Summary & Review

: introduction to genomics

Sung Wook Chi
Division of Life Sciences, Korea University

What we learned in the previous lecture

1. Introduction
   - What is genomics and functional genomics?

2. Architecture of Genome (Human, Eukaryote)
   - Chromosome, Gene Structure (genome browser)

3. Flow of genomic information
   - Gene expression, ESTs, epigenomic information

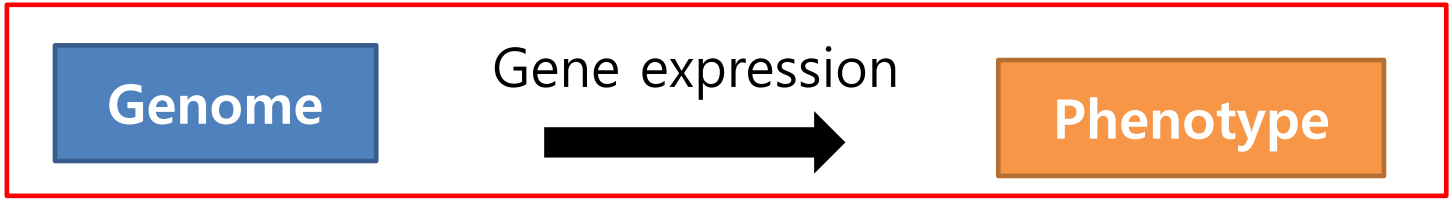
4. Today's Genomics
   - Human genome project, NGS, personal genomics

5. Human genome
   - Repeats (retrotransposons)

6. Open access of genome sequences
- Comparative genomics, Variations (CNV)

Functional Genomics

What I introduced in the previous lecture



1. Genome
2. Human Genome Project
3. Genome Annotation
4. Genomic Variation
5. Comparative Genomics

Genome-wide methods

1. Next-Generation Sequencing (NGS) (WGS, Exome-Seq)
2. Microarray
3. RNA-Seq
4. ChIP-Seq
5. Ribosomal Profiling
6. Bioinformatics

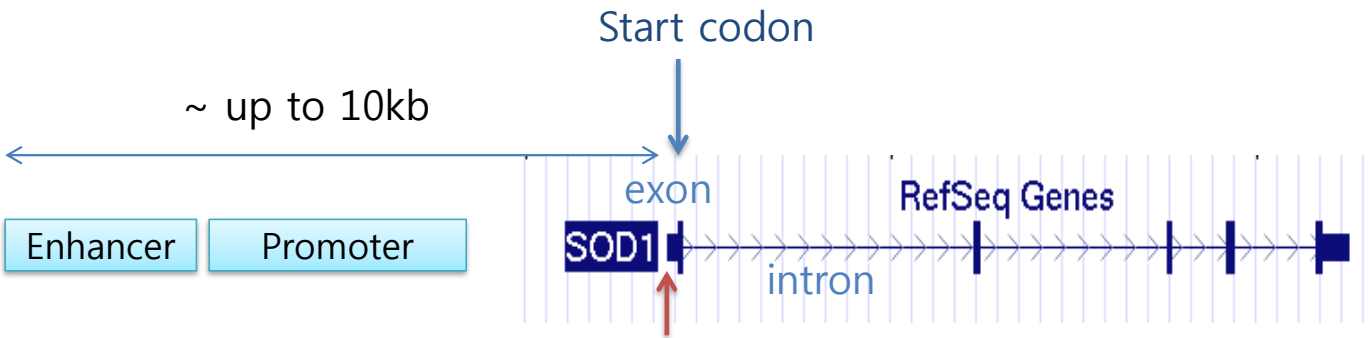
1. Biological function
2. Diseases
3. Quantitative trait /population

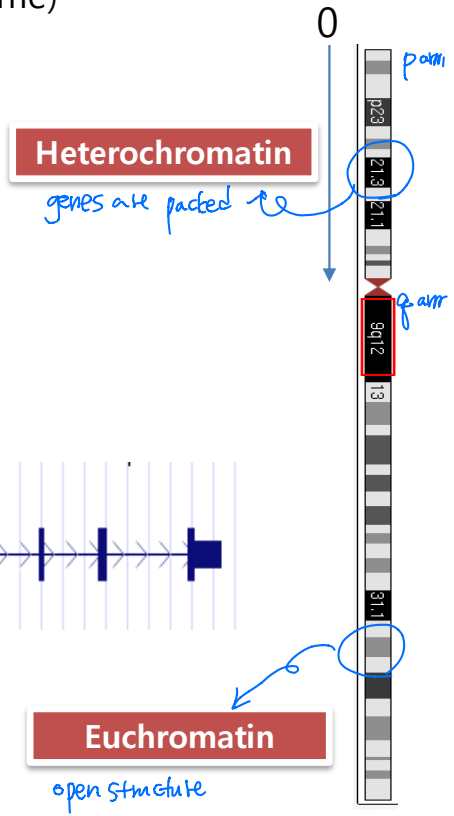
Genome and gene structure

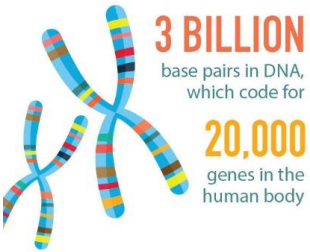
Genome: whole DNA sequences (Gene + Chromosome)

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

- Nucleosome > Chromatin > Chromosome







Human Genome Project (1990-2003)

http://genome.ucsc.edu/

Flow of genomics information

Phenotype = **genotype** + (environment + life history + **epigenetics**)

can affect

ex) chromatin methylation, str. of Cystein

★Central Dogma

Genome → Transcription → Transcriptome → Processing → Translation → Proteome

DNA ➡ Pre-mRNA ➡ mRNA ➡ protein

regulation point

1) initiation
2) elongation

3) Splicing / Poly (A)

