Functional Genomics

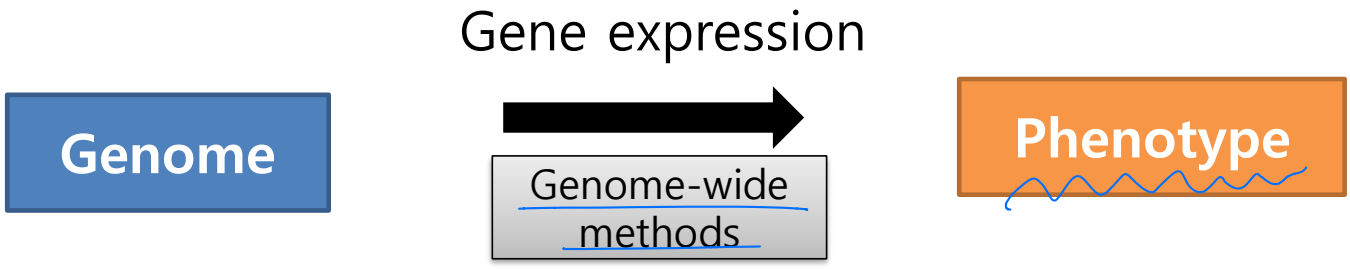
Review of Part I

Middle term exam : April 26th, Thursday, 5:00 pm

Sung Wook Chi

Division of Life Sciences, Korea University

Functional Genomics Class : part I

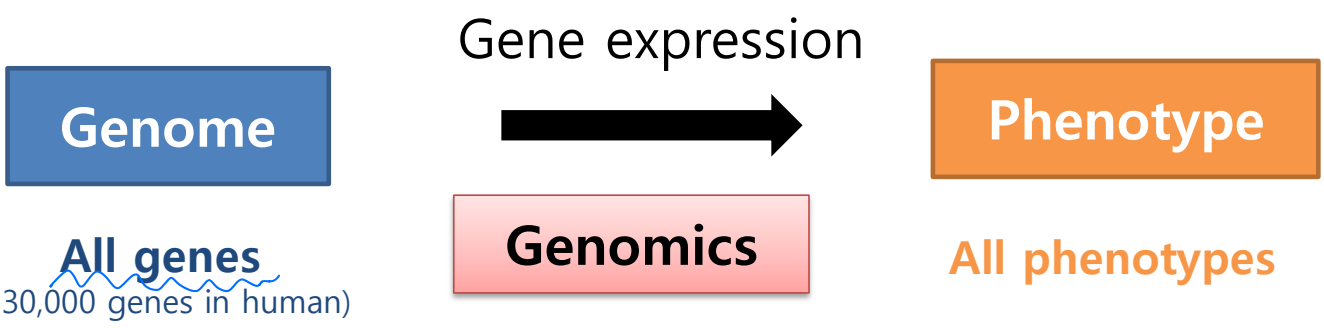


1. Human genome projects
2. Genome Sequencing
3. Sequence Alignment
4. Genomic variation

1. NGS
   - WGS
   - Exome-Seq
2. GWAS genomic variation & phenotype
3. Evolutional change (Phylogenetics)
4. Functional screening
   - Genetic interaction
   - RNAi screening
5. Regulation transcription
   - ChIP-Seq

1. Biological function
2. Diseases
3. Quantitative trait /population

1. Introduction to genomics/ functional genomics

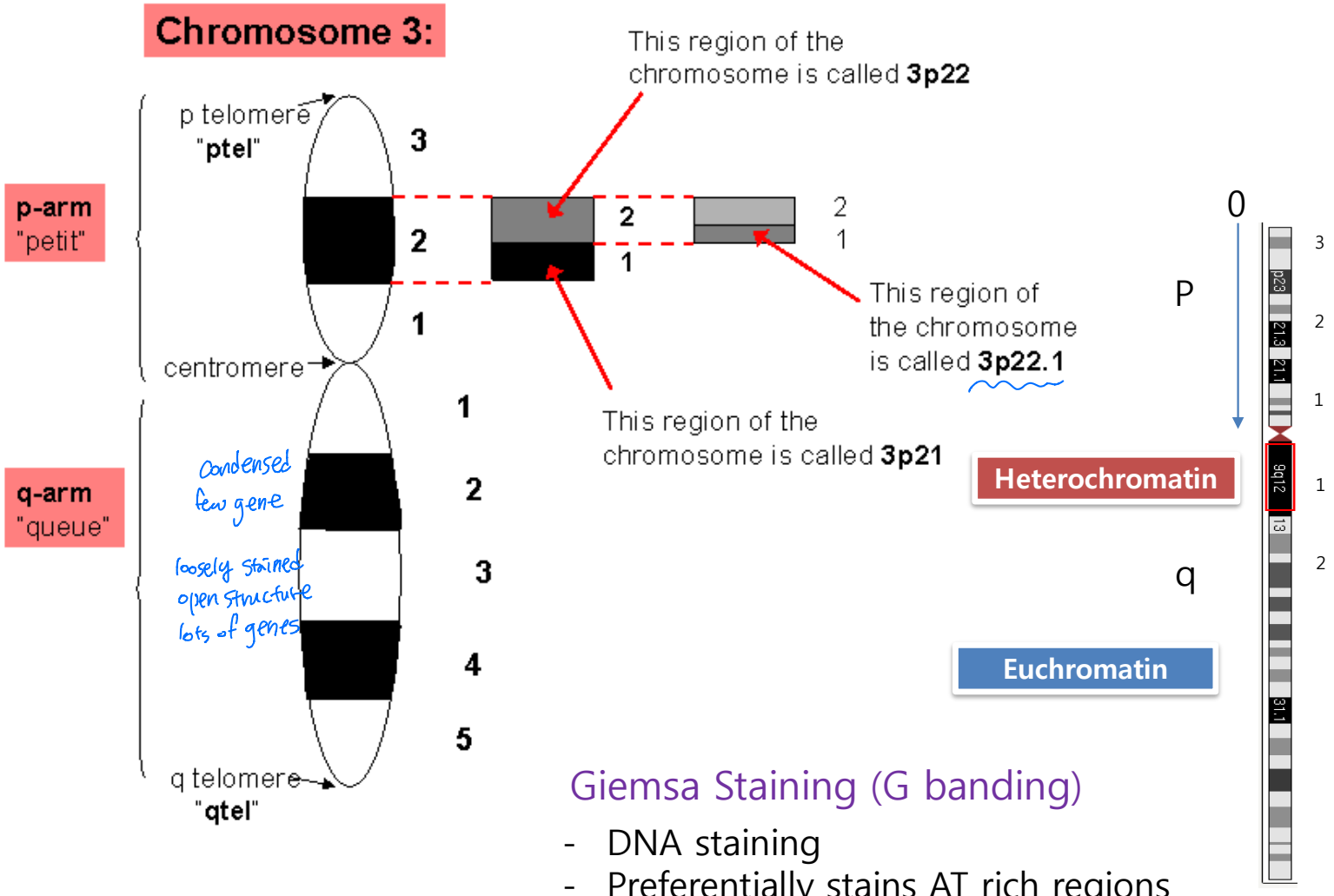


- **Holistic** (Global view of all genes) rather than focusing on one gene
- **High-throughput** (large amount of data in short time, Genome-wide methods)
  ↓ large data generated
- **Resource-generating** (database, analysis, data-driven research)
  Bioinformatics, Statistics, Interactions (network)

- Genome (Gene + chromosome) = DNA
  contain information

Phenotype = **genotype** + environment + life history + epigenetics
+ adding info.

Nomenclature of chromosome bands



Giemsa Staining (G banding)

- DNA staining
- Preferentially stains AT rich regions

Genome and gene structure

- DNA + Histone (2A, 2B, 3, 4) x 2 = Nucleosome

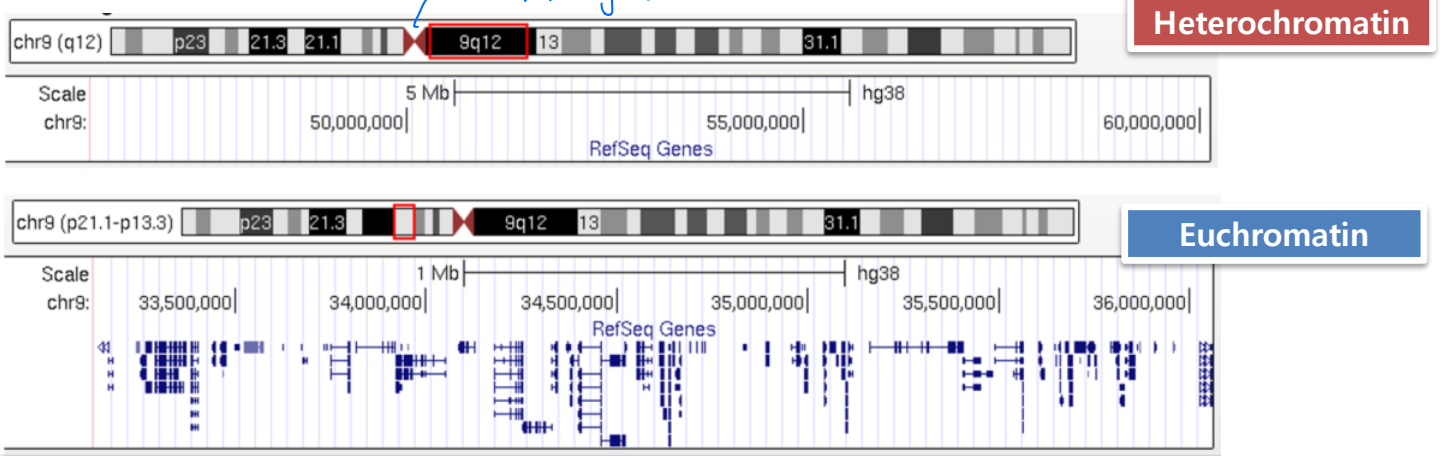
- Nucleosome > Chromatin > Chromosome

Chromatin (DNA + Structural information) Histone complex

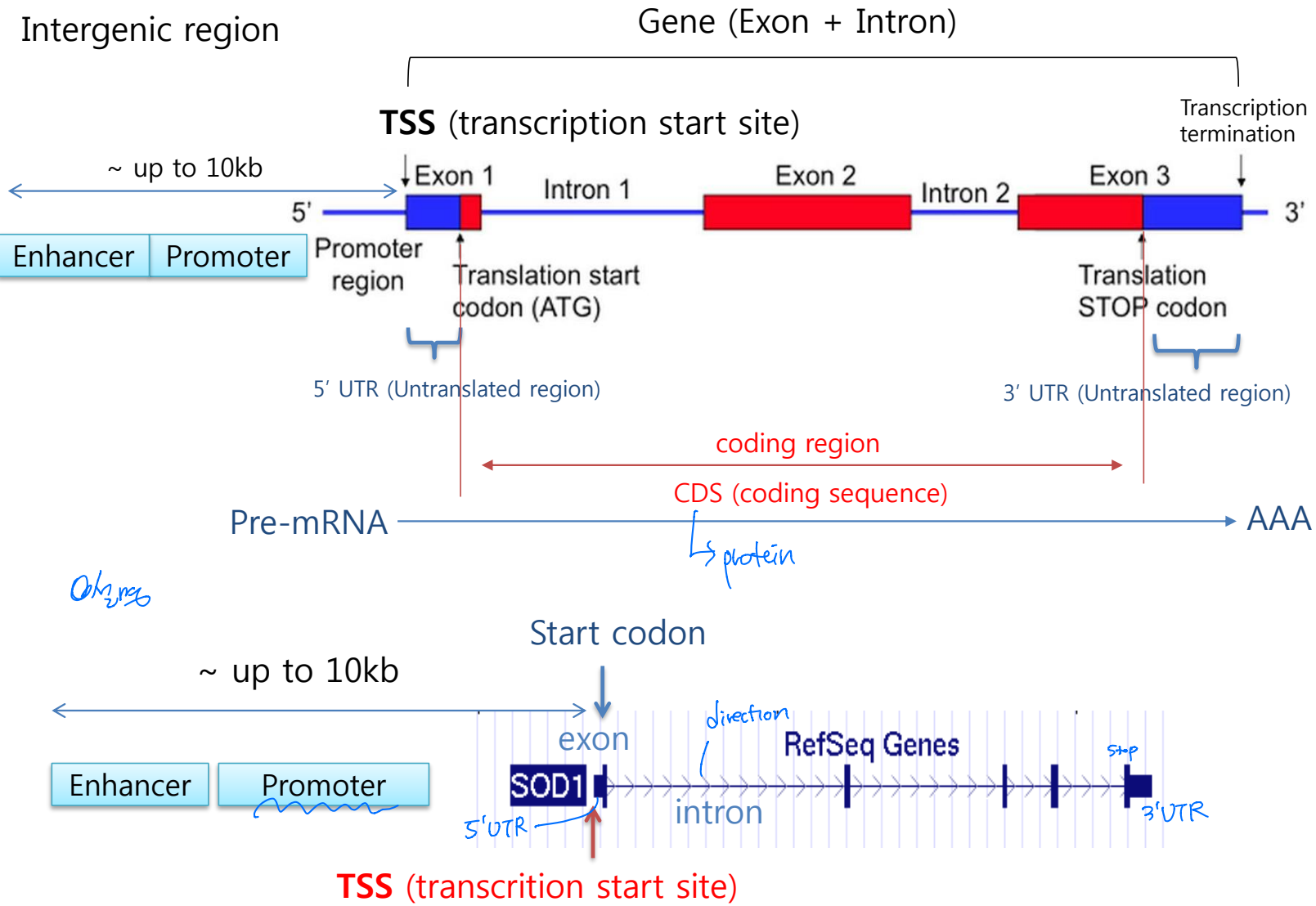
Human Genome = Gene (~5%) + intergenic regions (~95%)

(3 billion bp) ~ 22,000 coding genes (~1%)
+ non-coding genes (~4%)

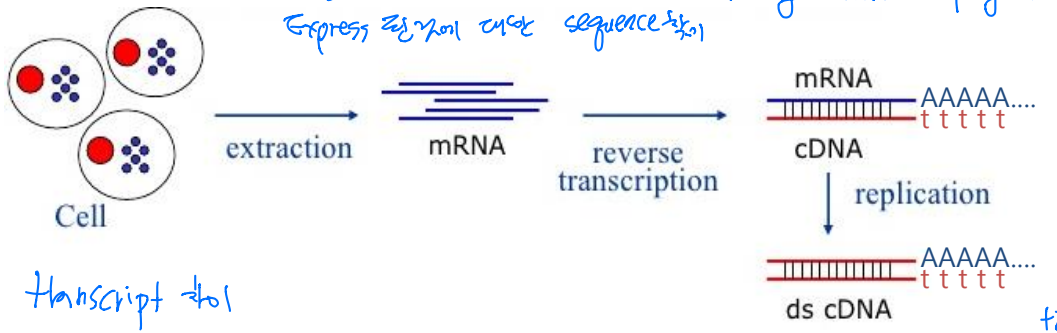
centromere, no genes

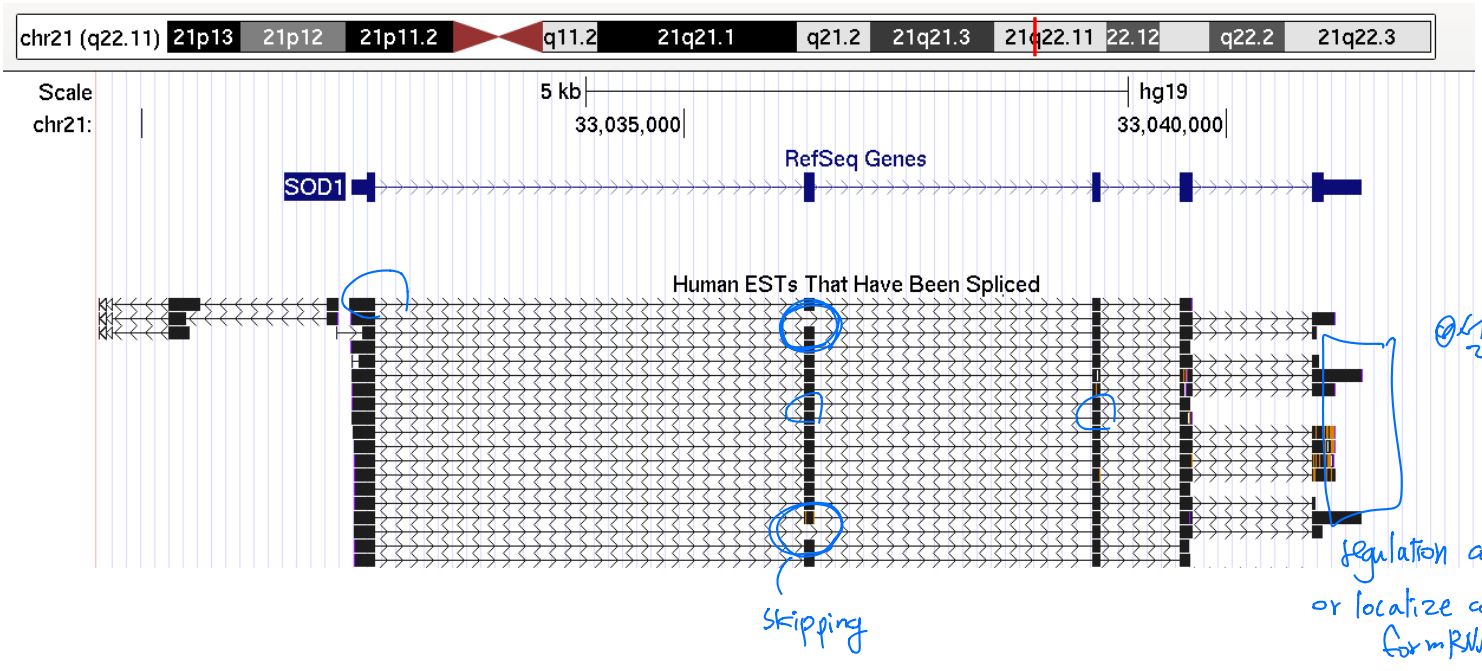


Gene structure

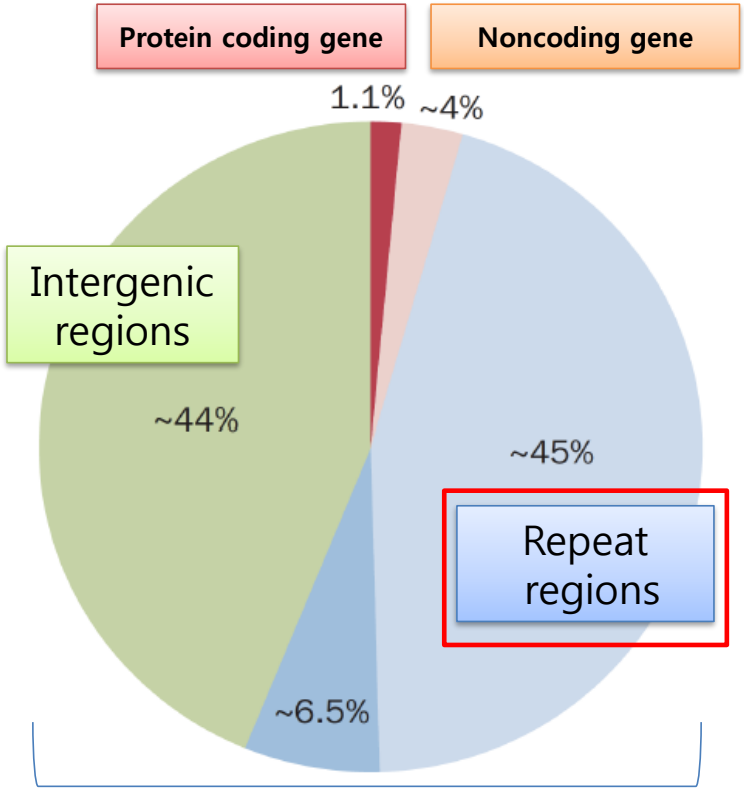


Expressed Sequence Tag (EST)





~50% of human genome is repeat sequences



~76% is transcribed at least once.

↳ ENCODE Project 결과

Table 1.2 Transposable elements in the human genome

Element	Estimated number	% of total genome
SINE + LINE	2.4×10^6	33.9
LTR	0.3×10^6	8.3
Transposons	0.3×10^6	2.8
Total	3.0×10^6	~45

Data from: Bannert, N. & Kurth, R. (2004). Retroelements and the human genome: New perspectives on an old relation. Proc. Natl. Acad. Sci. USA **101**, 14572–14579.

