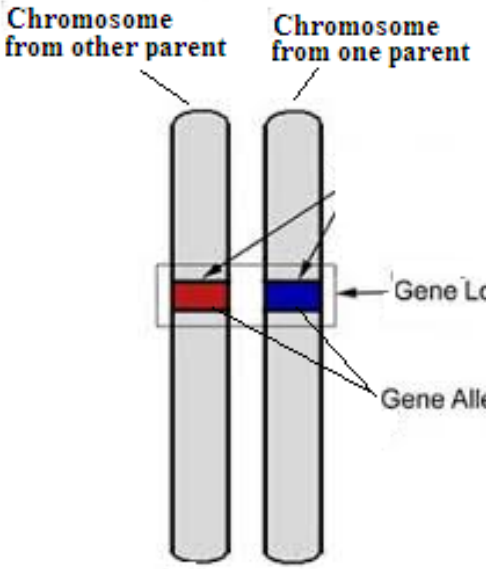
UNIVERSITY OF THESSALY
School of Medicine
Laboratory of Biomathematics

Analysis of Genetic Association and Pharmacogenetic Studies

(see papers by Zintzaras)

Genetics background

Chromosomes are very long molecule of double-stranded DNA and proteins

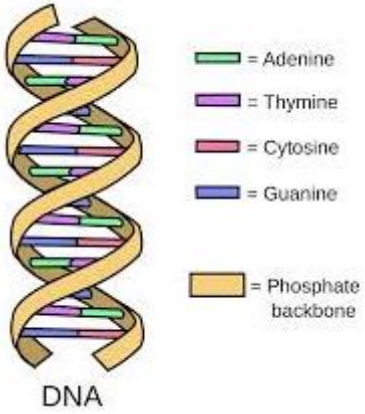


Genes are segments of DNA that code for proteins

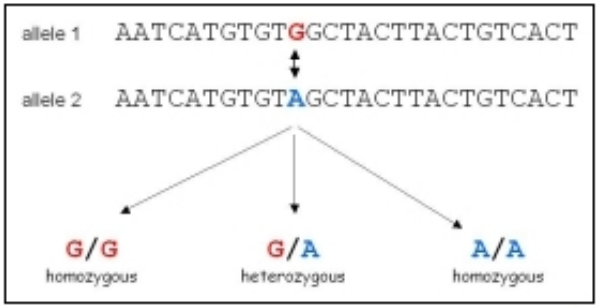
May be the same or different (mt or wt)

The genotype can be homozygous (wtwt or mtmt) or heterozygous (wtmt)

A locus may include multiple alleles (a subject possesses only two such alleles)

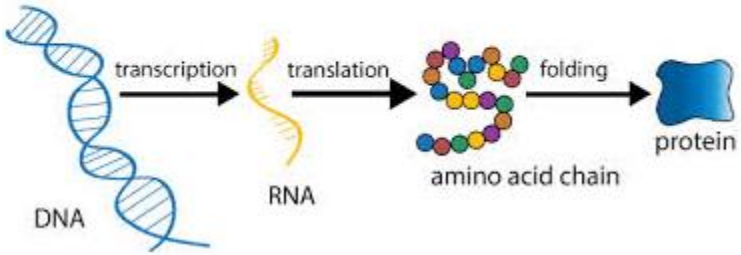


SNP or gene variant is an exchange of individual bases at the DNA level



This variation occurrs commonly in a population

Gene expression is the process by which genetic instructions are used to synthesize proteins.



Genetic association studies (GAS)

The evaluation of possible associations between phenotypic traits (diseases) and genetic variants (gene polymorphisms) is carried out using GAS

In the course, we will examine the following cases of GAS:

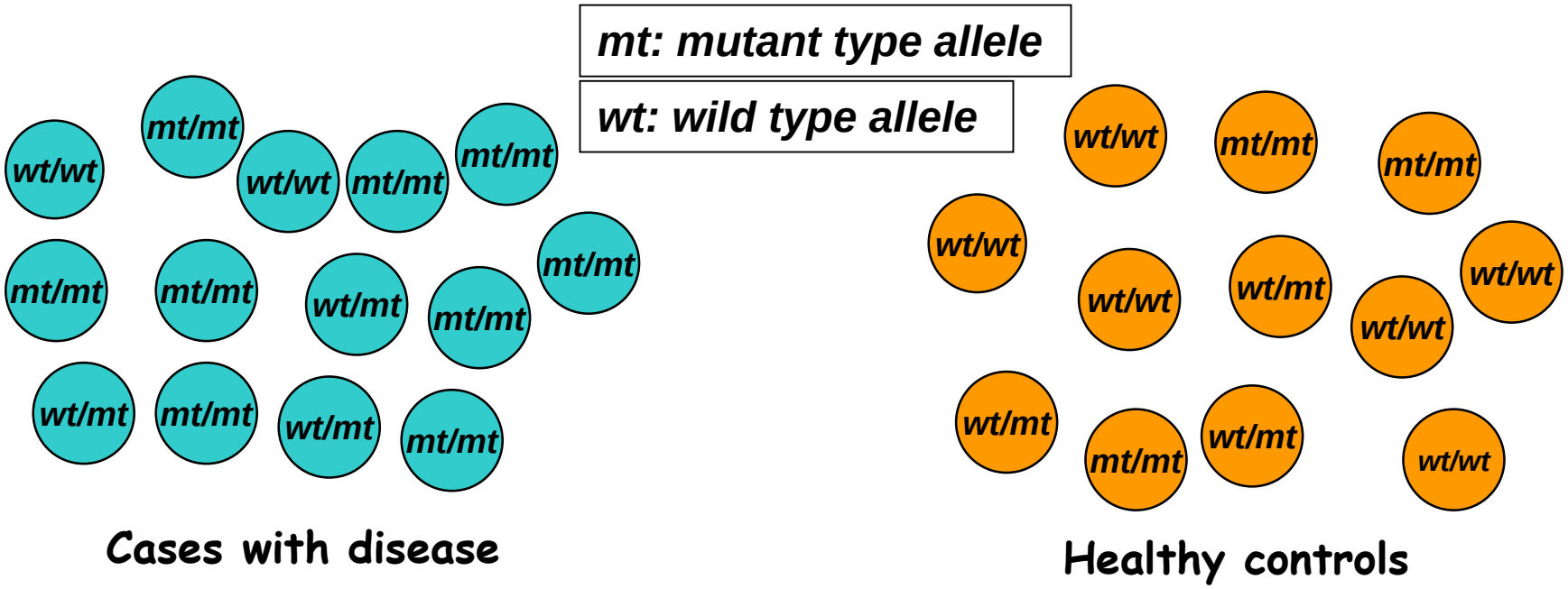
- **GAS for bi-allelic variant and binary outcome (healthy/diseased)**
- **GAS for bi-allelic variant and disease progression (healthy/diseased/diseased with complications)**
- **GAS for multi-allelic variant and binary outcome (healthy/diseased)**

and

Pharmacogenetic Studies

GAS for bi-allelic variant and binary outcome (healthy/diseased)

Study description



In the case of a (bi-allelic) genetic variant with two alleles (mutant type-*mt* and wild type-wt), where *mt* is thought to be associated with a disease, GAS will collect information on the numbers of diseased subjects and control subjects with each of the three genotypes (*wt/wt*, *wt/mt*, *mt/mt*)

Example

For example, in a GAS with 8261/4374 cases/controls, the association between *ACE D/I* (*wt/mt*) and CAD was investigated. The genotype distribution was as follows:

Genotype	Cases with CAD	Controls
mt/mt	1788	*874*
mt/wt	4145	*2165*
wt/wt	2328	*1335*

Genotype	Cases with CAD	Controls
mt/mt	1788	*874*
mt/wt	4145	*2165*
wt/wt	2328	*1335*

We would like to examine whether there is association between genotype distribution of ACE (D/I) (mt/mt, mt/wt, wt/wt) and susceptibility to CAD.

Study quality assessment

- **Prior to testing the association between a variant and the disease, the quality of a study should be assessed.**
- **A study quality surrogate point is whether the controls conform with the Hardy-Weinberg Equilibrium (HWE) rule in the controls**

- **Lack of HWE implies:**

  -genotyping errors and/or

  -structure in the population (i.e. non-unselected controls)

- **Even though the controls are not in HWE, the study can still be salvaged: a technique for adjusting for HWE departure exists.**

Hardy-Weinberg Equilibrium (HWE)

In our example, the genotype distribution is as follows:

Genotype	Cases with CAD	Controls
mt/mt	1788	*874*
mt/wt	4145	*2165*
wt/wt	2328	*1335*

Lets denote

the genotypic frequencies of the controls as

$P_{wt/wt}$, $P_{mt/mt}$ **and** $P_{mt/wt}$ **and**

the frequencies of the wt and mt alleles as

$f(wt)=p$ **and** $f(mt)=q=1-p$**.**

Genotype	Controls
mt/mt	*874*
mt/wt	*2165*
wt/wt	*1335*

When the controls are in HWE, we expect the genotypic frequencies to be equal to the products of corresponding allele frequencies:

$$P_{wt/wt} = p * p = p^2$$

$$P_{mt/mt} = q * q = q^2$$

$$P_{mt/wt} = 2 * p * q$$

Thus, the genotype distribution should follow the following HWE rule:

$$P_{wt/wt} : P_{mt/mt} : P_{mt/wt} = p^2 : 2pq : q^2$$

Genotype	Controls
mt/mt	*874*
mt/wt	*2165*
wt/wt	*1335*
Total (N)	*4374*

HWE rule

$P_{wt/wt}:P_{mt/mt}:P_{mt/wt} = p^2:2pq:q^2$

f(wt) = p = (2 * 1335 + 2165) / (2*4374) = 0.553

f(mt) = q = 1-p = 0.447

Thus, if the controls followed the HWE rule, we would expect the following distribution:

Genotype	Expected number of controls in HWE		
mt/mt	$E1 = q^2*N$	$= 0.447^2*4347$	$= 875$
mt/wt	$E2 = 2*p*q*N$	$= 2*0.553*0.447*4347$	$= 2163$
wt/wt	$E3 = p^2*N$	$= 0.553^2*4347$	$= 1336$

Controls		
Genotype	**Observed number**	**Expected number in HWE**
mt/mt	**O1 = *874***	**E1 = *875***
mt/wt	**O2 = *2165***	**E2 = 2163**
wt/wt	**O3 = 1335**	**E3 = 1336**

The departure from HWE (i.e. the differences between the observed and the expected values under the HWE rule) is tested using a x^2-test with (3-1-1)=1 df.

$$\mathbf{x}^2 = \frac{(\mathbf{O}_1 - \mathbf{E}_1)^2}{\mathbf{E}_1} + \frac{(\mathbf{O}_2 - \mathbf{E}_2)^2}{\mathbf{E}_2} + \frac{(\mathbf{O}_3 - \mathbf{E}_3)^2}{\mathbf{E}_3} =$$

$$\frac{(874-875)^2}{875} + \frac{(2165-2163)^2}{2163} + \frac{(1335-1336)^2}{1336} = 0.01$$

The x^2=0.01 is less than the 5% point of the x^2-distribution with 1 df which is 3.84.

Then, P>0.05 (P=0.944).

Thus, there is no real differences between the observed and the expected values under the HWE rule, i.e. the controls are in HWE.

