VarSim: A high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications

John C. Mu[1]*, Marghoob Mohiyuddin[2]*, Jian Li[2], Narges Bani Asadi[2], Mark B. Gerstein[3], Alexej Abyzov[4], Wing H. Wong[5,6], and Hugo Y.K. Lam [2]†

Motivation

- Lack of comprehensive simulation validation framework
- Multiple validation datasets critical for development of new secondary analysis methods
- Read simulation is a bottleneck when simulating high coverage
- Genome in a bottle consortium [1] generated a gold standard set of variants by combining multiple sequencing technologies. However, this is not scalable for generation of multiple datasets.
- SMASH [2] provides a simulation validation framework but only supports insertions and deletions
- RSVsim [3] simulates SVs but not small variants and does not allow validation of read alignments

Avaliability and Implementation

- Code implemented in Java and Python
- Source code available for download at **http://github.com/bioinform/varsim**
- Reads and variants for pre-constructed synthetic genomes are available at SRA

Our Contributions

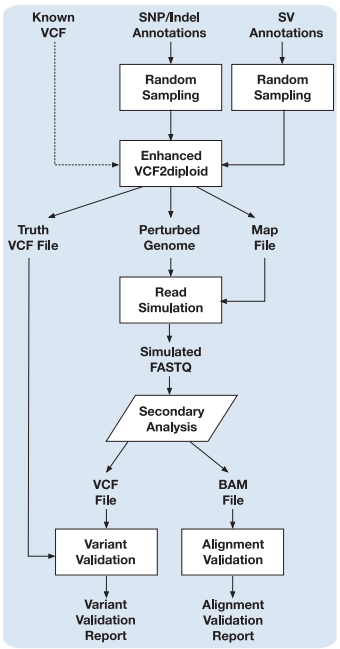
- Framework for validating both read alignment and variant calling accuracy through simulation
- Generates a map of reference genome to enable evaluation of reads overlapping structural variation
- Simulates diploid genome with both structural variations and small variants from a pre-defined set or databases
- Ability to support arbitrary read simulators (different platforms, exome, targeted)
- Parallel execution of read simulators and parallel compression of reads for high performance
- Somatic variant validation workflow
- Detailed and interactive analysis output as HTML page

Conclusions and Future Work

- Simulation is an important validation methodology as it allows for easy generation of multiple validation datasets and the availability of ground truth facilitates evalution
- VarSim provides a comprehensive simulation validation framework
- Support for translocations and interspersed duplications will be added

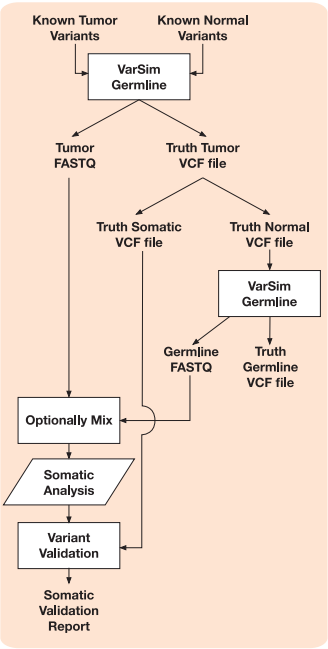
VarSim Germline Workflow

VarSim takes databases of known variants as input and randomly samples a specified number of variants. An enhanced version of vcf2diploid [4] is used to generate a diploid genome containing the specified variants, a VCF file with the true variants and a Map file that describes the structure relative to the reference genome. After alignment and variant calling, VarSim generates HTML reports showing the accuracy of the results.



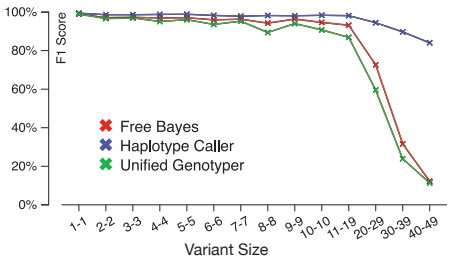
VarSim Somatic Workflow

VarSim is run twice to generate reads from both normal and tumor genomes. SVs are included in the normal variants and can be included in the tumor variants if available. The resulting reads can be mixed before input to a somatic analysis pipeline. An HTML accuracy report is generated based on the somatic variants called.

