Introduction to Genome Wide Association Study

Alam Ahmad Hidayat

December 2023

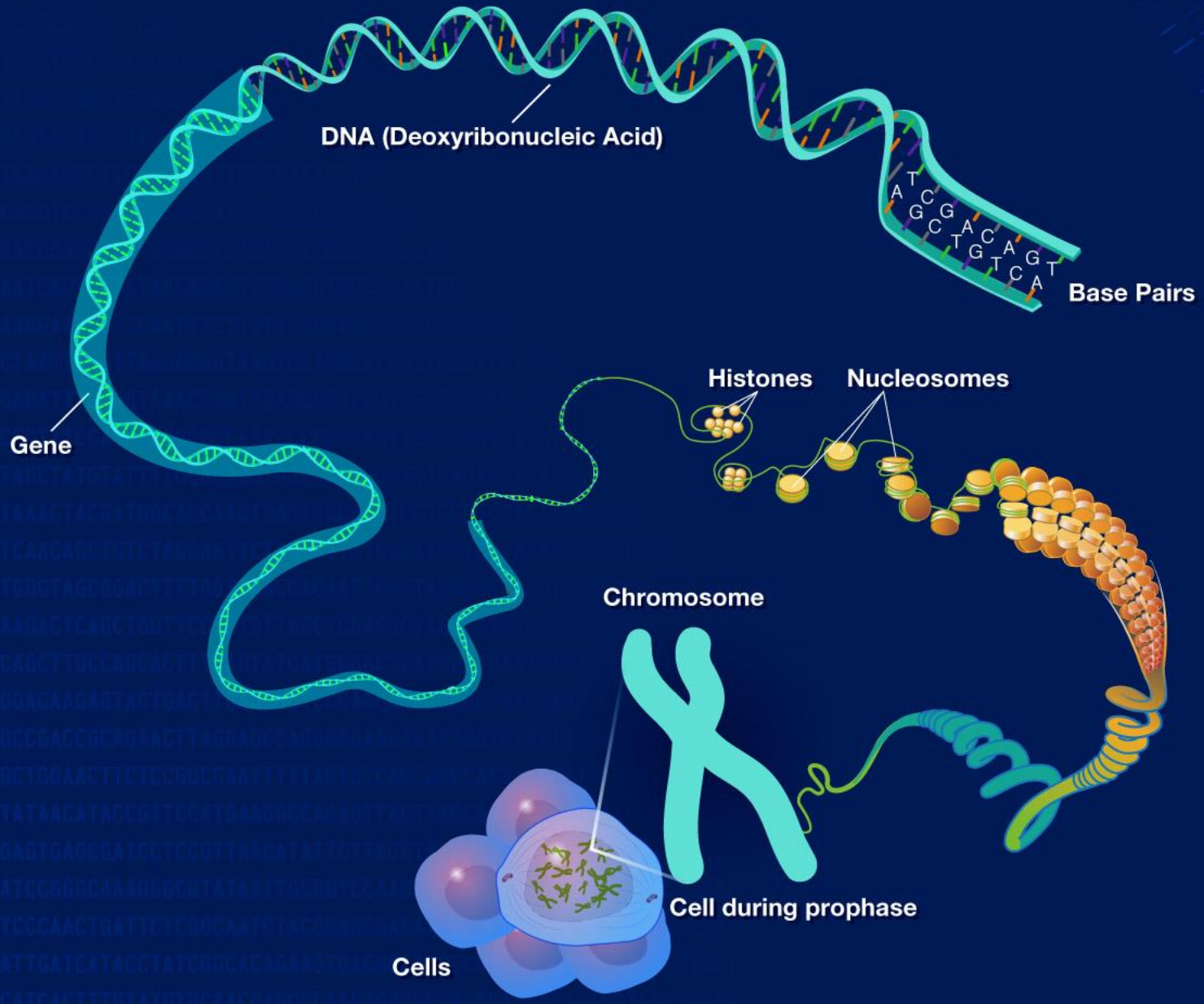
Introduction

- **Genome-wide association studies (GWAS)** test hundreds of thousands of genetic variants across many genomes to find those statistically associated with a specific trait or disease.
- GWAS applications: estimating its heritability, calculating genetic correlations, making clinical risk predictions, informing drug development programmes and inferring causal relationships between risk factors and health outcomes.
- More than 5,700 GWAS have now been conducted and a push for more statistical power has thrust GWAS sample sizes well beyond a million participants.

