

STATS M254 / BIOINFO M271 / BIOMATH M271
Statistical Methods in Computational Biology
Spring 2014

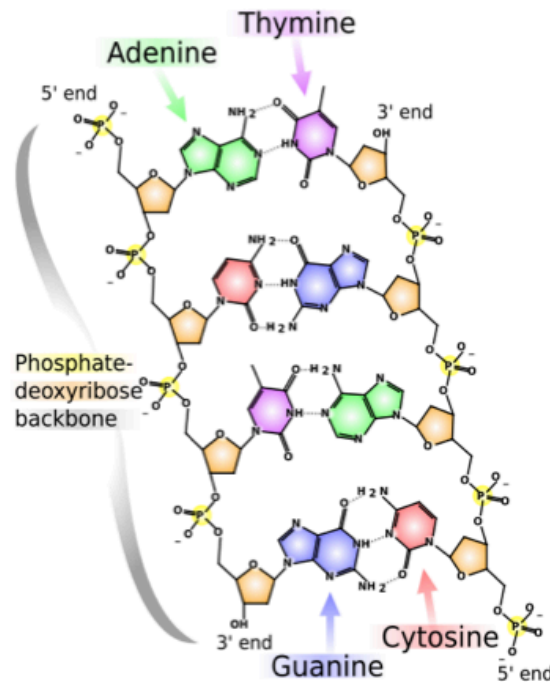
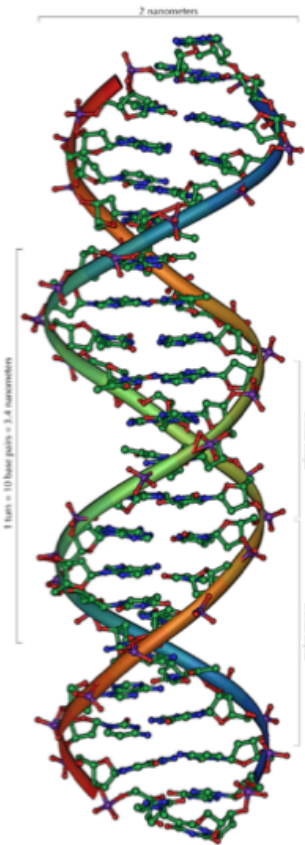
Lecture 1 Introduction and Data

Instructor: Jingyi Jessica Li

Outline

- Introduction to molecular biology
 - DNA, gene, RNA, protein, central dogma
- Typical data
 - Gene expression
 - RNA-seq
 - Regulatory sequences
 - ChIP-chip/seq

DNA

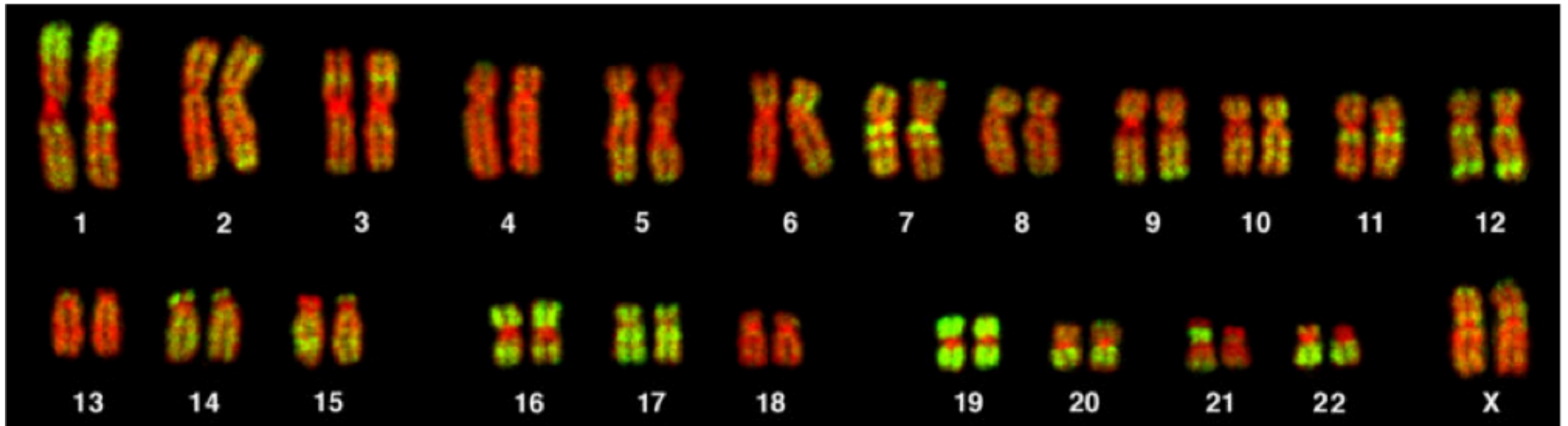


- DNA (Deoxyribonucleic acid) is a molecule to store genetic information of a living organism.
- DNA consists of two polymers made from four types of nucleotides: adenine (A), guanine (G), cytosine (C) and thymine (T).
- Purines: A, G; Pyrimidines: C, T
- Two polymers are complementary to each other and form a double-helix structure