The URL
http://www.had2know.com/academics/hardy-weinberg-equilibrium-calculator-2-alleles.html
provides a calculator for testing HWE

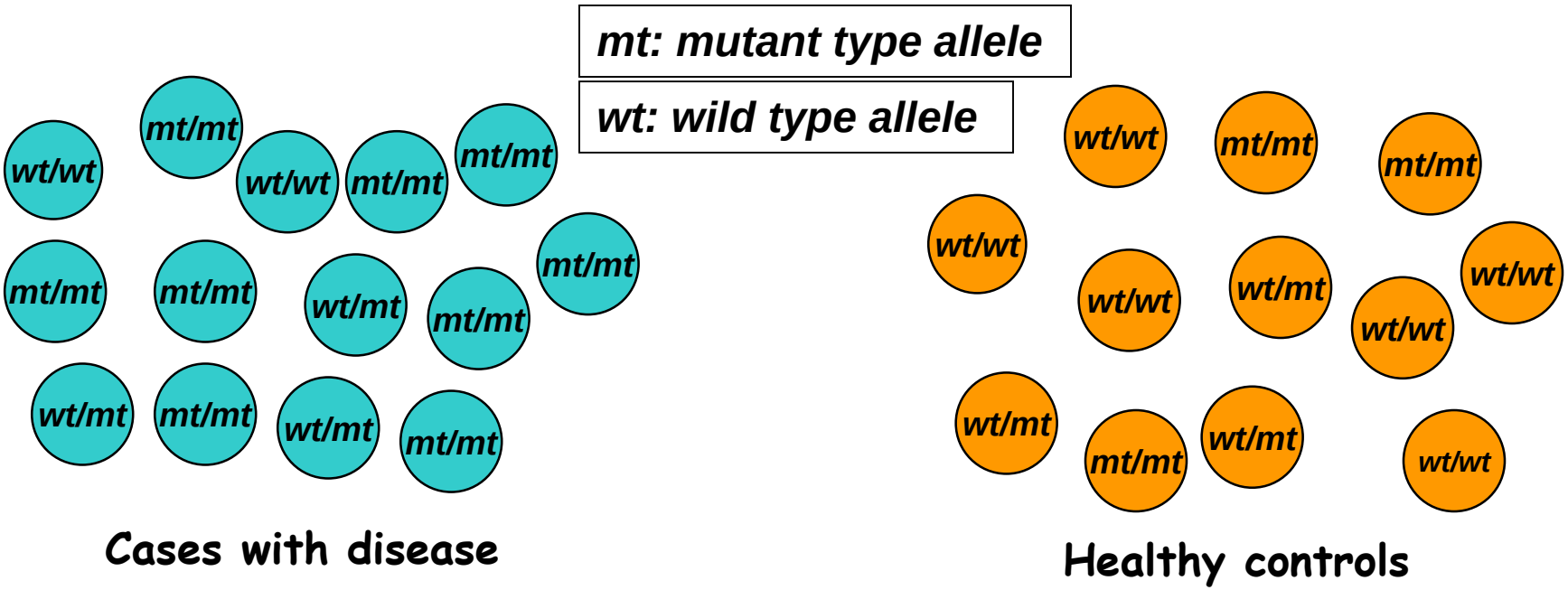
Practice

The distributions of genotypes of two MyD88 gene variants C938A and C1944G among cases with Hodgkin's lymphoma and healthy control subjects are the following.

Variant	Genotype	Cases N	Healthy controls N
MyD88 C938A	CC	46	74
	CA	50	15
	AA	5	3
MyD88 C1944G	CC	77	65
	CG	21	25
	GG	3	3

For each variant, test for HWE the controls.

Testing the association



The association between the genotype distribution and disease susceptibility is tested using a x^2-test.

Then, the magnitude of association is expressed in terms of odds ratio (OR) for various genetic models (contrasts)

For example, in a GAS with 8261/4374 cases/controls, the association between *ACE D/I* (*wt/mt*) and CAD was investigated. The genotype distribution was as follows:

Genotype	Cases with CAD	Controls
mt/mt	1788	*874*
mt/wt	4145	*2165*
wt/wt	2328	*1335*

The association between disease status and the genetic variant is tested using a x^2-test with (3-1)x(2-1)=2 df

x^2-test

Genotype	Cases with CAD	Controls	Total
mt/mt	O1 = 1788	O4 = 874	2662
mt/wt	O2 = 4145	O5 = 2165	6310
wt/wt	O3 = 2328	O6 = 1335	3663
Total	8261	4374	12635

The expected numbers of cases with CAD assuming no association between variant and disease are

E1 = (2662*8261) / 12635 = 1741

E2 = (6310*8261) / 12635 = 4125

E3 = (3663*8261) / 12635 = 2395

The respective expected numbers for controls are

E4 = (2662*4374) / 12635 = 921

E5 = (6310*4374) / 12635 = 2184

E6 = (3663*4374) / 12635 = 1268

Genotype	Cases with CAD	Controls
mt/mt	O1 = 1788 (E1 = 1741)	O4 = 874 (E4 = 921)
mt/wt	O2 = 4145 (E2 = 4125)	O5 = 2165 (E5 = 2184)
wt/wt	O3 = 2328 (E3 = 2395)	O6 = 1335 (E6 = 1268)

Then, the x^2-test is as follows:

$$x^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4} + \frac{(O_5 - E_5)^2}{E_5} + \frac{(O_6 - E_6)^2}{E_6} = 9.42$$

The x^2-test is 9.42 which is greater than 5.99, the 5% point of the x^2-distribution with (3-1)*(2-1)=2 df.

Then, P<0.05 (P=0.009).

Thus, there is significant association between *ACE D/I* gene variant and development of CAD

To perform a chi-squared test, you may use the following URL:

http://www.quantpsy.org/chisq/chisq.htm

Practice

The distributions of genotypes of two MyD88 gene variants C938A and C1944G among cases with Hodgkin's lymphoma and healthy control subjects are the following.

Variant	Genotype	Cases N	Healthy controls N
MyD88 C938A	CC	46	74
	CA	50	15
	AA	5	3
MyD88 C1944G	CC	77	65
	CG	21	25
	GG	3	3

Test the association between each variant and the disease.

Testing the genetic model (mode of inheritance)

In a GAS, when the association is significant, various genetic models of genotypes are tested by merging genotypes

These models include:

- **recessive model:** homozygous for mt vs. wt-carriers
- **dominant model:** mt-carriers vs. homozygous for wt
- **additive model:** homozygous for mt vs. homozygous for wt
- **co-dominant model:** heterozygous vs. all homozygotes

However, when the association is significant (x^2-test), you always expect at least the additive or co-dominant models to be significant (it can be significant both of them).

The significance of the genetic model is assessed using a x^2-test or, alternatively, the respective odds ratio (*OR*) and its 95% confidence interval (CI).

Note that the OR provides a measure of the magnitude of association between the genetic model and the disease and the 95% CI indicates the significance of this magnitude.

Recessive model

Genotype	Cases with CAD	Controls
mt/mt	1788	*874*
wt-carriers (mt/wt+wt/wt)	6473 (=4145+2328)	3500 (=2165+1335)

The association between variant and the disease for the recessive model is tested using a chi-squared (x2) test with (2-1)x(2-1)=1 df

Genotype	Cases with CAD	Controls	Total
mt/mt	O1=1788	O3=*874*	*2662*
wt-carriers (mt/wt+wt/wt)	O2=6473	O4=3500	9973
Total	8261	4374	12635

The expected numbers of cases with CAD assuming no association between variant and disease for the recessive model are

E1 = (2662*8261) / 12635 = 1741
E2 = (9973*8261) / 12635 = 6521

The respective expected numbers for controls are

E3 = (2662*4374) / 12635 = 921
E4 = (9973*4374) / 12635 = 3453

Genotype	Cases with CAD	Controls
mt/mt	O1=1788 (E1=1741)	O3=874 (E3=921)
wt-carriers (mt/wt+wt/wt)	O2=6473 (E2=6521)	O4=3500 (E4=3453)

Then, the x^2-test is as follows:

$$x^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4} = 4.75$$

The x^2-test is 4.75 which is greater than 3.84, the 5% point of the x^2-distribution with 1 df.

Then, P<0.05 (P=0.029).

Thus, there is significant association between *ACE D/I* gene variant and development of CAD for the recessive model.

To perform the chi-squared test, you may use the following URL: http://www.quantpsy.org/chisq/chisq.htm

Magnitude of association

The magnitude of the association for the recessive model is show with the OR:

$$OR = \frac{\text{" probability" a subject of being diseased when mtmt}}{\text{" probability" a subject of being diseased when wt - carriers}}$$

For *OR>1*: an *mt/mt* subject has greater chance of being diseased than a *wt-carrier* subject

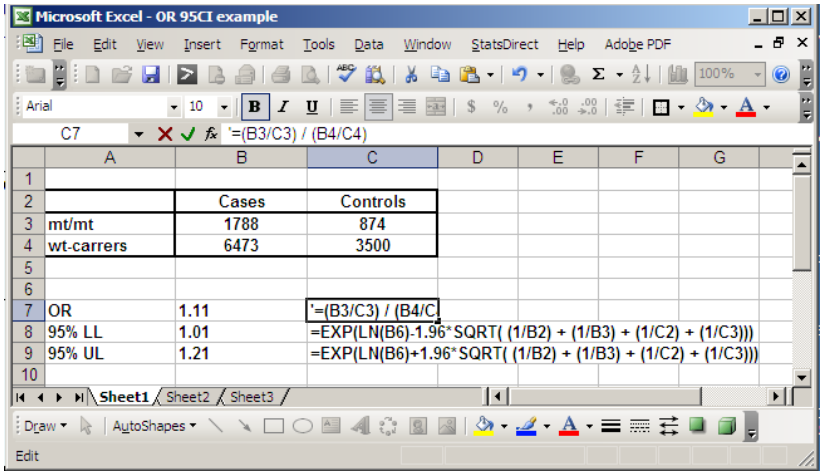
If the 95% CI does not include 1, then, the *OR* is significant (P<0.05).

Genotype	Cases with CAD	Controls
mt/mt	1788	*874*
wt-carriers (mt/wt+wt/wt)	6473	3500

$$OR = \frac{\text{" probability" a subject of being with CAD when mtmt}}{\text{" probability" a subject of being with CAD when wt-carrier}} = \frac{1788/874}{6473/3500} = 1.11$$

$$95\%CI = \left(e^{\ln(OR) - 1.96*\sqrt{\frac{1}{1788}+\frac{1}{6473}+\frac{1}{874}+\frac{1}{3500}}}, e^{\ln(OR) + 1.96*\sqrt{\frac{1}{1788}+\frac{1}{6473}+\frac{1}{874}+\frac{1}{3500}}}\right) =$$

$$= (1.01, 1.21)$$



OR = 1.11

95% CI = (1.01, 1.21)

Since OR=1.11, we conclude that homozygous for the mt allele (mt/mt) have 11% greater risk for CAD than wt-carriers.

Since "1" is not included in the 95% CI, we conclude that the OR is (marginally) significant (P<0.05).

