LIST307: Functional Genomics (Part I)

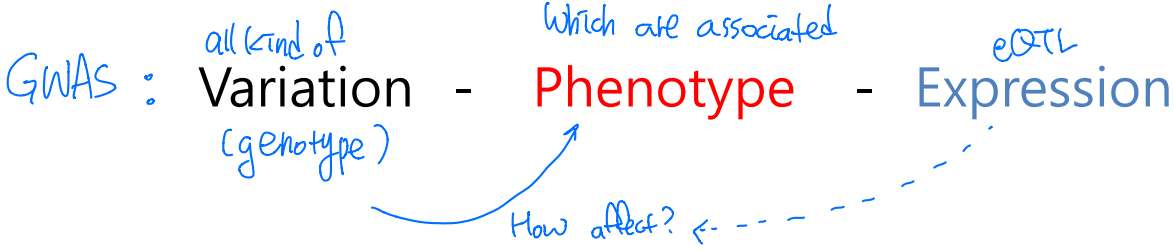
3/13 (Tue) : Introduction to Genomics
3/15 (Thu) : Genome Mapping
3/20 (Tue) : Human Genome Project
3/22 (Thu) : Next-Generation Sequencing
3/27 (Tue) : NGS & Sequence Alignment
3/29 (Thu) : Haplotyping & SNPs
4/3 (Tue) : Variation and Whole Genome Sequencing
4/5 (Thu) : Genome-wide association study
4/10 (Tue) : GWAS, eQTL, Phylogenetic analysis
4/12 (Thu) : Basics of Functional Genomics
4/17 (Tue) : Exome-Seq & Comparative Genomics

4/19 (Thu) : Review of functional genomics (part I)

Middle-term exam: 4/24 (Thu) or 4/26 (Tue) ?

Summary & More

: GWAS , eQTL



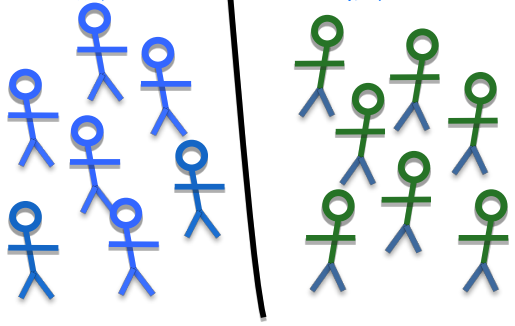
Sung Wook Chi

Division of Life Sciences, Korea University

GWAS

Control Population (unaffected)

Disease Population (affected)



SNP chip (disease와 관련된 유전자 찾아내는)

detect Tag SNP ⇒ enough to figure out region

NGS sequencing

P-value

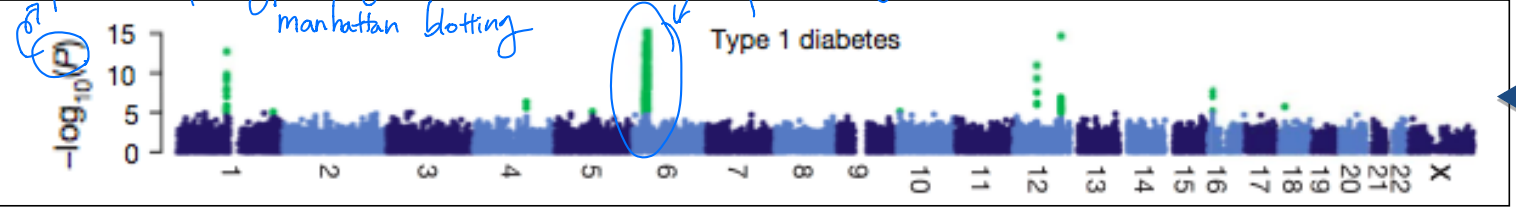
$$LOD = - \log_{10} \frac{\text{probability of birth sequence with a given linkage value}}{\text{probability of birth sequence with no linkage}}$$

(observed) / (expected)

(logarithm of the odds)

p value … phenotype이 significant한지 manhattan plotting

관련 있는 region



GWAS: Practice

sample	seq	type	grade
1	actgtacttc	normal	0
2	actgtacttc	normal	0
3	actgtacttc	normal	0
4	actgtccttc	normal	0
5	actgtccttc	normal	0
6	actgtccttc	tumor	2
7	actgtccttc	tumor	3
8	actctacttc	tumor	2
9	actgtccttc	tumor	3
10	actctccttc	tumor	4

	a	c
Normal	0.2 +D	0.3 - D
Tumor	0.2 - D	0.3 +D

D = 0.3 -0.2 = 0.1 > 0

Dmax= min (0.3, 0.2) =0.2

D' = 0.1/0.2 = 0.5 (?) significant 한지는 모르겠음

⇒ Chi square test (p=0.067889) : significant 하지 않음

Permutation
(False-discovery rate)

observed frequency

	a	c	Total
Normal	3	2	5
Tumor	1	4	5
Total	4	6	10

observed probability

	a	c
Normal	0.3	0.2
Tumor	0.1	0.4

little increased ⇒ associated 있다 하지만 significant 하지는 않음

Expected

Pn = 5/10 = 0.5 (normal 5명)
Pt = 5/10 = 0.5
Pa = 4/10 = 0.4 (4개)
Pc = 6/10 = 0.6

Expected

	a	c
Normal	0.5x0.4	0.5x0.6
Tumor	0.5x0.4	0.5x0.6

Expected

	a	c
Normal	0.2	0.3
Tumor	0.2	0.3

The Genome-wide Association Study (GWAS)

① design my set first

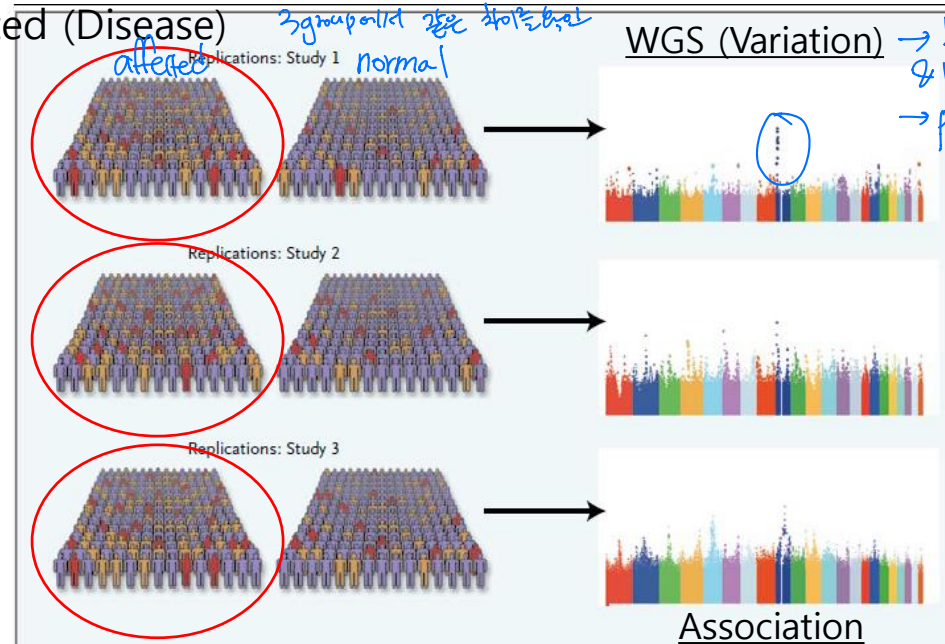
Affected (Disease)

Case-control designs

Disease

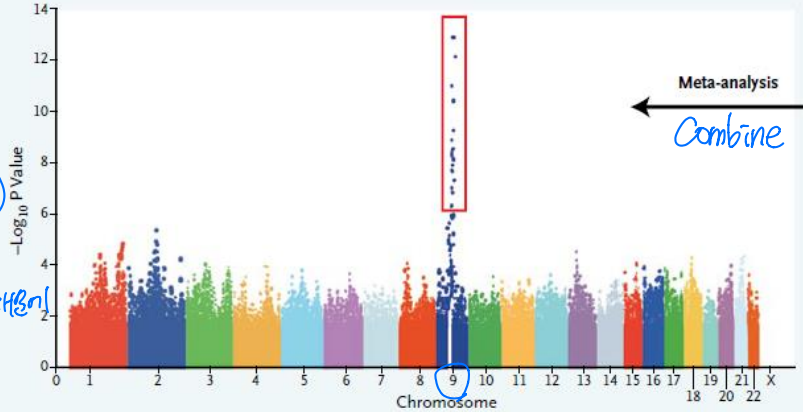
More defined case

Family based designs (Trio)



Quantitative trait loci (QTLs)

Statistical significance



Samples of GWAS

Case-control designs (Disease vs Normal)

easy way to grouping

- A difference in the frequency of an allele or genotype of the polymorphism under test between the two groups indicates that the genetic marker may increase risk of the disease or likelihood of the trait, or be in linkage disequilibrium with a polymorphism which does.
-Haplotypes can also show association with a disease or trait.

