Centre for Genomics and Personalised Health

Statistical Genomics and Bioinformatics Workshop 2025
Neurogenomics Program, CGPH

Mr Paul Ruiz Pinto, Dr Heidi Sutherland and Prof Divya Mehta

Outline

- About this workshop
- What will be taught – GWAS (Part 1) and DNAm (Part 2)
- Housekeeping - breaks and lunch break

Genome-wide association studies

Some biology

- Traits (includes characteristics as well as disorders) – monogenic or polygenic (*mono*: one, *poly*: many, *genic*: gene)
- Polygenic trait: When a trait or characteristic is influenced by two or more genes or genetic variants.
- Examples of polygenic traits include height, skin colour, hair colour

- Polygenic traits can be complex or multifactorial when caused by multiple genes as well as by environmental and lifestyle factors
- Examples include asthma, type-3 diabetes, PTSD, depression, cancer, metabolic syndrome

Some biology

- **Genetic variants**: Also known as single nucleotide polymorphisms (SNPs) is a single base-pair at which more than one nucleotide is observed (*poly*: many, *morphe*: form)
- **Example**: Position 100 on chromosome 1 has nucleotide A in majority of the people in a population but some have nucleotide G in the same position, then this position is called a SNP with alleles G and A
- **Allele frequency**: Indicates how common an allele is in a population. Calculated by counting the number of times we see an allele in a population and divided by the total number of copies of a gene
- **Minor allele frequency (MAF):** The frequency of the second most common allele in a population (minor allele)
- **Common SNPs**: SNPs with MAF > 0.05 or > 0.01

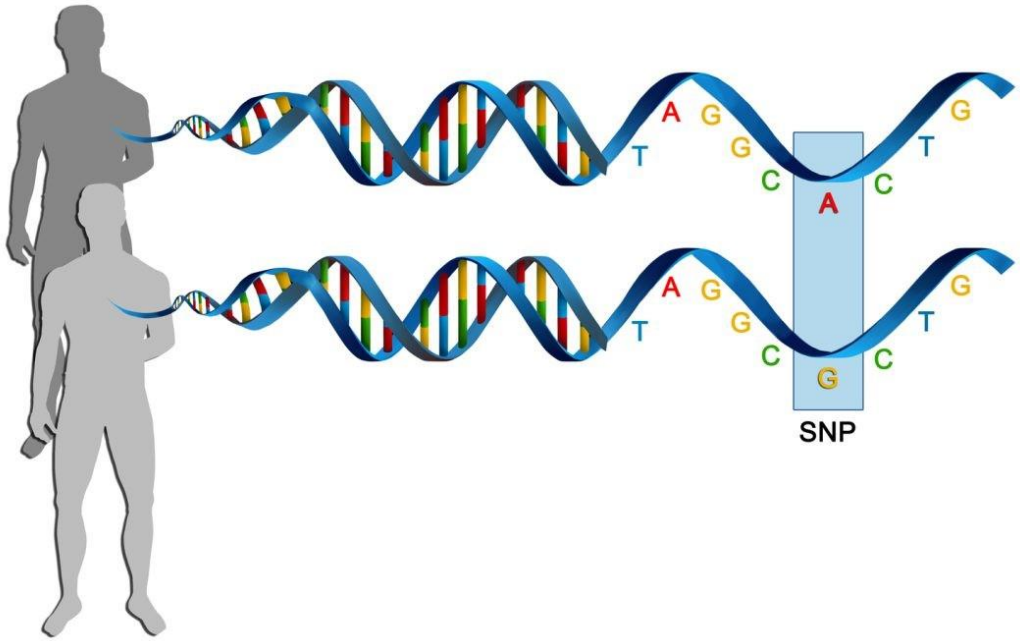


Image source: https://www.genengnews.com/topics/omics/study-finds-genetic-basis-of-common-diseases-may-span-tens-of-thousands-of-snps/

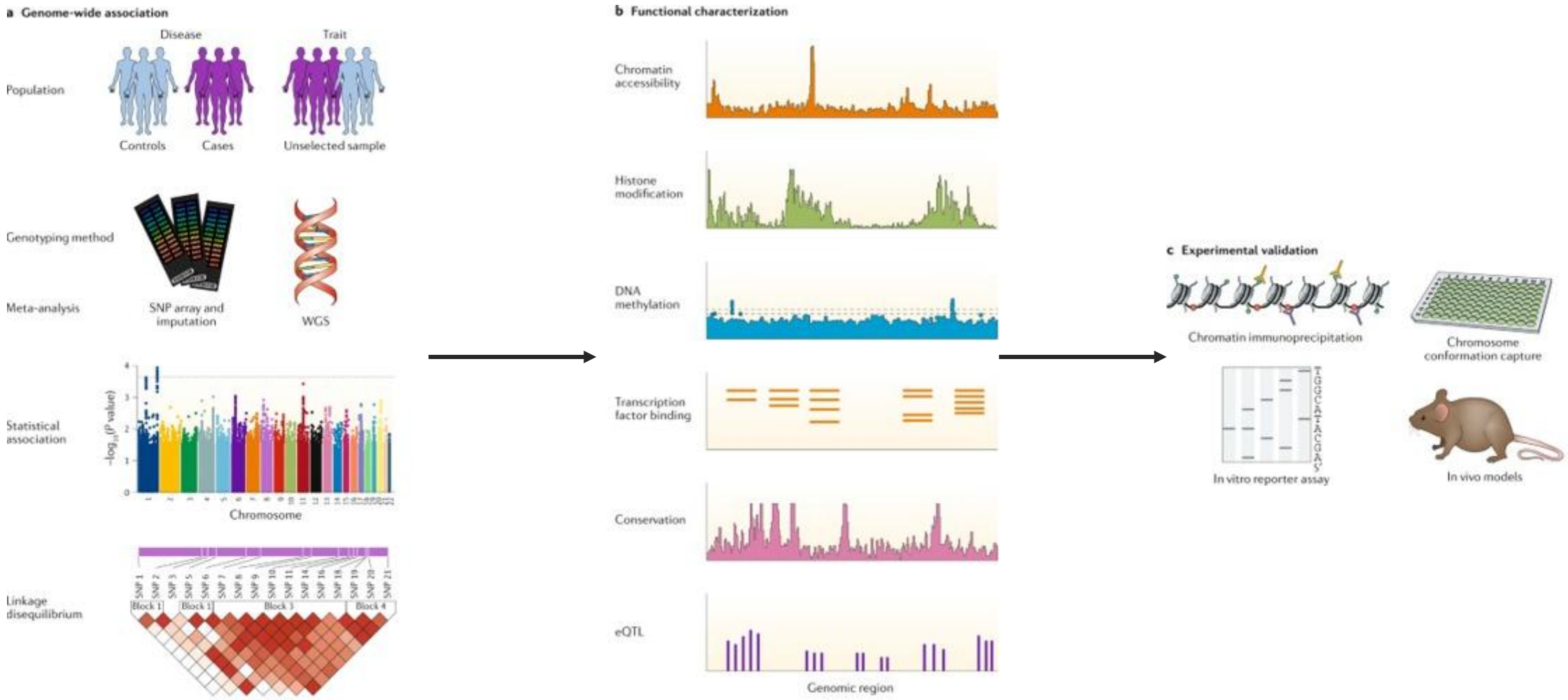
What is GWAS and why do it?

