

# Development of algorithms for the test of association between haplotypes and phenotypes using SNP data

Naoyuki Kamatani, M.D., Ph.D.

Division of Genomic Medicine, Department of Advanced Biomedical Engineering and Science, Institute of Rheumatology, Tokyo Women's Medical University

Algorithm Team, Genome Variation Model Project, JBIC  
Laboratory for Statistical Analysis, Group for Medical Informatics,  
RIKEN

# Difference in approaches between mathematical statistics and statistical genetics

## 1. Galton and Pearson's approach

Regression, Correlation

Truth only in mathematics but not in the real world

Model selection is the main approach

## 2. Fisher's approach

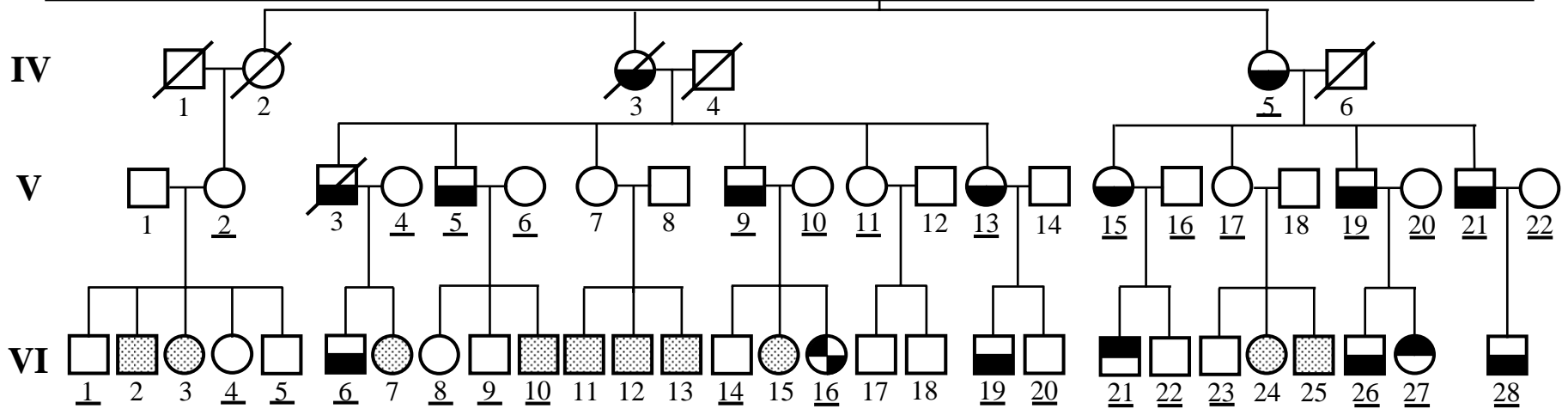
Variance-based, Maximum-likelihood

Truth not only in mathematics but also in the real world

Laws of inheritance that are expressed by probability functions are true.

Based on the data (familial relationship, genotypes, phenotypes), we can write the exact probability (or probability density) of the observed data only using laws of inheritance which are a set of probability functions.

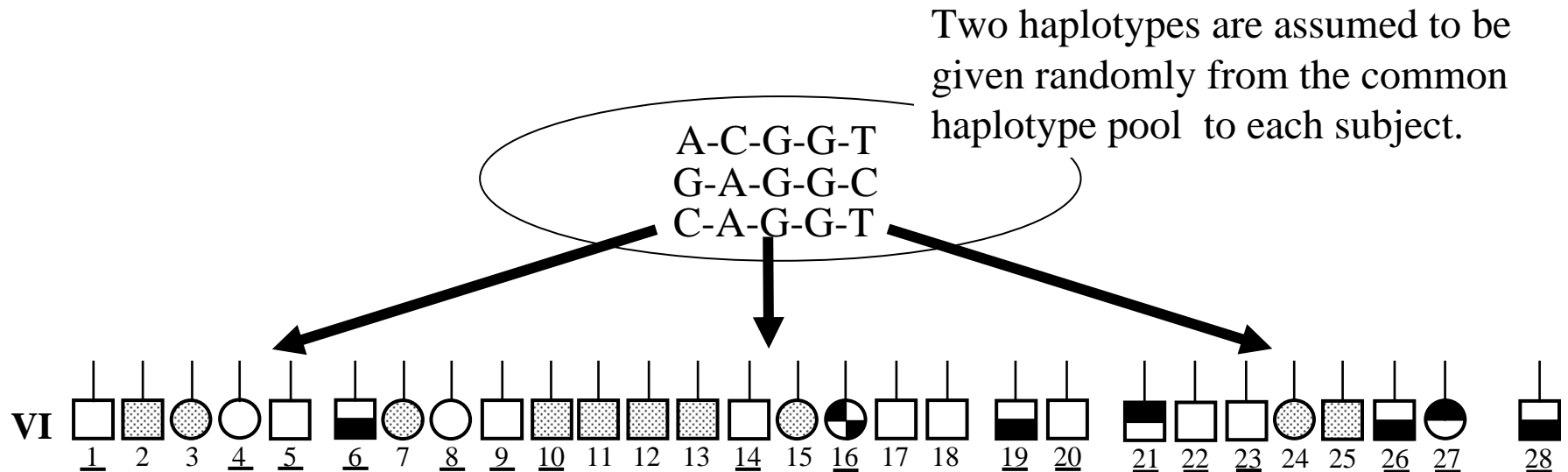
We can estimate by the maximum-likelihood method the parameters and test the hypothesis of the association between a phenotype and a locus → **Linkage analysis**.



However, when all the information is not available, we have to cope with the missing data problem

??????

Hardy-Weinberg's law is useful when information about the familial relationship is not available.



