Summary & Review

Sung Wook Chi
Division of Life Sciences, Korea University

What we learned in the previous lecture

- **Genomics**

- The Core Aims of Genome Science

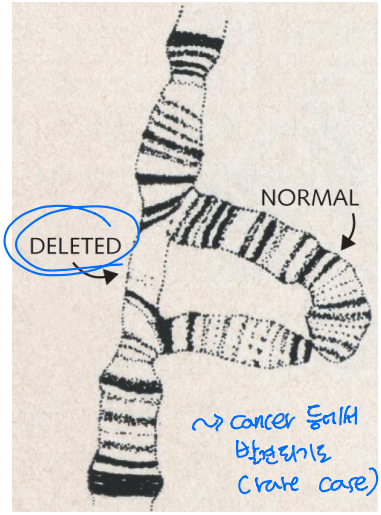
- **Mapping Genomes** ~> sequencing 위해서 필요했었던

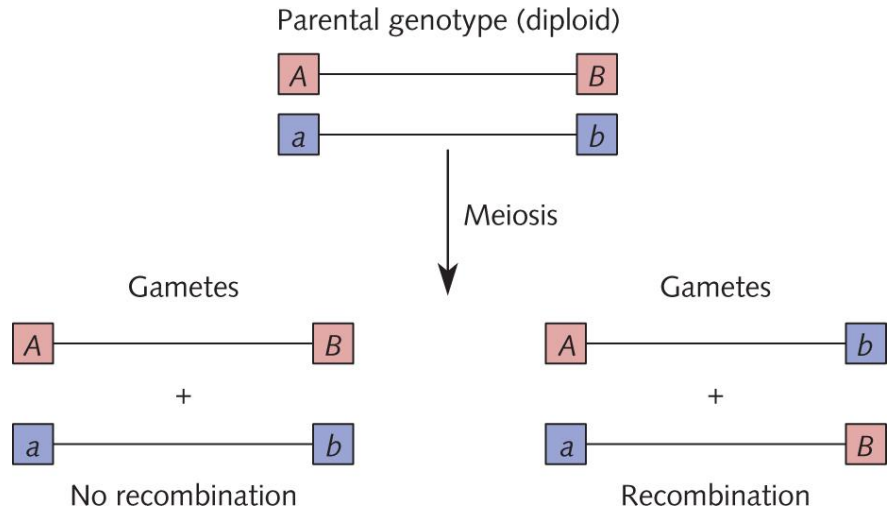
- Genetic Maps, Physical Maps, Cytological Maps
  size of fragment
- Comparative Genomics

- **The Human Genome Project**

- Objectives, Internet Resources

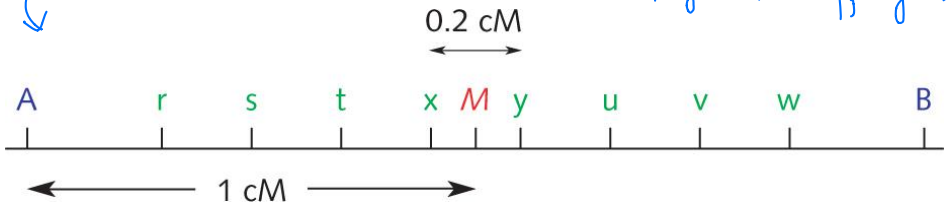
Genetic Map: recombination





often inside same chromosome

marker 이용하면 physical mapping 가능



↗ likely to be located in other chr.

1cM : 1% of recombination , 50 cM : maximum, 50% ⇒ 거리측정하등

실제 확률

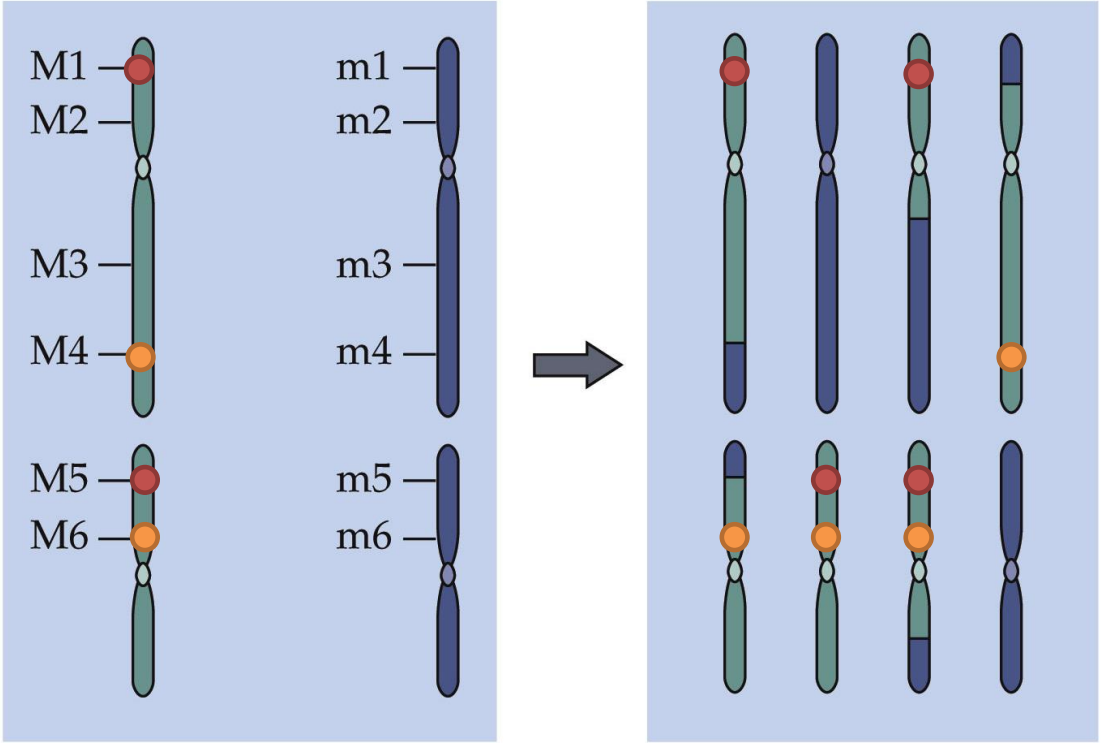
of genetic distance

Mapping Genomes : Genetic Maps

- **Genetic Maps**

should be located in same chromosome

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



Independent segregation

Located in different chromosome

: 50% chance of co-segregation

가까이있어서함께 chunk together & phenotype linked ⇒ linked.

Genetic Maps

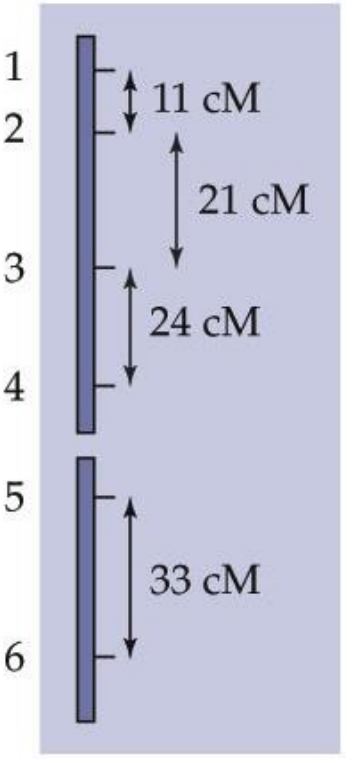
how we can construct genetic map

Linkage data -> Mapping function

(B)

	M1	M2	M3	M4	M5	M6
m1	–	.11	.31	.51	.49	.53
m2		–	.22	.46	.52	.48
m3			–	.25	.51	.50
m4				–	.49	.52
m5					–	.33
m6						–

(C)



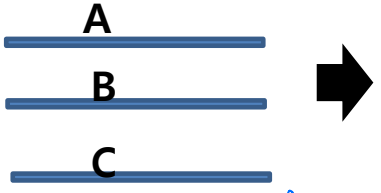
50cM – different chromosome

Mapping Genomes : Physical Maps

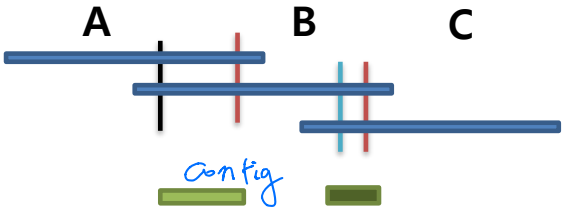
- **Physical Maps**

marker 간의 실제거리를 infer 하는방법
fragment analysis ⇒ arrange and find relative order

: an assembly of contiguous stretches of chromosomal DNA – **contigs** – in which the distance between landmark sequences of DNA is expressed in kilobases.



belong to some part of chromosome



how A,B,C are orientated (정보가 있으면 실제 order 추측가능)
from our genome

- Assemble contigs

1. Alignment of randomly isolated clones based on shared restriction fragment length profiles

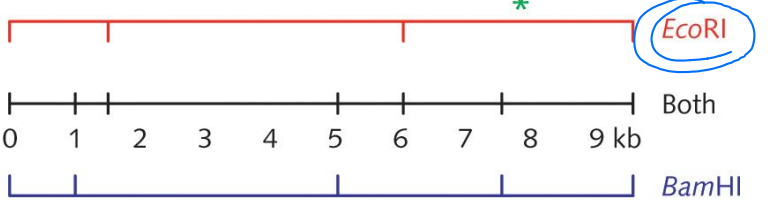
6n 7n 조합으로서의 ⇒ 일종의 landmarker

2. Hybridization-based approaches

: chromosome walking

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

특정 chromosome 에서만 발견되는 sequence가 있다면 act as landmarker → 추적해서 찾을것



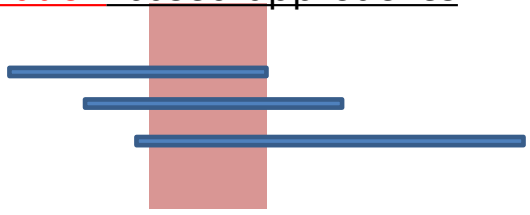
mapping based on length

Mapping Genomes : Physical Maps

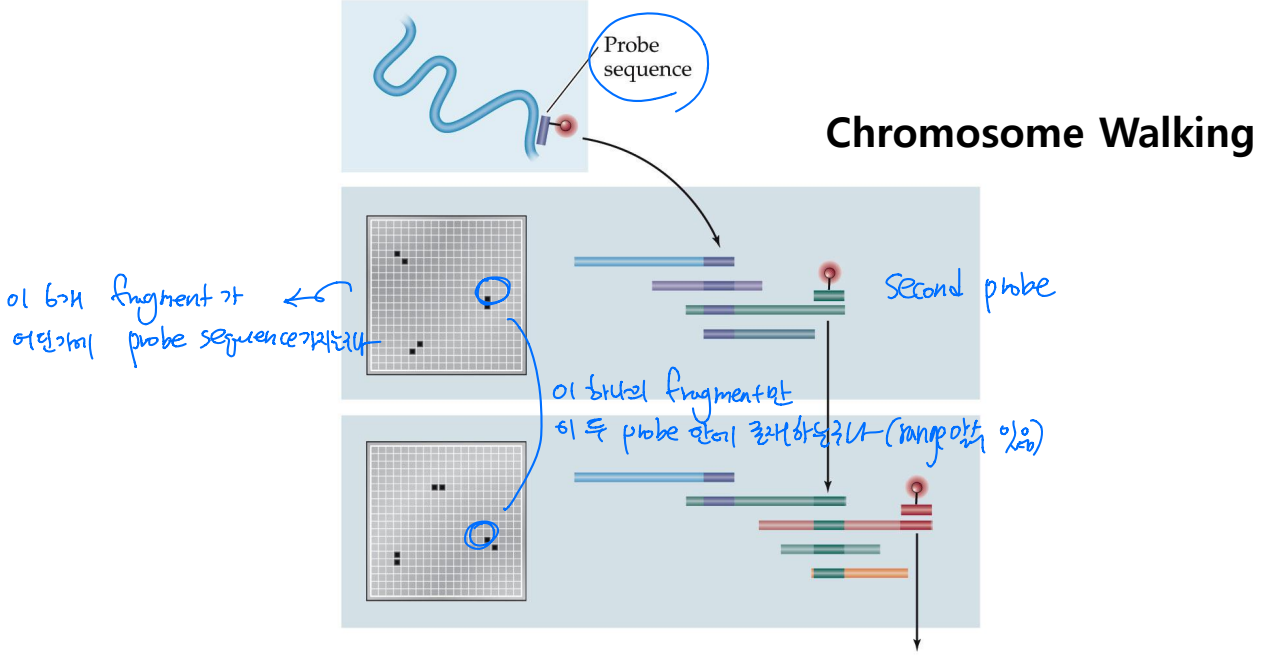
1. Assembling of contigs: Hybridization-based approaches