- **Genome-wide Association Study (GWAS):** Hypothesis free univariate association analysis of hundreds of thousands to millions of common genetic variants (MAF ≥ 0.01 or 0.05) across the genomes of many individuals to identify genotype–phenotype associations
- Phenotypes in GWAS are complex traits or diseases
- First GWAS was conducted in 2005 for age-related macular degeneration [Klein et al, 2005]. Since then >3,500 published GWASs for a wide variety of traits and disorders have been conducted

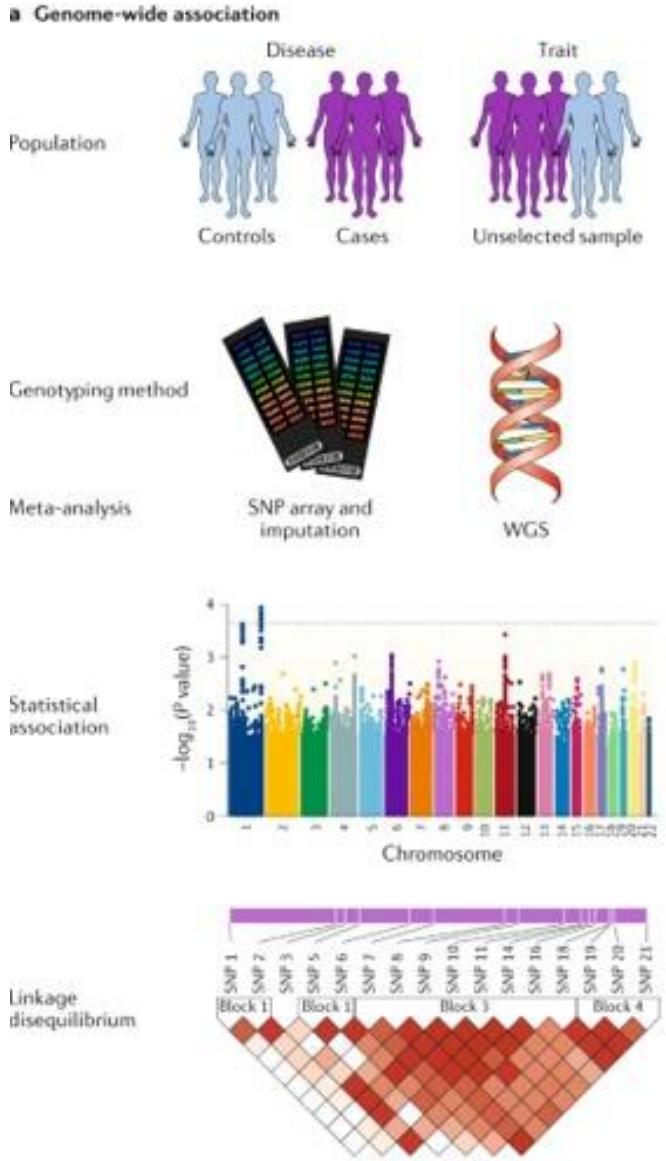
- **Why do it?**
  - Identification of novel disease-causing genes and mechanisms
  - Insights in genetic architecture of the trait
  - Identification of new drug targets and disease biomarkers
  - Risk prediction
  - Optimisation of therapies based on genotype

Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. Science 2005;308:385–389

GWAS workflow



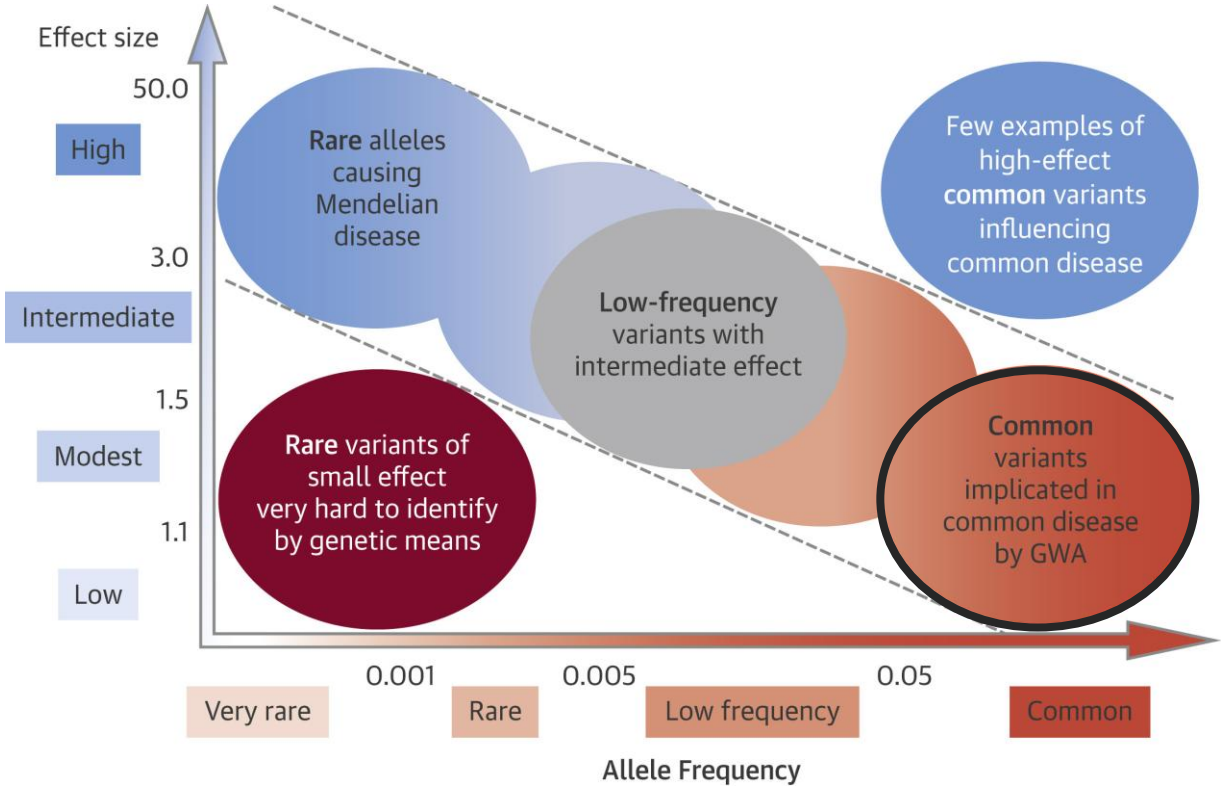
Step 1 - GWAS



1. Study design
   - Phenotype determination
   - Selecting an appropriate study population based on the phenotype – cases and controls or unselected population
2. Genotyping method
   - Whole genome sequencing
   - SNP array followed by imputation of remaining SNPs on genome using appropriate human reference genome
   - Meta-analysis: Combining the results from independent previous GWASs
3. Association analysis
   - Test for association of SNP with the phenotype
   - Linear regression for continuous phenotype or logistic regression for binary trait
4. Identify regions of interest
   - Regions that have reached genome-wide significance (association $p < 5 \times 10^{-8}$) or suggestive significance (association $p < 1 \times 10^{-5}$)
   - Account for linkage equilibrium

Note - **linkage disequilibrium** (**LD**) is the non-random association of alleles at different loci in a given population

