VISTA: An integrated framework for structural variant discovery

SEUNGMO LEE[1*], Varuni Sarwal[1*], Jianzhi Yang[2], Sriram Sankararaman[1], Mark Chaisson[2], Eleazar Eskin[1], Serghei Mangul[2,3]

[1] Department of Computer Science, University of California Los Angeles, 580 Portola Plaza, Los Angeles, CA 90095, USA
[2] Department of Quantitative and Computational Biology, Dana and David Dornsife College of Letters, Arts and Sciences University of Southern California, Los Angeles, California, 90089, United States
[3] Department of Clinical Pharmacy, Alfred E. Mann School of Pharmacy, University of Southern California, 1540 Alcazar Street, Los Angeles, CA 90033, USA

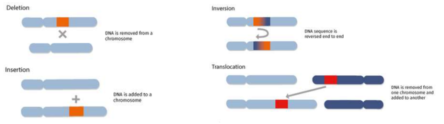
* Denotes equal contribution

Abstract

We report an integrated structural variant calling framework, VISTA (Variant Identification and Structural Variant Analysis) that leverages the results of individual callers using a novel and robust filtering and merging algorithm. In contrast to existing consensus based tools which ignore the length and coverage, VISTA overcomes this limitation by executing various combinations of top-performing callers based on variant length and genomic coverage to generate SV events with high accuracy. We evaluated the performance of VISTA on using comprehensive gold standard datasets across varying organisms and coverage. We benchmarked VISTA using the Genome-in-a-Bottle (GIAB) gold standard SV set, haplotype-resolved de novo assemblies from The Human Pangenome Reference Consortium (HPRC), along with an in-house PCR-validated mouse gold standard set. VISTA maintained the highest F1 score among top consensus based tools measured using a comprehensive gold standard across both mouse and human genomes. VISTA also has an optimized mode, where the calls can be optimized for precision or recall. VISTA-optimized is able to attain 100% precision and the highest sensitivity among other SV callers. In conclusion, VISTA represents a significant advancement in structural variant calling, offering a robust and accurate framework that outperforms existing consensus-based tools and sets a new standard for SV detection in genomic research.

Background

Structural variants (SVs) are genomic regions that contain an altered DNA sequence due to deletion, duplication, insertion, inversions and other complex rearrangements. SVs are present in approximately 1.5% of the human genome, but this small subset of genetic variation has been implicated in the pathogenesis of psoriasis, Crohn's disease and other autoimmune disorders, autism spectrum and other neurodevelopmental disorders, and schizophrenia. With advances in whole genome sequencing (WGS) technologies, a plethora of SV detection methods have been developed. However, dissecting SVs from WGS data remains a challenge, with the majority of SV detection methods prone to a high false-positive rate, and no existing method able to precisely detect a full range of SV's present in a sample.


Deletion
DNA is removed from a chromosome
Inversion
DNA sequence is reversed end to end
Insertion
DNA is added to a chromosome
Translocation
DNA is removed from one chromosome and added to another

Methods

- We used public benchmark data for the Ashkenazi Jewish Trio son (NA24385/HG002) from the Genome-in-a-Bottle (GIAB) consortium.
- The average depth of coverage was 45x and the reads were 2x250 bp paired-end reads.
- We used the Genome-in-a-Bottle preliminary variant set containing deletions in HG002 as our gold standard
- In order to demonstrate the scalability of VISTA across a large number of samples and gold standard sets, we tested VISTA on the following samples of human data: HG00733, HG00438, HG00621, HG00735, HG00741, HG01071, HG01106, HG01109, HG01243, HG01175.
- We used dipcall[31] to generate SVs of the human samples directly from haplotype-resolved de novo assemblies produced by The Human Pangenome Reference Consortium (HPRC).
- In order to demonstrate VISTA's generalizability across organisms, we used a set of homozygous deletions present in inbred mouse chromosomes.

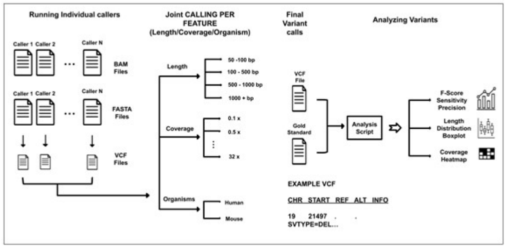


Figure 1: Overview of the approach implemented in VISTA.

