Summary

: GWAS, Phylogenetic analysis

Sung Wook Chi
Division of Life Sciences, Korea University

Overview of genome-wide association study (GWAS)

all type variation 과 phenotype 의 연관성을 분석

Sample design / Collection

WGS

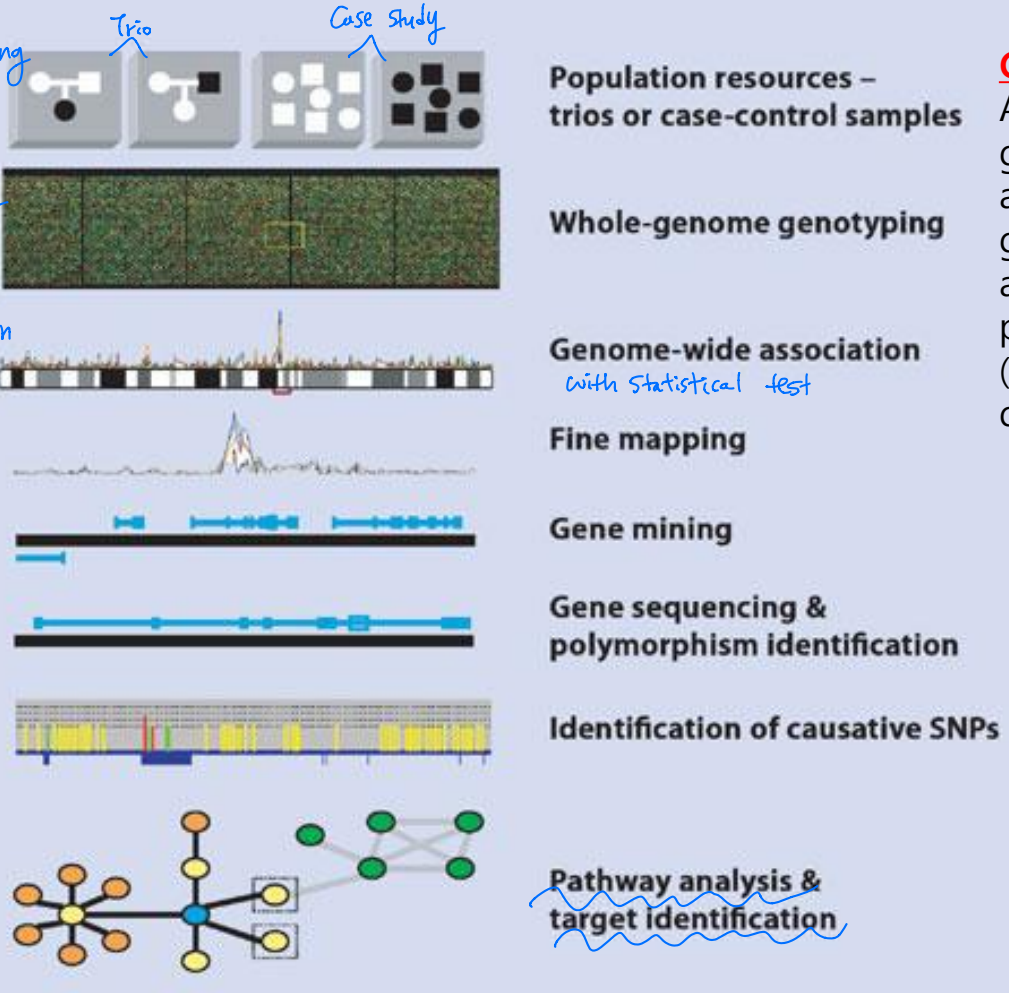
GWAS

Associated Loci QTL

Variation & Gene

Causative variation ?

Functional interpretation validation



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

Interpretation of phylogenetic tree: Orthologs and paralogs

How these diversity were generated through evolution event.

Variant sequences

↓

Sequence Alignment

↓ Calculate distance

Phylogenetic tree

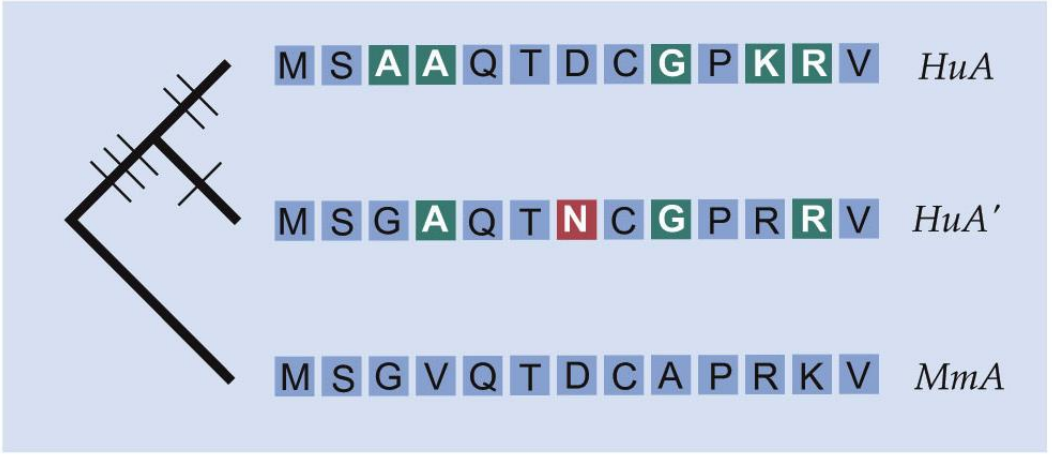
- Recalculate / Infer
- Distance or similarity b/w sequences

- Topology (order)
- Length (evolutional time)

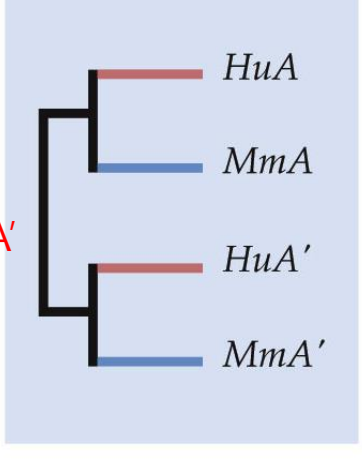
↓

Evolutional relationship / interpretation

(A)

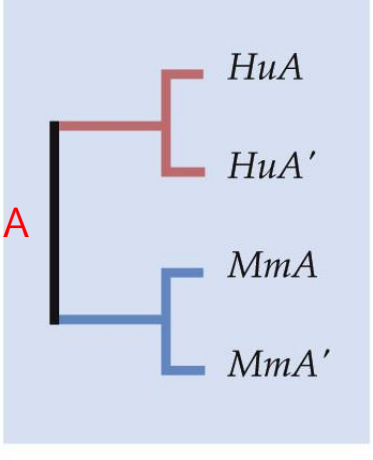


(B)



A, A′

(C)



A

: Whole genome sequencing (NGS)

1. Whole genome sequencing (WGS by NGS)
2. Variation Analysis for NGS data
3. Genome-wide association study (GWAS)
4. Phylogenetic analysis

Functional Genomics

: Exome-Seq and functional genomics

Sung Wook Chi
Division of Life Sciences, Korea University

Functional Genomics

-**Functional Genomics :** Studying functions of genes using massive data
1) genome-wide methods
2) Gene expression (transcription, translation, protein-protein interaction

*다른 gene : functional genetics

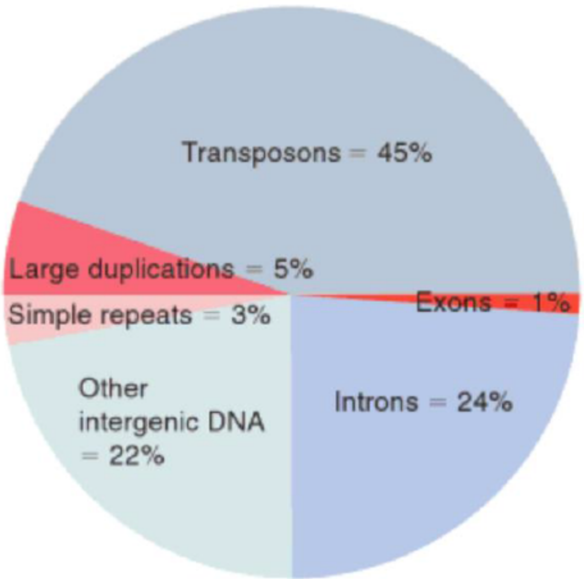
Variation → **GWAS** → Diseases, Quantitative trait

NGS
- WGS
- **Exome-Seq**

Whole Exome Sequencing (WES)

Whole Exome Sequencing, Why?

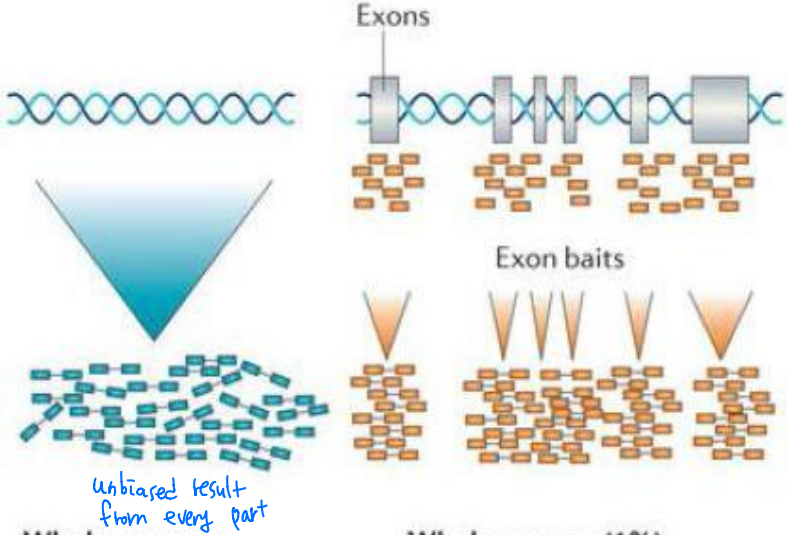
- Focuses on the part of the genome we understand best, the exons of genes
- Exomes are ideal to help us understand its relationship to phenotype.