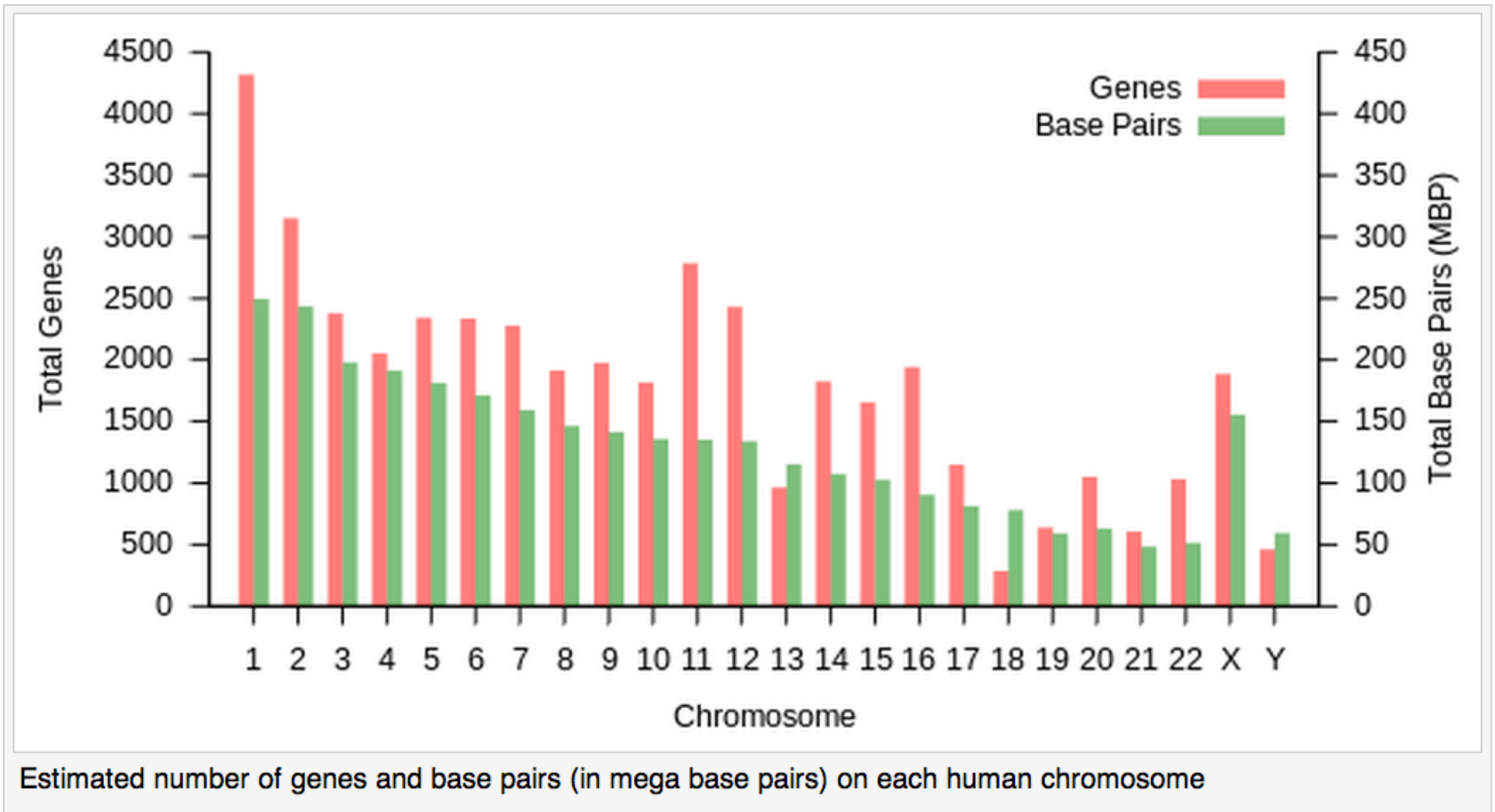
```

5' -ACCGTTCGACGGTAA-3'
    |||||
3' -TGGCAAGCTGCCATT-5'
    
```