https://doi.org/10.1038/s43586-021-00056-9

A Brief Guide to Genomics

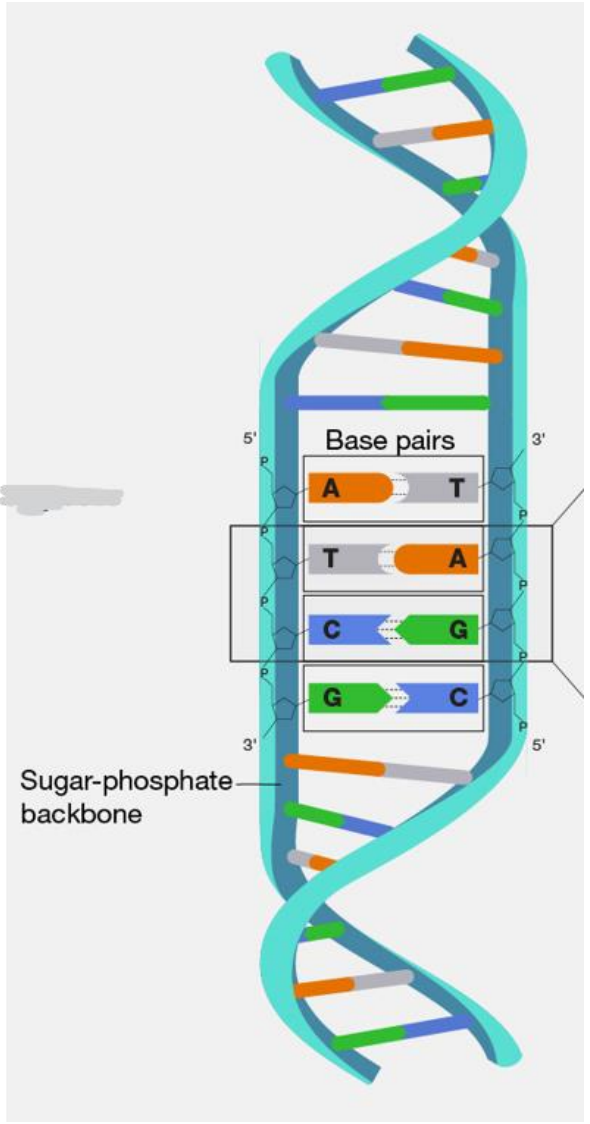
NHGRI FACT SHEETS
genome.gov



NIH National Human Genome Research Institute

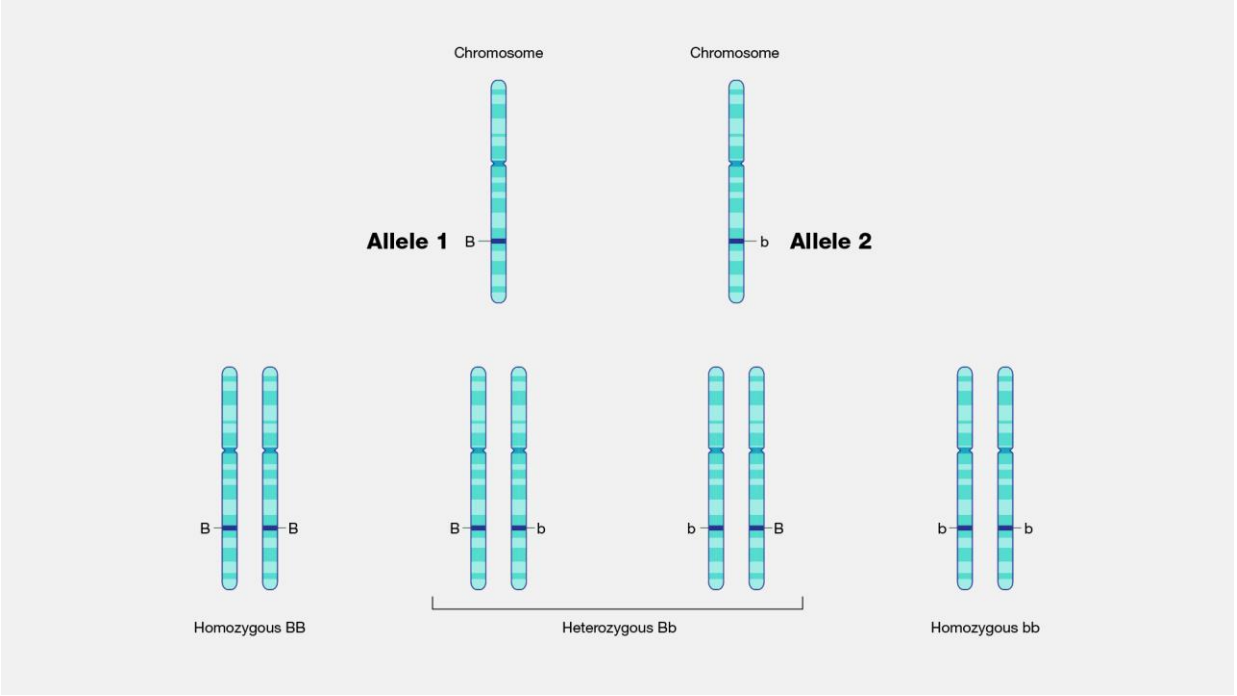
DNA

- Deoxyribonucleic acid (DNA) is the chemical compound that contains the instructions needed to develop and direct the activities of organisms.
- DNAs are made of two twisting, paired strands (double helix).
- Four nucleotide bases: adenine (A), thymine (T), guanine (G), and cytosine (C).
- A always pairs with a T; a C always pairs with a G.
- A gene is a unit of DNA that carries the instructions for making a specific protein or set of proteins.
- The chains of nucleotides in human DNA are wound up and compacted into 46 chromosomes (two sets of 23).
- An organism's complete set of DNA is called its genome (~3 billion DNA base pairs in humans).



Allele

- An allele is one of two or more versions of DNA sequence at a given genomic location
- An individual inherits two alleles, one from each parent.
- If the two alleles are the same: homozygous. If the alleles are different: heterozygous.