Adjusted OR

- **The OR can be found also using logistic regression in SPSS.**

- **In the logistic regression, we can include one or more covariates (e.g. age); then, the OR is adjusted for these covariates.**

Dominant model

Genotype	Cases with CAD	Controls
mt-carriers (mt/mt+mt/wt)	5933 (=1788+4145)	*3039 (=874+2165)*
wt/wt	2328	1335

The association between variant and the disease for the recessive model is tested using a x^2-test with (2-1)x(2-1)=1 df

Genotype	Cases with CAD	Controls	Total
mt-carriers (mt/mt+mt/wt)	O1=5933	O3=3039	*8972*
wt/wt	O2=2328	O4=1335	3663
Total	8261	4374	12635

The expected numbers of cases with CAD assuming no association between variant and disease for the dominant model are

E1 = (8972*8261) / 12635 = 5866

E2 = (3663*8261) / 12635 = 2395

The respective expected numbers for controls are

E3 = (8972*4374) / 12635 = 3106

E4 = (3663*4374) / 12635 = 1268

Genotype	Cases with CAD	Controls
mt-carriers (mt/mt+mt/wt)	O1=5933 (E1=5866)	O3=3039 (E3=3106)
wt/wt	O2=2328 (E2=2395)	O4=1335 (E4=1268)

Then, the x^2-test is as follows:

$$\mathbf{x}^2 = \frac{(\mathbf{O}_1 - \mathbf{E}_1)^2}{\mathbf{E}_1} + \frac{(\mathbf{O}_2 - \mathbf{E}_2)^2}{\mathbf{E}_2} + \frac{(\mathbf{O}_3 - \mathbf{E}_3)^2}{\mathbf{E}_3} + \frac{(\mathbf{O}_4 - \mathbf{E}_4)^2}{\mathbf{E}_4} = 7.61$$

The x^2-test is 7.61 which is greater than 3.84, the 5% point of the x^2-distribution with 1 df.

Then, P<0.05 (P=0.006).

Thus, there is significant association between *ACE D/I* gene variant and development of CAD for the dominant model.

OR = 1.11

95% CI = (1.01, 1.21)

Since OR=1.11, we conclude that homozygous for the mt allele (mt/mt) have 11% greater risk for CAD than wt-carriers.

Since "1" is not included in the 95% CI, we conclude that the OR is (marginally) significant (P<0.05).

Magnitude of association

The magnitude of the association for the dominant model is show with the OR:

$$OR = \frac{\text{" probability" a subject of being diseased when mt - carrier}}{\text{" probability" a subject of being diseased when wt / wt}}$$

For *OR>1*: an *mt-carrier* subject has greater chance of being diseased than an homozygous *wt/wt* subject

Genotype	Cases with CAD	Controls
mt-carriers (mt/mt+mt/wt)	**5933**	***3039***
wt/wt	**2328**	**1335**

$$OR = \frac{\text{" probability" a subject of being with CAD when mt-carrier}}{\text{" probability" a subject of being with CAD when wt / wt}} =$$

$$= \frac{5933/3039}{2328/1335} = 1.12$$

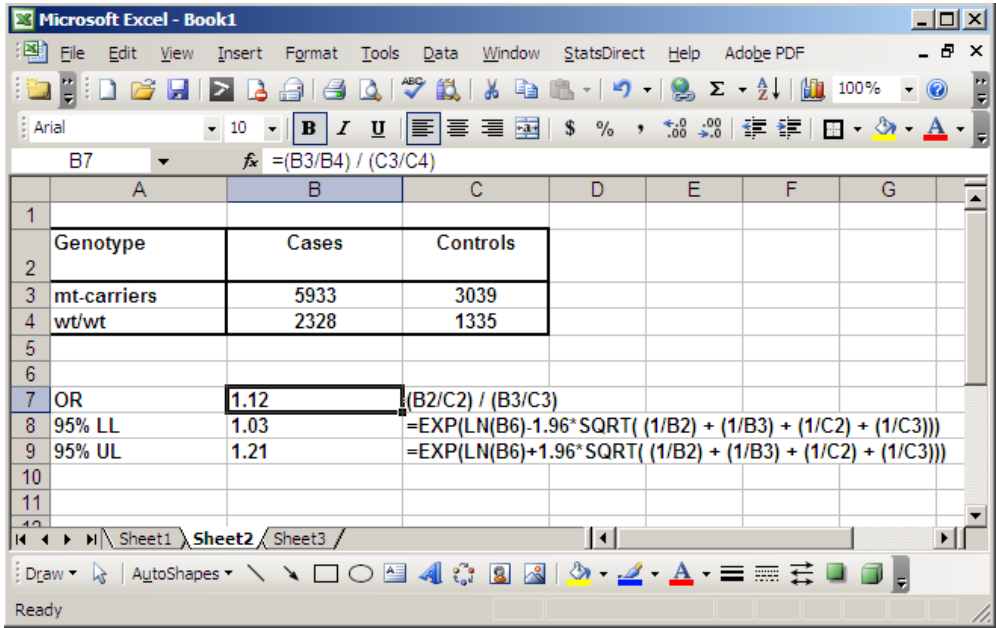
Since OR=1.12, **we conclude that wt-carriers have 12% greater risk for CAD than homozygous for the mt allele (mt/mt).**

Genotype	Cases with CAD	Controls
mt-carriers (mt/mt+mt/wt)	5933	*3039*
wt/wt	2328	1335

$$95\%CI = \left(e^{ln(OR)-1.96*\sqrt{\frac{1}{5933}+\frac{1}{2328}+\frac{1}{3039}+\frac{1}{1335}}}, e^{ln(OR)+1.96*\sqrt{\frac{1}{5933}+\frac{1}{2328}+\frac{1}{3039}+\frac{1}{1335}}}\right) =$$

$$= (1.03, 1.21)$$

Since "1" is not included in the 95% CI, we conclude that the OR is significant (P<0.05).



Additive model

Genotype	Cases with CAD	Controls
mt/mt	**1788**	***874***
wt/wt	**2328**	***1335***

$$OR = \frac{\text{" probability" a subject of being with CAD when mtmt}}{\text{" probability" a subject of being with CAD when wtwt}} = \frac{1788/874}{2328/1335} = 1.17$$

$$95\%CI = \left(e^{\ln(OR) - 1.96*\sqrt{\frac{1}{1788}+\frac{1}{2328}+\frac{1}{874}+\frac{1}{1335}}}, e^{\ln(OR) + 1.96*\sqrt{\frac{1}{1788}+\frac{1}{2328}+\frac{1}{874}+\frac{1}{1335}}}\right) =$$

$$= (1.06, 1.30)$$

$$OR = \frac{\text{" probability" a subject of being with CAD when mtmt}}{\text{" probability" a subject of being with CAD when wtwt}} = 1.17$$

$$95\%CI = (1.06, 1.30)$$

Since OR=1.17, we conclude that homozygous for the mt allele (mt/mt) have 17% greater risk for CAD than homozygous for the wt allele (wt/wt).

Since "1" is not included in the 95% CI, we conclude that the OR is significant (P<0.05).

Co-dominant model

Genotype	Cases with CAD	Controls
mt/wt	**4145**	***2165***
mt/mt+wt/wt	**1788+2328=4116**	**874+1335=2209**

$$OR = \frac{\text{" probability" a subject of being with CAD when mtwt}}{\text{" probability" a subject of being with CAD when wtwt} + \text{mtmt}} = \frac{4145/2165}{4116/2209} = 1.03$$

$$95\%CI = \left(e^{\ln(OR) - 1.96 * \sqrt{\frac{1}{4145}+\frac{1}{4116}+\frac{1}{2165}+\frac{1}{2209}}}, e^{\ln(OR) + 1.96 * \sqrt{\frac{1}{4145}+\frac{1}{4116}+\frac{1}{2165}+\frac{1}{2209}}}\right) =$$

$$= (0.96, 1.11)$$

$$OR = 1.03$$

$$95\%CI = (0.96, 1.11)$$

Since “1” is included in the 95% CI, we conclude that the OR is not significant (P≥0.05).

- **Recessive model: *OR*=1.11 (1.01-1.21), significant**

  Homozygous for the *mt* allele have greater risk than wt-carriers

- **Dominant model: *OR*=1.12 (1.03-1.21), significant**

  Carriers of the *mt* allele have greater risk than non-carriers

- **Additive model: *OR*=1.17 (1.06-1.30), significant**

  Homozygous for the *mt* allele have greater risk than homozygous for the *wt* allele

- **Co-dominant model: *OR*=1.03 (0.96-1.11), non-sign**

Practice

The distributions of genotypes of two MyD88 gene variants C938A and C1944G among cases with Hodgkin's lymphoma and healthy control subjects are the following.

Variant	Genotype	Cases N	Healthy controls N
MyD88 C938A	CC	46	74
	CA	50	15
	AA	5	3
MyD88 C1944G	CC	77	65
	CG	21	25
	GG	3	3

Explore the significance of possible genetic models and identify the mode of inheritance.

The degree of dominance index (h-index)

In our previous example, three different genetic models were significant:

recessive,
dominant and
additive

These models are not independent and the right choice of the genetic model (or mode of inheritance) is a difficult task.

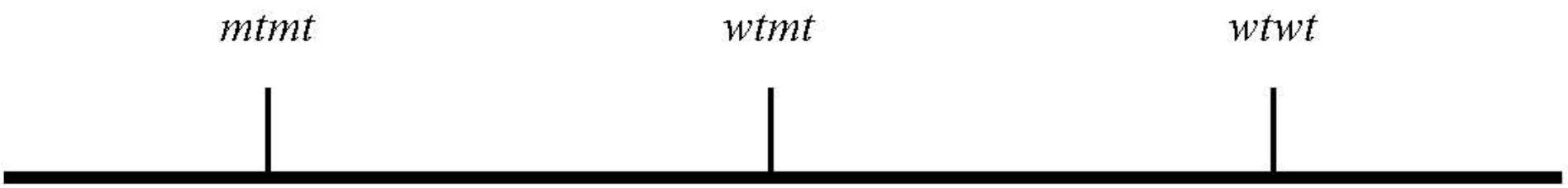
The identification of right genetic model is hard and the confusion cannot be avoided.

Also, the current genetic models do not quantify the mode of inheritance.

An alternative way to assess the mode of inheritance is to estimate the degree of dominance index (h-index).

The degree of dominance index answers the following question:

Where an heterozygous subject *wtmt* "lies" considering that an homozygous subject *mtmt* has the maximum susceptibility of being diseased and an homozygous subject *wtwt* has the least?

