Summary & Review

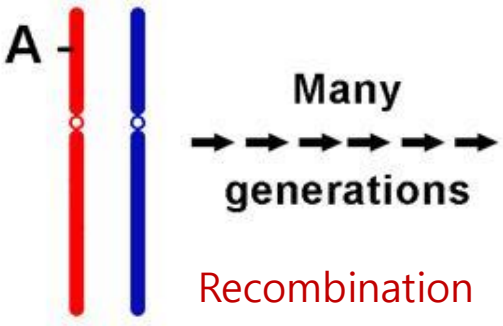
Basic knowledges for human genomic variation

- How to determine the haplotypes (tagged SNP) ?
- What is the method for haplotype maps (HapMap) ?
  : Linkage disequilibrium (LD)
- Why haplotype is important for genomic variation analysis?

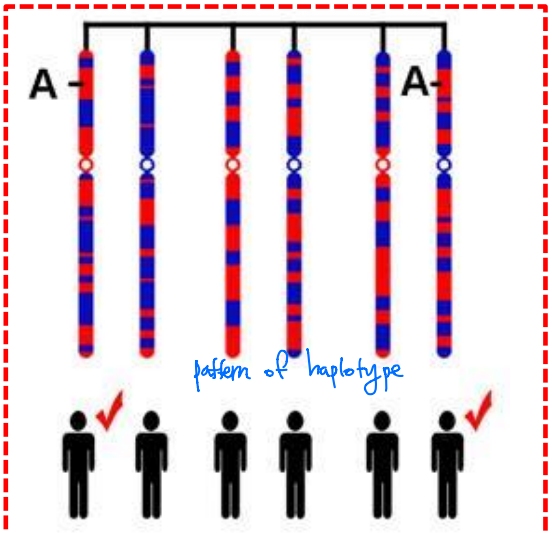
haplotype

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

염색체를 이루어 함께 유전되는 chromosome 단위

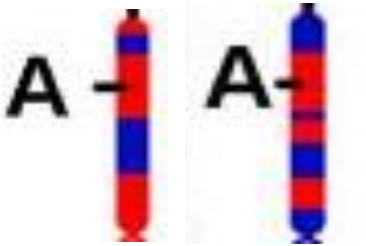


Human population



Limited set of haplotypes
(regional, environmental)

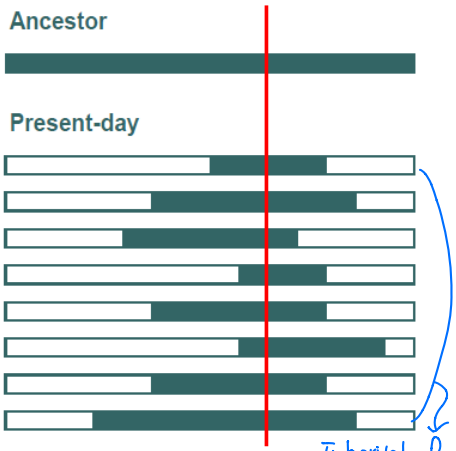
⇒ 이런 pattern이 random하지 않고 limited combination만 관찰됨



Only two haplotypes

> Easy to analyze

환경과 지역을 제한하면 haplotype 유사함

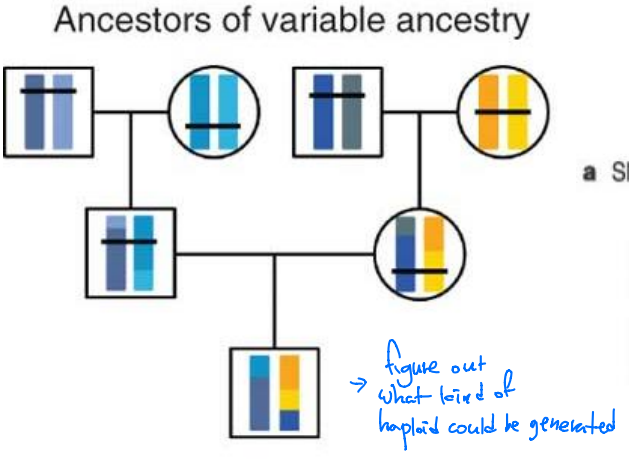


How to determine the same haplotype regions?

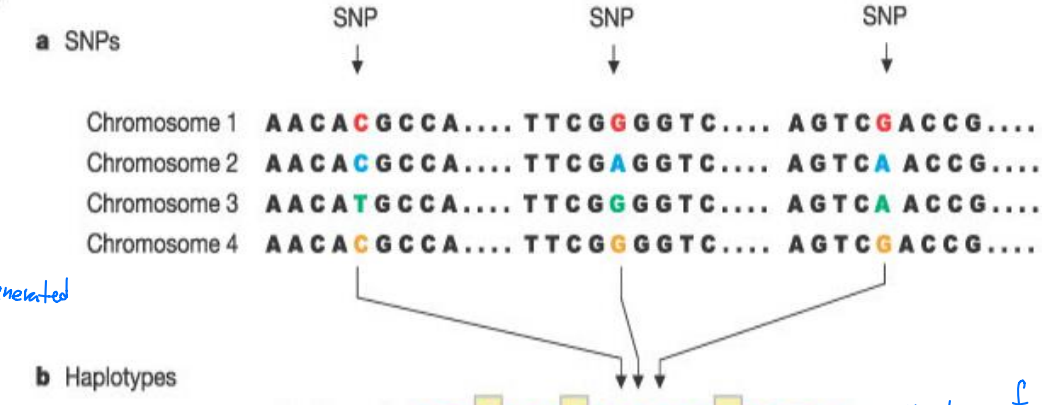
Variation (SNP: single nucleotide polymorphism)

SNP를 제외하고 나머지 region은 같음

Tagged SNP



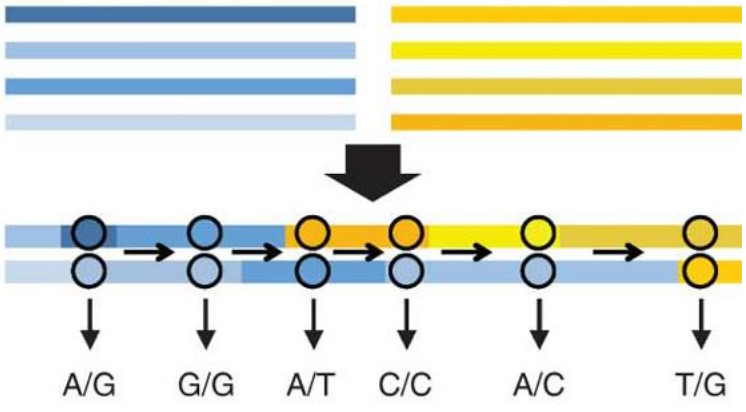
→ figure out what kind of haplotype could be generated



collection of SNP

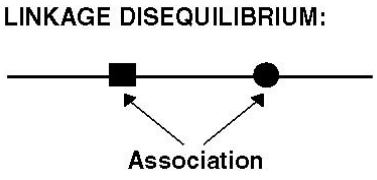
→ 여러 SNP 중에 what kind of haplotype 인지 알려줄 수 있는 SNP 지표

Tagged SNP



Only 4 haplotypes
: discriminate by 3 tagged SNPs

Linkage disequilibrium (LD)



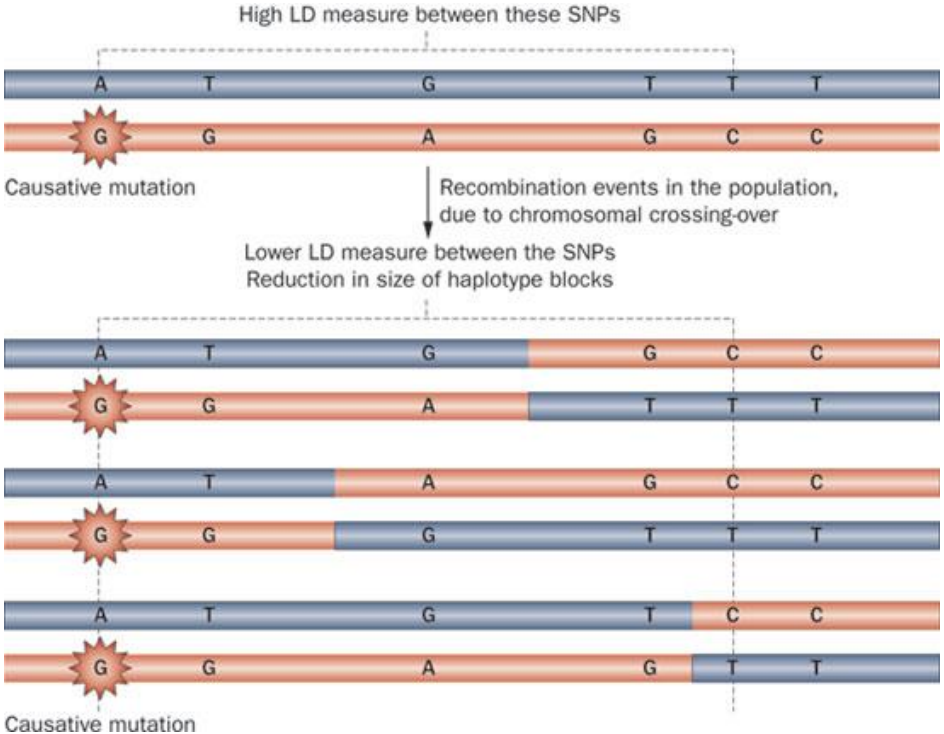
Sequencing results (population)

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

Two SNPs

(linked)

How much it becomes disequilibrium ?



