Genetic data analysis using R

Presenter: Ho Kim, Ph.D

Yoonhee Kim, Aekyung Park, Ho Kim

Biostatistics & Epidemiology Department
School of Public Health
Seoul National University

Outline

- R packages
- Genetic data analysis
  - Types of packages
- Association study
  - Genetics package ? data management
  - DGC.genetics package ? allele, genotype based tests
  - Haplo.stats package ? haplotype based tests
- Summary

Planning on vacation?!

Orange County, CA

Mt. Rainier, WA

Orlando, FL

Outline

- R packages
- Genetic data analysis
  - Types of packages
- Association study
  - Genetics package
  - DGC.genetics package
  - Haplo.stats package
- Summary

R packages ?!

- R is a tool to calculate.
- R asks us to write a program for running any kind of analysis.
- Somebody already wrote scripts and combined into a package for my routinely used analysis.
- Of course, you can also write your own packages and name after your name! it¡s up to you¡

Installation of packages 1

① http://cran.r-project.org ➔ Software ➔ packages

② Find the optimal packages for your analysis.

I. Ctrl +F ➔ Enter your keyword

II. Click the name of packages

III. Read the description of packages, and see the last update day and authors.

IV. For assuring the details of package, read the reference manual providing as pdf file.

V. You can also check the referenced papers if you want to know more specific usage.

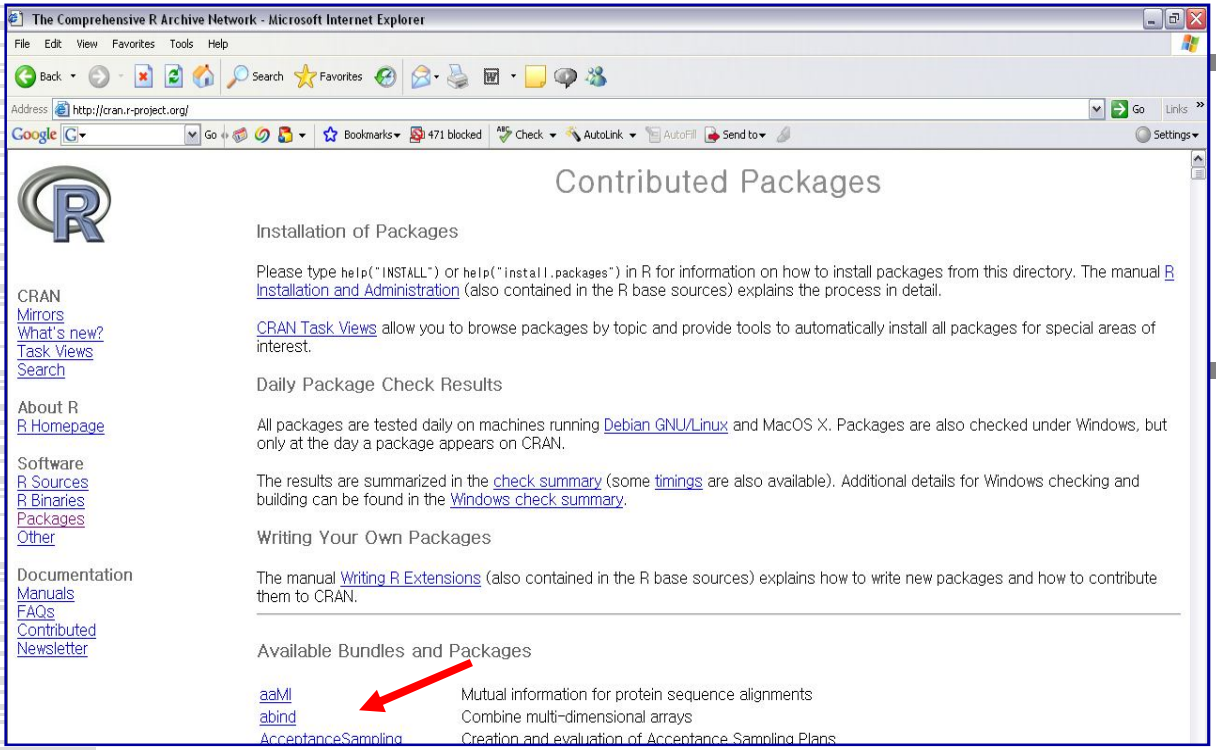
Installation of packages 2

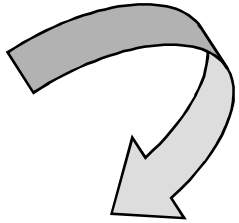
③ Download the package as *.tar.gz, *.tgz(MacOS), or*.zip(Windows) file to your directory ➔ *Start* R ➔ Packages ➔ Install Package(s) from local zip files¡

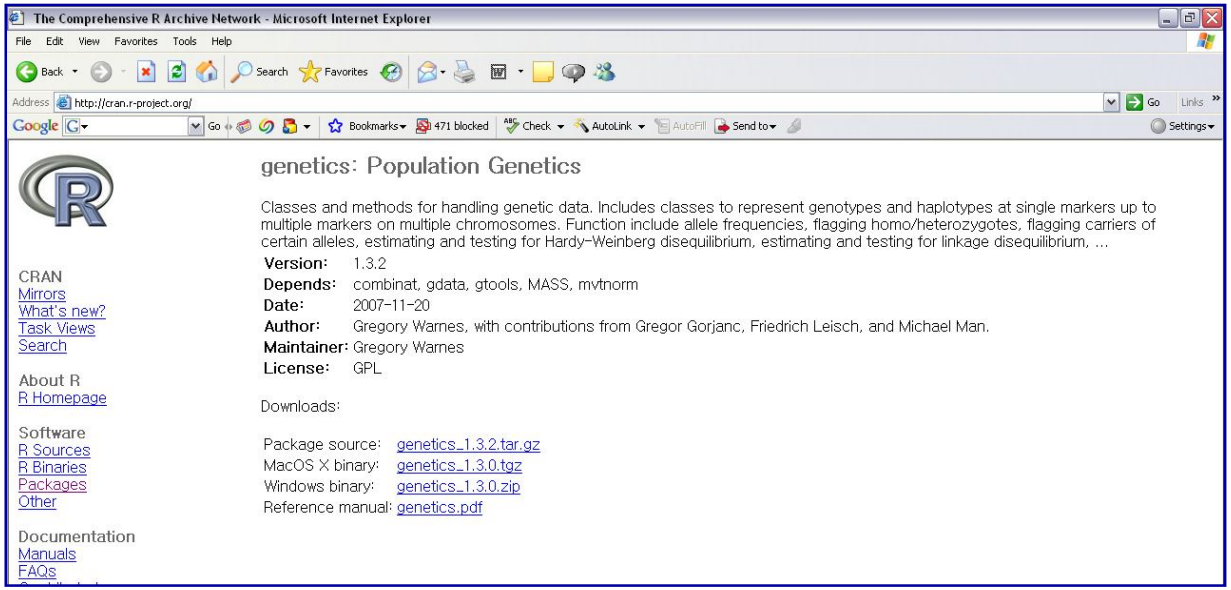
④ OR, (Need to internet connection)

