# Genome-wide association study

## Data formats in PLINK

Segun Fatumo PhD

UNIVERSITY OF CAMBRIDGE

wellcome trust sanger institute

# Data formats in PLINK

# PLINK Flat files (MAP/PED)

- PLINK is a very widely used application for analyzing genotypic data. It can be considered the “de-facto” standard of the field, although newer formats are starting to be widespread as well.

# MAP files

The **MAP** file describes the SNPs.

The fields in a MAP file are:

- Chromosome
- Marker ID
- Genetic distance
- Physical position

# MAP files

- Chromosome number *[integer]*
- SNP ID *[string]*
- SNP genetic position (cM) *[float]*
- SNP physical position (bp) *[integer]*

- This file should have L lines and 4 columns, where L is the number of SNPs contained in the dataset.

- Each SNP must have a unique physical position. All the SNPs must be ordered by physical position.

# Quick Exercise

- Download c_2150SNPs.map & c_2150SNPs.ped

- View this data

- how many individuals are in the datasets? How
- many SNPs? What is genotyping rate?

# MAP file - Example

| Example of a MAP file of the standard PLINK format: | | | |
|---|---|---|---|
| 21 | rs11511647 | 0 | 26765 |
| X | rs3883674 | 0 | 32380 |
| X | rs12218882 | 0 | 48172 |
| 9 | rs10904045 | 0 | 48426 |
| 9 | rs10751931 | 0 | 49949 |
| 8 | rs11252127 | 0 | 52087 |
| 10 | rs12775203 | 0 | 52277 |
| 8 | rs12255619 | 0 | 52481 |

# Ped (Pedigree)

- The **PED** file describes the individuals and the genetic data. The PED file corresponding to the example dataset is:

# Ped (Pedigree)

- This file can be *SPACE* or *TAB* delimited. Each line corresponds to a single individual. The first 6 columns are:
- **Family ID *[string]***
- **Individual ID *[string]***
- **Father ID *[string]***
- **Mother ID *[string]***
- **Sex *[integer]***
- **Phenotype *[float]***
- Columns 7 & 8 code for the observed alleles at SNP1, columns 9 & 10 code for the observed alleles at SNP2, and so on. Missing data are coded as "0 0" as for example for SNP3 of IND1. This file should have N lines and 2L+6 columns, where N and L are the numbers of individuals and SNPs contained in the dataset respectively.

# Ped (Pedigree)

- Pedigree Name: A unique alphanumeric identifier for this individual's family. Unrelated individuals should not share a pedigree name.
- Individual ID:  An alphanumeric identifier for this individual. Should be unique within his family (see above).
- Father's ID : Identifier corresponding to father's individual ID or "0" if unknown father. Note that if a father ID is specified, the father must also appear in the file.
- Mother's IDIdentifier corresponding to mother's individual ID or "0" if unknown mother Note that if a mother ID is specified, the mother must also appear in the file.
- SexIndividual's gender (1=MALE, 2=FEMALE).
- Affection status

# PED file - Example

| Example of a PED file of the standard PLINK format: | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FAM1 | NA06985 | 0 | 0 | 1 | 1 | A | T | T | T | G | G | C | C | A | T | T | T | G | G | C | C |
| FAM1 | NA06991 | 0 | 0 | 1 | 1 | C | T | T | T | G | G | C | C | C | T | T | T | G | G | C | C |
| 0 | NA06993 | 0 | 0 | 1 | 1 | C | T | T | T | G | G | C | T | C | T | T | T | G | G | C | T |
| 0 | NA06994 | 0 | 0 | 1 | 1 | C | T | T | T | G | G | C | C | C | T | T | T | G | G | C | C |
| 0 | NA07000 | 0 | 0 | 2 | 1 | C | T | T | T | G | G | C | T | C | T | T | T | G | G | C | T |
| 0 | NA07019 | 0 | 0 | 1 | 1 | C | T | T | T | G | G | C | C | C | T | T | T | G | G | C | C |
| 0 | NA07022 | 0 | 0 | 2 | 1 | C | T | T | T | G | G | 0 | 0 | C | T | T | T | G | G | 0 | 0 |
| 0 | NA07029 | 0 | 0 | 1 | 1 | C | T | T | T | G | G | C | C | C | T | T | T | G | G | C | C |
| FAM2 | NA07056 | 0 | 0 | 0 | 2 | C | T | T | T | A | G | C | T | C | T | T | T | A | G | C | T |
| FAM2 | NA07345 | 0 | 0 | 1 | 1 | C | T | T | T | G | G | C | C | C | T | T | T | G | G | C | C |

# PLINK Binary files (BED/BIM/FAM)

- The binary PLINK format contains the same information as the flat file PLINK format but in a compressed and significantly more efficient form.

