Summary & Review

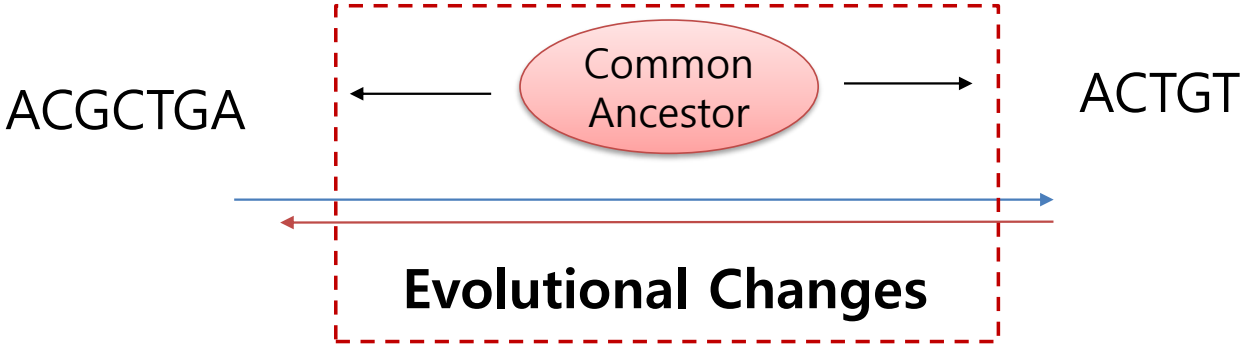
Genome Sequence Analysis

1) Sequence alignment
   - Purpose of sequence alignment
   - Principles of pair-wise sequence alignment
   - Sequence alignment for database search: BLAST

2) Sequence variation analysis haplotyping
   - Pre-required basic knowledge of human genome

Sung Wook Chi
Division of Life Sciences, Korea University

Biological question from sequence alignment



ACGCTGA

```
ACGCTGA
A--CTGT
```

Sequence alignment 1

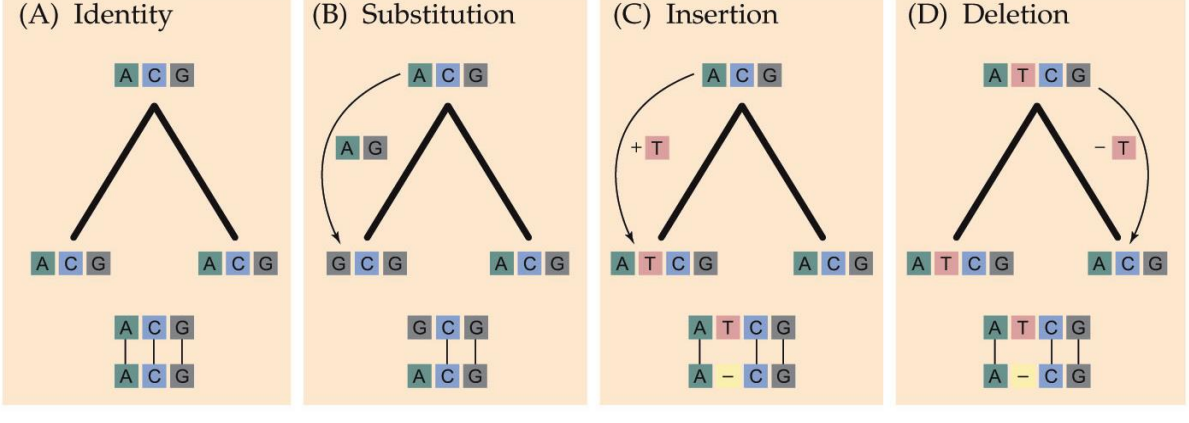
ACTGT

ACGCTGA

```
ACGCTGA
ACTGT--
```

Sequence alignment 2

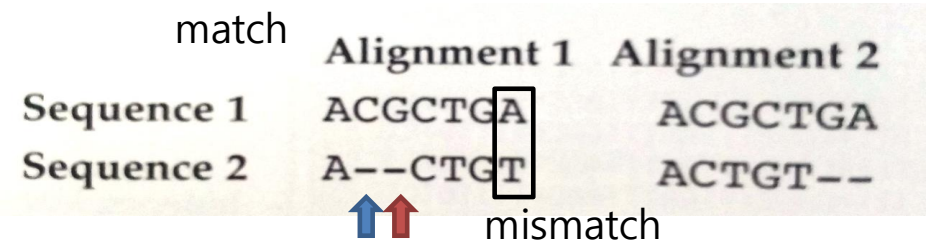
ACTGT



What is the best optimal alignment?

Sequence Alignment

Pairwise sequence alignment



Optimal sequence alignment

1. Evaluation of sequence similarity : **Distance**, **Score**
2. Performing optimal sequence alignment search: **Dynamic Programming**

➢ **Global Alignment**

```
TCAG--T-GTCGAAGT-TA
|  ||   |  |  |  |  ||  |
T-AGGCTAG-C-A-GTGTA
```

➢ **Local Alignment**

```
TCAGTGTCGAAGTTA
|||||||
TAGGCTAGCAGTGTA
```

Global alignment

Conserved region of sequence
> Functional domain / element

ex) kinase

비교법

Local alignment

High sequence similarity > Homolog > same function

Similar seq and

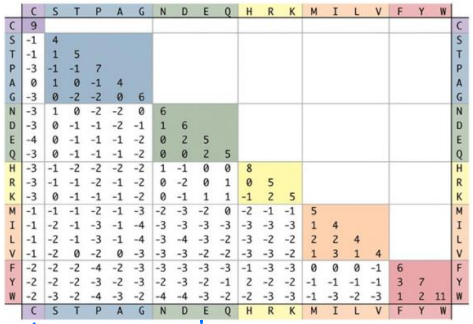
How to measure sequence similarity : Score

Score = (match or mismatch penalty) + gap penalty

Match = 3 , Mismatch = -1

Gap= -2

ACGCTGA
A--CTGT



	A	G	T	C
A	20	10	5	5
G	10	20	5	5
T	5	5	20	10
C	5	5	10	20

↳ amino acid

Affine gap penalty
(gap opening + (gap extension) x (n-1))

high low gap panelty ⇒ local

transition : 퓨린 → 퓨린
transversion : 퓨린 → 피리미딘

DOROTHY--------HODGKIN
DOROTHYCROWFOOTHODGKIN

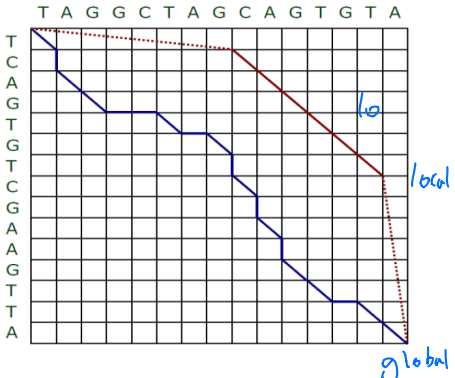
Local alignment
:Smith-Waterman algorithm

Global alignment
:Needleman-Wunsch algorithm

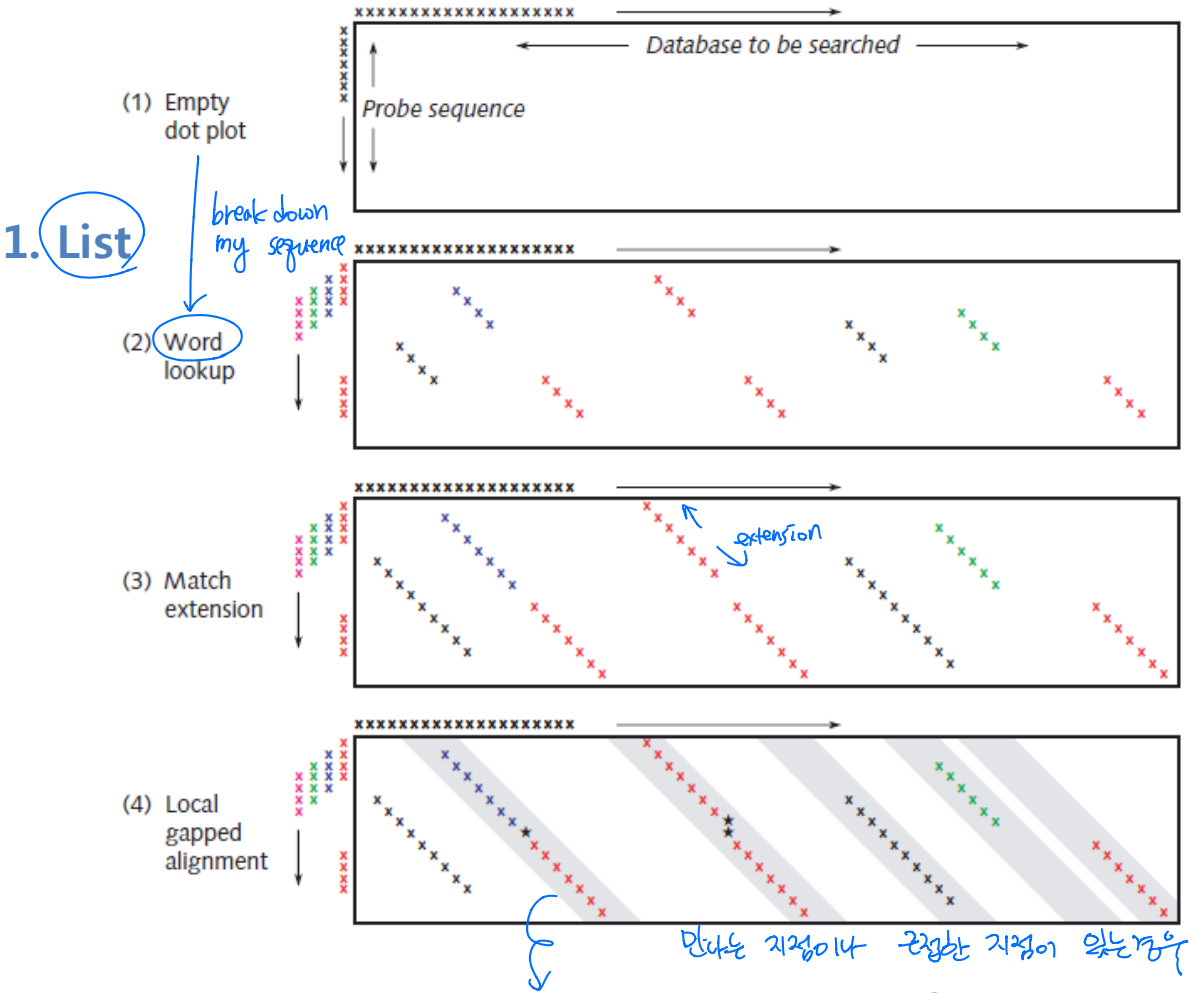
- Dot Plot : seq 형태를 비교하는데 유용

Repetitive sequence

-Dynamic Programming:



Sequence Database Search: BLAST



1. List

break down my sequence

2. Scan

3. Extend

until score is high

만나는 지점이나 근접한 지점이 있는경우

4. Statistical significance

→ 이런 애들이 랜덤하게 random sequences 에서 얼마나 자주 나타나는지

EXERCISE 2.3 *Perform a BLAST search* IL13

Using the gene that you identified in Exercise 1.1 perform a BLASTn search for homologs in other species. What conclusions can you reach regarding whether this gene is part of a gene family, or regarding its phylogenetic distribution?