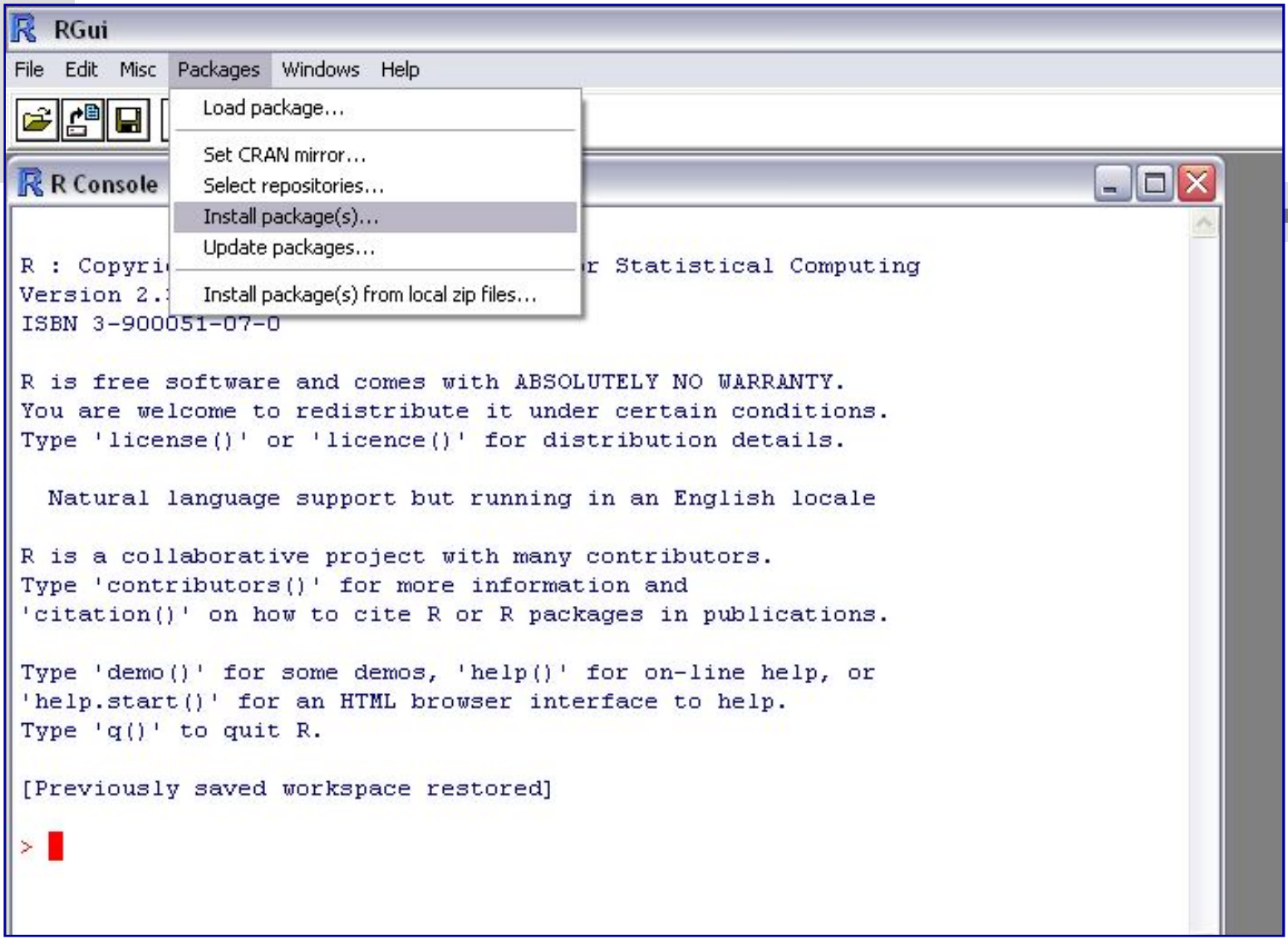
Start R ➔ Packages ➔ Install package(s)¡ ➔ *Select* CRAN mirror ➔ Install package(s)¡ ➔ *Select* package name on the list.

* *From time to time, click Packages ➔ Update packages¡ for maintaining up-to-date packages*


The Comprehensive R Archive Network - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://cran.r-project.org/
Google Go Bookmarks 471 blocked Check AutoLink AutoFill Send to Settings
Contributed Packages
Installation of Packages
Please type help("INSTALL") or help("install.packages") in R for information on how to install packages from this directory. The manual R Installation and Administration (also contained in the R base sources) explains the process in detail.
CRAN Task Views allow you to browse packages by topic and provide tools to automatically install all packages for special areas of interest.
Daily Package Check Results
All packages are tested daily on machines running Debian GNU/Linux and MacOS X. Packages are also checked under Windows, but only at the day a package appears on CRAN.
The results are summarized in the check summary (some timings are also available). Additional details for Windows checking and building can be found in the Windows check summary.
Writing Your Own Packages
The manual Writing R Extensions (also contained in the R base sources) explains how to write new packages and how to contribute them to CRAN.
Available Bundles and Packages
aaMI Mutual information for protein sequence alignments
abind Combine multi-dimensional arrays
AcceptanceSampling Creation and evaluation of Acceptance Sampling Plans
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed
Newsletter




The Comprehensive R Archive Network - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://cran.r-project.org/
Google Go Bookmarks 471 blocked Check AutoLink AutoFill Send to Settings
genetics: Population Genetics
Classes and methods for handling genetic data. Includes classes to represent genotypes and haplotypes at single markers up to multiple markers on multiple chromosomes. Function include allele frequencies, flagging homo/heterozygotes, flagging carriers of certain alleles, estimating and testing for Hardy-Weinberg disequilibrium, estimating and testing for linkage disequilibrium, ...
Version: 1.3.2
Depends: combinat, gdata, gtools, MASS, mvtnorm
Date: 2007-11-20
Author: Gregory Warnes, with contributions from Gregor Gorjanc, Friedrich Leisch, and Michael Man.
Maintainer: Gregory Warnes
License: GPL
Downloads:
Package source: genetics_1.3.2.tar.gz
MacOS X binary: genetics_1.3.0.tgz
Windows binary: genetics_1.3.0.zip
Reference manual: genetics.pdf
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs


RGui
File Edit Misc Packages Windows Help
Load package...
Set CRAN mirror...
Select repositories...
Install package(s)...
Update packages...
Install package(s) from local zip files...
R Console
R : Copyri r Statistical Computing
Version 2.
ISBN 3-900051-07-0
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.
Natural language support but running in an English locale
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.
[Previously saved workspace restored]
>

Calling packages

- *Start* R ➔ Packages ➔ Load package(s)¡ ➔ Select a package on the list in your directory
- OR, *Start* R ➔ *Type* library(PACKAGE NAME) whenever you want to use that package
- Tips !
  - Some packages are related to each other, so you need to install a required package as well for using a certain package.
  - Some packages are required a lower or higher R version.
  - ¡

Outline

- R packages
- Genetic data analysis
  - Types of packages
- Association study
  - Genetics package
  - DGC.genetics package
  - Haplo.stats package
- Summary

Genetic data analysis ?

- Statistical genetics
- Molecular genetics
- Genetic epidemiology
- Bioinformatics
- Genomics
- ¡

As of November, 22, 2007

Keyword : genetics

Package Name	Description
Adegenet	Genetic data handling for multivariate analysis
Gap	Genetic analysis package
Genetics	Population genetics
GSA	Gene set analysis
Hwde	Models and tests for departure from Hardy-Weinberg equilibrium
Ibdreg	Regression Methods for ibd linkage with covariates
Lodplot	Plot a genome scan
MasterBayes	ML and MCMC Methods for Pedigree Reconstruction and analysis
Multic	Quantitative linkage analysis tools using the variance components approach
Popgen	Statistical and population genetics
Qgen	Quantitative genetics using R
Qtl	Tools for analyzing QTL experiments
Qtlbim	QTL Bayesian interval mapping
Bqtl	Bayesian QTL mapping toolkit
SimHap	A comprehensive modeling framework to haplotypic analysis of population based data

As of November, 22, 2007

Keyword : SNP, CGH

Package Name	Description
SNP	
GenABEL	Genome-wide SNP association analysis
Haplo.stats	Haplotype analysis with traits and covariates in ambiguous linkage phase
Hapsim	Haplotype data simulation
ldDesign	Design of experiments for detections of linkage disequilibrium
Ldheatmap	Graphical display of pairwise linkage disequilibria between SNPs
mapLD	Linkage disequilibrium mapping
SNPassoc	SNPs based whole genome association studies
CGH	
ADaCGH	Analysis of data from aCGH experiments
Cgh	Microarray CGH analysis using the Smith-Waterman algorithm
cghFlasso	Detecting host spot on CGH array data with fused lasso regerssion
Clac	Clust along chromosome, a method to cell gains/losses in CGH array data

As of November, 22, 2007

Keyword : microarray, phylo

Package Name	Description
Microarray	
arrayImpute	Missing imputation for microarray data
ArrayMissPattern	Exploratory analysis of missing patterns for microarray data
Celsius	Retrieve affymetrix microarray measurements and metadata form celsuis
crpssjubDetectpr	Detection of cross-hybridization events in microarray experiments
Funcluster	Functional Profiling of microarray expression data
Limman	Linear models for microarray data
Maanova	Tools for analyzing micro array experiments
Sma	Statistical Microarray Analysis
Phylo	
ComPairWise	Compare phylogenetic or population genetic data alignments

Outline

- R packages
- Genetic data analysis
  - Types of packages
- Association study
  - Genetics package
  - DGC.genetics package
  - Haplo.stats package
- Summary

Association study

- A mechanism for detecting genetic variation that is associated with disease phenotypes across families
- To search for common alleles shared among unrelated individuals with similar phenotypes
- Statistical tests of whether or not affected individuals have distorted allele frequency
  - i.e.Chi-square test, logistic regression, machine learning methods ¡

Genome wide association study (GWAS)

- Due to high-throughput techniques, and the sequencing of the whole human genome (Human Genome Project, 1990-present), association mapping with genome wide genetic markers (100K, 500K, and even 1 million SNP chips) has been emerged.
- Several research groups have published the results of GWAS of complex diseases (Saxena, Voight et al. 2007, Scott, Mohleke et al. 2007, Sladek, Rocheleau et al. 2007) using chi-square test or logistic regression methods.

