An introduction to PLINK

Bärbel Maus
bmaus@ulg.ac.be
19/3/2013
Course “A tour in genetic epidemiology”

Motivation

- What is PLINK?
  - A software to analyse phenotype/genotype data
  - It is run from the command line
- Why should we use PLINK?
  - Perhaps the most common tool to analyse genome-wide genotyping data
  - It is free and open source
  - Designed to perform a wide range of basic, large-scale analyses in computationally efficient manner
  - Can be used on several platforms

Do I need to be afraid of PLINK?

Do I need to be afraid of PLINK?

NO!

- It is not necessary to know how to program to use PLINK
- This presentation will provide you with available documentation for PLINK
- PLINK commands have a clear and intuitive structure

How to get PLINK?

- Obtaining PLINK
  - `http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml`
- For Windows, choose MS-DOS

Download

PLINK is now available for free download. Below are links to ZIP files containing binaries compilied on various platforms as well as the C/C++ source code. Linux/Unix users should download the source code and compile (see notes below).

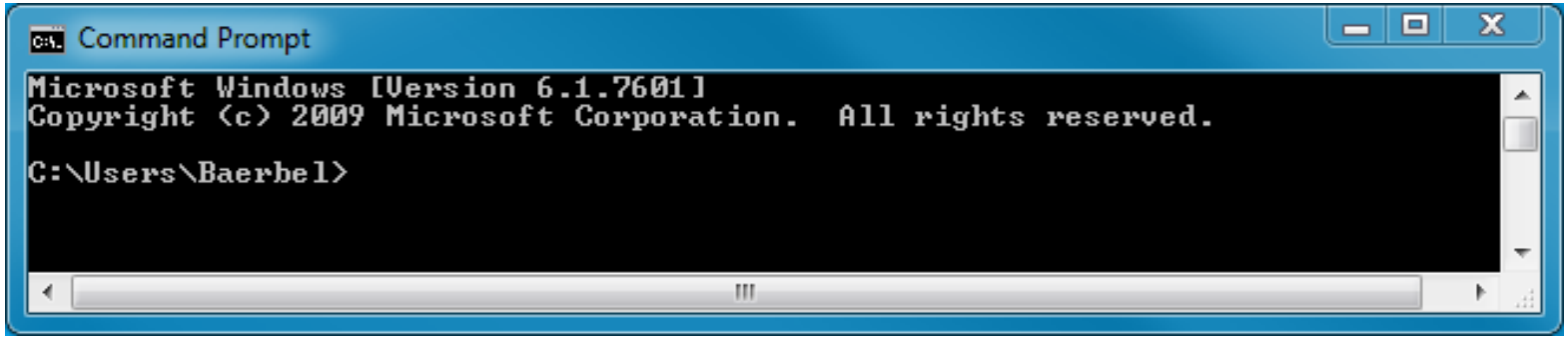
These downloads also contain a version of gPLINK, an (optional) GUI for PLINK. Please see these pages for instructions on use of gPLINK.

Remember This release is considered a *stable* release, although please remember that we cannot guarantee that it, just like most computer programs, does not contain bugs...

Platform	File	Version
Linux (x86_64)	plink-1.07-x86_64.zip	v1.07
Linux (i686)	plink-1.07-i686.zip	v1.07
MS-DOS	plink-1.07-dos.zip	v1.07 **(to be posted later today, 30-Oct)**
Apple Mac (PPC)	plink-1.07-mac.zip	v1.07 **(to be posted next week)**
Apple Mac (Intel)	plink-1.07-mac-intel.zip	v1.07
C/C++ source (.zip)	plink-1.07-src.zip	v1.07

How to get PLINK

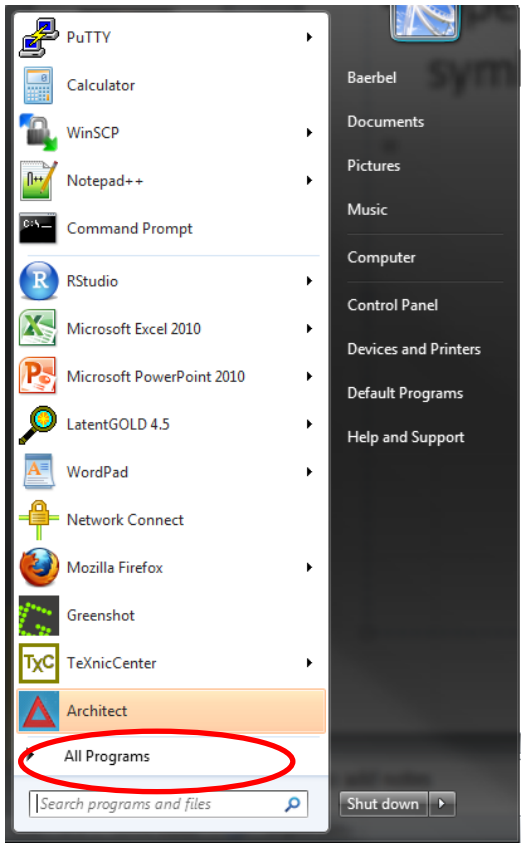
- Unzip zip file into directory, e.g., C:\Program Files\plink-1.07-dos
- You should be ready to go!
- Open command prompt

