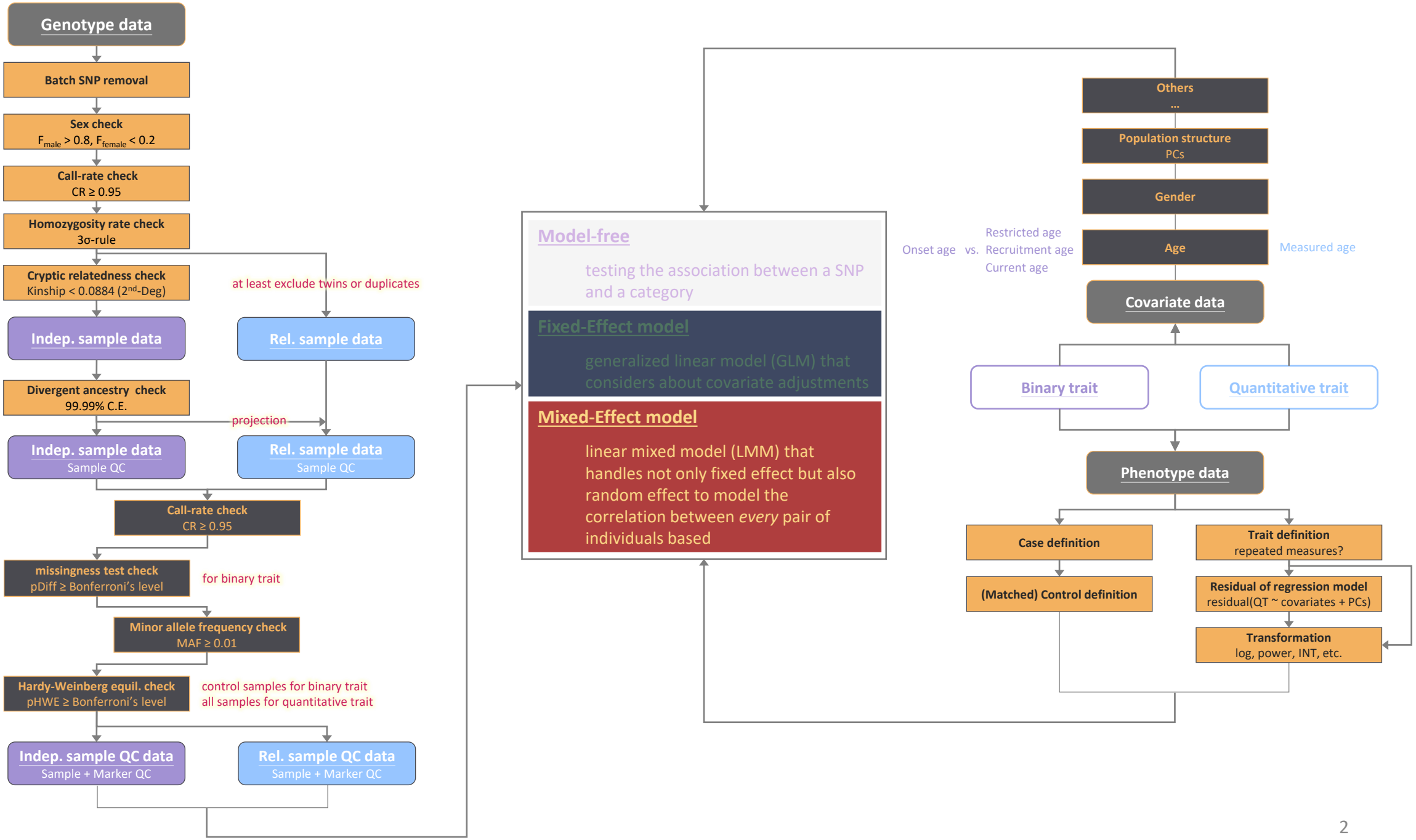
Re:ZERO

Re: Analysis in a genomic world from zero

- Starting Analysis in Genomic World -

陳佳煒 (Jia-Wei Chen)

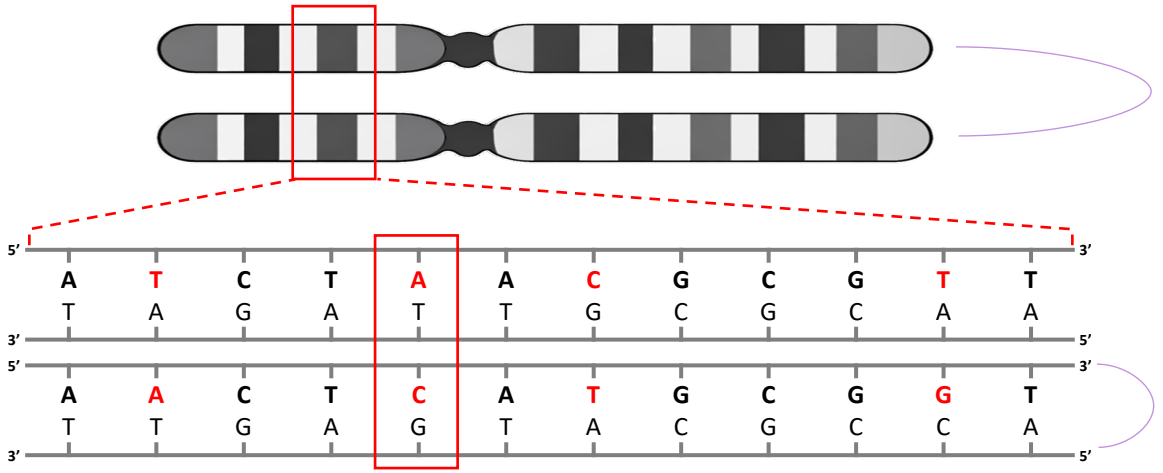
2026/04/07


Genotype data
Batch SNP removal
Sex check
$F_{male} > 0.8$, $F_{female} < 0.2$
Call-rate check
CR ≥ 0.95
Homozygosity rate check
3σ-rule
Cryptic relatedness check
Kinship < 0.0884 (2nd-Deg)
at least exclude twins or duplicates
Indep. sample data
Rel. sample data
Divergent ancestry check
99.99% C.E.
projection
Indep. sample data
Sample QC
Rel. sample data
Sample QC
Call-rate check
CR ≥ 0.95
missingness test check
pDiff ≥ Bonferroni's level
for binary trait
Minor allele frequency check
MAF ≥ 0.01
Hardy-Weinberg equil. check
pHWE ≥ Bonferroni's level
control samples for binary trait
all samples for quantitative trait
Indep. sample QC data
Sample + Marker QC
Rel. sample QC data
Sample + Marker QC
Model-free
testing the association between a SNP and a category
Fixed-Effect model
generalized linear model (GLM) that considers about covariate adjustments
Mixed-Effect model
linear mixed model (LMM) that handles not only fixed effect but also random effect to model the correlation between every pair of individuals based
Onset age
vs.
Restricted age
Recruitment age
Current age
Measured age
Others
...
Population structure
PCs
Gender
Age
Covariate data
Binary trait
Quantitative trait
Phenotype data
Case definition
(Matched) Control definition
Trait definition
repeated measures?
Residual of regression model
residual(QT ~ covariates + PCs)
Transformation
log, power, INT, etc.

:Terminology

Locus

The specific location (a base pair or a genomic region) on a chromosome



Diploid

Having two complete sets of chromosomes, one from each biological parent

Double strand

DNA molecule with two complementary polynucleotide chains held by base pairing

Marker

A DNA position used to track variation (could be SNV, indel, etc.)

Allele

A variant of the sequence of nucleotides at a particular location, or locus

* Allele 1 (A1) = A, Allele 2 (A2) = C
* Major / Minor allele
* Reference (Ref) / Alternative (ALT) allele
* Effect allele (EA)

Genotype

Combination of alleles at a locus

* Homozygous: AA, CC
* Heterozygous: AC

SNV

Single Nucleotide Variant: any single base change in DNA sequence

Common	MAF ≥ 0.05
SNP	MAF ≥ 0.01
Low-frequency	0.01 ≤ MAF < 0.05
Rare	0.001 ≤ MAF < 0.01
Ultra-rare	MAF < 0.001