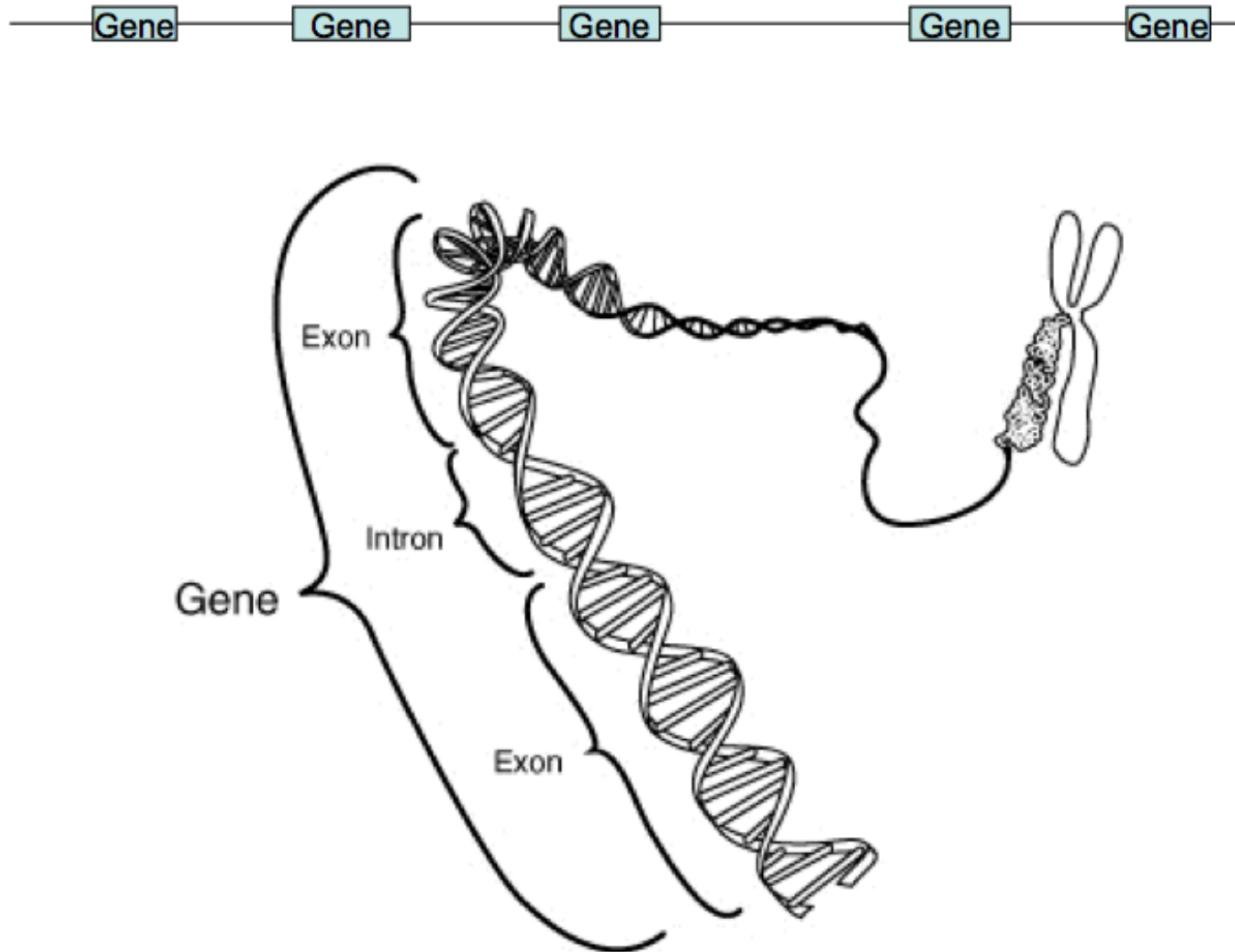
Chromosome



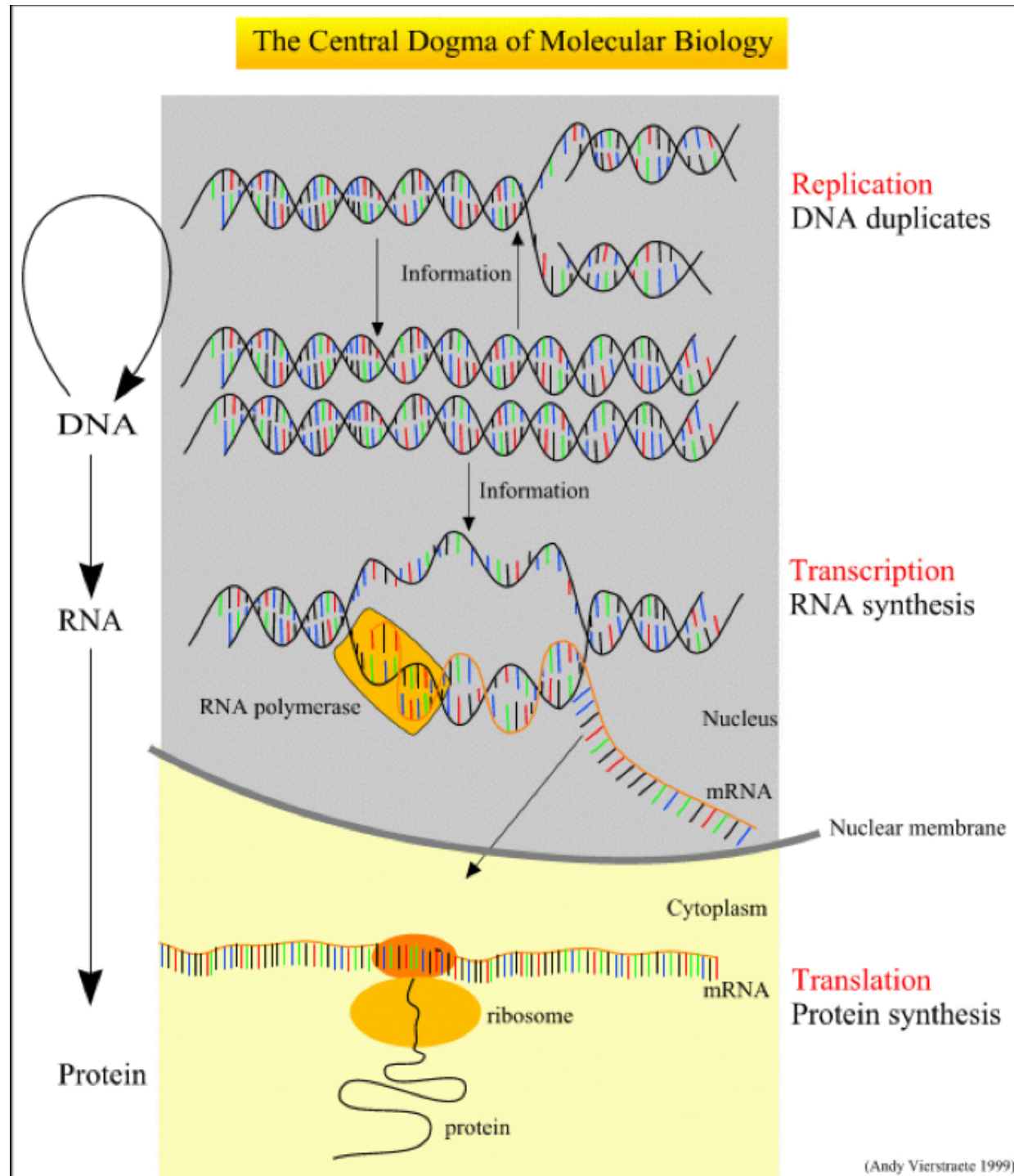
Chromosome



Gene



Central Dogma

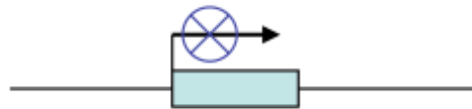


Data type 1: Gene expression data

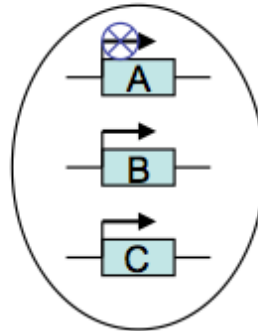
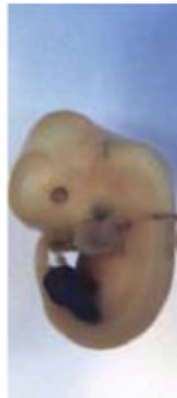
