npsv: A Simulation-Based Structural Variant Genotyping Tool

William Kelley ‘21, Musab Shakeel ‘21
Dept. of Computer Science

Advisor:
Michael Linderman

Abstract

Variants in the human genome range from single nucleotide variants to large structural variants (SVs) that can span 100s to millions of nucleotides. SVs, which have been associated with numerous genetic diseases, are challenging to discover and genotype in short-read next-generation sequencing (NGS) data. SVs are larger than the sequencer read length and so must be indirectly inferred from secondary features in the sequencing data, such as split reads, discordant read-pairs, and read depth (coverage). In this project, we are developing a software tool, npsv, to perform end-to-end genotyping (predict the zygosity) of deletion SVs in NGS data using a non-parametric, simulation-based approach that accounts for sample-, variant-, and pipeline-specific biases. We simulate NGS data for the putative SVs (using the same alignment pipeline as in the sample) and build sample- and/or variant-specific classifiers trained on features extracted from this simulated data. These classifiers are then applied to genotype putative deletions in the sample. We evaluate npsv using the Genome in a Bottle HG002 reference sample and tier 1 SV callset (14588 deletions).

Background

- “Genotyping” predicts variant zygosity, or the number of alternate alleles present, i.e.:
  - Homozygous reference, or 0 copies (0/0)
  - Heterozygous, or 1 copy (0/1)
  - Homozygous alternate, or 2 copies (1/1)
- SVs play a causal role in numerous diseases
- Improved genotyping accuracy will improve our understanding of disease and increase molecular diagnosis rate
- Existing tools use parametric models that do not factor in sample-, variant-, and/or pipeline-specific biases
- npsv aims to improve the genotyping accuracy by using a simulation-based approach to build per-sample and per-variant classifiers for SVs

References

[1] Alkan, C.; Coe, B. P.; Eichler, E. E. Genome Structural Variation Discovery and Genotyping. Nature Reviews Genetics 2011, 12, 363–376.
[2] Zook JM, et al.: A robust benchmark for germline structural variant detection. *bioRxiv* 2019:664623.

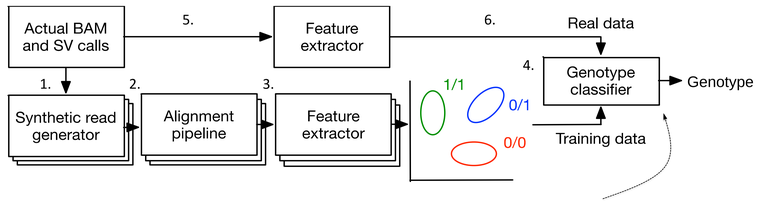
Acknowledgements

Research reported in this poster was supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103449. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NIGMS or NIH.

Methods

Workflow

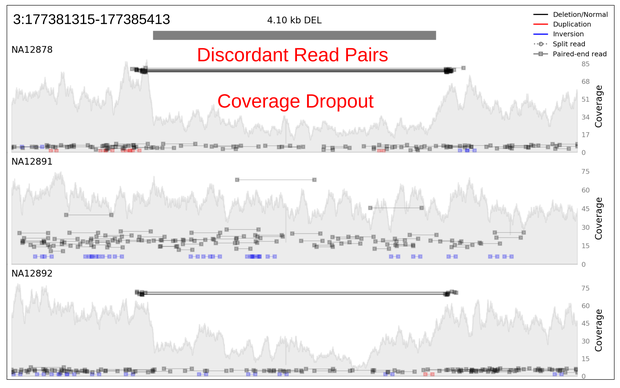
1. For each putative SV, for each of the three zygosities, generate synthetic reads or sample random SV
2. Align synthetic reads to reference genome using the same alignment pipeline used for the actual sample reads
3. Extract features from aligned synthetic reads
4. Train a classifier on simulated features
5. Extract features from aligned actual reads
6. Predict the genotype of the actual sample using the classifier and real features



- Single model: A single classifier trained on all variants
- Per-variant model: Individual classifiers for each variant trained on 300 replicates per variant

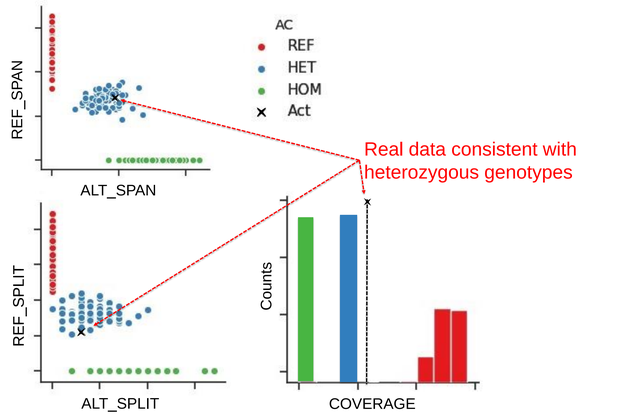
Pile-up data for an example variant

Region around an inherited heterozygous (0/1) 4.1kb deletion SV in NA12878 (child) and NA12891/2 (parents). Variant was incorrectly reported as *de novo* by existing tool.