GWAS results

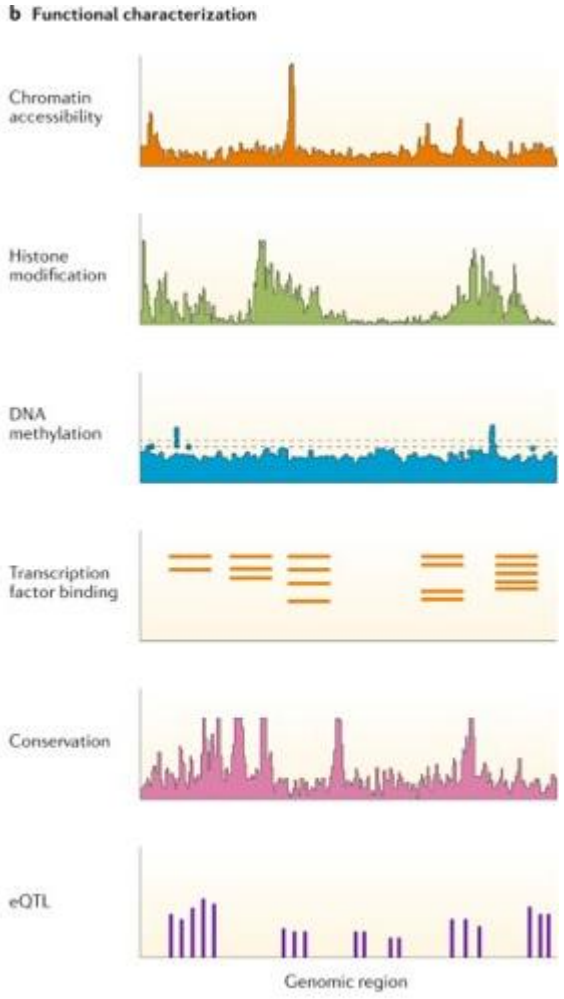


- GWAS analyses common variants (MAF > 0.05)
- The effect sizes of significant variants are < 1.5 (low to moderate)

An *allelic effect size* is the magnitude of the effect of an allele on a phenotype

Source: Assimes, Themistocles L., and Robert Roberts. "Genetics: implications for prevention and management of coronary artery disease." Journal of the American College of Cardiology 68.25 (2016): 2797-2818.

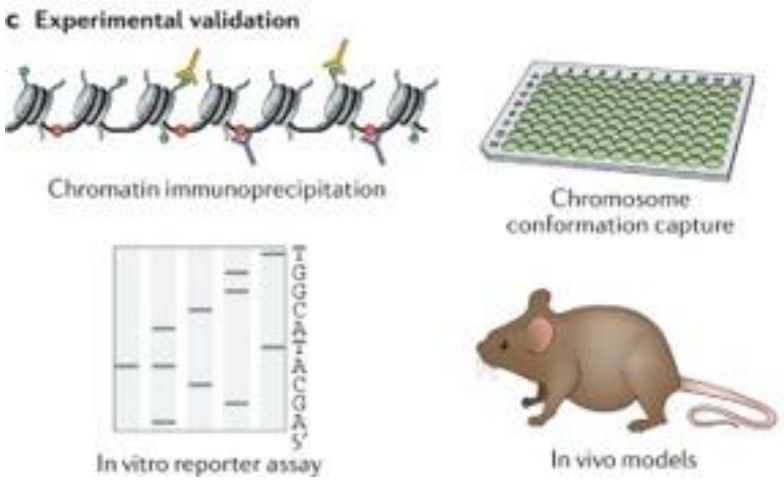
Step 2 – Functional characterisation (Post-GWAS analysis)



- Includes a wide variety of computational/bioinformatic approaches
- To understand the biological significance of the identified variants and prioritise variants and genes
- Approaches:
  1. Gene annotation of the SNPs
  2. Pathway analysis
  3. Fine-mapping to identify causal variants
  4. Integrative analysis with other omics to identify the mechanistic role of the variants
  5. Tissue enrichment
  6. Cross-disorder or cross-phenotype analysis
  7. Polygenic risk score analysis

Source: Tam, Vivian, et al. "Benefits and limitations of genome-wide association studies." Nature Reviews Genetics 20.8 (2019): 467-484.

Step 3 – Experimental validation



- Validation of GWAS results using cell-based systems and model organisms

Source: Tam, Vivian, et al. "Benefits and limitations of genome-wide association studies." Nature Reviews Genetics 20.8 (2019): 467-484.

PLINK

- An open-source software for whole genome data analysis
- Command-line program
- Performs
  - Data management
  - Summary statistics for quality control
  - Population stratification
  - Association testing
  - And many more…

Report

PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses

Shaun Purcell [b a], Benjamin Neale [b c], Kathe Todd-Brown [a], Lori Thomas [a], Manuel A.R. Ferreira [a], David Bender [b a], Julian Maller [b a], Pamela Sklar [b a a], Paul I.W. de Bakker [b a], Mark J. Daly [b a], Pak C. Sham [d]

https://zzz.bwh.harvard.edu/plink/

Genotype data

- Genotypes are stored in a matrix of size n × N with n >> N
- In PLINK, the bed file stores the genotype matrix in the form of 0, 1 and 2 or NA
- 0, 1 and 2 indicate the number of copies of A1 allele.

	SNP1	SNP2	SNP3	…	SNPn
Ind1	TT	AG	CT	…	AT
Ind2	TT	GG	CT	…	AA
Ind3	AT	GG	TT	…	TT
...	…	…	…	…	…
Indn	AA	AA	TT	…	AT

PLINK files – PED and MAP

- PED file -
  - Family ID (FID)
  - Individual ID (IID)
  - Paternal ID
  - Maternal ID
  - Sex (1=male; 2=female)
  - Phenotype (quantitative trait or affection status)
  - Genotype information
- MAP file -
  - Chromosome (1-22, X/23, Y/24, Mt/26)
  - rs# or SNP identifier
  - Genetic distance
  - Base-pair position (bp)

PED file

1	Sample1	0	0	2	0	G	G	A	C
2	Sample2	0	0	2	0	G	G	A	A
3	Sample3	0	0	2	0	G	G	A	A
4	Sample4	0	0	2	0	A	G	A	A
5	Sample5	0	0	1	0	G	G	A	A
6	Sample6	0	0	1	0	A	G	A	A
7	Sample7	0	0	2	0	G	G	A	A
8	Sample8	0	0	2	0	G	G	A	A
9	Sample9	0	0	2	0	G	G	A	A
10	Sample10	0	0	2	0	A	G	A	A

MAP file

16	rs11466023	9.45271	3299586
9	rs121908640	140.1591	133370370
7	rs121908764	123.6289	117267718
4	rs13117307	70.78986	56751740
11	rs137852761	95.60706	94180454
X	rs1972809	120.4023	119867475
19	rs2230267	80.15505	49469087
16	rs2270368	60.37151	50714335
22	rs2330809	19.06661	25002081
20	rs267606634	68.9663	43255169

PLINK files – Phenotype and covariates

- If present phenotype file has information on the phenotype studied
  - 2+ columns
  - Family ID
  - Individual ID
  - Phenotype 1
  - Phenotype 2
  - … Phenotype x
  - -9 or 0 is considered as missing
  - Case/Control as 1 = Control and 2 = Case

- Covariates file have the sample format
  - These can include information such as Age, principal components, etc

