CS681: Advanced Topics in Computational Biology

Can Alkan
EA509
calkan@cs.bilkent.edu.tr

http://www.cs.bilkent.edu.tr/~calkan/teaching/cs681/

SNP discovery with HTS data

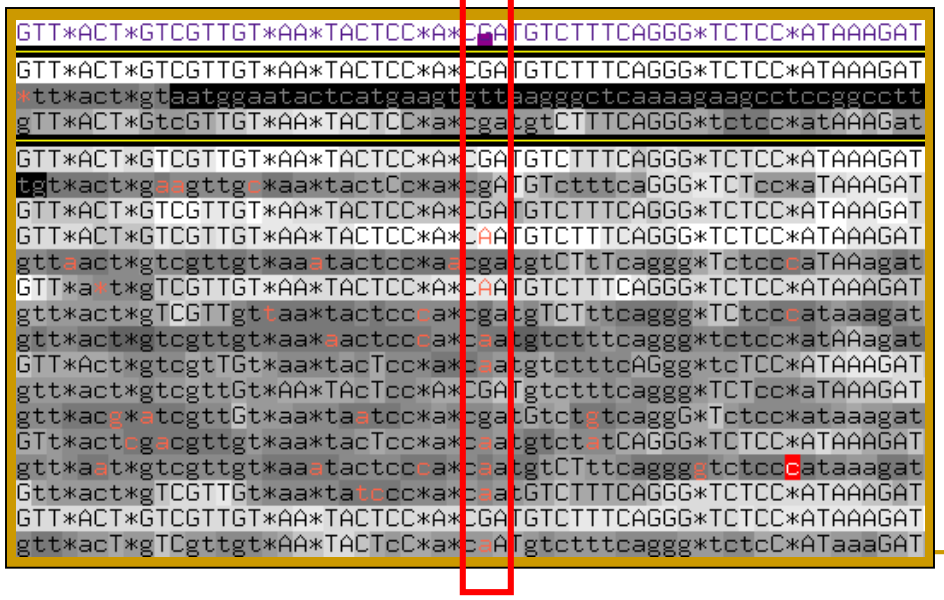
- SNP: single nucleotide polymorphism
  - Change of one nucleotide to another with respect to the reference genome
  - 3-4.5 million SNPs per person
  - Database: dbSNP http://www.ncbi.nlm.nih.gov/projects/SNP/
- Input: sequence data and reference genome
- Output: set of SNPs and their genotypes (homozygous/heterozygous)
- Often there are errors, filtering required
- SNP discovery algorithms are based on statistical analysis
- Non-unique mappings are often discarded since they have low MAPQ values

