유전체 연구 통계 기법

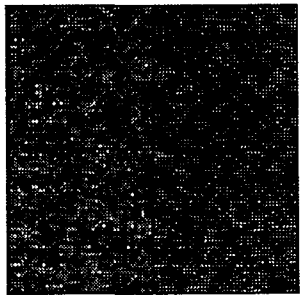
김 호
서울대학교 보건대학원

Contents

- Microarray Data Analysis
- Genetic Analysis: Linkage, LD, QTL
- Association Study using SNP & Haplotypes

I. Microarray Data Analysis

Microarrays

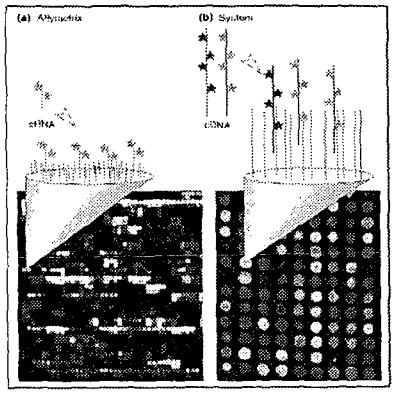


- 대량의 유전자에 대해 발현현상을 동시 관찰
- 유전자의 regulation과 interaction의 이해에 기여

Full Yeast Genome in a Chip
(Brown Lab, Stanford. Univ.)

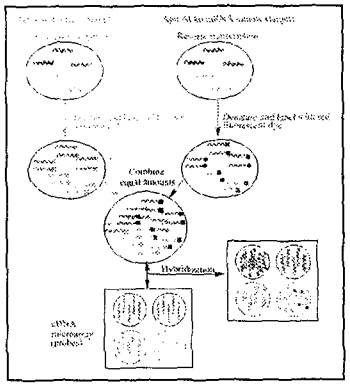
Types of Microarray

- Spotted arrays
  - highly specific complementary DNA (cDNA)
  - spots on glass slide as probes
  - two mRNA samples with different dyes
  - hundreds to thousands of spots per slide
- Oligonucleotide arrays (Affymetrix)
  - 20 sets of 25-mer probes per gene
  - pairs of positive match and mis-match
  - careful, secret selection of 25-mers
  - tens of thousands(6K-20K) of genes per chip



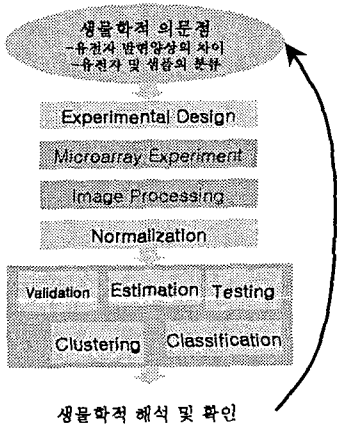
Hybridization samples indicated are cRNA or cDNA

Microarray Experiment



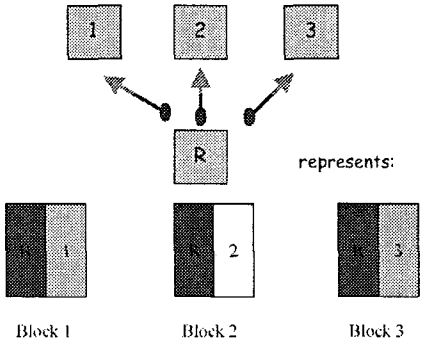
cDNA Microarray Experiment using Apo AI

Microarray & Statistical Method



Statistical problems in all the procedures

Experimental Design: Reference Design



Experimental Design: Loop Design

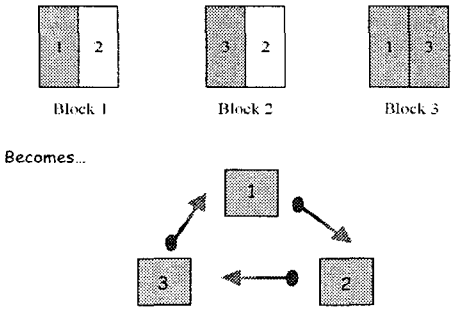
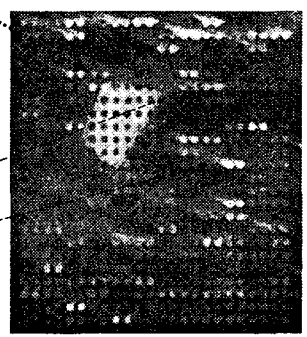


Image Analysis–Practical Problems

High Background
- 2 likely causes:
  - Insufficient blocking.
  - Precipitation of the lab eled probe.

Weak Signals
Comet tail



Steps in Images Processing

1. Addressing: locate centers

2. Segmentation: classification of pixels either as signal or background. using seeded region growing).

3. Information extraction: for each spot of the array, calculates signal intensity pairs, background and quality measures.

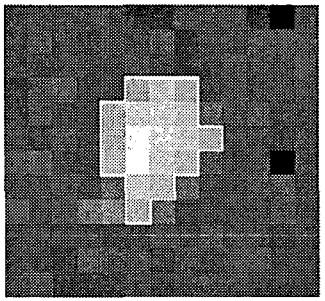


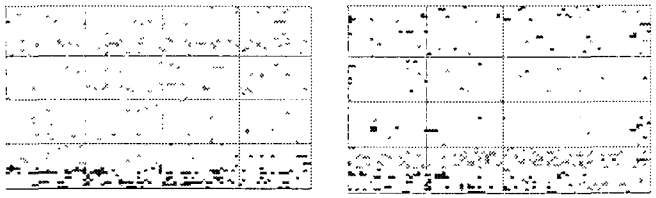
Figure 6: An example of a non-circular shaped spot. The yellow line shows the result of the SRG segmentation. The pixels inside the yellow line are classified as foreground and the other pixels are classified as background.