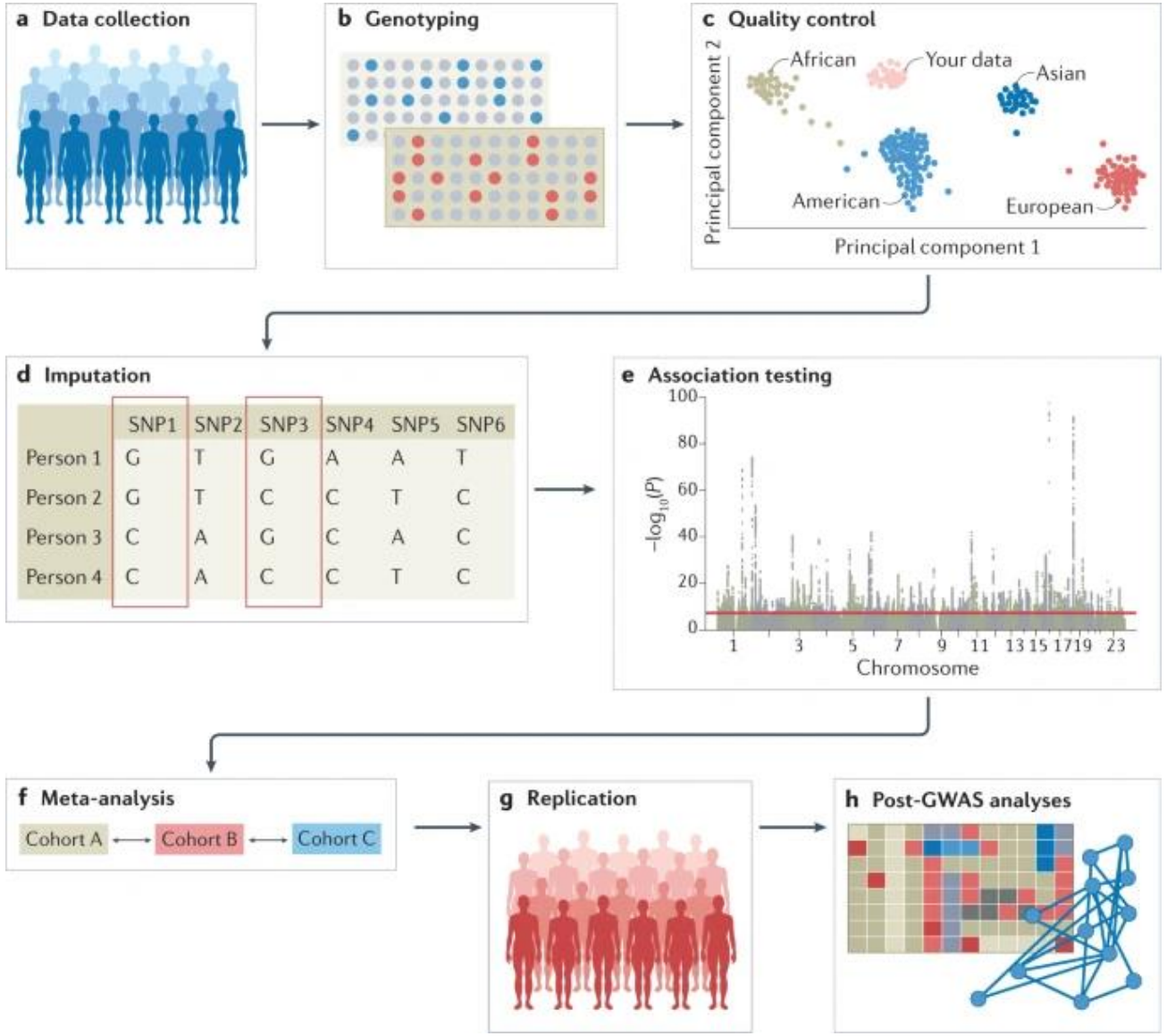
Single Nucleotide Polymorphisms

What is SNP?	https://learn.genetics.utah.edu/content/precision/snips
Allele Frequency	https://mr-dictionary.mrcieu.ac.uk/term/allele/#:~:text=At%20a%20given%20SNP%2C%20the,allele%20occurs%20within%20a%20population.

Sequencing

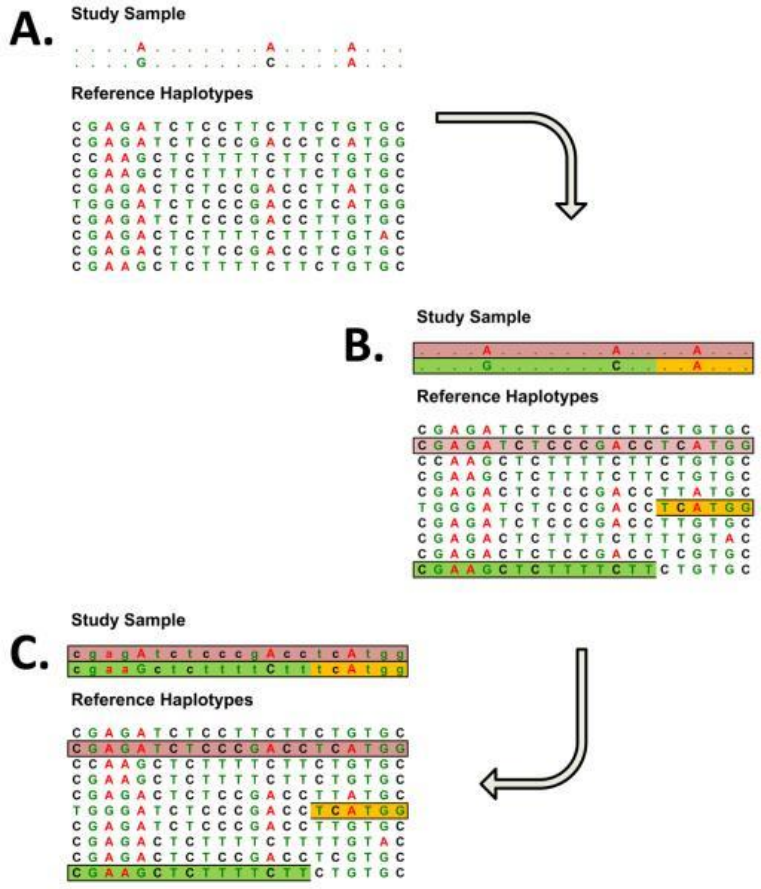
- The collection of DNA and phenotypic information from a group of individuals (such as disease status and demographic information such as age and sex);
- Genotyping of each individual using available GWAS arrays or sequencing strategies:
  - Whole Genome Sequencing
  - Whole Exome Sequencing
  - Microarray Genotyping