~40% of repeat regions is "**retrotransposon**" ← Retro Virus

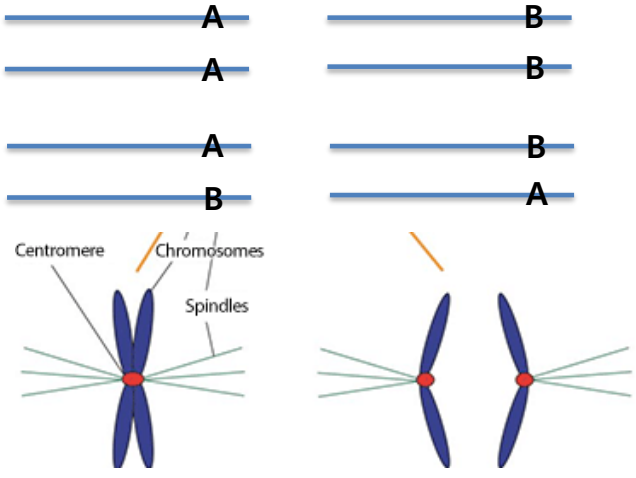
Majority of retrotransposon is LINE or SINE

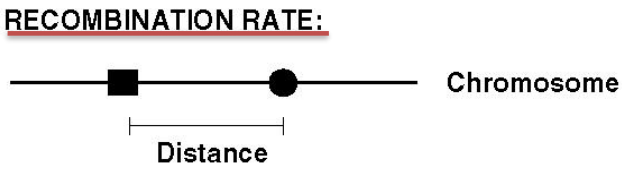
-> Cause genomic variation (disease)

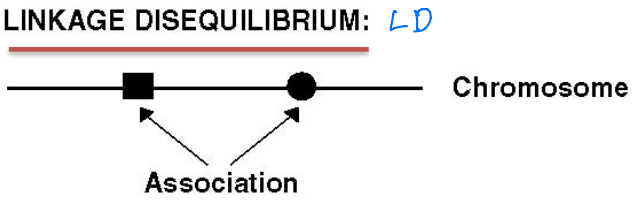
2. Mapping genome: terminology from genetics

Homozygote	An individual that has two identical alleles at some locus.
Heterozygote	An individual that has two different alleles at some locus.
Segregation	The separation of corresponding alleles during the reproductive process.

Linkage	Absence or reduction of independent assortment of parental genes, which are usually transmitted together because they lie on the same chromosome.





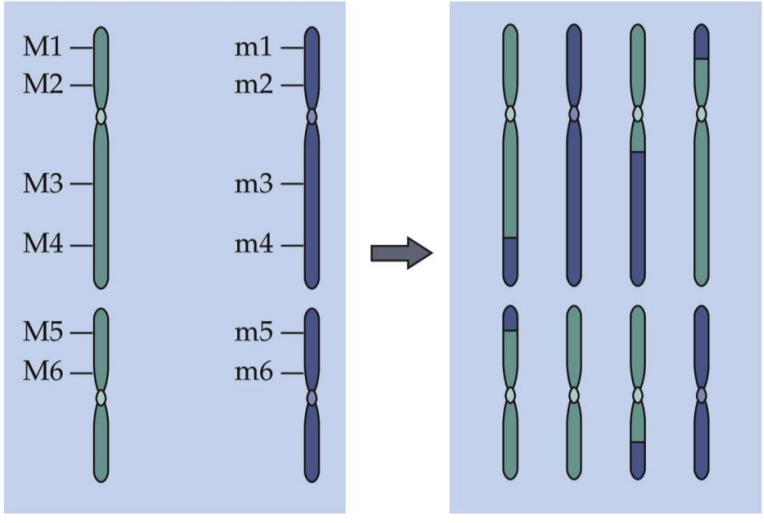


non-random association of alleles at different loci in a given population

Two markers are in **linkage disequilibrium if the** observed distribution of different combinations of alleles differs from that expected on the basis of independent hereditary transmission of the individual alleles.

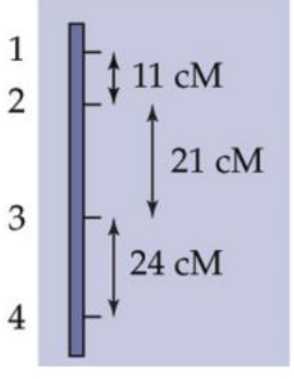
Genetic Maps

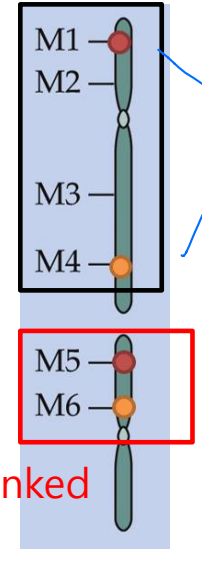
Genetic Maps
: relative order of genetics markers in linkage groups in which the distance between markers is expressed as units of recombination



Orange tree

		Normal			
		M1	M2	M3	M4
Thickskin	m1	–	.11	.31	.51
Reddish	m2		–	.22	.46
Sour	m3			–	.25
petite	m4				–





-centiMorgan (cM) : 1cM = recombination frequenct of 0.01

EXERCISE 1.1 *Constructing a genetic map*

Suppose that a breeder of orange trees begins to assemble a genetic map based on four recessive loci—thickskin, reddish, sour, and petite—named after the fruit phenotypes of homozygotes. After identifying two true-breeding trees that are either completely wild-type or are mutant for all four loci, the breeder crosses them and plants an orchard of the resultant F_2 trees. Based on the following frequencies of mutant classes, determine which loci are likely to be on the same chromosome and which are the most closely linked.

WT x Mut
ww mm

F2

ww wm wm mm

Total (n=968)

Normal	402	Thick	18
Petite	127	Red, sour, and petite	12
Sour	115	Red	11
Thick and red	108	Thick and petite	10
Thick, red, and petite	42	Thick, red, sour, and petite	8
Thick, red, and sour	41	Thick and sour	7
Sour and petite	38	Red and petite	3
Red and sour	24	Thick, sour, and petite	2

Thick
108+42+41+18+10+8+7 +2 = 236

Reddish = 249

Sour= 247

Petite = 242

Mendelian ratio

recessive (1/4) → 968 x 1/4 = 242

B/W two phenotypes (genes)

Unlinked
(Segregate independently)

→ 968 x 1/4 x 1/4 = ~60

abundantly observed = likely to be linked

far away or independent (다른 chromosome)

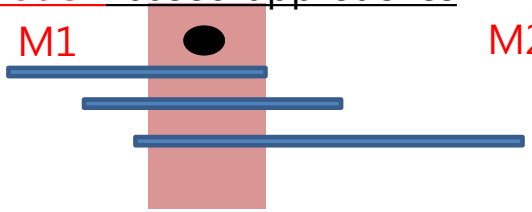
Thick and red	199	Thick and petite	62	Thick and sour	58
Red and petite	65	Red and sour	85	Sour and petite	60

약간 크

Mapping Genomes : Physical Maps

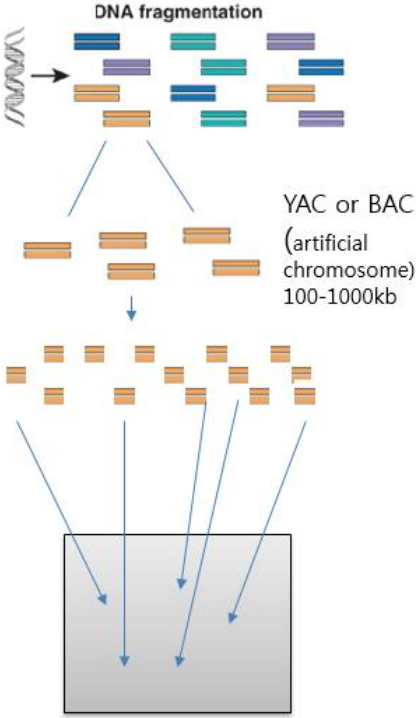
1. Assembling of contigs: Hybridization-based approaches

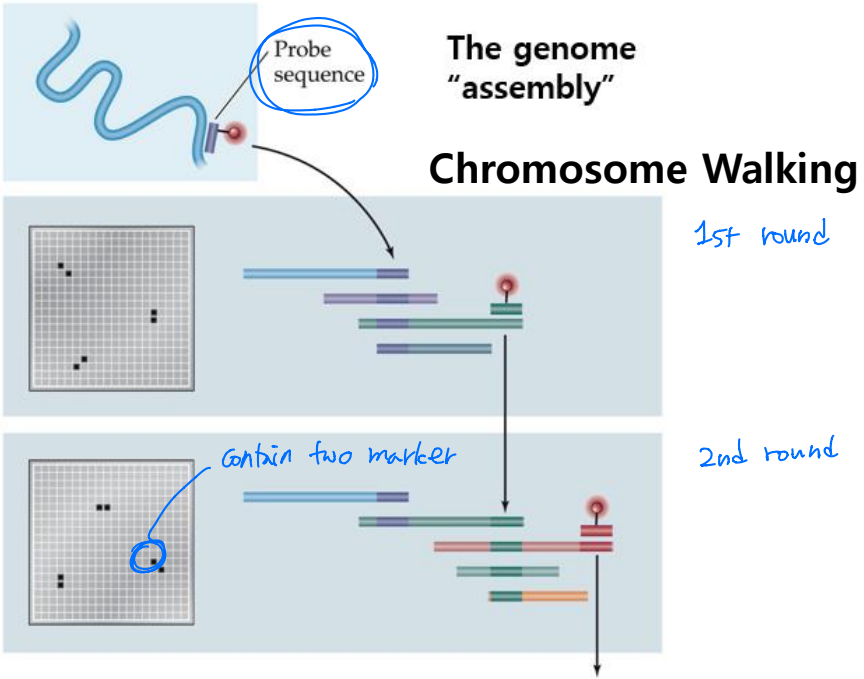
: chromosome walking



use large fragment & probe or STS

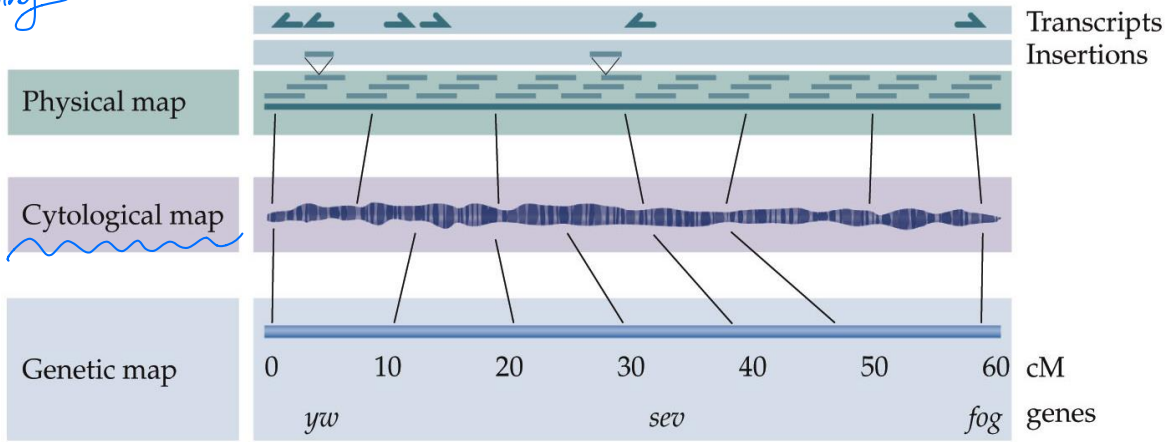
: sequence-tagged sites (STSs): single occurrence in the genome





Genomic maps

Cytosolic mapping



Synteny:
physical co-localization of genetic loci between chromosome segments of two of more organisms.

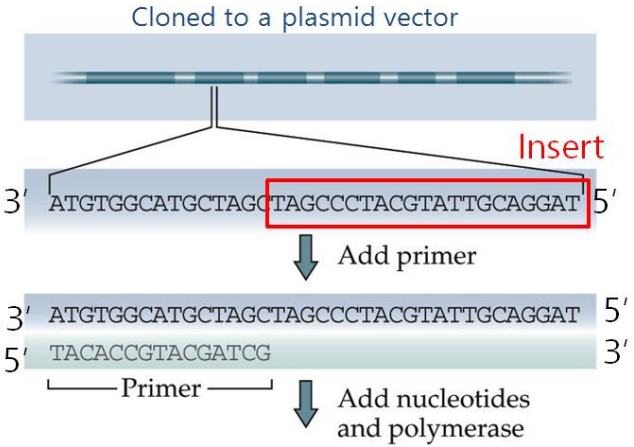
모양이 유지되는 block들
⇒ 같은 종이면

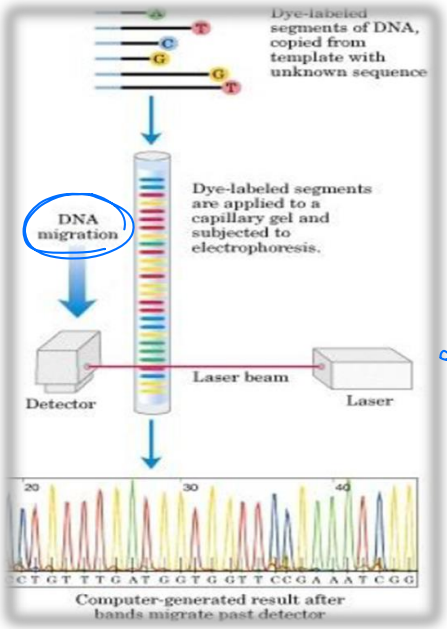
Homologs:
Genes with very similar sequences, evolved from a common ancestral DNA sequences

1) Orthologs: homologs in different species (evolution)
2) Paralogs: homologs within the same species (gene duplication)

3. Human Genome Project: Sanger Sequencing

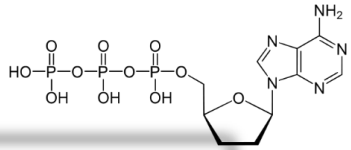
basic method





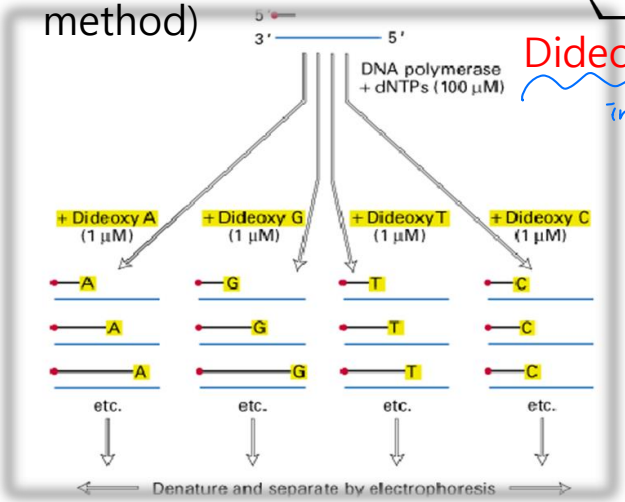
detect color

Sanger Sequencing (Chain termination method)



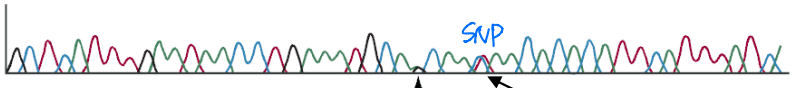
Dideoxy nucleotide

induce termination



EXERCISE 2.1 *Reading a sequence trace*

Read the following DNA sequence obtained by direct sequencing of a single individual organism, assuming that on the trace green is A, red is T, blue is C, and black is G. Remark on any ambiguities in the sequence.



Base calling

GCANCAA(C/T)AA

quality score 필요

not confident

↳ heterozygote 여러 allele 존재

Phred Scores: quality of base calling

Sanger Seq 에서 이용하기에도 NGS 도 이용

Quality score
(PHRED)
: base calling

$$q = -10 \log_{10}(p).$$

error ↓ q score ↑

⇒ based on shape of peak

Error rate
in base calling process
(estimated based on shape/height of peaks)

BOX 3.7 **Phred scores: a measure of quality of sequence determination**

The phred score of a sequence determination is a measure of sequence quality. It specifies the probability that the base reported is correct.

If p = the probability that a base is in error, then the corresponding phred score $q = -10 \log_{10}(p)$.

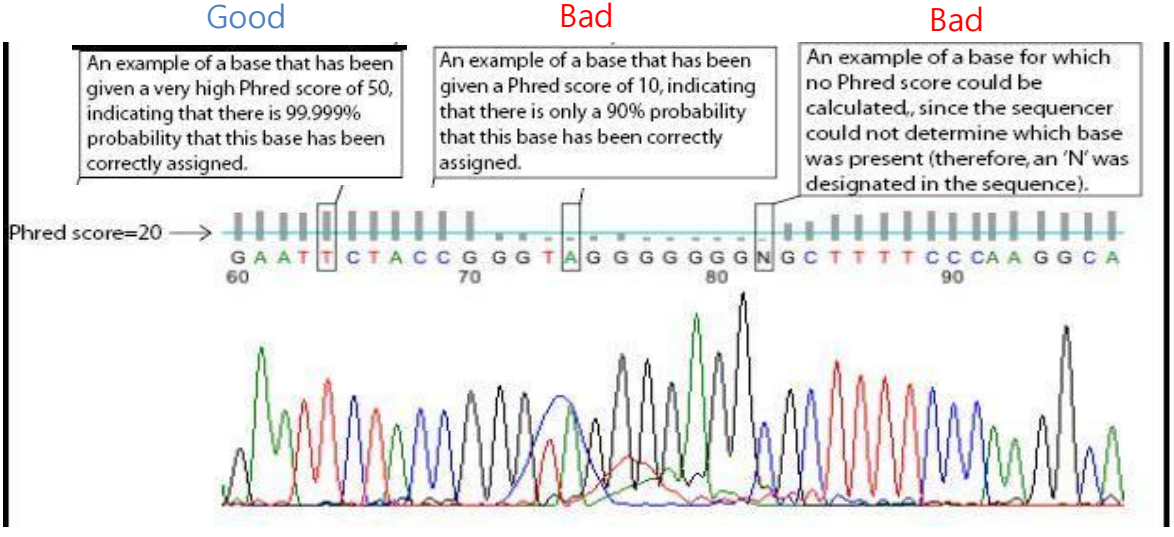
Here is a short table:

Quality score q	Probability of error	Error rate
10	0.1	1 base in 10 wrong
20	0.01	1 base in 100 wrong
30	0.001	1 base in 1000 wrong
40	0.0001	1 base in 10 000 wrong

Q>20,
Error rate < 0.01

Q=20 (p=0.01)

Q<20,
Error rate > 0.01



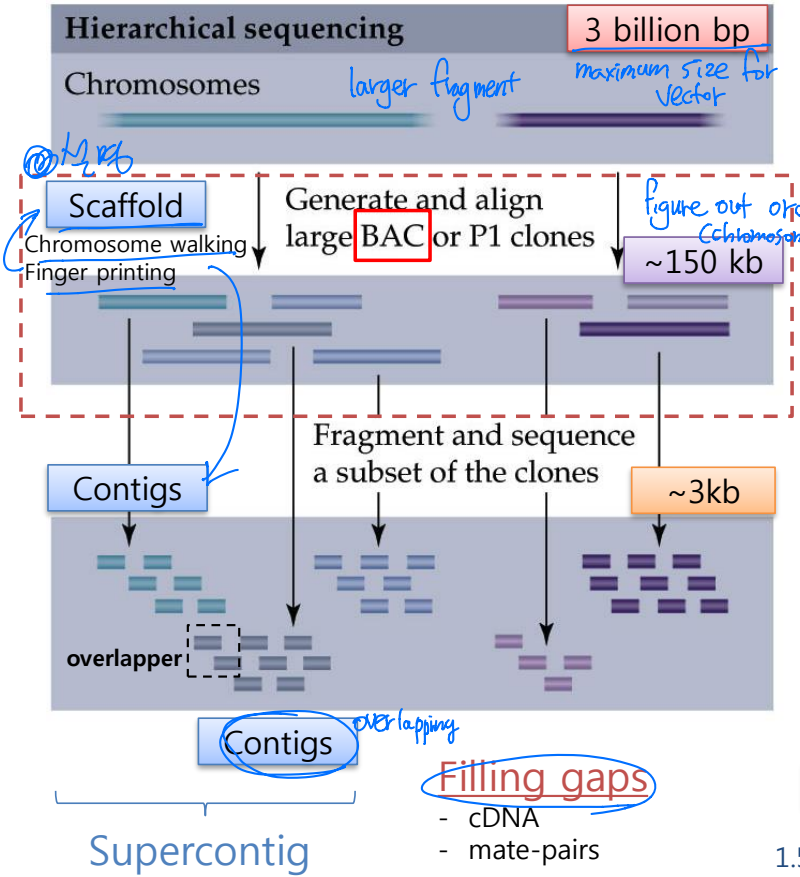
Hierarchical vs. shotgun sequencing

Human genome → Fragmentation → Sequencing → Assembly (De novo)

Hierarchical shotgun sequencing

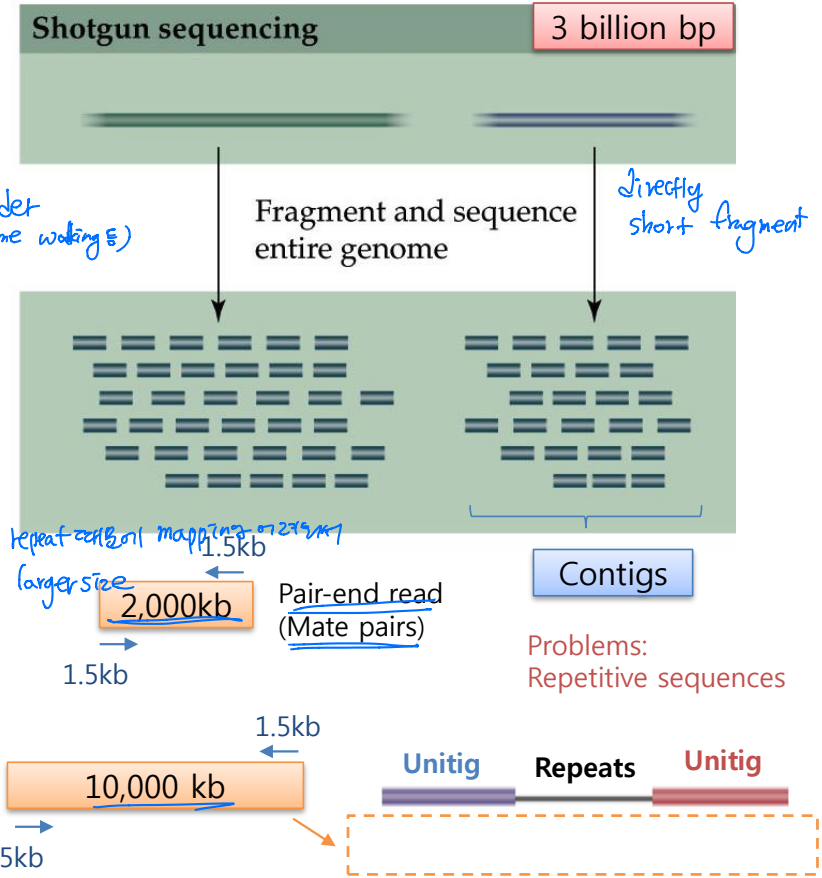
'BAC-to-BAC' method

(International consortium)



