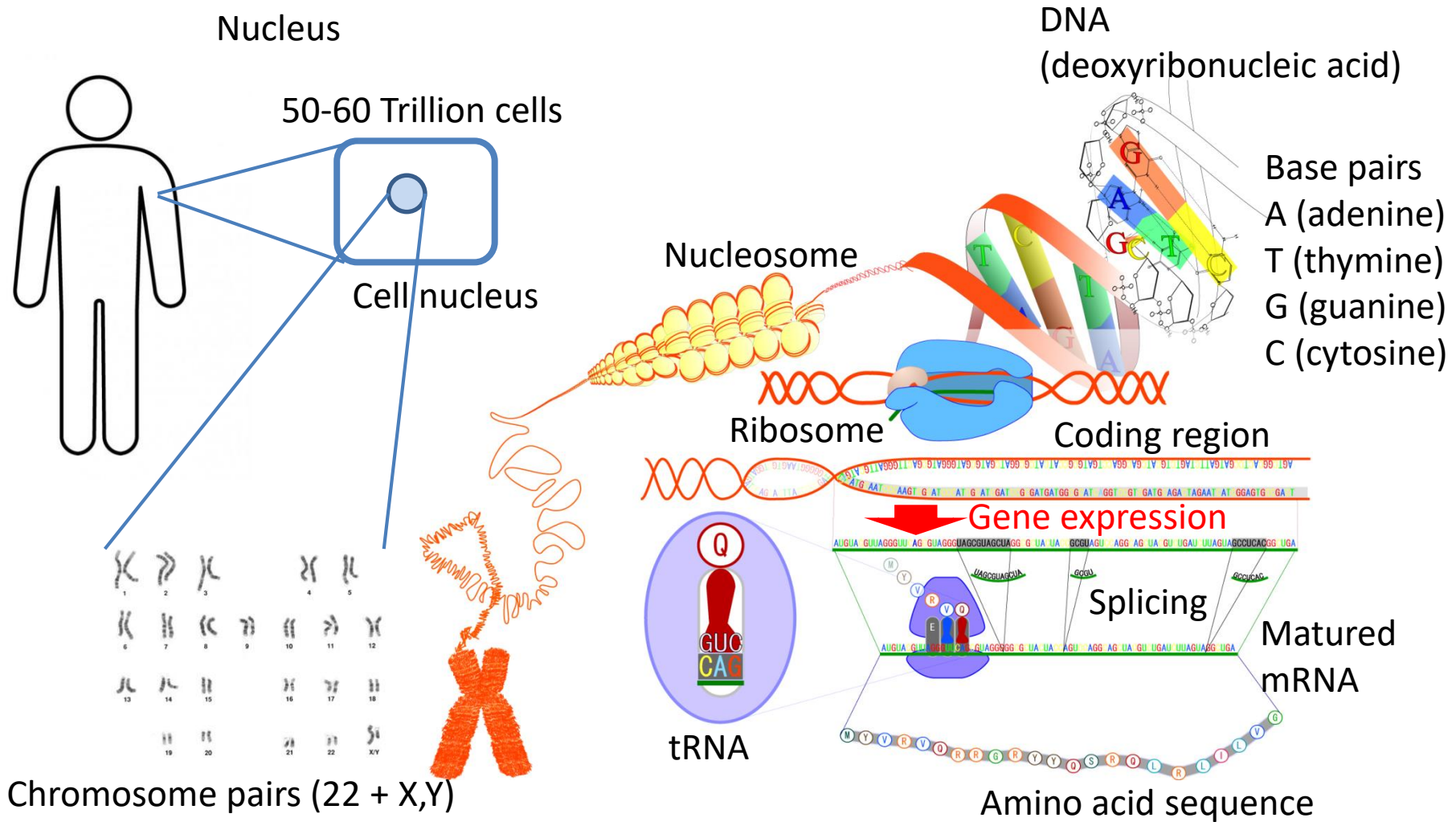# Gene Expression Regulatory Network Analysis

Akihiko Konagaya

School of Computing

Contents

- What is gene?

- What is gene expression?

- Gene expression regulatory network

- Complex Network Analysis

- Summary

What is gene?

Nucleus

50-60 Trillion cells

Cell nucleus

DNA
(deoxyribonucleic acid)

Base pairs
A (adenine)
T (thymine)
G (guanine)
C (cytosine)

Nucleosome

Ribosome

Coding region

Gene expression

Splicing

Matured
mRNA

tRNA

Amino acid sequence

Chromosome pairs (22 + X,Y)

Gene expression regulation

From DNA to Protein

https://www.youtube.com/watch?v=gG7uCskUOrA   2m:41s

Signal transduction => Gene expression

https://www.youtube.com/watch?v=vi-zWoobt_Q 3m:37s

Transcription factors
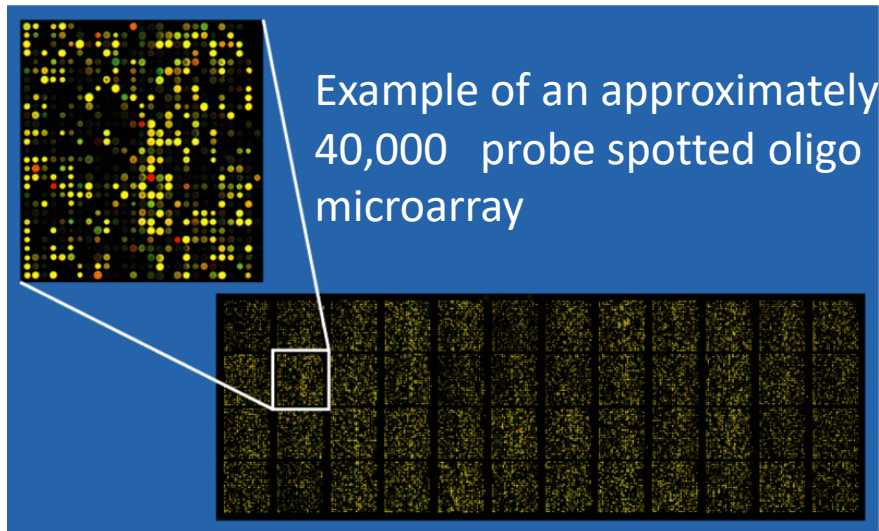
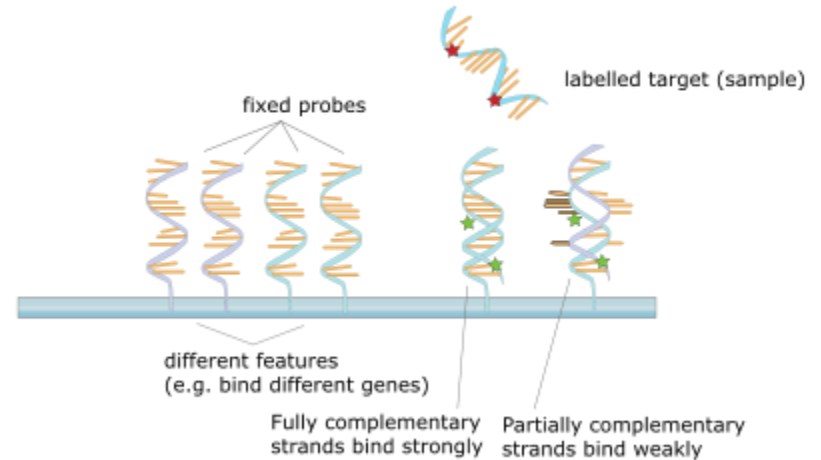https://www.youtube.com/watch?v=MkUgkDLp2iE 2m:55s