ANSWER: *Continuing with our example of the* IL13 *gene from Chapter 1, copy the transcript sequence of this gene either from the Ensembl or GenBank page. (For Ensembl, click on the "Transcript info" link and the sequence appears at the bottom of the page. For GenBank, first select "Nucleotide"; then type "IL13" into the search box and follow the link to the mRNA annotation for NM 002188.) Return to the NCBI home page, click on "BLAST," bring up the "Nucleotide-nucleotide (blastn)" page, and paste your sequence into the search box. Choose some search options (for example, "all RefSeq mRNA sequences") and submit the request. Depending on the number of queries in the queue, this could take a minute or two. In the case of* IL13, *alignments are shown for human, chimp, macaque, canine, and equine matches to the 1280 nt search. If you select optimization for somewhat similar sequences rather than highly similar ones, very high match scores are also seen for several other mammalian* IL13 *loci. The closest match for the* Drosophila *genome, however, has an E-value of just 0.17, and even this score is due to a single perfect match—a 23-bp element—that is probably not indicative of a homologous gene. Thus* IL13 *appears to be unique to vertebrates (and possibly even to mammals, since no fish or bird sequences have significant E-values), and to be present in a single copy per genome.*

NCBI에서 mRNA 서열을 FASTA 포맷으로 받아서 BLASTn 돌려봤다.

homolog 를 찾은 상황

What we will learn today

Genome Annotation

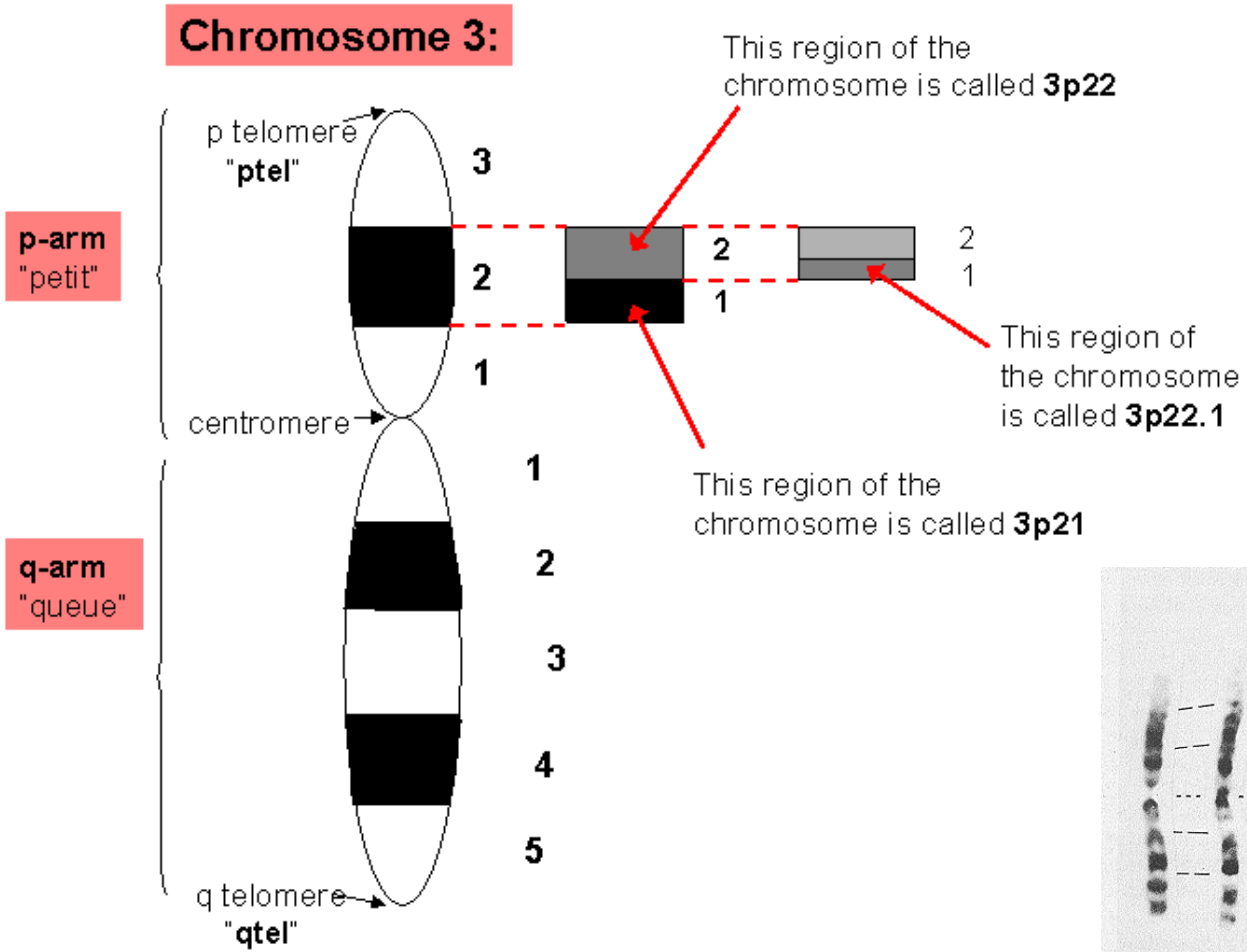
- Sequence alignment
- Linkage disequilibrium (LD), Haplotype, SNP

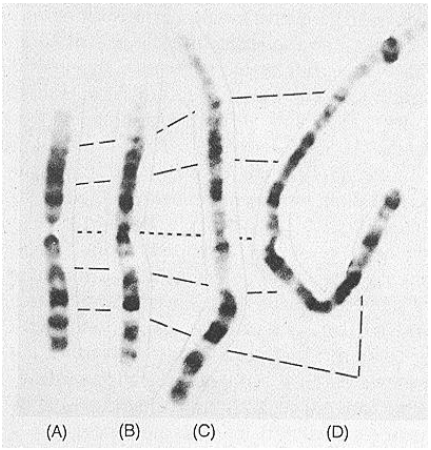
more focus on sequence

⇒ 표현형을 보는게 아니라 seq를 직접봐 linkage분석 (SNP 같이)

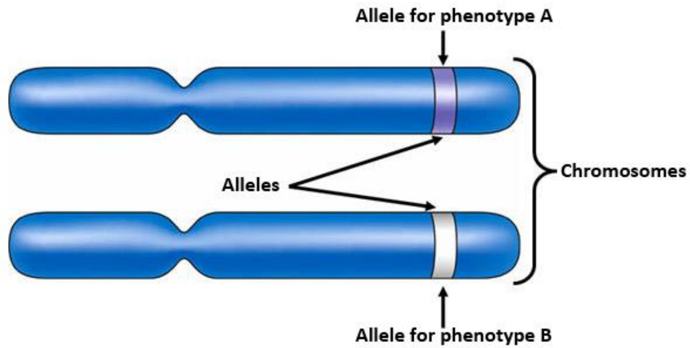
> Study of genomic variation
> (will be covered in next lecture)

Nomenclature of human chromosome



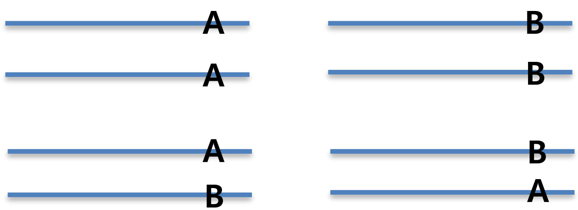


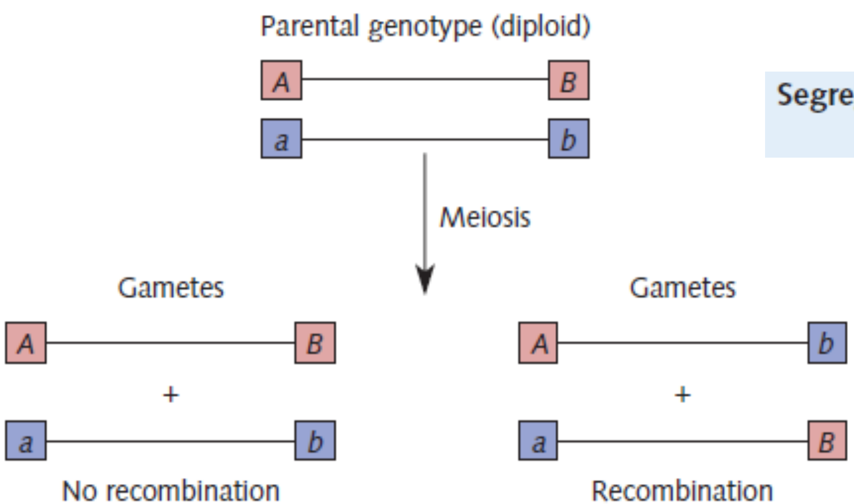
Human Genetics : recombination, linkage



- Allele, Genotype, Phenotype, Trait
- Homozygote, Heterozygote

AA / BB   AB / BA





Segregation	The separation of corresponding alleles during the reproductive process.

Independent assortment	The uncorrelated choices of genes for different characters that each parent transmits to children.
Linkage	Absence or reduction of independent assortment of parental genes, which are usually transmitted together because they lie on the same chromosome.

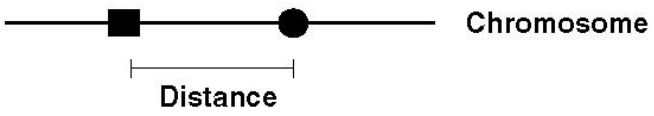
in same segregation block (chromosome)

서로다른 조합의 allele 들의 측정된 분포가 각각의
→ allele 독립유전 전달을 기반으로 했을 때 기대는 것과 다른 경우

Linkage disequilibrium (LD)

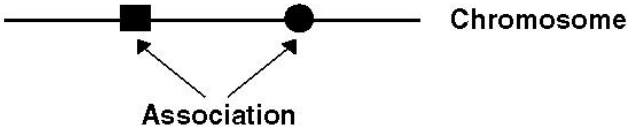
genetic map 지도 association → to understand gene block

RECOMBINATION RATE:



Mendelian inheritance
:alleles segregate independently

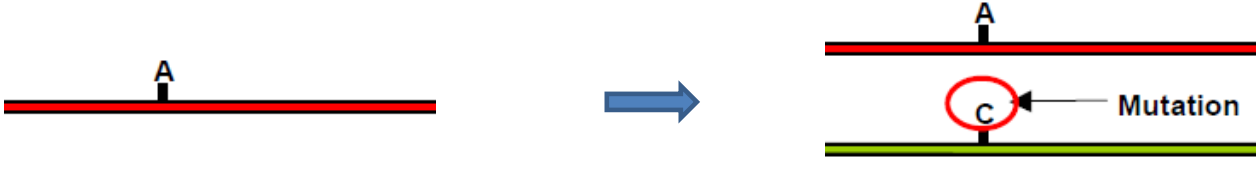
LINKAGE DISEQUILIBRIUM:



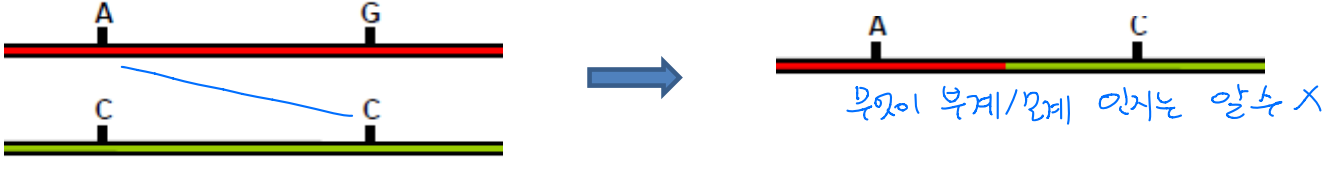
Two markers are in **linkage disequilibrium** when alleles at two or more loci **do not segregate independently**

Hardy-Weinberg disequilibrium

Alleles that exist today arose through ancient mutation events (at least inherited from parent)

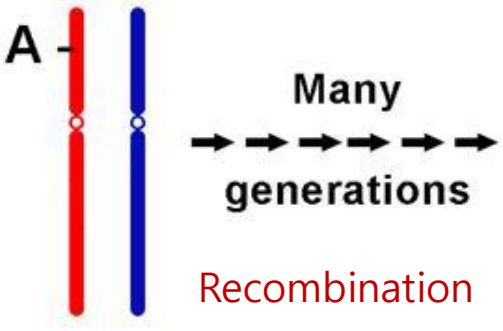


Recombination generates new arrangements for ancestral alleles

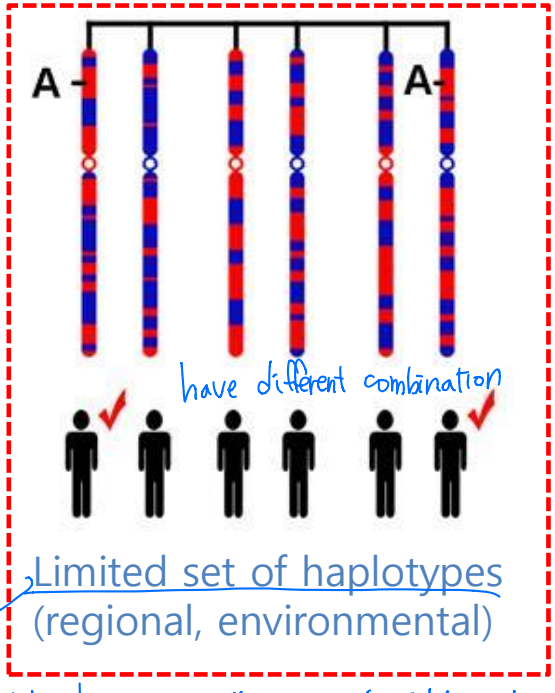


haplotype

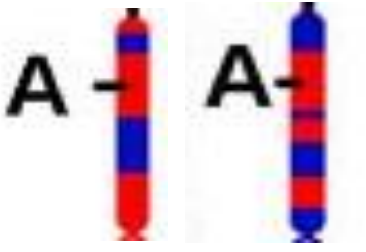
Haplotype: group of genes (DNA regions) in chromosome that are inherited (segregated) together from a single parent during recombination



Human population

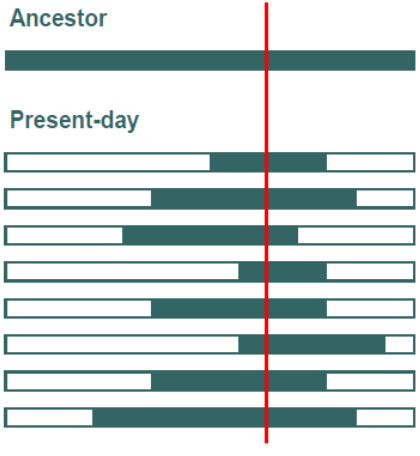


Limited set of haplotypes
(regional, environmental)



Only two haplotypes

> Easy to analyze

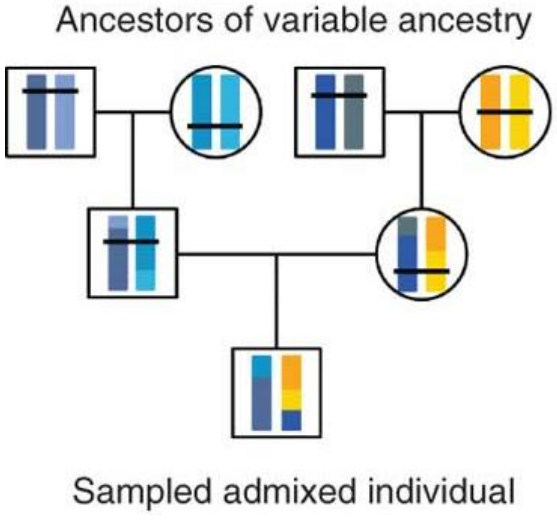


random 하게 생성되지 다양한 조합의 chromosome이 발견될것

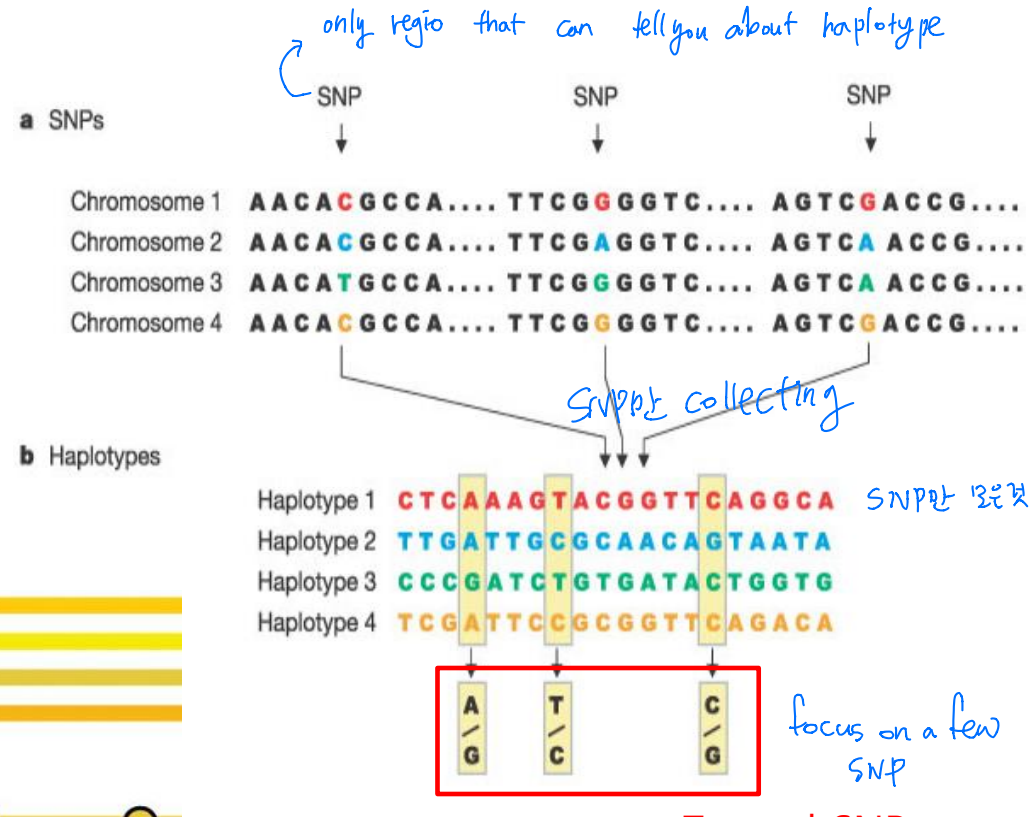
How to determine the same haplotype regions?

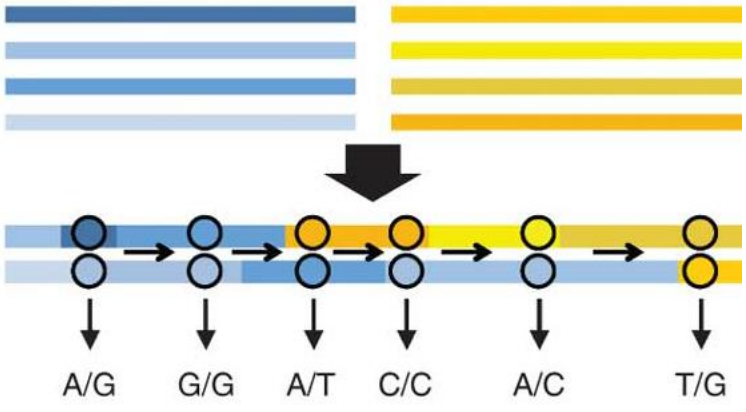
⇒ Variation (SNP)

Tagged SNP



0.1%의 SNP의 밀도 -> 약 1kb 당 1개의 SNP
특정 지역은 100kb당 5개 이내의 SNP -> recombination이 거의 일어나지 않는 지역 => haplotype

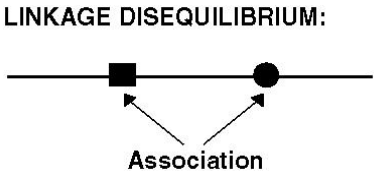




Tagged SNP
(적은 SNP 로써 haplotype을 figure out 하게 해주는 애들)

Only 4 haplotypes
: discriminate by 3 tagged SNPs

Linkage disequilibrium (LD)



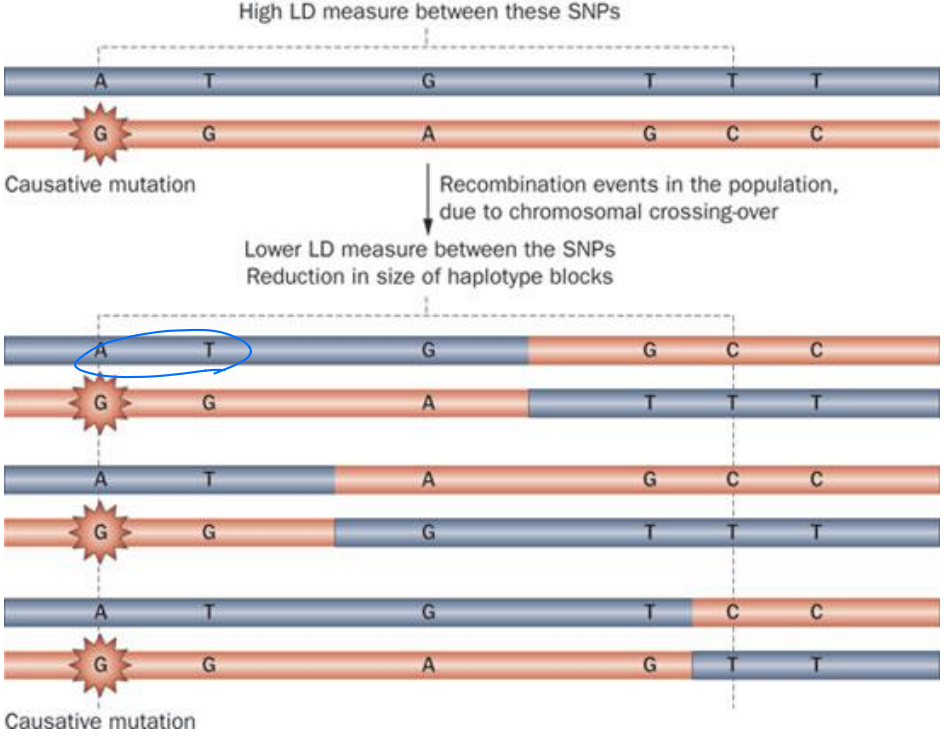
Sequencing results (population)

SNP끼리 haplotype을 어떻게 측정
⇒ SNP사이에 얼마나 recombination이 일어나는지

Observed frequency of co-occurrence ≠ Expected frequency of co-occurrence

Two SNPs

link가 있으면 같이 변함



How much it becomes disequilibrium ?

