Applied Bioinformatics

Reducing Biases in BAM files
October 2015, Belgrade

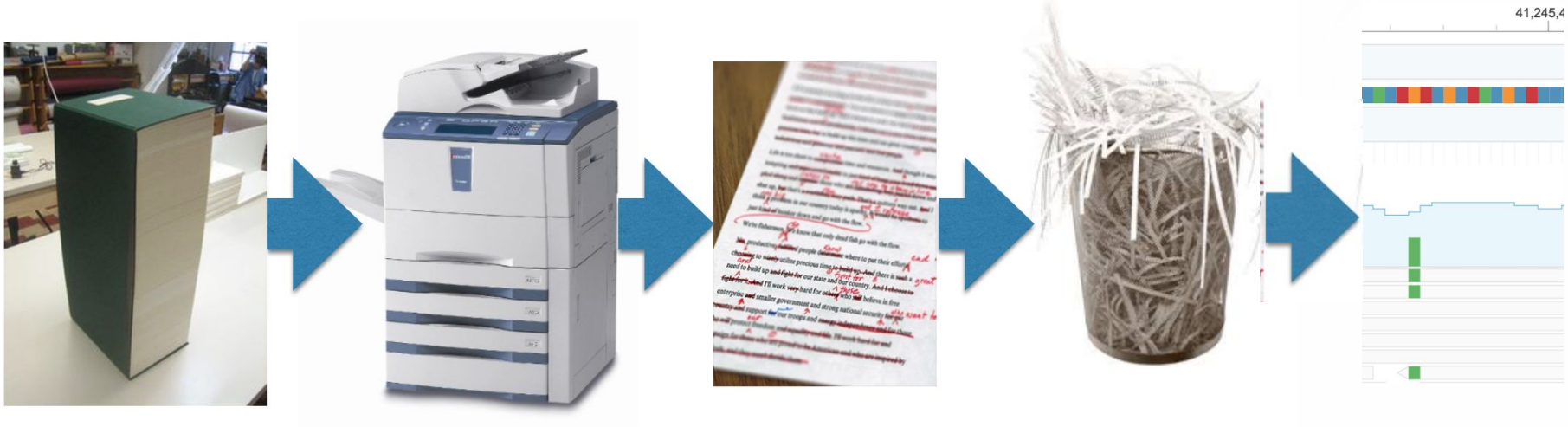
Goran Rakočević, PhD
goran.rakocevic@sbgenomics.com

Today's agenda

1. Marking duplicate reads
   a. The problem with duplicated reads
   b. Picard Mark Duplicates
2. Realignment around Indels
   a. The problem of alignment bias
   b. Idea behind Indel Realigner
   c. GATK IndelRealigner
3. Base quality score recalibration
   a. Biases in base qualities
   b. GATK Base Recalibrator

DNA Sequencing - Reminder

- We got a FASTQ file with the “reads” - little pieces of the genome



SevenBridges genomics

DNA Sequencing - Another view

- Started with many copies of the genome
- Shredded the into fragments ~several hundred basepairs in length
- Sequenced these fragments (got their sequences)

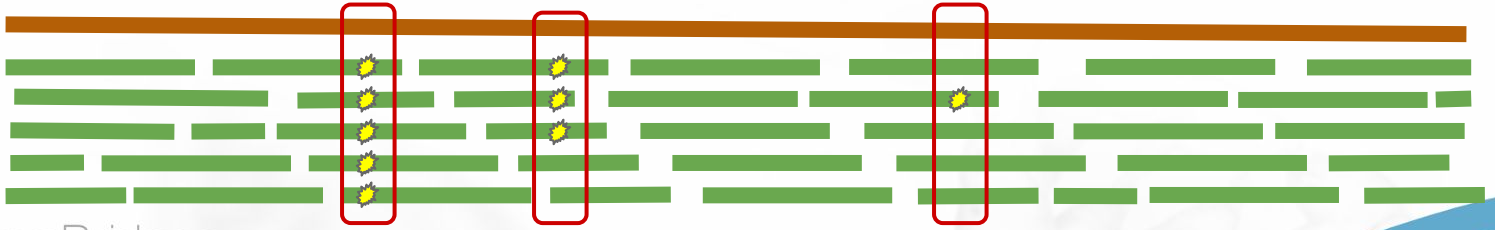
- Aligned each fragment against the reference genome
  to find the most probable location the fragment came from
- Focus on mismatching positions