4) Stability
shut down

5) Repression
on/off
repress expression

nucleus — **Complexity !** — cytoplasm

RNA polymerase II
Chromatin structure

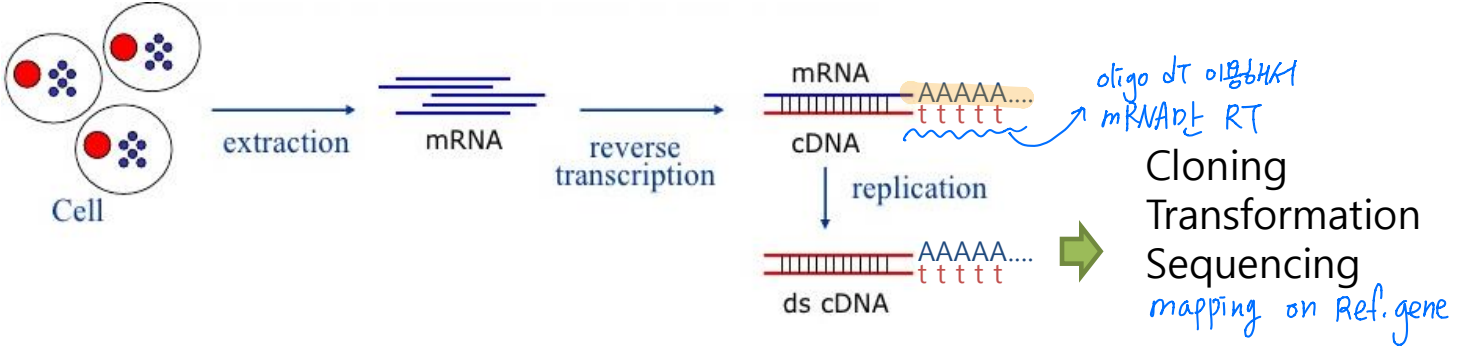
mRNA
Ribosome

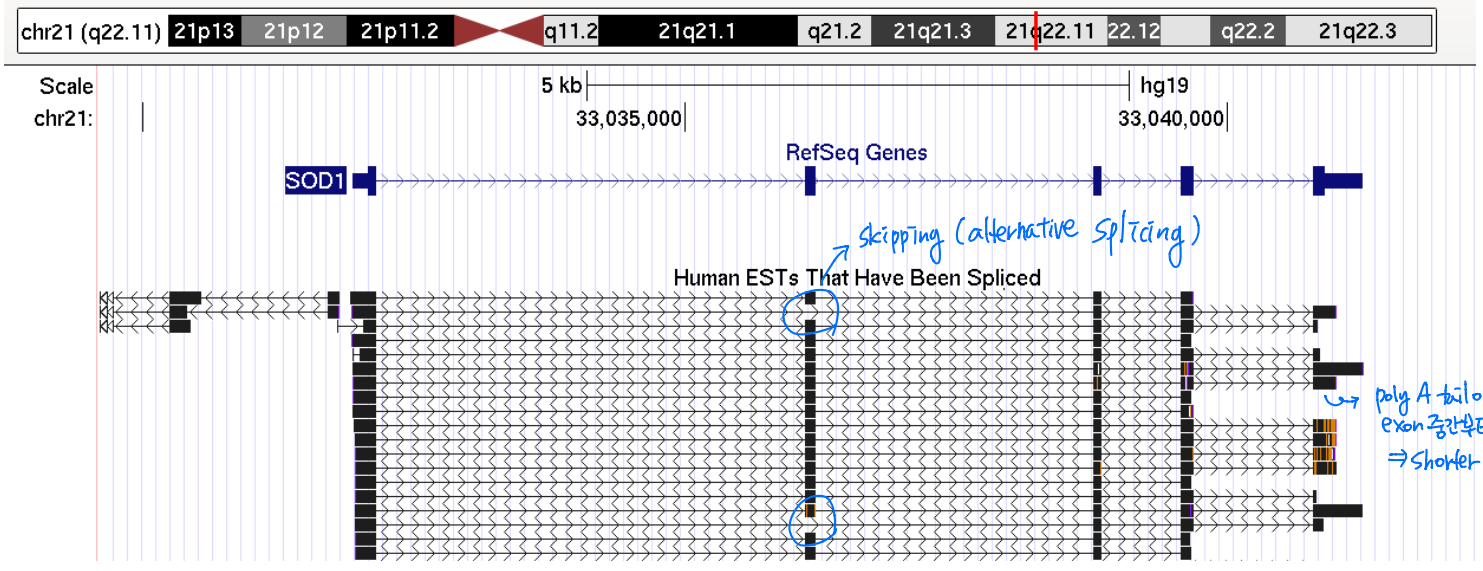
How to figure out expression? Expressed Sequence Tag (EST)

Gene expression

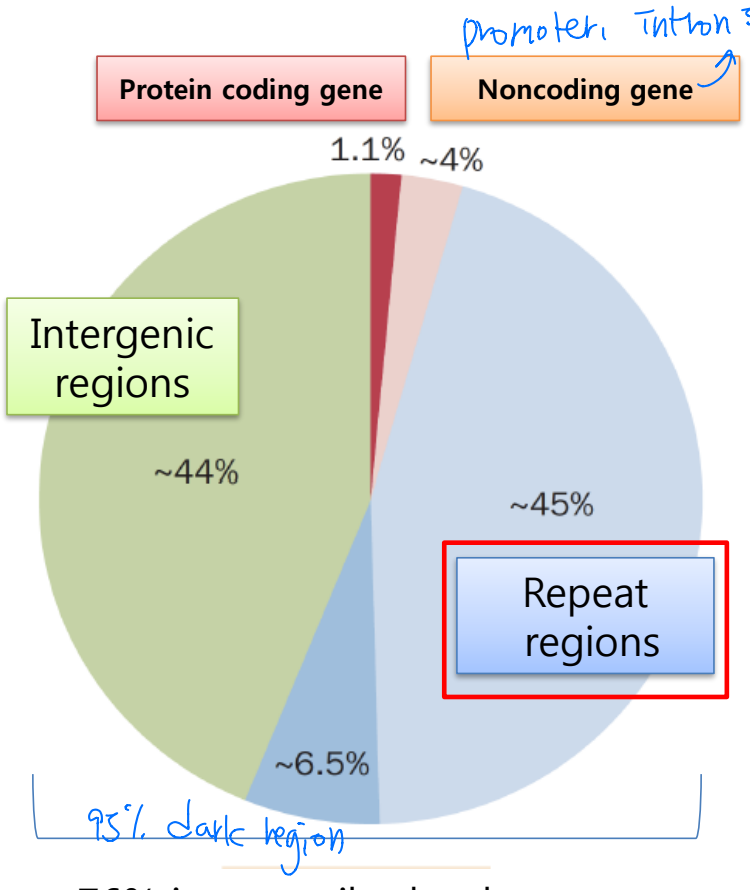
Genome → mRNAs

How?
EST clones





Human genome: ~50% is repeat sequence



~76% is transcribed at least once.

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

Retrotransposon

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"**

Majority of retrotransposon is LINE or SINE ⇒ Some of them are still active → can jump around

→ Cause genomic variation (disease)

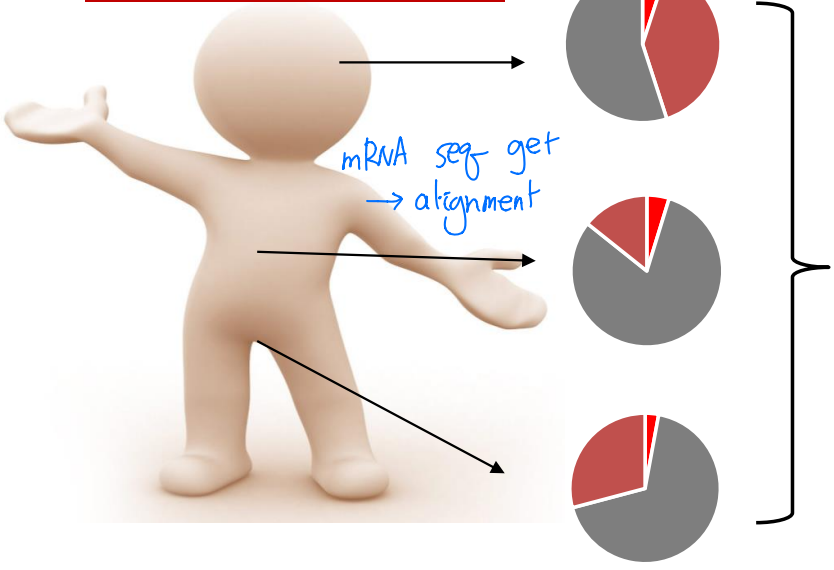
~75% of intergenic regions are transcribed !!!

ENCODE (Encyclopedia of DNA Elements) Projects

aim to sequence transcribed gene

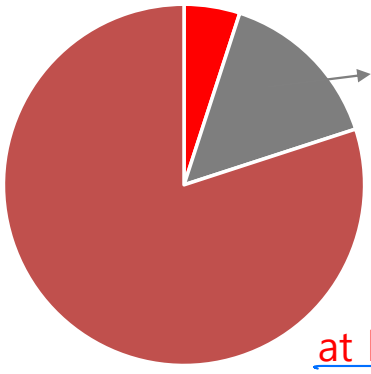
functional elements in the human genome
(regulatory elements, non-coding RNAs.....) (2012)

Coding gene



Non-coding genes

Human genome



at least once

~75%

Somehow made RNA in some cells

Importance of non-coding genes (RNAs) in human genome

Regulatory function ?

NGS & personal genomics

1990 - 2003

Human genome project
- Sequencing (3 billion bp)
- 13 years, 3 billion $

~2008

Next-generation sequencing
- 2 month, <1 million $

Now

NGS
- 1 week , ~1,000 $

2015~ (under develop)

Oxford Nanopore (MinION)

2017, Jan

- 2 days
~100 $
!!!!

Illumina (NovaSeq)