- Capturing and sequencing the **~1% of the human genome** that codes for protein sequences
- On average, whole-exome sequencing identifies **12,000 variants** in coding regions.
- **~90%** are found in publicly available databases.

- A whole exome is 1/6 the cost of whole genome and 1/15 the amount of data

Whole genome vs. Whole exome sequencing



Target Size:	3Gb	30Mb
Scale:	1	0.01

File size for human germline WGS (30X)

- Image Data 16TB
- BaseCall/Quality score data 200GB
- Final Alignment output 1 TB

Exome Sequencing Pipeline

Sample DNA Fragmentation

Illumina Library Preparation

Exome Enrichment ⇒ WGS에는 없는 과정

Cluster Generation

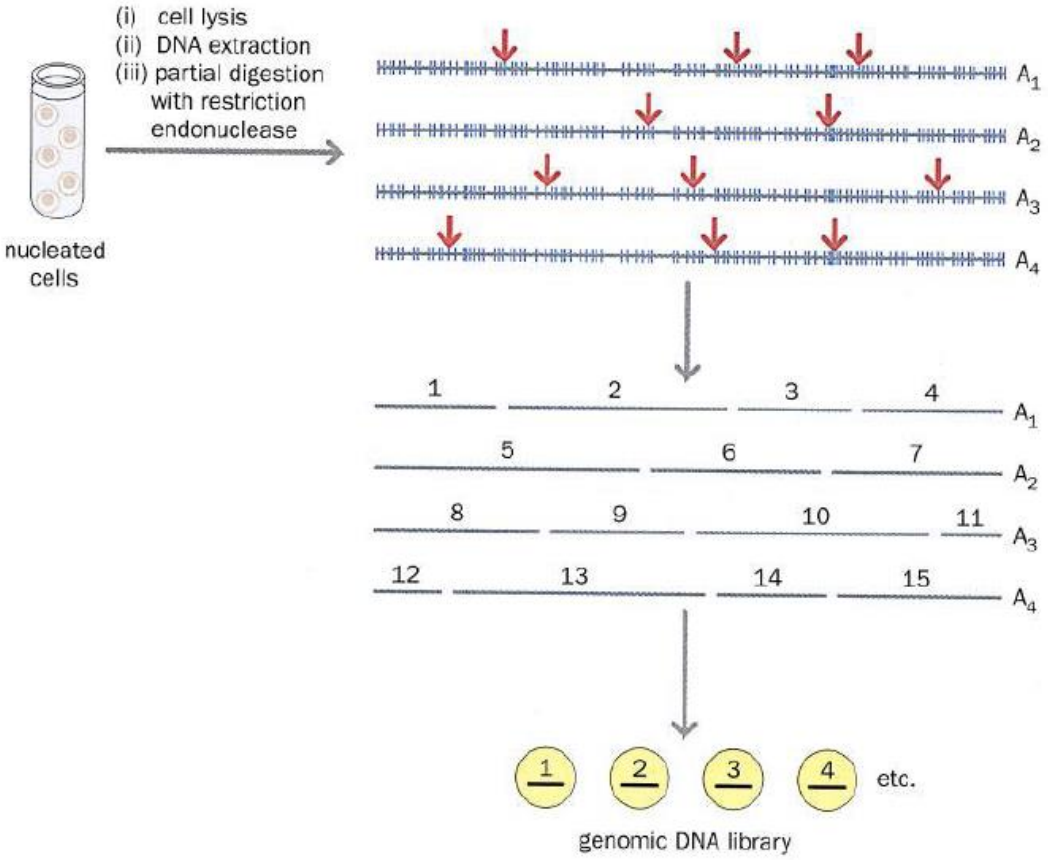
NGS

1. • **Analysis of raw data (images)**
2. • **Base Calling: Determination of sequences**
3. • **Mapping: Alignement to reference genome** variation detection
4. • **Variant calling:**
   - Substitutions
   - Indels

~30 000 to 50 000 variants /person

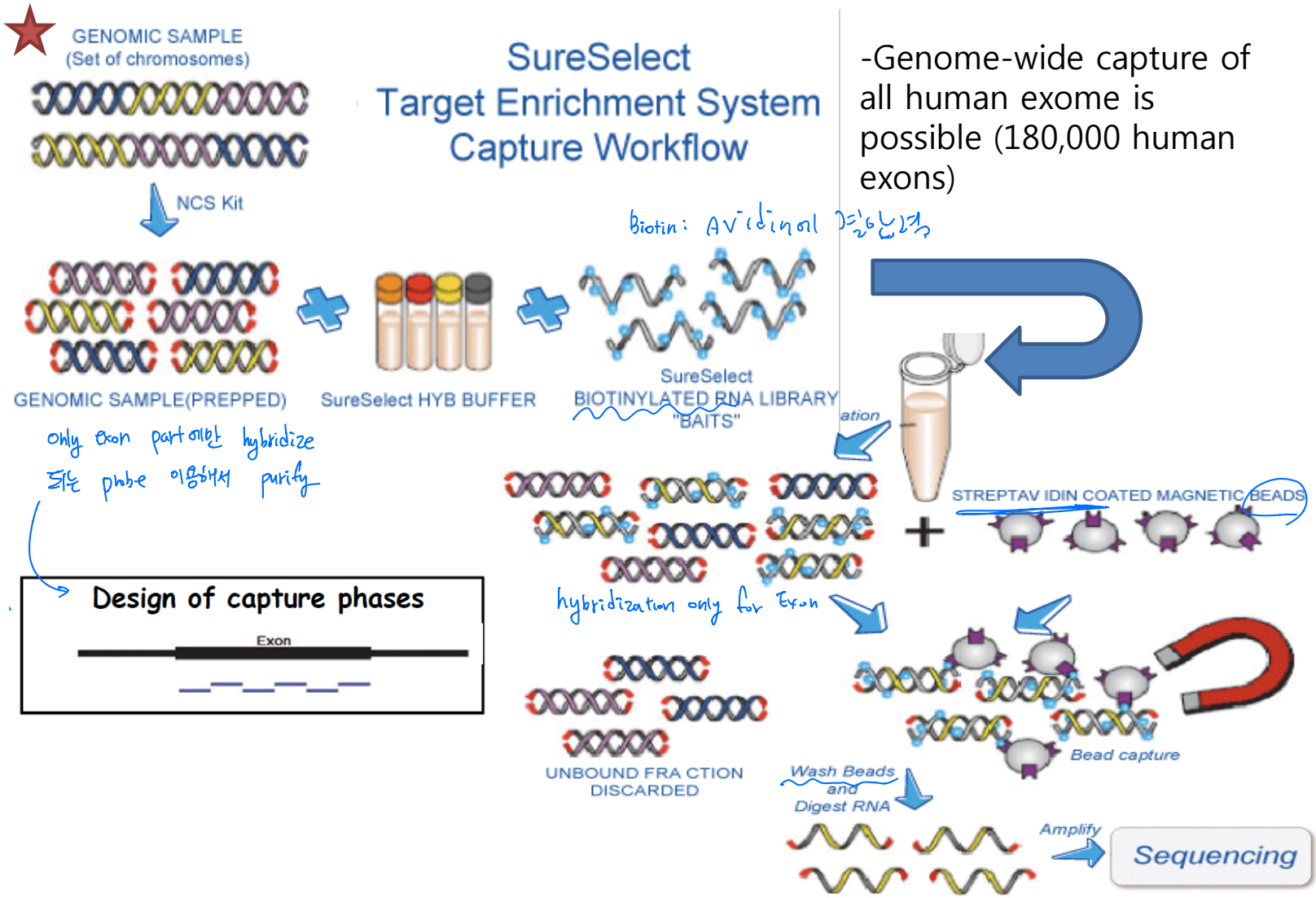
Filtering steps (artefacts, known variants,…)

Genomic DNA fragmentation (DNA libraries)