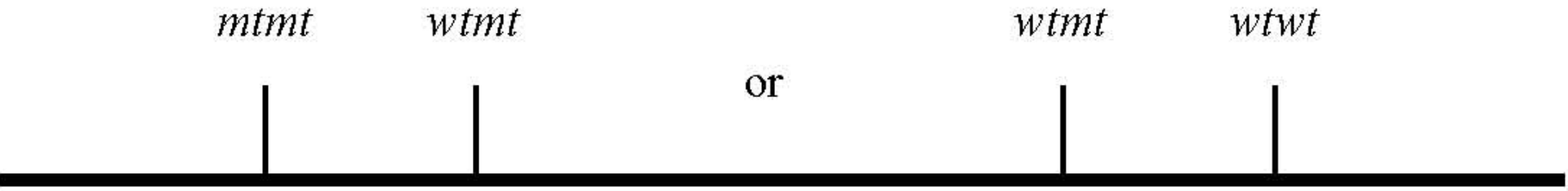


Risk for disease

(maximum risk of disease)

Protection for disease

(least risk of disease)



Maximum risk for disease

Least risk for disease

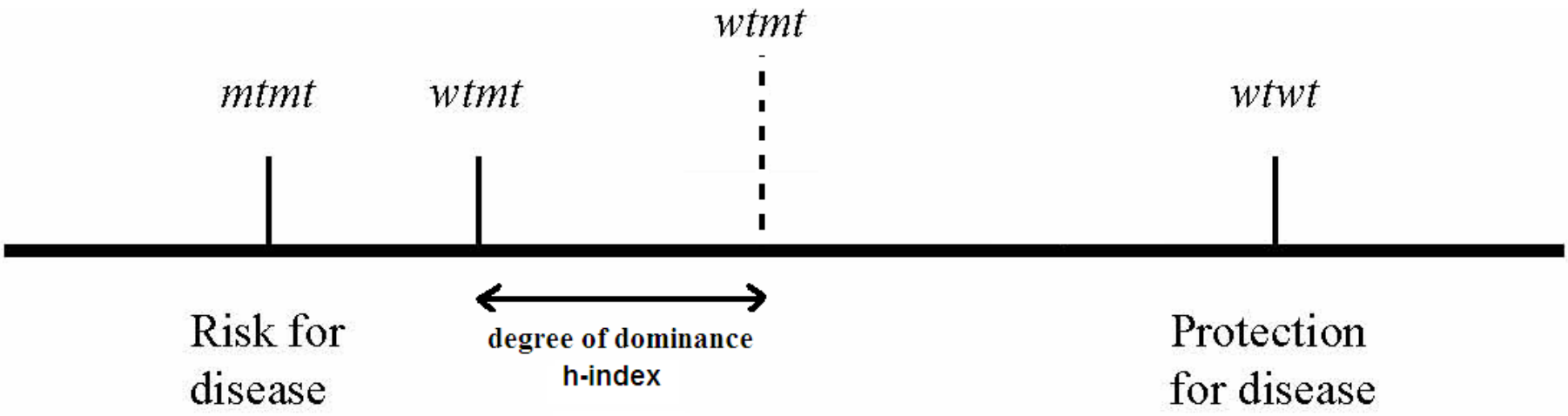
Then, the degree of dominance shows the "location" of *wtmt*, i.e. the mode of inheritance

Degree of dominance

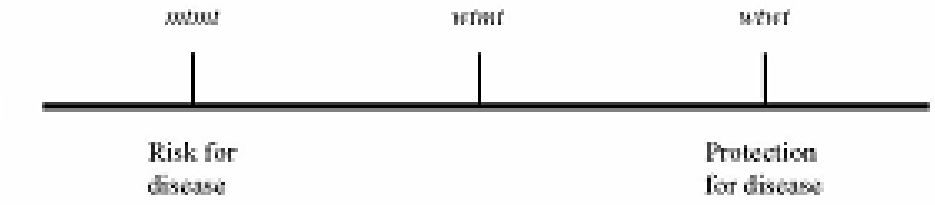
The degree of dominance could be derived from the ratio of the logarithms of the *OR* of co-dominant vs. the *OR* of the additive model:

$$h = \frac{\ln\left(OR_{co-dominant}\right)}{\left|\ln\left(OR_{additive}\right)\right|}$$

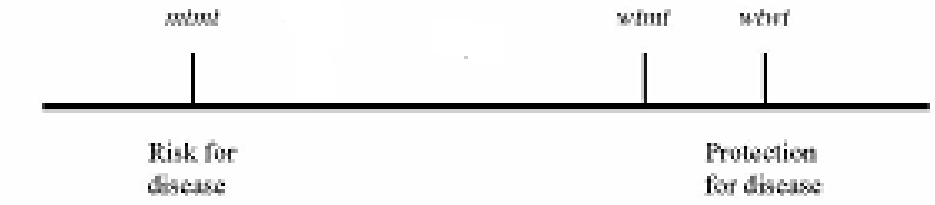
The h-index shows how much the heterozygous wtmt deviate from the middle of mtmt and wtwt



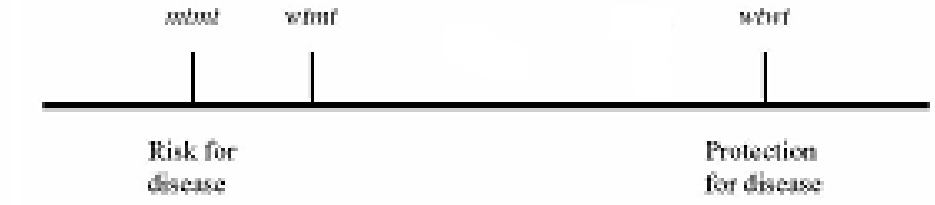
The different scenarios for assessing the genetic model based on the degree of dominance h-index are as follows:



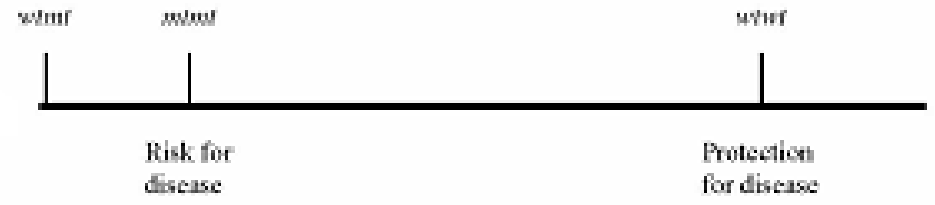
co-dominance
($h = 0$)



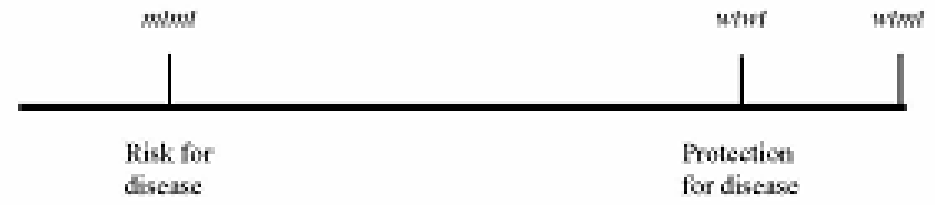
dominance
($-1 \leq h < 0$)



dominance
($0 < h \leq 1$)



over-dominance
($h > 1$)



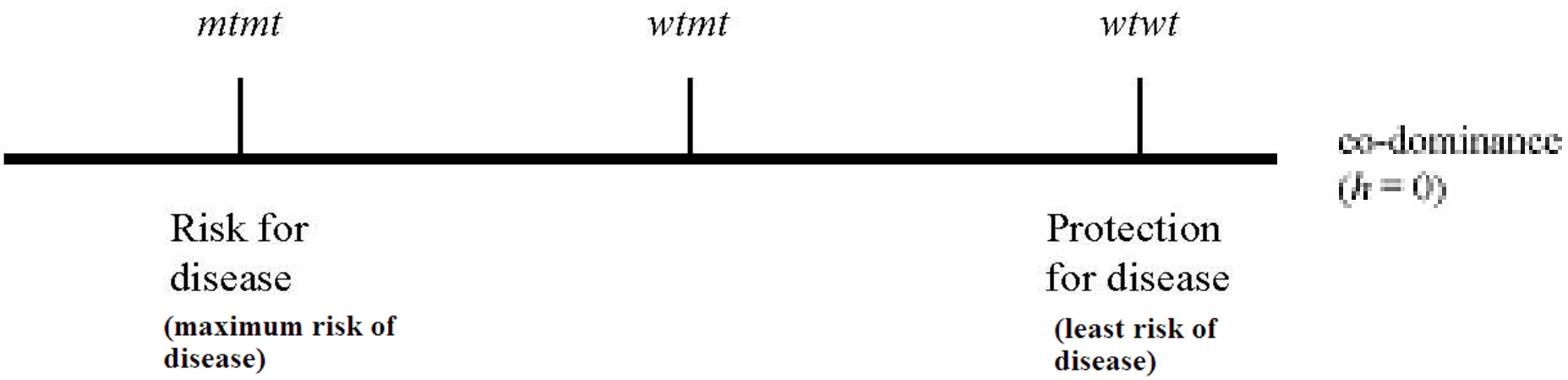
under-dominance
($h < -1$)

$$h = \frac{\ln\left(OR_{co-dominant}\right)}{\left|\ln\left(OR_{additive}\right)\right|}$$

When the co-dominant model is non-significant (i.e. $OR_{co}=1$ or $\ln(OR_{co})=0$) and the additive model is significant, the risk of disease for the heterozygote is in the middle of the two homozygotes.

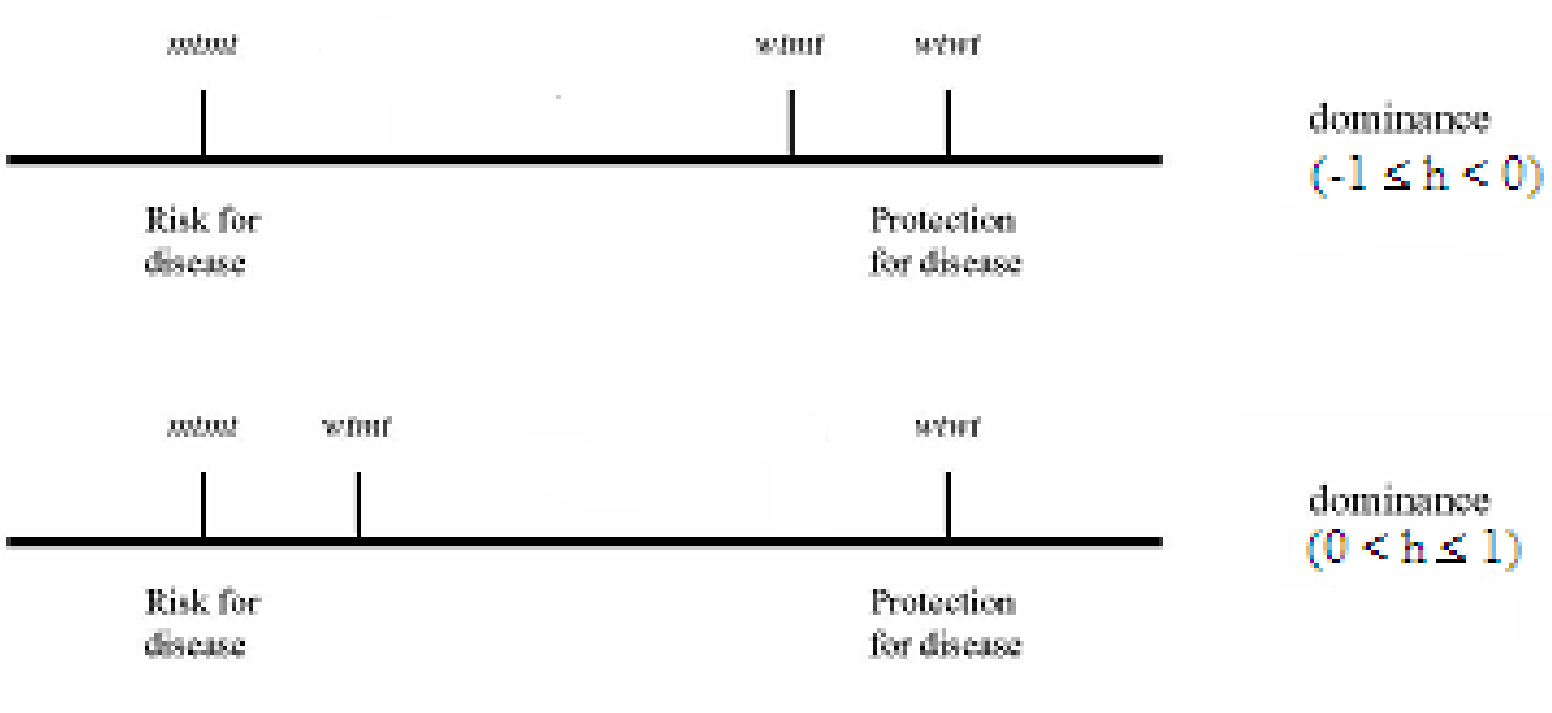
Then, it is assumed that *h*=0.

