HUMAN TECHNOPOLE

POLITECNICO
MILANO 1863

Data science for genetic data analysis and disease prediction

DS4HB 2025 Workshop

Tutorial 1.1

Andrea Lampis, Andrea Mario Vergani

Andrea Lampis

Andrea Mario Vergani

- **PhD students in Data Analytics and Decision Sciences**
    - *Di Angelantonio & Ieva Group* @ Health Data Science Centre, Human Technopole
    - Politecnico di Milano *(DEIB, DMAT)*
- **MSc in Computer Science and Engineering** @ Politecnico di Milano

Tutorial overview

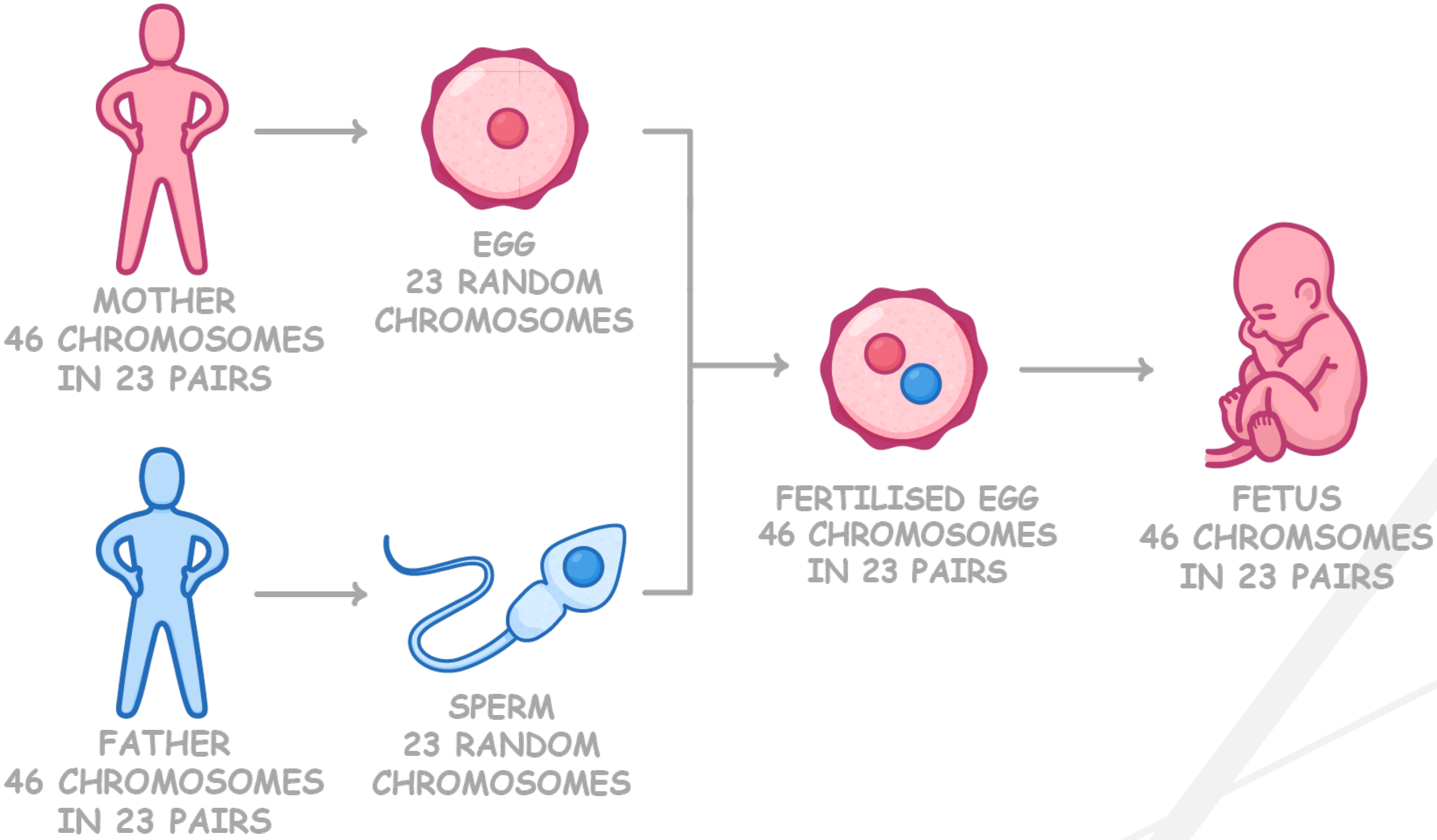
Objectives

- Learn to apply **data science** techniques to:
    - Analyse **genetic data**
    - Predict disease (genetic) **risk**
- Run your (first) statistical genetics **scripts**
- Access popular genetics databases and **web apps** to validate findings