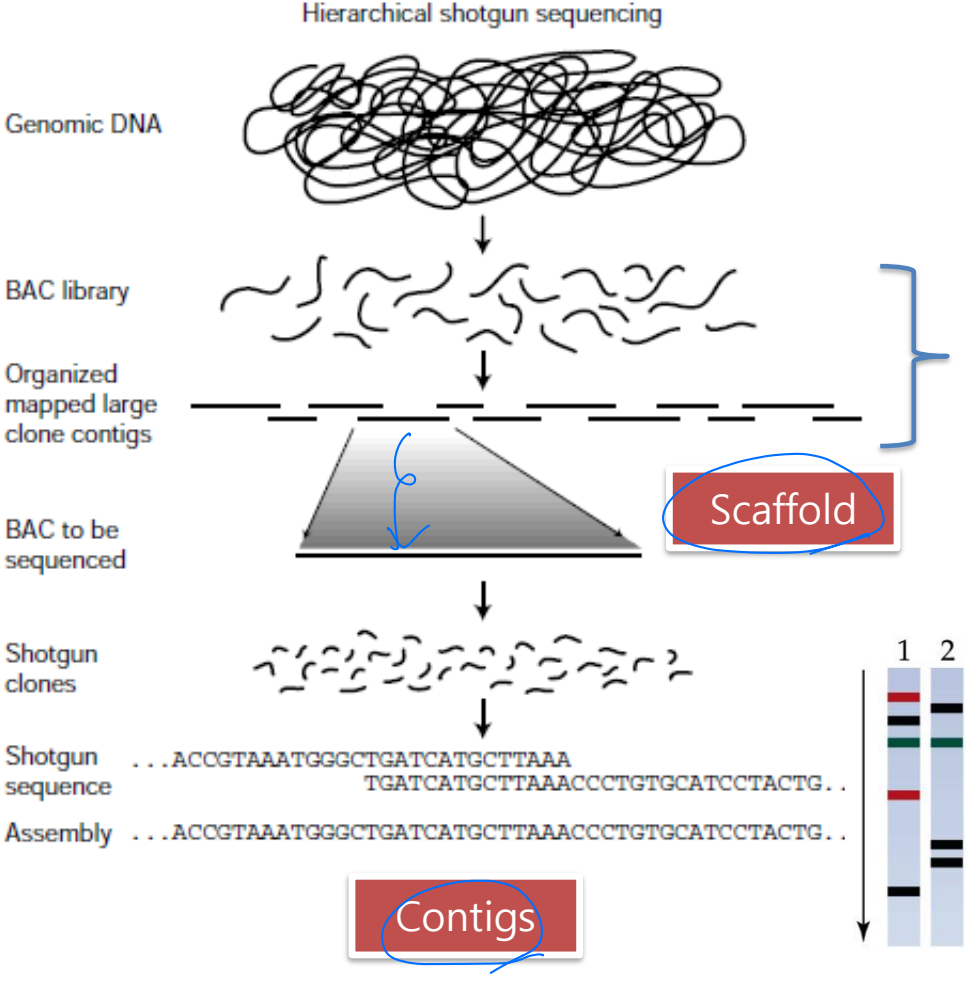
Whole-genome shotgun sequencing

(Cerela)

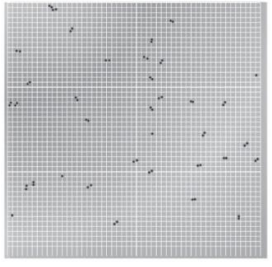


Hierarchical shotgun sequencing: Scaffold & Contigs

Aligning BAC clones by hybridization and fingerprinting



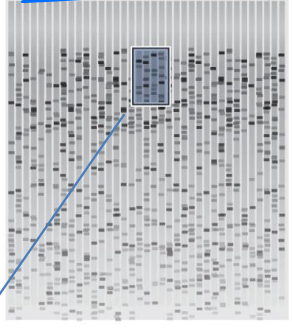
Chromosome walking ⇒ Contig 단계에서



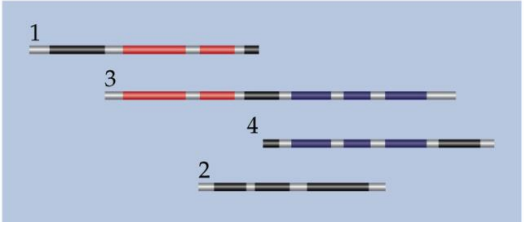
STS probes

Hybridization

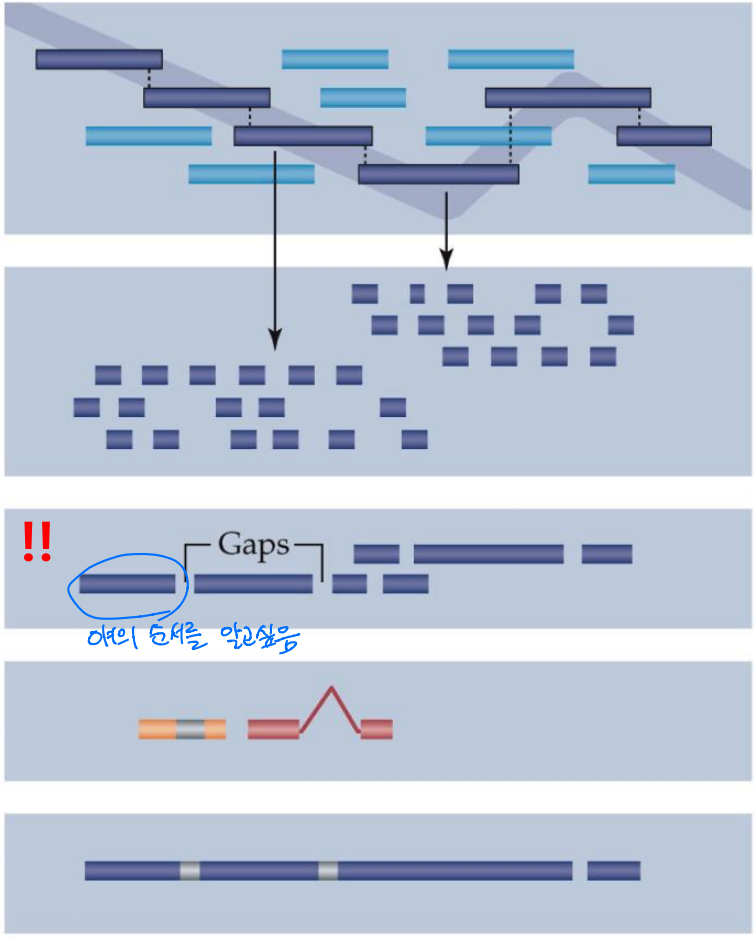
Fingerprinting



Fragments from restriction enzyme



Hierarchical assembly of a sequence-contig scaffold (supercontig)



Gaps !!

Aligned BACs with chosen tiling path

Scaffold

Fragmentation

Aligned shotgun sequences

Sequencing

Sequenced-clone contigs

Contigs

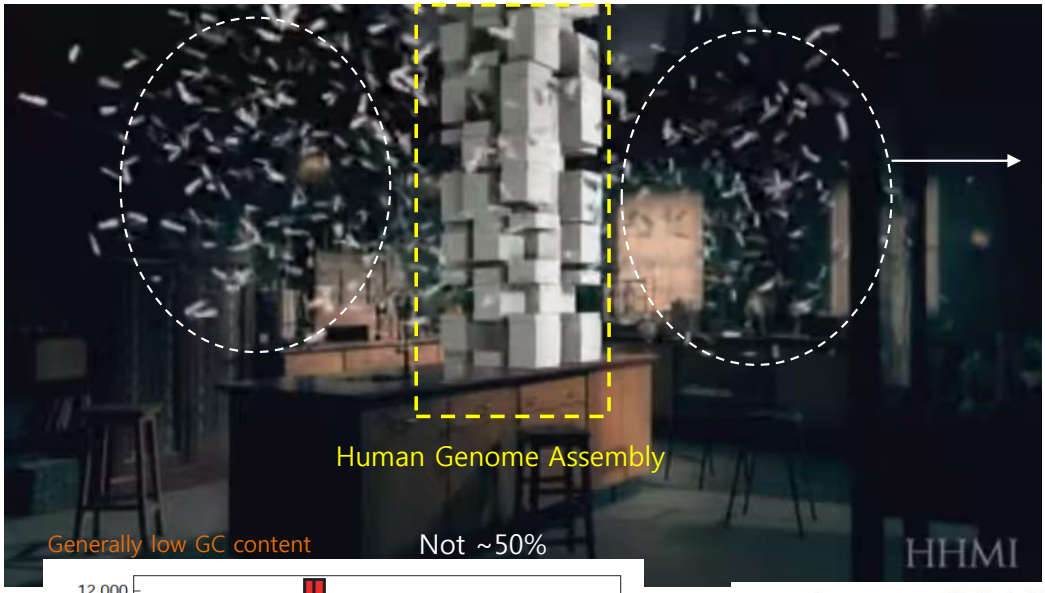
Add mate-pair and cDNA sequence information

Cover two part
⇒ figure out order

Sequence-contig scaffold

Supercontig

Human genome

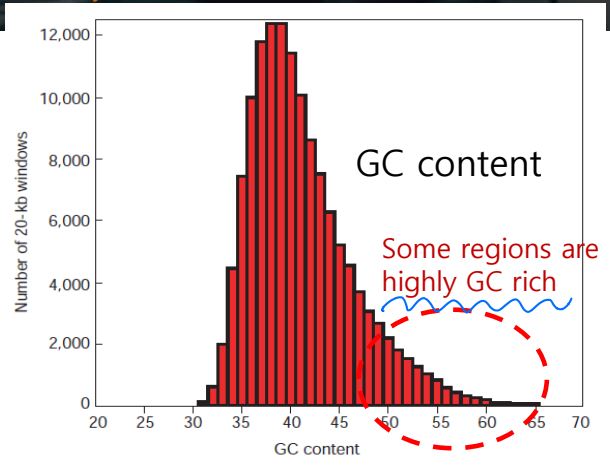


Unlocalized contigs

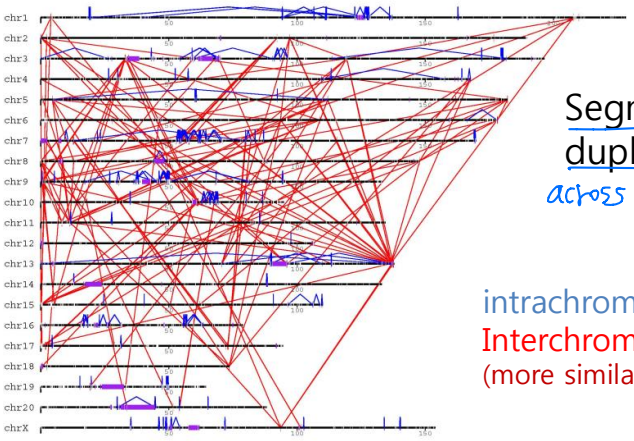
"chrUn"

Know the chromosome number, but unplaced

"Chr1_xxxx_random"



CpG island



Segmental duplication across genome

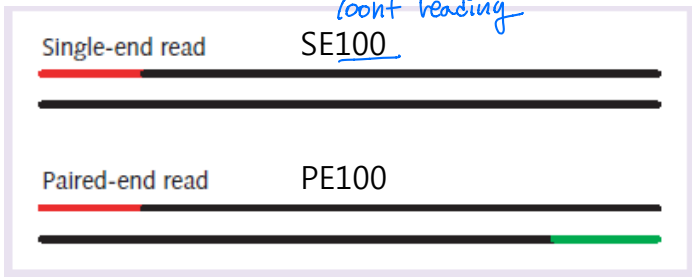
intrachromosomal
Interchromosomal
(more similar, larger)

4. Next-generation Sequencing : terminology

Fragment: a small piece of genomic DNA – typically several hundred bp in length – subject to an individual partial sequence determination, or *read*.

Single-end read: technique in which sequence is reported from only one end of a fragment (see Figure 1.7).

Paired-end read: technique in which sequence is reported from both ends of a fragment (with a number of undetermined bases between the reads that is known only approximately).



Paired-end sequence Raw sequence obtained from both ends of a cloned insert in any vector, such as a plasmid or bacterial artificial chromosome.

Coverage (or depth) The average number of times a nucleotide is represented by a high-quality base in a collection of random raw sequence. Operationally, a 'high-quality base' is defined as one with an accuracy of at least 99% (corresponding to a PHRED score of at least 20).

Read length: the number of bases reported from a single experiment on a single fragment.

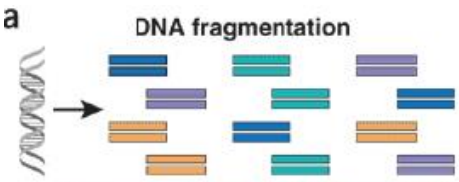
Assembly: the inference of the complete sequence of a region from the data on individual fragments from the region, by piecing together overlaps.

Contig: a partial assembly of data from overlapping fragments into a contiguous region of sequence.

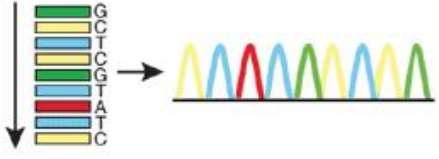
De novo *sequencing*: determination of a full-genome sequence without using a known reference sequence from an individual of the species to avoid the assembly step.

Resequencing: determination of the sequence of an individual of a species for which a reference genome sequence is known. The assembly process is replaced by mapping the fragments onto the reference genome.

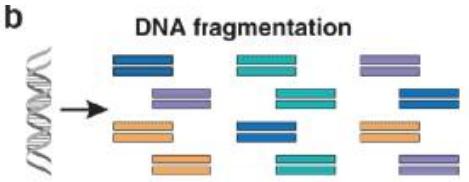
Sanger Method vs. NGS



Capillary array electrophoresis
(1 read/capillary)



100
per run



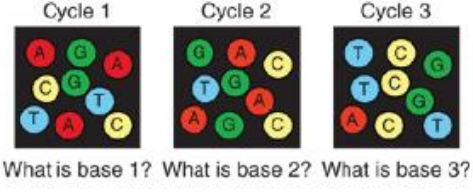
Library Preparation

Amplification

Sequencing Chemistry

>100 million
per run

Cyclic array sequencing
(>10^6 reads/array)

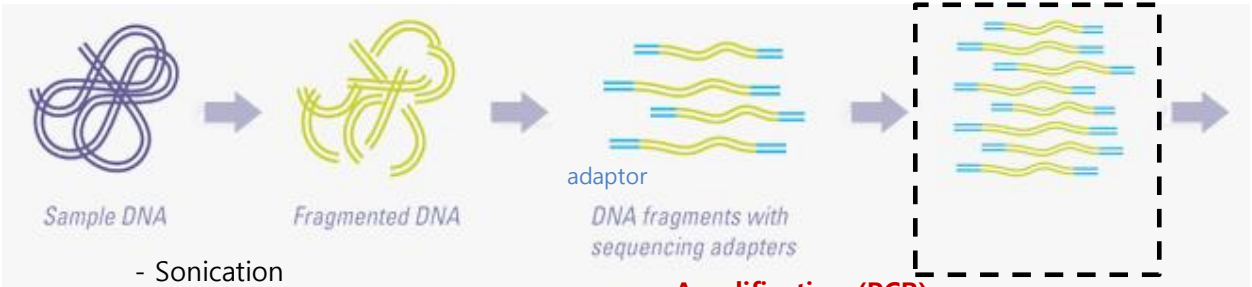


Signal detection

DNA Sequencing by NGS (flow & principle)

Preparation

Genome → **Fragmentation** → **Adaptor ligation** → **Sequencing Library**



- Sonication
- DNase

Amplification (PCR)

- Depending on NGS platform

NGS sequencing platform
(Sequencer)

Amplification

Emulsion PCR

Bridge amplification

No amplification
Single molecule

Sequencing chemistry

DNA polymerase reaction

Pyrosequencing

Chain termination method

-SBS (Sequencing by synthesis)
: reversible chain termination

Ligation, translocation (Channel)

Signal detection

→ Light (monocolor)

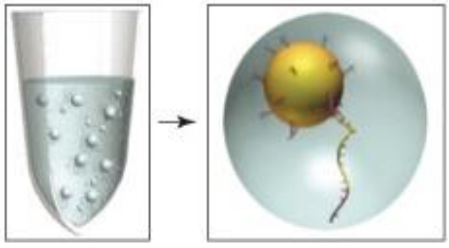
→ Fluorescence signal (4 colors)

→ Electric signal

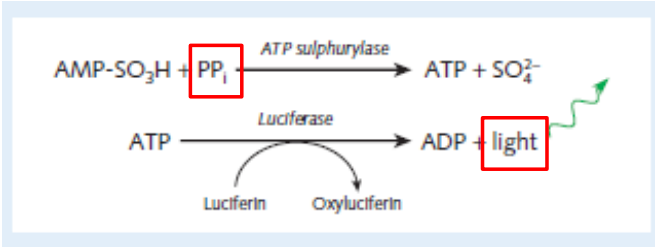
NGS platform comparison

Platform	Amplification	Chemistry	Detection	Description
454	emPCR	Pyrosequencing	Image (mono)	Low throuput, long read (~400), ~0.35 day
Abi SOLiD	emPCR	Seq by ligation	Image (color)	High throuput, short read (~100), ~2weeks
Illumina /Solexa	Solid-phase	Reversible teminator	Image (color)	High throuput, short read (~150), ~1 week
Ion torrent /proton	emPCR	Pyrosequencing	H+ (pH)	low-High throuput, short read (~100), 5 hours
Pacific Bioscience	Single molecule	realtime, polymerase	Image (color)	**longer read (3-15kb), Accuracy (~95%)**, 30min
Oxford Nanopore	Single molecule	realtime, polymerase	H+ (pH)	?