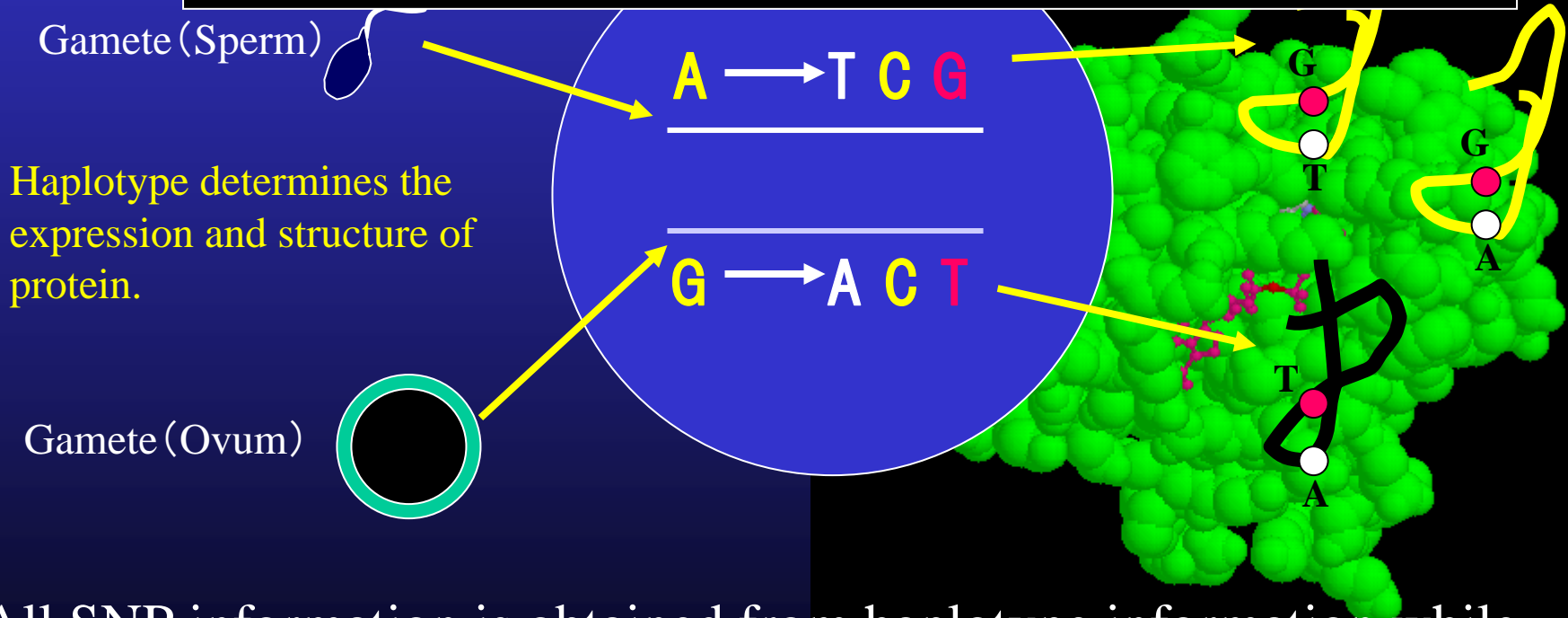
## A haplotype:

A list of alleles at linked loci derived from a parent

## A diplotype configuration:

Combination of two haplotypes in a subject

**Analysis based on haplotypes is necessary**



All SNP information is obtained from haplotype information while reverse is not true (Complete and incomplete information)

# Algorithms for haplotype analysis we constructed

## 1. Ldsupport (Kitamura et al. *Ann Hum Genet* 66: 183-193, 2002)

Inference of individual diplotype configurations

## 2. Ldpooled (Ito et al. *Am J Hum Genet* 72: 384-398, 2003)

Inference of haplotype frequencies using pooled DNA

## 3. Penhaplo (Ito et al. *Genetics* 168: 2339-2348, 2004)

Test of association between qualitative phenotype and diplotype configurations and inference of penetrances using the data from cohort, clinical trial and case-control studies.

## 4. QTLhaplo (Shibata et al. *Genetics* 168: 525-539, 2004)

Test of association between quantitative phenotypes and diplotype configurations and inference of parameters using the data from cohort, clinical trial and case-control studies.

In order to make targets of genetic information more flexible,  
we introduced a new

## Sample space based on haplotypes

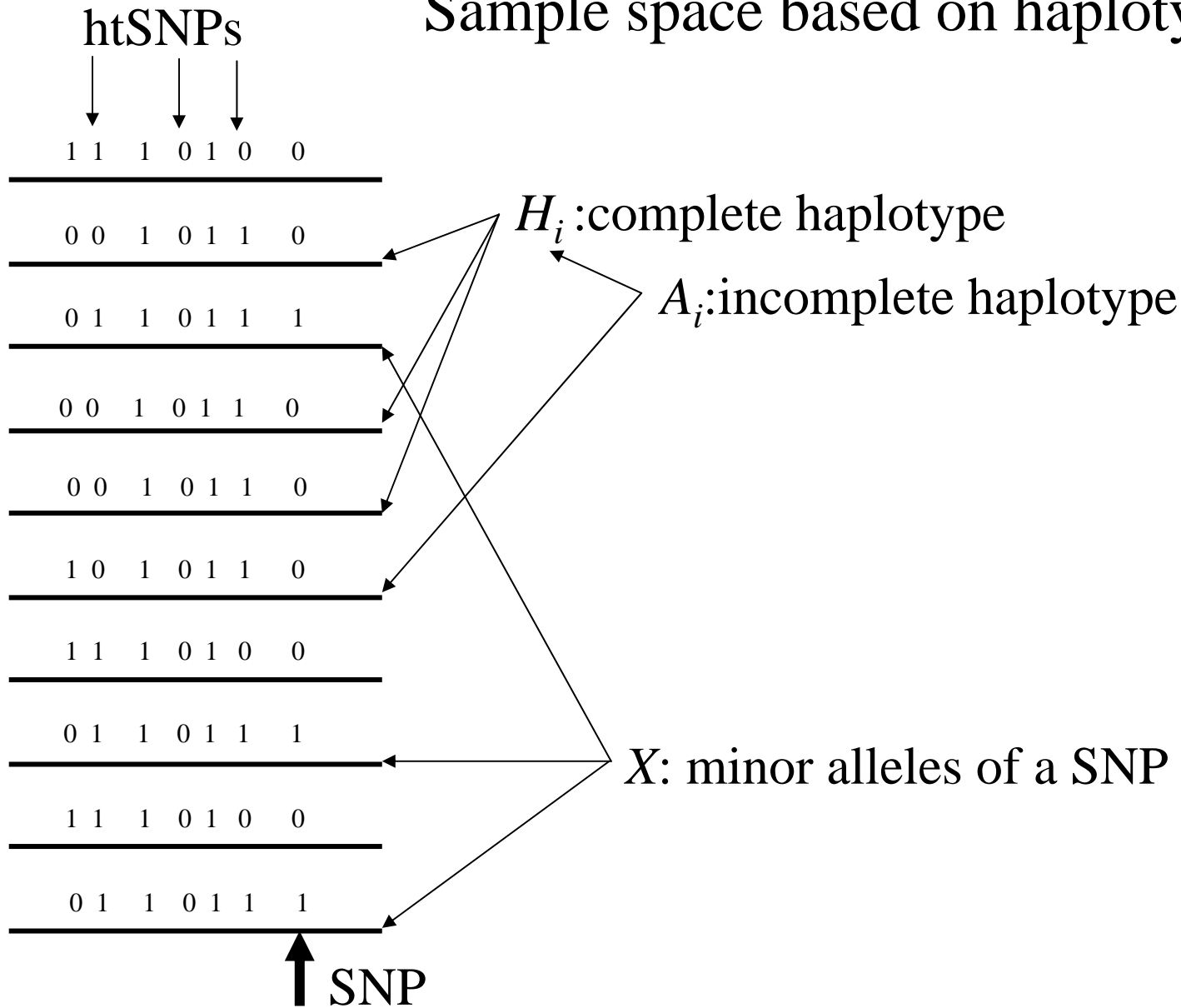
$\Omega$ : A set of all complete haplotypes