Normalization and Filtering

Filtering: 믿을 수 없는 값들을 제외하는 것, 그 결과는 missing이 됨

Normalization : Microarray자료에서 발현수준에 영향을 미치는 비생물적 기술적 변이(systematic variation)를 찾아내서 제거함

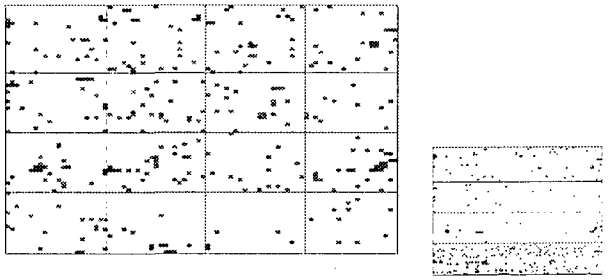
Effect of Location Normalization



Before normalization

After print-tip-group Normalization (just location)

Effect of location + scale Normalization



Just location

Testing Gene expression intensity

- Matrix X of log intensity ratios

  k(# of genes) by $n(n=n_1+n_2)$:

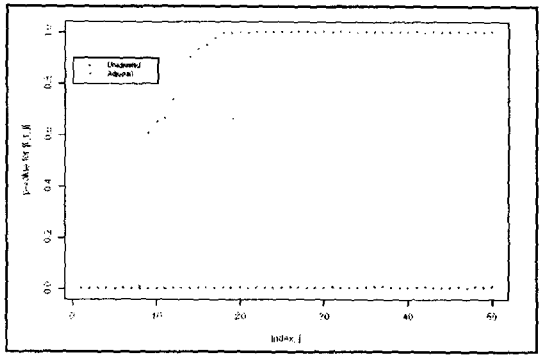
  Problem: Typically k is much larger than n

- Multiple Testing Problem

  When many hypotheses are tested, as is the case(# of gene=5548), the probability that at least one type I error is committed can increase sharply with the number of hypotheses.

  Westfall & Young step-down, Bonferroni single step, Sidak single step, Holm's step down, Young step down

Example



Apo AI

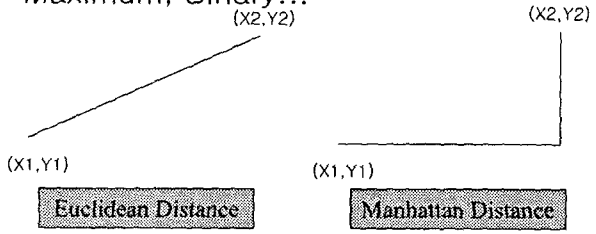
Top 50 genes

Cluster Analysis: intro

- Genes with similar function yield similar expression patterns in microarray experiments
- Searching for the groups (clusters) in the data, based on a measure or distance index of similarity or dissimilarity
- cluster – a set of entities which are alike

Kinds of Distance-1

- Euclidean : $\mathbf{d(x,y)=\{w_k(x_k-y_k)^2\}^{1/2}}$
- Manhattan : $\mathbf{d(x,y)=w_k|x_k-y_k|}$
- Maximum, Binary...

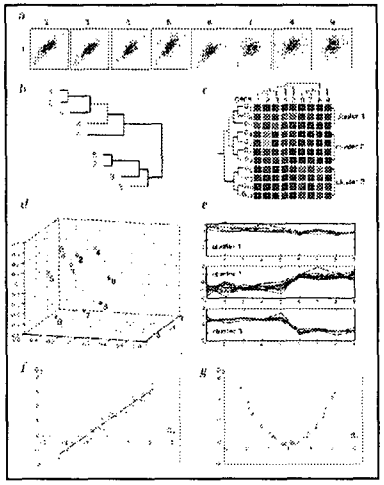


Kinds of Distance-2

- Correlation

$$r(x, y) = \frac{\sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{n} (x_k - \bar{x})^2 \sum_{k=1}^{n} (y_k - \bar{y})^2}}$$

$$d(x,y) = 1 - r(x,y)$$



a) data and partial scatterplot

b) distance between samples using hierarchical cluster dendrogram

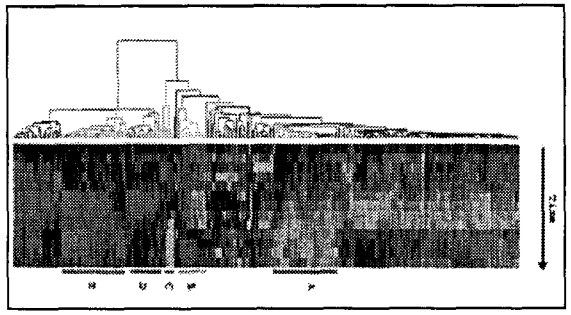
c) multidimensional method adding 2nd dimension of clustering

d) multidimensional scaling methods

e) similar fashion clustered by k-means clustering and SOM

f) linear correlation between individual expression level

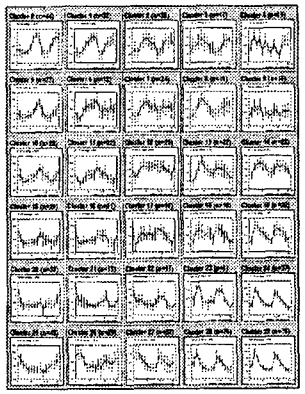
g) individual association using non-linear function