SevenBridges genomics

DNA Sequencing - Another view

- Started with many copies of the genome
- Shredded the into pieces ~several hundred basepairs in length
- Sequenced these pieces (got their sequences)

- Aligned each piece against the reference genome
  to find the most probable location the piece came from
- Focus on mismatching positions



SevenBridges genomics

Input DNA amounts

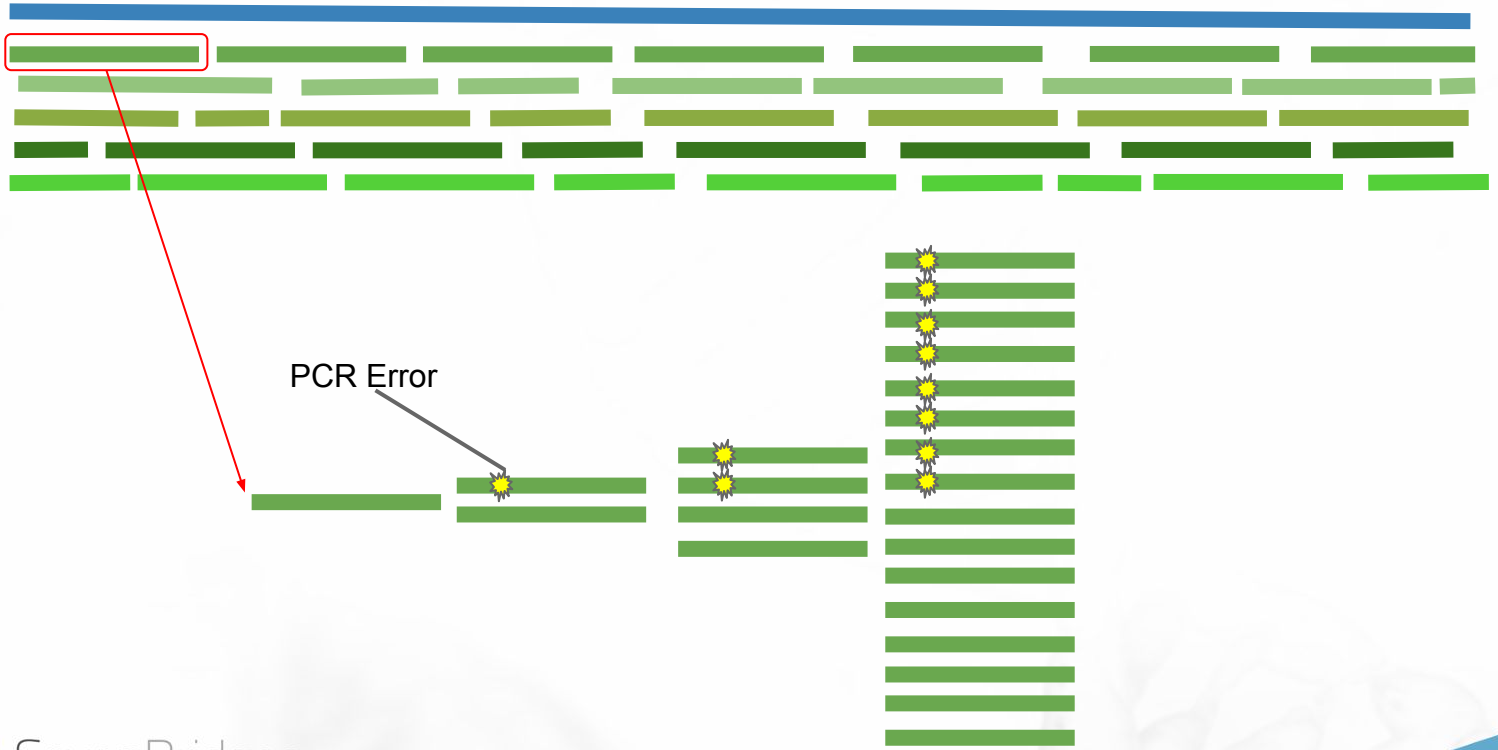
- Sequencer “catches” only a small percent of input fragments
- To cover the whole genome we need <u>a lot</u> of DNA!
- Two ways to achieve this:
  - Take a lot of starting material and extract DNA (get many copies of the genome from the organism)
  - Start with a moderate amount of DNA and make copies (use PCR process in the laboratory)
- First option can be expensive and difficult to do