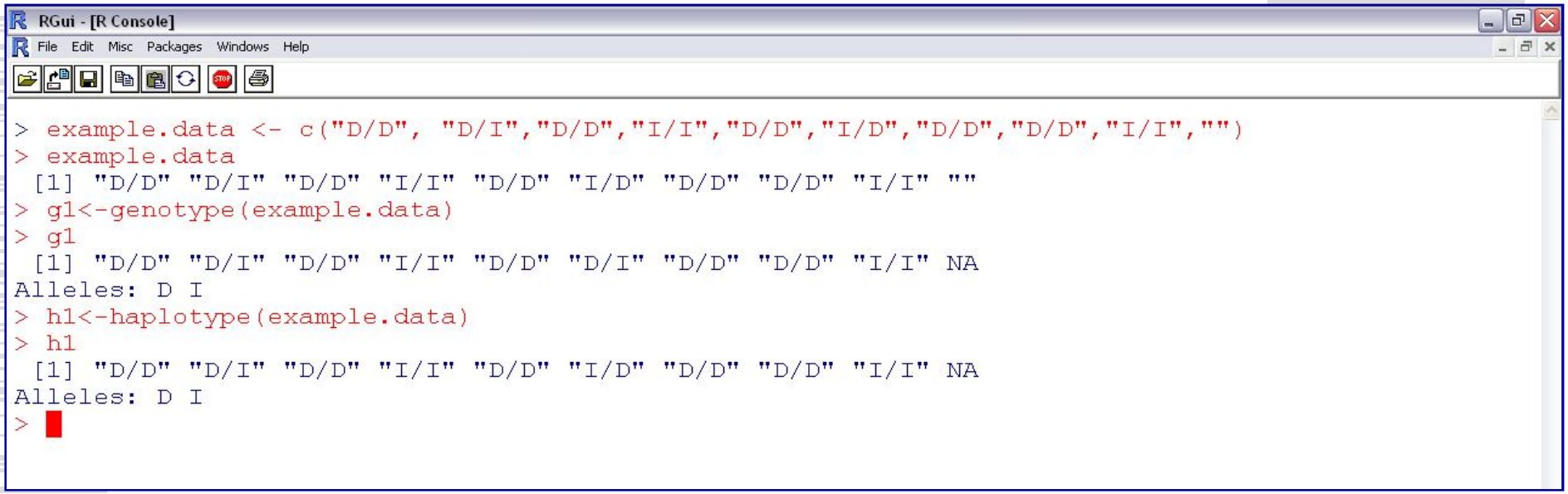
Outline

- R packages
- Genetic data analysis
  - Types of packages
- Association study
  - Genetics package ? data management
  - DGC.genetics package
  - Haplo.stats package
- Summary

The genetics Package

- Title : Population genetics
- Date : 2007-11-20
- Description : Classes and methods for handling genetic data. Includes classes to represent genotypes and haplotypes at single markers up to multiple markers on multiple chromosomes.
- Functions : allele frequencies, flagging homo/heterozygoes, flagging carriers of certain alleles, estimating and testing for Hardy-Weinberg disequilibrium, estimating and testing for linkage disequilibrium

Example
? *genotype, haplotype* ?



1) **Input observed gene or marker alleles for different individuals to ¡example.data¡**
2) **So, we have 10 individuals including 1 missing data.**
3) **Make the data into genotype format communicating in the genetics package using genotype function.**
4) **Now g1 is in a genotype format with ¡D¡ and ¡I¡ alleles as factors.**
5) **H1, using haplotype function, maintains the order information.**

Example

? *heterozygote, homozygote, carrier, allele.count*?

```
> heterozygote(g1)
 [1] FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE    NA
> homozygote(g1)
 [1]  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE    NA
> carrier(g1,"D")
 [1]  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE    NA
> allele.count(g1)
      D  I
 [1,] 2  0
 [2,] 1  1
 [3,] 2  0
 [4,] 0  2
 [5,] 2  0
 [6,] 1  1
 [7,] 2  0
 [8,] 2  0
 [9,] 0  2
[10,] NA NA
```

Example
? *summary* ?

```
>  example.data   <- c("D/D","D/I","D/D","I/I","D/D",
+                          "D/D","D/D","D/D","I/I","")
>       g1  <- genotype(example.data)
>       g1
 [1] "D/D" "D/I" "D/D" "I/I" "D/D" "D/D" "D/D" "D/D" "I/I" NA
Alleles: D I
> summary(g1)

Number of samples typed: 9 (90%)

Allele Frequency: (2 alleles)
   Count Proportion
D     13       0.72
I      5       0.28
NA     2         NA


Genotype Frequency:
    Count Proportion
D/D     6       0.67
D/I     1       0.11
I/I     2       0.22
NA      1         NA

Heterozygosity (Hu)   = 0.4248366
Poly. Inf. Content    = 0.32074

>
```

Aside to explain the output: Most of the output of this summary is self-explanatory. Two quantities that may not be familiar are the Heterozygosity(Hu) and Poly. Inf. Content. The heterozygosity is the expected proportion of heterozygotes, assuming HWE, given the observed allele frequencies or

$$1 - \sum_{i=1}^{k} p_i^2$$

where i indexes the k alleles (2 for our example diallelic locus) and p_i is the population frequency of the i^{th} allele (you should ask yourself why the above formula has the claimed interpretation). An unbiased estimate of this quantity is

$$H_u = \left(1 - \sum_{i=1}^{k} \hat{p}_i^2\right) \frac{2n}{2n-1}$$

where $\hat{p}_i$ are the estimated allele frequencies (i.e. the observed proportion in the sample) and n is the number of individuals sampled (Ott, 1992 Am J Hum Genet 51:283-290).

Poly. Inf. Content is an abbreviation of the Polymorphism Information Content or PIC (Botstein et al., 1980 Am J Hum Genet 32:314-331), another measure of informativeness of markers, with formula

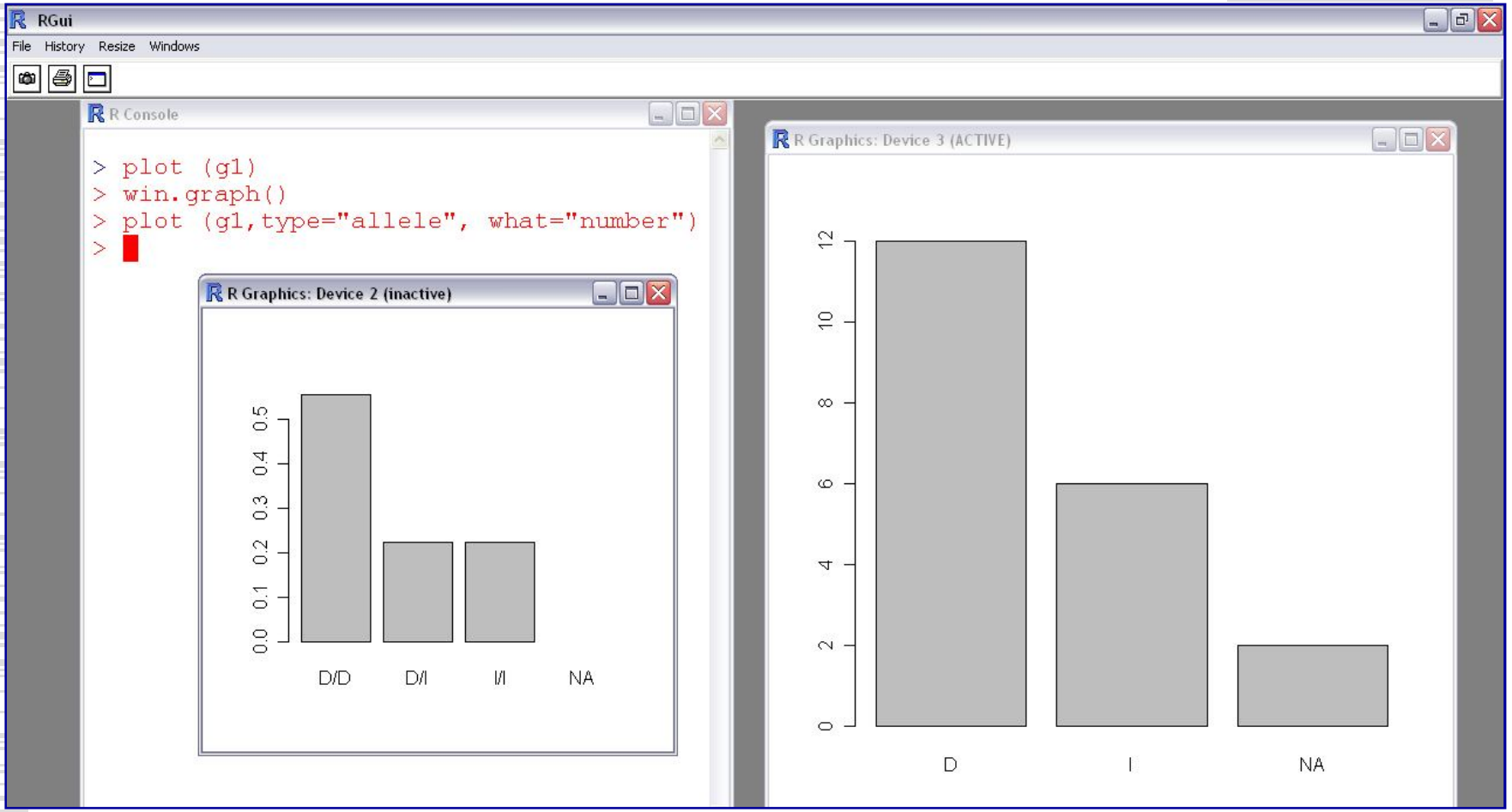
$$PIC = 1 - \sum_{i=1}^{k} p_i^2 - \sum_{i=1}^{k} \sum_{j=i+1}^{k} 2p_i^2 p_j^2$$

that can be estimated by

$$H_u - \sum_{i=1}^{k} \sum_{j=i+1}^{k} 2\hat{p}_i^2 \hat{p}_j^2$$

Example

? *plot.genotype*?



