UMR 1283 8199

Genome-Wide Association Study
(TD GWAS)

2020-11-06

Lijiao Ning

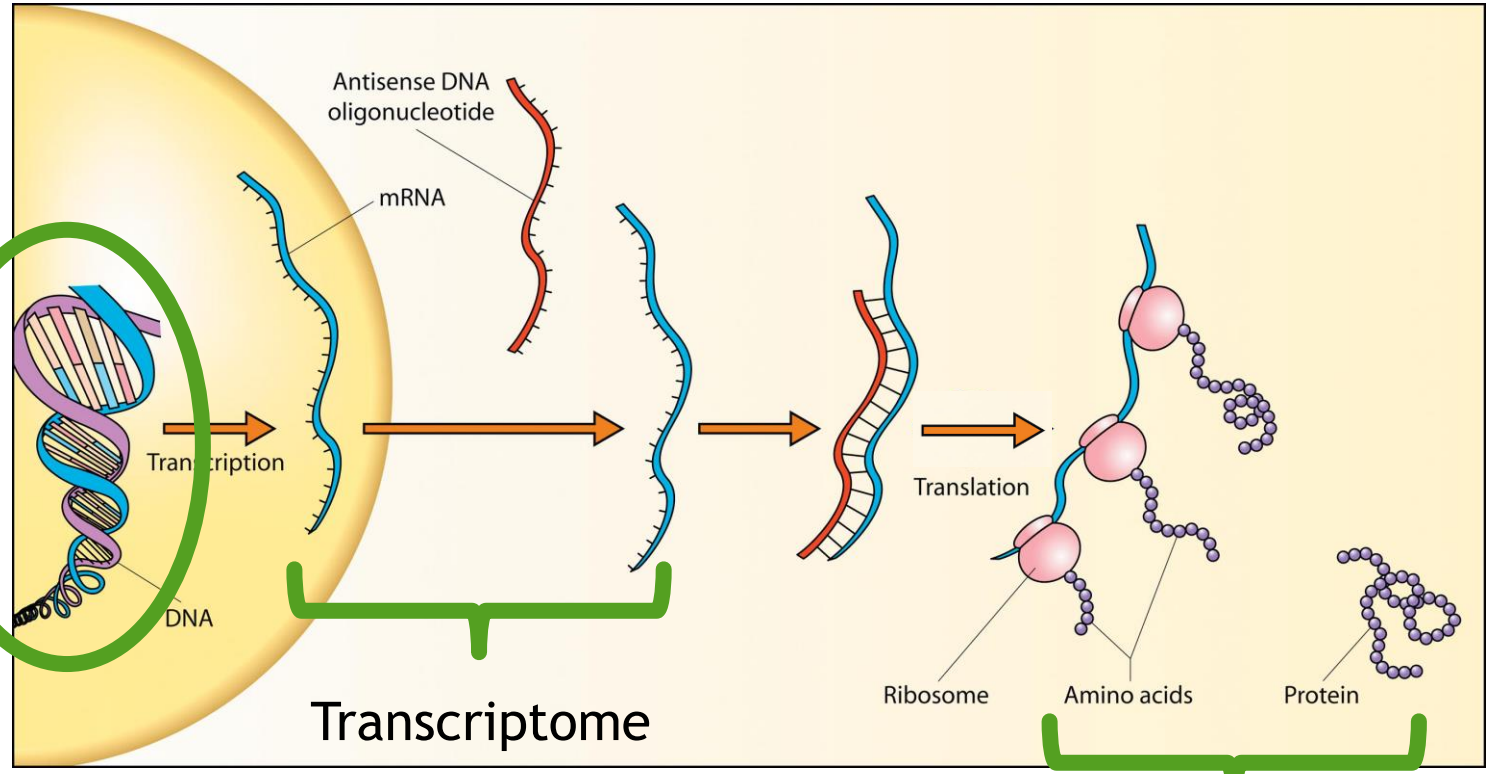
lijiao.ning@cnrs.fr

Inserm · Institut Pasteur de Lille — Vivre mieux, plus longtemps · Université de Lille · cnrs · E.g.i.d. European Genomic Institute for Diabetes

Course Outline

- Brief Introduction
  - The Omics
  - Some Notions In Genetics
- Basic Genetic Analyses
  - Linkage Analysis
  - Association Analysis
- Data Quality Control
- Association Test
- Visualization
- Practical Session

The Omics

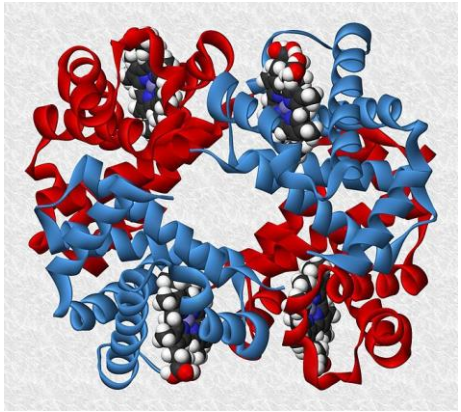


Genome

Proteome

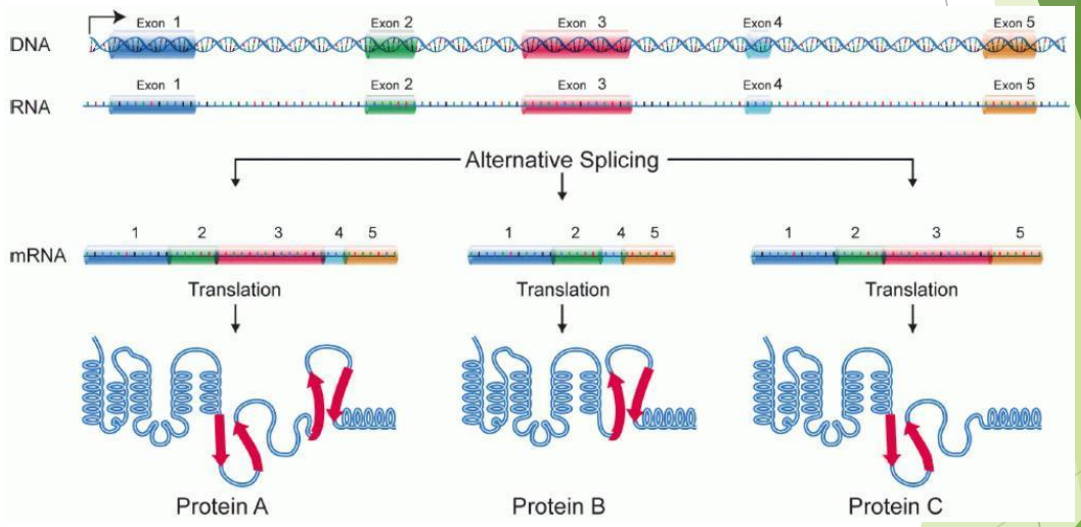
The Omics

- Proteomics
  - Proteome: the complete set of proteins synthesized by a cell or organism at a given time and under given conditions
  - Characterize biological information such as protein structure, function, location, interaction
  - Study methods
    - Electrophoresis
    - Mass spectrometry



