SVCROWS: A User-Defined Tool for Delineating Biologically Significant Structural Variants in Heterogeneous Datasets

Authors: 1Noah Brown*, 2Charles Danis, 3Vazira Ahmedjonova, +Jennifer L. Guler *

Abstract

Summary: SVCROWS is an option-rich R package that generates summarized and quantified structural variant (SV) regions from large and heterogeneous datasets. By using a size-weighted reciprocal overlap framework for comparisons, SVCROWS ensures that regions are delineated without size bias. User input directs stringency of comparisons, exhibiting different levels of resolution in complex regions of the genome.

Availability: SVCROWS can be downloaded as a Package or R-Markdown file at https://github.com/A-Crow-Nowhere/SVCROWS.git

Contact: jlg5fw@virginia.edu; njb8sg@virginia.edu

i. Introduction

Genomic structural variations (SVs) are large (>50 bp) duplications, deletions, inversions, insertions, and/or translocations[1]. SVs are evolutionarily important for organisms and causative factors in human disease[2–6]. Their abundance in all life makes them prime targets of study, which requires tools for their accurate delineation.

SV-calling applications determine SV properties using long or short-read sequencing data[7]. However, resolving SV location relies on precise sequence, mapping, and coverage across a genomic region. Such standards are hard to achieve, especially across repetitive regions like telomeric or rDNA arrays[8,9] or when using low-input methods like single-cell sequencing[10,11]. From an analysis perspective, artifactual variance occurs when comparing SV breakpoints across samples from large data sets[12] or using different SV-calling applications on the same data set[7]. These sources of SV breakpoint inconsistencies make it challenging to identify whether two SVs within a genomic neighborhood are distinct.

Applications exist to reconcile distinct SVs by creating so-called SV regions (SVRs). Construction of SVRs encompasses the variation in a genomic segment and simplifies the correlation between SVs and phenotypes[13,14]. However, SVR construction using this approach impacts quantification; SVR-calling in high variance datasets leads to under/over-simplification of SVs in a specific region. Inflation of final SVRs occurs from artificial truncation or elongation while a reduction of final SVRs arises from erroneous merging. Regardless of the impact, higher rates of genomic variance alter SVR construction and lead to false conclusions. Therefore, an approach that accounts for this variance will better accommodate the analysis of challenging genomic regions and heterogeneous datasets.

Here, we introduce a new application, 'SVCROWS' (Structural Variation Consensus with Reciprocal Overlap and Weighted Sizes). SVCROWS accounts for variation within the dataset while maintaining rigorous comparisons between SVs using a dynamic, user-defined approach. Because SVCROWS incorporates multiple lines of evidence to construct SVRs, this application is particularly useful for multi-sample datasets and parsing high-variance regions that require fine-tuned comparisons.

ii. Methods

SVCROWS is an accessible, R-based algorithm that incorporates multiple factors to delineate SV calls from datasets, produce a 'consensus list', and quantify matching SVRs (**Fig. 1A**). SVCROWS requires only a single text-based input table generated from any upstream SV-caller, which simplifies workflows to compare SV lists from different applications.

2.1 Selecting parameters to dictate stringency

SVCROWS uses the principle of reciprocal overlap (RO) thresholds to determine if SVs are placed in the same SVR (i.e. matching SVs). While the RO-based approach is used in previous applications[14–16], these rely on a single, fixed RO threshold value. This static method overlooks the disparity in SV lengths and their potential biological impacts, creating a false equivalency between small and large SVs. For instance, if we measure biological impact by the number of genes within an SV (**Fig. S1**), a larger SV is more likely to have a greater effect on the phenotype than a smaller SV. Using the static RO threshold approach of prior methods, larger and potentially more impactful SVs are more likely to be matched despite significant variance (**Fig. S1B**). Conversely, the variance in smaller, less impactful SVs prevents them from matching (**Fig. S1A**). Therefore, in regions of the genome where variance in SV length and position is high, static RO thresholds will not reflect the true diversity of SVs. To address this, the user directs SVCROWS to shift the RO threshold and other factors involved in the decision to match two SVs (i.e. stringency) according to the SV size.