Hierrarchical Clustering(Eisen et al. 1998)

Clustered display of data from time course of serum stimulation of primary human fibroblasts

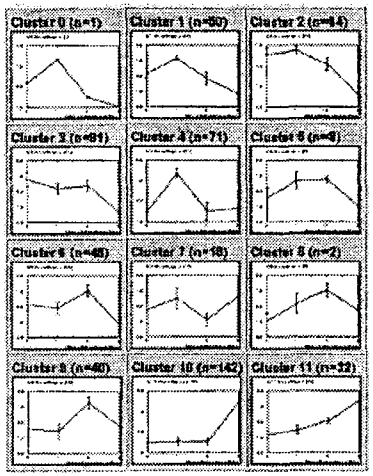
Example - Tamayo et al.(1999)



6 by 5 SOM.
The 828 genes that passed the variation filter were grouped into 30 clusters.

Each cluster is represented by the centroid (average pattern) for genes in the cluster.

Expression levels are shown on *y*-axis and time points on *x*-axis. Error bars indicate the SD of average expression. *n* indicates the number of genes within each cluster.



HL-60 SOM. HL-60 cells were treated with PMA for 0, 0.5, 4, or 24 hours,

and expression levels of more than 6,000 genes were measured at each time point.

The 567 genes passing the variation filter were grouped by a 4 3 3 SOM.

Software

상용화 S/W

NAME	COMPANY	FEATURE
ArraySuite	Affymetrix	plots, fold changes
ImaGene	Biodiscovery	quantification of gene expression value, constant-factor normalization
GeneSight	Biodiscovery	background adjustment, clustering(hierarchical, SOM)
GeneSpring	Silicon Genetics	normalization, clustering(hierarchical, SOM), fold-change
Spotfire	Spotfire	PCA, clustering, fold-change
Resolver	Rosetta	clustering, PCA, fold-change, plots
LifeArray	Incyte	clustering, PCA, fold-change
Expressionist	GeneData	clust, PCA, fold-change
GeneExpress	Gene Logic	clustering, PCA, fold-change
IPLab MicroArray Suit	Scanalytics	clustering, image, fold-changes

공개용 S/W

NAME	ORG	FEATURE
J-Express	U Bergen	clustering, PCA
UCI/NCGR	UCI/NCGR	t-test for fold-change
TreeView	Stanford	clustering, SOM, image analysis (similarity, Cluster,Lawrence Berkeley Lab, Eisen Lab)
EPCLUST	EBI	clustering
SOM	Whitehead Inst.	SOM

•http://home.cuhk.edu.hk/~b400559/arraysoft.html

II. Genetic Analysis

Linkage and Linkage Disequilibrium (1)

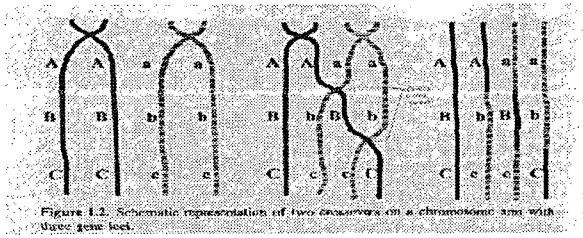
- Linkage: the tendency of genes or other DNA sequences at specific loci to be inherited together as a consequence of their physical proximity on a single chromosome.
- Linkage disequilibrium (allelic association): particular alleles at two or more neighboring loci show allelic association if they occur together with frequencies significantly different from those predicted from the individual allele frequencies.
- Linkage is a relation between loci, but association is a relation between alleles.

Linkage and Linkage Disequilibrium (2)

- Linkage: $0 \leq \theta < 0.5$
  (θ = recombination fraction)
  No linkage: $\theta = 0.5$
  Perfect linkage: $\theta = 0$
- Linkage disequilibrium: $0 \leq \rho \leq 1$
  (ρ = probability of allelic association)
  Linkage equilibrium: $\rho = 0$
  Complete linkage disequilibrium: $\rho = 1$

- GENETIC MAP DISTANCE (in units of Morgans)

  The expected number of crossovers occurring on a single chromosome (in a gamete) between loci.



- PARAMETRIC LINKAGE ANALYSIS

  To estimate the recombination fraction between markers and a hypothesized trait locus, where inheritance parameters of the trait locus (mode of inheritance, penetrance, phenocopy rate, allele frequencies etc) must be specified.

  Ex. Lod score method

- LOD SCORE

The common logarithm of the likelihood ratio:

$Z(\theta) = \log_{10} [L(\theta) / L(\frac{1}{2})]$

where θ is the recombination fraction between two loci

- Purpose Of The Lod Score Method

1. Estimation of the recombination fraction, θ

2. Hypothesis testing

H_0: $\theta = \frac{1}{2}$ (absence of linkage)

H_1: $\theta < \frac{1}{2}$ (linkage)

$$Z_{\max} = Z(\hat{\theta}) = \log_{10}[L(\hat{\theta}) / L(1/2)]$$

- Scale For Testing Linkage

$Z_{max} \geq 3$: Strong linkage

$Z_{max} > 0$: Support linkage

$Z_{max} < 0$: Against linkage

$Z_{max} = 0$: No support

(not related to recombination in linkage or no linkage)

- Asymptotic Distribution

$2 \ln [L(\theta) / L(\frac{1}{2})] = 4.6 \times Z_{max} \sim \chi^2_1$

under the null hypothesis of no linkage

$P(Z_{max} \geq 3) = P(\chi^2_1 \geq 13.8) = 0.0002$

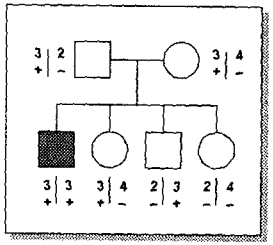
$\rightarrow \alpha = 0.0001$

- NONPARAMETRIC LINKAGE ANALYSIS

  Inheritance parameters of the trait locus are not specified. Rather, one focuses on pairs (or multiples) of affected individuals and investigates marker allele sharing among these individuals, contrasting observed allele sharing with that expected when the marker has nothing to do with the trait.