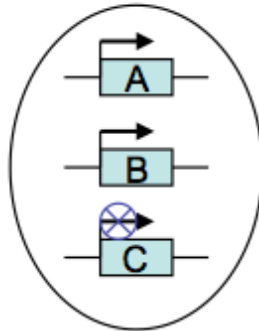
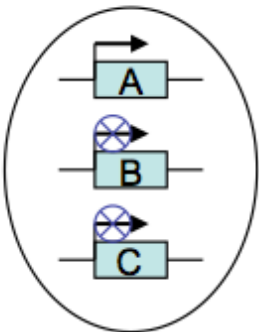
Expression



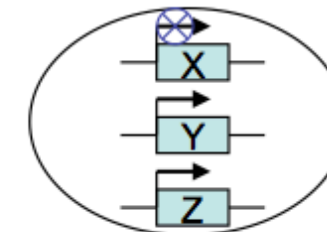
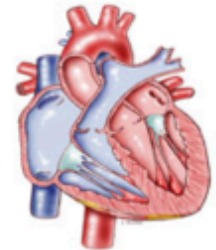
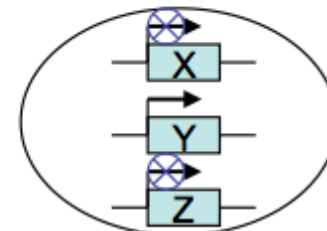
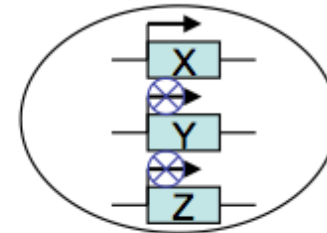
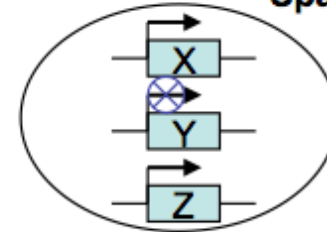
No Expression