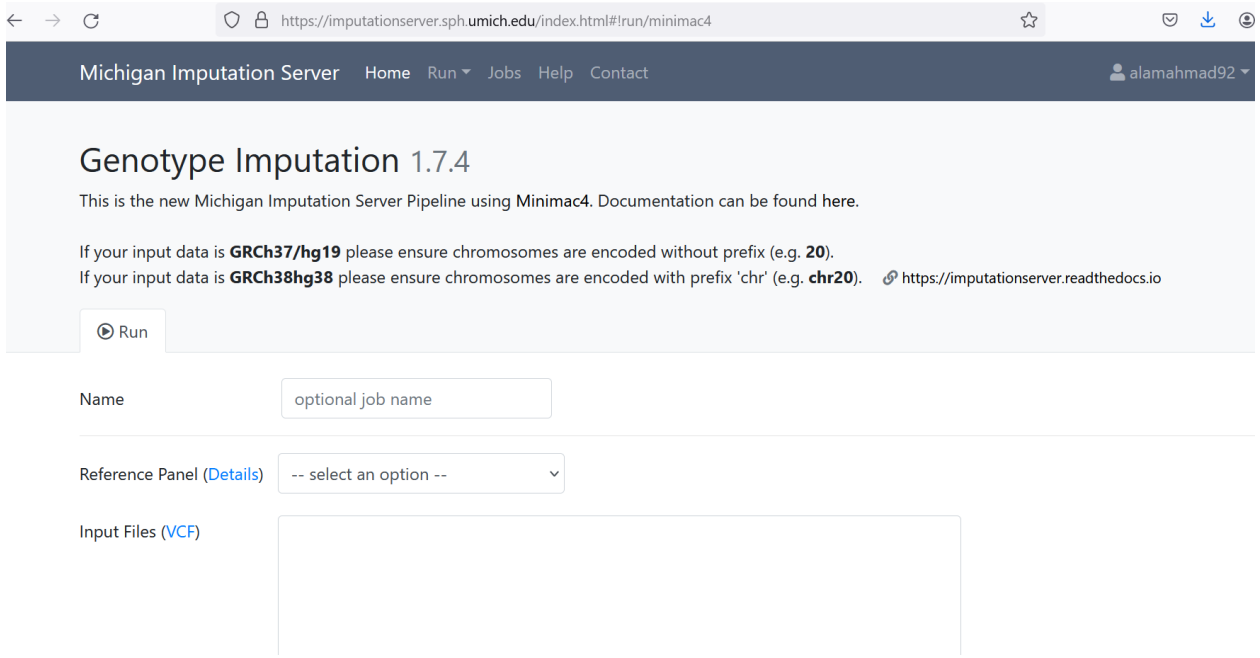
General Workflow of GWAS



https://doi.org/10.1038/s43586-021-00056-9

Genotyping Imputation





https://doi.org/10.1146%2Fannurev.genom.9.081307.164242

A Brief to Regressions

- Two different outcomes:
  - Continuous (height, blood pressure or BMI): Linear regression
  - Binary (the presence or absence of disease): Logistic regression
- Covariates: age, sex and ancestry are included to account for stratification and avoid confounding effects from demographic factors.
- The statistics of each SNP are computed by performing a number of independent regressions.

The general formulation of linear/logistic regressions in GWAS for a SNP (glm/statsmodels notation)

$$y_i \sim \Sigma_j \beta_{cov_j} X_{cov_j} + \beta_{SNP_i} G_{SNP_i}$$

y_i : the phenotypic outcome (binary or continuous)

X_{cov_j} : the j_{th} covariate/confounding variable

G_{SNP_i} : the genotypic value of the i_{th} SNP

Two key statistics:
β_{SNP_i} and its p-value ➡ $OR_{SNP_i} = e^{\beta_{SNP_i}}$

⬇

P-value is usually obtained from Wald's test

Odds Ratio

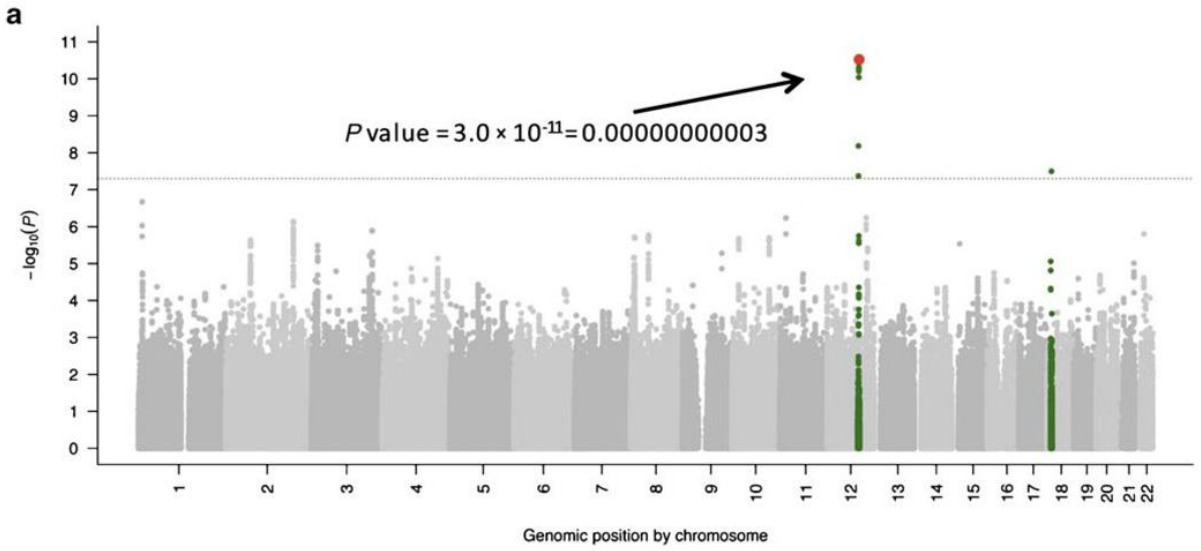
- The concept only works for binary outcomes (or binary phenotype): case/control, yes/no, etc
- An odds ratio (OR) is a measure of association between an exposure and a binary outcome.
- The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.

		Has cancer: Yes	Has cancer: No
Has the mutated gene?	Yes	23	117
	No	6	210

$$Odds\ Ratio = \frac{23/117}{6/210} = 6.88$$

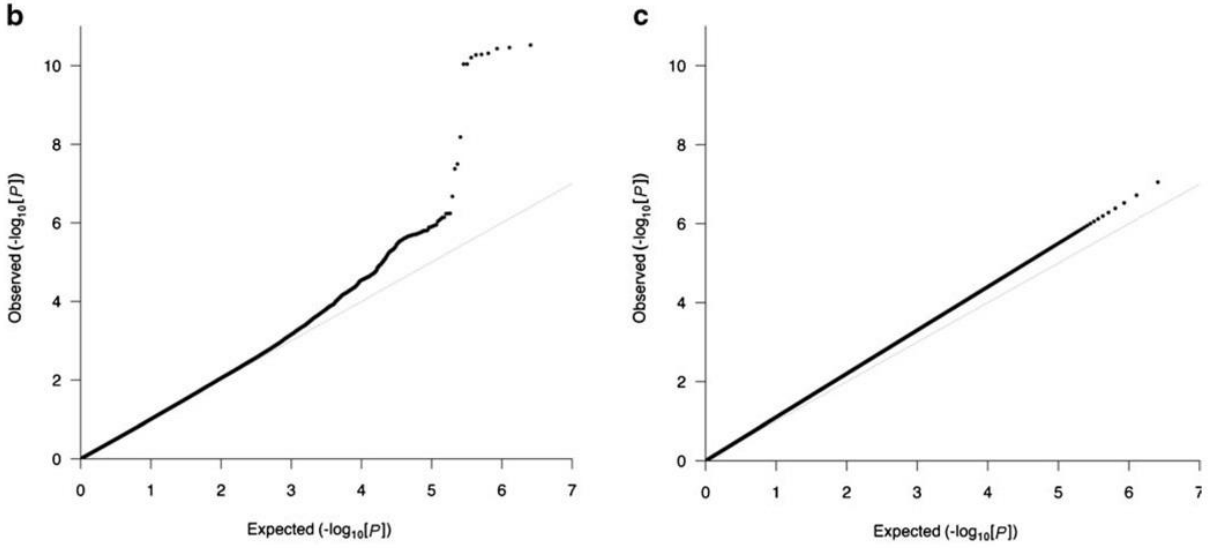
The odds are 6.88 greater that someone with mutated gene will also have a cancer

