Genome-Wide Association Studies

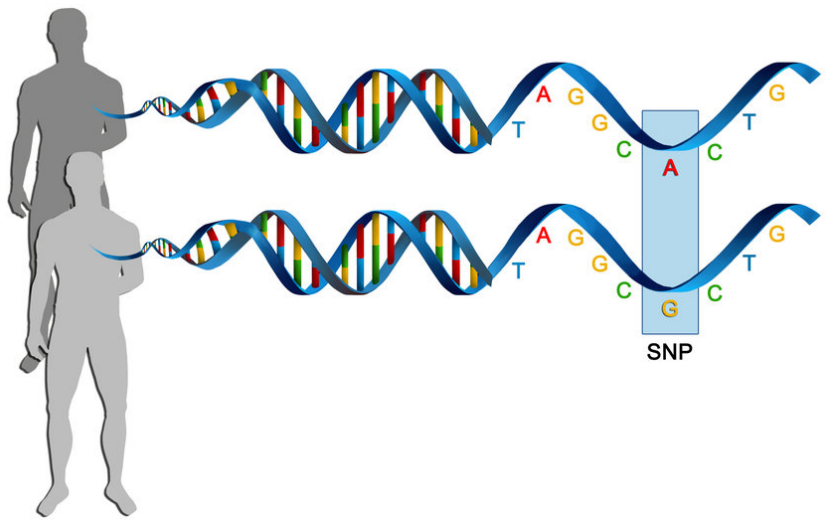
Data Processing and Imputation Pipeline

Beáta Éva Nagy
Supervisors: András Botond Nemes, Dr. Gábor Firneisz

January 2026

Genome-Wide Association Studies (GWAS)

- **Goal:** Find genetic variants statistically associated with diseases/traits (phenotypes)
  - Focus on Single-Nucleotide Polymorphisms (SNPs)



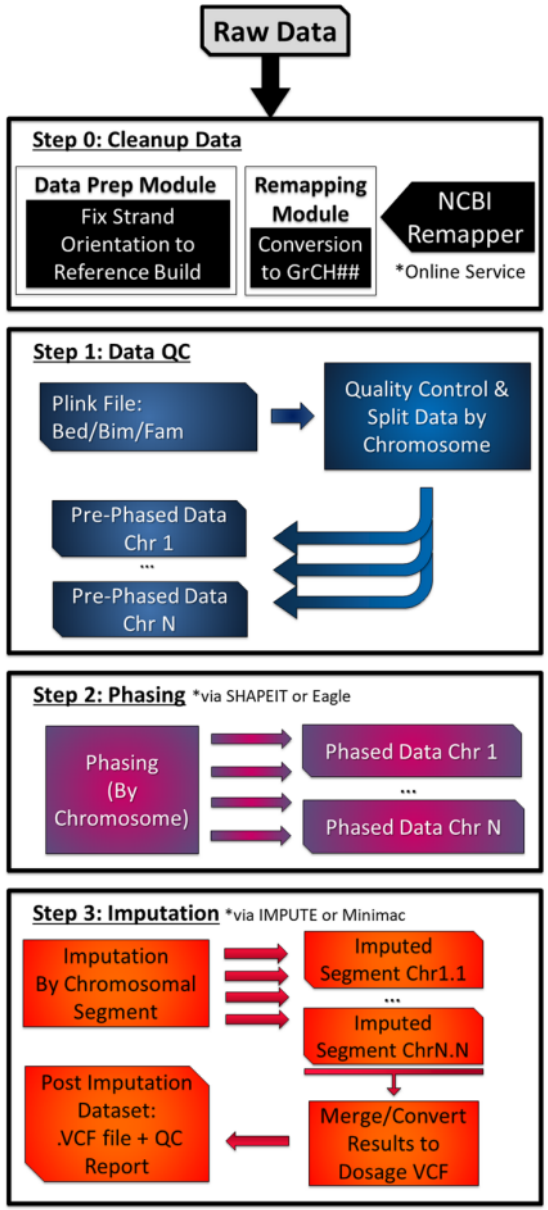
Source: https://www.researchgate.net/figure/A-single-nucleotide-polymorphism-SNP_fig1_359016330

Genome-Wide Association Studies (GWAS)

- **Method:** Statistical hypothesis testing
  - Millions of SNPs tested
  - Distinguish real biological associations from chance
  - Regression models adapted to the trait type (logistic/linear)

- **Impact:** identifying genetic risk factors
  - Since 2005, associations for type 2 diabetes, Crohn's disease, obesity etc.