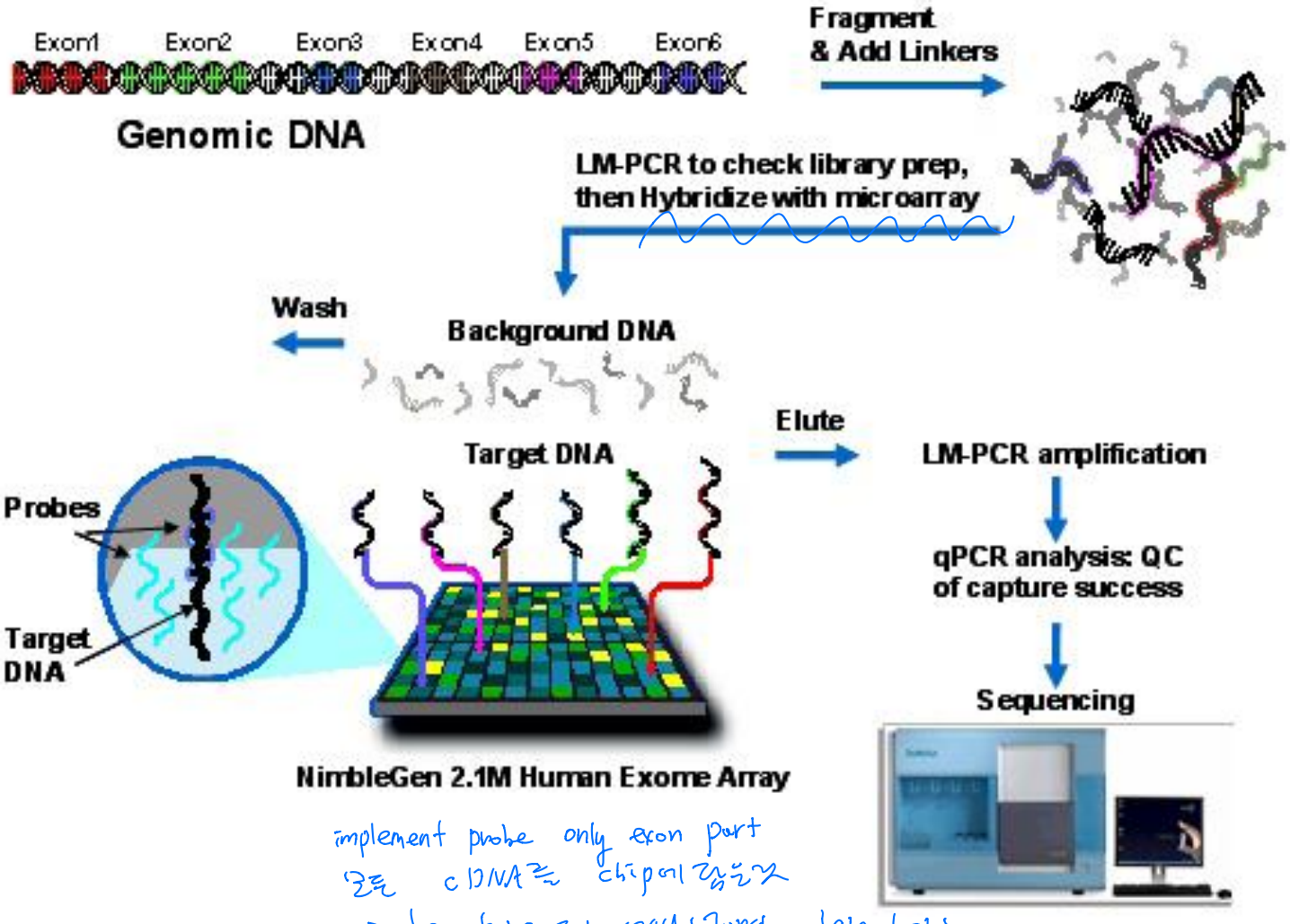
Exome capture, enrichment

Exome-capture (Agilent)



-Genome-wide capture of all human exome is possible (180,000 human exons)

Exome-capture (Nimblegen)



implement probe only exon part
근데 cDNA는 chip에 갈수가
⇒ denature 된 DNA 상태면 detect가능

Exome Sequencing

nature International weekly j

Access
To read this story in full you will need to login or make a payment

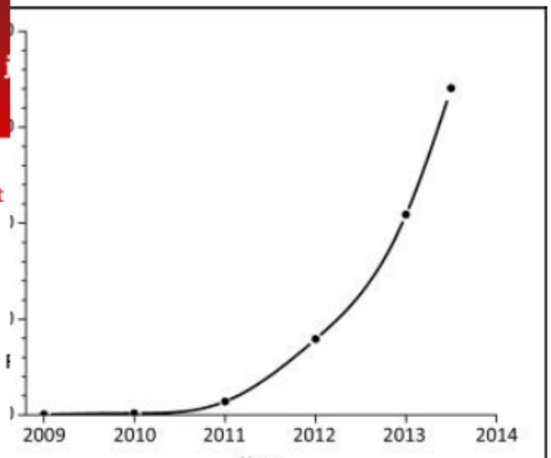
nature.com > Journal home > Table of Contents

Letter

Nature **461**, 272-276 (10 September 2009) | doi:10.1038/nature08250; R
June 2009; Published online 16 August 2009

Targeted capture and massively parallel sequencing of 12 human exomes

Sarah B. Ng[1], Emily H. Turner[1], Peggy D. Robertson[1], Steven D.
Abigail W. Bigham[2], Choli Lee[1], Tristan Shaffer[1], Michelle Wong[1]
Bhattacharjee[4], Evan E. Eichler[1,3], Michael Bamshad[2], Deborah A
Nickerson[1] & Jay Shendure[1]



Exon을 잡아내는 방법이 빠르게 증가

Advantages	Disadvantages
-Higher sequence coverage and less raw sequence and cost than WGS. -Point mutations -Small indels -CNAs (need further validation) Small indels	-Incomplete exome capturing -Limit to mutations in coding regions -Lack efficient methodology to detect structural variations

avoid noise from contamination

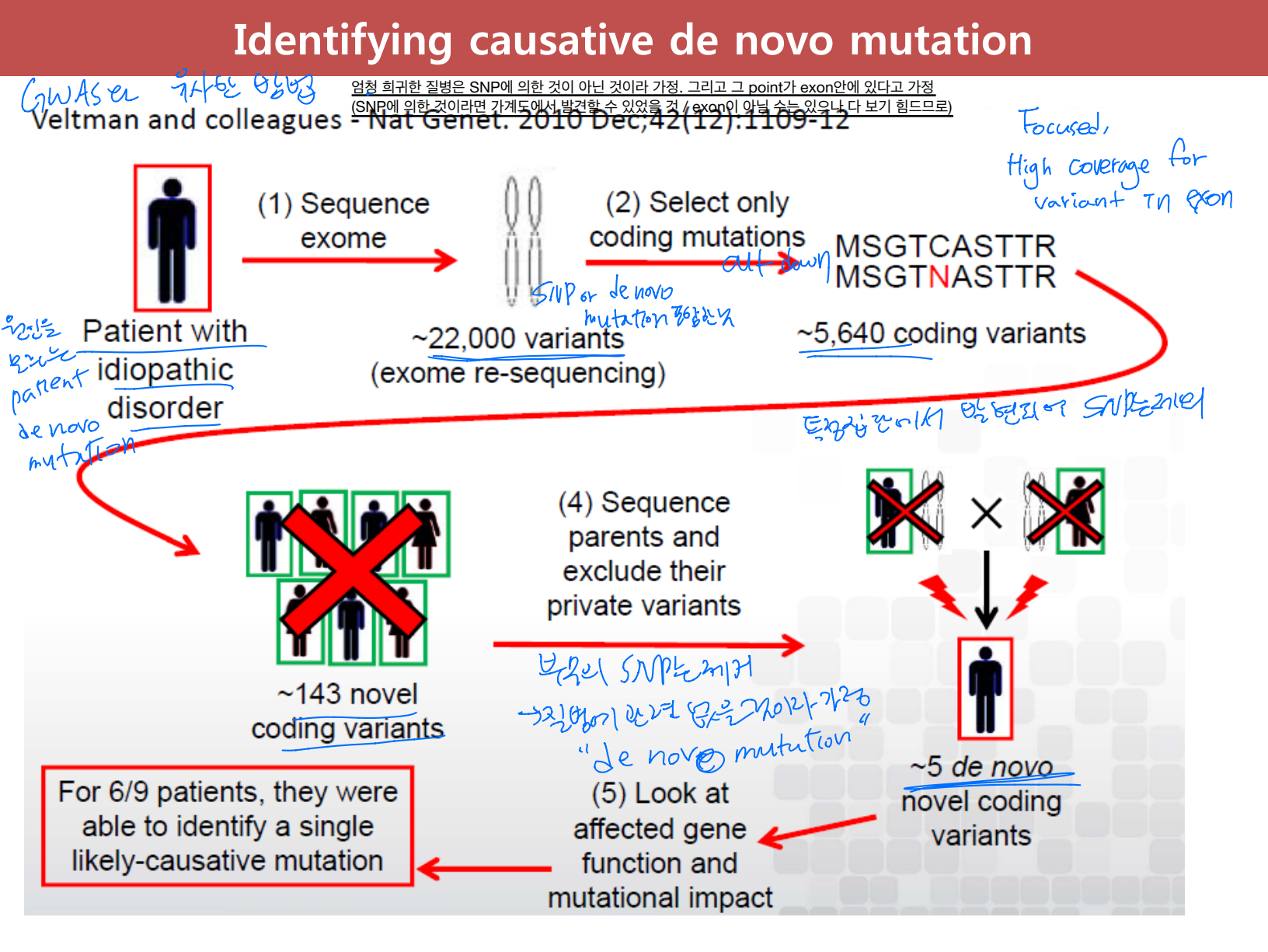
probe가 seq composition에 따라 affinity가 다름
(GC의 양 등에 영향을 받음
→ 특정 부분이 제대로 detect 되지 않을 가능성 있음)
① 이유에 좋았음

A

Identifying causative de novo mutation

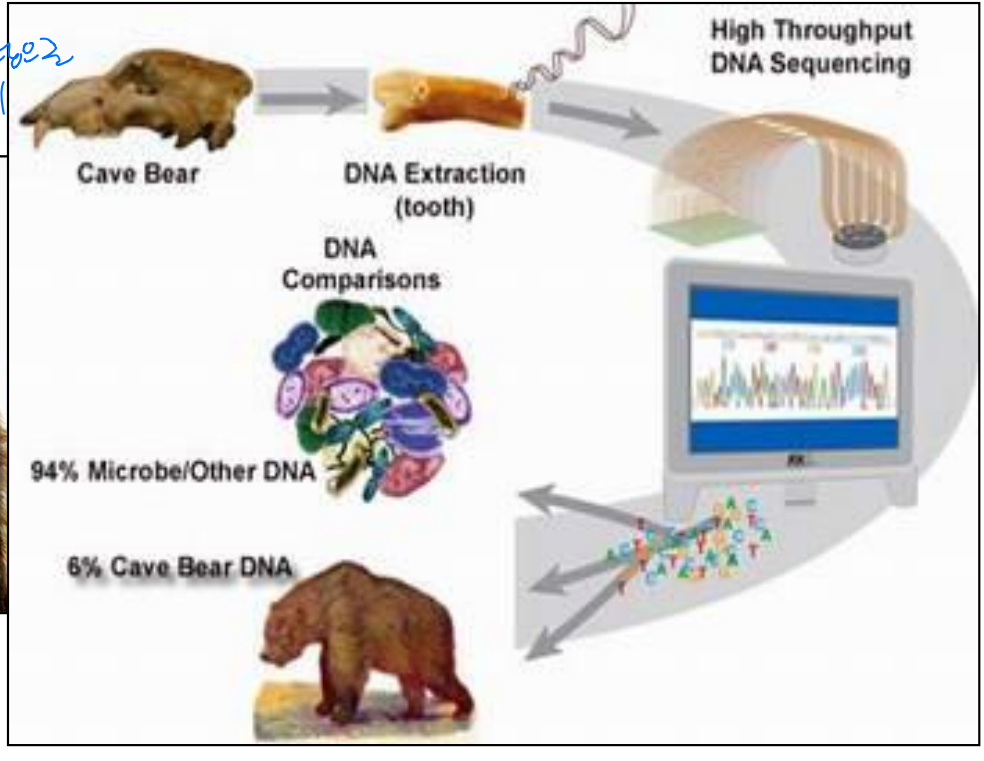
GWAS에 의하면 엄청 희귀한 질병은 SNP에 의한 것이 아닌 것이라 가정. 그리고 그 point가 exon안에 있다고 가정
(SNP에 의한 것이라면 가계도에서 발견할 수 있었을 것 / exon이 아닐 수는 있으나 다 보기 힘드므로)