Example
? *HWE.chisq, HWE.exact, HWE.test* ?

- HWE.chisq : Perform Chi-square test for Hardy-Weinberg Equilibrium
- HWE.exact : Exact test of HWE for 2-allele markers
- HWE.test : Estimate Disequilibrium and test for HWE

Test for Independence

빈도 / 백분율 / 행 백분율	1	2	총합
Patients	10	2	12
	55.56	11.11	66.67
	83.33	16.67	
Control	2	4	6
	11.11	22.22	33.33
	33.33	66.67	
총합	12	6	18
	66.67	33.33	100.00

통계량	자유도	값	확률값
카이제곱	1	4.5000	0.0339

경고: 셀들의 75%가 5보다 작은 기대도수를 가지고 있습니다.
카이제곱 검정은 올바르지 않을 수 있습니다.

Fisher¡s Exact test Pr <= P 0.1070 n = 18

Exact Test

Table Cell				
(1,1)	(1,2)	(2,1)	(2,2)	probabilities
12	0	0	6	.0001
11	1	1	5	.0039
10	2	2	4	.0533
9	3	3	3	.2370
8	4	4	2	.4000
7	5	5	1	.2560
6	6	6	0	.0498

Table Probabilities

- One-tailed p-value:

p=0.0533+0.0039+0.0001=0.0573

- Two-tailed p-value:

p= 0.0533+0.0039+0.0001+0.0498=0.1071

Example
Hardy Weinberg Equilibrium test

	D	I	
D	6 (4.6656)	1 (3.6288)	
I	.	2 (0.7056)	
			9

9
1

p=(6*2+1)/18=0.72
q=1-0.72=0.28
P(DD)=p*p=0.5184 (*9=4.6656)
P(DI)=2pg=0.4032(*9=3.6288)
P(II)=q*q=0.0784 (*9=0.7056)

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

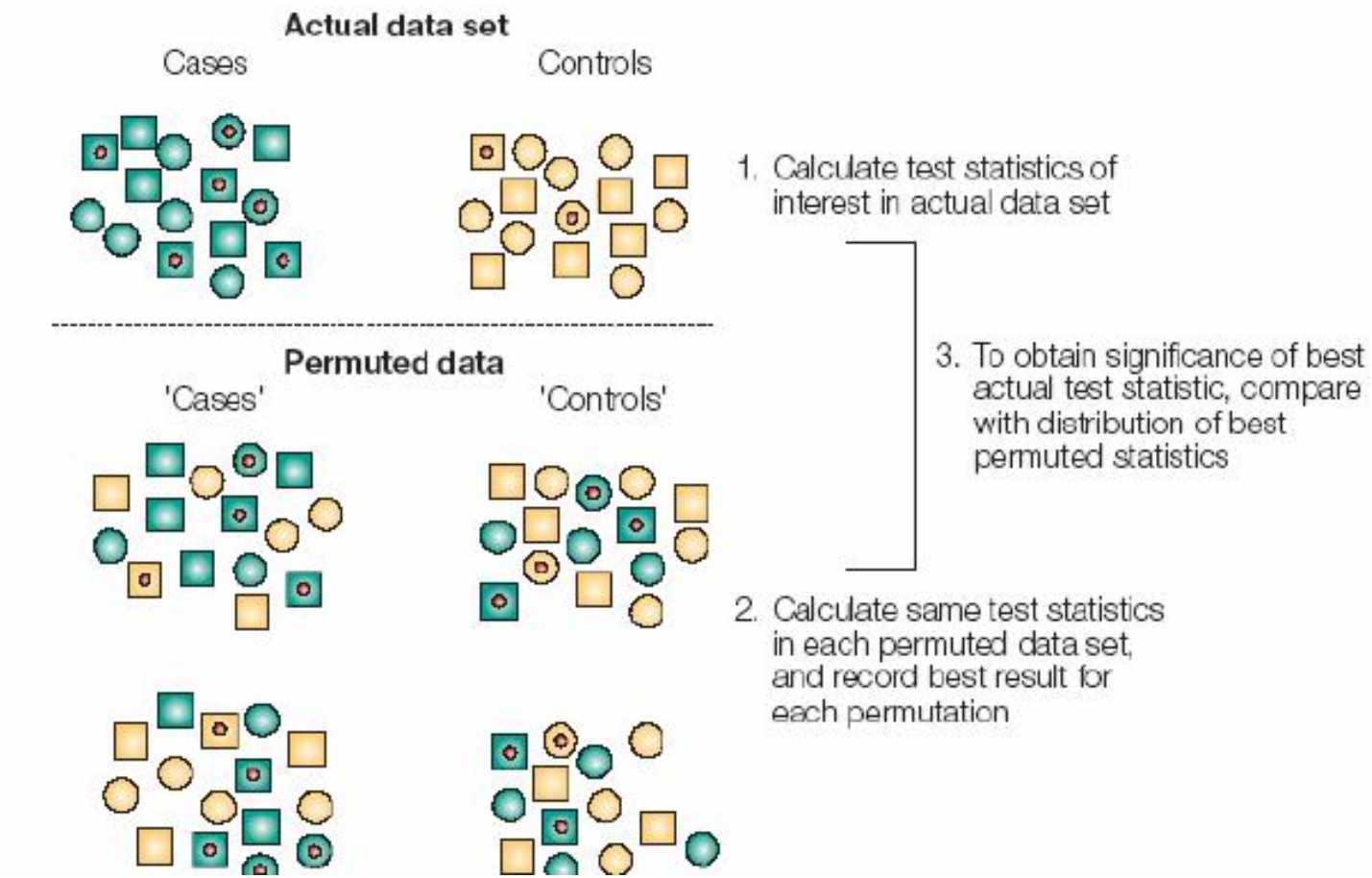
```
> ((6-4.6656)^2)/4.6656
[1] 0.3816494
> ((1-3.6288)^2)/3.6288
[1] 1.904373
> ((2-0.7056)^2)/0.7056
[1] 2.374534
> 0.3816494+1.904373+2.374534
[1] 4.660556
```

Exact test

- ***The probability of the observed sample is used to define the rejection zone, and the P-value of the test corresponds to the sum of the probabilities of all tables (with the same allelic counts) with the same or lower probability***.
- This is the "exact HW test" of Haldane (1954), Weir (1990b), Guo and Thompson (1992) and others. When the alternative hypothesis (H1) of interest is heterozygote excess or deficiency, more powerful tests than the probability-test can be used.

: The exact test does not rely on the chi-square distribution and thus offers a solution to the problem of sparse cells. For each combination of genotype and allele frequencies we can attribute a probability.

Permutation Test



Five steps to a permutation test (Good 1994)

1. Analyze the problem? identify the hypothesis and the alternative (s) of interest.
2. Choose a statistic.
3. Compute the test statistic for the original observations.
4. Generate the null reference distribution by

- rearranging the labels in a manner consistent with the randomization procedure
- compute the test statistics
- repeat these two steps until you obtain the distribution of the test statistic for all possible rearrangements.

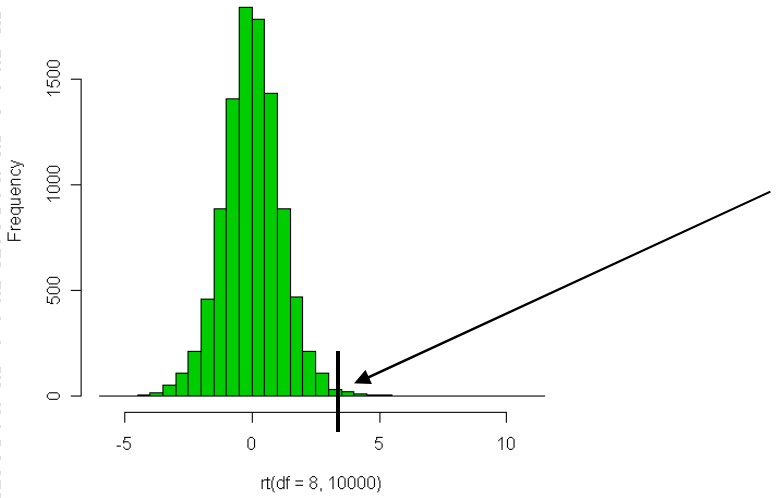
5. Accept or reject the hypothesis using this permutation distribution as a guide.

