BIO214 Lecture 11

Bioinformatics-II

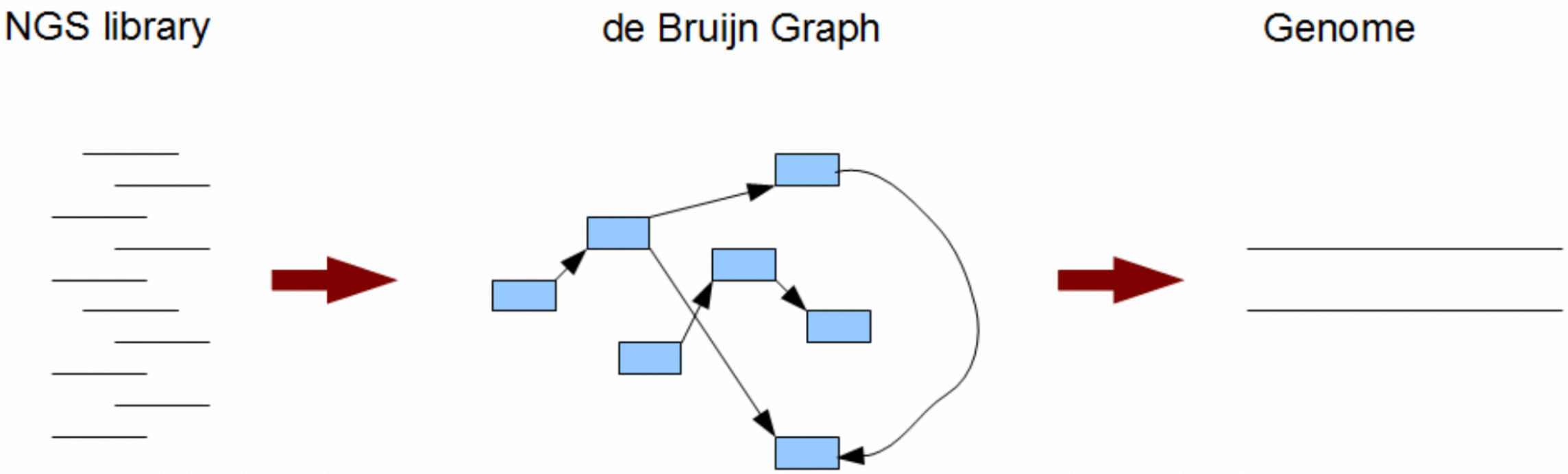
Genome Assembly & variant analysis

Zhen Wei; 2023-Feb-17

Outline

- Efficient assembly of short reads with De bruijn graph
- Detection of genetic variants
- GWAS
- eQTL

An efficient algorithm for short reads based genome assembly



De Bruijn graph-based genome assembly algorithm:

- Step 1: Short reads broken into small pieces (k-mers) and de Bruijn graph constructed.
- Step 2: Genome derived from de Bruijn graph by finding the longest possible path (Eulerian walks).