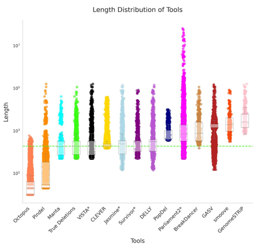
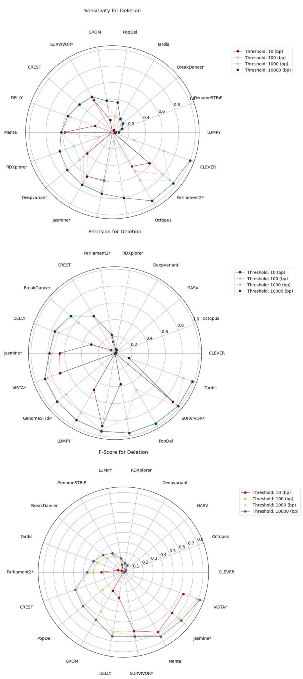


Figure 2. Length distribution and medians of deletions detected by VISTA and popular SV caller for human data. True deletions are indicated in green. The horizontal dashed line corresponds to the median value of true deletions. The asterisk represents consensus based callers.

Figure 3: Comparing the performance of VISTA with popular SV callers on human data sample HG002. A deletion is considered to be correctly predicted if the distance of right and left coordinates are within the threshold τ from the coordinates of a true deletion. Sensitivity, Precision
& F-Score of SV callers at different thresholds.



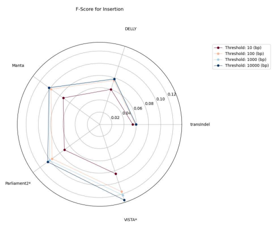


Figure 4: Comparing the performance of VISTA with popular SV callers on human data. An insertion is considered to be correctly predicted if the distance of right and left coordinates are within the threshold τ from the coordinates of a true insertion. F-score of SV callers at different thresholds. The asterisk represents consensus-based callers.

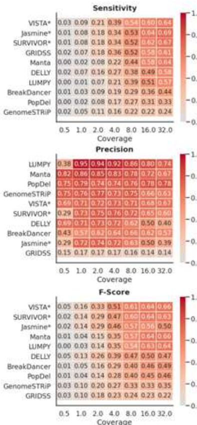


Figure 5: Performance of VISTA on low and ultra-low coverage mouse data. Heatmap depicting the sensitivity, precision, and F-Score based on 100 bp threshold across various levels of coverage

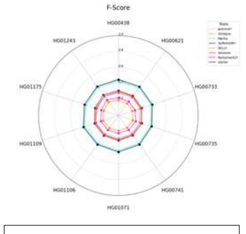


Figure 6: Comparing the performance of VISTA with popular SV callers on different human genomes using the dipcall gold standard set.

VISTA Algorithm

- VISTA takes the output of individual callers as inputs, and produces a highly accurate file predicting variants, in vcf format.
- Based on its organism type, coverage and length, VISTA bins the input vcfs into different categories. The comprehensive gold standard is used to evaluate metrics such as the sensitivity, precision and F-score for each bin. VISTA then uses a consensus-based approach, and decides the top performing caller for each bin.
- VISTA merges the outputs of each of the top performing callers into one output vcf file.
- For VISTA's discovery mode, where the ground truth is not know, we use the combination of callers identified during VISTA's pretraining, on the organism closest to the input sample: Octopus, Manta, DELLY, and GenomeSTRiP

Discussion

- We assessed the performance of VISTA in terms of precision (false discovery rate), recall (true-positive rate) based on the Genome in a Bottle (GIAB) v0.6 SV candidate truth set
- We compared the performance of VISTA with 16 individual SV callers and 3 consensus-based SV callers in terms of inferring deletions on mouse chr19 across seven mouse strains and HG002 Human sample. VISTA was the closest in terms of the length distribution of deletions as compared to the MusMusD gold standard
- **MusMusD**: VISTA (68%) has the highest F1 score (i.e., harmonic mean of precision and recall) for thresholds 100bp and above, followed by Manta[4](65.0%) and LUMPY[20] (64.8%)
- **HG002**: VISTA has achieved the highest recall at 100bp (72.4%) while having the fourth-highest precision (85.7%).
- **HG002**: Importantly, VISTA achieves the highest F1 score (i.e., harmonic mean of precision and recall) (78.5%) for thresholds 100bp and above, followed by Manta(73.8%) and Jasmine (72.4%).
- We also studied the robustness of VISTA across the 10 human samples. We observed all the callers to have a consistent trend across the samples, with a slightly elevated precision for samples HG00735 and HG01234. VISTA was able to consistently achieve the highest F-score across all samples, closely followed by Manta[4].
- VISTA was able to obtain the highest F-score consistently across all coverages from 0.5x to 32x, closely followed by Delly for 0.5x, SURVIVOR[22] and Jasmime[23] for 1x-8x.
- We have provided an "optimize mode" in VISTA, where the user can choose to optimize for either precision or recall
- As expected, the performance for insertions was reduced for insertions as compared to deletions given the limitations of short-read–based insertion detection algorithms.

Acknowledgments

Mangul Lab
University of Southern California
UCLA Computer Science
UCLA ZARLAB