Haplotype

Combination of alleles on a strand

A-T-C-T-A-A-C-G-C-G-T-T
A-A-C-T-C-A-T-G-C-G-G-T

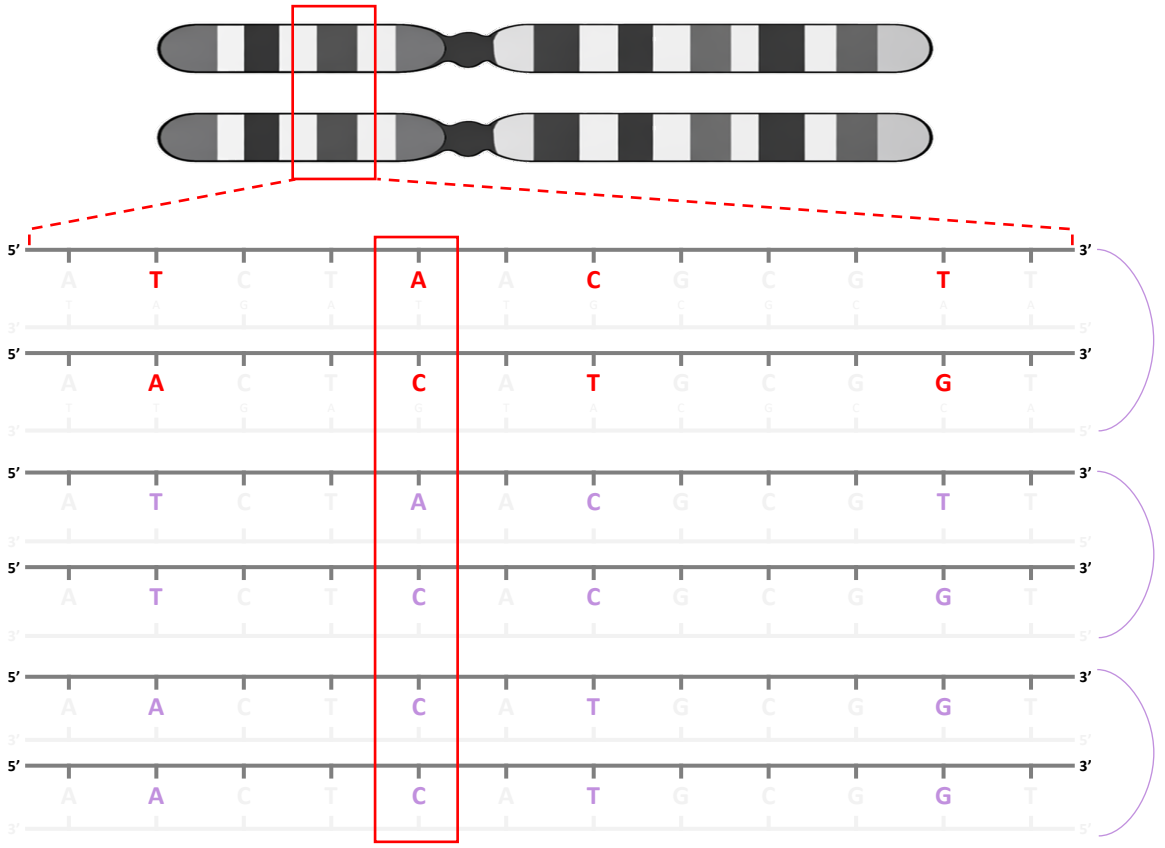
CR

Call-Rate: proportion of non-missing genotypes for a variant or a sample

HR

Homozygous/Heterozygous Rate: proportion of homozygous/heterozygous genotypes for a variant or a sample

Homozygous rate (variant level)
(0 + 0 + 1) / 3 = 0.33
Homozygous rate (genome-wide)
(0 + 0 + 0 + 0) / 4 = 0
(1 + 0 + 1 + 0) / 4 = 0.5
(1 + 1 + 0 + 1) / 4 = 0.75



IBD

Identity By Descent: DNA segments inherited from a recent common ancestor without recombination

IBS

Identity By State: DNA segments that match in sequence, regardless of shared ancestry (includes IBD + chance/population matches)

GRM

Genetic Relationship Matrix: pairwise genetic relatedness estimates from genome-wide SNP data

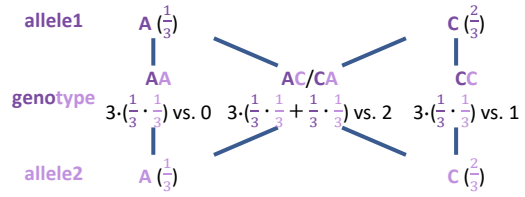
MAF

Minor Allele Frequency: frequency of the less common allele at a variant site

AF(A) = (1 + 1 + 0) / (3 × 2) = 0.33
AF(C) = (1 + 1 + 2) / (3 × 2) = 0.67
MAF = AF(A) = 0.33

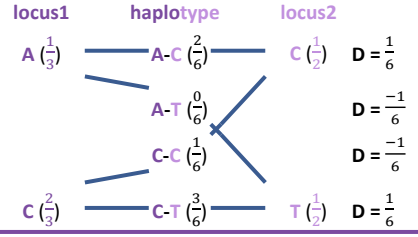
HWE

Hardy-Weinberg Equilibrium: allele/genotype frequencies stable across generations without evolutionary forces

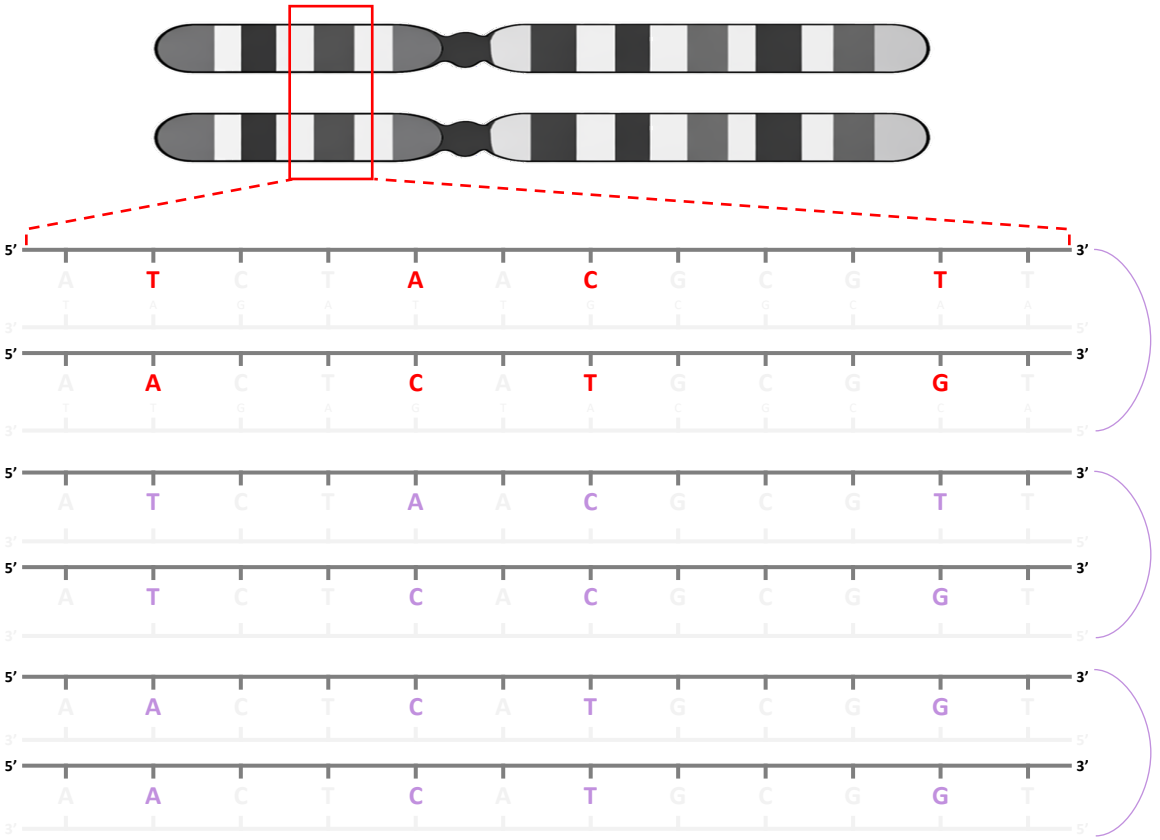


LD

Linkage Disequilibrium: non-random association of alleles at different loci due to shared inheritance history



	Binary trait	Quantitative trait
FID/IID	**FID** = 0 or **FID** = **IID**	
PID/MID	**PID** = **MID** = 0	
SEX	0 = unknown, 1 = male, 2 = female	
PHENO	-9/0 = missing 1 = unaffected 2 = affected	at least two different values
A1/A2	0 = missing 1 = A, 2 = C, 3 = G, 4 = T non-zero integers or any characters	



CHR	1-22, 23 = X, 24 = Y, 25 = XY, 26 = MT
ID	probe set ID (Ax-***): unique id for probe sequence affy SNP ID (Affy-***): unique id for CHR, POS, REF, and ALT dbSNP RS ID (rs***): unique id for a genome build
cM	genetic distance (centiMorgan)
POS	plain integers check genome build

.ped (pedigree + genotype)

Family ID **FID**	Individual ID **IID**	Paternal ID **PID**	Maternal ID **MID**	**SEX**	Phenotype **PHENO**	variant1 **A1**	variant1 **A2**	variant2 **A1**	variant2 **A2**	variant3 **A1**	variant3 **A2**	variant4 **A1**	variant4 **A2**	
FAM001	IND1	0	0	1	2	T	A	A	C	C	T	T	G	…
FAM001	IND2	0	0	1	2	T	T	A	C	C	C	T	G	
FAM001	IND3	0	0	2	1	A	A	C	C	T	T	G	G	
⋮														

.map

Chromosome **CHR**	Variant ID **ID**	Genetic distance **cM**	Physical position **POS**
1	var1	0	565433
1	var2	0	752566
1	var3	0	753541
⋮			

bfile

2-bit genotype codes

00	Homozygous for first allele in .bim file
01	Missing genotype
10	Heterozygous
11	Homozygous for second allele in .bim file

.bim (extended map)

A2	T	A	C	T	(major allele)
A1	T	A	C	T	(minor allele)
POS					
cM					
ID					
CHR					

FID	IID	PID	MID	SEX	PHENO

.fam (pedigree)

var1	var2	var3	var4	
10	10	10	10	...
00	10	00	10	
11	11	11	11	
⋮				

.bed (binary genotype)

pfile

.pvar (extended map)

ALT	T	A	C	T	(alternative allele)
REF	T	A	C	T	(reference allele)
POS					
cM					
ID					
CHR					

#FID	IID	PAT	MAT	SEX	PHENO

.psam (pedigree)

var1	var2	var3	var4

.pgen (binary genotype)

.ped (pedigree + genotype)

						var1		var2		var3		var4		
FID	IID	PID	MID	SEX	PHENO	A1	A2	A1	A2	A1	A2	A1	A2	
FAM001	IND1	0	0	1	2	T	A	A	C	C	T	T	G	...
FAM001	IND2	0	0	1	2	T	T	A	C	C	C	T	G	
FAM001	IND3	0	0	2	1	A	A	C	C	T	T	G	G	
⋮														

.map

CHR	ID	cM	POS
1	var1	0	565433
1	var2	0	752566
1	var3	0	753541
⋮			

bfile

2-bit genotype codes

00	Homozygous for first allele in .bim file
01	Missing genotype
10	Heterozygous
11	Homozygous for second allele in .bim file

.bim (extended map)

A2	A	C	T	(major allele)
A1	T	A	C	(minor allele)
POS				
cM				
ID				
CHR				

FID	IID	PID	MID	SEX	PHENO

.fam (pedigree)

var1	var2	var3	var4	
10	10	10	10	…
00	10	00	10	
11	11	11	11	
⋮				

.bed (binary genotype)

PHENO	**Binary** 2 = Case, 1 = Control, 0/-9/NA = Unknown (default) 1 = Case, 0 = Control, NA = Unknown **Quantitative** any numeric values
COVAR	**Demographic** age, gender **Subpopulation structure** PC1, PC2, … **Others** BMI, abc-covariates, etc.