PLINK files – the binary Ped files

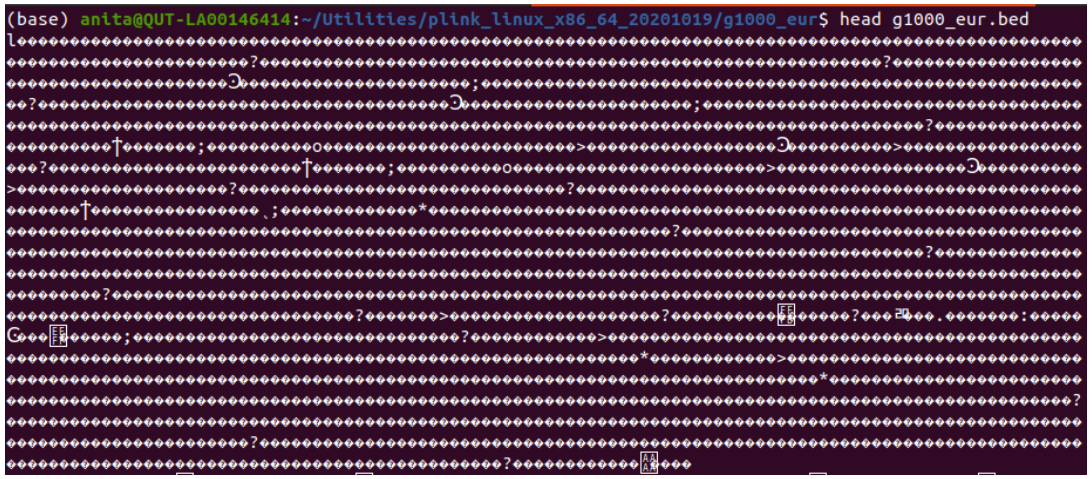
- PED and MAP are not efficient way to store data:
- For 503 individuals with 22,665,064 variants: PED file 43G and MAP file is 542M

- Efficient way to store whole genome data as using binary format of PED and MAP files
- The are a set of 3 files
  - bed (binary genotype) (2.7G)
  - bim (binary mapping) (629M)
  - fam (family details) (13K)

BED file

- Genotypes are stored in a matrix of size n × N with n >> N
- In PLINK, the bed file stores the genotype matrix in the form of 0, 1 and 2 or NA in binary format
- 0, 1 and 2 indicate the number of copies of A1 allele (default – minor allele).

Binary format of genotype file



	SNP1	SNP2	SNP3	…	SNPn
Ind1	TT	AG	CT	…	AT
Ind2	TT	GG	CT	…	AA
Ind3	AT	GG	TT	…	TT
…	…	…	…	…	…
Indn	AA	AA	TT	…	AT

	SNP1	SNP2	SNP3	…	SNPn
Ind1	2	1	1	…	1
Ind2	2	0	1	…	2
Ind3	1		0	…	0
…	…	…	…	…	…
Indn	0	2	0	…	1

Binary mapping (bim) and Family (fam) files

- The bim file contains information on the SNPs
  - Chromosome
  - Rs# or SNP identifier
  - Centimorgan
  - Base pair location
  - Minor allele in PLINK 1.9 (Alternate allele in PLINK 2)
  - Major allele in PLINK 1.9 (Reference allele in PLINK 2)

```
g1000_eur$ head g1000_eur.bim
1       rs537182016     0       10539   A       C
1       rs575272151     0       11008   G       C
1       rs544419019     0       11012   G       C
1       rs540538026     0       13110   A       G
1       rs62635286      0       13116   G       T
1       rs200579949     0       13118   G       A
1       rs531730856     0       13273   C       G
1       rs527952245     0       13313   G       T
1       rs558318514     0       13445   G       C
1       rs574697788     0       13494   G       A
```

- The fam file contains information on the individuals (first 6 columns of the PED file)
  - Family ID (FID)
  - Individual ID (IID)
  - Paternal ID
  - Maternal ID
  - Sex (1=male; 2=female)
  - Phenotype (quantitative trait or affection status)

```
g1000_eur$ head g1000_eur.fam
HG00096 HG00096 0 0 1 -9
HG00097 HG00097 0 0 2 -9
HG00099 HG00099 0 0 2 -9
HG00100 HG00100 0 0 2 -9
HG00101 HG00101 0 0 1 -9
HG00102 HG00102 0 0 2 -9
HG00103 HG00103 0 0 1 -9
HG00105 HG00105 0 0 1 -9
HG00106 HG00106 0 0 2 -9
HG00107 HG00107 0 0 1 -9
```

Simple commands

- Setting input
  - Note: All plink file (PED or binary have the same file name with different file extensions. E.g.: g1000_eur.bed, g1000_eur.bim, g1000_eur.fam
  - If input file is PED then: --file <filename (without extension)>
  - In input file is binary PED then: --bfile <filename (without extension)>

- Setting output
  - --out <filename>

- Convert PED to BED
  - plink --file g1000_eur --make-bed --out new_g1000
  - Here --make-bed if the flag to generate BED file

- Getting help
  - plink --help

Simple commands – selecting samples

- Selecting a set of samples for downstream analysis
  - plink --file g1000_eur --keep mysamples.txt --out new_g1000
  - --keep: Excludes all samples not named in the file

- Removing a set of samples
  - plink --file g1000_eur --remove removesamples.txt --out new_g1000
  - --remove: Excludes all samples named in the file

- Additional sample-level selection flags
  - --keep-fam: excludes all families not named in the file
  - --remove-fam: excludes all families named in the file

Simple commands – selecting SNPs

- Selecting a set of SNPs
  - Selecting by SNP id: plink --bfile g1000_eur --extract mysnps.txt --make-bed --out new_g1000
  - Selecting by genomic range: plink --bfile g1000_eur --extract range mysnprange.txt --make-bed --out new_g1000

- Removing a set of SNPs for downstream analysis
  - Excluding by SNP id: plink --file g1000_eur --exclude removesnps.txt --make-bed --out new_g1000
  - Excluding by genomic range: plink --file g1000_eur --exclude range removesnprange.txt --make-bed --out new_g1000

Quality control

- QC prior to GWAS analysis is very important

Published: 21 July 2011

Paper on genetics of longevity retracted

Heidi Ledford

Nature (2011) | Cite this article

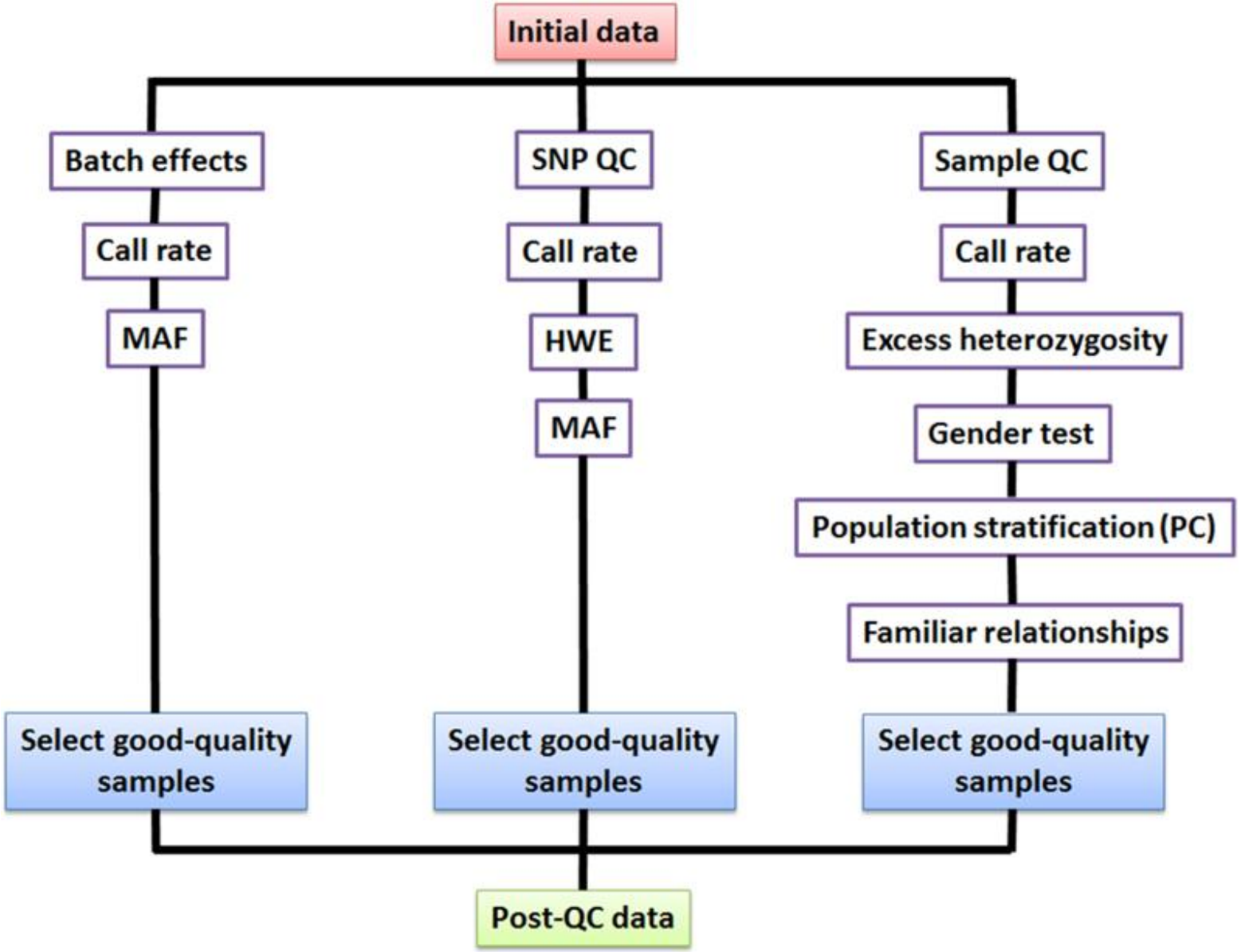
290 Accesses | **1** Citations | **120** Altmetric | Metrics