$H_i$ :  $i$ th complete haplotype

$X$ : minor allele of a SNP (a set of complete haplotypes  
with the minor allele at the SNP)

1. A complete haplotype  $H_i$ , an incomplete haplotype  $A_i$ , and a minor allele of a SNP  $X$  can be defined as events on the same sample space  $\Omega$ .
2. They can be targets to be associated with phenotypes.
3. Probability model can be applied to examine the relationship between those events.

# Sample space based on haplotypes





# Algorithms

# PENHAPLO

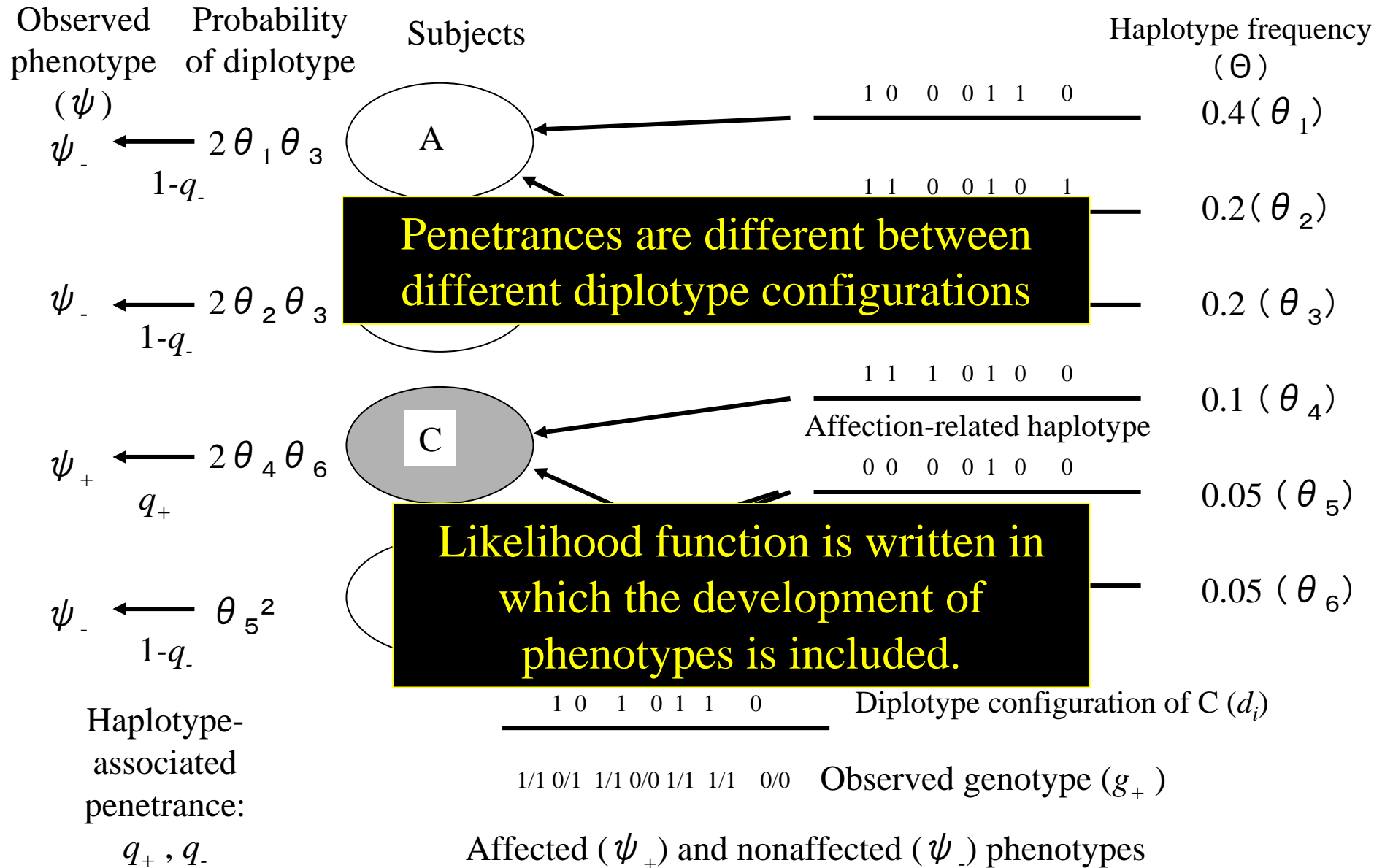
(Ito et al. Genetics, 2004)

**Algorithm:** **Infers** haplotype frequencies, diplotype configurations, and **penetrances** based on haplotypes, and **tests** the association between a **qualitative** phenotype and diplotype configurations. **SNPs, incomplete haplotypes and complete haplotypes** can be used as targets. **Dominant, recessive and genotype modes** can be used. **Ambiguous** diplotype configurations are allowed.

