

R packages and other notes

Some notes

- There are MANY, MANY R packages to conduct genetic analyses (some have been listed in class slides). There is very likely a package that does what you want to do.
- Take time to try and understand the overall structure of the code and what each line does (“uncover the black box”).
- Input data is key – make sure the format is correct!
- Spend time reading manuals, google examples, google errors.
- My most important advice is to take programming classes that focuses on programming only (doesn’t matter which language). Programming + content learning (e.g., genetics) might distract each other.
- There is no shortcut to learn programming
 - Even though you get sample code for one specific analysis/problem, chances are you won’t be able to use the same code next time.

General packages

- HardyWeinberg
 - Contains tools for exploring Hardy-Weinberg equilibrium for bi and multi-allelic genetic marker data.
 - <https://cran.r-project.org/web/packages/HardyWeinberg/index.html>
- SNPlocs.Hsapiens.dbSNP144.GRCh37
 - SNP locations and alleles for Homo sapiens extracted from NCBI dbSNP
 - <http://bioconductor.org/packages/release/data/annotation/html/SNPlocs.Hsapiens.dbSNP144.GRCh37.html>
 - <https://bioconductor.org/packages/release/data/annotation/html/SNPlocs.Hsapiens.dbSNP151.GRCh38.html>

Annotation resources

https://www.bioconductor.org/packages/release/workflows/vignettes/annotation/inst/doc/Annotation_Resources.html

Object Type	Example Package Name	Contents
TxDb	TxDb.Hsapiens.UCSC.hg19.knownGene	Transcriptome ranges for the known gene track of Homo sapiens, e.g., introns, exons, UTR regions.
OrgDb	org.Hs.eg.db	Gene-based information for Homo sapiens; useful for mapping between gene IDs, Names, Symbols, GO and KEGG identifiers, etc.
BSgenome	BSgenome.Hsapiens.UCSC.hg19	Full genome sequence for Homo sapiens.
Organism.dplyr	src_organism	Collection of multiple annotations for a common organism and genome build.
AnnotationHub	AnnotationHub	Provides a convenient interface to annotations from many different sources; objects are returned as fully parsed Bioconductor data objects or as the name of a file on disk.

GWAS

- GWAStools

- Classes for storing very large GWAS data sets and annotation, and functions for GWAS data cleaning and analysis.
- <https://www.bioconductor.org/packages/release/bioc/html/GWASTools.html>

- GENESIS

- Methodology for estimating, inferring, and accounting for population and pedigree structure in genetic analyses. Performs a Principal Components Analysis on genome-wide SNP data for the detection of population structure in a sample that may contain known or cryptic relatedness. Functions are provided to perform mixed model association testing for both quantitative and binary phenotypes.
- <https://bioconductor.org/packages/release/bioc/html/GENESIS.html>

Gene-Environment Interactions

- GxEScanR
 - Genome-wide association study (GWAS) and genome-wide by environmental interaction study (GWEIS) scans using imputed genotypes stored in the BinaryDosage format. The phenotype to be analyzed can either be a continuous or binary trait. The GWEIS scan performs multiple tests that can be used in two-step methods.
 - <https://github.com/USCbiostats/GxEScanR>

Rare variant analyses

- Rvtests

- <http://zhanxw.github.io/rvtests/>

- SKAT

- <https://cran.r-project.org/web/packages/SKAT/index.html>

- SAGE-GENIE

- <https://github.com/weizhouUMICH/SAIGE>

Mendelian Randomization

- Encodes several methods for performing Mendelian randomization analyses with summarized data. Summarized data on genetic associations with the exposure and with the outcome can be obtained from large consortia. These data can be used for obtaining causal estimates using instrumental variable methods.
- <https://cran.r-project.org/web/packages/MendelianRandomization/index.html>