.phe (phenotype)

FID	IID	PHENO1	PHENO2	…
FAM001	IND1			
FAM001	IND2			
FAM001	IND3			
⋮				

\+

.cov (covariate)

FID	IID	COVAR1	COVAR2	…
FAM001	IND1			
FAM001	IND2			
FAM001	IND3			
⋮				

→

.phecov (phenotype + covariate)

FID	IID	PHENO1	PHENO2	…	COVAR1	COVAR2	…
FAM001	IND1						
FAM001	IND2						
FAM001	IND3						
⋮							

:Toolset

LINUX

Terminal (終端機)

```
$    uname –m
x86_64    # 64-bit
i686 i386 # 32-bit
```

Terminal (終端機)

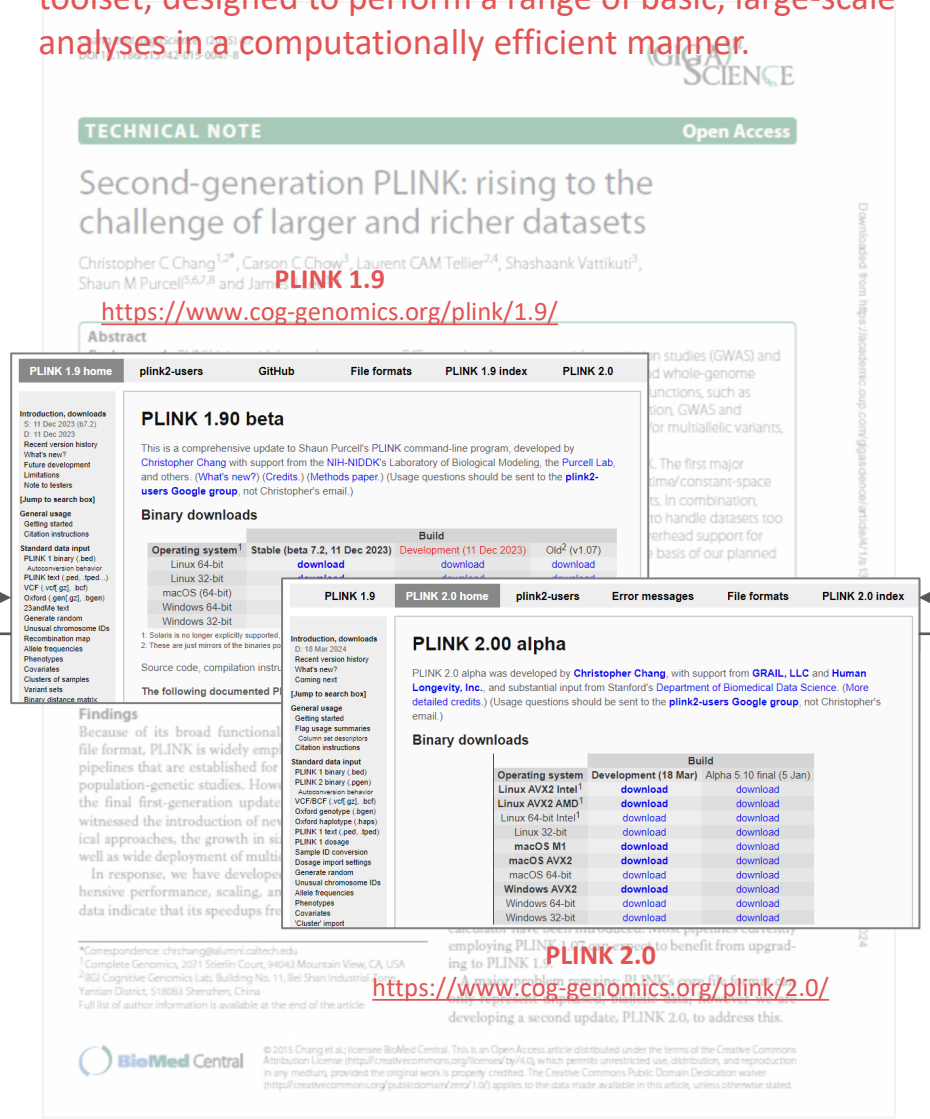
```
$    lscpu | grep avx2
… avx2 … # supports AVX2
         # not support AVX2
```

Terminal (終端機)

```
# execute from the specific folder
$    cd path/of/plink
$    plink

# execute everywhere
$    copy path/of/plink /usr/local/bin/
$    plink
```

PLINK

a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner.



PLINK 1.9
https://www.cog-genomics.org/plink/1.9/

PLINK 2.0
https://www.cog-genomics.org/plink/2.0/

WINDOWS

Command Prompt (命令提示字元)

```
>    wmic os get osarchitecture
64-bit # 64-bit
32-bit # 32-bit
```

Command Prompt (命令提示字元)

```
>    wmic cpu get caption
Family 6 Model 60 (Intel) # supports AVX2
```

Command Prompt (命令提示字元)

```
# execute from the specific folder
$    cd path/of/plink
$    plink
```

--keep / --remove: Filters individuals in/out based on a list of Family and Individual IDs

--pheno: Points to an external file containing your phenotype data

--glm: Performs association testing using Generalized Linear Models (linear, logistic, or Firth regression)

--covar: Loads a file with covariates to include in the regression model to prevent confounding

--threads: Sets the number of CPU cores for the process. Vital for speeding up calculations on large datasets like the 90k samples you're analyzing

--extract / --exclude: Filters variants (SNPs) in/out based on a list of IDs

--zero-cluster: Sets all genotypes to missing for members of a specific cluster (used for specific cleaning tasks)

--within: Loads family or cluster assignments from a file, ensuring analysis is performed within these specified groups

--het: Calculates the heterozygosity of individuals to identify sample contamination or inbreeding

--test-missing: Tests if missingness rates differ significantly between cases and controls (identifies technical bias)

--geno: Filters out SNPs with a missing call rate higher than a specified threshold (e.g., 0.05)

Terminal (終端機)

```
# basic execution
$     plink --bfile path/of/files --flags [arg.]
```

--missing: Generates statistics on missing data per individual and per SNP

--maf: Filters out SNPs with a Minor Allele Frequency below a specific threshold (e.g., 0.01)

--hwe: Filters out SNPs that deviate significantly from Hardy-Weinberg Equilibrium

--check-sex: Compares the sex assigned in the dataset to the actual heterozygosity on the X chromosome to identify mislabeled samples

--indep-pairwise: Performs Linkage Equilibrium (LD) pruning, identifying SNPs that are highly correlated to reduce redundancy

--king-cutoff: Uses the KING-robust kinship estimator to remove closely related individuals, ensuring sample independence

--score: Calculates Polygenic Risk Scores (PRS) by summing weighted alleles across the genome

--genomes: Calculates Identity-by-Descent (IBD) to identify cryptic relatedness or sample duplicates

--pca: Conducts Principal Component Analysis to visualize and adjust for population stratification

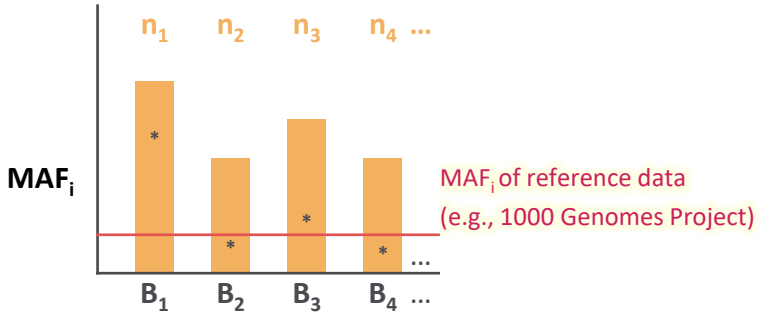
--make-bed / --make-pgen: Creates a PLINK 1.9 binary triplet (.bed, .bim, .fam) / PLINK 2.0 binary triplet (.pgen, .pvar, .psam)

:QualityControl

Batch SNP removal

Why batch SNP removal?

- False positive associations (e.g., cases and controls are in different batches)



How to do?