: chromosome walking

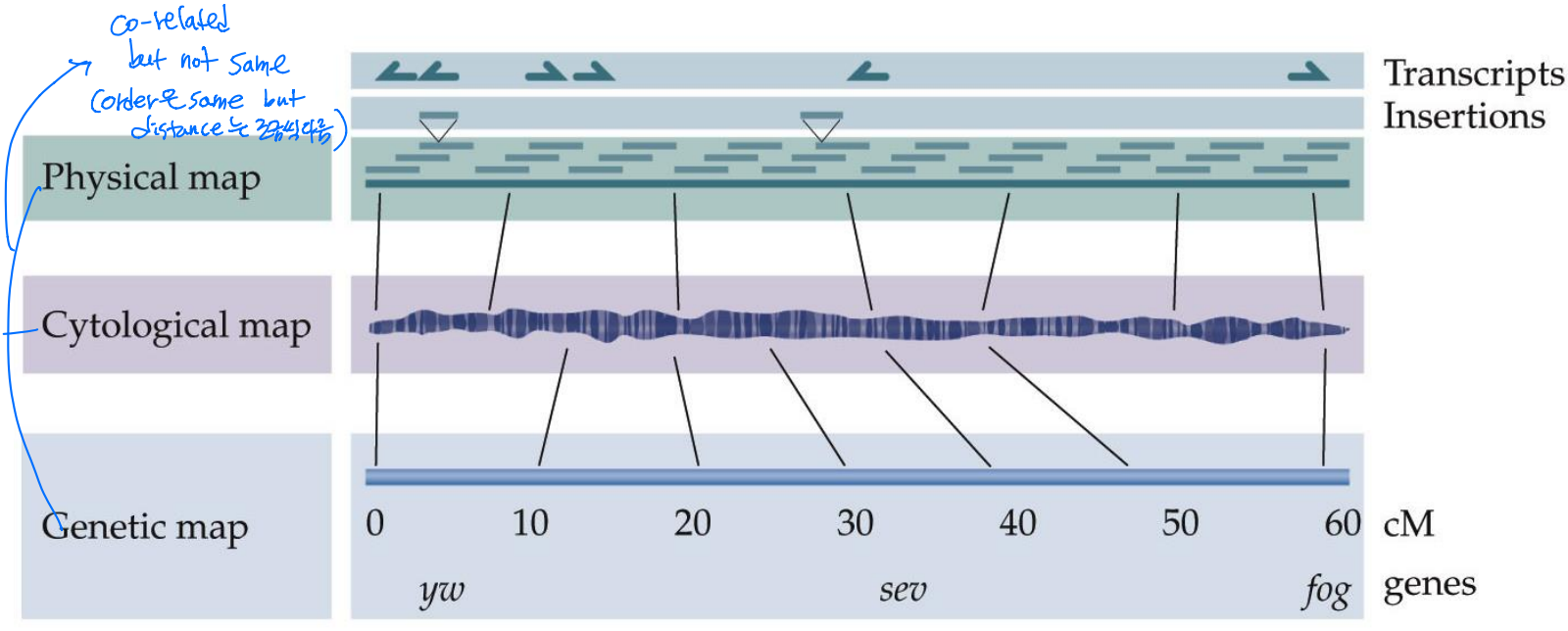
know sequence → design probe



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



Mapping Genomes :Cytological Map



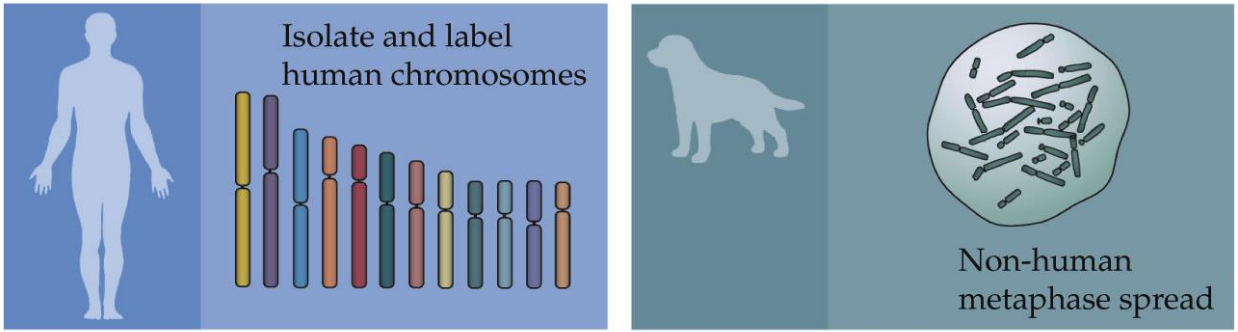
Physical map, Cytological map and Genetic map are correlated,

But not the same..


In situ hybridization

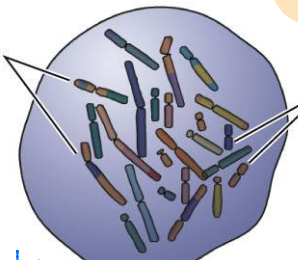
Comparative Genomics: Chromosome painting

paint each with 형광 probe



사람의 genome에 표지 해서 개의 genome에 FISH → 겹치는 부분 알 수 있음

FISH: fluorescence in situ hybridization



Multicolored chromosomes indicate breakage/fusion events

Single-color chromosomes indicate complete correspondence between species

Synteny

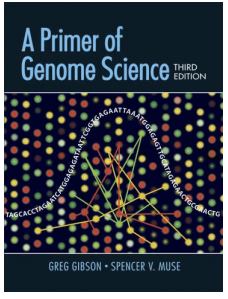
Compare large chunk of chromosome between diff. species

The Human Genome Project

Sung Wook Chi
Division of Life Sciences, Korea University

What we will learn today

- **Sanger Sequencing** / PHRED score
- **Human genome project**
  - What we know about the human genome



Chapter2. Genome Sequencing and annotation

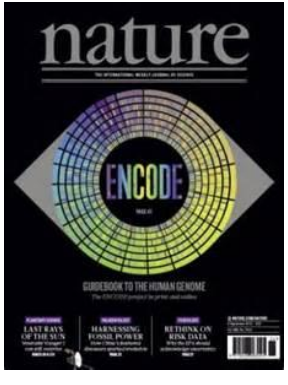
genome 관련 작업들

2001

Draft

2004

Completion



2007

Regulation

1000명의 sequence

2008

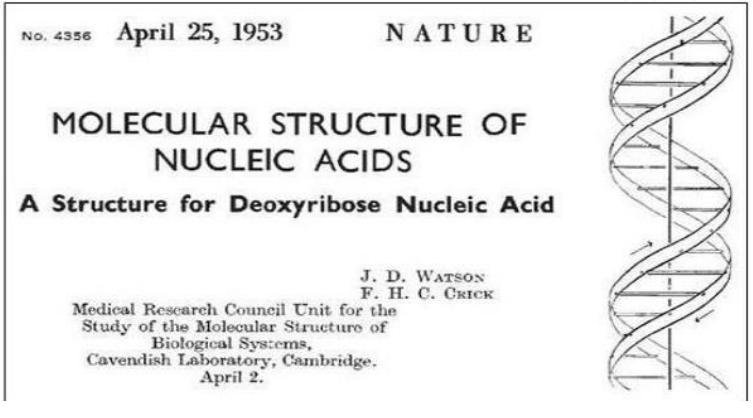
Variation

DNA sequencing (Sanger Sequencing)

DNA

(1928-)

A, T, G, C



No. 4356 April 25, 1953 NATURE

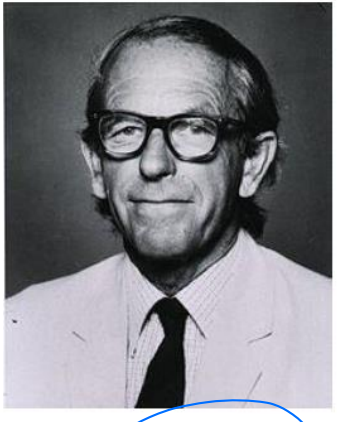
MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

J. D. Watson
F. H. C. Crick

Medical Research Council Unit for the Study of the Molecular Structure of Biological Systems, Cavendish Laboratory, Cambridge. April 2.

Sequencing



Fred Sanger
(1918-2013)

Protein Sequencing (1955, insulin)

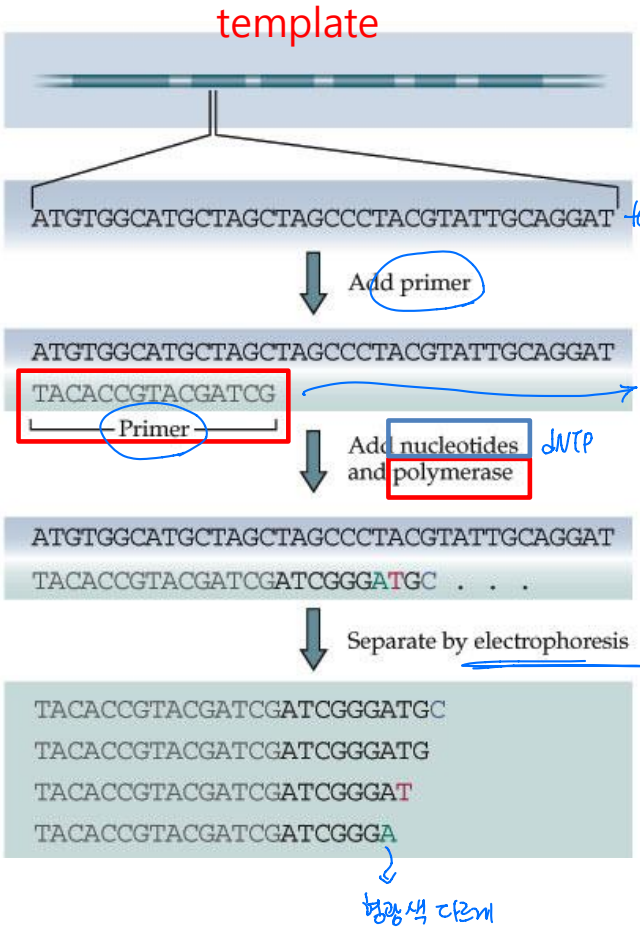
Nobel prize 1958

DNA Sequencing (Sanger sequencing) (1977)