The Epidemiologist R Handbook

- <https://epirhandbook.com/index.html>
- Serve as a quick R code reference manual
- Provide task-centered examples addressing common epidemiological problems
- Assist epidemiologists transitioning to R
- Be accessible in settings with low internet-connectivity via an [offline version](#)
- Basics, Data Management, Analysis, Data Visualization, Reports and dashboards, Miscellaneous: writing functions, directory interactions, version control and collaboration with Git and Github, common errors, getting help, R on network drives, data table

The Epidemiologist R Handbook

Table of contents

About this book

- 1 Editorial and technical notes
- 2 Download handbook and data

Basics

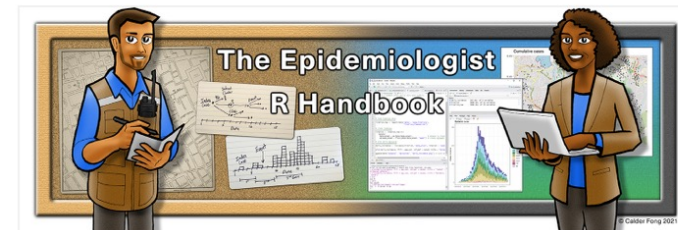
- 3 R Basics
- 4 Transition to R
- 5 Suggested packages
- 6 R projects
- 7 Import and export

Data Management

- 8 Cleaning data and core functions
- 9 Working with dates
- 10 Characters and strings
- 11 Factors
- 12 Pivoting data
- 13 Grouping data
- 14 Joining data
- 15 De-duplication
- 16 Iteration, loops, and lists

Analysis

- 17 Descriptive tables
- 18 Simple statistical tests
- 19 Univariate and multivariable regression
- 20 Missing data
- 21 Standardised rates



R for applied epidemiology and public health

This handbook strives to:

- Serve as a quick R code reference manual
- Provide task-centered examples addressing common epidemiological problems
- Assist epidemiologists transitioning to R
- Be accessible in settings with low internet-connectivity via an **offline version**



Written by epidemiologists, for epidemiologists

We are applied epis from around the world, writing in our spare time to offer this resource to the community. Your encouragement and feedback is most welcome:

- Structured **feedback form**
- Email epiRhandbook@gmail.com or tweet [@epiRhandbook](https://twitter.com/epiRhandbook)
- Submit issues to our **GitHub repository**

How to use this handbook

- Browse the pages in the Table of Contents, or use the search box
- Click the "copy" icons to copy code
- You can follow-along with the **example data**
- See the "Resources" section of each page for further material