Standard GWAS threshold: **p-value** $\sim 5 \times 10^{-8}$! -> usually due to Bonferroni's correction: α/N where N is the number of tests (i.e., the number of included SNPs). Visualized via Manhattan Plot.



The QQ plot is a graphical representation of the deviation of the observed P values from the null hypothesis (a theoretical χ2-distribution).

-> A separation may suggest <u>population stratification</u>



Case Study:
Colorectal Cancer Study in Makassar

Data

- Processed data (PLINK input): .ped and .map files
- Binary version: .bed, .fam, and bim

1. Family ID (if unknown use the same id as for the sample id in column two
2. Sample ID
3. Paternal ID (if unknown use 0)
4. Maternal ID (if unknown use 0)
5. Sex (if unknown use 0)
6. Not used, set to 0
7. Rest of the columns: SNPs

```
4304 4304 0 0 0 0 C C C C G G G G G G C C G G C C T T T T
6925 6925 0 0 0 0 C C C C T T G G A A C C G G C C T T T T
7319 7319 0 0 0 0 C C C C G G G G G G C C G G C C T T T T
6963 6963 0 0 0 0 A A C C T T G G A A C C G G C C T T T T
6968 6968 0 0 0 0 C C C C G G G G G G G G G G C C T T T T
```

1. Chromosome ID (e.g. Chr1 for Chromosome 1)
2. Unique SNP identifier
3. Genomic distance (if unknown use 0)
4. SNP Position

```
Chr1 Chr1_314 0 314
Chr1 Chr1_317 0 317
Chr1 Chr1_323 0 323
Chr1 Chr1_324 0 324
Chr1 Chr1_332 0 332
Chr1 Chr1_334 0 334
Chr1 Chr1_342 0 342
Chr1 Chr1_346 0 346
Chr1 Chr1_348 0 348
Chr1 Chr1_349 0 349
```

Our raw data: 181 individuals and 733293 variants

Software

- PLINK 1.9 (https://www.cog-genomics.org/plink/)
- bcftools (https://www.htslib.org/download/)
- vcftools (https://vcftools.sourceforge.net/)
- conform-gt (https://faculty.washington.edu/browning/conform-gt.html)
- ADMIXTURE/fastStructure (https://dalexander.github.io/admixture/ or https://rajanil.github.io/fastStructure/)
- Hail (https://hail.is/docs/0.2/install/linux.html)
- Numpy/Pandas/Scipy/Matplotlib/statsmodels
- Some R packages

Quality Control Steps

QC Using PLINK:

Input: .ped and .map **and suggest** mkdir results

Missing rate per sample, impose 95% call rate

plink --file Smokescreen_Biorealm_p9-10 --mind 0.05 --recode --out results/crc

Missing rate per snp, impose 95% call rate

plink --file results/crc --geno 0.05 --recode --out results/crc

Only filter MAF > 1%

plink --file results/crc --maf 0.01 --recode --out results/crc

Perform Hardy Weinberg Equilibrium test and report the statistics (p-value < 1e-6)

plink --file results/crc --hardy midp --hwe 1e-6 midp --recode --out results/crc

Check heterozygosity

plink --file results/crc --het small-sample --out het

Create txt file to save HET information

echo "FID IID obs_HOM N_SNPs prop_HET" > het.txt
awk 'NR>1{print $1,$2,$3,$5,($5-$3)/$5}' het.het >> het.txt

Determine 3SD of heterozygosity rates (HR)

awk 'NR>1{sum+=$5;sq+=$5^2}END{avg=sum/(NR-1);print avg-3*(sqrt(sq/(NR-2)-2*avg*(sum/(NR-2))+(((NR-1)*(avg^2))/(NR-2)))),avg+3*(sqrt(sq/(NR-2)-2*avg*(sum/(NR-2))+(((NR-1)*(avg^2))/(NR-2))))}' het.txt

Create a list of samples whose HR values are outside of 3SD range

awk '$5<=<lower-limit> || $5>= <upper-limit>' het.txt> het.drop

Remove the samples

plink --file results/crc --remove het.drop --recode --out results/crc

https://github.com/alamahmadh/gwas_pipeline/blob/main/quality_control_gwas.txt

Complete Plink tutorial https://zzz.bwh.harvard.edu/plink/index.shtml

Remove duplicates

plink --file results/crc --list-duplicate-vars 'ids-only' 'suppress-first' --out results/crc.dupvar

plink --file results/crc --exclude results/crc.dupvar --recode --out results/crc

Remove indels

plink --file results/crc --snps-only 'just-acgt' --recode --out results/crc

Convert to vcf

plink --file results/crc --recode vcf --out results/crc

Convert to vcf.gz

bcftools sort results/crc.vcf -Oz -o results/crc.vcf.gz

Slice by chromosome

bcftools index -s results/crc.vcf.gz | cut -f 1 | while read C; do bcftools view -Oz -o results/vcf_per_chr/chr${C}.crc.vcf.gz results/crc.vcf.gz "${C}" ; done

At least for the CRC data I found that after this it should be ready for the proper submission into Michigan Imputation Server. You can additionally run conform-gt to fix some strand issue, especially if the Imputation Server asks us to resolve that problem first

Retain only SNPs (after the imputation)