  Ex. IBD (identical by descent) test

AN EXAMPLE FAMILY WITH DISEASE LOCUS AT THE MARKER



Sib-Pair Markers		Disease Status		# of Shared i.b.d.	C
sib1	sib2	d1	d2		
3|3	3|3	+	+	2	1
3|3	3|4	+	-	1	0.25
3|3	2|3	+	-	1	0.25
3|3	2|4	+	-	0	0.25
3|4	3|4	-	-	2	0.5
3|4	2|3	-	-	0	0.5
3|4	2|4	-	-	1	0.5
2|3	2|3	-	-	2	0.5
2|3	2|4	-	-	1	0.5
2|4	2|4	-	-	2	0.5

- Only '+ +' indicates as "affected" ('+' is recessive to '-')

** Qualitative Trait

- $C_j = (d_{j1} - \mu)(d_{j2} - \mu)$
  $= \alpha + \beta IBD_j + \varepsilon_j$

Allelic Association (LD)

Morton et al. (2001)

Locus A	Locus B: Allele 1	Locus B: Allele 2	Allele frequency
Allele 1	$\pi_{11} = Q\rho + QR(1-\rho)$	$\pi_{12} = (1-\rho)Q(1-R)$	$Q = \pi_{11} + \pi_{12}$
Allele 2	$\pi_{21} = (R-Q)\rho + R(1-Q)(1-\rho)$	$\pi_{22} = (1-R)\rho + (1-Q)(1-R)(1-\rho)$	$1-Q = \pi_{21} + \pi_{22}$
Allele frequency	$R = \pi_{11} + \pi_{21}$	$1-R = \pi_{12} + \pi_{22}$	1

A, B: diallelic loci; π_{11}, π_{12}, π_{21}, π_{22}: haplotypes; ρ: association probability

Measures of LD

- Covariance

  $D = | \pi_{11}\pi_{22} - \pi_{12}\pi_{21} |$
- Association

  $\rho = D/Q(1-R)$
- All other measures are functions of Q, R, ρ.

- Let $P(1 \text{ allele}) = p_1$, $P(2 \text{ allele}) = 1 - p_1$, $P(T \text{ allele}) = p_T$ and $P(+ \text{ allele}) = 1 - p_T$
  With LD, the *T* and *1* alleles appear positively associated so that

  $P(T1 \text{ haplotype}) = P_{T1} = p_T p_1 + D$
  where D ($D = p_{T1} - p_T p_1$) is the *disequilibrium parameter.*

  → $D = 0$: Gametic phase equilibrium or linkage equilibrium

  $D > 0$: Positive disequilibrium or association

- Linkage And LD

  - The two loci can be assumed to reside on different chromosomes.

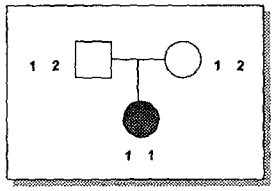
  → The presence of LD does not necessarily imply linkage between the loci considered.

  - Although LD originally referred to an association of alleles at different loci, it has become customary to take LD to mean association among alleles due to close linkage. "*allelic association*"

- Another Approach To LD Analysis ("Family-Based Study")

1. Haplotype relative risk (HRR) method
   : Falk and Rubinstein (1987)
2. Haplotype-based haplotype relative risk (HHRR) method: Terwilliger and Ott (1992)
3. Transmission/ disequilibrium test (TDT)
   : Spielman *et al.* (1993)
4. Sib-Transmission/ disequilibrium test (S-TDT): Spielman and Ewens (1998)

- Transmission/ disequilibrium test



		Not transmitted	
		Allele1 (A1)	Allele2 (A2)
Transmitted	Allele 1 (A1)	0 (a)	2(b)
	Allele2 (A2)	0 (c)	0(d)

- Focus on heterozygous parents only, and allow the use of multiple affected siblings.
- McNemar's test (standard χ^2 test) H_0: $b = c$
  The TDT statistic: $\chi_1^2 = \frac{(b-c)^2}{b+c}$
- Powerful only in the presence of LD.

- QUANTITATIVE TRAIT

  A phenotype with a continuous (normal/ lognormal) distribution.

  Ex. Height, blood pressure, head circumstance and the cholesterol level in the blood

- QUALITATIVE TRAIT

  A phenotype with a discrete distribution.

  Ex. Signs and symptoms indicate whether a disease state is present or absent.

- Effects On The Quantitative Traits

  – The mean phenotype for a trait in a population:

  $P = G + E + GE$

  – The variation in the trait:

  $V_p = V_g + V_e$

  $V_g = V_a + V_d + V_i$

  with genetic variance (V_g), environmental variance (V_e), additive variance (V_a), dominance variance (V_d) and interaction genetic (epistatic) variance (V_i).