# BED File

- The BED files are encoded in binary format.
- Binary PED (BED) files.

- We can inspect the BED file with the Unix xxd command, to view a binary file

- ***xxd -b test.bed***
- which generates:

```
0000000: 01101100 00011011 00000001 11011100
00001111 11100111 l..... 0000006: 00001111 01101011
00000001
```

# BIM files

The fields in a BIM file are:

- Chromosome
- Marker ID
- Genetic distance
- Physical position
- Allele 1
- Allele 2

***How has BIM file different from MAP file ?***

# BIM File

**Example of a BIM file of the binary PLINK format:**

| | | | | |
|---|---|---|---|---|
| • 21 | rs115116470 | 26765 | A | T |
| • X | rs3883674 0 | 32380 | C | G |
| • X | rs122188820 | 48172 | T | T |
| • 9 | rs109040450 | 48426 | A | T |
| • 9 | rs107519310 | 49949 | C | T |
| • 8 | rs112521270 | 52087 | A | C |
| • 10 | rs127752030 | 52277 | A | A |
| • 8 | rs122556190 | 52481 | G | T |

# FAM files

- **The fields in a FAM file are**
- Family ID
- Sample ID
- Paternal ID
- Maternal ID
- Sex (1=male; 2=female; other=unknown)
- Affection (0=unknown; 1=unaffected; 2=affected)