Completely linked

실제로는 딱 두가지로 나뉘는 게 아니라 이 사이에 위치

Linked (located in same cosegregated region)

Completely independent

observed = expected

Do given sequence results contain enough variations to determine haplotype ?

diversity를 설정할 value가 필요 → SNP

(length=50 nt)

9 SNPs

10 sequence results (human)

each indicate one haplotype

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 ⇒ minor layer frequency

0.4 quite frequent

0.1 rare

0.1보다 0.4가 더 많다

Diversity 얼마나 diverse한가?

$$[\ (41 \times 0) + (3 \times 0.1) + (3 \times 0.4) + (2 \times 0.2) + 0.5\ ] / 50 = \mathbf{0.048}$$

0 : 똑같

과거의 염색체(diploid)가 서로다른 정도를 가지고 있을 확률이 얼마나 되는지 계산

actual length of sequence

0.048 x 50 = 2.4

random하게 잡았을때 다를 확률

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

0.9X0.1이 되는곳 있으니까 (가능한 짝 x diploid)

$$[\ (41 \times 0) + (3 \times 0.1 \times 0.9 \times 2) + (3 \times 0.4 \times 0.6 \times 2) + (2 \times 0.2 \times 0.8 \times 2) + 0.5 \times 0.5 \times 2\ ] / 50 = \mathbf{0.0624}$$

0.0624 x 50 = 3.12 ⇒ 이 population에서 과거의 haplotype 골랐을때 최소 2개 location은 heterozygote 일것이라 예상

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 frequency

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

≠

빨간색, 파란색이 independent 하다고 생각

expected

확률적으로 계산한 것

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

parameter of diverse

D : linkage disequilibrium coefficient

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

if two loci are in linkage equilibrium ⇒ D=0

if two loci are in linkage disequili- ⇒ D≠0

probability이기 때문에 0이하가 될수 X

	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

원래 table과 같도록

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

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

Completely linked
(D=Dmax) (D'=1)

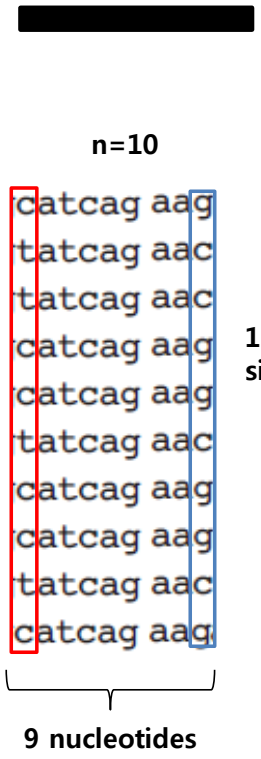
min{x_{21}, x_{12}}

Linked

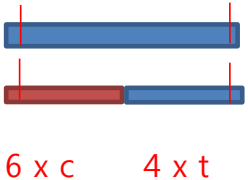
Completely independent
(D=0) (D'=0)

observed = expected

Linkage Disequilibrium (LD) analysis > haplotyping

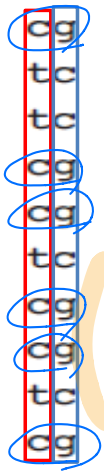


1. Segregation sites ?



6 x c 4 x t

SNP1



SNP2

5 x g 5 x c

2. Minor allele frequencies?

SNP1 = 0.4

SNP2 = 0.5

3.Expected average heterozygosity (H)?

Diversity?

(0.4 x1 + 0.5 x1 + 0 x7) / 9 = 0.1

H= (0.4 x 0.6 x 2 + 0.5 x 0.5 x 2 + 0 x7) / 9 = 0.06

4. Level of LD in SNP?

SNP1 (c) Vs. SNP2 (g)

- Observed freq

Pcg =6/10 Pcc =0

Ptg =0 Ptc =4/10

- Expected freq

Pc =6/10 Pt =4/10

Pg =6/10 Pc =4/10

Pc x Pg =0.6 x 0.6 = 0.36 +D

Pc x Pc =0.6 x 0.4 = 0.24 -D

Pt x Pg =0.4 x 0.6 = 0.24 -D

Pt x Pc =0.4 x 0.4 = 0.16 +D

{0.36, 0.24}

observed - expected의 차이

6/10 = 0.36 + D D = 0.24

D_{max} = min (0.24, 0.24) = 0.24

D' = 0.24/0.24 =1

Complete LD !!!

한쪽 ~ 다른 region이

```
 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
```

5 13 26 independent 36

frequency가 낮으면 information이 별로 없음. 귀 때문이다 라고 단정짓기 어려움 → frequency 높은 SNP 채택 to analyze haplotype

같이

반반 → 여기서부터 뭔가 recombination이 일어났을 것

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

(5t, 13c) → completely linked

(13c, 26g) → 반반

(26c, 36a) → independent

13	26
g	g
c	g
c	a
g	a
g	a
c	g
g	a
g	a
c	g
g	g

observed

	g	a
g	0,2	0,4
c	0,3	0,1

expected

	g	a
g	0,30	0,30
c	0,20	0,20

D>0 ⇒ min {0,3, 0,2}

g = 0,6 g = 0,5

c = 0,4 a = 0,5

Finding D

	g	a
g	0,2 = 0,3 - D	0,4 = 0,3 + D
c	0,3 = 0,2 + D	0,1 = 0,2 - D

$D_{max} = \min\{0,1, 0,1\} = 0,1$

D = 0,1 ⇒ (...?

---

t	26
c	g
t	g
t	a
c	a
c	a
t	g
c	a
c	a
t	g
c	g

	g	a
t	0,3	0,1
c	0,2	0,4

	0,2	0,2
	0,3	0,3

D = 0,1

$D_{max} = 0,1$?

t = 0,4 g = 0,5

c = 0,6 a = 0,5

5t	36a
c	a
t	a
t	a
c	c
c	c
t	c
c	c
c	a
t	c
c	c

observed

	a	c
c	0.2	0.4
t	0.2	0.2

c = 0.6 a = 0.4

t = 0.4 c = 0.6

Haplotype Map (HapMap)

African

ancestor가 다른지 대체로 haplotype도 다름

European Ancestry

Asian

(2008)

(2005)

Genome Variation Analysis (review and more)

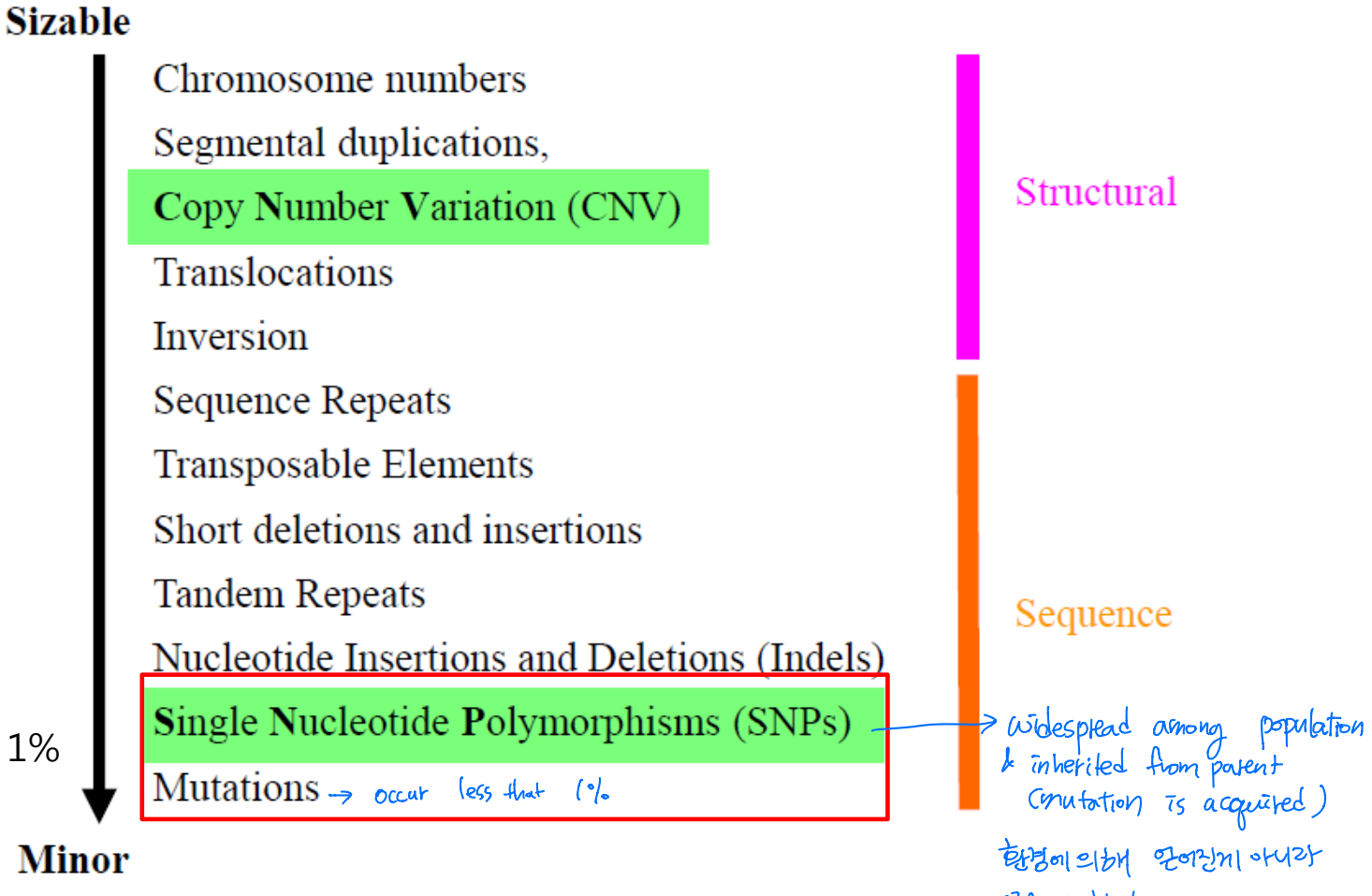
Introduction to genomic variation

focus on cancer thing

Sung Wook Chi

Division of Life Sciences, Korea University

Genetic Variations



Oncogenes vs. Tumor Suppressor genes

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

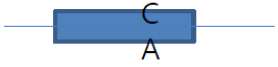
tumor : want to amplify

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

want to loss or stop

Corrupted Genomes in Cancer Cells

Amplification

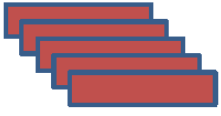


Point mutation

↝ loss function

Translocation

different chromosome 이 있다라는

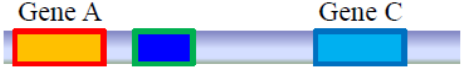


Increased signal

Abnormal signal

Different Types of CNVs

copy number variation

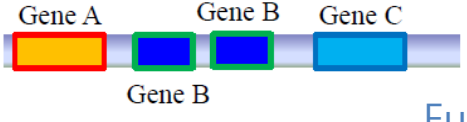


Deletion of gene (Loss)

Deletion (loss)

ex. tumor suppressor

Duplication of gene (Gain)



