

# **Session 6:**

# **Analysis of Association**

# **Studies**

---

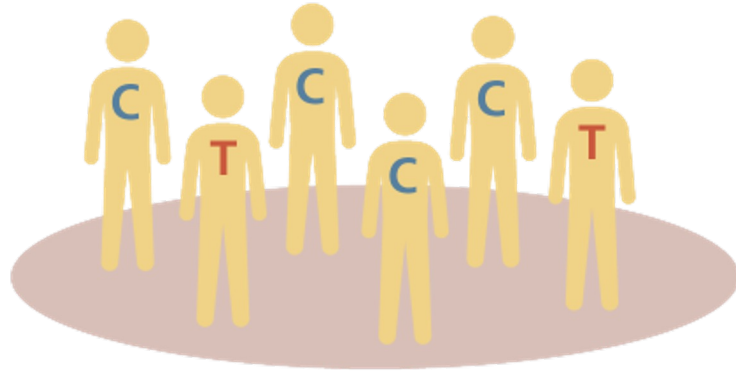


# Association studies

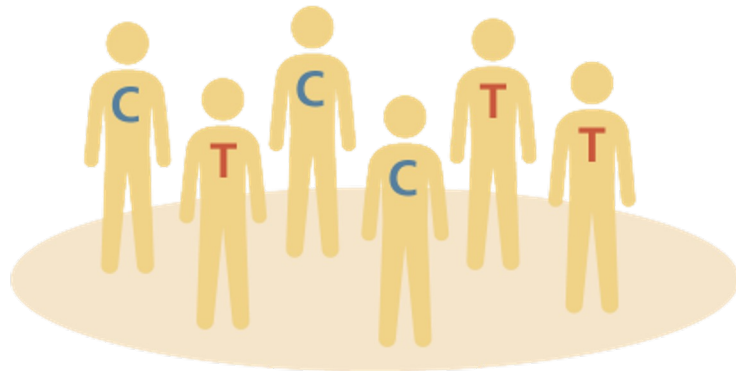
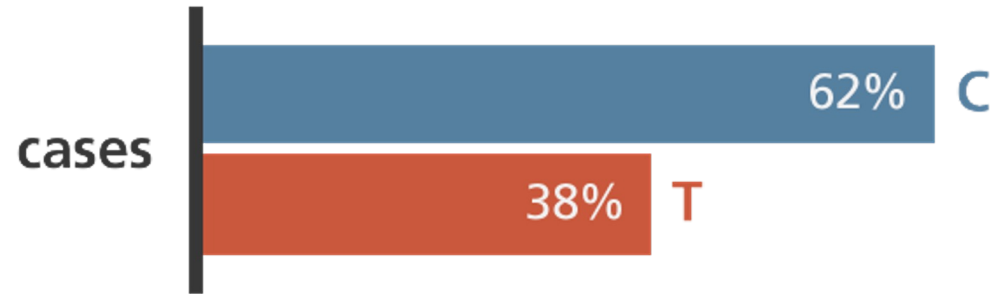
---

- > **Determine if a particular genetic variant (exposure) co-occurs with a trait more often than would be expected by chance.**
- > **Binary outcomes:**
  - Calculate the odds ratio (OR), which represents the odds that an outcome will occur given a particular exposure (effect allele), compared to the odds of the outcome occurring in the absence of that exposure (reference allele).
- > **Quantitative outcomes:**
  - Calculate the change in an outcome for every unit increase of an exposure (effect allele).

# SNP



**cases (n=1,000)**  
people with heart disease



**controls (n=1,000)**  
people without heart disease



\*Still probabilistic – for every C allele, you are MORE LIKELY to develop heart disease, but it is not guaranteed that you will, or that you won't if you have the T allele.

