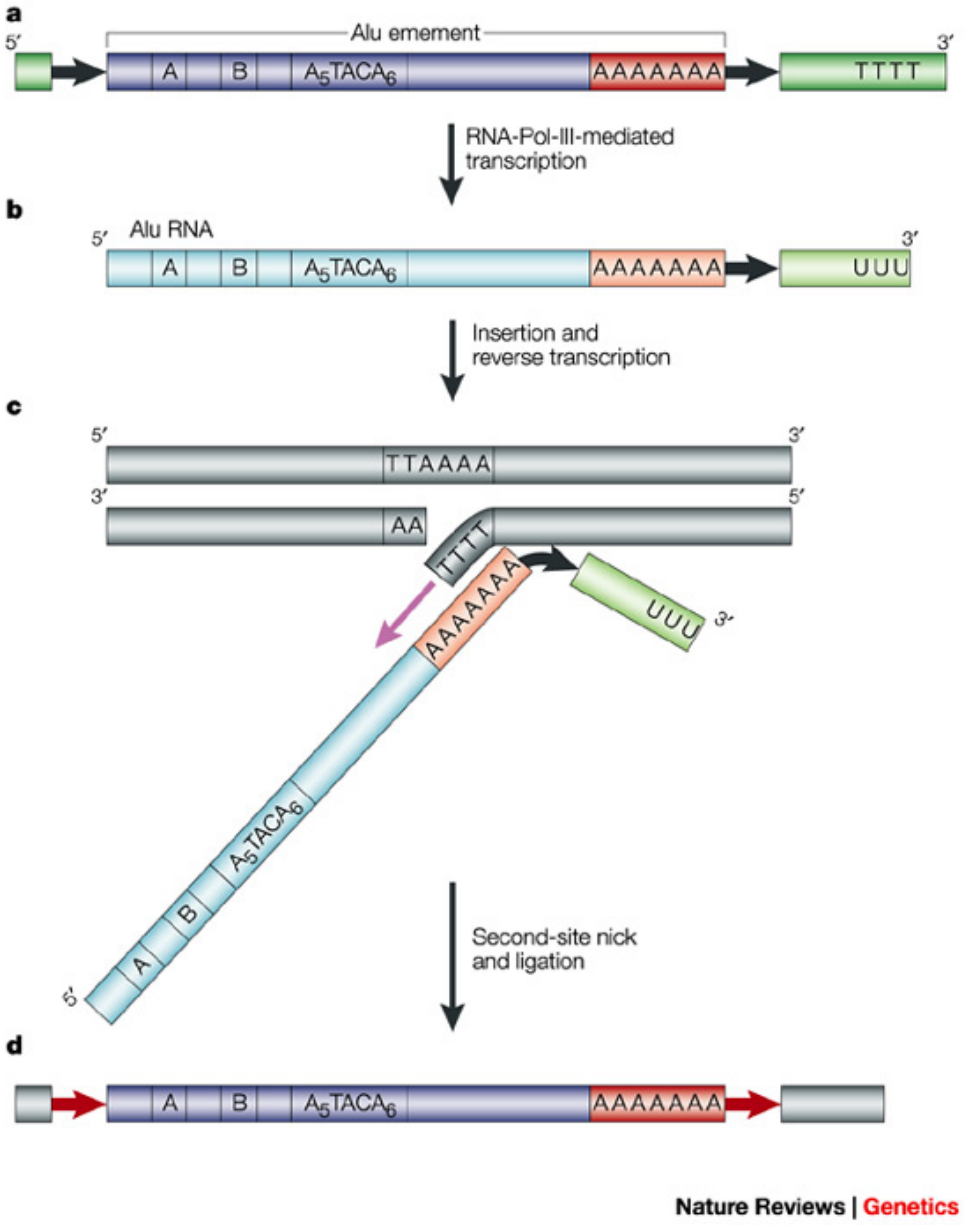
Alu Polymorphism Detection

Petko Fiziev, Bioinformatics IDP

What are Alu repeats?