Resequencing-based SNP discovery

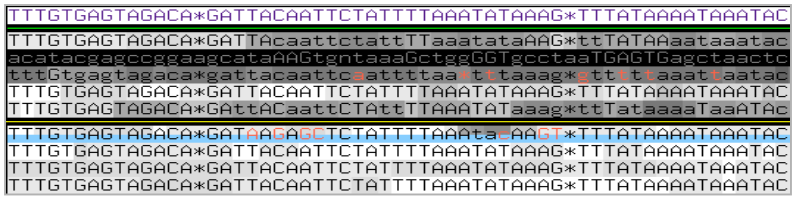
genome reference sequence

Read mapping

Read alignment



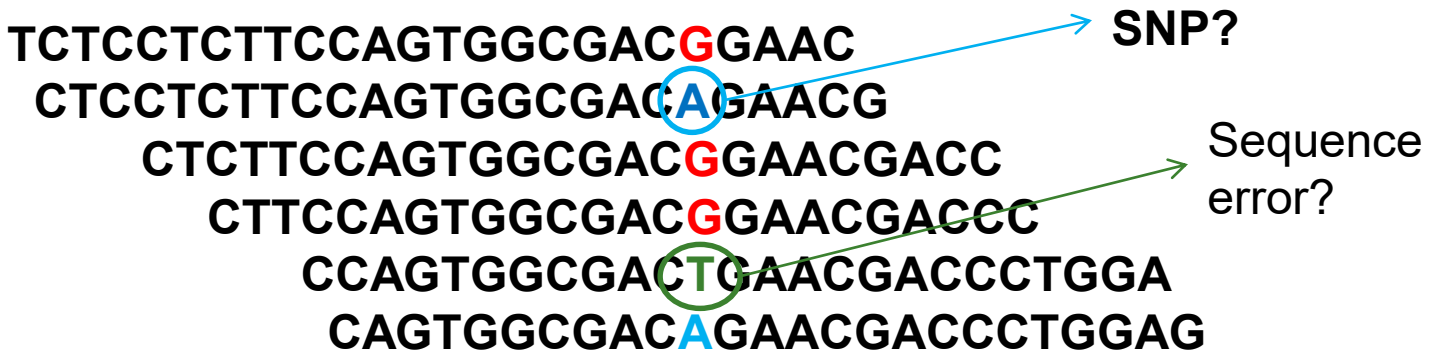
Paralog identification



SNP detection + inspection

Goal

- Given aligned short reads to a reference genome, is a read position a SNP, PSV or error?