## “Odds”

the likelihood of something happening

## “Odds Ratio”

the likelihood of something happening in one group in relation to the likelihood of something happening in another group

# Odds ratio

---

The odds ratio is our measure of association for a case-control study. It tells us whether and how much an exposure increases the likelihood of our outcome of interest. We often look at two things:

**The estimate (and standard error)** -- the odds ratio itself. How big in the connection between an exposure and an outcome? Are those with an exposure more likely to have the outcome?

**The p-value** -- how certain are we that the odds ratio didn't just happen by chance?

# Association analysis in case-control studies (the 2x2 table)

		Disease status		
		Cases	Controls	Total
Genotype	AA/AC	a	b	a+b
	CC	c	d	c+d
Total		a+c	b+d	

# Association analysis in case-control studies

		Disease status		
		Cases	Controls	Total
Genotype	AA/AC	a	b	a+b
	CC	c	d	c+d
Total		a+c	b+d	

Calculate Odds Ratio (OR) as the odds of being a case among genotype AA/AC divided by the odds of being a case among genotype CC.

$$\frac{a/b}{c/d} = \frac{ad}{bc}$$

# Association analysis in case-control studies

		Disease status		
		Cases	Controls	Total
Genotype	AA/AC	a	b	a+b
	CC	c	d	c+d
Total		a+c	b+d	

$$OR = \frac{ad}{bc}$$

$H_0: OR = 1$  (no association)

OR > 1 indicates increased odds

OR < 1 indicates decreased odds (protective)



# Confidence intervals for odds ratios

		Disease status	
		Cases	Controls
Genotype	AA/AC	a	b
	CC	c	d

$$OR = \frac{ad}{bc}, \quad s.e(\log(OR)) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Confidence interval:  $e^{\log(OR) \pm z_{\alpha/2} \times s.e(\log(OR))}$

Lower limit of 95% confidence interval:  $e^{\log(OR) - 1.96 \times s.e}$

Upper limit of 95% confidence interval:  $e^{\log(OR) + 1.96 \times s.e}$

# Important to set your reference allele!

	Cases	Controls
CC	2	1
AA	3	3

## Assume we have an A/C SNP

- > Odds ratio when AA is reference (CC is the "exposure"):  $(2*3)/(1*3) = 2$
- > The odds of the outcome is 2x more likely among those with CC genotype compared to among those with the AA genotype.
- > Odds ratio when CC is reference (AA is the exposure):  $(3*1)/(2*3) = 0.5$
- > The odds of the outcome is ½ as likely among those with AA genotype compared to among those with the CC genotype.
- > These are the same thing! But the language matters.

# BREAKOUT ACTIVITY

You conduct a case-control study among 478 cases and 178 controls and want to calculate the odds ratio for the outcome among those homozygous for the C allele vs. those with at least one T allele. You genotype everyone and observe the following genotype counts among your cases and controls.

	Cases	Controls	Total
CC	86	20	106
TT+TC	392	158	550
Total	478	178	656

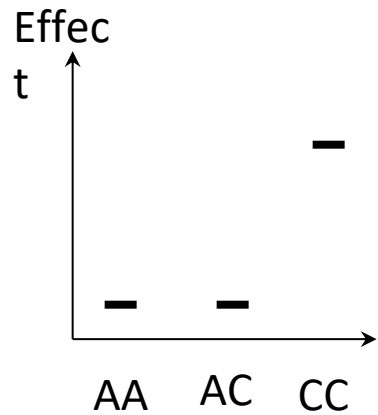
- Calculate the frequency of the CC genotype for cases and controls, respectively
- Calculate the odds ratio associated with carrying the CC genotype using the TT/TC genotypes as reference.