Veltman and colleagues - Nat Genet. 2010 Dec;42(12):1109-12



Ancient Genomes Resurrected

- **Nuclear genomes of ancient remains: cave bear, mommoth,**
- **Neanderthal genome (10^6 bp)**



Problems: contamination modern humans and coisolation bacterial DNA
Solution : Exome Capture based on Homo Sapiens

The Neanderthal Genome

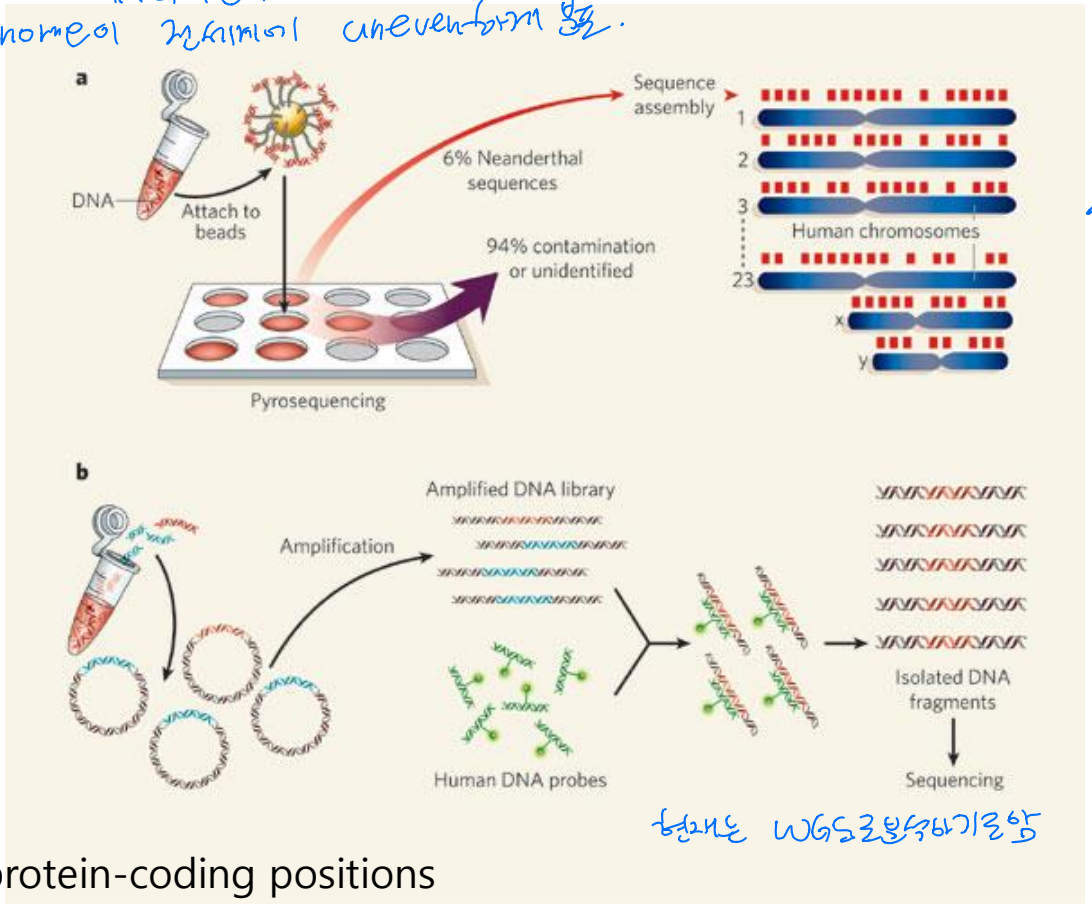
인간 genome 이랑 네안데르탈인과 얼마나 다른가?

→ 네안데르탈인의 genome이 현재에서 sequencing 됨.

- Neandertal genome composed of more than 4 billion nucleotides from three individuals

- Gene flow has occurred from Neandertals to humans of Eurasian descent, but not to Africans

(2010)



현재는 WGS로 분석하기 용이

- sequenced ~14,000 protein-coding positions

- identified 88 amino acid substitutions that have become fixed in humans since our divergence from the Neandertals

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

Functional Genomics

네안데르탈인은 호모 사피엔스에서는 흔히 발견되지않으나 , 현생 아시아, 유럽인에서 발견되는 것들이 많이 발견
-> 아프리카에서 시작한 호모사피엔스와 섞였을 것으로 추측 (?)

Gene expression

Genome → Genome-wide methods → **Phenotype**

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

NGS

Genetic interaction
RNAi screening

DNA
Chip-Seq / ENCODE

RNA
RNA-Seq /microarray

Protein
Translation

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

Forward and Reverse Genetics

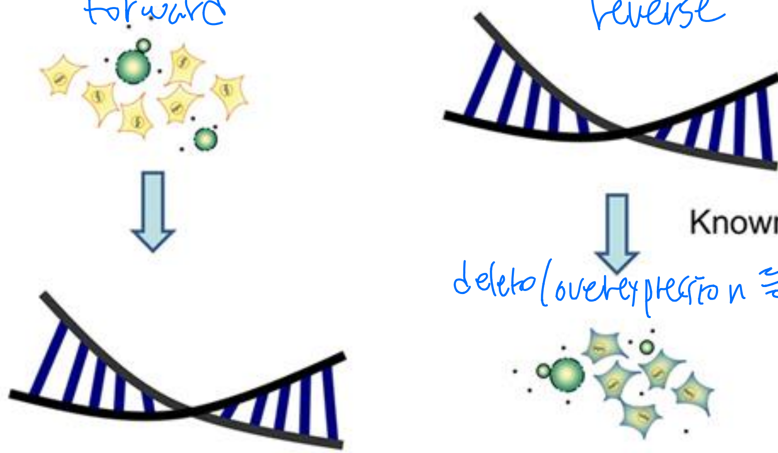
Clues to gene function can be inferred from different types of genetic manipulation

- Two major types of genetic analysis have been applied in model systems

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

forward

reverse



delete/overexpression 등

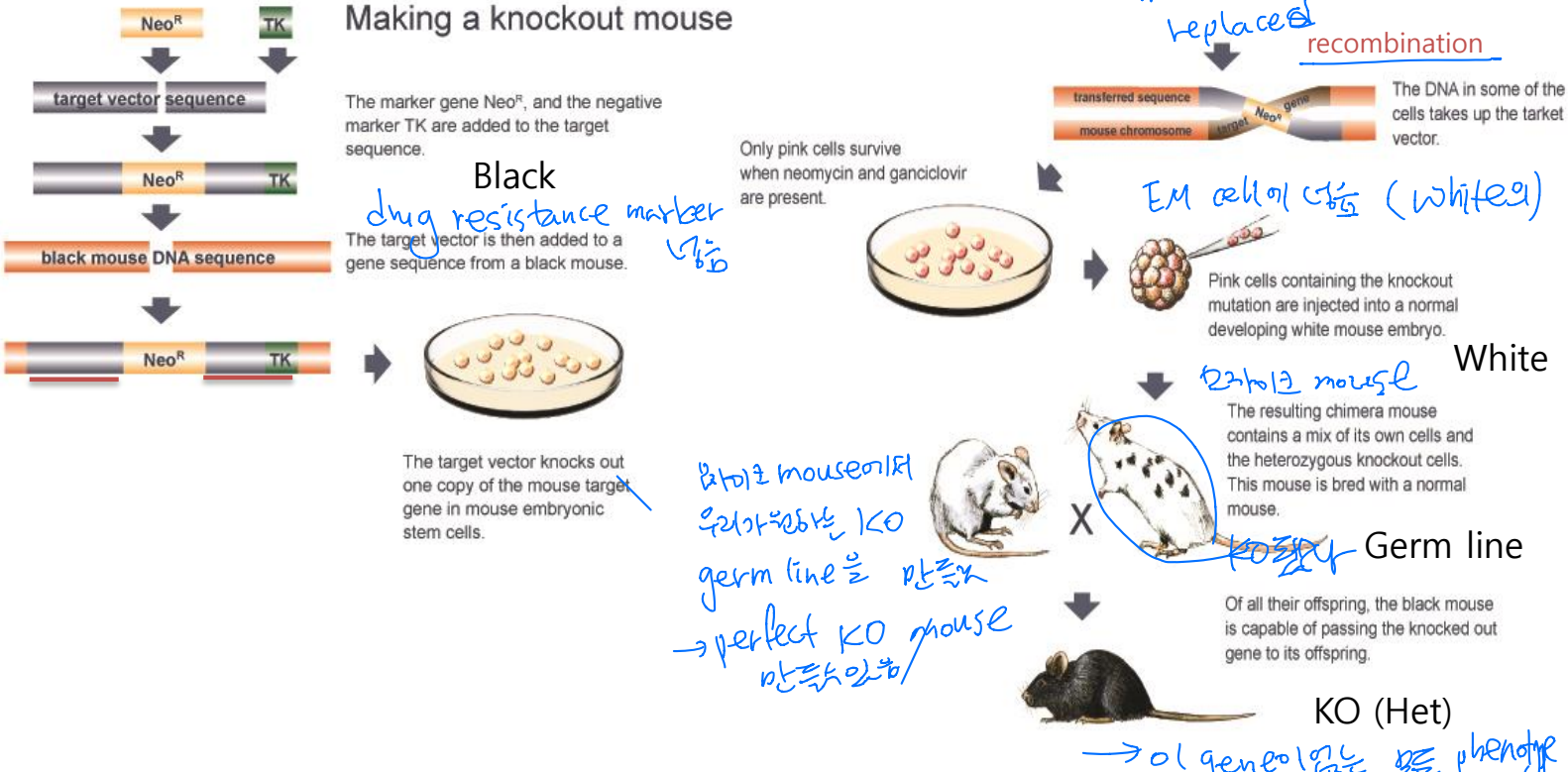
ptn 알기 때문에 gene의 seq 먼저 알게 된다 → mutagenesis 발생 후 phenotype 본다

