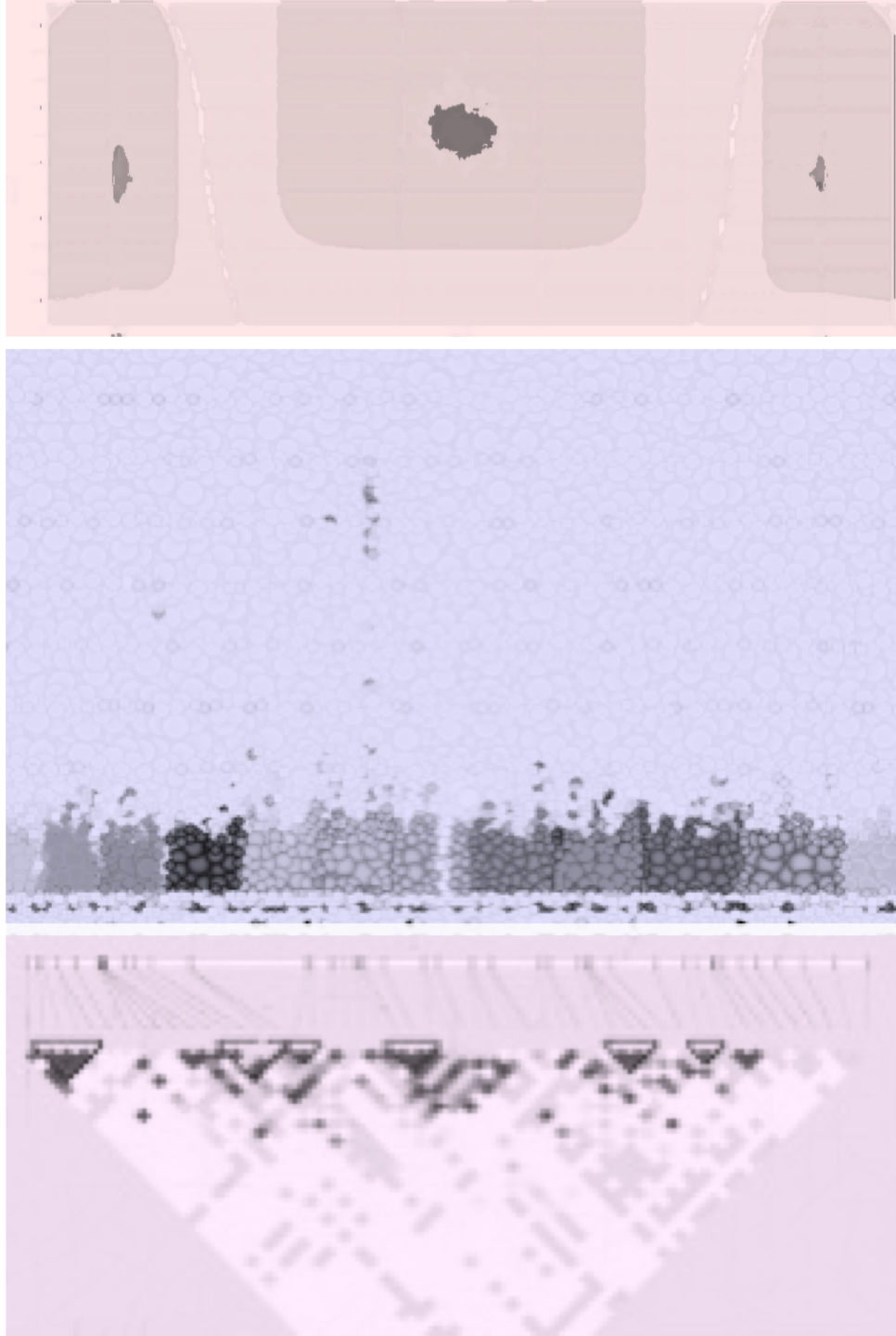
HKU Med School of Clinical Medicine
Department of Surgery
香港大學外科學系