# Solution

	Cases	Controls	Total
CC	86	20	106
TT+TC	392	158	550
Total	478	178	656

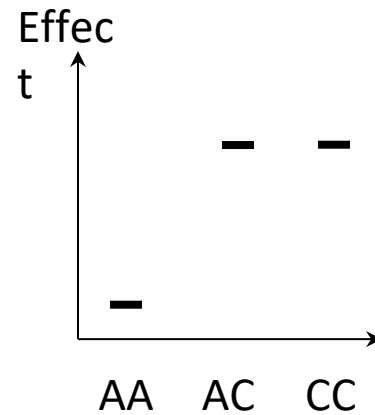
- Calculate the frequency of the CC genotype for cases and controls, respectively
  - Cases:  $86/478=0.18$ ; Controls:  $20/178=0.11$
- Calculate the odds ratio associated with carrying the CC genotype using the TT/TC genotypes as reference.
  - Odds ratio =  $(86 \times 158) / (392 \times 20) = 1.73$
  - $\ln(\text{OR}) = 0.55$
  - $\text{SE}(\ln(\text{OR})) = \sqrt{1/158 + 1/392 + 1/20 + 1/86} = 0.2655$
  - Lower limit of the CI:  $\exp(0.55 - 1.96 \times 0.2655) = 1.03$
  - Upper limit of the CI:  $\exp(0.55 + 1.96 \times 0.2655) = 2.92$
- OR (95% CI): 1.73 (1.03-2.92)

# Common models of penetrance



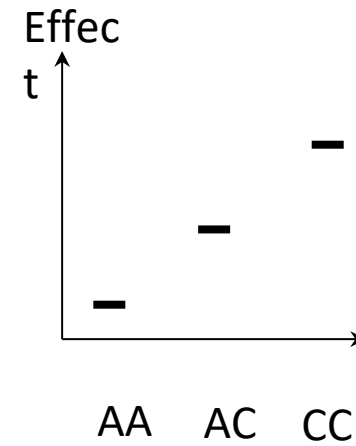
Recessive

Genotype coding: 0,0,1



Dominant

Genotype coding: 0,1,1



Additive

Genotype coding: 0,1,2

Effect = mean of continuous trait or log(OR) of binary trait

# In practice, we often use regression analyses

➤ **Allows you to adjust for relevant factors**

$$Y = \alpha + \beta_1 \mathbf{g} + \beta_2 \mathbf{x}_1 + \dots + \beta_{k+1} \mathbf{x}_k$$

- $\beta_1$  = SNP effect (for every SNP, one unit increase in outcome)
  - $g$  = genotype (often assume additive model so coded (0,1,2))
  - $X$  = additional covariates (e.g., study, age, PCs, matching factors)
- > Coefficients are estimated using maximum likelihood estimation (MLE)
- > Test  $H_0: \beta_1 = 0$  (likelihood ratio test, wald test, score test)

# Quantitative outcome genetic association

---

- ▶ Linear regression
- Additive coding of SNP (0,1,2) most common

$$Y = \alpha + \beta * SNP + X$$

- $\beta$  = SNP effect (for every SNP, unit increase in outcome)
  - We do not need to use the exponent in quantitative outcomes
- SNP = covariate coded (0,1,2)
- X = additional covariates (e.g.. study, age, PCs from population stratification)

# Always know and be purposeful of your reference

In epidemiology, the reference group always matters.

**Exposure** (gene allele reference)

**Outcome** (some outcomes have no “direction”) blue vs. green eyes

**Population** (other factors are always involved, i.e., age, cultural practices, access to care).



Additional useful software  
packages for genetic  
analyses and R coding  
resources

# Software Packages for Managing Genetic Data

---

## > PLINK

- <https://www.cog-genomics.org/plink/>
- <https://www.cog-genomics.org/plink/2.0/>

## > VCFtools

- VCFtools is a program package designed for working with VCF files, such as those generated by the 1000 Genomes Project. The aim of VCFtools is to provide easily accessible methods for working with complex genetic variation data in the form of VCF files.
- <https://vcftools.sourceforge.net/>

## > BCFtools

- Utilities for variant calling and manipulating VCFs and BCFs.
- <https://samtools.github.io/bcftools/bcftools.html>