표현형으로부터 gene을 나타내는 방법

특정 seq에 대한 정보가 있을 때, function 추정

Selective gene inactivation and modification

- Knockout of a specific gene results in a loss of gene function
  - Knockout is commonly used in mouse models, but the procedures are laborious and time-consuming



- Recently, several methods have been systematically applied to the gene function studies with cultured cells, although they have theoretical disadvantage

Selective gene inactivation and modification

- Four different functional classes of mutation can be generated, and different screens can be conducted depending on the desired effect on the target gene

TABLE 12.3 DIFFERENT TYPES OF REVERSE GENETIC SCREEN IN CULTURED MAMMALIAN CELLS

Type of screen	Basis of method
Loss-of-function	usually, the RNA interference (RNAi) pathway is induced to selectively degrade RNA transcripts of a specific target gene (see Figures 12.3 and 12.4)
Gain-of-function	involves transfection of an exogenous gene copy driven by a suitable promoter to overexpress or misexpress that gene in a desired cell type
Dominant-negative	relies on producing a mutant gene product that interferes with the normal product of a specific target gene; similar to loss-of-function screen but suppresses gene activity more efficiently than RNA; often the mutant product is designed to be a truncated version of the wild type that competes with wild-type protein in some way, or the mutant protein becomes incorporated with wild-type proteins in multisubunit complexes, thereby inactivating the complex
Modifier	seeks to identify genes that enhance or suppress a specific phenotype; the initial phenotype is produced by one of the three methods above; thereafter, the mutant cells are subjected to a second genetic modification (which can again be any of the three methods above); if altered expression of a second gene enhances or suppresses the initial phenotype, the second gene is considered to be a modifier of the gene that caused the initial phenotype; it then remains to be worked out how the two genes interact

→ infer relationship

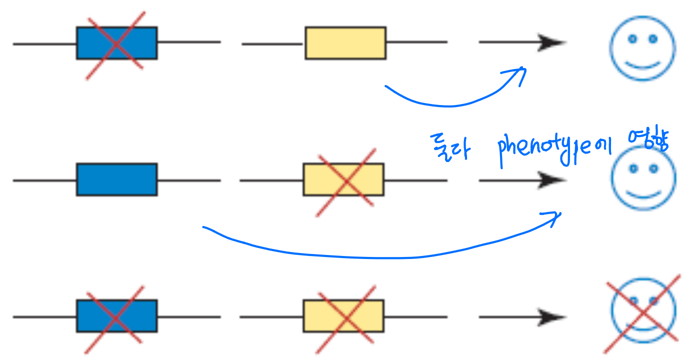
Genetic interaction

↳ 반대방향의 Loss of function 과 gain control 관찰

Genetic interaction

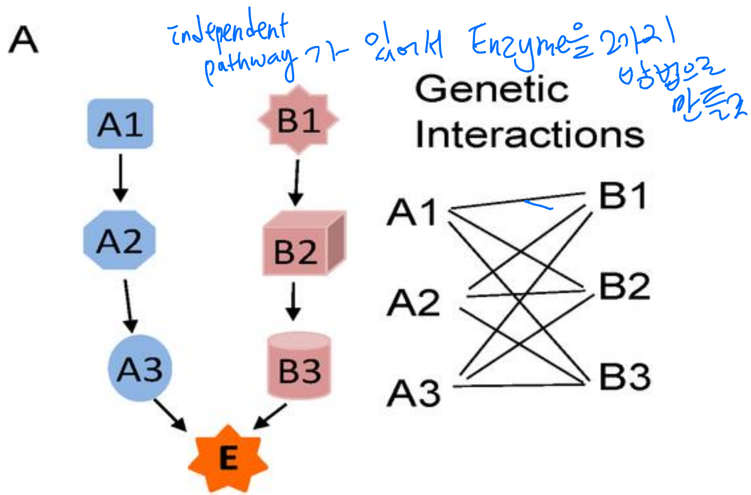
Genetic interaction
(e.g. synthetic sick or lethal interaction)

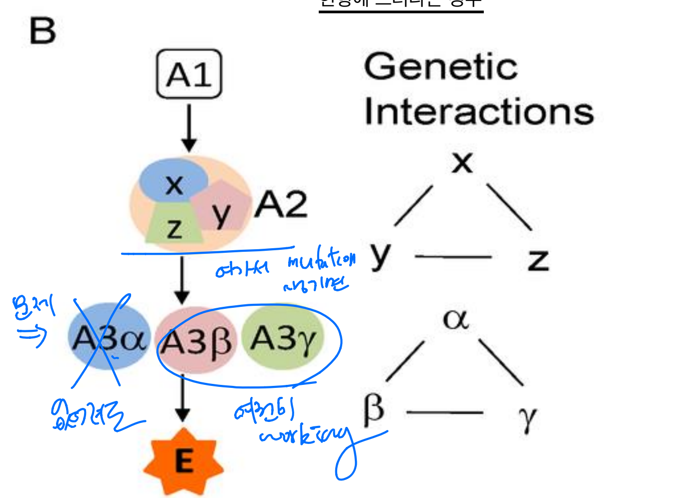
different mutation
same phenotype
or working together



one mutation is not enough
↳ still working

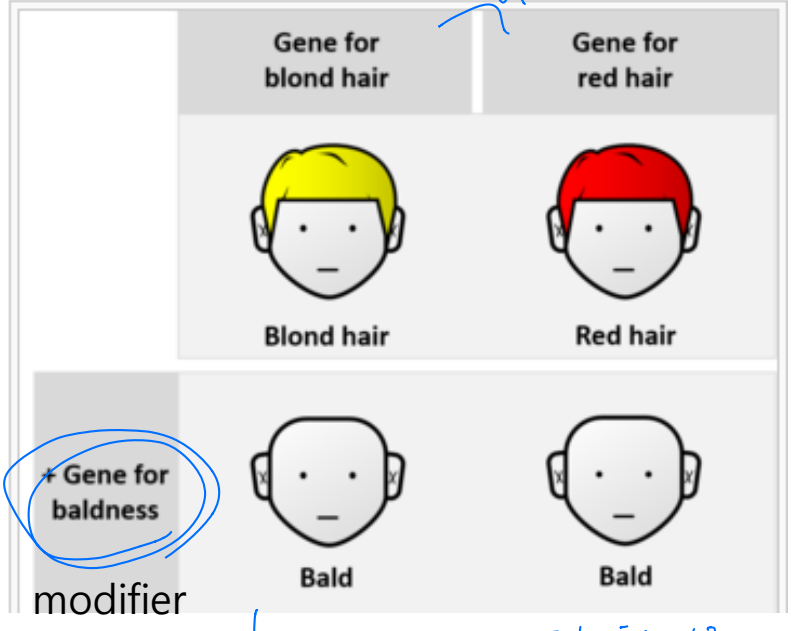
ptn. complex를 만들어 각 complex의 한부분이 없더라도 나머지가 보완할 수 있는 경우 -> 2개 다 없애야지만 표현형에 드러나는 경우





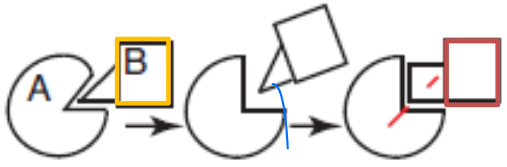
Epistasis

Epistasis is a phenomenon that consists of the effect of one gene being dependent on the presence of one or more 'modifier genes' (genetic background). Similarly, epistatic mutations have different effects in combination than individually

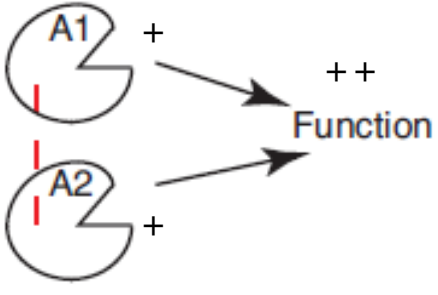


modifier

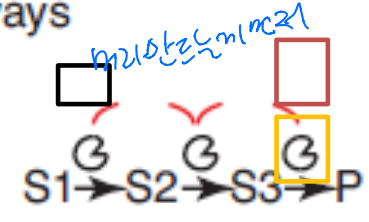
Molecular recognition



Redundancy



Positive interactions in linear pathways



Dominant Negative Phenotype

loss of function을 만들어내는 것으로 확인하는 방법

homo zygote인 loss-of-ftn을 만드는 것이 어려운데 ,
dimer을 형성하는 놈은 한 부분만 mutation을 넣어줘도 그 기능을 상실함
-> 이 형태로 loss-of-ftn에서의 표현형을 추측할 수 있음.

- A mutation whose gene product adversely affects the normal, wild-type gene product within the same cell, usually by dimerizing (combining) with it

dimer를 형성하는것

Functional	Non-Functional	Non-Functional
		
25%	50%	25%
Dimer of two normal Factor XI monomers. The dimer has normal function. About one fourth the molecules are such homodimers.	A dimer of one normal (green) and one monomer encoded by the dominant negative gene (red). The dominant-negative gene product inhibits the function of the product of the normal gene rendering the entire molecule functionless. About half the molecules are heterodimers.	A dimer of two dominant negative monomers that is functionless. About one fourth of the molecules are such homodimers.