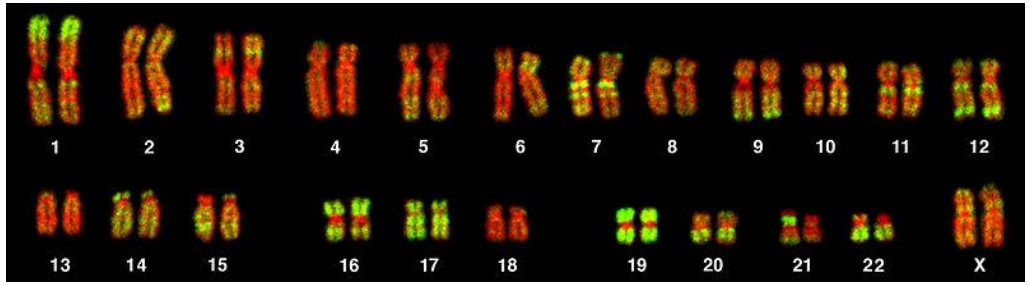
The Omics

- Transcriptomics
  - Transcriptome: the total set of transcripts (RNA) produced in a cell or organism
  - Reflect the level of gene expression
  - Characteristics: dynamic
  - Study methods
    - DNA microarrays
    - NGS (Next generation sequencing) for RNA sequencing

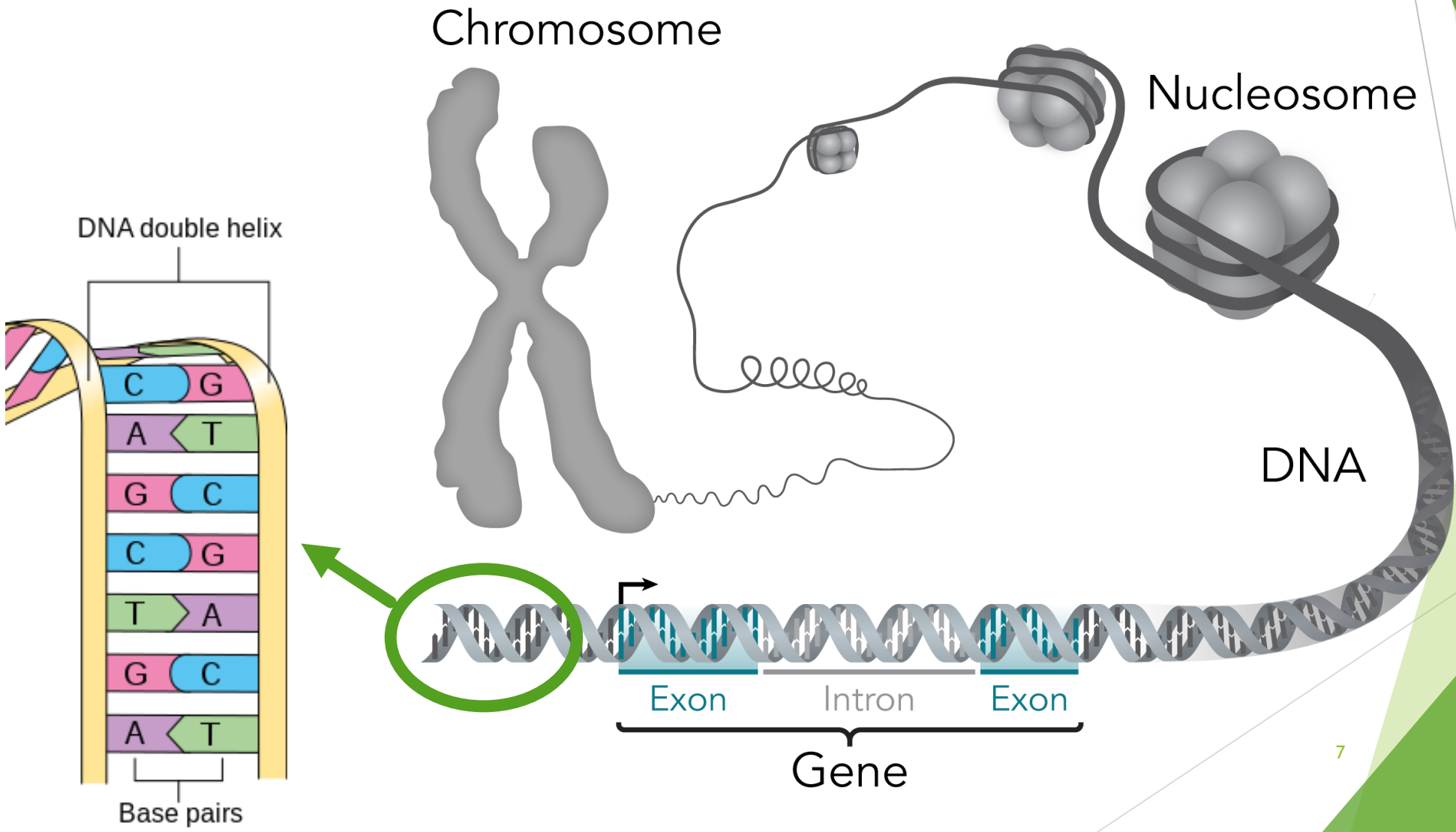


The Omics

- Genomics
  - Genome: the whole set of genetic information (DNA) in a cell or organism
  - Human genome
    - About 3 billion DNA base pairs
    - More than 20,000 protein coding genes
    - Diploid (2n)
    - 23 pairs of chromosomes
  - Study method
    - DNA microarray
    - NGS for genotyping