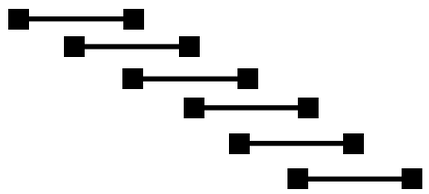
De Brujin Graph

Genome： AAABBBBA

https://www.youtube.com/watch?v=TNYZZKrjCSk

De Brujin Graph

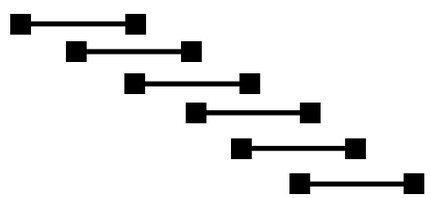
Genome： AAABBBBA



3-mers： AAA AAB ABB BBB BBB BBA

De Brujin Graph

Genome： AAABBBBA

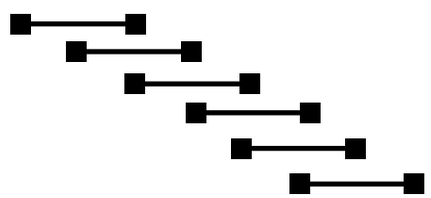


3-mers： AAA AAB ABB BBB BBB BBA

L-R 3-1mers： AA, AA

De Brujin Graph

Genome： AAABBBBA

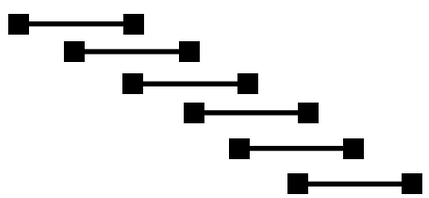


3-mers： AAA AAB ABB BBB BBB BBA

L-R 3-1mers： AA, AA AA, AB AB, BB BB, BB BB, BB BB, BA

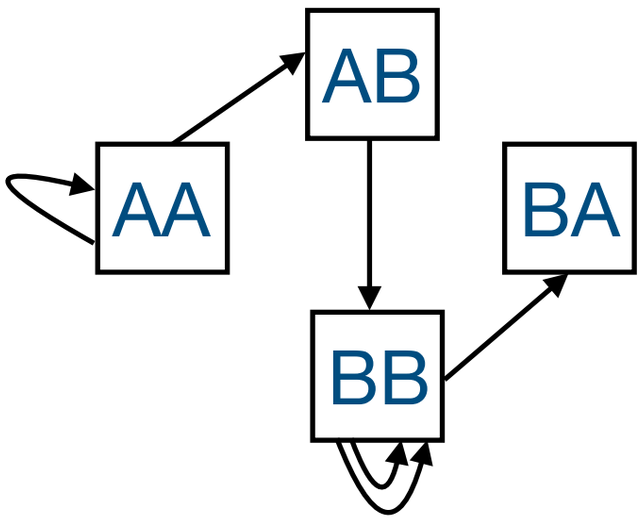
De Brujin Graph

Genome： AAABBBBA



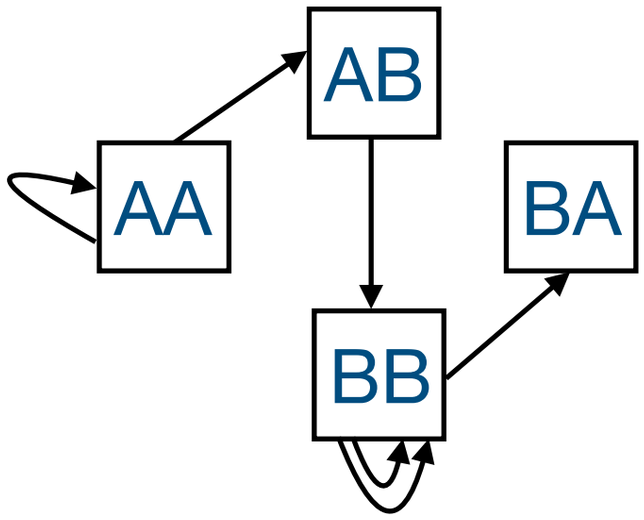
3-mers： AAA AAB ABB BBB BBB BBA

L-R 3-1mers： AA, AA AA, AB AB, BB BB, BB BB, BB BB, BA



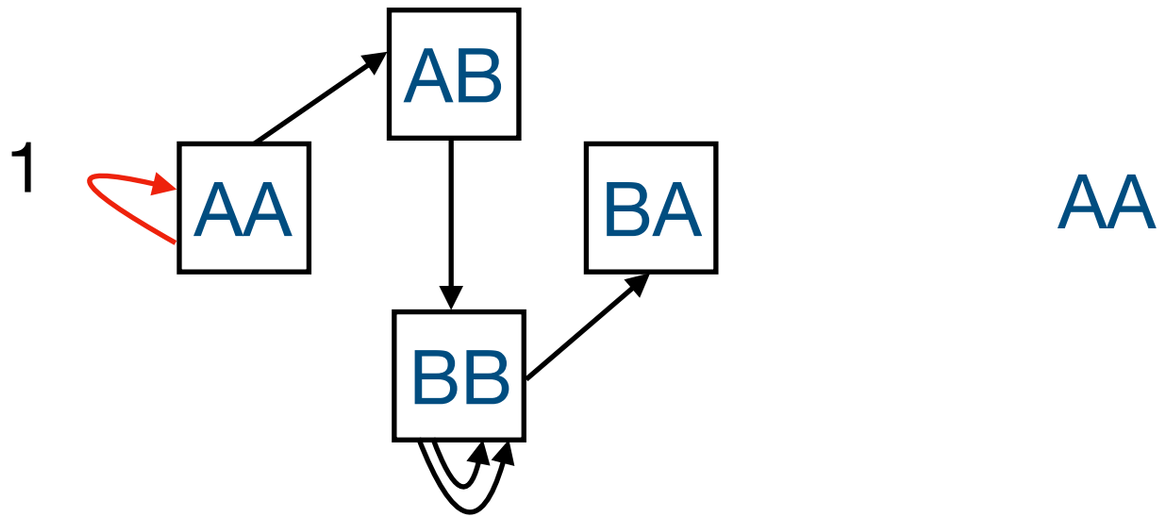
- One edge per k-mer
- One node per distinct k-1 mer

Eulerian walk



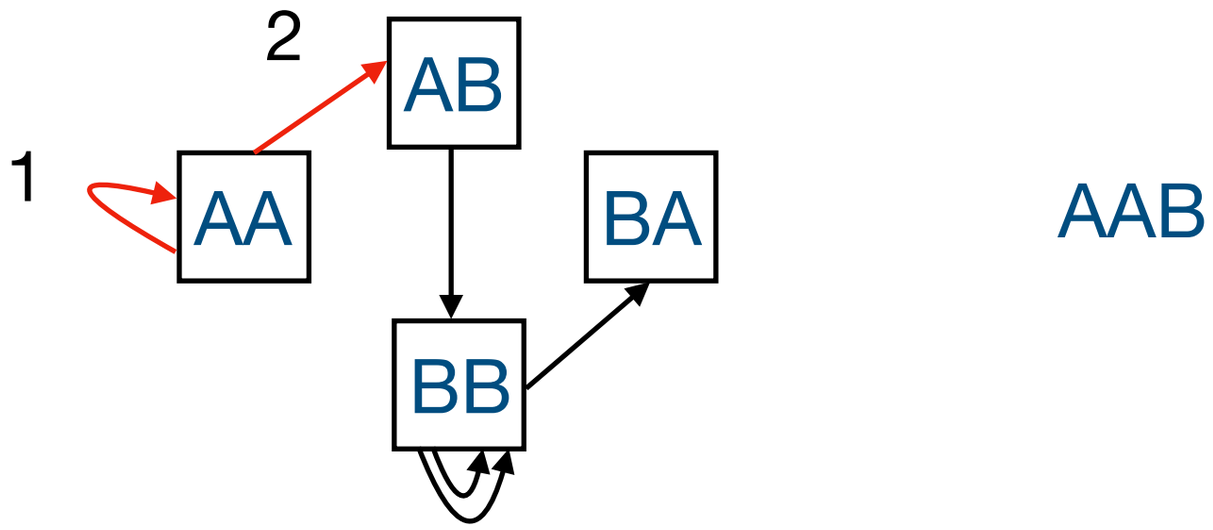
- Walk crossing each edge exactly once gives a reconstruction of the genome.

Eulerian walk



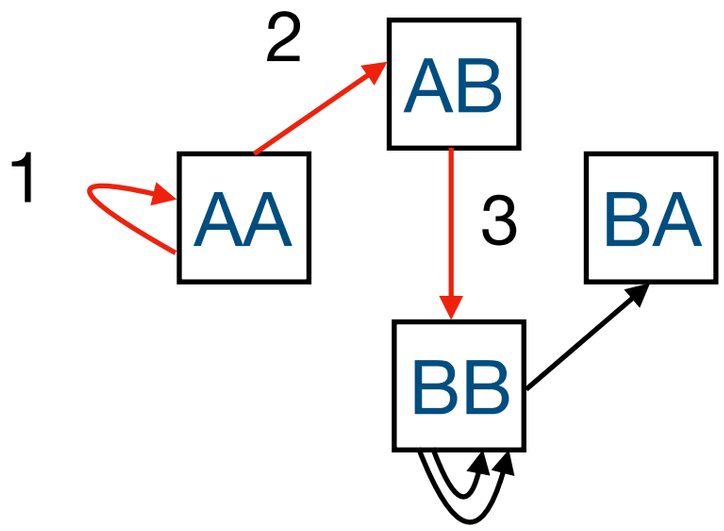
- Walk crossing each edge exactly once gives a reconstruction of the genome.

Eulerian walk



- Walk crossing each edge exactly once gives a reconstruction of the genome.

