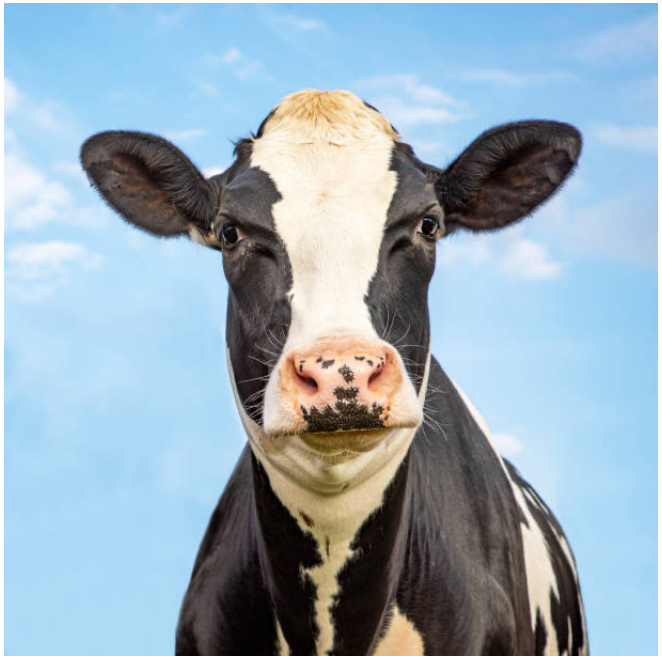
# Introduction to Whole Genome Sequencing

## From FASTQ to VARIANTS

**Rodrigo Pelicioni Savegnago**
**Department of Animal Science**

# From Fastq to VCF file

MICHIGAN STATE UNIVERSITY



# From Fastq to VCF file

MICHIGAN STATE UNIVERSITY

Hum Genet (2012) 131:1541–1554
DOI 10.1007/s00439-012-1213-z

REVIEW PAPER

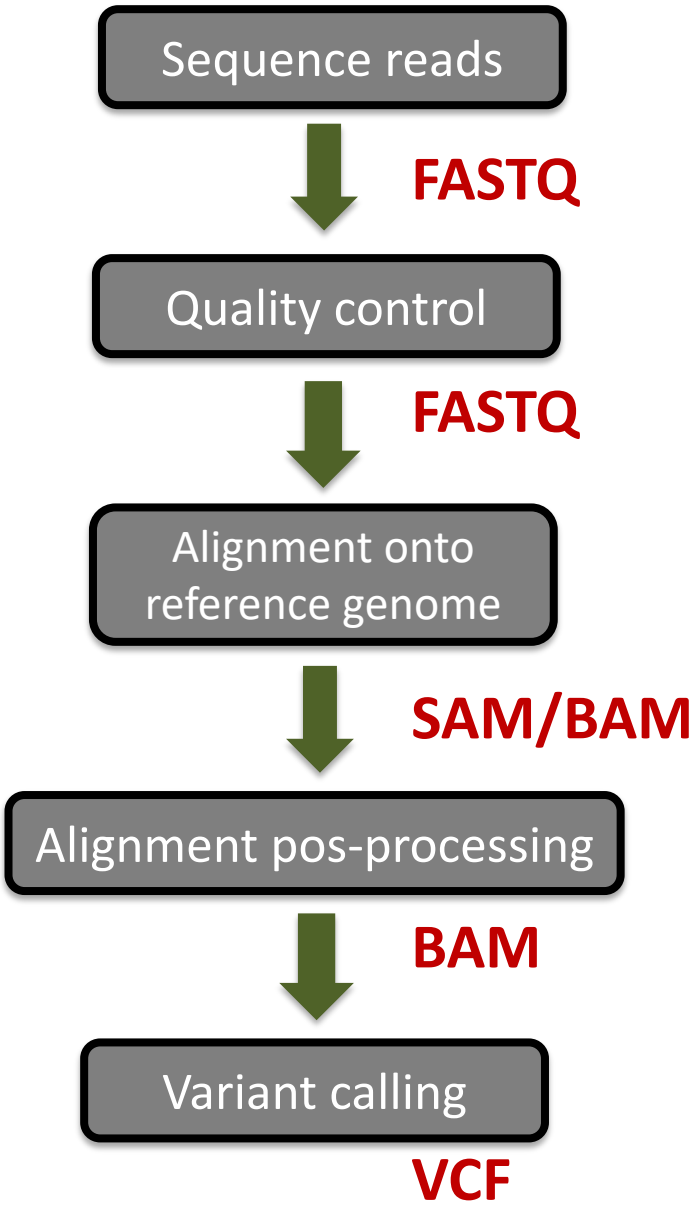
## A beginners guide to SNP calling from high-throughput DNA-sequencing data

**André Altmann · Peter Weber · Daniel Bader · Michael Preuß · Elisabeth B. Binder · Bertram Müller-Myhsok**

Received: 31 January 2012 / Accepted: 31 July 2012 / Published online: 11 August 2012