Project Overview



Source: https://doi.org/10.1186/s12859-019-2964-5

Computational Tools

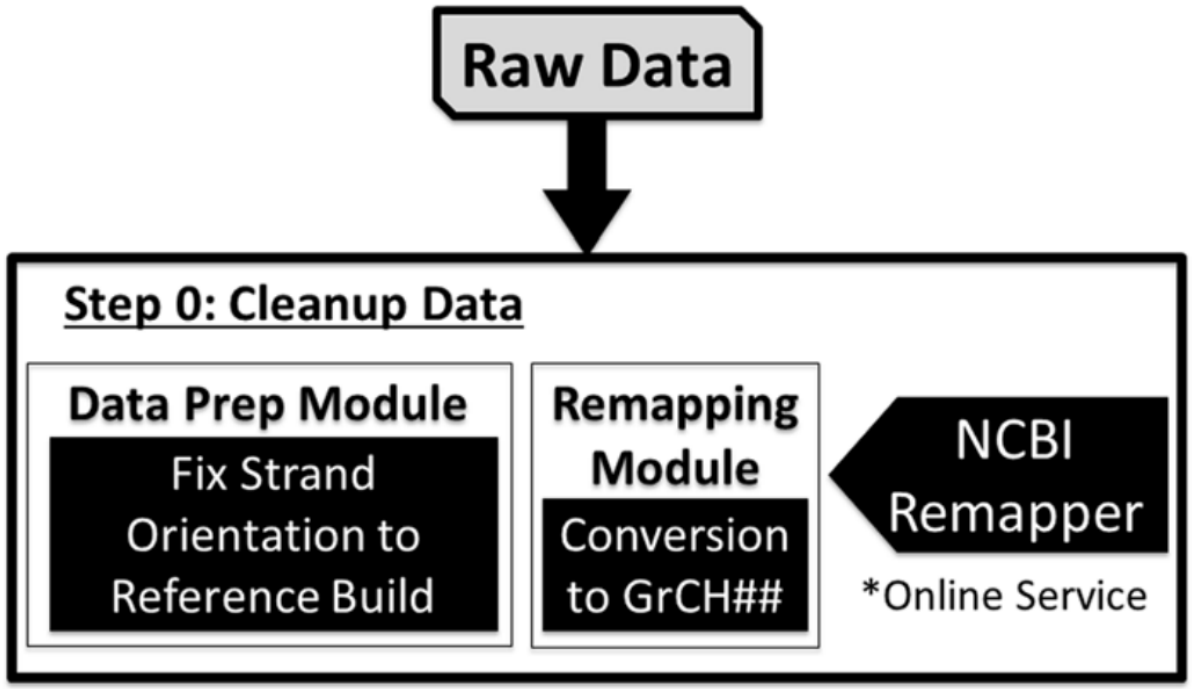
- PLINK: Quality control + statistical analysis
- BCFtools: Efficient data manipulation
- CrossMap: Coordinate conversion between reference assemblies
- SHAPEIT4: Haplotype phasing
- Minimac4: Genotype imputation

Raw Data

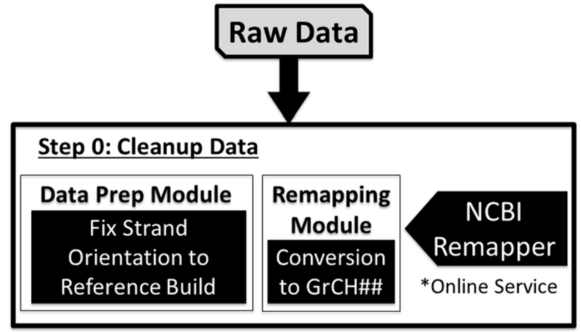
- Genetic databases
  - 4 ethnic groups, 5,000 mother – child pairs
  - Genotype data: 600,000 – 1.2 million SNPs per person
  - + Phenotype data

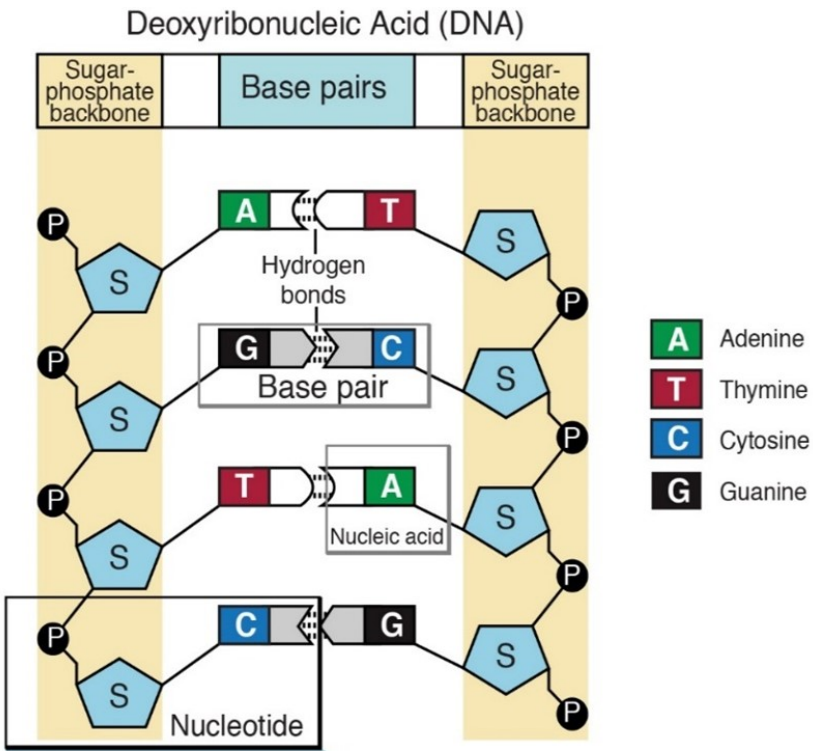
- **Aim:** build two multi-ethnic datasets (mothers + children)