Introduction to data formats and PLINK

Clara Tang

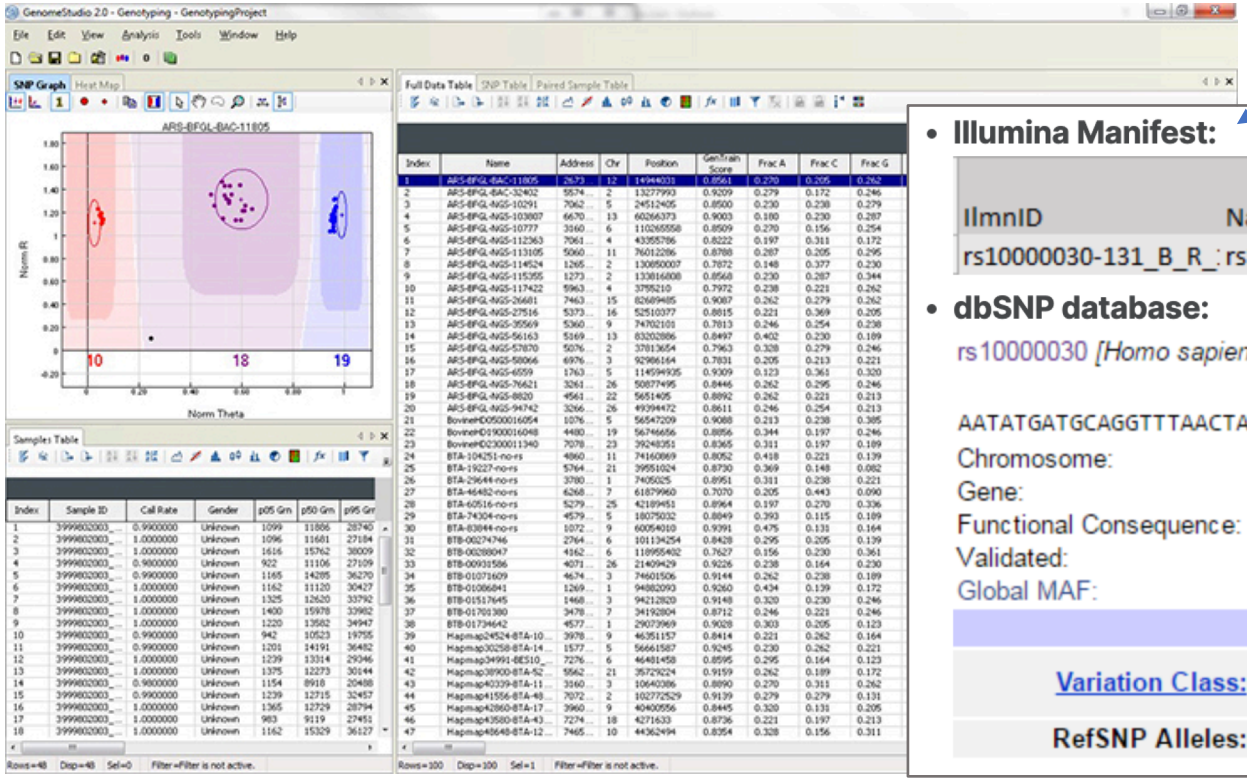
claratang@hku.hk

June 2022



Genotyping

Illumina GenomeStudio



Choose the build (GRCh37/38) that matched with your preferred imputation server

- **Illumina Manifest:**

IlmnID	Name	Ilmn Strand	SNP	Genome Build	Chr	MapInfo	Source Strand	SourceSeq	Ref Strand
rs10000030-131_B_R_:	rs10000030	BOT	[T/C]	37	4	1E+08	TOP	AGC[A/G]GGC	-

- **dbSNP database:**

rs10000030 *[Homo sapiens]*

AATATGATGCAGGTTTAACTACAGC[A/G]GGCTTAAGAAAAGATACCAATATCT

Chromosome:	4:102452997
Gene:	LOC105377621 (GeneView)
Functional Consequence:	intron variant
Validated:	by 1000G,by 2hit 2allele,by cluster,by frequency,by hapmap
Global MAF:	A=0.1671/837
Allele	
Variation Class:	SNV: single nucleotide variation
RefSNP Alleles:	A/G (FWD)