Source: Batzer et al. 2002

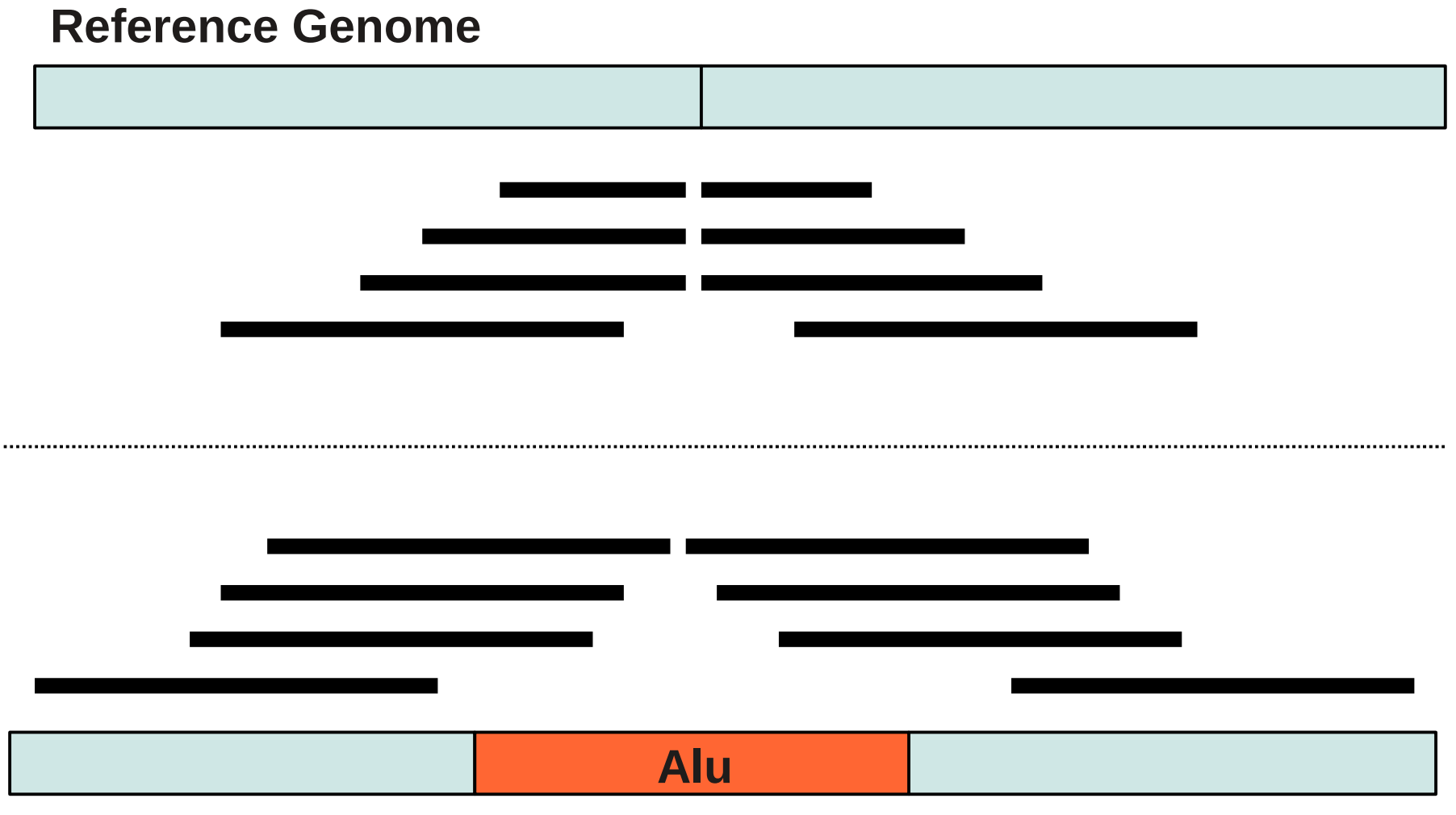
Why do we care about Alus?