0. Data Cleanup



0.1 Data Cleanup - Strand Orientation Fix

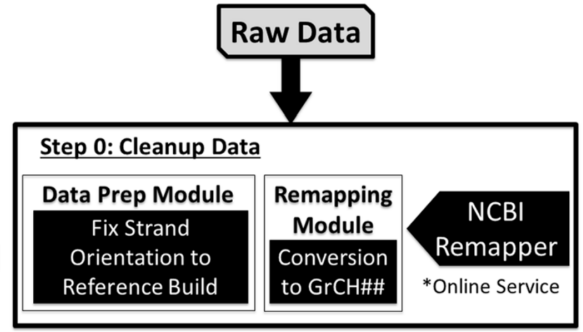


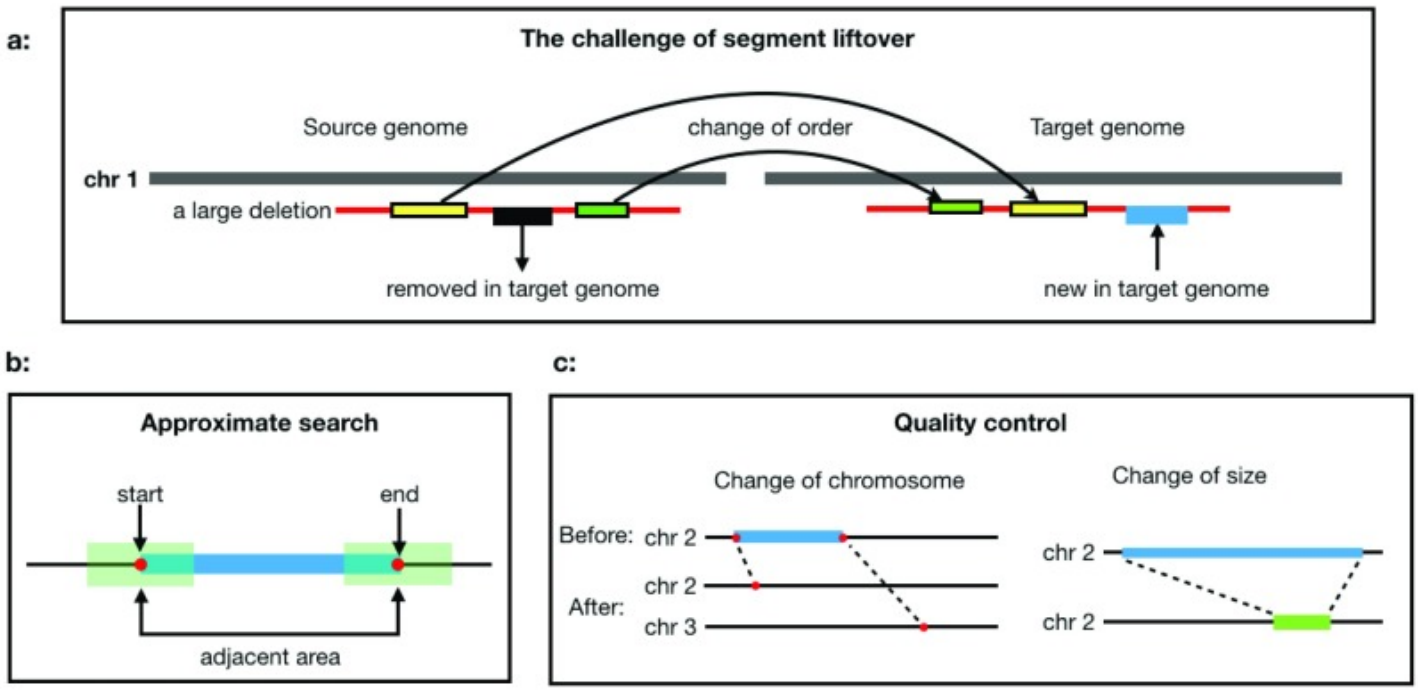


Source: `https://passel2.unl.edu/view/lesson/6f214d098527/4`

DNA: forward + reverse strand orientation $\rightarrow$ one consistent representation