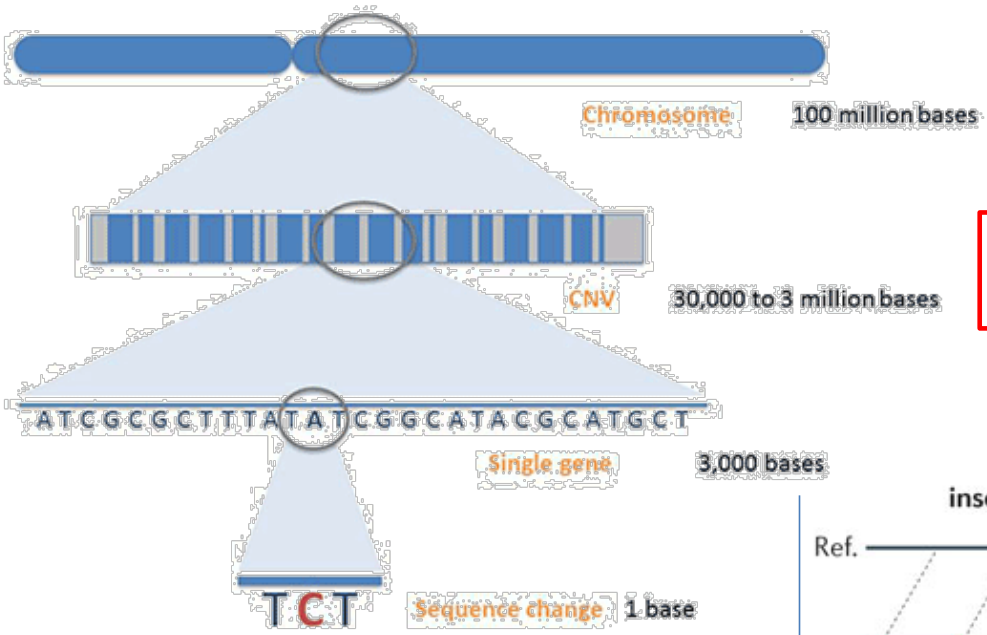
6 TB in 2 days

- Personal Genomics

- Single Cell Genomics

질병의 원인 밝히는 등에서 중요한 역할

Genomic Variation



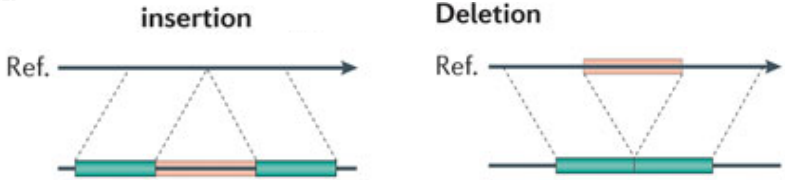
Structural variation

DNA recombination

→ meiosis 중 일어남

1. Gene duplication
(CNV: Copy number variation)

2. Large insertion / deletion



Sequence variation

⇒ phenotype 에 영향

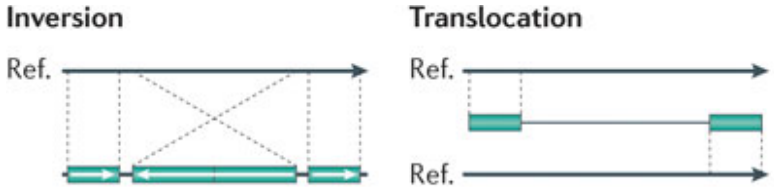
1. Single nucleotide change
: SNP (Single Nucleotide Polymorphism)

≠ mutation 이미 일어나 있는 것

→ 인종에 따라

2. Small nucleotide deletion or insertion
: Indel

3. Inversion, Translocation



Genome Projects

: Organization and Objectives

sequencing 없이 두 유전자의 거리를 알아내는 법

sequencing이 없던 시절엔 어떻게 genome을 알아냈는가

Sung Wook Chi

Division of Life Sciences, Korea University

What we will learn today

- **Genomics**

- The Core Aims of Genome Science

- **Mapping Genomes**

figure out location & distance w/o sequencing

- Genetic Maps, Physical Maps, Cytological Maps

genetic recombination

marker → order

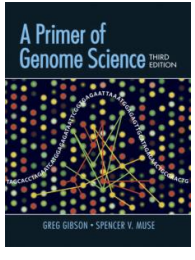
현미경통해 staining을 보고 band를 통해유추

- Comparative Genomics

- **The Human Genome Project**

- Objectives, Internet Resources

Details about human genome project will be on next lecture



Chapter 1. Genome Projects: Organization and Objectives

Genomics

Genome → Sequencing → phenotype

Mapping

- Identification of genes
- Location of genes
- Relation of genes

- Mapping
- Assembly 조립
- Annotation 주석

Variation

- Variation (sequence, structure)
- Comparative genomics

- Variant identification (SNP) (Translocation)
- Phylogenetic analysis

Functional genomics

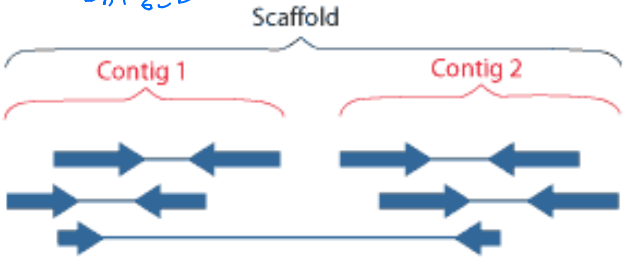
Gene expression analysis

Analysis (Bioinformatics)

The Core Aims of Genome Science

Mapping

<physical mapping>
큰조각들로 DNA를 먼저 잘라서 분석



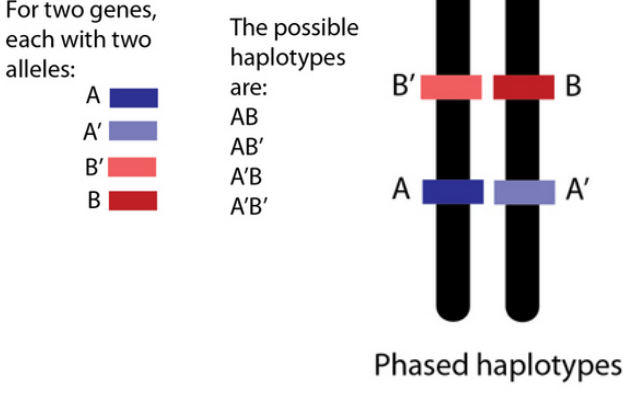
1. **Assembly of genetic and physical maps**
   - linkage, recombination frequency
   - contig, scaffold

2. **Mapping expressed gene sequences in the order of genome**
   - cDNA clones, ESTs

3. **Annotate complete set of genes**
   - Sequence search, annotation
   - ORF (open reading frame)

실제로 어떤 유전자가 어디있는지 확인

Linked (A and B') (A' and B)
⇒연관분석, genetic mapping



Variation

mapping 다음 variation 분석

4. **Sequence variation**
   - SNPs, haplotype, linkage disequilibrium

5. **Comparative genomics**
   - To provide the resources for comparison with other genomes (**synteny**)

The Core Aims of Genome Science

Functional genomics

6. Functional genomics Screening

- accumulate functional data, including biochemical and phenotypic properties of genes
  : genomics (*near-saturation mutagenesis, high-throughput reverse genetics*)
  : transcriptomics
  : proteomics, structural genomics)

7. Compiling gene expression

- SAGE (serial analysis of gene expression)
- transcription profiling (microarray, RNA-Seq)

Analysis

8. Bioinformatics (Computational genomics)

- To establish an integrated Web-based database and research interface

9. Systems biology (Integrated genomics)

- Understanding whole interactions of biological components (genes, proteins.....)
- Understanding emerging properties of network interactions : Network biology

What we will learn today

- **Genomics**

- The Core Aims of Genome Science

- **Mapping Genomes**

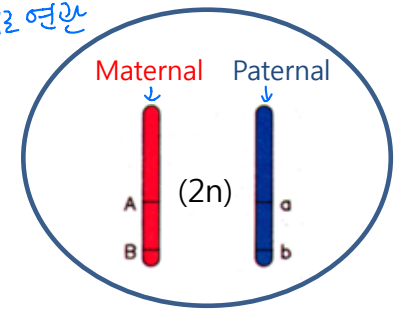
- Genetic Maps, Physical Maps, Cytological Maps
- Comparative Genomics