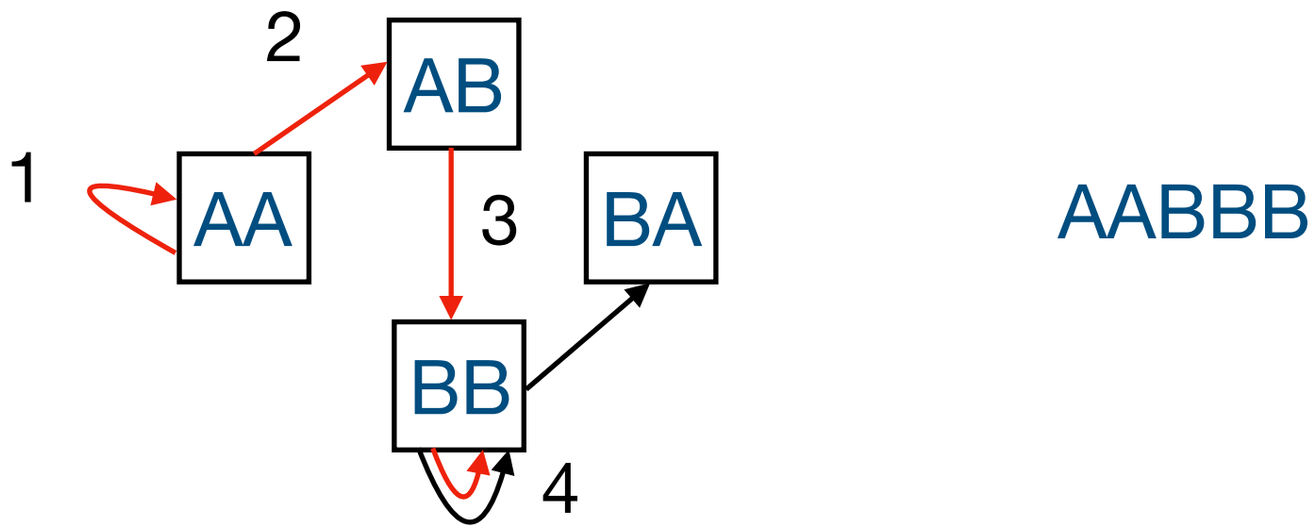
Eulerian walk



AABB

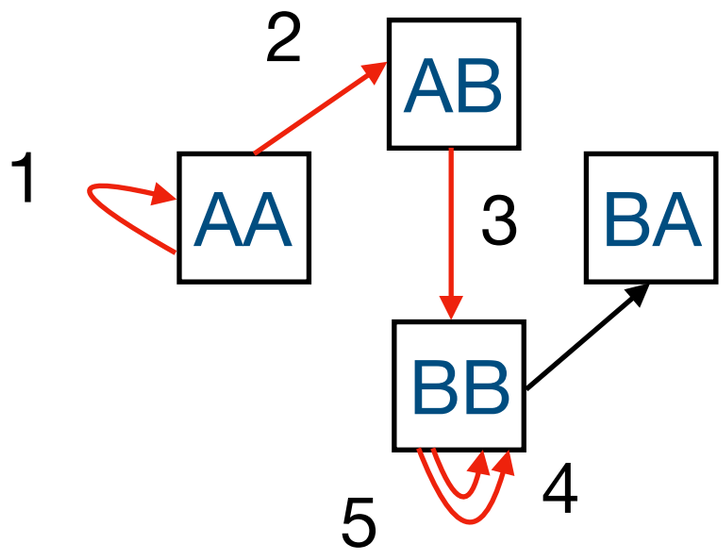
- Walk crossing each edge exactly once gives a reconstruction of the genome.

Eulerian walk



- Walk crossing each edge exactly once gives a reconstruction of the genome.

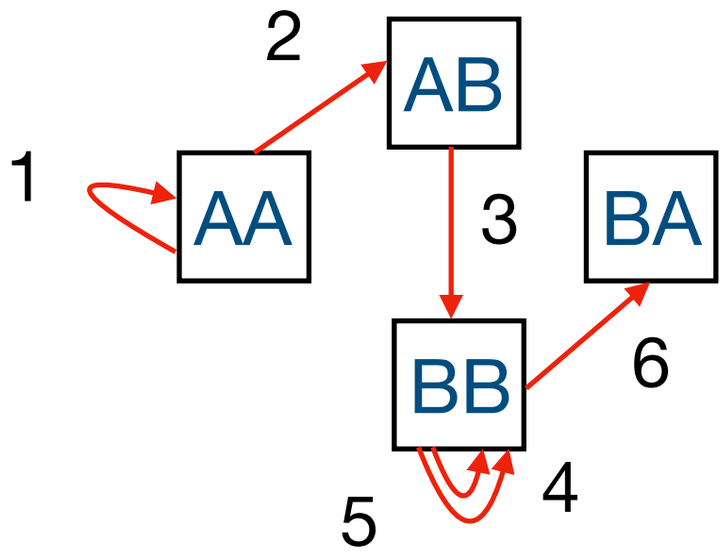
Eulerian walk



AABBBB

- Walk crossing each edge exactly once gives a reconstruction of the genome.

Eulerian walk



AABBBBA

- Walk crossing each edge exactly once gives a reconstruction of the genome.