- Compare AFs of SNPs to ones in public database (e.g., 1000 Genomes Project) to identify the SNPs with weird Afs
- Instead of excluding the SNP, marking the genotype calls of samples from problematic batches as missing for this specific SNP

Input

dat.clust

FID	IID	Batch
FAM001	IND1	1
FAM001	IND2	1
FAM001	IND3	2
⋮		

The batch information can be requested from typing center

batchSNPs.zero

SNP	Batch
var1	1
var2	3
var3	3
⋮	

It may be a challenge for non-programmer to find out batch SNPs

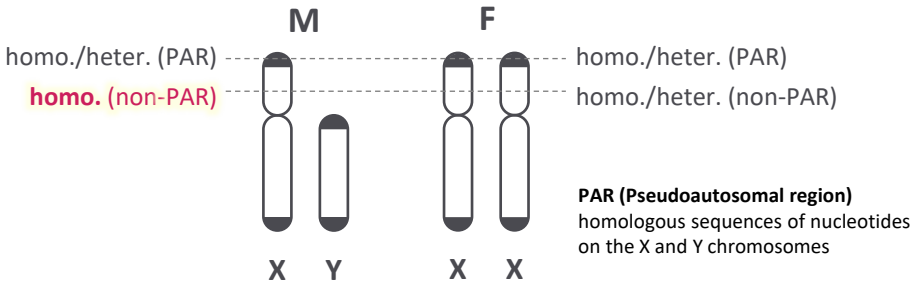
 # batch SNP removal (by batches)

```
$ plink --bfile path/of/your/file --zero-cluster batchSNPs.zero --within dat.clst --make-bed --out dat_wg
$ plink --bfile dat_wg --chr 1-22 --make-bed --out dat_as
```

Sex check

Why sex check?

- Chromosome aberration
- Covariate identification
- Sample misidentification



How to do?

- Using either
  homozygosity rate : M (≥ 0.9) & F (< 0.9) or
  inbreeding coefficient : M (> 0.8) & F (< 0.2)
  of X chromosome to check for the gender
- If EHR data is accessed, directly comparing EHR gender to genetic gender

sex check

```
$ plink --bfile dat_wg --check-sex --out dat_wg
$ grep "PROBLEM" dat_wg.sexcheck > rmInd_sex.txt
$ plink2 --bfile dat_as --remove rmInd_sex.txt --write-samples --out dat_as_1
```

Output

dat_wg.**sexcheck**

FID	IID	PEDSEX	SNPSEX	**STATUS**	F
FAM001	IND1	1	1	OK	0.9588
FAM001	IND2	1	1	OK	0.9616
FAM001	IND3	1	1	OK	0.9588
⋮					
FAM001	IND17	2	1	PROBLEM	0.9539
⋮					
FAM001	IND28	1	2	PROBLEM	-0.05586
⋮					

→

rmInd_sex.txt

FAM001	ind17
FAM001	ind28

Remove samples with inconsistent genders (**STATUS = PROBLEM**)

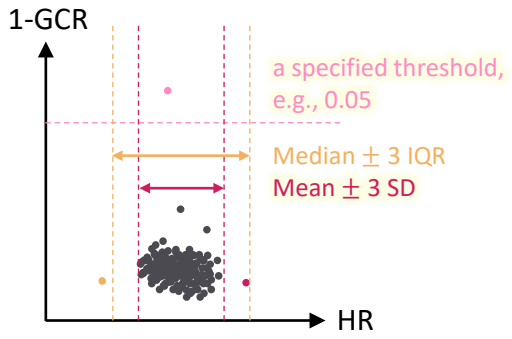
Call-rate & Homozygosity rate check

Why genotype call-rate check?

- Low DNA quality or concentration

Why homozygosity rate check?

- Inbreeding (excess Homozygoisty)
- DNA contamination (excess Heterozygosity)



N = # of markers

N_i = # of nonmissing markers for individual i

O_i = # of homozygous markers for individual i

$GCR_i = \frac{N_i}{N}$

$HR_i = \frac{N_i - O_i}{N_i}$

How to do?

- Calculate genotype call-rate (GCR) and heterozygosity rate (HR) for **autosomes**
- A specific cutoff (95% or even 99%) for CR and use 3-σ rule or IQR rule for HR

call-rate & homozygosity rate check

```
$   plink --bfile dat_as --keep dat_as_1.id --missing --het --out dat_as_1
$   plink –bfile dat_as --keep dat_as_1.id --remove rmInd_missing_het.txt --write-samples --out dat_as_2
```

Output

dat_as_1.**imiss**

FID	IID	MISS_PHENO	N_MISS	N_GENO	**F_MISS**
FAM001	IND1	Y			0.00158
FAM001	IND2	Y			0.00181
FAM001	IND3	Y			0.1375
⋮					

Remove samples with CR = 1 − F_MISS smaller than a threshold, say 0.95

dat_as_1.**het**

FID	IID	**O(HOM)**	E(HOM)	**N(NM)**	F
FAM001	IND1	426964		624027	
FAM001	IND2	588153		623882	
FAM001	IND3	425156		624154	
⋮					

Remove samples with HR = $\frac{\text{O(HOM)}}{\text{N(NM)}}$ out of mean(HR) ± 3 sd(HR)

rmInd_missing_het.txt

FAM001	IND3	CR
FAM001	IND2	HET

Cryptic relatedness check

Why cryptic relatedness check?

- Biostatistical Independence (e.g., assumption of GLM, bias AF estimation)
- Pedigree structure (e.g., family-based study)

	Proportion IBD	Kinship coeff.
duplicate/MZ twin	1	> 0.354
1st degree	0.5	[0.177, 0.354]
2nd degree	0.25	[0.0884, 0.177]
halfway of 2nd & 3rd degrees	> 0.1875	
3rd degree	0.125	[0.0442, 0.0884]

How to do?

- using **independent** markers (pair correlation $r^2 < 0.2$) to calculate the relatedness of individuals by either **proportion IBD** or **kinship coefficient** to check for the relatedness

 # cryptic relatedness check (preferred)

```
$ plink --bfile dat_as --keep dat_as_2.id --indep-pairwise 50 5 0.2 --out dat_as_2
$ plink --bfile dat_as --keep dat_as_2.id --extract dat_as_2.prune.in --king-cutoff 0.0442 --out dat_as_2
$ plink --bfile dat_as --keep dat_as_2.id --remove dat_as_2.king.cutoff.out.id --write-samples --out dat_as_3
```

 # cryptic relatedness check

```
$ plink --bfile dat_as --keep dat_as_2.id --indep-pairwise 50 5 0.2 --out dat_as_2
$ plink --bfile dat_as --keep dat_as_2.id --extract dat_as_2.prune.in --genome --out dat_as_2
$ plink --bfile dat_as --keep dat_as_2.id --remove rmInd_rel.txt --write-samples --out dat_as_3
```

Output

dat_auto_2.**genome**

FID1	IID1	FID2	IID2	RT	EZ	Z0	Z1	Z2	**PI_HAT**	PHE	DST	PPC	RATIO
⋮													

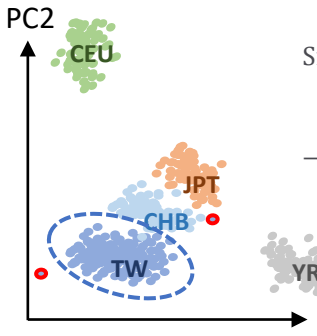
Remove samples with lower CR in a pair of samples if **PI_HAT** > 0.1875

It may be a challenge for a non-programmer to find the optimal sample set

Divergent ancestry check

Why divergent ancestry check?

- Population stratification (e.g., Hamer, D. , & Sirota, L. (2000))
- Pedigree structure (e.g., family-based study)



Singular Value Decomposition (SVD)

$$G_{I\times S} = U_{I\times \min(I,S)}\,\Sigma_{\min(I,S)\times\min(I,S)}\,V^T_{S\times\min(I,S)}$$

$$\rightarrow G\,V = (U\,\Sigma\,V^T)V = U\,\Sigma \quad (V \text{ is orthogonal } \rightarrow V^TV = I)$$

V : loadings (relationship) of markers

$G\,V = U\,\Sigma$: PC scores of individuals

How to do?

- merge study and reference data
  → exclude ambiguous markers → prune correlated markers
  → PCA → exclude outliers
- PCA on reference data → extract loadings (marker weights)
  → project study data onto reference data → exclude outliers

divergent ancestry check

```
$  https://www.cog-genomics.org/plink/2.0/resources#phase3_1kg # 1000 genomes phase 3 data

$  awk '{print $1,$4,$4,$2}' dat_as.bim > lst_var.txt
$  plink2 --zst-decompress all_hg38.pgen.zst > all_hg38.pgen
$  plink2 --pfile all_hg38 vzs --allow-extra-chr --extract bed1 lst_var.txt --snps-only --max-alleles 2 --set-all-var-ids @:# --rm-dup exclude-all --make-bed --out all_hg38_extract
$  sed 's/#IID/IID/g;s/Population/population/g' all_hg38.psam > relationships_w_pops.txt

$  plink2 --bfile all_hg38_extract --maf 0.01 --freq counts --pca biallelic-var-wts --out all_hg38_extract
$  plink2 --bfile dat_as --keep dat_as_3.id --read-freq all_hg38_extract.accounts --score all_hg38_extract.eigenvec.var 2 4 header read variance-standardize no-mean-imputation --score-col-nums 5-14 --out dat_as_3
$  plink --bfile dat_as --keep dat_as_3.id --remove rmInd_divAncestry.txt --make-bed --out dat_as_sampQC
```

Output

all_hg38_extract.**account**

#CHROM	ID	REF	ALT	ALT_CTS	OBS_CT
	var1				
	var2				
	var3				
	⋮				

Allele counts of variants

all_hg38_extract.**eigenvec.var**

#CHROM	ID	MAJ	NONMAJ	**PC1**	**PC2**	...	**PC10**
	var1						
	var2						
	var3						
	⋮						