Dynamic RO thresholding by SVCROWS involves 6 input parameters to define the stringency based on the analysis needs (**Fig. 1B**). Firstly, users provide size categories that group SVs based on perceived biological impact (**Fig. 1B,** blue inputs 1&2). Secondly, users dictate bounds that allow SVCROWS to match breakpoints between input SVs (**Fig. 1B,** green inputs 3&4, **Fig. S2A**). Finally, users define RO thresholds at the boundaries of the size categories (**Fig. 1B,** yellow inputs 5&6). Collectively, these inputs set the SVCROWS algorithm to respond to SV size (**Fig. 1B** red line). We provide more details on parameters and functions in the *Supplementary Methods*).

2.2 SVR consensus generation

In its nominal "Scavenge" function, SVCROWS organizes SV data into a table of distinct SVRs. The algorithm processes SVRs in descending order of length (**Fig. 1A, top**). By default, the final SVR position is determined based on the original breakpoints provided by the upstream SV-caller, with the boundaries dictated by the largest SV in the region. To be matched, two SVs must reside on the same chromosome, share the same type (e.g., "duplication"; see *Supplementary Methods*), and meet the user-defined stringency requirements (see section 2.1). When used as recommended, the stringency is higher for larger SVs than smaller SVs (**Fig 1B**). SVCROWS quantifies matching SVs on the number of RO and SV breakpoint (BP) matches at 5' and 3' ends of an SV (**Fig. 1A, bottom**).

2.3 Optional inputs and functions

In addition to size-weighted comparisons (see section 2.1), SVCROWS incorporates several features to construct and quantify SVRs. While some features mirror existing functionality, SVCROWS integrates these functions into a single, user-friendly platform and provides new functionality (* below). Below, we outline six major features, with detailed descriptions available in the *Supplementary Methods*:

i. <u>*Input parameter calculation</u>: SVCROWS calculates six input parameters based on the second and fourth quartiles of input SV sizes, standardizing comparisons across datasets from different sources.
ii. <u>*Breakpoints as a secondary SV matching factor</u>: Users can opt to incorporate breakpoint matching to aid SV matching (**Fig. S2A**). When enabled, SVCROWS reduces the required RO threshold to its minimum value (user input 5) for SVs with matching breakpoints. This feature is useful for identifying matches in repetitive genomic regions where one breakpoint is consistent, but the other is misaligned.
iii. <u>*Use of 'Known' regions in input</u>: A user can define input 'known' SVs, which drops all requirements for matching in that region to a single base pair (**Fig. S2B**). For example, by defining the region of *AMY1* in the human genome as a 'known' CNV[17], a user can remove all variation in this genomic neighborhood.
iv. <u>Expansion of SVRs to maximum size</u>: The user can employ this option to set SVR size based on the maximum and minimum matching breakpoint of all SVRs in a region (**Fig. S2C**). This functionality may be useful when downstream applications benefit from larger SVRs (e.g., gene ontology studies).
v. <u>Querying against a list of features</u>: The user can alternatively use the "Hunt" function, which takes a secondary input, a 'feature list', for comparison with an input list of SVs (or SVRs). For example, this functionality allows the user to compare a set of genes to SVRs determined from "Scavenge mode".
vi. <u>Enumeration of other variables</u>: As SVs match, the user can provide numeric information that 1) takes the number or reads originally used to call the SV and adds them together and 2) takes the quality score of each SV and averages them together.

