Scotch: A Novel Method to Detect Insertions and Deletions from Next-Generation DNA Sequencing Data

Rachel L. Goldfeder[1,2] and Euan A. Ashley[2]

[1] Biomedical Informatics Training Program, Stanford, CA 94305

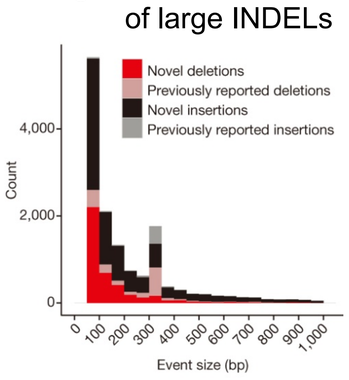
[2] Division of Cardiovascular Medicine, Department of Medicine, Stanford, CA 94305

Abstract

Clinical-grade genome sequencing and interpretation requires accurate and complete genotype calls for all interrogated positions. While single nucleotide variant detection is highly accurate and consistent, these variants explain only a small fraction of disease risk. Other types of variation that disrupt the open reading frame, such as insertions and deletions (INDELs), have systematically been shown to have dramatic effects on phenotype. However, current methods have low sensitivity for larger INDELs (>= five bases), primarily due to challenges surrounding aligning sequence reads that span complex loci. We present Scotch, a novel INDEL detection method that leverages signatures of poor read alignment, through machine learning approaches, to accurately identify INDELs from next-generation DNA sequencing data. Using biologically realistic simulated genomes and sequence reads with technologically representative error profiles (generated by ART), we evaluate Scotch and several currently available INDEL callers. We show that Scotch outperforms current methods, particularly for larger INDELs. This method will enable researchers and clinicians to more accurately identify larger INDELs, which will in turn improve patient care and our understanding of human traits and diseases.

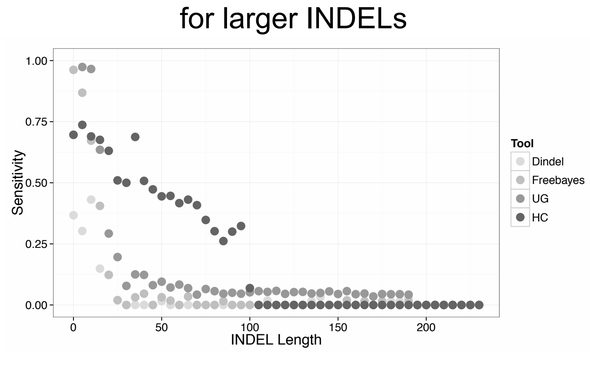
Background

Our genomes contain thousands of large INDELs



MJP Chaisson *et al. Nature* **000**, 1-4 (2014) doi:10.1038/nature13907

Current tools have low sensitivity for larger INDELs



RL Goldfeder *et al.* (In preparation)

Scotch Methodology

Malt

1. Extract Features from alignment

Position specific
- Depth of coverage
- Base Quality
- Mapping Quality

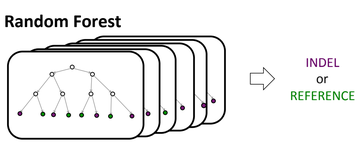
Alignment
- Insertions
- Deletions
- Soft Clipping

Genomic Region
- GC content
- Repeats
- Sequence uniqueness

Brew

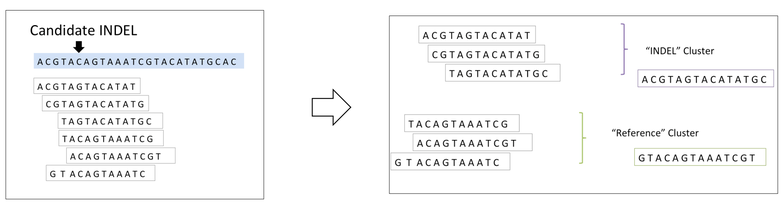
2. Use machine learning approaches to find candidate INDEL locations

	Feature 1	...	Feature N
Position 1	40	...	0.5
Position 2	20	...	1
...	...		
Position N			