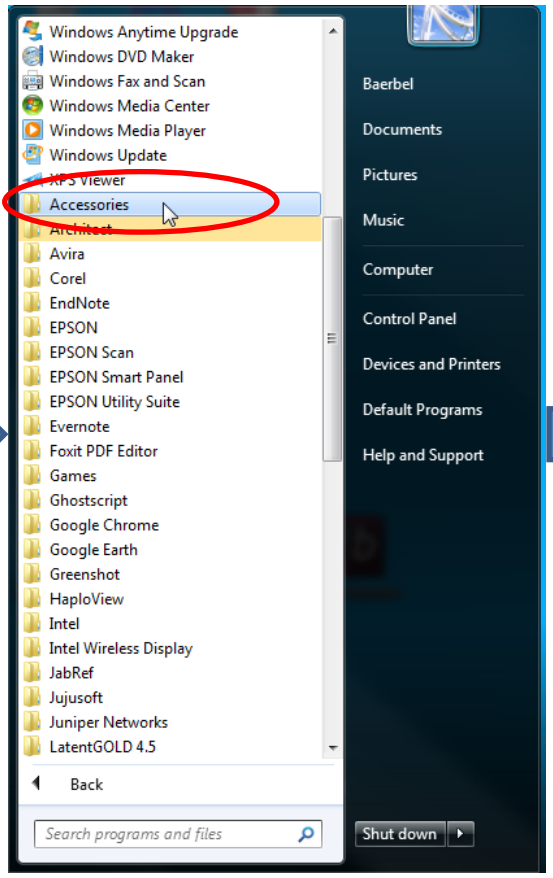


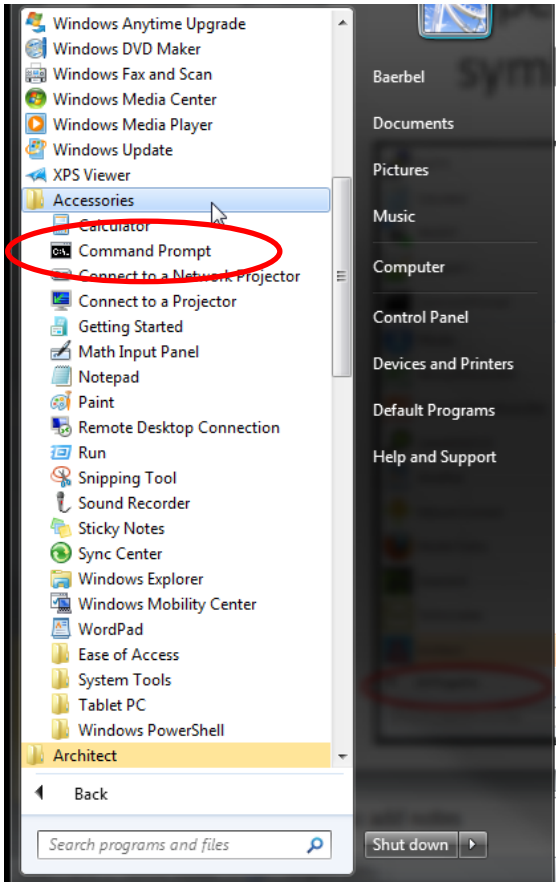
- Please mail me if any problems (bmaus@ulg.ac.be)