Issues

- Measurement of genome wide expression data

- Construction of gene expression regulatory network

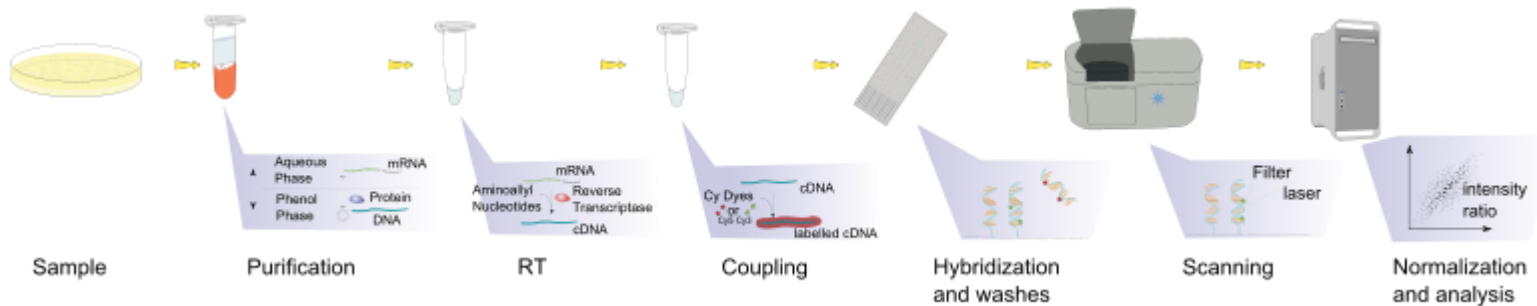- Biologically meaningful interpretation of the gene expression regulation network
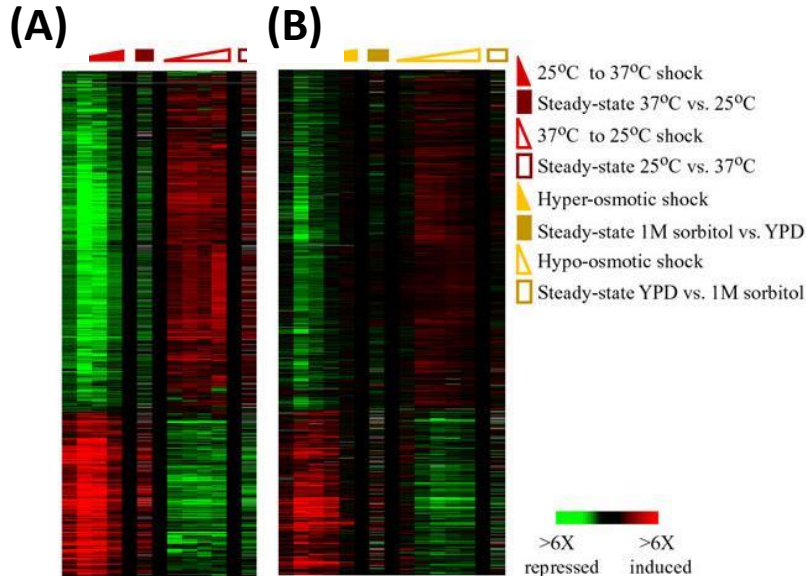
DNA Micro Array

Example of an approximately 40,000 probe spotted oligo microarray

DNA Hybridization

fixed probes

labelled target (sample)

different features
(e.g. bind different genes)

Fully complementary
strands bind strongly

Partially complementary
strands bind weakly

Experimental Protocol

Sample | Purification | RT | Coupling | Hybridization and washes | Scanning | Normalization and analysis

http://en.wikipedia.org/wiki/DNA_microarray

Genome-wide Gene Expression Analysis

(A) **(B)**

25°C to 37°C shock
Steady-state 37°C vs. 25°C
37°C to 25°C shock
Steady-state 25°C vs. 37°C
Hyper-osmotic shock
Steady-state 1M sorbitol vs. YPD
Hypo-osmotic shock
Steady-state YPD vs. 1M sorbitol

>6X repressed   >6X induced

(C) **(D)**

Average expression ratio — Minutes

Gene Expression Regulation

(A) 25°C to 37°C
37°C to 25°C

(B) Low Osmotic Pressure ⇒ High
High Osmotic Pressure ⇒ Low

(C) Average Expression Levels in（A）

(D) Average Expression Levels in（B）
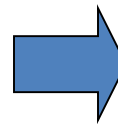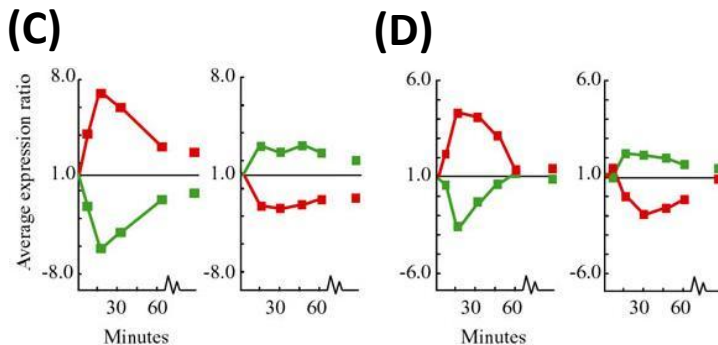
Clustering is useful to identify genes responsible for the stimuli

Gasch AP, Spellman PT,…, Botstein D, Brown PO : Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes, Molecular Biology of the Cell, vol. 11, 4241-4257, (2000)

Limitation of Co-expression Analysis

Genes which have a similar expression profile may be regulated by the same transcription factors.
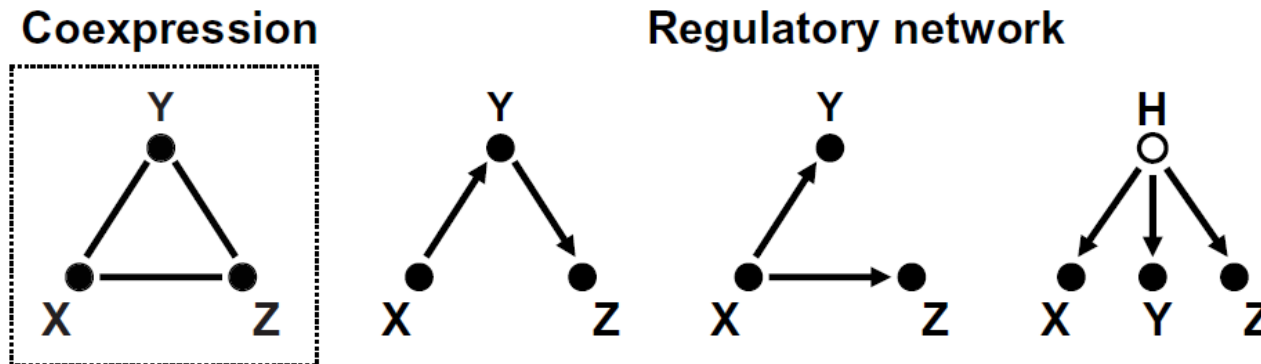
Issues: There are at least three regulation mechanisms even if X, Y and Z show the same expression profile.

Discovery of Driver Genes

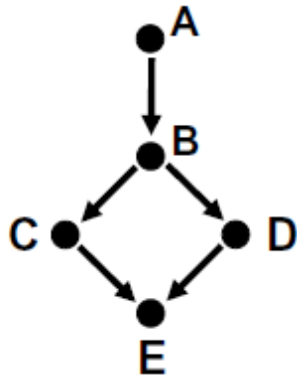How to identify driver genes such as
Yap1p (Transcription factor for oxidant stress)
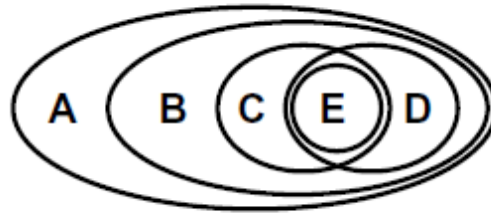**Msn2p** / **Msn4p** (Transcription factor for heat and osmotic stress)
from Gene Expression Data