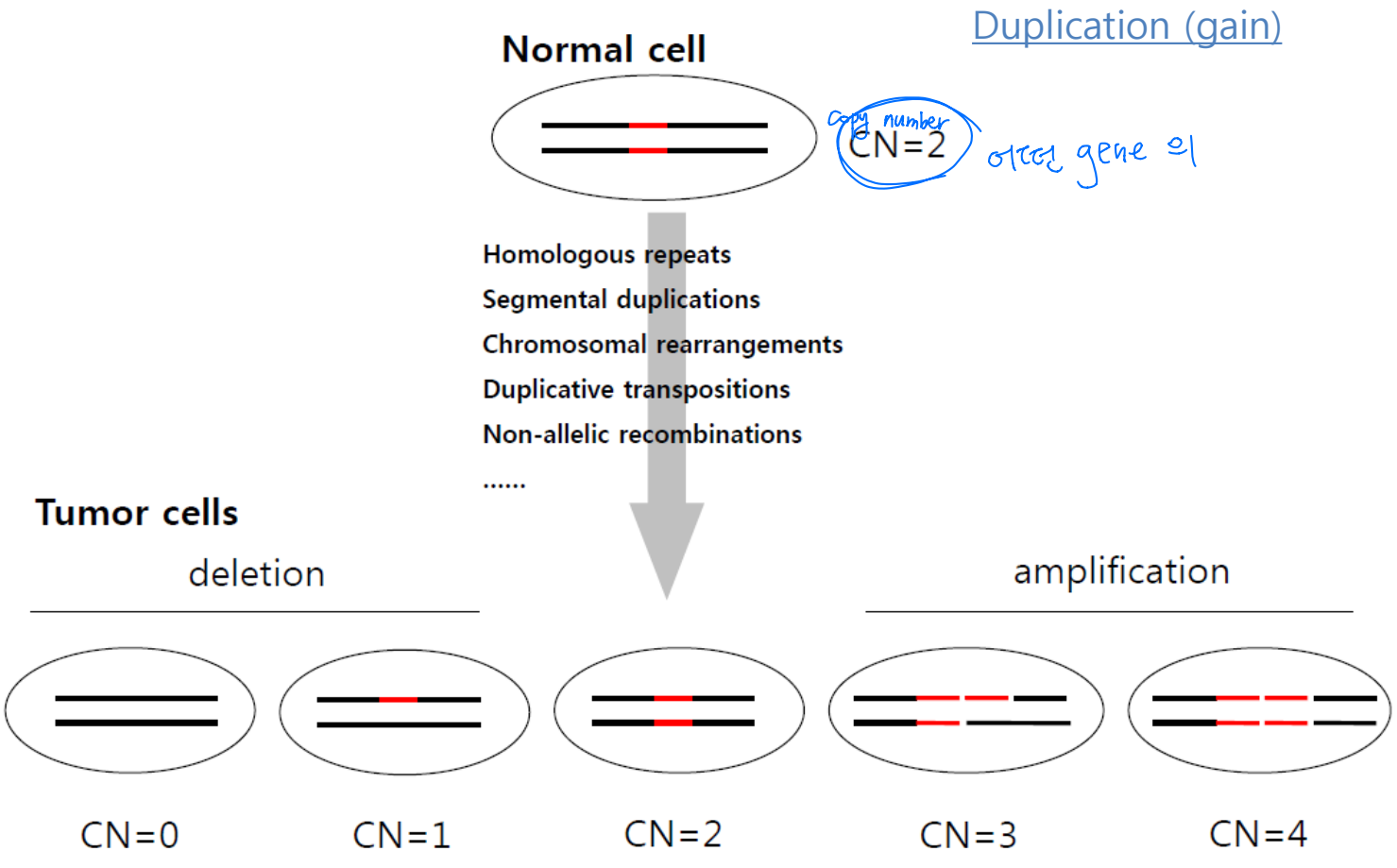
Duplication (gain)

ex oncogene

Deletion generated new fused gene

Fusion (deletion, new function)

Genetic Alterations in Tumor (Copy number changes)



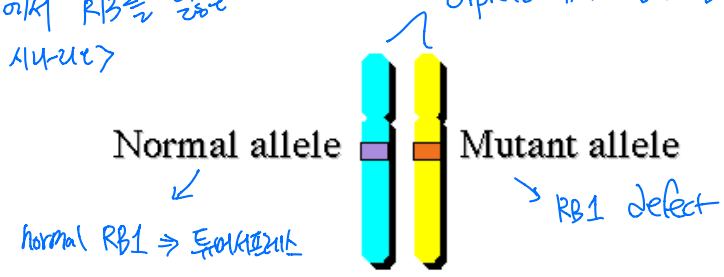
Loss of heterozygosity (LOH)

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 (well known tumor suppress gene)
- By screening paired blood and tumor samples with markers spaced across the genome, we may discover the locations of tumor suppressor genes

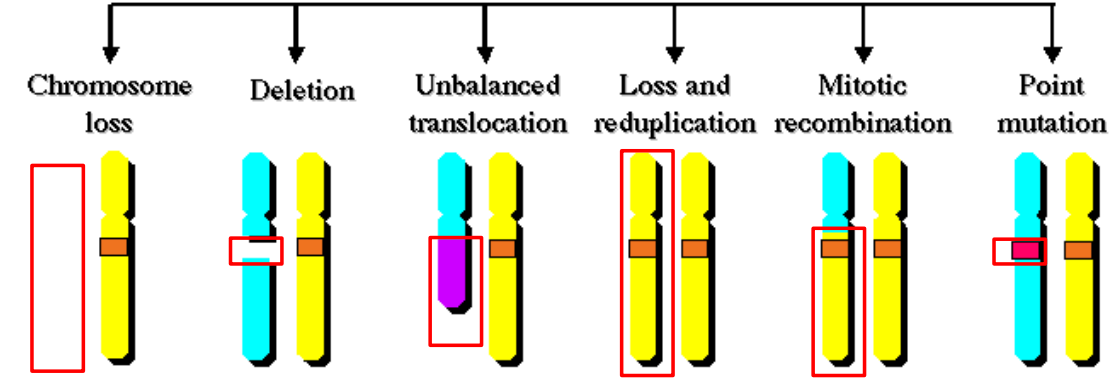
<Tumor 에서 RB를 잃는 여러 시나리오>

diploid 에서 양이나 변화하지 않았음

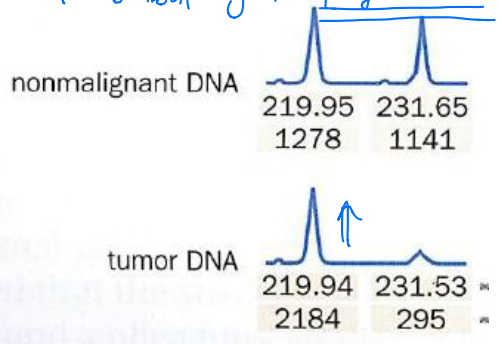


normal RB1 ⇒ 특이적인지

RB1 defect



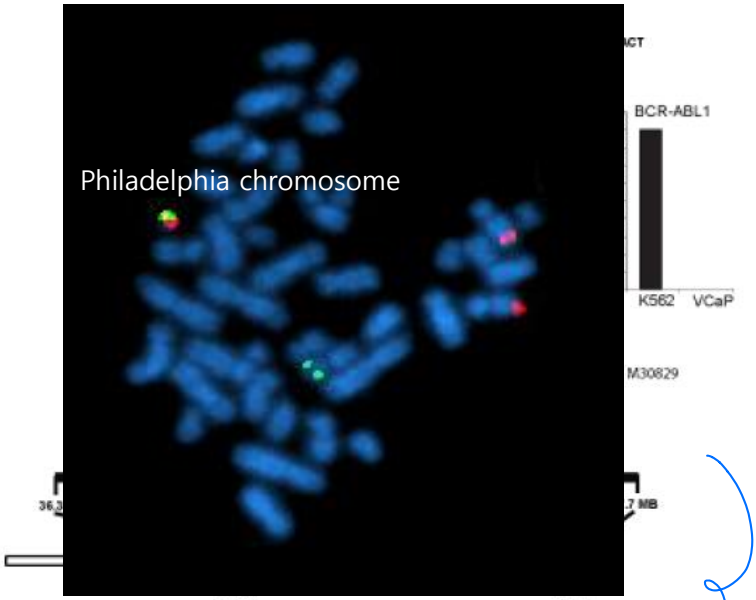
PCR 등 이용해 signal amplify



normal allele 기능 있음

Sequencing 이전에 특정 region의 변화를 비교하여 tumor 관련 gene 들의 위치적 특징 찾는다

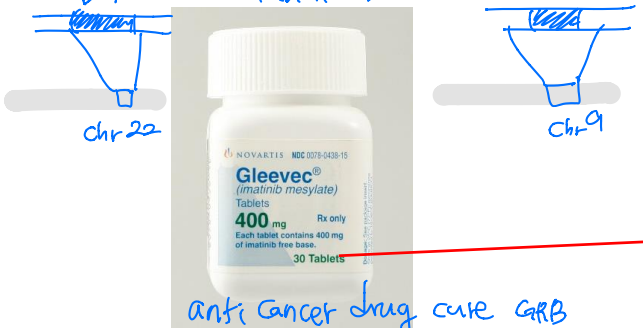
Fusion Gene in Cancer



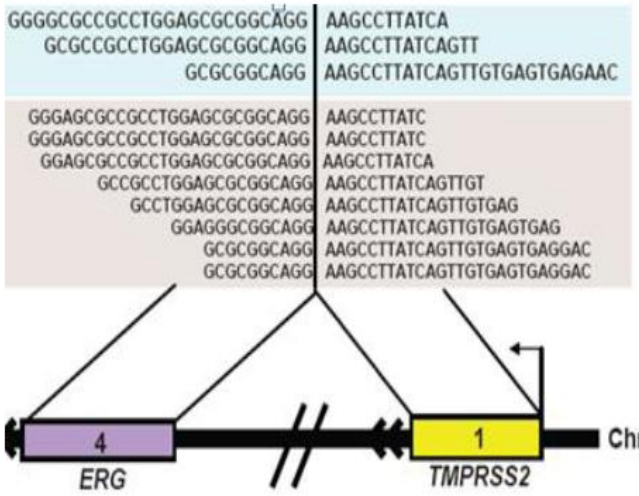
BCR-ABL1 caused by translocation
Chronic myeloid leukemia