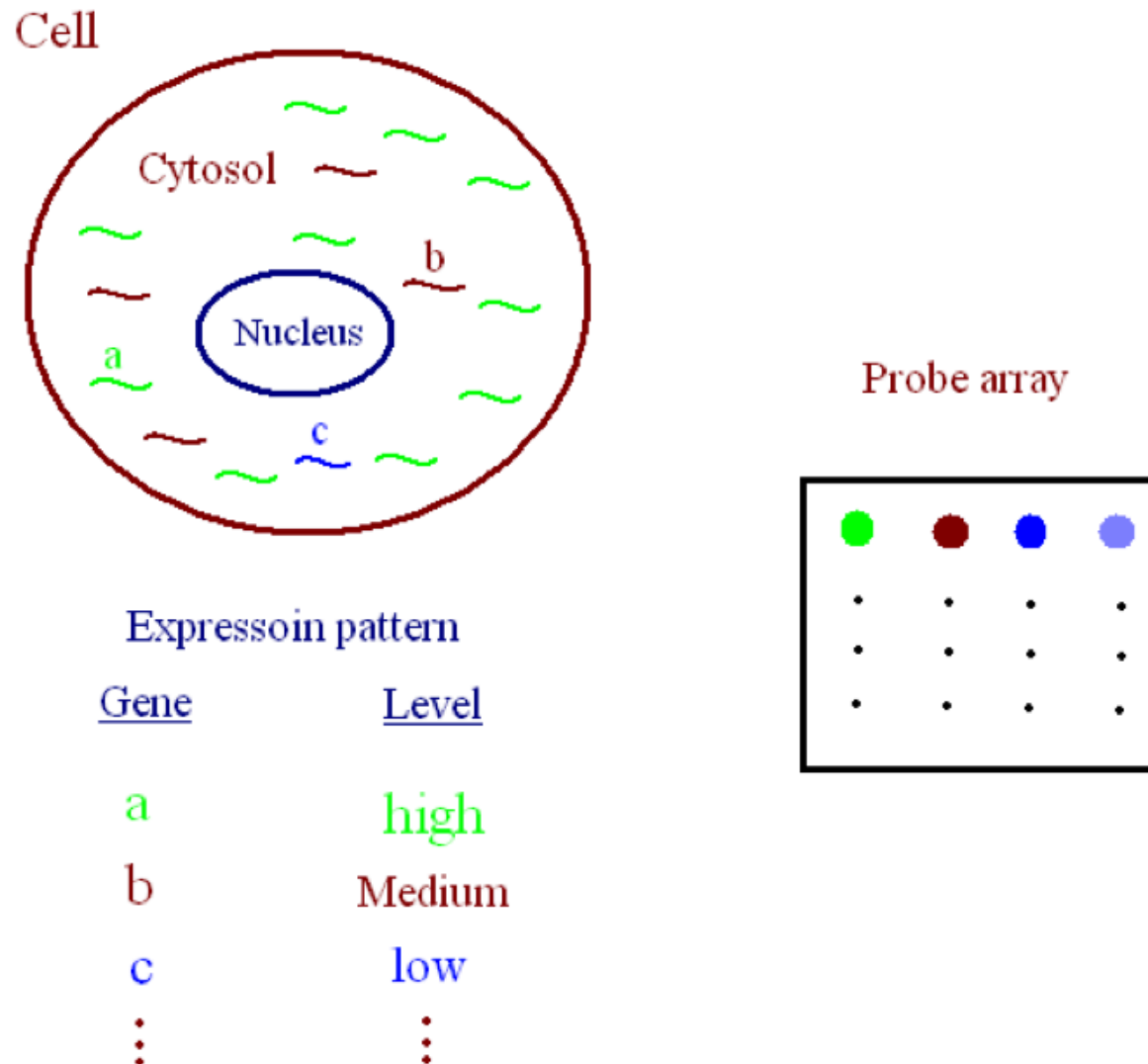
Temporally

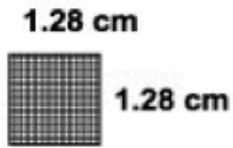


Spatially

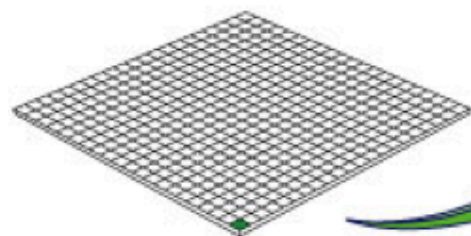
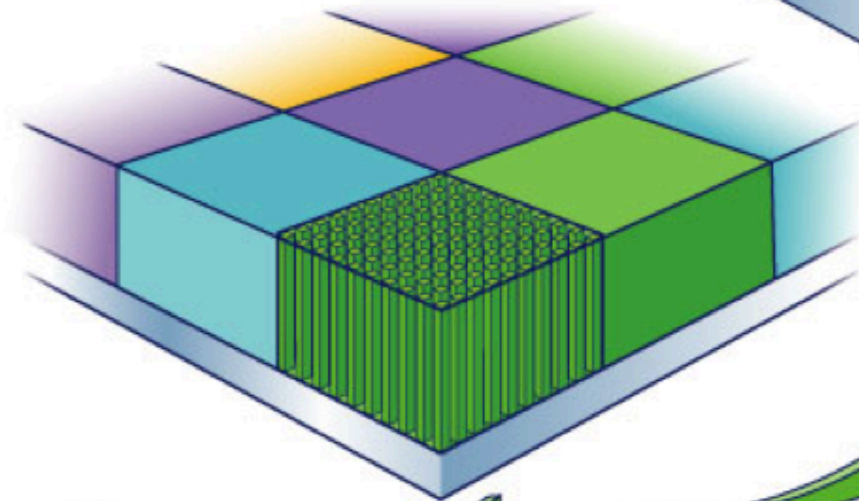
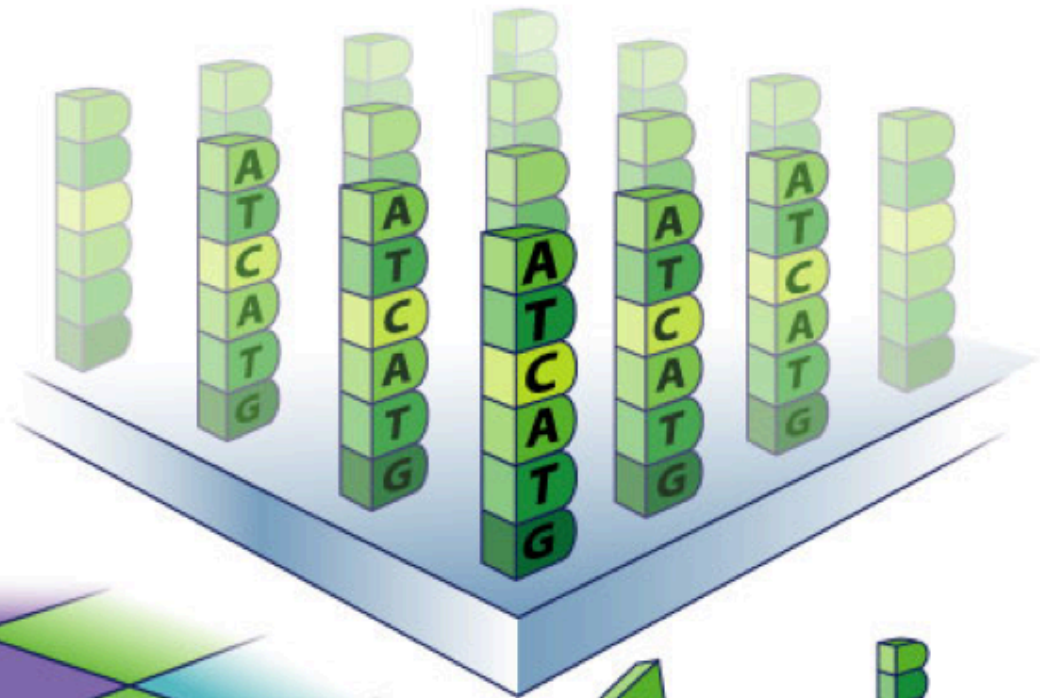