- **test statistic based on:**
- **a) the actual observations (Fisher-Pitman)**
- **b) their ranks**

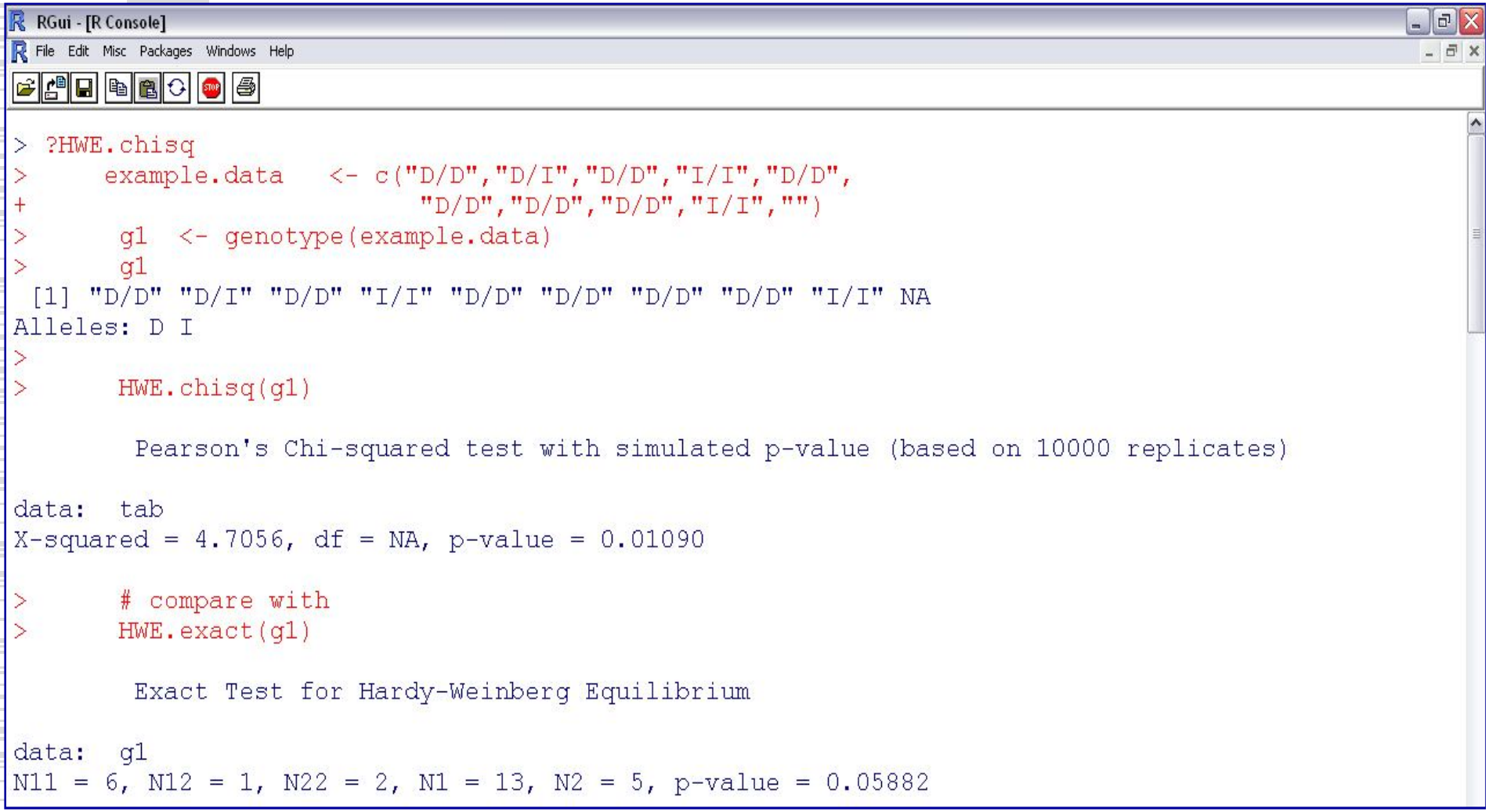
A numerical Example

- Gene Expression data for two groups
- Test statistic=t statistics (comparing two means).

	1	2	T
원자료	1,2,3,4	0,-1,-2,-3	4.38
P1	1,0, -2,-1	2,3,4,-3,	-1.19
¡ .	¡ .	¡ .	



One-sided p-value


RGui - [R Console]
File Edit Misc Packages Windows Help
> ?HWE.chisq
> example.data <- c("D/D","D/I","D/D","I/I","D/D",
+ "D/D","D/D","D/D","I/I","")
> g1 <- genotype(example.data)
> g1
[1] "D/D" "D/I" "D/D" "I/I" "D/D" "D/D" "D/D" "D/D" "I/I" NA
Alleles: D I
>
> HWE.chisq(g1)
Pearson's Chi-squared test with simulated p-value (based on 10000 replicates)
data: tab
X-squared = 4.7056, df = NA, p-value = 0.01090
> # compare with
> HWE.exact(g1)
Exact Test for Hardy-Weinberg Equilibrium
data: g1
N11 = 6, N12 = 1, N22 = 2, N1 = 13, N2 = 5, p-value = 0.05882

RGui - [R Console]

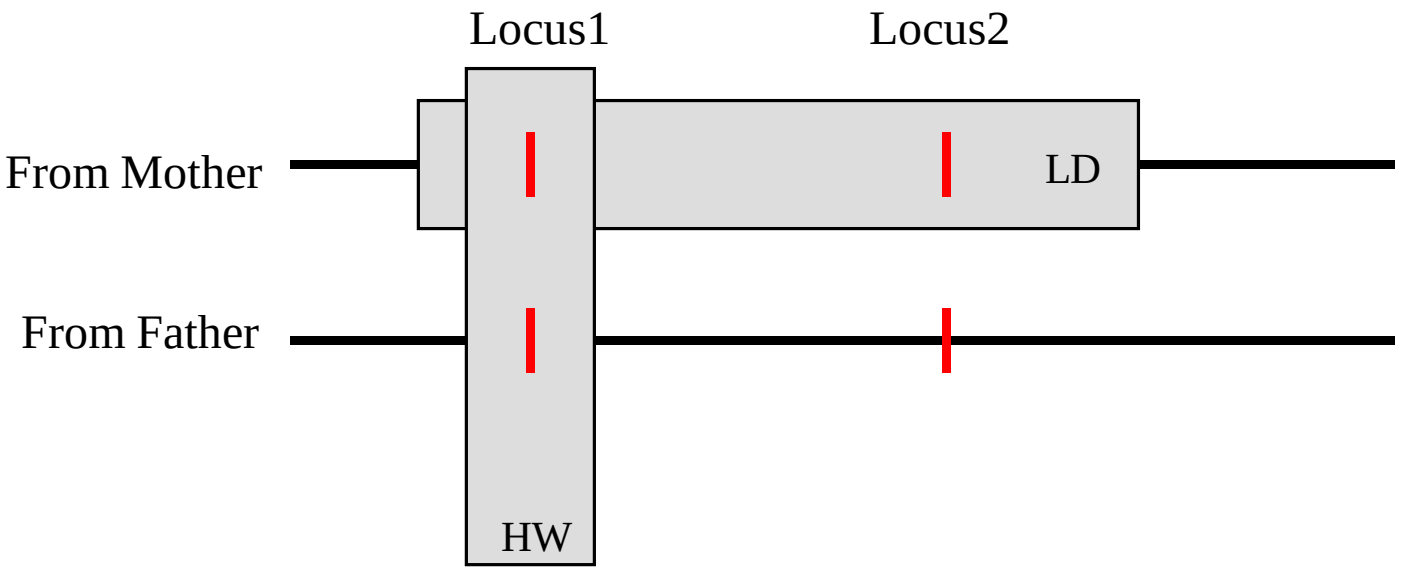
File Edit Misc Packages Windows Help

```
        ------------------------------------
        Test for Hardy-Weinberg-Equilibrium
        ------------------------------------

Call:
HWE.test.genotype(x = g1)

Raw Disequlibrium for each allele pair (D)


             D          I
  D            -0.1450617
  I -0.1450617

Scaled Disequlibrium for each allele pair (D')


        D     I
  D       -1.88
  I -1.88

Correlation coefficient for each allele pair (r)


            D         I
  D           0.7230769
  I 0.7230769

Overall Values

          Value
  D  -0.1450617
  D' -1.8800000
  r   0.7230769
```

> HWE.test(g1)

```
Confidence intervals computed via bootstrap using 1000 samples

    * WARNING: The R^2 disequlibrium statistics is bounded between [0,1].  The
    * confidence intervals for R^2 values near 0 and 1 are ill-behaved. A rough
    * correction has been applied, but the intervals still may not be correct for R^2
    * values near 0 or 1.


             Observed    95% CI                     NA's Contains Zero?
  Overall D   -0.14506173 (-0.24691358,  0.01234568) 0    YES
  Overall D'  -1.88000000 (-8.00000000,  0.12500000) 32   YES
  Overall r    0.72307692 (-0.12500000,  1.00000000) 32   YES
  Overall R^2  0.52284024 ( 0.00000000,  1.00000000) 32   YES

Significance Test:

        Exact Test for Hardy-Weinberg Equilibrium

data:  g1
N11 = 6, N12 = 1, N22 = 2, N1 = 13, N2 = 5, p-value = 0.05882



Warning message:
NAs returned from diseq call in: diseq.ci(x, R = ci.B, conf = conf)
```

Comparing HW & LD



LD (Linkage Disequilibrium)

Frequency		Site j		
		0 wild	1 variant	
Site i	0 wild	p_{00}	p_{oj}	P_{0+}
	1 variant	p_{i0}	p_{ij}	p_i
		P_{+0}	p_j	

D	$p_{ij} ? p_i p_j$
|D'|	$d_{ij}/d_{ij,max}$, $d_{ij,max}=max(p_i p_j,(1-p_i)(1-p_j))$, $d_{ij}<0$ $min((1-p_i)p_j, p_i(1-p_j))$, $d_{ij}>0$
r^2	$(p_{ij}p_{00}-p_{i0}p_{0j})^2/p_i p_{0+} p_j p_{+0}$

Example

? *LD , print.LD* ?