Genomic Sequence



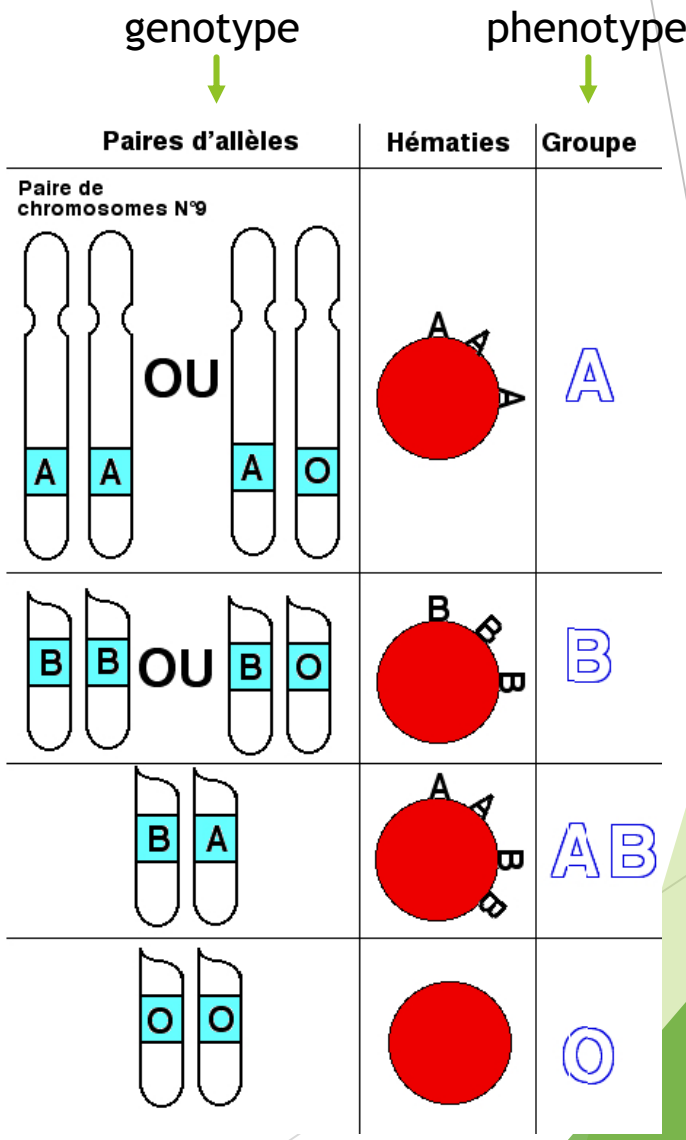
Alleles

- Allele: different versions of the same gene which located at the same genetic position (locus)
  - Homozygotes: two copies of the same allele
  - Heterozygotes: one of each type of allele

  - Dominant: the presence of a single allele is sufficient for the phenotype to be expressed
  - Recessive: need a pair of alleles for the phenotype to be expressed
  - Codominant: simultaneous expression of both alleles

Genotype & Phenotype

- Genotype: all genes carried by an individual
- Phenotype: observable characteristics of an individual
- *E.g.*: Blood group gene on chromosome 9



Mutation & Polymorphism

- Mutation: changes in DNA sequence compared to a “normal” form (reference genome check: https://genome.ucsc.edu/), naturally occurring but rare

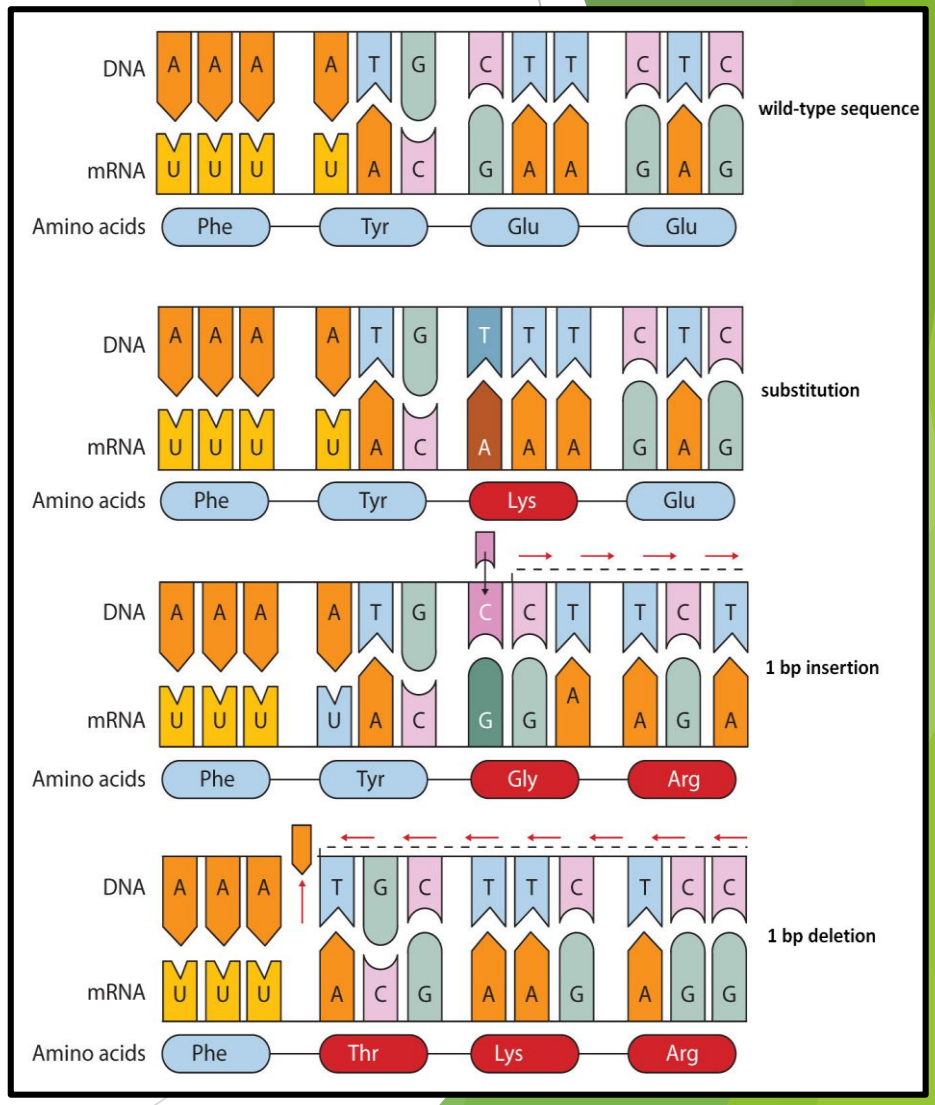
- Polymorphism: variations in DNA sequence, relatively more frequent in a population