SevenBridges genomics

PCR Duplication process

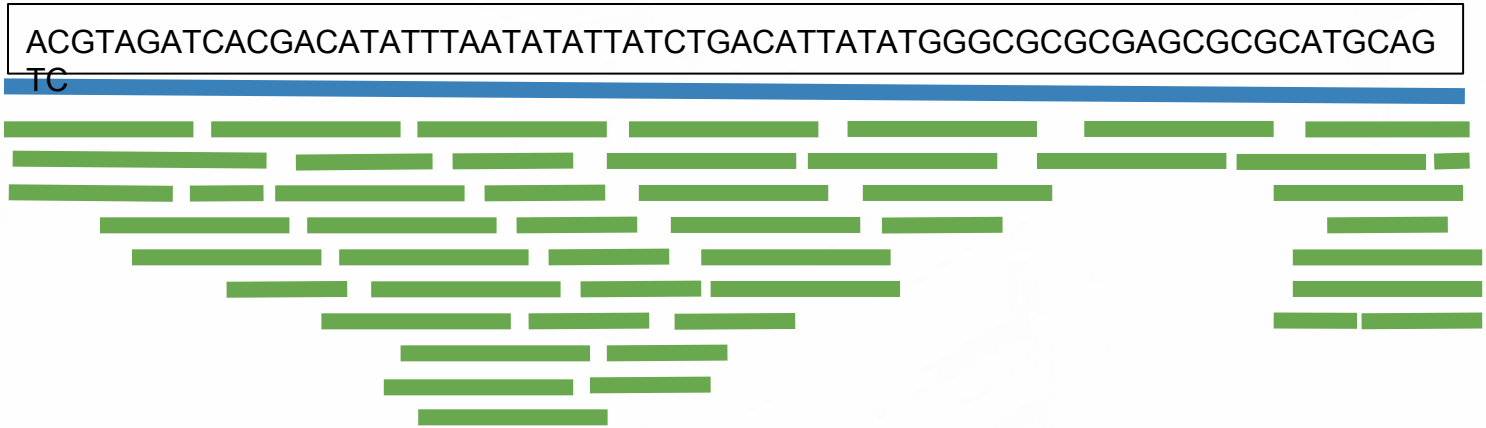
SevenBridges genomics

PCR Duplication process - Errors



SevenBridges genomics

PCR Duplication process - GC Bias



- Regions with less GC bases get higher amplification
- Uneven coverage across the genome
- More PCR is needed to get GC-rich regions to desired coverage (But that over-amplifies AT-rich regions)

SevenBridges genomics

Dealing with the effects of PCR

- GC-bias and the stochastic nature of PCR
  make precise estimations of duplications difficult
- Removing duplicates can reduce the problem of PCR errors
- A simple, conservative criterion is used by Picard Markduplicates tool:
  - If two fragments map to the sample place
    and have the length, consider them duplicates
- On average ~5% duplicates reads for WGS (up to 10%)
- Some protocols deliberately make ~90% duplicates
  - Do not dedupe those!

Realignment around indels

- Alignments are based on penalties (mismatch, gap)
- Penalties are based on statistics on general DNA sequences, so that we correctly align most of the time
- In some cases, local base distribution is such, that these penalties do not hold
- Indel can particularly cause misalignments (we get several mismatches instead of an indel)
- This causes false positives SNPs and missed Indels

Realignment around indels (2)

Delete this A

GTACACACACACGG

GTACAC—CACACGG (score = -3)

GTACACCACACGG (score = -6)

Mismatch = -1
Indel = -3

Delete this A

GTACACAAAACGG

GTACAC—AAAACGG (score = -3)

GTACACAAAACGG (score = -2)