- HERITABILITY Of The Trait (H^2)

  The fraction of the variation caused by genetic variation.

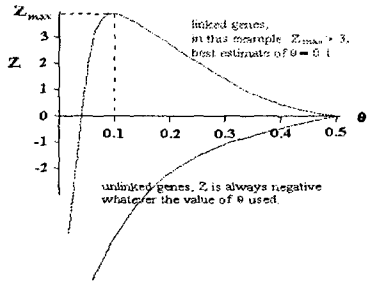
  $H^2 = V_g / V_p = V_g / (V_g + V_e)$

- QUANTITATIVE TRAIT LOCI (QTL)

  The location of a gene that affects a trait that is measured on a quantitative (linear) scale. The loci that are determinants of quantitative trait expression.

- Statistical QTL Technique

1. Least squares method



2. Bayesian linkage analysis (Markov chain Monte Carlo method)
3. Variance component approach

- General Procedure For Mapping QTL

Step1: Identify the molecular genotype of each marker for a quantitative trait.

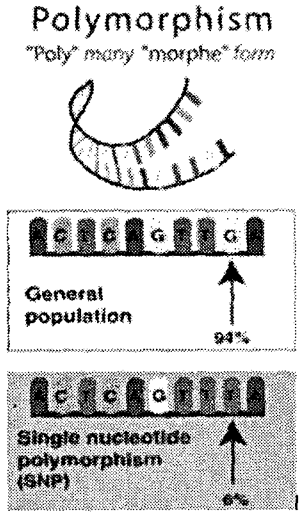
Step2: Determine if an association exists between any of the markers and the quantitative trait, by the methods of one-way ANOVA and regression analysis by using phenotypic and genotypic data.

Step3: Take those molecular marker loci that are associated the quantitative trait and perform a multiple regression analysis.

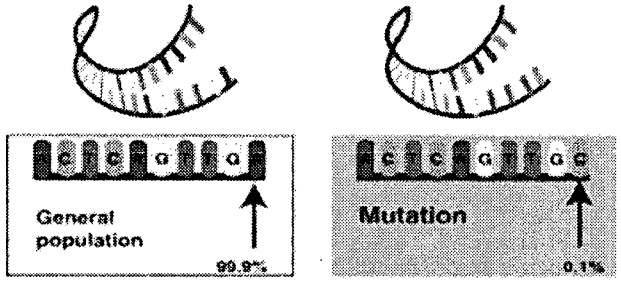
- Findings Of The QTL Analysis
  - # of genes involved to the quantitative trait
  - The location of the loci determining the quantitative trait
  - How much of the phenotypic (& genotypic) variation is accounted for
  - What parent has the worthful alleles

III. Association Study Using SNP and haplotype

SNPs
(pronounced snips)



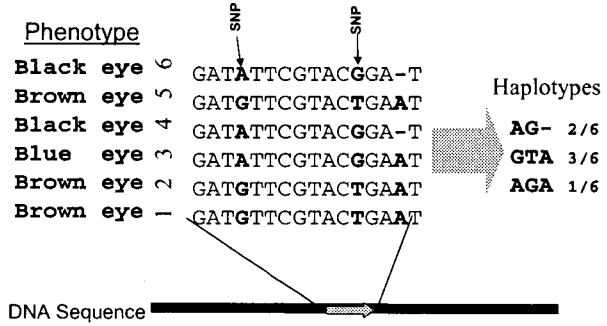
Mutation



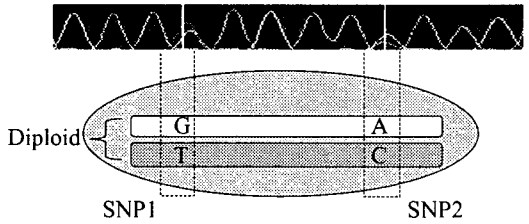
Polymorphism – Definition

- Polymorphism
  - A sequence variation that occurs at least 1 percent of the time (> 1%)
  - 90% of variations are SNPs
- Mutation
  - If the variation is present less than 1 percent of the time (<= 1%)

From SNP to Haplotype



Haplotyping: Phase Problem



Observed: SNP1 G/T SNP2 A/C
Possible Haplotypes: GA, TC or GC, TA

n SNPs → 2^n possible haplotypes

In-silico Haplotyping: Two Tasks

I. Reconstruction of the haplotypes of the sampled individuals

II. Estimation of haplotypes frequencies in a population

In-silico Haplotyping: Approaches

1) Clark's algorithm

2) E-M algorithm (expectation-maximization algorithm)

3) Bayesian algorithm

Clark's Algorithm

1) Find Homozygotes or heterozygotes at one locus

SNP1	T T	T-A-C	
SNP2	A A	T-A-C	Unambiguously defined
SNP3	C C		
SNP1	T T	T-A-C	
SNP2	A A	T-A-G	
SNP3	C G		

Clark's Algorithm

2) Try to solve ambiguous haplotype as a combination of solved ones

SNP1	A T	T-A-C : solved one
SNP2	A A	A-A-G
SNP3	C G	

.....

Continue until either all haplotypes have been solved or until no more haplotypes can be found in this way

Clark's Algorithm problems

- No homozygotes or single SNP heterozygotes -> chain might never get started
- Many unsolved haplotypes left at the end
- Quite useful in practice !!