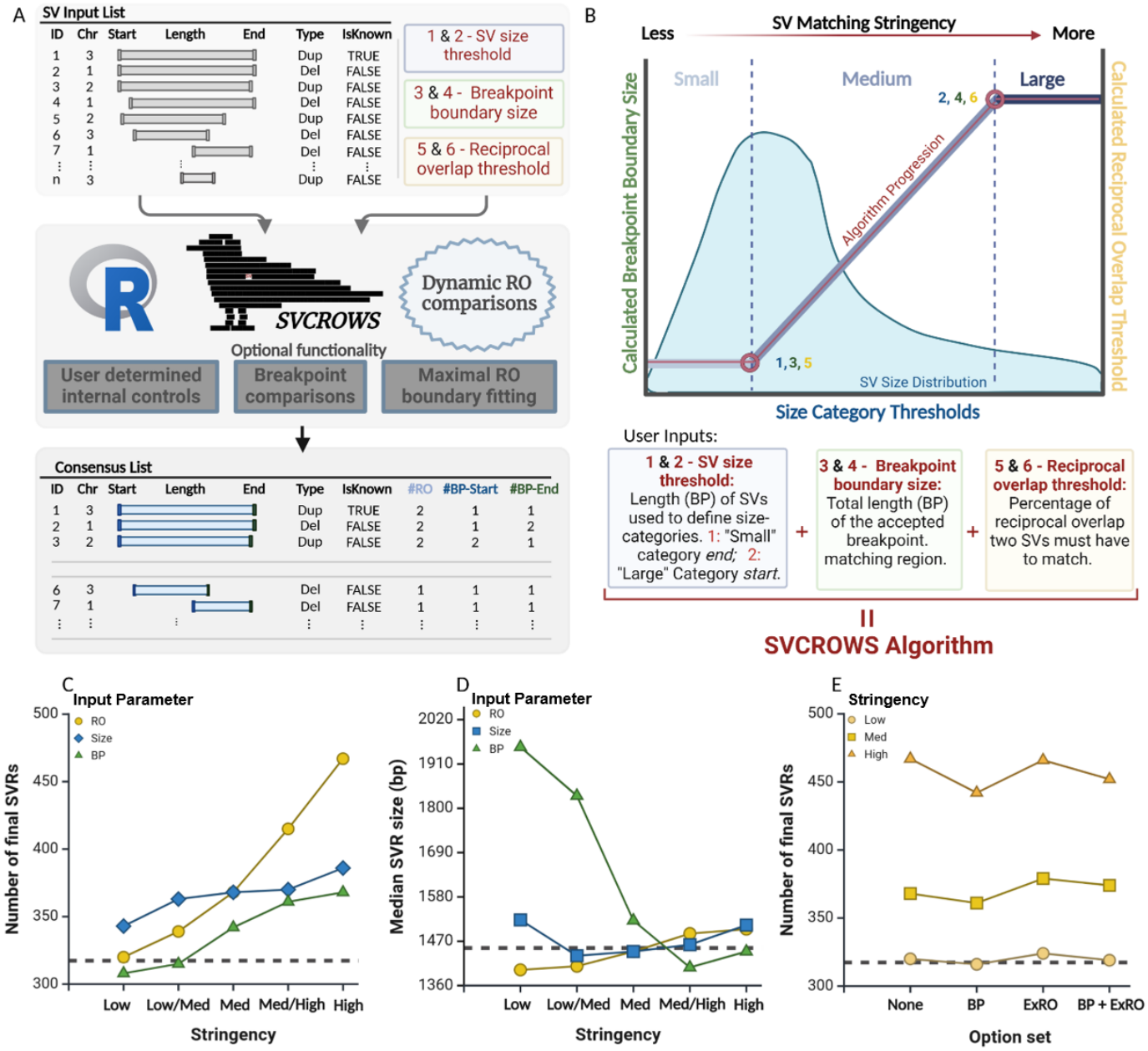

Figure 1. User input determines the stringency of SV comparisons using widely simple inputs. A) Logical workflow of SVCROWS "scavenge" mode. A list of SVRs is produced, from a single input. The number of matches is tabulated (bottom). 'Dup' = duplication, 'Del' = Deletion, '#RO' = Number of matching SVs, 'BP-Start/End' = matching breakpoints on the 3'/5' ends of the SVs. **B)** Six inputs defined by the user refine reciprocal overlap determination. **x-axis:** "Small" and "Large" SV category thresholds (variables 1&2). **Left y-axis:** The width (in base pairs) of BP-Matching region (see **fig. 2C**, variables 3&4 respectively). **Right y-axis:** The required level of RO-threshold (as a percentage) for SV matching (variables 3&4 respectively). These inputs generate the SVCROWS Calculation Trajectory (red line) which defines the dynamic RO stringencies. **C-E)** Different applications of SVCROWS on the EGAD00001009455 single-cell ovarian cancer dataset. Comparable data acquired from CNVRuler[14] is shown by the gray dashed line. **C, D)** Different options of the 6 principle inputs affect the total SVRs (**C**) and median lengths (**D**). **E)** The effect of the breakpoint factor (BP) and expanding RO regions (ExRO) has on the total number of final SVRs over 3 levels of stringency of the RO threshold. Exact run inputs are provided in **Supplementary file 1**.

iii. Results

Inputs and option-rich functions of SVCROWS provide many ways for the user to tailor the application to the goals of their investigation. We tested the application on real-world data to evaluate how specific factors drive differences in the final output.

3.1 Input dataset and testing approach

For this analysis, we utilized the EGAD00001009455 single-cell severe ovarian cancer dataset from the European Genome Archive (EGA)[18]. This dataset comprises 326 complete single-genome sequences from a single patient, characterized by high frequency SVs with variable breakpoints— a common feature in cancers[19]. We used the pre-aligned sequences as input for the Lumpy SV caller (see *Supplementary Methods*), known for its precision in breakpoint detection[20]. This process produced an input list of 4,771 CNVs from 112 cells (**Supplementary File 2**).

We assessed the three SVCROWS user input parameters using this dataset: size categories (blue lines, **Figs. 1C-D**), breakpoint (BP) overlapping region sizes (green lines, **Figs. 1C-D**), and RO threshold (yellow lines, **Figs. C-E**). We used the average size of a human exon (~1.3kb[21]) and the average size of a gene-plus-intergenic region (~38kb[22–24]) to standardize comparisons to lengths relevant to a small and large biological change respectively (see **Supplementary file 1** for details of specific run parameters).

