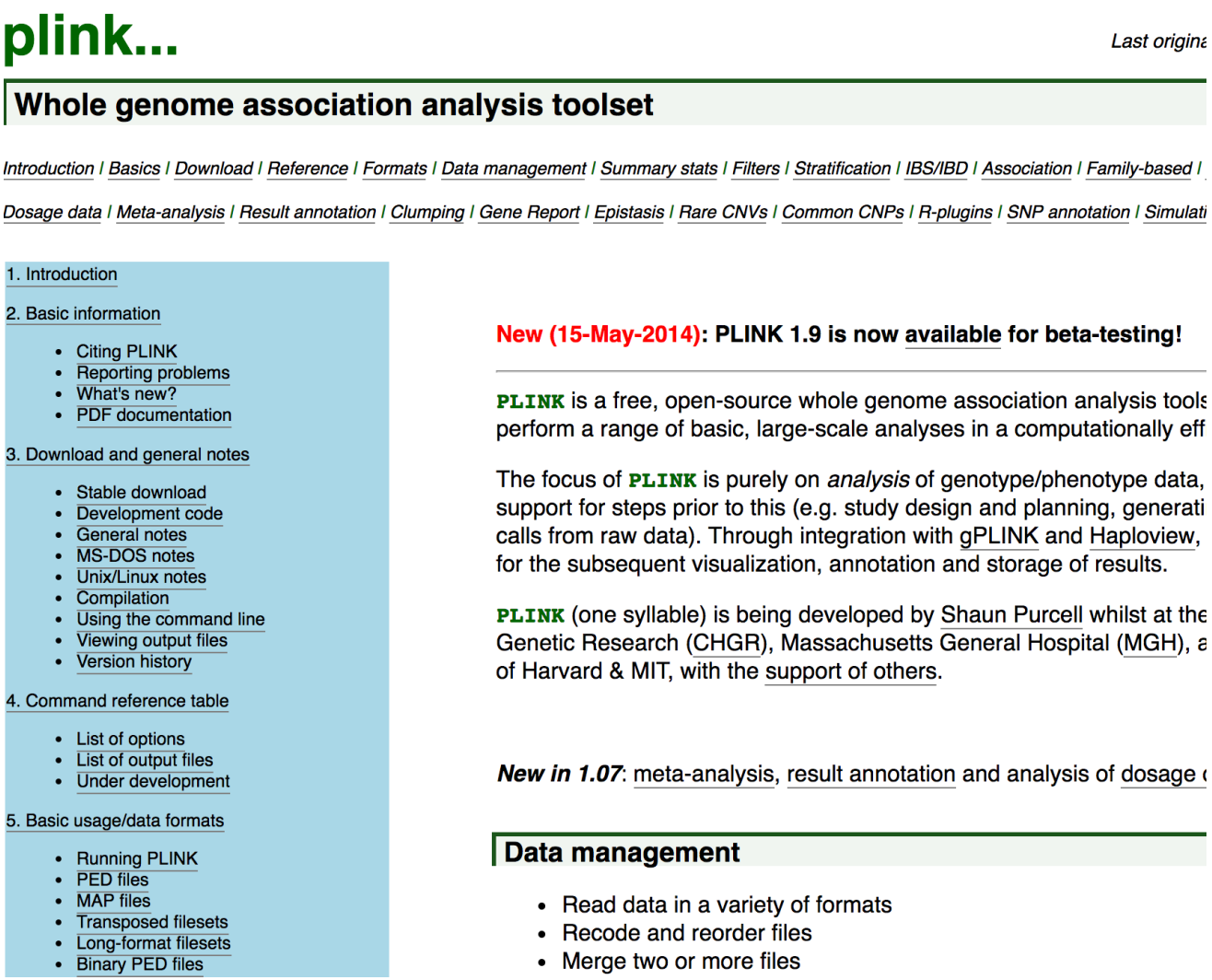
Computer Session 2: Introduction to Genetic Data and PLINK

Felix Tropf & Charlie Rahal

Things you want to do with genetic data

- Clean data/Quality control data (QC)
- Manage data (formats)
- Data description/ Summary statistics
- Conduct genetic (association) analyses
- Generate genetic related matrix
- Generate genetic principal components
-
-
-

http://zzz.bwh.harvard.edu/plink/



plink...