-Problem: genotype and haplotype frequencies vary between ethnic or geographic populations, possibly due to different ancestry.
-> population stratification , population structure

↳ different sampling 때문에 발생 : background frequency가 달라짐 (인종, 지역 등)

Family based designs (Trio)

↳ Because all kind of SNPs are inherited from parent

안양방식 ⇒ Collect patient & their genotype

Cohort: a group of subjects with a common defining characteristic

limiting area and relate with genotype

Tracks two or more groups forward from exposure to outcome

코호트연구 :특정 요인에 노출된 집단과 노출되지 않은 집단을 추적하고 연구 대상 질병의 발생률을 비교하여 요인과 질병 발생 관계를 조사하는 연구방법

Population stratification

Population 1	Population 2
freq $(A) = 0.8$	freq $(A) = 0.3$
freq $(a) = 0.2$	freq $(a) = 0.7$
Diabetics = 20%	Diabetics = 10%

combined case

over rated

under rated

0.8/(0.8+0.3)

$$P(A) \text{ diabetic} = (0.73 \times 20\% + 0.27 \times 10\%) = 17.3\%$$

$$P(a) \text{ diabetic} = (0.22 \times 20\% + 0.78 \times 10\%) = 12.2\%$$

> group에 따라 달라짐

Stratification may be environmental, cultural, or genetic

-Problem: genotype and haplotype frequencies vary between ethnic or geographic populations, possibly due to different ancestry.
-> population stratification

EXERCISE 3.3 *Perform a case-control association test*

Two parallel association studies between a candidate gene and skin cancer are performed in New York City and Miami, Florida. The numbers of cases and controls who are homozygous for a 3-bp insertion or deletion polymorphism or who are heterozygous are as follows:

Q. Is this deletion related to skin cancer?

	New York			**Miami**		
	Insertion	*Hetero-zygous*	*Deletion*	*Insertion*	*Hetero-zygous*	*Deletion*
w/ cancer *Case*	221	198	55	171	236	83
w/o cancer *Control*	279	212	35	189	244	77

Calculate the probability of association between the deletion polymorphism and the disease for each population separately and for the combined population, using the formula:

Chi square test

$$\chi^2 = N\,[N(r_1 + 2r_2) - R(n_1 + 2n_2)]^2 / RS[N(n_1 + 4n_2) - (n_1 + 2n_2)^2]$$ → 식을 알려주는 X

Discuss the possible reasons for any discrepancy in the conclusions from the two studies.

N=1,000

New York observed **Miami**

P(in) =0.5

474		Hetero-zygous	Deletion
Case	221	198	55
Control	279	212	35

490		Hetero-zygous	Deletion
Case	171	236	83
Control	189	244	77

P(in) =0.36

$$\chi^2 = N\,[N(r_1 + 2r_2) - R(n_1 + 2n_2)]^2 / RS[N(n_1 + 4n_2) - (n_1 + 2n_2)^2]$$

Quantity	New York	Miami	Combined
r_1	198	236	434
r_2	55	83	138
n_1	410	480	890
n_2	90	160	250
N	1,000	1,000	2,000
R	474	490	964
S	526	510	1,036
	7.6	0.8	6.9
	P<0.01		P<0.01

Deletion 이 Cancer 연관있다 하기 어려움

Plugging these values into the formula results in χ^2 values of 7.6, 0.8, and 6.9 for New York, Miami, and the combined population, respectively. Prima facie this suggests little effect of the polymorphism in Miami, but suggests a significant effect at p < 0.01 in New York and across the two populations.

New York

P(in)=((221+279)/1000) =0.5 (insertion)
P(het)=((198+212)/1000) =0.41 (heterozygote)
P(del)=((55+35)/1000)=0.09 (deletion)
P(case)=((221+198+55)/1000) =0.474 (case)
P(cont)=((279+212+35)/1000) =0.526 (control)

Expected

P(case)xP(in)	P(case)xP(het)	P(case)xP(del)
P(cont)xP(in)	P(cont)xP(het)	P(cont)xP(del)

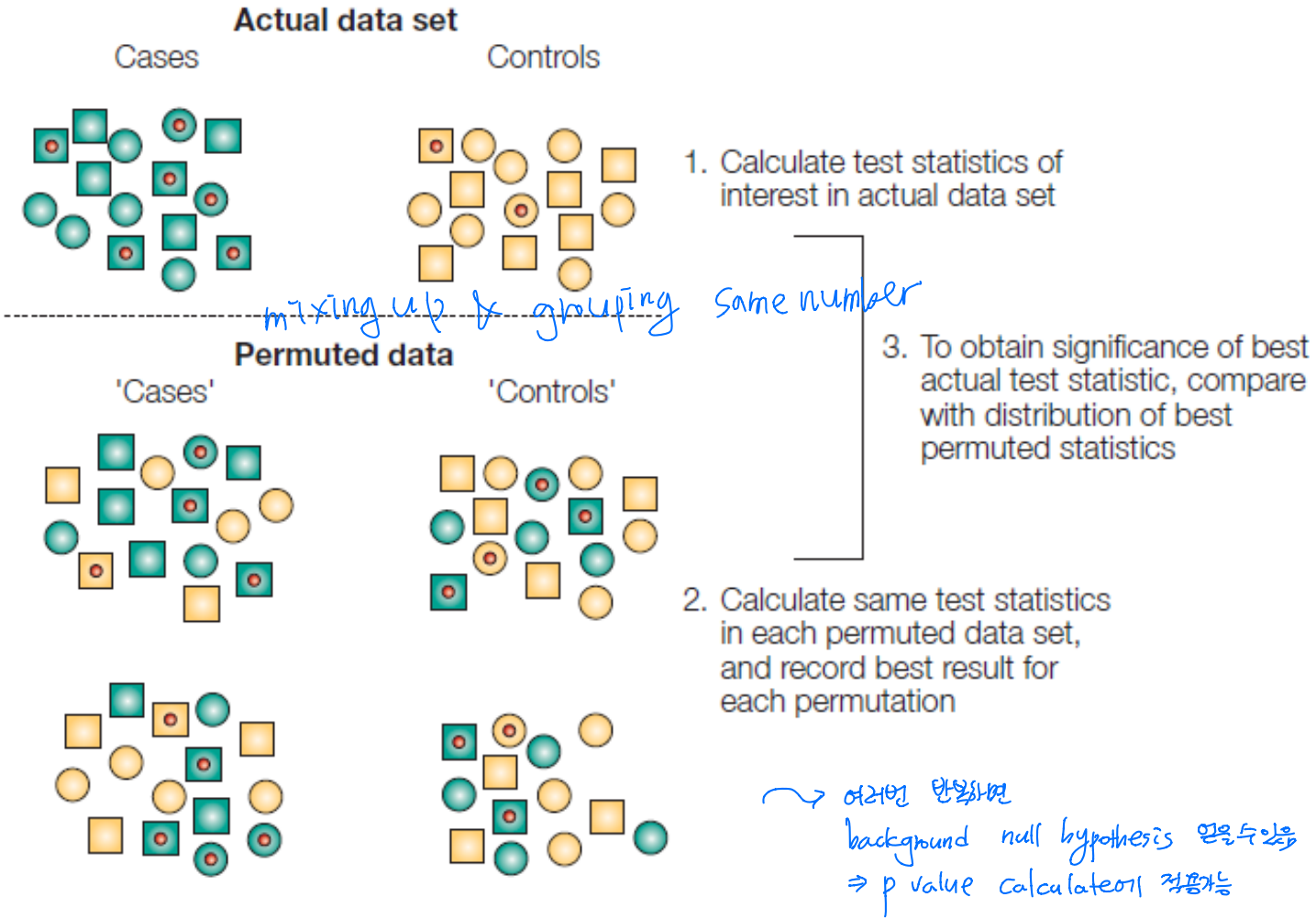
237	194	42
263	216	47

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

χ^2 = the test statistic $\sum$ = the sum of

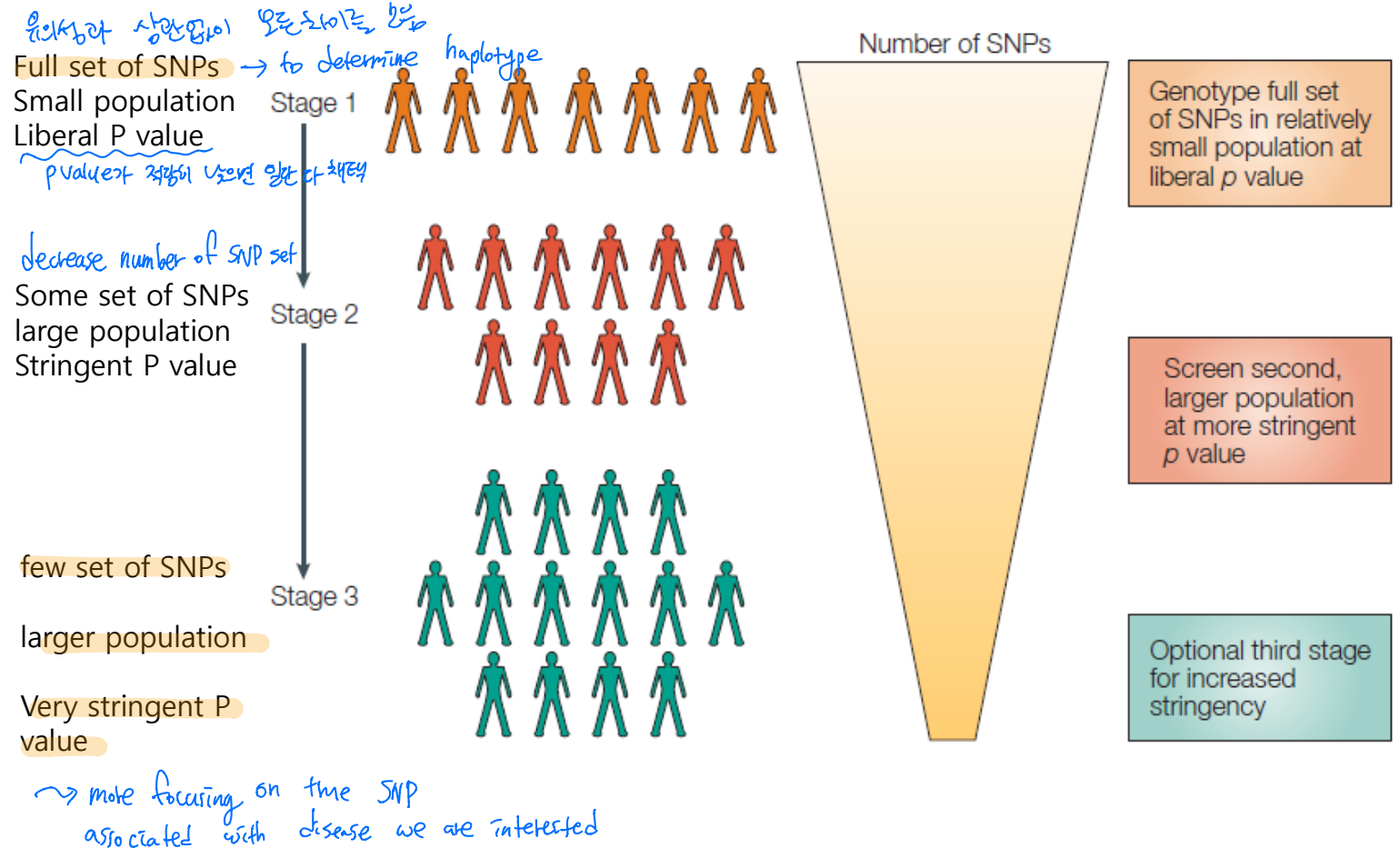
O = Observed frequencies E = Expected frequencies