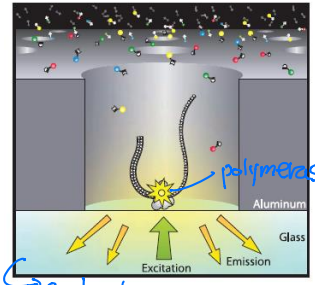
Emulsion PCR



pyrosequencing



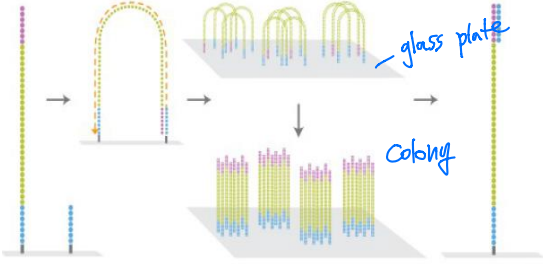
Pacific Bioscience



good at

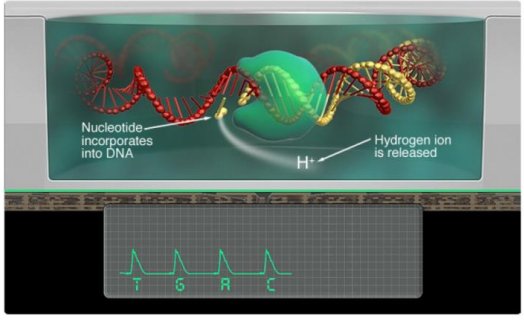
Useful for long read sequencing (resequencing genomes)

Solid-phase (bridge amplification)

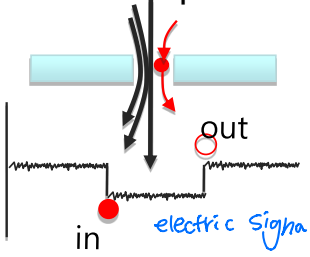


SBS : reversible terminator

Proton (pH meter)



Nanopore



5. Sequence Alignment

Sequence change

ACGCTGA

Common Ancestor

ACTGT

Evolutional Changes

ACGCTGA

```
ACGCTGA
A--CTGT
```

Sequence alignment 1

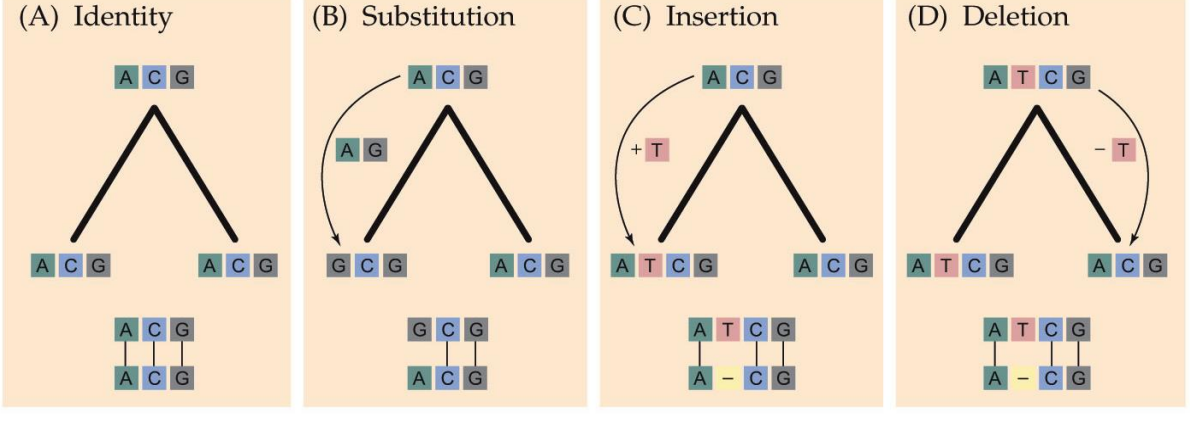
ACTGT

ACGCTGA

```
ACGCTGA
ACTGT--
```

Sequence alignment 2

ACTGT

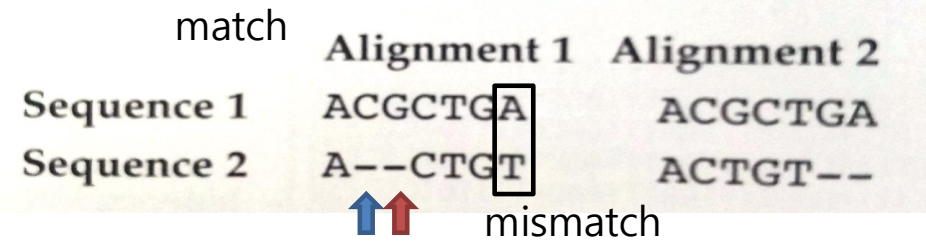


alignment ⇒ indicating evolutionary event

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**

same start, end

```
T C A G - - T - G T C G A A G T - T A
|   | |     |   |   |   |   | |   |
T - A G G C T A G - C - A - G T G T A
```

➢ **Local Alignment**

```
T C A G T G T C G A A G T T A
|   | | | | |
T A G G C T A G C A G T G T A
```

Global alignment

Conserved region of sequence
> Functional domain / element

Local alignment

High sequence similarity > Homolog > same function

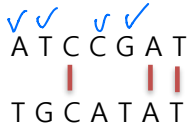
How to measure sequence similarity : Distance

- (Method 1) Counting identical letters on each position

  → **Hamming distance**

  (같은 - match)

  4 (7-3)



- (Method 2) Inserting gaps to maximize the number of identical letters

  → **Edit distance**

  (Levenshtein distance)

  5

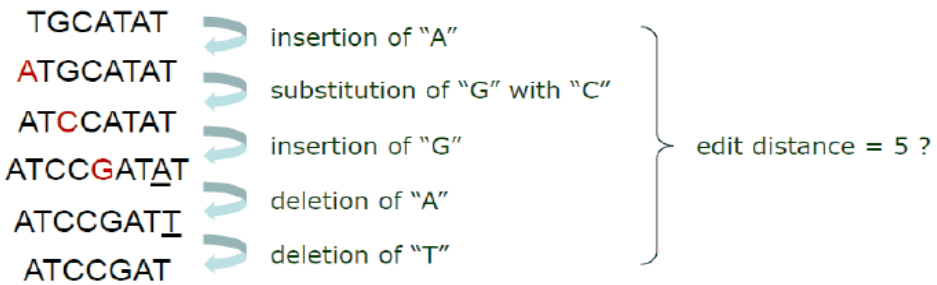
Hamming distance=3

➢ **Definition**

- Edit distance between two sequences x and y : the minimum number of editing operations (insertion, deletion, substitution) to transform x into y

➢ **Example**

- x="TGCATAT" (m=7), y="ATCCGAT" (n=7)



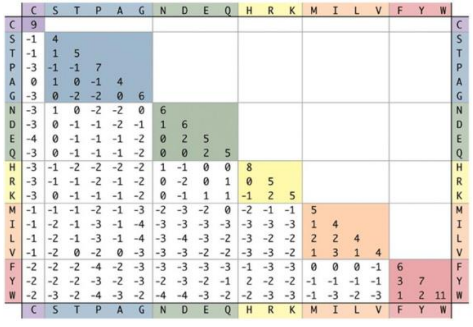
How to measure sequence similarity : Score

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

Affine gap penalty

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

Local alignment

:Smith-Waterman algorithm

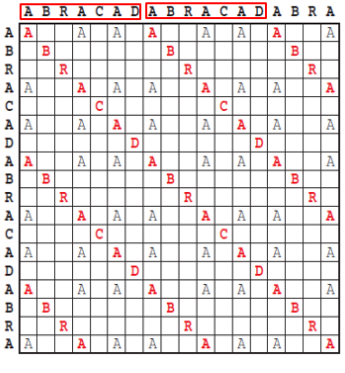
Global alignment

:Needleman-Wunsch algorithm

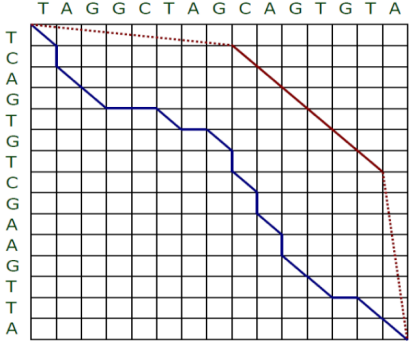
```
DOROTHY--------HODGKIN
DOROTHYCROWFOOTHODGKIN
```

- Dot Plot

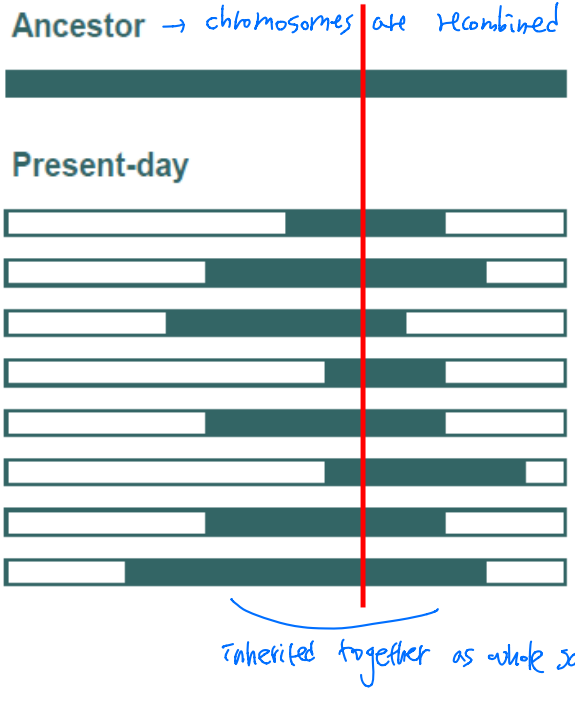
Repetitive sequence



-Dynamic Programming:



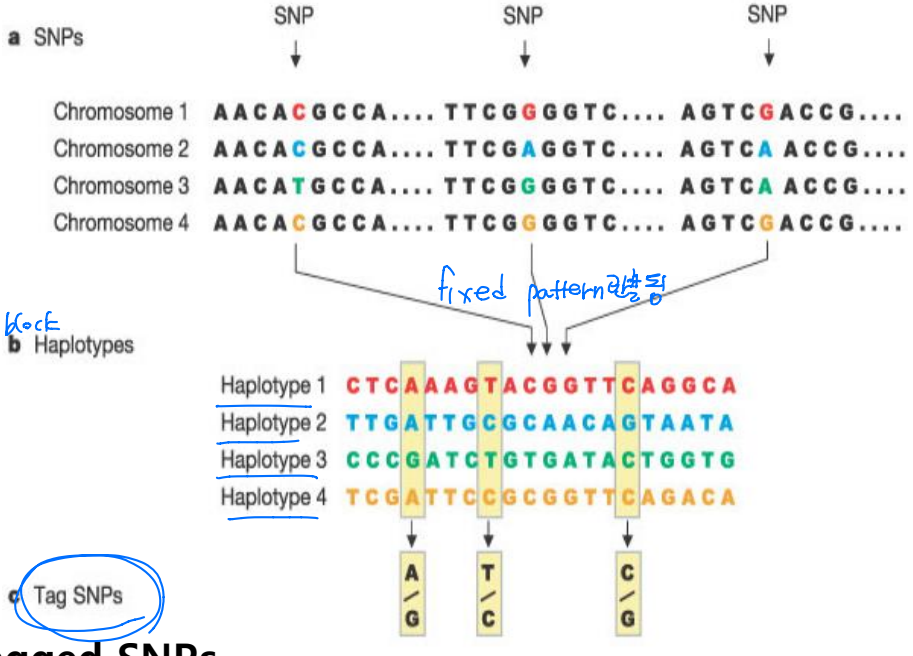
5. Haplotyping : LD analysis



- **Haplotypes**
  : local combinations of genetic polymorphisms that tend to be co-inherited

- A collection of specific alleles (SNPs) in a cluster of tightly-liked genes on a chromosome

99%는 같기때문에 SNP를 marker로 이용



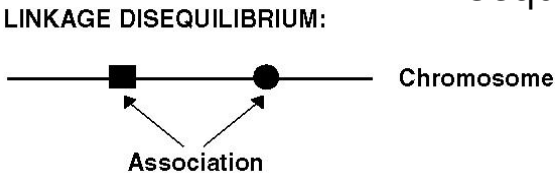
Construction of HapMap

Linkage Disequilibrium analysis

Tagged SNPs

Linkage disequilibrium (LD)

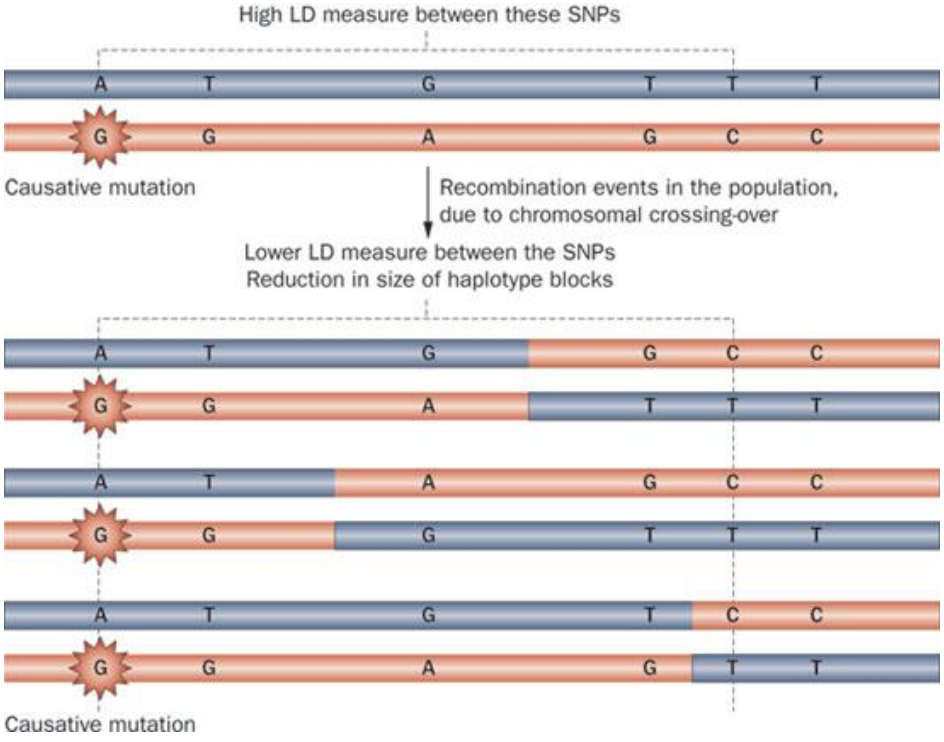
Sequencing results (population)



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

Two SNPs

linked



How much it becomes disequilibrium ?

Completely linked

항상같이 관찰됨

Linked

Completely independent

observed = expected

→ 많은 recombination 때문에

Calculation of linkage disequilibrium (LD)

catcag aag
tatcag aac
tatcag aac
catcag aag
catcag aag
tatcag aac
catcag aag
catcag aag
tatcag aac
catcag aag

observed

	c	t	total
g	Pcg	Ptg	Pg
c	Pcc	Ptc	Pc
	Pc	Pt	1

≠ not independence

expected

	c	t	total
g	Pc X Pg	Pt X Pg	Pg
c	Pc X Pc	Pt X Pc	Pc
	Pc	Pt	1

D : linkage disequilibrium coefficient

	c	t	total
g	Pcg = Pc X Pg + D		Pg
c			Pc
	Pc	Pt	1

observed = expected + D

D = 0 ~ 1

	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

how much they are linked

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

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

Completely linked
(D=Dmax) (D′=1)

Linked

min{ p2q1 , p1q2 }

Completely independent
(D=0) (D′=0)

observed = expected

EXERCISE 3.1 *Quantifying heterozygosity and LD*

Using the following ten sequences:

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

(a) Count the number of segregating sites.

(b) Calculate the expected average heterozygosity per nucleotide.

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

Pick two haplotypes -> heterozygote ?

	A	a
	p	$1-p$
A p	p^2	pq
a $1-p$	pq	$(1-p)^2$

Do given sequence results contain enough variations to determine haplotype ?

(length=50 nt)

9 SNPs

10 sequence results (human)

10 sequence of haploid

```
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
```

0.1 0.4 | 0.4 0.1 | 0.5 0.1 | 0.2 0.4 | 0.2

Diversity [(41 x 0) + (3 x 0.1) + (3 x 0.4) + (2 x 0.2) + 0.5] / 50 = **0.048**

0.048 x 50 = 2.4

Expected average heterozygosity per nucleotide (H)

Select two haploid sequences from those above to make diploid,
then see how much their sequence is expected to be different to form heterozygosity

[(41 x 0) + (3 x 0.1 x 0.9 x 2) + (3 x 0.4 x 0.6 x 2) + (2 x 0.2 x 0.8 x 2) + 0.5 x 0.5 x 2] / 50 = **0.0624**

0.0624 x 50 = 3.12

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

cgga
tcga
tcaa
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:

Tagged SNP

	Site 5	Site 13	
		c ($q_1 = 0.4$)	**g** ($q_2 = 0.6$)
t ($p_1 = 0.4$)		$p_{11} = 0.4$	$p_{12} = 0.0$
c ($p_2 = 0.6$)		$p_{21} = 0.0$	$p_{22} = 0.6$

	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

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

0.16

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

completely linked ⇒ 5t, 13c는 항상 same block

0.4 x 0.6 , 0.6 x 0.4

D′ = 0.24/0.24 =1

Complete LD !!!

completely independent

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.

Haplotyping : linkage disequilibrium

A	B
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	2
2	1
2	1
2	2
2	2
2	2
2	2

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

Expected

p1=10/16 = 0.625

q1=11/16 = 0.629 p1q1=0.39

observed

	B_1	B_2	Total
A_1	9/16 $= p_1q_1 + D$	1/16 $= p_1q_2 - D$	p_1
A_2	2/16 $= p_2q_1 - D$	4/16 $= p_2q_2 + D$	p_2
Total	q_1	q_2	1

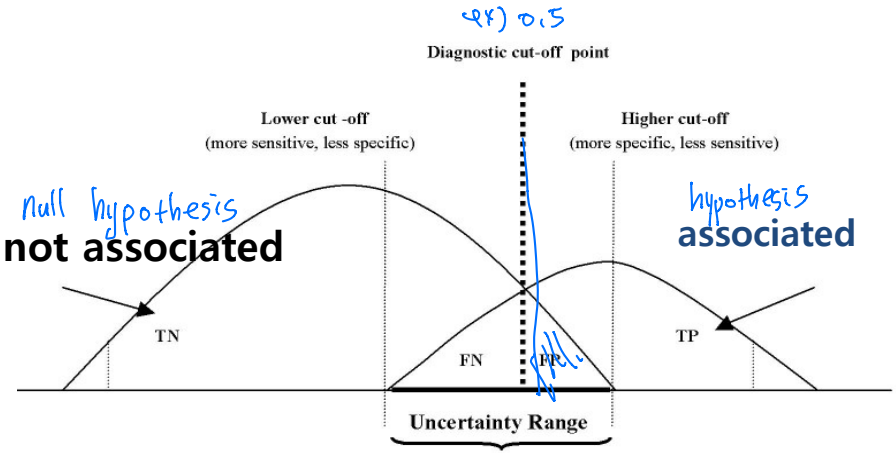
D=0.133

Dmax
=min (0.625 x 0.371, 0.375 x 0.629)
= 0.23

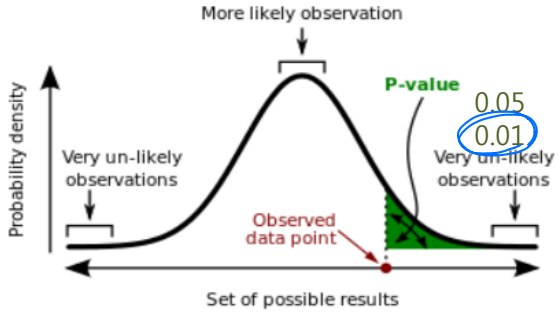
D'=0.133 / 0.23 = 0.58

Whether it is significant ?

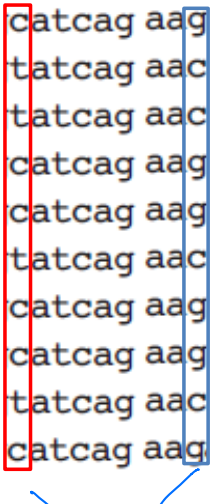
Statistical significance : Chi square test



Null hypothesis (H_0) Alternative hypothesis (H_1)



Null hypothesis (H_0) **Observation**



Degrees of freedom (df) = n-1 = 4-1 = 3

observed

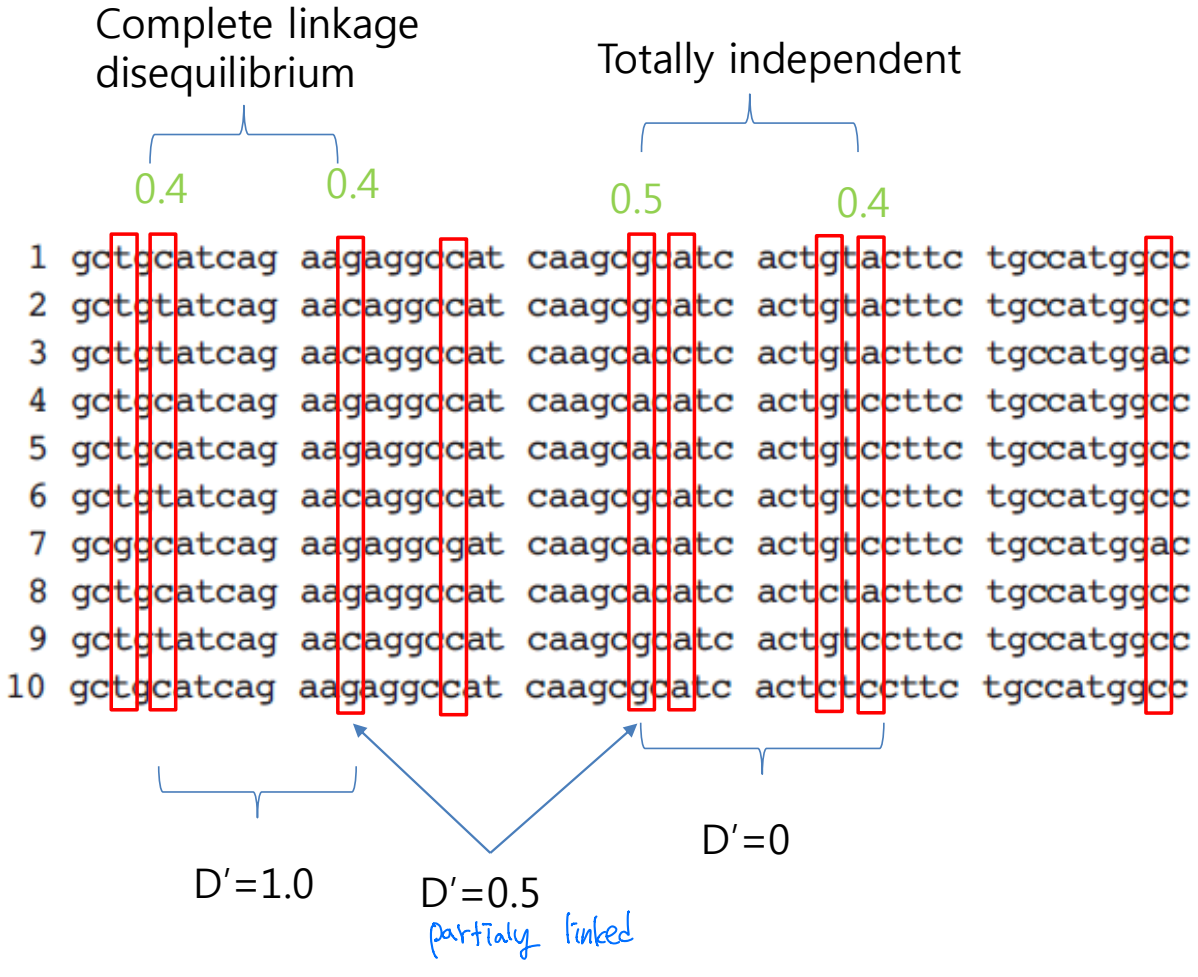
	c	t	total
g	Pcg	Ptg	Pg
c	Pcc	Ptc	Pc
	Pc	Pt	1