- insertional mutations
- gene conversion
- recombination
- alterations in gene expression,
- pseudogenization
- structural variation
- formation of segmental duplications
- genetic marker for diseases
- track down human evolution

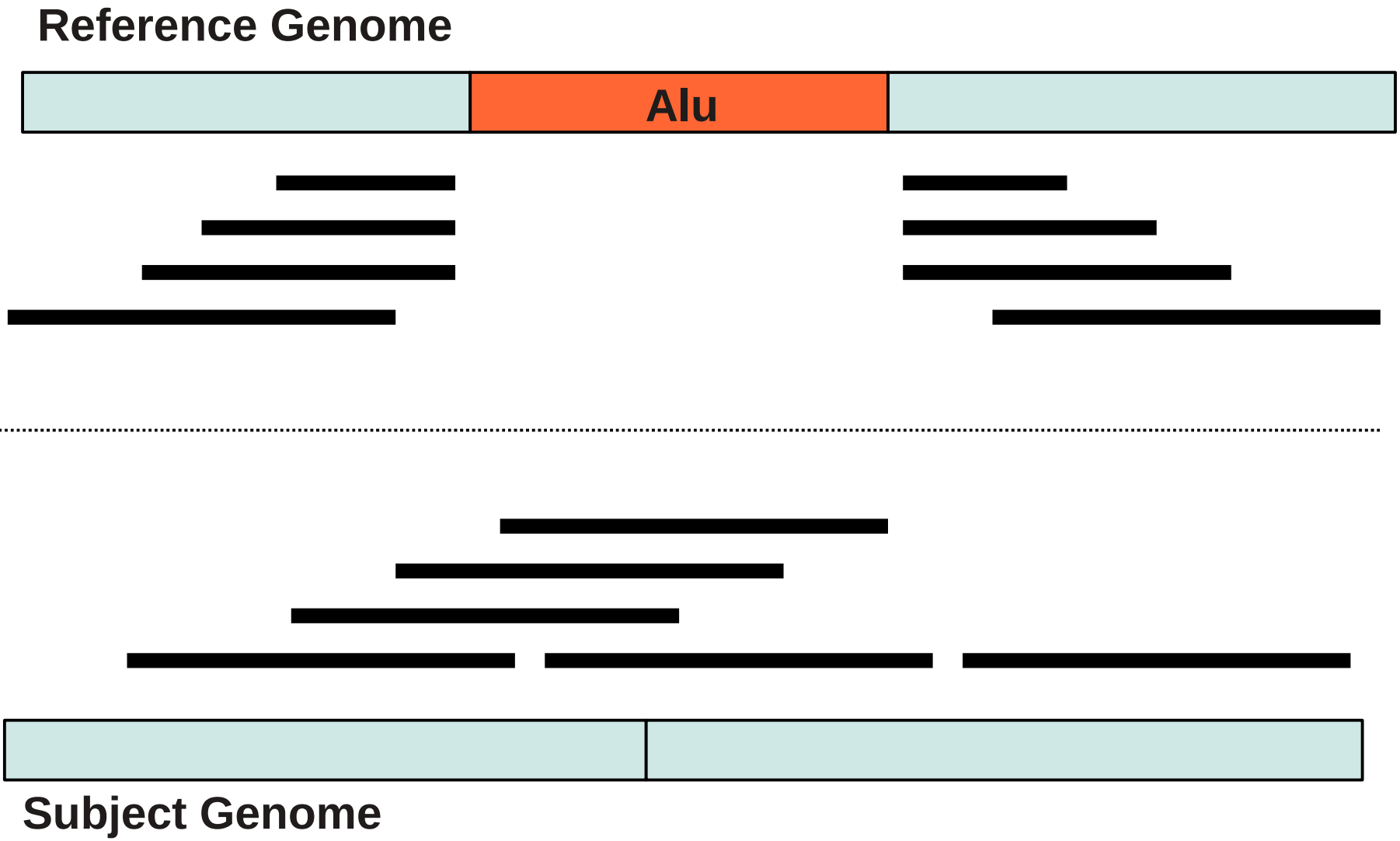
Project goals

- Given:
  - a reference genome + Alu annotation
  - a set of short reads from a subject genome
  - DB of known Alu sequences
- Detect Alu insertions and deletions
- Diploid human genome
- SNPs and small indels
- No other large structural variations
- Sequencing errors