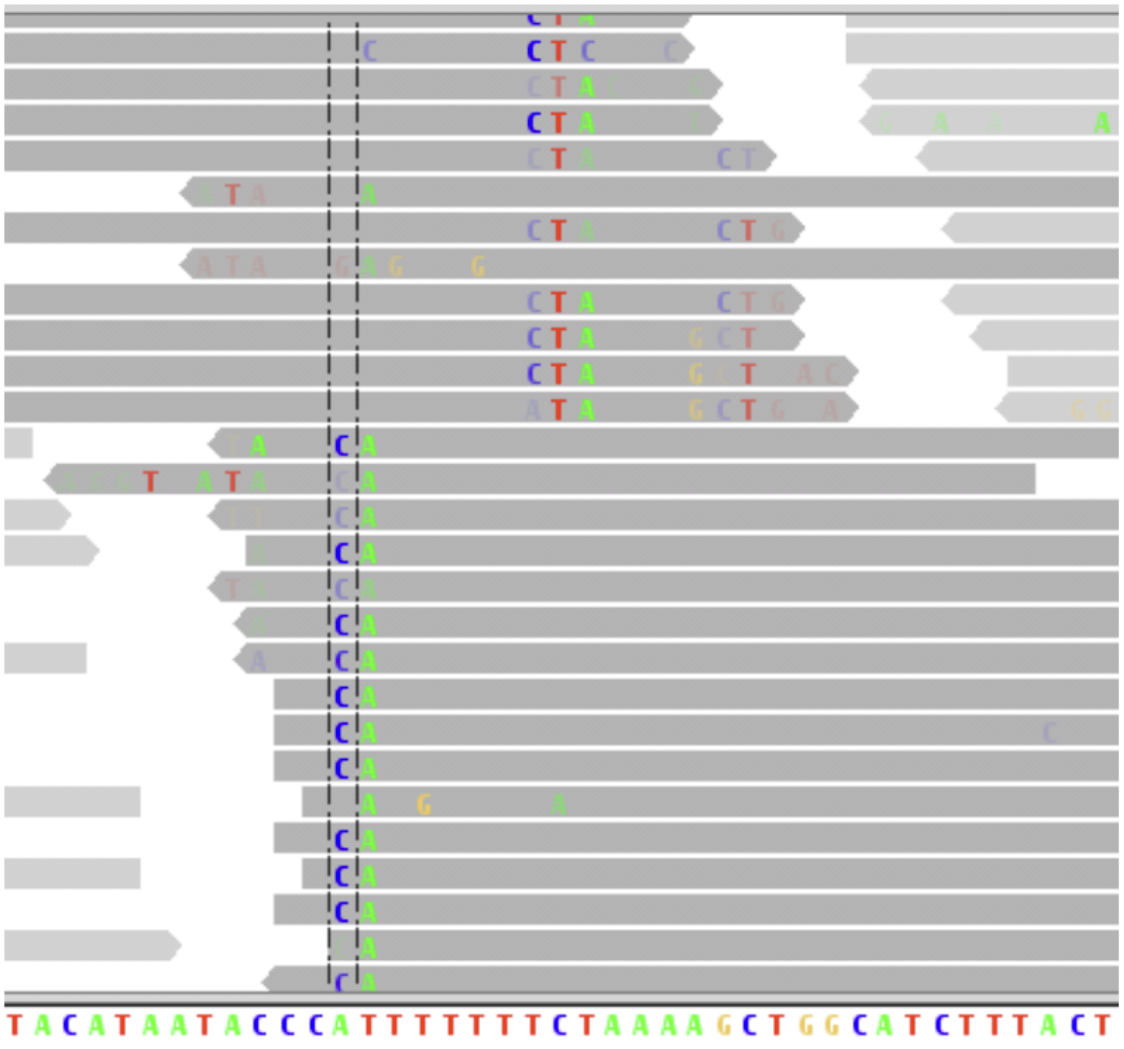


Reference **TCTCCTCTTCCAGTGGCGACGGAACGACCCTGGAGCCAAGT**

Challenges

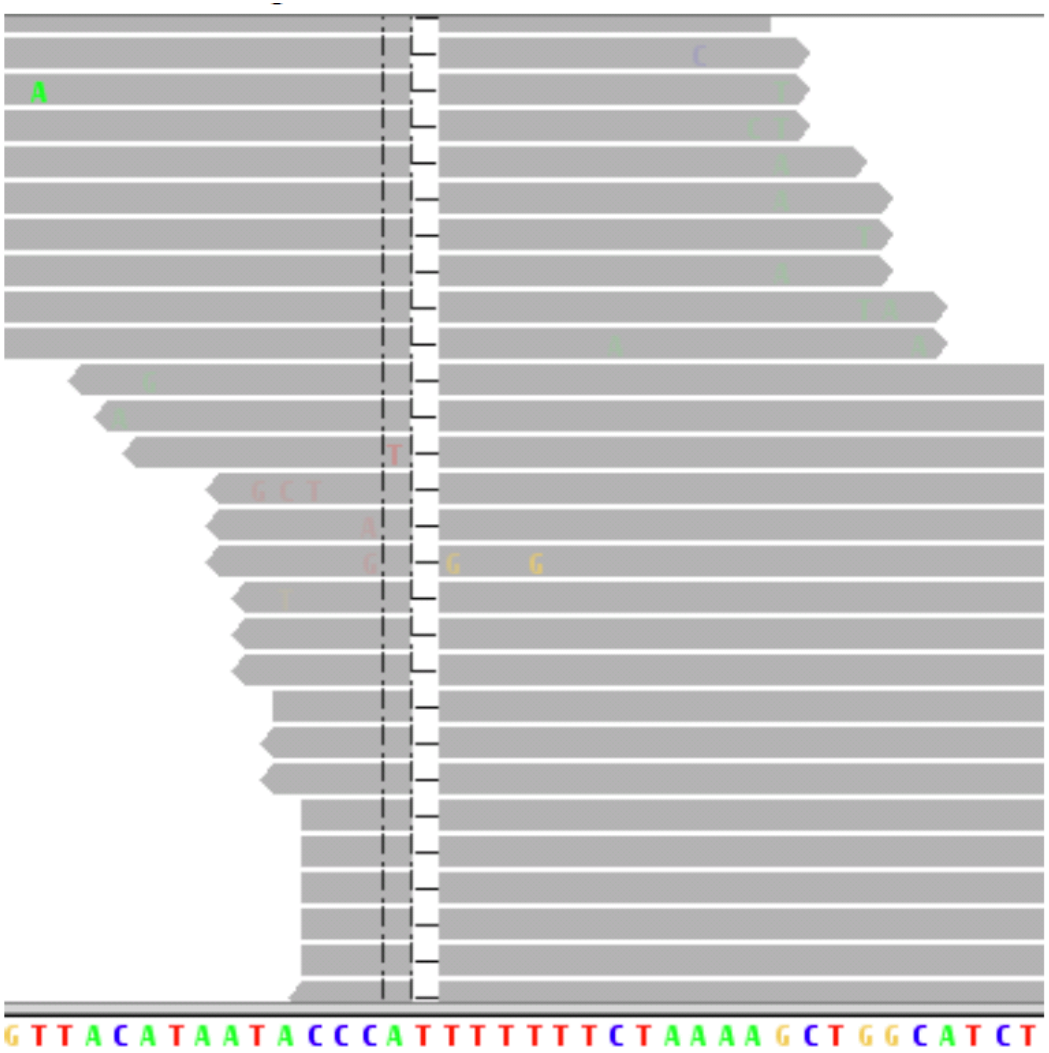
- Sequencing errors
- Paralogous sequence variants (PSVs) due to repeats and duplications
- Misalignments
  - Indels vs SNPs, there might be more than one optimal trace path in the DP table
  - Short tandem repeats
  - Need to generate multiple sequence alignments (MSA) to correct