SPAdes

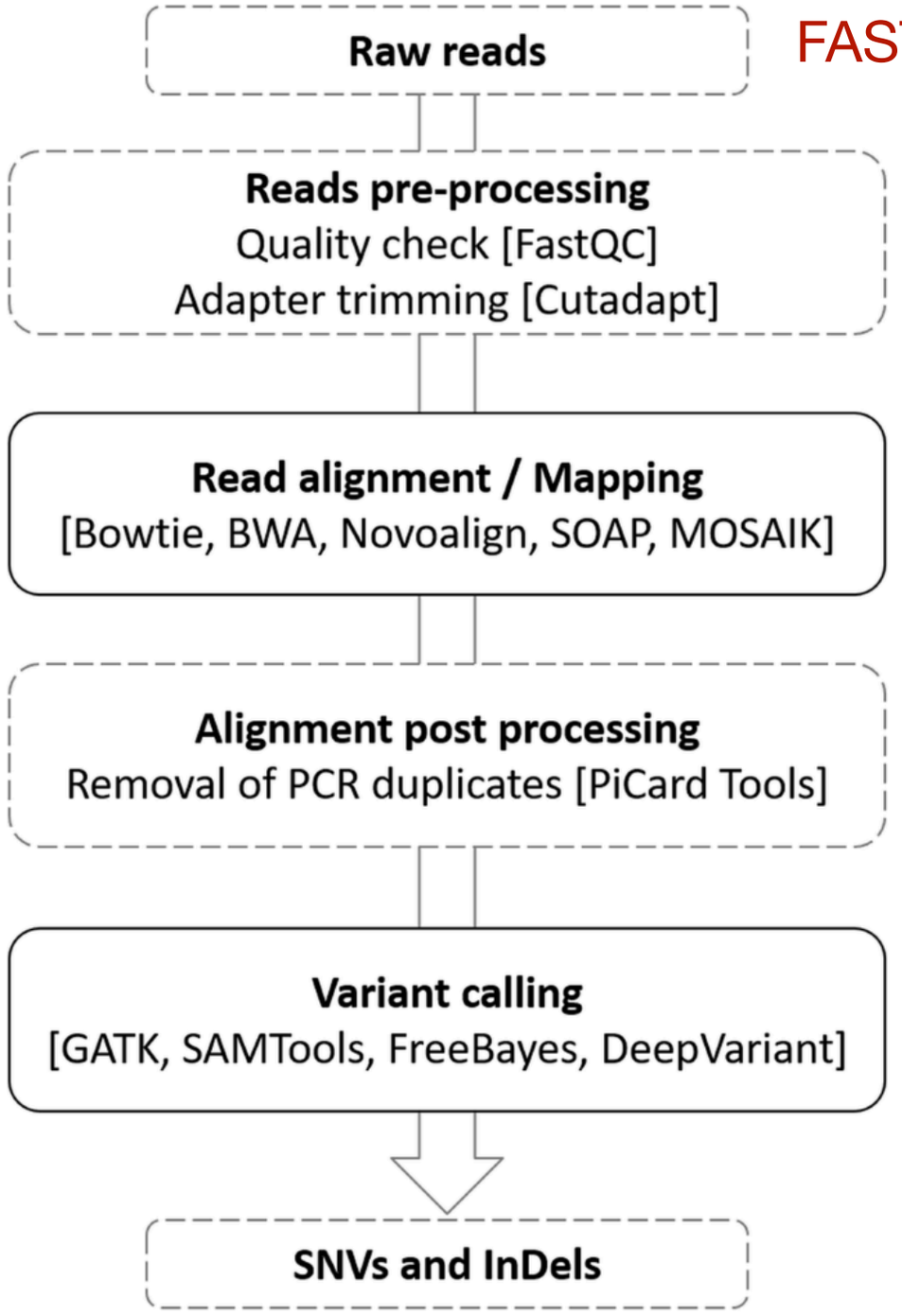
Assembly	NG50	# contigs	Largest	Total length	MA	MM	IND	GF (%)	# genes
Single-cell *E. coli*									
A5	14399	745	101584	4441145	3	11.92	0.19	89.867	3443
ABySS	68534	**179**	178720	4345617	6	3.49	0.83	88.265	3704
CLC	32506	503	113285	4656964	**1**	5.54	1.00	92.286	3767
EULER-SR	26662	429	140518	4248713	12	9.98	20.17	84.846	3410
Ray	45448	361	210820	4379139	16	5.29	1.24	88.345	3634
SOAPdenovo	1540	1166	51517	2958144	**1**	**1.49**	**0.11**	57.668	1766
Velvet	22648	261	132865	3501984	2	2.19	1.17	73.761	3079
E+V-SC	32051	344	132865	4540286	2	2.26	0.70	91.727	3767
IDBA-UD contigs	98306	244	**284464**	**4814043**	3	4.37	0.23	**95.158**	**4041**
IDBA-UD scaffolds	109057	229	**284464**	**4813609**	3	4.42	0.75	**95.145**	**4046**
SPAdes 3.12 contigs	105885	231	268283	**4795250**	3	2.02	0.30	**94.853**	**4028**
SPAdes 3.12 scaffolds	**117600**	214	**285212**	**4800301**	3	2.41	0.61	**94.886**	**4030**

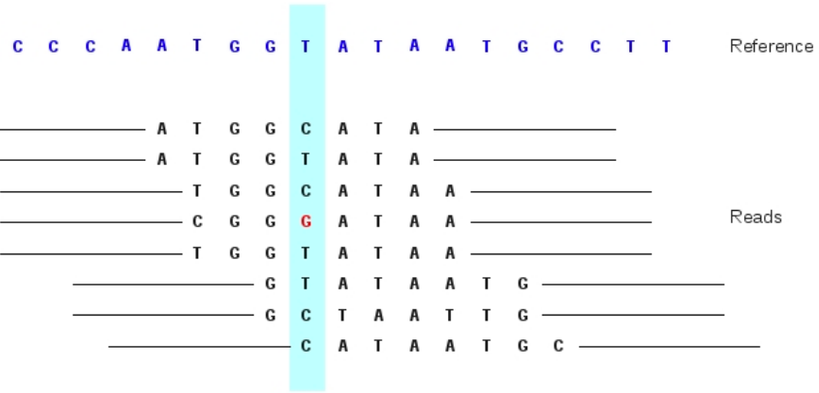
https://cab.spbu.ru/software/spades/#examples

- SPAdes is a de-brujin graph based genome assembler.
- By default SPAdes assembles using kmers of lengths 21, 33, and 55 and chooses the assembly with the best N50 score.
- N50 can be understood as the median contig length in the assembly.