# From Fastq to VCF file

MICHIGAN STATE UNIVERSITY



# 01. QC of FASTQ

MICHIGAN STATE UNIVERSITY

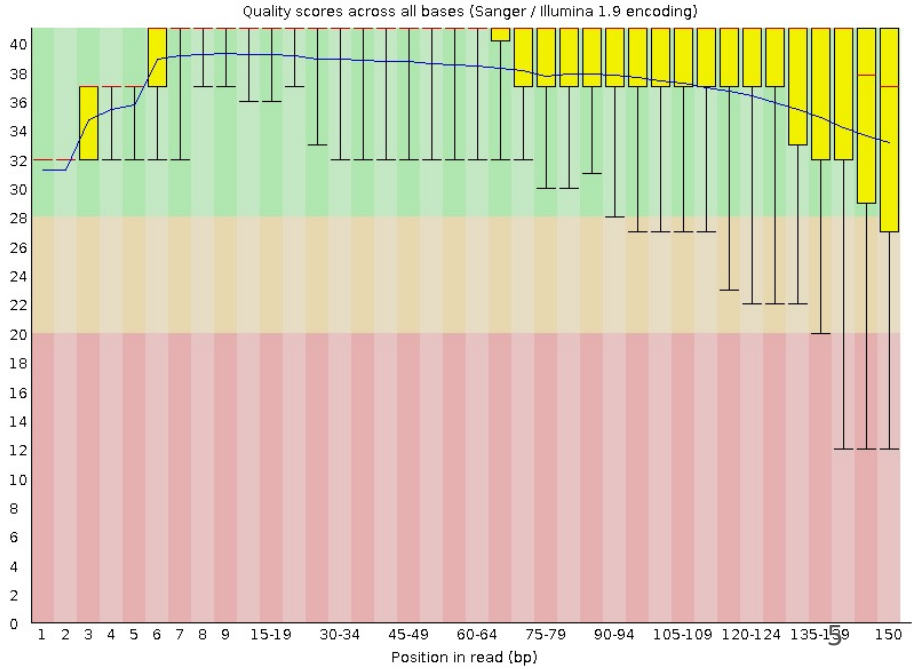
- **Initial visual appraisal**

- Are all files present?
- Are the files of expected size?
- Check the names of FASTQ files

- **Other quality controls**

- Phred score distribution
- GC content

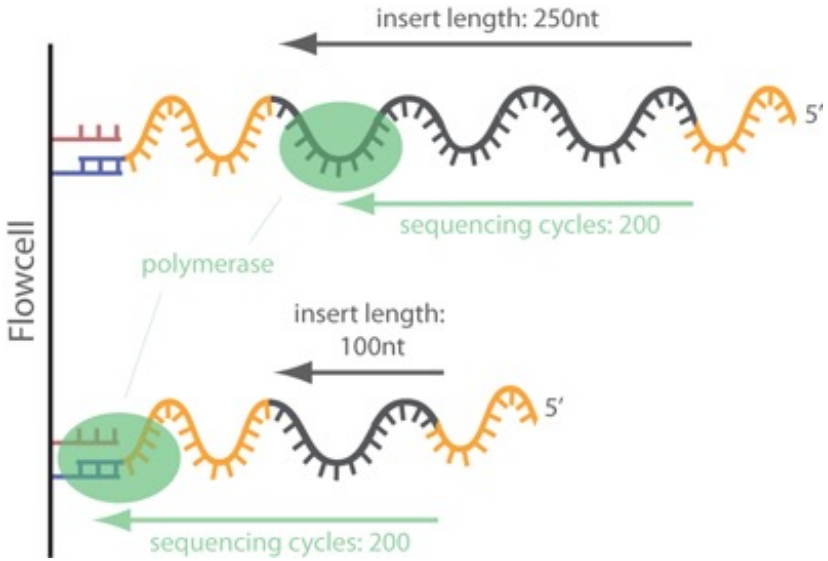
- **FASTQC**



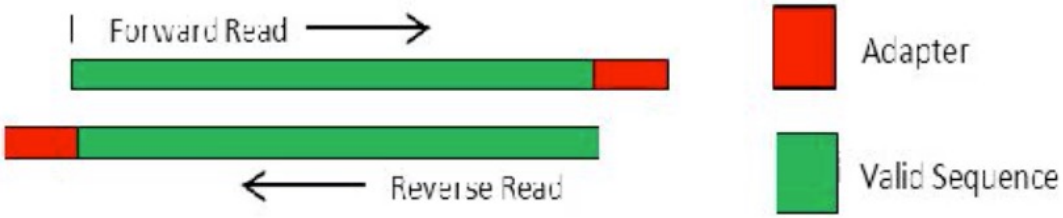
# 02. Adapter trimming

MICHIGAN STATE UNIVERSITY

- **Trimmomatic** is one option



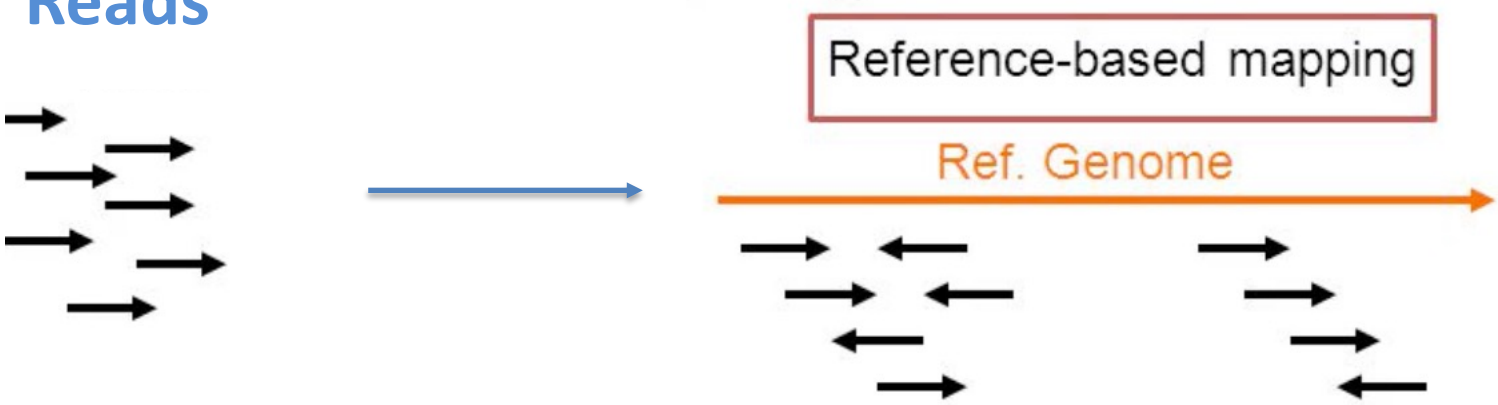

- Clip Illumina **adapters**:



# 03. Read mapping



**Reads**



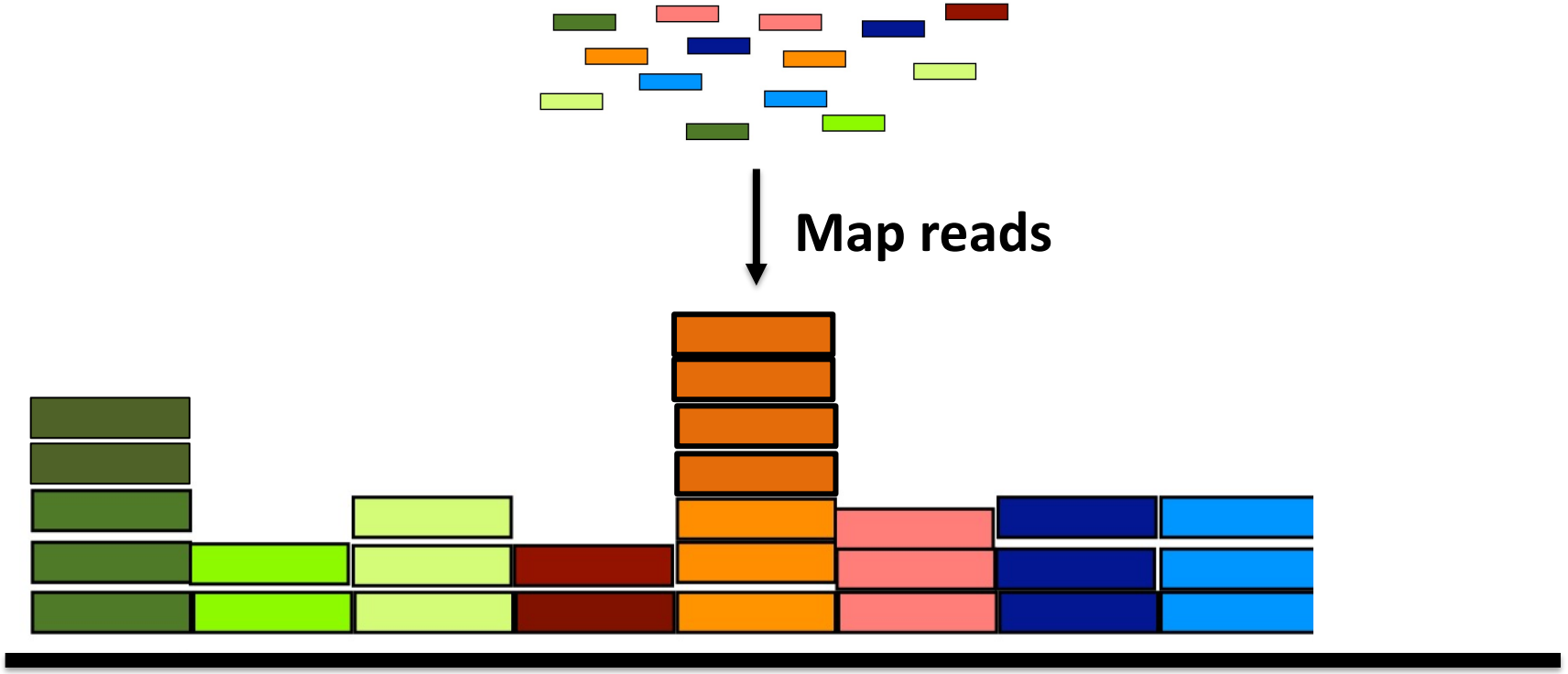

**FASTQ**

**SAM = text file**

**BAM = binary file**

# 03. Read mapping

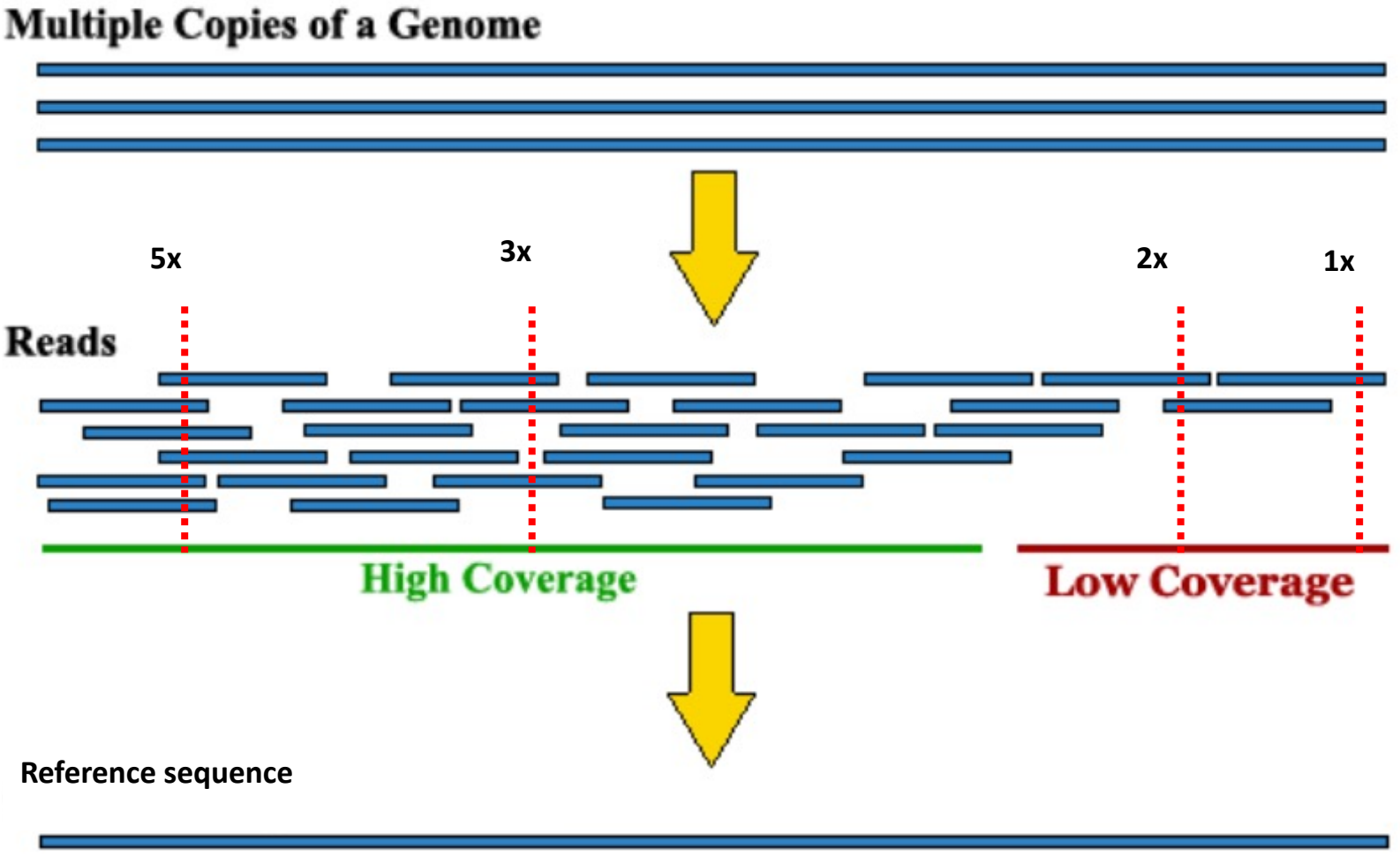





Seesi et al (2016). Genomics-Guided Immunotherapy of Human Epithelial Ovarian Cancer. In Rodriguez-Oquendo (Eds.) **Translational Cardiometabolic Genomic Medicine.** pp. 237-250.

# 03. Read mapping: coverage depth

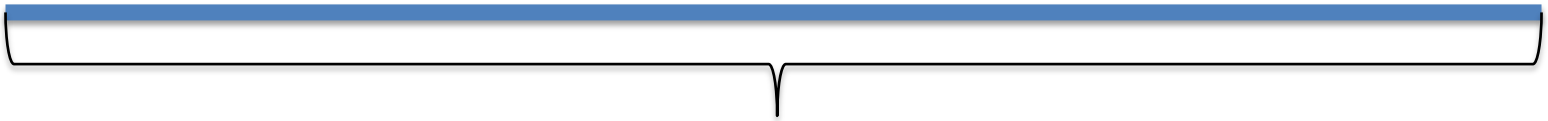




# 03. Read mapping: coverage depth

MICHIGAN STATE UNIVERSITY

**Reference sequence**



G = 8’’

length = 1’’