Permutation testing

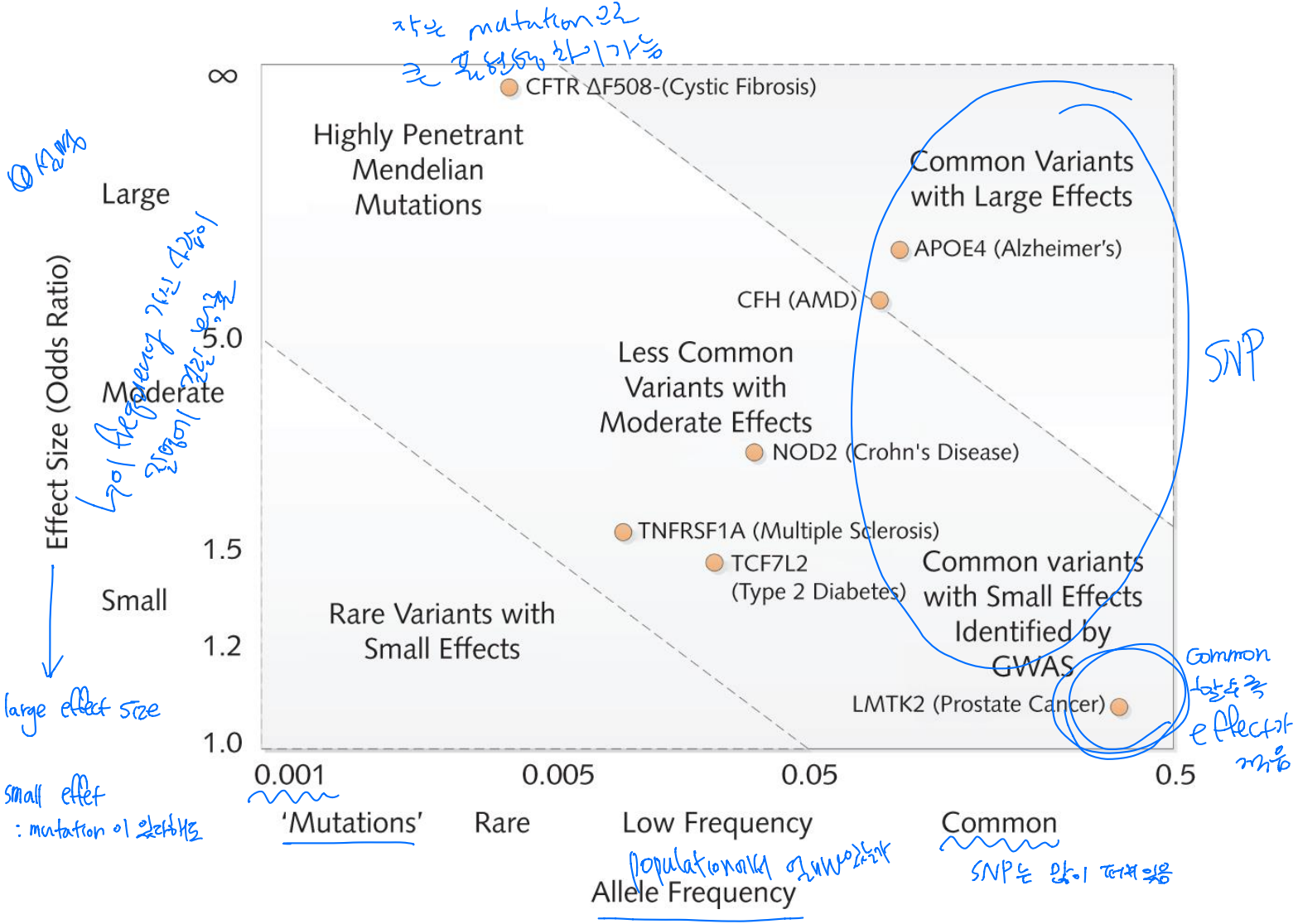


↳ 여러번 반복하면 background null hypothesis 만족두만족
⇒ p value calculation 가능

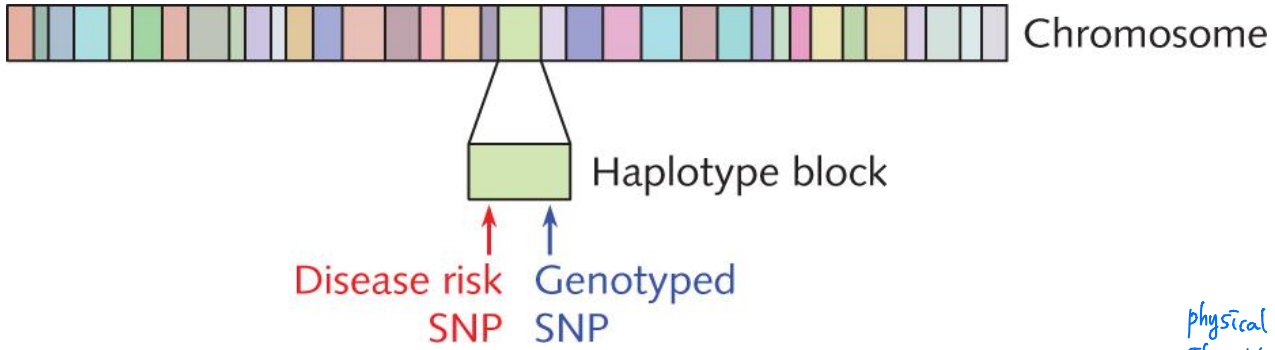
Using a multistage approach to minimize sample sizes.



Allele frequency (variation rate) vs. effect size (penetrance)



GWAS : loci associated with traits



physical characteristic (height, blood pressure)

Color codes refer to categories of traits

Ansan Cohort (GWAS in Korea)

Korea Association REsource (KARE)

NGS가 없을때에는 SNP chip을 이용해 SNP tracking

	Baseline	1st follow-up	2nd follow-up	3rd follow-up	4th follow-up	5th follow-up
Ansung	5,018	4,580 (91.3)	3,975 (79.2)	3,434 (68.4)	3,403 (67.8)	3,186 (63.5)
Ansan	5,020	4,023 (80.1)	3,540 (70.5)	3,255 (64.8)	3,262 (65.0)	3,052 (60.8)
Total	10,038	8,601 (85.7)	7,594 (74.9)	7,594 (66.6)	6,688 (66.4)	6,688 (62.1)

- Participants (follow-up rate)
- Two cohorts: Ansung (rural area) & Ansan (urban area)

height, blood pressure 등 all kind of data following study

nature genetics

A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits

Yoon Shin Cho[1], Min Jin Go[1], Young Jin Kim[1], Jee Yeon Heo[1], Ji Hee Oh[1], Hyo-Jeong Ban[1], Dankyu Yoon[2], Mi Hee Lee[1], Dong-Joon Kim[1], Miey Park[1], Seung-Hun Cha[1], Jun-Woo Kim[1], Bok-Ghee Han[1], Haesook Min[1], Younjhin Ahn[1], Man Suk Park[1], Hye Ree Han[1], Hye-Yoon Jang[3], Eun Young Cho[3], Jong-Eun Lee[3], Nam H Cho[4], Chol Shin[5], Taesung Park[2,6], Ji Wan Park[7], Jong-Keuk Lee[8], Lon Cardon[9], Geraldine Clarke[10], Mark I McCarthy[10,11], Jong-Young Lee[1], Jong-Koo Lee[12], Bermseok Oh[1,13] & Hyung-Lae Kim[1]

To identify genetic factors influencing quantitative traits of biomedical importance, we conducted a genome-wide association study in 8,842 samples from population-based cohorts recruited in Korea. For height and body mass index, most variants detected overlapped those reported in European samples. For the other traits examined, replication of promising GWAS signals in 7,861 independent Korean samples identified six previously unknown loci. For pulse rate, signals reaching genome-wide significance mapped to chromosomes 1q32 (rs12731740, $P = 2.9 \times 10^{-9}$) and 6q22 (rs12110693, $P = 1.6 \times 10^{-9}$), with the latter ~400 kb from the coding sequence of *GJA1*. For systolic blood pressure, the most compelling association involved chromosome 12q21 and variants near the *ATP2B1* gene (rs17249754, $P = 1.3 \times 10^{-7}$). For waist-hip ratio, variants on chromosome 12q24 (rs2074356, $P = 7.8 \times 10^{-12}$) showed convincing associations, although no regional transcript has strong biological candidacy. Finally, we identified two loci influencing bone mineral density at multiple sites. On chromosome 7q31, rs7776725 (within the *FAM3C* gene) was associated with bone density at the radius ($P = 1.0 \times 10^{-11}$), tibia ($P = 1.6 \times 10^{-6}$) and heel ($P = 1.9 \times 10^{-10}$). On chromosome 7p14, rs1721400 (mapping close to *SFRP4*, a frizzled protein gene) showed consistent associations at the same three sites ($P = 2.2 \times 10^{-3}$, $P = 1.4 \times 10^{-7}$ and $P = 6.0 \times 10^{-4}$, respectively). This large-scale GWA analysis of well-characterized Korean population-based samples highlights previously unknown biological pathways.