When *h*=0, we argue that "co-dominance" (or "additiviness") exists.



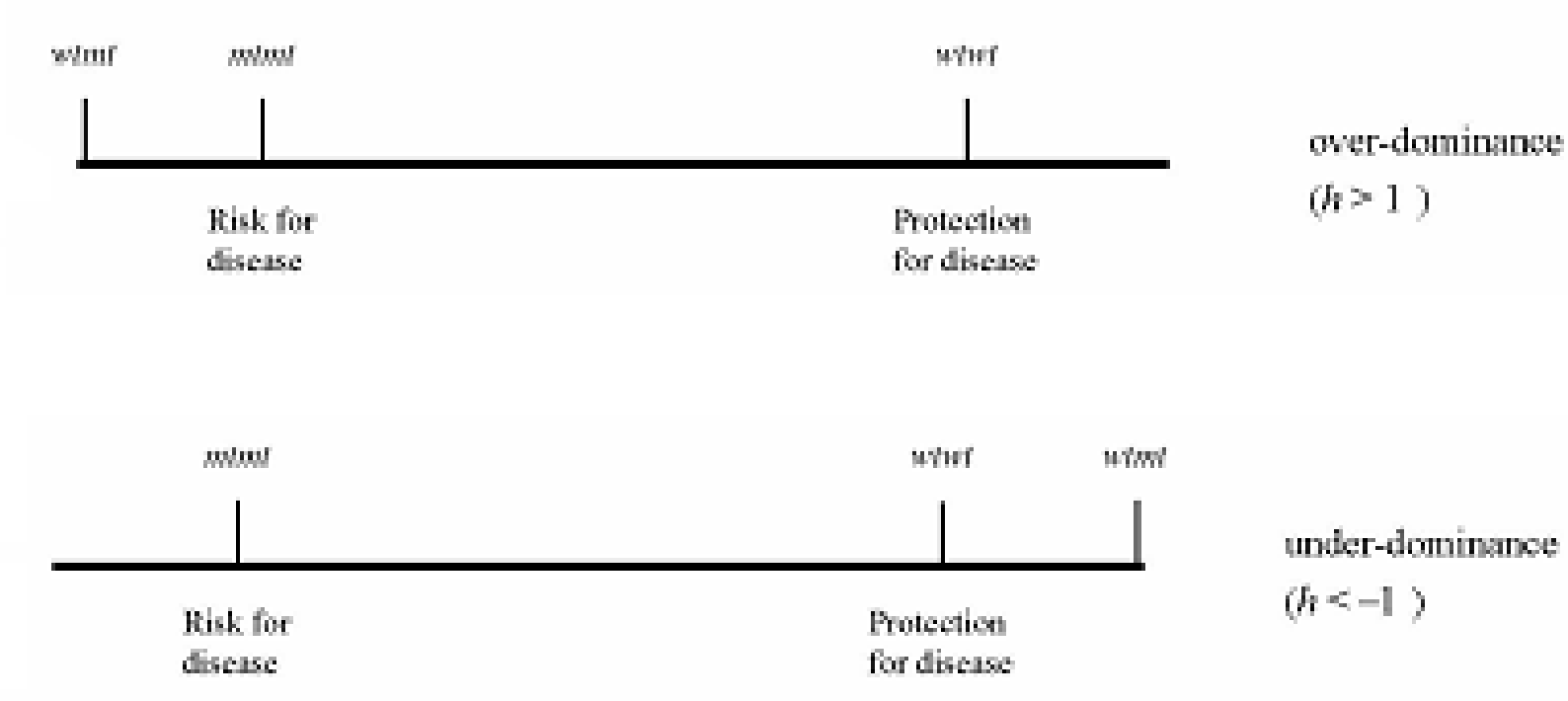
-1≤*h*<0: *wtmt* is expected to have a risk of being diseased somewhere in between the middle of the two homozygotes and towards to *wtwt*

0<*h* ≤1: *wtmt* is expected to have a risk of being diseased somewhere in between the middle of the two homozygotes and towards to *mtmt*

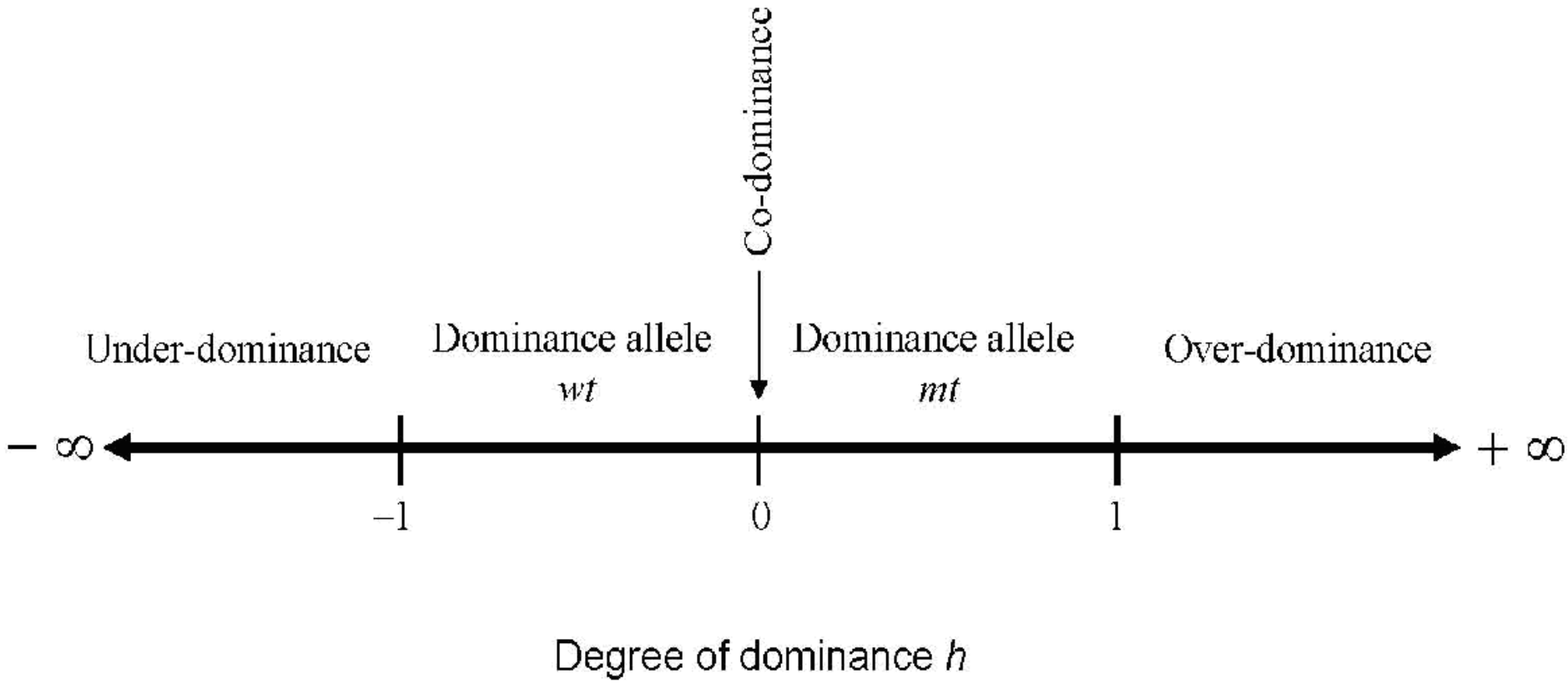


h*>1: *wtmt* has a higher risk of being diseased than *mtmt

h*<-1: *wtmt* has least chance of being diseased than *wtwt



Alternatively, we could talk for the dominance of the wt or mt allele as follows:



In the ACE D/I (wt=D/mt=I) vs. CAD example:

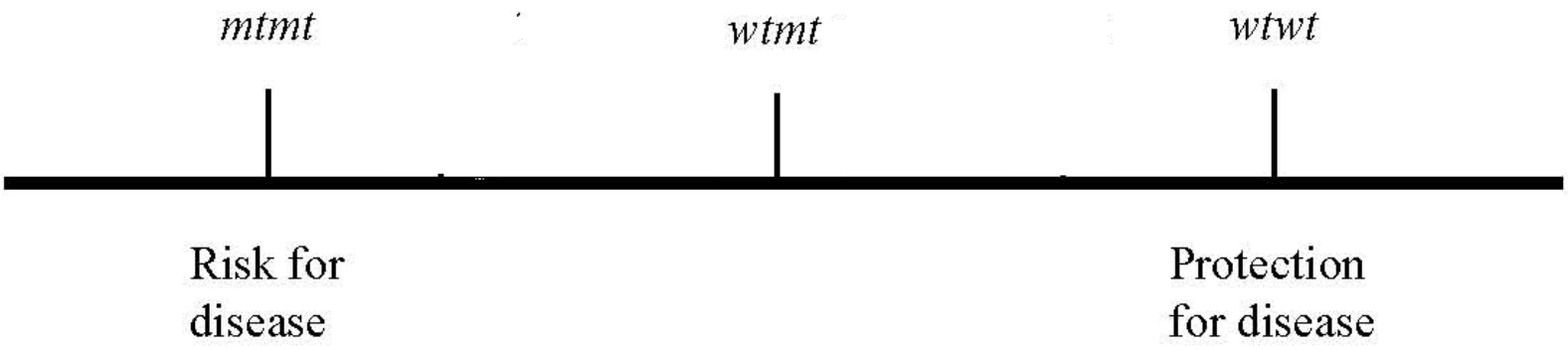
Genotype	Cases with CAD	Controls
mt/mt	1788	*874*
mt/wt	4145	*2165*
wt/wt	2328	*1335*

The co-dominant model is not significant

The additive model is significant

Thus, h=0

***h*=0: the risk of disease for the heterozygote wtmt is in the middle of the two homozygotes.**



Practice

A GAS investigating the association between the variant ADH2 (that has two alleles **1-mt* and **2-wt)* with alcoholism produced the following genotype distributions:

Genotype	Controls	Cases
*2/*2	448	238
*2/*1	93	85
*1/*1	4	17

- **Calculate the ORs and the 95% CIs for the co-dominant and additive models.**
- **Then, calculate the *h*-index.**
- **Interpret the results**

ADH2 vs. alcoholism – Co-dominant model:

Genotype	Cases	Controls
*2/*1	85	93
*1/*1+*2/*2	255	452

The OR= (85/93)/(255/452)=1.62

The 95% CI is (1.16, 2.26)

Thus, the co-dominant model is significant (i.e. h≠0).

ADH2 vs. alcoholism – Additive model:

Genotype	Cases	Controls
*1/*1	17	4
*2/*2	238	448

The OR= (17/4)/(238/448)=8.00

The 95% CI is (2.66, 24.05)

Thus, the additive model is significant.