Translocation

BCR → fusion ↷ ! ABL1

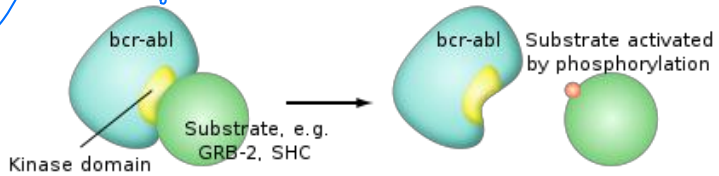


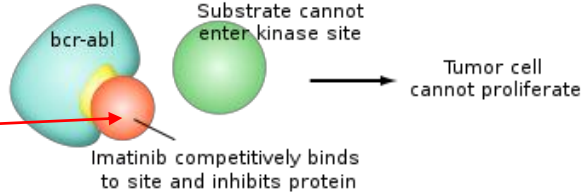
anti cancer drug cure GRB



TMPRSS2-ERG caused by deletion
Prostate cancer

working as unusual

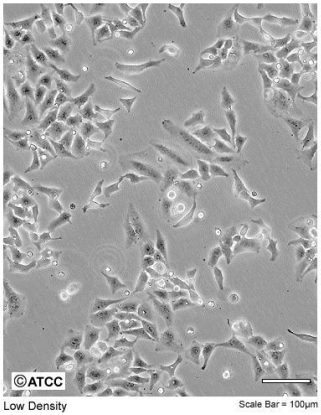


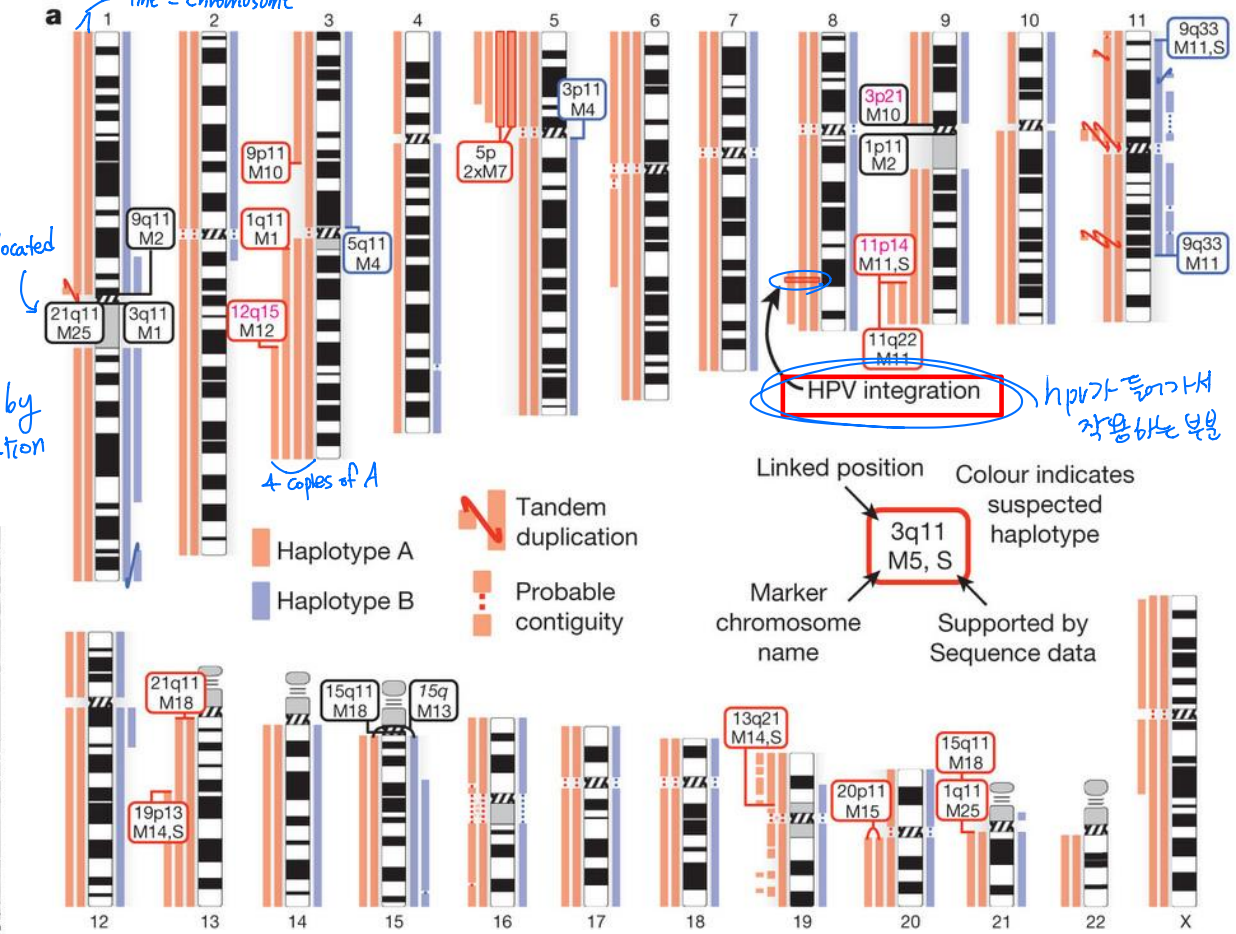


Analysis of Cancer Genome (example of HeLa cell)

Henrietta Lacks
1945–1951

ATCC Number: CCL-2
Designation: HeLa





NGS로 cervical cancer chromosome 분석 normal과 비교

line = chromosome

translocated

messed up by translocation CNV

4 copies of A

hpv가 들어가서 작용하는 부분

HPV integration > Cervical cancer
자궁경부암

Nature Vol: 500, Pages:207–211, 2013

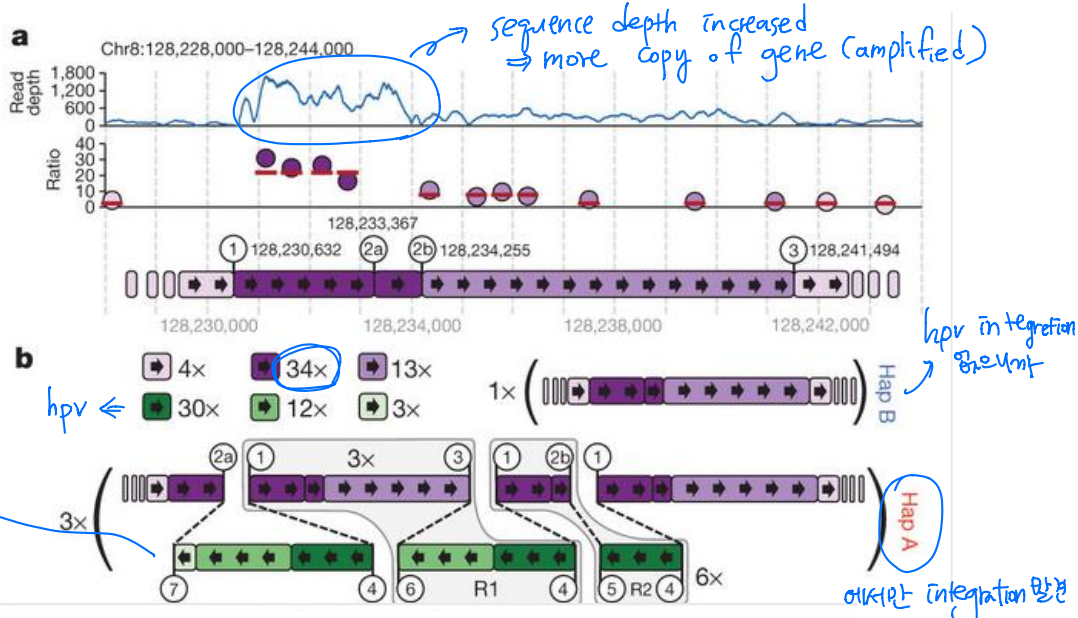
Analysis of haplotype, HPV CNVs, effect on Myc oncogene expression

HPV integration

virus는 강력한 promoter 가지고있음
여러 copy가 integrated
= 강력한 promoter 여러copy
= 근처 gene도 expression↑

> Myc oncogene overexpression

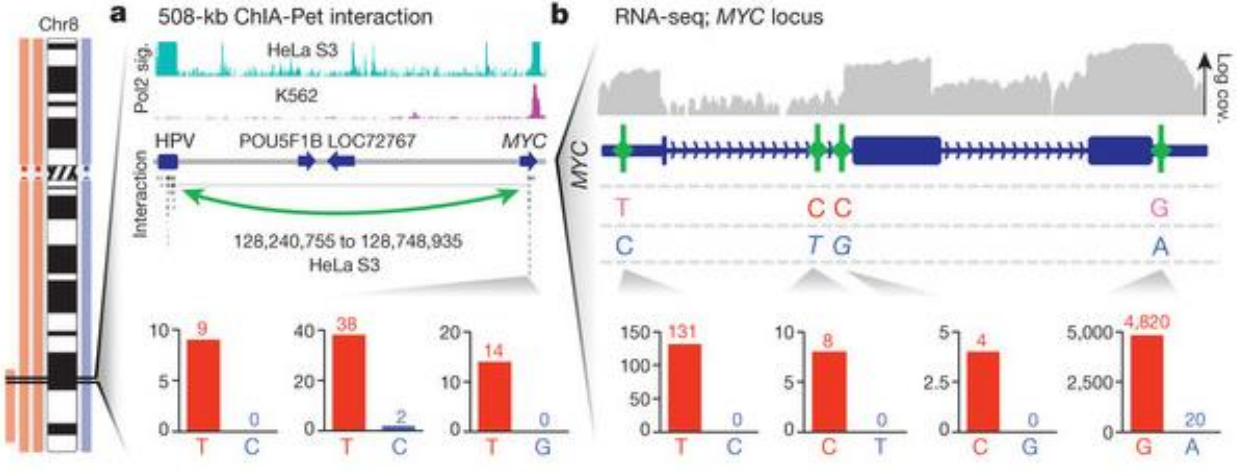
> **Cervical cancer** (HeLa)



sequence depth increased
⇒ more copy of gene (amplified)

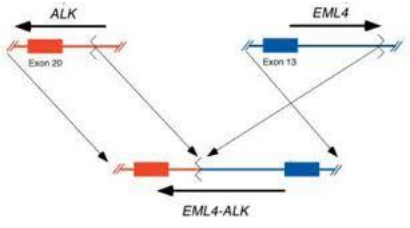
hpv integration 확인가능

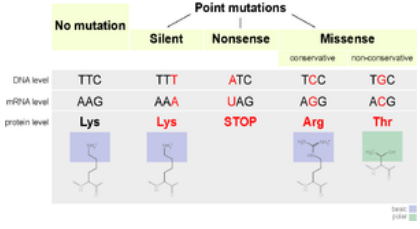
여러번 integration 발생

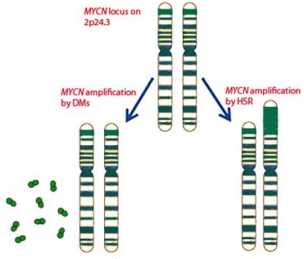


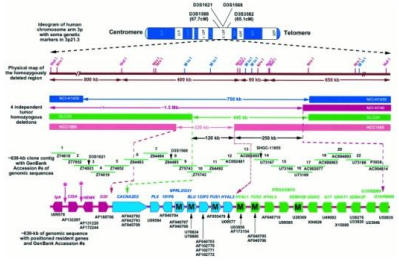
hap A에서만 많이 발현됨
(이 말이 아닌듯함)