Positional cloning of a candidate complex disease

noninsulin-dependent diabetes mellitus

(A) low resolution

Linkage mapping	7 cM at Chr 2q37
Construct physical contig	~1.7 Mb, 22 genes
Low resolution case-control association scan	~100 kb, 3 genes
Sequence 10 individuals over 66 kb	179 variants, including 161 SNPs
Strong association with SNP 43	Calpain 10, Intron 3
Haplotype tests → complexity	3 SNP heterozygotes
Population attributable risk	4% Northern Europeans 14% Mexican Americans

large chunk of genetic region

narrow down region

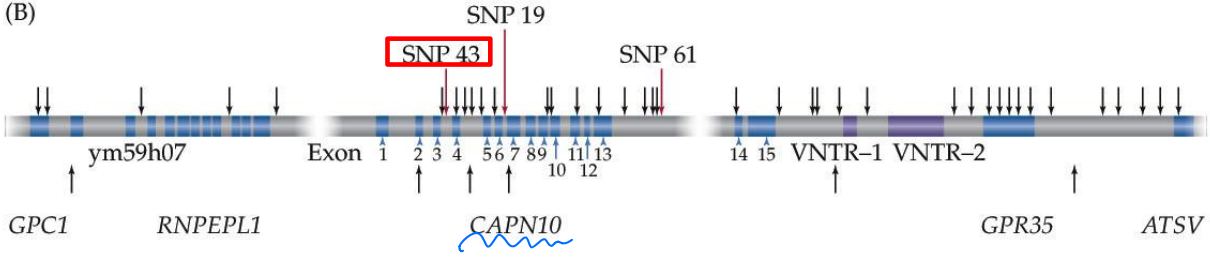
positional cloning : classical way

recent way to identify SNPs - disease

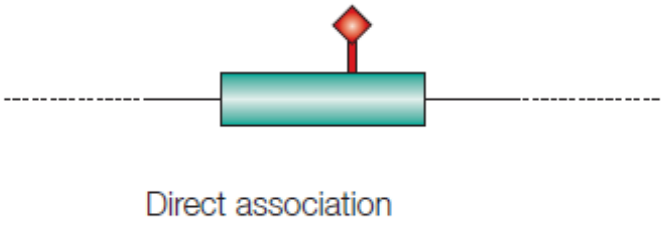
SNP affects binding to nuclear factor in human pancreatic extracts and transcriptional activation of a reporter gene in a cell line

Hypothesis: Calpain 10 protease involvement in NIDDM

(B)

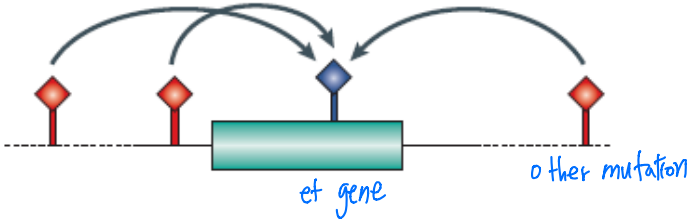


Testing SNPs for association by direct and indirect method



Direct association

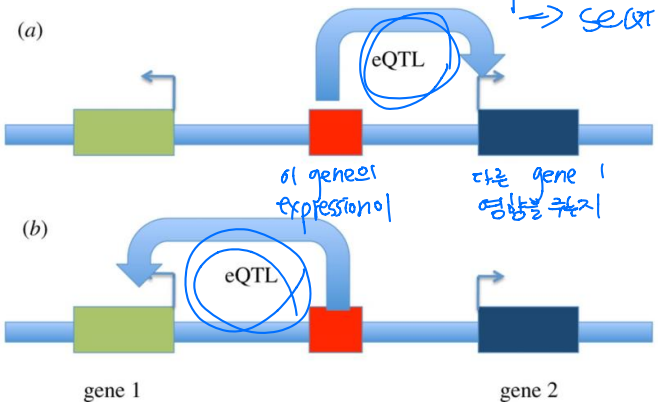
Alteration of amino acids
Change in protein stability & function



Indirect association

Alteration of neighboring gene expression

Alteration of splicing of neighboring genes.



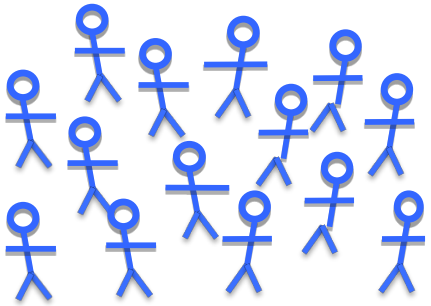
Expression quantitative trait loci (eQTLs) are genomic loci that contribute to variation in expression levels of mRNAs

Variation ↔ (linked) Phenotype

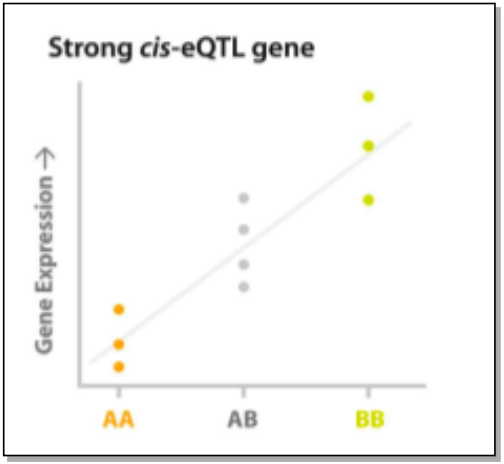
↕ linked

Gene expression – eQTLs

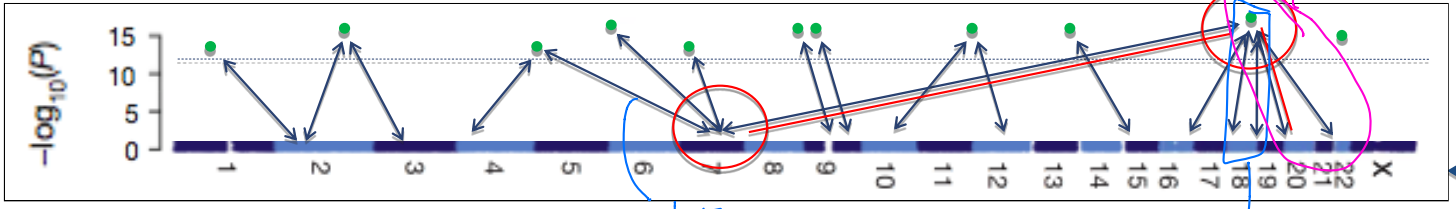
A Quantitative Gene-Expression Association (eQTL)



cDNA Levels



Expression Quantitative Trait Loci (eQTLs)