- **The Human Genome Project**

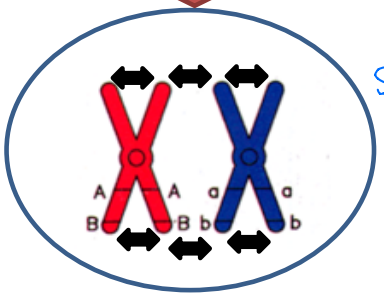
- Objectives, Internet Resources

Recombination: Genetic maps

location에 대한 정의 없을때
① 가까울 수록 연관
② 멀수록 멀리 연관
③ 독립

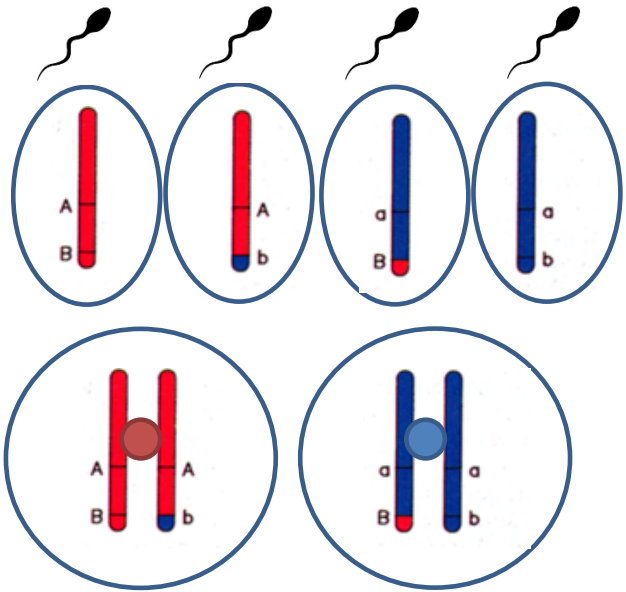


germ line cell
⇒ miosis 함



Sister chromatin 만듦

적도판에 배열하며 recombination → gene material 교환



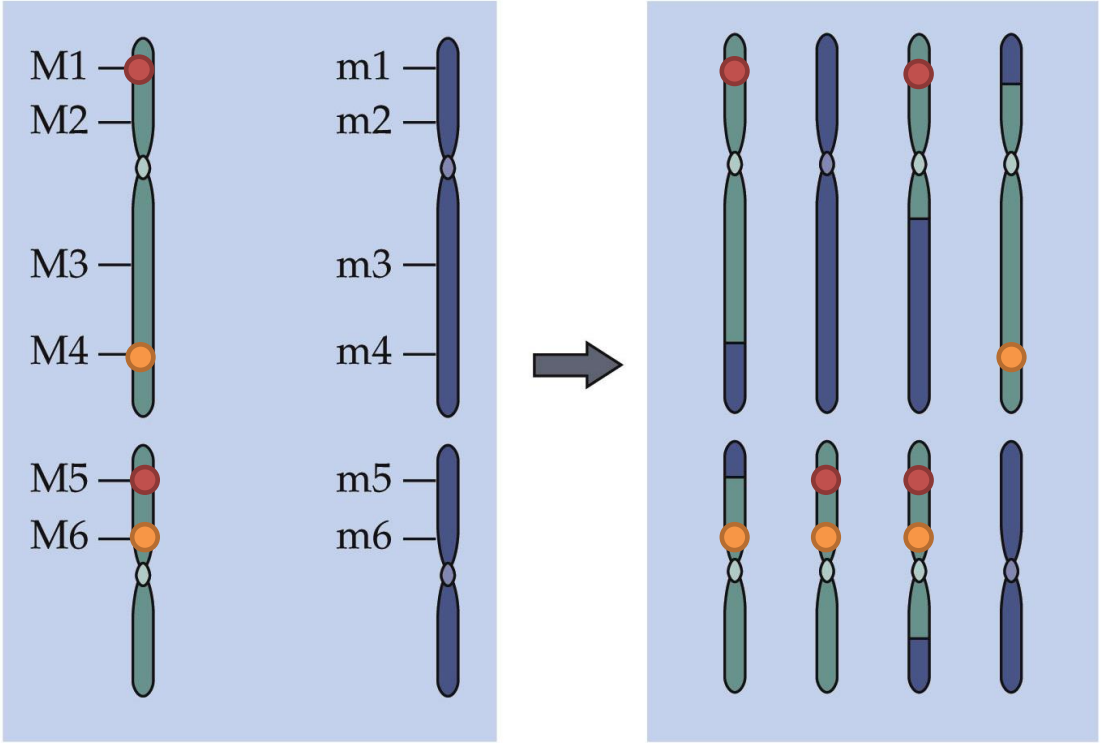
Recombination

Mapping Genomes : Genetic Maps

- **Genetic Maps**

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

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

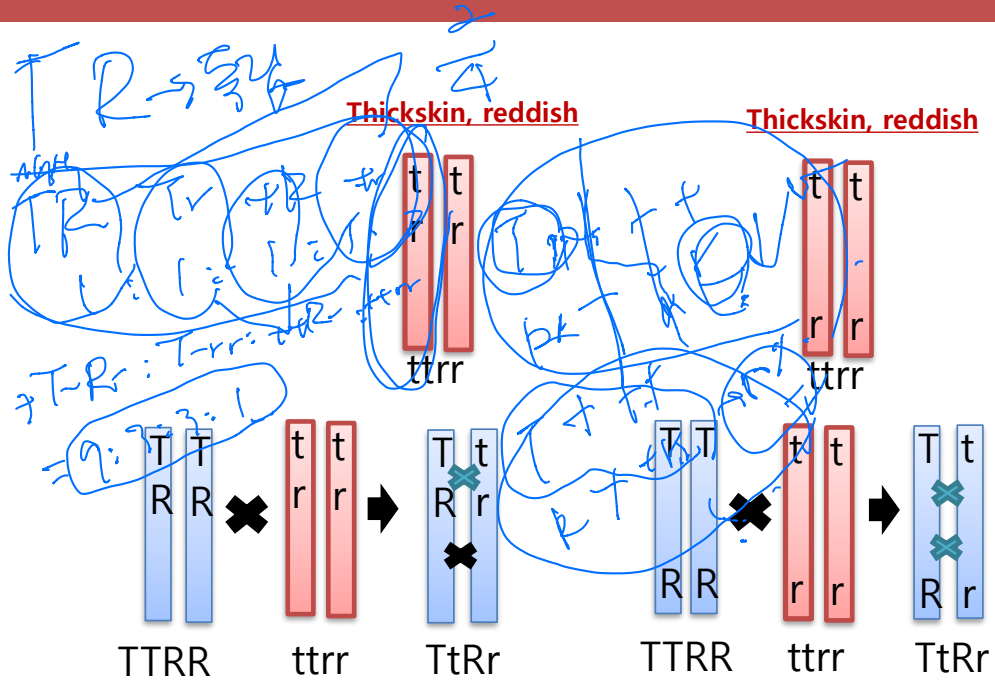


Located in different chromosome

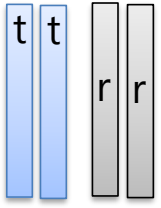
: 50% chance of co-segregation

독립된 경우

Genetic distance

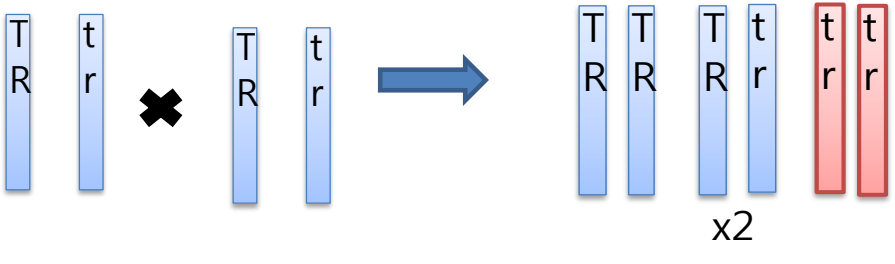


Thickskin, reddish



ttrr

: 50% chance
of co-segregation



Recombination rate !!!!

Genetic Maps

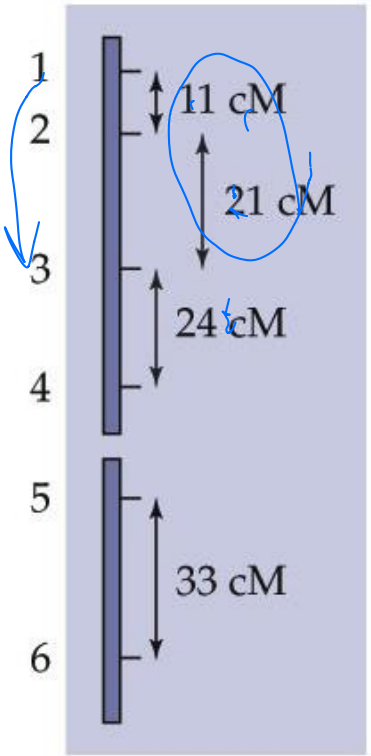
Linkage data -> Mapping function

(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						–

각각의 상관관계 염색체 거리를 측정

(C)



50cM – different chromosome

random packing 해서 두 유전자 이상했을 때

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.

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

Mendelian ratio 독립유전의 법칙

recessive (1/4) → 968 x 1/4 = 242

WT x Mut

ww mm

wm wm

F2

ww wm wm mm

Thick

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

Reddish = 249

Sour= 247

Petite = 242

$\frac{4 \times 3}{2}$ ⇒ 6개 조합

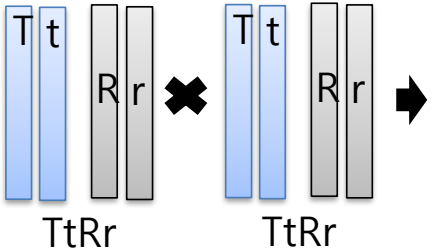
Normal	402	Thick	18
Petite	127	Red, sour, and petite	12
Sour	115	Red	11
Thick and red	108	Thick and petite	10
Thick, red, and petite	42	Thick, red, sour, and petite	8
Thick, red, and sour	41	Thick and sour	7
Sour and petite	38	Red and petite	3
Red and sour	24	Thick, sour, and petite	2

Total (n=968)

B/W two phenotypes (genes)

Unlinked
(Segregate independently)

Thickskin, reddish



968 x 1/4 x 1/4 = ~60

B/W two phenotypes (genes)

Unlinked
(Segregate independently)

독립이라는 가정했을때 멘델비율수치 ⇒ 이것을 기준으로 연관여부파악

968 x 1/4 x 1/4 = ~60

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

highly linked

likely linked

likely independent
or 멘델비율

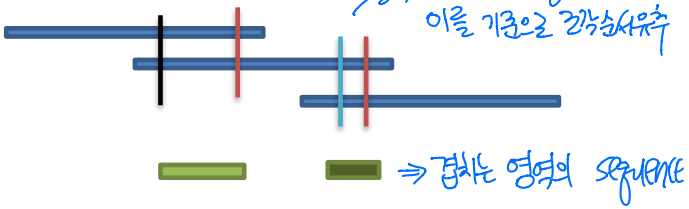
Mapping Genomes : Physical Maps

- **Physical Maps**

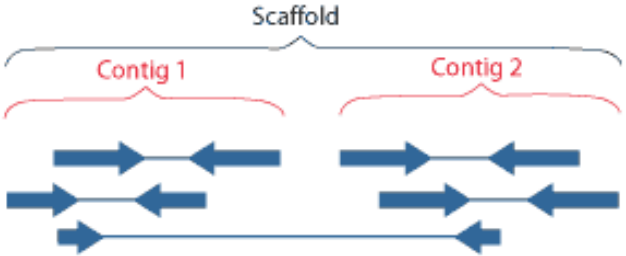
: an assembly of contiguous stretches of chromosomal DNA – **contigs**- in which the distance between landmark sequences of DNA is expressed in kilobases.