Technical problems mar study of centenarians.

After online publication of our Report "Genetic signatures of exceptional longevity in humans" (1), we discovered that technical errors in the Illumina 610 array and an inadequate quality control protocol introduced false-positive single-nucleotide polymorphisms (SNPs) in our findings. An independent laboratory subsequently performed stringent quality control measures, ambiguous SNPs were then removed, and resultant genotype data were validated using an independent platform. We then reanalyzed the reduced data set using the same methodology as in the published paper. We feel the main scientific findings remain supported by the available data: (i) A model consisting of multiple specific SNPs accurately differentiates between centenarians and controls; (ii) genetic profiles cluster into specific signatures; and (iii) signatures are associated with ages of onset of specific age-related diseases and subjects with the oldest ages. However, the specific details of the new analysis change substantially from those originally published online to the point of becoming a new report. Therefore, we retract the original manuscript and will pursue alternative publication of the new findings.

Quality control

- QC done at
  - Individual-level or Sample-level
  - Marker-level or SNP-level



Sample-level QC

1. Exclude samples with low genotyping rate (possibly due to low quality DNA)
   - --mind <threshold>
   - Example: --mind 0.1 remove individuals with missing genotype rate more than 10%
2. Exclude samples that exhibit discrepancy between recorded sex with genotyped sex
   - --check-sex
3. Exclude samples with high heterozygosity rate
   - --het
   - Remove samples that deviate ±3 SD from sample heterozygosity
4. Remove samples that are related
   - --genome --min <threshold>
   - Done using pairwise identify-by-decent (IBD)
   - IBD = 1 => identical; 0.5 => 1st degree relatives; 0.25 => 2nd degree; 0.125 => 3rd degree
5. Exclude samples with genetic ancestry inconsistent with the ancestry of the population being studied

Note: Steps 3 - 5 are generally done with high quality pruned SNPs (to be discussed soon)

SNP-level QC

- Remove SNPs with MAF < 0.05 or 0.01
  - --maf <threshold>
  - E.g: --maf 0.05

- Removing missing SNPs
  - --geno
  - The SNP is missing in samples

- Remove SNPs not in Hardy-Weinberg equilibrium (i.e., HWE p < 1e-6)
  - --het <threshold>
  - E.g.: --het 1e-06

QC

- Heterozygosity:
  - Very high or low heterozygosity rates in individuals could be due to DNA contamination or high levels of inbreeding
  - Therefore, samples with extreme heterozygosity are typically removed as QC
- HWE:
  - Poor-quality genotyping can result in heterozygotes being called as homozygotes, generating more homozygotes than expected
  - Setting $P < 10^{-6}$ as the threshold implies one SNP per million will be removed when HWE holds
  - Only SNPs extremely discordant with HWE should be removed as mild HWE may also be due to processes related to disease

SNP Pruning

- Process of removing highly correlated SNPs
- We do this when we want to examine heterozygosity or ancestry or create genomic principal components
- Command:
  - plink --indep-pairwise <window> <step> <rsq> --bfile g1000_eur --make-just-fam --out g1000_eur_prunedsnps
  - Window: fixed number of SNPs assessed at a time
  - Step: Sliding window number of SNPs
  - Rsq or r^2: squared correlations

Step 1: window = 6

Step 2: window = 6
Step = 3

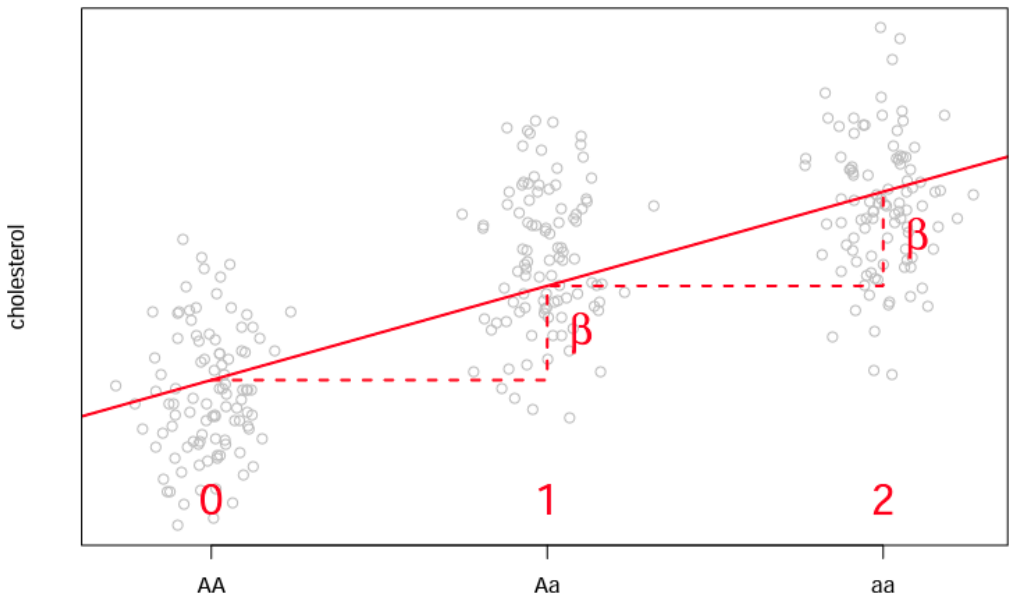
Step 3: window = 6
Step = 3

Association testing

- GWAS models a single genetic marker (e.g., a SNP) as predictor in the model and the quantitative phenotype as the response along with other relevant covariates (such as age, sex, etc.) using regression
- Linear regression
  - Conducted when the phenotype is continuous variable (height)
  - --linear
- Logistic regression
  - Conducted when the phenotype is binary (case/control)
  - --logistic