[T/C]: BOT, minus (-), REV
[**A/G**]: TOP, **plus (+)**, FWD

PLINK Input Report Plug-in

Top (TOP) strand

Plus (+) strand

Forward (FWD) strand

PLINK 1 binary file

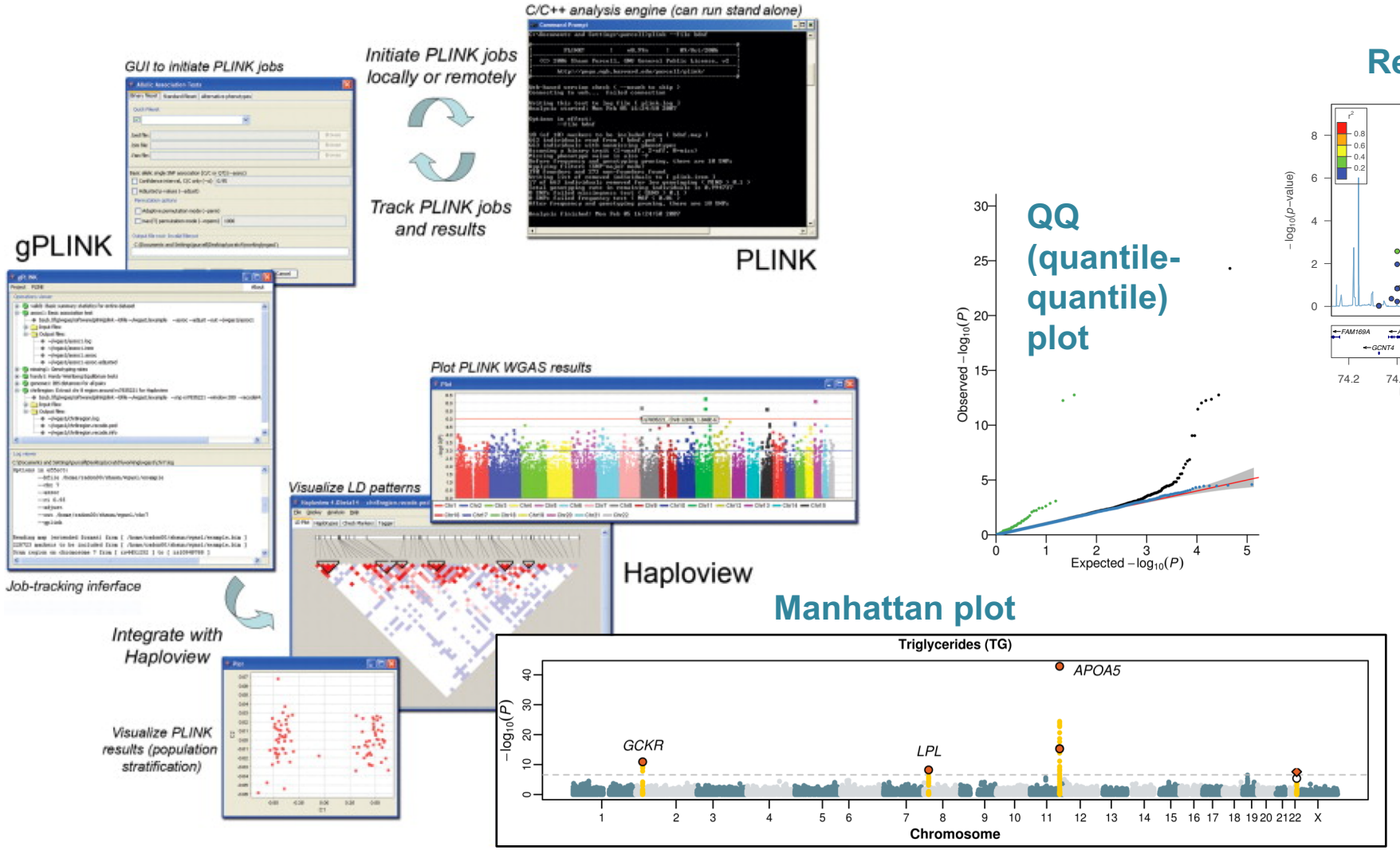
.bed
.bim
.fam

.bim

```
4 rs1000030 102452997 G A
```

PLINK: Population-based LINKage analyses

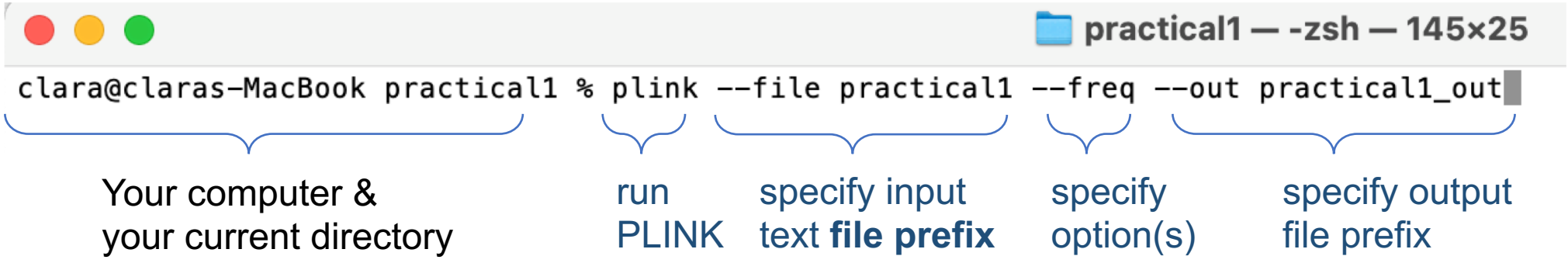
- Command-line tool for genotype data management and genome-wide association analyses



Purcell, Sham *et al. AJHG* **81(3):**559-575 (2007)