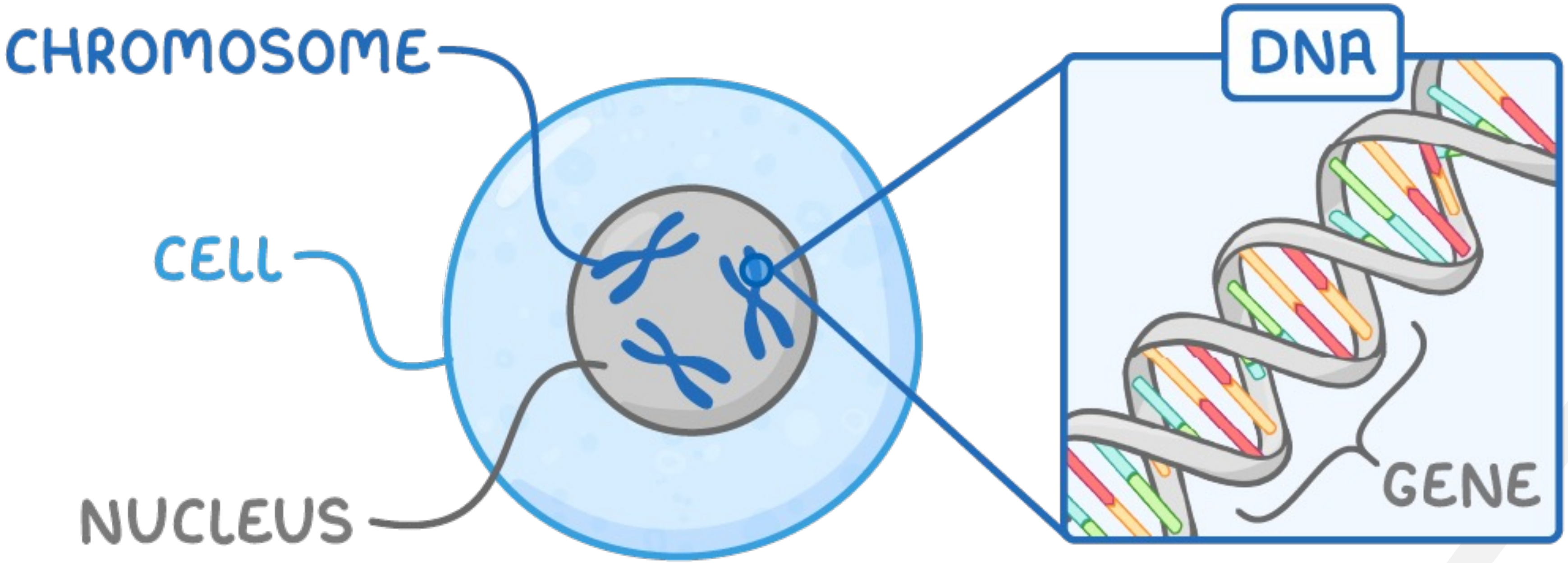
Organization

- **Background** *(Andrea Lampis)*
- **Part 1** *(Andrea Lampis)* - R notebook
    - Genetic data quality control
    - Genome-Wide Association Studies (GWAS)
    - Polygenic Risk Score (PRS)
- **Part 2** *(Andrea Mario Vergani)* - Python notebook
    - GWAS databases and summary statistics
    - Biological relevance of GWAS findings