1개만 mutation 되어도 dimer로 작용하는 경우는 loss of function

mutation으로인해 loss of funct

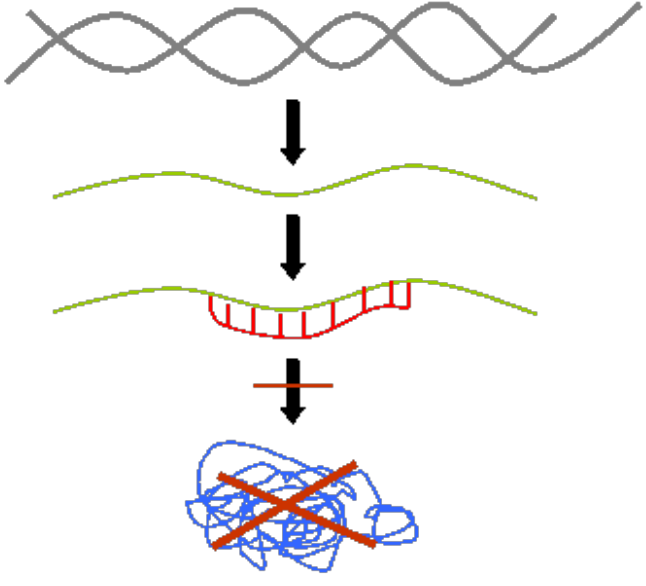
Selective gene inactivation and modification

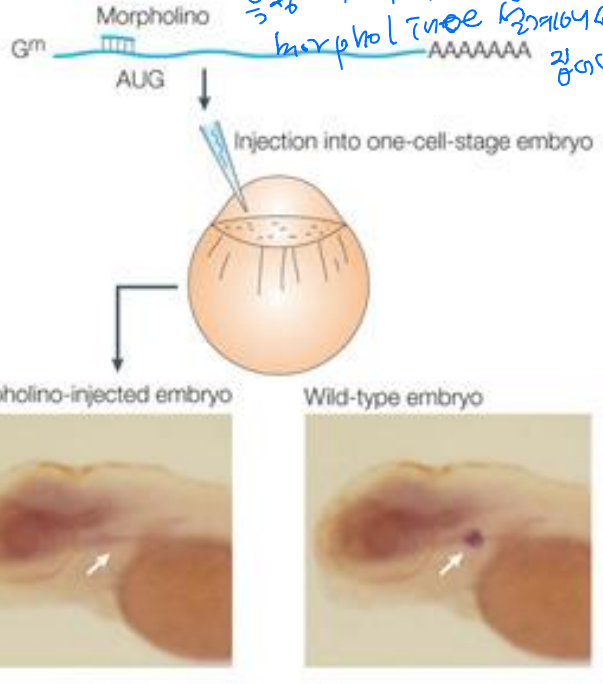
Clues to gene function can be inferred from different types of genetic manipulation

- Antisense technology: an early general approach to inhibit gene expression using the specificity of base pairing
  - RNA or oligonucleotide constructs are designed to have a sequence that is complementary to that of RNA transcripts from a gene of interest (knockdown)
  - This technology was first developed in the 1980s (antisense RNA)
  - Subsequent way was use of antisense oligonucleotides
  - More recent was was use of morpholino antisense oligonucleotides
    : vastly more stable and robust structure than conventional oligonucleotides, and more consistent in producing significant inhibition of gene expression

Antisense technology : Morpholino

Morpholino oligos are a class of antisense, a technology used to block access of other molecules to specific sequences within nucleic acids. Morpholinos block small (~25 base) regions of the base-pairing surfaces of RNA and are used as research tools for reverse genetics by knocking down gene function.





RNA interference (RNAi)

RNA interference (**RNAi**) is a biological process in which RNA molecules inhibit gene expression, typically by causing the destruction of specific mRNA molecules

Phenomena first observed in petunia (1990)
Attempted to overexpress chalone synthase (anthrocyanin pigment gene) in petunia (trying to darken flower color) -> Caused the loss of pigment.

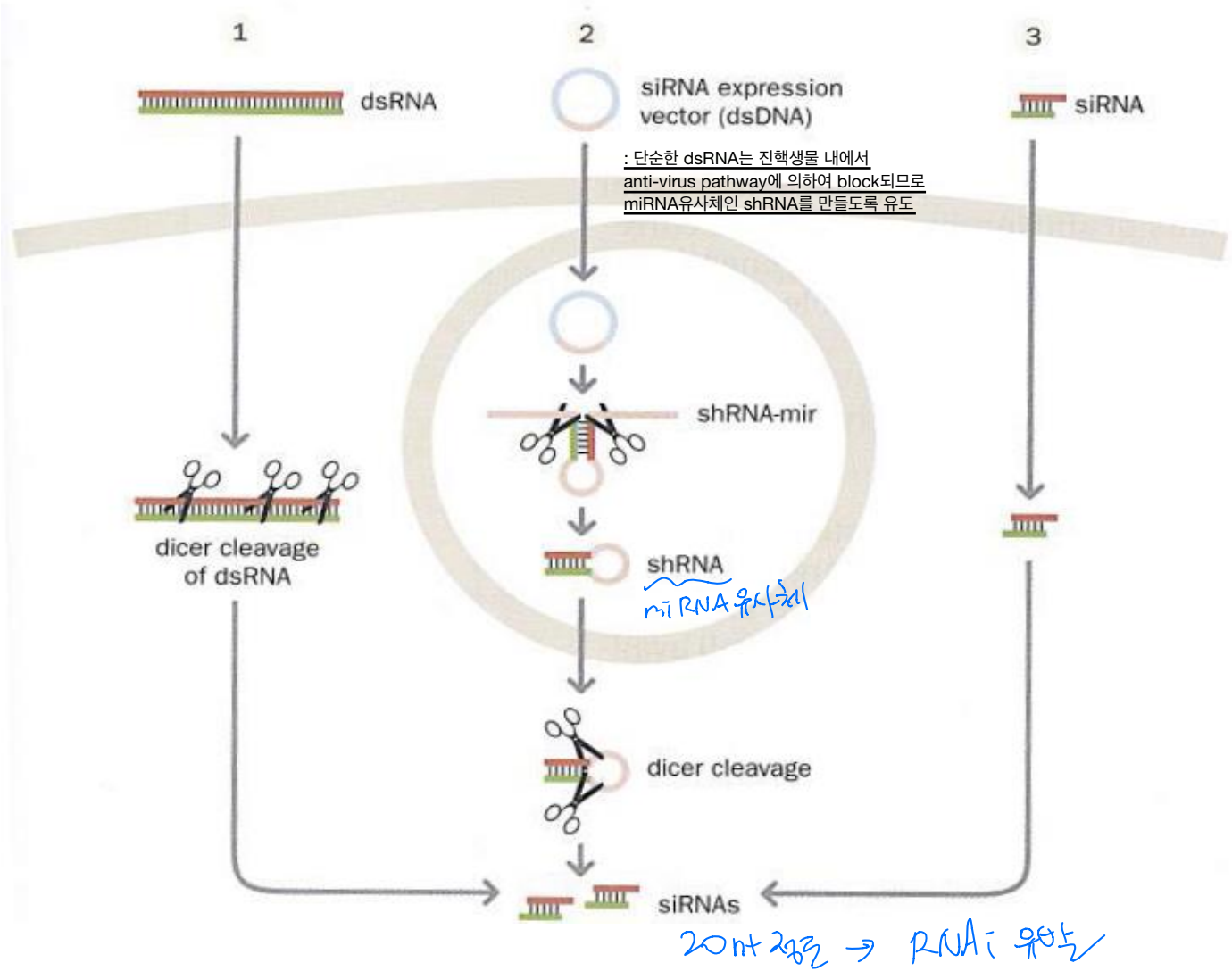
=> ds 를 만들어서 특정 gene의 활성을 낮춰버림
(cosuppression : transfaction한 gene과 원래 페튜니아가 가진 gene이 모두 억제되버림)

2006 Nobel prize in physiology or Medicine

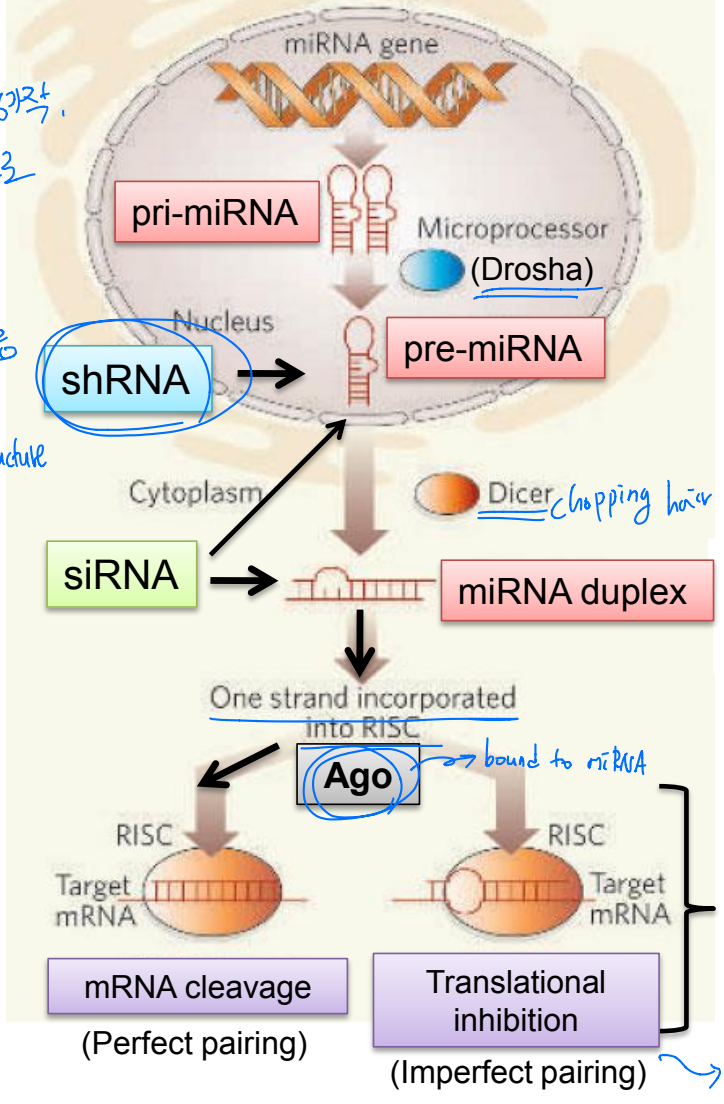
Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*