PLINK: Population-based LINKage analyses

- Command-line tool for genotype data management and genome-wide association analyses
  - Windows : Putty
  - **Linux** and Mac : **Terminal**




```
PLINK v1.90b6.6 64-bit (10 Oct 2018)           www.cog-genomics.org/plink/1.9/
(C) 2005-2018 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to practical1_out.log.
Options in effect:
  --file practical1
  --freq
  --out practical1_out

8192 MB RAM detected; reserving 4096 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (59 variants, 2506 people).
--file: practical1_out-temporary.bed + practical1_out-temporary.bim +
practical1_out-temporary.fam written.
59 variants loaded from .bim file.
2506 people (1233 males, 1273 females) loaded from .fam.
2506 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 2504 founders and 2 nonfounders present.
Calculating allele frequencies... done.
--freq: Allele frequencies (founders only) written to practical1_out.frq .
```

log file
`.log`

```
--file: practical1_out-temporary.bed
+ practical1_out-temporary.bim
+ practical1_out-temporary.fam written
```

PLINK v1.07

https://zzz.bwh.harvard.edu/plink/download.shtml

plink...

*Last original PLINK release is **v1.07** (10-Oct-2009);*

Whole genome association analysis toolset

Introduction / *Basics* / *Download* / *Reference* / *Formats* / *Data management* / *Summary stats* / *Filters* / *Stratification* / *IBS/IBD* / *Association* / *Family-based* / *Permutation* / *LD calcualtions* / *Haplotypes* / *Conditional tests* / *Proxy association* / *Im*

Result annotation / *Clumping* / *Gene Report* / *Epistasis* / *Rare CNVs* / *Common CNPs* / *R-plugins* / *SNP annotation* / *Simulation* / *Profiles* / *ID helper* / *Resources* / *Flow chart* / *Misc.* / *FAQ* / *gPLINK*

Getting started with PLINK

This page contains some important information on learning to use PLINK and how to handle any problems you encounter.

We suggest that after downloading PLINK you first try the tutorial. This will familiarize you with the basic PLINK commands.

Citing PLINK

If you use PLINK in any published work, please cite both the software (as an electronic resource/URL) and the manuscript descr

```
Package:     PLINK (including version number)
Author:      Shaun Purcell
URL:         http://pngu.mgh.harvard.edu/purcell/plink/

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR,
Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007)
PLINK: a toolset for whole-genome association and population-based
linkage analysis. American Journal of Human Genetics, 81.
```

PLINK 1.9 and 2.0

https://www.cog-genomics.org/plink2/

The following documented PLINK 1.07 flags are not supported by 1.90 beta 6:

- --qual-geno-scores[3]
- --segment[4]
- --dfam
- --tucc
- --p2, --genedrop
- --hap, --hap-window, --hap-snps[5]
- --proxy-assoc, --proxy-impute[5]
- --cnv-list, --cfile, --gfile
- --id-dict, --id-match[6]
- --compress, --decompress[7]

PLINK 1.9 home | plink2-users

Introduction, downloads
S: 2 Apr 2022 (b6.26)
D: 2 Apr 2022
Recent version history
What's new?
Future development
Limitations
Note to testers

[Jump to search box]

General usage
Getting started
Citation instructions

Standard data input
PLINK 1 binary (.bed)
Autoconversion behavior
PLINK text (.ped, .tped...)
VCF (.vcf[.gz], .bcf)
Oxford (.gen[.gz], .bgen)
23andMe text
Generate random
Unusual chromosome IDs
Recombination map
Allele frequencies
Phenotypes

PLINK 1.9

This is a comprehen
Christopher Chang
and others. (What's
users Google grou

Binary downloads

	Build		
Operating system[1]	Stable (beta 6.26, 2 Apr)	Development (2 Apr)	Old[2] (v1.07)
Linux 64-bit	**download**	download	download
Linux 32-bit	**download**	download	download
macOS (64-bit)[3]	**download**	download	download (32-bit)
Windows 64-bit	**download**	download	download
Windows 32-bit	**download**	download	

1: Solaris is no longer explicitly supported, but it should be able to run the Linux binaries.
2: These are just mirrors of the binaries posted at https://zzz.bwh.harvard.edu/plink/download.shtml.
3: You need to have Rosetta 2 installed to run this on M1 Macs.

PLINK 1.9 and 2.0

PLINK automatically converts most other formats to binary files before executing the commands

- PLINK 1.9: https://www.cog-genomics.org/plink/1.9/ ⇦ Relatively stable
- PLINK 2.0: https://www.cog-genomics.org/plink/2.0/ ⇦ Under development; keep updated with new features

PLINK 1 binary

bfile	bed / bim / fam

.bed

Contains binary version of the SNP info of the *.ped file. (not in a format readable for humans)

.bim

Chr	SNP	GD	BPP	Allele 1	Allele 2
1	rs1	0	870000	C	T
1	rs2	0	880000	A	G
1	rs3	0	890000	A	C