Principle of gene expression microarray



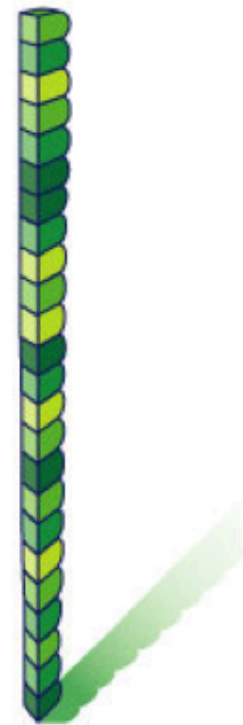


Actual size of GeneChip™



500,000 cells on each GeneChip™ array

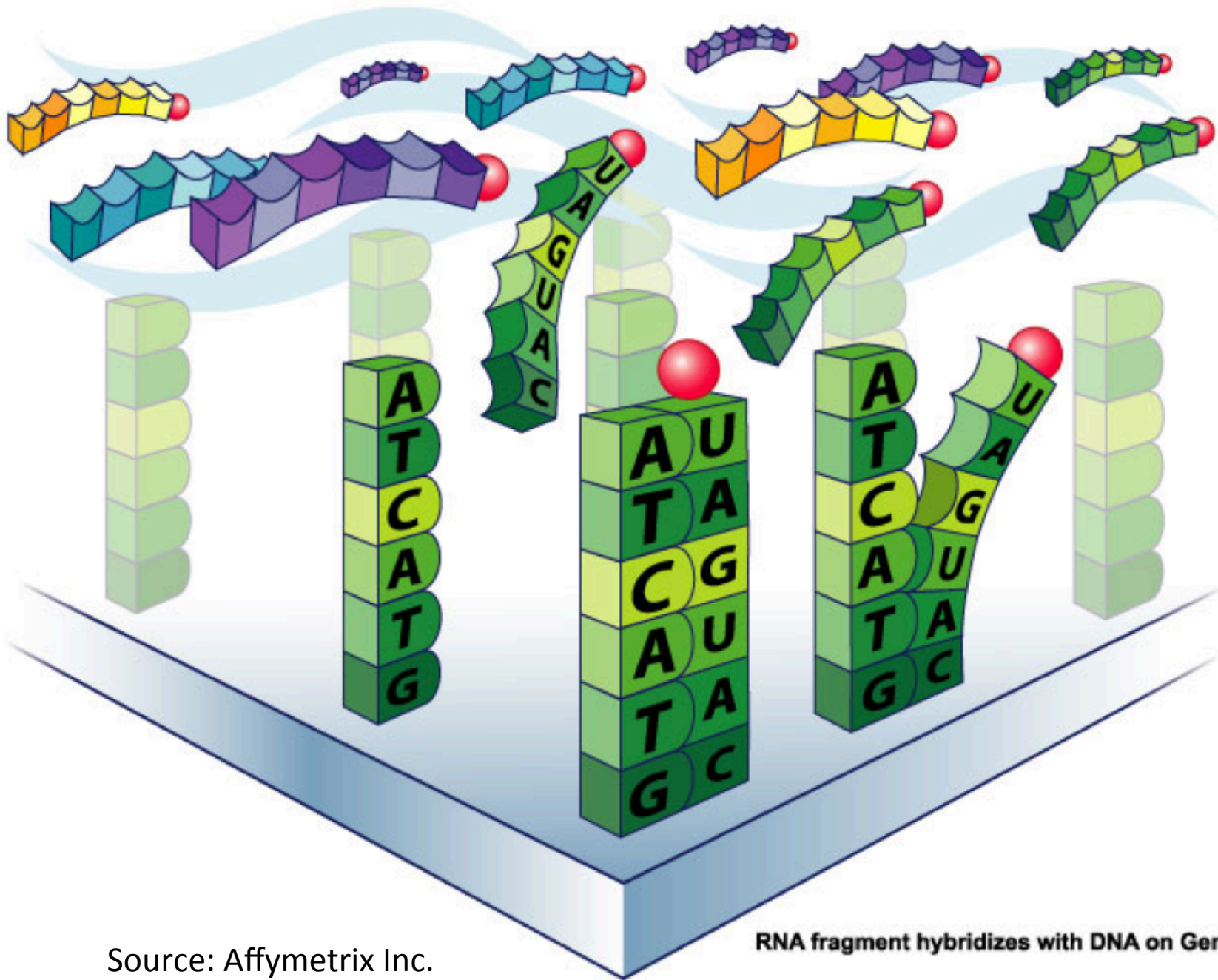