- Note: The **FAM file** is just the first six columns of the **PED file**

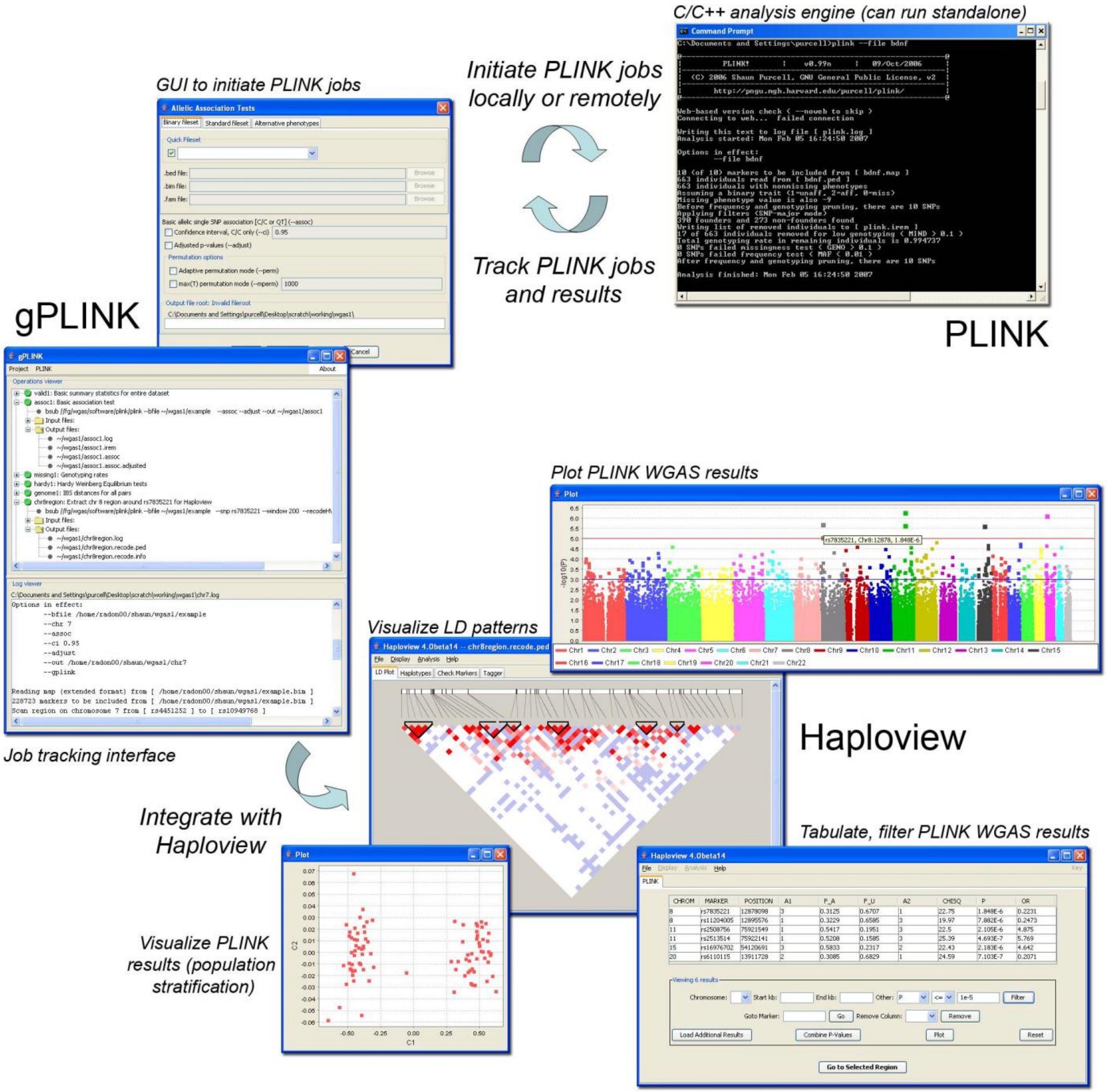
# FAM file

| Example of a FAM file of the binary PLINK format: | | | | | |
|---|---|---|---|---|---|
| FAM1 | NA06985 | 0 | 0 | 1 | 1 |
| FAM1 | NA06991 | 0 | 0 | 1 | 1 |
| 0 | NA06993 | 0 | 0 | 1 | 1 |
| 0 | NA06994 | 0 | 0 | 1 | 1 |
| 0 | NA07000 | 0 | 0 | 2 | 1 |
| 0 | NA07019 | 0 | 0 | 1 | 1 |
| 0 | NA07022 | 0 | 0 | 2 | 1 |
| 0 | NA07029 | 0 | 0 | 1 | 1 |
| FAM2 | NA07056 | 0 | 0 | 0 | 2 |
| FAM2 | NA07345 | 0 | 0 | 1 | 1 |

# PED/MAP to BED/BIM/FAM conversion

- To convert *myPlinkTextData.ped* and *myPlinkTextData.map* in Plink binary format, use Plink as follows:

- plink --file myPlinkTextData --make-bed --out myPlinkBinaryData



C/C++ analysis engine (can run standalone)

```
C:\Documents and Settings\purcell>plink --file bdnf

@----------------------------------------------------------@
|        PLINK!       |     v0.99n      |   09/Oct/2006     |
|----------------------------------------------------------|
|  (C) 2006 Shaun Purcell, GNU General Public License, v2  |
|----------------------------------------------------------|
|       http://pngu.mgh.harvard.edu/purcell/plink/         |
@----------------------------------------------------------@

Web-based version check ( --noweb to skip )
Connecting to web...  failed connection

Writing this text to log file [ plink.log ]
Analysis started: Mon Feb 05 16:24:50 2007

Options in effect:
        --file bdnf

10 (of 10) markers to be included from [ bdnf.map ]
663 individuals read from [ bdnf.ped ]
663 individuals with nonmissing phenotypes
Assuming a binary trait (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
Before frequency and genotyping pruning, there are 10 SNPs
Applying filters (SNP-major mode)
390 founders and 273 non-founders found
Writing list of removed individuals to [ plink.irem ]
17 of 663 individuals removed for low genotyping ( MIND > 0.1 )
Total genotyping rate in remaining individuals is 0.994737
0 SNPs failed missingness test ( GENO > 0.1 )
0 SNPs failed frequency test ( MAF < 0.01 )
After frequency and genotyping pruning, there are 10 SNPs

Analysis finished: Mon Feb 05 16:24:50 2007
```

