Evaluation of INDEL Callers for Next-Generation DNA Sequencing Data

Rachel L. Goldfeder, BS[1], and Euan A. Ashley MRCP, DPhil[1]
[1]Stanford University, Stanford, CA

Abstract

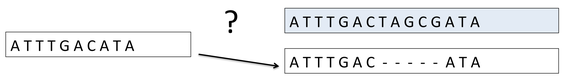
Small insertions and deletions (INDELs) in the human genome play a significant role in disease and genetic variation. In both research and clinical settings, the ability to detect INDELs is critical for disease understanding, diagnosis, and treatment. However, detecting INDELs from next-generation DNA sequencing data is still a major challenge. In order to better understand the current state of the art, we evaluated five commonly used INDEL-calling pipelines: Pindel, Dindel, SOAPindel, UnifiedGenotyper and HaplotypeCaller. To evaluate each pipeline, we computationally generated a genome that contains INDELs of varying sizes. We created paired-end artificial sequencing reads for this genome, incorporating an error profile similar to that of a HiSeq2000 using ART. Finally, we processed the reads with each of the five pipelines using "best practice" parameters and filters and calculated recall and precision. We found greater variability in recall than precision across the pipelines. We also evaluated how robust each pipeline was to changes in aligner, read length, coverage, and variant type (homozygous vs heterozygous, with SNPs present or absent). In general, the pipelines perform better with Novoalign, longer reads, higher coverage, and homozygous INDELs. Our results suggest that longer reads and high coverage are necessary for clinical-grade INDEL detection.

Background

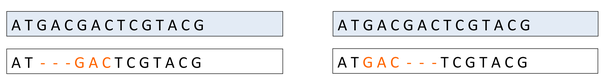
Challenges in INDEL calling:

1. Chemistry

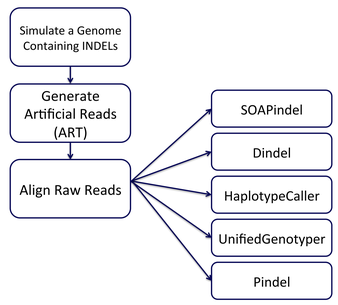
2. Alignment to the Reference Genome

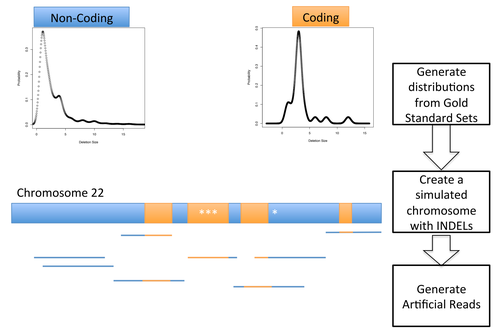


3. Identification of INDEL Location

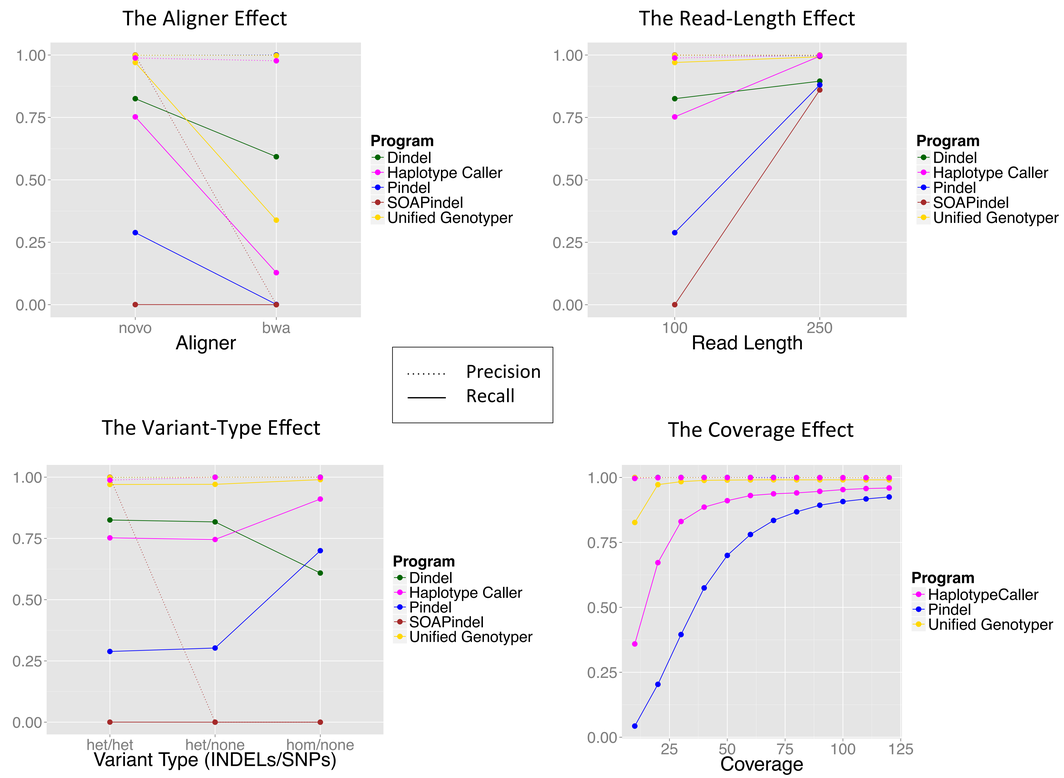


Methods





Results



Conclusions

- Most programs call few false positives; recall is more variable across the programs than precision.
- For most programs, aligning reads with Novoalign rather than BWA provides better recall.
- Longer reads provide much better recall.
- Most programs perform better when detecting homozygous INDELs than heterozygous INDELs.
- Higher coverage provides better recall.

Future Directions

- Determine the similarity of the call sets
- Characterize features of INDELs detected by all 5 programs or none of the programs
- Examine the concordance of genotypes
- Study the effect of varying insert size between paired end reads
- Perform these analyses on multiple simulated genomes

Acknowledgements

This research is funded in part by the NLM Informatics Training Grant and National Science Foundation Graduate Research Fellowship. Thanks to the entire Ashley lab for support, specially Pablo Cordero for managing the lab's computing resources.