Last origina

Whole genome association analysis toolset

Introduction / *Basics* / *Download* / *Reference* / *Formats* / *Data management* / *Summary stats* / *Filters* / *Stratification* / *IBS/IBD* / *Association* / *Family-based* /

Dosage data / *Meta-analysis* / *Result annotation* / *Clumping* / *Gene Report* / *Epistasis* / *Rare CNVs* / *Common CNPs* / *R-plugins* / *SNP annotation* / *Simulati*

1. Introduction

2. Basic information
   - Citing PLINK
   - Reporting problems
   - What's new?
   - PDF documentation

3. Download and general notes
   - Stable download
   - Development code
   - General notes
   - MS-DOS notes
   - Unix/Linux notes
   - Compilation
   - Using the command line
   - Viewing output files
   - Version history

4. Command reference table
   - List of options
   - List of output files
   - Under development

5. Basic usage/data formats
   - Running PLINK
   - PED files
   - MAP files
   - Transposed filesets
   - Long-format filesets
   - Binary PED files

New (15-May-2014): PLINK 1.9 is now available for beta-testing!

PLINK is a free, open-source whole genome association analysis tools perform a range of basic, large-scale analyses in a computationally eff

The focus of **PLINK** is purely on *analysis* of genotype/phenotype data, support for steps prior to this (e.g. study design and planning, generati calls from raw data). Through integration with gPLINK and Haploview, for the subsequent visualization, annotation and storage of results.

PLINK (one syllable) is being developed by Shaun Purcell whilst at the Genetic Research (CHGR), Massachusetts General Hospital (MGH), a of Harvard & MIT, with the support of others.

New in 1.07: meta-analysis, result annotation and analysis of dosage

Data management

- Read data in a variety of formats
- Recode and reorder files
- Merge two or more files

Shaun Purcell, Ben Neale et al. 2007 (Citations >12K)

AJHG

Cell PRESS

Volume 81, Issue 3, September 2007, Pages 559-575

Report

PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses

Shaun Purcell[b, a], Benjamin Neale[b, c], Kathe Todd-Brown[a], Lori Thomas[a], Manuel A.R. Ferreira[a], David Bender[b, a], Julian Maller[b, a], Pamela Sklar[b, a, a], Paul I.W. de Bakker[b, a], Mark J. Daly[b, a], Pak C. Sham[d]

https://doi.org/10.1086/519795

Whole-genome association studies (WGAS) bring new computational, as well as analytic, challenges to researchers. Many existing genetic-analysis tools are not designed to handle such large data sets in a convenient manner and do not necessarily exploit the new opportunities that whole-genome data bring. To address these issues,

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

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 at Mount Sinai School of Medicine, and others. (What's new?) (Credits.) (Methods paper.)

Binary downloads

	Build		
Operating system[1]	**Stable (beta 4.4, 21 May)**	Development (31 May)	Old[2] (v1.07)
Linux 64-bit	**download**	download	download
Linux 32-bit	**download**	download	download
OS X (64-bit)	**download**	download	download
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 http://zzz.bwh.harvard.edu/plink/download.shtml.

Source code, compilation instructions, and the like are on the developer page.

The following documented PLINK 1.07 flags are not supported by 1.90 beta 4:

- --qual-geno-scores[3]
- --segment[4]

What we do now

1) Start Plink

2) (Manage) genetic data formats

3) Manage genetic data

4) Create summary statistics (descriptives)

Check

Do you have on your VM a folder containing:

1) plink
2) plink_mac
3) HapMap1.ped
4) HapMap1.map
5) rm_ind.txt
6) extract_snp.txt

Can you navigate to the folder with the command line? (cd ../; ls)

1) Do the Plink

1) Do the Plink

Type:

./plink

This is plink

```
[nuff1148@login11(arcus-b) 2pratise]$ ./plink
PLINK v1.90b4.4 64-bit (21 May 2017)           www.cog-genomics.org/plink/1.9/
(C) 2005-2017 Shaun Purcell, Christopher Chang   GNU General Public License v3

  plink [input flag(s)...] {command flag(s)...} {other flag(s)...}
  plink --help {flag name(s)...}

Commands include --make-bed, --recode, --flip-scan, --merge-list,
--write-snplist, --list-duplicate-vars, --freqx, --missing, --test-mishap,
--hardy, --mendel, --ibc, --impute-sex, --indep-pairphase, --r2, --show-tags,
--blocks, --distance, --genome, --homozyg, --make-rel, --make-grm-gz,
--rel-cutoff, --cluster, --pca, --neighbour, --ibs-test, --regress-distance,
--model, --bd, --gxe, --logistic, --dosage, --lasso, --test-missing,
--make-perm-pheno, --tdt, --qfam, --annotate, --clump, --gene-report,
--meta-analysis, --epistasis, --fast-epistasis, and --score.

'plink --help | more' describes all functions (warning: long).
[nuff1148@login11(arcus-b) 2pratise]$
```

1) Do the Plink

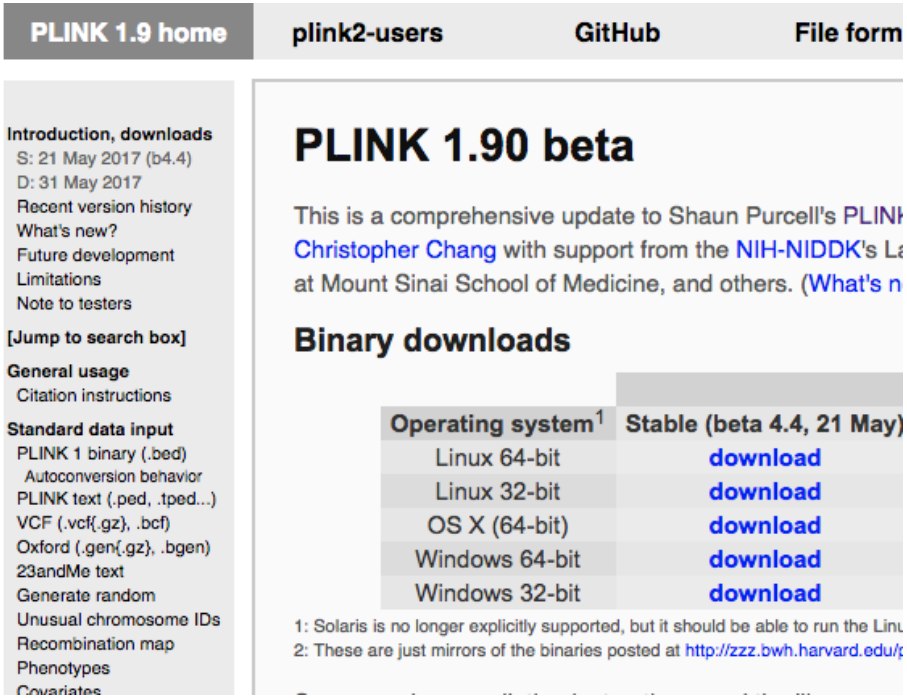
Typical plink command

```
./plink \
--file filename \
.                \
.                \
--out outputname
```

./plink calls the software;
-- is a 'flag';
followed by an option;
Linked to an argument