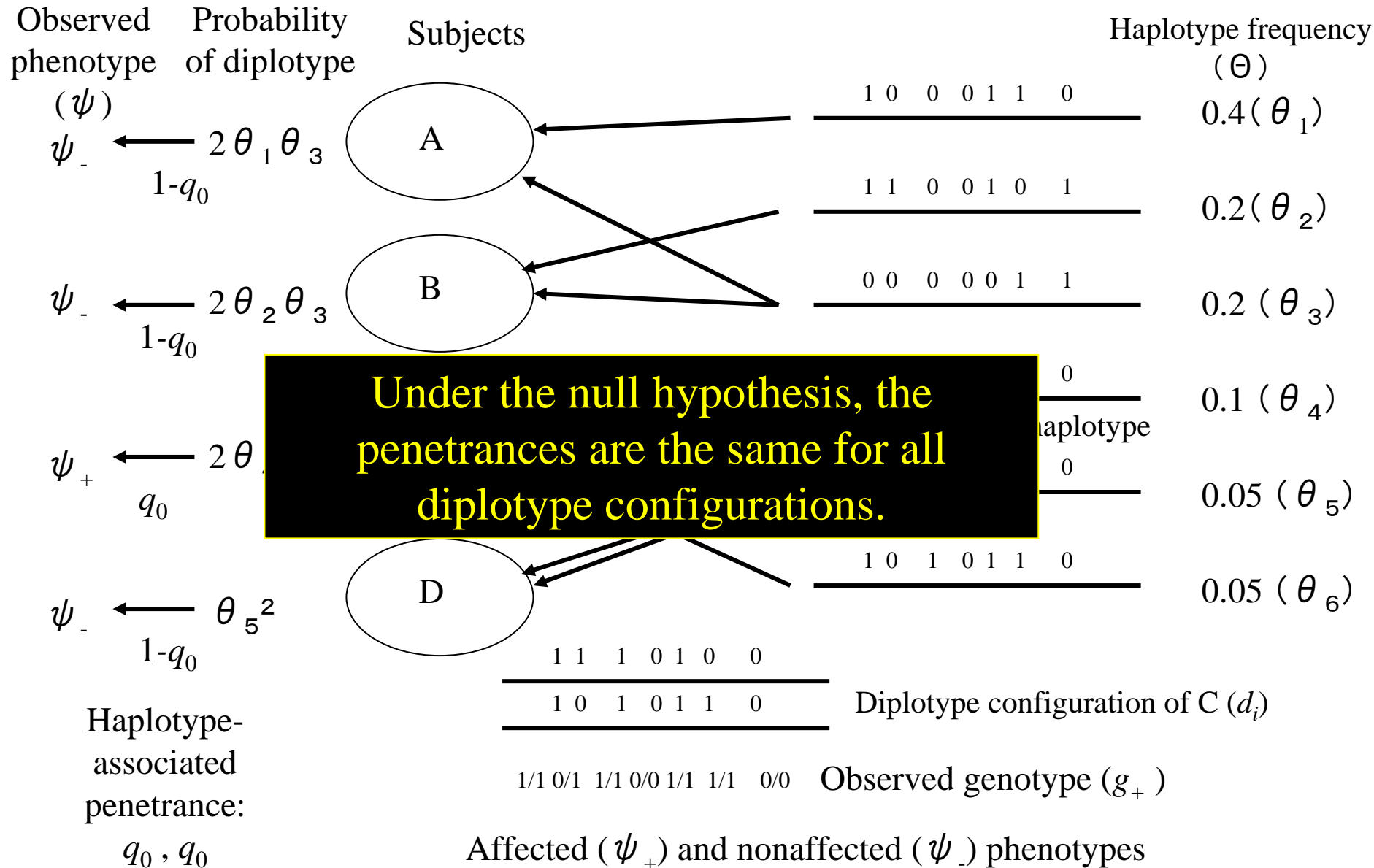
**Input data:** **Qualitative** phenotypes and genotype data for linked loci from many subjects

**Output data:** Maximum likelihood estimated **penetrances** for different diplotype configurations and **P-values** for the test of association between diplotype configurations and phenotypes.

# Sample space for PenHaplo (for alternative hypothesis)



# Sample space for PenHaplo (for null hypothesis)



Probability that  $i$ th subject gets a diplotype configuration under haplotype frequencies  $\Theta$

function under alternative

Probability that  $i$ th subject develops a phenotype under a diplotype configuration and penetrances

$$L(\Theta, q_+, q_-) \propto \prod_{i=1}^n \sum_{a_k \in A_i} P(d_i = a_k | \Theta, q_+, q_-) P(\psi_i = w_i | d_i = a_k, \Theta, q_+, q_-)$$

$A_i$ : A set of possible diplotype configurations for  $i$ th subject consistent with the observed genotypes.

$d_i$ : diplotype configuration of  $i$ th subject

$a_k$ :  $k$ th diplotype configuration

$q_+, q_-$ : Penetrances for a subset of diplotype configurations and the complement of the subset, respectively.