# PLINK

---

- > Statistical software for analyzing phenotype/genotype data
  - Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559-75.
  - Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015 Feb 25;4:7.
- > It is free and open source
- > It is fast
- > It is designed to conduct data quality control steps as well as generate descriptive statistics and run association analysis (to mention a few things)
- > Many genotype datasets are delivered in PLINK format

# Where do you find PLINK?

- <http://zzz.bwh.harvard.edu//plink/>
- <https://www.cog-genomics.org/plink/1.9/>

## 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

Operating system <sup>1</sup>	Build		
	Stable (beta 7, 16 Jan)	Development (13 Feb)	Old <sup>2</sup> (v1.07)
Linux 64-bit	<a href="#">download</a>	<a href="#">download</a>	<a href="#">download</a>
Linux 32-bit	<a href="#">download</a>	<a href="#">download</a>	<a href="#">download</a>
macOS (64-bit) <sup>3</sup>	<a href="#">download</a>	<a href="#">download</a>	<a href="#">download (32-bit)</a>
Windows 64-bit	<a href="#">download</a>	<a href="#">download</a>	<a href="#">download</a>
Windows 32-bit	<a href="#">download</a>	<a href="#">download</a>	<a href="#">download</a>

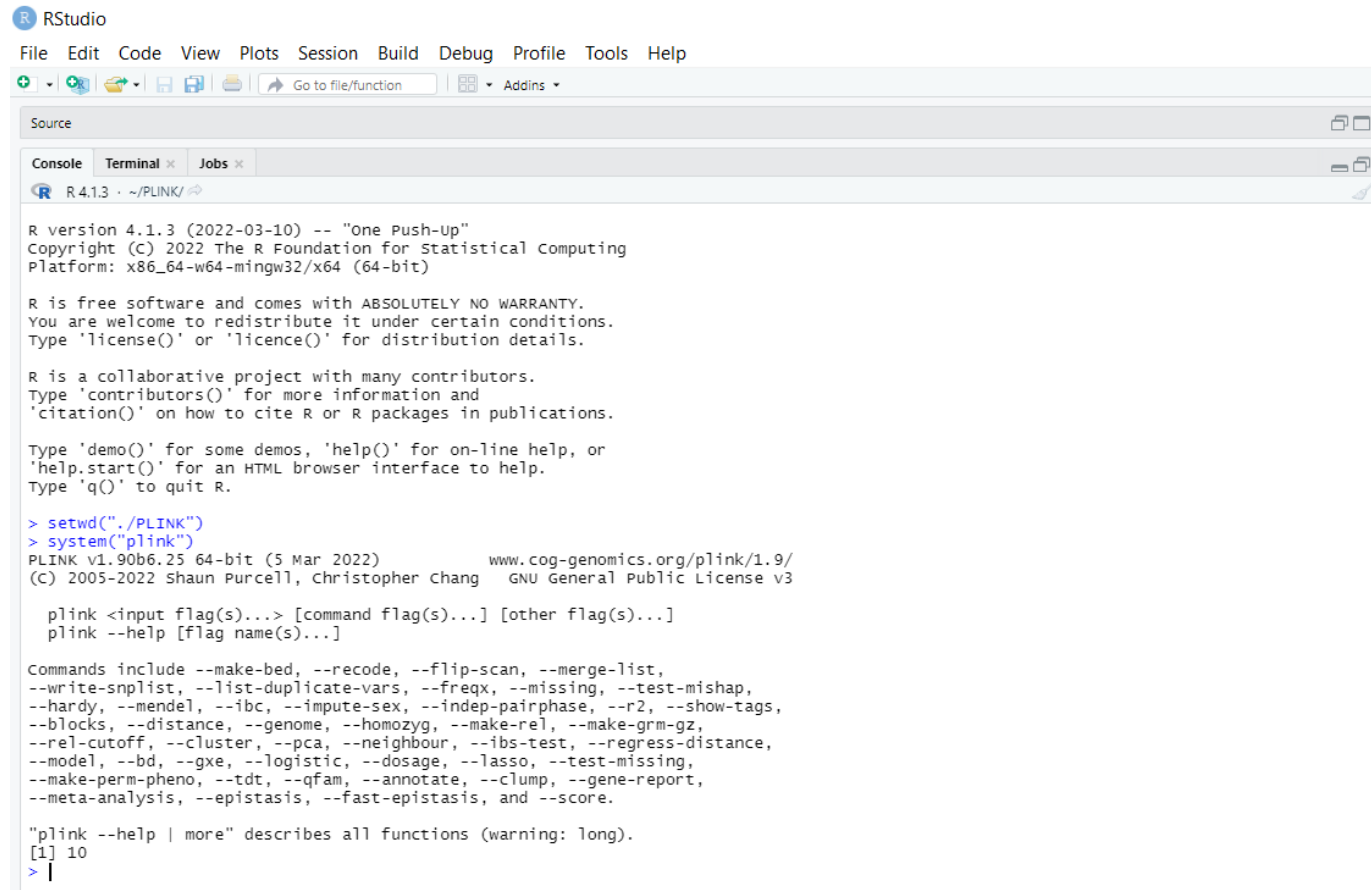
1: Solaris is no longer explicitly supported, but it should be able to run the Linux binaries.