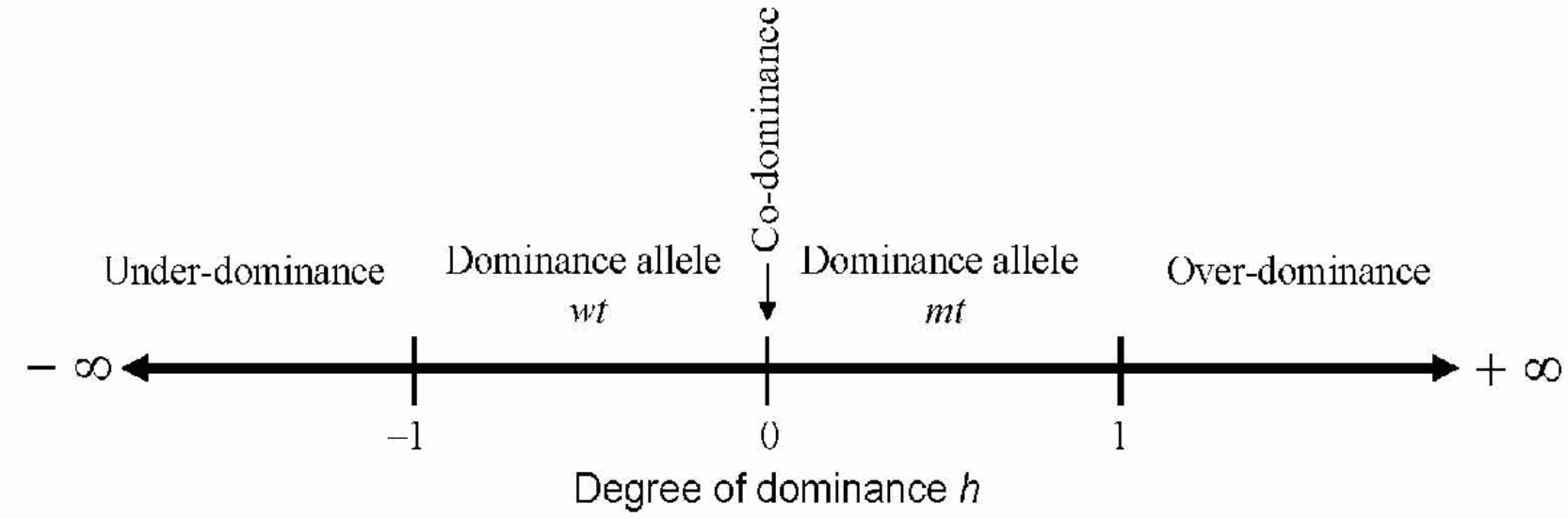
ADH2 vs. alcoholism:

Genotype	Controls	Cases
*2/*2	448	238
*2/*1	93	85
*1/*1	4	17

The degree of dominance is

$$h = \frac{\ln\left(OR_{co-dominant}\right)}{\left|\ln\left(OR_{additive}\right)\right|} = \frac{\ln(1.62)}{\left|\ln(8.00)\right|} = 0.23$$

Since h≠0 and 0<h ≤1 (h=0.23), there is indication that dominance exists for allele *1 (mt allele).



In other words, the homozygous *1/*1 (mt/mt) has a greater risk of being alcoholic than the homozygous *2/*2 (wt/wt), and the heterozygote *2/*1 has a risk of alcoholism closer to the homozygote *1/*1 than to the midpoint between the two homozygotes.

A genetic model-free approach

A genetic model-free approach for testing the association between disease status (disease vs. healthy) and genotype is the generalized odds ratio (OR_G).

The OR_G is a single statistic that utilizes the complete genotype distribution (not merging genotypes like in the co-dominant model) and provides an estimate of the overall risk effect.

Definition of OR_G

***OR_G* is the probability of a subject being diseased relative to probability of being free of disease, given that the diseased subject has a higher mutational load than the non-diseased.**

$$\mathrm{OR_G} = \frac{\text{Probability being diseased, diseased has high mutational load}}{\text{Probability of being non-diseased, non-diseased has low mutational load}}$$

When *OR_G*>1 then an increased genetic exposure (mutational load) implies disease.

Mathematical definition

In mathematical terms, the generalized odds ratio is defined as:

$$OR_G = \frac{P(D_i, D_j^{'} / i > j)}{P(D_i^{'}, D_j / i > j)}$$

where *i*=1-3 (*i*=1 for wt/wt, *i*=2 for wt/mt and *i*=3 for mt/mt) when *D* denote diseased (cases) and D' the non-diseased (control).

The SE of ln(OR_G) is given by:

$$se\left[\ln\left(OR_G\right)\right] = \frac{1}{PQ}\left(\sqrt{\sum_i \sum_s n_{is}\left(2Q(A_{is}+D_{is}) - 2P(B_{is}+C_{is})\right)^2}\right)$$

$$P = \sum_{i<j}\sum_{s<t} n_{is}n_{jt} \quad Q = \sum_{i<j}\sum_{s>t} n_{is}n_{jt} \quad A_{is} = \sum_{j<i}\sum_{t<s} n_{jt} \quad D_{is} = \sum_{j>i}\sum_{t>s} n_{jt} \quad B_{is} = \sum_{j<i}\sum_{t>s} n_{jt} \quad C_{is} = \sum_{j>i}\sum_{t<s} n_{jt}$$

ORGGASMA

"ORGGASMA": a software for implementing the generalized odds ratio methodology for the analysis and meta-analysis of GAS.

The software "ORGGASMA" (together with instructions how to operate it) is freely available and it can be downloaded form the web site http://biomath.med.uth.gr

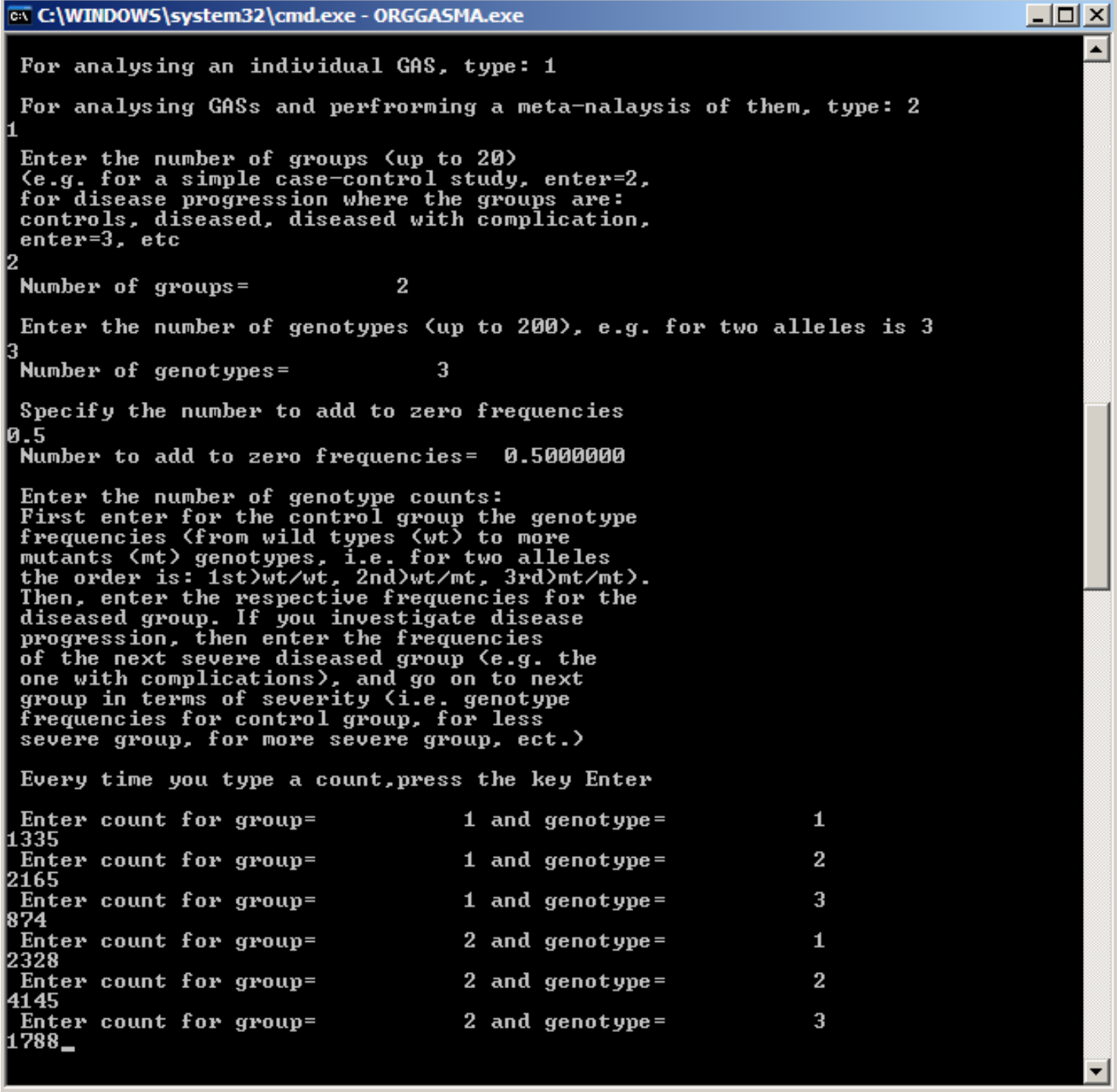
Download the "ORGGASMA" software and operated it only for the "cmd" command of windows (do not double click the icon).

Example: ACE D/I (wt=D/mt=I) vs. CAD

Genotype	Cases with CAD	Controls
mt/mt	1788	*874*
mt/wt	4145	*2165*
wt/wt	2328	*1335*

Assumption: Subjects who are homozygous for *I* allele (I/I) have the highest mutational load, those homozygous for *D* allele (D/D) have the lowest, and heterozygous (D/I) have an intermediate level.

In ORGGASMA, the data are entered as follows:



The ORGGASMA software for ACE (D/I) vs. CAD study produces the following results:

***OR*$_G$=1.102 with 95% CI: (1.04-1.17)**

Since 1 is not included in the 95% CI the OR$_G$ is significant.

OR_G=1.102

The interpretation of the finding is as follows:

For any two subjects, diseased with CAD and healthy, the probability of being diseased is 10% higher (relative to the probability of being non-diseased) given that the diseased subject has higher mutational load for the variant ACE (I/D) than the healthy one.

Practice

A GAS investigating the association between the alleles *ADH2*1* (mt) and *ADH2*2* (wt) with alcoholism produced the following genotype distributions:

Genotype	Controls	Cases
*2/*2	448	238
*2/*1	93	85
*1/*1	4	17

Calculate the OR_G and the respective 95% CI. Then, interpret the results.

The ORGGASMA software for ADH2 (*2/*1) vs. alcoholism study produces the following results:

OR_G=2.02 with 95% CI: (1.47-2.76)

The interpretation of the finding is as follows:

For any two subjects, diseased and healthy, the probability of being alcoholic is 2-fold higher (relative to the probability of being non-alcoholic) given that the alcoholic subject has higher mutational load for the variant ADH2 (*2/*1) than the non-alcoholic.

Analysis of GAS for bi-allelic variant and disease progression

(healthy/diseased/diseased with complications)

The association between a variant and disease progression is examined using the Generalized Odds Ratio (ORG).

The *ORG* expresses the probability of a subject being more diseased (disease progression) relative to the probability of being less diseased, given that a more diseased subject has a higher mutational load.

$$OR_G = \frac{\text{Probability of being more diseased with higher mutational load}}{\text{Probability of being less diseased with lower mutational load}}$$

Example

A genetic association study (GAS) was conducted to investigate the association between five *AKR1B1* gene variants (rs2259458 G/T) and diabetes progression.

The cohort consisted of 169 diabetic cases with microvascular complications, 107 diseased controls (diabetics without microvascular complications) and 315 healthy controls. The genotype distribution was the following:

Variant	Genotype	Cases	Diseased Controls	Healthy controls
rs2259458 G/T	TT	54	33	53
	GT	75	65	138
	GG	36	43	115

Is the variant associated with disease progression?