A1 **A2**

.fam

FID	IID	PID	MID	Sex	P
1	1	0	0	2	1
2	2	0	0	1	0
3	3	0	0	1	1

No header

.bim

1. Chromosome
2. Variant id
3. Position (in morgans or centimorgans)
4. Base-pair coordinate
5. **A1:** minor allele
6. **A2:** major allele

A1 allele is the allele that most PLINK functions make reference to

PLINK 1.9 and 2.0

PLINK automatically converts most other formats to binary files before executing the commands

- PLINK 1.9: https://www.cog-genomics.org/plink/1.9/ ⇦ Relatively stable
- PLINK 2.0: https://www.cog-genomics.org/plink/2.0/ ⇦ Under development; keep updated with new features

PLINK 1 binary

bfile	bed / bim / fam

.bed

Contains binary version of the SNP info of the *.ped file. (not in a format readable for humans)

No header

.bim

Chr	SNP	GD	BPP	Allele 1	Allele 2
1	rs1	0	870000	C	T
1	rs2	0	880000	A	G
1	rs3	0	890000	A	C

.fam

FID	IID	PID	MID	Sex	P
1	1	0	0	2	1
2	2	0	0	1	0
3	3	0	0	1	1

.fam

1. Family ID
2. Individual ID
3. Paternal ID
4. Maternal ID
5. Sex (2=Female; 1=Male)
6. Phenotype

```
FAM1    IND1    0       0       1       1
FAM1    IND2    0       0       2       1
FAM1    IND3    IND1    IND2    1       2
```

Father, Mother] founder

Son] nonfounder

PLINK 1.9 and 2.0

PLINK automatically converts most other formats to binary files before executing the commands

- PLINK 1.9: https://www.cog-genomics.org/plink/1.9/ ⇦ Relatively stable
- PLINK 2.0: https://www.cog-genomics.org/plink/2.0/ ⇦ Under development; keep updated with new features

PLINK 1 binary

bfile	bed / bim / fam

PLINK 2 binary

pfile	pgen / pvar / psam

.bed

Contains binary version of the SNP info of the *.ped file. (not in a format readable for humans)

.bim

Chr	SNP	GD	BPP	Allele 1	Allele 2
1	rs1	0	870000	C	T
1	rs2	0	880000	A	G
1	rs3	0	890000	A	C

A1 A2

.fam

FID	IID	PID	MID	Sex	P
1	1	0	0	2	1
2	2	0	0	1	0
3	3	0	0	1	1

No header

.pgen

.pvar

- No GD
- #Header

.psam

- #Header

Pros for next generation sequencing data

- Reliable tracking of **REF** vs. **ALT** alleles
- Compression of low-MAF variants
- Phased genotypes.
- Dosages.
- VCF-style header information
- Multiallelic variants.

PLINK functions

- Data management
  - Read genotype data in binary or multiple text formats

PLINK functions

- Data management
  - Read genotype data in binary or multiple text formats

PLINK 1 text and binary

plink		
	`--file`	`ped + map`
	`--tfile`	`tped + tfam`
	`--lfile`	`lgen + map + ped`
	`--bfile`	**`bed + bim + fam`**

`.ped`

FID	IID	PID	MID	Sex	P	rs1	rs2	rs3
1	1	0	0	2	1	CT	AG	AA
2	2	0	0	1	0	CC	AA	AC
3	3	0	0	1	1	CC	AA	AC

`.map`

Chr	SNP	GD	BPP
1	rs1	0	870000
1	rs2	0	880000
1	rs3	0	890000

`--file`

- one individual per row
- one variant per column

`--tfile (transposed)`

- one individual per row
- one variant per column

`--lfile (long format)`

- one individual one variant per row/column

PLINK functions

- Data management
  - Read genotype data in binary or multiple text formats
    https://www.cog-genomics.org/plink/1.9/formats

```
plink   --vcf     vcf / vcf.gz
        --bcf     bcf
```

Variant Calling Format (VCF) / binary VCF
- Standard format for sequencing data
- e.g. 1000 Genomes Project genotypes
- Dosage file after imputation from imputation servers

```
        --data    gen / bgen + sample
```

Oxford format
- Format for Oxford statistical genetics tools
- e.g. imputation-related tools IMPUTE2 and SNPTEST
- Triplet of values indicate likelihoods of homozygote A1, heterozygote, and homozygote A2 genotypes

PLINK functions

- Data management
  - Read genotype data in binary or multiple text formats
  - Output genotype data in **binary** or multiple **text** formats

```
plink --bfile <filename> --make-bed --out outfile
```

make-bed	bed + bim + fam

```
plink --bfile <filename> --recode --out outfile
```

recode	ped + map
recode transpose	tped + tfam
recode lgen	lgen + map + ped
recode vcf recode vcf-fid recode vcf-iid	vcf
recode A	additive (0/1/2)