SevenBridges genomics

A small example (1)

- Let’s look at a small region that me might be trying to align to:

GTACACACACACGG

A small example (1)

- Let's look at a small region that me might be trying to align to:

Delete this A

GTACACACACACGG

SevenBridges genomics

A small example (1)

- Let's look at a small region that me might be trying to align to:

Delete this A

GTACACACACACGG

Mismatch = -1
Indel = -3

SevenBridges genomics

A small example (1)

- Let's look at a small region that me might be trying to align to:

Delete this A

GTACAC**A**CACACGG
GTACAC—CACACGG

(Score = -3)

Mismatch = -1
Indel = -3

SevenBridges genomics

A small example (1)

- Let's look at a small region that me might be trying to align to:

Delete this A

GTACAC**A**CACACGG
GTACAACACACGG

(Score = -6)

Mismatch = -1
Indel = -3

SevenBridges genomics

A small example (2)

- Let’s look at another region that me might be trying to align to:

GTACACAAAAAAGG

A small example (2)

- Let's look at another region that me might be trying to align to:

Delete this A

GTACACAAAAAAGG

SevenBridges genomics

A small example (2)

- Let's look at another region that me might be trying to align to:

Delete this A

GTACACAAAAAAGG

Mismatch = -1
Indel = -3

SevenBridges genomics

A small example (2)

- Let's look at another region that me might be trying to align to:

Delete this A

GTACACAAAAAAGG
GTACAC—AAAAAGG

(Score = -3)

Mismatch = -1
Indel = -3

SevenBridges genomics

A small example (3)

- Let's look at another region that me might be trying to align to:

Delete this A

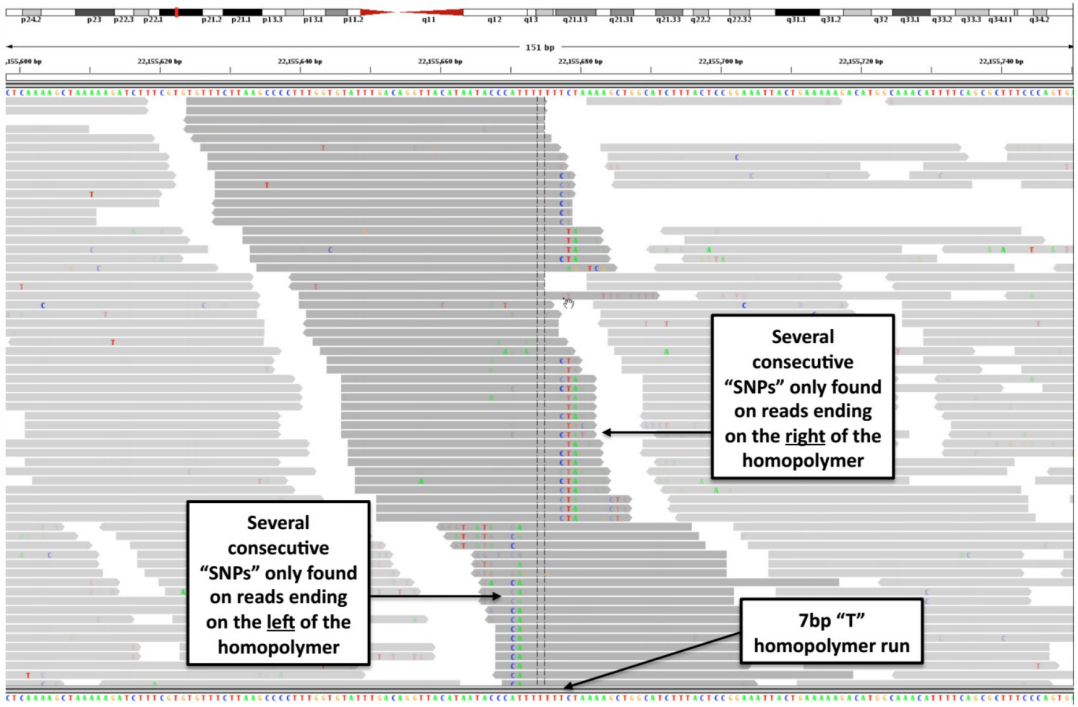
GTACACAAAAACGG
GTACACAAAACGG

(Score = -2)