Chromosomes

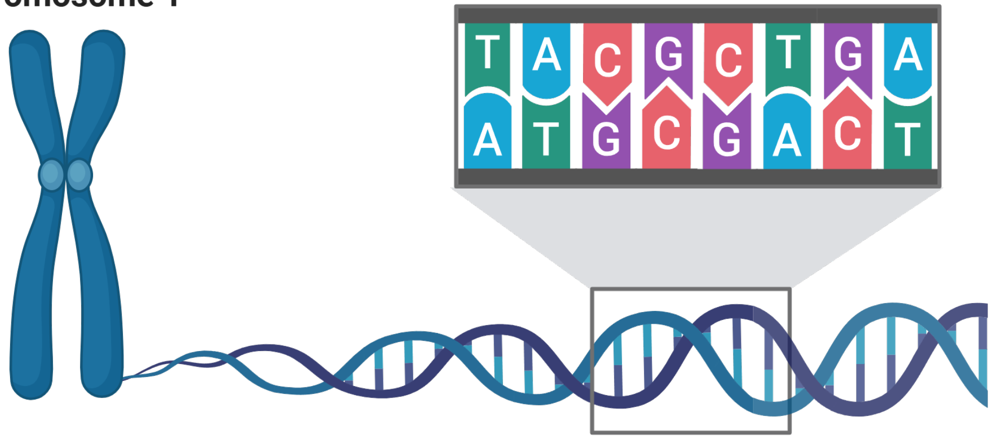


Cells



DNA

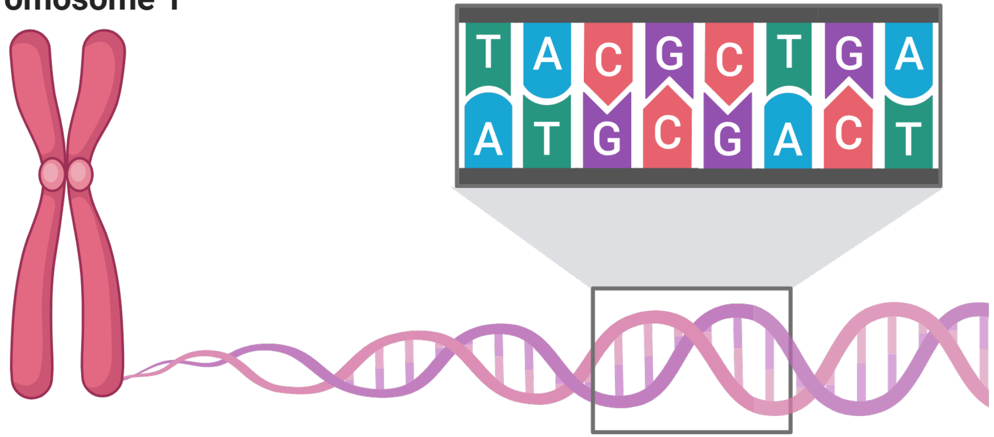
Paternal Chromosome 1



T A C G C T G A

A T G C G A C T

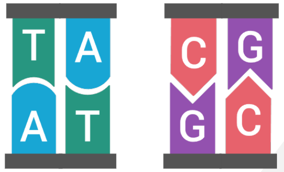