Completely linked

Linked

A T

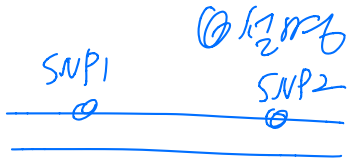
Completely independent

observed = expected

Calculation of linkage disequilibrium

To understand the calculation of linkage disequilibrium consider following example

Suppose there are two genes on Chromosome 5 of apple, each with two alleles



ACTGGTAT.....GATCAACCAG
ACTCGTAT.....GATCAACCAG
ACTCGTAT.....GATCATCCAG

SNP1 **SNP2**

Showing only alleles for both SNPs

Alleles	SNP1	SNP2
Allele 1	G	A
Allele 2	C	T

1. Observed haplotype frequencies

Haplotype	Frequency
A_1B_1	x_{11}
A_1B_2	x_{12}
A_2B_1	x_{21}
A_2B_2	x_{22}

2. Calculate allele frequencies (expected)

If p1 and p2 =frequency of the alleles at SNP1 and q1 and q2 =frequency of the alleles at SNP2 then in tabular form it could be written as follows

SNP1		SNP2	
Allele	**Frequency**	**Allele**	**Frequency**
G	p1	A	q1
C	p2	T	q2

Allele	Frequency
A_1	$p_1 = x_{11} + x_{12}$
A_2	$p_2 = x_{21} + x_{22}$
B_1	$q_1 = x_{11} + x_{21}$
B_2	$q_2 = x_{12} + x_{22}$

Calculation of linkage disequilibrium

SNP1		SNP2	
Allele	Frequency	Allele	Frequency
G	p1	A	q1
C	p2	T	q2

Haplotype	Frequency
A_1B_1	x_{11}
A_1B_2	x_{12}
A_2B_1	x_{21}
A_2B_2	x_{22}

When haplotype frequencies are equal to the product of their corresponding allele frequencies, it means the loci are in linkage equilibrium

observed — expected

Haplotype frequency		Product of allelic frequency
p11 GA 1/3	=	p1q1 2/9
p12 GT 0	=	p1q2 1/9
p21 CA 1/3	=	p2q1 4/9
p22 CT 1/3	=	p2q2 2/9

D = 1/9
D = -1/9
D = -1/9
D = 1/9

Allele	Frequency
A_1	$p_1 = x_{11} + x_{12}$
A_2	$p_2 = x_{21} + x_{22}$
B_1	$q_1 = x_{11} + x_{21}$
B_2	$q_2 = x_{12} + x_{22}$

	A_1	A_2	Total
B_1	$x_{11} = p_1q_1 + D$	$x_{21} = p_2q_1 - D$	q_1
B_2	$x_{12} = p_1q_2 - D$	$x_{22} = p_2q_2 + D$	q_2
Total	p_1	p_2	1

observed frequency가 0 이될수 X
minimum

observed — expected

0.2+D — 0.3-D
0.2-D — 0.3+D

⇒ Dmax는 0.2

SNP1: G G C C
SNP2: A A A T

GA = 2/4 ⇒ 0.5
G, A → 2/4 × 3/4 = 0.375
차이가 0.125
이게 어디쯤이냐~
0.2랑 0.125랑
100
한 어쯤~
어떻게
0 0~1값으로

D : linkage disequilibrium coefficient

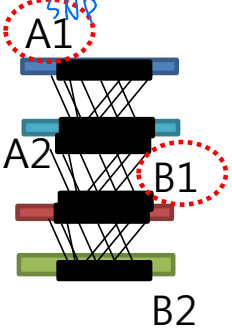
Linkage Disequilibrium(LD)

D_{max}: maximum possible value of D

$$D' = \frac{D}{D_{max}}$$

$$D_{max} = \begin{cases} \min(p_1q_1,\ p_2q_2) & \text{when } D < 0 \\ \min(p_1q_2,\ p_2q_1) & \text{when } D > 0 \end{cases}$$

D_{max}: maximum possible value of D given the allele frequency



Equilibrium
D' = 0.17

A1B1	A1B2
A1B1	A1B2
A1B1	A1B1
A2B1	A2B1
A2B2	A2B2

$p_{11}=p_1q_1+D$

$D=p_{11}-p_1q_1$

observed =4/10 – 6/10X6/10 expected

=0.4- 0.36 = 0.04

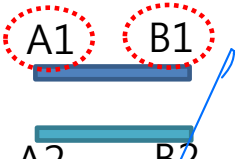
Dmax=min (p1q2, p2q1)

= min (6/10X4/10, 4/10x6/10) = 0.24

D' = 0.04 / 0.24 =~0.17

Coupling
D' = 1

A1B1	A1B1
A1B1	A1B1
A1B1	A1B1
A2B2	A2B2
A2B2	A2B2

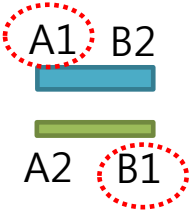


D=0.24

D=6/10 – 6/10 X6/10

Repulsion
D' = -1

A1B2	A1B2
A1B2	A1B2
A1B1	A1B1
A2B1	A2B1
A2B1	A2B1



D=-0.16

D= 2/10 – 6/10 x 6/10

	A_1	A_2	Total
B_1	$x_{11} = p_1q_1 + D$	$x_{21} = p_2q_1 - D$	q_1
B_2	$x_{12} = p_1q_2 - D$	$x_{22} = p_2q_2 + D$	q_2
Total	p_1	p_2	1

Exercise

EXERCISE 3.1 *Quantifying heterozygosity and LD*

Using the following ten sequences:

```
 1 gctgcatcag aagaggccat caagcggatc actgtacttc tgccatggcc
 2 gctgtatcag aacaggccat caagcggatc actgtacttc tgccatggcc
 3 gctgtatcag aacaggccat caagcacctc actgtacttc tgccatggac
 4 gctgcatcag aagaggccat caagcacatc actgtccttc tgccatggcc
 5 gctgcatcag aagaggccat caagcacatc actgtccttc tgccatggcc
 6 gctgtatcag aacaggccat caagcggatc actgtccttc tgccatggcc
 7 gcggcatcag aagaggcgat caagcacatc actgtccttc tgccatggac
 8 gctgcatcag aagaggccat caagcacatc actctacttc tgccatggcc
 9 gctgtatcag aacaggccat caagcggatc actgtccttc tgccatggcc
10 gctgcatcag aagaggccat caagcggatc actctccttc tgccatggcc
```

(a) Count the number of segregating sites. SNPs

(b) Calculate the expected average heterozygosity per nucleotide.

(c) Determine the level of linkage disequilibrium between the common polymorphisms.

SNPs

9 SNPs

(n=50)

```
 1  gctgcatcag aagaggccat caagcggatc actgtacttc tgccatggcc
 2  gctgtatcag aacaggccat caagcggatc actgtacttc tgccatggcc
 3  gctgtatcag aacaggccat caagcagctc actgtacttc tgccatggac
 4  gctgcatcag aagaggccat caagcagatc actgtccttc tgccatggcc
 5  gctgcatcag aagaggccat caagcagatc actgtccttc tgccatggcc
 6  gctgtatcag aacaggccat caagcggatc actgtccttc tgccatggcc
 7  gcggcatcag aagaggcgat caagcagatc actgtccttc tgccatggac
 8  gctgcatcag aagaggccat caagcagatc actctacttc tgccatggcc
 9  gctgtatcag aacaggccat caagcggatc actgtccttc tgccatggcc
10  gctgcatcag aagaggccat caagcggatc actctccttc tgccatggcc
```

0.1 # 3
0.4 # 3
0.5 # 1
0.2 # 2

ANSWER: *(a) There are 9 segregating sites, at positions 3, 5, 13, 18, 26, 28, 34, 36, and 49.*

(b) The minor allele frequencies at these 9 sites are, respectively, 0.1, 0.4, 0.4, 0.1, 0.5, 0.1, 0.2, 0.4, and 0.2. Since the expected heterozygosity of each nucleotide is given by 2pq where p is one allele frequency and q (= 1 – p) is the other allele frequency, the expected average heterozgosity H (including the 41 nonpolymorphic sites) is

0.1 x 0.9 x 2 =0.18

$$H = [(41 \times 0) + (3 \times 0.18) + (3 \times 0.48) + (2 \times 0.32) + 0.50]/50 = 0.0624$$

0.48
0.32
0.5

This means that, on average, just over three site differences are expected between any pair of randomly chosen alleles.

0.0624 X 50 =~ 3.1

~> 이 원리까지 중요

(c) There are four common polymorphisms, at positions 5, 13, 26, and 36.

cgga
tcga
tcga
cgac
cgac
tcgc
cgaa
cgaa
tcgc
cggc

If we extract these sites, it is easier to see how they are related. Following the procedure in Box 3.1, draw a table of haplotype frequencies for each pairwise combination. For example, for sites 5 and 13:

Site 5

t ($p_1 = 0.4$)
c ($p_2 = 0.6$)

Site 13

c ($q_1 = 0.4$)	**g** ($q_2 = 0.6$)
$p_{11} = 0.4$	$p_{12} = 0.0$
$p_{21} = 0.0$	$p_{22} = 0.6$

observed

Since $D = p_{11} - p_1q_1$, *for this pair* $D_{5,13} = 0.4 - (0.4 \times 0.4) = 0.24$.

expected

not zero → likely to be linked

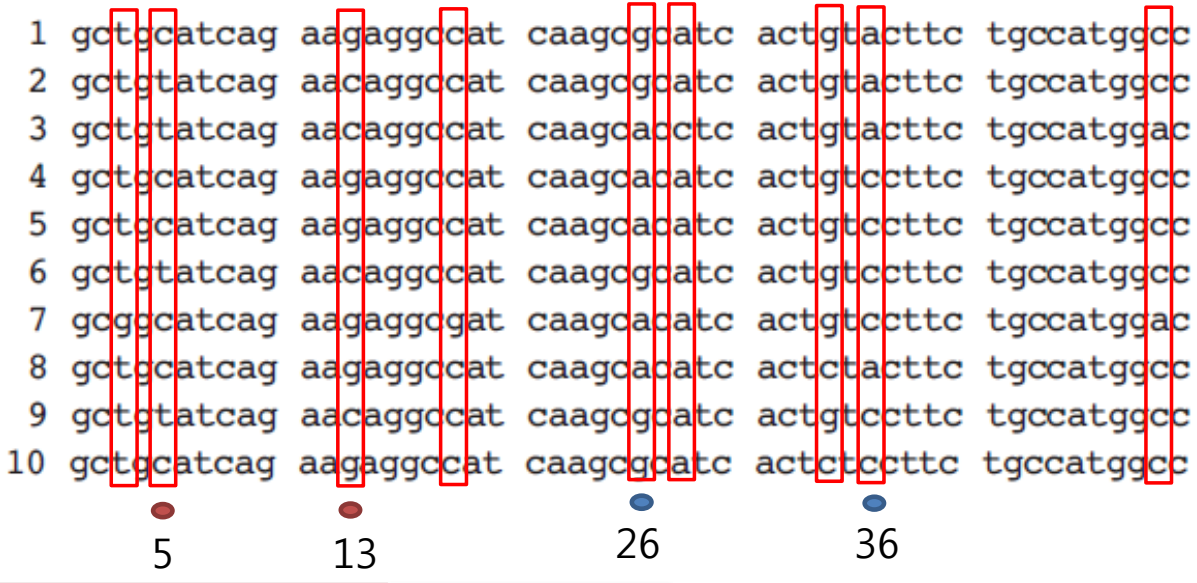
This is also the maximal value, D_{max}, *since all of the less common alleles at both sites always segregate together. Consequently, D′ is equal to 1. You should be able to calculate the following table for the other linkage disequilibrium estimates.* D_{max} *is just the maximum value* p_{11} *could take, minus* p_1q_1.

