Indel Cleaning and Calling

NGS Analysis and Visualization Workshop
February 4, 2010
Andrey Sivachenko, Eric Banks
Motivation

• Sequence aligners are often unable to perfectly map reads containing insertions or deletions (indels)
  • Indel-containing reads can be either left unmapped or arranged in gapless alignments
  • Mismatches in a particular read can interfere with the gap, esp. in low-complexity regions
  • Single-read alignments are “correct” in a sense that they do provide the best guess given the (limited!) information and constrains.

• Major issues:
  • Indel detection becomes difficult with so many missing reads
  • Indels can be overlooked or misplaced in individual reads
  • Artifacts introduced by the gapless alignments cause the appearance of false positive SNPs (usually in clusters)
Example: SNP clusters are really a hidden indel

Before MSA realignment:

- Notice that the “SNP”s are all found in clusters.
- Notice that the “SNP”s change depending on which end of the read span them.
- Most likely what you’re looking at is a 1bp deletion (see next slide); the aligner is unable to accurately align the reads here.
Example: SNP clusters are really a hidden indel

• SNP clusters disappear when it is run through our MSA realigner...
Example: Indel “scatter”

Even when aligner detects indels in individual reads successfully, they can be scattered around (e.g. due to additional mismatches in the read)

- A (heterogeneous) **insertion** + adjacent **insertion** → clean homogeneous (?) **insertion**
SNP Calling: Bayesian SNP caller on 10Mb of merged pilot 1 (low coverage) reads for CEU individuals

- There were **74,363** total SNPs called in the region

- Of those SNPS, **36,438 (49%) occurred in clusters** (cluster = 2 or more SNPs within 10bp)
  - About half the SNP calls are ignored with naïve filtering!

- Nearby clusters (i.e. less than a read length away) were merged
  - There were 3,356 total clusters after merging

- 30% of the SNP clusters were removed by realignment
MSA for Resequencing Applications

- We have the reference and (approximate) placement
- Departures from the reference are small
- Generate alt reference as suggested by each non-matching read (Smith-Waterman)
- Test each non-matching read against each alt reference candidate
- Select alt reference consensus: best “home” for all non-matching reads
- Why is it MSA: look for improvement in overall placement score (sum across reads)
- Optimizations and constrains:
  - Expect two alleles
  - Expect a single indel
  - Downsample in regions of very deep coverage
  - Alignment has an indel: use that indel as an alt. ref candidate
GATK Realignment Pipeline Implementation

**JAVA_CMD** = java –Xmx4096m –jar <path to GenomeAnalysisTK.jar>

**REF** = reference fasta file
(e.g. /seq/references/Homo_sapiens_assembly18/v0/Homo_sapiens_assembly18.fasta)

```
JAVA_CMD \n-T RealignerTargetCreator \n-R REF \n-I <my_reads.recal.bam> \n-o <target.intervals>
```

```
JAVA_CMD \n-R REF \n-I INFO \n-T IntervalCleaner \n-I <my_reads.recal.bam> \n-L <target.intervals> \n-O <cleaned.bam> \ncleanedOnly [-compress 1]
```

```
JAVA_CMD
-T CleanedReadInjector \n-R REF \n-I <my_reads.recal.bam> \n--cleaned_reads <cleaned.bam> \n--output_bam <final.cleaned.bam> \n[-compress 1]
```
Indel Calling

• Current procedure: pure cutoff-based heuristics using counts
  – Min. coverage
  – Min. fraction of alignments supporting the indel
  – Max. fraction of residual mismatches around the indel
  – For somatic: a call in tumor without any evidence for germline event (some min. coverage in normal is required)

```
JAVA_CMD \n-T IndelGenotyperV2 \n-R REF \n-I <final.cleaned.bam> \n[-I <final.tumor.cleaned.bam> --somatic] \n[--verbose --o <indels.with.stats.txt>] \n[--refseq /humgen/gsa-scr1/GATK_Data/refGene.sorted.txt] \n[--blacklistedlanes <lane_blacklist.txt>] \n-outputFile <indels.bed> \n
# cleaned bam or cleaned normal bam (for somatic)
# cleaned tumor bam (if calling somatic; MUST be specified AFTER normal)
# indels annotated with additional stats and (optionally) gene loci
# annotate indels with gene loci
# completely ignore reads coming from the specified lanes
# print just indels in simple bed format
```

• Additional post-filtering
Advice: Post-filtering

- Filter out indels when:
  - Indel-containing reads carry too many mismatches, on average
  - High fraction of mismatches in a small window around the indel
  - Somatic event is called within 10 bp of germline event called in another sample

- Validation rates
  - Up to 95% for germline events
  - Estimated 30-40% for somatic events [more validation data needed]
Future Directions

- Clean multiple files jointly - SOON
  - Sample(s) may have low coverage at the site: a read that could inform MSA about the correct alt consensus is missing
  - Pooling and cleaning reads from N samples jointly: xN coverage and better chance to sample correct alt consensus
- When cleaning an interval, also try known alt consensus variants (dbSNP, HapMap, 1kG, custom…) – SOON
- Create complete final bam file on the fly, while cleaning - SOON
  - No need for the separate CleanedReadInjector final stage
- Move away from cutoffs towards statistically based scoring system for the calls - TBD