- Different types of mutation / polymorphism
  - Germline *vs.* Somatic
  - Chromosomal modification: translocation, inversion, fission, fusion
  - Punctual modification: substitution, insertion, deletion, duplication

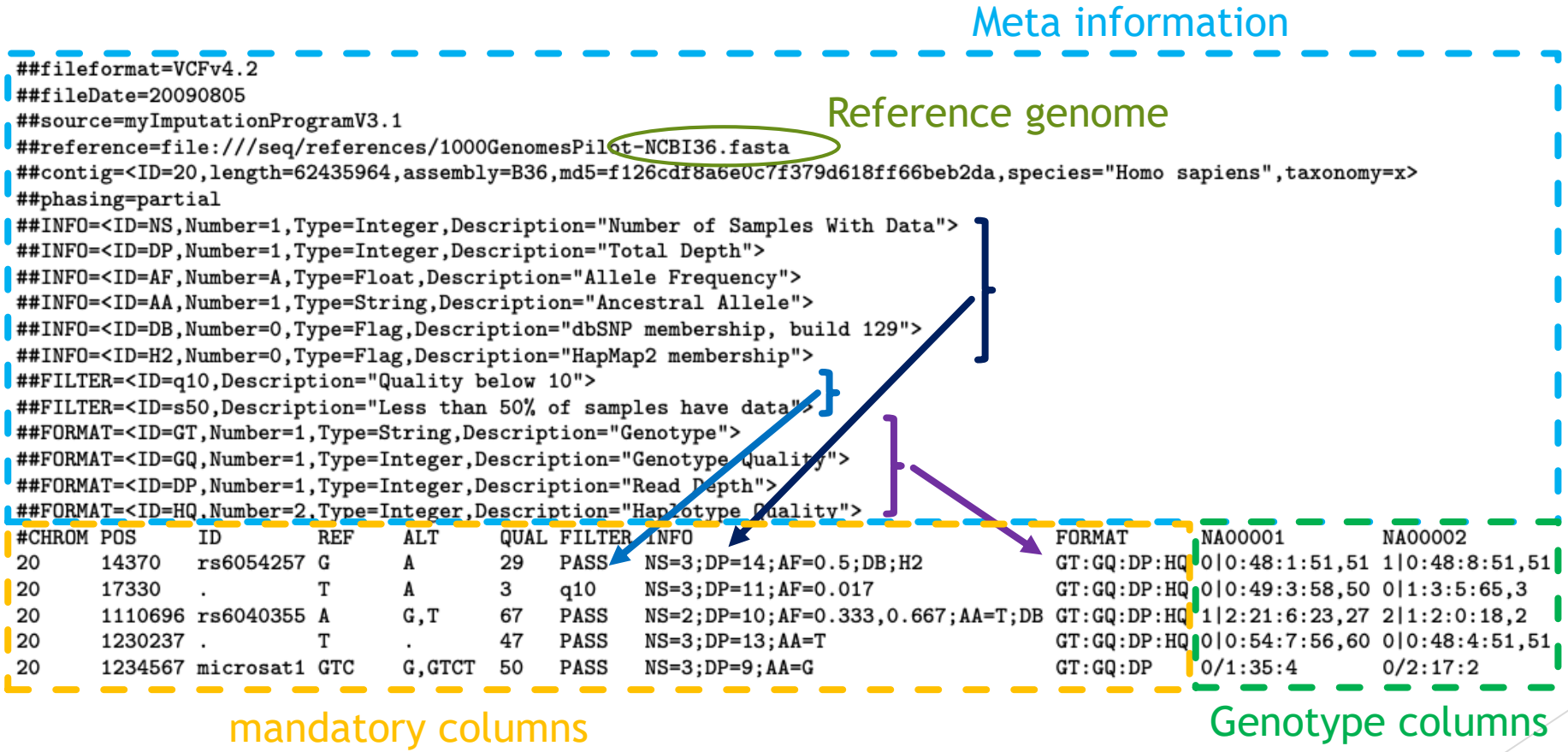
Punctual Mutations

- SNP (single nucleotide polymorphism)
  - Synonym: no change in produced amino acid
  - Missense: results in produced of another amino acid
  - Nonsense: results in a premature stop codon

- INDEL (insertion or deletion)
  - In-frame: insertion or deletion of a multiple of 3 bases, no shift in the reading frame
  - Frameshift



VCF (Variant Call Format)



(Source: https://samtools.github.io/hts-specs/VCFv4.2.pdf)

Basic Genetic Analyses

- Linkage Analysis
  - Applied to family data
  - Aims to find alleles whose transmission is not independent in the family
- Association Analysis
  - Applied to population data
  - Search for alleles significantly associated with the phenotype of interest
  - Two main types
    - Candidate gene study
    - Genome-wide

Association Test

- Chi2 test

	AA	Aa	aa
Status = 1 (case)	$O_{AA, 1}$	$O_{aa, 1}$	$O_{aa, 1}$
Status = 0 (control)	$O_{AA, 0}$	$O_{aa, 0}$	$O_{aa, 0}$

- Generalized linear regression
  - Logistic model for discrete trait
  - Linear model for continuous trait
  - genotype coding (assuming “a” is the risk allele)
    - Additive: aa *vs.* Aa *vs.* AA (aa = 2, Aa = 1, AA = 0)
    - Recessive: aa *vs.* (AA + Aa) (aa = 1, Aa = AA = 0)
    - Dominant: (Aa + aa) *vs.* AA (aa = Aa = 1, AA = 0)