Knock out of **A**
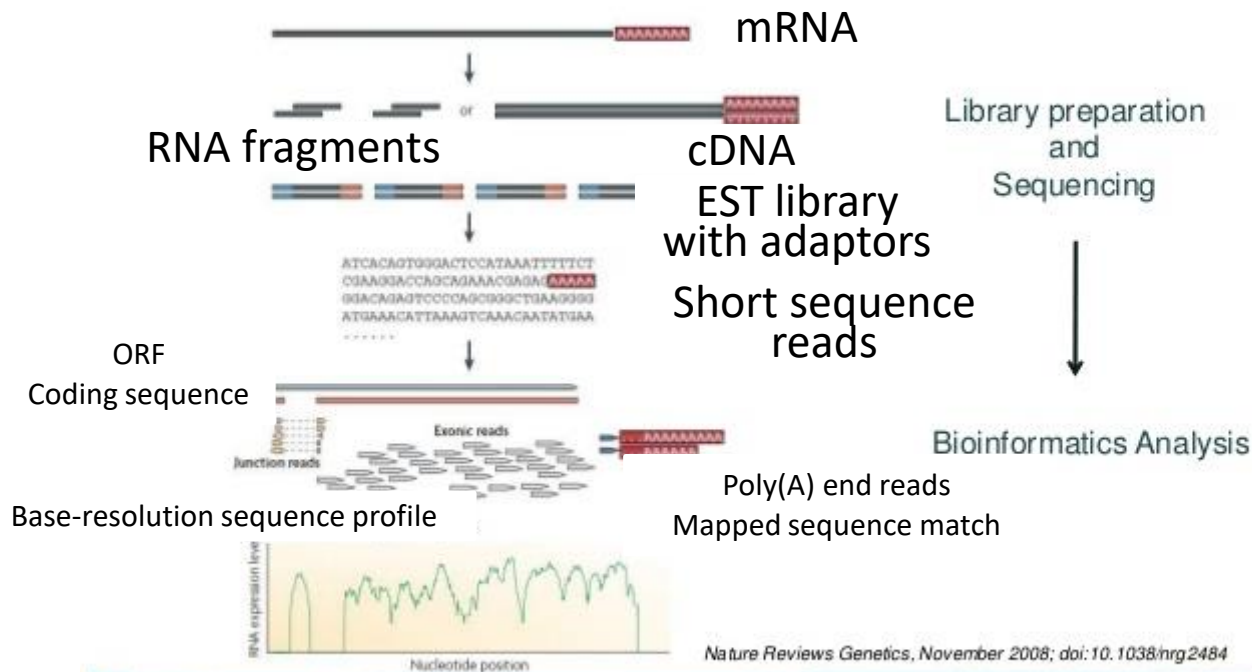affects to **B, C, D and E**

Pathway structure

**Downstream effects of interventions**

RNA Sequencing

Ex) Next Generation Sequencer illumine NextSeq 550

A typical RNA-seq experiment

mRNA

RNA fragments

cDNA
EST library
with adaptors

Short sequence reads

ORF
Coding sequence

Base-resolution sequence profile

ATCACAGTGGGACTCCATAAATTTTTCT
CGAAGGACCAGCAGAAACGAGAGAAAAA
GGACAGAGTCCCCAGCGGGCTGAAGGGG
ATGAAACATTAAAGTCAAACAATATGAA

Library preparation and Sequencing

Bioinformatics Analysis

Poly(A) end reads
Mapped sequence match

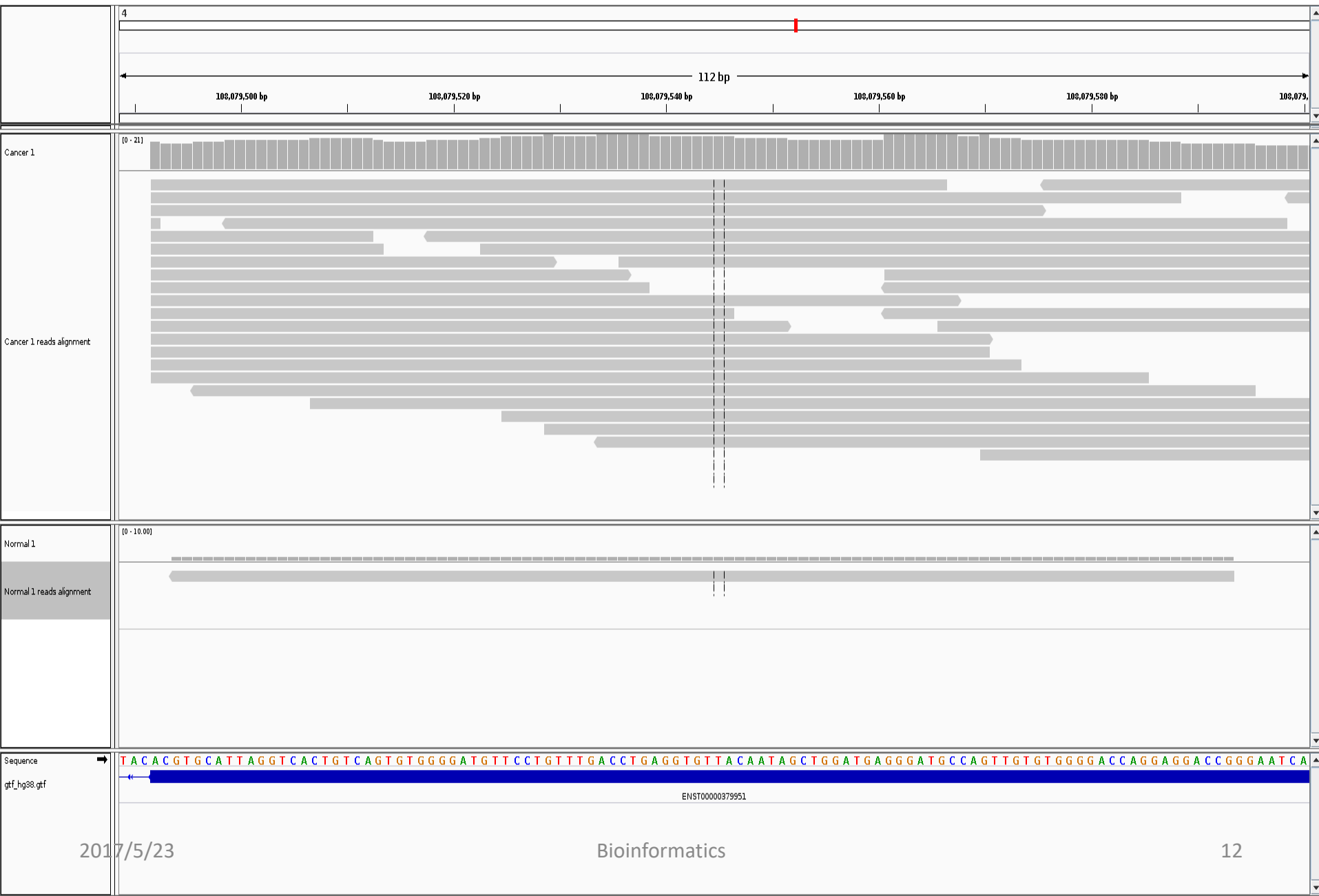Nature Reviews Genetics, November 2008; doi:10.1038/nrg2484

>10 M Reads (75bp)
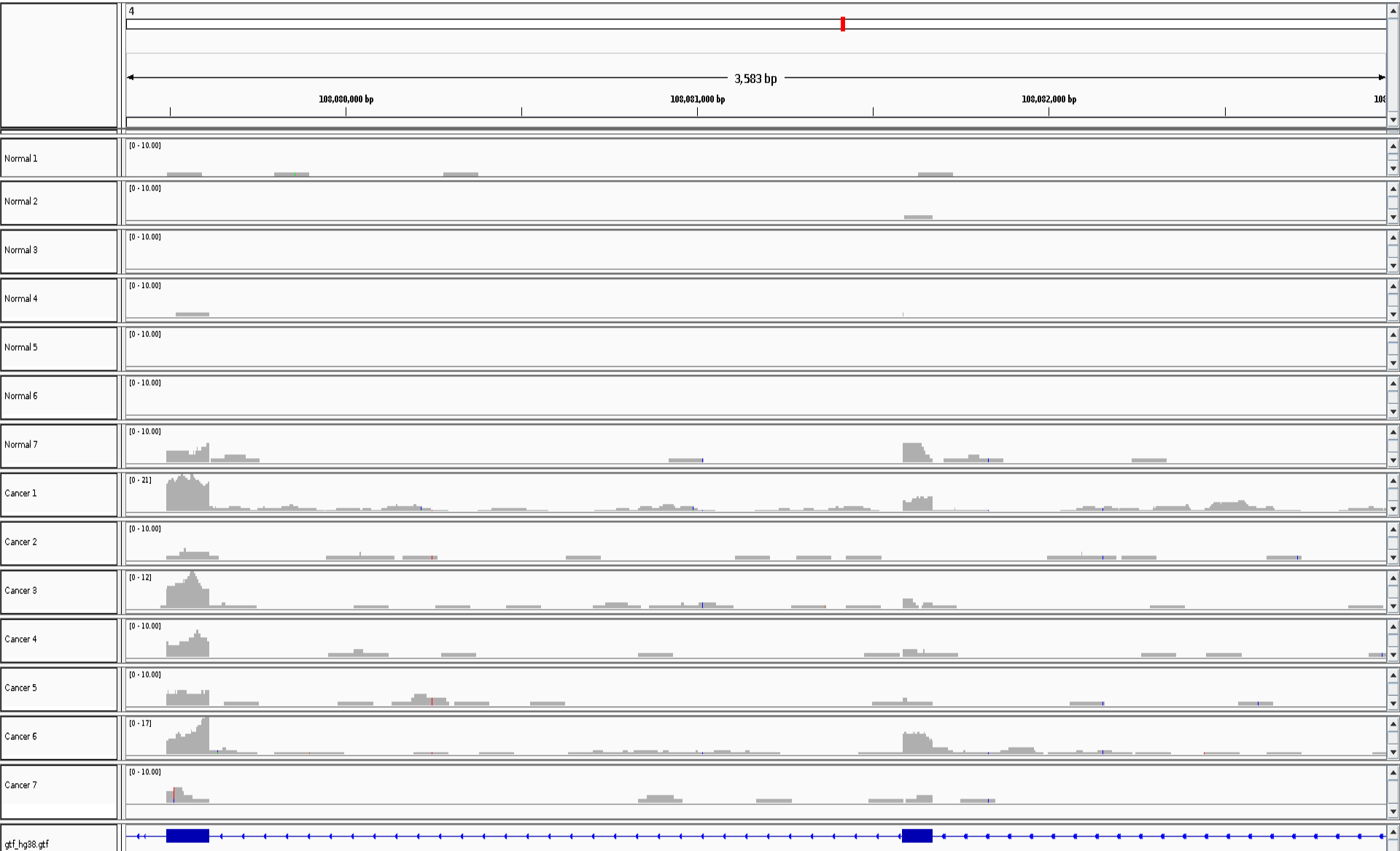40 samples
11 hours for sequencing