Application of NGS for translational genomics









Translocation

BCR-ABL
EML4-ALK
RET/PTC

Point mutation

TP53
KRAS
BRAF

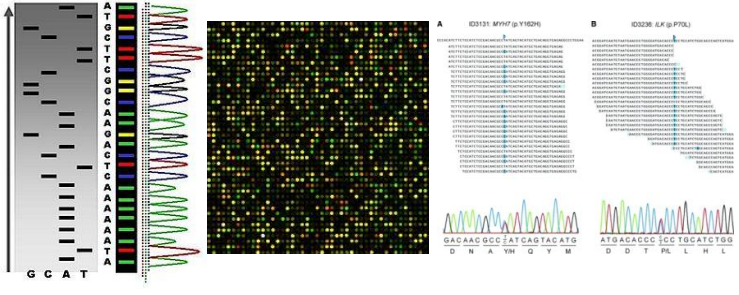
Amplification

EGFR
Myc

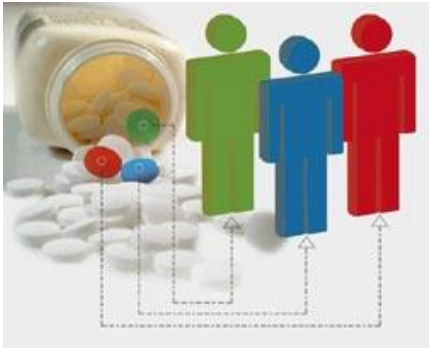
Deletion

Rb
TP53
PTEN

Clinical sequencing



Personalized cancer care



Genomics and Medicine : 100,000 Genome Project in England

need lots of database

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.**

100000명으로 유전병과 질병 나라을

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

still working on

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

Whole genome sequencing (NGS)

for analyzing sequence variations

Sung Wook Chi

Division of Life Sciences, Korea University

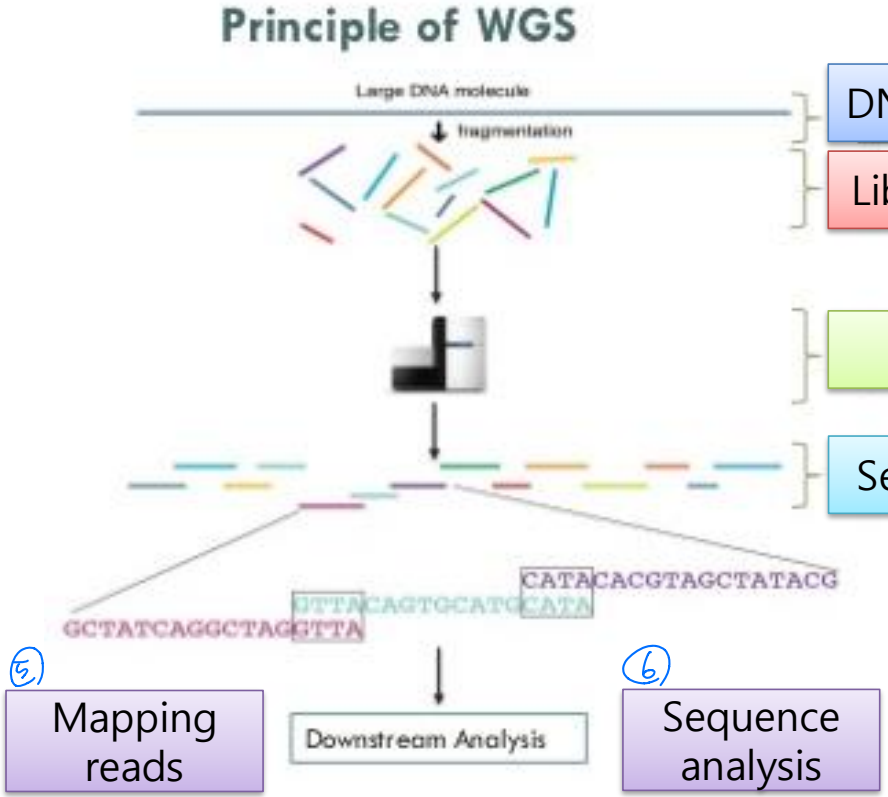
What we will learn today

Whole genome sequencing (NGS)
for analyzing sequence variations

1. Whole genome sequencing (WGS by NGS)

2. Variation Analysis for NGS data how we can analyze

Whole Genome Sequencing (NGS) for variant identification



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

Genomic variation analysis by NGS

→ increase price

Preprocessing

QC
: Quality check

Reads Mapping

Analysis

Library preparation

Sample preparation
- DNA
- RNA
- Small RNA

Library preparation
- Target enrichment
- Sample barcoding

Sequence generation

Primary Data analysis
- Image processing
- Signal processing

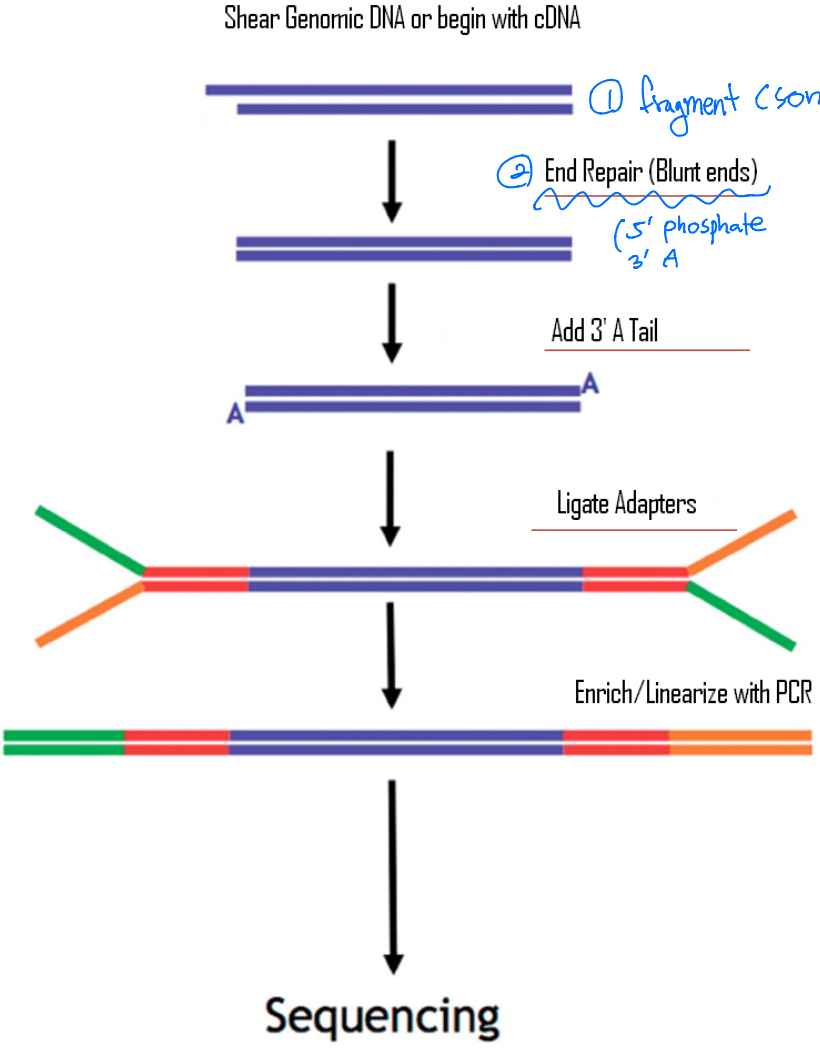
Raw sequence data QC

Secondary data analysis
- Assembly
- Mapping

Application specific analysis
- SNP calling
- Structural variation
- Read counting

NGS data analysis for variant identification

NGS Library preparation



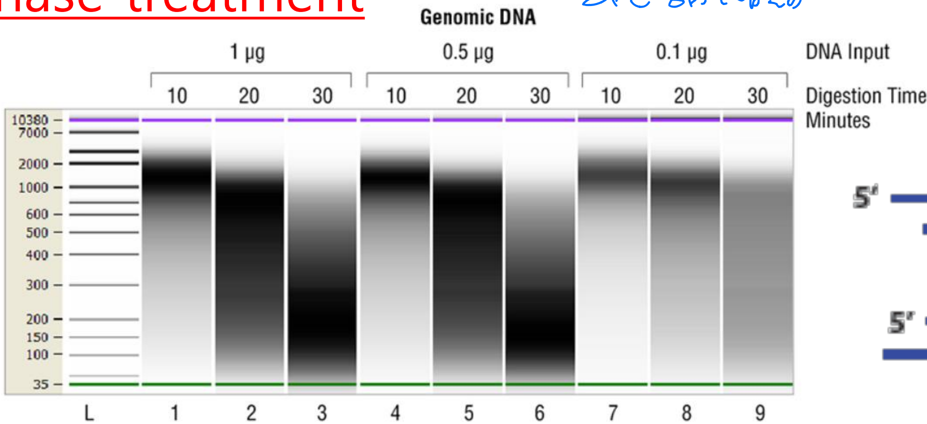
1. DNA fragmentation / Target Selection
   PCR amplification for specific region of interest
2. Adapter ligation
3. Size selection
4. Library quantification (QC)

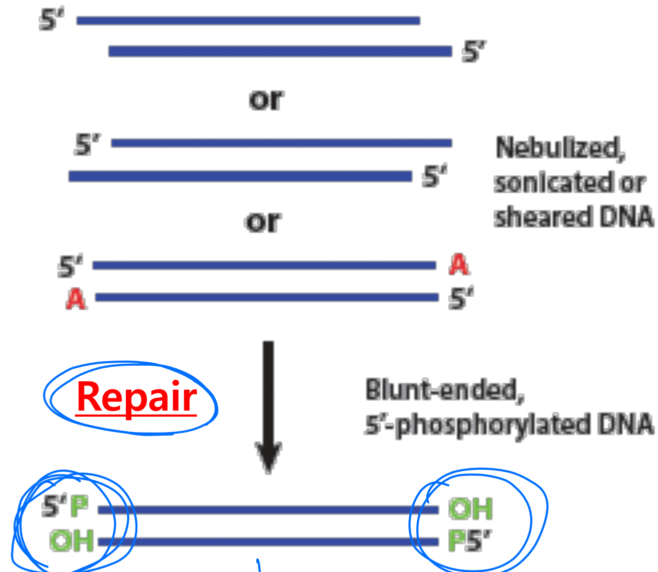
DNA fragmentation

DNA fragmentation for library construction

Dnase treatment

→ euchromatin의 절편이 더 좋아서 잘잘림 → random process
자른 차이 있음





↓ adaptor ligation

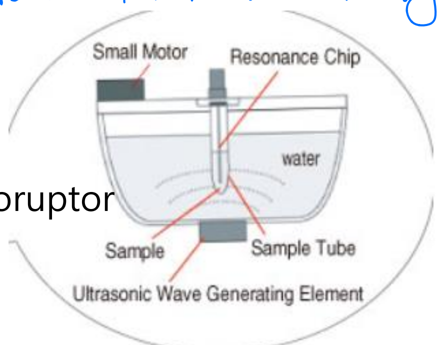
Double strand : Polymerase (Klenow fragment)