PLINK: A GWAS Toolset

- Open-source whole genome association analysis toolset
- Special input formats for PLINK
  - For PLINK 1.x: .bim, .bed, .fam
  - For PLINK 2.0: .pvar, .pgen, .psam
  - Default coding:
    - male = 1, female = 2, unknown = 0
    - control = 1, case = 2, missing = -9 or 0
- More details: https://www.cog-genomics.org/plink2/formats

```
| > head -n 3 1_QC_GWAS/HapMap_3_r3_1.bim
1       rs2185539       0       556738  T       C
1       rs11510103      0       557616  G       A
1       rs11240767      0       718814  T       C
```

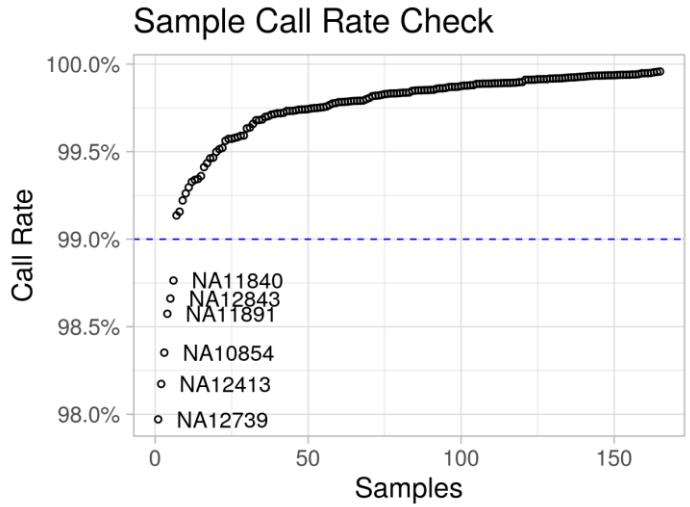
```
| /disks/DATATMP/SB_lning/TD_GWAS @ R402 (lning)
| > head -n 3 HapMap_3_r3_1.pvar
#CHROM  POS     ID      REF     ALT
1       556738  rs2185539       C       T
1       557616  rs11510103      A       G
```

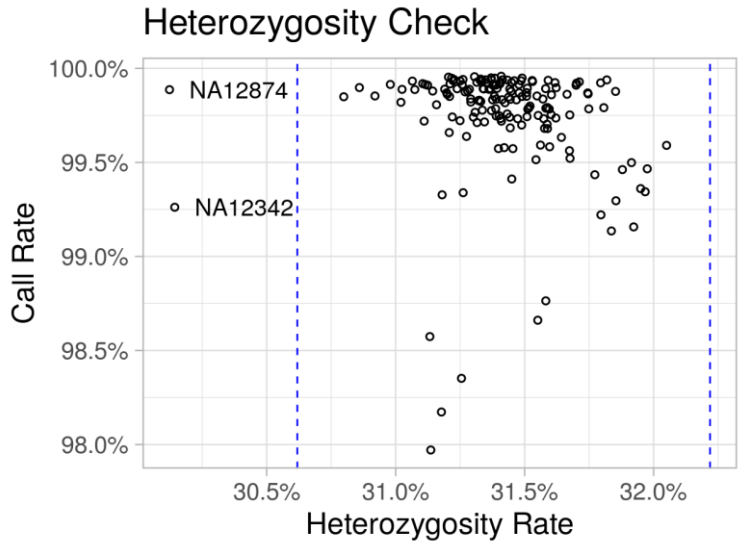
```
| > head -n 3 1_QC_GWAS/HapMap_3_r3_1.fam
1328 NA06989 0 0 2 2
1377 NA11891 0 0 1 2
1349 NA11843 0 0 1 1
```

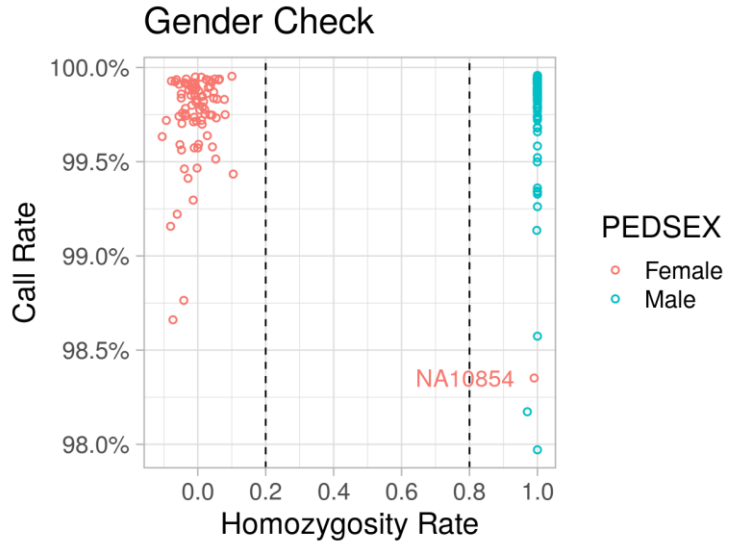
```
| /disks/DATATMP/SB_lning/TD_GWAS @ R402 (lning)
| > head -n 3 HapMap_3_r3_1.psam
#FID    IID     PAT     MAT     SEX     PHENO1
1328    NA06989 0       0       2       2
1377    NA11891 0       0       1       2
```

Data Quality Control

- Sample-based
  - Sample call rate (--missing)
  - Heterozygosity (--het)
  - Gender discordant (--check-sex)