On this page

R for applied epidemiology and public health

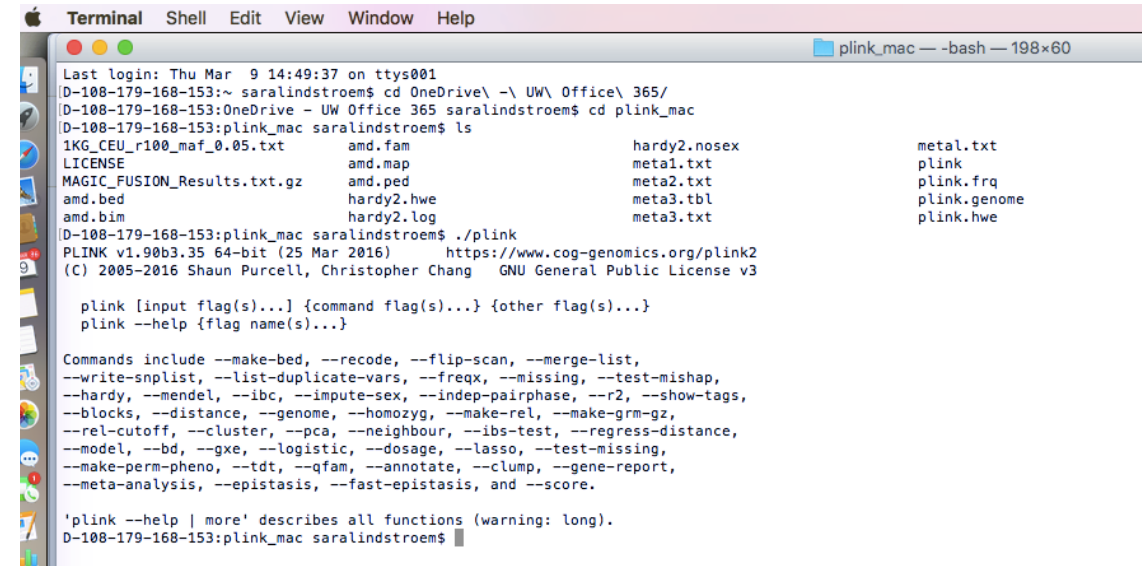
How to use this handbook

Acknowledgements

Terms of Use and Contribution

Some additional software beyond R: PLINK

- What is PLINK?
 - Statistical software for analyzing phenotype/genotype data
 - Purcell S, et al. [PLINK: a tool set for whole-genome association and population-based linkage analyses](#). Am J Hum Genet. 2007.
 - Chang CC, et al. [Second-generation PLINK: rising to the challenge of larger and richer datasets](#). Gigascience. 2015.
- Why PLINK?
 - It is arguably the most commonly used software for large-scale genetic association studies
 - It is fast
 - It is designed to conduct data quality control steps as well as generate descriptive statistics and run association analysis (just to mention a few things)
 - Many GWAS datasets are created in PLINK files (.bed, .bim, .fam)



```
Terminal Shell Edit View Window Help
Last login: Thu Mar  9 14:49:37 on ttys001
D-108-179-168-153:~ saralindstroem$ cd OneDrive\ -\ UW\ Office\ 365\
D-108-179-168-153:OneDrive - UW Office 365 saralindstroem$ cd plink_mac
D-108-179-168-153:plink_mac saralindstroem$ ls
1KG_CEU_r100_maf_0.05.txt      amd.fam                      hardy2.nosex                metal.txt
LICENSE                       amd.map                      meta1.txt                   plink
MAGIC_FUSION_Results.txt.gz  amd.ped                      meta2.txt                   plink.frq
amd.bed                       hardy2.hwe                   meta3.tbl                   plink.genome
amd.bim                       hardy2.log                   meta3.txt                   plink.hwe
D-108-179-168-153:plink_mac saralindstroem$ ./plink
PLINK v1.90b3.35 64-bit (25 Mar 2016)      https://www.cog-genomics.org/plink2
(C) 2005-2016 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-snp-list, --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).
D-108-179-168-153:plink_mac saralindstroem$
```

Risk Prediction

- LDPred
 - LDpred is a Python based software package that adjusts GWAS summary statistics for the effects of linkage disequilibrium (LD).
 - <https://github.com/bvilhjal/ldpred>
- PRSics-2
 - PRSice (pronounced 'precise') is a Polygenic Risk Score software for calculating, applying, evaluating and plotting the results of polygenic risk scores (PRS) analyses.
 - <http://www.prsice.info>

Some additional software: Large scale data

- GATK
 - Variant Discovery in High-Throughput Sequencing Data. Includes multiple tools with a primary focus on variant discovery and genotyping.
 - <https://gatk.broadinstitute.org/hc/en-us>
- Hail
 - Python library that simplifies genomic data analysis. It provides powerful, easy-to-use data science tools that can be used to interrogate even biobank-scale genomic data (e.g., UK Biobank, [gnomAD](#), TopMed, FinnGen, and Biobank Japan).
 - <https://hail.is>
- BOLT-LMM
 - The BOLT-LMM software package currently consists of two main algorithms, the BOLT-LMM algorithm for mixed model association testing, and the BOLT-REML algorithm for variance components analysis (i.e., partitioning of SNP-heritability and estimation of genetic correlations).
 - https://alkesgroup.broadinstitute.org/BOLT-LMM/BOLT-LMM_manual.html