- Assemble contigs

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



2. Hybridization-based approaches
   : chromosome walking
   : **sequence-tagged sites (STSs)** single occurrence in the genome



Sequence-Tagged Sites (STSs) 서열표지부위

- Short (200 to 500 base pair) DNA sequence

Can be used as marker of location

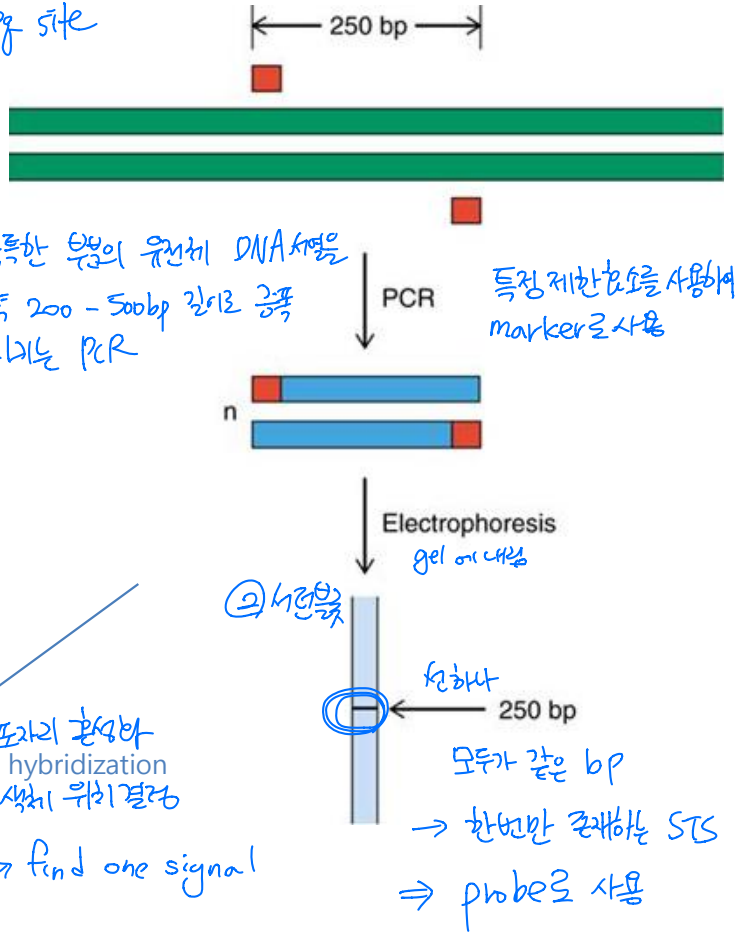
- Single occurrence in the genome whose location and base sequence are known.

Marker 역할을 해야하므로 딱 한번만 있는 seq 이다

- Detectable by PCR
- Design short primers

: Hybridize few hundred bp apart

: Amplify a predictable length of DNA

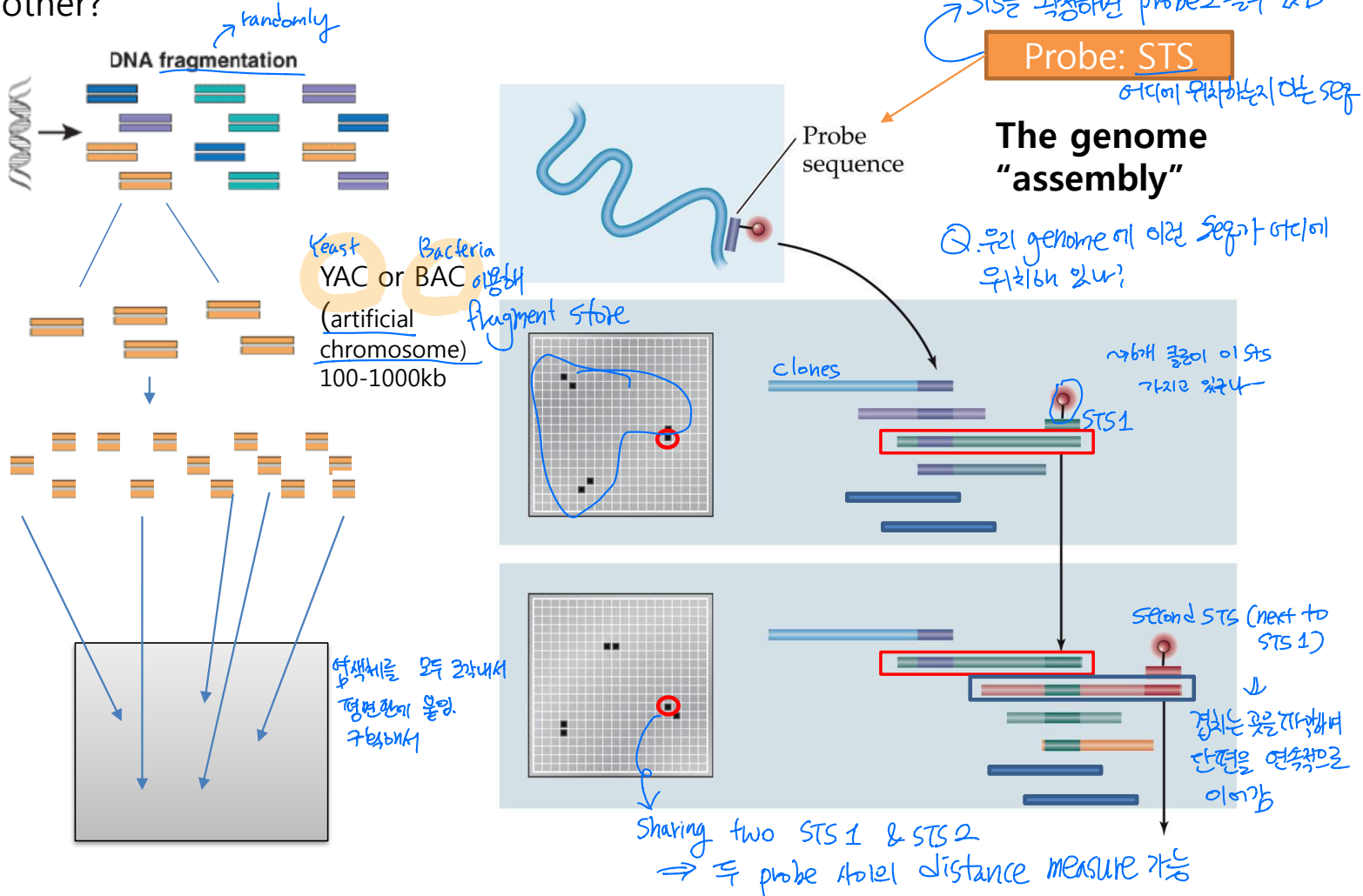


세포자리 분석법
In situ hybridization
⇒염색체 위치결정

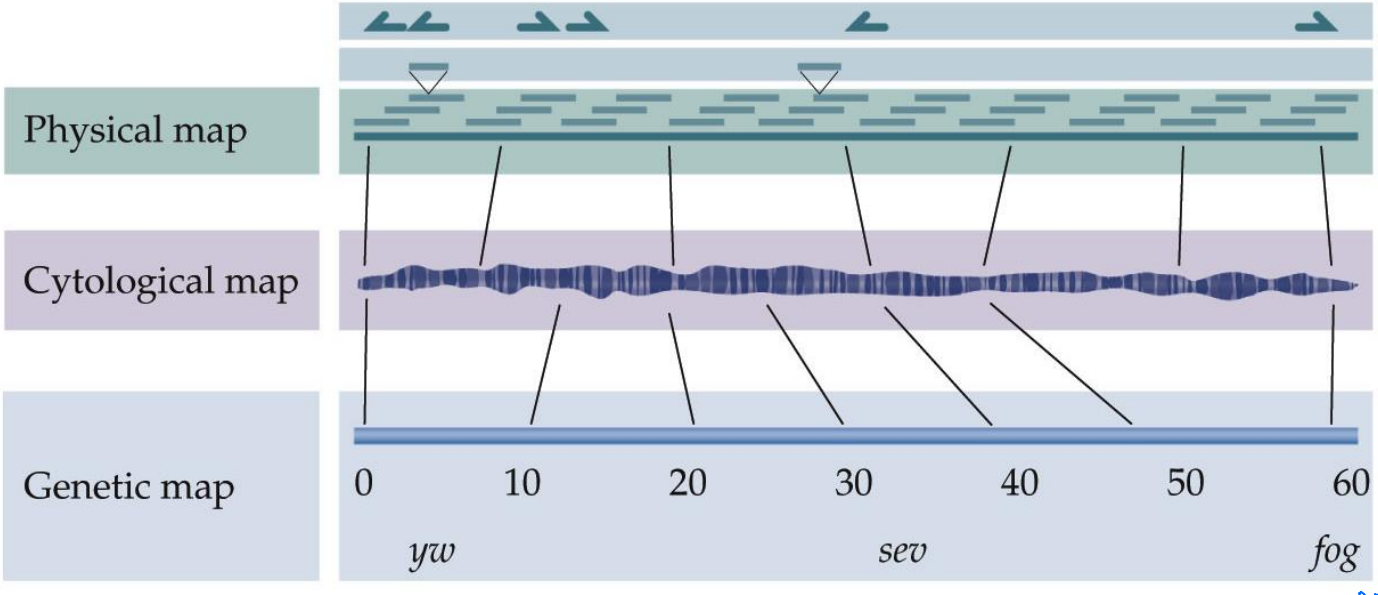
find one signal

Chromosome Walking

How are individual clones in a genomic library positioned relative to each other?