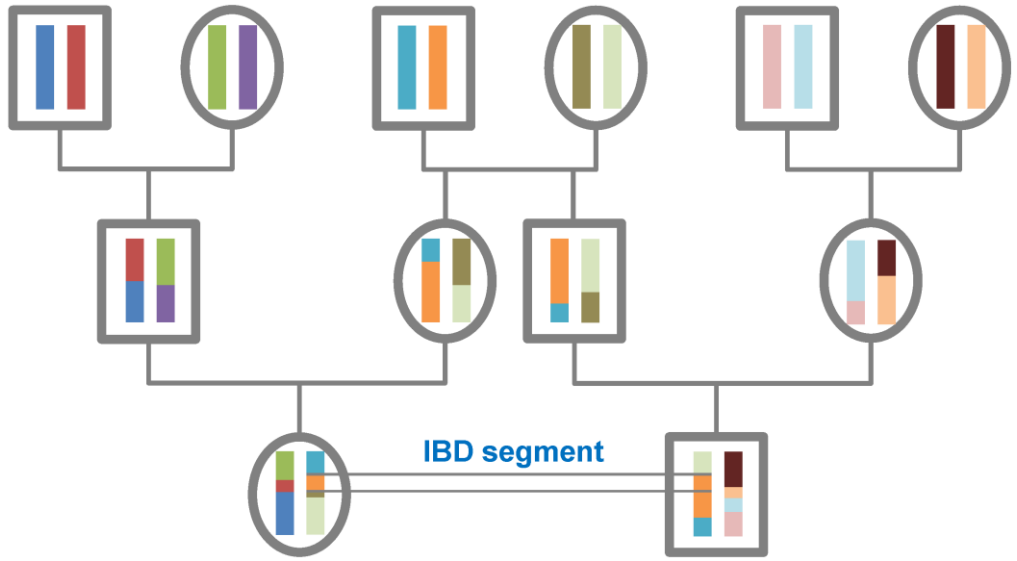






Data Quality Control

- Sample-based
  - Sample call rate
  - Heterozygosity
  - Gender discordant
  - Relatedness (--genome)

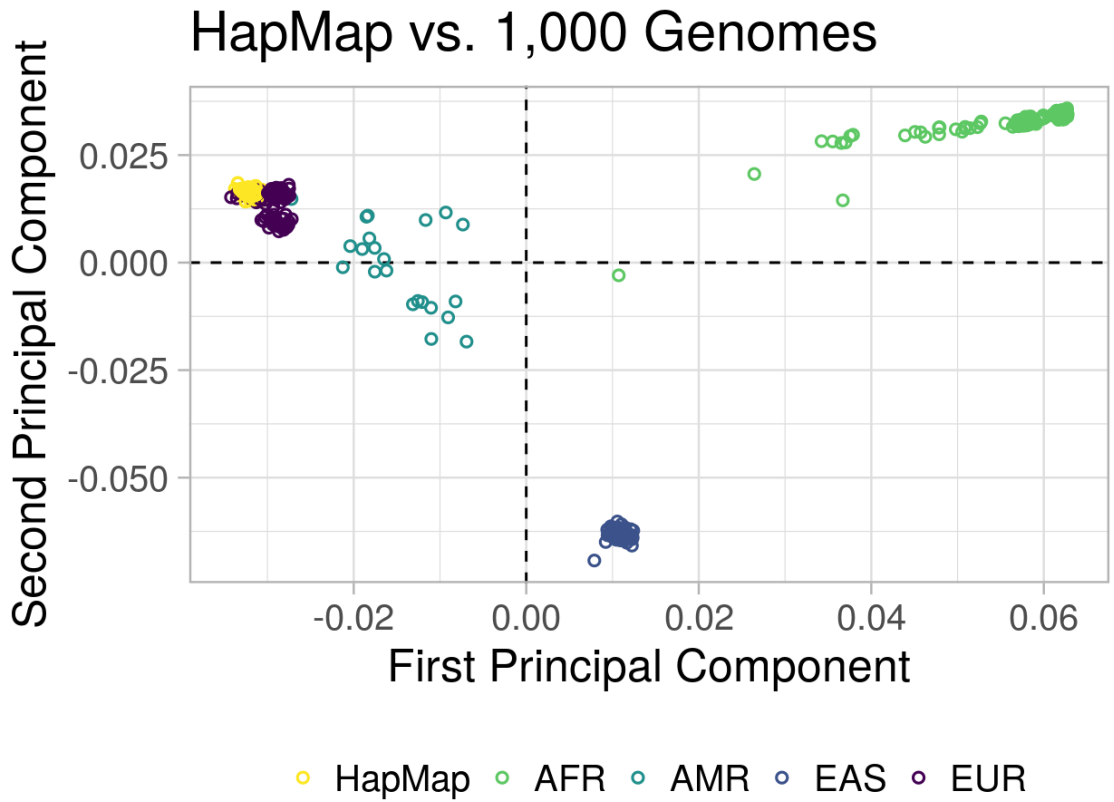


	IBD windows	Pair count
1	[0.2,0.3] = Second degree relatives	2
2	(0.3,0.4]	0
3	(0.4,0.6] = First degree relatives	97
4	(0.6,0.8]	0
5	(0.8,1] = MZ twins/duplicates	0

Data Quality Control

- Sample-based
  - Sample call rate
  - Heterozygosity
  - Gender discordant
  - Relatedness
  - Population structure

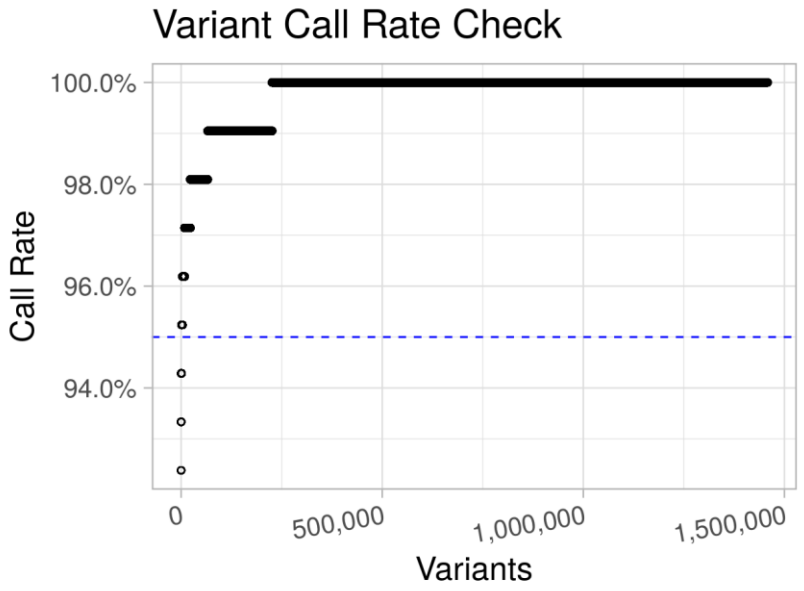
  (--make-grm-bin -- pca)