Need to realign



Slide from Andrey Sivachenko

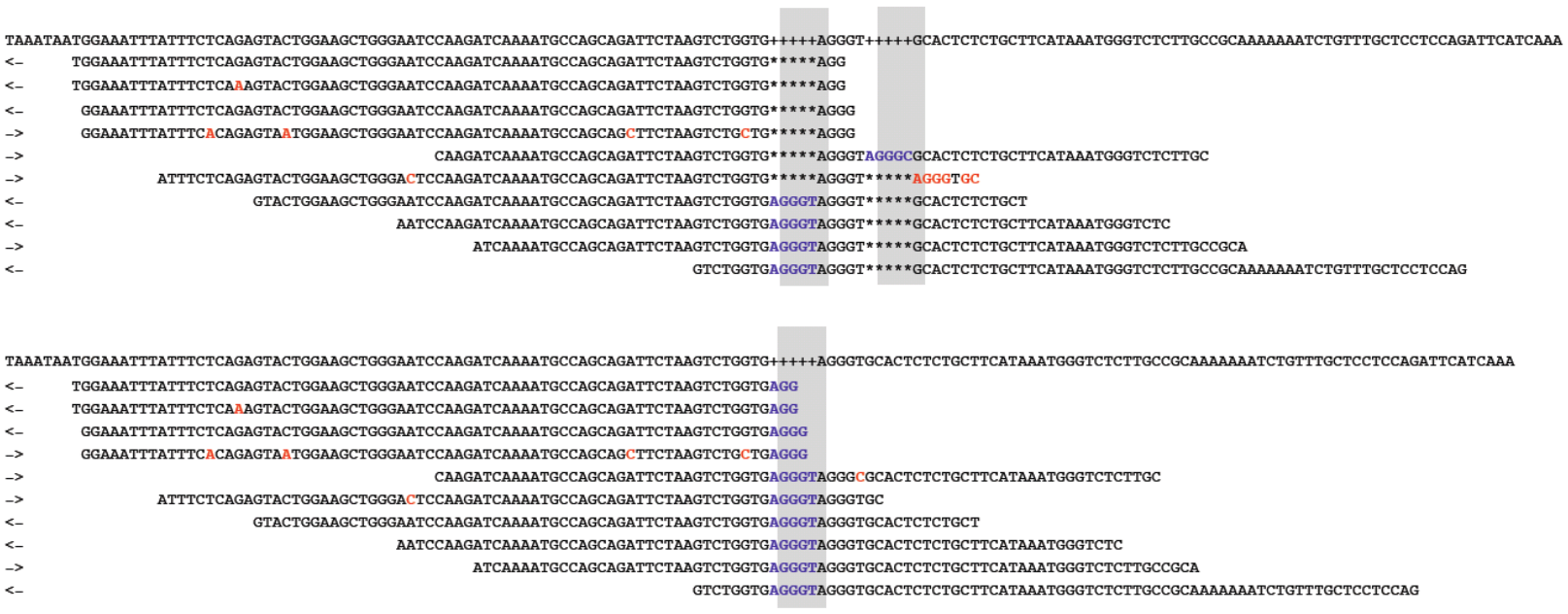
After MSA



Slide from Andrey Sivachenko

Indel scatter

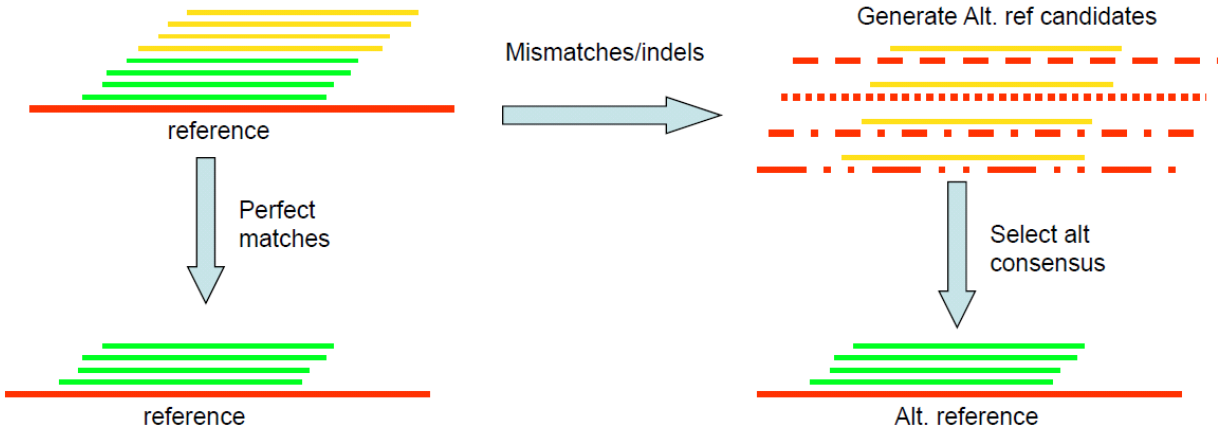
Even when read mapper detects indels in individual reads successfully, they can be scattered around (due to additional mismatches in the read)