3.2 Higher stringency levels allow higher resolution in complex regions

We observed that increasing stringency across all SVCROWS parameters leads to a rise in the total number of final SVRs (**Fig. 1C**), with RO thresholds exerting the greatest influence on stringency. This positive trend for each parameter highlights that some SVRs are more variable than others, necessitating a multifaceted approach like SVCROWS for effective delineation. Additionally, when compared to another SVR-generation application that employs a static 50%/50% RO threshold (CNVRuler[14], **Fig. 1C**), SVCROWS achieves superior SVR delineation at certain stringency levels. This enhanced SVR resolution by SVCROWS can contribute to a better appreciation of SV diversity within variable genomic regions.

We observed two distinct trends for median SVR size relative to parameter stringency (**Fig. 1D**). Increasing stringency of the 'Size' and 'RO' parameters increased median size due to stricter small SV matching. Conversely, decreasing stringency of the 'BP' parameter leads to facile large SV matching (**Supplementary file 1**), which exaggerated differences in median SVR size at lower 'BP' stringencies. This pattern indicates that larger SVs are responsible for more of the variance in this particular dataset and improves our understanding of the nature of SV variance.

SVCROWS has two optional features that directly impact SVR construction: BP matching, and expanding the RO boundary (ExRO). We noticed that higher stringency parameters were more affected by these SVCROWS features (**Fig. 1E**). While not as impactful as input stringency, BP matching helps offset false positive SVR decisions at high stringencies or when used in conjunction with ExRO. Altogether, the user can employ combinations of stringencies and optional functions to explore variance and patterns within complex SVRs.

IV. Conclusion

Genomic SVs are challenging to delineate across heterogeneous regions, multiple datasets, and disparate applications; we need flexible tools that provide a robust framework for SV comparison. SVCROWS is a size-weighted reciprocal overlap-based method designed to account for SV diversity. The application integrates multiple user-defined inputs, allowing the stringency of comparisons to be tailored to refine SVR construction. Compared to existing applications, SVCROWS offers higher resolution SVR construction, enabling detailed investigation of complex genomic regions across samples. While SVCROWS' size-weighted comparisons are unique to this class of applications, the current version of SVCROWS lacks a statistical framework for SVR generation and instead operates on the principle of SV similarity for matching. Future developments will focus on incorporating statistical analyses of results and visualization tools into SVCROWS to further enhance its utility.