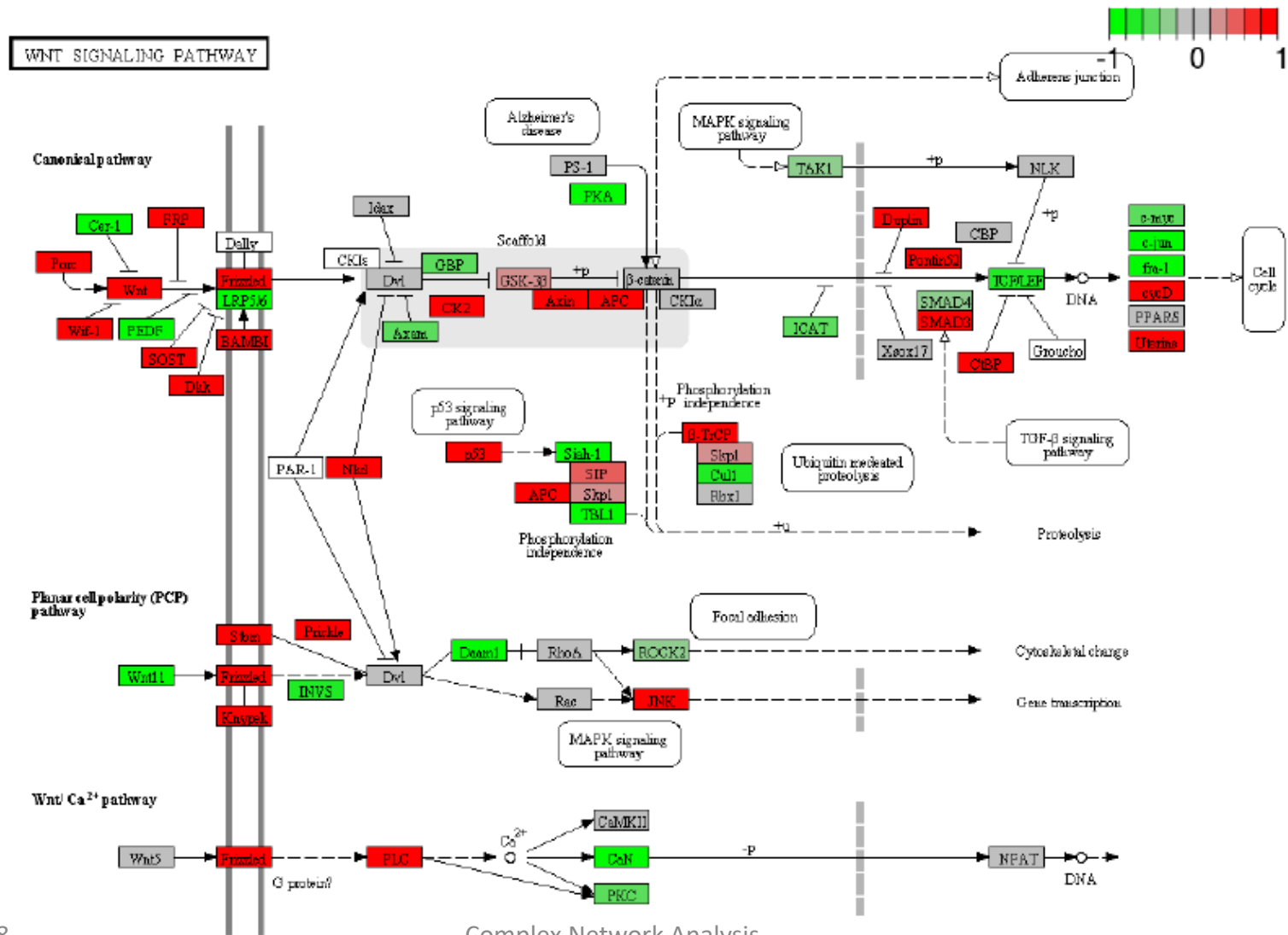# Visualization of reads alignment with genome reference
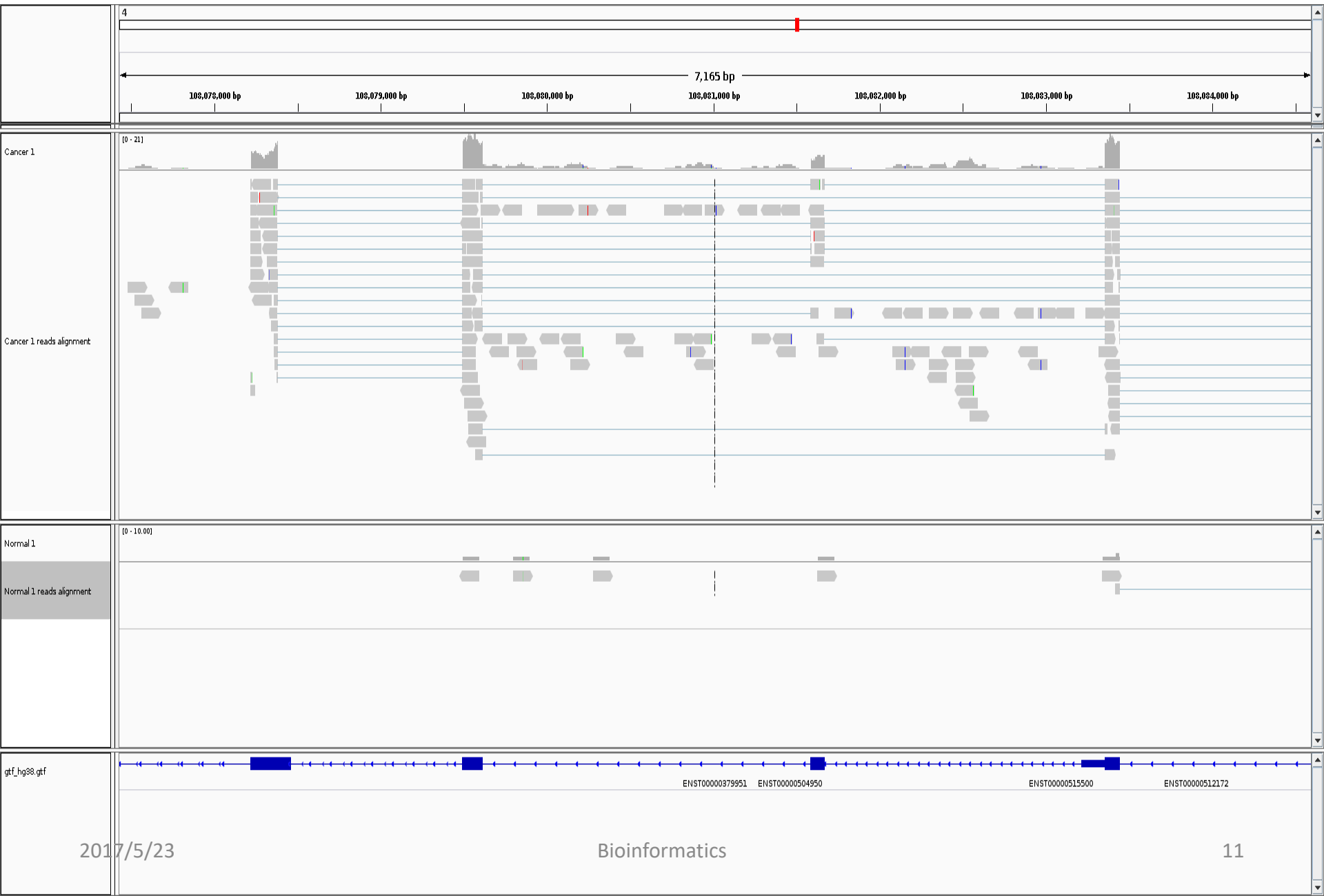
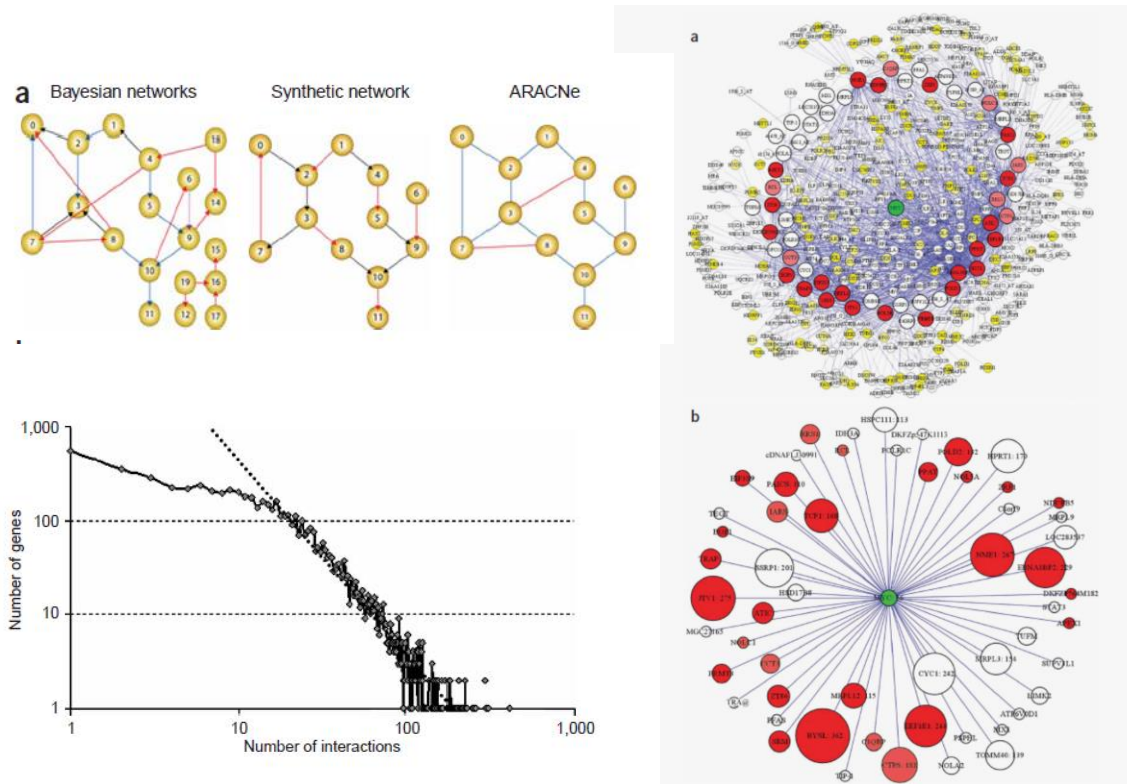# Visualization of reads alignment with genome reference (zoomed)