0.2 Data Cleanup - Genome Remapping

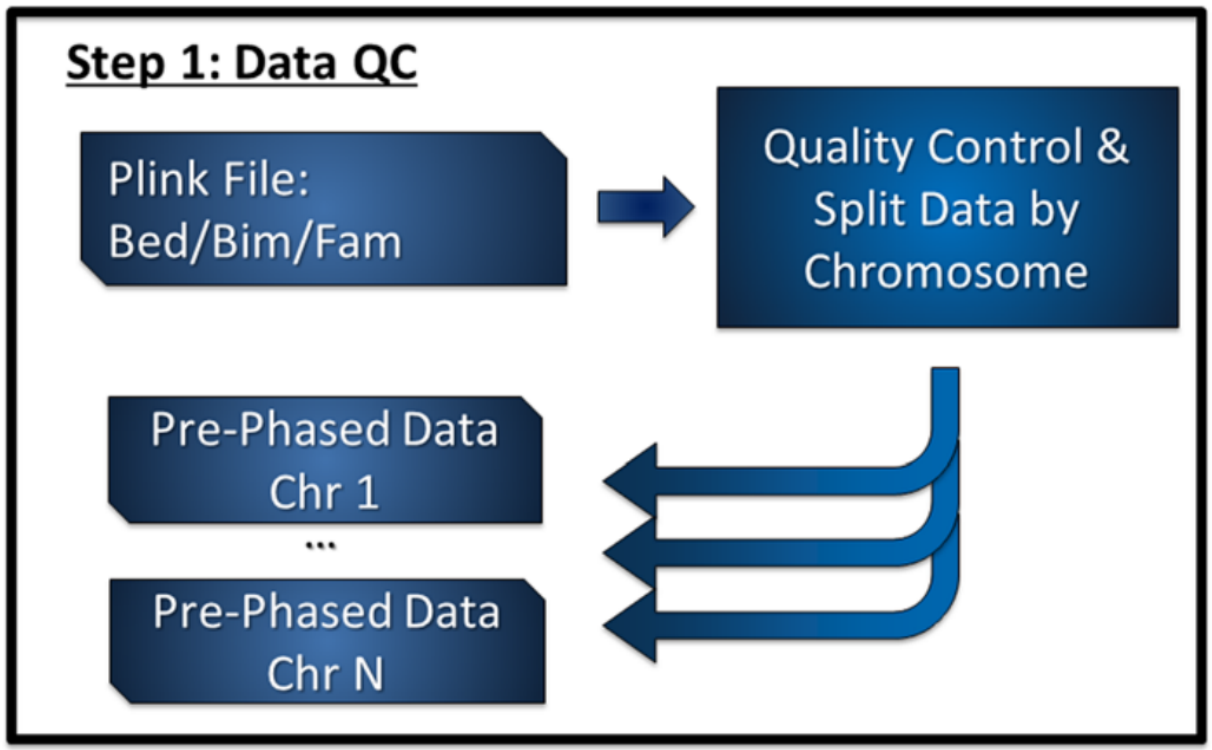




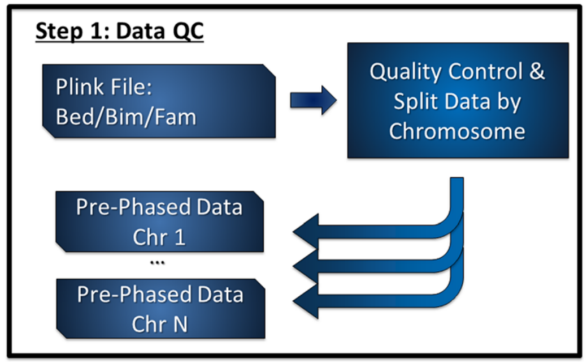
Source: `https://pmc.ncbi.nlm.nih.gov/articles/PMC5998006/`

Chromosomal coordinates change over time $\rightarrow$ genome build version update

1. Quality Control



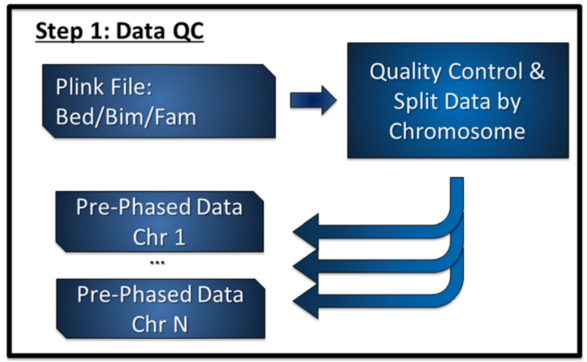
1.0 Input Data



Three standardized binary files:

- BED file: Compressed binary, **genotype data**
- BIM file: Genetic variant information (chromosome, **SNP** ID, position, alleles)
- FAM file: Individual information (**family** identification, sex, **phenotype**)

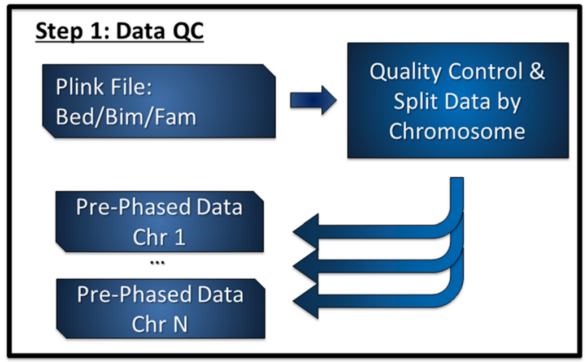
1.1 Quality Control Steps



Filtering human + genotyping errors:

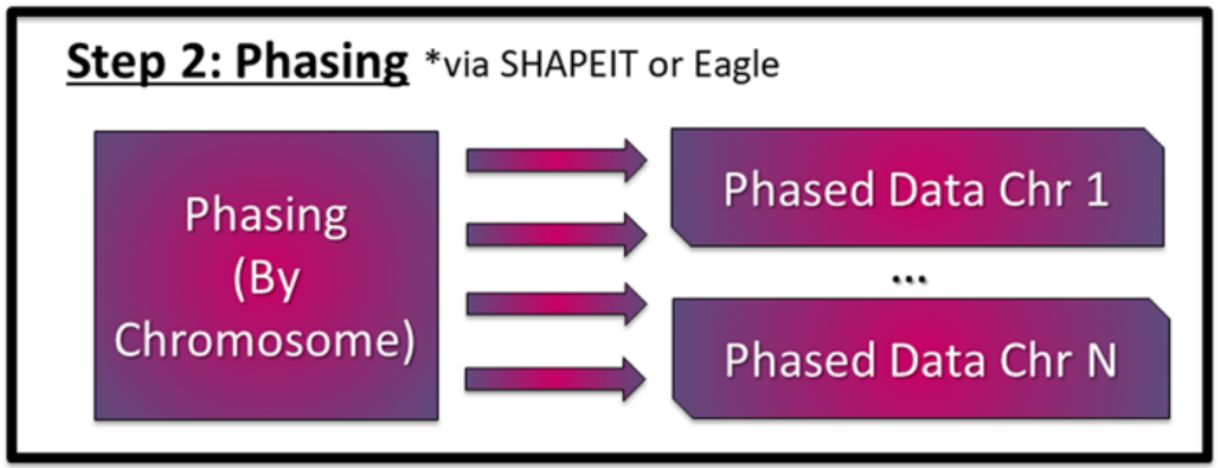
1. Remove SNPs with $> 2\%$ missingness
2. Remove individuals with $> 2\%$ missingness
3. Keep only biallelic SNPs (A/C/G/T substitution)

1.1 Quality Control Steps

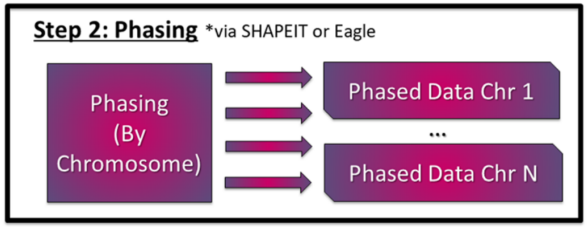


4. Chromosome Filtering: Remove sex + mitochondrial
5. Sorting and Indexing: Order by position within chromosomes
   - Indexing: Faster processing of large databases
6. Variant Normalization: Consistent representation
7. Minor (less frequent) Allele Count > 5 filter
   - Remove rare variants