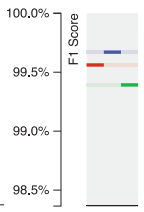


Variant Calling Validation Results

Small Deletions (20 bp wiggle, 80% reciprocal overlap)

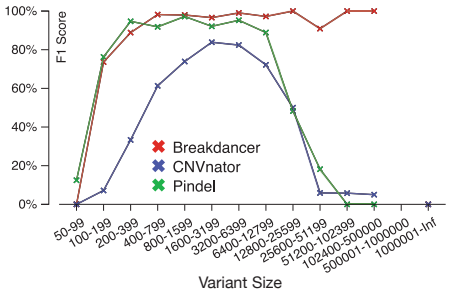


SNVs (Exact match)



We compared the accuracy of several variant callers on variants from NA12878 GiaB high confidence regions. For all variant calling comparisons alignments from BWA-MEM after re-alignment and re-calibration with GATK was used. We simulated a total of 2,799,440 SNVs, 366,247 indels and 3,531 SVs

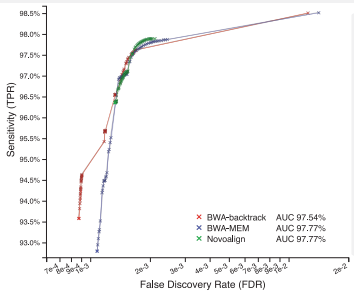
SV Deletions (100 bp wiggle, 50% reciprocal overlap)



HaplotypeCaller performs very well and was superior to both UnifiedGenotyper and FreeBayes, especially for larger deletions. However, we note that all callers suffered a loss in accuracy for indels greater than 10 bp. The results show a similar pattern for insertions. All tools perform similarly for SNVs.

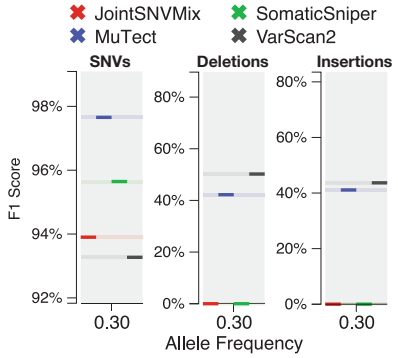
All tools performed well for moderate-sized deletion SVs. Only BreakDancer was able to recover larger deletion SVs. However, it was not able to recover exact breakpoints. All tools failed to adequately recover deletion SVs in the smaller range. Note that each tool has a specific range where it performs well, this suggests that a merged meta-calling approach is appropriate.

Alignment Validation Results



VarSim was used to compare the alignment accuracy of BWA-backtrack, BWA-MEM and Novoalign before realignment. Overall they all performed very well on the 100 bp paired-end reads. Novoalign and BWA-MEM were slightly more accurate compared to BWA-backtrack in terms of area under the curve. However, BWA-backtrack is able to achieve a lower error floor. VarSim is capable of outputting a similar plot for reads overlapping each type of variant.

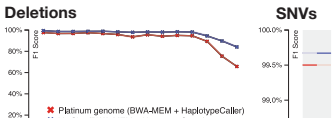
Somatic Analysis Validation Results



We analyzed a simulated tumor genome with VarSim. This genome was constructed by adding somatic variants from the COSMIC database to NA12878. The somatic variant callers MuTect, VarScan2, JointSNVMix and Somatic Sniper were compared based on both somatic SNV and indel calling accuracy. We used a pure normal sample and a tumor sample with 0.3 somatic allele frequency for this analysis.

Overall, MuTect was superior to the other tools for SNV calling. Only MuTect and VarScan2 were able to call somatic indels. At 0.3 allele frequency, the F1 score for insertions was 0.41 for MuTect and 0.43 for VarScan2. For deletions the F1 score was 0.42 for MuTect and 0.50 for VarScan2. If we consider the details, VarScan2 had a higher sensitivity at the cost of lower precision, while MuTect had lower sensitivity and higher precision.

Real Data Comparison



We compared variants called from Illumina platinum genome reads to the the calls from VarSim simulated reads. Overall, the F1 scores were close. We found that the differences in insertions and deletions were mostly due to limitations in the read simulator. In particular, ART does not account for the low quality bases typically found around homopolymers for Illumina reads.

Bibliography

[1] J. M. Zook, et, al., Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls, Nat. Biotechnol., 2014.
[2] A. Talwalkar et. al., SMaSH: A Benchmarking Toolkit for Human Genome Variant Calling, arXiv, 2014
[3] C. Bartenhagen and M. Dugas, RSVSim: an R/Bioconductor package for the simulation of structural variations Bioinformatics, 2013
[4] J. Rozowsky, et.al., AlleleSeq: analysis of allele-specific expression and binding in a network framework, Molecular Systems Biology, 2011.

Affiliations
1. Department of Electrical Engineering, Stanford University, Stanford, CA 94305.
2. Department of Bioinformatics, Bina Technologies, Redwood City, CA 94065.
3. Program in Computation Biology and Bioinformatics, Yale University, New Haven, CT 06511
4. Mayo Clinics, Department of Health Sciences Research, Rochester, MN 55905.
5. Department of Statistics, Stanford University, Stanford, CA 94305.
6. Department of Health Research and Policy, Stanford University, Stanford, CA 94305.
* The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
† To whom correspondence should be addressed.

www.bina.com

Contact us
rd@bina.com