- Most GWAS conduct single SNP association testing with linear regression assuming an additive model

$$y = \beta_0 + \beta \times \#\text{minor alleles}$$



Linear regression of SNPs : Additive model

Results

- PLINK association testing produces a file with the following columns

CHR	Chromosome
SNP	SNP identifier
BP	Physical position (base-pair)
A1	Tested allele (minor allele by default)
TEST	Code for the test (see below)
NMISS	Number of non-missing individuals included in analysis
BETA/OR	Regression coefficient (--linear) or odds ratio (--logistic)
STAT	Coefficient t-statistic
P	Asymptotic p-value for t-statistic

Benefits and limitations of GWAS

Benefits

- Successful in identifying novel variant–trait associations
- Provides insights into novel biological mechanisms
- Aid in clinical translation
- Provide insight into ethnic variation of complex traits
- Enable study of low-frequency and rare variants
- Can identify novel monogenic and oligogenic disease genes

Limitations

- Penalised by multiple testing burden
- Explain only a modest fraction of the missing heritability
- Do not necessarily pinpoint causal variants and genes
- Cannot identify all genetic determinants of complex traits
- Limited clinical predictive value
- Affected by population structure and cryptic-relatedness
- Majority are focussed on European ancestry

nature > nature reviews genetics > review articles > article

Review Article | Published: 08 May 2019

Benefits and limitations of genome-wide association studies

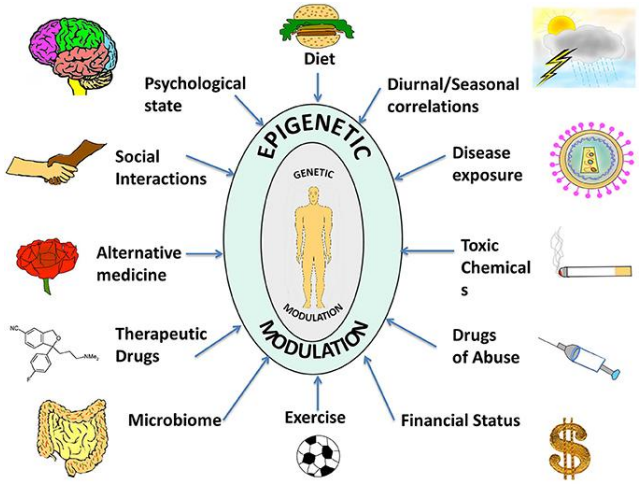
Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré & David Meyre

Nature Reviews Genetics **20**, 467–484 (2019) | Cite this article

87k Accesses | **728** Citations | **194** Altmetric | Metrics

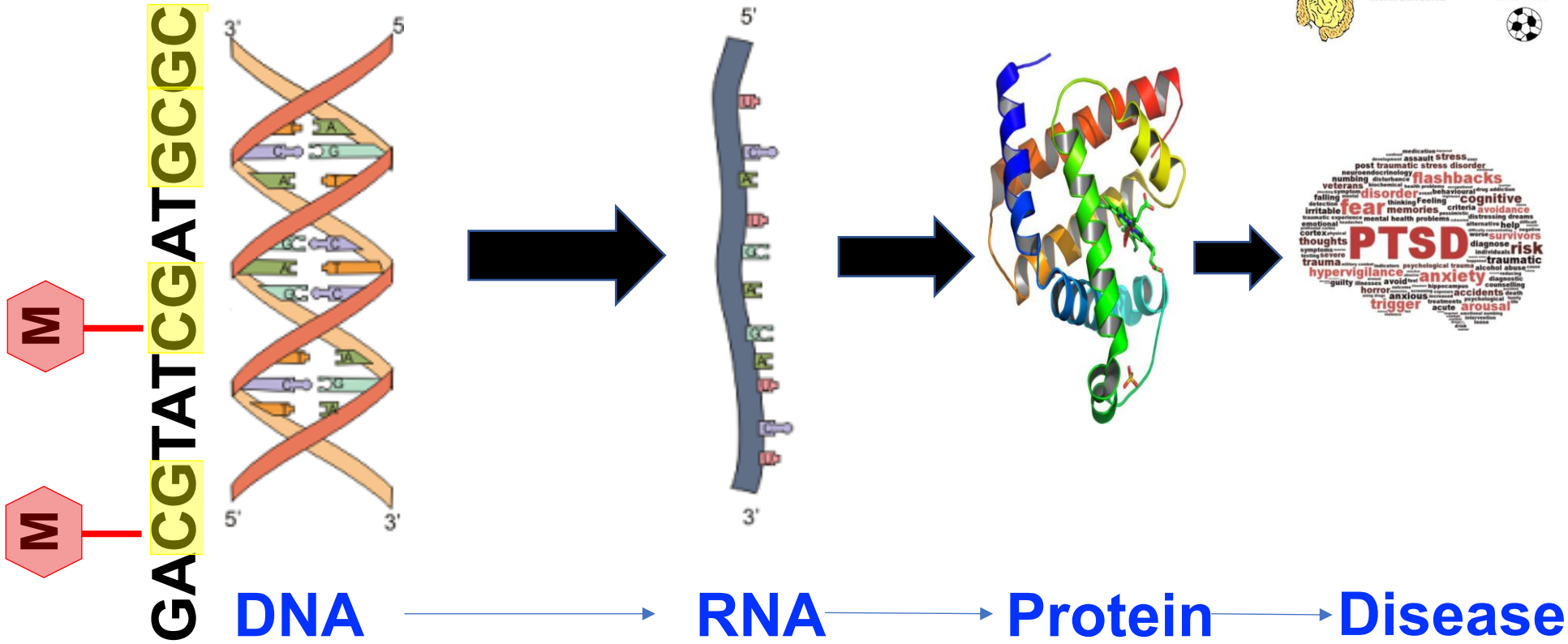
DNA methylation

Measuring the influence of the environment on the genome

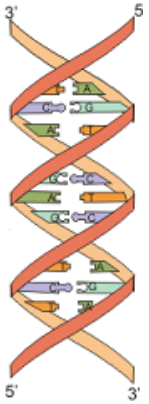


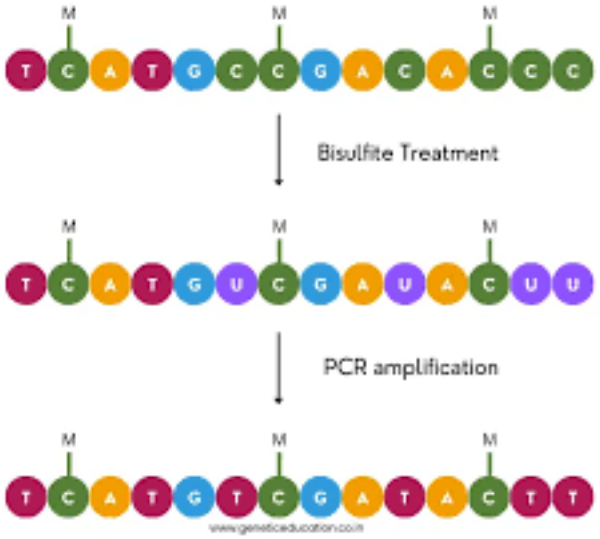
DNA methylation (Epigenetics)

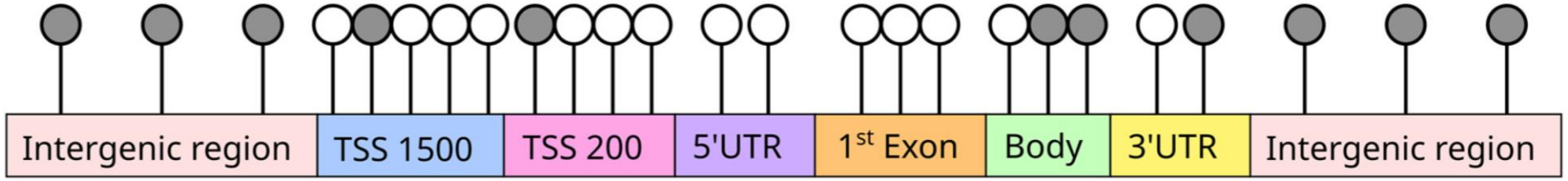
Gene-expression



Illumina EPIC array (DNAm)