Nobel prize 1980

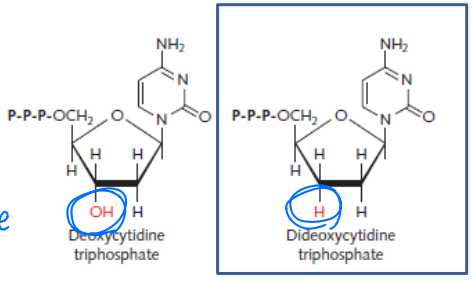
Sanger Sequencing: Principle

DNA polymerase reaction

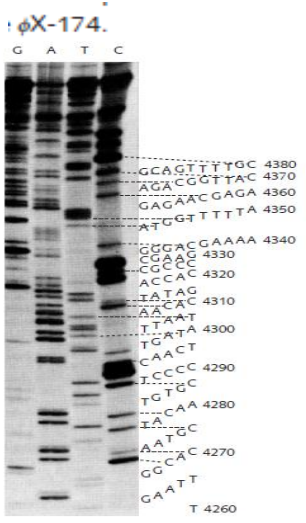


ddNTP

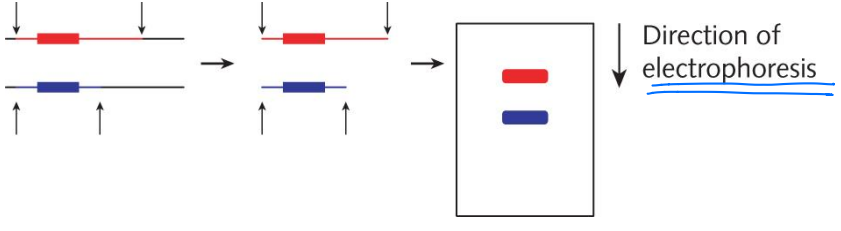
Nucleotide : Dideoxy NTP



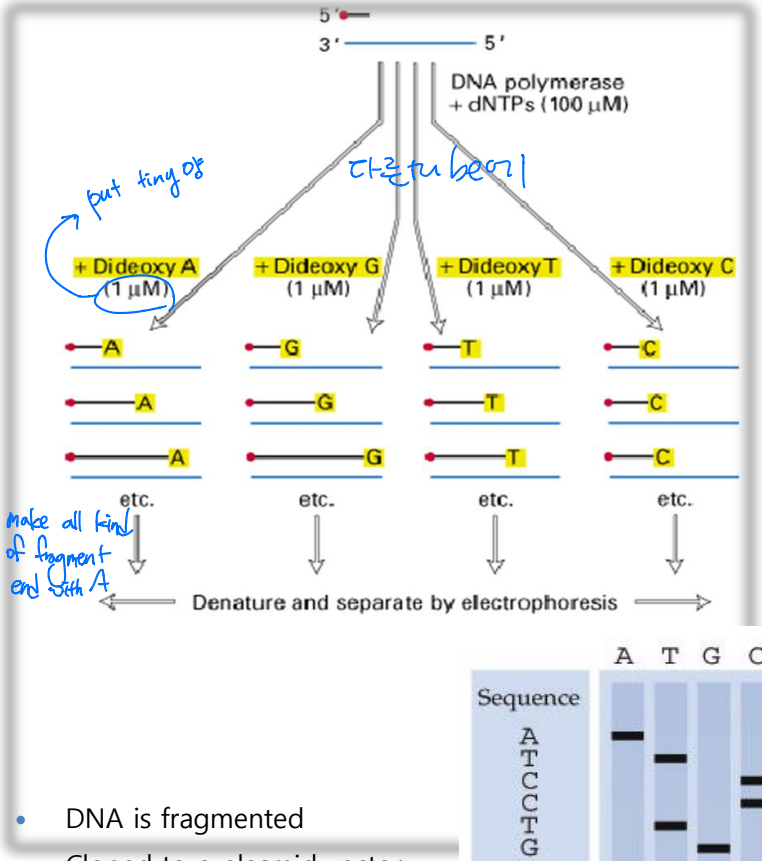
Chain termination



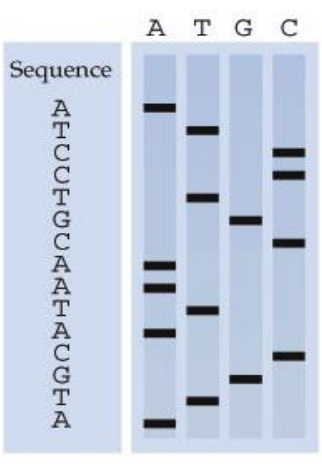
From: Sanger, F., Nicklen, S., & Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**, 5463–5467.



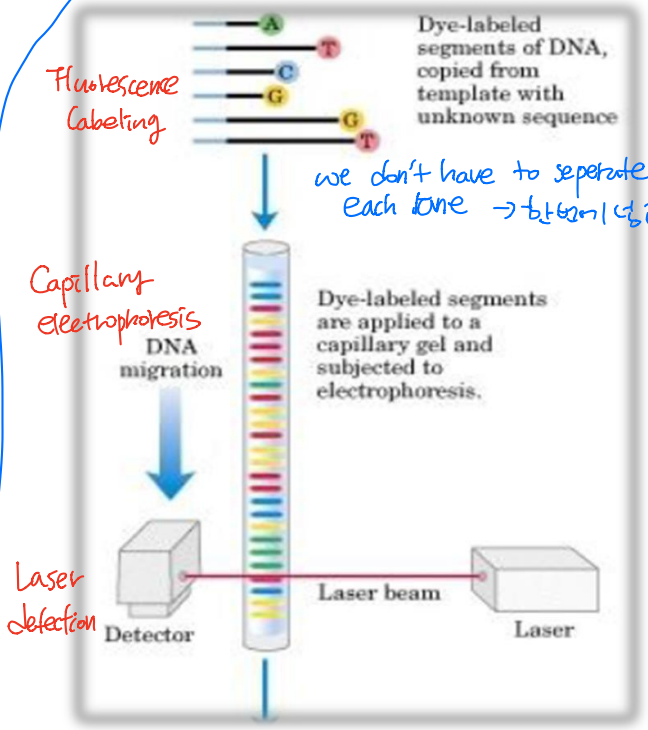
Sanger Sequencing : Chain termination method

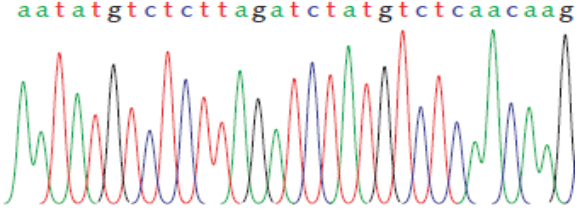


- DNA is fragmented
- Cloned to a plasmid vector
- Cyclic sequencing reaction
- Separation by electrophoresis



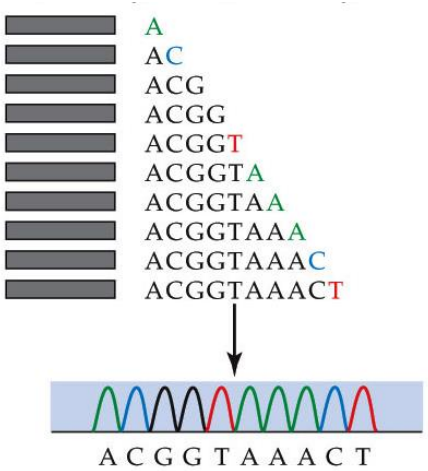
Automation of Sanger method





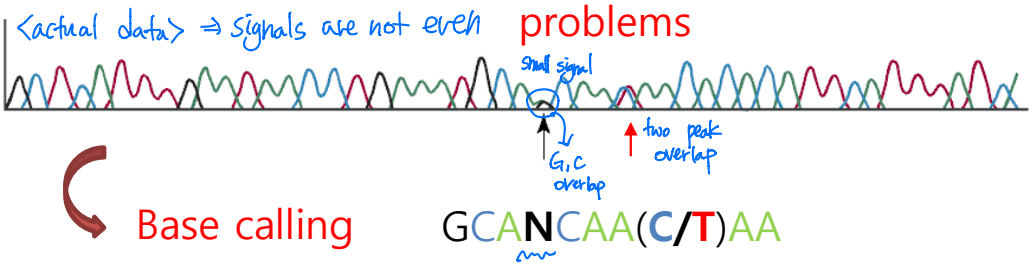
Base calling : read peak signal

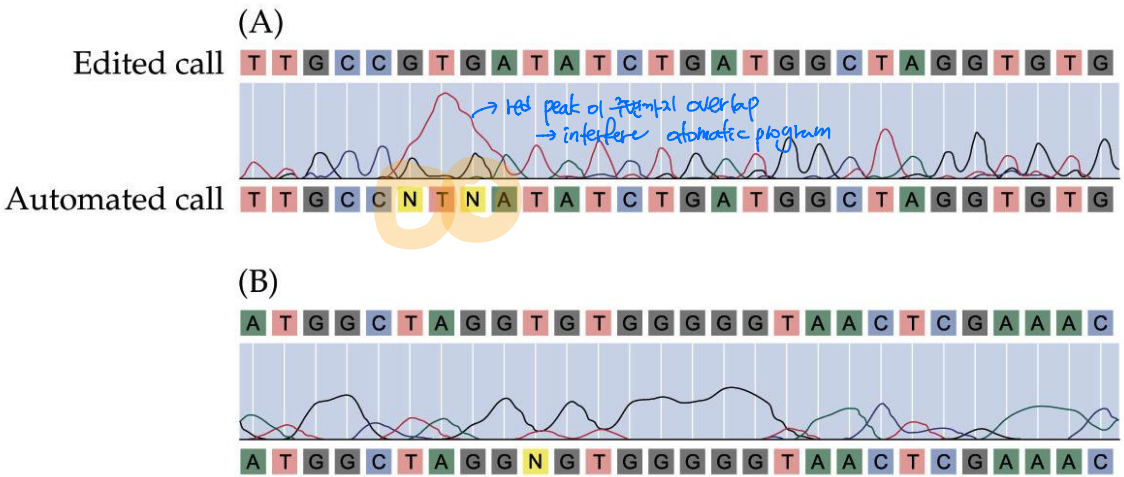
Sanger Sequencing : Base calling



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.





- Shape of signal is import
- How to evaluate quality of base calling?
- Need quality score !!

https://www.youtube.com/watch?v=e2G5zx-OJIw

PHRED score: quality score of base calling

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

Phillip Green

Phred Base-calling Algorithm (1998)

$Q = -\log_{10} P$

P: probability of error

Fluorescence peaks

Locate predicted peaks, using Fourier methods to fit best distribution

Locate observed peaks, for which the area under the concavity exceeds 10% of the previous 10 peaks or 5% of the previous one

Match observed and predicted peaks using a three-stage shifting algorithm

Find missing peaks

Assess error probabilities of each peak according to four-parameter model

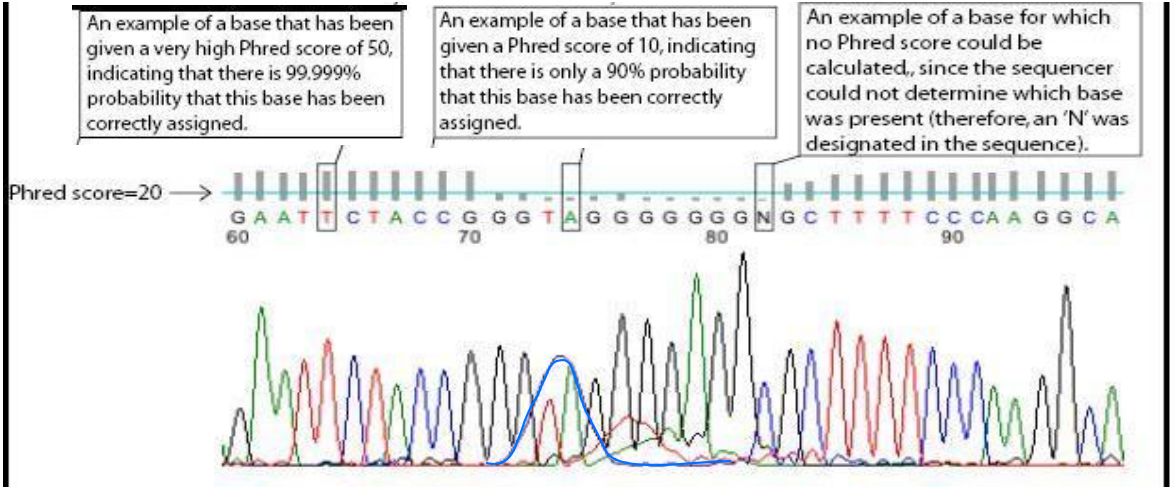
Phred Base-calling algorithm

Sequence reads

Q>20, Error rate < 0.01 good

Q=20 (p=0.01)

Q<20, Error rate > 0.01 bad



Human genome sequencing with Sanger method

Automation of sanger sequencing method

Vector – insert (~1.5kb) ⇒ need to be fragmented under this size

- Sanger Sequecing
- Base-calling

Sequence reads (~1.5kb)
Quality scores

3 billion base pairs

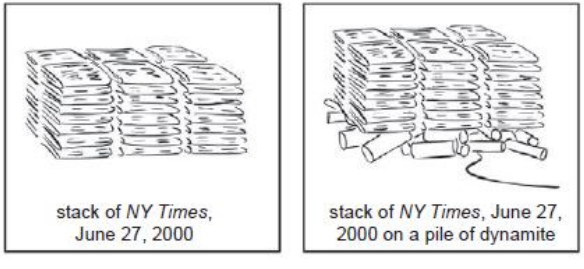
Shotgun Sequencing Strategy

Human genome

Fragmentation (random)