Slide from Andrey Sivachenko

MSA for resequencing

- 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



Slide from Andrey Sivachenko

GATK HaplotypeCaller

- No MSA needed
- All reads around a candidate region is assembled
  - into two haplotypes when possible
- Phasing is possible

SNP callers

- Genome Analysis Tool Kit (GATK; Broad Inst.)
  - UnifiedGenotyper (deprecated)
  - HaplotypeCaller (standard)
- Samtools (Sanger Centre)
- FreeBayes (Boston College)
- SOAPsnp (BGI)
- VARiD (U. Toronto)
- ....

Base quality recalibration

- The quality values determined by sequencers are not optimal
- There might be sequencing errors with high quality score; or correct basecalls with low quality score
- Base quality recalibration: after mapping correct for base qualities using:
  - Known systematic errors
  - Reference alleles
  - Real variants (dbSNP, microarray results, etc.)
- Most sequencing platforms come with recalibration tools
- In addition, GATK & Picard have recalibration built in

GATK SNP calling

$$P(G \mid D) = \frac{P(G)P(D \mid G)}{\sum_i P(G_i)P(D \mid G_i)}$$

$$P(D \mid G) = \prod_j \left(\frac{P(D_j \mid H_1)}{2} + \frac{P(D_j \mid H_2)}{2} \right), where\ G = H_1 H_2$$

$$P(D_j \mid H) = P(D_j \mid b)$$

$$P(Dj \mid b) = \begin{cases} 1 - \varepsilon_j & D_j = b \\ \varepsilon_j & otherwise \end{cases}$$

G: genotype
D: data
H: haplotype
b: base

GATK genotype likelihoods

Likelihood for the genotype

Prior for the genotype

Likelihood for the data given genotype

Independent base model

$$L(G \mid D) = P(G)P(D \mid G) = \prod_{b \in \{good\_bases\}} P(b \mid G)$$

- Likelihood of data computed using pileup of bases and associated quality scores at given locus
- Only “good bases” are included: those satisfying minimum base quality, mapping read quality, pair mapping quality
- *P(b | G)* uses platform-specific confusion matrices
- *L(G|D)* is computed for all 10 genotypes

Slide from Mark Depristo

SNP calling artifacts

- SNP calls are generally infested with false positives
  - From systematic machine artifacts, mismapped reads, aligned indels/CNV
  - Raw/unfiltered SNP calls might have between 5-20% FPs among novel calls
- Separating true variation from artifacts depends very much on the particulars of one's data and project goals
  - Whole genome deep coverage data, whole genome low-pass, hybrid capture, pooled PCR are have significantly different error models

Slide from Mark Depristo

Filtering

- Hard filters based on
  - Read depth (low and high coverage are suspect)
  - Allele balance
  - Mapping quality
  - Base quality
  - Number of reads with MAPQ=0 overlapping the call
  - Strand bias
  - SNP clusters in short windows