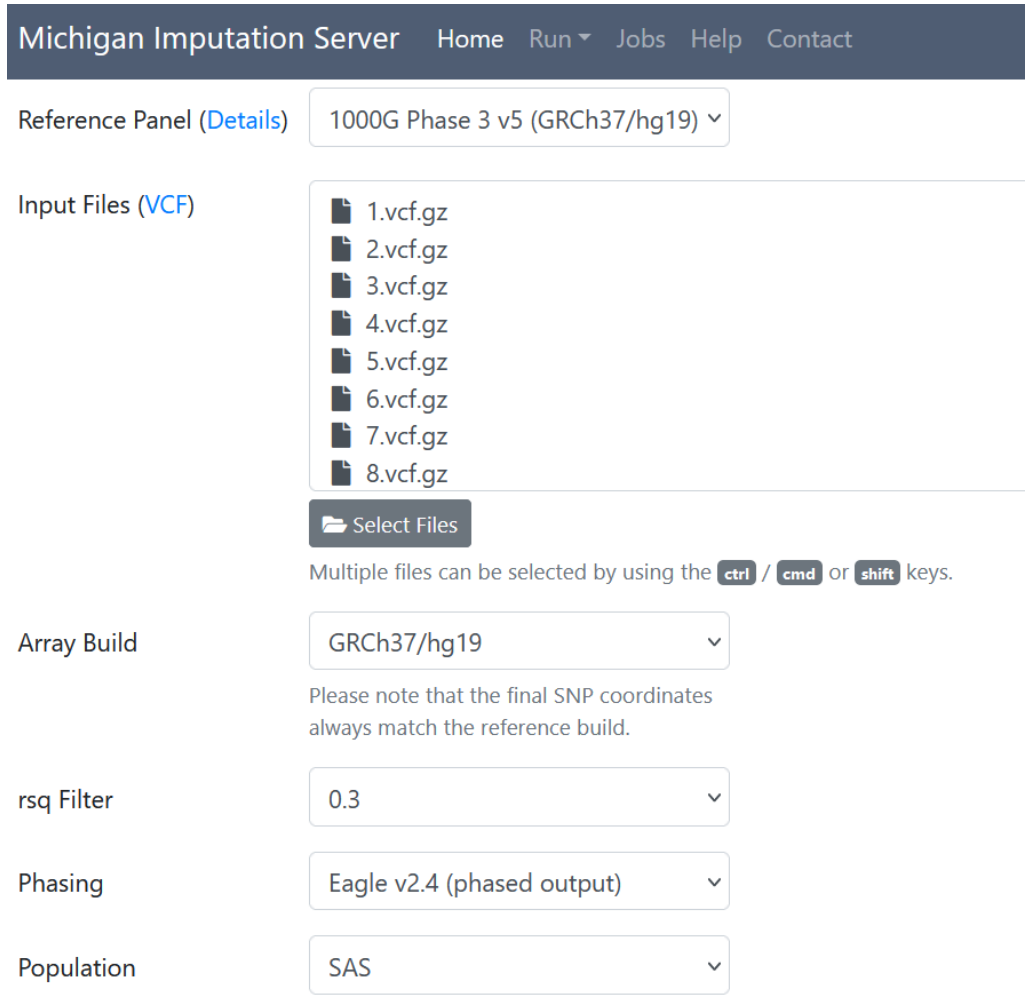
for CHR in {1..22}; do bcftools view -O z -o chr${CHR}.snps.vcf.gz -v snps chr${CHR}.dose.vcf.gz; done

Genotyping Imputation

- We can use Michigan Imputation Server.

https://imputationserver.sph.umich.edu/index.html

- Sign in/log in, upload separate vcf files (22 autosomal chr), set the parameters, and submit a job
- Once the job is finished download all files (the size of all imputed files will be larger than before!).



Variant Calling Format (VCF)

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

GENOMICS

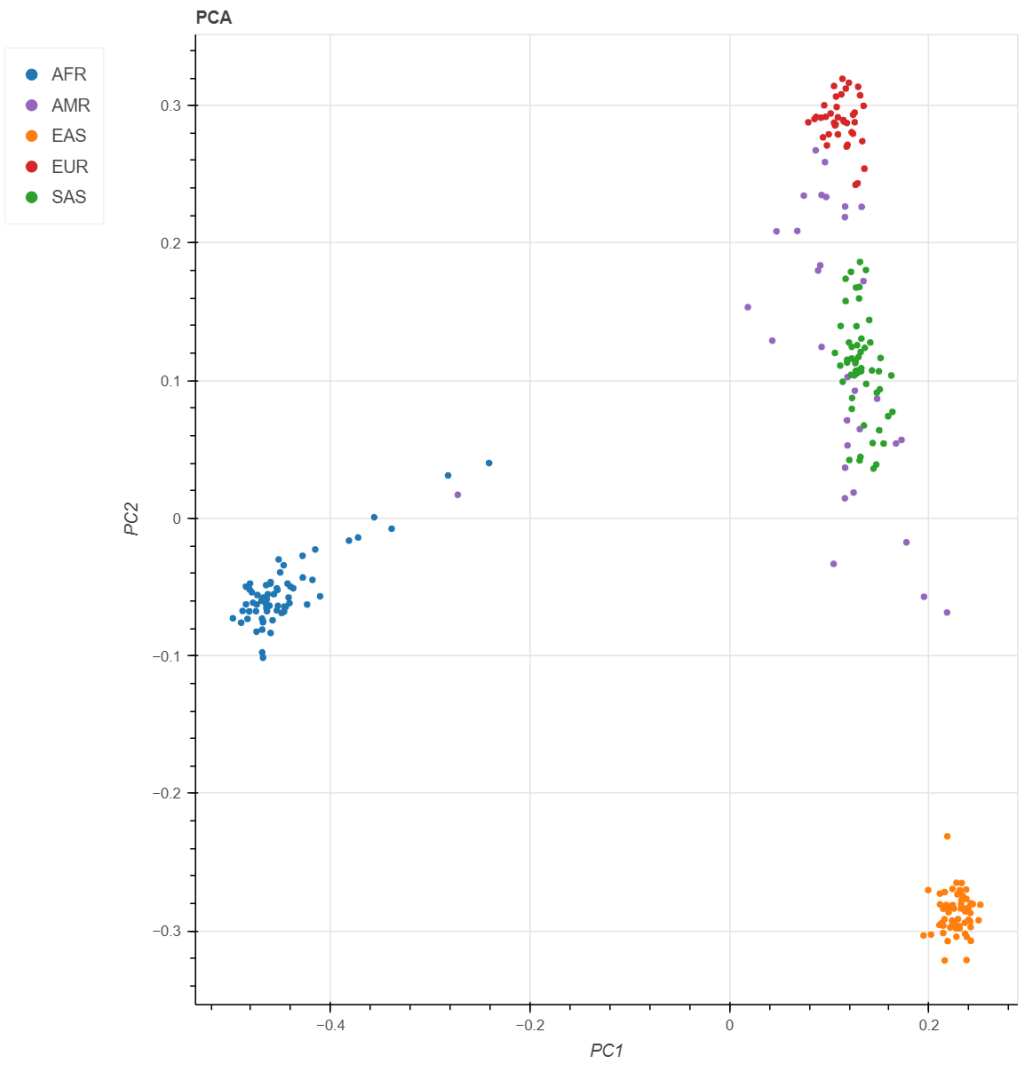
Hail: An Introduction to an Efficient Genomic Analysis Tool