PLINK

Initiate PLINK jobs locally or remotely

Track PLINK jobs and results

GUI to initiate PLINK jobs

☑

.bed file: Browse
.bim file: Browse
.fam file: Browse

Basic allelic single SNP association [C/C or QT] (--assoc)

☐ Confidence interval, C/C only (--ci) 0.95

☐ Adjusted p-values (--adjust)

Permutation options

☐ Adaptive permutation mode (--perm)

☐ max(T) permutation mode (--mperm) 1000

Output file root: Invalid fileroot

C:\Documents and Settings\purcell\Desktop\scratch\working\wgas1\

gPLINK

```
valid1: Basic summary statistics for entire dataset
assoc1: Basic association test
    bsub //fg/wgas/software/plink/plink --bfile ~/wgas1/example   --assoc --adjust --out ~/wgas1/assoc1
    Input files:
    Output files:
        ~/wgas1/assoc1.log
        ~/wgas1/assoc1.irem
        ~/wgas1/assoc1.assoc
        ~/wgas1/assoc1.assoc.adjusted
missing1: Genotyping rates
hardy1: Hardy Weinberg Equilibrium tests
genome1: IBS distances for all pairs
chr8region: Extract chr 8 region around rs7835221 for Haploview
    bsub //fg/wgas/software/plink/plink --bfile ~/wgas1/example   --snp rs7835221 --window 200  --recodeHV
    Input files:
    Output files:
        ~/wgas1/chr8region.log
        ~/wgas1/chr8region.recode.ped
        ~/wgas1/chr8region.recode.info
```

Log viewer

C:\Documents and Settings\purcell\Desktop\scratch\working\wgas1\chr7.log

```
Options in effect:
        --bfile /home/radon00/shaun/wgas1/example
        --chr 7
        --assoc
        --ci 0.95
        --adjust
        --out /home/radon00/shaun/wgas1/chr7
        --gplink

Reading map (extended format) from [ /home/radon00/shaun/wgas1/example.bim ]
228723 markers to be included from [ /home/radon00/shaun/wgas1/example.bim ]
Scan region on chromosome 7 from [ rs4451252 ] to [ rs10949768 ]
```

Job tracking interface

Plot PLINK WGAS results

Visualize LD patterns

Haploview

Integrate with Haploview

Visualize PLINK results (population stratification)

Tabulate, filter PLINK WGAS results

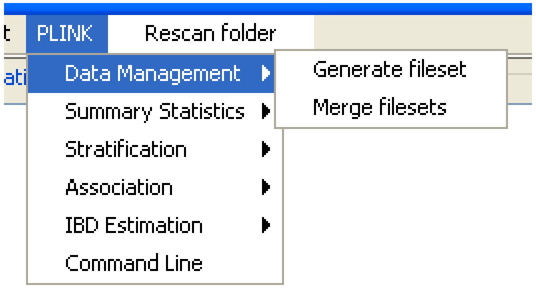
| CHROM | MARKER | POSITION | A1 | F_A | F_U | A2 | CHISQ | P | OR |
|---|---|---|---|---|---|---|---|---|---|
| 8 | rs7835221 | 12878098 | 3 | 0.3125 | 0.6707 | 1 | 22.75 | 1.848E-6 | 0.2231 |
| 8 | rs11204005 | 12895576 | 1 | 0.3229 | 0.6585 | 3 | 19.97 | 7.882E-6 | 0.2473 |
| 11 | rs2508756 | 75921549 | 1 | 0.5417 | 0.1951 | 3 | 22.5 | 2.105E-6 | 4.875 |
| 11 | rs2513514 | 75922141 | 1 | 0.5208 | 0.1585 | 3 | 25.39 | 4.693E-7 | 5.769 |
| 15 | rs16976702 | 54120691 | 3 | 0.5833 | 0.2317 | 2 | 22.43 | 2.183E-6 | 4.642 |
| 20 | rs6110115 | 13911728 | 2 | 0.3085 | 0.6829 | 1 | 24.59 | 7.103E-7 | 0.2071 |