→ 다시 조각

Sequencing



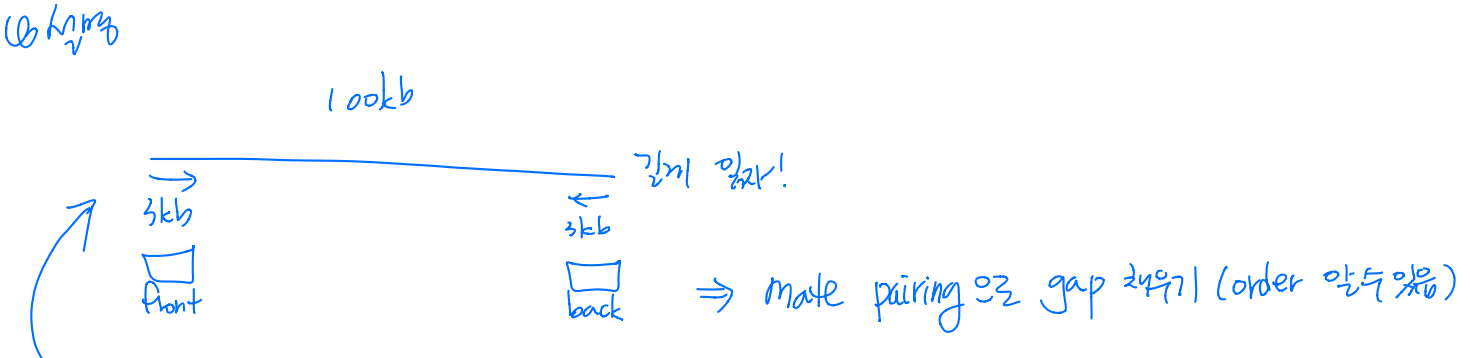
Assembly

조합.

need to recombine

What we will learn today

- **Sanger Sequencing** / PHRED score
- **Human genome project**
  - Hierarchical shotgun sequencing
  - Whole genome shotgun sequencing
  - What we know about the human genome



Human Genome Projects

Celera

Craig Ventor

Vs.

competetion

International Human Genome Sequencing Consortium

Francis Collins NIH 책임자

Whole-genome shotgun sequencing

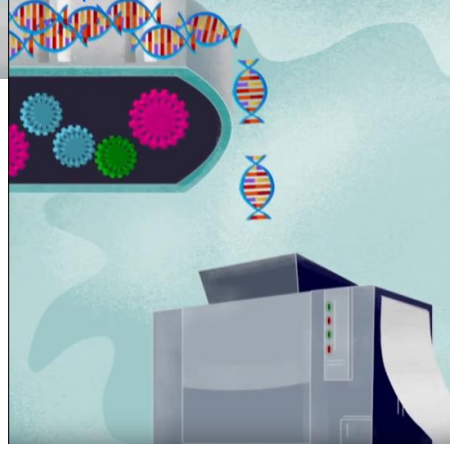
break into small piece

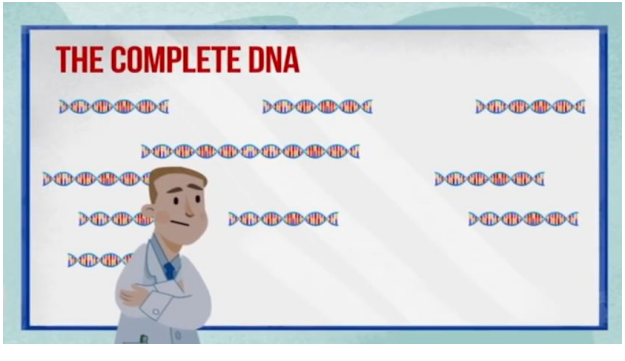
Hierarchical shotgun sequencing

1. Short (~kb) random **fragmentation** of genome

Sanger sequencing 이라는 방법으로 각각 읽어냄

2. Sanger Sequencing





3. Assembly (Contigs)

small fragment → back into original

Hierarchical Shotgun Sequencing

↳ used 시리얼 fragmentation

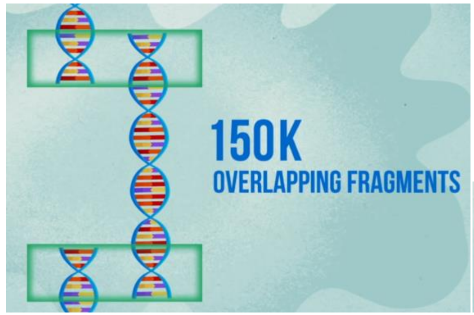
Fragmentation of genome

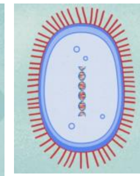
~150Kb

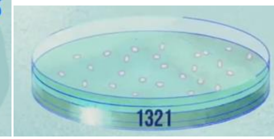
↗ Store / amplify 가능

BAC/YAC

Scaffold (Physical maps)







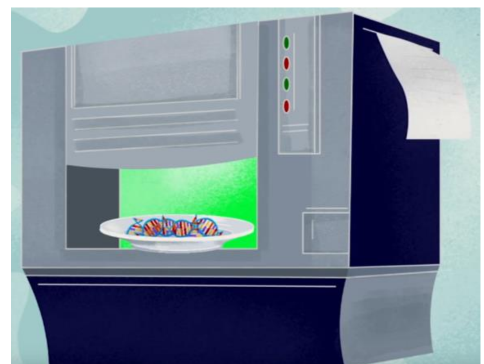
Chromosome walking

named all the clones with chromosome

150kb 사이즈의 클론을 만들어 BAC / YAC에 넣어 관리

Fragmentation of BAC/YAC clone

Sanger Sequencing



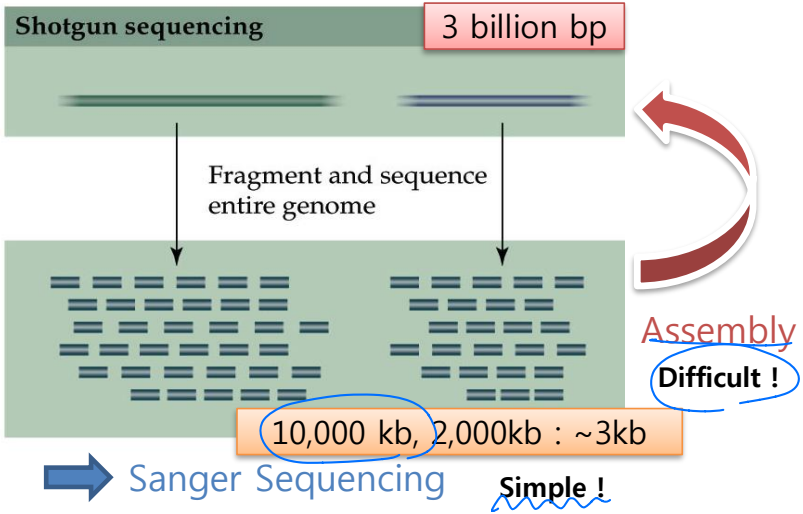
각 fragment를 작은사이즈로 만들어 (kb 규모) contig의 서열을 알아냄 → 물리적지도 순으로 이어붙임

3. Assembly (Contigs)

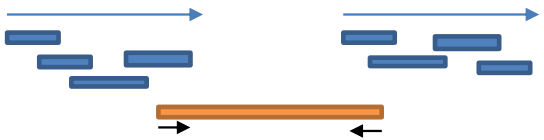
TTTGCTCGGA
TCCTGGCTCC
AGGCCGCGTC

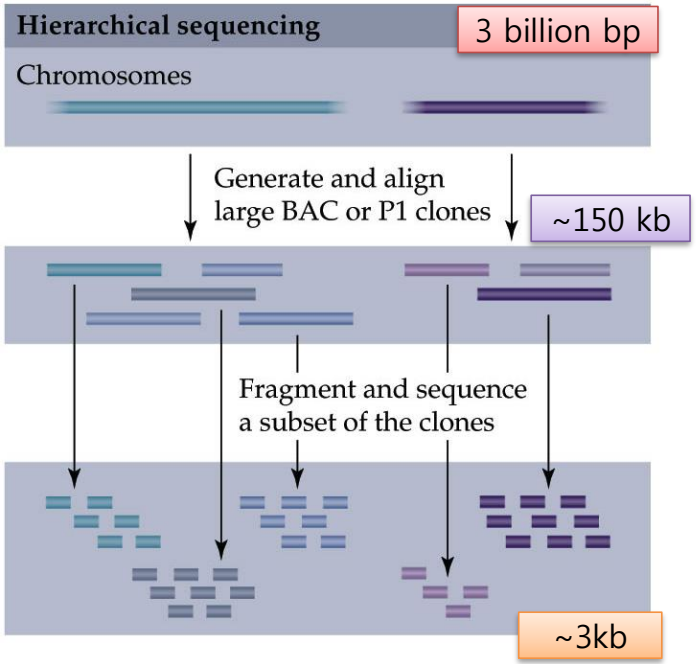
Hierarchical vs. whole genome shotgun sequencing

initially attempting to fragment



Overlapper > contigs > Scaffolds

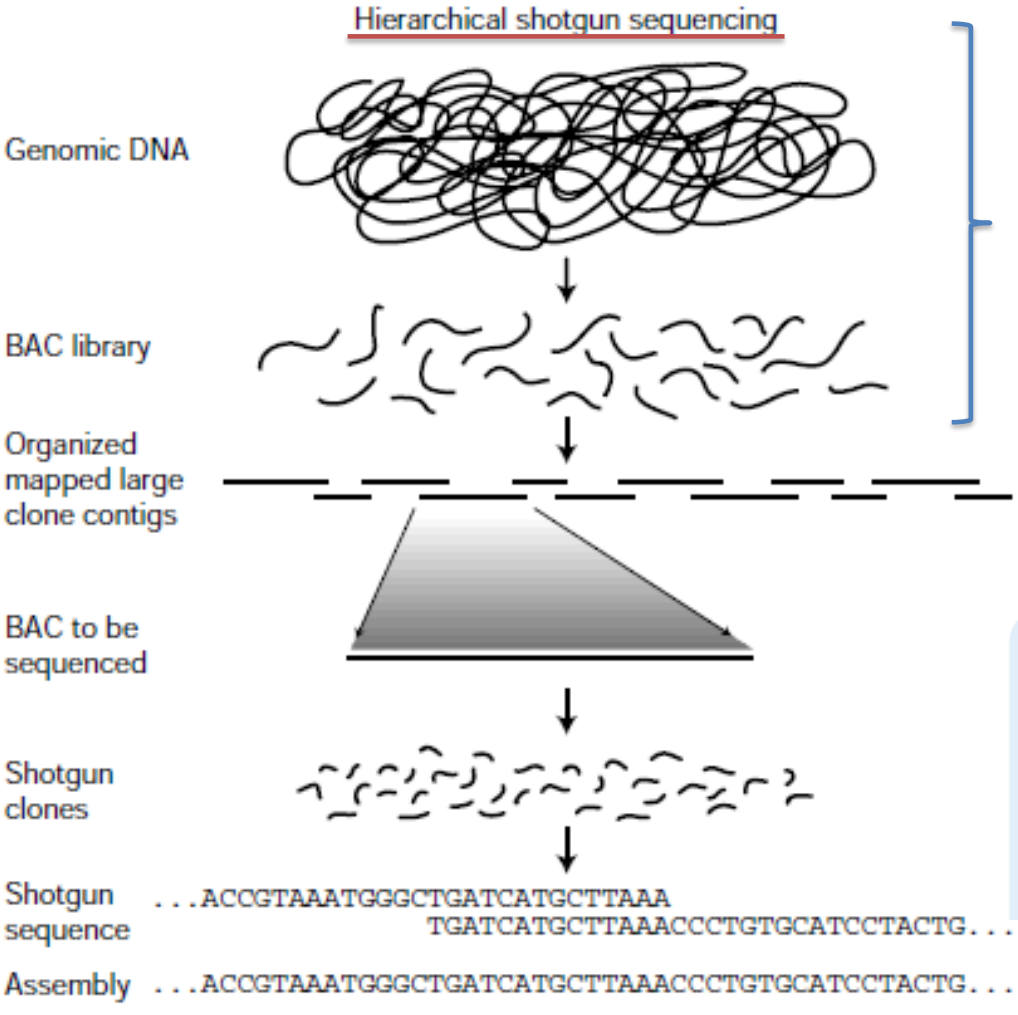




- Scaffold
- Overlapper> Contigs > Supercontig

예외 ⇒ gap이 생기는 이유는 repeat sequence

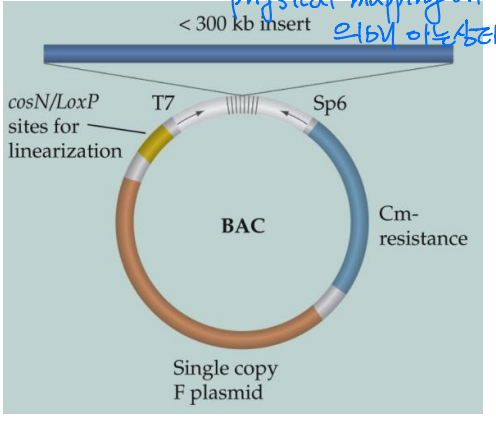
Hierarchical shotgun sequencing strategy