Mismatch = -1
Indel = -3

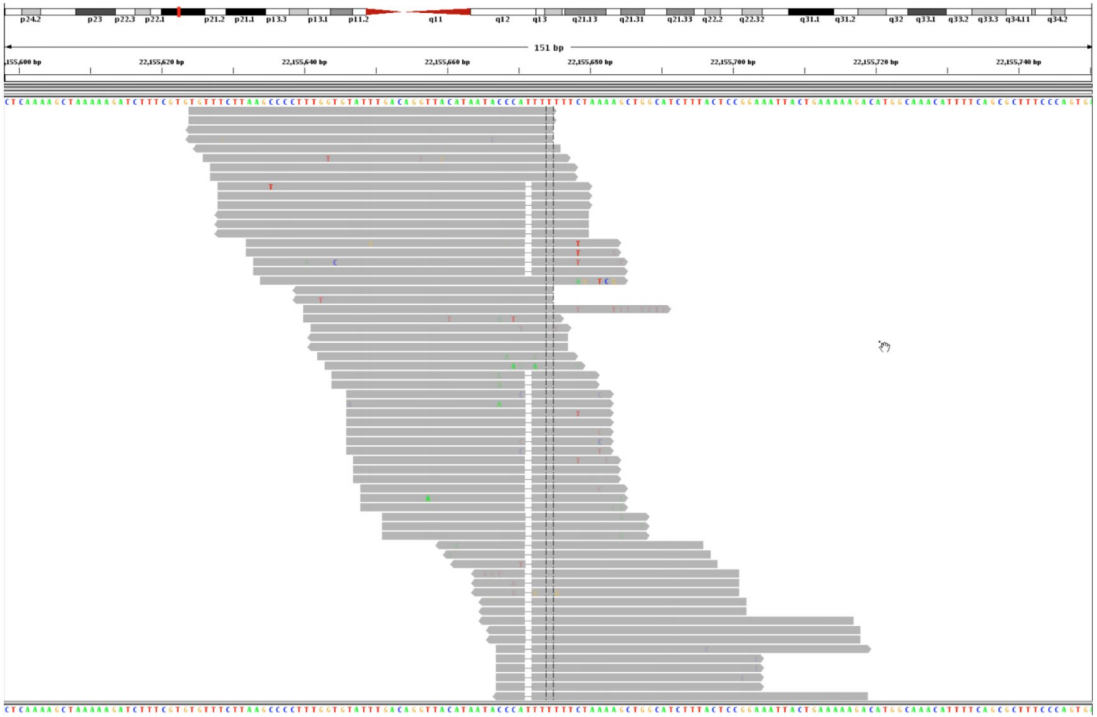
SevenBridges genomics

Real Example: A misaligned region



Source: BroadE: Best practices using GATK

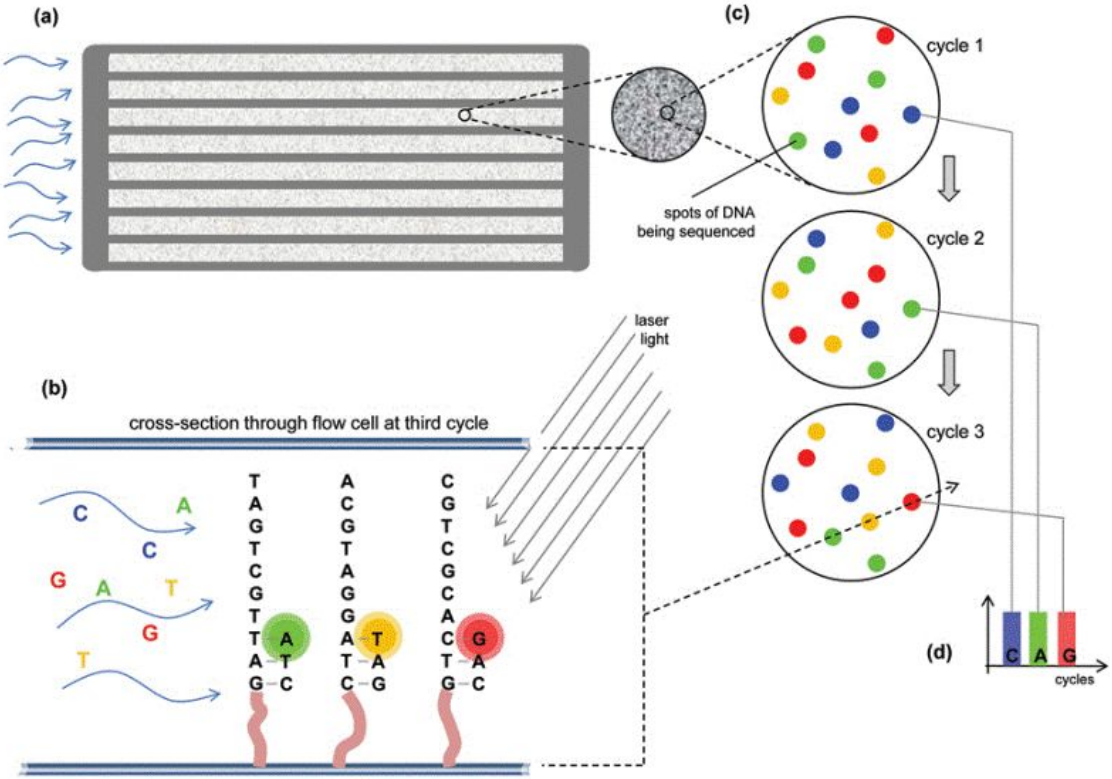
Real Example: After realignment

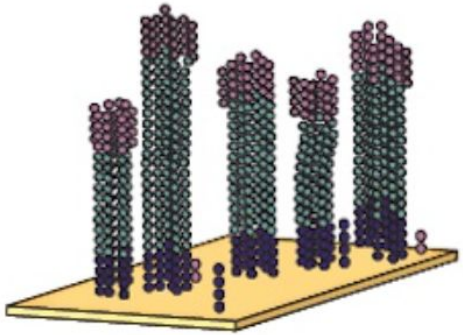


Source: BroadE: Best practices using GATK