PLINK functions

- Data management
  - Read genotype data in binary or multiple text formats
  - Output genotype data in binary or multiple text formats
  - Merge multiple **binary** or multiple **text** formats with various merge modes

```
plink --bfile <filename> --bmerge --out outfile
plink --bfile <filename> --merge --out outfile
plink --bfile <filename> --merge-list --out outfile
```

--merge-mode <mode number>

1. (default) Ignore missing calls, otherwise set mismatches to missing
2. Only overwrite calls which are missing in the original file
3. Only overwrite calls which are nonmissing in the new file
4. Never overwrite
5. Always overwrite
6. **(no merge) Report all mismatching calls**
7. **(no merge) Report mismatching nonmissing calls**

if use with --merge-list, genotypes replacement depends on the ordering in the list file

PLINK functions

- Data management
  - Read genotype data in binary or multiple text formats
  - Output genotype data in binary or multiple text formats
  - Merge multiple binary or multiple text formats with various merge modes
  - Update variant and pedigree file(s)

```
--update-name <filename>
--update-map <filename>
--update-alleles <filename>
```

PLINK functions

- Data management
  - Read genotype data in binary or multiple text formats
  - Output genotype data in binary or multiple text formats
  - Merge multiple binary or multiple text formats with various merge modes
  - Update variant and pedigree file(s)
  - Flip DNA strand of variants

PLINK functions

- Data management
  - Read genotype data in binary or multiple text formats
  - Output genotype data in binary or multiple text formats
  - Merge multiple binary or multiple text formats with various merge modes
  - Update variant and pedigree file(s)
  - Flip DNA strand of variants
  - Read phenotype and covariate data

`--pheno <file>`	Read phenotype file
`--mpheno <col num>`	Specify column of phenotype to be used
`--pheno-name <name>`	Specify header name of phenotype to be used
`--covar <file>`	Read covariate file
`--covar-name <col num(s)>`	Specify column number of covariate(s) to be used
`--covar-name <name(s)>`	Specify header name of covariate(s) to be used

PLINK functions

- Data management
  - Read genotype data in binary or multiple text formats
  - Output genotype data in binary or multiple text formats
  - Merge multiple binary or multiple text formats with various merge modes
  - Update variant and pedigree file(s)
  - Flip DNA strand of variants
  - Read phenotype and covariate data
- Input filtering
- Basic summary statistics
- Relatedness matrices
- Population stratification
- Association test

PLINK functions : input filtering

--keep <file>	Keep samples in specified file
--remove <file>	Remove samples in specified file
--filter-cases --filter-controls	Keep only cases (or controls)
--mind <max>	Remove samples with missingness > max threshold
--extract <file>	Extract variants in specified file
--exclude <file>	Exclude variants in specified file
--autosome --chr <chr(s)> --from <num> --to <num>	Specify the chromosome or chromosomal positions to be extracted
--snp <name> --snps <names>	Extract only specified variant(s)
--geno <max>	Exclude variants with missingness > max threshold
--maf <min>	Exclude variants with maf < min threshold
--hwe <min>	Extract variants with Hardy-Weinberg equilibrium exact test p-value < min threshold in controls

PLINK functions: Basic summary statistics

--freq	Output minor allele frequency (MAF)
--missing	Output missingness information per variant and per sample
--hardy	Output a list of genotype counts and Hardy-Weinberg equilibrium exact test statistics • by phenotype ('ALL', 'AFF', and 'UNAFF')
--mendel	Output summaries of Mendelian errors per sample and per variant
--het	Output observed and expected autosomal homozygous genotype counts and inbreeding coefficient
--check-sex	Output comparison of sex assignment in pedigree and genotype data

PLINK other functions

PLINK 1.90 beta

This is a comprehensive update to Shaun Purcell's PLINK command-line program, developed by Christopher Chang with support from the NIH-NIDDK's Laboratory of Biological Modeling, the Purcell Lab, and others. (What's new?) (Credits.) (Methods paper.) (Usage questions should be sent to the **plink2-users Google group**, not Christopher's email.)