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

1. Download
2. Open
3. Look at it
4. delete



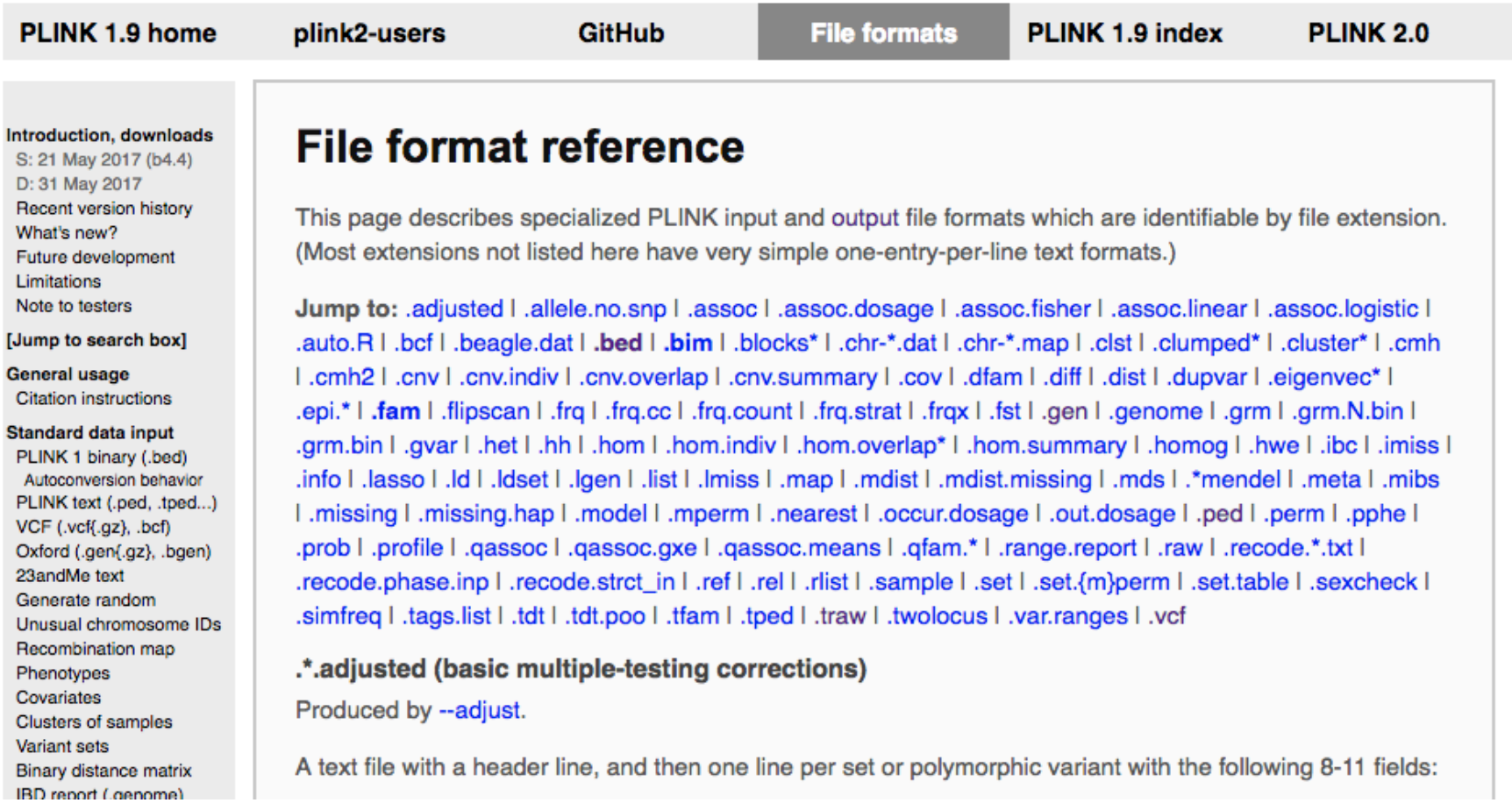
2) Genetic data

2) Genetic data

	SNP 1	SNP 2	$\cdots$	SNP 1,000,000
P1	0	1	$\cdots$	2
P2	1	0	$\cdots$	0
P3	1	2	$\cdots$	1
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
P1000	2	1	$\cdots$	2

$1,000 \times 1,000,000$ matrix; each cell $\in \{0, 1, 2\}$.

Plink knows it all



PLINK 1.9 home | plink2-users | GitHub | File formats | PLINK 1.9 index | PLINK 2.0

Introduction, downloads
S: 21 May 2017 (b4.4)
D: 31 May 2017
Recent version history
What's new?
Future development
Limitations
Note to testers

[Jump to search box]

General usage
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
Phenotypes
Covariates
Clusters of samples
Variant sets
Binary distance matrix
IBD report (.genome)

File format reference

This page describes specialized PLINK input and output file formats which are identifiable by file extension. (Most extensions not listed here have very simple one-entry-per-line text formats.)