$\psi_i$ : Qualitative phenotype of  $i$ th subject

$w_i$ : Observed phenotype of  $i$ th subject

Ito et al. Genetics 2004

Likelihood

Probability that  $i$ th subject develops a phenotype under a a diplotype configuration and a penetrance

Probability that  $i$ th subject gets a diplotype configuration under haplotype frequencies  $\Theta$

null hypothesis

$$L(\Theta, q_0) \propto \prod_{i=1}^n \sum_{a_k \in A_i} P(\psi_i = w_i \mid d_i = a_k, q_0) P(d_i = a_k \mid \Theta)$$

$q_0$ : Penetrance common to all diplotype configurations

$$L_{max} = L(\hat{\Theta}, \hat{q}_+, \hat{q}_-)$$

$$L_{0max} = L(\hat{\Theta}, \hat{q}_0)$$

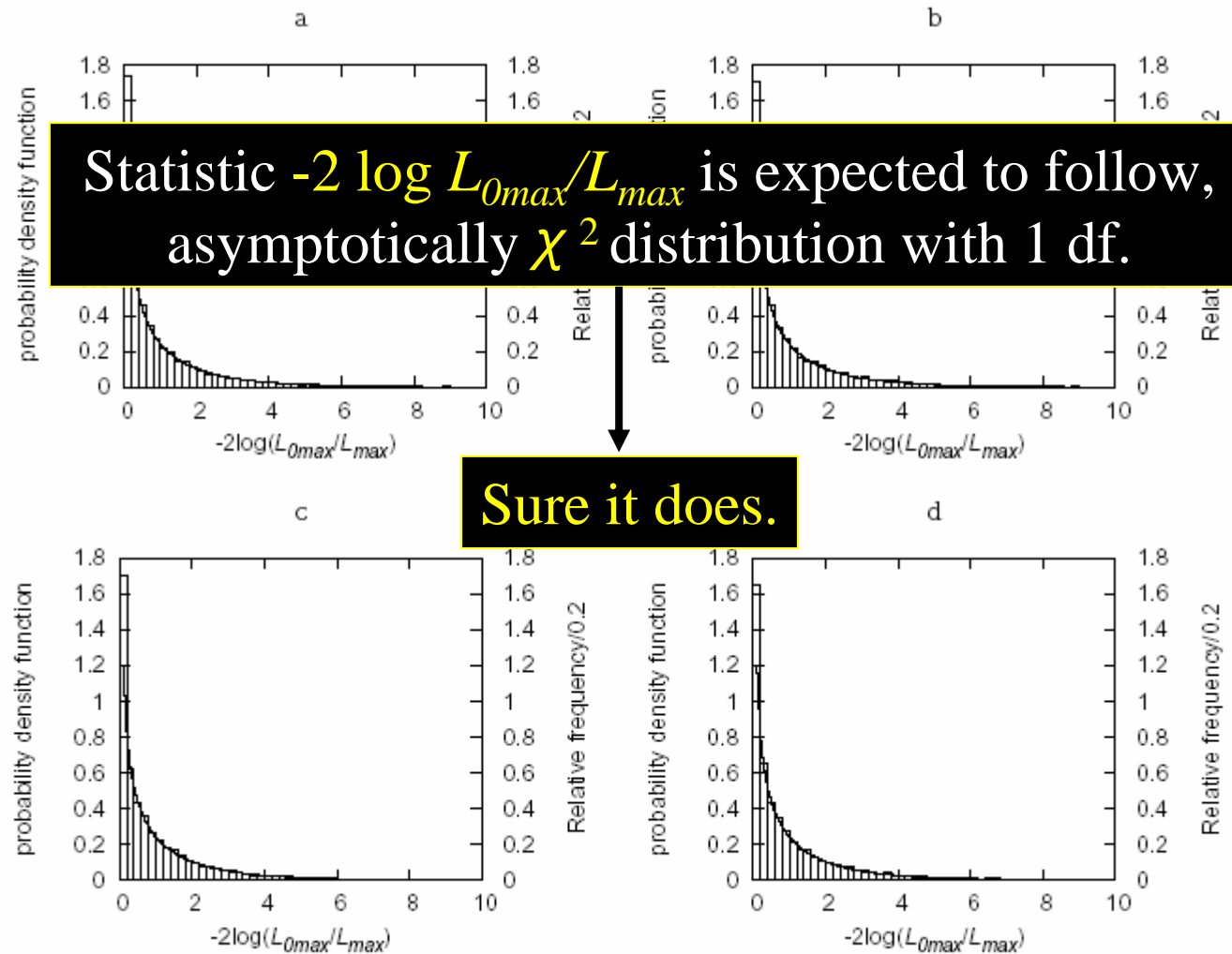
Parameters that maximize the likelihood functions in alternative and null hypotheses, respectively, are determined using EM algorithm.