'BAC-to-BAC' method

- Scaffold: decide order/orientation
- Chromosome walking
  Fingerprinting

BAC clone의 특성은 physical mapping의 이론적 이득성태

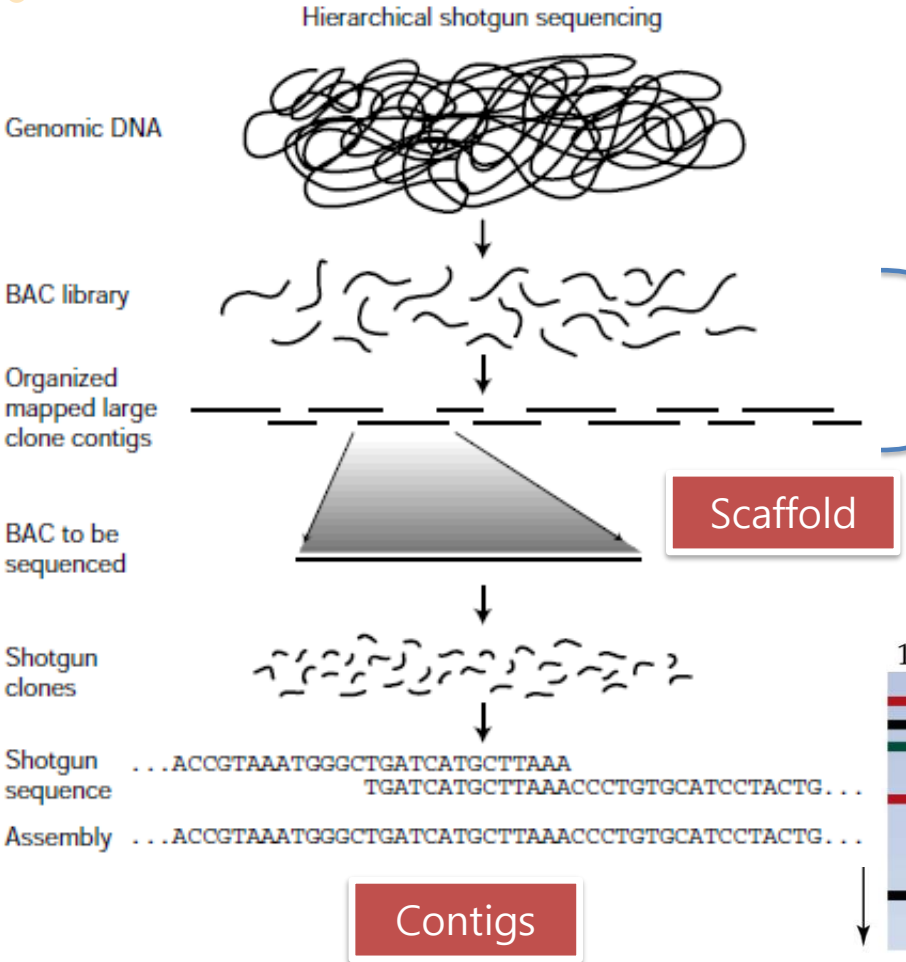


BOX 3.8 **BACs: bacterial artificial chromosomes**

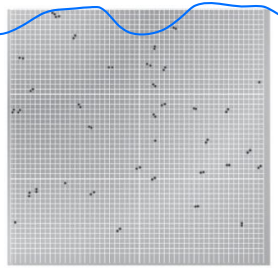
A plasmid is a small piece of double-stranded DNA in a bacterial cell, in addition to the main genome. A bacterial artificial chromosome, or BAC, is a plasmid containing foreign DNA – for instance, a fragment of the human genome. A typical BAC, in an *Escherichia coli* cell, can carry about 250 000 bp.

Hierarchical shotgun sequencing: Scaffold & Contigs

Aligning BAC clones by hybridization and fingerprinting



Chromosome walking

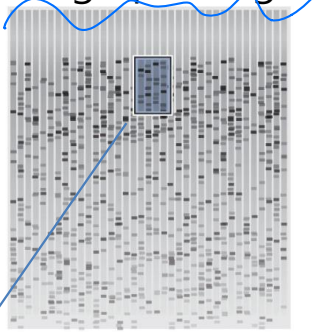


STS probes

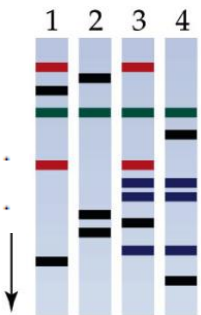
Hybridization

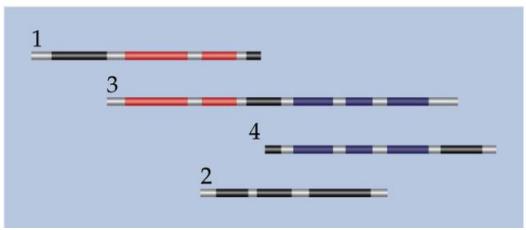
Scaffold 만드는 방법

Fingerprinting



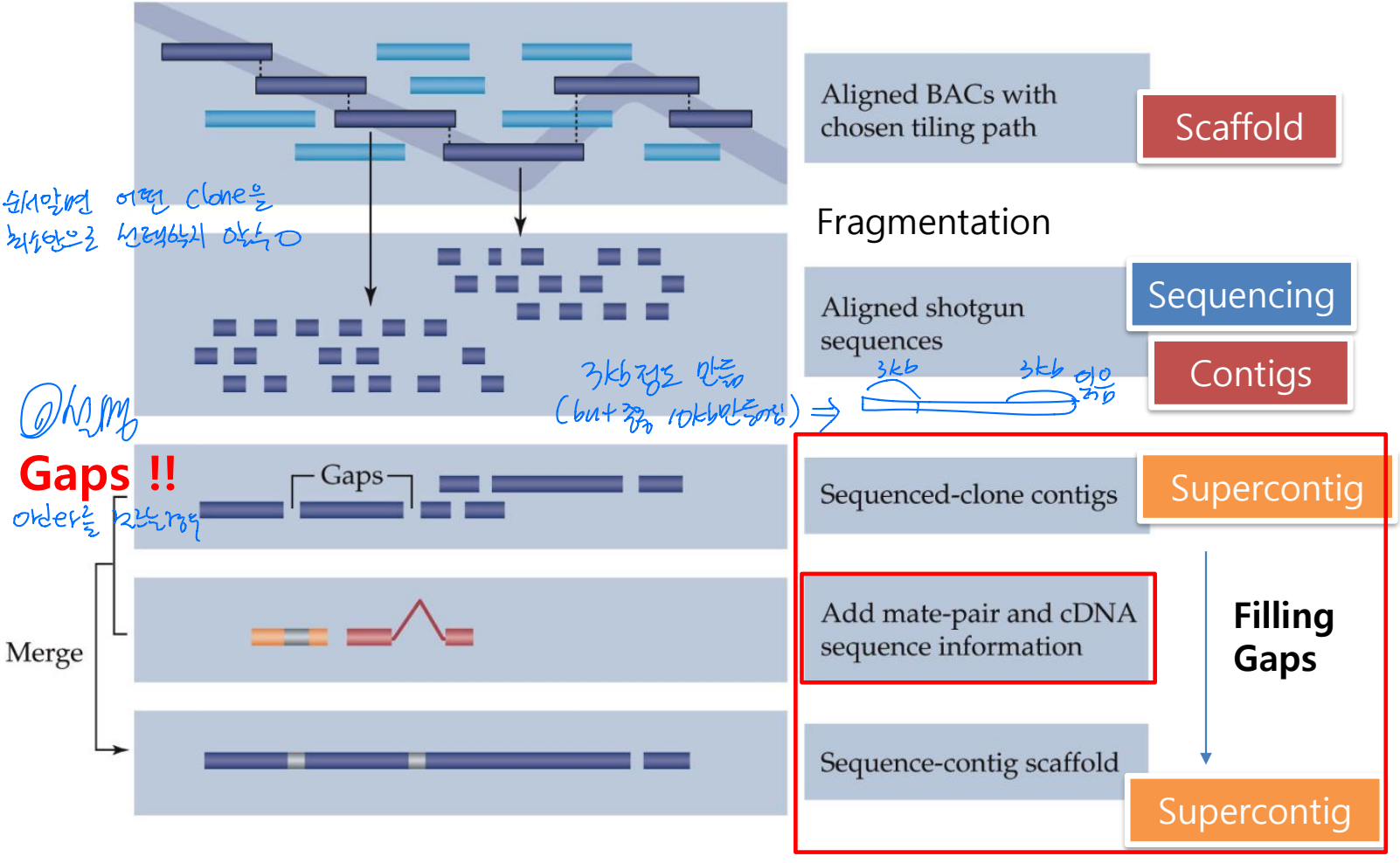
Fragments from restriction enzyme





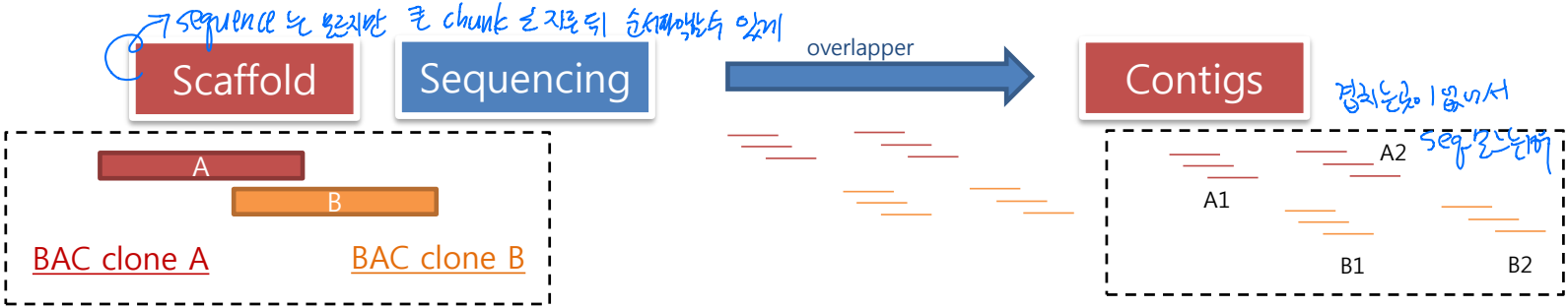
여기도 좋았다 ~

Hierarchical assembly of a sequence-contig scaffold (supercontig)

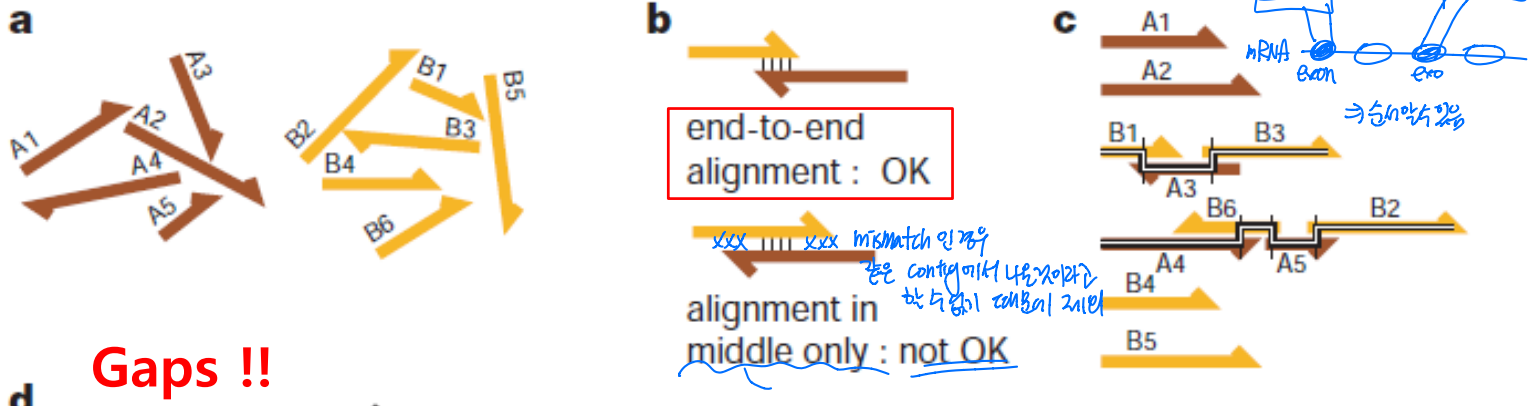


Challenge : How to merge contigs (filling gaps) to generate supercontig?