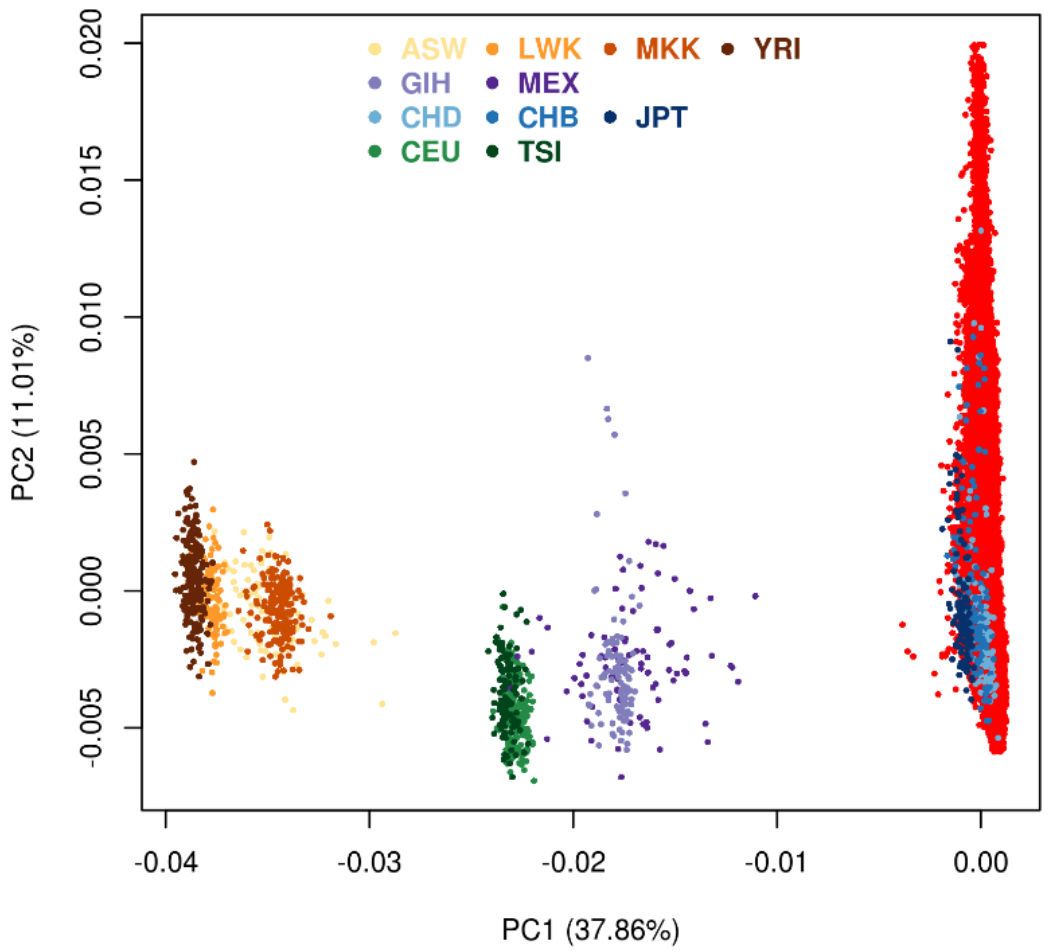
PC loadings of variants

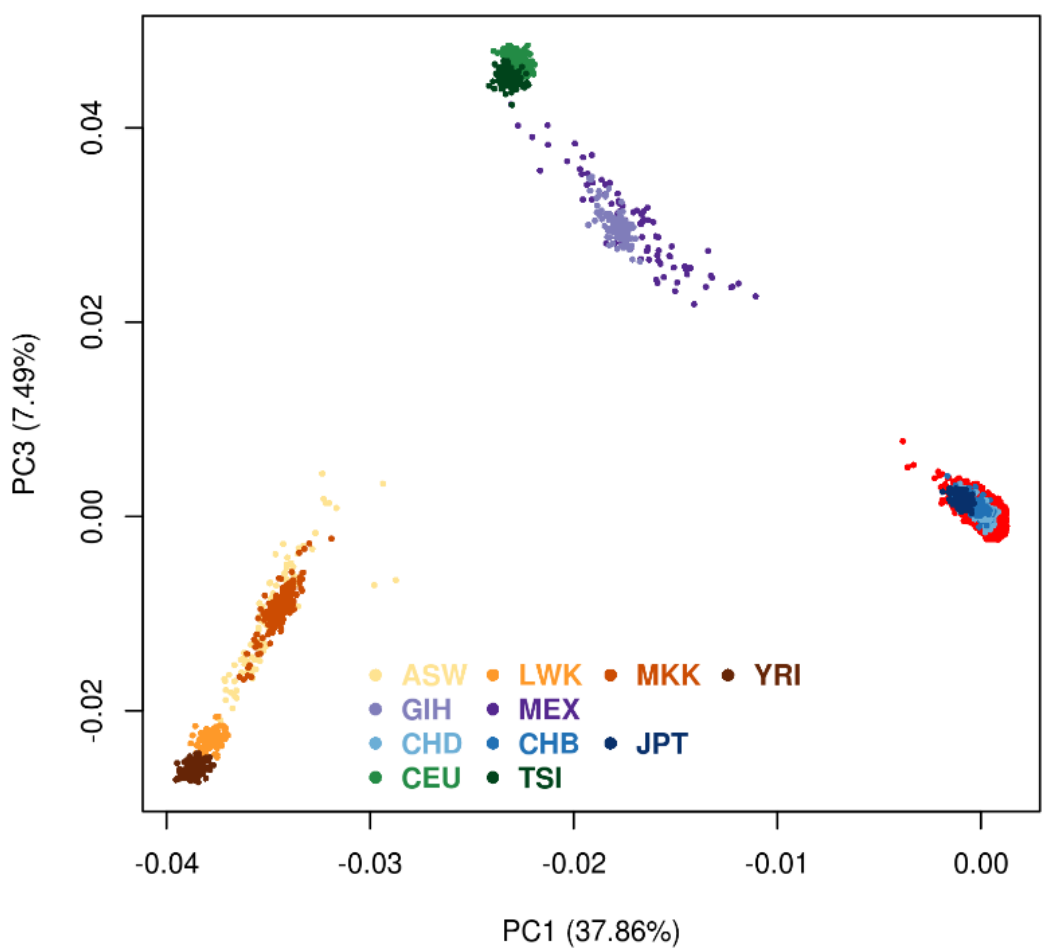
dat_as_3.**score**

#FID	IID	ALLELE_CT	NAMED_ALLELE_DOSAGE_SUM	**SCORE1_AVG**	...	**SCORE10_AVG**
FAM001	IND1					
FAM001	IND2					
FAM001	IND3					
⋮						

PC scores of individuals

Divergent ancestry check



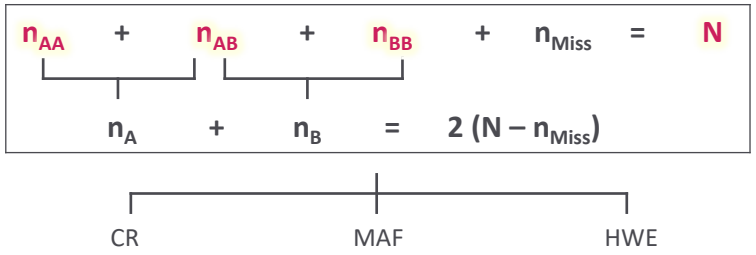


When sample size of target data is much larger than that of ancestry data,
one of the first two PCs may be dominant by the variation in the target data

Marker QC

How to do?

- A specific cutoff (95% or even 99%) for CR
- A specific cutoff (0.01 or 0.05) for MAF
- A specific cutoff or Bonferroni's level for HWE; HWE test is applied to controls for case/control study and to all for quantitative-trait study



Why marker QC?

- Technical artifacts (lower CR reflects assay issues since a failure across many samples)
- Statistical power(low MAF yields unstable variance and spurious associations)
- Biological Plausibility (HWE violation suggests miscalling or hidden population structures)

 # call-rate check

```
$   plink --bfile dat_as_sampQC --geno 0.05 --write-snplist --out dat_as_5

# case/control nonrandom missingness check
$   plink --bfile dat_as_sampQC --keep dat_as_5.snplist --test-missing --out dat_as_5
$   n=$(wc -l < dat_as_5.missing)
$   awk -v n=$n '$5 < 0.05/(n-1) {print $2}' dat_as_5.missing > dat_as_5.snplist
```

 # MAF check

```
$   plink --bfile dat_as_sampQC --keep dat_as_5.snplist --maf 0.01 --write-snplist --out
    dat_as_6
```

 # HWE check

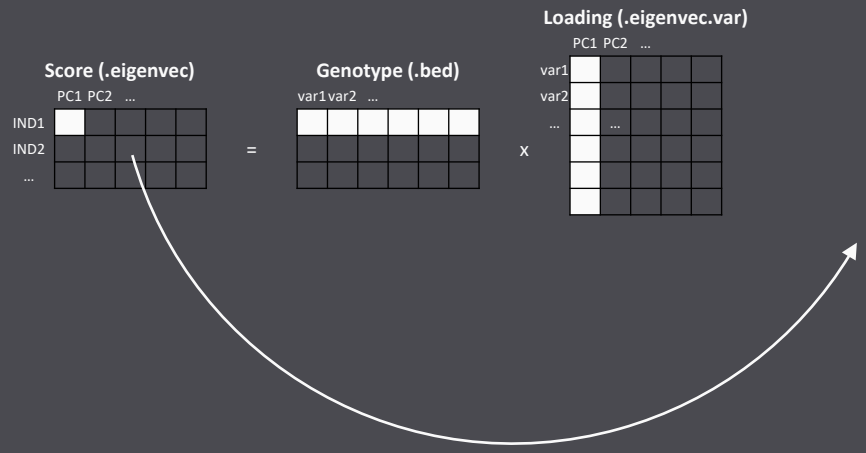
```
# quantitative trait
$   bonf=$(bc -l <<< "0.05/$(wc -l < dat_as_6.snplist)")
$   plink --bfile dat_as_sampQC --keep dat_as_6.snplist --hwe ${bonf} --make-bed –out
    dat_as_sampQC_varQC

# binary trait
$   bonf=$(bc -l <<< "0.05/$(wc -l < dat_as_6.snplist)")
$   plink2 --bfile dat_as_sampQC --pheno dat.phecov --1 --keep-if "PHENO==0"
    dat_as_6.snplist --hwe ${bonf} --write-snplist --out dat_as_7
$   plink2 --bfile dat_as_sampQC --keep dat_as_7.snplist --make-bed --out
    dat_as_sampQC_varQC
```