Expected

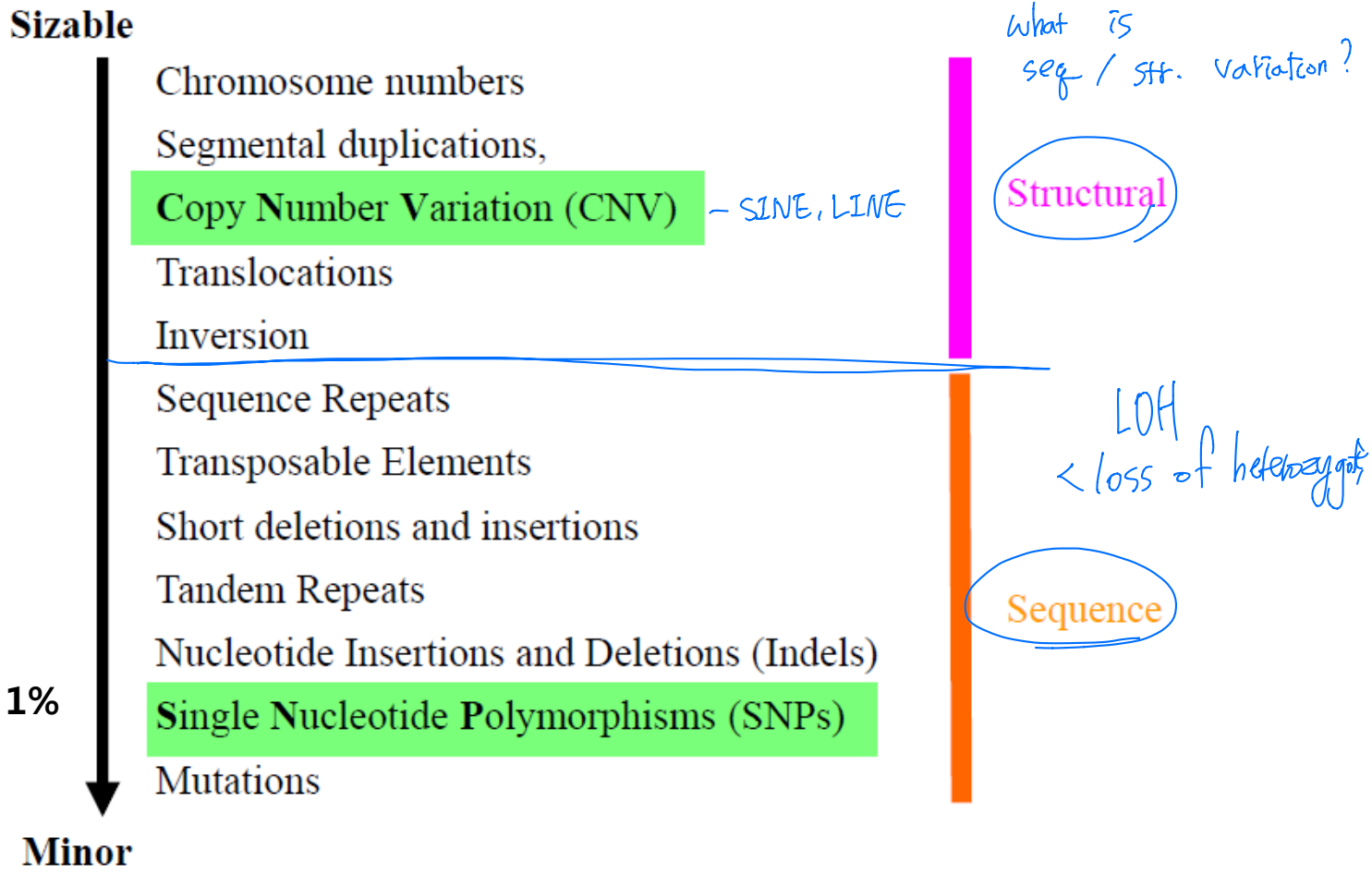
	c	t	total
g	Pc X Pg	Pt X Pg	Pg
c	Pc X Pc	Pt X Pc	Pc
	Pc	Pt	1

Chi square test (p=0.067889)

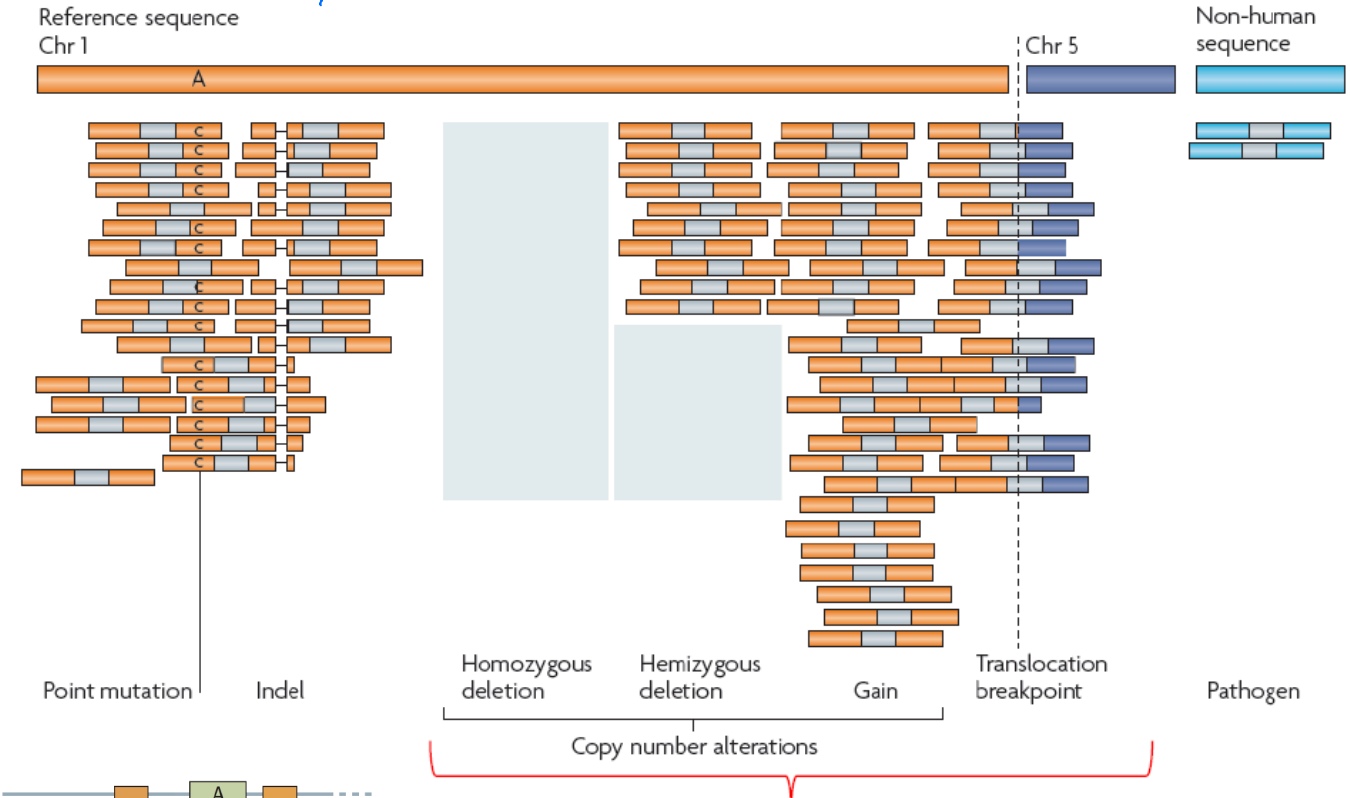
linked.


Complete linkage disequilibrium
Totally independent
0.4
0.4
0.5
0.4
1 gctgcatcag aagaggccat caagcgcatc actgtacttc tgccatggcc
2 gctgtatcag aacaggccat caagcgcatc actgtacttc tgccatggcc
3 gctgtatcag aacaggccat caagcacctc actgtacttc tgccatggac
4 gctgcatcag aagaggccat caagcacatc actgtccttc tgccatggcc
5 gctgcatcag aagaggccat caagcacatc actgtccttc tgccatggcc
6 gctgtatcag aacaggccat caagcgcatc actgtccttc tgccatggcc
7 gcggcatcag aagaggcgat caagcacatc actgtccttc tgccatggac
8 gctgcatcag aagaggccat caagcacatc actctacttc tgccatggcc
9 gctgtatcag aacaggccat caagcgcatc actgtccttc tgccatggcc
10 gctgcatcag aagaggccat caagcgcatc actctccttc tgccatggcc
D'=1.0
D'=0.5
partialy linked
D'=0

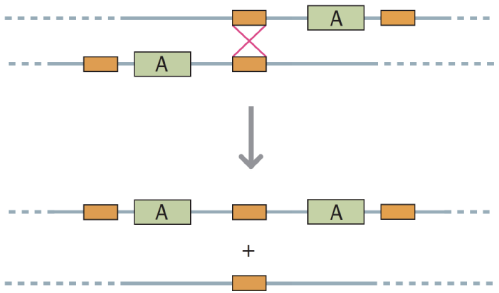
6. Genomic Variations



Structural Variations (SVs)

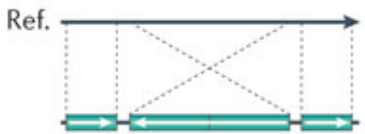


Structural variations



Gene duplication
>
copy number variation

Inversion

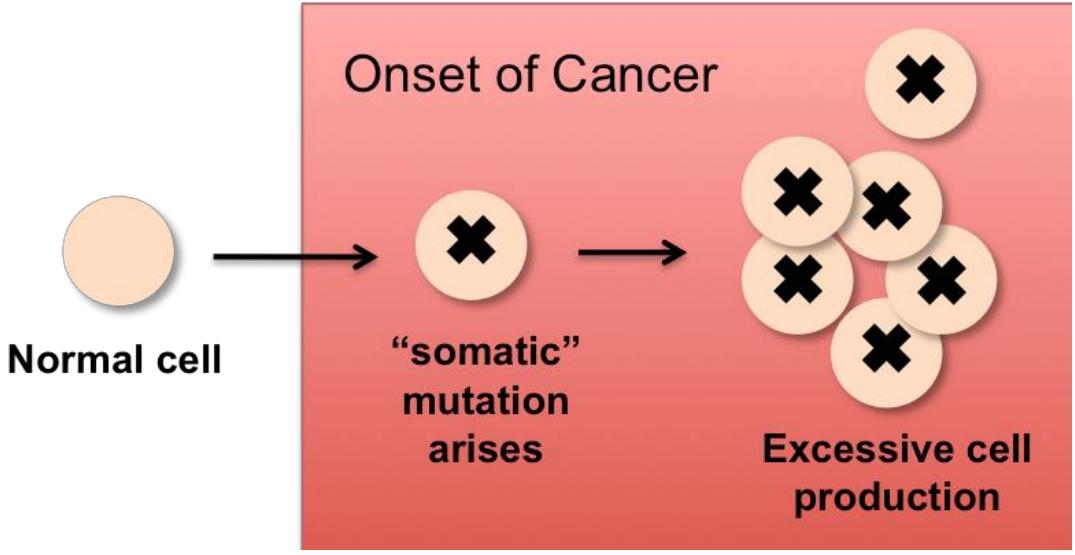


Translocation

Genetic Variations

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

- Somatic variation (Tumor tissue)
  : Cancer genomics

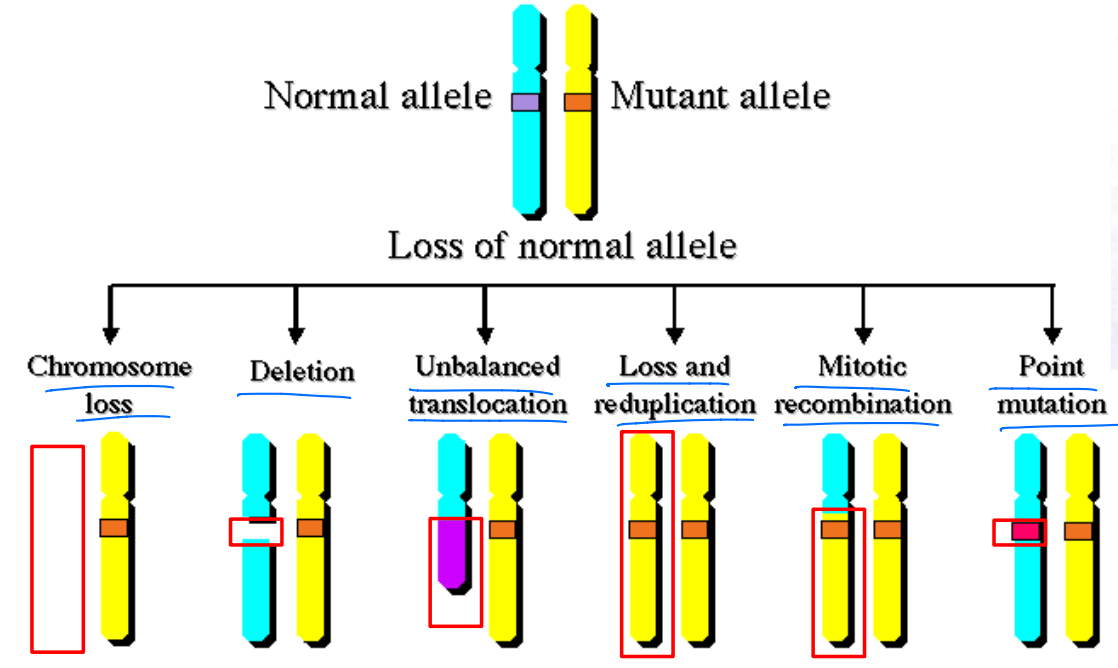


Loss of heterozygosity (LOH)

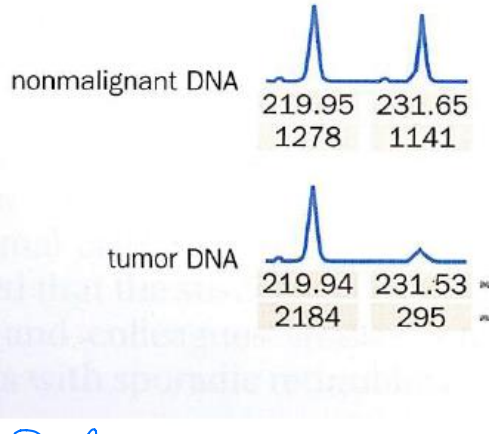
tumor에서 잘 관찰됨

Loss of heterozygosity as a marker to locate tumor suppressor genes

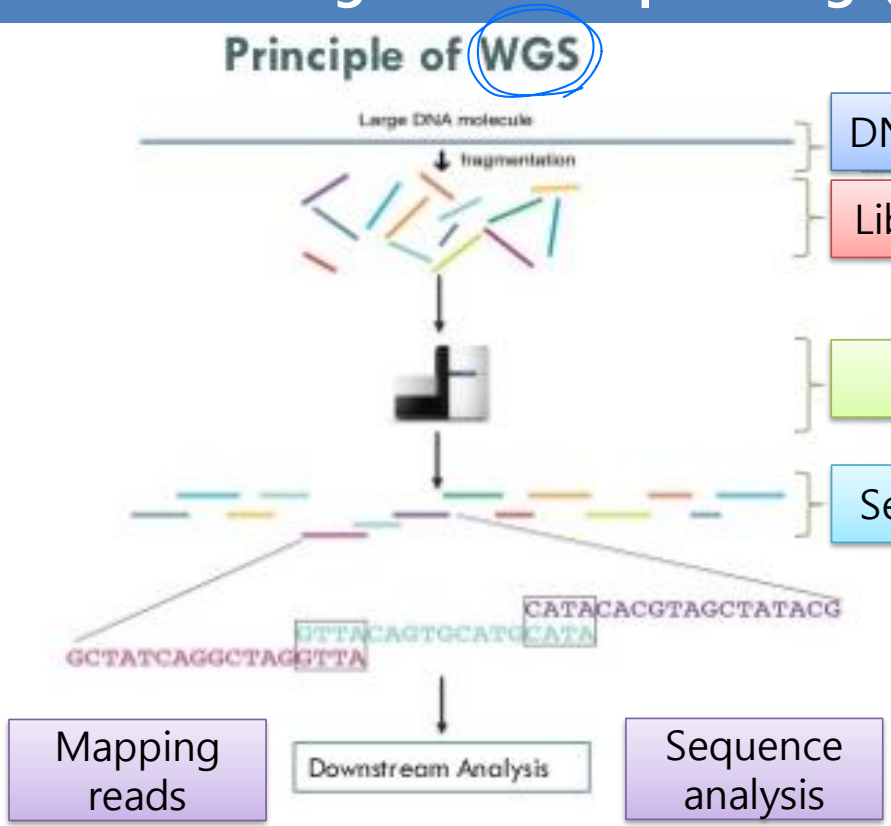
- Somatic genetic changes in retinoblastoma caused loss of heterozygosity (LOH) at markers close to the RB1 locus
- By screening paired blood and tumor samples with markers spaced across the genome, we may discover the locations of tumor suppressor genes



few case에서 loss of heterozygosity 있음



7. Whole genome sequencing (WGS by NGS) for variant

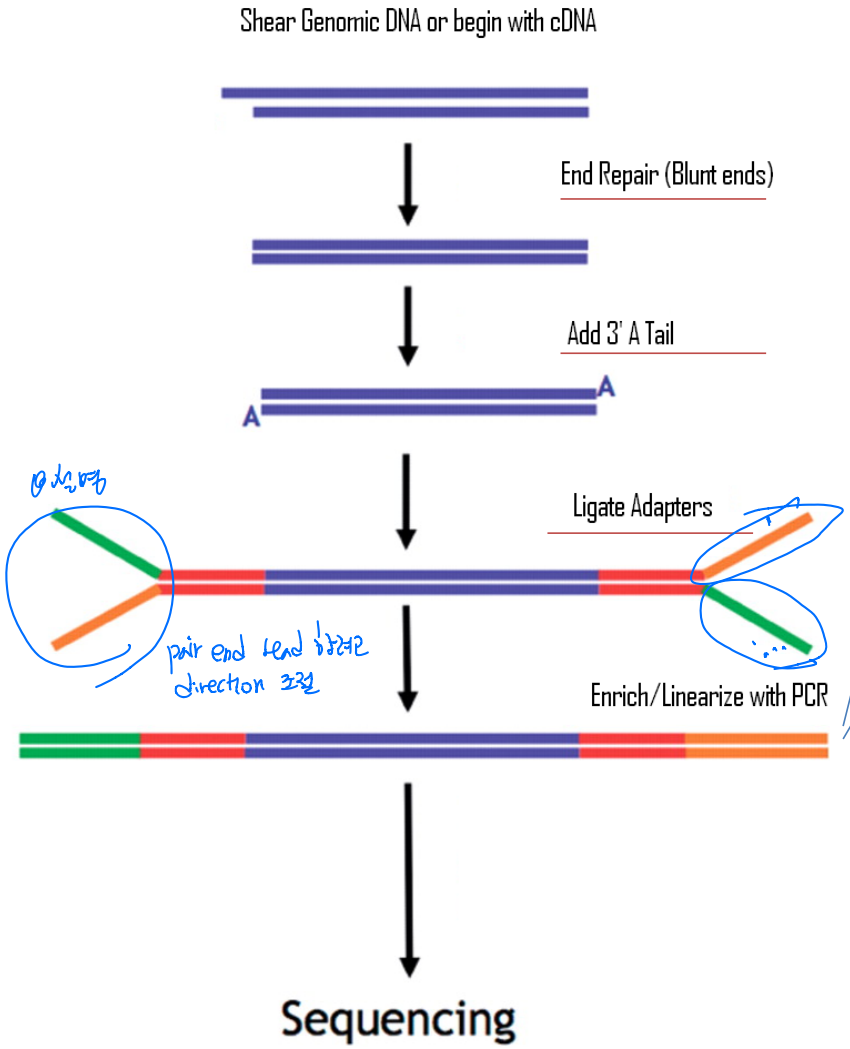


Sequencer - Illumina HiSeq 1500
Technique - Paired-End sequencing
Coverage - 100x
Read Length – 100-250 basepairs

Require high coverage for variant identification

Advantages	Disadvantages
-Point mutations -CNVs -Small indels -Rearrangements -Somatic mutations in non-coding regions (promoters, enhancers, and non-coding RNAs)	-Point mutations and indels: >30-fold haploid coverage -Rearrangements: >10-fold physical coverage

NGS Library preparation



1. DNA fragmentation / Target Selection
   PCR amplification for specific region of interest
2. Adapter ligation
3. Size selection
4. Library quantification (QC)