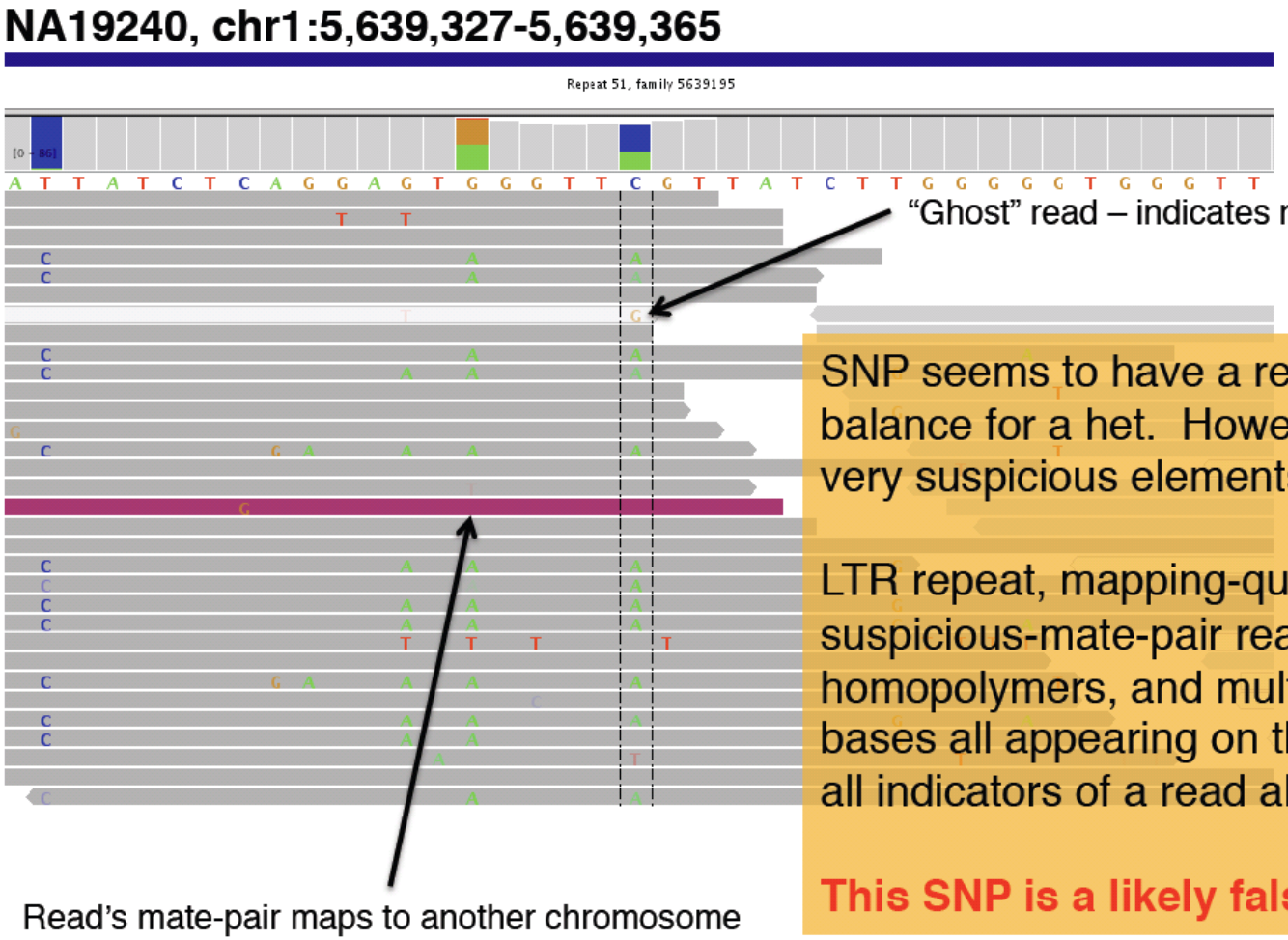
Filtering

- Statistical determination of filtering parameters:
  - Training data: dbSNP, HapMap, microarray experiments, other published results
  - Based on the distribution of values over the training data adjust cut off parameters depending on the sequence context
    - VQSR: Variant Quality Score Recalibration

Indicators of call set quality

- Number of variants
  - Europeans and Asians: ~3 million; Africans: ~4-4.5 million
- Transition/transversion ratio
  - Ideally Ti/Tv= 2.1
- Hardy Weinberg equilibrium
  - Allele and genotype frequencies in a population remain constant
  - For alleles **A** and **a**; freq(**A**)=p and freq(**a**)=q; $p+q=1$
  - If a population is in equilibrium then
    - freq(**AA**) = p^2
    - freq(**aa**) = q^2
    - freq(**Aa**) = $2pq$
- Presence in databases: dbSNP, HapMap, array data
- Visualization

Validation through visualization



SNP seems to have a reasonable allele balance for a het. However, there are some very suspicious elements in the vicinity.

LTR repeat, mapping-quality-zero read, suspicious-mate-pair read, short homopolymers, and multiple non-reference bases all appearing on the same reads are all indicators of a read alignment error.

This SNP is a likely false-positive.

Slide from Kiran Garimella

Pooled sequencing

- When sequence coverage is low, pool mapping of data from multiple samples (ideally from the same population) into a single file
- SNP calling is more challenging
  - Allele frequencies close to error rate
  - Track which read comes from which individual

NEXT: INDELS

Indel discovery with HTS data

- Indels: insertions and deletions < 50 bp.
  - ~0.5 million indels per person
  - Database: dbSNP http://www.ncbi.nlm.nih.gov/projects/SNP/
- Input: sequence data and reference genome
- Output: set of indels and their genotypes (homozygous/heterozygous)
- Often there are errors, filtering required
- Most indel detection methods are based on statistical analysis
- Tools: GATK, Dindel, Pindel, SAMtools, SPLITREAD, PolyScan, VarScan, etc.