(1998)

RNA interference (RNAi)



RNA interference (RNAi)



http://www.youtube.com/watch?v=H5udFjWDM3E

RNAi screening

design 20nt siRNA → hybrid to RNA

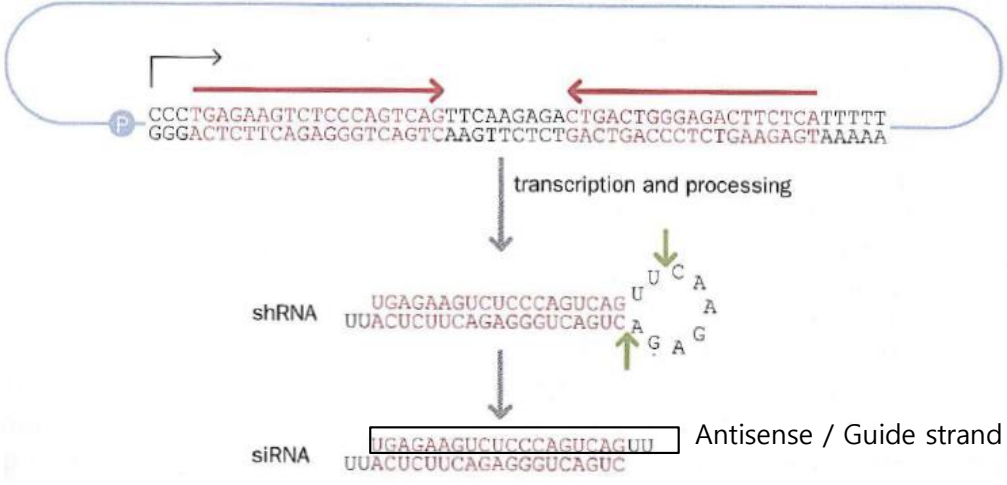
Global RNAi screens provides a systems-level approach to studying gene function in cells

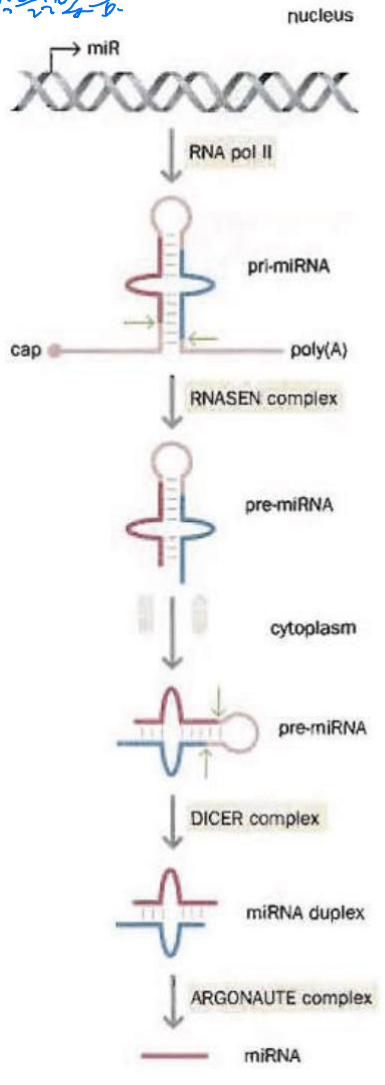
우리 genome에 붙는 여러 종류의 RNA를 디자인 -> 개별적으로 각 gene 발현을 막음
*siRNA는 cis로 작용하기 때문에 단순히 관심있는 gene을 shRNA로 만들 수 있도록 설계하여 클로닝하여 넣어도 효과를 볼 수 있음

- RNAi studies began by attempting to assess the function of individual genes
- In the post-genome era, it became possible to perform large-scale, and eventually genomewide, analyses of gene function (1)
- To perform global RNAi screens, suitably large nucleic acid libraries (siRNA, shRNA libraries) need to be made (2)
- Short hairpin RNA (shRNA) libraries
  - A short transcript that spontaneously forms a hairpin RNA, which then undergoes cleavage → inverted-repeat로 hairpin structure 만들어짐
  - Pairs of long complementary oligonucleotides are designed in such a way that, when annealed and cloned into a suitable expression vector, a transcript can be produced with inverted repeats that base -pair

shRNA library

1 probe RNA와 같은 gene 발현 억제함.

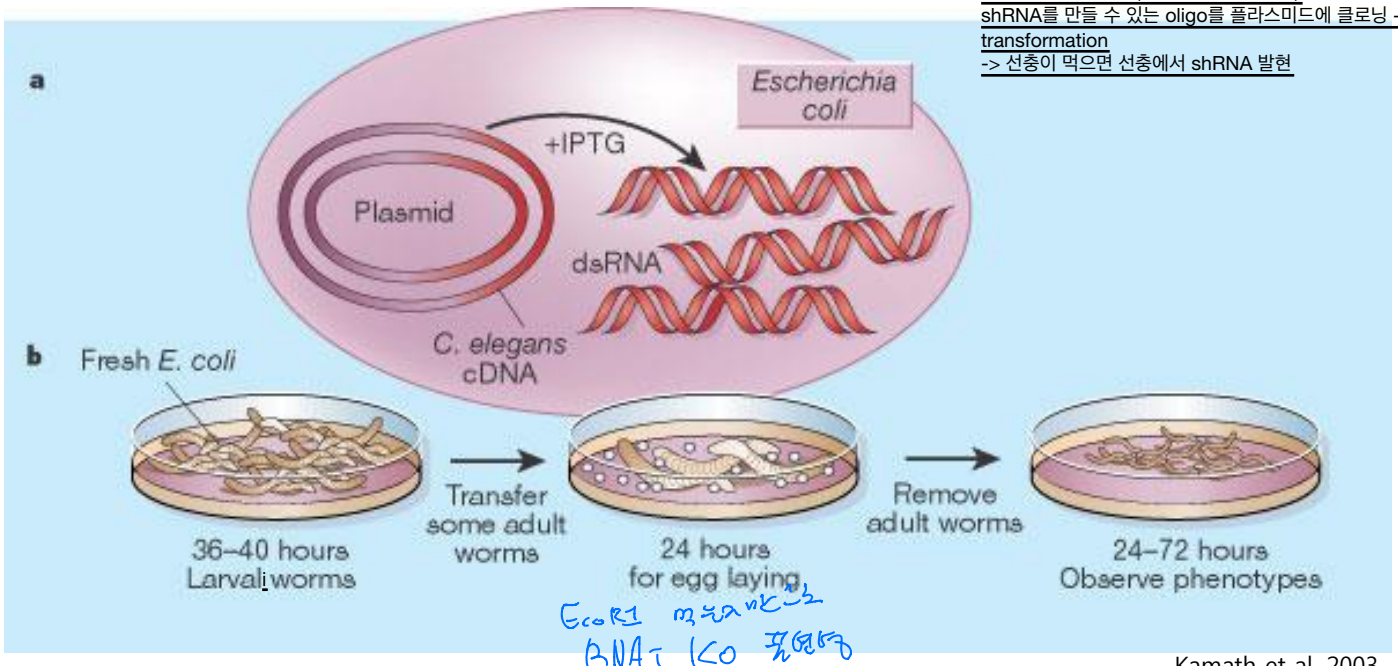




RNAi Screening

=> 특정 gene이 표현형에 어떤 영향을 끼치는지 알 수 있음.

선충의 먹이 -> 세균(실험실에서는 E.coli)
shRNA를 만들 수 있는 oligo를 플라스미드에 클로닝 -> E.coli transformation
-> 선충이 먹으면 선충에서 shRNA 발현



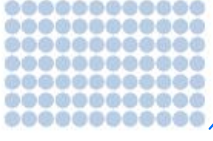
Kamath et al. 2003

16,757 strains = 86% of predicted ORFs
Looked for sterility or lethality(Nonv), slow growth (Gro) or defects (Vpep)
1,722 strains (10.3% had such phenotypes)

High-throughput shRNA screening

Arrayed Screens

- Hypothesis driven gene sets
- Robust interrogation of each gene
- High Content Screening



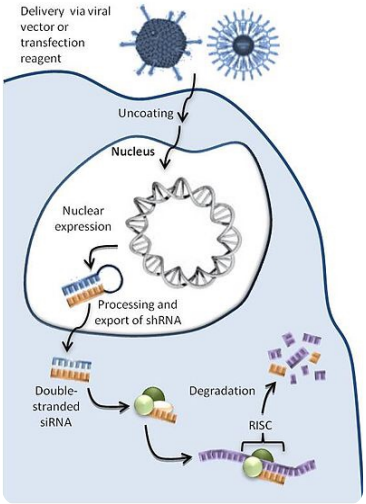
plate

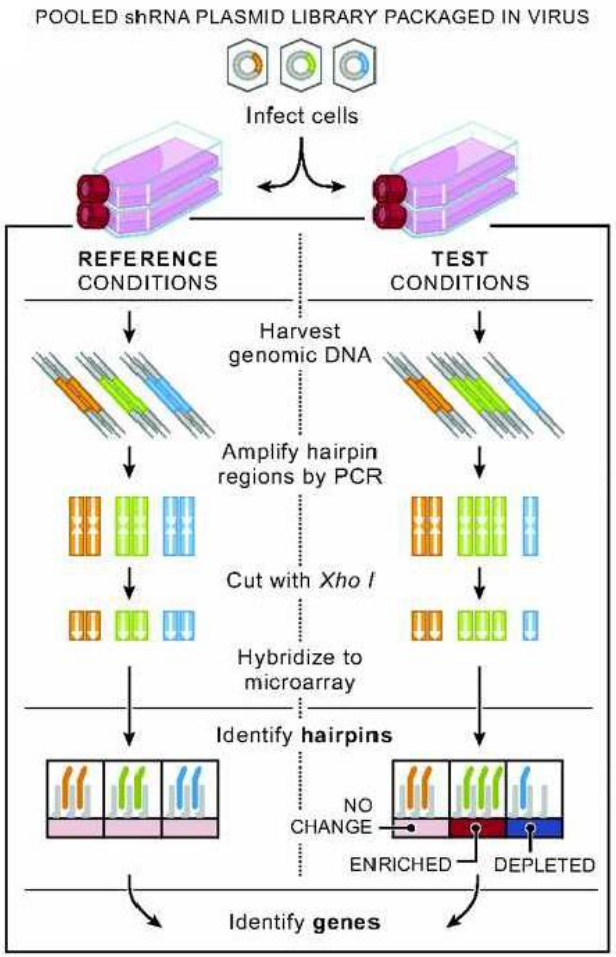
Pooled Screens

- Genome-scale screening
- Many conditions: drug doses, time points, cell types etc.
- Long time course assays - proliferation and survival
- Lower resolution – readout is differential representation of hairpin measured by microarray or sequencing