Sequencing data from NGS : FASTQ format & Q score

FASTQ

Label

Sequence

```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

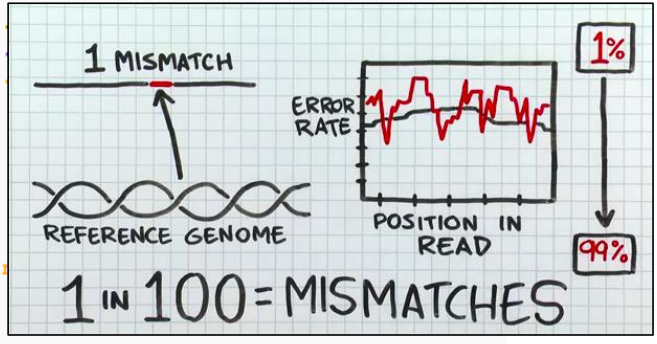
25

Q scores (as ASCII chars)

PHRED Q score :

$Q=-10\log_{10}(P)$

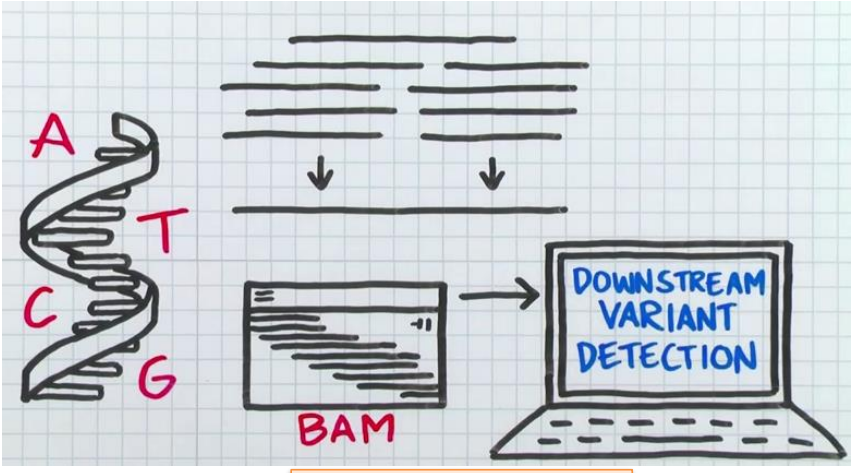
```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
.........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
.............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
.............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                                                  |                       |
33                        59   64       73                                                 104                     126
 0.....................26...31.......40
                     -5....0........9.....................
                           0........9.....................
                             3.....9.....................
  0.2....................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control In
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```



NGS WGS analysis for variation detection

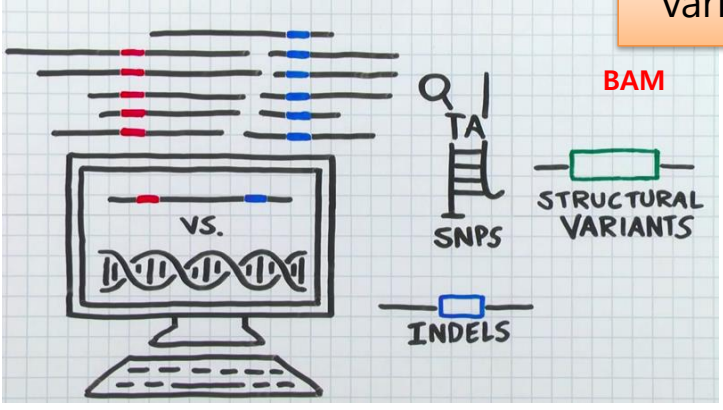
Library preparation → NGS → FASTQ → Mapping → BAM

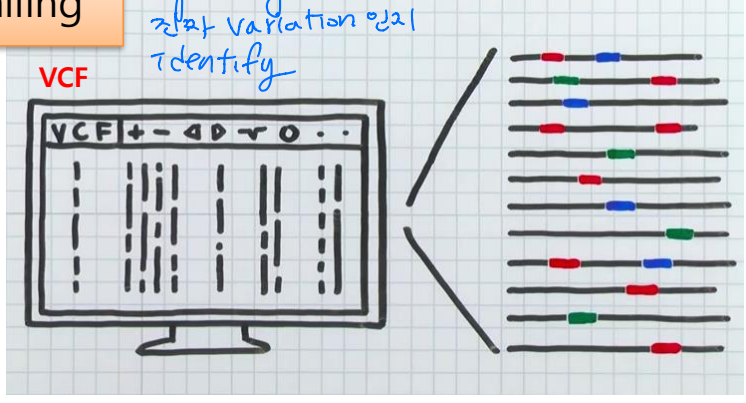
way to identify Variation



Variant calling

Sequencing error 인지 진짜 variation 인지 identify





https://www.youtube.com/watch?v=6JfXU5iOAD0&list=PLGlvFEwL2wDFVLBA1egJFwHyetpVDPExK&index=11

8. Genome-wide association study (GWAS)

1. Noncoding SNPs, 5'UTR, 3'UTR, intron, intergenic regions;

2. Coding SNPs,

- nonsynonymous or missense or replacement polymorphism
- Synonymous or sense polymorphism
- Regulatory polymorphism : Synonymous and noncoding polymorphism

→ **Phenotype**

.

Genome-wide association study (GWAS)

• An examination of *genetic variation* across a given genome

• Designed to identify genetic associations with observable traits –Such as blood pressure or weight, –or why some people get a disease or condition

Quantitative trait loci (QTLs) are stretches of DNA linked to, or containing, the genes that underlie a quantitative trait.

-Mapping regions of the genome that contain genes involved in specifying a quantitative trait is done using molecular tags, SNPs
(scan genome with SNPs for linkage to QTL)

QTL linkage mapping (QTL analysis)

Overview of genome-wide association study (GWAS)

Sample design / Collection

② WGS

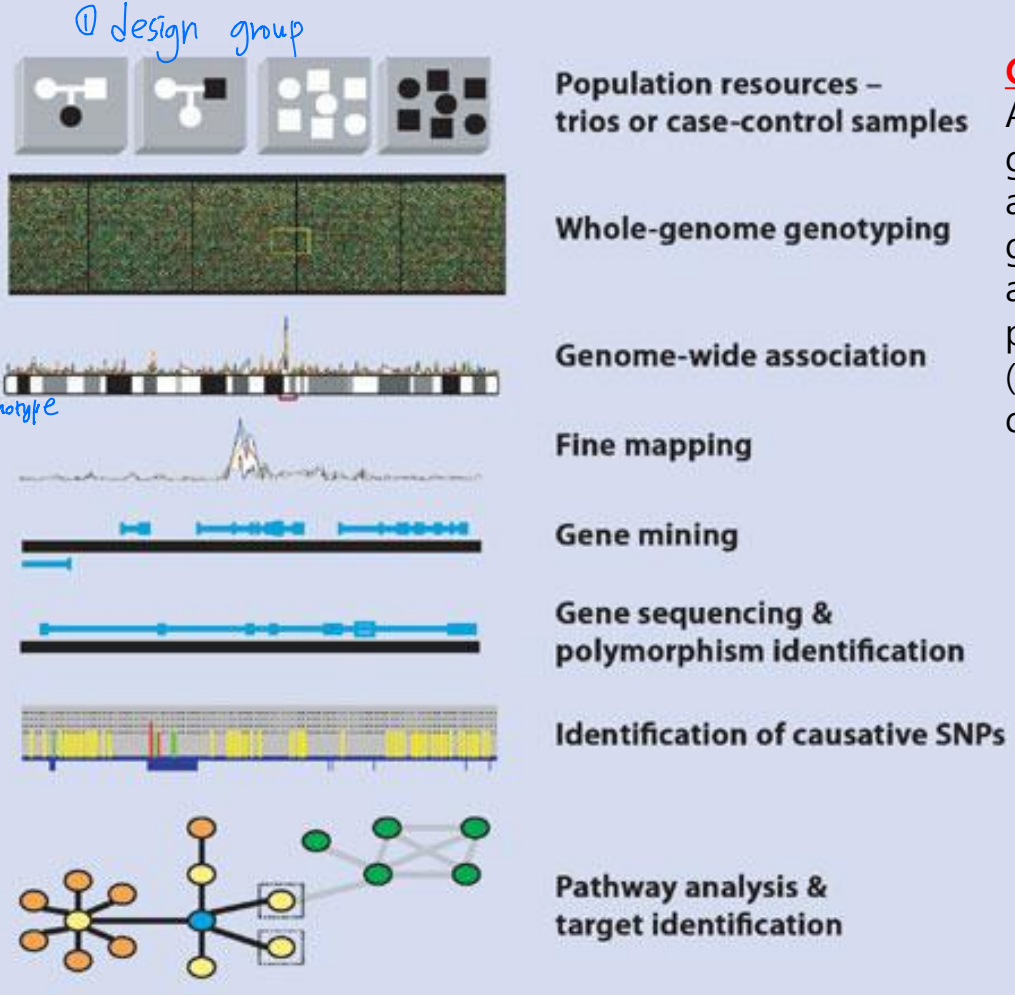
③ GWAS
to find variation associated with phenotype

Associated Loci
QTL

Variation & Gene

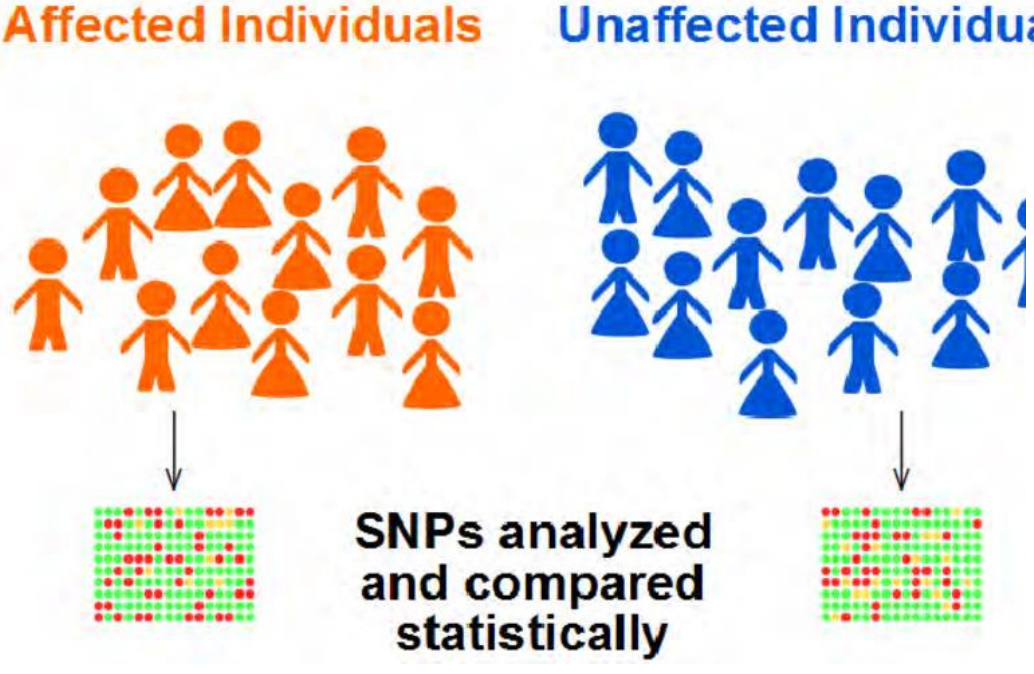
Causative variation ?

Functional interpretation validation



GWAS:
An examination of genetic variation across a given genome whether it associated with phenotypes (quantitative traits, diseases)

GWAS: Genome-wide association study



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	actctactta	tumor	2
9	actgtccttc	tumor	3
10	actctcctta	tumor	4

이 SNP가 disease랑 상관있나?
NI랑 비슷

Distribution of SNPs : Population genetics

- SNPs – phenotypic variation (Quantitative genetics)
- QTL (quantitative trait loci) = polygenic phenotypic variation

Linked?

SNPs ⟷ Phenotype

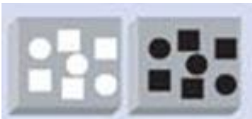
The Genome-wide Association Study (GWAS)

Affected (Disease)

WGS (Variation)

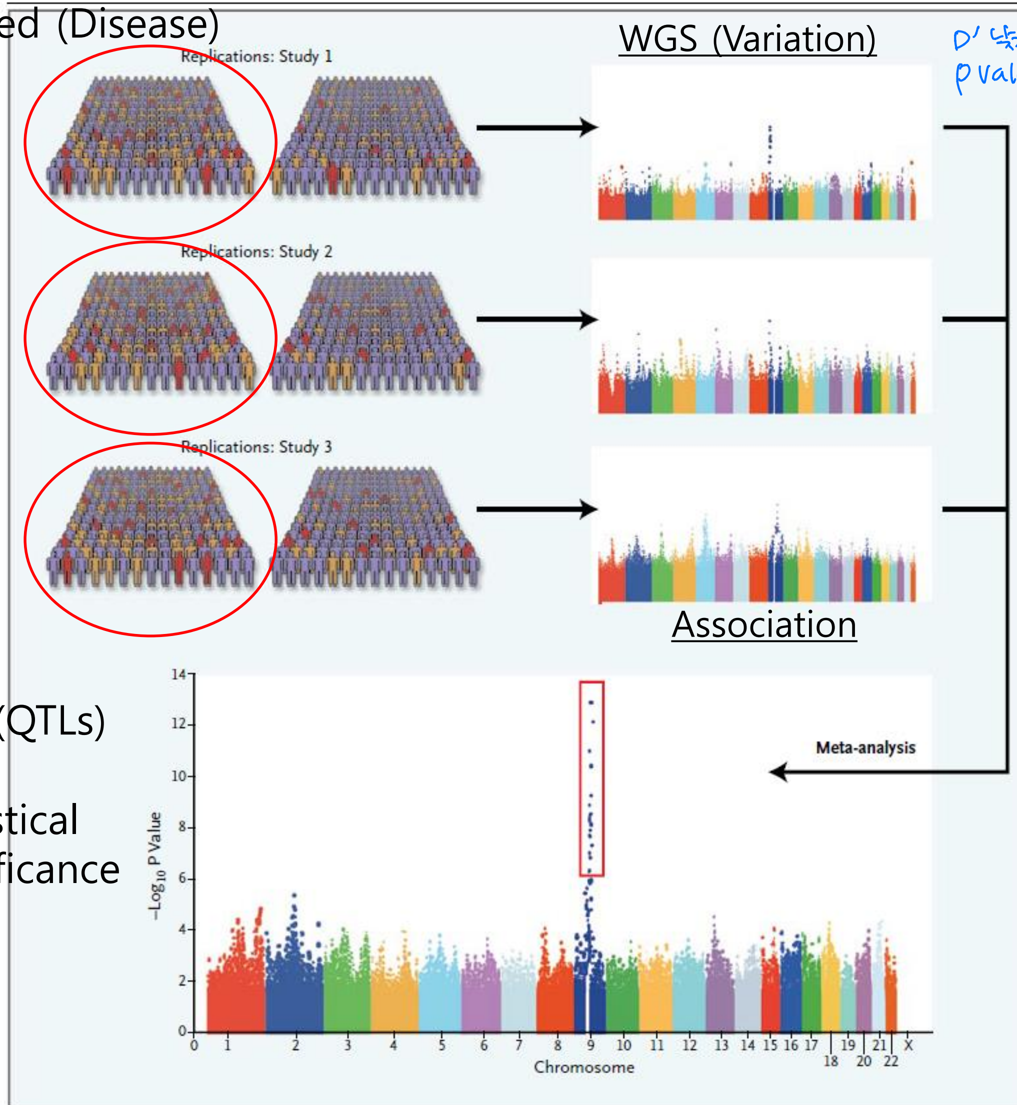
D' 낮기때문에 pvalue이득

Case-control designs



Disease

Family based designs (Trio)



Association

Quantitative trait loci (QTLs)

Statistical significance

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	actctactta	tumor	2
9	actgtccttc	tumor	3
10	actctcctta	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

Dmax= min (0.3, 0.2) =0.2

D′ = 0.1/0.2 = 0.5 (?)

Chi square test (p=0.067889)

Permutation
(False-discovery rate)

observed

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

observed

	a	c
Normal	0.3	0.2
Tumor	0.1	0.4

Expected

Pn = 5/10 = 0.5 Pa = 4/10 = 0.4

Pt = 5/10 = 0.5 Pc = 6/10 = 0.6

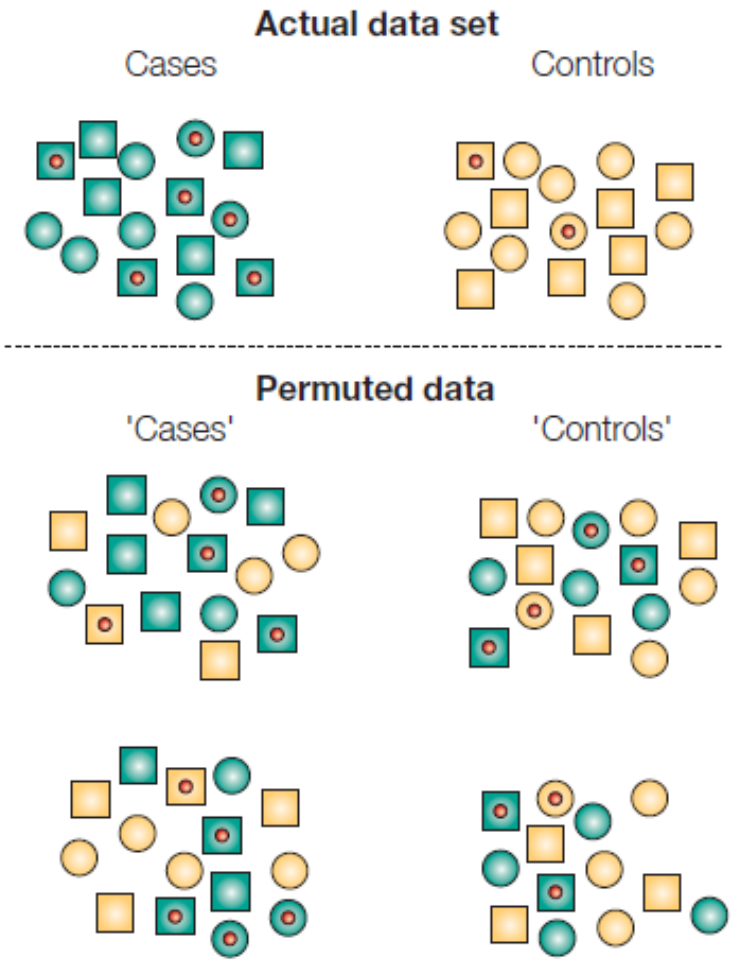
Expected

	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

Permutation testing



1. Calculate test statistics of interest in actual data set

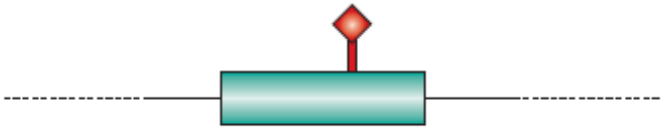
mixing

2. Calculate same test statistics in each permuted data set, and record best result for each permutation

3. To obtain significance of best actual test statistic, compare with distribution of best permuted statistics

⇒ 인위적인 random distribution that can be used as background or null hypothesis

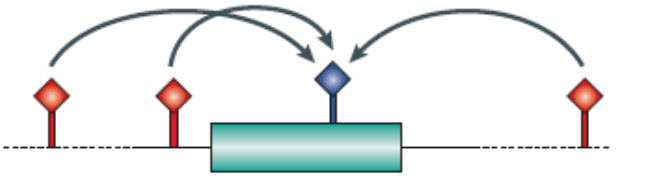
Testing SNPs for association by direct and indirect method



Direct association

Alteration of amino acids
Change in protein stability & function

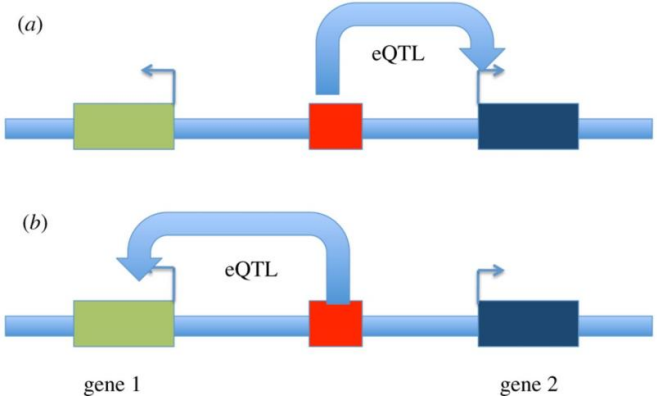
affect to function or expression



Indirect association

Alteration of neighboring gene expression

Alteration of splicing of neighboring genes.