Challenges (reminder)

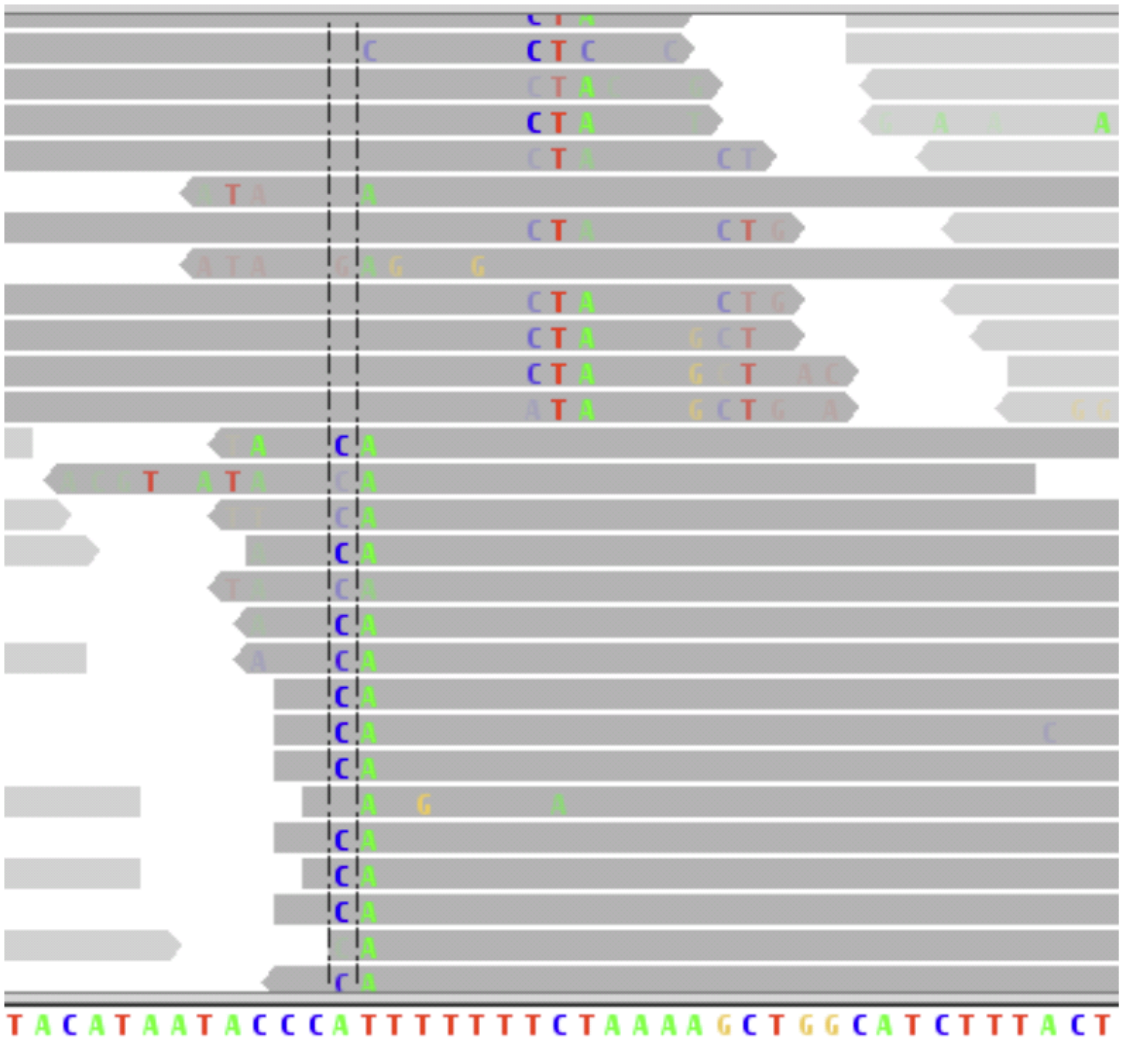
- Sequencing errors
- Paralogous sequence variants (PSVs) due to repeats and duplications
- Misalignments
  - Indels vs SNPs, there might be more than one optimal trace path in the DP table
  - Short tandem repeats
  - Need to generate multiple sequence alignments (MSA) to correct

Finding indels

- 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 constraints.