$$\min(p_1q_2,\ p_2q_1) \quad \text{when } D > 0$$

Allele pair	p_1	q_1	p_{11}	D	D_{max}	D′
5t, 13c	0.4	0.4	0.4	0.24	0.24	1.00
5t, 26g	0.4	0.5	0.3	0.10	0.20	0.50
5t, 36a	0.4	0.4	0.2	0.04	0.24	0.17
13c, 26g	0.4	0.5	0.3	0.10	0.20	0.50
13c, 36a	0.4	0.4	0.2	0.04	0.24	0.17
26c, 36a	0.5	0.4	0.2	0.00	0.20	0.00

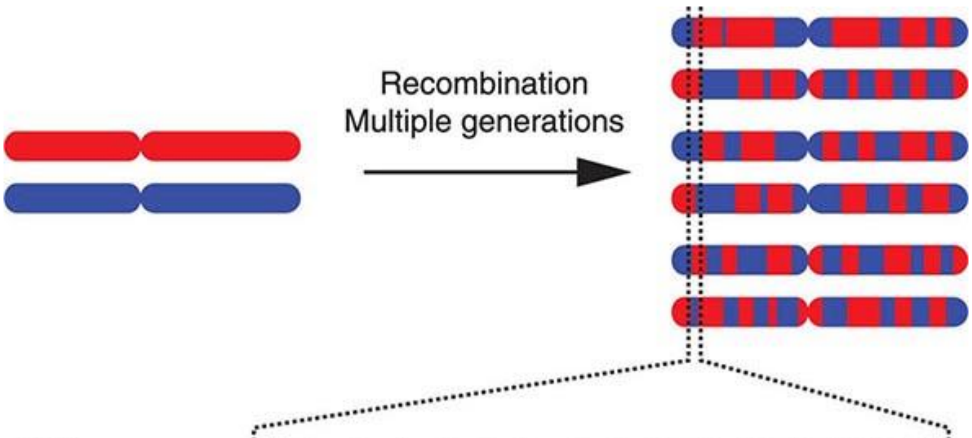
0.4 x 0.6 , 0.6 x 0.4

This means that there is complete LD between the first two sites, but linkage equilibrium between the last two sites, with partial LD for all other comparisons.



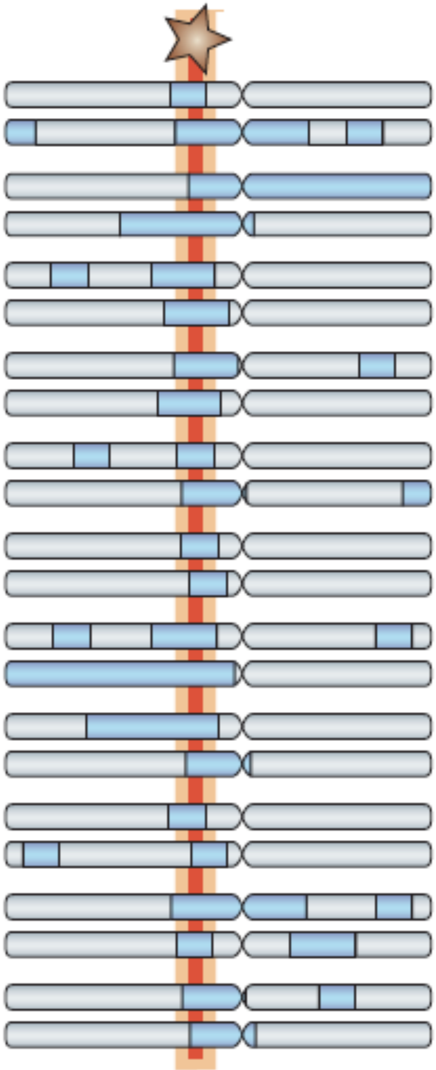
Allele pair	p_1	q_1	p_{11}	D	D_{max}	D′
5t, 13c	0.4	0.4	0.4	0.24	0.24	1.00
5t, 26g	0.4	0.5	0.3	0.10	0.20	0.50
5t, 36a	0.4	0.4	0.2	0.04	0.24	0.17
13c, 26g	0.4	0.5	0.3	0.10	0.20	0.50
13c, 36a	0.4	0.4	0.2	0.04	0.24	0.17
26c, 36a	0.5	0.4	0.2	0.00	0.20	0.00

Haplotype: identification of associated regions



SNP	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Haplotype 1	A	A	G	T	C	A	G	G	C	A	C	T	C	A	A	G
Haplotype 2	G	G	T	C	A	G	A	T	C	A	C	T	C	A	A	G
Haplotype 3	A	A	G	T	C	A	G	G	T	G	G	C	A	G	C	T
Haplotype 4	G	G	T	C	A	G	A	T	C	A	G	C	A	G	C	T
Haplotype 5	A	A	G	T	C	A	A	T	C	A	C	T	C	A	A	G
Haplotype 6	G	G	T	C	A	G	G	G	T	G	C	T	C	A	A	G
Tag SNP			G/T						C/T					A/G		

~> 여기위치로 쭉 잡



Haplotype Map (HapMap)

African

European Ancestry

Asian

질병원인 / 개별개체의 차이를 파악하는데 중요

(2008)

(2005)

Genome Variation Analysis

Introduction to genomic variation

Sung Wook Chi
Division of Life Sciences, Korea University

Genomic Variant Analysis

Human

DNA Sample

Animal/Plant

Collection, Analysis & Management of Genomic Information

Genomic Short Read Data

Genomic Sequence Alignment

Genomic Variant Analysis

Disease Association Analysis

Data Collection

Analysis

Mining

Personalized Medicare

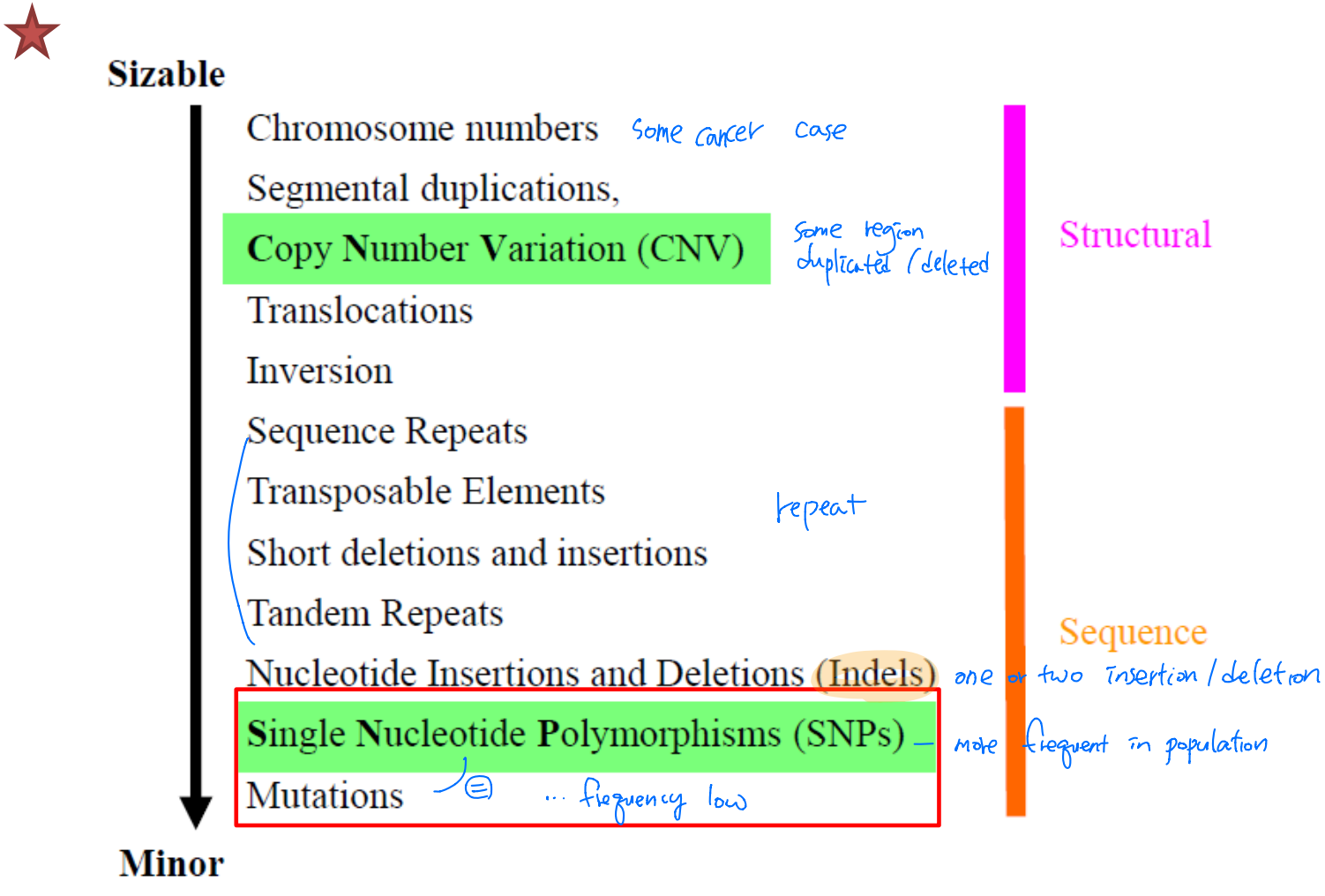
Biomarker/Drug Discovery

Molecular Breeding

SNP의 따라 특정 사람에 게는 drug 효능들이 달라진다로

Variation → **Genome** → Gene expression → **Phenotype** → **Change**

Genetic Variations



Sequence Variations (SNP, mutation)

POLYMORPHISM		MUTATION	
93%	C T A A G T A	99.7%	C T A A G T A
7%	C T A C G T A	0.3%	C T A C G T A

Widespread

Accepted mutation

Quantitative trait

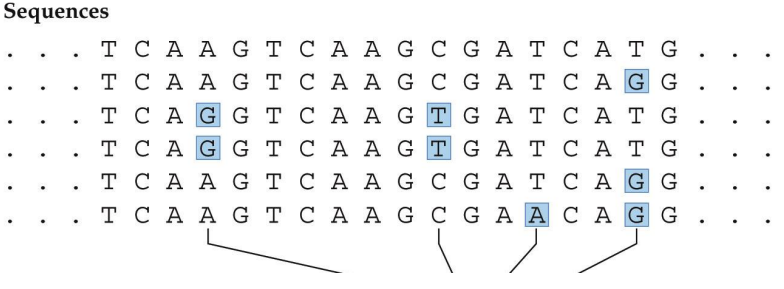
Rare

Diseases (more lethal)