Binary downloads

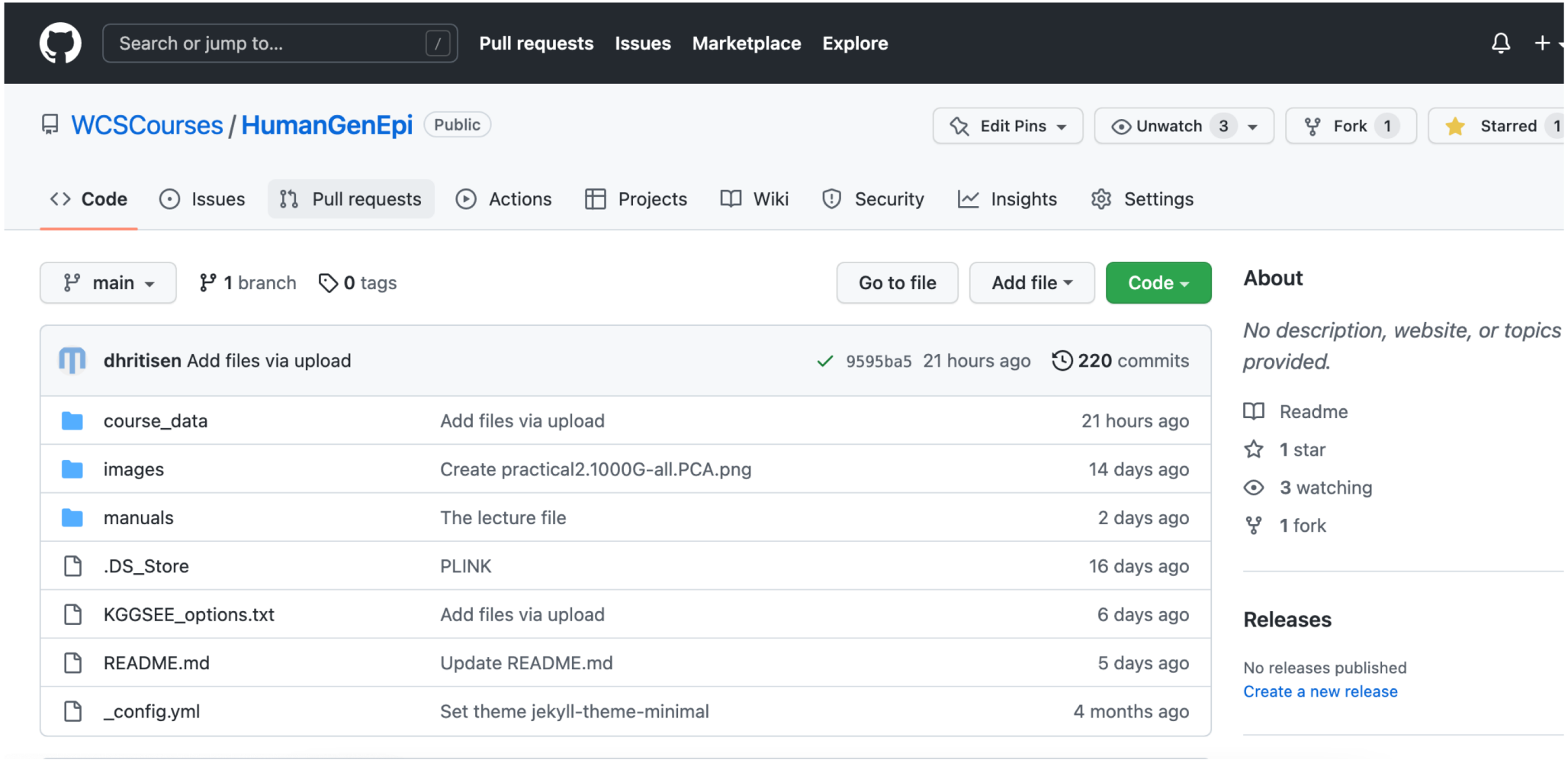
	Build		
Operating system[1]	**Stable (beta 6.26, 2 Apr)**	Development (2 Apr)	Old[2] (v1.07)
Linux 64-bit	**download**	download	download
Linux 32-bit	**download**	download	download
macOS (64-bit)[3]	**download**	download	download (32-bit)
Windows 64-bit	**download**	download	download
Windows 32-bit	**download**	download	

1: Solaris is no longer explicitly supported, but it should be able to run the Linux binaries.
2: These are just mirrors of the binaries posted at https://zzz.bwh.harvard.edu/plink/download.shtml.
3: You need to have Rosetta 2 installed to run this on M1 Macs.

Practical on PLINK

https://github.com/WCSCourses/HumanGenEpi

https://github.com/WCSCourses/HumanGenEpi/tree/main/manuals/Introduction_to_data_formats