Known haplotypes:

22544
22544
22544
22544
33334
33334
23233
14234

Ambiguous individual 1:

Genotype
32344 / 23534
→ 33334
→ 22544

Ambiguous individual 2:

Genotype
32444 / 23434
→ 33434
→ 22444

Figure 1 Illustration of how our method uses the fact that unresolved haplotypes tend to be similar to known haplotypes. Suppose that we have a list of haplotypes, as shown on the left side of the figure, that are known without error (e.g., from family data or because some individuals are homozygous). Then, intuitively, the most likely pair of haplotypes for ambiguous individual 1 consists of two haplotypes that have high population frequency, as shown. All methods considered here will correctly identify this as the most likely reconstruction. However, ambiguous individual 2 cannot possess any of the haplotypes in the known list, and the most plausible reconstruction for this individual consists of two haplotypes that are *similar, but not identical*, to two haplotypes that have high population frequency, as shown. Of the methods considered here, only our method uses this kind of information, leading to the improved performance we observed.

Haplotype Inference

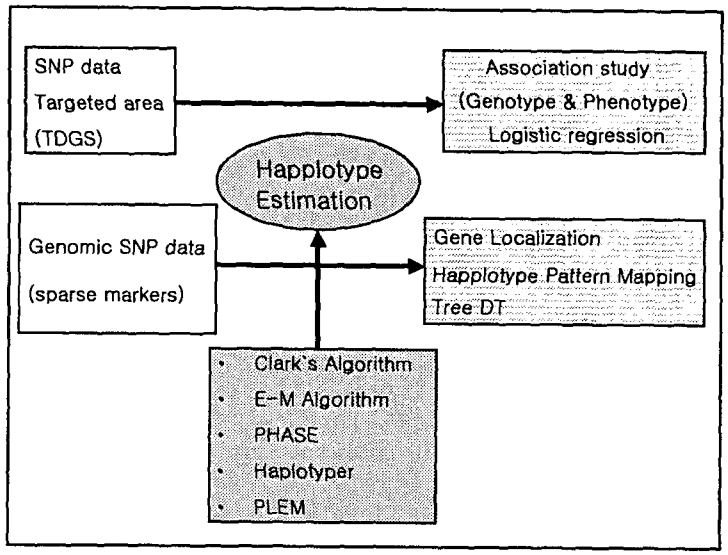
A: SNP data: 0 (MM), 1 (Mm), 2 (mm) for a single locus

B: Haplotype data: 0(M), 1 (m) for a single locus

A

B

```
#1 1, 2
00000
00100
#2 1, 3
00000
00010
#3 1, 4
00000
01001
#4 1, 5
00000
00001
#5 1, 1
00000
00000
#6 1, 1
00000
00000
```



Haplotype Pattern Mining

A new method for linkage disequilibrium map.

Based on discovering recurrent patterns inspired by data mining

Define a class of useful haplotype pattern in genetic case-control data

Give an algorithm for finding disease-associated haplotypes

Haplotyes exceeding a threshold level are used for prediction of disease susceptibility gene location.

- Traditional linkage studies > use recombination information only in pedigrees
- Association methods > use recombination information at the population level
- Association methods have greater power to detect small and moderate genetic effects than does linkage analysis (Risch and Merikangas 1996)

- SNP markers are preferred over microsatellite markers for asso st

① High abundance in human genome

② Low mutation rate

③ Accessibility of high-throughput genotyping

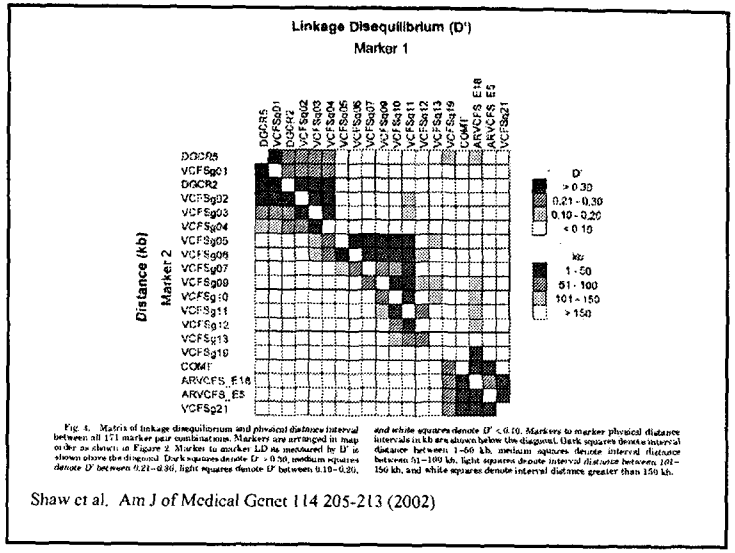
- Power of asso st depends on

① Sample size and density of the marker

② age and freq of the disease mutations and SNPs, LD in the region

LD pattern

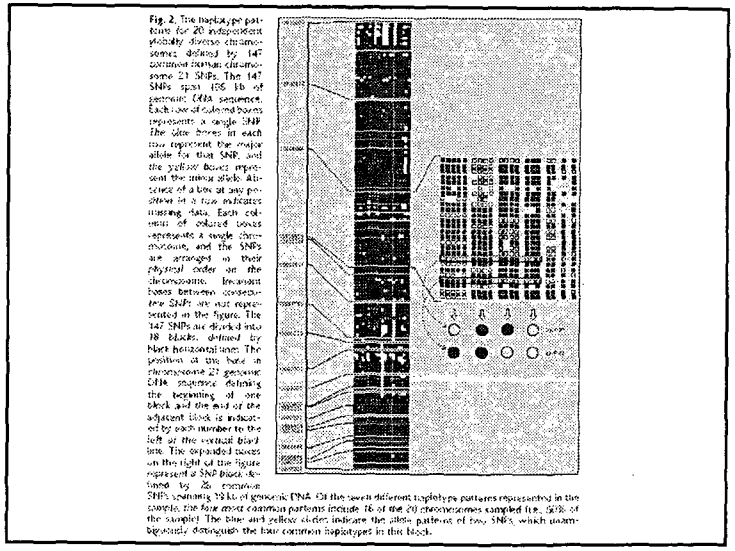
- Kruglyak (1999) : LD < 3kb so 500,000 SNPs required
- Reich et al (2001) 60 kb 50,000 SNPs
- Substantial variation (region, pop) : several hundred kb to several kb