N = 28 reads

$$Cov\ Depth = \frac{length * N}{G} = \frac{1 * 28}{8} = 3.5\text{x}$$

# 03. Read mapping: coverage depth

MICHIGAN STATE UNIVERSITY

**ALIGNERS**

**WGS**

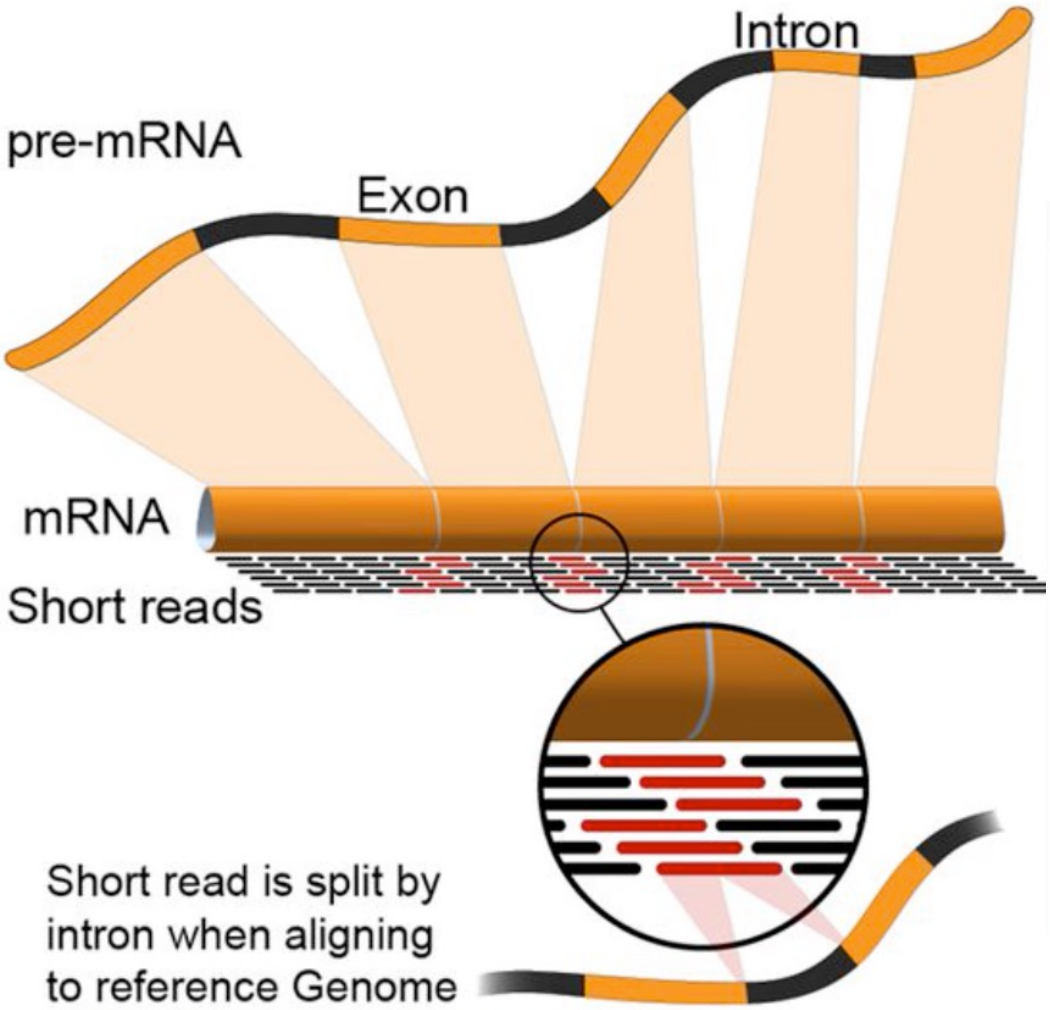
- BWA
- BOWTIE
- BOWTIE2

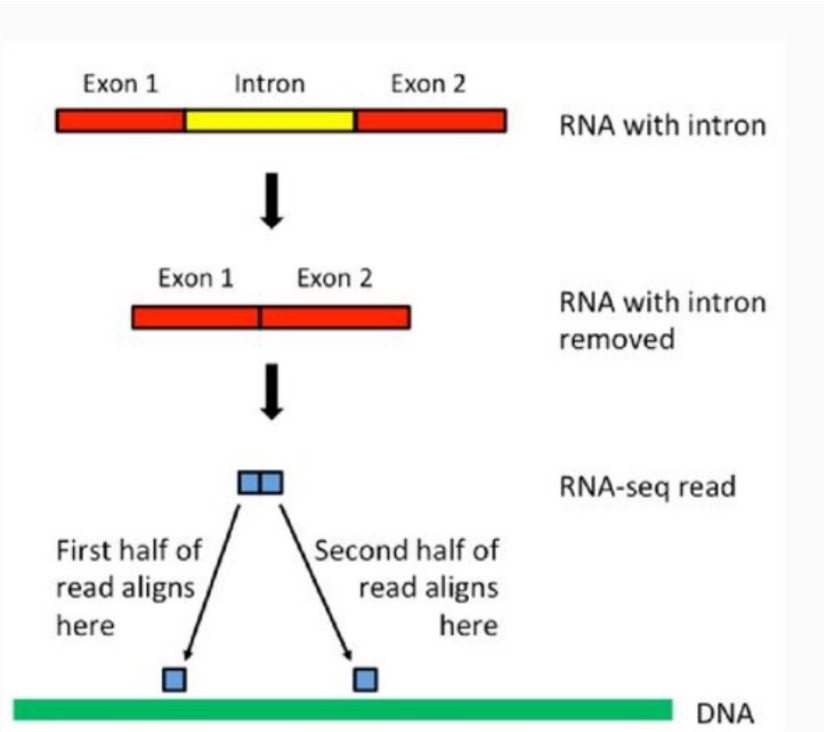
**RNAseq**

- TOPHAT
- HISAT2
- STAR