Study quality assessment

- **Prior to testing the association between the variant and disease progression, the quality of the study will be assessed by testing the controls for Hardy-Weinberg Equilibrium (HWE).**
- **A calculator for testing for HWE is proving at the following URL http://www.had2know.com/academics/hardy-weinberg-equilibrium-calculator-2-alleles.html**
- **The P-value for the HWE testing is P=0.30 (P≥0.05).**
- **Thus, the controls are in HWE.**

Assessing the association using the OR_G

In order to testing the association between the variant and disease progression, we applied the ORGGASMA software (http://biomath.med.uth.gr).

Then, the following results were produced:

OR_G=1.65 with 95% CI (1.32, 2.08)

Since “1” is included in the 95% CI, the ORG is significant (P<0.05).

ORG=1.65 with 95% CI (1.32, 2.08)

Thus, the variant rs2259458 G/T is associated to disease severity and the risk of disease progression is related to mutational load of the variant.

Alternatively, a subject has 65% higher risk of disease progression (i.e. of being more diseased) relative to the risk of no progression (i.e. of being less diseased) given that the subject with disease progression has a higher mutational load than the subject without disease progression.

Practice

A GAS was conducted to examining the association between the His159Tyr mutation of the BAFF-R gene and the progression of Sjogren's Syndrome (SS) which leads to lymphoma.

For this purpose, three groups of subjects were genotyped: healthy controls, diseased controls (SS without lymphoma) and cases (SS with lymphoma). The genotypic data were as follows:

	Baff-R mt	
Disease progression	not	yes
Healthy controls	177	3
Diseased controls (SSnoL)	166	11
Cases (SSL)	64	6

Is the BAFF-R mutation associated with SS progression?

OR_G=2.75 (1.36-5.57)

A subject has almost a 3-fold higher risk of disease progression (relative to the risk of not progressing) given that the subject is a mutant-carrier.

Analysis of GAS for multi-allelic variant and binary outcome (healthy/diseased)

The association between the genotype distribution of a multi-allelic variant and disease development is tested using a x^2-test.

The magnitude of association for a specific genotypic contrast can be expressed in terms of odds ratio (OR).

Example

A GAS for investigating the association between apoE (apolipoprotein E) genotype and colorectal cancer (CRC) was contacted.

The apoE appears with three alleles: e2, e3, e4. The derived genotype distribution was the following:

apoE	Cases (CRC)	Controls (healthy)
e3/e3	894	930
e3/e2	206	962
e3/e4	361	242
e2/e2	11	34
e2/e4	44	4
e4/e4	35	66

Is the apoE variant associate with CRC development?

Study quality assessment

Prior to testing the association between the variant and disease progression, the quality of the study will be assessed by testing the controls for HWE.

A calculator for testing for HWE is proving at the following URL
http://www.had2know.com/academics/hardy-weinberg-equilibrium-calculator-3-alleles.html

The P-value for the HWE testing is P=0.10 (P≥0.05).

Thus, the controls are in HWE.

Testing the association

The association between the genotype distribution and disease susceptibility is tested using a x^2-test with (6-1)x(2-1)=5 df.

To perform a chi-squared test, you may use the following URL:
http://www.quantpsy.org/chisq/chisq.htm

Testing the association

The x^2-test is 458.64 which is greater than 11.1, the 5% point of the x^2-distribution with 5 df.

Then, P<0.05 (P<0.0001).

Thus, there is significant association between apoE gene variant and development of CRC.

Genetic contrast

We may investigate whether specific genetic contrasts are associated with CRC development.

One contrast is the apoE e4-carrier vs. CRC development.

Then, we can examine whether an apoE e4-carrier has a greater risk in developing CRC.

In order to examining the contrast apoE e4-carrier vs. CRC, the genotypic data are merged as follows:

apoE	Cases (CRC)	Controls (healthy)
e4-carriers (e3/e4+e2/e4+e4/e4)	440	332
non-e4-carriers (e3/e3+e3/e2+e2/e2)	1111	1926

The association between e4-carriers and CRC development can be tested using a x2-test with 1 df (http://www.quantpsy.org/chisq/chisq.htm).

Then, the P-value of the x^2-test is P<0.05.

Thus, e4-carriers are associated with CRC development.

Magnitude of association

apoE	Cases (CRC)	Controls (healthy)
e4-carriers (e3/e4+e2/e4+e4/e4)	440	332
non-e4-carriers (e3/e3+e3/e2+e2/e2)	1111	1926

Then, the magnitude of association for e4-carriers relative to non-e4-carriers in developing CRC can be expressed using the following OR:

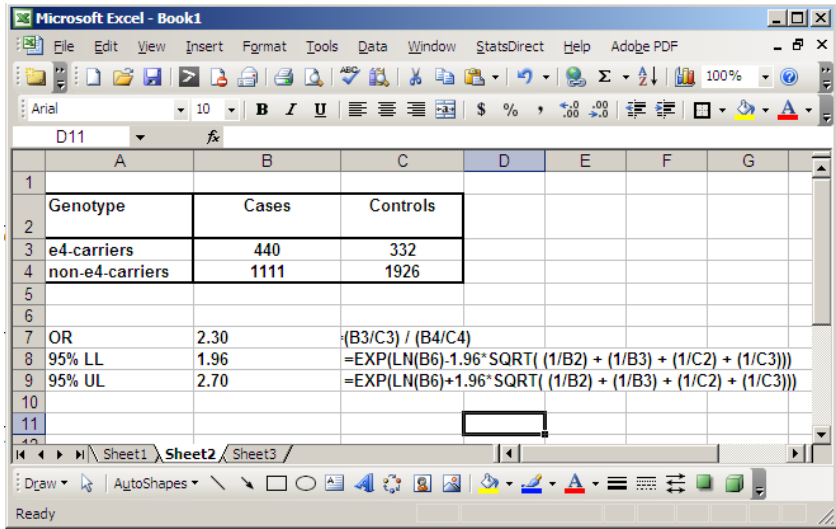
$$OR = \frac{"\ probability"\ of\ developing\ CRC\ when\ e4-carrier}{"\ probability"\ of\ developing\ CRC\ when\ non-e4-carrier}$$

apoE	Cases (CRC)	Controls (healthy)
e4-carriers (e3/e4+e2/e4+e4/e4)	440	332
non-e4-carriers (e3/e3+e3/e2+e2/e2)	1111	1926

$$OR = \frac{\text{" probability" of developing CRC when } e4-\text{carrier}}{\text{" probability" of developing CRC when non}-e4-\text{carrier}} = \frac{440/332}{1111/1926} = 2.30$$

$$95\%CI = \left(e^{\ln(OR)-1.96*\sqrt{\frac{1}{440}+\frac{1}{1111}+\frac{1}{332}+\frac{1}{1926}}}, e^{\ln(OR)+1.96*\sqrt{\frac{1}{440}+\frac{1}{1111}+\frac{1}{332}+\frac{1}{1926}}}\right) =$$

$$= (1.96, 2.70)$$



apoE	Cases (CRC)	Controls (healthy)
e4-carriers (e3/e4+e2/e4+e4/e4)	440	332
non-e4-carriers (e3/e3+e3/e2+e2/e2)	1111	1926

OR= 2.30 and 95% CI is (1.96, 2.70)

Since "1" is not included in the 95% CI, we conclude that the OR is significant (P<0.05).

Thus, apoE e4-carriers have 2 times greater risk for developing CRC than non-e4-carriers.

Pharmacogenetic (PG) Studies

In therapeutics (and personalized medicine), some patients respond to treatment and other patients not.

Therefore, it is believed that certain genes play role in the response to therapy.

PG studies investigate the association between a gene (variant or expression) and the response to therapy.

In this course we will examine the following cases of PG studies:

- **PG studies with gene expression as a binary variable and binary response to therapy (response/no response)**
- **PG studies with gene expression as a continuous variable and binary response to therapy (response/no response)**
- **PG studies with gene polymorphism and binary response to therapy (response/no response)**

PG studies with gene expression as a continuous variable and binary response to therapy (response/no response)

Example

MDR1 gene overexpression is considered to be a major cause of multidrug resistance and it is implicated in the response to chemotherapy in AML patients.

In a PG study, the association of MDR1 gene expression and response to chemotherapy in patients with AML has been investigated. The results were as follows:

	mRNA expression levels mean±SD
Responders (N=37):	**1.4±2.7**
Non-responders (N=15):	**0.3±0.5**

Is the response to treatment associated with MDR1 mRNA expression levels?

We can test whether the outcome of response is related to gene expression levels by testing the significance of the differences in average gene expression between responders and non-responders.

We can simply test the difference using a t-test.

We may use the online t-test at the URL

http://www.quantitativeskills.com/sisa/statistics/t-test.htm

for testing the equality of the two means.

If we apply the t-test, the P-value is P=0.057.

Thus, there is indication that MDR1 expression levels are (marginally) significant different between Responders and Non-responders.

Then, MDR1 may be implicated in response to chemotherapy in AML patients

PG studies with gene expression as a binary variable and binary response to therapy (response/no response)

Example

In a PG study, the association of MDR1 gene expression and response to chemotherapy in patients with AML has been investigated. The results were as follows:

	mRNA expression levels	
	+ve	*-ve*
Responders	**27**	**22**
Non-responders	**21**	**5**

Is the response to treatment associated with MDR1 mRNA expression levels?

We may examine the association between response to treatment and MDR1 mRNA expression levels using a x^2-test with 1 df.

To perform the chi-squared test, use the following URL:
http://www.quantpsy.org/chisq/chisq.htm

The P-value is P<0.05 (P=0.027).

Thus, there is significant association between response to treatment and MDR1 mRNA expression levels.

The magnitude of association

Once the association is significant, we can estimate the magnitude of association by calculating the OR and the respective 95% CI

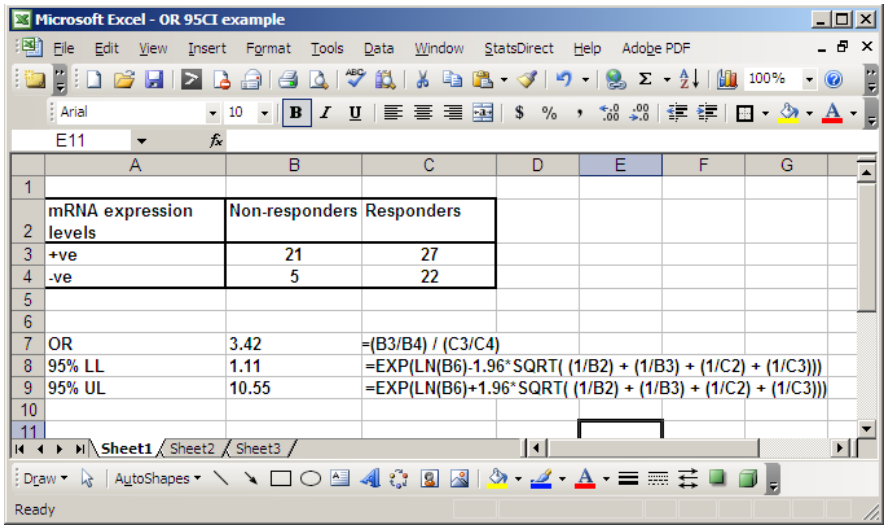
mRNA expression levels	Non-responders	Responders
+ve	21	27
-ve	5	22

The magnitude of association (OR) is as follows:

$$OR = \frac{\text{"probability" of being non-responder when + ve mRNA expression}}{\text{"probability" of being responders when + ve mRNA expression}} = \frac{21/5}{27/22} = 3.42$$

$$95\%\,CI = \left(e^{\ln(OR)-1.96*\sqrt{\frac{1}{27}+\frac{1}{21}+\frac{1}{22}+\frac{1}{5}}},\ e^{\ln(OR)+1.96*\sqrt{\frac{1}{27}+\frac{1}{21}+\frac{1}{22}+\frac{1}{5}}}\right) =$$

$$= (1.11, 10.56)$$



mRNA expression levels	Non-responders	Responders
+ve	21	27
-ve	5	22

$$OR = 3.42 \qquad 95\%\ CI = (1.11, 10.56)$$

An OR=3.42 implies that a patient has more than 3 times (3.4 times) probability for non-responding to chemotherapy when the mRNA expression is +ve.