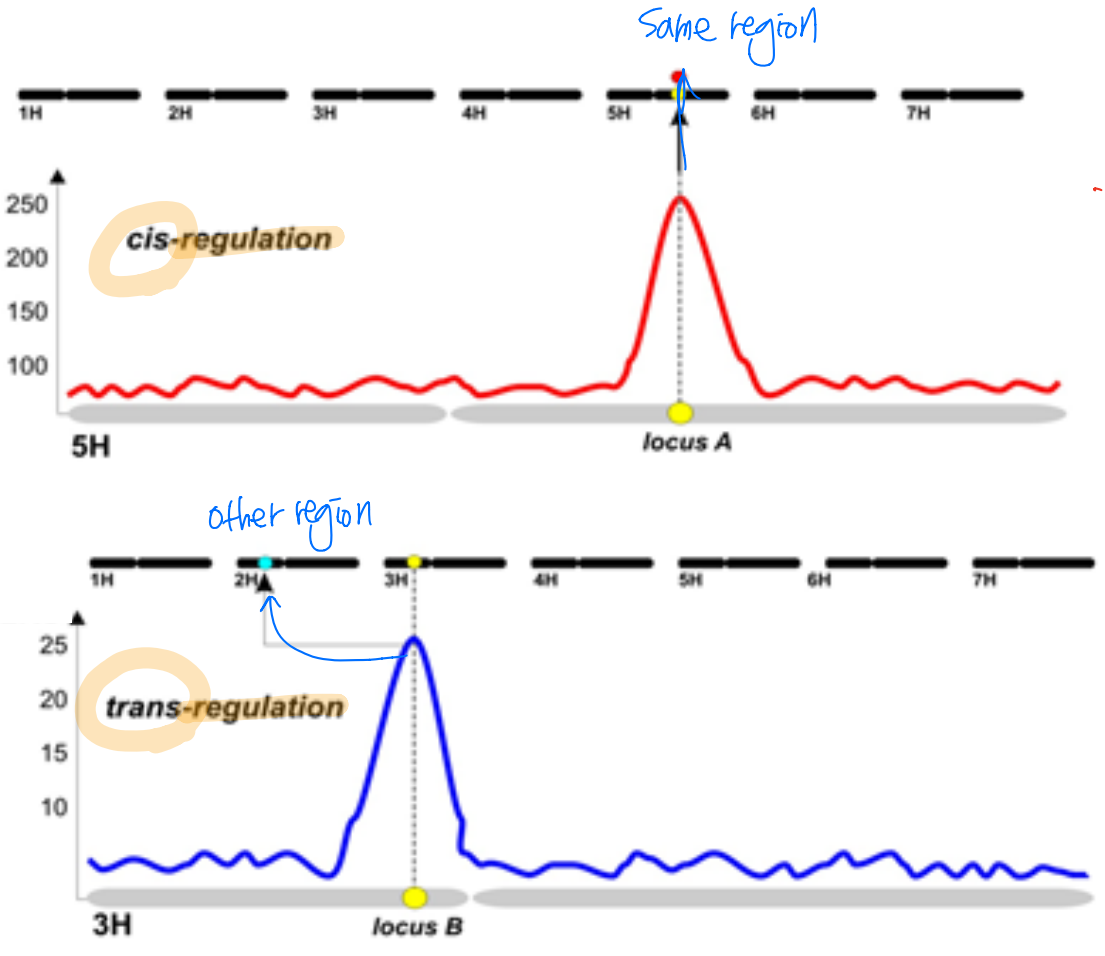


다른 chr에 있는 gene들

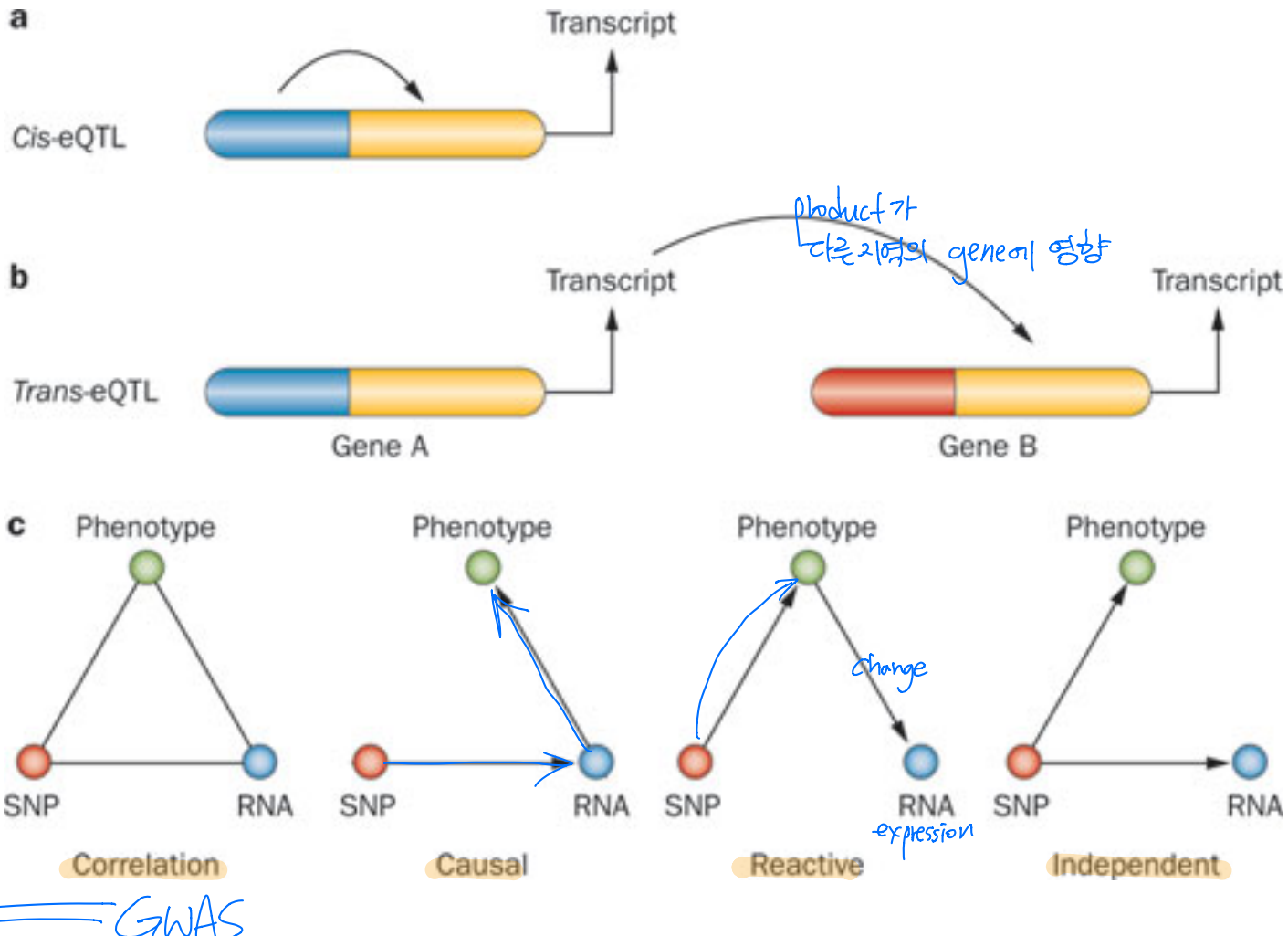
연관됨 : associated 되어있다.

같은 chromosome에 있는 gene을 affect 하기도하지만

A Quantitative Gene-Expression Association (eQTL)



A Quantitative Gene-Expression Association (eQTL)



Phylogenetic analysis

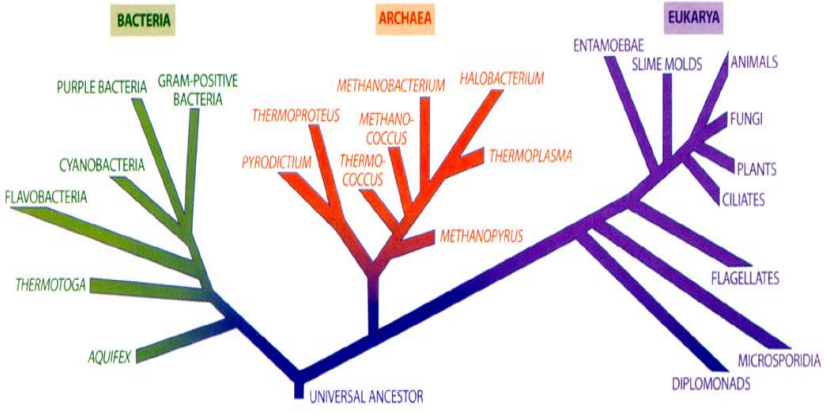
Sung Wook Chi
Division of Life Sciences, Korea University

Phylogenetics

phylogenetics is the study of evolutionary relationships among groups of organisms (Phylogeny), which are discovered through molecular sequencing data