Polymorphism: Single DNA base change found in >1% of population

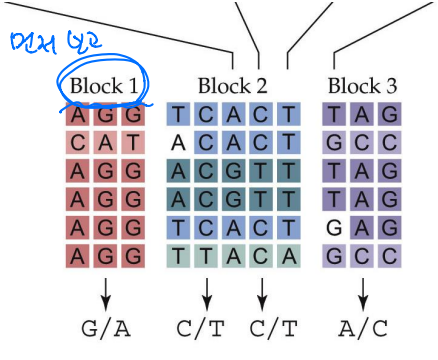
Mutation: Single DNA base change found in <1% of population

Genetic mutations are one type of genetic polymorphism



Haplotypes

Tagging SNPs



↳ large haplotype region 찾는 데 이용

Genetic Variations

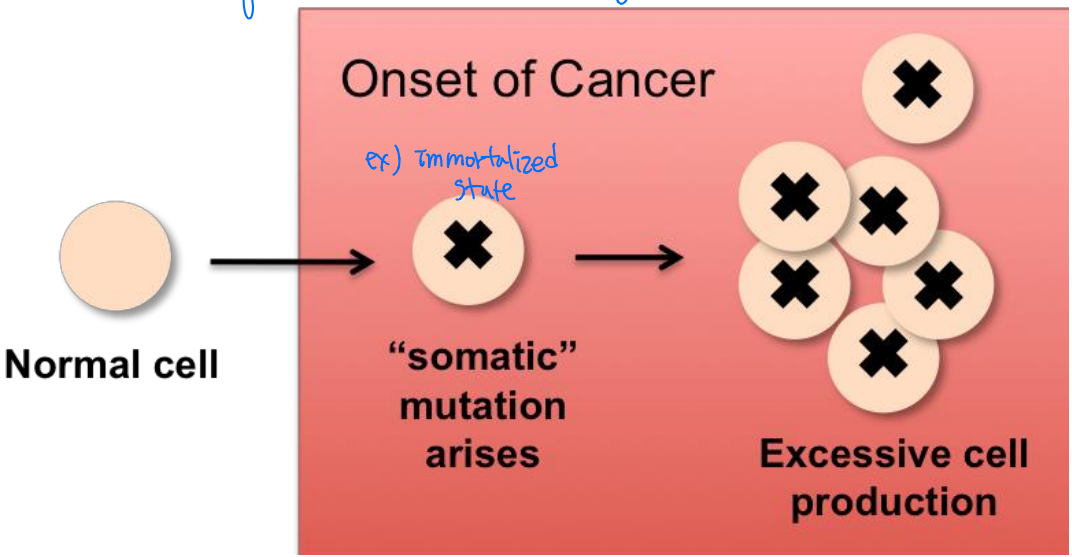
↳ all inherited & every tissue of my body have same variation

- Germline variation (Blood, normal tissue)
  : GWAS, Targeted genetic study

- Somatic variation (Tumor tissue)
  : Cancer genomics

→ variation이 tumor tissue에서만

↳ acquired mutation during cell division



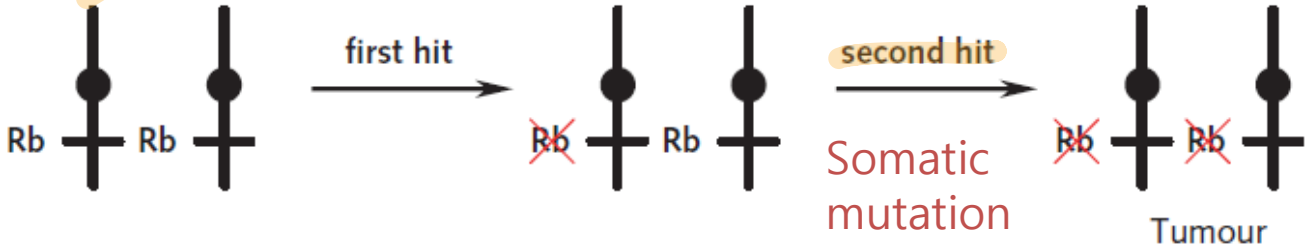
Mutation and Cancer

Somatic mutation

Two hits theory

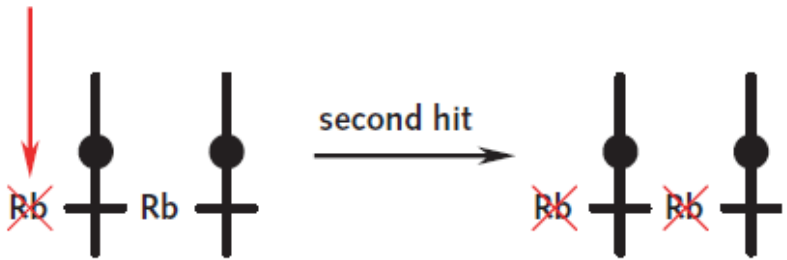
산발적인, 이래저 일어나는

Sporadic retinoblastoma: two hits required



Familial retinoblastoma: first hit inherited, one more required

Germ line mutation



already exist
⇒ reason why some people are more easy to get cancer

Mutation and Cancer

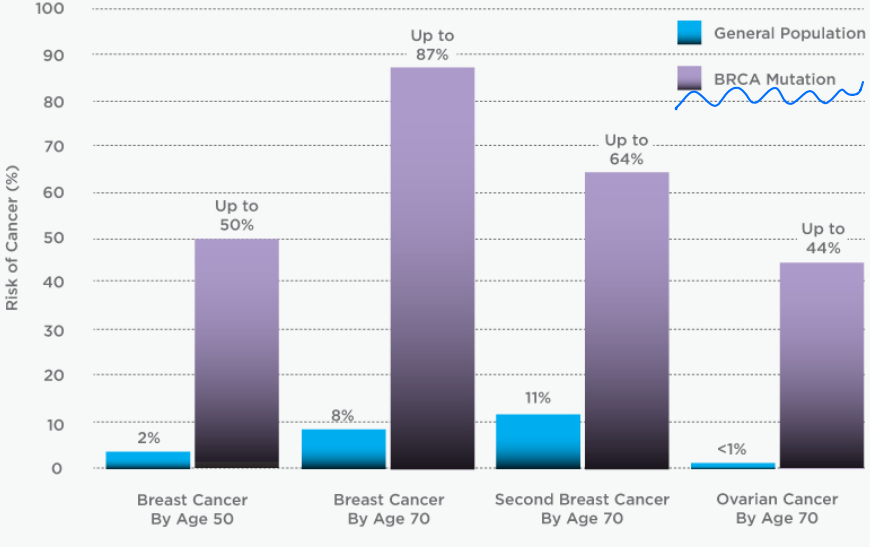
(1) Mutations detectable in the genome indicate propensity for development of cancers. Mutations in BRCA1 and BRCA2, as indicators for likelihood of breast and ovarian cancer development, are probably the best known.

(2) Sequence analysis can predict disease progression and outcome.

(3) Sequence analysis can help choose optimal treatment.

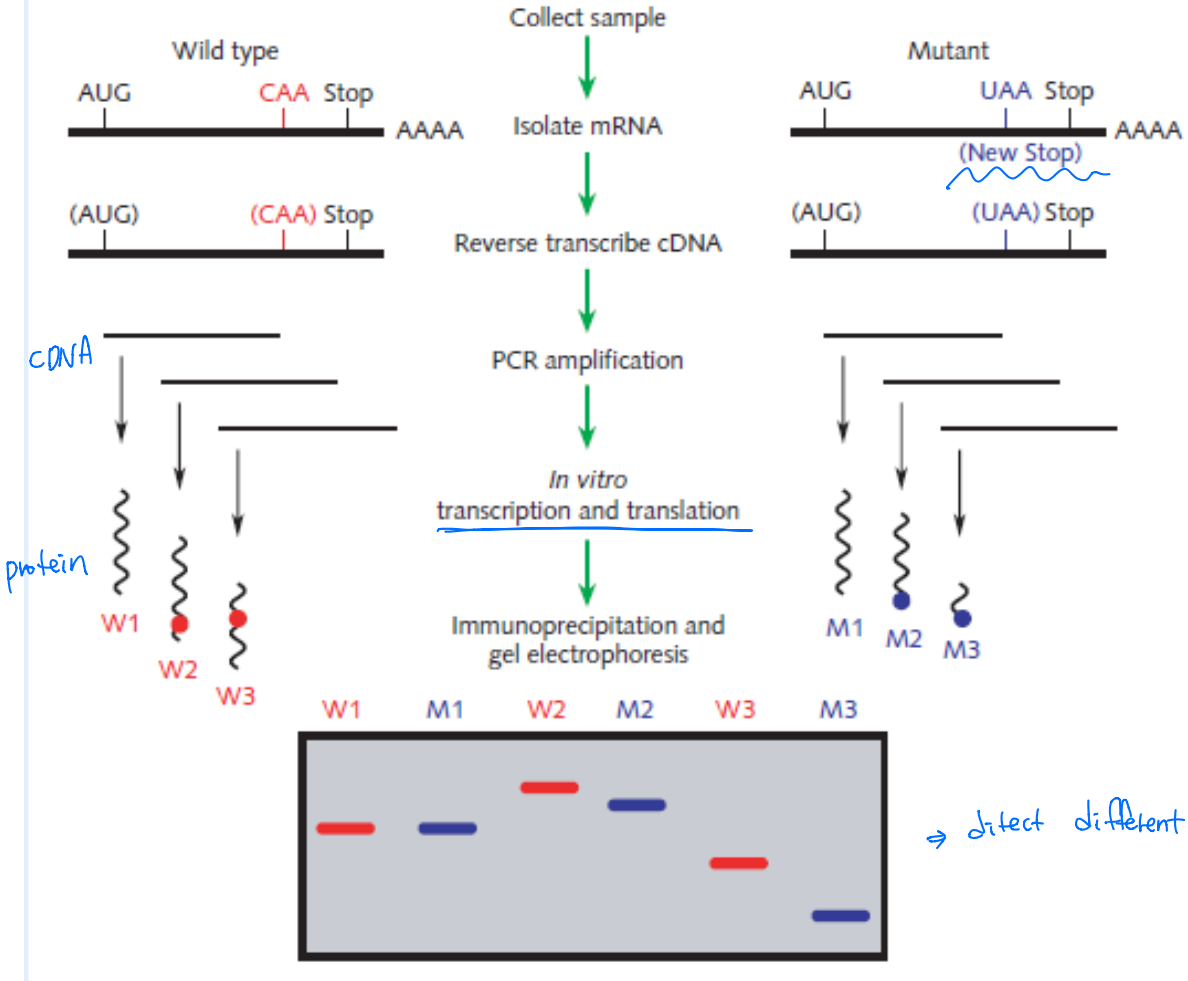
Table 2.1 Common *BRCA1* and *BRCA2* mutations

Population	Common *BRCA1* mutations	Common *BRCA2* mutations
Ashkenazi Jews	185delAG, 5382insC	6174delT
Iceland		999del5
Denmark	2594delC, 5208T→C	
Lithuania	4153delA, 5382insC, 61G→C	
China	589delCT, IVS7–27del10, 1081delG, 2371–2372delTG	3337C→T