SevenBridges genomics

Base Quality Scores - Illumina

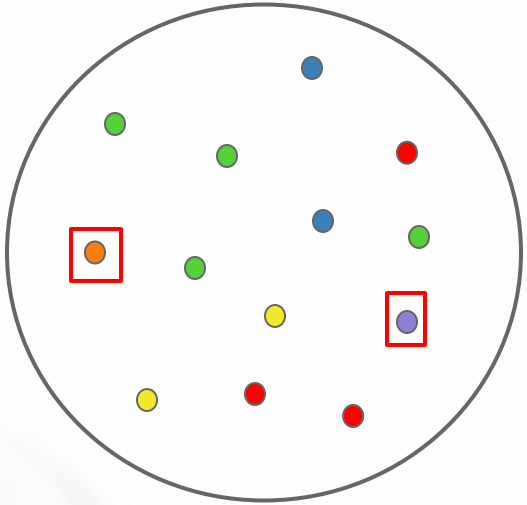




Base quality
=
Probability of seeing a color

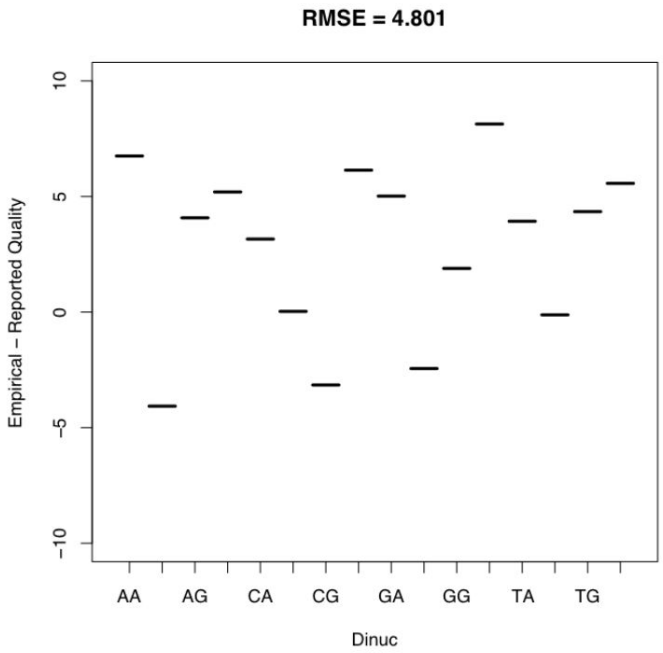
Base Quality Scores

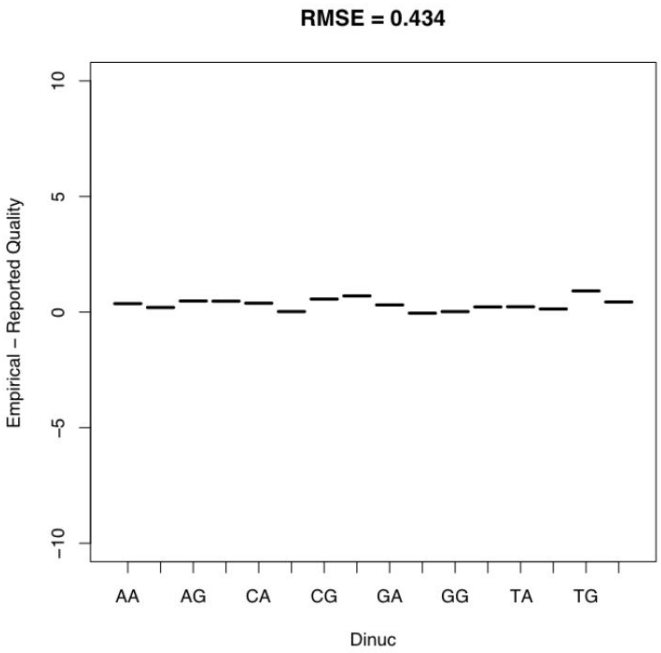
- Base quality represents the probability of a correct base-call