Millions of DNA strands built up in each cell



Actual strand = 25 base pairs

Source: Affymetrix Inc.

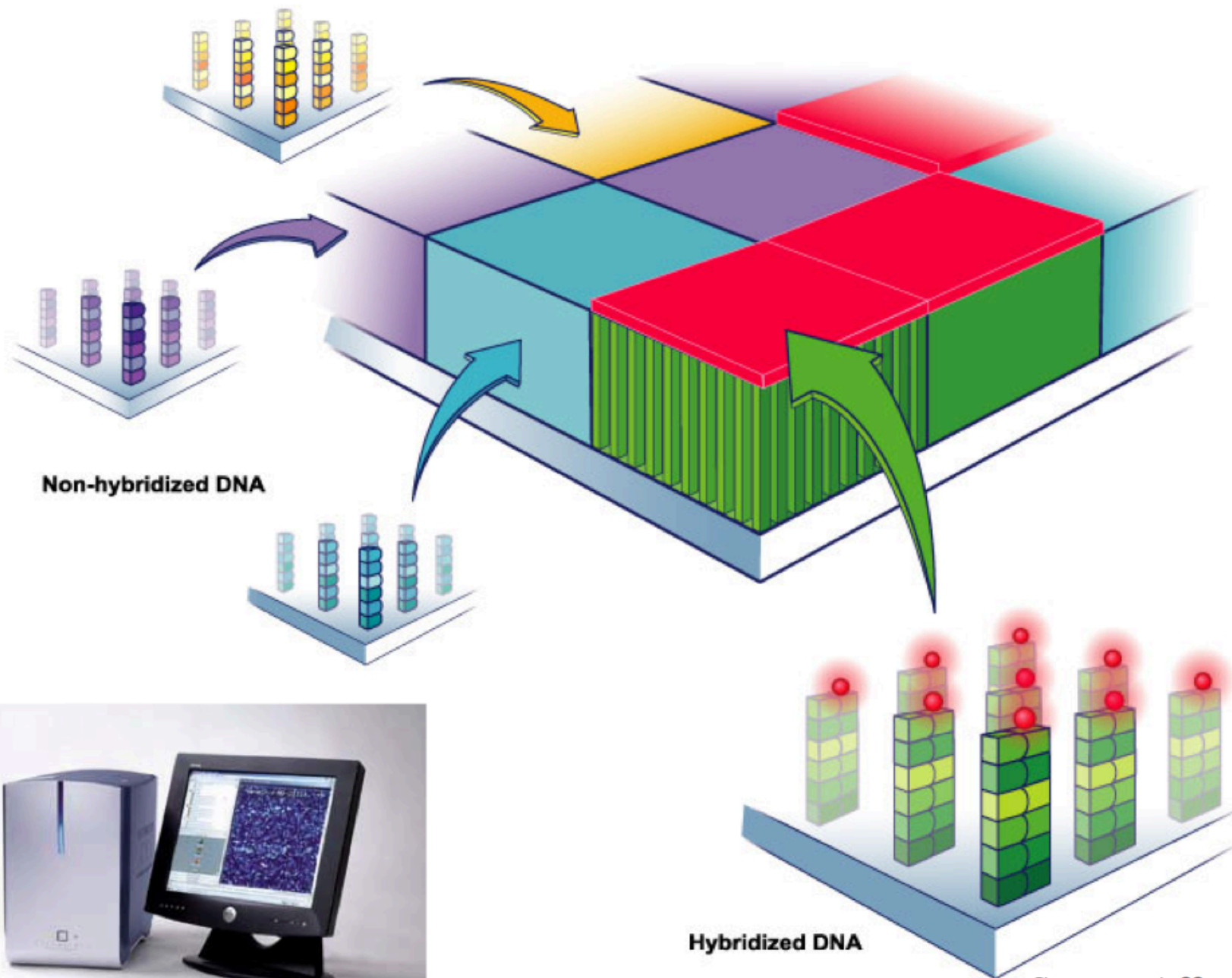
RNA fragments with fluorescent tags from sample to be tested



