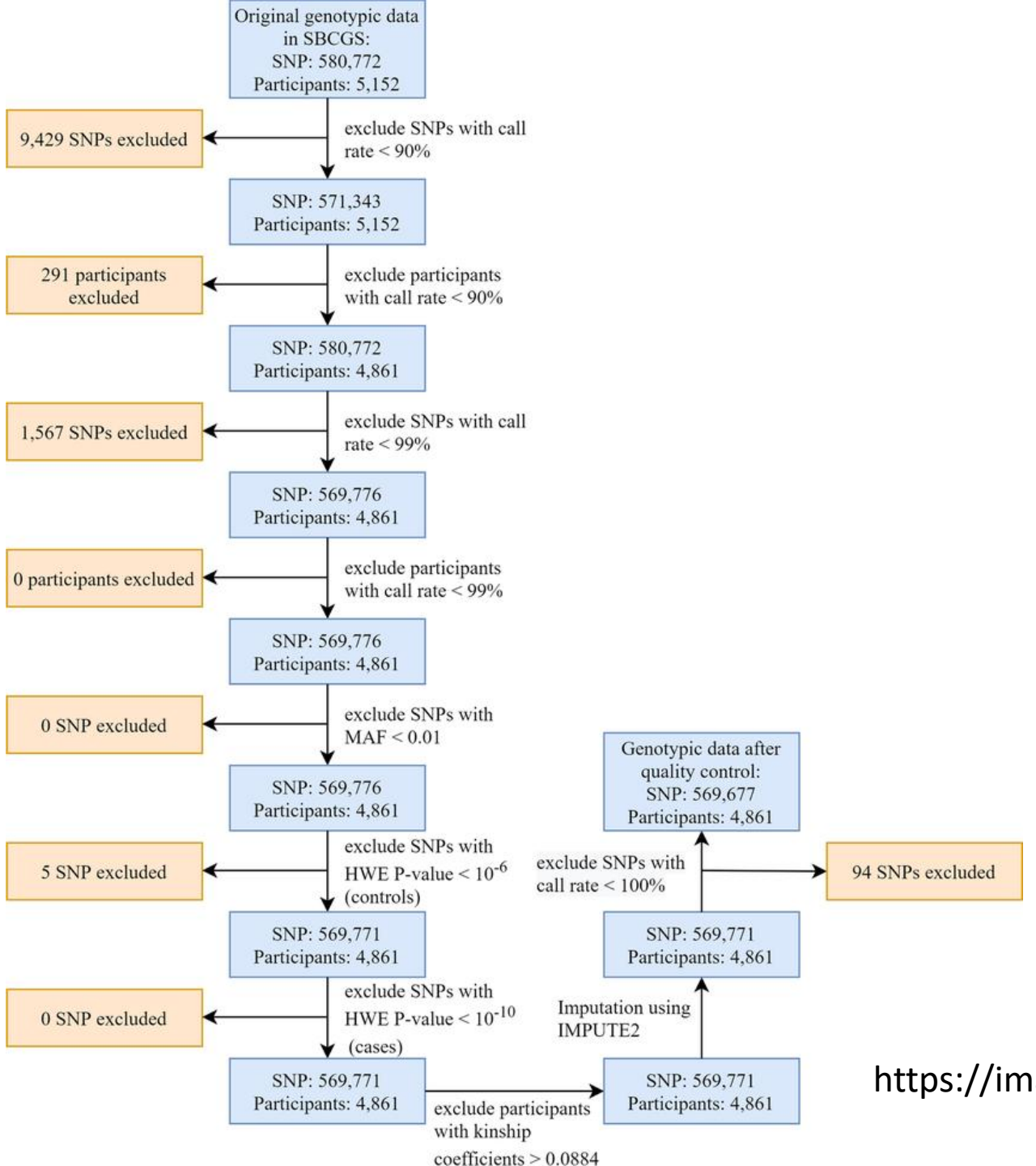
Week 7-GWAS QC



https://images.app.goo.gl/Lv7qnRfiZjXvgqKp7

QC steps

1. Missingness (call rate)
2. Sex discrepancy
3. MAF
4. HWE
5. Heterozygosity
6. Relatedness
7. Population stratification