2: These are just mirrors of the binaries posted at <https://zzz.bwh.harvard.edu/plink/download.shtml>.

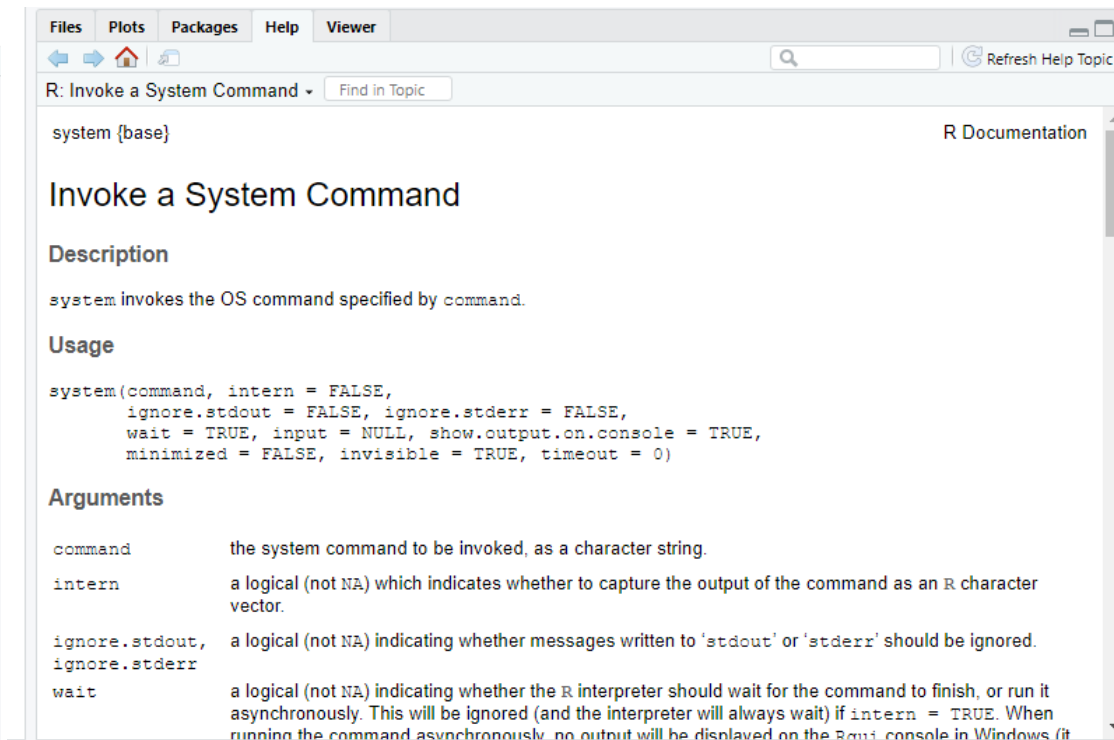
3: You need to have [Rosetta 2](#) installed to run this on M1 Macs.