Assembling individual sequenced clones into the draft genome sequence



Assembly (Contigs > Supercontig)



The key steps (a–d) in assembling individual sequenced clones into the draft genome sequence. A1–A5 represent initial sequence contigs derived from shotgun sequencing of clone A, and 1–B6 are from clone B.

Human genome project: hierarchical shotgun sequencing

The New York Times

(Feb, 13, 2001)

articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

(First draft, 2001)

READING THE BOOK OF LIFE

READING THE BOOK OF LIFE; Grad Student Becomes Gene Effort's Unlikely Hero

By NICHOLAS WADE
Published: February 13, 2001

A surprising hero helped the consortium of academic scientists decoding the human genome to avoid a drubbing by its rival, the Celera Genomics company. Scientists throughout the world are now beating an electronic path to his Web site, where they can analyze and download the human genome sequence. He is a graduate student at the University of California at Santa Cruz, and his name is not Clark but James Kent.

FACEBOOK
TWITTER
GOOGLE+
EMAIL
SHARE
PRINT

Jim Kent

David Haussler

Methods

Assembly of the Working Draft of the Human Genome with GigAssembler

W. James Kent[1,3] and David Haussler[2]

Genome Res. 2001 September; 11(9): 1541–1548

cDNA를 통해 과학으로 gap 메워주는 프로그램이다

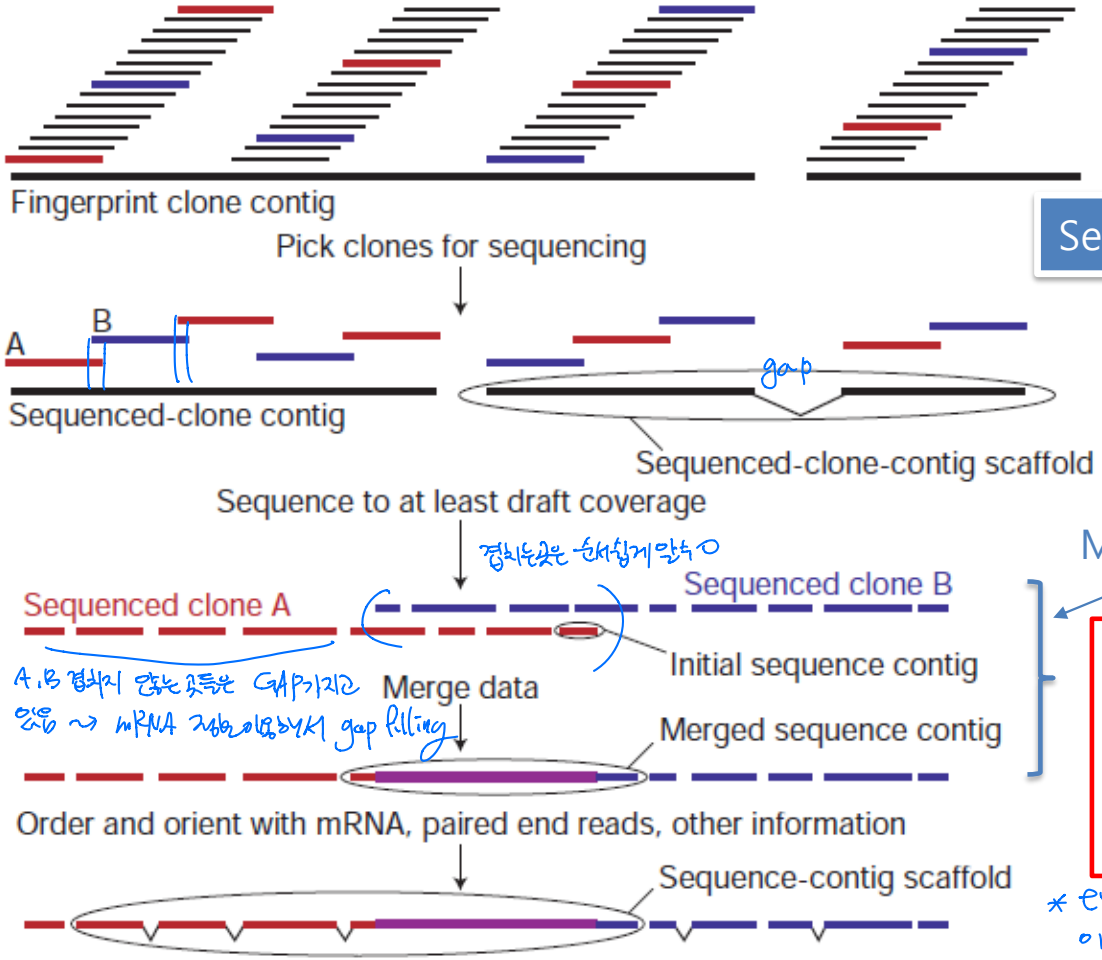
can used to fill the gap

UCSC Genome Bioinformatics

UCSC genome browser http://genome.ucsc.edu/

Genome Res. 2002 **12** (6): 996–1006

Genome sequence assembly (Levels of clone)



Scaffold

Sequencing

Contigs

Supercontig

Merge

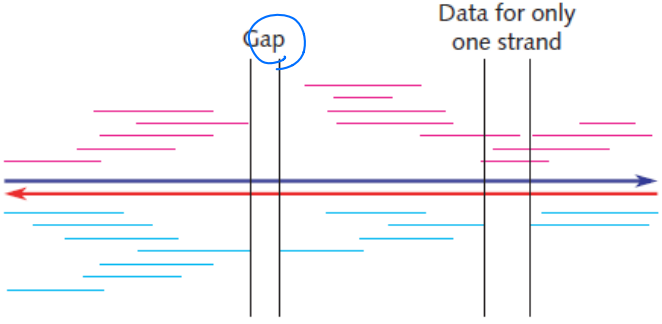
Filling Gaps !!

Order/ orient

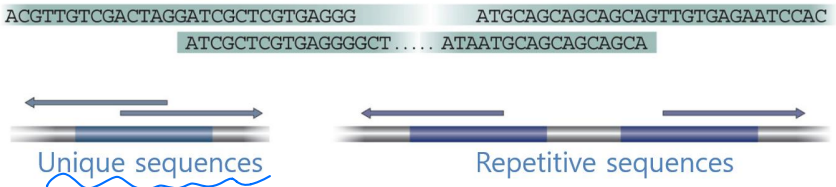
- mRNA
- Paired end reads

Supercontig

Whole genome shotgun sequencing (Celera's approach)

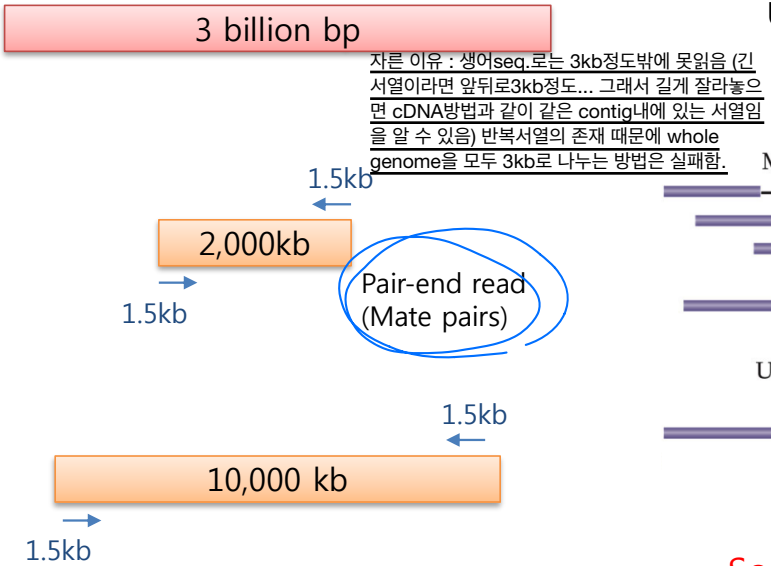


Problems: Repetitive sequences

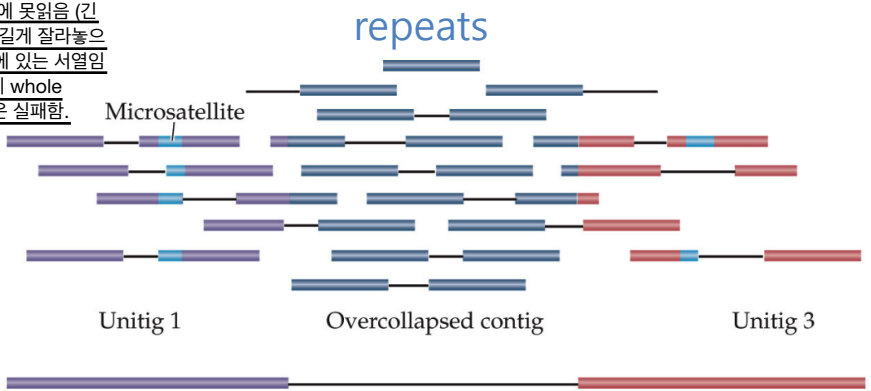


Unitigs and repeat resolution

Unique contiguous sequence alignments



자른 이유 : 생어seq.로는 3kb정도밖에 못읽음 (긴 서열이라면 앞뒤로3kb정도... 그래서 길게 잘라놓으면 cDNA방법과 같이 같은 contig내에 있는 서열임을 알 수 있음) 반복서열의 존재 때문에 whole genome을 모두 3kb로 나누는 방법은 실패함.



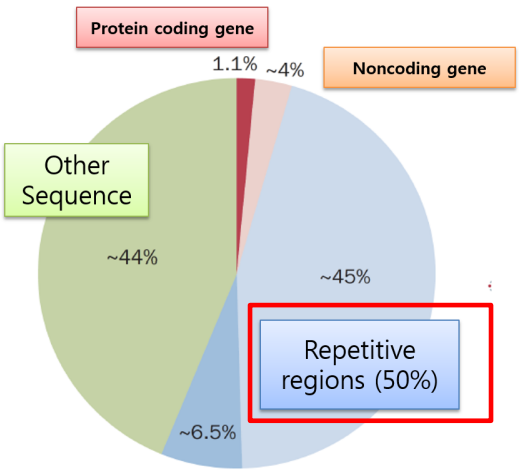
Solution: Mate pairs from larger fragments

Completion of human genome project

BOX 3.9 **Common and different steps in 'BAC-to-BAC' and whole-genome shotgun methods**

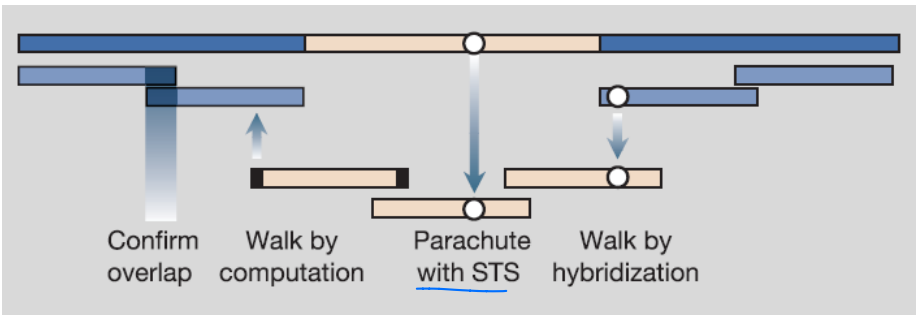
	'BAC-to-BAC' method	Whole-genome shotgun method
1.	Make random cuts to produce fragments of:	
	~150 kb	~2000 kb and 10 000 kb
2.	Make plasmid library in BACs.	
3.	Fingerprint, overlap, and order BAC clones.	3. Skip this step.
4.	Partially sequence 1500 bp subfragments of individual clones.	
5.	Assemble overlaps by computer.	