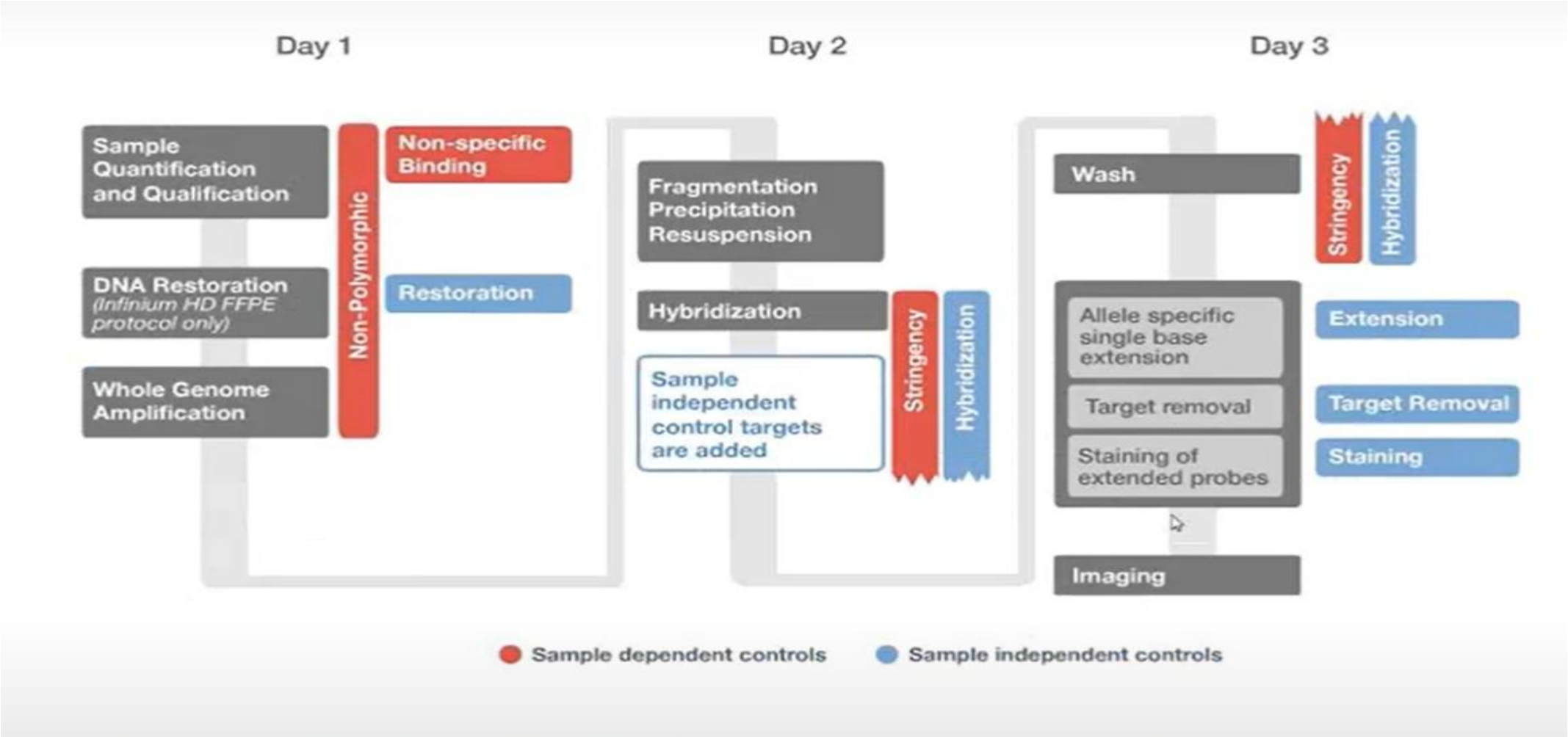




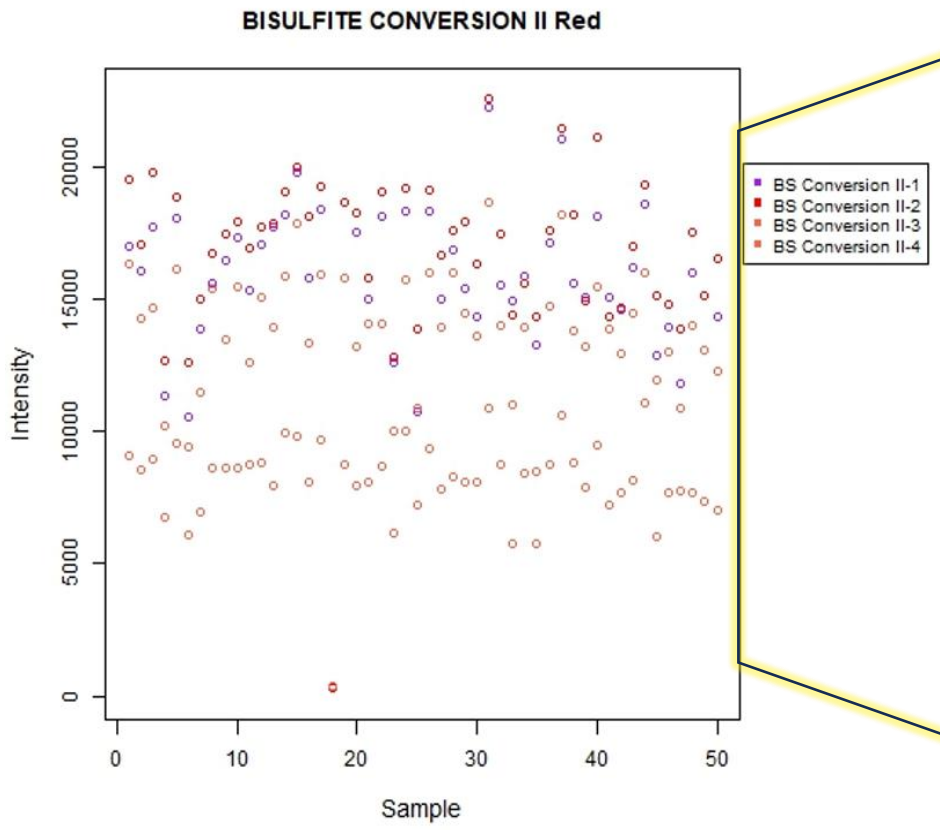


http://journal.frontiersin.org/article/10.3389/fcell.2014.00049/full

Infinium Assay Workflow



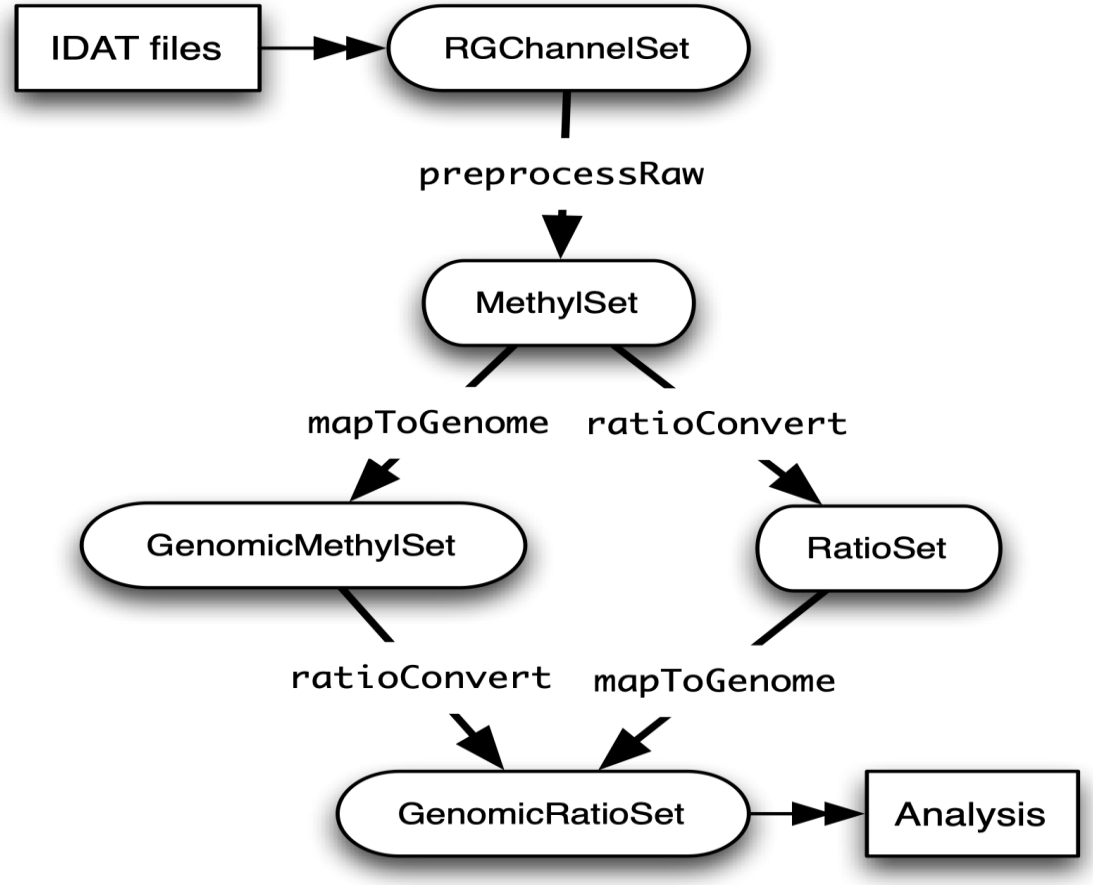
System Controls: ENmix



Purpose	Name	Number on the Array	Evaluate Green (GRN)	Evaluate Red (RED)	Expected Intensity
Bisulfite conversion I	BC conversion I C1, C2, C3	3	+	-	High
Bisulfite conversion I	BC conversion I U1, U2, U3	3	+	-	Background
Bisulfite conversion I	BC conversion I C4, C5, C6	3	-	+	High
Bisulfite conversion I	BC conversion I U4, U5, U6	3	-	+	Background
Bisulfite conversion II	BC conversion II 1, 2, 3, 4	4	-	+	High
Specificity I	GT perfect match 1, 2, 3 (PM)	3	+	-	High
Specificity I	GT mismatch 1, 2, 3 (MM)	3	+	-	Background
Specificity II	Specificity 1, 2, 3	3	-	+	High
Non-Polymorphic	NP (A), (T)	2	-	+	High
Non-Polymorphic	NP (C), (G)	2	+	-	High
Negative	Average[1]	600	+	+	Background
Negative	StdDev[2]		+	+	Background

Infinium controls do not use specific thresholds.
To remove samples, we rely on multiple metrics.

Class Structure: Minfi



Generalized Linear Model (GLM): Limma

Methylation ~ Age + ε

```
# Create a design matrix for the linear model:
# Columns:
# # [1] "SID"          "Tissue"       "Sex"          "Race"         "Ethnicity"
#  [6] "Age"          "GID"          "Description" "Basename"     "filenames"
# [11] "CD8T"         "CD4T"         "NK"           "Bcell"        "Mono"
# [16] "Neu"          "CD8T.1"       "CD4T.1"       "NK.1"         "Bcell.1"
# [21] "Mono.1"       "Neu.1"
design <- model.matrix(~ Age + Sex + Race + CD4T + Mono,
                       data = phenoFV)

fit <- lmFit(beta, design)
fit <- eBayes(fit, trend= TRUE)

tt <- topTable(fit, coef = 2, number = Inf)
tt
```

References

1. Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Usinglme4. Journal of Statistical Software, 67(1). https://doi.org/10.18637/jss.v067.i01
2. Du, P., Zhang, X., Huang, CC. et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 11, 587 (2010). https://doi.org/10.1186/1471-2105-11-587
3. Fortin, JP., Labbe, A., Lemire, M. et al. Functional normalisation of 450k methylation array data improves replication in large cancer studies. Genome Biol 15, 503 (2014). https://doi.org/10.1186/s13059-014-0503-2
