Detecting SVs in Cancer Genomes Direct Cancer/Somatic Comparison

Cody J. Weinberger[1,*] and Wen-Hsiung Li, PhD[2]

[1] The Department of Ecology and Evolution, University of Chicago, 1101 E 57th Street, Chicago, IL 60615, USA; email: cweinberger@uchicago.edu

[2] Biodiversity Research Center, Academia Sinica, 128 Academia Road Sec. 2, Nankang Taipei 115, Taiwan

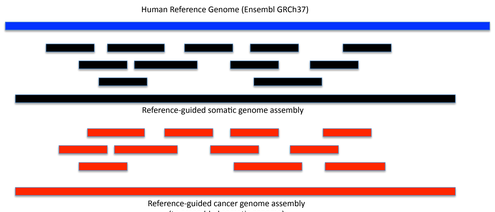
Introduction

As next generation sequencing technology (NGS) increasingly allows for cheap sequencing of entire genomes, the challenge of obtaining accurate sequences is largely computational. While it has become routine to detect single nucleotide polymorphisms, detecting large scale structural variants (SVs) is a much more complex issue primarily due to short read length. A promising avenue of SV detection is using soft-clipped reads, which allows for single nucleotide resolution of potential breakpoints. When using NGS to detect somatic mutations, the traditional approach has been to compare both somatic and cancer soft-clipped reads to a reference and subtract the somatic breakpoints from the cancer breakpoints to obtain those that have occurred only in somatic cells. However, comparing the cancer and somatic genomes indirectly prevents detection of some types of mutations, e.g. any mutation nested within a novel insertion. There is also a large bias towards detecting deletions over insertions using this method, as novel regions of soft-clipped reads cannot map to the reference. A method for comparing cancer and somatic genomes directly is presented here to avoid these issues.

Method

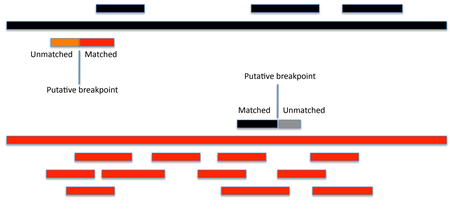
Step 1

- Reference guided assembly of somatic reads to human reference genome (using SHEAR)
- Reference guided assembly of cancer reads to resulting assembled somatic genome



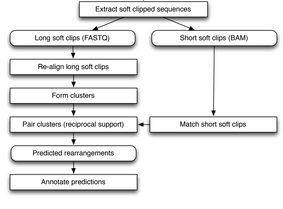
Step 2

- Build alignments of reads to opposing genome



Step 3

- Run Socrates on resulting SAMs to output putative breakpoints

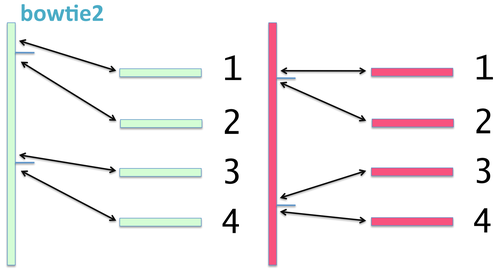


Socrates

Schroeder and Hsu *et. al.* 2014

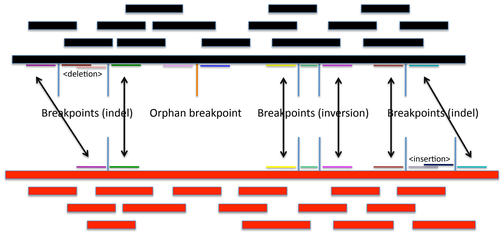
Step 4

- Output flanking bases to FASTA from each set of breakpoints (using java) and index with bowtie2



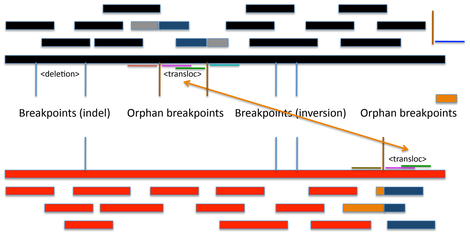
Step 5

- Crossmap flanking sequences (short fasta sequences) to look for matches



Step 6

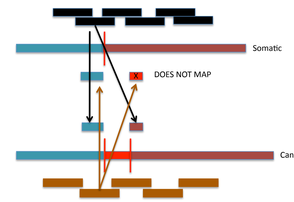
- Sort and output matched and unmatched breakpoints (using java)



Advantages

1) Detect More Novel Insertions

- Socrates only detected 65% of novel insertions
  - PRISM detected 97%
  - Pindel detected 93%
- CREST, another soft-clip cancer genome tool, has same difficulty with novel insertions
- By mapping onto both genomes, novel insertions will also be detected as a deletion in one genome

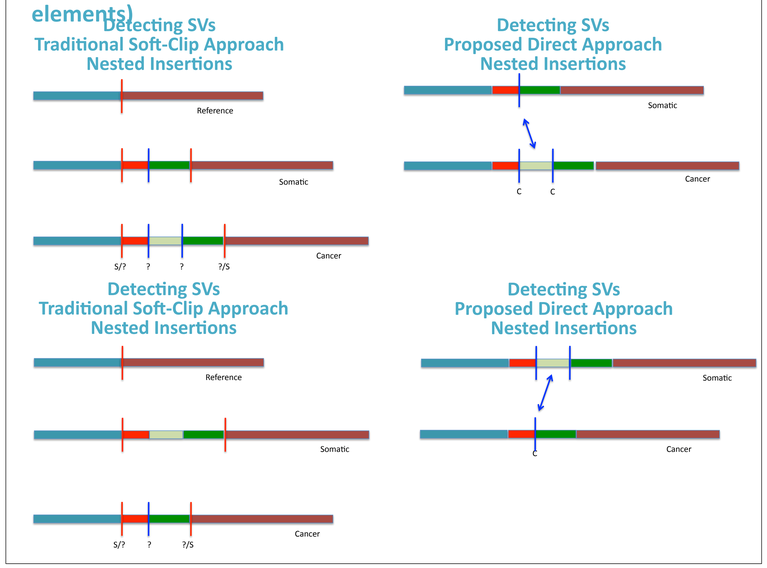


2) Improve Confidence of Putative Breakpoints

- Calling 2 of every breakpoint (1 for each genome)
- Greater confidence in breakpoints predicted by both

3) Detect nested SVs

- Any indels in a region inserted in the somatic genome relative to (human reference) cannot be detected indirectly
- But they can be easily detected directly comparing somatic to cancer genome
- Expect preference for same regions to vary (e.g. transposable elements)



4) Output breakpoints on full cancer genome

- Traditionally only view breakpoints in the context of a reference genome which contains extraneous regions and lacks all novel regions of the individual of interest
- Instead obtain somatic and cancer genomes with breakpoints annotated on each – much more information and all pertinent to the sequenced individual

THE UNIVERSITY OF CHICAGO BIOLOGICAL SCIENCES