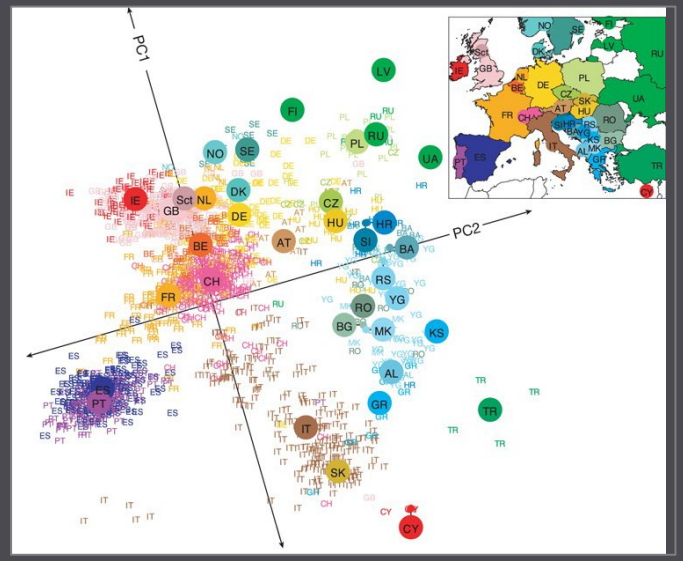
:Population Structure

Subpopulation structure

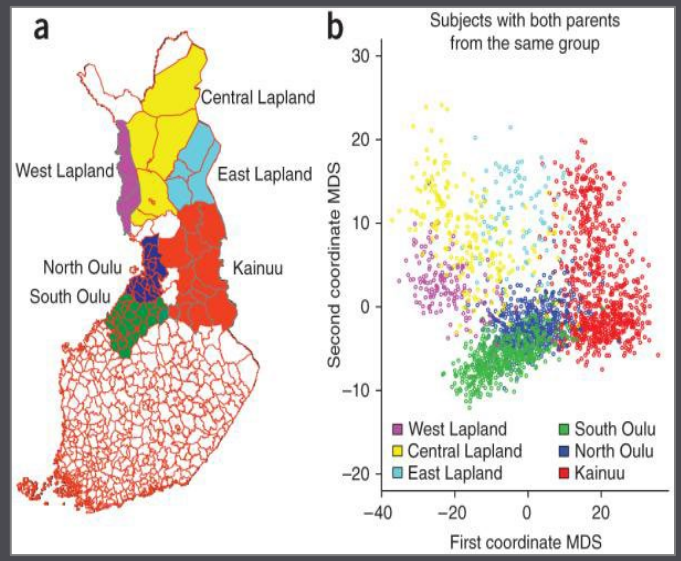
Why subpopulation structure?

- A **confounder** effect that AF differences between subpopulations are correlated with trait differences can induce spurious associations

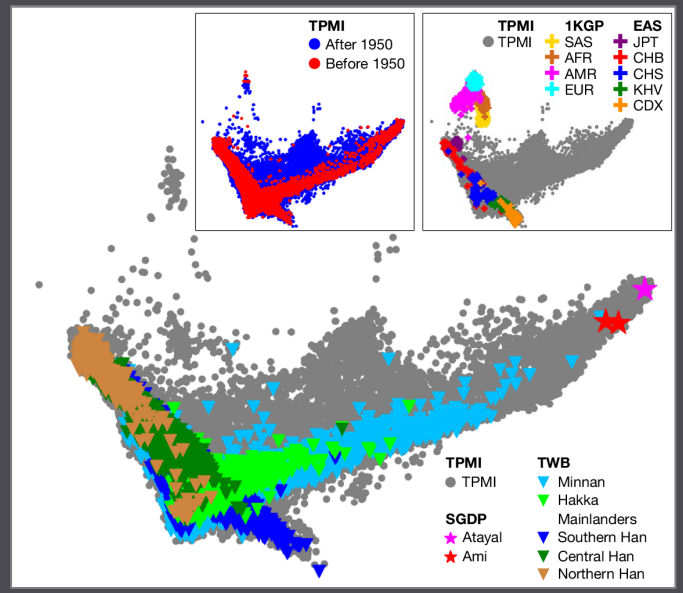




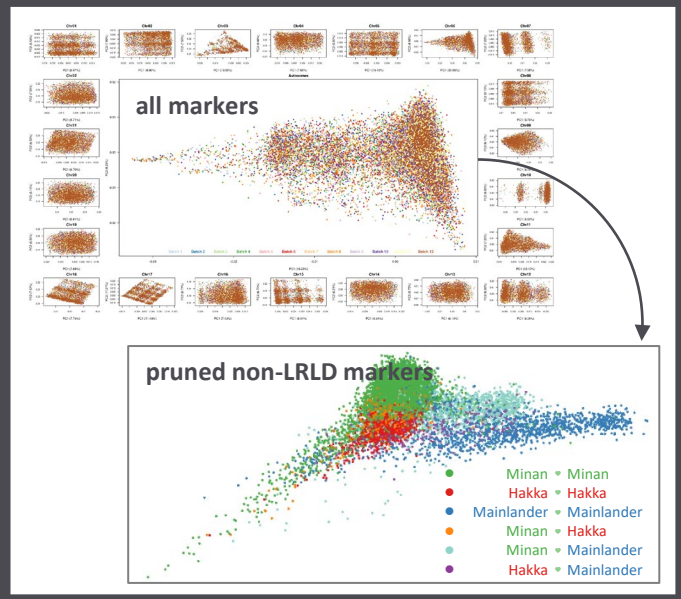
Novembre, J. et al. *Nat.* (2008) **European**



Sabatti, C. et al. *Nat.Genet.* (2009) **Finnland**



Yang et al. *Nat.* (2025) **T**aiwan **P**recision **M**edicine **I**nitiative (**TPMI**)

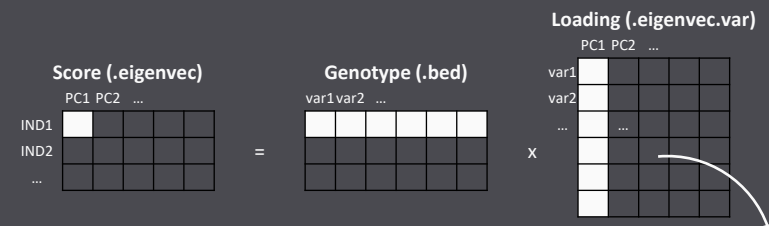


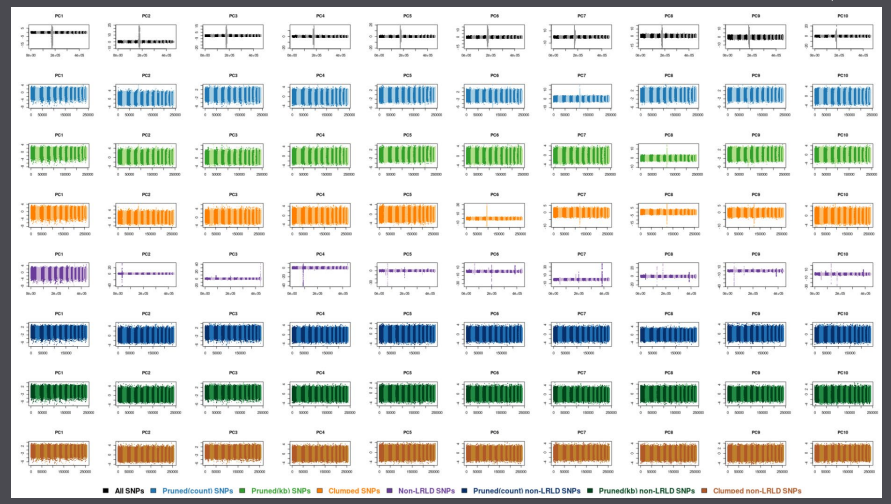
Tai**W**an **B**iobank (TWB1)

Subpopulation structure

How to do?