Under null hypothesis

$$-2 \log(L_{0max}/L_{max}) \sim \chi^2$$

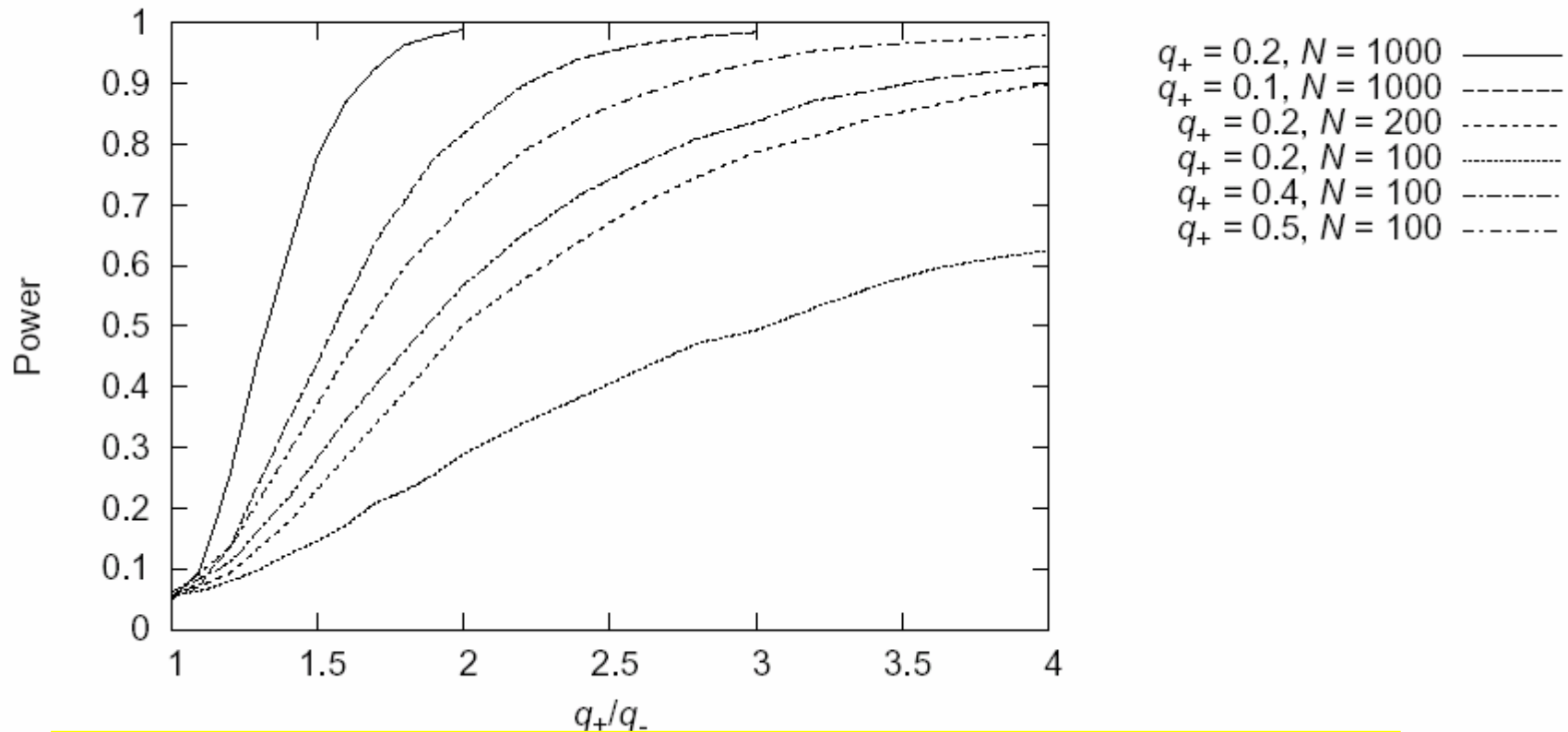
Statistic  $-2 \log L_{0max}/L_{max}$  is expected to follow, asymptotically  $\chi^2$  distribution with 1 df.

Expected and empirical distributions of statistic  
 $-2 \log L_{0max}/L_{max}$  under the null hypothesis



Test statistic  $-2 \log L_{0max}/L_{max}$  is expected to follow, under the null hypothesis,  $\chi^2$  distribution with 1 degree of freedom.

## Empirical power at various values of $q_+/q_-$ and the sample size



Power increases with increasing sample size  $N$  and penetrance ratio  $q_+/q_-$ .



**Estimated probability that a subject with certain genotypes develops a phenotype**

$$P(\psi_{N+1} = \psi_+ \mid g_{N+1}, \hat{\Theta}) \\ = \hat{q}_+ \sum_{a_k \in D_+} P(d_{N+1} = a_k \mid g_{N+1}, \hat{\Theta}) + \hat{q}_- \sum_{a_k \in D_-} P(d_{N+1} = a_k \mid g_{N+1}, \hat{\Theta})$$

$\psi_{N+1}$

$\psi_+$ : A

$g_{N+1}$

The probability that a subject with known genotypes develops a phenotype is estimated using maximum likelihood estimated penetrances ( $\hat{q}_+, \hat{q}_-$ ) and haplotype frequencies ( $\hat{\Theta}$ ).

$D_+$ : A set of certain diplotype configurations

$d_{N+1}$ : Diplotype configuration of  $N + 1$ th subject

# Conditions necessary for personalized medicine (Translating genomic evidence to the clinical practice)

- 1st step **Hypothesis testing**  
Is a phenotype (adverse events or efficacy) associated with genotypes?
- 2nd step **Replication (validation)**  
Is the association replicated in the test using independent samples?
- 3rd step **Algorithm for the intervention**  
Can the algorithm for the medical intervention be constructed, and is the outcome expected to be beneficial to the patients?

## Personalized drug delivery in Institute of Rheumatology, Tokyo Women's Medical University

1. Prediction of the adverse events of sulfasalazine
2. Prediction of the adverse events of methotrexate
3. Prediction of the efficacy of methotrexate
4. Prediction of the complication of amyloidosis

Institute of Rheumatology, Tokyo Women's Medical University

Largest rheumatology institution in the world

6,000 RA (rheumatoid arthritis) outpatients

44 permanent rheumatologists (quality controlled)

5-year cohort study enrolling 4,800 RA patients are on-going

Association between adverse events by  
sulfasalazine and haplotypes of  
N-acetyltransferase 2 (NAT2) gene

# Why is haplotype analysis necessary for NAT2 gene?



Slow acetylator



Rapid acetylator



Slow acetylator

## Association between haplotypes and adverse events by sulfasalazine

Penetrance for a set of diplotype configurations

Penetrance for the complement set

Incomplete haplotype	inheritance	P-value	q+	q-	RR
CCGG	Dominant	0.004	0.1512	0.5	0.30
*CG*	Dominant	0.007	0.1611	0.6667	0.24
TC*C	Recessive	0.007	0.6667	0.1611	0.24
C	Dominant	0.014	0.1561	0.4615	0.34
T***	Recessive	0.014	0.4615	0.1561	0.34
*CGG	Dominant	0.014	0.1561	0.4615	0.34

**Haplotypes should be considered for NAT2 gene**

Association between haplotypes (C677T-A1298C in MTHFR gene)  
and efficacy and adverse events by methotrexate (analysis by PENHAPLO)

Association between haplotype and risk of high dosage

Incomplete haplotype	inheritance	P-value	q+	q-	RR
*A	Recessive	0.007	0.476	0.259	1.836
CC	Dominant	0.007	0.259	0.476	0.545