Alignment coverage of normal and cancer if gene LEF1 (upregulated in cancer)

Gene Regulatory Network of WNT Signaling pathway with RNA Seq. Data

Complex Network Analysis

Reverse engineering of regulatory networks in human B cells

Identification of a master transcription regulation factor (MYC) with ARACNe (algorithm for the reconstruction of accurate cellular networks) which uses mutual information to eliminate indirect interactions.

Basso K, Margolin AA, …, Dalla-Favera R, Califano A: Reverse Engineering of Regulatory Networks in Human B Cells, Nature Genetics, vol.37, no.4, pp.382-389 (2005)

Comparison between Bayesian network and ARACNe on the synthetic network

a — Bayesian networks, Synthetic network, ARACNe

b — BN sensitivity, BN precision, ARACNe sensitivity, ARACNe precision
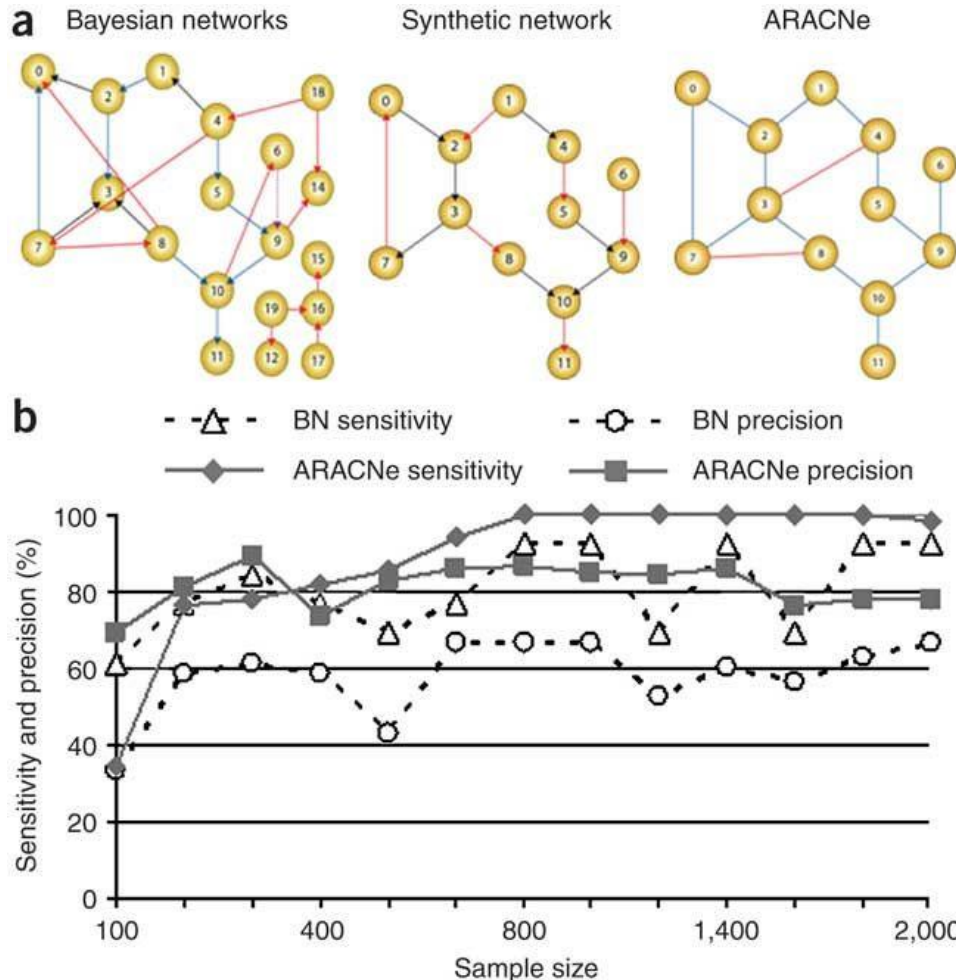
Bayesian network:

Blue arrows: correct detection
Red arrows: incorrect detection
Black arrows: correct but opposite direction

Synthetic network:

Black arrows: up regulation
Red arrows: down regulation

ARACNe:

Blue lines: correct detection
Red lines: incorrect detection
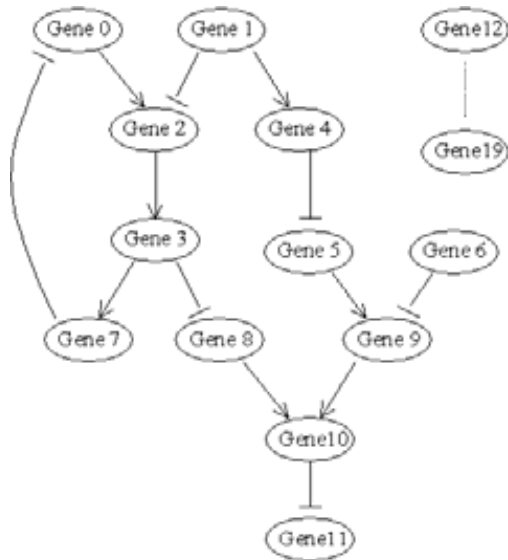
Gene simulator

$$Y_t+1 = f(Y_t) = A(Y_t - T) + \varepsilon$$

Y_t : gene expression level at t (from 0 to 100)
A : a matrix for gene interaction
T : a vector of threshold regulating values (50)
ε : white noise (-10 to 10)