- Remove long-range LD (LRLD) (ref.)
- Pruning or clumping of markers
- Check the loading plots of PCs





subpopulation structure

```
$   plink2 --bfile dat_as_sampQC_varQC --pca --out dat_as_sampQC_varQC

# pruning by counts of SNPs
$   plink --bfile dat_as_sampQC_varQC --indep-pairwise 5000 500 0.2 --out pruneCnt
$   plink2 –bfile dat_as_sampQC_varQC --extract pruneCnt.prune.in --pca biallelic-var-wts --out dat_as_sampQC_varQC_pruneCnt

# pruning by distance
$   plink --bfile dat_as_sampQC_varQC --indep-pairwise 5000kb 1 0.2 --out pruneKb
$   plink2 –bfile dat_as_sampQC_varQC –extract pruneKb.prune.in --pca biallelic-var-wts --out dat_as_sampQC_varQC_pruneKb

# clumping according to MAF
$   plink2 --bfile dat_as_sampQC_varQC --freq --out dat_qc
$   awk 'NR==1{$(NF+1)="pseudoMAF"} NR>1{$(NF+1)=($5>0.5?$5:(1-$5))}1' dat_as_sampQC_varQC.afreq > dat_as_sampQC_varQC.afreq_
$   plink --bfile dat_as_sampQC_varQC --clump dat_qc.afreq_ --clump-snp-field ID --clump-field pseudoMAF --out dat_qc
$   awk '{print $3}' dat_qc.clumped > dat_qc.clumped.in
$   plink2 --bfile dat_as_sampQC_varQC --extract dat_qc.clumped.in --pca biallelic-var-wts --out dat_as_sampQC_varQC_clump

# remove long-range LD (LRLD) region
$   plink2 --bfile dat_as_sampQC_varQC --exclude range LRLD_GRCh38.txt --pca biallelic-var-wts --out dat_as_sampQC_varQC_rmLRLD
```

:AssociationTest

Binary trait

A clear dichotomous outcome (or state)

* combine ICD codes with lab tests, medications, procedures to improve phenotype accuracy

* super control vs. standard control vs. matched control

* extreme case/control imbalance may require firth logistic regression

Quantitative trait

A continuous (or near-continuous) values where distances are meaningful

* technically discrete values may treat as quantitative if it has many levels

* under model assumption, transformation is usually needed but may not be easy for interpretation

Inverse normal transformation (INT):

$$\mathrm{INT}(x) = \Phi^{-1}\left(\frac{rank(x)-c}{n-2c+1}\right)\begin{cases} c = \frac{3}{8}\ (Blom, 1958) \\ c = \frac{1}{3}\ (Tukey, 1962) \\ c = \frac{1}{2}\ (Bliss, 1967) \end{cases}$$

Covariates

Variables are included in the model to control confounding or improve accuracy

* age: onset age vs. restricted age / recruitment age / current age

* gender

* subpopulation structure: PCs

* others

Model free

Compare **case vs control frequencies directly**, without specifying a regression model

Total (alleles)	A	B		AA $x_0 = 0$	AB $x_1 = 1$	BB $x_2 = 2$	Total (genotypes)
2R	$2r_0 + r_1$	$r_1 + 2r_2$	**case**	r_0	r_1	r_2	R
2S	$2s_0 + s_1$	$s_1 + 2r_2$	**control**	s_0	s_1	s_2	S
2N	$2n_0 + n_1$	$n_1 + 2n_2$		n_0	n_1	n_2	N

Allele-based test

$$\chi^2 = \frac{2N[(2r_0+r_1)(s_1+2r_2)-(r_1+2r_2)(2s_0+s_1)]}{(2R)(2S)(2n_0+n_1)(2n_2+n_1)} \sim \chi^2(1)$$

Genotype-based test

$$\chi^2 = \frac{1}{RS}\sum_{j=0}^{2}\frac{(r_j S - s_j R)^2}{n_j} \sim \chi^2(2)$$

Cochran–Armitage trend test

$$T = \sum_{j=0}^{2} \boldsymbol{x_j}(r_j S - s_j R), \frac{T^2}{Var(T)} \sim \chi^2(1)$$

Fixed-effect model

Account for **covariates** like Age, Sex, or Principal Components (PCs), and assume all predictors' impact are constant across the entire study population (**fixed Effects**)

covariates target SNP

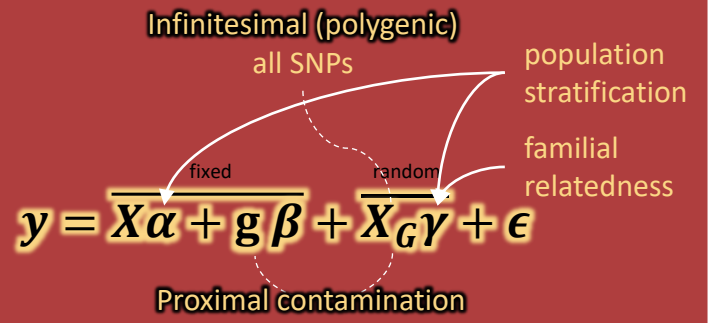
$$y = X\alpha + g\beta + \epsilon$$

BT: Firth logistic regression

QT: Linear regression

Mixed-effect model

Extend linear model by adding **random effects** to model **genetic similarity** (complex population structures and cryptic relatedness)



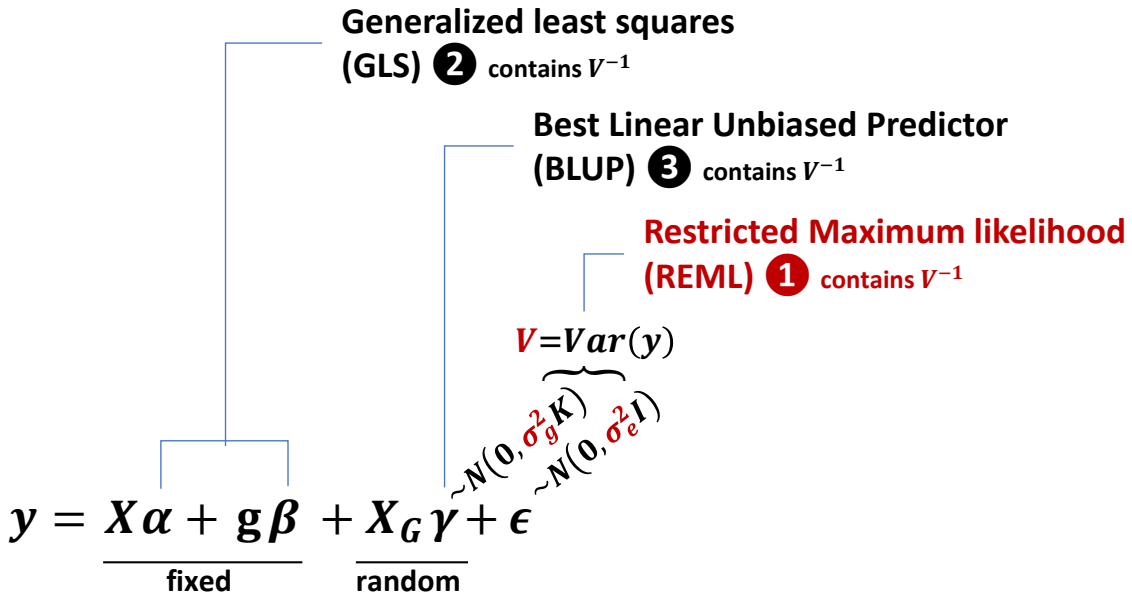
BT: SAIGE REGENIE

QT: BOLT-LMM SAIGE REGENIE

Genomic Control

To reduce false positives inflated by population structure in GWAS (Devlin B & Roeder K. 1999)

$$\lambda_{GC} = \frac{median(\chi^2_{obs})}{median(\chi^2_{null})} \rightarrow \chi^2_{adj} = \frac{\chi^2_{obs}}{\lambda_{GC}}$$



BOLT-LMM	SAIGE	REGENIE
Loh P.R., et al. (2015) Nat Genet.	Zhou W., et al. (2018) Nat Genet.	Mbatchou, J., et al. (2021) Nat Genet.
* a matrix-vector product + conjugate gradient * REML (no explicit **V**)	* sparse GRM * approximate REML	* no GRM * block ridge regression prediction
* Quantitative trait (transformed by yourself) * Binary trait (not recommended) * Precompiled executable	* Quantitative trait (INT provided) * Imbalance case/control (SPA) * Install from source codes	* Quantitative trait (INT provided) * Imbalance case/control (Firth logistic regression) * Precompiled executable

```
plink --bfile dat_as_sampQC_varQC --pheno
dat.phecov --1 --pheno-name PHENO1 --assoc --
out result_bt_allele

plink --bfile dat_as_sampQC_varQC --pheno
dat.phecov --1 --pheno-name PHENO1 --model --
out result_bt_model
```

```
regenie \
  --step 1 \
  --bed dat_as_sampQC_varQC \
  --phenoFile  dat.phecov \
  --covarFile  dat.phecov \
  --covarColList age,sex,PC1,PC2,PC3,PC4,PC5 \
  --bt \
  --bsize  1000
  --lowmem --lowmem-prefix tmp_regenie \
  --out regenie_step1
```

```
regenie \
  --step 2 \
  --bed dat_as_sampQC_varQC \
  --phenoFile  dat.phecov \
  --covarFile  dat.phecov \
  --covarColList age,sex,PC1,PC2,PC3,PC4,PC5 \
  --bt --firth --approx --pThresh 0.01 \
  --bsize  400 \
  --pred regenie_step1_pred.list \
  --out result_bt_regenie
```

```
plink2 \
  --bfile dat_as_sampQC_varQC \
  --pheno dat.phecov --1 \
  --covar dat.phecov \
  --covar-name age,sex,PC1,PC2,PC3,PC4,PC5 \
  --covar-variance-standardize \
  --glm no-x-sex hide-covar
cols=chrom,pos,ref,alt,a1freq,a1freqcc,gcountcc,n
obs,orbeta,se,tz,p,firth,err \
  --out result_bt_plink2
```