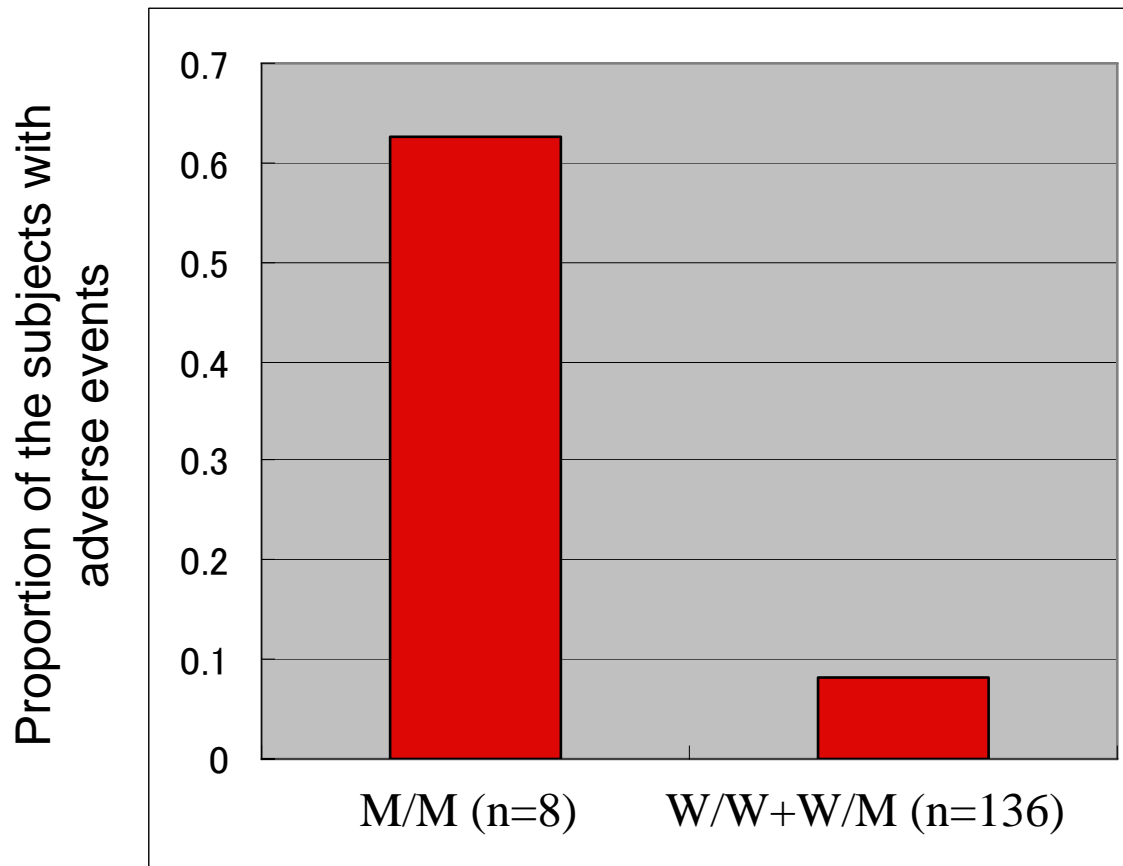
**Haplotype analysis is not necessary for MTHFR gene**

Association between haplotype and risk of adverse events

Incomplete haplotype	inheritance	P-value	q+	q-	RR
C*	Recessive	0.005	0.167	0.367	0.455
TA	Dominant	0.005	0.367	0.167	2.200
*A	Dominant	0.019	0.297	0.000	—
CC	Recessive	0.019	0.000	0.297	—

Since TC is not present, CC is the complement of \*A, and TA is the complement of C\*.

Association between diplotype configurations of NAT2 gene and  
adverse events by sulfasalazine  
(Cohort study, 144 subjects)



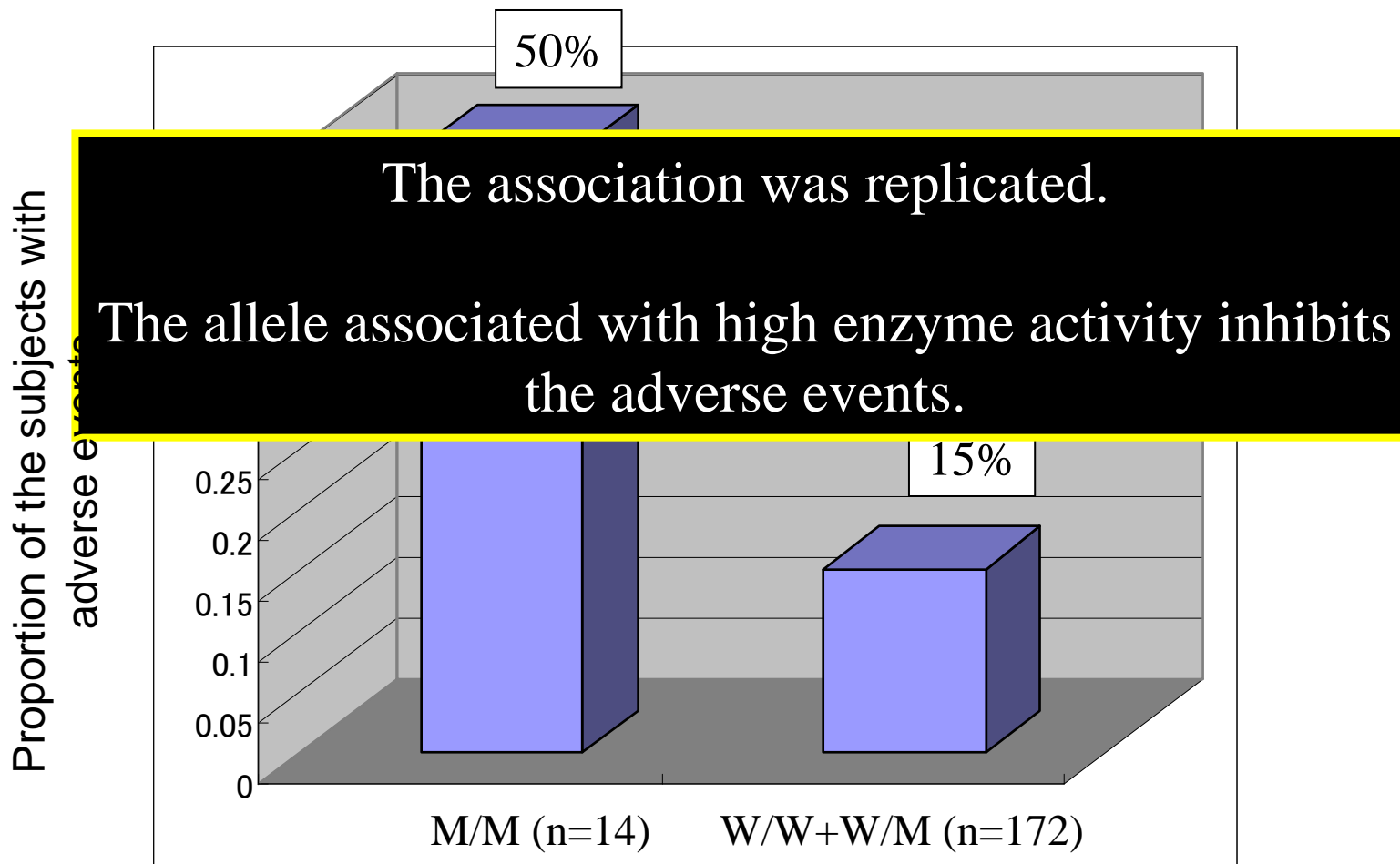
Absence of W haplotype is  
associated with adverse events