- These are produced by the sequencing machine



SevenBridges genomics

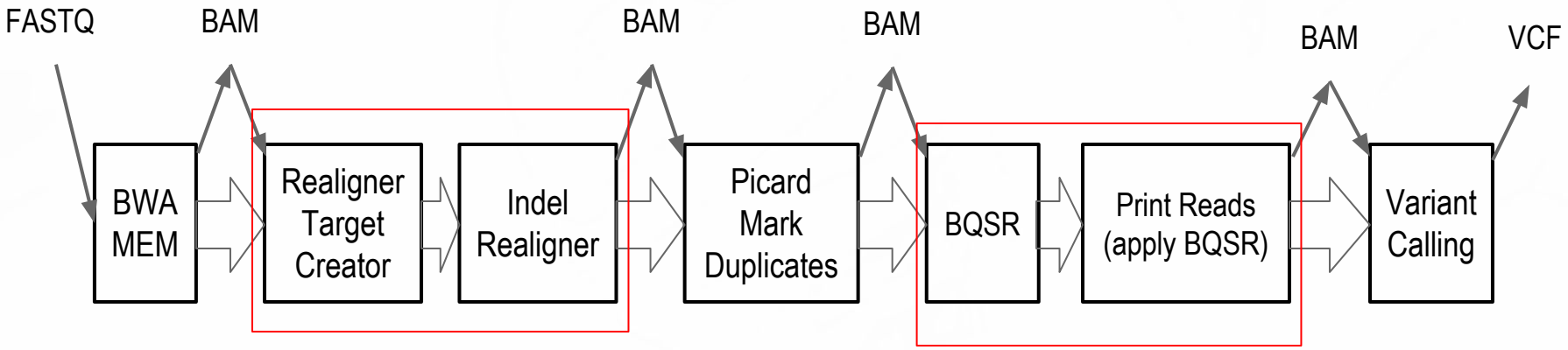
Biases in base Quality Scores





SevenBridges genomics

Full Pipeline (Broad Best Practices)



https://www.broadinstitute.org/gatk/guide/best-practices.php

SevenBridges genomics

Questions?

SevenBridges genomics