Binary trait

Quantitative trait

```
bolt \
  --bfile=dat_as_sampQC_varQC \
  --phenoFile=dat.phecov  --phenoCol=PHENO1 \
  --covarFile=dat.phecov \
  --covarCol=sex --qCovarCol=age,PC{1:5} \
  --lmm \
  --LDscoresFile=LDSCORE.1000G_EAS_m.tab \
  --geneticMapFile=genetic_map_hg38_withX.txt
  --numThreads=8 \
  --statsFile=result_qt_bolt.txt \
```

```
Rscript step1_fitNULLGLMM.R \
  --plinkFile=dat_as_sampQC_varQC \
  --phenoFile=dat.phecov --phenoCol=PHENO1 \
  --covarColList=age,sex,PC1,PC2,PC3,PC4,PC5 \
  --traitType=binary \
  --outputPrefix=saige_null \
  --nThreads=8
```

```
Rscript step2_SPAtests.R \
  --bgenFile=dat_as_sampQC_varQC.bgen \
  --sampleFile=dat_as_sampQC_varQC.sample \
  --GMMATmodelFile=saige_null.rda \
  --varianceRatioFile=saige_null.varianceRatio.txt \
  --LOCO=TRUE \
  --SAIGEOutputFile=result_bt_saige.txt
```

```
plink2 \
  --bfile dat_as_sampQC_varQC \
  --pheno dat.phecov \
  --covar dat.phecov \
  --covar-name age,sex,PC1,PC2,PC3,PC4,PC5 \
 --covar-variance-standardize \
  --glm no-x-sex hide-covar
cols=chrom,pos,ref,alt,a1freq,nobs,orbeta,se,tz,p \
  --out result_qt_plink2
```

.model

CHR SNP A1 A2
TEST (GENO, TREND, ALLELIC, DOM, REC)
AFF UNAFF
CHISQ DF **P**

.regenie

CHROM GENPOS ID ALLELE0 **ALLELE1**
A1FREQ INFO
N TEST **BETA** **SE** CHISQ **LOG10P**
EXTRA

.glm.logistic.hybrid

#CHROM POS ID REF **ALT**
CASE_NON_A1_CT CASE_HET_A1_CT
CASE_HOM_A1_CT CTRL_NON_A1_CT
CTRL_HET_A1_CT CTRL_HOM_A1_CT
A1_FREQ **A1_CASE_FREQ** **A1_CTRL_FREQ**
FIRTH? OBS_CT **OR** **LOG(OR)_SE** Z_STAT **P**
ERRCODE

Binary trait

Quantitative trait

SNP CHR BP GENPOS **ALLELE1** ALLELE0
A1FREQ F_MISS
CHISQ_LIINREG P_LIINREG
BETA **SE**
CHISQ_BOLT_LMM_INF P_BOLT_LMM_INF
CHISQ_BOLT_LMM **P_BOLT_LMM**

(binary trait)

CHR POS SNPID Allele1 **Allele2**
AC_Allele2 **AF_Allele2** imputationInfo
N **BETA** **SE** Tstat Var
p.value (Saddlepoint approximation, SPA)
p.value.NA (Normal approximation)
Is.SPA.converge
varT varTstar **AF.Cases** **AF.Controls**

(quantitative trait)

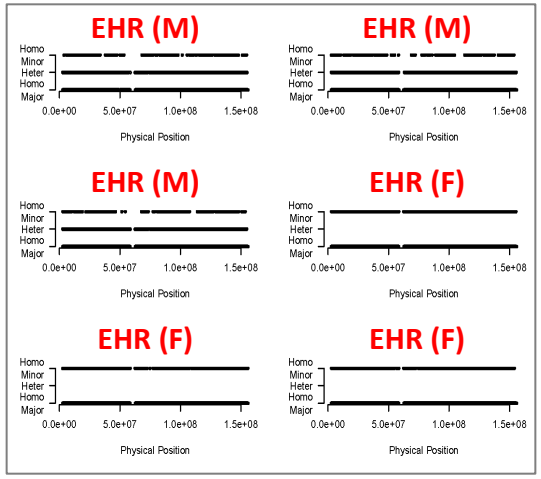
CHR POS SNPID Allele1 **Allele2**
AC_Allele2 **AF_Allele2** imputationInfo
N **BETA** **SE** Tstat **p.value** varT varTstar

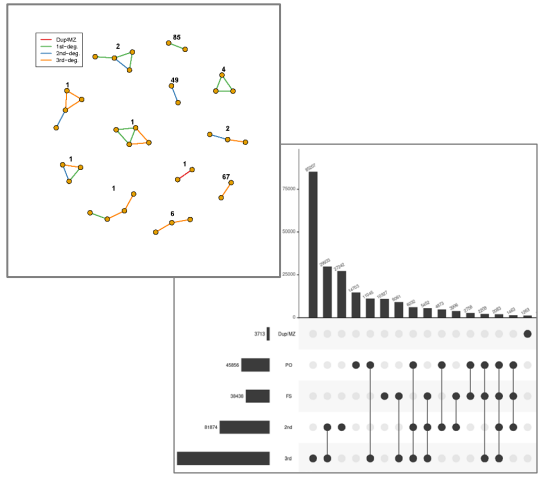
.glm.linear

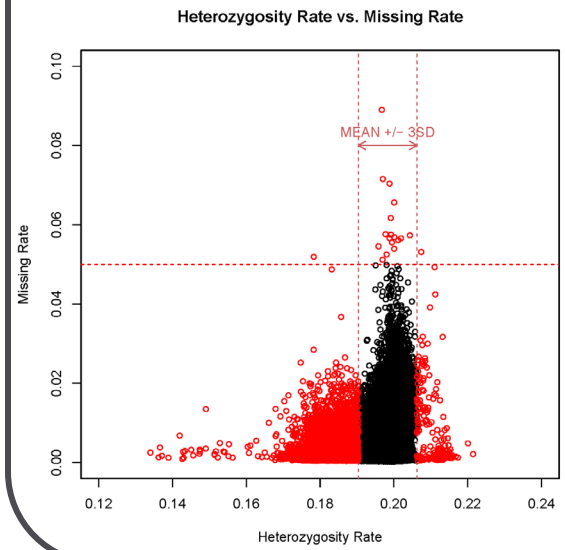
#CHROM POS ID REF **ALT**
A1_FREQ
OBS_CT **BETA** **SE** T_STAT **P**

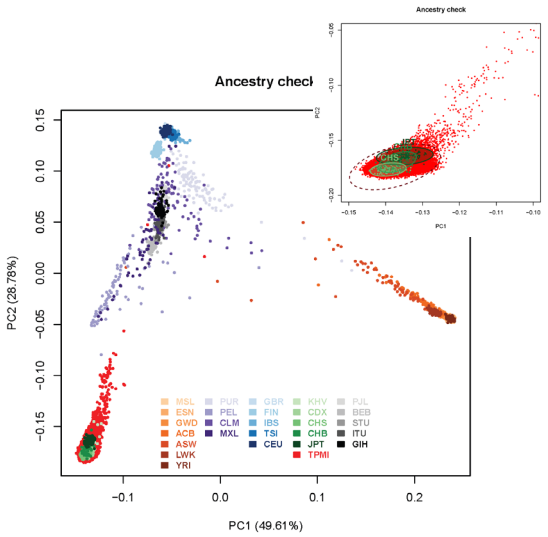
:Figure

Sample QC



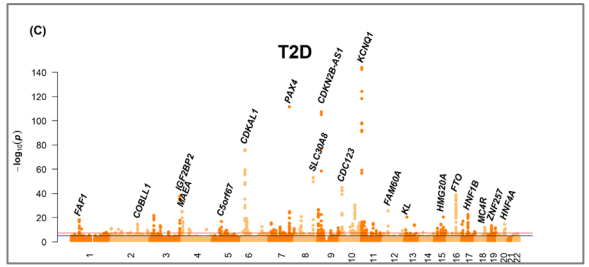




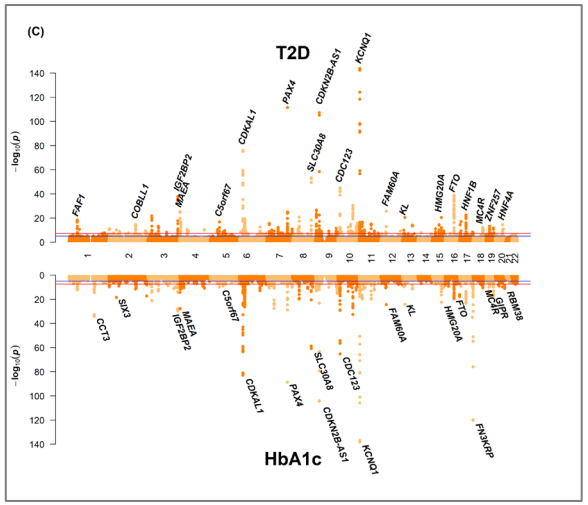


Association test

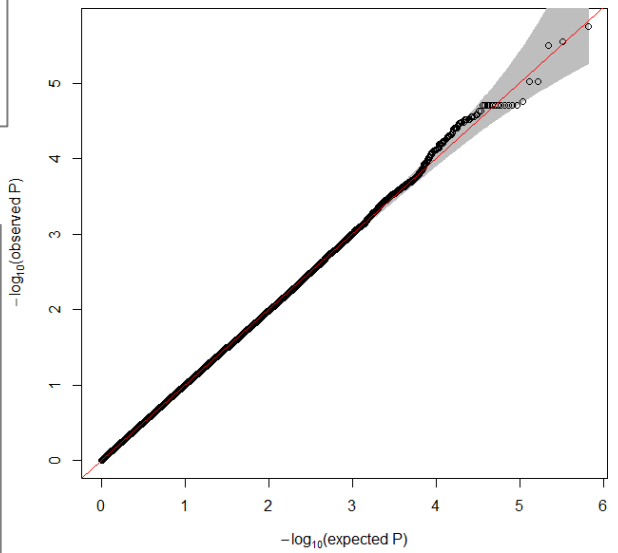
Manhattan plot



Miami plot



QQ plot



Thanks for your attention!!
<(_ _)>