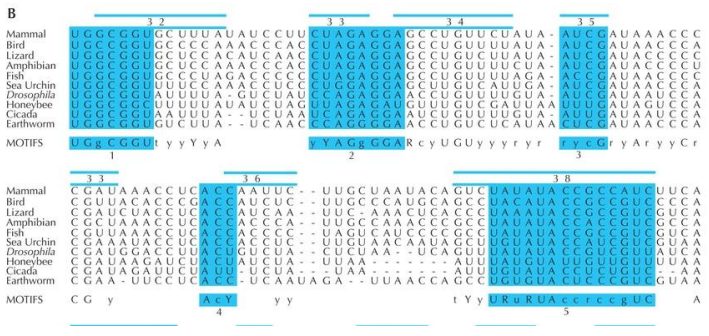
use to categorize

A version of the "tree of life" : Phylogenetic tree



DNA : morphological

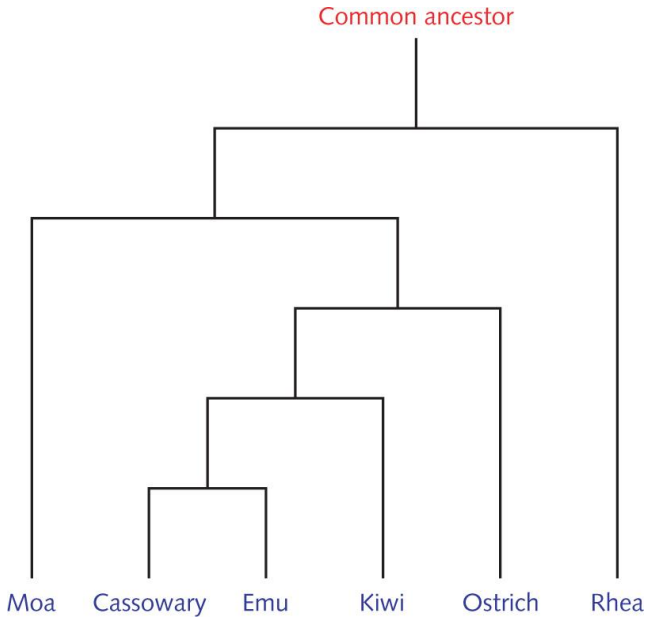
M. Madigan and B. Marrs, 1997

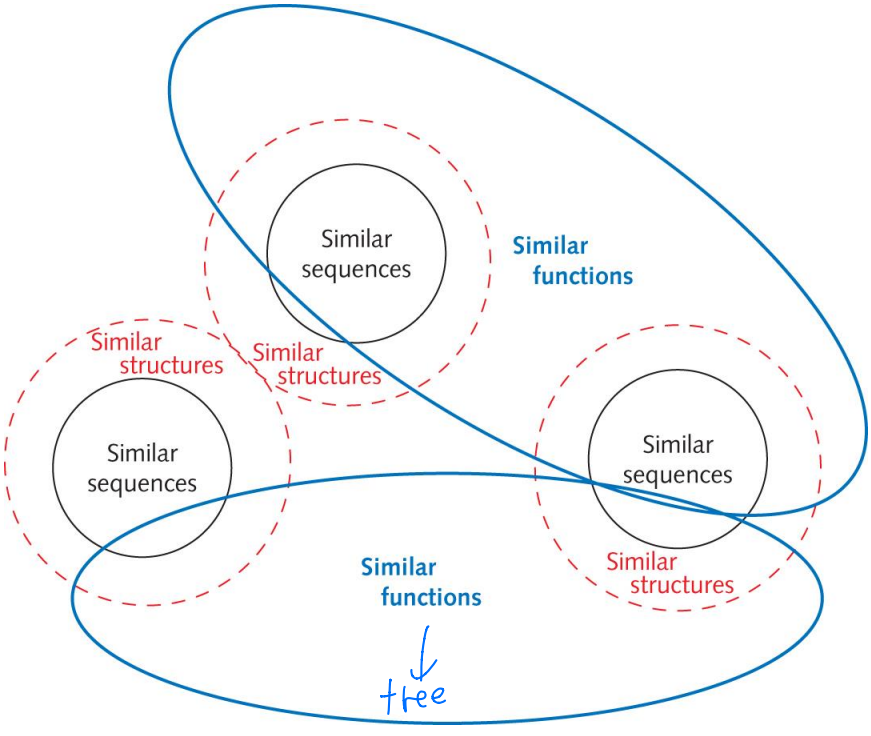


DNA information 바탕으로 계통분류

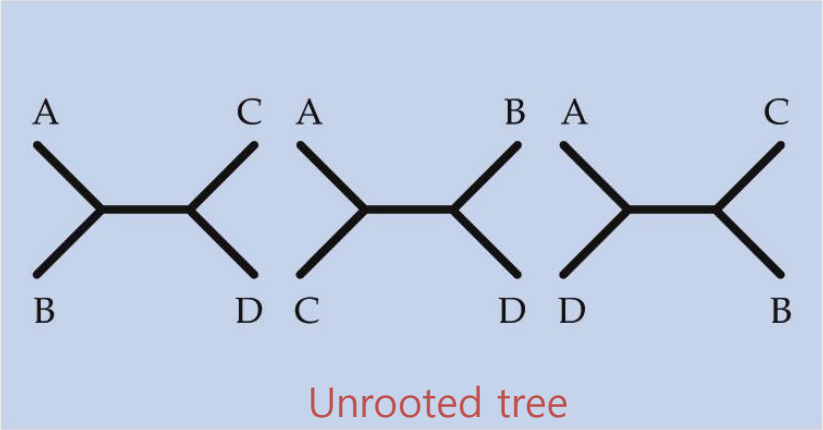
Obtained from aligned sequences of ribosomal RNA

Phylogenetic analysis

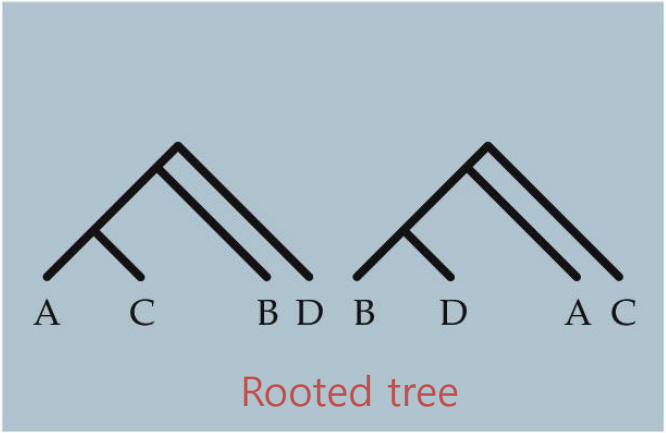




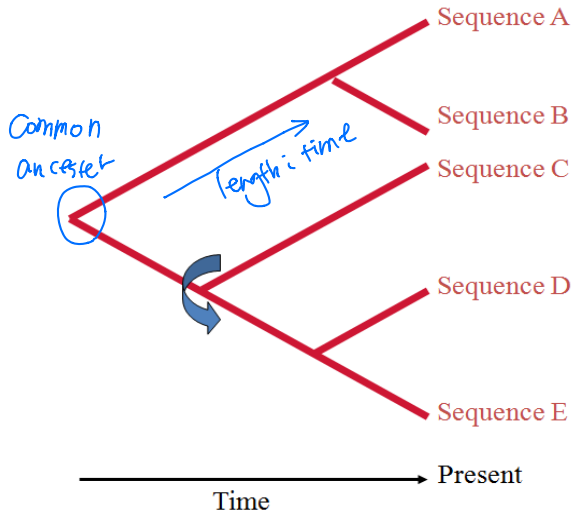
Phylogenetics (evolutional tree)

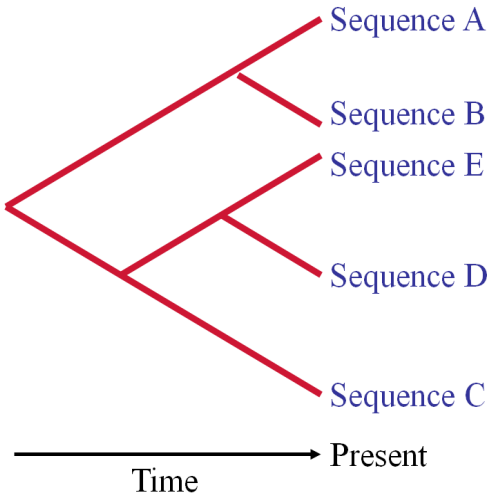


Unrooted tree



Rooted tree





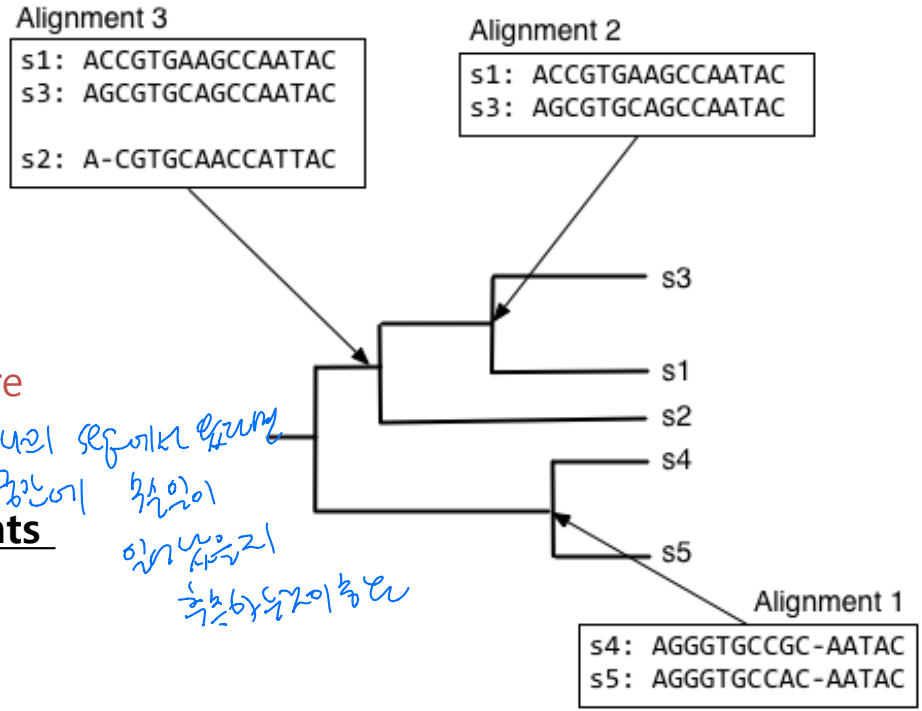
Estimated from mutation frequency / distance

Sequence Alignment, distances & tree

```
s1: ACCGTGAAGCCAATAC
s3: AGCGTGCAGCCAATAC

s2: A-CGTGCAACCATTAC

s4: AGGGTGCCGC-AATAC
s5: AGGGTGCCAC-AATAC
```

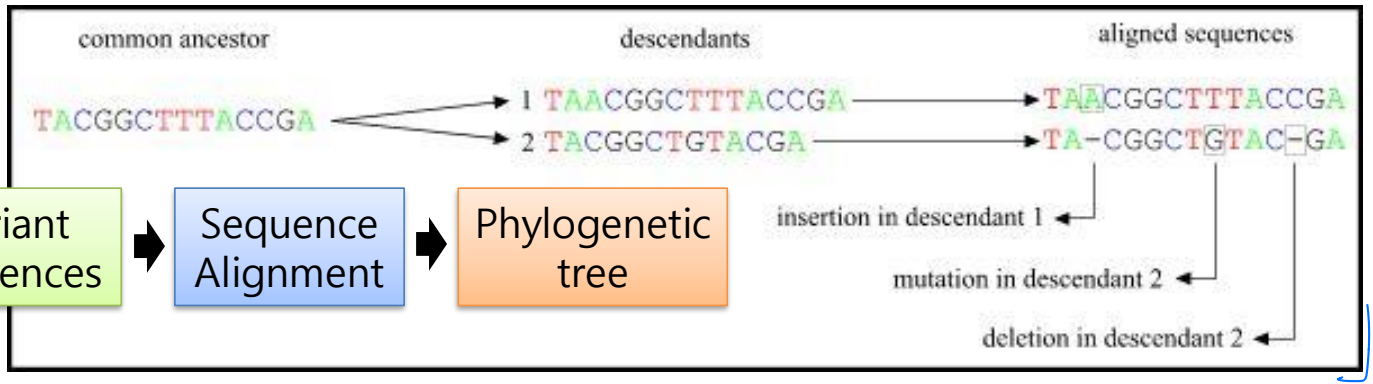


- Sequences to be aligned are phylogenetically related

Perform pair-wise alignments

Measure distances

Generate tree



Variant sequences → Sequence Alignment → Phylogenetic tree

Phylogenetic tree reconstruction

- Clustering Method: distance-based
  - UPGMA
  - Neighbor-joining

> distance를 활용한 일반적인 방법

- Cladistic method: Character state-based methods
  - Maximum parsimony

- Maximum Likelihood

여러분께 맡기겠다~

Clustering Method : UPGMA

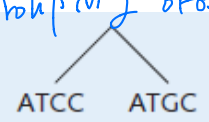
UPGMA (unweighted pair group method with arithmetic mean)