Viewing 6 results

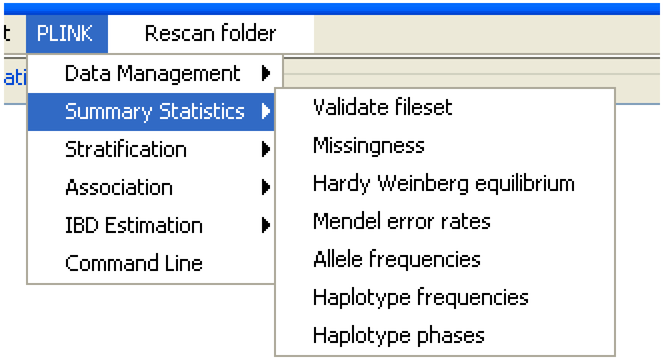
Chromosome: Start kb: End kb: Other: P <= 1e-5

# GUI for many **PLINK** analyses

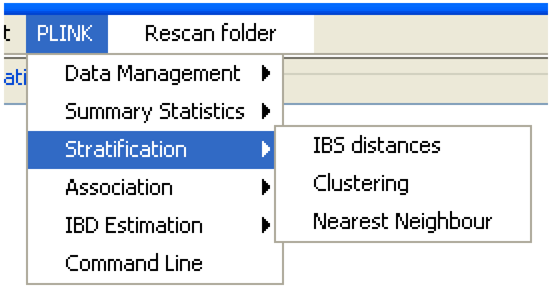
### *Data management*



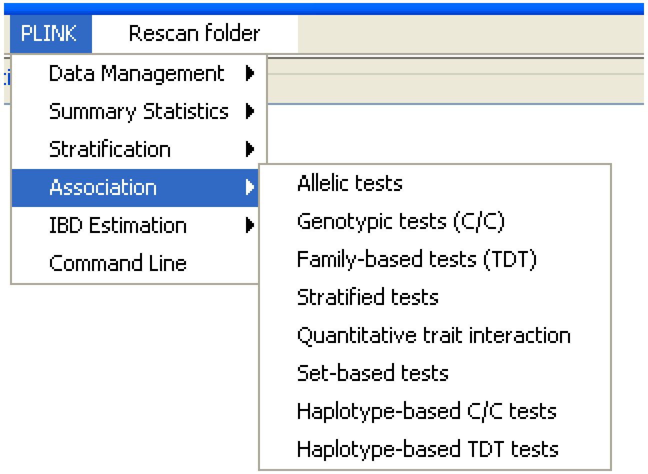
### *Summary statistics*



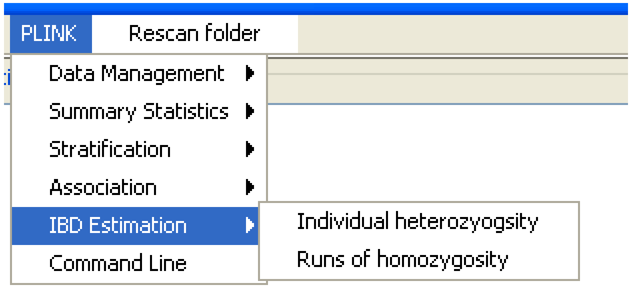