Shaw et al. Am J of Medical Genet 114 205-213 (2002)

Hap Block Structure

- Hap block, tag SNP
- Patil et al (2001): chromosome 21 for 24,047 SNPs (>=10% minor allele freq)
- 20 haps were partitioned into 4,135 hap blocks (repeated hap accounted for 80% of the observed hap)
- 4,563 SNPs (tag SNPs)
- Zhang et al (2002) 2,575 blocks, 3,582 tag SNPs (15% is sufficient)



Strategy for Suggested Asso St for complex disease by Zhang et al. (2002) AJHG 71:1386

1. Small # of people (10–20) genotyped at a very dense SNP map, haps also determined
2. Hap block partitioning algorithm : hap block and tag SNPs
3. Large # of people genotyped at tag SNP marker loci
4. Association study analysis

- The above approach will clearly reduce the genotyping cost
- Power ? Research needed
- Zhang et al (2002) AJHG 71:1386
- Tag SNPs (25%) power loss by 4% ~12% by randomly chosen SNPs in a two locus hap analysis
- Tag SNPs (14%) 9 % power loss comparing to 21% power loss by randomly chosen SNPs

Other Examples

Table 1. Localization of SNPs and identification of haplotypes of the β2AR gene

Nucleotide:	-1023	-709	-654	-468	-406	-367	-47	-20	46	79	252	491	523	Frequency (%)			
Allele:	G/A	C/A	G/A	C/G	C/T	T/C	T/C	T/C	G/A	C/G	G/A	C/T	C/A	Ca	A-A	As	H-L
Haplotype																	
1	A	C	G	C	C	T	T	T	A	C	G	C	C	0.7	25.0	12.5	10.0
2	A	C	G	G	C	C	C	C	G	G	G	C	C	48.3	6.3	10.0	26.7
3	G	A	A	C	C	T	T	T	A	C	G	C	C	0.7	0.0	0.0	0.0
4	G	C	A	C	C	T	T	T	A	C	G	C	C	33.0	29.7	45.0	40.0
5	G	C	A	C	C	T	T	T	G	C	G	C	C	1.4	0.0	0.0	0.0
6	G	C	G	C	C	T	T	T	G	C	A	C	A	13.2	31.3	30.0	13.3
7	G	C	G	C	C	T	T	T	G	C	A	T	A	1.0	1.6	0.0	3.3
8	G	C	A	C	C	T	T	T	A	C	A	C	A	0.7	0.0	0.0	0.0
9	A	C	G	C	T	T	T	T	A	C	G	C	C	0.0	4.7	0.0	0.0
10	G	C	G	C	C	T	T	T	G	C	A	C	C	0.7	0.0	0.0	3.3
11	G	C	G	C	C	T	T	T	G	C	G	C	C	0.3	0.0	2.5	0.0
12	A	C	G	G	C	T	T	T	A	C	G	C	C	0.0	1.6	0.0	3.3
Location:	5'	5'	5'	5'	5'	5'	AA16 Arg/Cys	5'	AA16 Gly/Arg	AA27 Glu/Gln	syn	AA164 Ile/Val	syn				

Drysdale et al. PNAS 2000, 97(19) 10483-10488

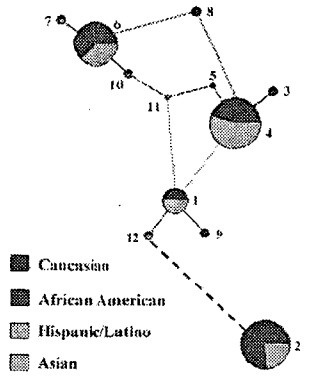


Fig. 1. Phylogeny of β_2AR haplotypes. Each haplotype is represented by a circle whose area represents the overall frequency of that haplotype in the sample. Lines connecting haplotypes are solid black for single-site differences, solid blue for two-site differences, and dashed for more than two differences. Each circle is subdivided to show the proportion of the individual haplotype frequency found in each of the four population groups as represented by the indicated colors.