Maternal Chromosome 1



T A C G C T G A

A T G C G A C T

Possible base pairings



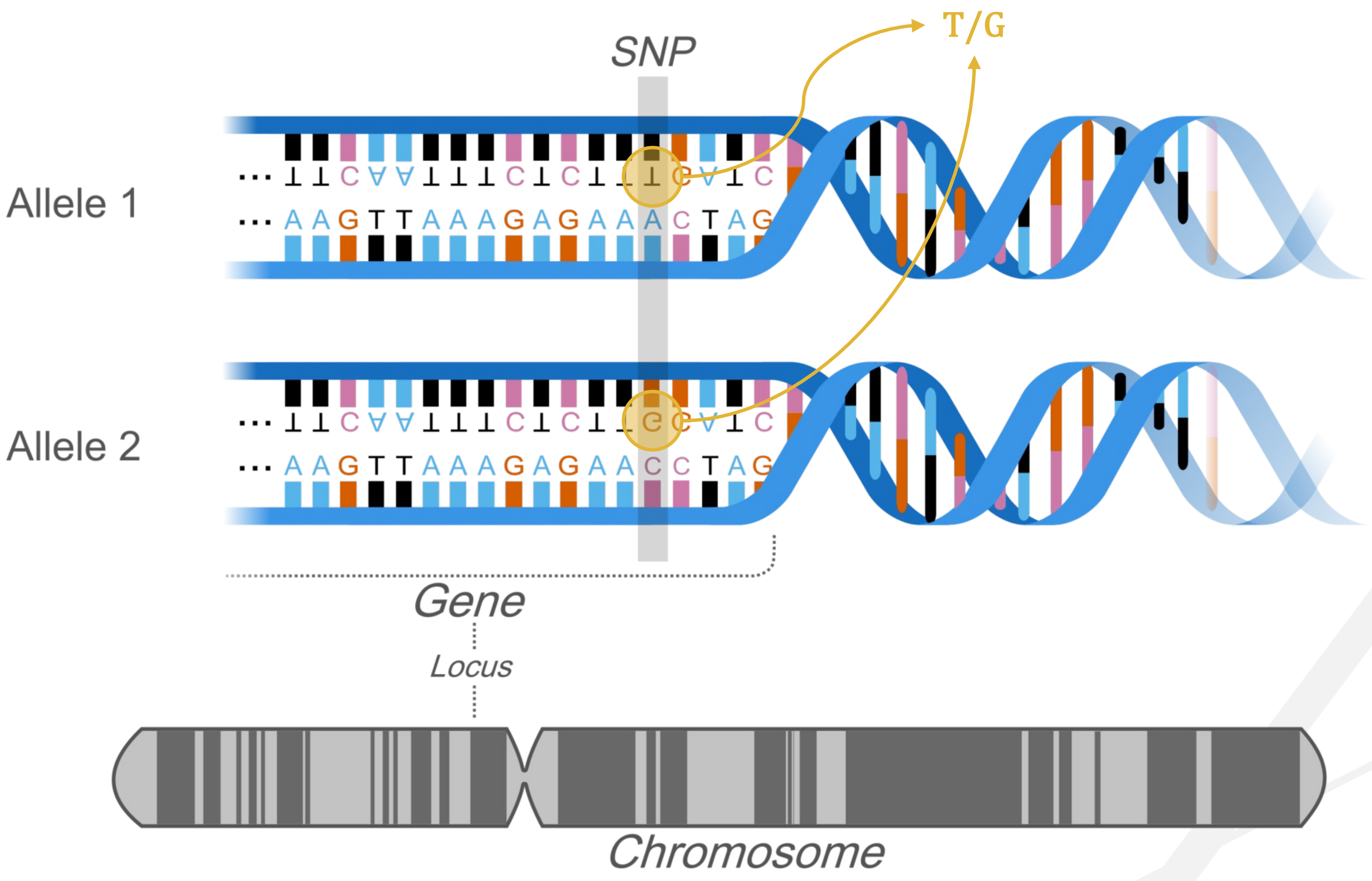
T A

A T

C G

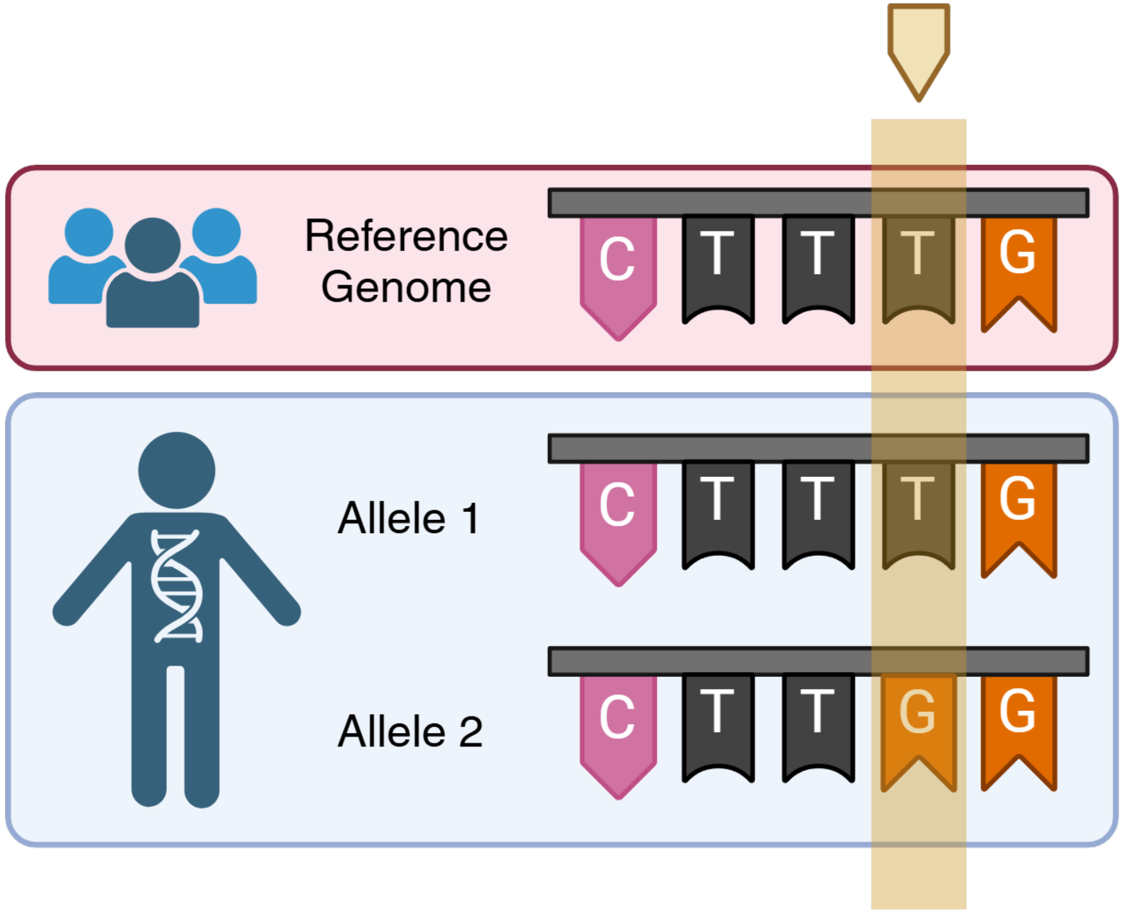
G C

Single Nucleotide Polymorphism (SNP)