1. Missingness/Call rate

- Identification of individuals with poor genotype quality

Missing genotypes

To generate a list genotyping/missingness rate statistics:

```
plink --file data --missing
```

This option creates two files:

```
plink.imiss
plink.lmiss
```

which detail missingness by individual and by SNP (locus), respectively. For individuals, the format is:

```
FID             Family ID
IID             Individual ID
MISS_PHENO      Missing phenotype? (Y/N)
N_MISS          Number of missing SNPs
N_GENO          Number of non-obligatory missing genotypes
F_MISS          Proportion of missing SNPs
```

For each SNP, the format is:

```
SNP             SNP identifier
CHR             Chromosome number
N_MISS          Number of individuals missing this SNP
N_GENO          Number of non-obligatory missing genotypes
F_MISS          Proportion of sample missing for this SNP
```

→ --mind 0.05 個人 genotyping 成功率 > 95%

→

Inclusion thresholds

This secion describes options that can be used to filter out individuals or SNPs on the basis of the summary statistic measures described in the previous summary statistics page.

Summary statistics versus inclusion criteria

The following table summarizes the relationship between the commands to generate summary statistics (as described on the previous page, versus the commands to exclude individuals and/or markers, which are described on this page.

Feature	As summary statistic	As inclusion criteria
Missingness per individual	`--missing`	`--mind` *N*
Missingness per marker	`--missing`	`--geno` *N*
Allele frequency	`--freq`	`--maf` *N*
Hardy-Weinberg equilibrium	`--hardy`	`--hwe` *N*
Mendel error rates	`--mendel`	`--me` *N M*

Default threshold values

By default, `PLINK` does not impose any filters on minor allele frequency or genotyping rate. (Note that versions prior to 1.04 use to have thresholds of 0.01 for frequency and 0.1 for individual and SNP missing rate -- this is no longer the case, i.e. it is as if the `--all` keyword is always specified).

To perform an analysis, or generate a new dataset, with filters applied, add the `--mind`, `--geno` or `--maf` options are to the command line, for example, when the `--remove` command is given.

2. Sex discrepancy

Sex check

This option uses X chromosome data to determine sex (i.e. based on heterozygosity rates) and flags individuals for whom the reported sex in the PED file does not match the estimated sex (given genomic data). To run this analysis, use the flag:

```
plink --bfile data --check-sex
```

which generates a file

```
plink.sexcheck
```

which contains the fields

```
FID     Family ID
IID     Individual ID
PEDSEX  Sex as determined in pedigree file (1=male, 2=female)
SNPSEX  Sex as determined by X chromosome
STATUS  Displays "PROBLEM" or "OK" for each individual
F       The actual X chromosome inbreeding (homozygosity) estimate
```

A `PROBLEM` arises if the two sexes do not match, or if the SNP data or pedigree data are ambiguous with regard to sex. A male call is made if `F` is more than 0.8; a femle call is made if `F` is less than 0.2.

The command

```
plink --bfile data --impute-sex --make-bed --out newfile
```

will impute the sex codes based on the SNP data, and create a new file with the revised assignments, in this case a new binary fileset.

5. Heterozygosity

Inbreeding

```
--het ['small-sample'] ['gz']
--ibc
```

--het computes observed and expected autosomal homozygous genotype counts for each sample, and reports method-of-moments F coefficient estimates (i.e. (<observed hom. count> - <expected count>) / (<total observations> - <expected count>)) to `plink.het`. (The '**gz**' modifier has the usual effect.)

Expected counts are based on loaded (via --read-freq) or imputed MAFs; if there are very few samples in your immediate fileset, --read-freq is practically mandatory since imputed MAFs are wildly inaccurate in that case. Also, due to the use of allele frequencies, if your dataset has a highly imbalanced ancestry distribution (e.g. >90% EUR but a few samples with ancestry primarily from other continents), you may need to process the rare-ancestry samples separately.