Detection of genetic variants

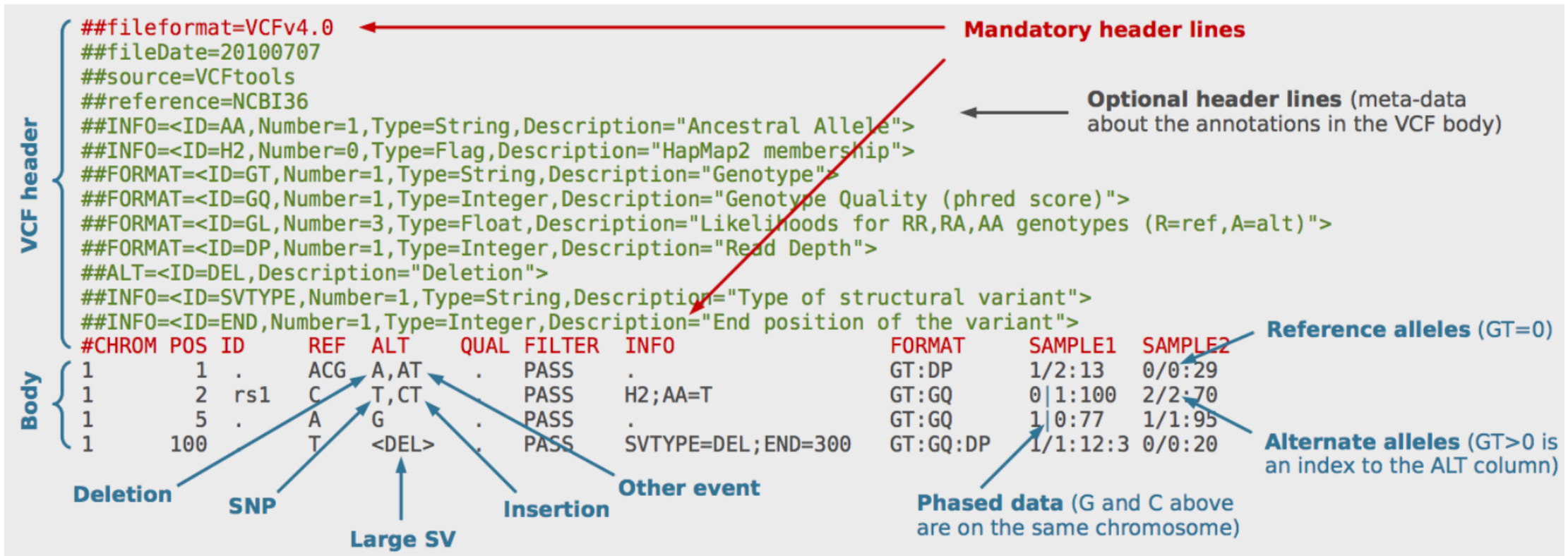
Variant calling pipeline





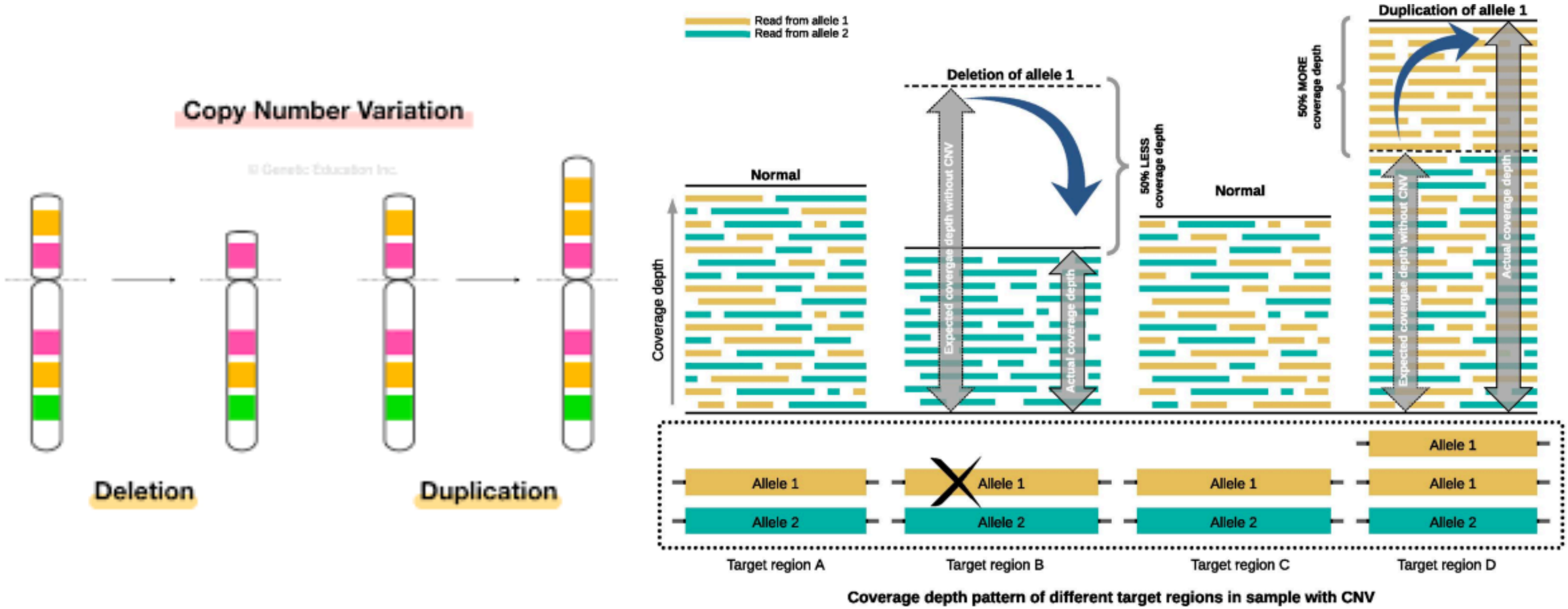
Kumaran, M., Subramanian, U., & Devarajan, B. (2019). Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data. BMC bioinformatics, 20(1), 1-11.

VCF format



- The **Variant Call Format (VCF)** is the standard file format for storing genetic variation and was developed as part of the 1000 Genomes Project.