Classical method detecting mutation-induced truncated protein

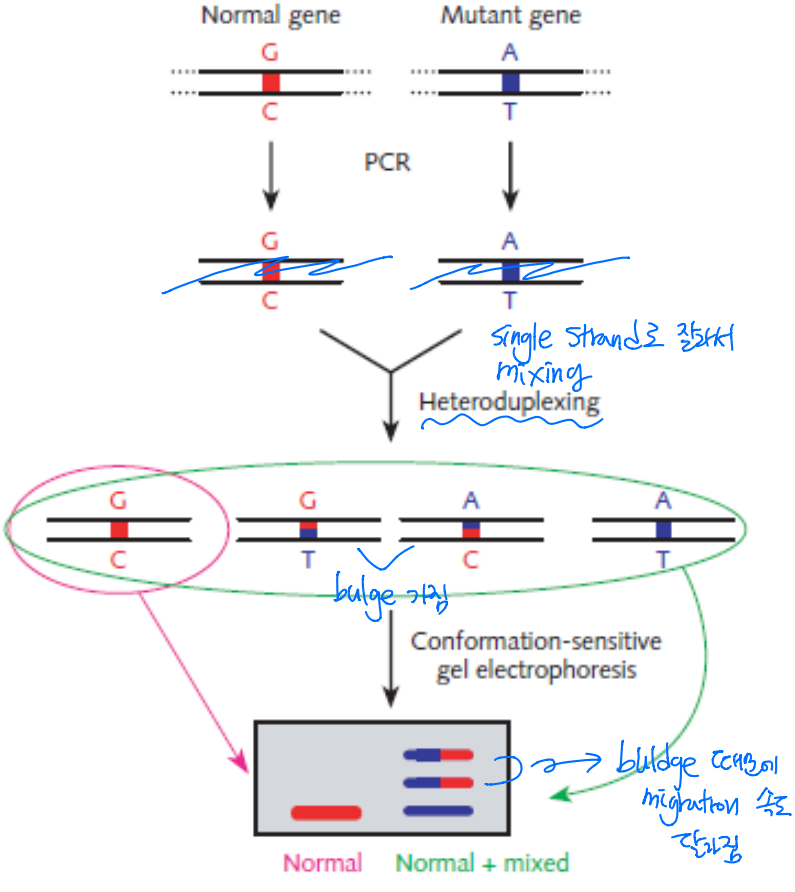
잘린 ptn을 만드는 mutation detecting



⇒ detect different protein size

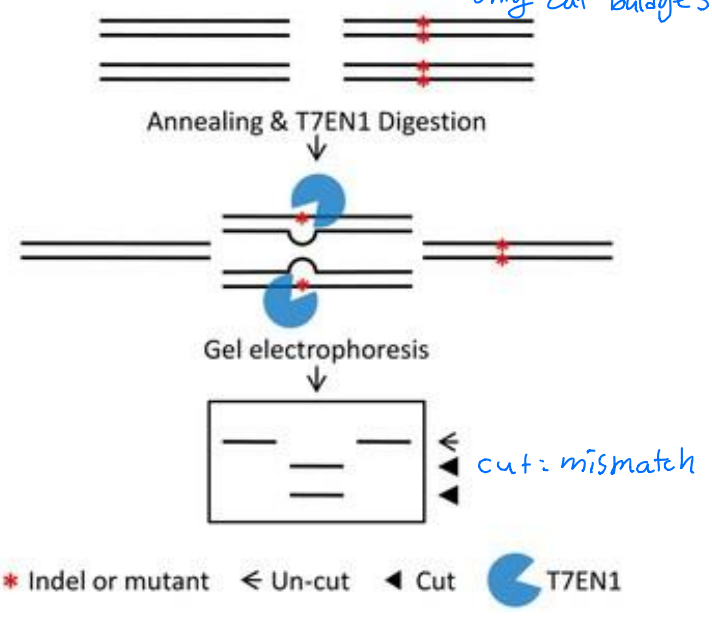
How to experimentally detect mutation?

Conformation-sensitive gel electrophoresis



T7E1 assay

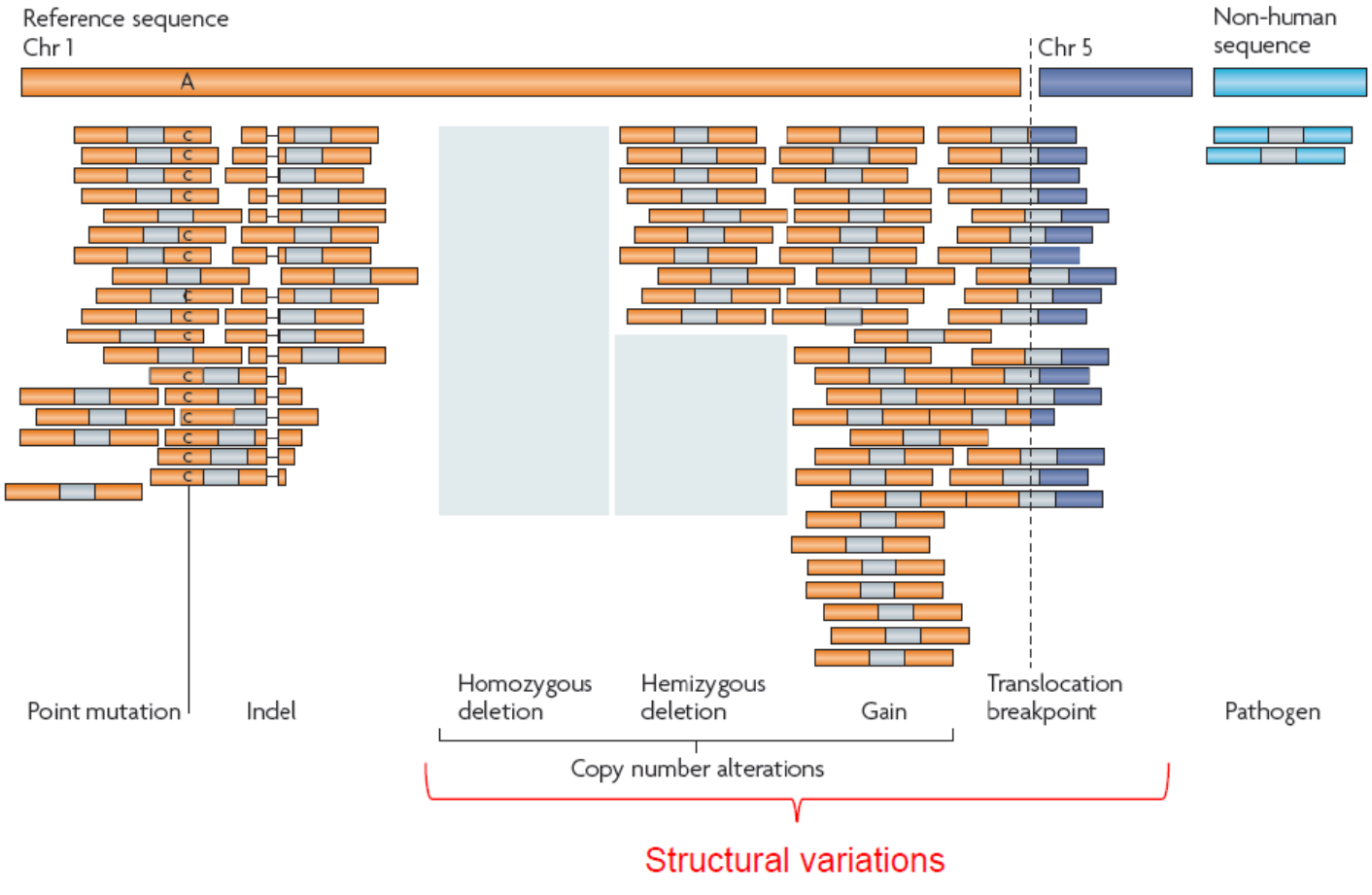
enzyme derived from T7 (endonuclease)
only cut bulge site



T7E1 enzyme:
cut mismatched DNA duplex site

Now days we are using sequencing (NGS)

Structural Variations (SVs)

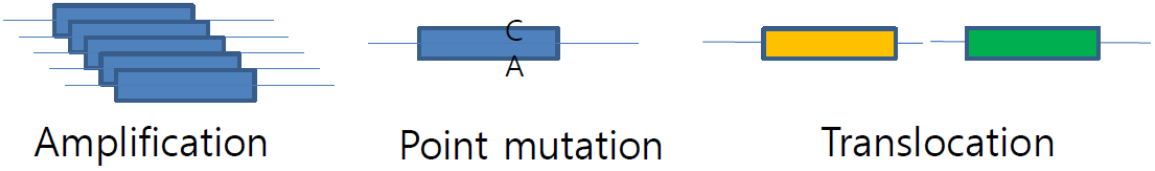


Oncogenes vs. Tumor Suppressor genes

- Oncogenes
  - Growth signals
  - Cell multiplication
  - Activated in cancer

- Tumor Suppressor genes
  - Growth suppressive signals
  - Cell stop dividing
  - Inactivated in in cancer

Corrupted Genomes in Cancer Cells



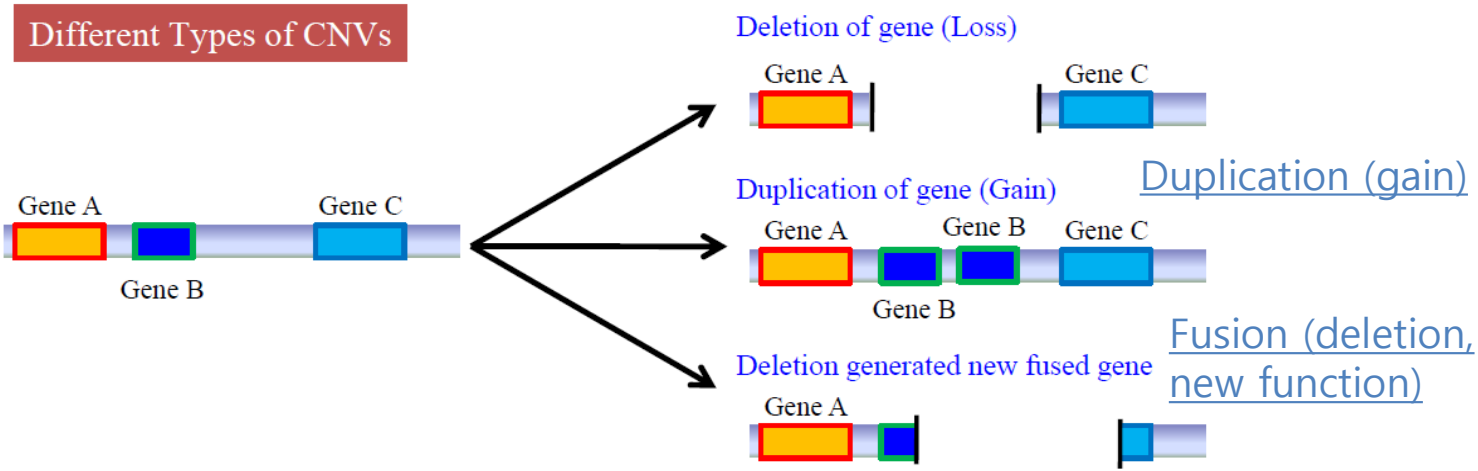
Amplification Point mutation Translocation