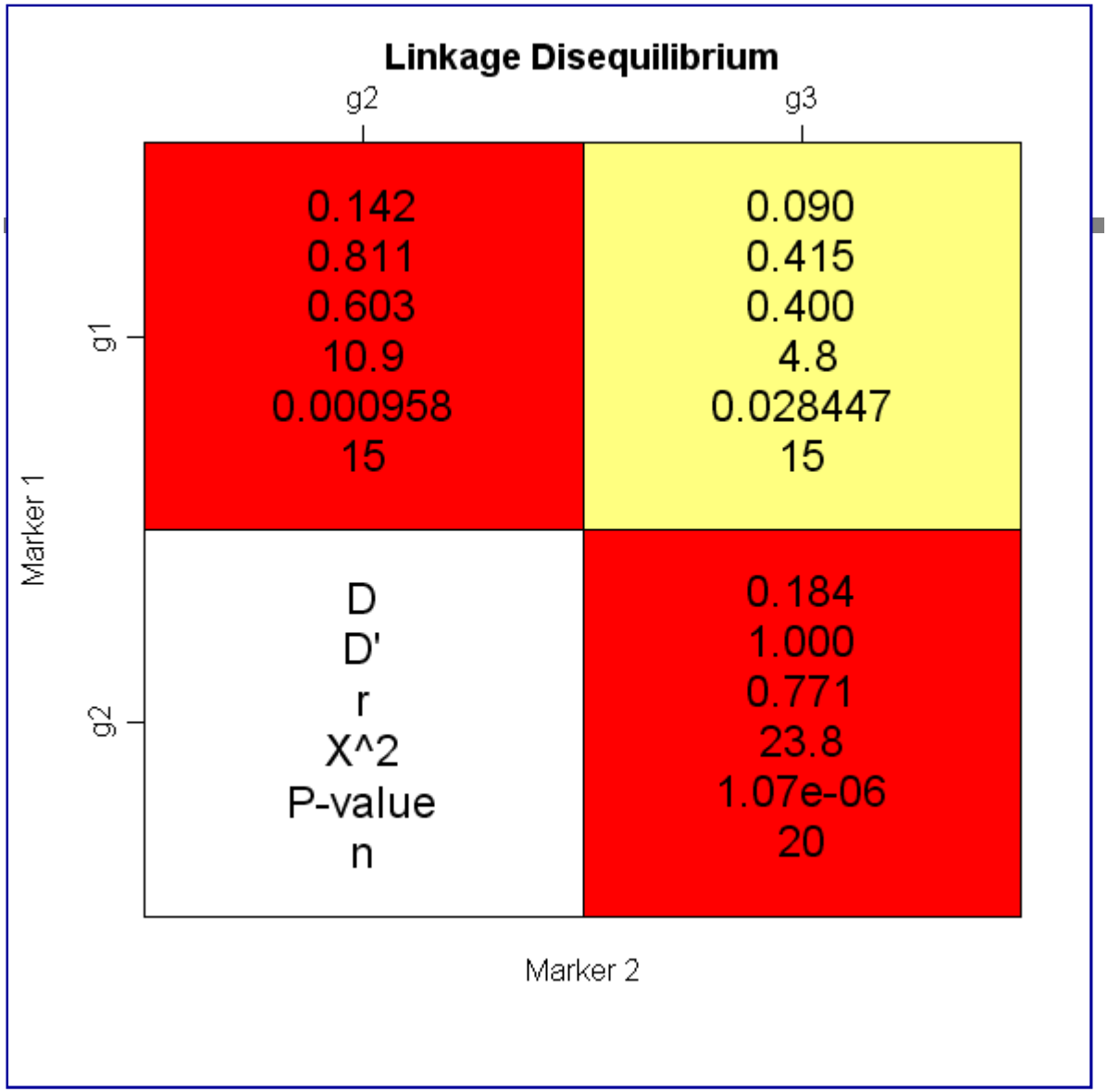
```
>      g1 <- genotype( c('T/A',    NA, 'T/T',    NA, 'T/A',    NA, 'T/T', 'T/A',
+                        'T/T', 'T/T', 'T/A', 'A/A', 'T/T', 'T/A', 'T/A', 'T/T',
+                          NA, 'T/A', 'T/A',   NA) )
>
>       g2 <- genotype( c('C/A', 'C/A', 'C/C', 'C/A', 'C/C', 'C/A', 'C/A', 'C/A',
+                        'C/A', 'C/C', 'C/A', 'A/A', 'C/A', 'A/A', 'C/A', 'C/C',
+                        'C/A', 'C/A', 'C/A', 'A/A') )
>
>       g3 <- genotype( c('T/A', 'T/A', 'T/T', 'T/A', 'T/T', 'T/A', 'T/A', 'T/A',
+                        'T/A', 'T/T', 'T/A', 'T/T', 'T/A', 'T/A', 'T/A', 'T/T',
+                        'T/A', 'T/A', 'T/A', 'T/T') )
>       data <- makeGenotypes(data.frame(g1,g2,g3))
>
>       # Compute & display  LD for one marker pair
>       ld <- LD(g1,g2)
>       print(ld)

Pairwise LD
-----------
                    D         D'
Estimates: 0.1419402 0.8110866
                 Corr
Estimates: 0.6029553

               X^2      P-value  N
LD Test: 10.90665 0.0009581958 15
```

? For genotype data, AB/ab cannot be distinguished from aB/Ab.

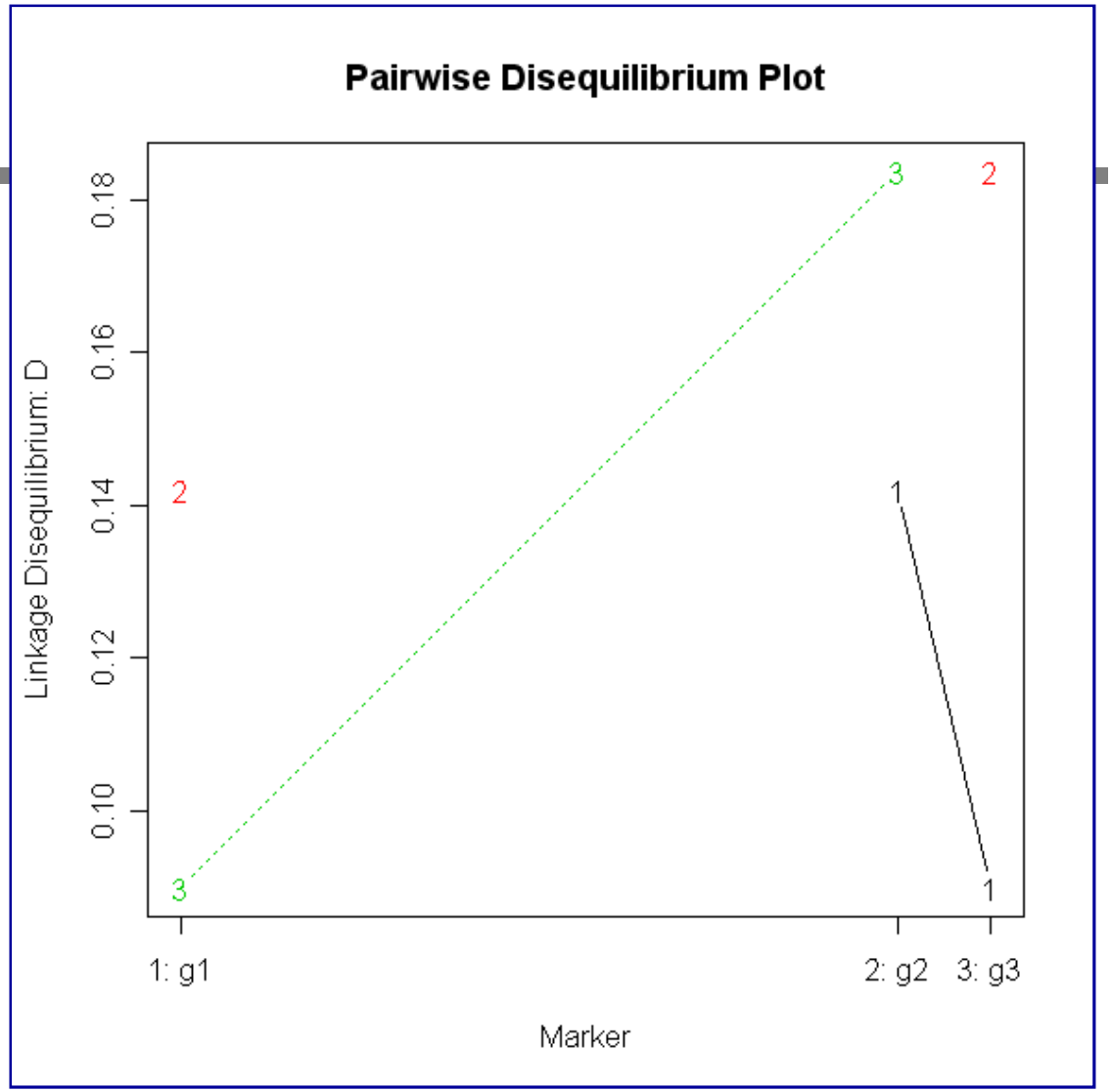
? Consequently, we estimate p(AB) using maximum likelihood and use this value in the computations.

```
>       # Compute LD table for all 3 genotypes
>       ldt <- LD(data)
>
>       # display the results
>       print(ldt)                                  # textual display

Pairwise LD
-----------
                       g2           g3
g1 D           0.1419402    0.0899457
g1 D'          0.8110866    0.4151342
g1 Corr.       0.6029553    0.4000333
g1 X^2        10.9066513    4.8007990
g1 P-value 0.0009581958   0.02844654
g1 n                  15           15

g2 D                       0.1836927
g2 D'                      0.9996881
g2 Corr.                   0.7712137
g2 X^2                    23.7908198
g2 P-value              1.073934e-06
g2 n                              20
```