to find scaffold



2004년에 gap 없이 염색체 아래까지도 다 gap은 남음

Remaining gaps > finishing the physical maps



iterative 'walking' from the ends of contigs

What we will learn today

- **Sanger Sequencing** / PHRED score
- **Human genome project**

- Hierarchical shotgun sequencing
- Whole genome shotgun sequencing
- What we know about the human genome

2004

centromere, telomere, repeat. seq은 다음에 진행

Finishing the euchromatic sequence of the human genome

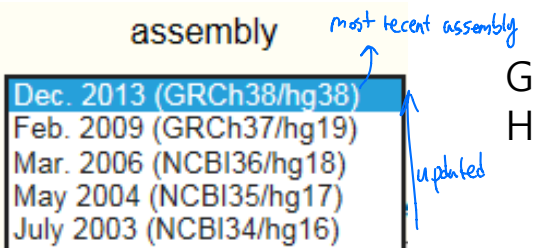
International Human Genome Sequencing Consortium*

* *A list of authors and their affiliations appears in the* Supplementary Information

The sequence of the human genome encodes the genetic instructions for human physiology, as well as rich information about human evolution. In 2001, the International Human Genome Sequencing Consortium reported a draft sequence of the euchromatic portion of the human genome. Since then, the international collaboration has worked to convert this draft into a genome sequence with high accuracy and nearly complete coverage. Here, we report the result of this finishing process. The current genome sequence (Build 35) contains 2.85 billion nucleotides interrupted by only 341 gaps. It covers ~99% of the euchromatic genome and is accurate to an error rate of ~1 event per 100,000 bases. Many of the remaining euchromatic gaps are associated with segmental duplications and will require focused work with new methods. The near-complete sequence, the first for a vertebrate, greatly improves the precision of biological analyses of the human genome including studies of gene number, birth and death. Notably, the human genome seems to encode only 20,000–25,000 protein-coding genes. The genome sequence reported here should serve as a firm foundation for biomedical research in the decades ahead.

↳ 염은별리를 바탕으로 유전자 개수 확인

Current state of human genome sequences



GRCh38 Genome Reference Consortium
Human Reference 38 (GRCh 38/ hg38)

3.2 Gb text file (Fasta format)

chr10.fa one file
chr11.fa
chr11_gl000202_random.fa
chr12.fa
chr13.fa
chr14.fa
chr15.fa
chr16.fa
chr17_ctg5_hap1.fa
chr17.fa
chr17_gl000203_random.fa
chr17_gl000204_random.fa
chr17_gl000205_random.fa
chr17_gl000206_random.fa
chr18.fa
chr18_gl000207_random.fa
chr19.fa
chr19_gl000208_random.fa
chr19_gl000209_random.fa

chr1.fa
chr1_gl000191_random.fa
chr1_gl000192_random.fa
chr20.fa
chr21.fa
chr21_gl000210_random.fa
chr22.fa
chr2.fa
chr3.fa
chr4_ctg9_hap1.fa
chr4.fa
chr4_gl000193_random.fa
chr4_gl000194_random.fa
chr5.fa
chr6_apd_hap1.fa
chr6_cox_hap2.fa
chr6_dbb_hap3.fa
chr6.fa
chr6_mann_hap4.fa

chr6_mcf_hap5.fa
chr6_qbl_hap6.fa
chr6_ssto_hap7.fa
chr7.fa
chr7_gl000195_random.fa
chr8.fa
chr8_gl000196_random.fa
chr8_gl000197_random.fa
chr9.fa
chr9_gl000198_random.fa
chr9_gl000199_random.fa
chr9_gl000200_random.fa
chr9_gl000201_random.fa
chrM.fa
chrUn_gl000211.fa
chrUn_gl000212.fa
chrUn_gl000213.fa
chrUn_gl000214.fa
chrUn_gl000215.fa

chrUn_gl000216.fa
chrUn_gl000217.fa
chrUn_gl000218.fa
chrUn_gl000219.fa
chrUn_gl000220.fa
chrUn_gl000221.fa
chrUn_gl000222.fa
chrUn_gl000223.fa
chrUn_gl000224.fa
chrUn_gl000225.fa
chrUn_gl000226.fa
chrUn_gl000227.fa
chrUn_gl000228.fa
chrUn_gl000229.fa
chrUn_gl000230.fa
chrUn_gl000231.fa
chrUn_gl000232.fa
chrUn_gl000233.fa
chrUn_gl000234.fa

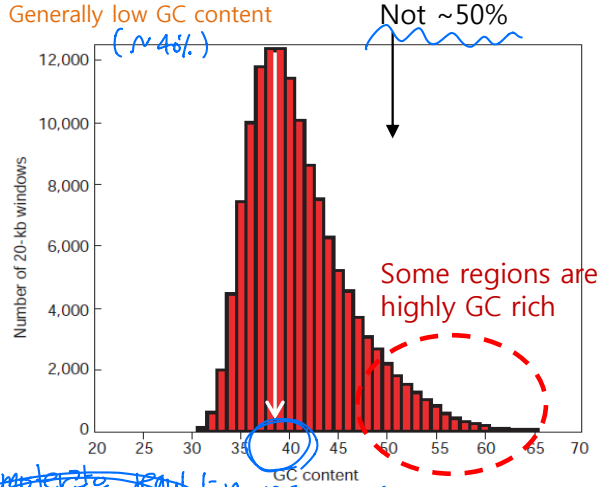
chrUn_gl000235.fa
chrUn_gl000236.fa
chrUn_gl000237.fa
chrUn_gl000238.fa
chrUn_gl000239.fa
chrUn_gl000240.fa
chrUn_gl000241.fa
chrUn_gl000242.fa
chrUn_gl000243.fa
chrUn_gl000244.fa
chrUn_gl000245.fa
chrUn_gl000246.fa
chrUn_gl000247.fa
chrUn_gl000248.fa
chrUn_gl000249.fa
chrX.fa
chrY.fa

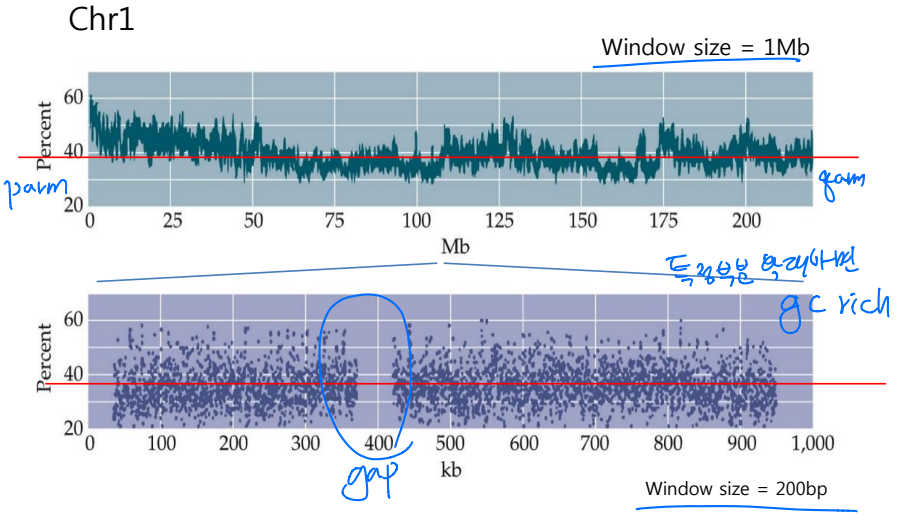
- Unplaced contig names consist of the chromosome number, followed by the NCBI accession number, followed by "random"

- Unlocalized contig names (contigs whose associated chromosome is not known) consist of "chrUn" followed by the NCBI accession number

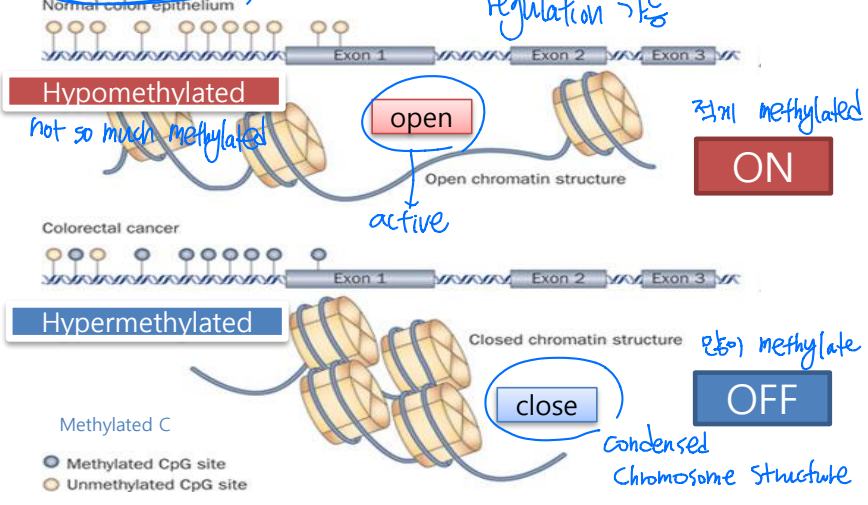
https://www.youtube.com/watch?v=-gVh3z6MwdU