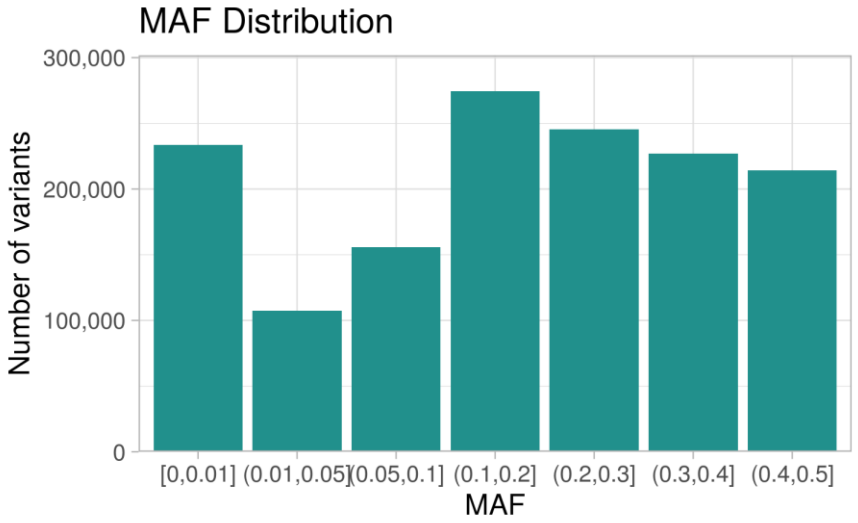


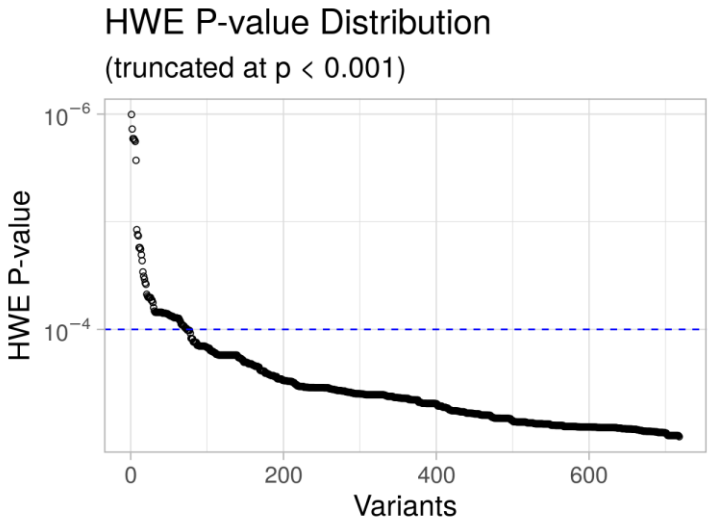
(1,000 Genomes Project: https://www.internationalgenome.org)

Data Quality Control

- Variant-based
  - Variant call rate (--missing)
  - Mendel errors (--mendel)
  - Minor allele frequency (MAF) distribution (--freq)
  - Hardy-Weinberg Equilibrium (--hardy)







Generalize Linear Regression

- Run analyses
  - --assoc / --logistic / --linear
  - --covar

- Results of logistic model

Multiple testing correction

```
##     CHR         SNP        BP A1 TEST NMISS     OR   STAT      P       FDR
## 1:   22 rs11089128 14560203  G  ADD   105 2.3200  1.497 0.1344 0.9558372
## 2:   22 rs11167319 14850625  G  ADD   105 2.0380  1.532 0.1256 0.9558372
## 3:   22  rs2027649 14880040  G  ADD   105 1.5530  1.242 0.2143 0.9558372
## 4:   22  rs3016104 15063831  G  ADD   105 1.6510  1.096 0.2730 0.9558372
## 5:   22  rs9680776 15380277  T  ADD   103 0.7106 -1.075 0.2825 0.9558372
## 6:   22 rs12106650 15551377  T  ADD   104 0.6899 -1.057 0.2905 0.9558372
```

Generalize Linear Regression

- Why multiple testing is a problem?

  If we do 100 tests simultaneously and set and use significance level at 0.05,

  P(at least 1 significant result by change) = 1 - P(non significant results)

  $= 1 - (1 - 0.05)^{100} = 0.99$

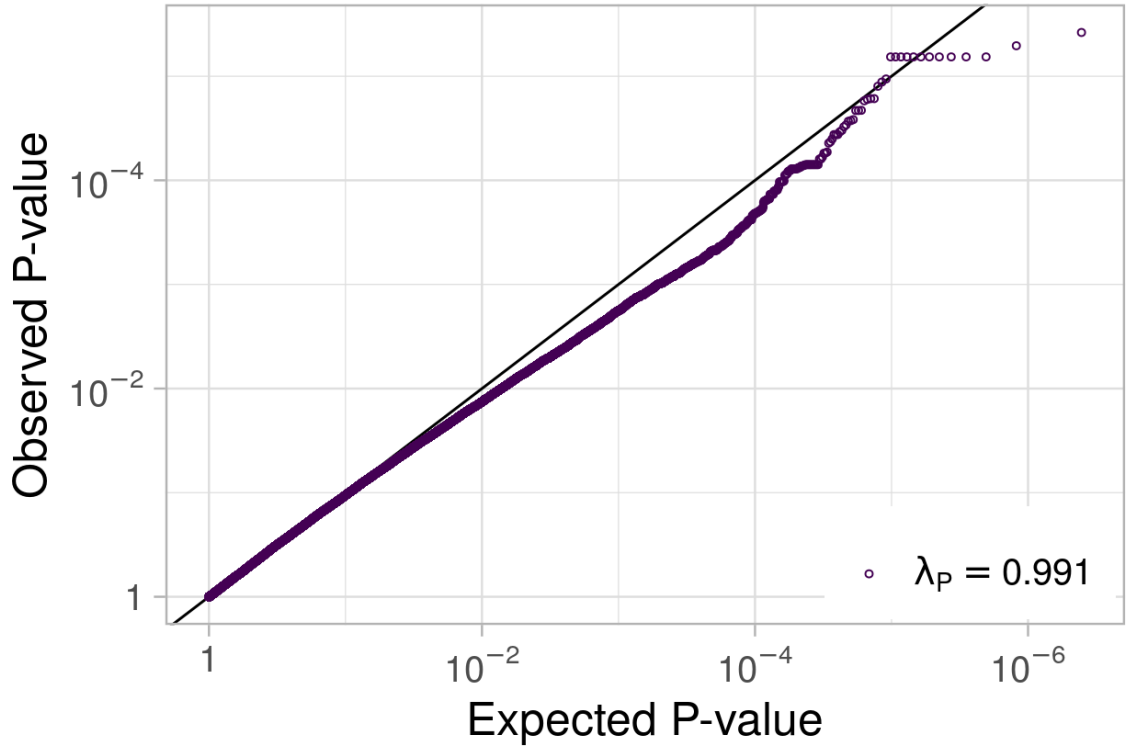
- Bonferroni method

  P(at least 1 significant result by change) = $1 - (1 - 0.05 / 100)^{100} = 0.049$