Good for test data generation and
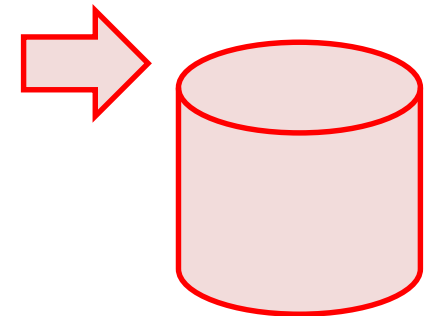evaluation of network extraction algorithms

Given gene network

Generated gene expression
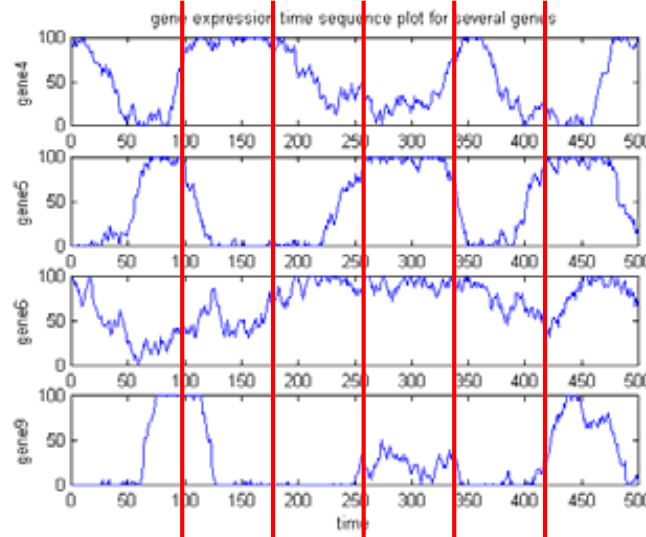
Generated
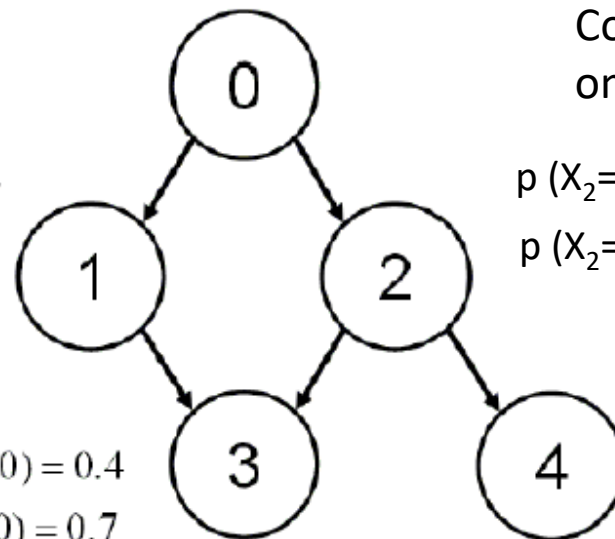Time-course
Gene expression
samples

Jing Yu, et al: Using Bayesian Network Inference Algorithms
to Recover Molecular Genetic Regulatory Networks, in 3rd International
Conference on Systems Biology, Sweden (2002)

Bayesian Network

$$p(x_1, x_2, \cdots, x_N) = \prod_{i=i}^{N} p(x_i | x_1, x_2, \cdots, x_{i-1})$$

Good for node dependencies

$$p(x_0 = 1) = 0.5$$

Conditional probability of node 2
only depend on the state of node 1.

$p(x_1 = 1 | x_0 = 0) = 0.2$

$p(x_1 = 1 | x_0 = 1) = 0.6$

p (X_2=1| x_0 = 0) = 0.5

p (X_2=1| x_0 = 1) = 0.2

Conditional probability of node 4
only depend on the state of node 2.

$p(x_3 = 1 | x_1 = 0, x_2 = 0) = 0.4$

$p(x_3 = 1 | x_1 = 1, x_2 = 0) = 0.7$

$p(x_3 = 1 | x_1 = 0, x_2 = 1) = 0.2$
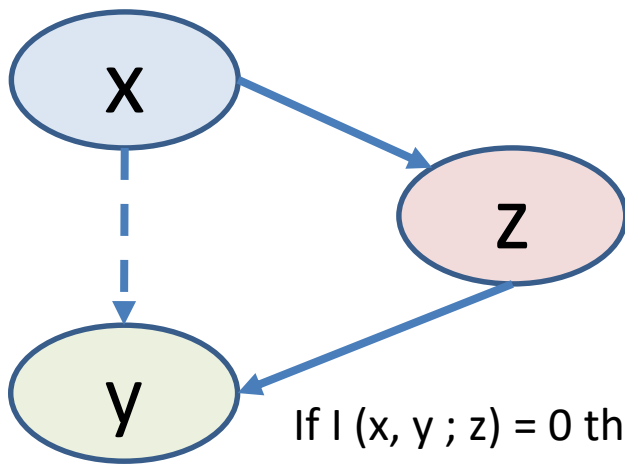
$p(x_3 = 1 | x_1 = 1, x_2 = 1) = 0.4$

$p(x_4 = 1 | x_2 = 0) = 0.7$

$p(x_4 = 1 | x_2 = 1) = 0.4$

Conditional probability of node 3
only depend on the state of node 1
and node 2.

ARACNe : Improves gene network considering "indirect interaction" between genes

Two genes, x and y, as indirectly interacting through a third gene, z, if the conditional mutual information $I(x,y|z)$ is equal to zero.

If I (x, y ; z) = 0 there is an indirect link between X and Y

$I(X; Y) = H(X) - H(X|Y)$
$= H(Y) - H(Y|X)$
$= H(X) + H(Y) - H(X,Y)$

H(X) : entropy of X; ambiguity of X
H(X|Y) : conditional entropy of X of Y; ambiguity of X when Y is available