V. References

1. Mahmoud, M. *et al.* Structural variant calling: the long and the short of it. *Genome Biol.* **20**, 246 (2019).
2. Flint, J. *et al.* High frequencies of alpha-thalassaemia are the result of natural selection by malaria. *Nature* **321**, 744–750 (1986).
3. Gonzalez, E. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**, 1434–1440 (2005).
4. Lukácová, M., Barák, I. & Kazár, J. Role of structural variations of polysaccharide antigens in the pathogenicity of Gram-negative bacteria. *Clin. Microbiol. Infect. Off. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis.* **14**, 200–206 (2008).
5. Marshall, C. R. *et al.* Structural Variation of Chromosomes in Autism Spectrum Disorder. *Am. J. Hum. Genet.* **82**, 477 (2008).
6. Songsomboon, K. *et al.* Genomic patterns of structural variation among diverse genotypes of Sorghum bicolor and a potential role for deletions in local adaptation. *G3 GenesGenomesGenetics* **11**, jkab154 (2021).
7. Joe, S. *et al.* Comparison of structural variant callers for massive whole-genome sequence data. *BMC Genomics* **25**, 318 (2024).
8. Levy-Sakin, M. *et al.* Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nat. Commun.* **10**, 1025 (2019).
9. Rothschild, D. *et al.* Diversity of ribosomes at the level of rRNA variation associated with human health and disease. *bioRxiv* 2023.01.30.526360 (2024) doi:10.1101/2023.01.30.526360.
10. Mallory, X. F., Edrisi, M., Navin, N. & Nakhleh, L. Methods for copy number aberration detection from single-cell DNA-sequencing data. *Genome Biol.* **21**, 1–22 (2020).
11. Ning, L. *et al.* Current Challenges in the Bioinformatics of Single Cell Genomics. *Front. Oncol.* **4**, 7 (2014).
12. Conrad, D. F. & Hurles, M. E. The population genetics of structural variation. *Nat. Genet.* **39**, S30–S36 (2007).
13. da Silva, V. *et al.* CNVRanger: association analysis of CNVs with gene expression and quantitative phenotypes. *Bioinformatics* **36**, 972–973 (2020).
14. CNVRuler: a copy number variation-based case–control association analysis tool | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/28/13/1790/234696.
15. Handsaker, R. E. *et al.* Large multiallelic copy number variations in humans. *Nat. Genet.* **47**, 296–303 (2015).

16. Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**, 269–276 (2011).

17. Venkatapoorna, C. M. K. *et al.* Association of Salivary Amylase (AMY1) Gene Copy Number with Obesity in Alabama Elementary School Children. *Nutrients* **11**, 1379 (2019).

18. Freeberg, M. A. *et al.* The European Genome-phenome Archive in 2021. *Nucleic Acids Res.* **50**, D980–D987 (2022).

19. T, F. *et al.* Single-cell genomic variation induced by mutational processes in cancer. - Abstract - Europe PMC.

20. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).

21. Human molecular genetics - NLM Catalog - NCBI. https://www.ncbi.nlm.nih.gov/nlmcatalog/101523906.

22. Shabalina, S. A., Ogurtsov, A. Y., Kondrashov, V. A. & Kondrashov, A. S. Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet.* **17**, 373–376 (2001).

23. Kołomański, M., Szyda, J., Frąszczak, M. & Mielczarek, M. DNA sequence features underlying large-scale duplications and deletions in human. *J. Appl. Genet.* **63**, 527–533 (2022).

24. Rajic, Z. A. *et al.* Size of the protein-coding genome and rate of molecular evolution. *J. Hum. Genet.* **50**, 217–229 (2005).