5′P : PNK (T4 Polynucleotide Kinase)
→ attach phosphate to 5′

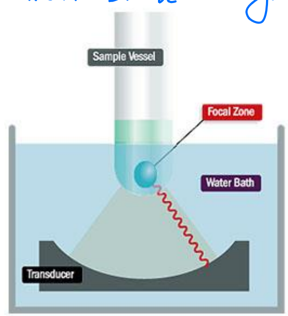
초음파

Sonication (shearing)

prefered ⇒ more random fragment than Dnase enzyme

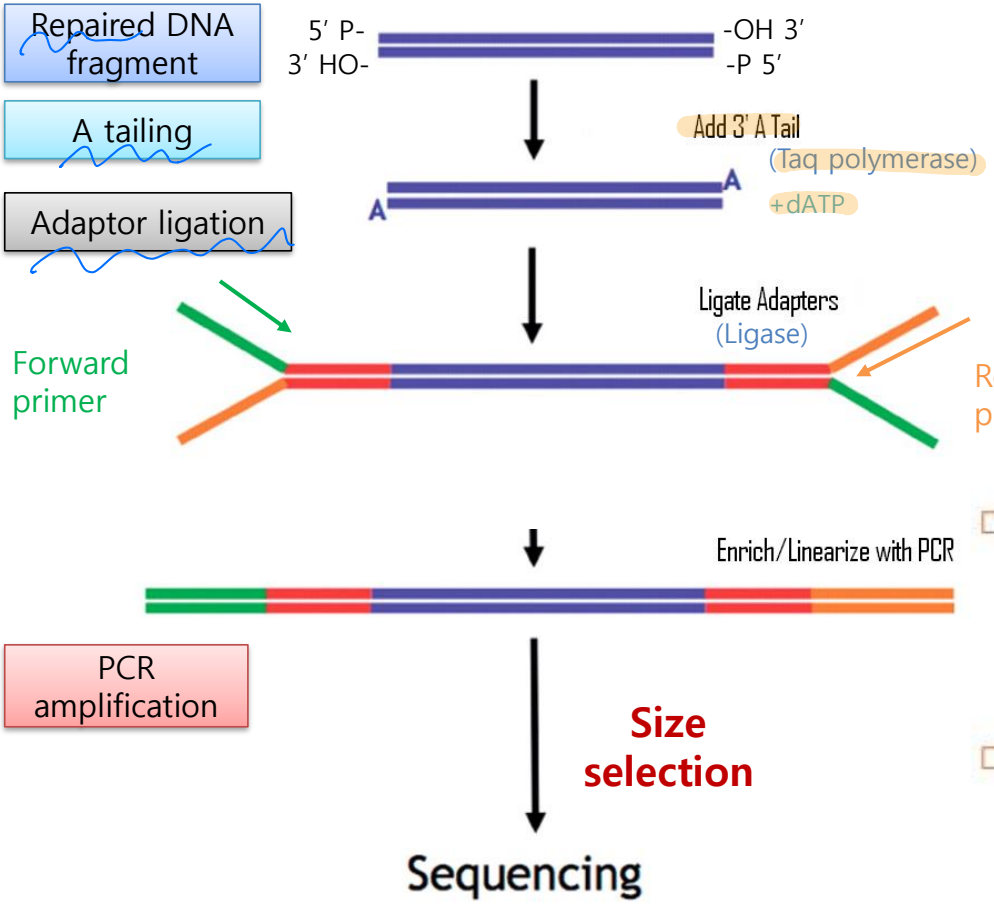


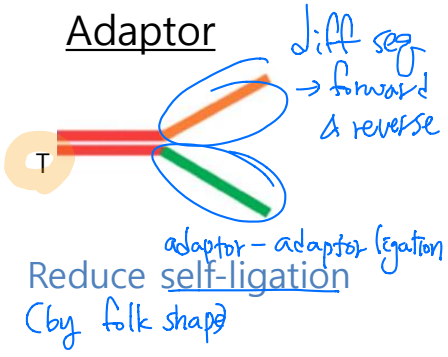
Bioruptor



Covaris

A-tailing, ligation, PCR amplification : Library preparation





- **Adapter ligation**
  - T-overhangs
  - Forked structure controls orientation
- **Library amplification**
  - Few cycles
  - Enrich for correctly-adapted fragments

NGS library preparation & Illumina Sequencing

Size selection

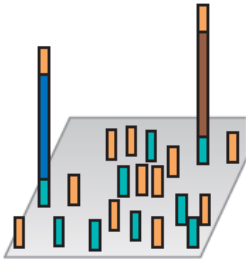
- Gel electrophoresis
- Bead based method

중요 과정!

library quantification

glass plate에 spread 하기 적합한 concentration 과정!

Illumina NGS sequencing



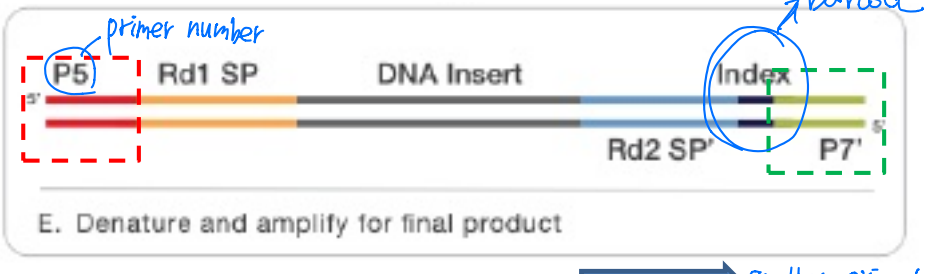
P5 P7

SBS (Sequencing by synthesis)

prime that initiate polymerase reaction

Sequencing primer (SE)

Sequencing primer (PE)



another primer

Sequencing primer for barcode reading (Only to read barcode sequence)

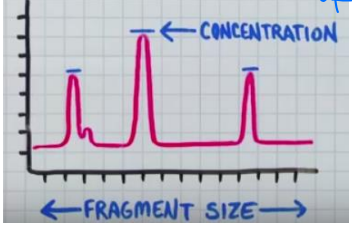
DNA quantification

: Accurate concentration (Mole number)

1) Spectrophotometer

- Absorbance : amount of DNA

2) Bioanalyzer ⇒ measure concentration of library as accurate as possible



3) qPCR

https://www.youtube.com/watch?v=_yC0Bzw3WbQ

Sequencing data from NGS : FASTQ format & Q score

FASTQ

fast는 '>' 사용

Label

Sequence — base calling result

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

base calling

\+ (next line)

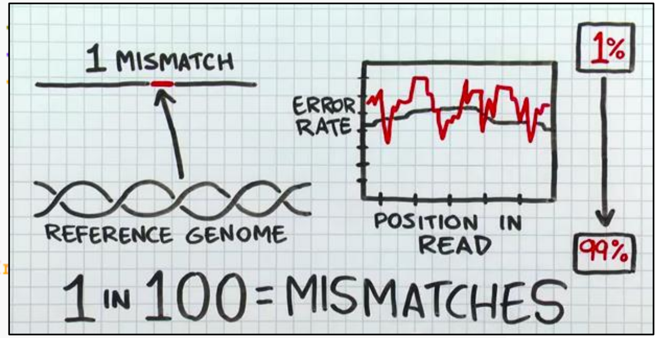
Q scores (as ASCII chars)