Increased signal Abnormal signal

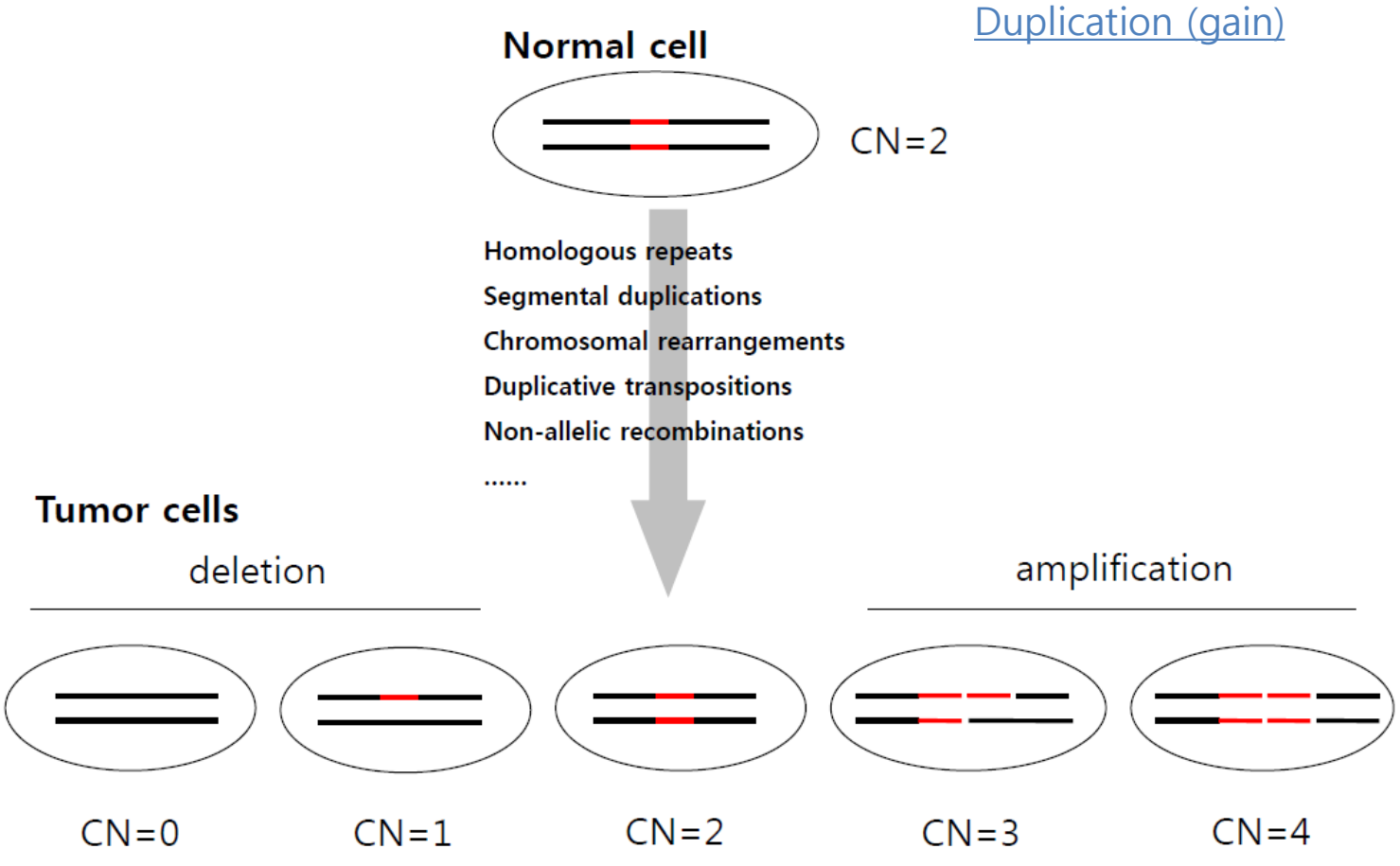
Deletion (loss)

Duplication (gain)

Fusion (deletion, new function)



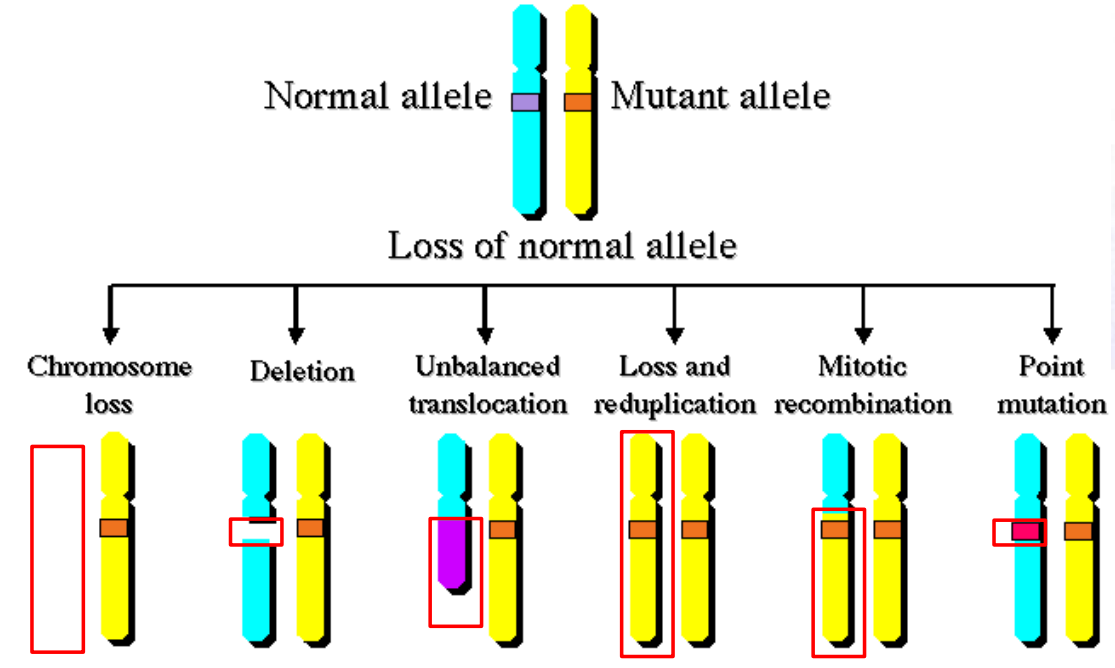
Genetic Alterations in Tumor (Copy number changes)

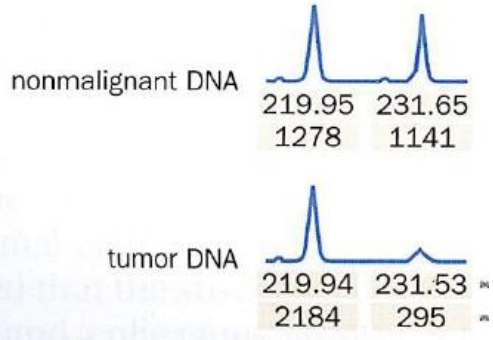


Loss of heterozygosity (LOH)

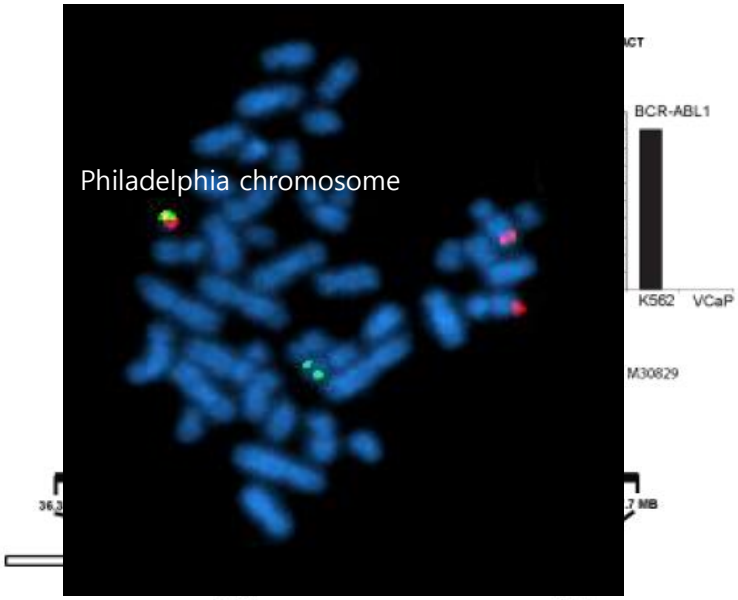
Loss of heterozygosity as a marker to locate tumor suppressor genes

- Somatic genetic changes in retinoblastoma caused loss of heterozygosity (LOH) at markers close to the RB1 locus
- By screening paired blood and tumor samples with markers spaced across the genome, we may discover the locations of tumor suppressor genes

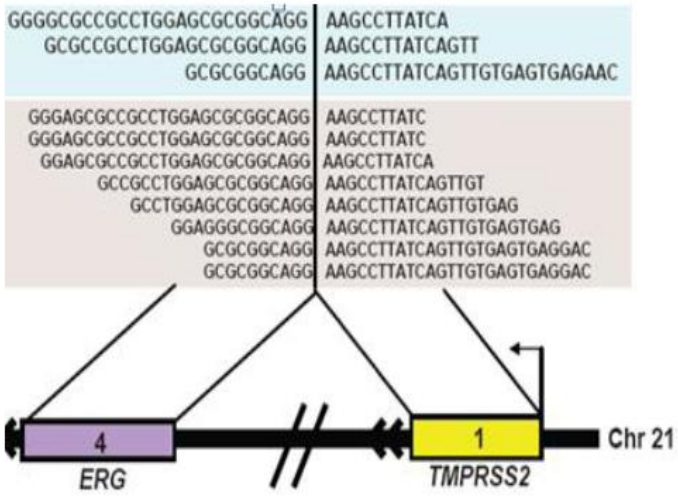




Fusion Gene in Cancer

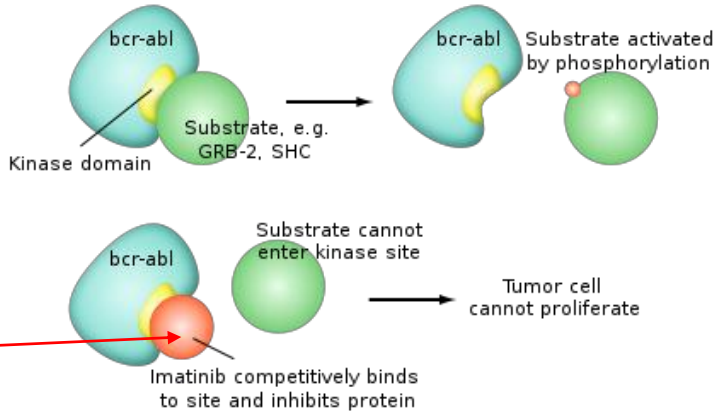


BCR-ABL1 caused by translocation
Chronic myeloid leukemia



TMPRSS2-ERG caused by deletion
Prostate cancer

Translocation



Genomics and Medicine : 100,000 Genome Project in England

Genomics england

Cancer and rare diseases

About Us | 100,000 Genomes Project | Taking Part | For Healthcare Professionals | Research | Industry Partnerships | News & Events

Genomics England is delivering the **100,000 Genomes Project.**

We are creating a new genomic medicine service with the NHS – to support **better diagnosis and better treatments** for patients. We are also enabling medical research.

More information about the 100,000 Genomes Project

Start in late 2012, aim to finish by 2017

https://www.youtube.com/watch?v=hxou7ayQSZQ