Hail is an open-source Python library for genomic data manipulation and analysis. Five years in the making, we want to (re)introduce our actively developed tool to you, our users!

KUMAR VEERAPEN
23 JUN 2020 • 6 MIN READ

Ancestry Estimation

- Standard ancestral estimation software:
  - ADMIXTURE
  - Structure/fastStructure
  - EIGENSTRAT etc.
- For simplicity, I used **hail package** in Python to obtain the principal component analysis (PCAs) from the genotyping data as a “ancestry” covariate (out of memory problem ☹)



Genotype Encoding

Several different genetic models:

- **Additive models (common)**
- Dominant models
- Recessive models

Genotype	Score
aa	0
Aa	1
AA	2

Default mode in **hail**

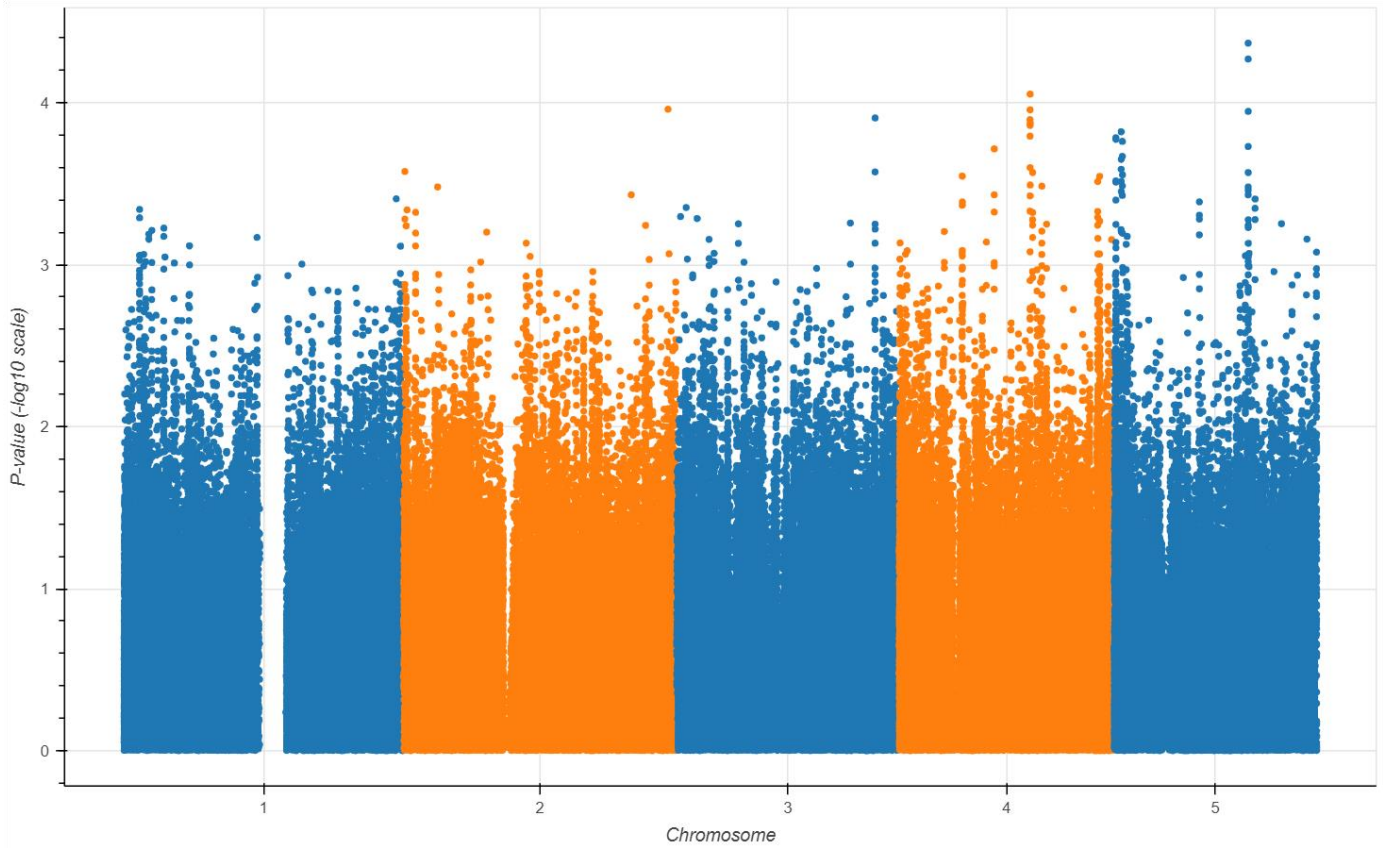
Regression Steps

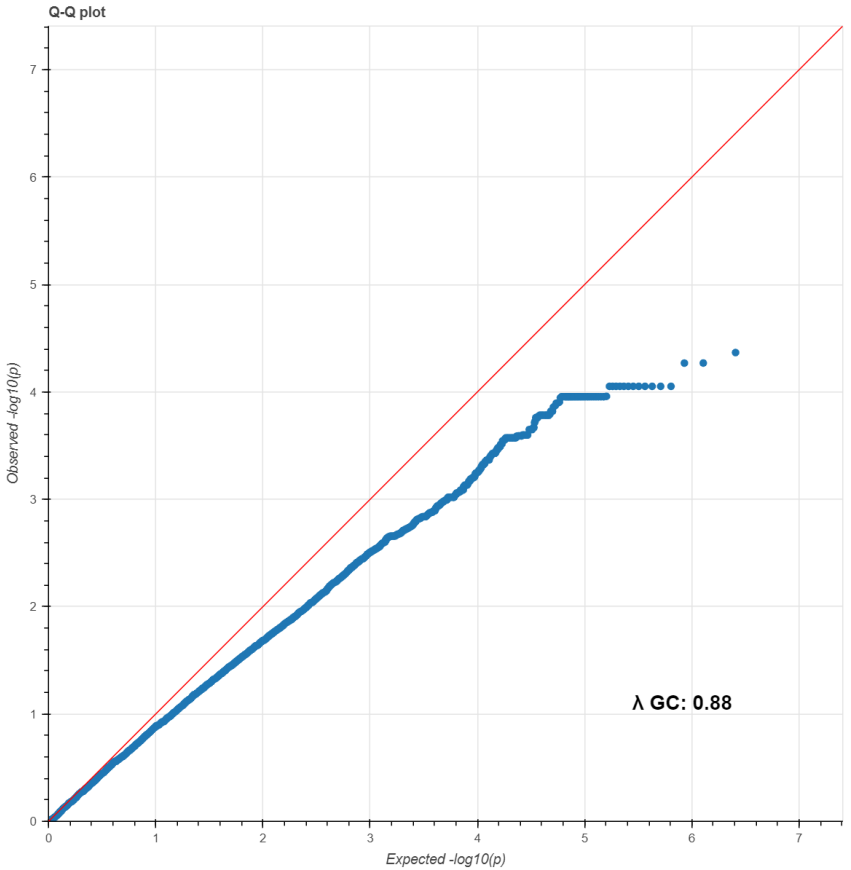
- In the CRC data we employed Case-Control study (binary outcomes) and hence we used a logistic regression.
- The regression process along with the preparation steps and visualization will be done in hail

Follow along with the tutorial

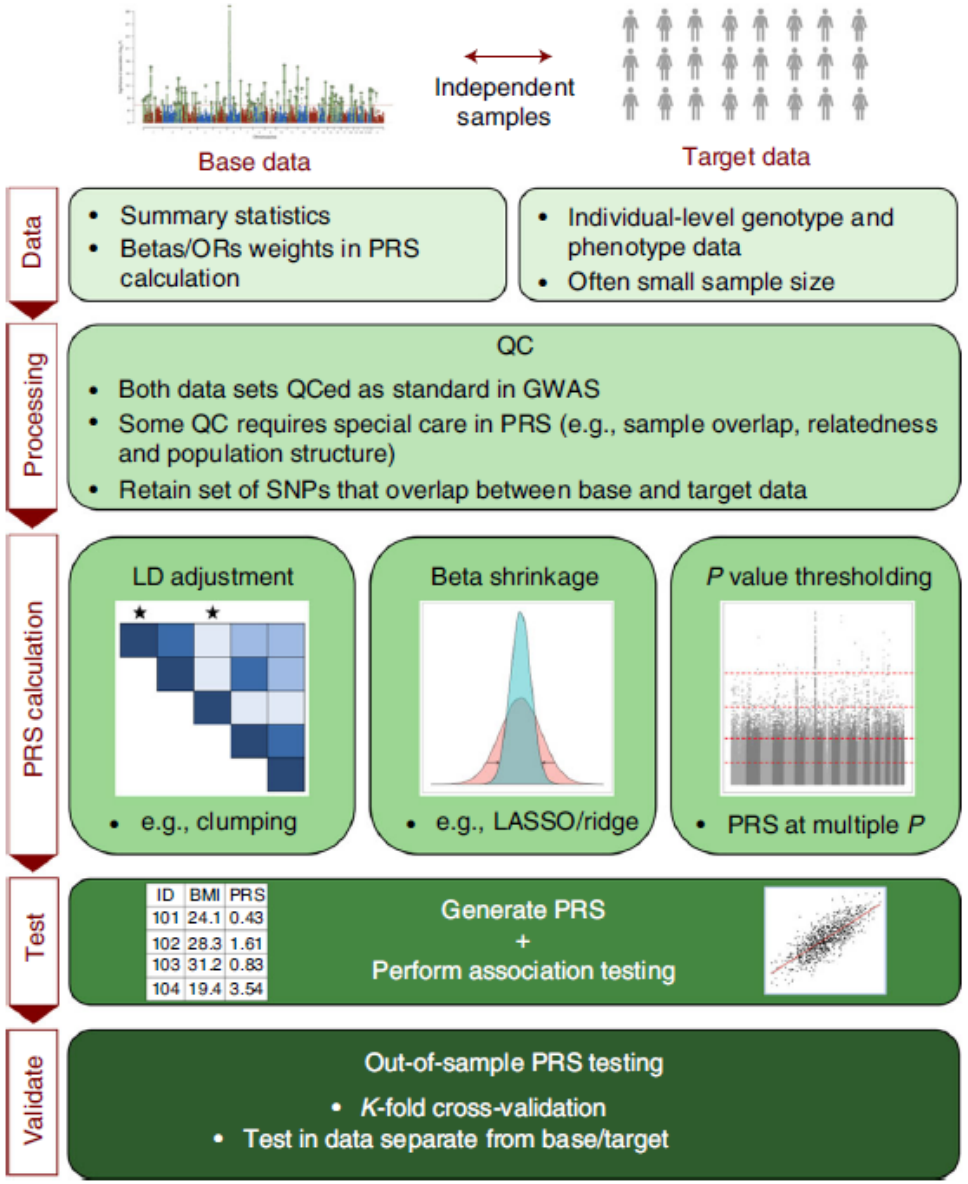
https://hail.is/docs/0.2/tutorials/01-genome-wide-association-study.html

Results


P-value (-log10 scale)
Chromosome
0
1
2
3
4
1
2
3
4
5


Q-Q plot
Observed -log10(p)
Expected -log10(p)
λ GC: 0.88
0
1
2
3
4
5
6
7

Polygenic Risk Score (PRS)



https://doi.org/10.1038/s41596-020-0353-1

Thank You