PHRED Q score : Q score ↑ = ↓ error

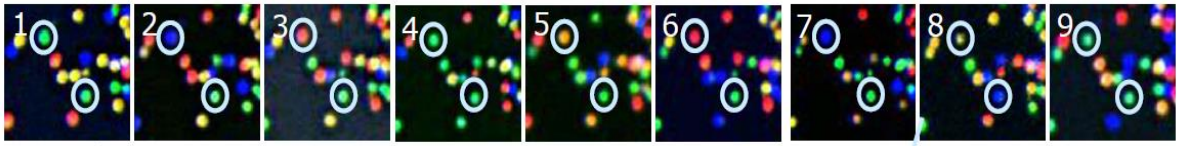
$Q=-10\log_{10}(P)$

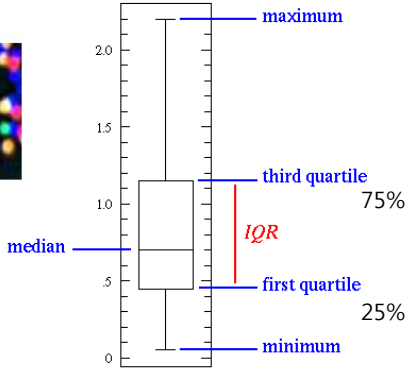
```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
.................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL...............................................
!"#$%&'()*+,-./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 I
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```



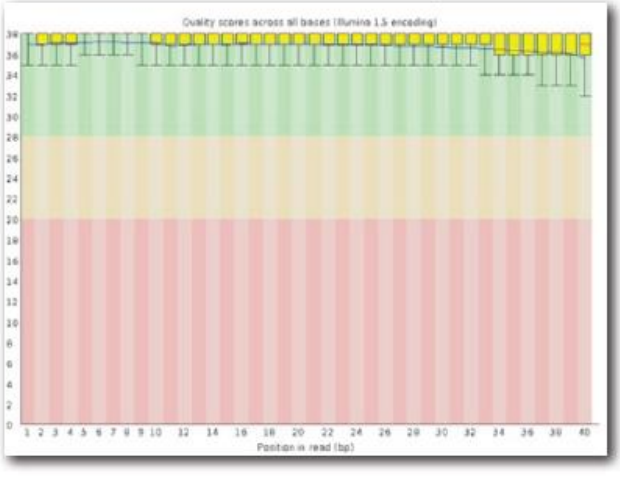
Sequence Quality Control





per base sequence quality

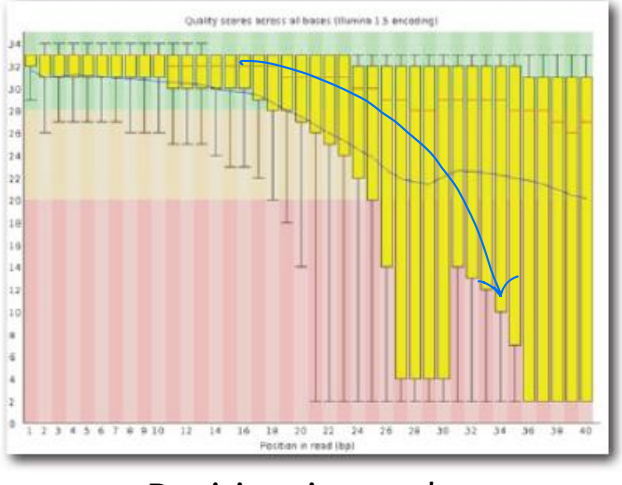
good



Q score

Position in reads

bad



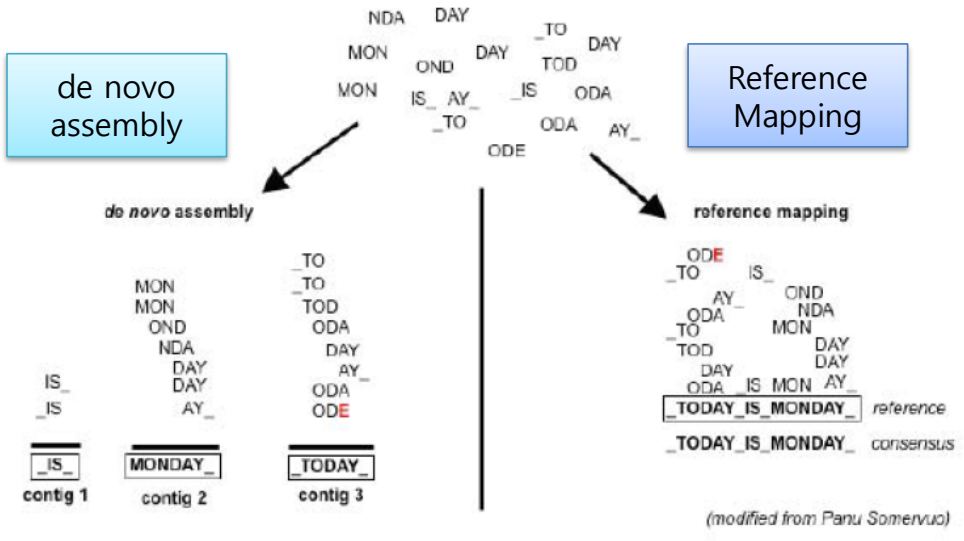
Q score

Position in reads

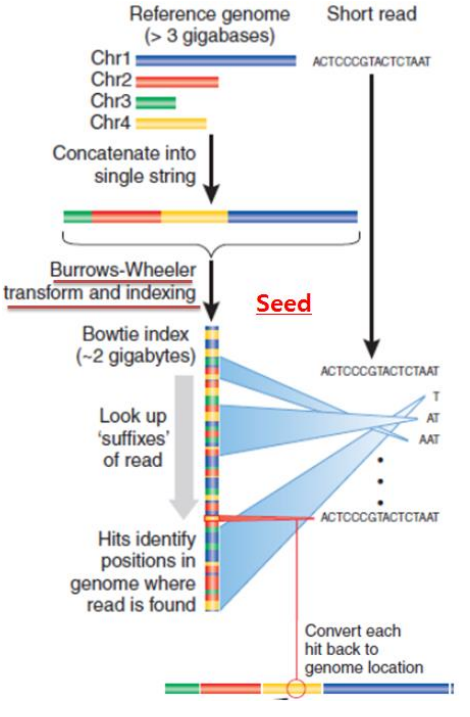
illumina chemistry 때문에

Q score tends to decrease depending on the increment of position

NGS Sequence mapping



Mapping billions of short reads on genome

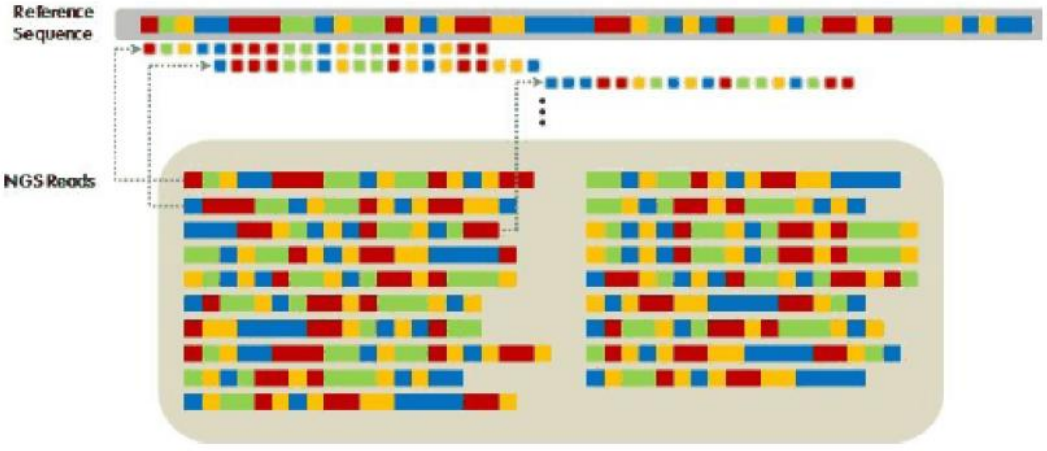


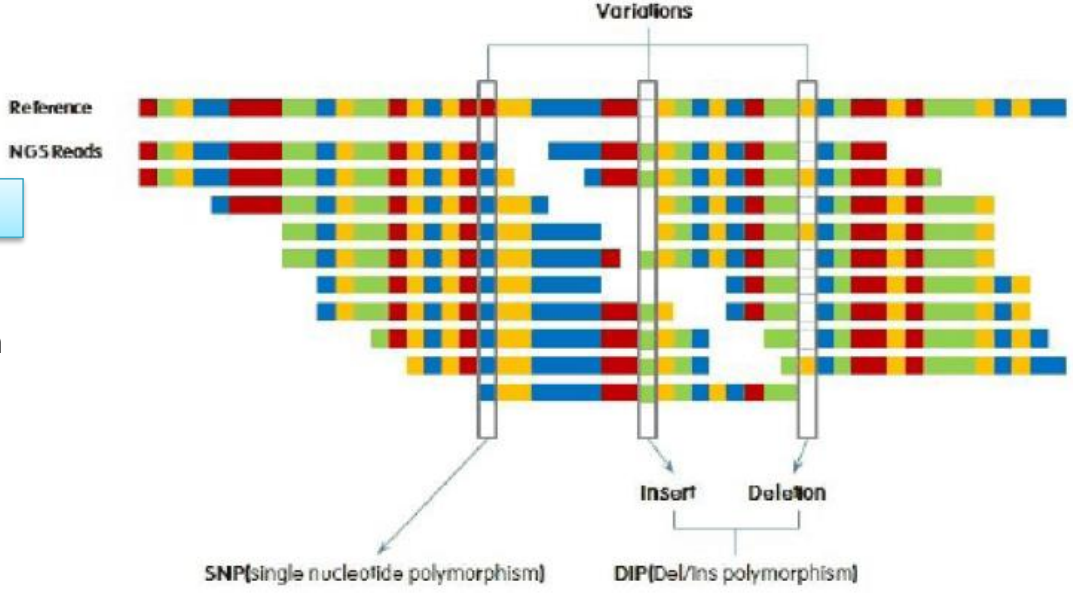
NGS WGS reads mapping (BWT based method)

- Transform genome sequences into data structure (condensed, indexed, fast search)
- Find billions of short reads sequence there (near perfect match)
- Location of matched reads (Genomic coordination, Alignment)

BWT (Burrow-Wheeler transform) based mapping algorithm