- False discovery rate (FDR): the proportion of false positive amongst all significant results

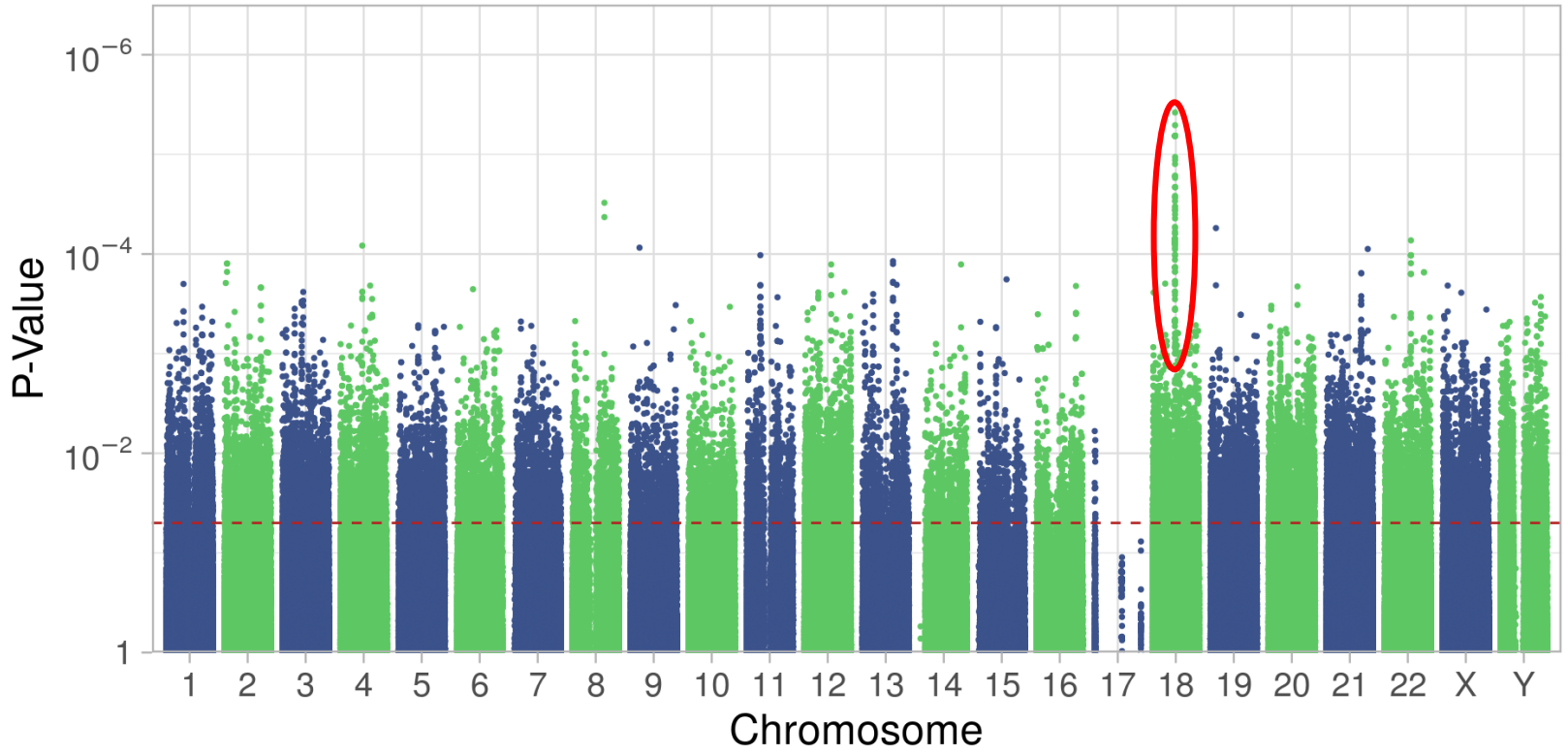
Visualization

- Quantile-Quantile Plot
  - Genomic inflation factor λ



Visualization

- Manhattan Plot
  - Adapt for huge data-points visualization



References

- Agler, Cary S et al. “Protocols, Methods, and Tools for Genome-Wide Association Studies (GWAS) of Dental Traits.” *Methods in molecular biology (Clifton, N.J.)* vol. 1922 (2019): 493-509. doi:10.1007/978-1-4939-9012-2_38
- Tam, Vivian et al. “Benefits and limitations of genome-wide association studies.” *Nature reviews. Genetics* vol. 20,8 (2019): 467-484. doi:10.1038/s41576-019-0127-1
- Marees, Andries T et al. “A tutorial on conducting genome-wide association studies: Quality control and statistical analysis.” *International journal of methods in psychiatric research* vol. 27,2 (2018): e1608. doi:10.1002/mpr.1608
- Chang, Christopher C et al. “Second-generation PLINK: rising to the challenge of larger and richer datasets.” *GigaScience* vol. 4 7. 25 Feb. 2015, doi:10.1186/s13742-015-0047-8
- Chang C.C. (2020) Data Management and Summary Statistics with PLINK. In: Dutheil J. (eds) Statistical Population Genomics. Methods in Molecular Biology, vol 2090. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-0199-0_3
- Zhang, Xiang et al. “Chapter 10: Mining genome-wide genetic markers.” *PLoS computational biology* vol. 8,12 (2012): e1002828. doi:10.1371/journal.pcbi.1002828

Practical Session

- QC + Association test on HapMap data:

https://share-good.egid.fr/fop/VZfxQvAD/TD_GWAS_data.zip

- Data description can be found here:

https://github.com/Ning-L/TD_GWAS