Distance matrix

hamming distance ⇒ 그냥 다른거

	ATCC	ATGC	TTCG	TCGG
ATCC	0	1	2	4
ATGC		0	3	3
TTCG			0	2
TCGG				0

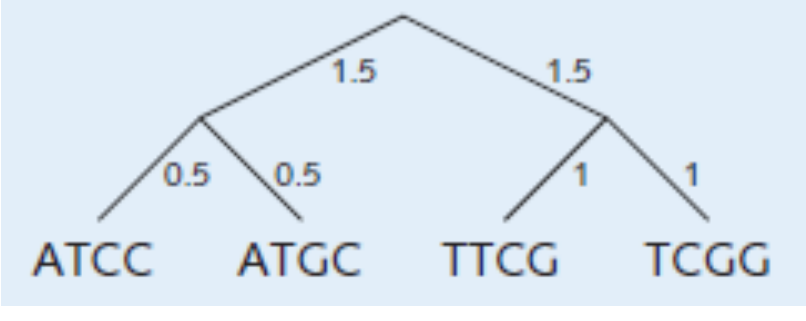
제일 grouping 하기 쉬운 것



The reduced distance matrix is:

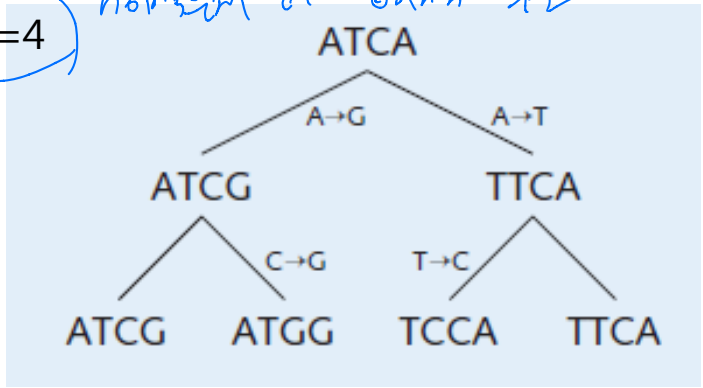
	{ATCC, ATGC}	TTCG	TCGG
{ATCC, ATGC}	0	$\frac{1}{2}(2+3)=2.5$	$\frac{1}{2}(4+3)=3.5$
TTCG		0	2
TCGG			0

두 그룹에서의 거리의 산술적평균

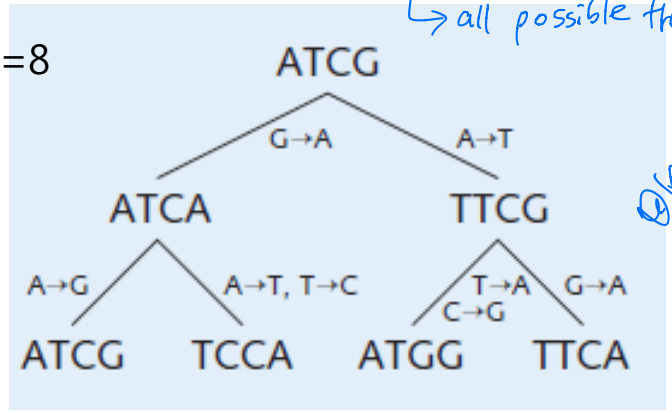


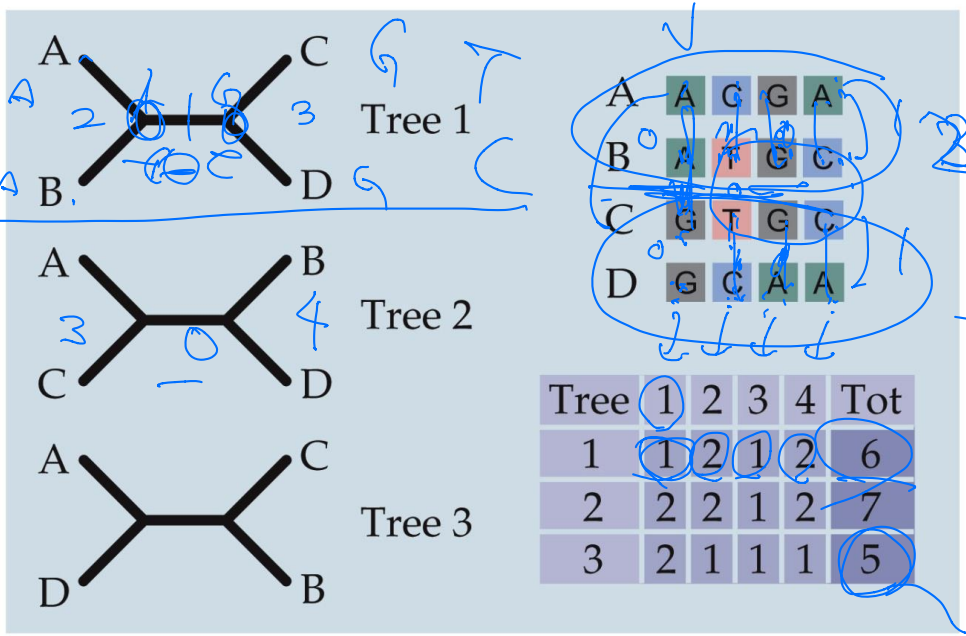
Cladistic Method : Maximum Parsimony

n=4



n=8





Maximum Parsimony (exercise)

EXERCISE 2.4 *A simple phylogenetic analysis*

A few years ago, a pressing question in the area of human evolution was the exact phylogenetic relationship between humans, chimpanzees, and gorillas. Fossil, morphological, and early molecular data provided conflicting and inconclusive results as to which pair of organisms were most closely related. Suppose the following four short DNA sequences are available, including a sequence from the outgroup organism, orangutan:

↳ most divergent
⇒ root

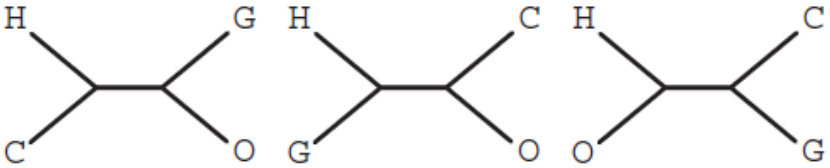
Human	CGAAATGCAT
Chimpanzee	CGGACTGTAT
Gorilla	CAAGCTGTAC
Orangutan	TAAGACGTAG

Carry out a phylogenetic analysis to infer the rooted evolutionary tree of human, chimpanzee, and gorilla.

Maximum Parsimony (exercise)

Human	CGAAATGCAT
Chimpanzee	CGGACTGTAT
Gorilla	CAAGCTGTAC
Orangutan	TAAGACGTAG

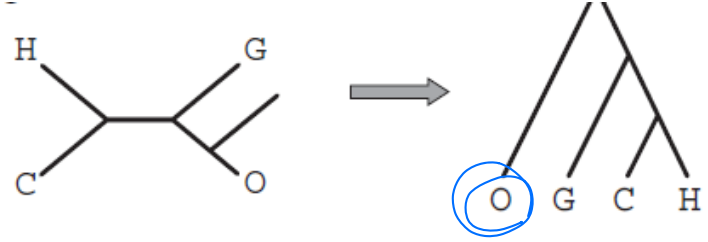
all possible unrooted tree



	1234567890	Total
Tree A	1111210102	10
Tree B	1212210102	12
Tree C	1212110102	11

like to be true

calculate edit distance



rooted tree

considering outgroup

Multiple Sequence Alignment

- Idea: Sequences to be aligned are phylogenetically related
  - these relationships are used to guide the alignment
- Popular implementations: **CLUSTALW,**

1. **Perform pair-wise alignments** between all pairs of sequences :(n x (n-1)/2 possibilities)

2. **Generate distance matrix.**
   - Distance between a pair = number of mismatched positions in alignment divided by total number of matched positions

3. Generate a Neighbor-Joining '**guide tree**' from distance table

4. Use guide tree to progressively **align sequences in pairs from tips to root of tree.**

Multiple Sequence Alignment

Hbb_Human	1	-			
Hbb_Horse	2	.17	-		
Hba_Human	3	.59	.60	-	
Hba_Horse	4	.59	.59	.13	-
Myg_Whale	5	.77	.77	.75	.75	-

① Calculate distance

CLUSTAL W

Quick pair wise alignment
calculate distance matrix

② making tree

pairwise alignment

Neighbour-joining tree
(guide tree)

alpha-helices

```
1  PEEKSAVTALWGKVN--VDEVGG
2  GEEKAAVLALWDKVN--EEEVGG
3  PADKTNVKAAWGKVGAHAGEYGA
4  AADKTNVKAAWSKVGGHAGEYGA
5  EHEWQLVLHVWAKVEADVAGHGQ
```

2 1 3 4

Progressive alignment
following guide tree

Interpretation of phylogenetic tree: Orthologs and paralogs

Variant sequences

↓

Sequence Alignment

↓

Phylogenetic tree

- Recalculate / Infer
  - Distance or similarity b/w sequences

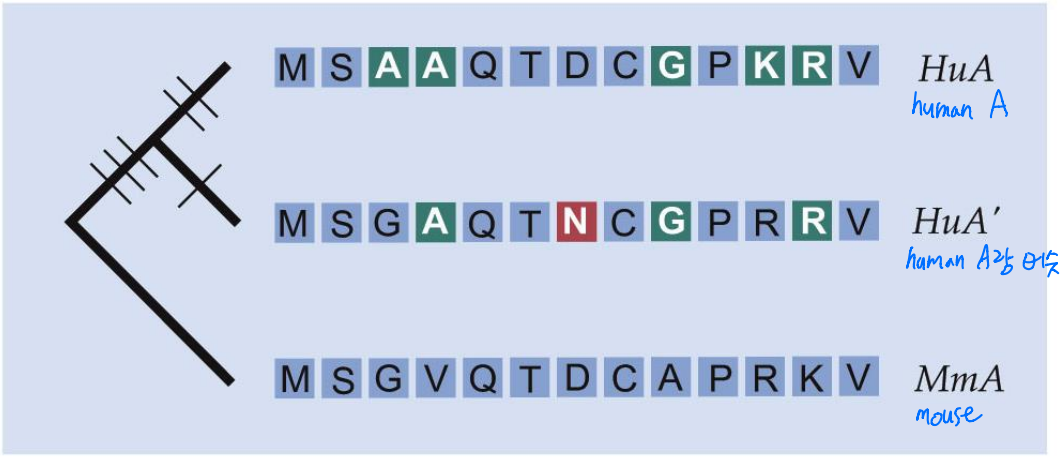
- Topology (order)
- Length (evolutional time)