High-throughput shRNA screening

Pooled RNAi screening strategy and performance using pools of 45,000 shRNA-expressing viruses.





cancer cell을 가지고
drug sensitive gene을 찾고자함
hairpin을 만들어 주는 vector 설계
→ 주입한 cell에 발현
→ 특정 gene KD

PGK — HI — ═══ (antisense) - ═══ (sense) —

Pool of siRNA
+ / - Treatment (condition)

- **Identification of shRNA clones**
  shRNA가 특이적 농축되었는가
- **Samples vs. controls (Ratio)**

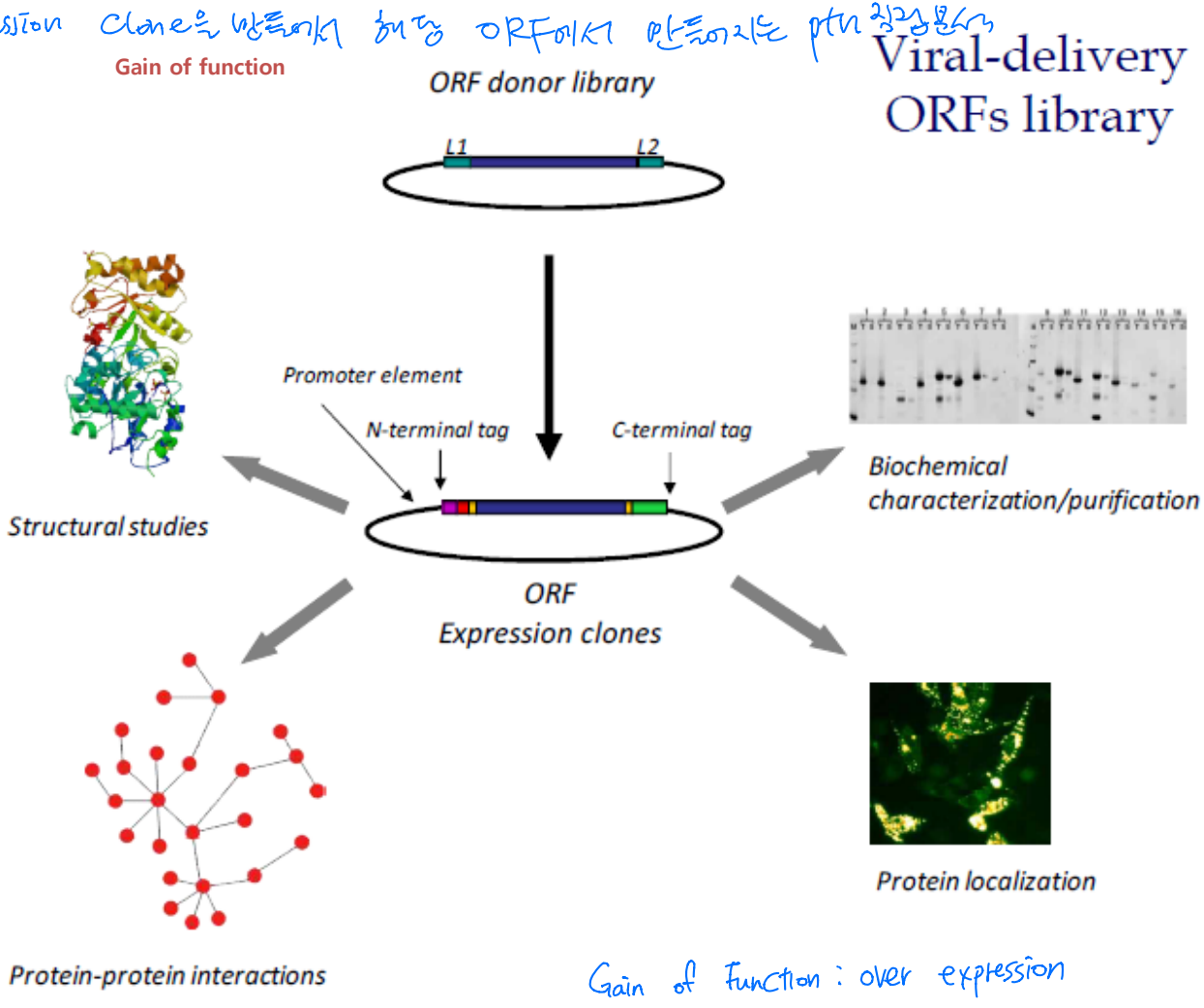
- **Identification of genes**

drug이 많이 들어갔을 때
시간이 많이 지났을 때
= shRNA가 특이적
→ 즉 drug에 반응하는 sensitive

Open Reading Frames (ORFs) expression

ORF expression clone은 vector에서 해당 ORF에서 만들어지는 ptn 과발현주

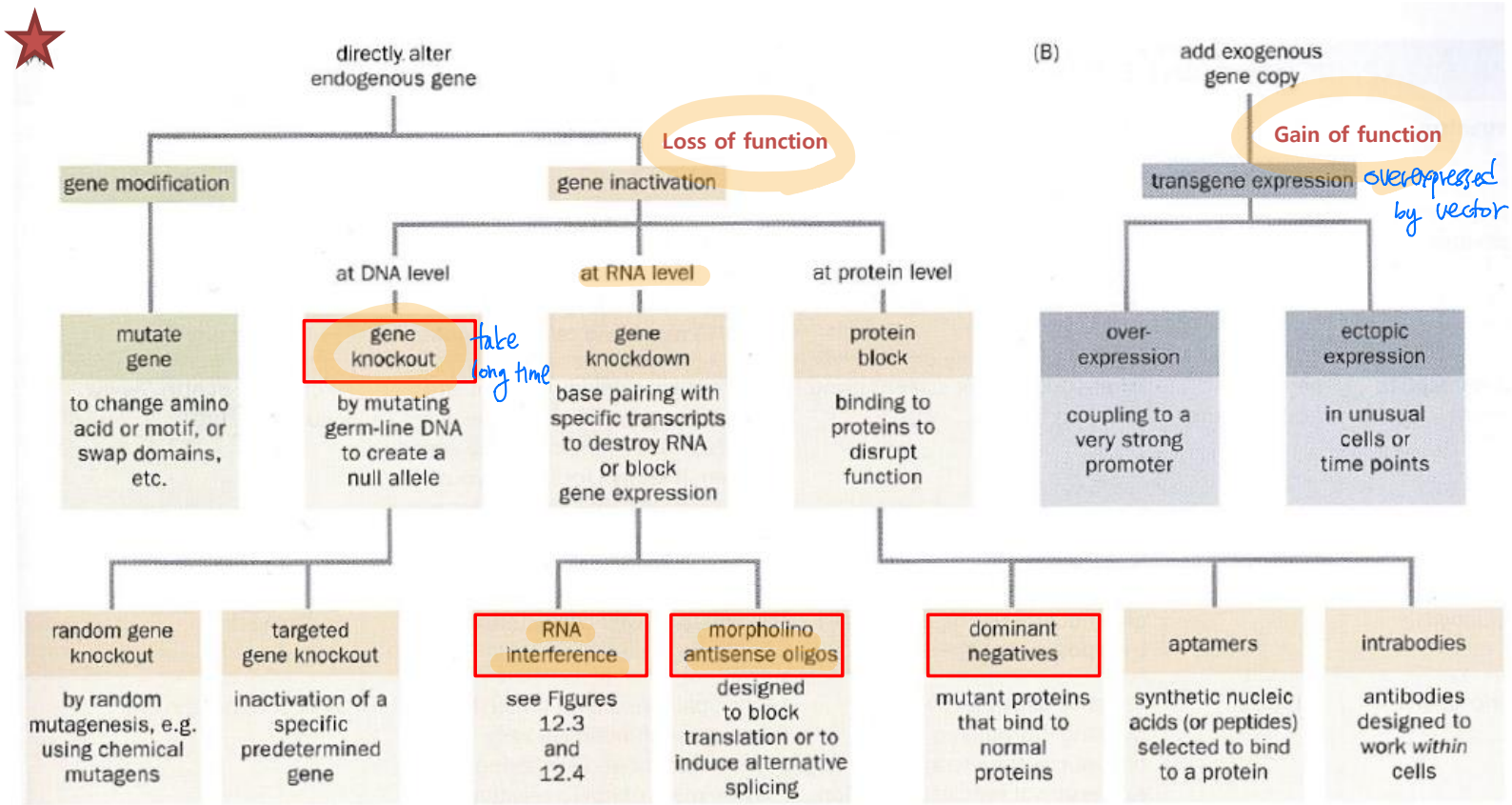
Gain of function



Gain of function : over expression

Selective gene inactivation and modification

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



youtube : Perkin Elmer Screening