Jump to: .adjusted | .allele.no.snp | .assoc | .assoc.dosage | .assoc.fisher | .assoc.linear | .assoc.logistic | .auto.R | .bcf | .beagle.dat | **.bed** | **.bim** | .blocks* | .chr-*.dat | .chr-*.map | .clst | .clumped* | .cluster* | .cmh | .cmh2 | .cnv | .cnv.indiv | .cnv.overlap | .cnv.summary | .cov | .dfam | .diff | .dist | .dupvar | .eigenvec* | .epi.* | **.fam** | .flipscan | .frq | .frq.cc | .frq.count | .frq.strat | .frqx | .fst | .gen | .genome | .grm | .grm.N.bin | .grm.bin | .gvar | .het | .hh | .hom | .hom.indiv | .hom.overlap* | .hom.summary | .homog | .hwe | .ibc | .imiss | .info | .lasso | .ld | .ldset | .lgen | .list | .lmiss | .map | .mdist | .mdist.missing | .mds | .*mendel | .meta | .mibs | .missing | .missing.hap | .model | .mperm | .nearest | .occur.dosage | .out.dosage | .ped | .perm | .pphe | .prob | .profile | .qassoc | .qassoc.gxe | .qassoc.means | .qfam.* | .range.report | .raw | .recode.*.txt | .recode.phase.inp | .recode.strct_in | .ref | .rel | .rlist | .sample | .set | .set.{m}perm | .set.table | .sexcheck | .simfreq | .tags.list | .tdt | .tdt.poo | .tfam | .tped | .traw | .twolocus | .var.ranges | .vcf

.*.adjusted (basic multiple-testing corrections)

Produced by --adjust.

A text file with a header line, and then one line per set or polymorphic variant with the following 8-11 fields:

Genetic data formats

a) Plink .ped and .map files
b) Plink .bed, .bim and .fam files
c) .vcf files (1000Genome)

a) Plink .ped and .map files

Type:

```
head hapmap1.ped
```

Check your files

Type:

less –S hapmap1.ped

Type:

q

Two files: PED/MAP

- Family ID
- Individual ID
- Paternal ID
- Maternal ID
- Sex (1=male; 2=female; other=unknown)
- Phenotype
- Alleles (coded AGCT or 1, 2, 3, 4)

Two files: PED/MAP

- chromosome (1-22, X, Y or 0 if unplaced)
- rs# or snp identifier
- Genetic distance (morgan)
- Base-pair position (bp units)

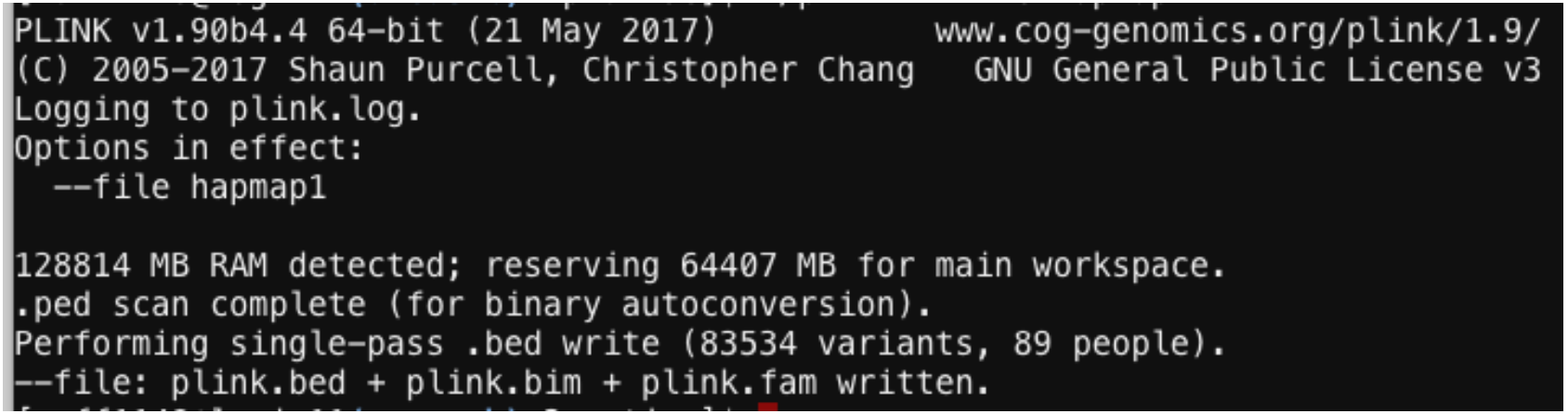
```
1 rs6681049 0 1
1 rs4074137 0 2
1 rs7540009 0 3
1 rs1891905 0 4
1 rs9729550 0 5
1 rs3813196 0 6
1 rs6704013 0 7
1 rs307347 0 8
1 rs9439440 0 9
1 rs3128342 0 10
```