Sequence Alignment (Mapping)





Location

Genomic coordination

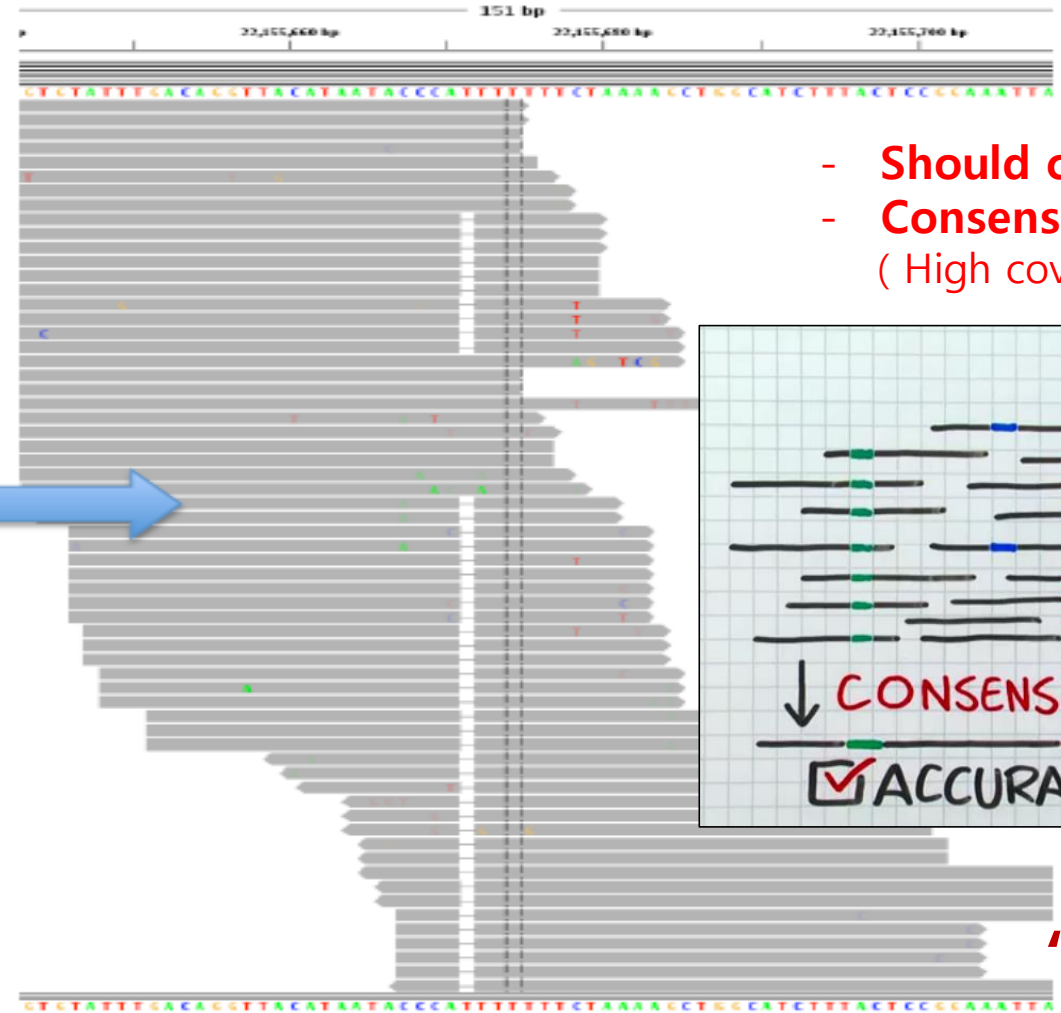
Alignment

Match
Mismatch
Deletion
Insertion

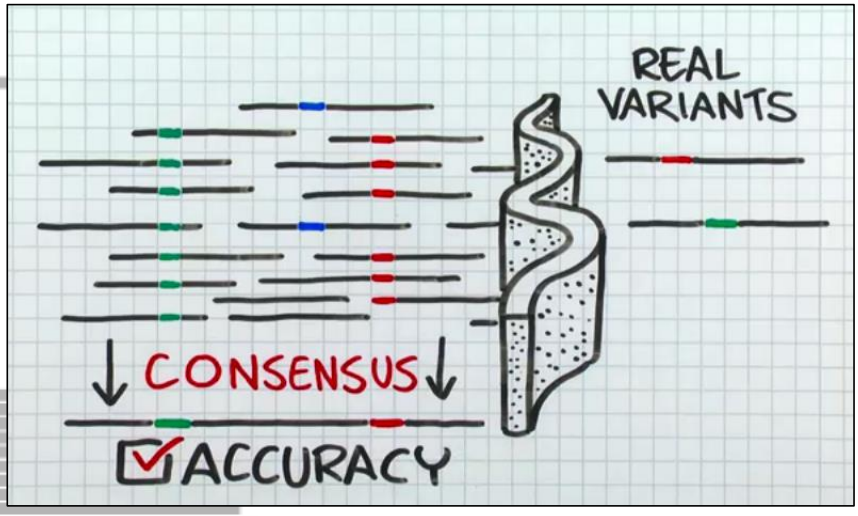
[BAM format]

Variation

Variation in NGS reads : Real or not ?



- **Should consider Q-score**
- **Consensus**
  (High coverage > Statistics)



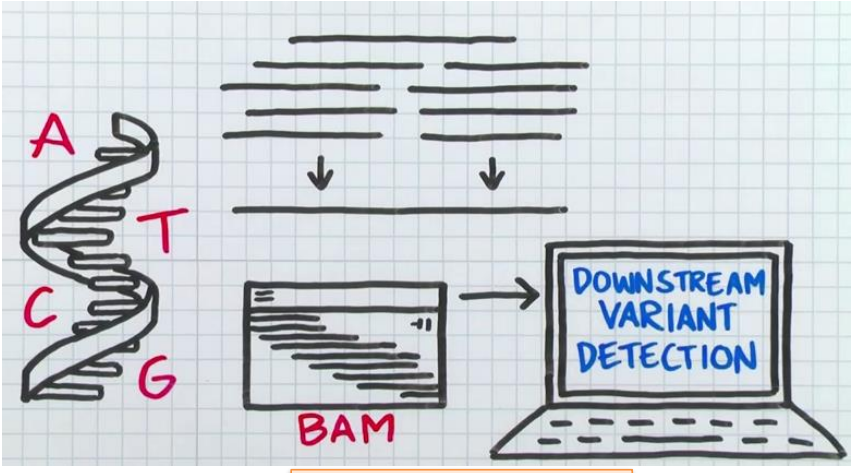
"Variant calling"

https://www.youtube.com/watch?v=0-5XtQOyVOA

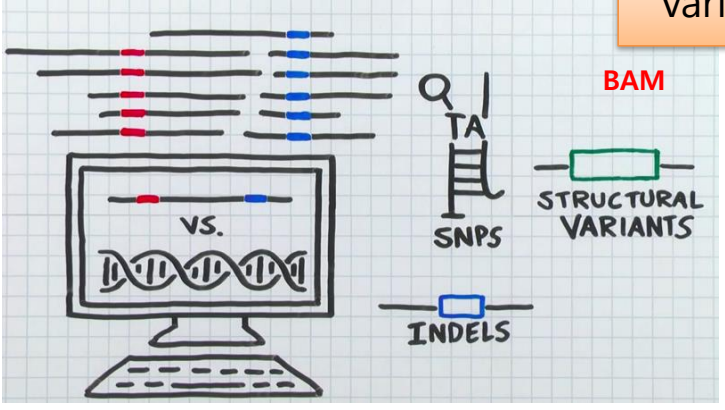
[VCF format]

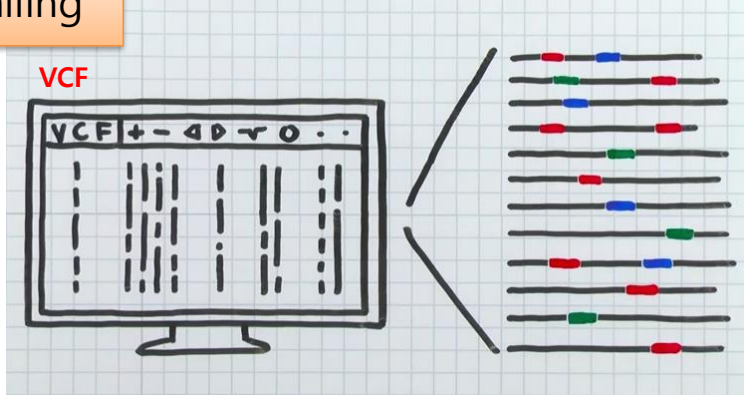
NGS WGS analysis for variation detection

Library preparation → NGS → FASTQ → Mapping → BAM



Variant calling



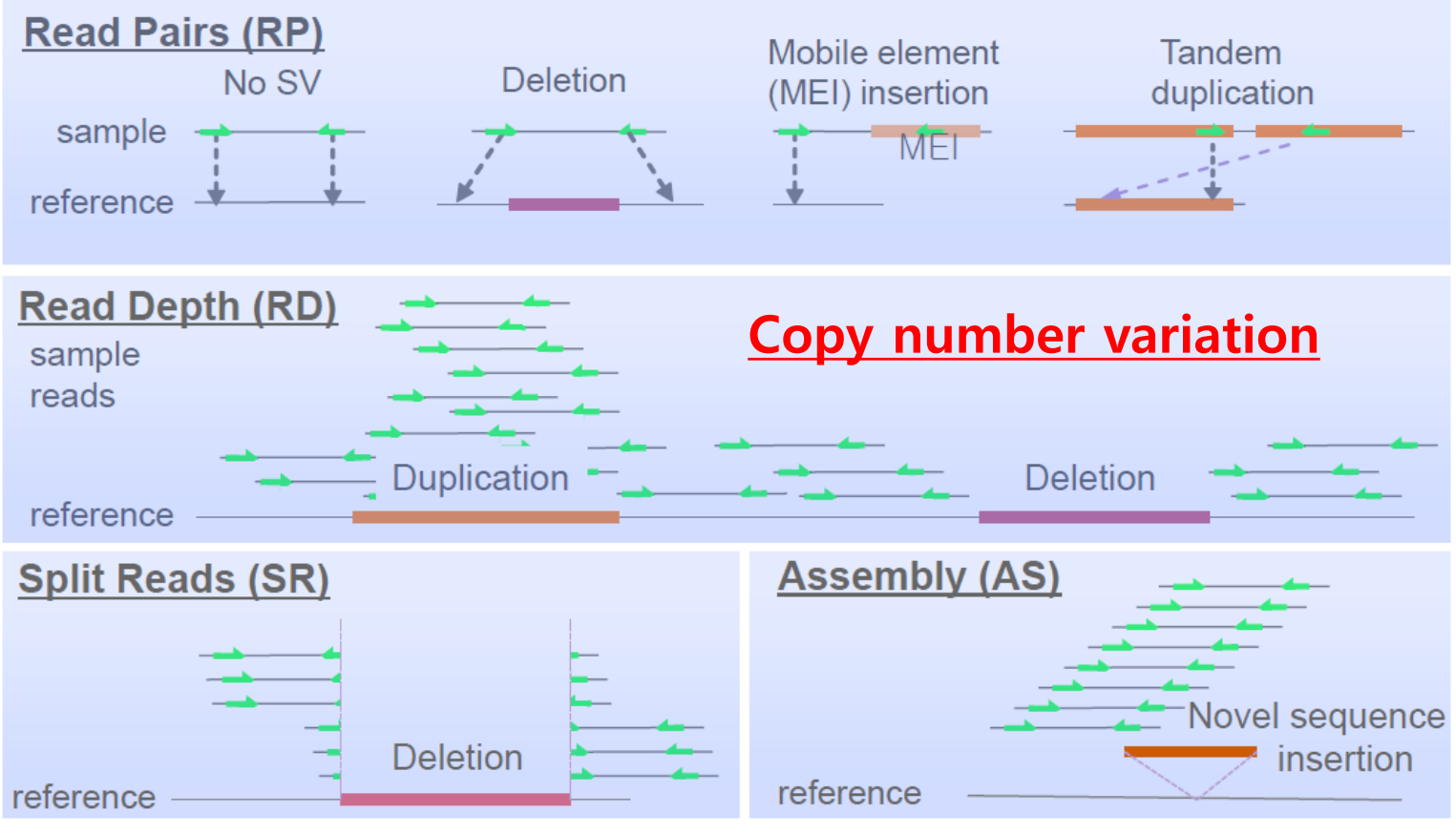


https://www.youtube.com/watch?v=6JfXU5iOAD0&list=PLGlvFEwL2wDFVLBA1egJFwHyetpVDPExK&index=11

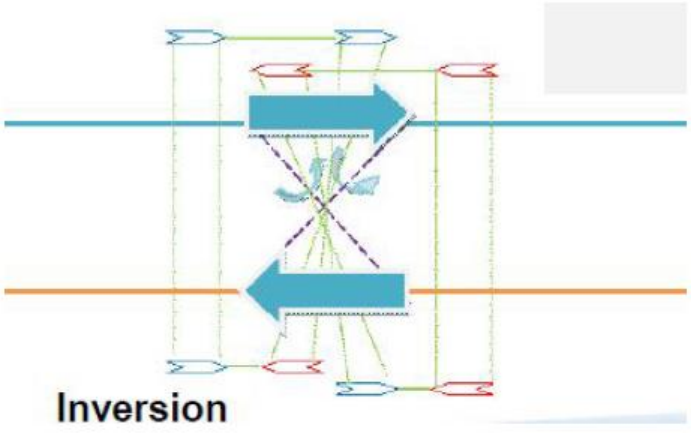
SV Discovery with diverse resources

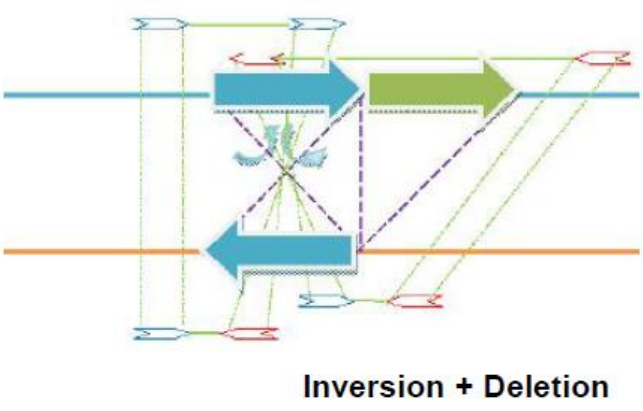
Discovery of Structural Variation

- Challenging in mapping
- Difficult for interpretation



SV Discovery with diverse resources





Faulty mapping of pair-end reads could be interpreted as "inversion" event !!

- **Inversion**
  - Does not involve a loss of genetic information
  - Simply rearranges the linear gene sequence
  - Generally considered to have no deleterious or harmful effects