4. Hansen, K. D., & Fortin, J.-P. (2025). The minfi User's Guide. In https://bioconductor.org/packages/devel/bioc/vignettes/minfi/inst/doc/minfi.html
5. Heiss, J. (2013). Recommended Work Flow. In https://hhhh5.github.io/ewastools/articles/exemplary_ewas.htmlHorvath, S. (2013). DNA methylation age of human tissues and cell types. Genome Biol, 14(10), R115. https://doi.org/10.1186/gb-2013-14-10-r115
6. Koestler, D.C., Jones, M.J., Usset, J. et al. Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). BMC Bioinformatics 17, 120 (2016). https://doi.org/10.1186/s12859-016-0943-7

References

7. Maksimovic, J., Phipson, B., & Oshlack, A. (2017). A cross-package Bioconductor workflow for analysing methylation array data. F1000Research, 5. https://f1000research.com/articles/5-1281
8. Marschner, I. C. (2011). glm2: Fitting Generalized Linear Models with Convergence Problems. R Journal, 3(2), 12-15. https://journal.r-project.org/archive/2011/RJ-2011-012/RJ-2011-012.pdf9. Pelegri , D., & Gonzalez, J. R. (2015). Chronological and gestational DNAm age estimation using different methylation-based clocks. In https://bioconductor.org/packages/release/bioc/vignettes/methylclock/inst/doc/methylclock.html
9. Peters TJ, Meyer B, Ryan L, Achinger-Kawecka J, Song J, Campbell EM, Qu W, Nair S, Loi-Luu P, Stricker P, Lim E, Stirzaker C, Clark SJ, Pidsley R. Characterisation and reproducibility of the HumanMethylationEPIC v2.0 BeadChip for DNA methylation profiling. BMC Genomics. 2024 Mar 6;25(1):251. doi: 10.1186/s12864-024-10027-5. PMID: 38448820; PMCID: PMC10916044.
10. Touleimat, N., & Tost, J. (2012). Complete Pipeline for Infinium® Human Methylation 450K BeadChip Data Processing Using Subset Quantile Normalisation for Accurate DNA Methylation Estimation. Epigenomics, 4(3), 325–341. https://doi.org/10.2217/epi.12.21
11. Wang Y, Hannon E, Grant OA, Gorrie-Stone TJ, Kumari M, Mill J, Zhai X, McDonald-Maier KD, Schalkwyk LC. DNA methylation-based sex classifier to predict sex and identify sex chromosome aneuploidy. BMC Genomics. 2021 Jun 28;22(1):484. doi: 10.1186/s12864-021-07675-2. PMID: 34182928; PMCID: PMC8240370.