Call the genotype files

Type:

./plink --file hapmap1

How many people and how many genetic marker are in the data?


PLINK v1.90b4.4 64-bit (21 May 2017) www.cog-genomics.org/plink/1.9/
(C) 2005-2017 Shaun Purcell, Christopher Chang GNU General Public License v3
Logging to plink.log.
Options in effect:
--file hapmap1
128814 MB RAM detected; reserving 64407 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (83534 variants, 89 people).
--file: plink.bed + plink.bim + plink.fam written.

b) Plink .bed, .bim and .fam files

Change data format to binary

Type:

```
./plink --file hapmap1 \
--make-bed             \
--out test
```

Check the log

Type:

ls

What do you see?

Type:

less test.log

So what does it say?

```
PLINK v1.90b4.4 64-bit (21 May 2017)
Options in effect:
  --file hapmap1
  --make-bed
  --out test

Hostname: login11
Working directory: /panfs/pan01/vol037/data/sfos-reprogene/gwas/Felix/SummerSchool/2pratise
Start time: Wed Jun 21 11:34:58 2017

Random number seed: 1498041298
128814 MB RAM detected; reserving 64407 MB for main workspace.
Scanning .ped file... done.
Performing single-pass .bed write (83534 variants, 89 people).
--file: test-temporary.bed + test-temporary.bim + test-temporary.fam written.
83534 variants loaded from .bim file.
89 people (89 males, 0 females) loaded from .fam.
89 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 89 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.99441.
83534 variants and 89 people pass filters and QC.
Among remaining phenotypes, 44 are cases and 45 are controls.
--make-bed to test.bed + test.bim + test.fam ... done.

End time: Wed Jun 21 11:34:58 2017
```

Check the log

Type:

ls

We generated files are:

- test.bed (binary file, genotype information)
- test.fam (first six columns of mydata.ped)
- test.bim (extended MAP file)

Check the .bim file

Type:

head test.bim

- chromosome (1-22, X, Y or 0 if unplaced)
- rs# or snp identifier
- Genetic distance (morgan)
- Base-pair position (bp units)
- Allele 1
- Allele 2

```
[nuff1148@login11(arcus-b) 2pratise]$ head test.bim
1       rs6681049       0       1       1       2
1       rs4074137       0       2       1       2
1       rs7540009       0       3       0       2
1       rs1891905       0       4       1       2
1       rs9729550       0       5       1       2
1       rs3813196       0       6       1       2
1       rs6704013       0       7       0       2
1       rs307347        0       8       0       2
1       rs9439440       0       9       0       2
1       rs3128342       0       10      1       2
```

Genetic data

	SNP 1	SNP 2	$\cdots$	SNP 1,000,000
P1	0	1	$\cdots$	2
P2	1	0	$\cdots$	0
P3	1	2	$\cdots$	1
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
P1000	2	1	$\cdots$	2

$1,000 \times 1,000,000$ matrix; each cell $\in \{0, 1, 2\}$.

c) .vcf files (1000G)