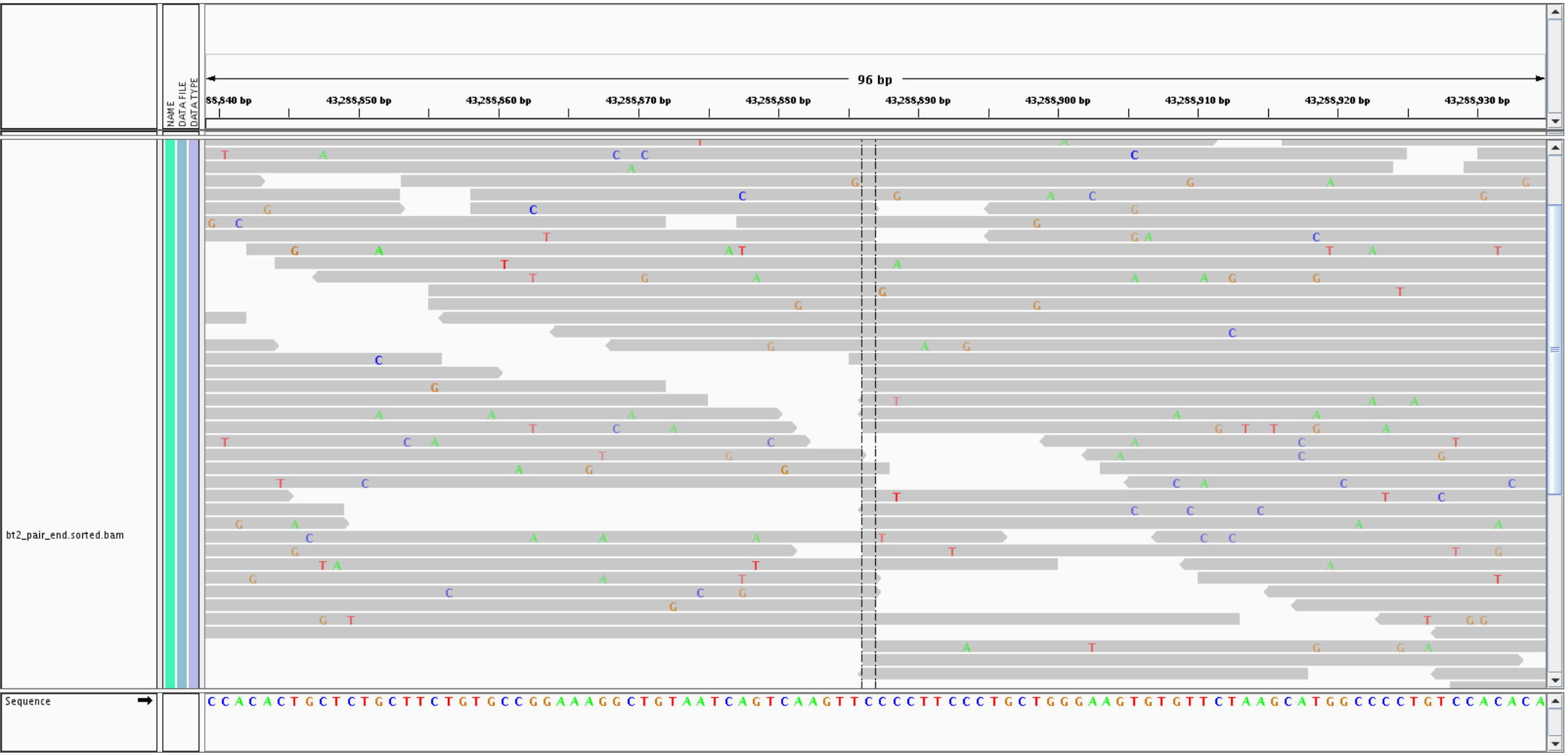
The idea (insertions)



The idea (deletions)



How real heterozygous insertions look like



The algorithm

- Map all reads to the reference (MAPQ > 10)
- Map all reads to a DB with Alu sequences to tag Alu reads
- Detect potential inserts
  - Sliding window of 30 bp
    - Are there any Alu reads in the window?
    - Count how many reads start at each position
    - Count how many reads end at each position
    - Are those counts high enough somewhere in the window?