Mapping Genomes :Cytological Map

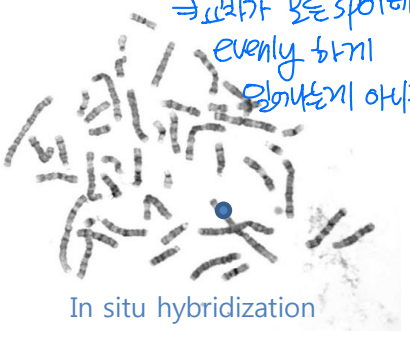


특정 probe가 chromosome에 가서 붙으면 어디 붙었는지 현미경으로 확인가능

Physical map, Cytological map and Genetic map are correlated,

But not the same..

⇒균일하게 evenly 하게 일어나는게 아니겠네



In situ hybridization

Comparative Genomics

- **Synteny** 두 종 이상의 유전자가 다른 종에 의해 공유되는 동일한 염색체에있는 경우

: conservation of gene order between chromosome segments of two or more organisms. or species

- **Homologs** ↗ some kind of genes have same sequence → same function

: orthologs, paralogs

1) Orthologs : homologs in different species (evolution) ↗ comparative genomics에서 관심

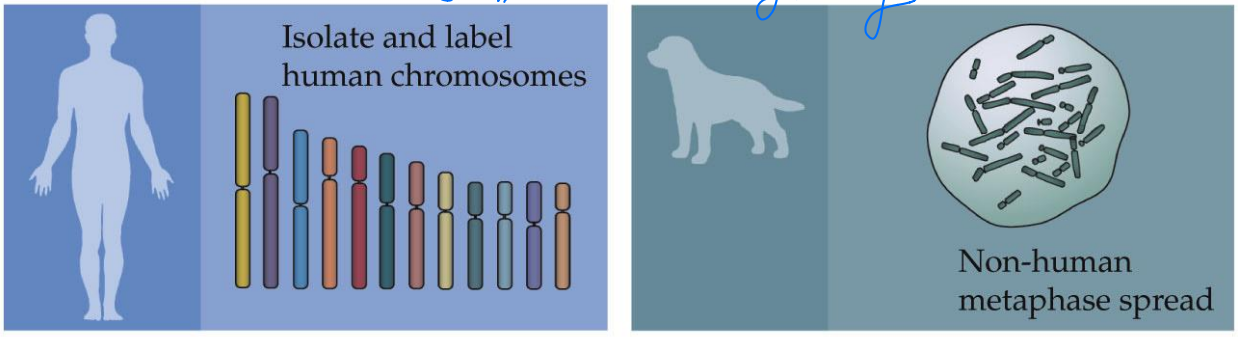
2) Paralogs : homologs within the same species (gene duplication)

- **Chromosome painting**

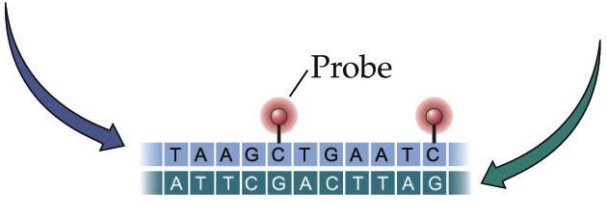
- Rearrangement of gene order:
  Chromosome fusions and splits, reciprocal translocations and inversions.
- each chromosome of one species is separately labeled with a set of fluorescent dye

Comparative Genomics: Chromosome painting

각 chr. 대표하는 STS probe를 design 할 수 있음 ⇒ 각 chromosome 대표하는 probe를 (or same color)
개의 유전체에 붙여봄 ⇒ 연관성이 synthesy



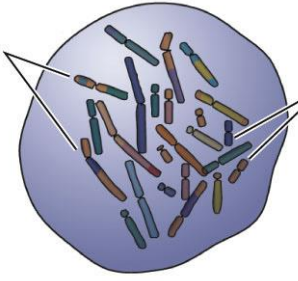
Different color combination for each chromosome probe



Hybridize with fluorescent probe

FISH: fluorescence in situ hybridization

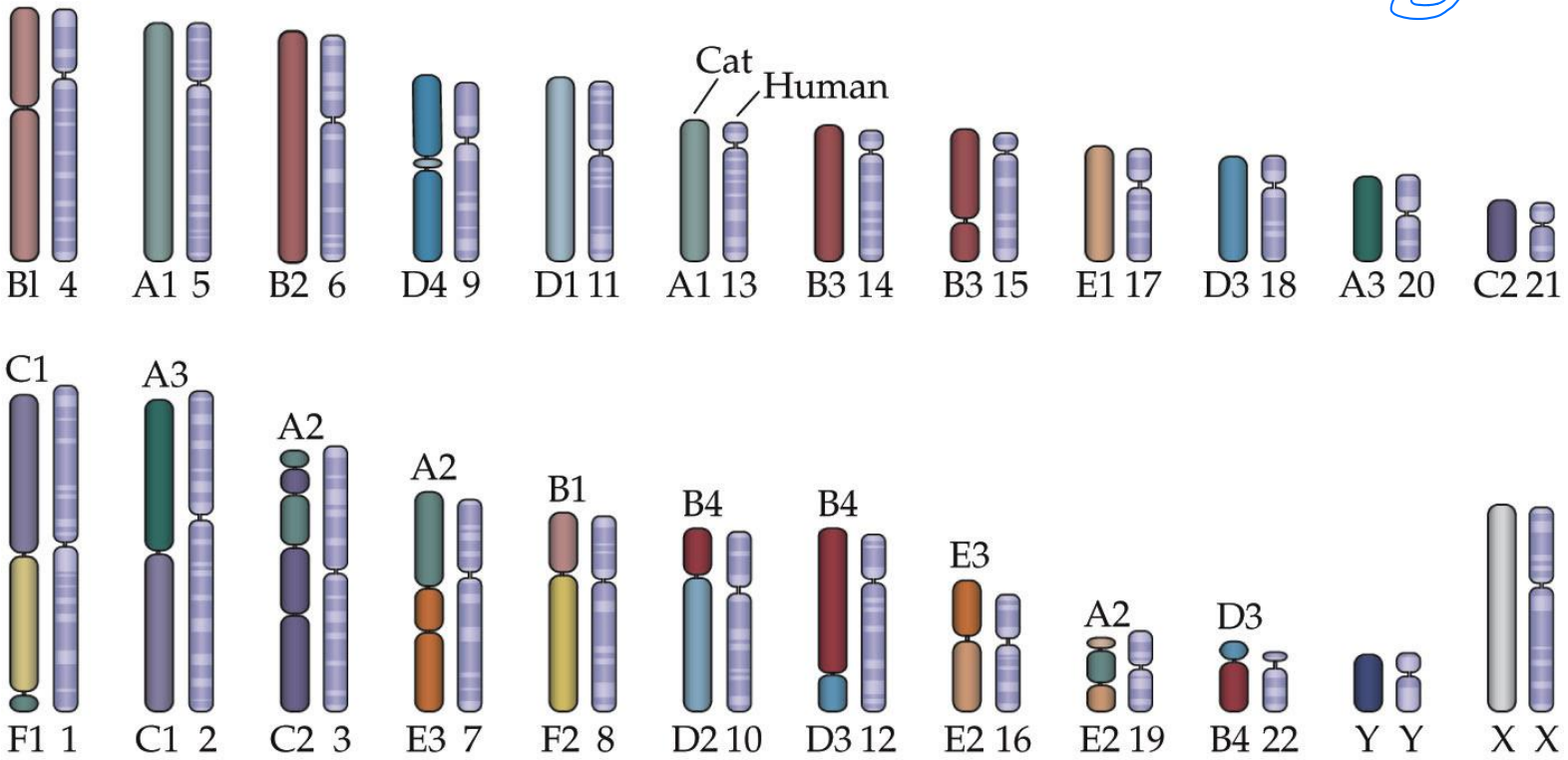
Multicolored chromosomes indicate breakage/fusion events