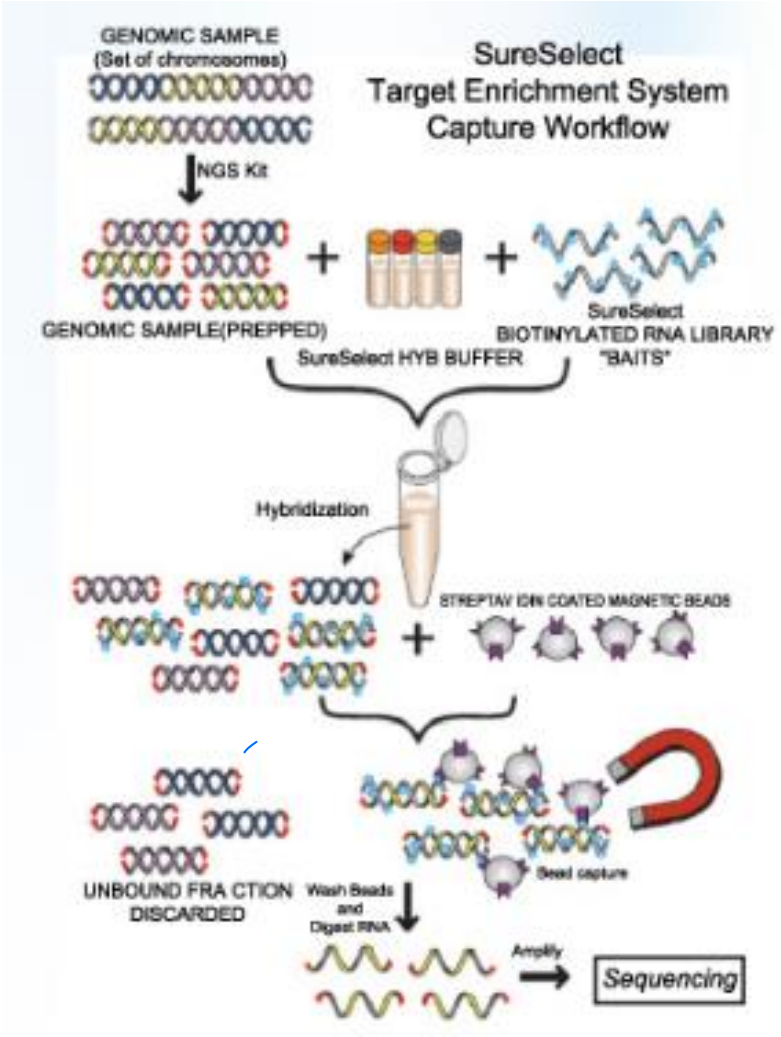


other kind of GWAS study

Concept

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

9. Exome-Seq & Functional genomics



Exome Capture by hybridizing with probes from exons

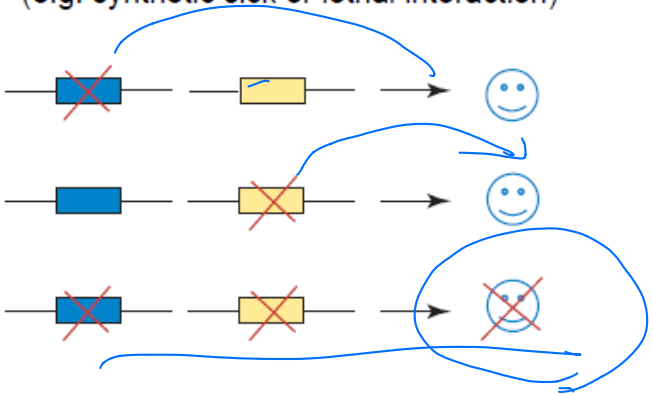
•Advantages
-Higher coverage
(More confident in variations)
-less raw sequences and cost
-Elimination of background noise

•Disadvantages
-Incomplete exome
(depending on annotation)
-Limited to mutations in only CDS
-Difficult to detect structural variations
-Intrinsic variations caused by probe hybridization

Genetic interaction

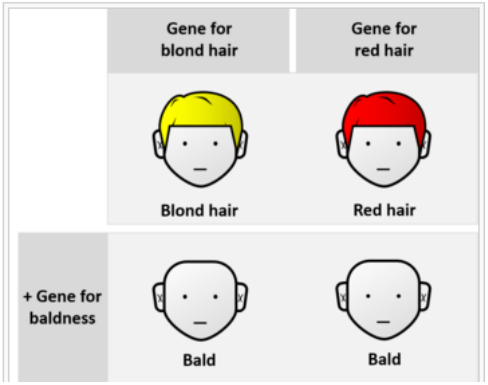
- **Forward genetics** (phenotype→ gene) : random mutagenesis
- **Reverse genetics** (gene → phenotype) : targeted mutagenesis

Genetic interaction
(e.g. synthetic sick or lethal interaction)



Type of screen
Loss-of-function
Gain-of-function
Dominant-negative
Modifier

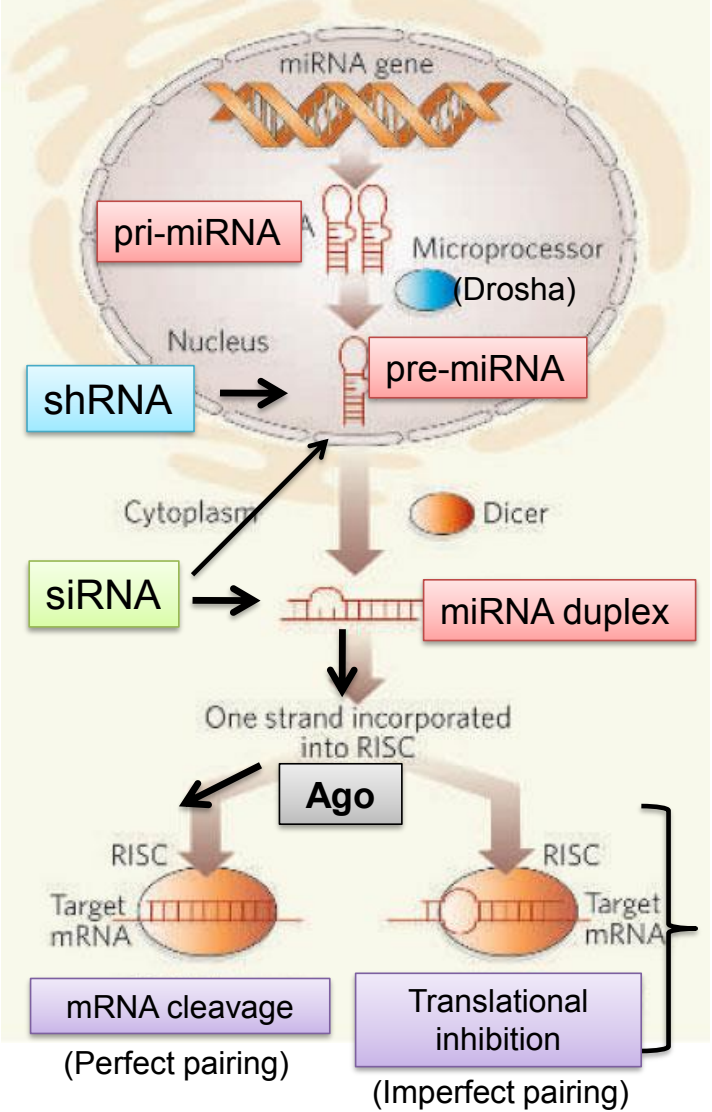
Epistasis



Dominant Negative

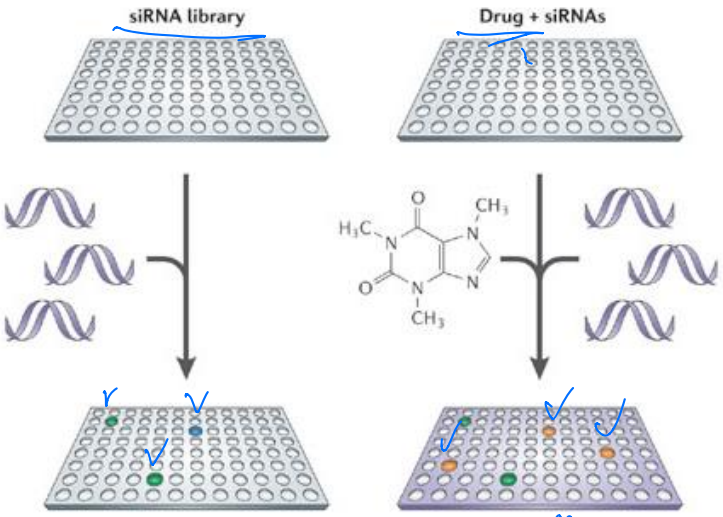
Non-Functional

RNAi screnning : functional genomics



Genome

RNAi : loss of function

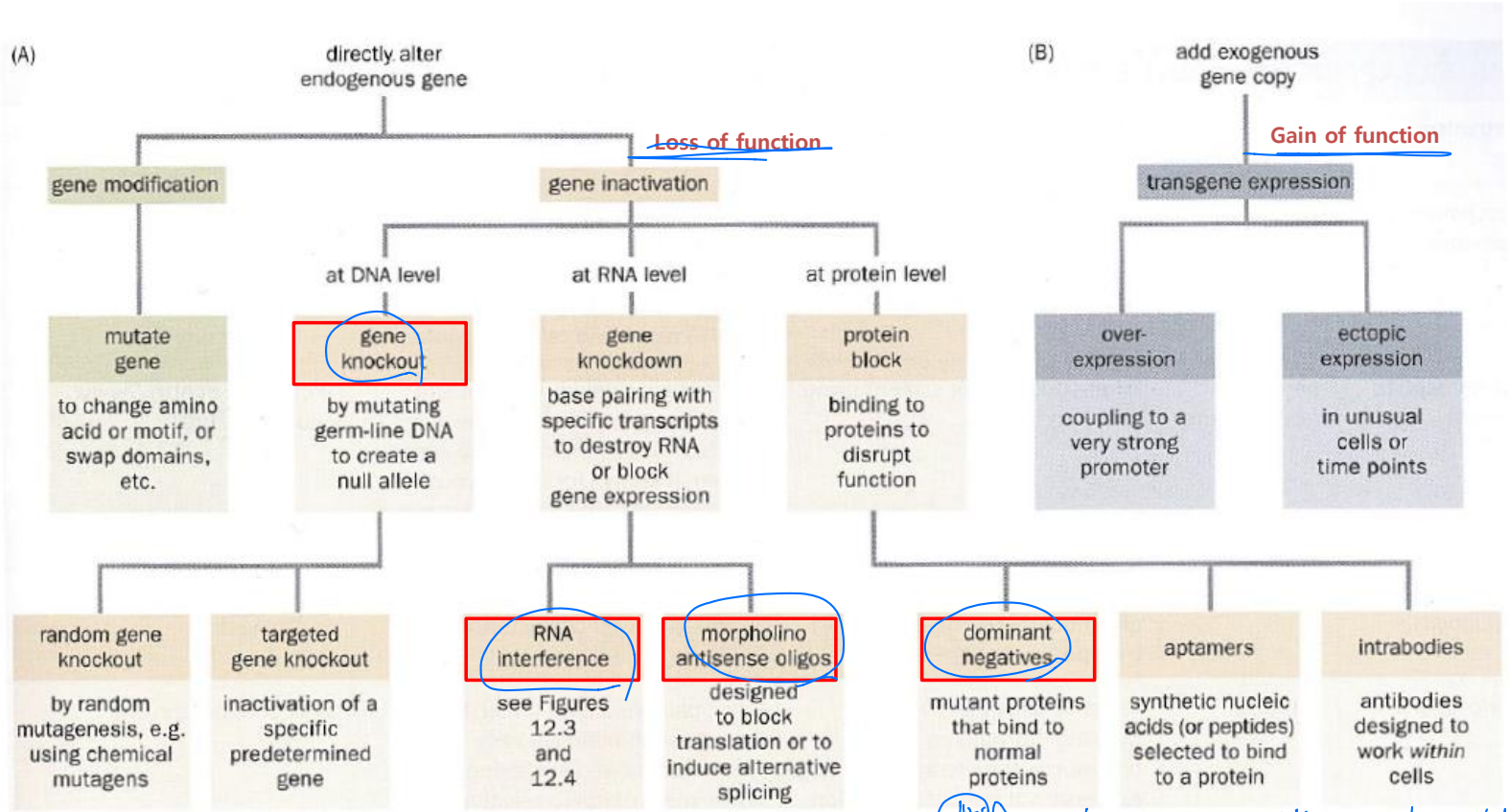


screening ⇒ find gene affecting

Phenotype

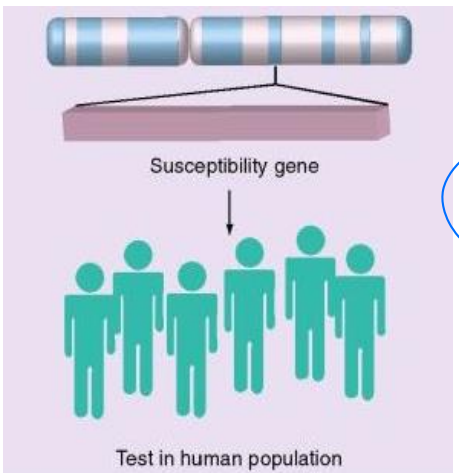
Selective gene inactivation and modification

- Study of gene function in cultured cells or using cell extracts has limitations
- Defining gene function in this wider context requires the genetic manipulation of model organisms



forward : phenotype 관찰 → random mutation
reverse : gene(screening) → phenotype

Functional genomics studies : Forward genetics & GWAS



Gamma ray, ENU

Saturation of forward genetics

Treating mutagen on germ line

Human disease

Animal model

Linkage analysis

Pathway analysis

The gene

X

F1

F2

LOD

Mapping, sequencing, function

Congenic strains

Relative oxygen burst

6

4

2

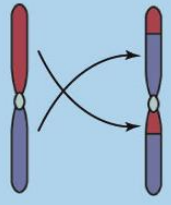
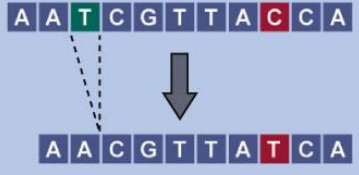
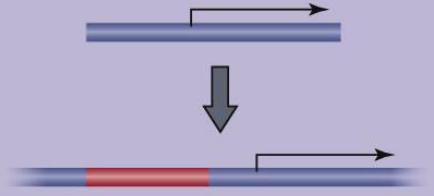
0

a/a

b/b

The three major types of mutagen

forward 에서 random mutation 쓰는 방법
* reverse 에서는 siRNA 이용

Gamma rays	Chemicals (e.g., ENU)	Insertions
		
▶ Strong mutations ▶ May disrupt multiple genes ▶ Laborious cloning	▶ Full spectrum of mutations ▶ Random distribution ▶ Mutation detection difficult	▶ Mutation is tagged ▶ Reversible ▶ Nonrandom distribution

Double-strand breakage of DNA

ENU, also known as N-ethyl-N-nitrosourea
- Alkylating agent
: transferring the ethyl group of ENU to nucleobases (usually thymine)

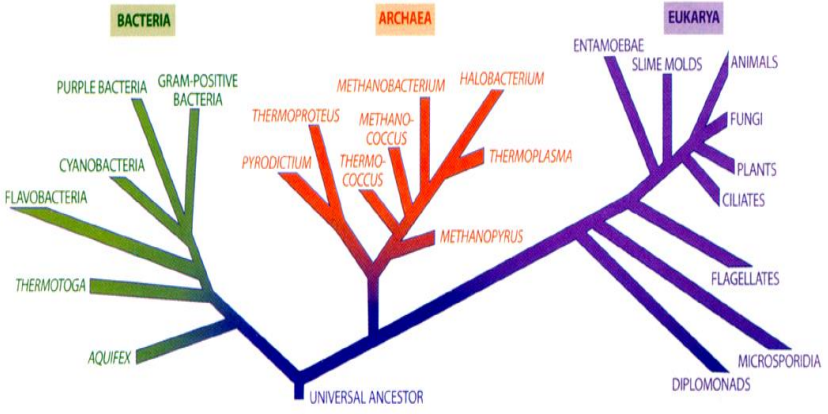
P element in drosophila (transposon)

10. Phylogenetics

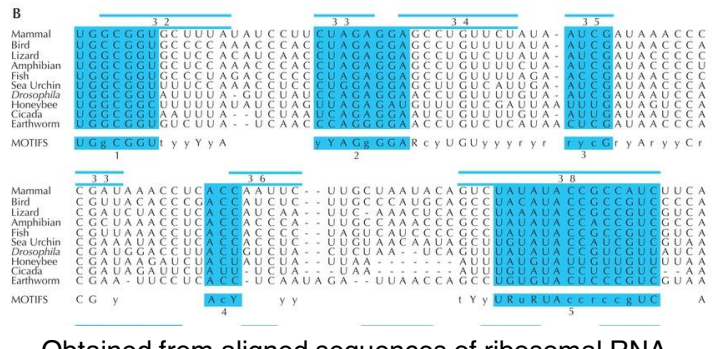
like sequence alignment

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

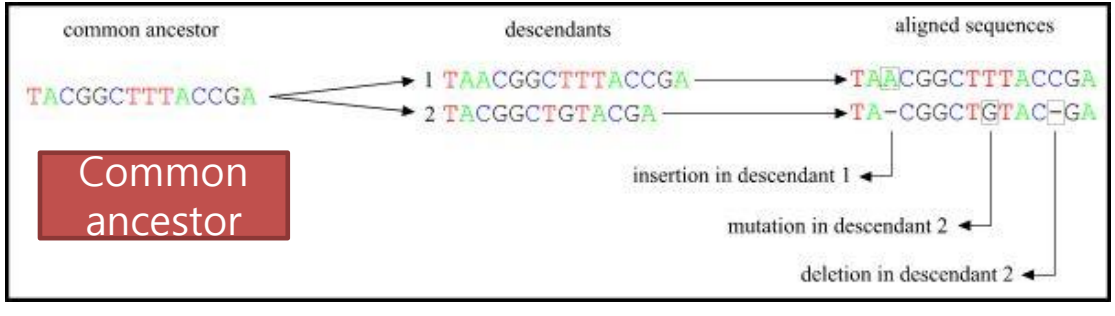
A version of the "tree of life" : Phylogenetic tree



M. Madigan and B. Marrs, 1997



Obtained from aligned sequences of ribosomal RNA



Perform pair-wise alignments

Measure distances

Generate tree

Interpretation of phylogenetic tree: Orthologs and paralogs

Variant sequences

↓

Sequence Alignment

↓

Phylogenetic tree

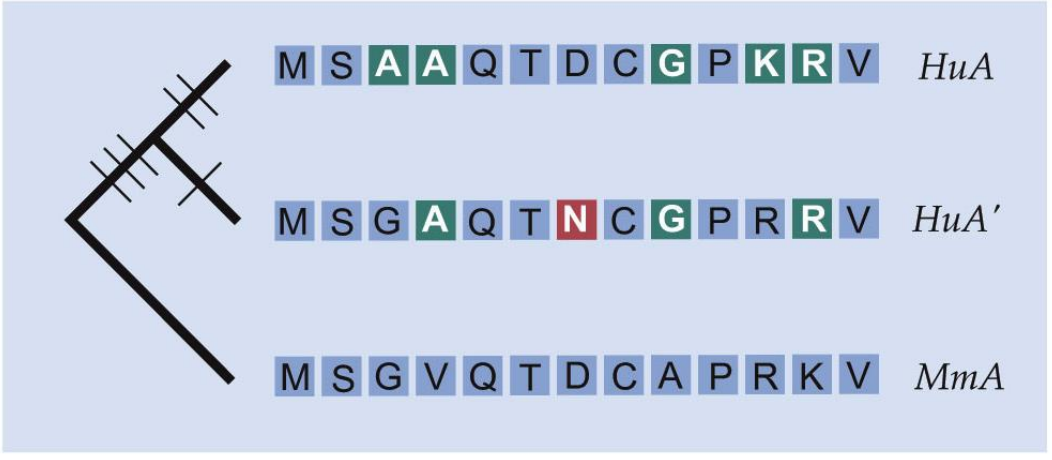
- Recalculate / Infer
- Distance or similarity b/w sequences

- Topology (order)
- Length (evolutional time)

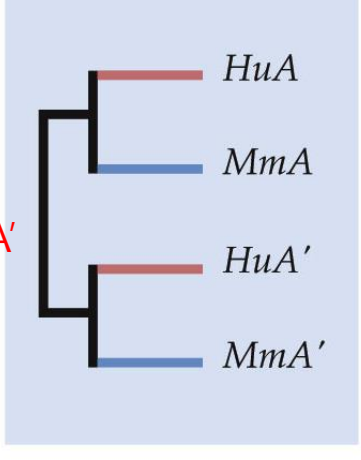
↓

Evolutional relationship / interpretation

(A)

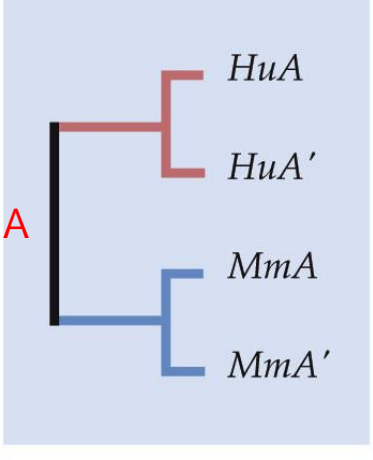


(B)