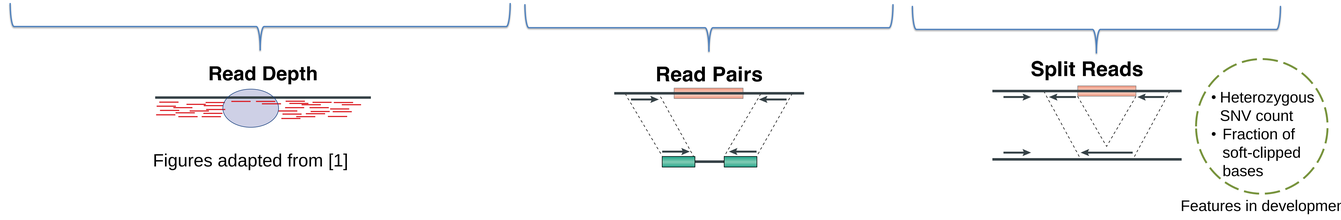
Classification intuition

Plot of absolute features for actual data (black) and 100 iterations of sampled homozygous reference genotypes (red), and simulated heterozygous (blue) and homozygous alternate (green) genotypes.



SV Features

Coverage	DHFC	DHBFC	DHFFC	RefAlt Span	Fraction of Ref/Alt Span Reads	Insert Upper/Lower	RefAlt Split	Fraction of Ref/Alt Split Reads	Binomial Probabilities of Genotype
Mean number of reads overlapping positions in the event	Coverage relative to chromosome	Coverage relative to other regions with similar GC content	Coverage relative to the flanks of the SV	P_{insert}-weighted counts of ref. and alt. spanning reads	Balance of ref/alt spanning reads	Fraction of spanning reads with Z-scores < -1.5 and > 1.5	Count of reads aligned to ref. and alt. alleles by Paragraph aligner	Balance of ref/alt reads	Statistical likelihood of each genotype given read counts



Figures adapted from [1]

Features in development

Results

Evaluation

We evaluated npsv with 14588 deletions from the Genome in a Bottle HG002 deletion SV callset (> 50bp)[2].

	Test 0/0	Test 0/1	Test 1/1
True 0/0	1997	305	26
True 0/1	433	2119	61
True 1/1	189	196	1264

Concordance = ∑green / all

Non-reference Concordance = (∑green + ∑blue) / all

Single model npsv concordance results

	Only high-quality 0/1 and 1/1 variants		High-quality 0/0, 0/1 and 1/1 variants	
Features	**Concordance**	**Non-reference Concordance**	**Concordance**	**Non-reference Concordance**
Absolute features	70.0%	86.7%	73.2%	84.1%
All features	77.3%	90.6%	77.4%	86.4%
Relative features only (with SV length)	**83.0%**	**91.2%**	**80.9%**	**86.2%**

Single model npsv compared with other tools

	Only high-quality 0/1 and 1/1 variants		High-quality 0/0, 0/1 and 1/1 variants	
Tool	**Concordance**	**Non-reference Concordance**	**Concordance**	**Non-reference Concordance**
SVTyper	49.9%	56.1%		
SV2	44.9%	53.6%		
Paragraph	76.8%	81.8%		
svviz2 (count)	87.7%	94.1%	84.7%	88.7%
NPSV (relative features)	**83.0%**	**91.2%**	**80.9%**	**86.2%**

Discussion

- Simulating the 0/0 genotypes (instead of sampling) improved accuracy for variant-only test set, but reduced accuracy when 0/0 calls were included
- Concordance increases with variant size; smaller variants are challenging because the read-pair evidence becomes inconclusive
- 85% of discordant calls overlap Variable Number Tandem Repeats (VNTRs). These repetitive sequences appear to reduce the sensitivity of the graph aligner used to detect split reads. As a solution, we are testing other aligners.
- Some VNTR calls appear to be offset from true variant

Ongoing and Future Work

- Replace graph aligner read counts with svviz2 counts and re-evaluate single model npsv
- Add additional features such as counts of heterozygous SNVs in event and fraction of soft-clipped bases to further improve genotyping accuracy
- Use similarity of real and simulated data to identify better (best) variant representation in repetitive regions
- Compare single model and per-variant approaches: Is the additional time cost for the per-variant model worth it?
- Update BAM file processing method to make simulated reference and alternate contigs the same length
- Detailed review of putative Mendelian violations