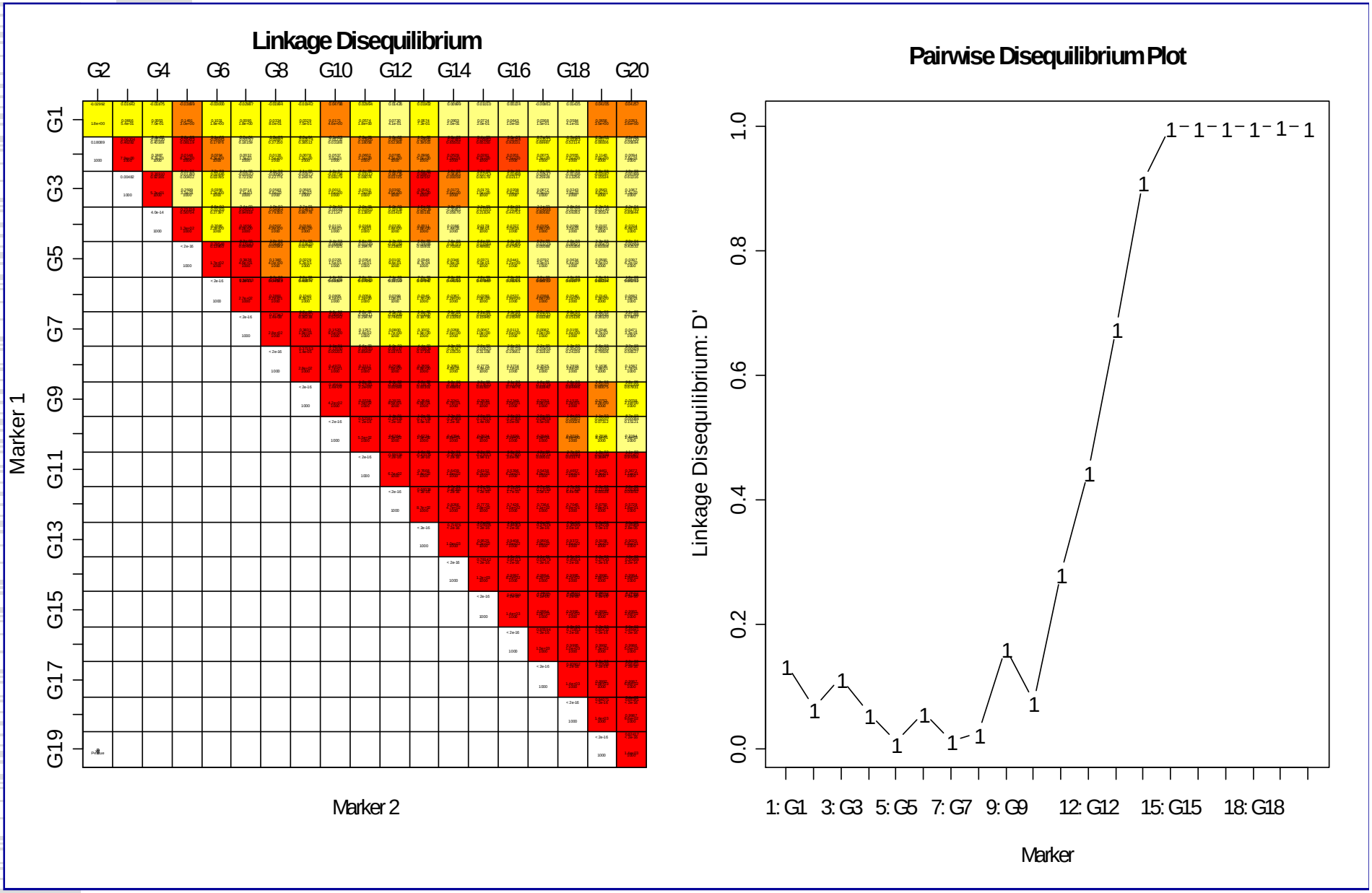
> LDtable(ldt)


Linkage Disequilibrium
g2
g3
g1
g2
Marker 1
Marker 2
0.142
0.811
0.603
10.9
0.000958
15
0.090
0.415
0.400
4.8
0.028447
15
D
D'
r
X^2
P-value
n
0.184
1.000
0.771
23.8
1.07e-06
20

> LDplot(ldt, distance=c(124, 834, 927))



plot(ldt, digits=2, marker=19)



Outline

- R packages
- Genetic data analysis
  - Types of packages
- Association study
  - Genetics package
  - DGC.genetics package - allele, genotype based tests
  - Haplo.stats package
- Summary

The dgc.genetics package

- Add-in package downloaded from http://www-gene.cimr.cam.ac.uk/clayton/software/
- Download the package as DGCgenetics_1.0.zip file to your directory ➔ *Start* R ➔ Packages ➔ Install Package(s) from local zip files¡
- Functions : allele.table, gcontrasts, logit

Using exercises from Dr. Heather Cordell website

```
> casecon <- read.table("casecondata.Rped", header=T)
> casecon[1:5,]
  pedigree id id.father id.mother sex affected loc1_1 loc1_2 loc2_1 loc2_2 loc3_1 loc3_2 loc4_1 loc4_2
1        1  1        NA        NA   1        2      1      1      2      2      1      1      2      2
2        2  1        NA        NA   1        2      1      2      2      1      1      2      2      1
3        3  1        NA        NA   1        2      1      2      2      1      1      2      2      1
4        4  1        NA        NA   1        2      1      1      2      2      1      1      2      2
5        5  1        NA        NA   1        2      1      1      2      1      1      1      2      2
> nrow(casecon)
[1] 1056
> ncol(casecon)
[1] 14
➢attach(casecon)
> case <- affected-1
> g1 <- genotype(loc1_1, loc1_2)
> g2 <- genotype(loc2_1, loc2_2)
> g3 <- genotype(loc3_1, loc3_2)
> g4 <- genotype(loc4_1, loc4_2)
```

*** Case control data**

➔ 1,056 unrelated persons

➔ 4 markers

http://www.staff.ncl.ac.uk/heather.cordell/WTACdancaseconexercise2005.html

Using exercises from Dr. Heather Cordell website

```
> table(g1,case)
    case
g1     0   1
 1/1 196 159
 1/2 323 178
 2/2 131  47
> chisq.test(g1,case)
        Pearson's Chi-squared test
data:  g1 and case
X-squared = 18.2406, df = 2, p-value = 0.0001094
> allele.table(g1,case)
   case
g1    0   1
 1 715 496
 2 585 272
> chisq.test(allele.table(g1,case))
        Pearson's Chi-squared test with Yates' continuity correction
data:  allele.table(g1, case)
X-squared = 17.8783, df = 1, p-value = 2.355e-05
```

Genotype based test

Allele based test

http://www.staff.ncl.ac.uk/heather.cordell/WTACdancaseconexercise2005.html

Using exercises from Dr. Heather Cordell website

```
> gcontrasts(g1) <- "genotype"
> logit (case ~ g1)
Logistic regression:  case ~ g1

Odds ratios (1 unit change), lower and upper confidence limits, and tests:

        OR    Lower    Upper   z-test    P-value
G1 1/1 1.4720534 1.1148766 1.943660  2.726918 0.006392893
G1 2/2 0.6510421 0.4451425 0.952180 -2.212623 0.026923639

> anova(logit (case ~ g1))
Analysis of Deviance Table
Model: binomial, link: logit
Response: case
Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev
NULL              1033   1364.22
g1    2   18.49    1031   1345.73
```