### *Population stratification*



### *Association analysis*



### *IBD-based analysis*

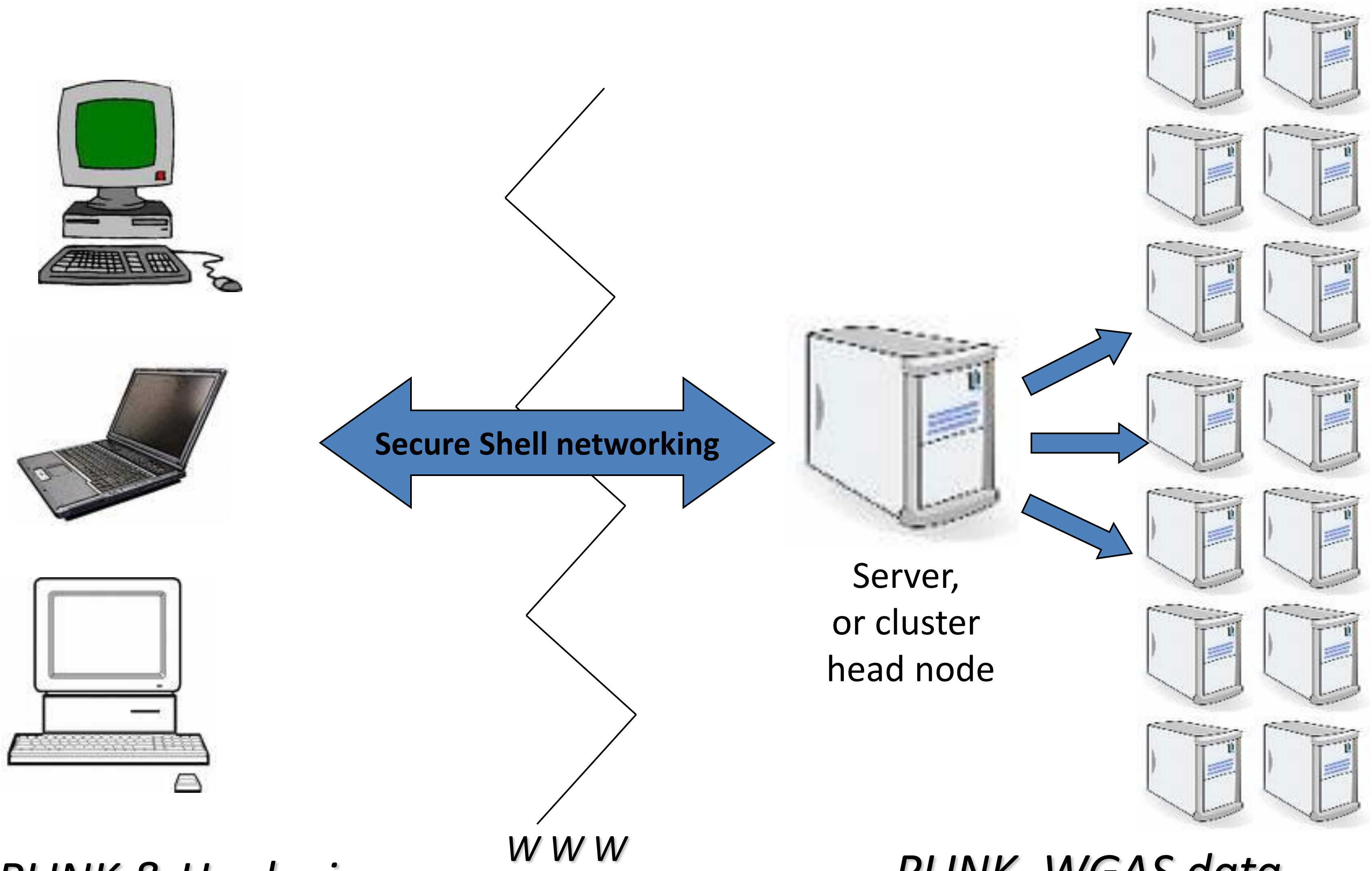


# Computational efficiency

*350 individuals genotyped on 100,000 SNPs*

| | |
|---|---|
| Load, filter and analyze | ~12 seconds |
| 1 permutation (all SNPs) | ~1.6 seconds |