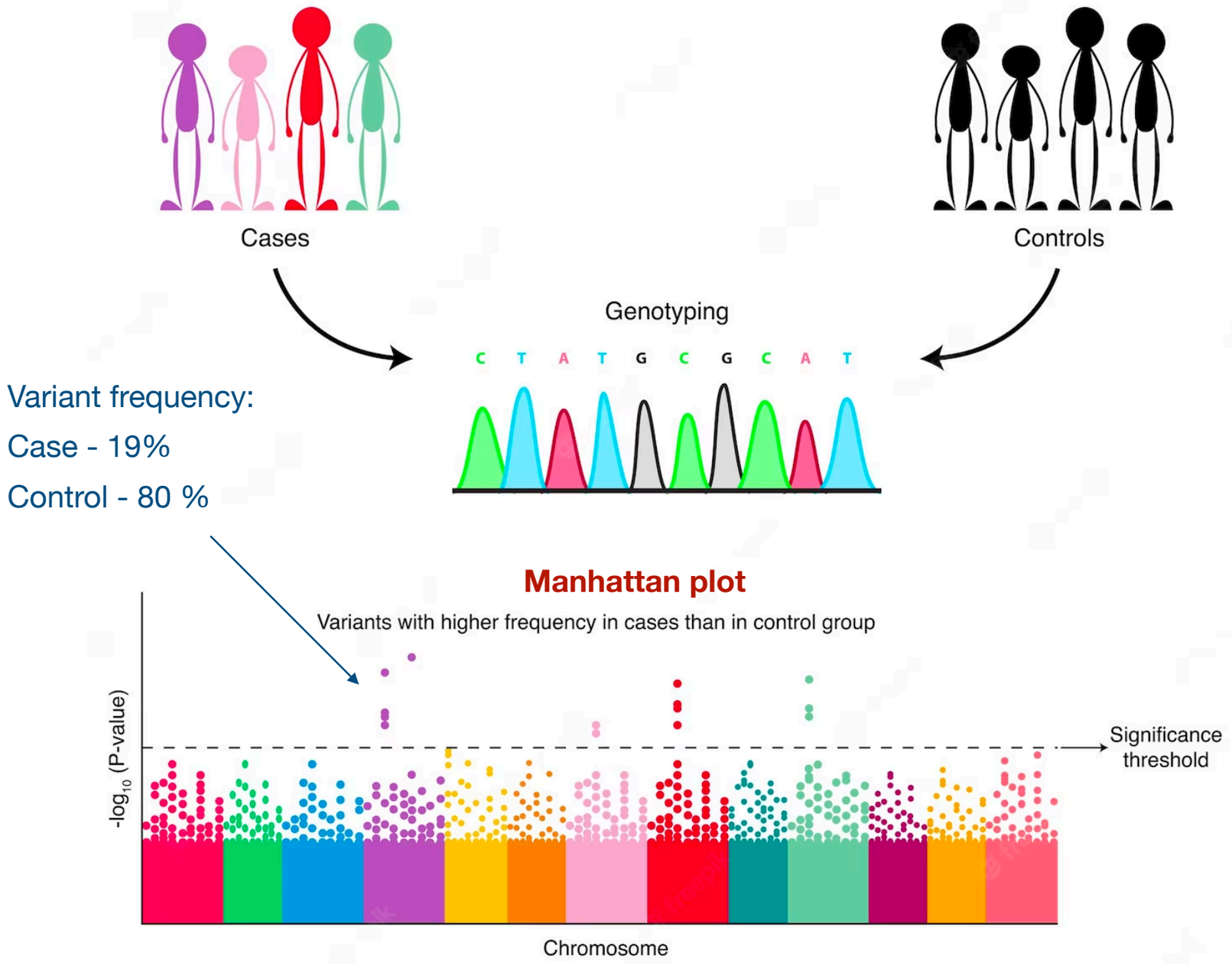
Copy number variation detection



- CNVs are regions of the genome with variable number of copies.
- DNA-Seq can detect CNVs by analyzing the number of sequencing reads that map to a genomic region.
- Higher reads suggest a duplication, while lower reads suggest a deletion.
- CNV detection requires careful normalization and calibration, as read depth can be affected by factors such as GC content, mappability, and sequencing bias.

Genome-wide association analysis

GWAS at a glance

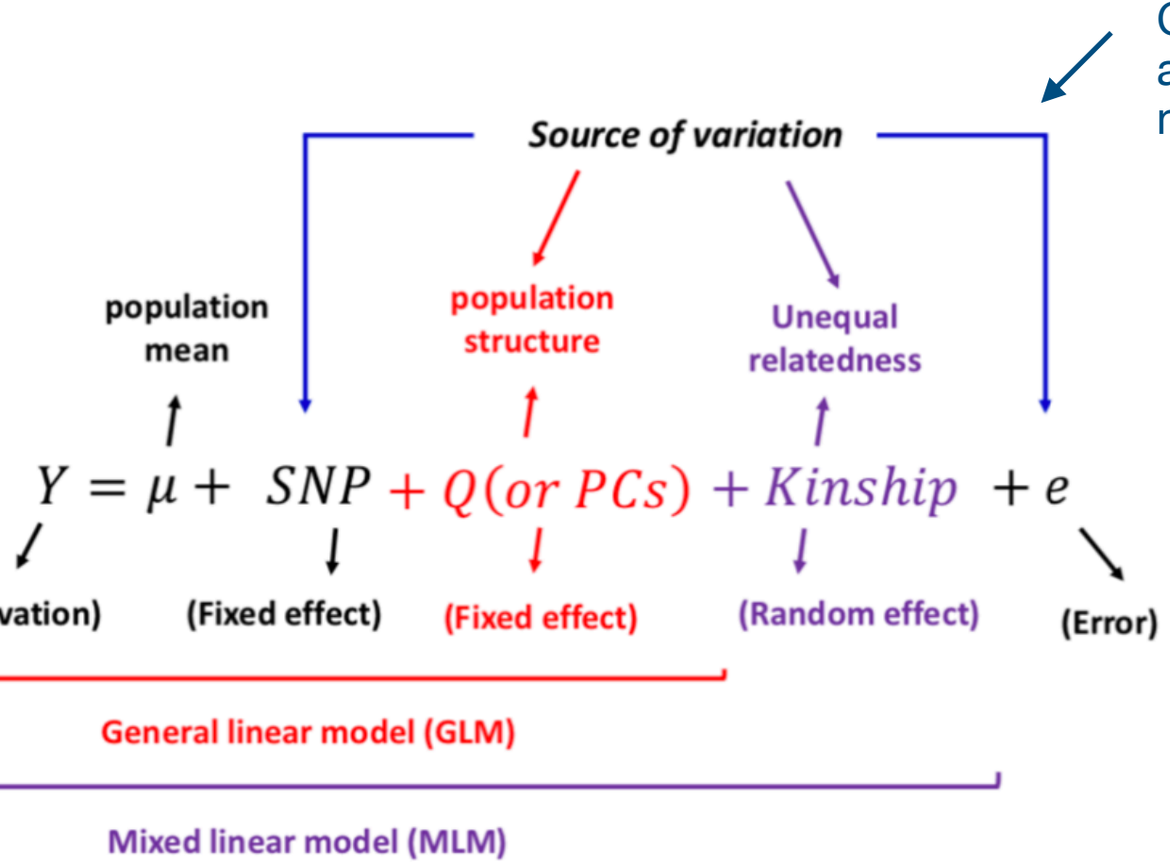


GWAS aim to identify if any of the millions of SNPs are associated with a specific disease (e.g. Cancer) or trait (e.g. Hight / intelligence).

GWAS test using linear regression

Account for confounding variables

Can lead to spurious associations in GWAS if not accounted for.



Population Structure:

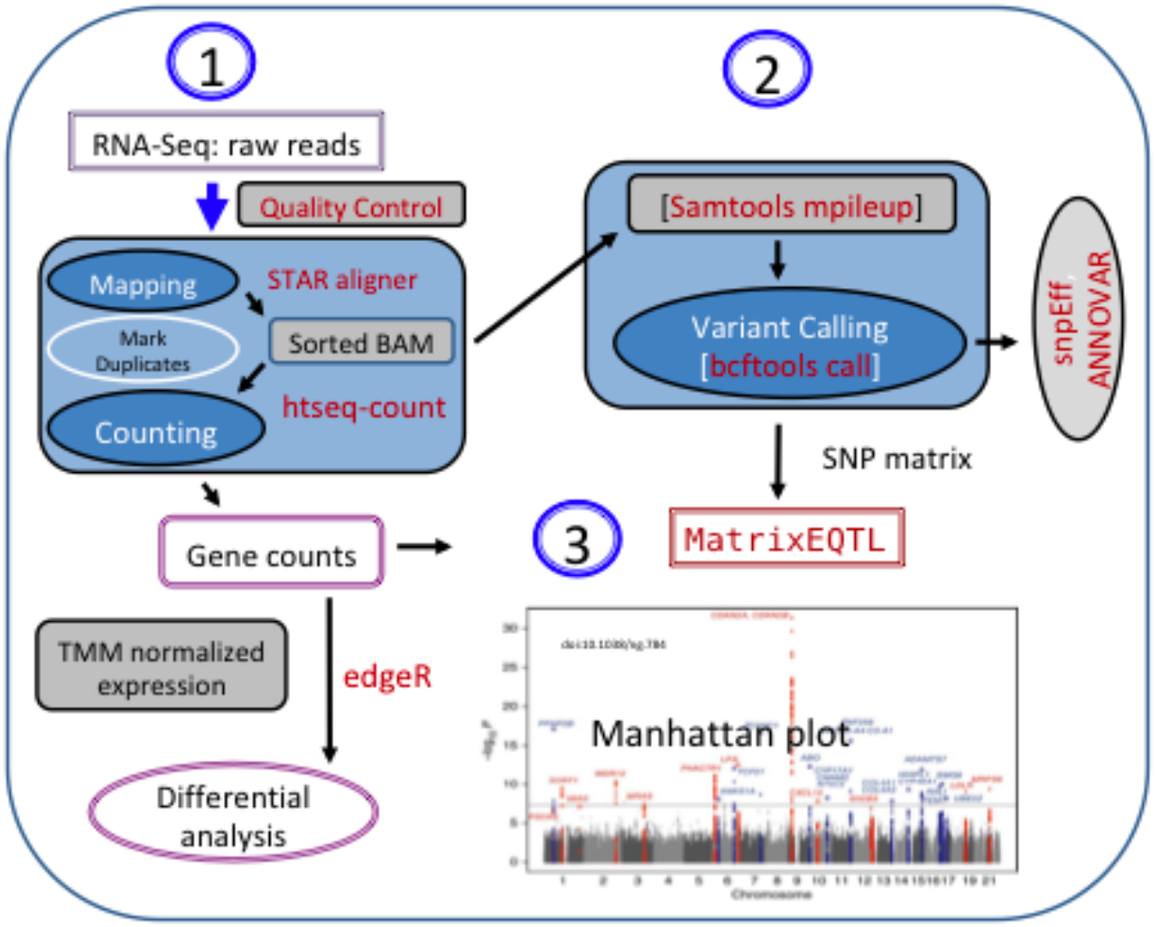
- Systematic differences in allele frequencies between subgroups in a population due to non-random mating between individuals.
- Can be estimated from data using statistical methods such as PCA.