By default, the **n**/(**n**-1) multiplier in Nei's expected homozygosity formula is now omitted, since **n** may be unknown when using --read-freq. The '**small-sample**' modifier causes the multiplier to be included, while forcing --het to use imputed MAFs (and known **n**s) from founders in the immediate dataset. (--maf-succ is not applied here.)

.het (method-of-moments F coefficient estimates)

Produced by --het.

A text file with a header line, and one line per sample with the following six fields:

FID	Family ID
IID	Within-family ID
O(HOM)	Observed number of homozygotes
E(HOM)	Expected number of homozygotes
N(NM)	Number of (nonmissing, non-monomorphic) autosomal genotype observations
F	Method-of-moments F coefficient estimate

Merge files: https://zzz.bwh.harvard.edu/plink/dataman.shtml#mergelist

Merge multiple filesets

To merge more than two standard and/or binary filesets, it is often more convenient to specify a single file that contains a list of PED/MAP and/or BED/BIM/FAM files and use the `--merge-list` option. Consider, for an extreme example, the case where each fileset contains only a single SNP, and that there are thousands of these files -- this option would help build a single fileset, in this case.

For example, consider we had 4 PED/MAP filesets (labelled `fA.*` through `fD.*`) and 4 binary filesets, labelled `fE.*` through `fH.*`). Then using the command

```
plink --file fA --merge-list allfiles.txt --make-bed --out mynewdata
```

would create the binary fileset

```
mynewdata.bed
mynewdata.bim
mynewdata.fam
```

(alternatively, the `--recode` option could have been used instead of `--make-bed` to generate a standard ASCII PED/MAP fileset). In this case, the file `allfiles.txt` was a list of the to-be-merged files, one set per row:

```
fB.ped fB.map
fC.ped fC.map
fD.ped fD.map
fE.bed fE.bim fE.fam
fF.bed fF.bim fF.fam
fG.bed fG.bim fG.fam
fH.bed fH.bim fH.fam
```

Important Each fileset must be on a line by itself: lines with two files are interpreted as PED/MAP filesets; lines with three files are interpreted as binary BED/BIM/FAM filesets. The files on a line must always be in this order (PED then MAP; BED then BIM then FAM)

Note In this case the first of the 8 files must be the starting file, i.e. associated with `--file` on the command line; this file only contains the 8-1 remaining files therefore. The final `mynewdata.*` files will contain information from all 8 files.

The `--merge-mode` option can also be used with the `--merge-list` option, as described above: however, it is not possible to specify the "diff" features (i.e. modes 6 and 7).

GWAS

Linear and logistic models

These two features allow for multiple covariates when testing for both quantitative trait and disease trait SNP association, and for interactions with those covariates. The covariates can either be continuous or binary (i.e. for categorical covariates, you must first make a set of binary dummy variables).

WARNING! These commands are in some ways more flexible than the standard `--assoc` command, but this comes with a price: namely, these run more slowly...

In this section we consider:

- Basic uasge
- Covariate and interactions
- Flexibly specifying the precise model
- Flexibly specifying joint tests

Basic usage

For quantitative traits, use

```
plink --bfile mydata --linear
```

For disease traits, specify logistic regression with

```
plink --bfile mydaya --logistic
```

instead. All other commands in this section apply equally to both these models.

These commands will either generate the output file

```
plink.assoc.linear
```

or

```
plink.assoc.logistic
```

depending on the phenotype/command used. The basic format is:

```
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
```

For the additive effects of SNPs, the direction of the regression coefficient represents the effect of each extra **minor allele** (i.e. a positive regression coefficient means that the minor allele increases risk/phenotype mean). If the `--beta` command is added along with `--logistic`, then the regression coefficients rather than the odds ratios will be returned.

Ref.

- https://zzz.bwh.harvard.edu/plink/
- https://lsl.sinica.edu.tw/Activities/class/files/202404021007.pdf