# 03. Read mapping: splice aware aligners





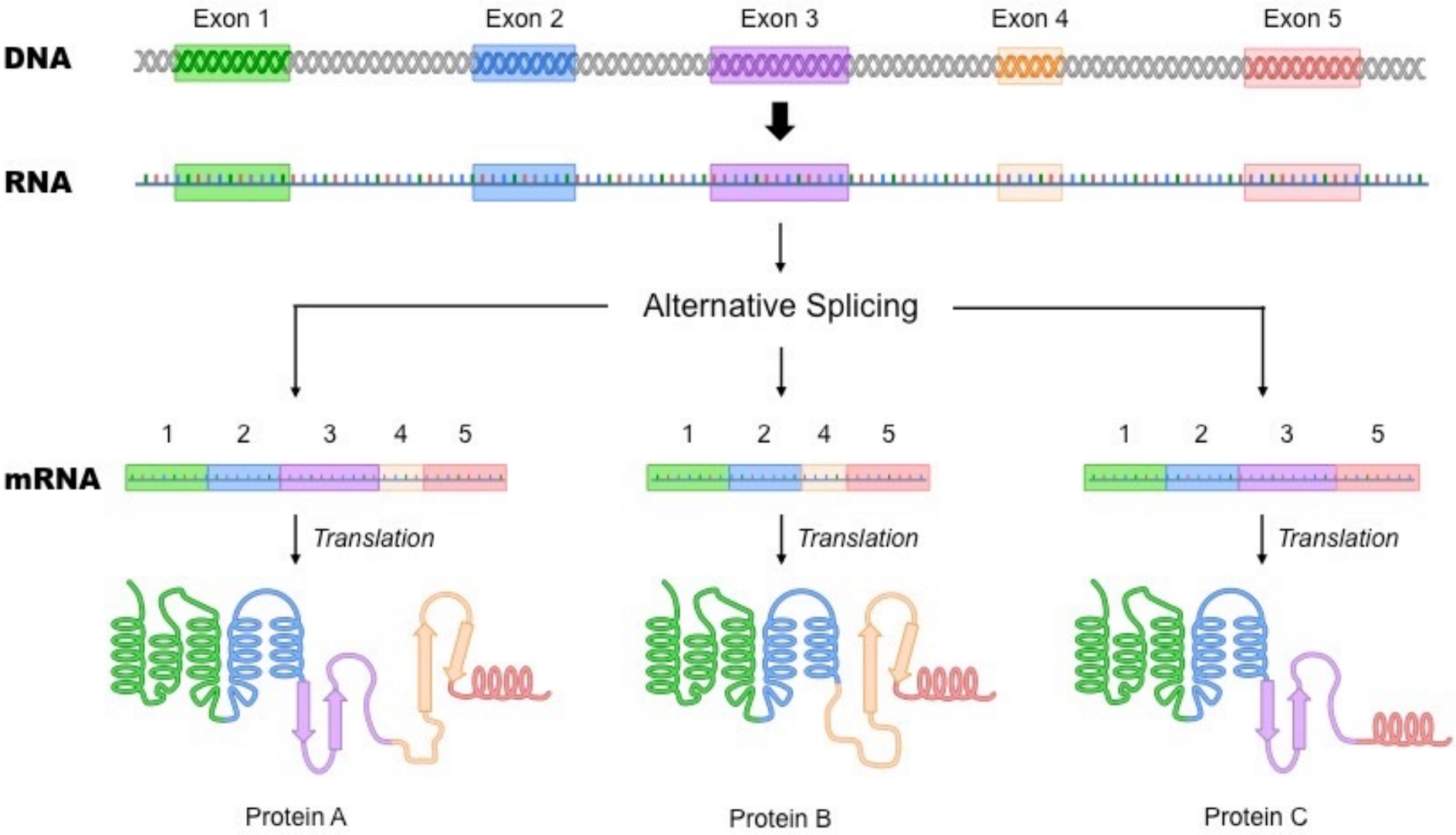

http://en.wikipedia.org/wiki/File:RNA-Seq-alignment.png

***"Systematic evaluation of spliced alignment programs for RNA-seq data"***

*Nature Methods, 2013*

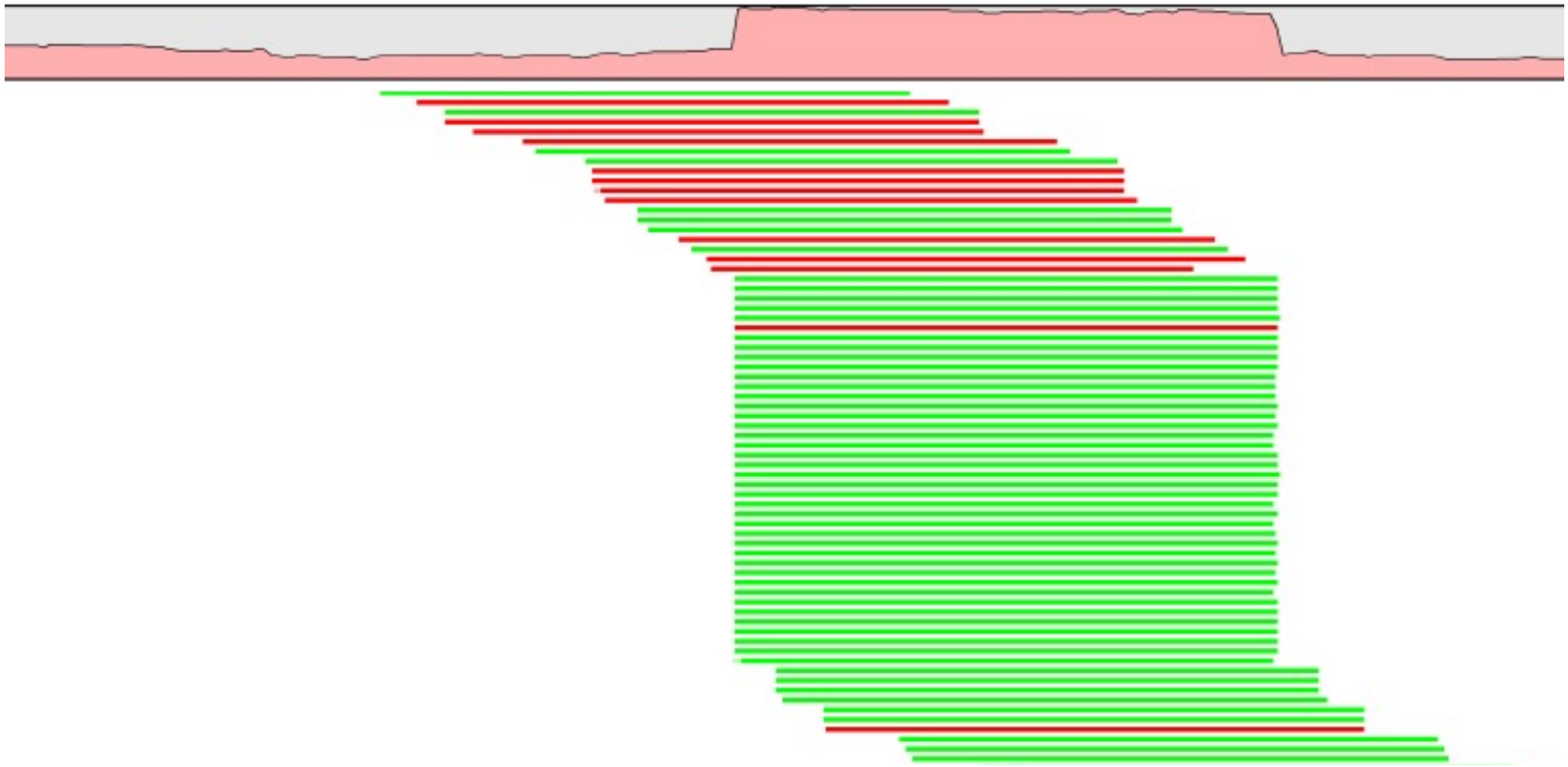
# 03. Read mapping: RNA alternative splicing