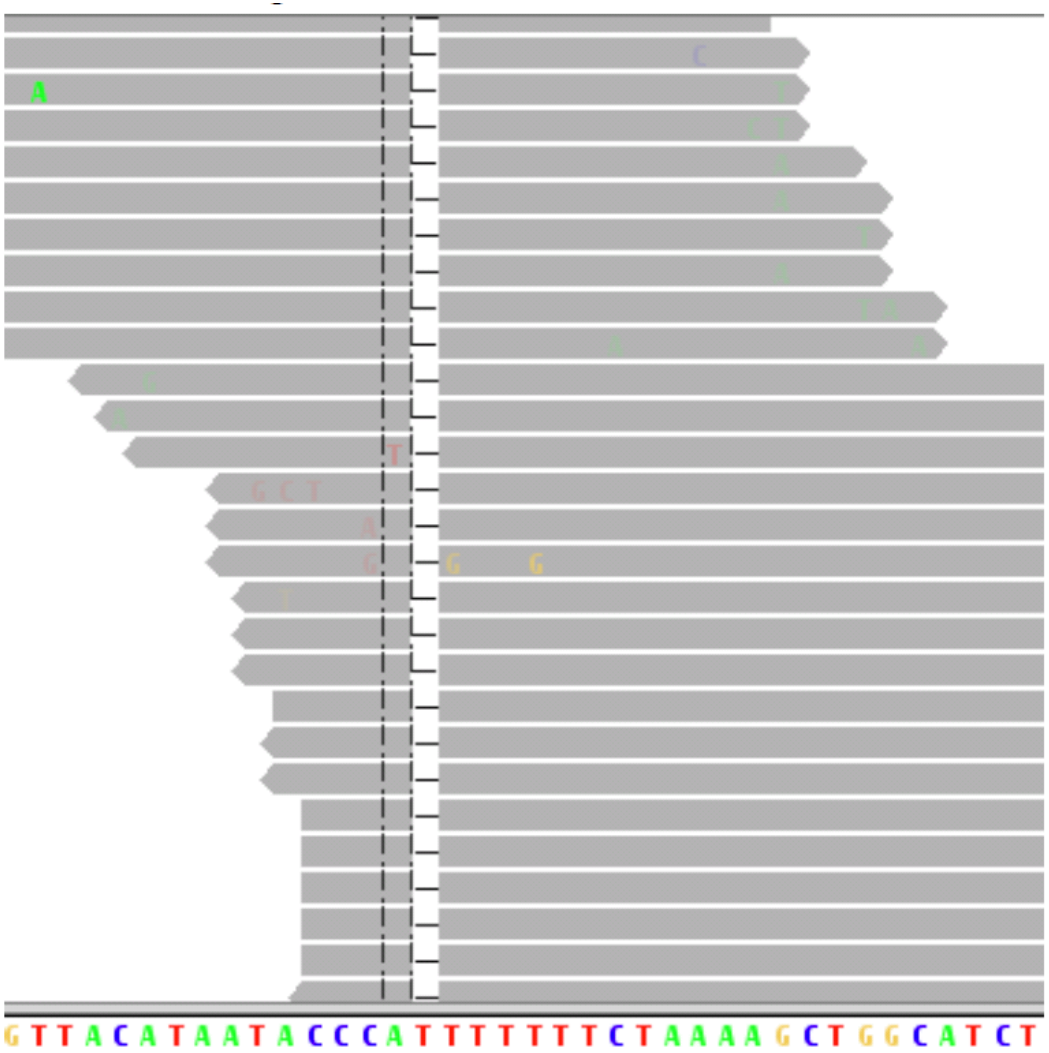
Slide from Andrey Sivachenko

Need to realign



Slide from Andrey Sivachenko

After MSA



Slide from Andrey Sivachenko

Left alignment of indels

- If there is a short repeat, there might be more than one alternative alignments of indels
  - Common practice is to select the “left aligned” version

CGTATGATCTAGCGCGCTAGCTAGCTAGC
CGTATGATCTA - - GCGCTAGCTAGCTAGC

CGTATGATCTAGCGCGCTAGCTAGCTAGC
CGTATGATCTAGC - - GCTAGCTAGCTAGC

CGTATGATCTAGCGCGCTAGCTAGCTAGC
CGTATGATCTAGCGC - -TAGCTAGCTAGC

Left aligned