X, Y : independent
⇔ H(X|Y) = H(X)
⇔ I(X; Y) = 0

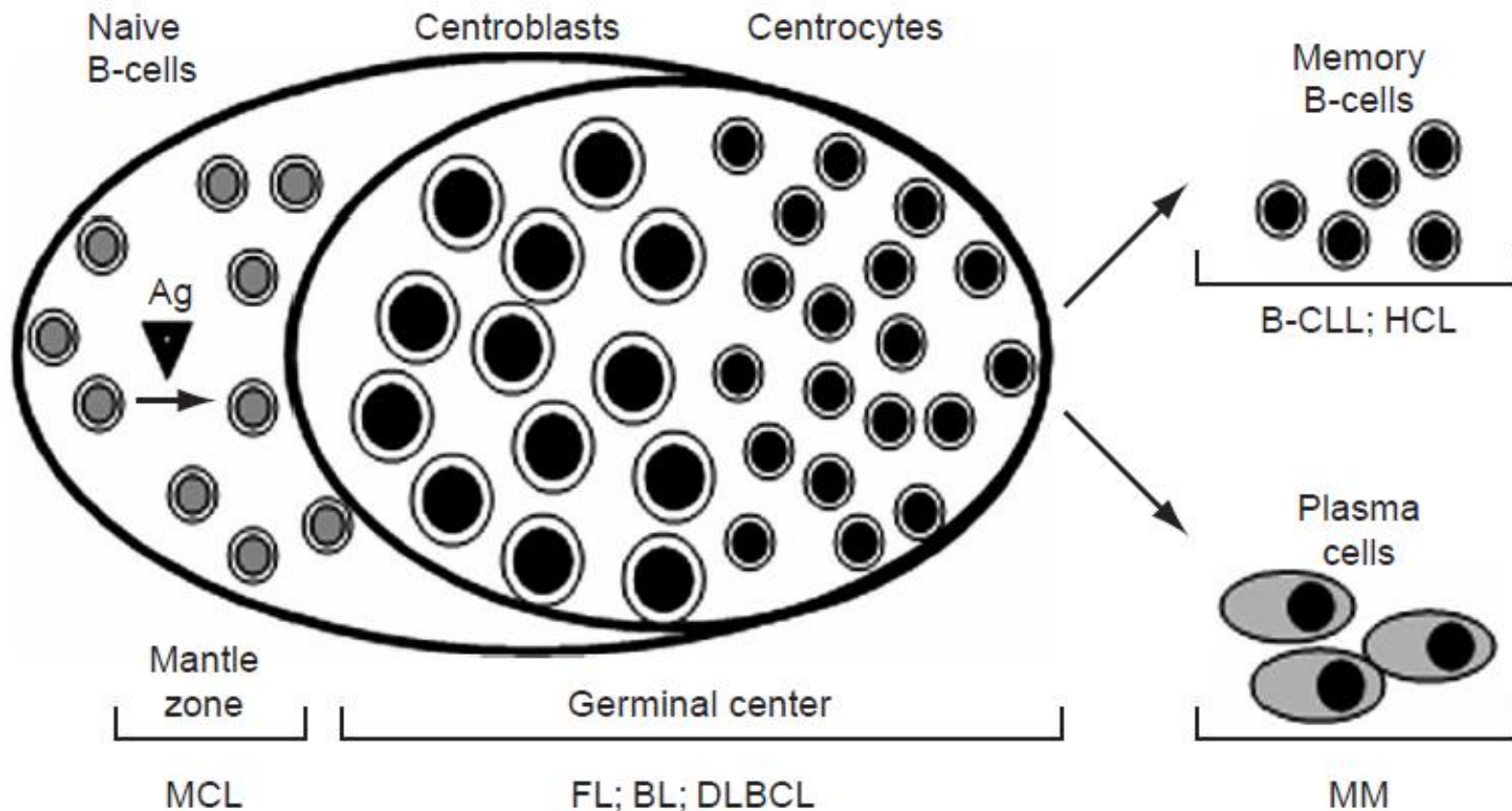Ambiguity of X does not decrease even if we know Y

B-cell lymphoma（B細胞リンパ腫）

https://www.youtube.com/watch?v=Y-HaTwHdSM4 (1m 37s)

Data set
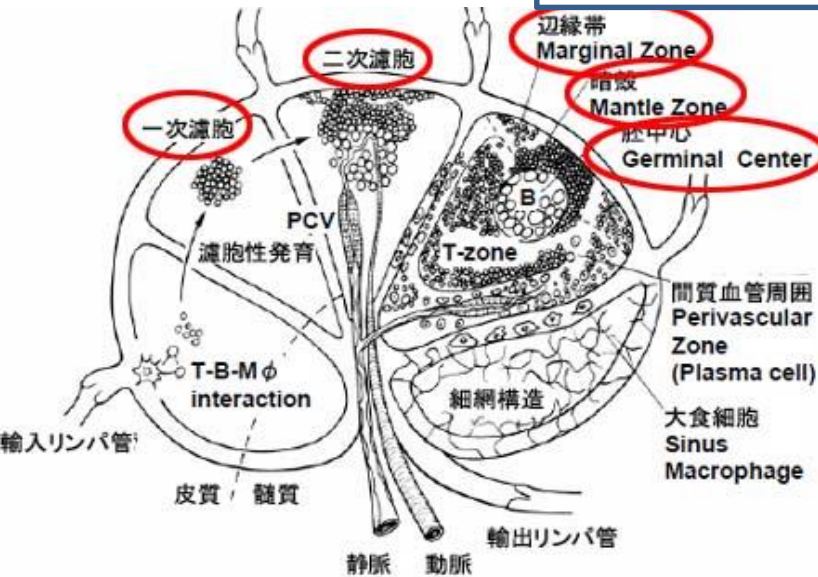
Genome-wide expression profiles from a panel of 336 B cell phenotypes to produce antibodies with high affinity

Normal　B cell
正常B細胞

Cancer　B cell
癌B細胞

Centroblasts
胚中心細胞

Centrocites
中心細胞

Mantle zone
マントル細胞

Naive B cell
抗原未認識B細胞

Cord blood
へその緒

B cell　unmutated
非変異B細胞

Purified B cell unmutated
精製非変異B細胞

B cell　mutated
変異B細胞

Supplementary Table 1. List of samples included in the gene expression profiles database	
Sample code	Description
Normal B cells	
CB 2-23	purified CB
CB 3-10	purified CB
CB 3-30	purified CB
CB 3-7	purified CB
CB 6-8	purified CB
CC 3-28	purified CC
CC 4-14	purified CC
CC 4-6	purified CC
CC 7-25	purified CC
CC 7-7	purified CC
M 4-12	purified M
M 4-14	purified M
M 4-26	purified M
M 5-2	purified M
M 6-8	purified M
N 4-13	purified N
N 4-14	purified N
N 4-7	purified N
N1 4-21	purified N
N2 4-21	purified N
Cord 26-2	purified B cells from Cord blood
Cord 511	purified B cells from Cord blood
Cord 512	purified B cells from Cord blood
Cord 1171	purified B cells from Cord blood
Cord 1193	purified B cells from Cord blood
Primary tumor samples and cell lines	
CLL1269	B-CLL unmutated
CLL1275	B-CLL unmutated
CLL1282	B-CLL unmutated
CLL1318	B-CLL unmutated
CLL1310	B-CLL unmutated
CLL4	B-CLL unmutated
CLL14	B-CLL unmutated
CLL1304	B-CLL unmutated
P-CLL1268	purified B-CLL unmutated
P-CLL1285	purified B-CLL unmutated
P-CLL1302	purified B-CLL unmutated
P-CLLB3	purified B-CLL unmutated
P-CLLB4	purified B-CLL unmutated
P-CLLB7	purified B-CLL unmutated
P-CLLB14	purified B-CLL unmutated
P-CLLB15	purified B-CLL unmutated
CLL1208	B-CLL mutated
CLL1255	B-CLL mutated
CLL1259	B-CLL mutated
CLL1333	B-CLL mutated
CLL1270	B-CLL mutated

A MYC-specific subnetwork was obtained by including all the genes that have $P < 10^{-7}$
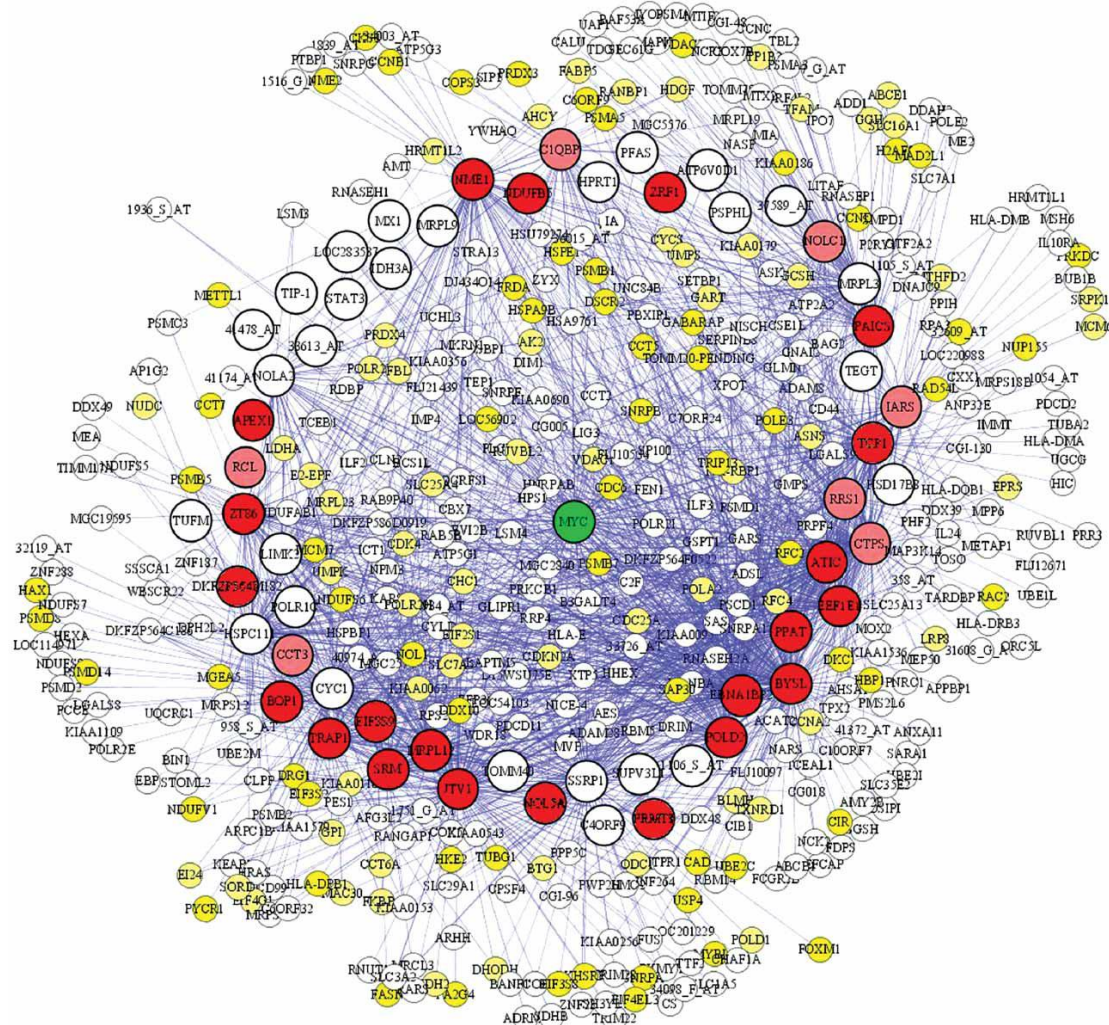
56 genes directly connected (first neighbors; red, pink, white)