MICHIGAN STATE UNIVERSITY



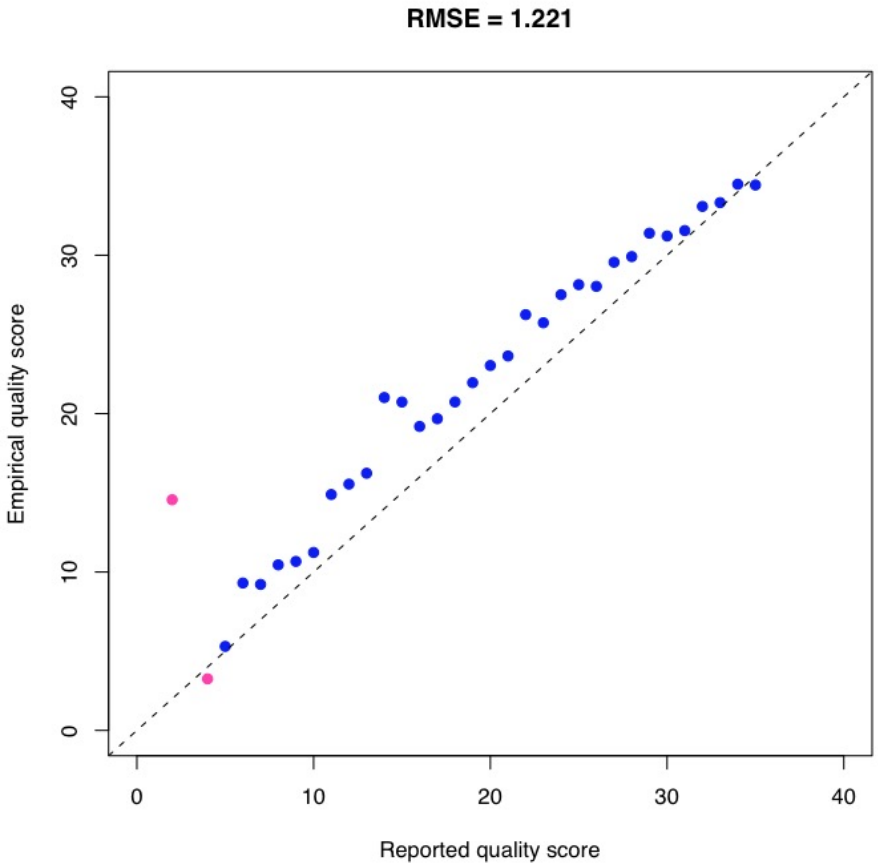
## PCR amplification bias

- Some reads are better amplified than others → bias!
- Keep only one (with highest mapping quality)

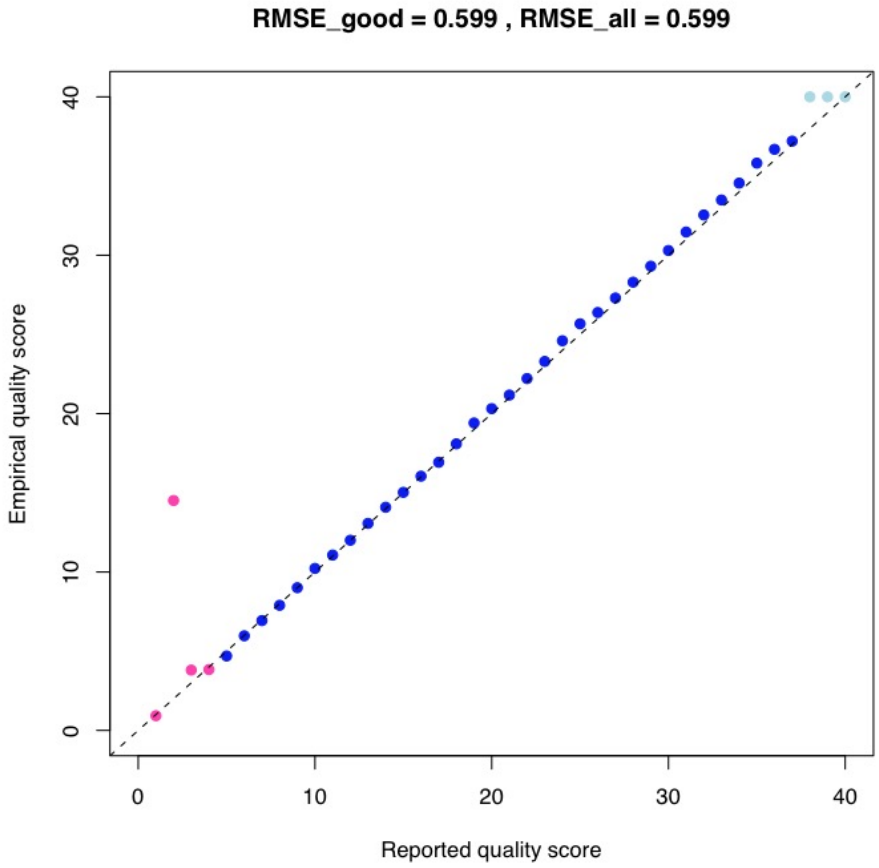


# 05. BQSR – Reported vs Empirical Quality

MICHIGAN STATE UNIVERSITY




Original Data




After GATK Recalibration

# 06. Variant Calling



- SAMTOOLS mpileup + BCFTOOLS call (deprecated)
- BCFTOOLS mpileup + BCFTOOLS call