Distill

3. Cluster reads to determine haplotypes



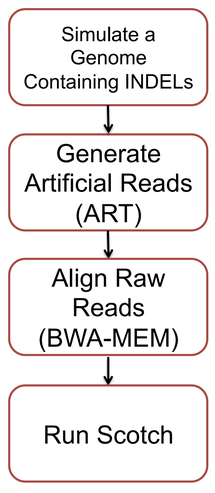
Age

4. Determine consensus sequence to identify INDELs

ACGTACAGTAAATCGTACATATGCAC

ACGTA - - - - - - - - - GTACATATGC

Evaluation with Simulated Data



Dataset:

50x coverage, 100bp paired-end reads

41,000 INDELs

Sizes uniformly distributed 1-200bp

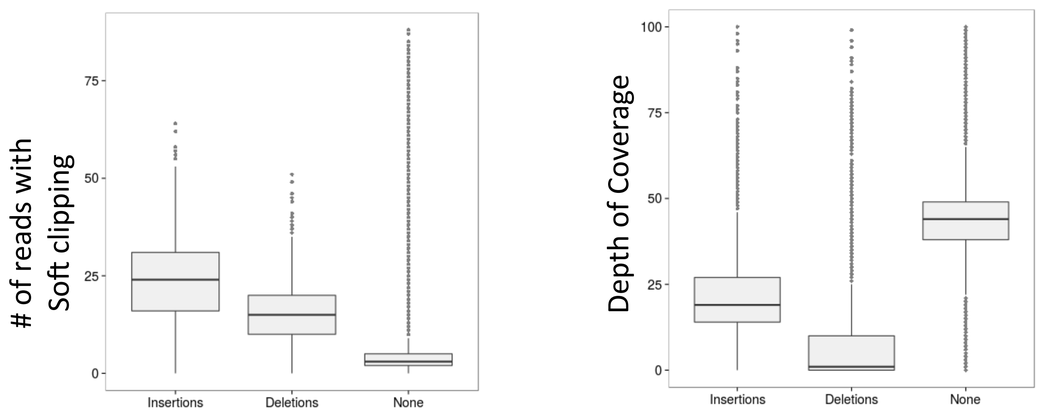
41,000 non-INDELs

Results:

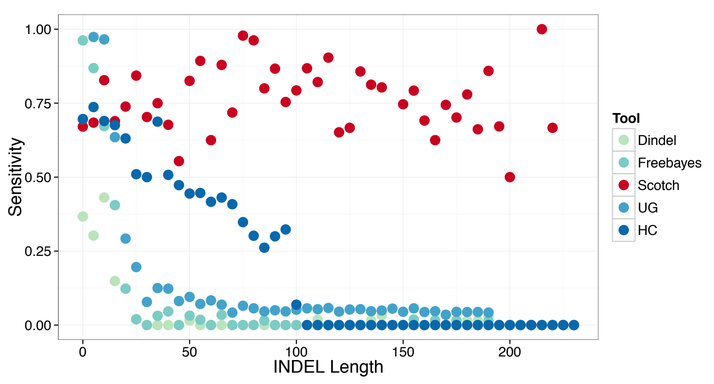
Scotch has high accuracy for identifying INDEL coordinates (5 fold CV)

Accuracy	Value
Training Set Accuracy	99.96%
Test Set Accuracy	99.35%

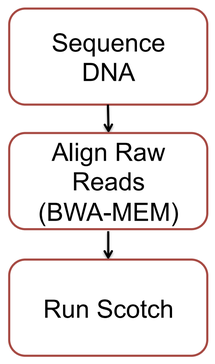
Important Features



Scotch has high sensitivity for larger INDELs



Evaluation with Real Data



Dataset:

NA12878

50x coverage

150bp paired-end reads

Results:

Scotch identified 1,434 INDELs on chr16

We randomly selected 10 for Sanger Sequencing

10 out of 10 putative INDELs confirmed

None of these were present in the GIAB dataset

Conclusions & Future Directions

Conclusions
- Scotch has higher sensitivity for larger INDELs (simulated data) and high PPV for real data
- Depth of coverage and soft clipping are important features in INDEL detection

Future Directions:
- Benchmark Scotch in “difficult to analyze” regions of the genome
- Apply to clinical sample to identify medically relevant large INDELs

Acknowledgments

This work was supported by a National Science Foundation Graduate Research Fellowship, CEHG Fellowship, CERSI Scholarship, and NLM Training Grant T15 LM7033. We thank Daryl Waggott and the rest of the Ashley lab for useful discussions.