# You can also run PLINK from R or R Studio console using the `system()` function

- This function takes a character string and executes it as a command in the command-line



```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Source
Console Terminal Jobs
R 4.1.3 ~./PLINK/
R version 4.1.3 (2022-03-10) -- "One Push-Up"
Copyright (c) 2022 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
> setwd("./PLINK")
> system("plink")
PLINK v1.90b6.25 64-bit (5 Mar 2022) www.cog-genomics.org/plink/1.9/
(c) 2005-2022 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).
[i] 10
> |
```



Files Plots Packages Help Viewer

R: Invoke a System Command Find in Topic Refresh Help Topic

system {base} R Documentation

## Invoke a System Command

### Description

system invokes the OS command specified by `command`.

### Usage

```
system(command, intern = FALSE,
        ignore.stdout = FALSE, ignore.stderr = FALSE,
        wait = TRUE, input = NULL, show.output.on.console = TRUE,
        minimized = FALSE, invisible = TRUE, timeout = 0)
```

### Arguments

<code>command</code>	the system command to be invoked, as a character string.
<code>intern</code>	a logical (not NA) which indicates whether to capture the output of the command as an R character vector.
<code>ignore.stdout</code> , <code>ignore.stderr</code>	a logical (not NA) indicating whether messages written to 'stdout' or 'stderr' should be ignored.
<code>wait</code>	a logical (not NA) indicating whether the R interpreter should wait for the command to finish, or run it asynchronously. This will be ignored (and the interpreter will always wait) if <code>intern = TRUE</code> . When running the command asynchronously, no output will be displayed on the R console in Windows (if

# Things that PLINK can do for you

---

- > Recode files to different formats (e.g., create input files for other programs)
- > Merge files
- > Extract subsets (SNPs and/or subjects)
- > Compress data in binary formats
- > Conduct quality control steps and filtering
- > Provide statistics on minor allele frequencies, Hardy-Weinberg Equilibrium, missing rates etc.
- > Conduct certain association analyses

# Limitations of PLINK

---

- > PLINK cannot create plots, tables, or visualizations of your results
  - We will provide instructions for installing R packages that can be used to visualize results from genetic analyses
- > More advanced analyses may require other specialized tools
  - For example:
    - > Regression models with complex parameterizations
    - > Rare variant analyses

# R Packages

---

## > **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>

## > **Hardy-Weinberg Equilibrium**

- 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>

## > **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>



# GWAS: Other Software Packages

---

## > GCTA

- GCTA (Genome-wide Complex Trait Analysis) is a software package initially developed to estimate the proportion of phenotypic variance explained by all genome-wide SNPs for a complex trait but has been greatly extended for many other analyses of data from genome-wide association studies (GWASs).
- <https://yanglab.westlake.edu.cn/software/gcta/#Overview>

## > METAL

- METAL is a tool for the meta-analysis of genome-wide association studies
- <https://github.com/statgen/METAL>

# Other Software Packages for Analyzing 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>

# R: 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

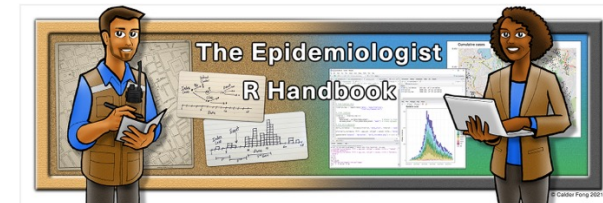
- 3 R Basics
- 4 Transition to R
- 5 Suggested packages
- 6 R projects
- 7 Import and export

#### 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](mailto: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

# A 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