHub repository, allowing both data and app versions to be tracked. An extensive User Guide is provided with Shiny-SoSV release, detailing the user-interface and available options. To help further enhance usability, multiple end-user cases also accompany Shiny-SoSV.

To illustrate the utility of Shiny-SoSV here, we present a common example 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.

As demonstrated in **Fig 4**, the Shiny-SoSV user can select to evaluate the effect of "Tumour purity/VAF" ("Evaluation across") on either or both "Sensitivity" and "Precision" ("evaluation measurements"). On the side bar, the user can select individual or combination(s) of SV callers (e.g. Manta, Lumpy, and their union set) for comparison and 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 (saturation) with this combination of parameters for this caller. The table below the plots provides the user with the information required to make decisions on sequencing depth, which is further dependent on his/her budget and objective (i.e. acceptable sensitivity and precision level). For example, for a desired sensitivity and precision of 80%, the user may elect to sequence at 60x using the SV caller(s) chosen (e.g. union set of Manta and Lumpy) for all samples. However, if the resulting cost is prohibitive, the user may choose to sequence only the subset of samples with tumour purity > 29% to up to 30x depth of coverage whilst still achieving the desired 80% sensitivity and precision levels.

Conclusions

Shiny-SoSV provides an easy to use and interactive graphical user interface for evaluating the effects of multiple variables impacting somatic SV detection. The current release allows evaluation of common effectors of SV detection using five 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 efforts. In sum, we believe Shiny-SoSV will enable bioinformaticians, and importantly also non-bioinformaticians, to optimally design WGS experiments for detecting SVs in cancer genomes.

Supporting information

S1 Fig. The joint impact of tumour purity/VAF and breakpoint precision threshold on sensitivity and precision. Shown are the effect of the interaction of VAF and breakpoint precision threshold (by colour) on sensitivity (a) and precision (b) for five SV callers (Manta, Lumpy, GRIDSS, SvABA, Delly). Results are based on simulation data Set 1 with tumour coverage of 20x, 60x and 90x and normal coverage of 60x.

S2 Fig. The joint impact of tumour coverage and tumour purity/VAF on sensitivity and precision. Shown are the effect of the interaction of tumour coverage and VAF (by colour) on sensitivity (a) and precision (b) for five SV callers (Manta, Lumpy, GRIDSS, SvABA, Delly). Results are based on simulation data Set 1 with normal coverage of 60x and breakpoint precision threshold of 5bp, 100bp and 200bp.

S3 Fig. The joint impact of breakpoint precision threshold and tumour coverage on sensitivity and precision. Shown are the effect of the interaction of breakpoint precision threshold and tumour coverage (by colour) on sensitivity (a) and precision (b) for five SV callers (Manta, Lumpy, GRIDSS, SvABA, Delly). Results are based on simulation data Set 1 with normal coverage of 60x and VAF of 0.2, 0.5 and 1.

S4 Fig. The joint impact of normal coverage and VAF on sensitivity and precision. Shown are the effect of the interaction of normal coverage and VAF (by colour) on sensitivity (a) and precision (b) for five SV callers (Manta, Lumpy, GRIDSS, SvABA, Delly). Results are based on simulation data Set 1with tumour coverage of 60x and breakpoint precision threshold of 5bp, 100bp and 200bp.

S1 Table. Simulation SV sets.

Acknowledgements

We gratefully acknowledge the comments and suggestions on the manuscript from Dr Timothy Peters at Garvan Institute of Medical Research.

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).

T.G. is funded by an Australian Government Research Training Program Scholarship, E.K.F.C is funded in part by a Movember Revolutionary Team Award (PRoMis) from Movember Australia and the Prostate Cancer Foundation Australia, and V.M.H. is funded by the University of Sydney Foundation and Petre Foundation, Australia.

REFERENCES

1. Tubio JM. Somatic structural variation and cancer. Brief Funct Genomics. 2015;14(5):339-51. doi: 10.1093/bfgp/elv016. PubMed PMID: 25903743.
2. Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, et al. Punctuated evolution of prostate cancer genomes. Cell. 2013;153(3):666-77. doi: 10.1016/j.cell.2013.03.021. PubMed PMID: 23622249.
3. Northcott PA, Lee C, Zichner T, Stütz AM, Erkek S, Kawauchi D, et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. Nature. 2014;511(7510):428-34. doi: 10.1038/nature13379.
4. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Kallberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016;32(8):1220-2. doi: 10.1093/bioinformatics/btv710. PubMed PMID: 26647377.
5. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014;15(6):R84. doi: 10.1186/gb-2014-15-6-r84. PubMed PMID: 24970577; PubMed Central PMCID: PMCPMC4197822.
6. Cameron DL, Schroder J, Penington JS, Do H, Molania R, Dobrovic A, et al. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res. 2017;27(12):2050-60. doi: 10.1101/gr.222109.117. PubMed PMID: 29097403; PubMed Central PMCID: PMCPMC5741059.
7. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics (Oxford, England). 2012;28(18):i333-i9. Epub 09/03. doi: 10.1093/bioinformatics/bts378. PubMed PMID: 22962449.
8. Wala JA, Bandopadhayay P, Greenwald NF, O'Rourke R, Sharpe T, Stewart C, et al. SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Research. 2018;28(4):581-91. doi: 10.1101/gr.221028.117.
9. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biology. 2019;20(1):117. doi: 10.1186/s13059-019-1720-5.
10. Lee AY, Ewing AD, Ellrott K, Hu Y, Houlahan KE, Bare JC, et al. Combining accurate tumor genome simulation with crowdsourcing to benchmark somatic structural variant detection. Genome Biology. 2018;19(1):188. doi: 10.1186/s13059-018-1539-5.
11. Cameron DL, Di Stefano L, Papenfuss AT. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. Nature communications. 2019;10(1):3240-. doi: 10.1038/s41467-019-11146-4. PubMed PMID: 31324872.
12. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nature Communications. 2019;10(1):1784. doi: 10.1038/s41467-018-08148-z.
13. Xia LC, Ai D, Lee H, Andor N, Li C, Zhang NR, et al. SVEngine: an efficient and versatile simulator of genome structural variations with features of cancer clonal evolution. Gigascience. 2018;7(7). doi: 10.1093/gigascience/giy081. PubMed PMID: 29982625; PubMed Central PMCID: PMCPMC6057526.

14. Gong T, Hayes VM, Chan EKF. Detection of somatic structural variants from short-read next-generation sequencing data. Briefings in Bioinformatics. In press. doi: 10.1093/bib/bbaa056.
15. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60. doi: 10.1093/bioinformatics/btp324. PubMed PMID: 19451168; PubMed Central PMCID: PMCPMC2705234.
16. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society (B). 2011;73(1):3-36.
17. Wickham H. ggplot2: Elegent Graphics for Data Analysis: Springer-Verlag New York; 2016.
18. Liu B, Conroy JM, Morrison CD, Odunsi AO, Qin M, Wei L, et al. Structural variation discovery in the cancer genome using next generation sequencing: computational solutions and perspectives. Oncotarget. 2015;6(8):5477-89. doi: 10.18632/oncotarget.3491. PubMed PMID: 25849937; PubMed Central PMCID: PMCPMC4467381.