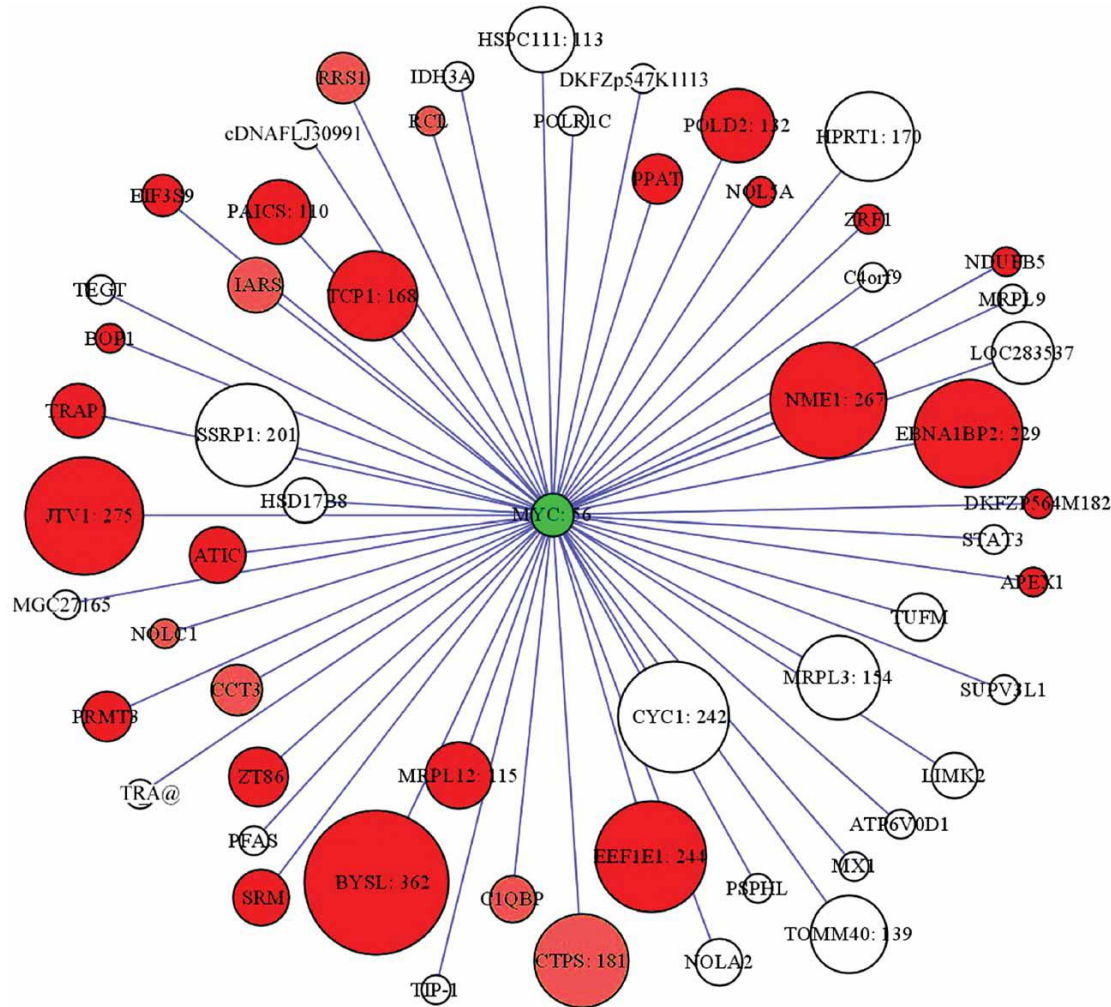444 most statistically significant genes (second neighbors; yellow)

Green node: MYC

Red node: ChIP seq data available
Pink node: not available

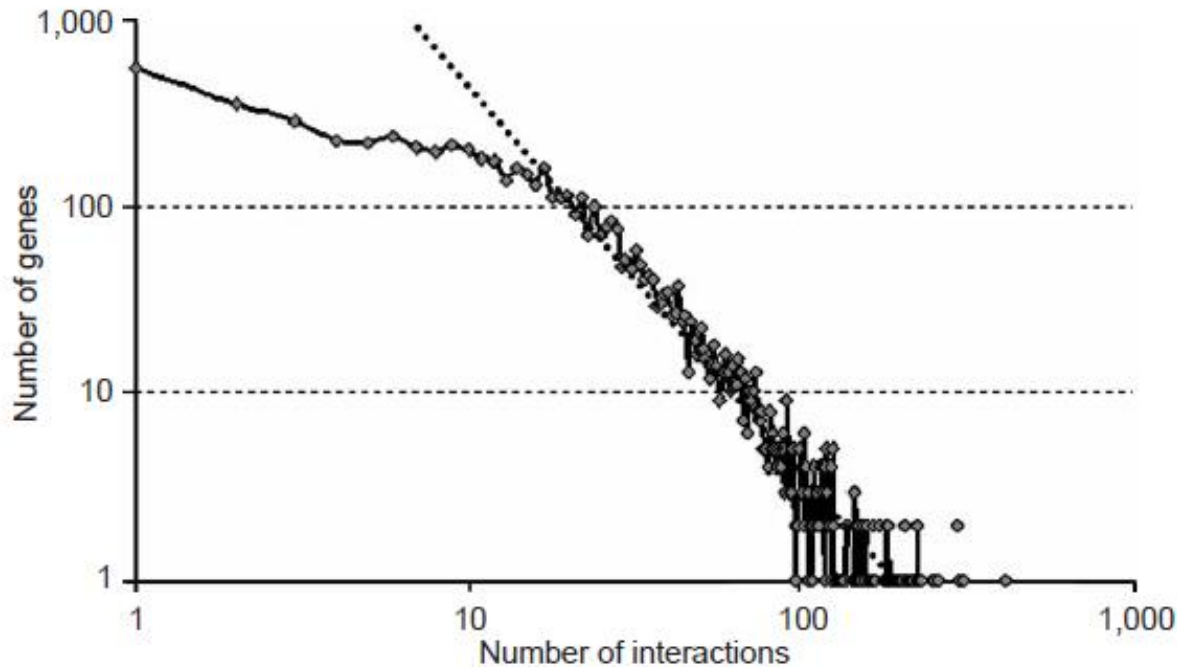White nodes: no information available about MYC relationship

Hub of hubs

MYC plays a role of hub of hubs which regulates more than 2000 genes in total.

The sizes of circles are proportional to the number of connections.

Power law in gene expression regulatory network

Log-log scale plot with regards to the number of genes and interactions

Summary

- Gene expression is regulated by complex network.

- Estimation of causality is more important than finding co-expressions.

- Bayesian Network and mutual information are effective to extract gene expression regulatory networks.

- Complex network analysis is also applicable to gene expression regulatory networks.

Report Issues

- Due date　　Jan 25th（Thu）Noon
- Submission through OCW
- Subject　　Protein-protein interaction network
- Style　a PDF file A4　1 or 2 pages

Issue
  - Impression or investigation report on the issue of gene expression regulatory network held on Jan. 18th.