Table 2. β_2AR haplotype pairs found in the asthmatic cohort

Haplotype pair	Chromosome A haplotype		Chromosome B haplotype		n	%
2/4	A C G G C C C C G G G C C	/	G C A C C T T T A C G C C		37	30.6
2/2	A C G G C C C C G G G C C	/	A C G G C C C C G G G C C		25	20.7
2/6	A C G G C C C C G G G C C	/	G C G C C T T G C A C A		22	18.2
4/4	G C A C C T T T A C G C C	/	G C A C C T T T A C G C C		14	11.6
4/6	G C A C C T T T A C G C C	/	G C G C C T T G C A C A		8	6.6
2/5	A C G G C C C C G G G C C	/	G C A C C T T T G C G C C		2	1.7
4/10	G C A C C T T T A C G C C	/	G C G C C T T T G C A C C		2	1.7
1/4	A C G C C T T T A C G C C	/	G C A C C T T T A C G C C		1	0.8
1/6	A C G C C T T T A C G C C	/	G C G C C T T T G C A C A		1	0.8
2/11	A C G G C C C C G G G C C	/	G C G C C T T T G C G C C		1	0.8
2/3	A C G G C C C C G G G C C	/	G A A C C T T T A C G C C		1	0.8
2/7	A C G G C C C C G G G C C	/	G C G C C T T T G C A T A		1	0.8
2/8	A C G G C C C C G G G C C	/	G C A C C T T T A C A C A		1	0.8
3/4	G A A C C T T T A C G C C	/	G C A C C T T T A C G C C		1	0.8
4/5	G C A C C T T T A C G C C	/	G C A C C T T T G C G C C		1	0.8
4/7	G C A C C T T T A C G C C	/	G C G C C T T T G C A T A		1	0.8
4/8	G C A C C T T T A C G C C	/	G C A C C T T T A C A C A		1	0.8
6/7	G C G C C T T T G C A C A	/	G C G C C T T T G C A T A		1	0.8

Nucleotide positions are omitted for clarity but are the same as in Table 1. Chromosomes A and B are arbitrarily

TABLE I. Cross-Tabulation of Genotype by Disease Group[a]

Genotype pattern	Number of outcomes		Number of haplotypes			
	Cases	Controls	1-1-1	2-1-1	2-2-2	Other
1	31	70	2	0	0	0
2	10	22	1	1	0	0
3	16	20	1	0	1	0
4	1	1	1	0	0	1
5	0	1	0	2	0	0
6	3	2	0	1	1	0
7	1	0	0	1	0	1
8	2	0	0	0	2	0
9	1	0	0	0	1	1
10	0	1	0	0	0	2
Total	63	117				

[a] In pattern 7, other = 2-2-1, for all other patterns, "other" refers to 1-2-1. As previously noted, the nomenclature is not consistent with Sjölander et al., or other previous papers. The first and third indices are switched, so what we call 1-1-1 Sjölander et al. [1995a,b; 1996] call 2-1-2, our 2-1-1 is their 1-1-2, and our 2-2-2 is their 1-2-1.

Wallenstein, Hodge, and Weston, Genetic Epidemiology 15:173–181 (1998)

TABLE II. Cross-Tabulation of Haplotype by Disease Group

	1-1-1	2-1-1	2-2-2	Other	Total
Cases	89 (68%)	14 (11%)	24 (18%)	3 (2%)	130
Controls	183 (78%)	26 (11%)	22 (9%)	3 (1%)	234
					364

Cohort study

TABLE III. Parameter Estimates Obtained by Logistic Regression

Parameter	Estimate	Std Err
"Homozygous" parameterization with no intercept		
θ (1,1,1)	−.431	.105
θ (2,1,1)	−.323	.336
θ (2,2,2)	.425	.298
θ (other)	.006	.717
"Baseline-haplotype" parameterization with the (1-1-1) haplotype as the base haplotype		
Intercept	−.862	.211
β (2,1,1)	.108	.323
β (2,2,2)	.850	.339
β (other)	.527	.726

Case-control study

References

• Pan W (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics 18(4):546-54.

•Clark (1990). Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Bio Evol 7: 111-122

• Escoffier and Slatkin (1995). Maximum likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Bio Evol 12: 921-927.

• Stephens, Smith, and Donnelly (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68, 978-989.

• Niu, Qin, Xu and Liu (2002) Bayesian haplotype inference for multiple linked single-nucleotide ploymorphisms. Am J Hum Genet 70;157-169

• Wallenstein, Hodge, Weston (1998) Logistic regression model for analyzing extended haplotype data. Genet Epidemiol 15:173-181.

•Http://www.genome.helsinki.fi/eng/research/projects/DM/index.html

•ZHAOHUI S. QIN, TIANHUA NIU, JUN S. LIU (2002) Partition-Ligation–Expectation-Maximization Algorithm for Haplotype Inference with Single-Nucleotide Polymorphisms *Am. J. Hum. Genet. 71:1242–1247, 2002*

Special thanks to

Microarray : 고대통계학과 이재원 교수

Genetics: 서울의대 의학연구원 서영주 박사

SNP & haplotype: 서울대 보건대학원 조성일 교수

References

•Baron M (2001) The search for complex disease genes: fault by linkage or fault by association? Mol Psychiatry 6(2):143-9.

• Escoffier and Slatkin (1995). Maximum likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Bio Evol 12: 921-927.

• Stephens, Smith, and Donnelly (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68, 978-989.

• Niu, Qin, Xu and Liu (2002) Bayesian haplotype inference for multiple linked single-nucleotide ploymorphisms. Am J Hum Genet 70;157-169

•Toivonen et al. (2000) Data Mining Applied to Linkage Disequilibrium Mapping. AM J Hum Genet 67: 133-145

•Petteri Sevon, Hannu T.T. Toivonen, Vesa Ollikainen. **TreeDT: Gene Mapping by Tree Disequilibrium Test**. The Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001), pp. 365-370. San Francisco, California, August 2001.

•Petteri Sevon, Vesa Ollikainen, Päivi Onkamo, Hannu Toivonen, Heikki Mannila, and Juha Kere. **Mining Associations Between Genetic Markers, Phenotypes and Covariates**. Genetic Analysis Workshop 12, Genetic Epidemiology, 21 (Suppl. 1), 2001.