How to open command prompt

- Open start menu by clicking on window symbol in left corner 







Tip

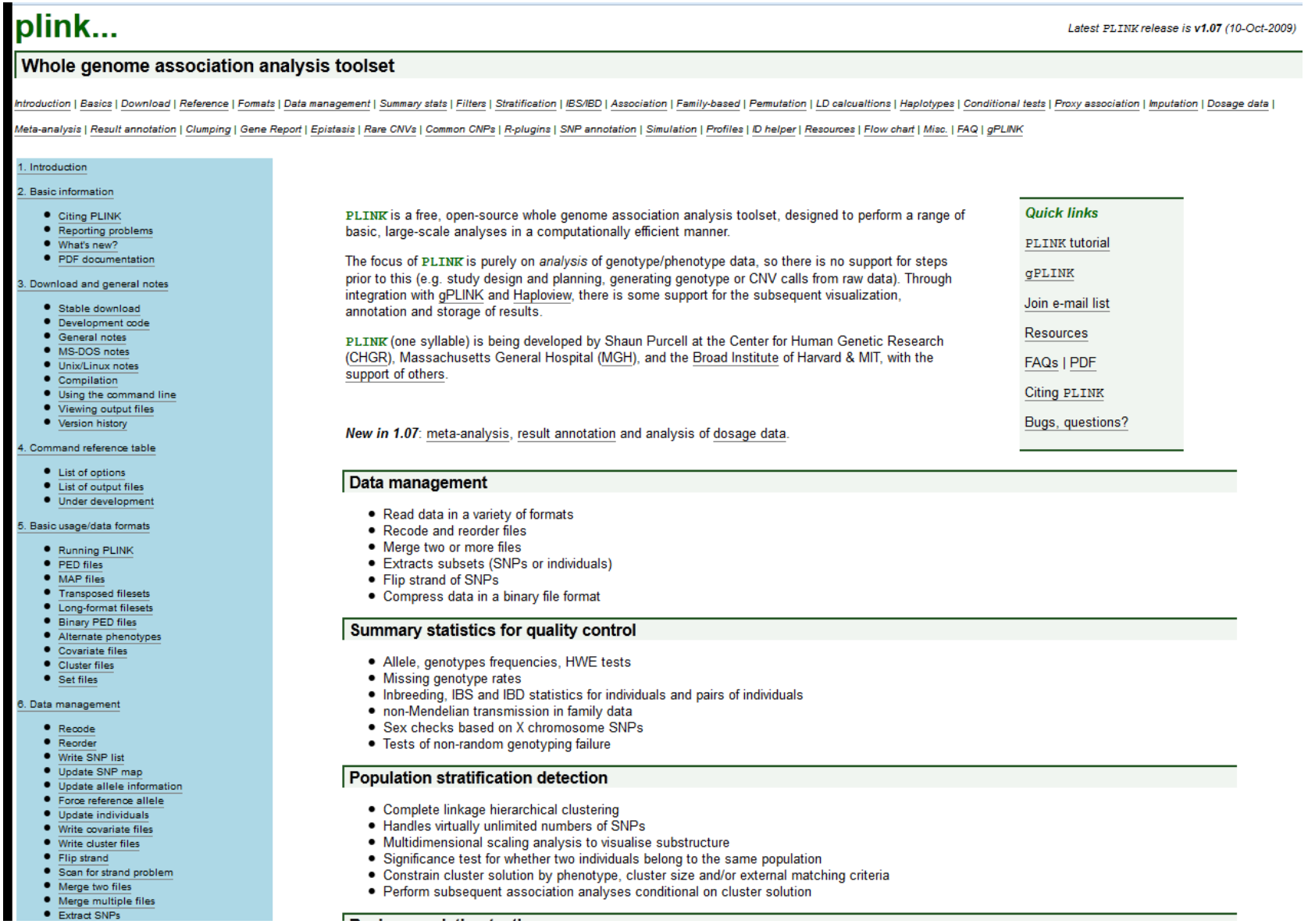
- After placing PLINK in a convenient place, put the path of PLINK's location in your environment path to make it easier to call. You can call plink then from wherever.
- This process is temporary and will only work for the current window.
- "PLINK_location" is the folder where PLINK is located.
- **Windows** in a command prompt:

```
> echo %PATH%
> path = C:\PLINK_location;%PATH%
```

Further information

- Documentation
  - PDF document
  - `http://pngu.mgh.harvard.edu/~purcell/plink/dist/plink-doc-1.07.pdf`
  - Tutorial
  - `http://pngu.mgh.harvard.edu/~purcell/plink/tutorial.shtml`

PLINK website



plink...

Latest 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 | Imputation | Dosage data | Meta-analysis | Result annotation | Clumping | Gene Report | Epistasis | Rare CNVs | Common CNPs | R-plugins | SNP annotation | Simulation | Profiles | ID helper | Resources | Flow chart | Misc. | FAQ | gPLINK

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
   - Alternate phenotypes
   - Covariate files
   - Cluster files
   - Set files
6. Data management
   - Recode
   - Reorder
   - Write SNP list
   - Update SNP map
   - Update allele information
   - Force reference allele
   - Update individuals
   - Write covariate files
   - Write cluster files
   - Flip strand
   - Scan for strand problem
   - Merge two files
   - Merge multiple files
   - Extract SNPs

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

The focus of PLINK is purely on *analysis* of genotype/phenotype data, so there is no support for steps prior to this (e.g. study design and planning, generating genotype or CNV calls from raw data). Through integration with gPLINK and Haploview, there is some support for the subsequent visualization, annotation and storage of results.