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HCB181_1		HCB182_1		HCB183_1		HCB18
1	1	rs6681049		2	1	.	.	PR	GT	0/0	0/0	0/0	0/0	0/0	0/0
1	2	rs4074137		2	1	.	.	PR	GT	0/0	0/1	0/1	0/1	0/1	0/0
1	3	rs7540009		2	.	.	.	PR	GT	0/0	0/0	0/0	0/0	0/0	0/0
1	4	rs1891905		2	1	.	.	PR	GT	0/1	0/1	0/1	1/1	0/0	1/1
1	5	rs9729550		2	1	.	.	PR	GT	0/0	0/1	1/1	0/0	0/0	0/0
1	6	rs3813196		2	1	.	.	PR	GT	0/0	0/0	0/0	0/0	0/0	0/0
1	7	rs6704013		2	.	.	.	PR	GT	0/0	0/0	0/0	0/0	0/0	0/0
1	8	rs307347		2	.	.	.	PR	GT	./.	./.	./.	./.	0/0	0/0
1	9	rs9439440		2	.	.	.	PR	GT	0/0	0/0	0/0	0/0	0/0	0/0
1	10	rs3128342		2	1	.	.	PR	GT	0/0	0/0	0/0	0/0	0/0	0/0
1	11	rs12044597		2	1	.	.	PR	GT	1/1	0/0	0/1	1/1	1/1	0/0
1	12	rs10907185		2	1	.	.	PR	GT	0/0	1/1	0/1	0/0	0/0	1/1

2) Manage genetic data

3) Manage genetic data

a) Interested in subgroup? Remove individuals

b) Interested in specific genetic variants? Extract a SNP

a) Remove individuals

Type:

head test.fam

Type:

head rm_ind.txt

```
HCB181 1 0 0 1 1
HCB182 1 0 0 1 1
HCB183 1 0 0 1 2
HCB184 1 0 0 1 1
HCB185 1 0 0 1 1
HCB186 1 0 0 1 1
HCB187 1 0 0 1 1
HCB188 1 0 0 1 1
HCB189 1 0 0 1 1
HCB190 1 0 0 1 1
```

```
HCB181 1
HCB182 1
```

Syntax for removing individuals

Type:

```
./plink --bfile test        \
--remove rm_ind.txt         \
--make-bed                  \
--out test_rm_ind
```

Check the log

Type:

ls

Do you see the files?

Type:

less test_rm_ind.log

How many individuals have been removed?

What happened?

```
PLINK v1.90b4.4 64-bit (21 May 2017)
Options in effect:
  --bfile test
  --make-bed
  --out test_rm_ind
  --remove rm_ind.txt

Hostname: login11
Working directory: /panfs/pan01/vol037/data/sfos-reprogene/gwas/Felix/SummerSchool/2pratise
Start time: Wed Jun 21 12:04:55 2017

Random number seed: 1498043095
128814 MB RAM detected; reserving 64407 MB for main workspace.
83534 variants loaded from .bim file.
89 people (89 males, 0 females) loaded from .fam.
89 phenotype values loaded from .fam.
--remove: 87 people remaining.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 87 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate in remaining samples is 0.994533.
83534 variants and 87 people pass filters and QC.
Among remaining phenotypes, 44 are cases and 43 are controls.
--make-bed to test_rm_ind.bed + test_rm_ind.bim + test_rm_ind.fam ... done.

End time: Wed Jun 21 12:04:56 2017
```

b) Extract SNP(s)

Type:

head test.bim

Type:

head extract_snp.txt

```
1       rs6681049       0       1       1       2
1       rs4074137       0       2       1       2
1       rs7540009       0       3       0       2
1       rs1891905       0       4       1       2
1       rs9729550       0       5       1       2
1       rs3813196       0       6       1       2
1       rs6704013       0       7       0       2
1       rs307347        0       8       0       2
1       rs9439440       0       9       0       2
1       rs3128342       0       10      1       2
```

```
rs6681049
```

Call the genotype files

Type:

```
./plink --file test                    \
--extract extract_snp.txt      \
--out test_extract_snp
```

Call the genotype files

Type:

```
./plink --bfile test                    \
--extract extract_snp.txt               \
--out test_extract_snp
```

Call the genotype files

Type:

```
./plink --bfile test                    \
--extract extract_snp.txt               \
--make-bed                              \
--out test_extract_snp
```

Check the log

Type:

ls

Do you see the files?