A, A′

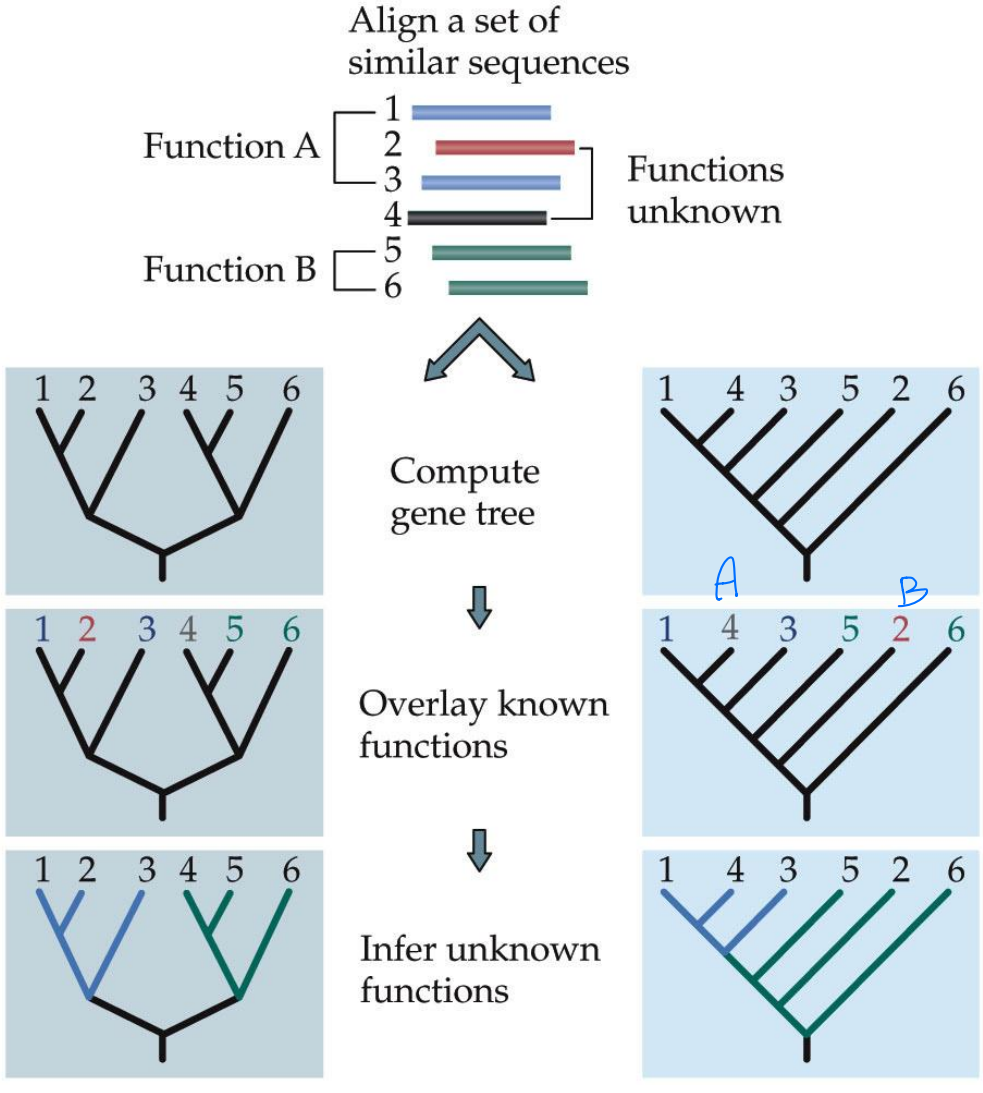
(C)



A

free shape
→ evolution event 에 상관없이 다른형태임

Inferring gene function from phylogenetic analysis



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:

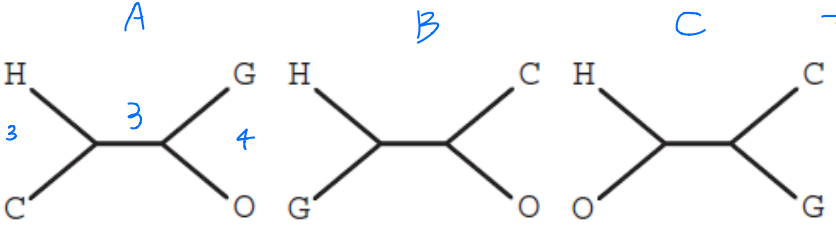
Human	`CGAAATGCAT`
Chimpanzee	`CGGACTGTAT`
Gorilla	`CAAGCTGTAC`
Orangutan	`TAAGACGTAG`

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

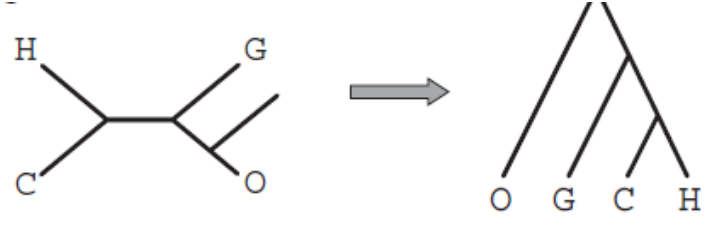
(1) all possible unrooted tree



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

(2) calculate edit distance

minimum = more likely to be happened



rooted treee considering outgroup

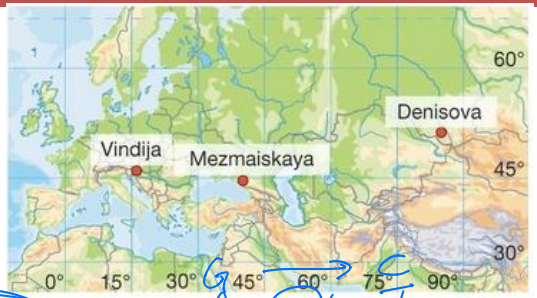
Ancient human genome (Phylogenetic tree)

ARTICLE

doi:10.1038/nature12886

The complete genome sequence of a Neanderthal from the Altai Mountains

2014, Nature

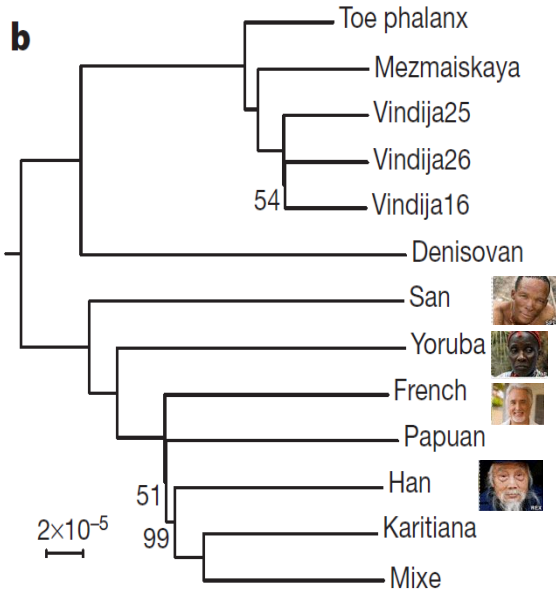


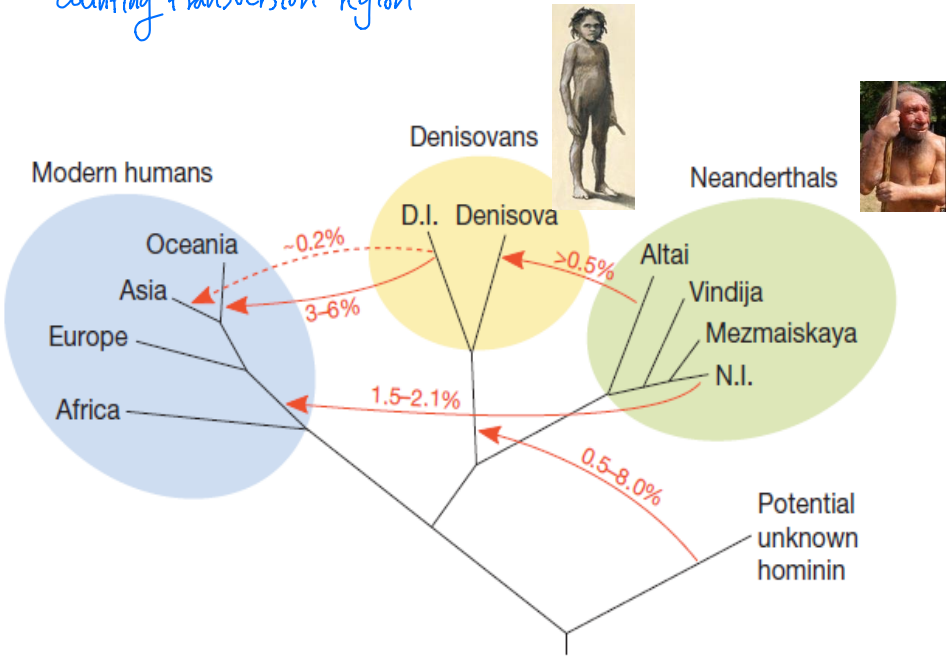
Found in 2010
Woman X

Covering larger region

A neighbour-joining tree (Fig. 2b) based on transversions, that is, purine–pyrimidine differences, among 7 present-day humans

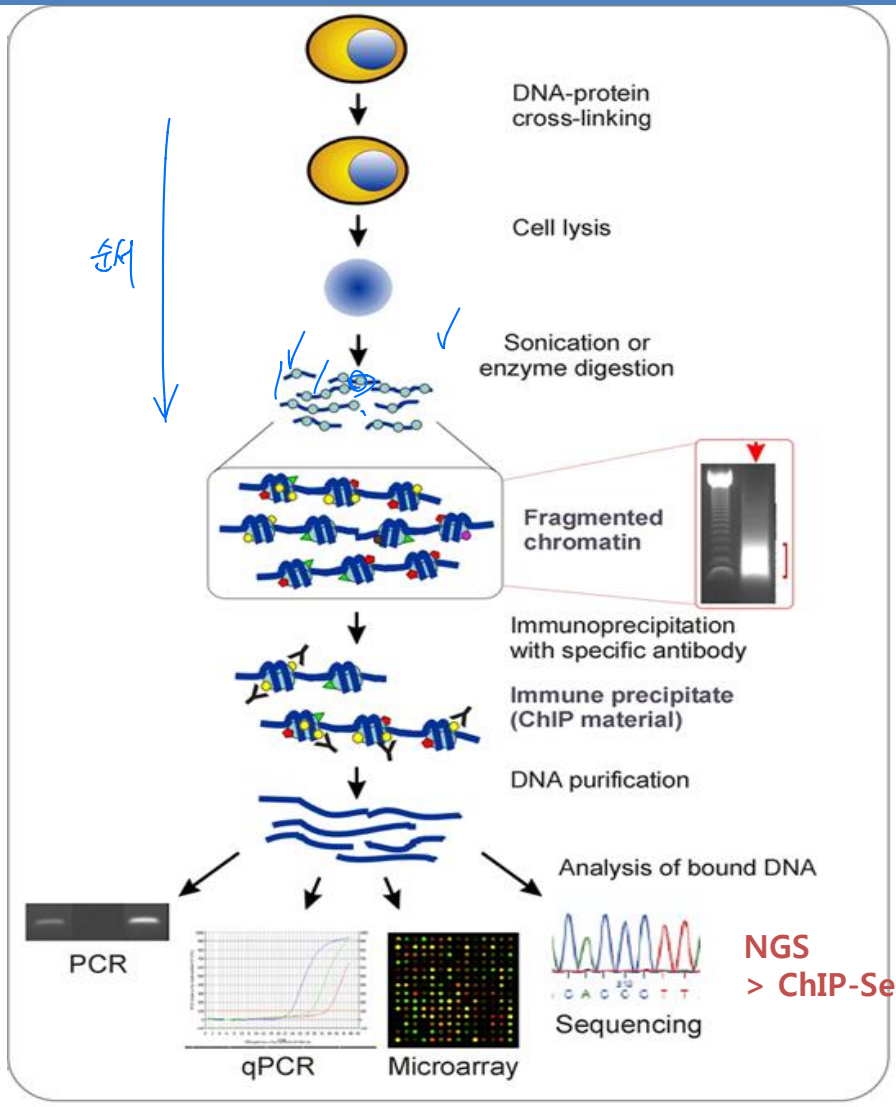
counting transversion region





https://www.youtube.com/watch?v=W7XKxT5xG_I

11. Chromatin immunoprecipitation (ChIP)



- Protein cross-linked to DNA *in vivo* by treating cells with formaldehyde

- Fragmentation
  Shear chromatin (sonication)

- Immuniprecipitation
  IP with specific antibody

- Reverse cross-links, purify DNA

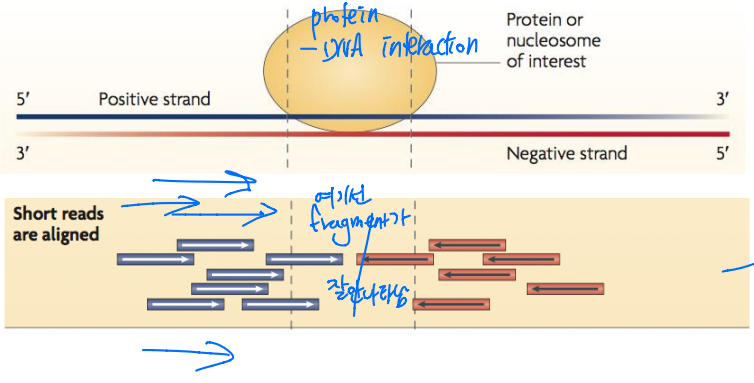
- PCR amplification

- Sequencing:
  Identify sequences

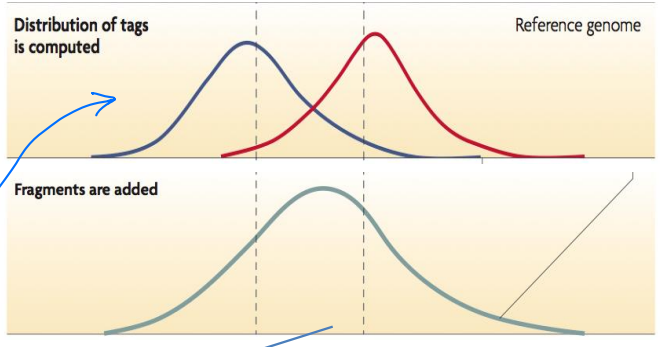
- Genome-wide association map

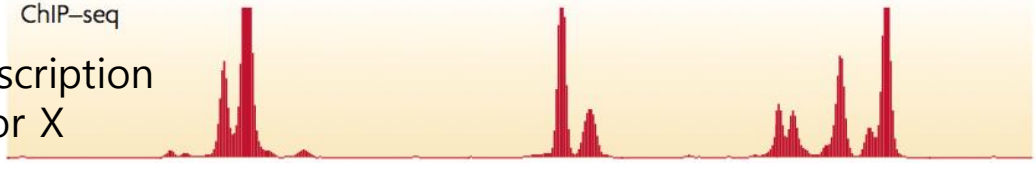
ChIP-Seq analysis

Mapping the reads from ChIP-Seq



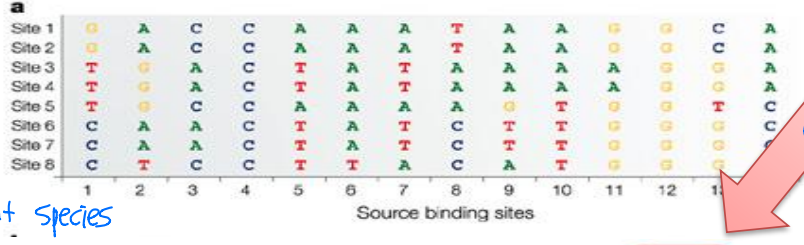
Peak analysis





Transcription factor X

figureout seq



Binding site information

across different species

Sequence conservation

Conservation score across multi-genomes

Combine 해서 연구 가능

GWAS, eQTL

Disease associate mutation

Mechanism : transcriptional dysregulation

Phylogenetic footprinting

evolutional footprinting

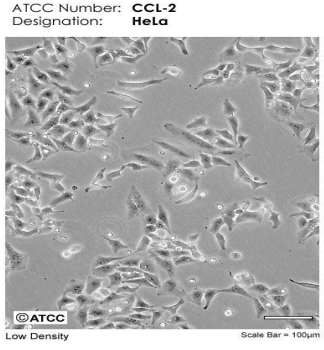
Practical exercise from the recent research

Genome sequence of Cancer

Henrietta Lacks
1945–1951

Cervical cancer

HPV infection



HeLa cell line

2013

WGS by NGS

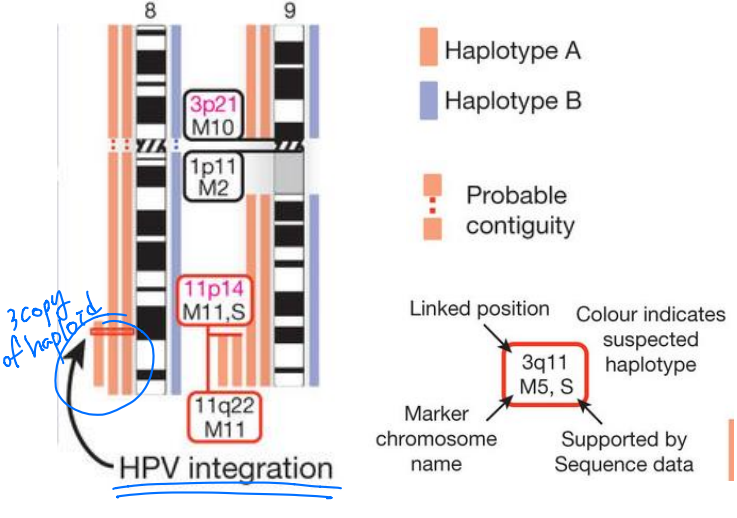
Structural variation

Haplotype maps (Tagged SNP)

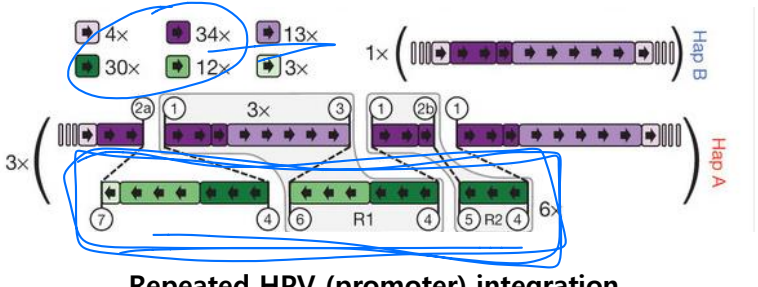
Nature Vol: 500, Pages:207–211, 2013

genomes are all messed up

HPV integration in only Haplotype A



HPV repeats



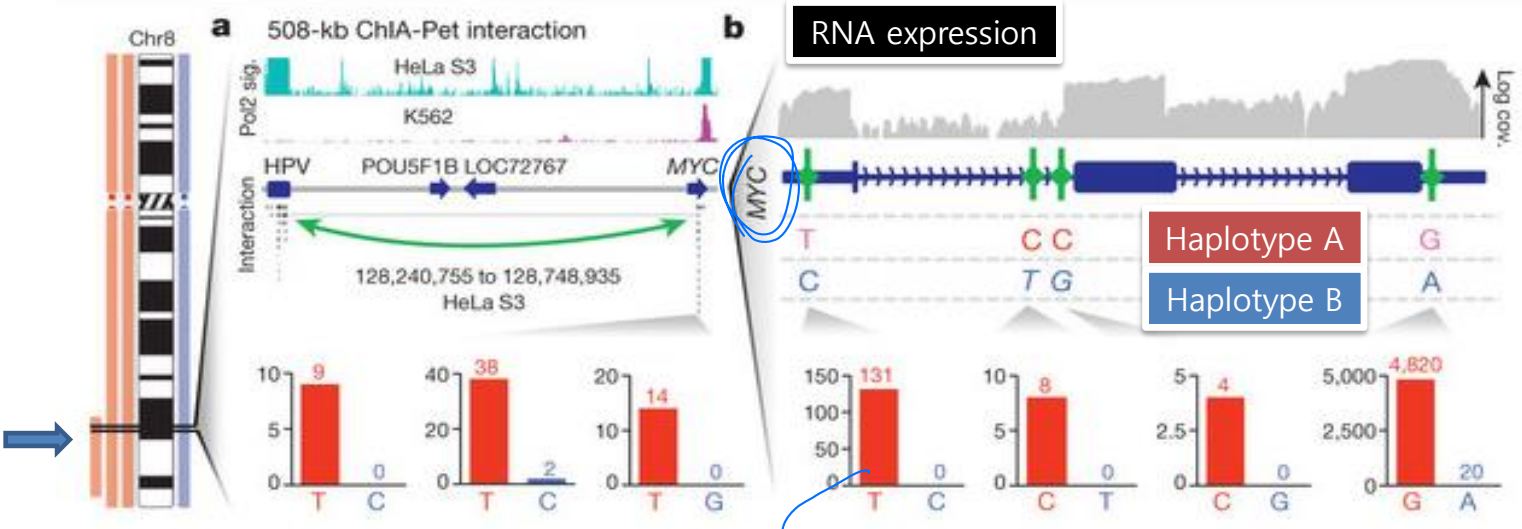
Repeated HPV (promoter) integration in heterochromatin region

Insertion > indirect interaction > alter gene expression

Haplotype : tagged SNP

Haplotype specific HPV – Myc interaction

Haplotype specific transcriptional activation of Myc by HPV integration



HPV integration
Copy number variation

RNA expression separately measured for each haplotype

A안 over expressed

수업에서 배운것을 모두 활용하세요

Functional Genomics Class : part I

~ 11 questions will be in the middle-term exam !!!

~ 6 questions will be multiple choice problems

1. Introduction to genomics
2. Mapping genome
3. Human genome project and sequencing
4. Next-generation sequencing
5. Sequence alignment, haplotyping (LD analysis)
6. Genomic variation
7. Whole genome sequencing (NGS)
8. Genome-wide association study
9. Exome-seq and functional genenomics
10. Phylogenetics
11. ChIP-Seq

Most of questions will low or middle level

Thinking of making 1 question at middle-high level