PLINK (one syllable) is being developed by Shaun Purcell at the Center for Human Genetic Research (CHGR), Massachusetts General Hospital (MGH), and the Broad Institute of Harvard & MIT, with the support of others.

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

Quick links

PLINK tutorial

gPLINK

Join e-mail list

Resources

FAQs | PDF

Citing PLINK

Bugs, questions?

Data management

- Read data in a variety of formats
- Recode and reorder files
- Merge two or more files
- Extracts subsets (SNPs or individuals)
- Flip strand of SNPs
- Compress data in a binary file format

Summary statistics for quality control

- Allele, genotypes frequencies, HWE tests
- Missing genotype rates
- Inbreeding, IBS and IBD statistics for individuals and pairs of individuals
- non-Mendelian transmission in family data
- Sex checks based on X chromosome SNPs
- Tests of non-random genotyping failure

Population stratification detection

- Complete linkage hierarchical clustering
- Handles virtually unlimited numbers of SNPs
- Multidimensional scaling analysis to visualise substructure
- Significance test for whether two individuals belong to the same population
- Constrain cluster solution by phenotype, cluster size and/or external matching criteria
- Perform subsequent association analyses conditional on cluster solution

http://pngu.mgh.harvard.edu/~purcell/plink/

PLINK files

- There are two standard file types for PLINK: ped and map files, e.g., filename.ped and filename.map
- Ped files contain information about the family, phenotype and genotype
- Map files contain information about the genetic markers

Example Map file:

Chr	Marker	cM	Position
1	rs9729550	0	11352421
1	rs6603788	0	1218086

Example:

- Genotypes are stored in: `chr1.ped`
- The markers map is described in: `chr1.map`

- At the command prompt

```
>plink --file chr1.ped
```

Output

All results are written to files with specific suffices, depending on the type of the performed operation(s).

Examples for standard suffixes for PLINK output:

Type of operation	**Suffix**
Association	plink.assoc
Logistic regression model	plink.assoc.logistic
Hardy-Weinberg test statistics	plink.hwe

… and many, many more (see documentation)

Specify root name (this replaces ‘plink’ in filenames; suffix unchanged):

```
>plink --out name
```

Plink commands

```
>plink --file filename --options
```

filename	without extension, PLINK will look for filename.ped and filename.map
options	various kind of options, see the following slides and documentation

Several options can be combined and position of options is not fixed! For example:

```
>plink --noweb --file ibdrelease5_QCI --remove related.indiv.txt
```

But the order in which commands are executed is fixed and may not correspond to the order in which they are entered!

Example analysis command

```
> plink --file filename --hardy
```

Creates text file plink.hwe (use text editor, e.g., notepad to open)

CHR	SNP	TEST	A1	A2	GENO	O(HET)	E(HET)	P
1	snp1	ALL	A	C	1/2/3	0.3333	0.4444	1
1	snp1	AFF	A	C	0/1/2	0.3333	0.2778	1
1	snp1	UNAFF	A	C	1/1/1	0.3333	0.5	1
1	snp2	ALL	G	T	1/3/2	0.5	0.4861	1
1	snp2	AFF	G	T	0/1/2	0.3333	0.2778	1
1	snp2	UNAFF	G	T	1/2/0	0.6667	0.4444	1

```
> plink --file filename --hardy --out filename2
```

Creates text file filename2.hwe

Example data management command

```
> plink --file filename --extract mysnps.txt --recode
```

Creates file plink.ped and plink.map (use text editor, e.g., notepad to open)

```
> plink --file filename --extract mysnps.txt --recode
--out filename2
```

Creates file filename2.ped and filename2.map

Working with PLINK

- Type all options on a single line
- If necessary add “\” to include more options
- Ensure exact syntax and spelling!
- Always check the logfile!
- Check if PLINK options can really be combined
- Check the order in which PLINK options are executed, see

`http://pngu.mgh.harvard.edu/~purcell/plink/flow.shtml`

Homework 2

- Form subgroups
  - One uses PLINK and one uses R
- Every group will present its homework in class. When?
- Write a report of at maximum six pages, excluding figures, tables and bibliography (single-spaced and typed)
- See Guidelines in document “Organization of homework and project_NONPH_1213.pdf”

Let me know if you have questions or problems:
Send a mail to bmaus@ulg.ac.be or visit me in Bât. 37, Chambre 1/6 during my consultation hours (Thursday 9 – 11 AM)