One note about big data analyses

- Many research groups do their analysis on unix-based clusters
 - Firewalls, computational capacity, data storage
- Much of genetic analysis software is in python or perl
- Often packages come with great tutorials for analyses. The challenge will be for you to figure out how to run it on your cluster environment
- These are often specific things you will learn when you join a particular research group

Programming courses

- SISG 2021
 - <https://si.biostat.washington.edu/suminst/sisg2021/modules>
 - [Module 3: Introduction to R](#)
 - [Module 13: Association Mapping: GWAS and Sequencing Data](#)
 - [Module 16: Computational Pipeline for WGS Data](#)
- edX
 - <https://www.edx.org/learn/computer-programming>

R exercises

- <http://faculty.washington.edu/tathornt/SISG2020.html>

Time	Topic	Lecture	Exercises/Discussion
8:00am-9:20am	1. Introduction, Case Control Association Testing	Slides (Intro) [.pdf], (Lecture) [.pdf], video	Exercises [.pdf], video , R Script:[.R] , Key: [.html], [Rmd]
9:40am-11:00am	2. Association Testing with Quantitative Traits	Slides [.pdf], video	Exercises: [.pdf], video , R Script:[.R], Key: [.html], [Rmd]
11:30am-12:50pm	3. Introduction to the PLINK Software for GWAS	Slides [.pdf], video	Exercises [.pdf], video , Plink Script: [.txt], R Script:[.R], Key (Rscript) : [.R]
1:10am-2:30pm	4. Gene and Pathway Level Analysis of Genetic Association Studies.	Slides [.pdf], video	Exercises [.pdf], video , Plink and R Script: [.txt]
Tuesday, July 28th			
Time	Topic	Lecture	Exercises/Discussion
8:00am-9:20am	5. Population Structure Inference	Slides [.pdf], video	Exercises [.pdf], video , R Script: [.R], Key: [.html], [Rmd]
9:40am-11:00am	6. GWAS in Samples with Structure	Slides [.pdf], video	Exercises [.pdf], video , R Script:[.R]
11:30am-12:50pm	7. Interaction Analysis	Slides [.pdf], video	Exercises [.pdf], video , R Script:[.txt]
1:10am-2:30pm	8. Introduction to Rare Variant Analysis and Collapsing Tests	Slides [.pdf], video	Exercises [.pdf], video , Key: R Script:[.txt]
Wednesday, July 29th			
Time	Topic	Lecture	Exercises/Discussion
8:00am-9:20am	9. Rare Variant Analysis: Kernel (Variance Component) Tests and Omnibus Tests	Slides [.pdf], video	Exercises [.pdf], video , R Script:[.txt]
9:40am-11:00am	10. Power and Sample Size, Design Considerations, and Emerging Issues	Slides [.pdf], video	Exercises [.pdf], video , R Script:[.txt]

Datasets

A zipped folder with the genetic relatedness matrix (GRM) and other files for exercise 6, where a linear mixed model analysis is performed, can be \$

Note: New link to zipped file with LMM files on dropbox posted below. The previously posted LMM zipped file was corrupted.

[LMM_FILES_NEW.zip](#)

All individual data files below can be downloaded as a single zipped folder from dropbox. This file can be downloaded here:

Note: New link to zipped file with the Data files on dropbox are posted below. The previously posted zipped file was corrupted.

[SIGG2020Data_NEW.zip](#)

Alternatively, you can download each of the data files below. Before trying to read data into an R or PLINK session, we recommend looking at it first, in a text editor. Is the data comma- or tab delimited? Does it have a 'header' row containing variable names?

- [bpdata.csv](#)
- [Ht.pheno](#)
- [LHON.txt](#)
- [Population_Sample_Info.txt](#)
- [SNPlistHeight.txt](#)
- [SNPlistTransferrin.txt](#)
- [Tr.pheno](#)
- [YRI_CEU_ASW_MEX_NAM.bed](#)
- [YRI_CEU_ASW_MEX_NAM.bim](#)
- [YRI_CEU_ASW_MEX_NAM.fam](#)