# gPLINK / PLINK in “remote mode”



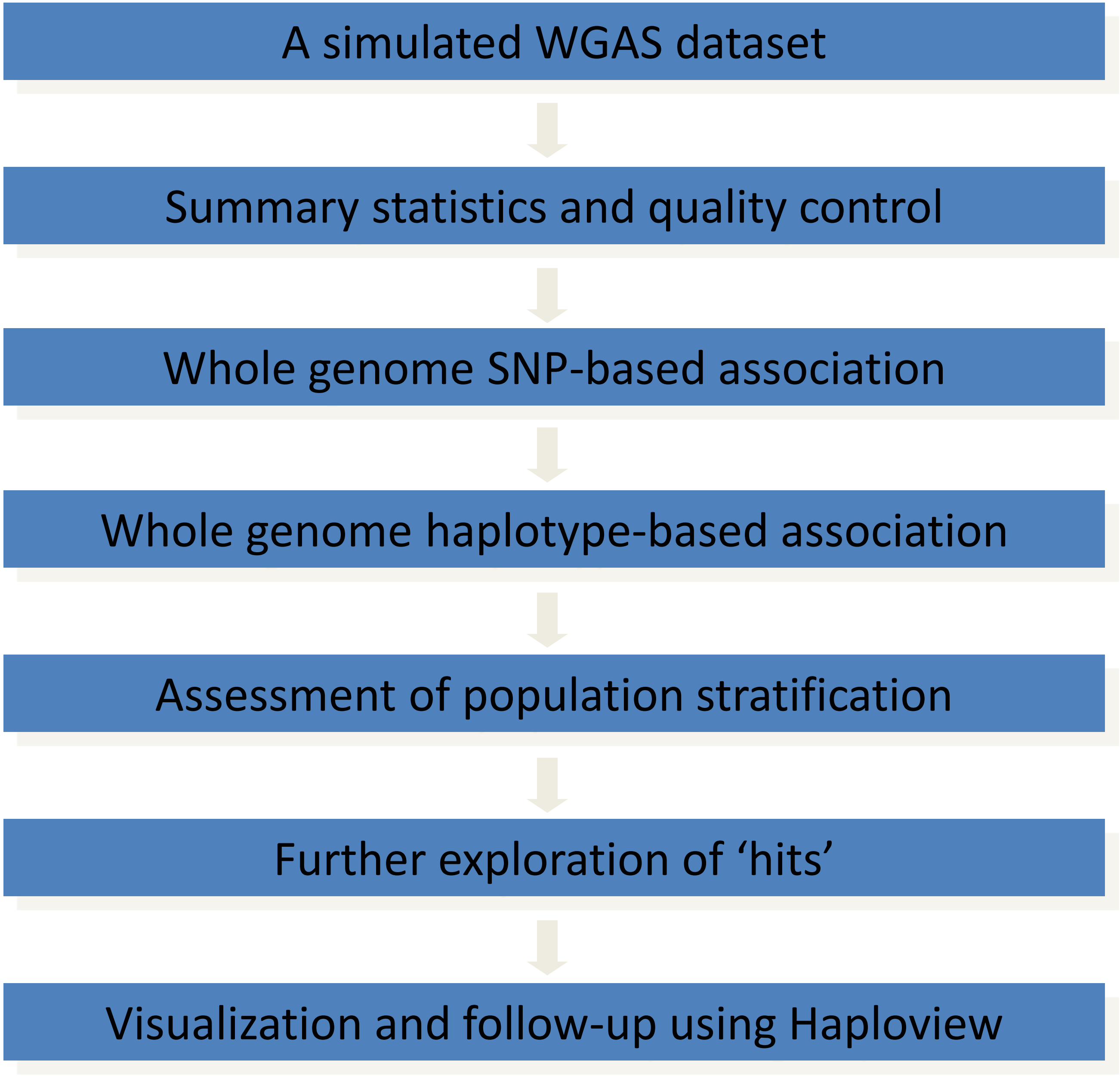
Secure Shell networking

Server,
or cluster
head node

W W W

*gPLINK & Haploview:*
*initiating and viewing jobs*

*PLINK, WGAS data*
*& computation*



A simulated WGAS dataset

Summary statistics and quality control

Whole genome SNP-based association

Whole genome haplotype-based association

Assessment of population stratification

Further exploration of ‘hits’

Visualization and follow-up using Haploview

# Exercise

- Create a simulated class.ped and class.med files
- Convert your ped & map file to binary file

- **Assumption:**

-You may assume that there are 20 students in this class.

-You may assume also that two students are brothers

# Manipulating the data files

- Get only the genotypes for a single chromosome or a region around a
- SNP
- --chr 13
- Exercise: Get data from chromosome 13 and write to a new BED file. If
- you are having trouble running the full dataset, you can use this fileset

instead of fulldataset.

- plink --bfile bipolar --chr 13 --make-bed --out bipolar_chr13