GC contents in human genome

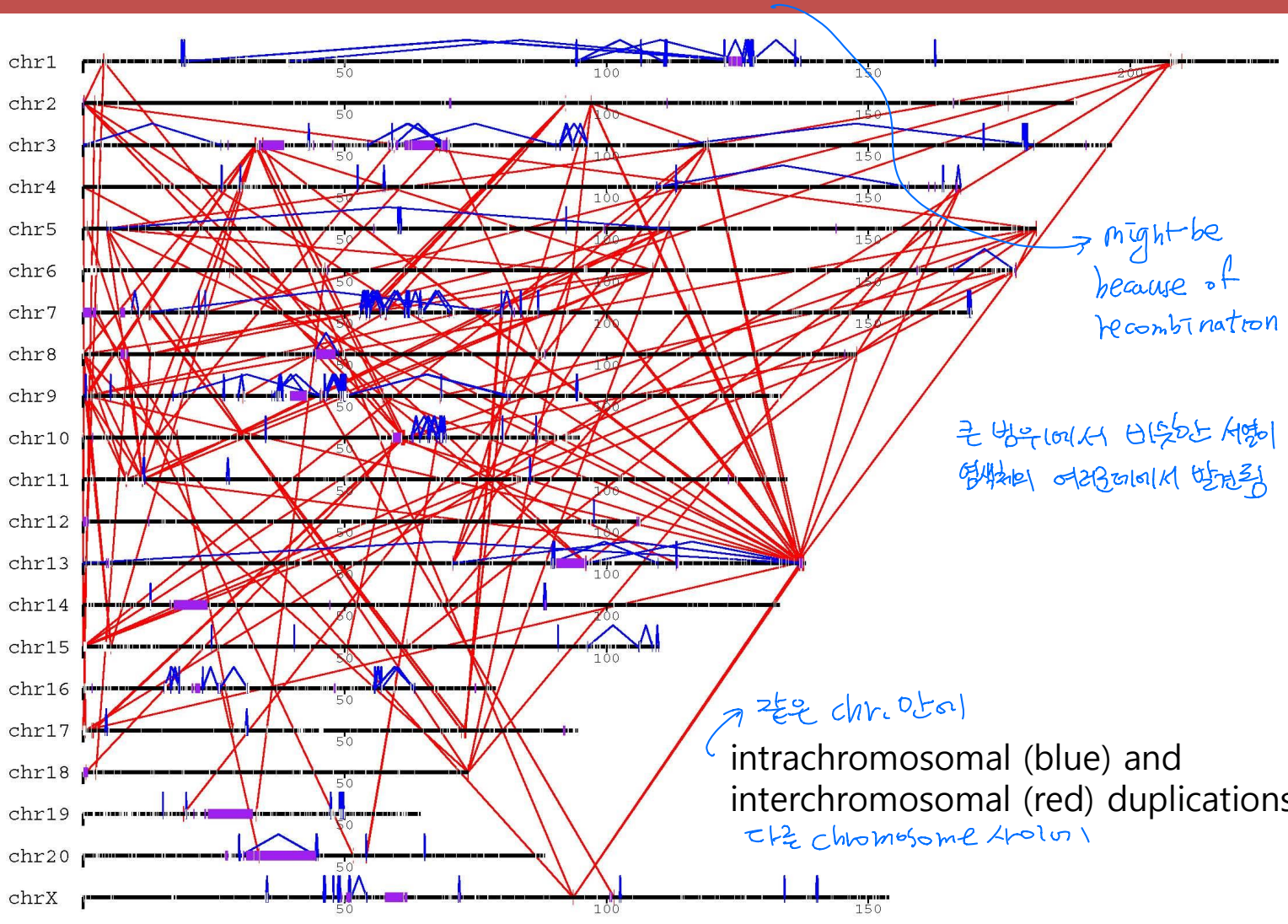




CpG island



Segmental duplication

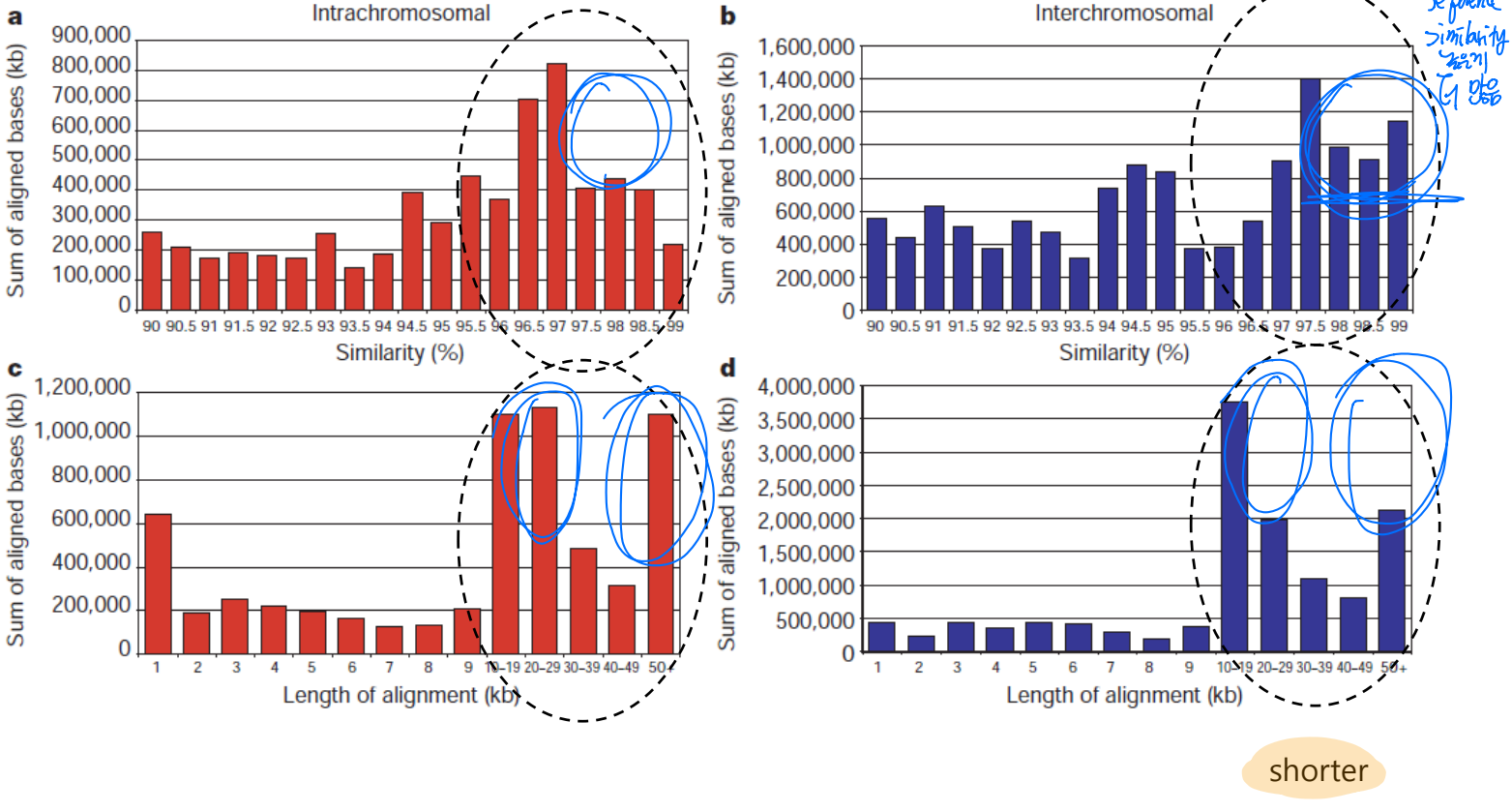


Segmental duplication

Intra-chromosomal

Inter-chromosomal

More similar

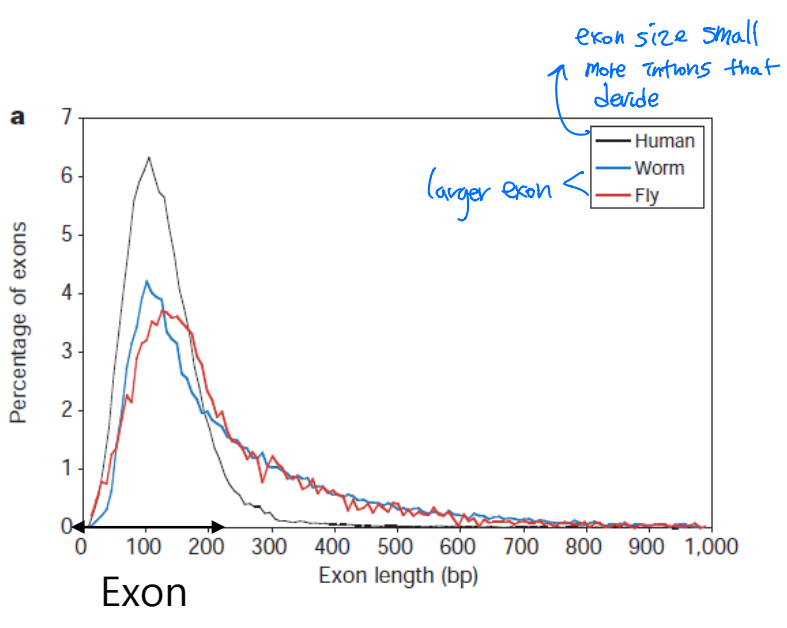


shorter

Characteristics of human genes

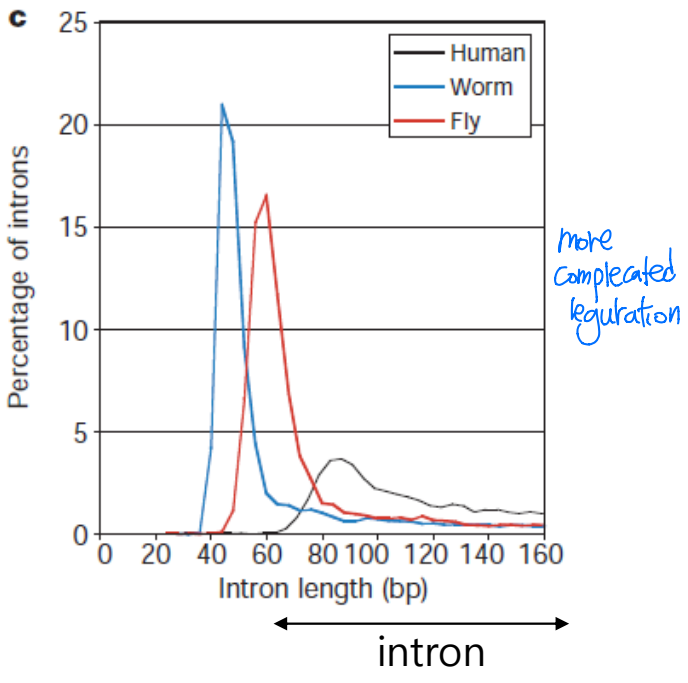
	Median	Mean
Internal exon	122 bp	145 bp
Exon number	7	8.8
Introns	1,023 bp	3,365 bp
3′ UTR	400 bp	770 bp
5′ UTR	240 bp	300 bp
Coding sequence	1,100 bp	1,340 bp
(CDS)	367 aa	447 aa
Genomic extent	14 kb	27 kb

≈ 1kb



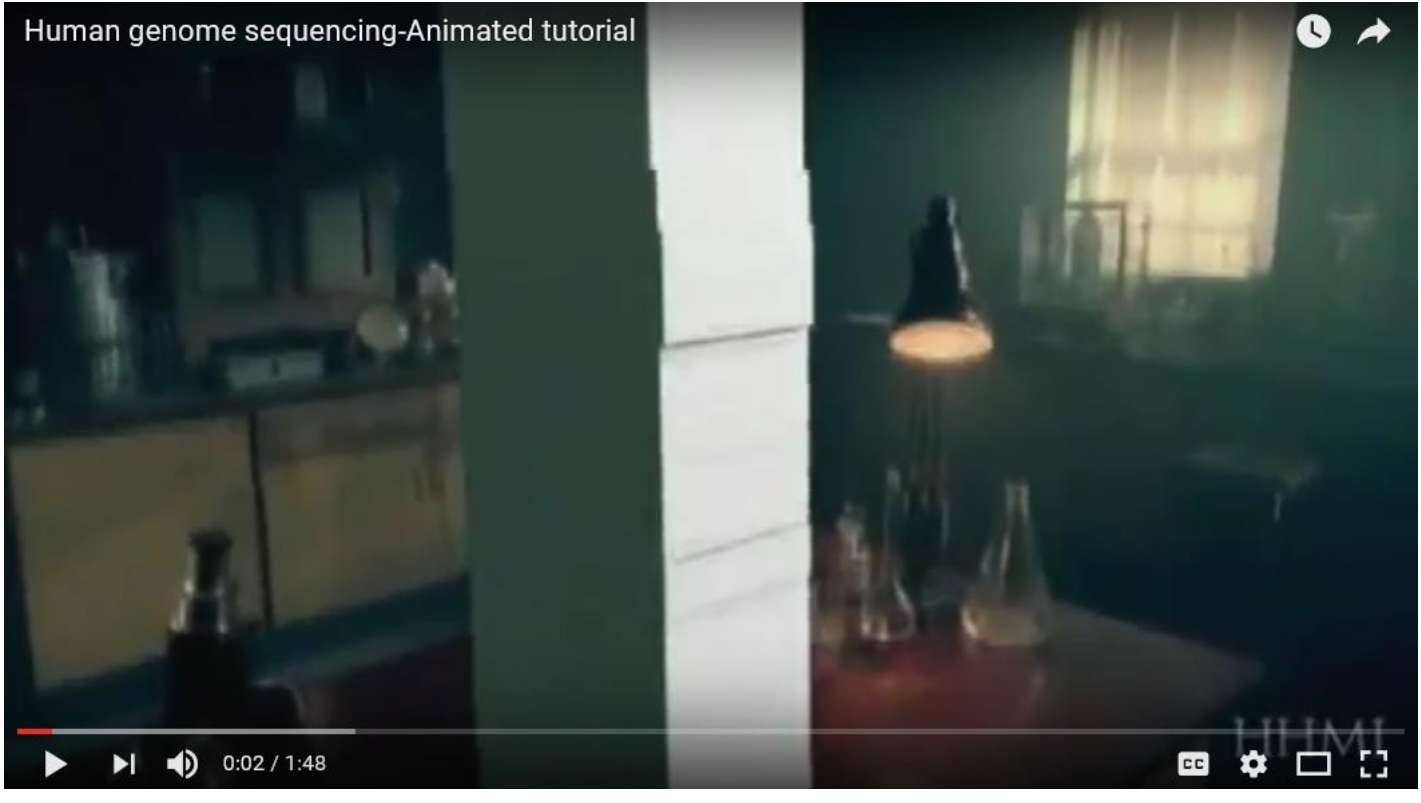
exon size small
more introns that
devide

larger exon <



more complecated regulation

Summary: Human Genome Project



https://www.youtube.com/watch?v=-gVh3z6MwdU