T/G
SNP
Allele 1
… T T C A A T T T C T C T T T G A T C
… A A G T T A A A G A G A A A C T A G
Allele 2
… T T C A A T T T C T C T T G G A T C
… A A G T T A A A G A G A A C C T A G
Gene
Locus
Chromosome

Adapted from Wikimedia

Reference Genome



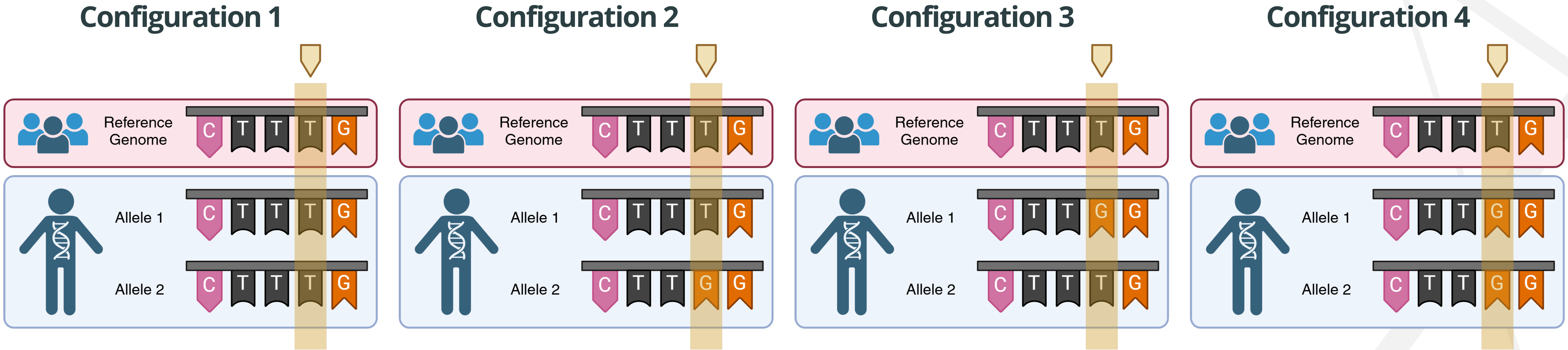
Biallelic variants:

- T **Reference Allele:**
  The nucleotide found in the reference genome, representing the baseline sequence.

- G **Alternate Allele:**
  A variant nucleotide found in some individuals, differing from the reference sequence.

0/1/2 Encoding

T Reference Allele G Alternate Allele



Configuration 1

Reference Genome C T T T G

Allele 1 C T T T G

Allele 2 C T T T G

Configuration 2

Reference Genome C T T T G

Allele 1 C T T T G

Allele 2 C T T G G

Configuration 3

Reference Genome C T T T G

Allele 1 C T T G G

Allele 2 C T T T G

Configuration 4

Reference Genome C T T T G

Allele 1 C T T G G

Allele 2 C T T G G

T/T: Homozygous for T

0

T/G: Heterozygous

1

G/T: Heterozygous

1

G/G: Homozygous for G

2

Count of alternate allele

Why 0/1/2 Matters

Genome-wide Association Studies (GWAS):

- Quantifies statistical association between genetic variants (e.g., SNPs) and traits or diseases.
- Scans the genomes of many individuals to find variants linked to specific outcomes.

GWAS Summary Statistics

SNP	…	Beta (effect size)	…
rs10212		-0.0912	
rs21210		0.7895	
…			
rs20192		0.0245	

GWAS typically use additive models:

- Assumes each minor/alternate allele contributes additively to the trait or disease risk.
- Allows researchers to treat genotype effects as linear, so the effect size (often derived from regression coefficients) represents the change per additional risk allele.

Why use 0/1/2 encoding?

- Reduces complex genotype data to a single number.
- Enables efficient testing of millions of SNPs for associations.
- Facilitates the use of standard statistical tools like regression.

Google Colab notebooks

Part 1

Part 2

https://github.com/ht-diva/ds4hb_workshop_t1_1