Kinship:

- Describes the genetic relatedness between individuals in a population.
- Kinship matrix is often modeled as the covariance of the random effect term.

Expression quantitative trait loci

General workflow



https://adinasarapu.github.io/posts/2017/12/blog-post-eqtl/

- An expression quantitative trait locus (**eQTL**) is a genetic locus that affects gene expression.
- eQTL mapping studies use RNA-Seq data to identify eQTLs.
- Variants are called either from DNA-Seq / RNA-Seq; expression levels are quantified via the regular pipeline, and differential analysis is performed between genotypes.

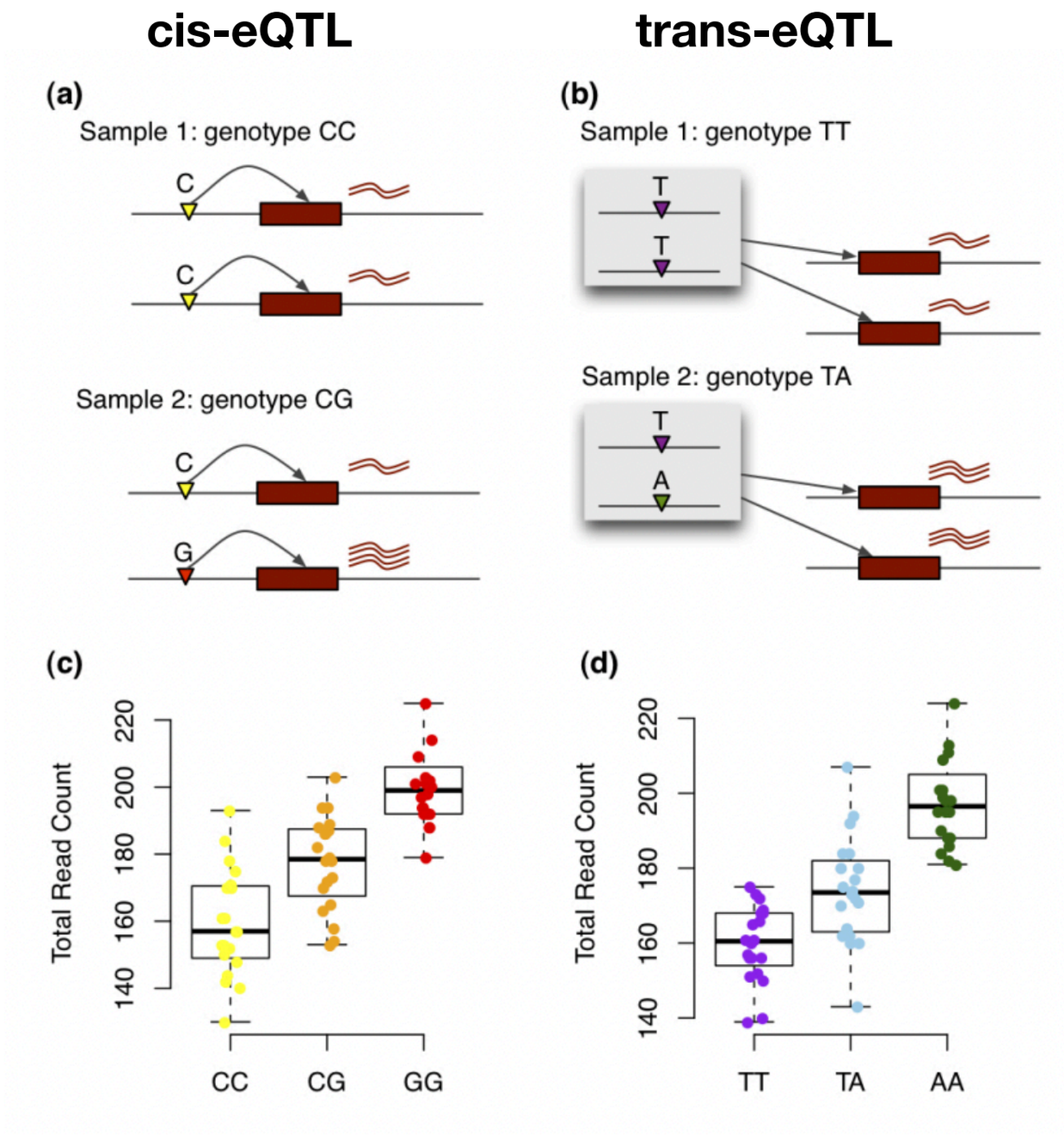
Cis & trans QTL

cis-eQTL:

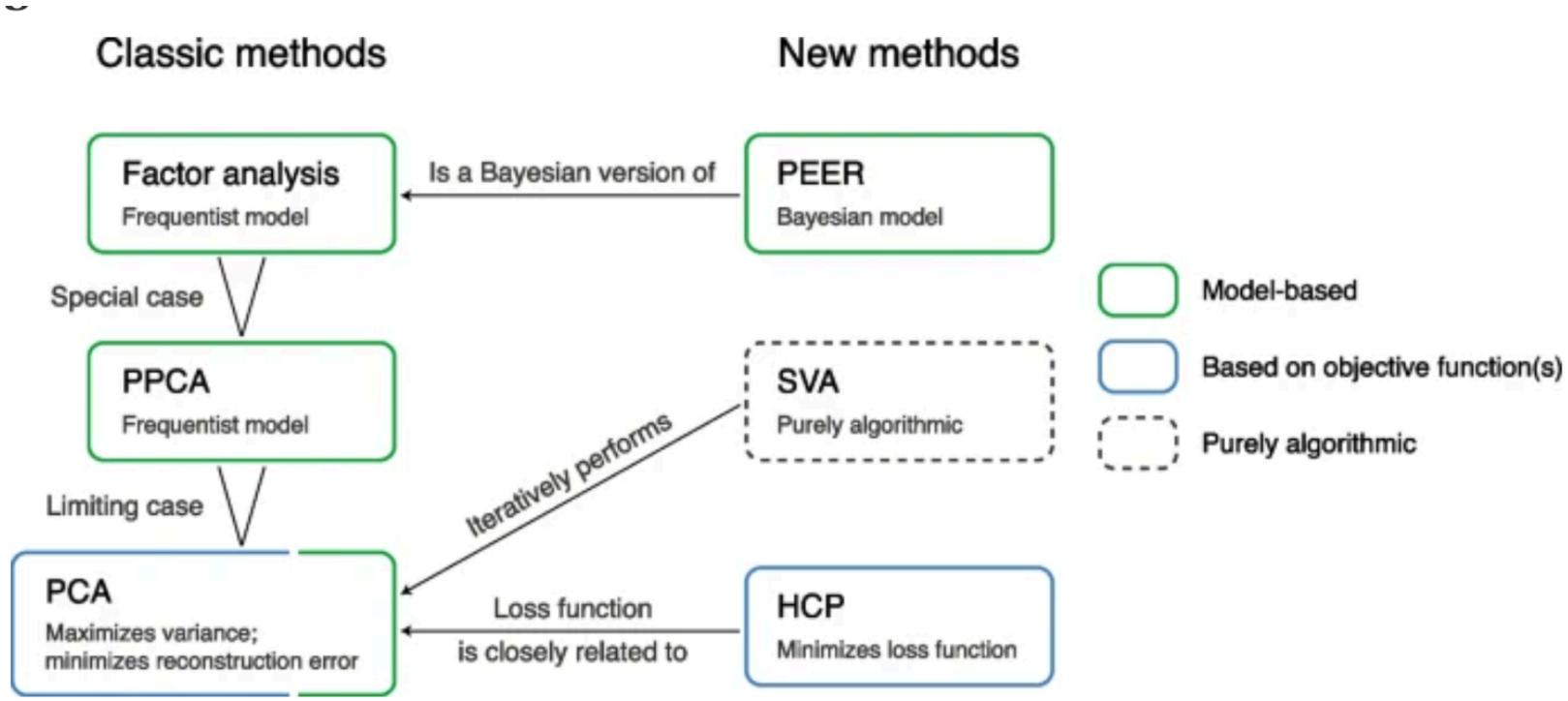
- Variants affecting expression of local genes.
- Found in promoter and gene body of the effected genes.

trans-eQTL:

- Variants affecting expression of distal genes.
- Found in other regions.



Accounting for hidden batches



- P values of eQTL association are calculated from the linear regression tests with the following regression equation.
- expression = $\alpha + \sum_k \beta_k \cdot \text{covariate}_k + \gamma \cdot \text{genotype}$
- The *K* covariates are estimated by hidden variable inference methods such as PCA, SVA, PEER or HCP.

Zhou, Heather J., et al. "PCA outperforms popular hidden variable inference methods for molecular QTL mapping." Genome Biology 23.1 (2022): 1-17.

Accounting for hidden batches

PCA outperform the rests …

- P values of eQTL association are calculated from the linear regression tests with the following regression equation.
- expression = $\alpha + \sum_k \beta_k \cdot \text{covariate}_k + \gamma \cdot \text{genotype}$
- The K covariates are estimated by hidden variable inference methods such as PCA, SVA, PEER or HCP.

Zhou, Heather J., et al. "PCA outperforms popular hidden variable inference methods for molecular QTL mapping." Genome Biology 23.1 (2022): 1-17.

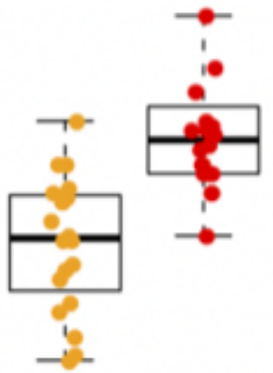
In silico mutation analysis

Infer the consequence of mutation by predictive modeling

Only correlation !

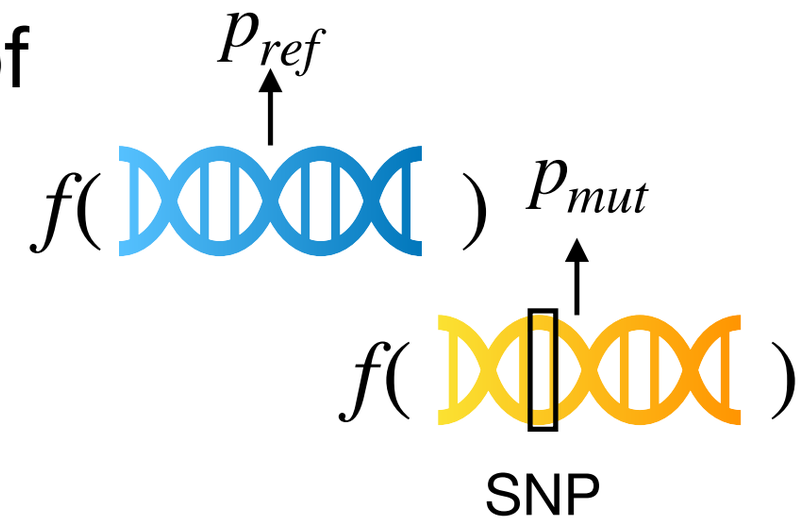
GWAS: Variants <——> Phenotype

eQTL: Variants <——> Gene expression



In-silico mutation: Variants <——> Change of Prob(epigenetic marker)

Can reveal causality !



In-silico mutation

- $f()$ is a sequence based predictive model, it accepts an input of a DNA string and output a probability of the string being a functional epigenetic modification or protein.

- Calculate the probabilities of WT sequence and mutated sequence (e.g. caused by a SNP)

$f($ WT sequence $)$ -> prob1

$f($ mutated sequence $)$ -> prob2

- Inference of SNP function:

  - prob1 >> prob2: **loss of function** mutation
  - prob1 << prob2: **gain of function** mutation