Since the 95% CI does not include 1, the OR is significant (P<0.05)

PG studies with gene polymorphism and binary response to therapy (response/no response)

Example

The association of MDR1 C3435T gene polymorphism and response to chemotherapy in patients with AML has been investigated. The results were as follows:

	MDR1 C3435T genotype		
	CC	***CT***	***TT***
Responders	**158**	**65**	**39**
Non-responders	**13**	**18**	**9**

Is the response to treatment associated with the MDR1 C3435T gene polymorphism?

	MDR1 C3435T genotype		
	CC	***CT***	***TT***
Responders	**158**	**65**	**39**
Non-responders	**13**	**18**	**9**

The association between response to treatment and MDR1 C3435T gene polymorphism is tested using a x^2-test with (3-1)*(2-1)=2 df.

To perform the chi-squared test, we may use the following URL:
http://www.quantpsy.org/chisq/chisq.htm

The P-value of the x^2-test is P<0.05 (P=0.004).

Thus, there is significant association between response to treatment and MDR1 C3435T gene polymorphism.

Genetic contrasts

Once the association between the genotype distribution and the response outcome is significant, we can examine the magnitude of association for various genetic contrasts (models):

- **Additive**
- **Co-dominant**
- **Recessive**
- **Dominant**

In the course, we will focus on the dominant contrast (the other contrasts can be examined in a similar way.

Dominant contrast

MDR1 C3435T	Non-Responders	Responders
T-carriers	27 (=18+9)	104 (=39+65)
Non-T-carriers	13	158

The significance of association for the dominant model can be examined using a x^2-test with 1 df.

To perform the chi-squared test, we may use the following URL: http://www.quantpsy.org/chisq/chisq.htm

Dominant contrast

MDR1 C3435T	Non-Responders	Responders
T-carriers	27 (=18+9)	104 (=39+65)
Non-T-carriers	13	158

The P-value of the x^2-test is $P<0.05$ ($P=0.004$).

Thus, there is significant association between response to treatment and MDR1 C3435T gene polymorphism for the Dominant contrast.

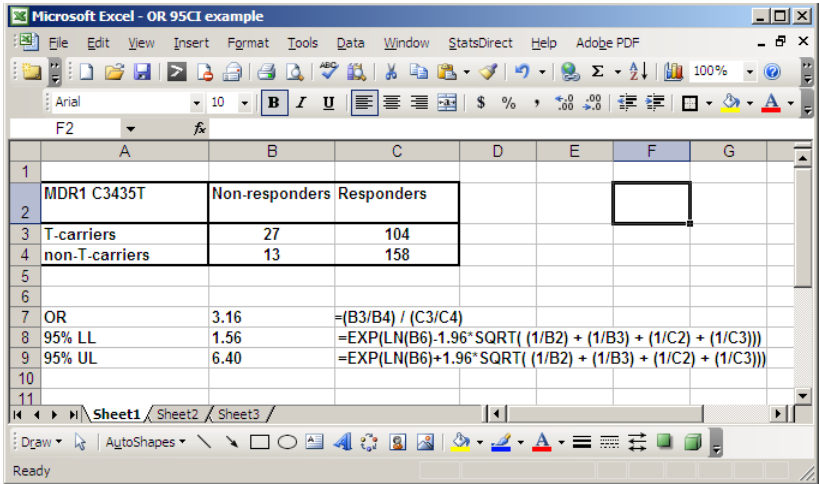
MDR1 C3435T	Non-Responders	Responders
T-carriers	27 (=18+9)	104 (=39+65)
Non-T-carriers	13	158

The magnitude of association can be assessed using the OR as follows:

$$OR = \frac{\text{"probability" of being non-responder when T-carrier}}{\text{"probability" of being responder when T-carrier}} = \frac{27/13}{104/158} = 3.16$$

$$OR\ 95\%\ CI = \left(e^{\ln(OR)-1.96*\sqrt{\frac{1}{27}+\frac{1}{13}+\frac{1}{104}+\frac{1}{158}}},\ e^{\ln(OR)+1.96*\sqrt{\frac{1}{27}+\frac{1}{13}+\frac{1}{104}+\frac{1}{158}}}\right) =$$

$$= (1.56, 6.40)$$



OR=3.16 95% CI (1.56, 6.40)

An OR=3.16 implies that a T-carrier patient has 3 times more probability for non-responding to chemotherapy than responding.

Since the 95% CI does not include 1, the OR is significant (P<0.05)

PG studies with gene polymorphism and binary response to therapy (response/no response)

Model-free approach

A genetic model-free approach for testing the association between a gene polymorphism (variant) and a binary response to therapy (response/no response) is the generalized odds ratio (OR_G).

The OR_G is a single statistic that utilizes the complete genotype distribution (not merging genotypes like in the dominant model) and provides an estimate of the overall risk of response according to the mutational load of the variant.

Definition of OR_G

***OR_G* is the probability of a subject being non-responder relative to probability of being responder, given that the non-responder subject has a higher mutational load than the responder**

$$\mathrm{OR_G} = \frac{\text{Probability of being non-responder, non-responder has high mutational load}}{\text{Probability of being responder, responder has low mutational load}}$$

When *OR_G*>1 then an increased genetic exposure (mutational load) implies non-responsiveness to treatment.

“ORGGASMA”: a software for implementing the generalized odds ratio methodology for the analysis and meta-analysis of PG studies.

Now, the “diseased” are replaced by the “non-responders” and the “controls” by the “responders”.

The software “ORGGASMA” (together with instructions how to operate it) is freely available and it can be downloaded form the web site http://biomath.med.uth.gr

Download the “ORGGASMA” software and operated it only for the “cmd” command of windows (do not double click the icon)

In our example, the data are as follows:

	MDR1 C3435T genotype		
	CC	***CT***	***TT***
Responders	**158**	**65**	**39**
Non-responders	**13**	**18**	**9**

OR_G=2.36 with 95% CI: (1.40-3.98)

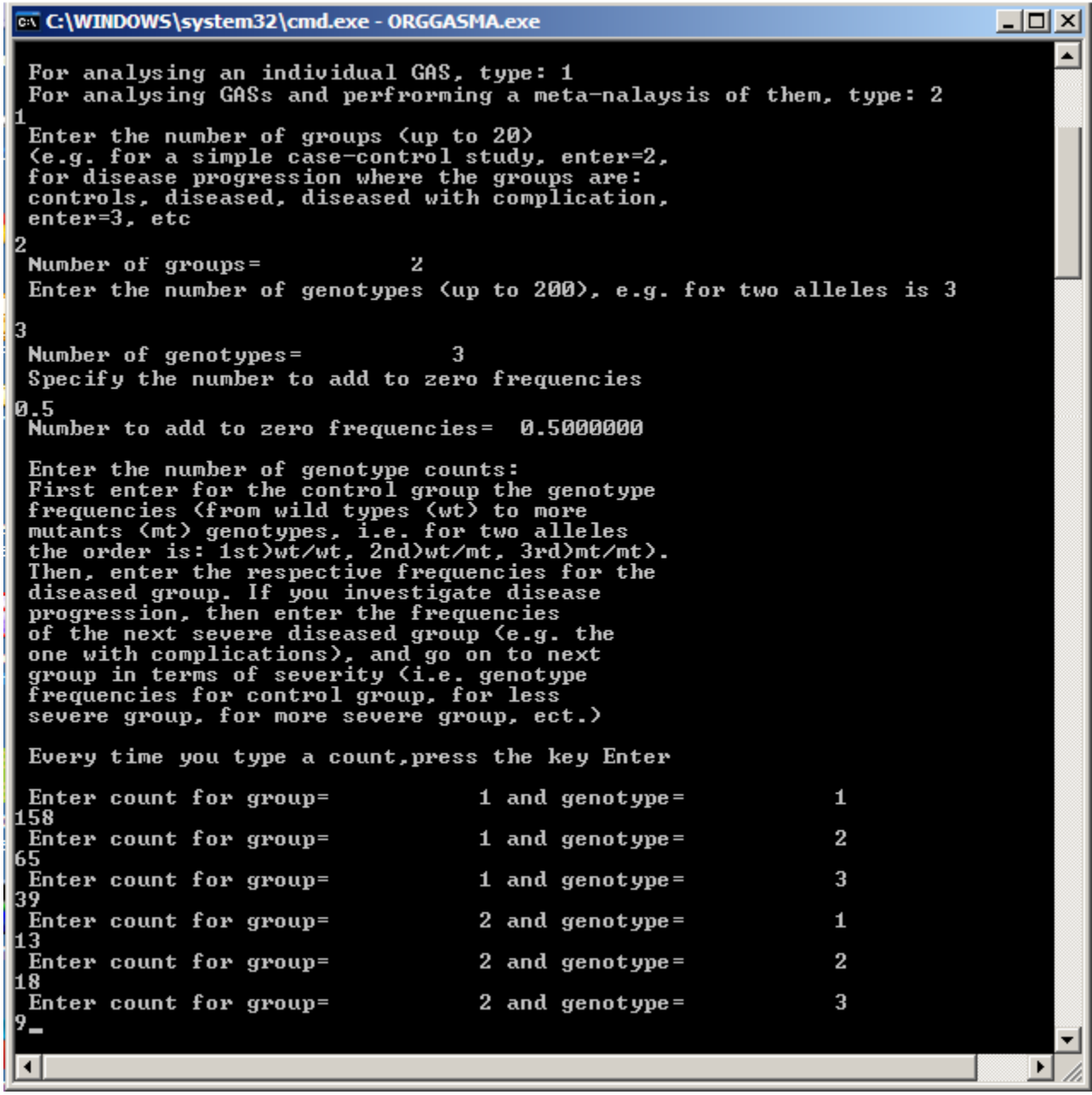
There is 2-fold probability of being non-responder relative to probability of being responder, given that the non-responder has higher mutational load than the responder.

Thus, the risk of non-responsiveness is proportional to the increased genetic exposure.

In our example, the data are as follows:

	MDR1 C3435T genotype		
	CC	*CT*	*TT*
Responders	158	65	39
Non-responders	13	18	9

We may use ORGGASMA to derive the ORG as previously



Practice

Imatinib resistance is major cause of imatinib mesylate (IM) treatment failure in chronic myeloid leukemia (CML) patients. Prove whether the ABCG2 gene expression or the ABCG2 A/G variant is significantly associated with poor response to IM. The data is as follows:

Case 1:

	ABCG2 *mRNA expression levels*
	mean±SD
Responders (N=202):	4.72±2.8
Non-responders (N=67):	1.35±1.5

Case 2:

	ABCG2 *mRNA expression levels*	
	+ve	*-ve*
Responders	28	34
Non-responders	51	12

Case 3:

	ABCG2 A/G *genotype*		
	CC	*CT*	*TT*
Responders	99	144	123
Non-responders	113	108	59