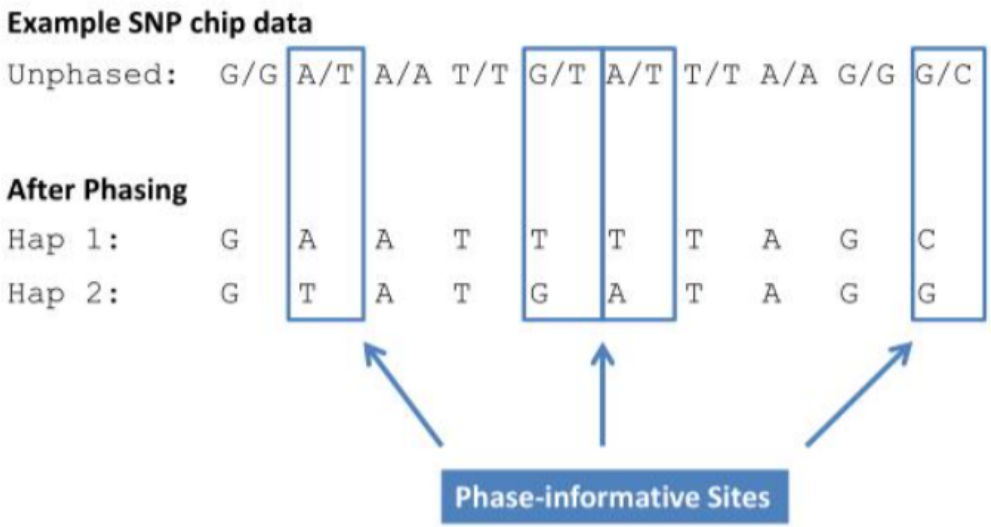
2. Haplotype Phasing



2. Haplotype Phasing

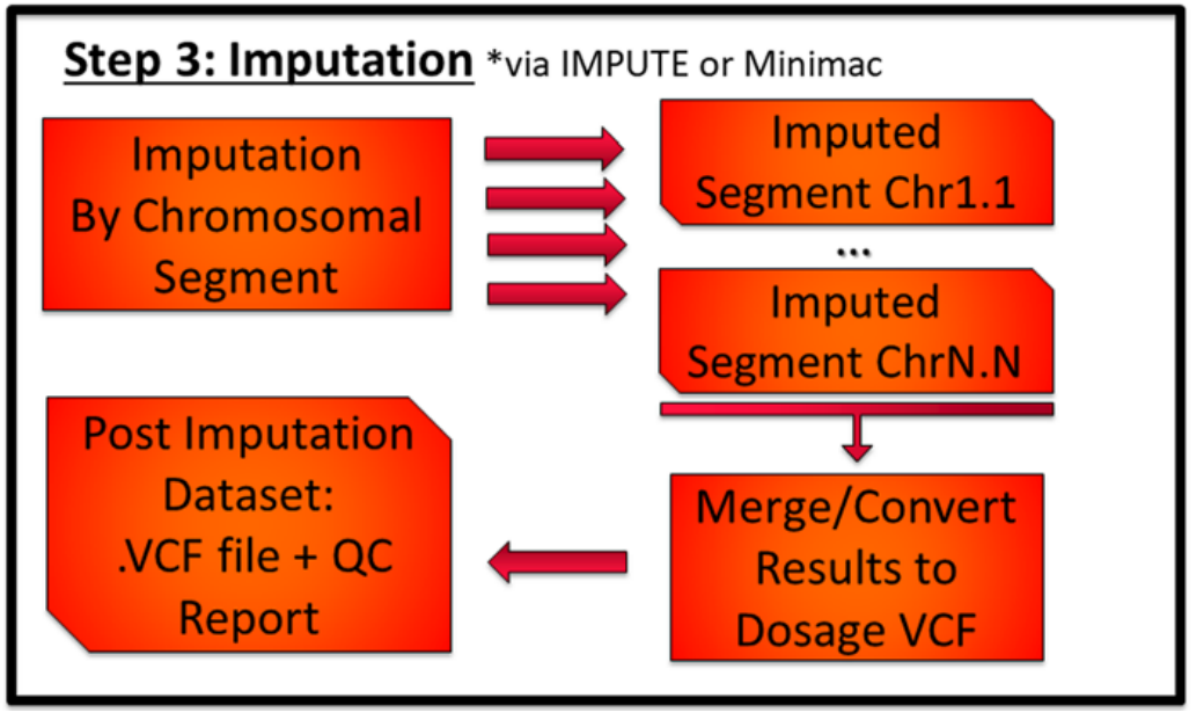


- Haplotype: Variants inherited together
- Linkage disequilibrium: Non-random association of nearby variants
- $\longrightarrow$ splitting: one haplotype from each parent

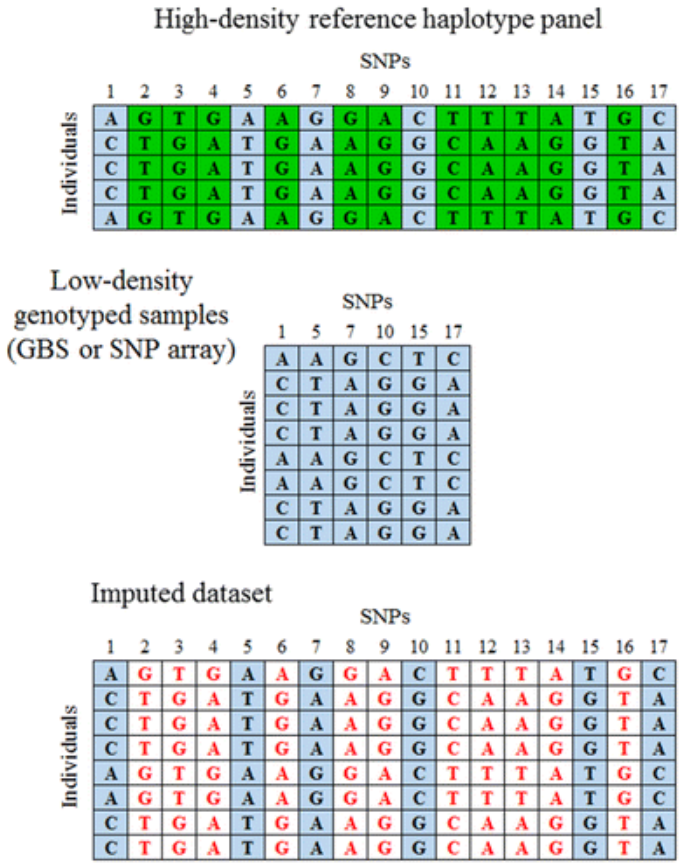


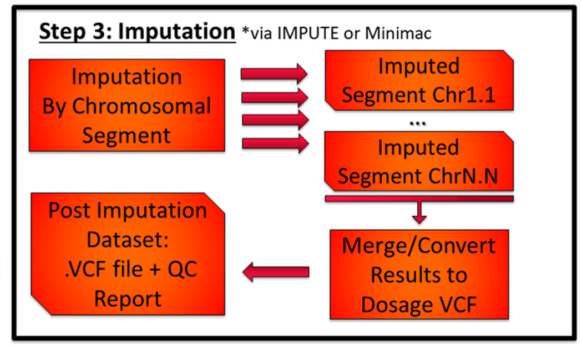
Source: https://www.slideserve.com/beyla/genotype-phasing-and-imputation-in-1x-sequencing-data

3. Imputation



3.1 Imputation



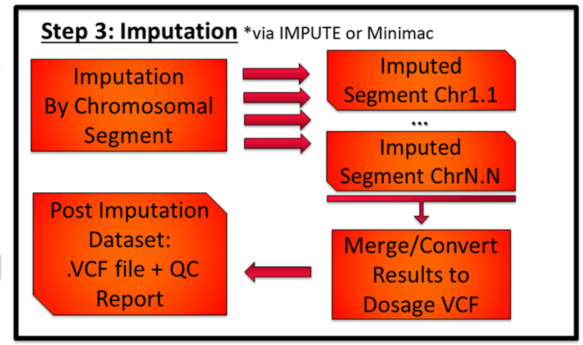


Source:

`https://www.researchgate.net/figure/Untyped-genotype-imputation-using-haplotype-reference-panel_fig4_322610020`

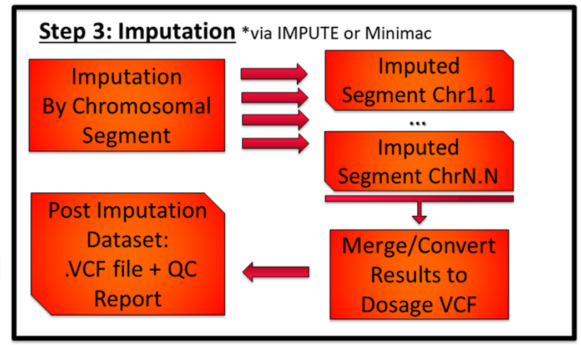
Predict missing variants based on known reference panels

3.2 Imputation Post-Processing



1. Quality metric: R^2 (squared correlation between true and imputed genotypes) $\rightarrow$ SNPs with $R^2 < 0.3$ excluded
2. Remove SNPs not in all ethnic groups
3. Merge ethnicities $\rightarrow$ one multi-ethnic mothers + one multi-ethnic children dataset
4. remove SNPs with less frequent variant frequency < 0.025 (ethnicity-specific)

3.3 Imputation Results

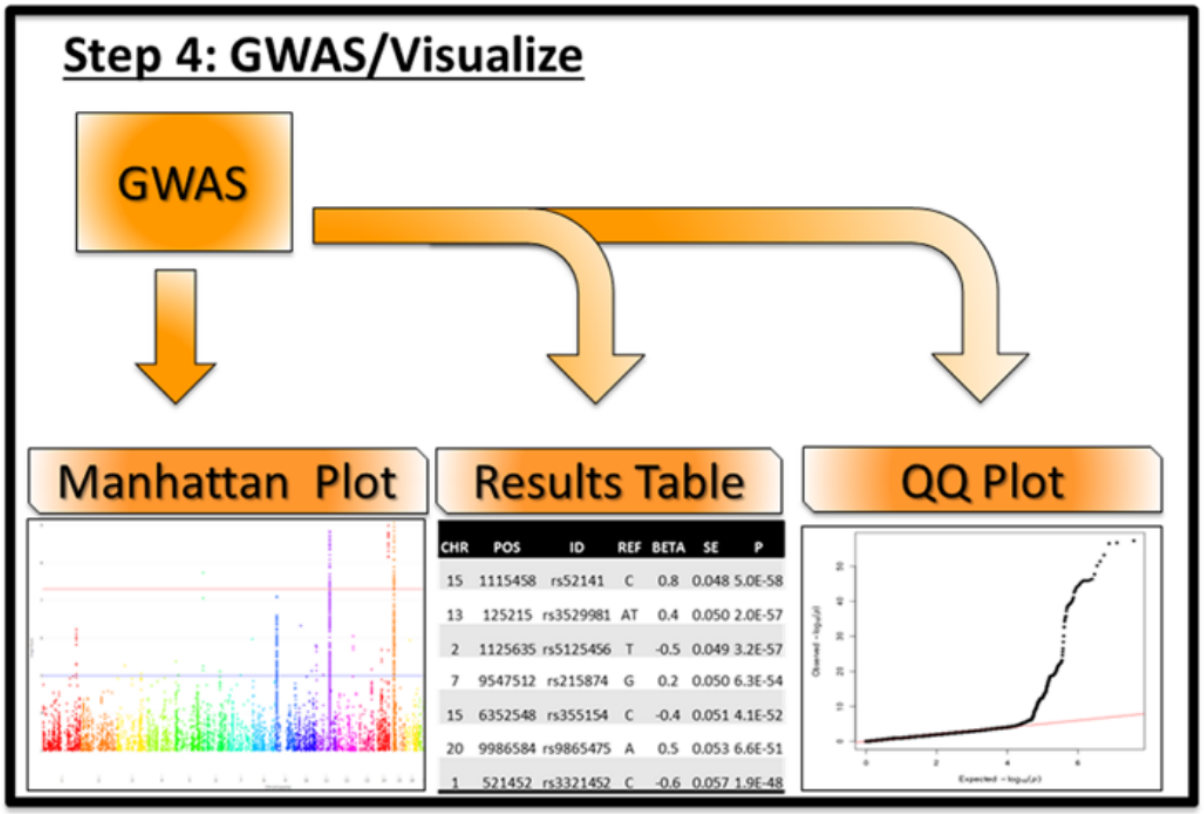


SNP counts:

	European	Hispanic	Thai	Afro-Caribbean
Initial	0.6M	1.0M	1.1M	1.1M
$R^2 = 0.3$	29.0M	32.6M	26.2M	38.6M
$R^2 = 0.8$	11.0M	18.0M	9.4M	23.6M