Single-color chromosomes indicate complete correspondence between species

Synteny b/w cat & human genomes

Old school: Chromosome painting (Labelled probe)



New school: Genome sequence alignment across multi-species

What we will learn today

- **Genomics**

- The Core Aims of Genome Science

- **Mapping Genomes**

- Genetic Maps, Physical Maps, Cytological Maps
- Comparative Genomics

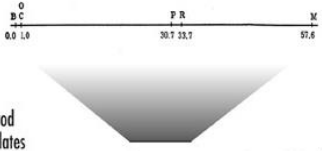
- **The Human Genome Project**

- Objectives, Internet Resources

Human Genome Projects : History

Mendel discovers laws of genetics
1865

Rediscovery of Mendel's work
1900



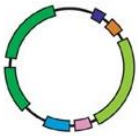
Garrod formulates the concept of human inborn errors of metabolism
1905

Sturtevant makes the first linear map of genes
1913

Avery, McLeod, and McCarty demonstrate DNA is the hereditary material
1944

Watson and Crick describe the double helical structure of DNA
1953

Nirenberg, Khorana and Holley determine the genetic code
1966



Cohen and Boyer develop recombinant DNA technology
1972

Mendel's law

Genetic map

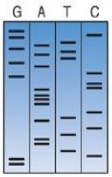
DNA

Genetic code

Recombinant DNA

huntington 병과 관련된 염색체 위치로 detect

first 시도 to figure out



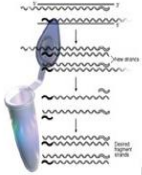
Issuing of Belmont Report on the use of human subjects in research
1974

Sanger and Maxan & Gilbert develop DNA sequencing methods
1977

GenBank database established
1982

First human disease gene mapped with DNA markers --Huntington disease
1983

First public discussion of sequencing the human genome
1984



PCR invented
1985

International Nucleotide Sequence Database Consortium formed

Muscular dystrophy gene identified by positional cloning

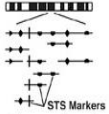
First automated DNA sequencing instrument developed
1986

First-generation human genetic map developed
1987

National Research Council (U.S.) issues report on *Mapping and Sequencing the Human Genome*

Development of yeast artificial chromosome (YAC) cloning

Human Genome Organization (HUGO) formed
1988



Sequence-Tagged Sites (STS) mapping concept established

Cystic fibrosis gene identified by positional cloning
1989

HGP

1990

2003

Sanger sequencing

GenBank

PCR

Human genetic map

YAC

STS

STS 기술의 genome project의 지대한 영향 기억

Human Genome Projects : History

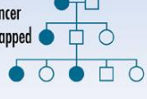
HGP

figure out order of fragment

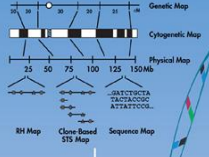
1990
Human Genome Project (HGP) launched in the U.S.

Ethical, Legal, and Social Implications (ELSI) programs founded at NIH and DOE

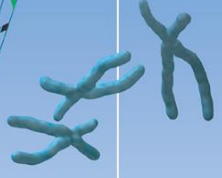
First gene for breast cancer (BRCA1) mapped



1991
First U.S. Genome Centers established



1992
Second-generation human genetic map developed



Rapid data release guidelines established by NIH and DOE

1993
New five-year plan for the HGP in the U.S. published

Sanger Centre founded (later renamed Wellcome Trust Sanger Institute)

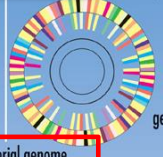
The Wellcome Trust

1994
HGP's human genetic mapping goal achieved

Human genetic map

order

1995
Human Physical map



First bacterial genome (*H. influenzae*) sequenced

virus seq. 성공

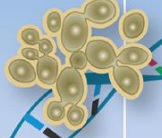
U.S. Equal Employment Opportunity Commission issues policy on genetic discrimination in the workplace

1996
First human gene map established

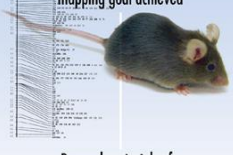
Pilot projects for human genome sequencing begin in U.S.

First archael genome sequenced

Yeast (*S. cerevisiae*) genome sequenced



HGP's mouse genetic mapping goal achieved



Bermuda principles for rapid and open data release established

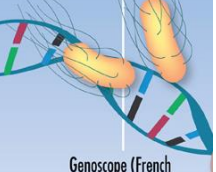
1997
DOE forms Joint Genome Institute

JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

NCHGR becomes NHGRI

National Human Genome Research Institute

E. coli genome sequenced



Genoscope (French National Genome Sequencing Center) founded

Human Genome Projects : History

HGP

Sanger Seq

1998

Incorporation of 30,000 genes into human genome map

New five-year plan for the HGP in the U.S. published

RIKEN Genomic Sciences Center (Japan) established

Roundworm (*C. elegans*) genome sequenced

SNP initiative begins

GTGCT
GTCCT

Chinese National Human Genome Centers (in Beijing and Shanghai) established

1999

Full-scale human sequencing begins

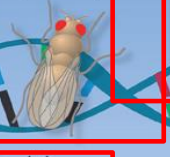
Sequence of first human chromosome (chromosome 22) completed

2000

Draft version of human genome sequence completed

President Clinton and Prime Minister Blair support free access to genome information

Fruit fly (*D. melanogaster*) genome sequenced



Mustard cress (*A. thaliana*) genome sequenced

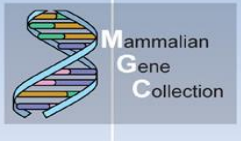
Executive order bans genetic discrimination in U.S. federal workplace

2001

Draft version of human genome sequence published

Draft

10,000 full-length human cDNAs sequenced



2002

Draft version of mouse genome sequence completed and published

Draft version of rat genome sequence completed

Draft version of rice genome sequence completed and published

2003

Finished version of human genome sequence completed

HGP ends with all goals achieved

to be continued..

NGS

The Human Genome Projects : Objectives

TABLE 1.1 ***Initial Goals of the Human Genome Project*** (Part 1)

From the First 5-Year Plan: 1993–1998

1. THE GENETIC MAP
   - Complete 2 to 5 cM map by 1995
   - Develop new technology for rapid and efficient genotyping
2. THE PHYSICAL MAP
   - Complete STS map to 100 kb resolution
3. DNA SEQUENCING
   - Develop approaches to sequence highly interesting regions on Mb scale
   - Develop technology for automated high throughput sequencing
   - Attain sequencing capacity of 50 Mb per year; sequence 80 Mb by 1998
4. GENE IDENTIFICATION
   - Develop efficient methods for gene identification and placement on maps
5. TECHNOLOGY DEVELOPMENT
   - Substantially expand support for innovative genome technology research
6. MODEL ORGANISMS
   - Finish STS map of mouse genome to 300 kb resolution
   - Obtain complete sequence of biologically interesting regions of mouse genome
   - Finish sequences of *E. coli* and *S. cerevisiae* genomes
   - Substantial progress on complete sequencing of *C. elegans* and *D. melanogaster*

The Human Genome Projects :objectives

TABLE 1.1 ***Initial Goals of the Human Genome Project*** (Part 2)

From the First 5-Year Plan: 1993–1998

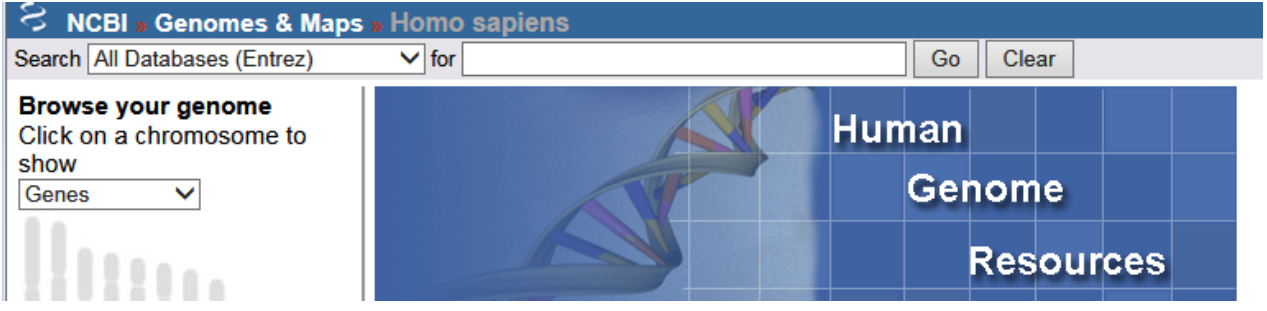
7. INFORMATICS
   - Continue to create, develop, and operate databases and database tools
   - Consolidate, distribute, and develop software for genome projects
   - Continue to develop tools for comparison and interpretation of genome information