The algorithm (part 2)

- Detect potential deletions
  - Scan all known Alus in the reference genome
  - What is the overall coverage within the Alu?
  - How many reads end at the beginning of a known Alu?
  - How many reads start at the end of a known Alu?

How much is high enough?

- N – genome size, M – # reads
- H – # reads that start at a certain position
- T – # reads that end at a certain position
- H and T ~ Poisson(λ), $\lambda = \frac{M}{N}$

$$P(Obs|H_0) = C * P(H \geq K|\lambda) * P(T \geq K|\lambda)$$

C – some constant

K = 4 or K = 5 was enough for my testing data

Is it homozygous or heterozygous?

- Calculate the ratio of spanning reads to all reads
- Insertion is on both chromosomes if the ratio is 0

Is it homozygous or heterozygous?

- Insertion is only on one of the chromosome copies if the **ratio is equal to what you expect divided by 2**. Estimate it!

Test data (insertions)

- Human chromosome 22 (50 Mbp, 23361 Alu instances from 37 families)
- Generate 2 chromosome copies:
  - Add a SNP every 1,000 bp
  - Add an indel every 10,000 bp
  - Insert 1,000 Alus per chromosome copy in random positions and orientation
  - New Alus cannot be inserted into each other, but can land inside existing Alu sequences.
  - Mutate the new Alus in 5% of their bases **(a higher mutation rate actually helps)**

Test data (short reads)

- 100 bp pair-end reads
  - Fragment lengths: 150 bp to 500 bp, uniform
  - Average overall error rate ~ Normal (mean = 5%, var = 0.01%)
  - Lower error rate at the beginning, increases gradually towards the end of the reads
  - Generate fastq with quality scores
- Generate 30x, 20x and 10x total coverage

Results (insertions)

- Total inserted Alus: 2000

Total Coverage (both chromosomes)	30x (K=5)	30x (K=4)	20x (K=4)	10x (K=3)
True Positives	1761 (88%)	1847 (92%)	1515 (75%)	817 (40%)
False Positives	2 (FDR: 0.1%)	155 (FDR: 7%)	15 (FDR: 1%)	29 (FDR: 3.43%)
False Negatives	239 (11%)	148 (7%)	470 (23%)	1174(58.70%)

- Running time < 1 hour

Test data (deletions)

- Human chromosome 22 (50 Mbp, 23361 native Alu instances from 37 families + 1000 artificially inserted Alus)
- Generate 2 chromosome copies:
  - Add a SNP every 1,000 bp
  - Add an indel every 10,000 bp
  - **Delete randomly 500 Alus** per chromosome copy from the set of artificial Alus
- Generate 30x total coverage (15x per chromosome)

Results (deletions)

- Accuracy:
  - Total deleted Alus: 1000, Alus tested: 24361
  - True positives: 850 (85%)
  - False positives: 142 (FDR: 14%)
  - False negatives: 150
- Running time ~ 1 hour

Improvements and ideas

- Improve detection of insertions within existing Alus
  - Use RepeatMasker annotation
  - Use insert size deviations
- Improve detection resolution
- Pool data from multiple individuals for higher coverage.
  - Population specific Alus (European, Asian, African etc)

References

- **Alu repeats and human genomic diversity.** Batzer et al., Nature Genetics, 2002
- **Alu repeat discovery and characterization within human genomes.** Hormozdiari et al. Genome Research, 2011
- **PAIR: polymorphic Alu insertion recognition.** Sveinbjörnsson et al., Bioinformatics 2012

Thank you!