Practical on PLINK

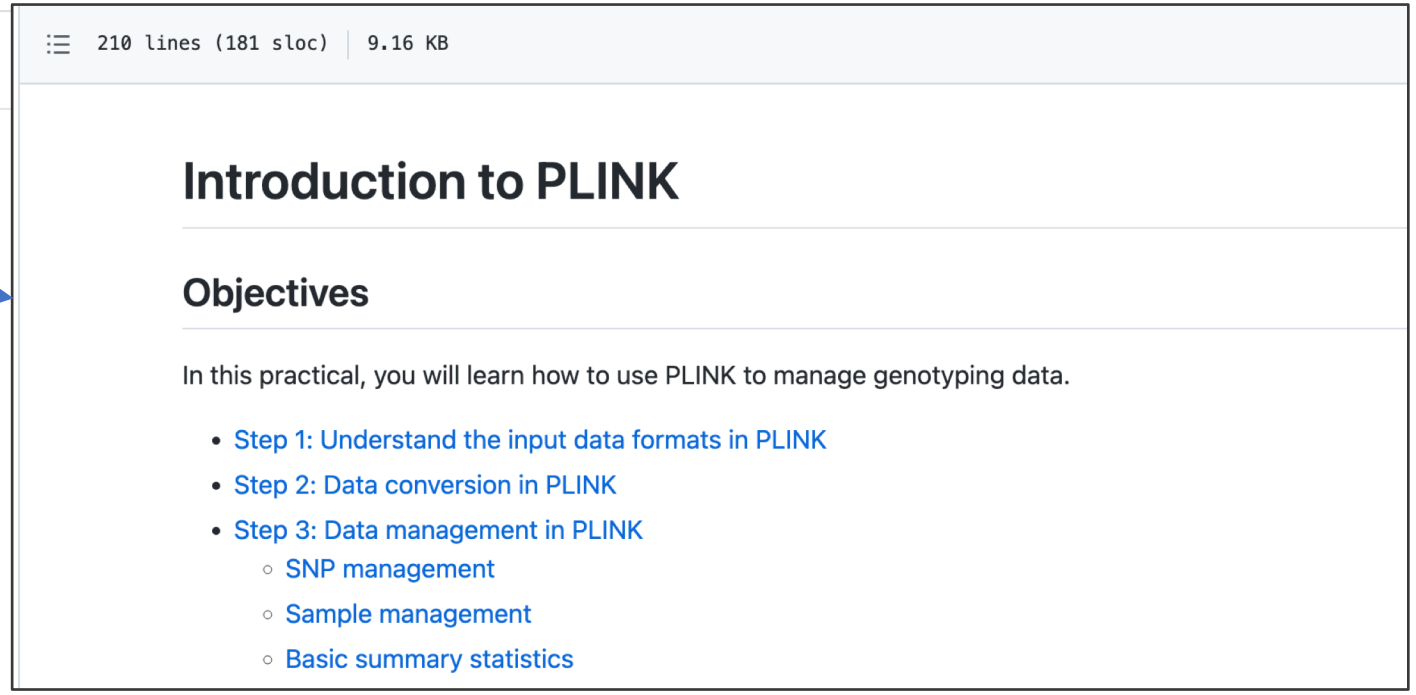
https://github.com/WCSCourses/HumanGenEpi

https://github.com/WCSCourses/HumanGenEpi/tree/main/manuals/Introduction_to_data_formats

README.md

Course manual

- Computational Resources
- Introduction to Plink
- Sample Array QC
- Variant Level Association Analysis
- Population Stratification
- Meta Analysis and Replication
- Imputation
- Independence Signals and Fine Mapping
- Gene and Pathway Association Analysis
- Polygenic Risk Scores
- Data sharing and Ethical Legal and Social Issues (ELSI)



210 lines (181 sloc) | 9.16 KB

Introduction to PLINK

Objectives

In this practical, you will learn how to use PLINK to manage genotyping data.

- Step 1: Understand the input data formats in PLINK
- Step 2: Data conversion in PLINK
- Step 3: Data management in PLINK
  - SNP management
  - Sample management
  - Basic summary statistics

Practical on PLINK

📕 **Q:**

What are the minor allele frequencies for the two SNPs, **rs693** and **exm175886**, in *APOB* across all unrelated samples?

```
plink   --bfile practical1_1 \
        --remove related.indiv \
        --snps rs693,exm175886 \
        --freq \
        --out practical1_1.unrelated.APOB_rs693_exm175886
```

```
cat practical1_1.unrelated.APOB_rs693_exm175886.frq
```

CHR	SNP	A1	A2	MAF	NCHROBS
2	rs693	A	G	0.251	5008
2	exm175886	T	C	0.003994	5008

Practical on PLINK

📕 **Q2:**

How many ***LDLR*** variant(s) has/have **MAF<0.01** among the **unrelated female** samples?

Practical on PLINK

📕 Q2:

How many ***LDLR*** variant(s) has/have **MAF<0.01** among the **unrelated female** samples?

```
awk '$5==2 { print $1,$2 }' practical1_1.fam > females.indiv
egrep LDLR LDLgenes.set > LDLR.set

plink  --bfile practical1_1 \
       --remove related.indiv \
       --keep females.indiv \
       --extract range LDLR.set \
       --freq \
       --out practical1_1.unrelated.female.LDLR
```

--filter-founders

--filter-females

--chr 19

```
awk '$5<0.01' practical1_1.unrelated.female.LDLR.frq
```

```
19        19:11205754    T    C       0.004324      2544
```

THANK YOU