Type:

less test_extract_snp.log

Type:

head test_extract_snp.bim

```
(C) 2005-2017 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to extracted.log.
Options in effect:
  --bfile test
  --extract extract_snp.txt
  --make-bed
  --out extracted

128814 MB RAM detected; reserving 64407 MB for main workspace.
83534 variants loaded from .bim file.
89 people (89 males, 0 females) loaded from .fam.
89 phenotype values loaded from .fam.
--extract: 1 variant remaining.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 89 founders and 0 nonfounders present.
Calculating allele frequencies... done.
1 variant and 89 people pass filters and QC.
Among remaining phenotypes, 44 are cases and 45 are controls.
--make-bed to extracted.bed + extracted.bim + extracted.fam ... done.
```

3) Data description

We will create files with summary statistics of missing data

Generate a subfolder: descriptives

Do you remember?

3) Data description

Type:

mkdir descriptives

Type:

ls

3) Data description

Type:

cd descriptives

Type:

ls

3) Data description

Type:

cd ..

Type:

ls

Check missing

Type:

```
./plink --bfile test          \
--missing                     \
--out descriptives/miss
```

Check missing

Type:

ls descriptives

You find three files:

missing.imiss

missing.lmiss

missing.log

Check missing

Type:

nano descriptives/*missing.imiss*

Missing SNPs
by individual

FID	IID	MISS_PHENO	N_MISS	N_GENO	F_MISS
HCB181	1	N	671	83534	0.008033
HCB182	1	N	1156	83534	0.01384
HCB183	1	N	498	83534	0.005962
HCB184	1	N	412	83534	0.004932
HCB185	1	N	329	83534	0.003939
HCB186	1	N	1233	83534	0.01476
HCB187	1	N	258	83534	0.003089
HCB188	1	N	864	83534	0.01034
HCB189	1	N	517	83534	0.006189
HCB190	1	N	519	83534	0.006213
HCB191	1	N	303	83534	0.003627
HCB192	1	N	319	83534	0.003819
HCB193	1	N	401	83534	0.0048
HCB194	1	N	411	83534	0.00492
HCB195	1	N	667	83534	0.007985
HCB196	1	N	308	83534	0.003687
HCB197	1	N	271	83534	0.003244
HCB198	1	N	506	83534	0.006057
HCB199	1	N	300	83534	0.003591
HCB200	1	N	412	83534	0.004932
HCB201	1	N	332	83534	0.003974
HCB202	1	N	281	83534	0.003364
HCB203	1	N	700	83534	0.00838

Check missing

Type:

nano descriptives/*missing.imiss*

Missing individuals by SNP

CHR	SNP	N_MISS	N_GENO	F_MISS
1	rs6681049	0	89	0
1	rs4074137	0	89	0
1	rs7540009	0	89	0
1	rs1891905	0	89	0
1	rs9729550	0	89	0
1	rs3813196	0	89	0
1	rs6704013	2	89	0.02247
1	rs307347	12	89	0.1348
1	rs9439440	2	89	0.02247
1	rs3128342	1	89	0.01124
1	rs12044597	0	89	0
1	rs10907185	0	89	0
1	rs11260616	1	89	0.01124
1	rs745910	2	89	0.02247
1	rs2803291	12	89	0.1348
1	rs7531342	12	89	0.1348
1	rs262688	0	89	0
1	rs2460000	0	89	0
1	rs260509	0	89	0
1	rs2645091	0	89	0
1	rs2643895	0	89	0
1	rs2840529	0	89	0

Does plink make my life harder?

1) It uses many files
2) No interface
3) Usually separate files for data exclusion, covariates, phenotypes
4) Why should I use this?

Why plink is awesome

1) It is fast – especially plink 1.9
2) It works with huge data
3) It is safe: log-files, data not overwritten, explicit language
4) Its file format is largely supported
5) Several default options for (complex) genetic data description and data cleaning (e. g. hwe)
6) Excellent documentation

Thanks for your attention!

Questions?