Logistic regression

```
> contrasts(g1)
    1/1 2/2
1/1   1   0
1/2   0   0
2/2   0   1
```

http://www.staff.ncl.ac.uk/heather.cordell/WTACdancaseconexercise2005.html

Using exercises from Dr. Heather Cordell website

> gcontrasts(g1) <- ¡additive"

> logit (case ~ g1)

Logistic regression

Logistic regression: case ~ g1

Odds ratios (1 unit change), lower and upper confidence limits, and tests:

	OR	Lower	Upper	z-test	P-value
g1:a:2	0.668319	0.5548023	0.805062	-4.243053	2.204990e-05

> contrasts(g1)
:a:2
1/1 0
1/2 1
2/2 2

> gcontrasts(g1)<-"dominance"

> logit(case~g1)

Logistic regression: case ~ g1

Odds ratios (1 unit change), lower and upper confidence limits, and tests:

	OR	Lower	Upper	z-test	P-value
g1:a:2	0.6650323	0.5462618	0.8096264	-4.0639100	4.825744e-05
g1:d:1:2	1.0214889	0.7808252	1.3363295	0.1551081	8.767361e-01

> contrasts(g1)

	:a:2	:d:1:2
1/1	0	0
1/2	1	1
2/2	2	0

http://www.staff.ncl.ac.uk/heather.cordell/WTACdancaseconexercise2005.html

Outline

- R packages
- Genetic data analysis
  - Types of packages
- Association study
  - Genetics package
  - DGC.genetics package
  - Haplo.stats package ? haplotype based tests
- Summary

The haplo.stats Package

- Title : Statistical Analysis of Haplotypes with Traits and Covariates when Linkage Phase is Ambiguous
- Date : 2007-6-14
- Description : A suite of S-PLUS/R routines for the analysis of indirectly measured haplotypes. The statistical methods assume that all subjects are unrelated and that haplotypes are ambiguous (due to unknown linkage phase of the genetic markers).
- Functions : haplo.em, haplo.glm and haplo.score.

Example

? haplo.group, haplo.glm, haplo.score, haplo.cc ?

? GLM Regression of Trait on Ambiguous Haplotypes

? **Description:** Perform glm regression of a trait on haplotype effects, allowing for ambiguous haplotypes. This method performs an iterative two-step EM, with the posterior probabilities of pairs of haplotypes per subject used as weights to update the regression coefficients, and the regression coefficients used to update the posterior probabilities.

? *Continuing the case control data from previous slides¡ .*

Making dataset

```
> locus<-as.matrix(casecon[,7:14])
> locus[1:2,]
  loc1_1 loc1_2 loc2_1 loc2_2 loc3_1 loc3_2 loc4_1 loc4_2
1      1      1      2      2      1      1      2      2
2      1      2      2      1      1      2      2      1
> hap<-setupGeno(locus,miss.val=c(0,NA))
> hap[1:2,]
     loc-1.a1 loc-1.a2 loc-2.a1 loc-2.a2 loc-3.a1 loc-3.a2 loc-4.a1 loc-4.a2
[1,]        1        1        2        2        1        1        2        2
[2,]        1        2        2        1        1        2        2        1
>  casecon$y<-casecon$affected -1
>  my.data <- data.frame(hap=hap, y=casecon$y)
```

> haplo.group(casecon$y,hap)

Counts per Grouping Variable Value

---

group
 0 1
672 384

Frequencies for Haplotypes by Grouping Variable using maximum likelihood estimates

---

Haplotype Frequencies By Group

---

	loc.1	loc.2	loc.3	loc.4	Total	casecon.y.0	casecon.y.1
1	1	1	1	1	0.01880	0.01838	0.01963
2	1	1	1	2	0.16092	0.14803	0.18205
3	1	1	2	1	0.02623	0.03009	0.01967
4	1	2	1	1	0.00173	0.00174	0.00169
5	1	2	1	2	0.37837	0.35259	0.42279
6	2	1	1	1	0.00132	0.00064	0.00237
7	2	1	1	2	0.00256	0.00118	0.00558
8	2	1	2	1	0.40690	0.44269	0.34621
9	2	1	2	2	0.00049	0.00078	NA
10	2	2	1	2	0.00208	0.00295	0.00000
11	2	2	2	1	0.00059	0.00093	NA
12	2	2	2	2	0.00000	0.00000	NA

> fit.hap <- haplo.glm(y ~ hap, family = binomial, allele.lev= attributes(hap)$unique.alleles, data=my.data)

> fit.hap

Call:

haplo.glm(formula = y ~ hap, family = binomial, data = my.data, allele.lev = attributes(hap)$unique.alleles)

coefficients:

	coef	se	t.stat	pval
(Intercept)	-1.049	0.135	-7.791	1.60e-14
hap.1	0.332	0.347	0.957	3.39e-01
hap.2	0.432	0.131	3.306	9.77e-04
hap.3	-0.124	0.311	-0.400	6.89e-01
hap.5	0.429	0.107	4.012	6.46e-05
hap.7	1.753	0.203	8.629	0.00e+00
hap.rare	-0.335	0.807	-0.415	6.78e-01

Haplotypes:

	loc.1	loc.2	loc.3	loc.4	hap.freq
hap.1	1	1	1	1	0.01892
hap.2	1	1	1	2	0.16043
hap.3	1	1	2	1	0.02630
hap.5	1	2	1	2	0.37802
hap.7	2	1	1	2	0.00277
hap.rare	*	*	*	*	0.00598
haplo.base	**2**	**1**	**2**	**1**	**0.40758**

hap.8

Haplotype based logistic regression

Score Statistics for Association of Traits with Haplotypes

```
> score.hap <- haplo.score(casecon$y, hap,trait.type="binomial")
> score.hap
```

Global Score Statistics

---

global-stat = 25.37852, df = 6, p-val = 0.00029

Haplotype-specific Scores

---

	loc-1	loc-2	loc-3	loc-4	Hap-Freq	Hap-Score	p-val
[1,]	2	1	2	1	0.4069	-4.31276	2e-05
[2,]	1	1	2	1	0.02623	-1.40452	0.16016
[3,]	1	1	1	1	0.0188	0.24407	0.80717
[4,]	2	1	1	2	0.00256	1.60225	0.1091
[5,]	1	1	1	2	0.16092	2.07692	0.03781
[6,]	1	2	1	2	0.37837	3.16712	0.00154

```
> cc.test <- haplo.cc(casecon$y, hap)
> cc.test
```

Global Score Statistics

---

global-stat = 25.346, df = 6, p-val = 0.00029458

Counts for Cases and Controls

---

control case
652 384

Haplotype Association Analysis in a Case-Control design
Combine results from haplo.score, haplo.group, and haplo.glm for case-control study designs.

Haplotype Scores, p-values, Hap-Frequencies (hf), and Odds Ratios (95% CI)

---

	loc-1	loc-2	loc-3	loc-4	Hap-Score	p-val	pool.hf	control.hf	case.hf	glm.eff	OR.lower	OR	OR.upper
8	2	1	2	1	-4.31314	1.61E-05	4.07E-01	4.43E-01	3.46E-01	Base	NA	1	NA
3	1	1	2	1	-1.40442	1.60E-01	2.62E-02	3.01E-02	1.97E-02	Eff	0.48098	0.88384	1.6241
1	1	1	1	1	0.24406	8.07E-01	1.88E-02	1.84E-02	1.96E-02	Eff	0.70602	1.39392	2.7521
7	2	1	1	2	1.59757	1.10E-01	2.56E-03	1.18E-03	5.58E-03	Eff	3.87403	5.77298	8.6027
2	1	1	1	2	2.07665	3.78E-02	1.61E-01	1.48E-01	1.82E-01	Eff	1.19211	1.53994	1.9893
5	1	2	1	2	3.16786	1.54E-03	3.78E-01	3.53E-01	4.23E-01	Eff	1.24542	1.53598	1.8943
4	1	2	1	1	NA	NA	1.73E-03	1.74E-03	1.69E-03	R	0.14708	0.71474	3.4733
6	2	1	1	1	NA	NA	1.32E-03	6.37E-04	2.37E-03	R	0.14708	0.71474	3.4733
9	2	1	2	2	NA	NA	4.94E-04	7.85E-04	NA	<NA>	NA	NA	NA
10	2	2	1	2	NA	NA	2.08E-03	2.95E-03	1.89E-09	R	0.14708	0.71474	3.4733
11	2	2	2	1	NA	NA	5.93E-04	9.35E-04	NA	<NA>	NA	NA	NA
12	2	2	2	2	NA	NA	1.05E-09	1.51E-09	NA	<NA>	NA	NA	NA

Outline

- R packages
- Genetic data analysis
  - Types of packages
- Association study
  - Genetics package
  - DGC.genetics package
  - Haplo.stats package
- Summary

Take home messages

- There is no gold standard tools to analyze genetic data. There are other several similar packages, and statistical genetics programs which we did not deal with today.
- R wants to communicate with you. Try to understand their languages and logics to program.
- Try to understand the usage of functions by running examples in popping up manual. *(?function name*)

Thank you !

- Questions & comments to

  - Ho Kim hokim@snu.ac.kr
  - Aekyung Park parkak11@snu.ac.kr
  - Yoonhee Kim nina78@snu.ac.kr