↓

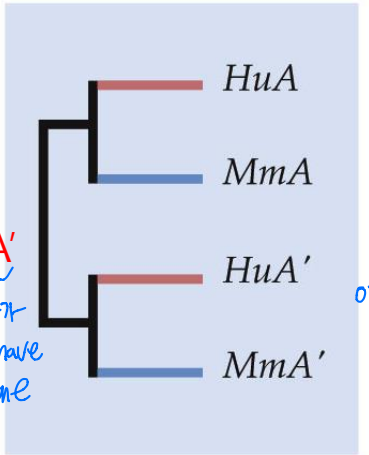
Evolutional relationship / interpretation

evolutionary event 에 관해 알려줌

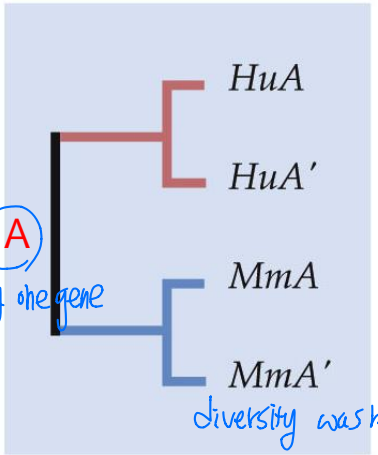
(A)



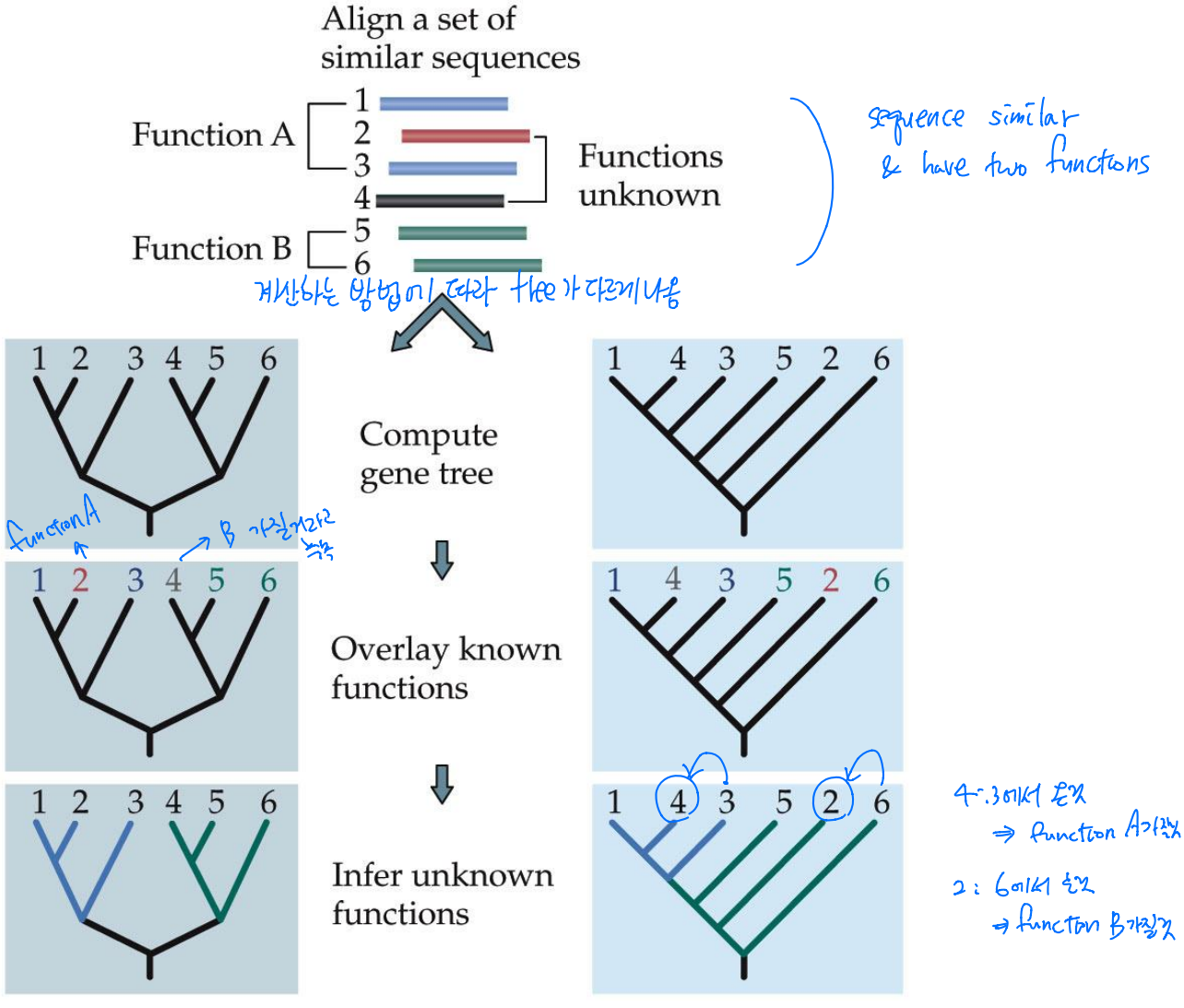
(B)



(C)

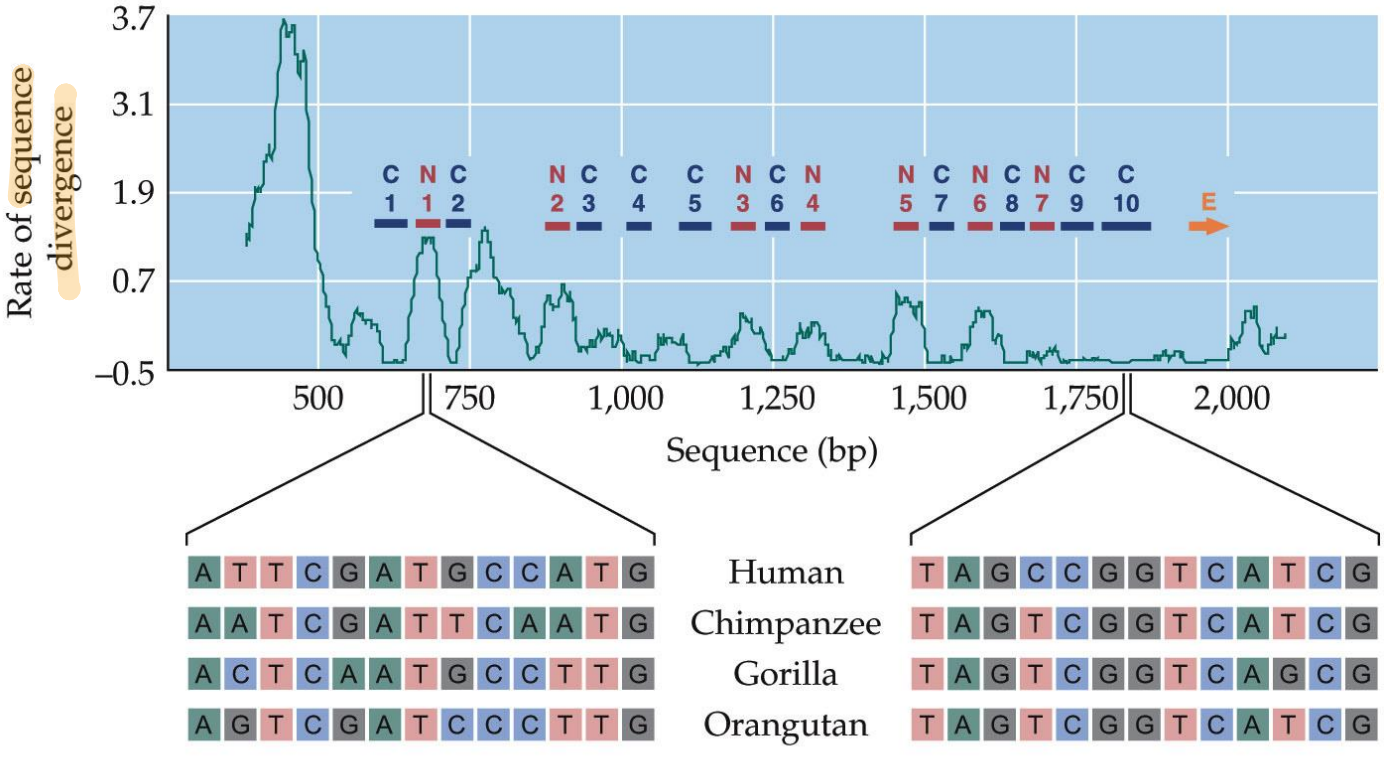


Inferring gene function from phylogenetic analysis



Phylogenetic shadowing

Within closely related species



진화적 보존을 확인 가능

: more essential

↳ regulatory element 조절하는 지역일 것

Phylogenetic footprinting / evolutional footprinting

Within distantly related species