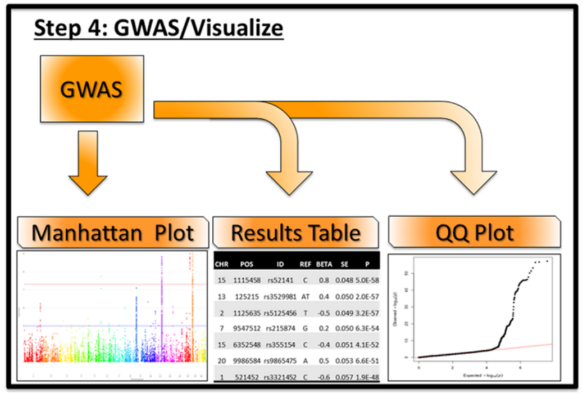
Merged dataset with $R^2 = 0.8$ filter: 18M overlapping SNPs

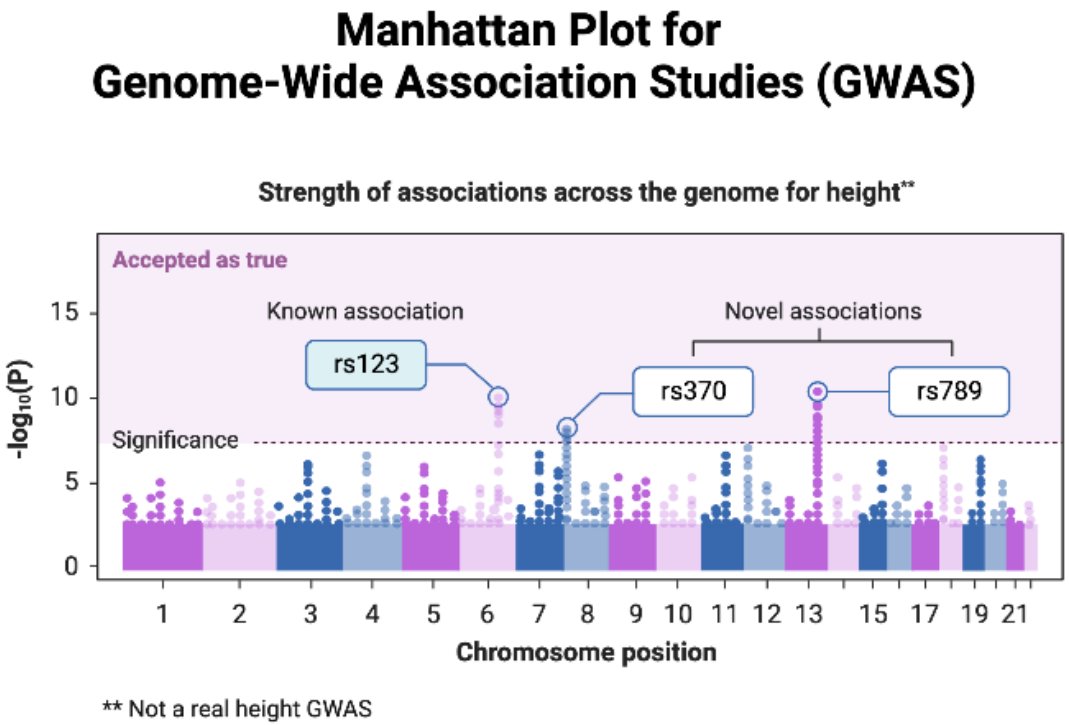
Next Steps


Step 4: GWAS/Visualize
GWAS
Manhattan Plot
Results Table
QQ Plot
CHR POS ID REF BETA SE P
15 1115458 rs52141 C 0.8 0.048 5.0E-58
13 125215 rs3529981 AT 0.4 0.050 2.0E-57
2 1125635 rs5125456 T -0.5 0.049 3.2E-57
7 9547512 rs215874 G 0.2 0.050 6.3E-54
15 6352548 rs355154 C -0.4 0.051 4.1E-52
20 9986584 rs9865475 A 0.5 0.053 6.6E-51
1 521452 rs3321452 C -0.6 0.057 1.9E-48

Next Steps

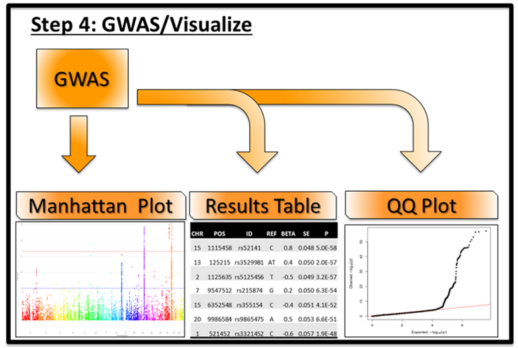
- Run GWAS
- Create Manhattan plot





Source: hhttps://www.biorender.com/template/manhattan-plot-for-genome-wide-association-studies-gwas

Next Steps



- Record significant SNPs ($p < 5 \times 10^{-8}$)
- Analyze using linear or logistic regression
- More complex, interaction GWAS analysis

Thank You for Your Attention!

Declaration of the Use of AI

I used AI tools to make my semester work more accurate in the following aspects:

- Research of biological concepts, reliable sources for the written report
- Review of biological and grammatical accuracy + understandability of the written report
- Troubleshooting during the data engineering steps (during the use of the listed computational tools)
- Improving the visual outline of the written report and presentation