RR=7.73,  $P < 0.001$

Tanaka et al J Rheumatol, 2002



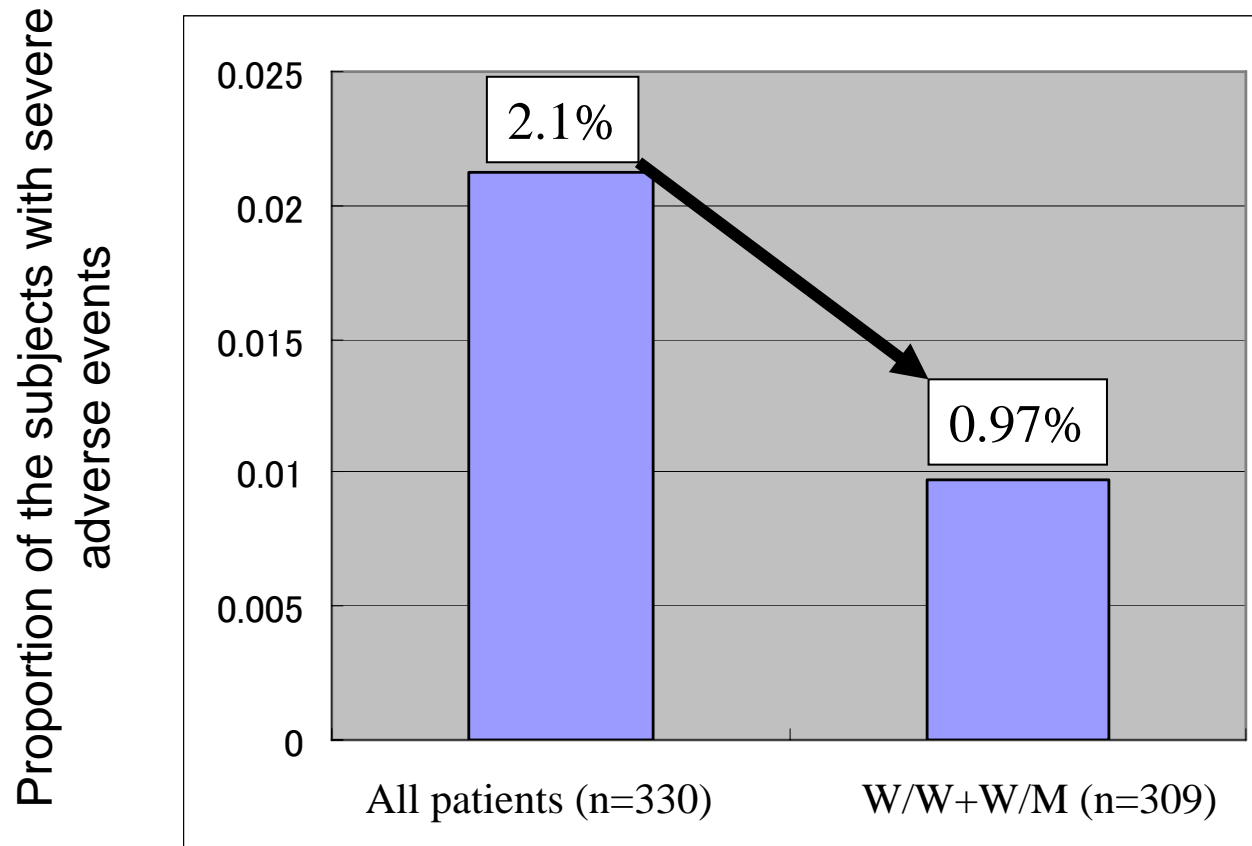
Association between diplotype configurations of NAT2 gene and adverse events  
(Replication study: 186 subjects)



RR=3.3 (1.8-6.2),  $P < 0.001$

**Taniguchi et al**

## Association between diplotype configurations of NAT2 gene and severe adverse events (Cohort 330 subjects)



If we exclude 21 subjects (6.4%) from the treatment with sulfasalazine, the proportion of the severe adverse events would be reduced by 54%.

After 10-20 years people would say..

以前の医師は患者さんの最も基本的な情報である、  
ゲノム配列も調べず治療をしていたのですか？  
何と勇敢な、そして何と危険な。

Had you been treating a patient without the fundamental  
information of the genome sequence?

How brave, and how risky!