- FREEBAYES

- GATK

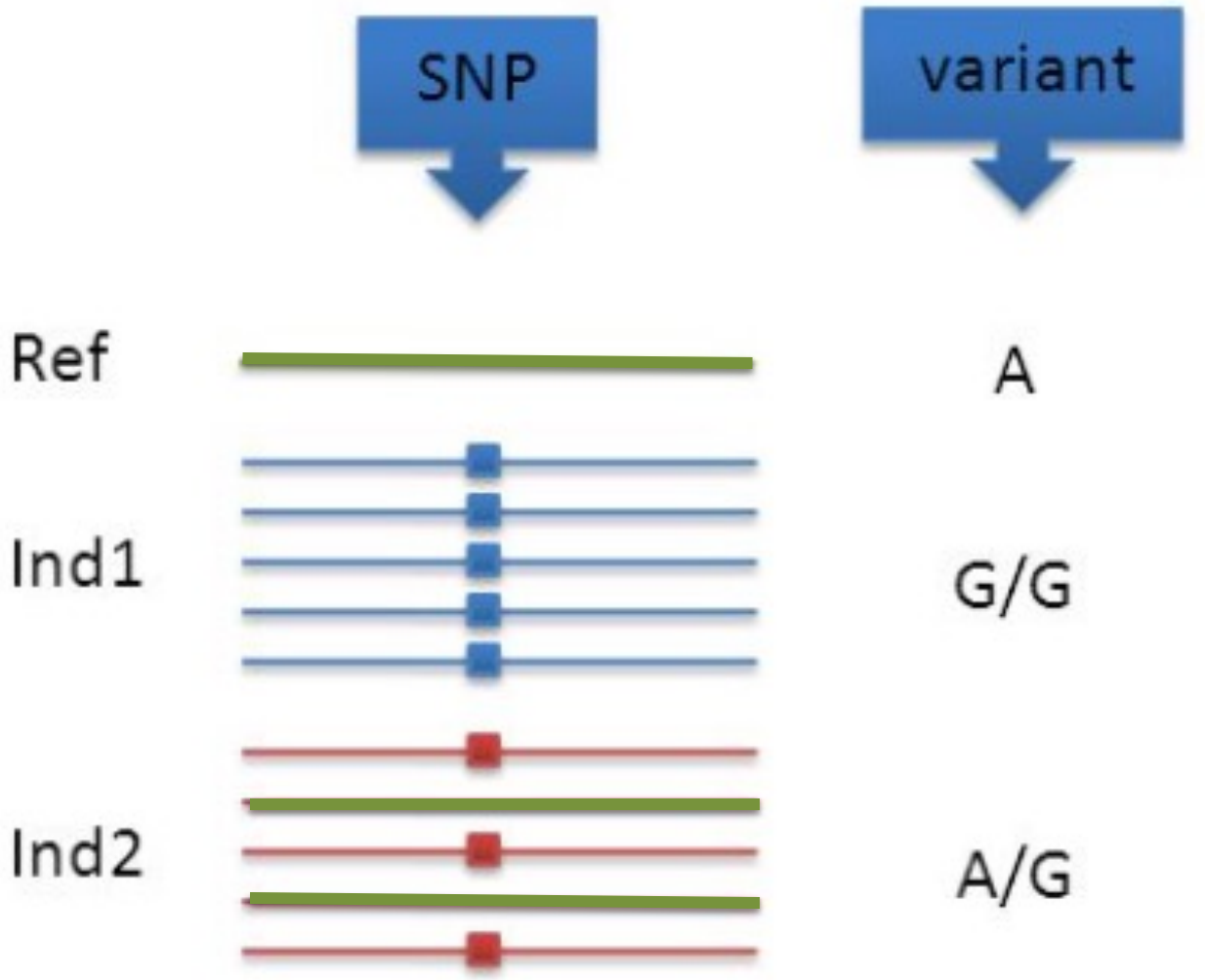
- VARSCAN2

- DeepVariant

... others

# 06. Variant Calling





MICHIGAN STATE UNIVERSITY

## Coverage is important!

Assuming a true heterozygous variant C/T (expected 50% C and 50% T)

**Reference**

... GTGC**C**AGGACCAGATCG ...

GTGC**C**AGGACCAGATCG
GTGC**C**AGGACCAGATCG
GTGC**C**AGGACCAGATCG
GTGC**C**AGGACCAGATCG
GTGC**C**AGGACCAGATCG

@1X 50% chance not to sample the T-allele
@2X 25%
@3X 12.5%
@4X 6.25%
@5X 3.125%
...
@15X 0.003%

## Coverage is important!

Assuming a true heterozygous variant C/T (expected 50% C and 50% T)

**Increase the converage will increase detection of all alleles**
(Assuming na unbiased approach which is not necessarily the case: PCR amplification errors, reference errors, mapping errors.

# Hands on



- Step-by-step from FASTQ to VCF
- Specific bioinformatic programs
- Linux language