8. ETHICAL, LEGAL, AND SOCIAL IMPLICATIONS (ELSI)
   - Continue to identify and define issues and develop policy options
   - Develop and disseminate policy regarding genetic testing
   - Foster greater understanding of human genetic variation
   - Enhance public and professional education programs on sensitive issues

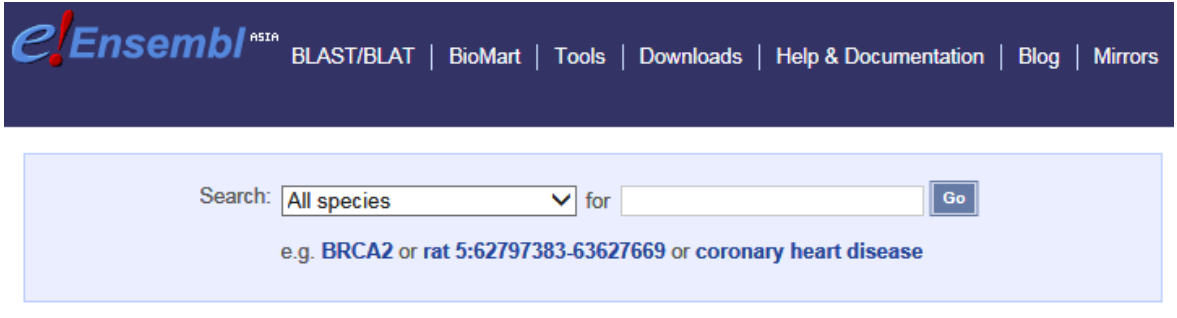
9. OTHER
   - Training of interdisciplinary genome researchers
   - Technology transfer into and out of genome centers
   - Outreach

How to access genome sequencing data

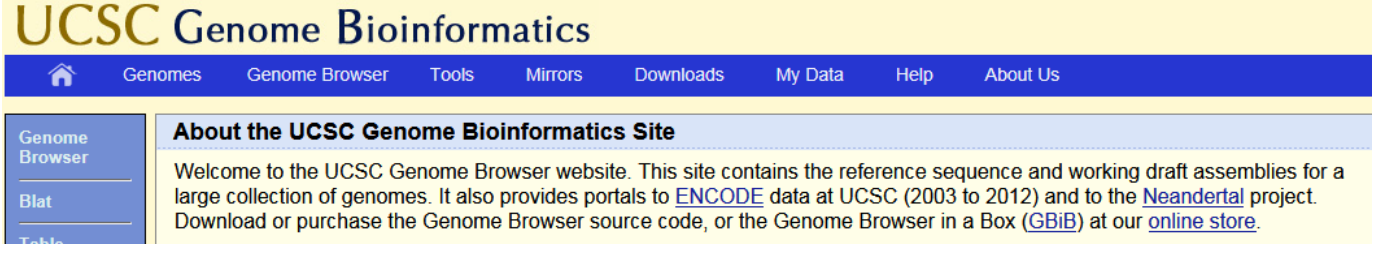
1. NCBI Genome



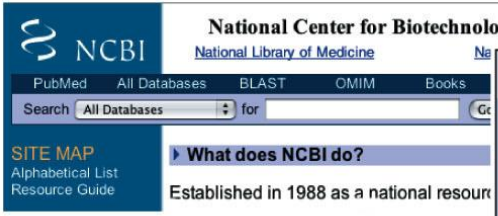
2. Ensembl



3. UCSC Genome



Internet Resources



NCBI
(National Center for Biotechnology Information)

http://www.ncbi.nlm.nih.gov

- Online Mendelian Inheritance in Man
- Map viewer
- Basic Local Alignment Search Tool
- Links to literature
- Entrez data-mining tools
- Cancer Genome Anatomy Project
- Gene annotation
- GenBank

Online Mendelian Inheritance in Man (OMIM) Web site

http://www.ncbi.nlm.nih.gov

- Gene Map (cytogenetic location of human Mendelian loci)
- Genes and Disease
- FAQs and Statistics
- Morbidity Map: Alphabetical listing of diseases and corresponding loci
- Allied Resources

cytosol information, disease

OMIM

Use genome browser to examine a human disease gene

특정질병관련 유전자를 web에서 찾기

EXERCISE 1.2 *Use the NCBI and Ensembl genome browsers to examine a human disease gene*

Choose a human disease of interest to you, and then use the OMIM site to identify a gene that is implicated in the etiology of the disease. Then use the NCBI, UCSC, and Ensembl genome browsers to answer the following questions about the gene:

Asthma / IL13

a. *What are the various identifiers (aliases) for your gene?*
b. *Where is the gene located on the chromosome (cytologically and physically)?*
c. *What is the Reference Sequence (RefSeq) for the gene?*
d. *How many exons are there in the major transcript, and how long is it?*
e. *What is known about the function of the gene?*
f. *Do the three annotations agree? Which browser do you prefer, and why?*

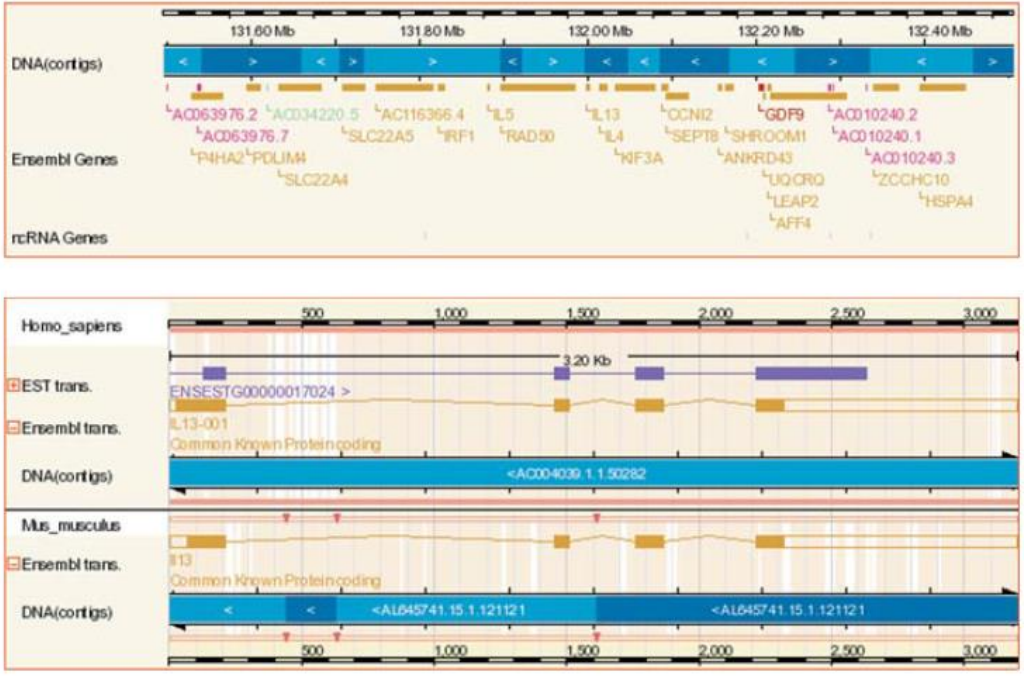
Compare the structure of a gene in a mouse and a human

EXERCISE 1.3 *Compare the structure of a gene in a mouse and a human*

Using either the NCBI or Ensmbl Browser, explore the structure of the gene you used in Exercise 1.2 in a mouse and a human (and if possible, in other vertebrates).

IL13

genome 정보찾아서 보기

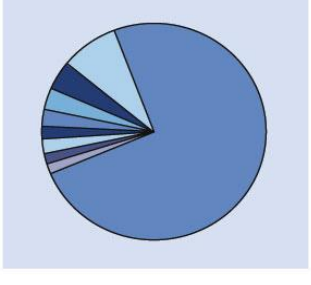


Human genome project race

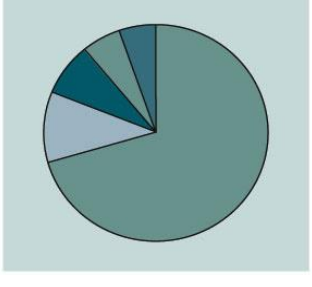
IHGSC	Celera
Attract >50 ethnically diverse volunteer donors	Attract 21 ethnically diverse volunteer donors
Obtain informed consent	Obtain informed consent
Construct PAC/BAC libraries; assess quality	Establish permanent cell lines and 2-, 10-, and 50-kb libraries
Choose 8 libraries for assembly of physical map: All male for first draft, ethnic identities unknown	Choose 5 libraries for sequencing: 2 male, 3 female; 1 African, 1 Asian, 1 Hispanic, 2 Caucasian
Hierarchical sequencing	Shotgun sequencing

Francis Collins

International Human Genome Sequencing Consortium



Vs.



Craig Ventor

Celera

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