Source: Affymetrix Inc.

RNA fragment hybridizes with DNA on GeneChip¹¹

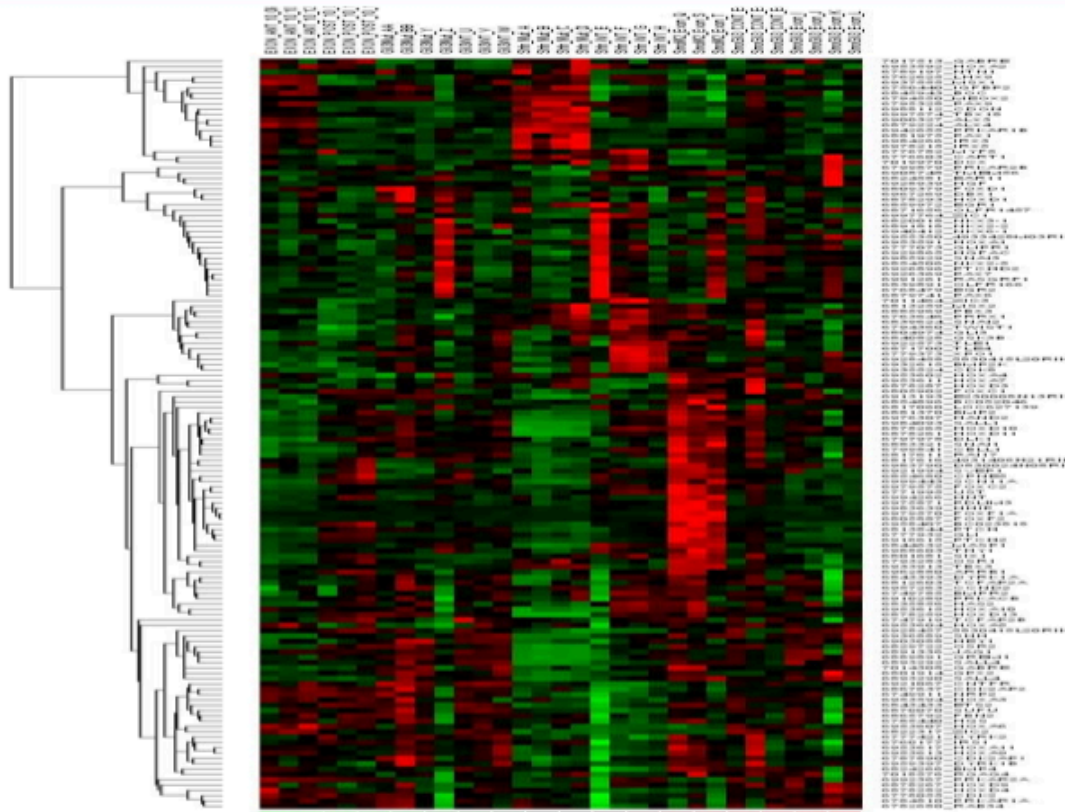
Shining a laser light at GeneChip causes tagged DNA fragments that hybridized to glow



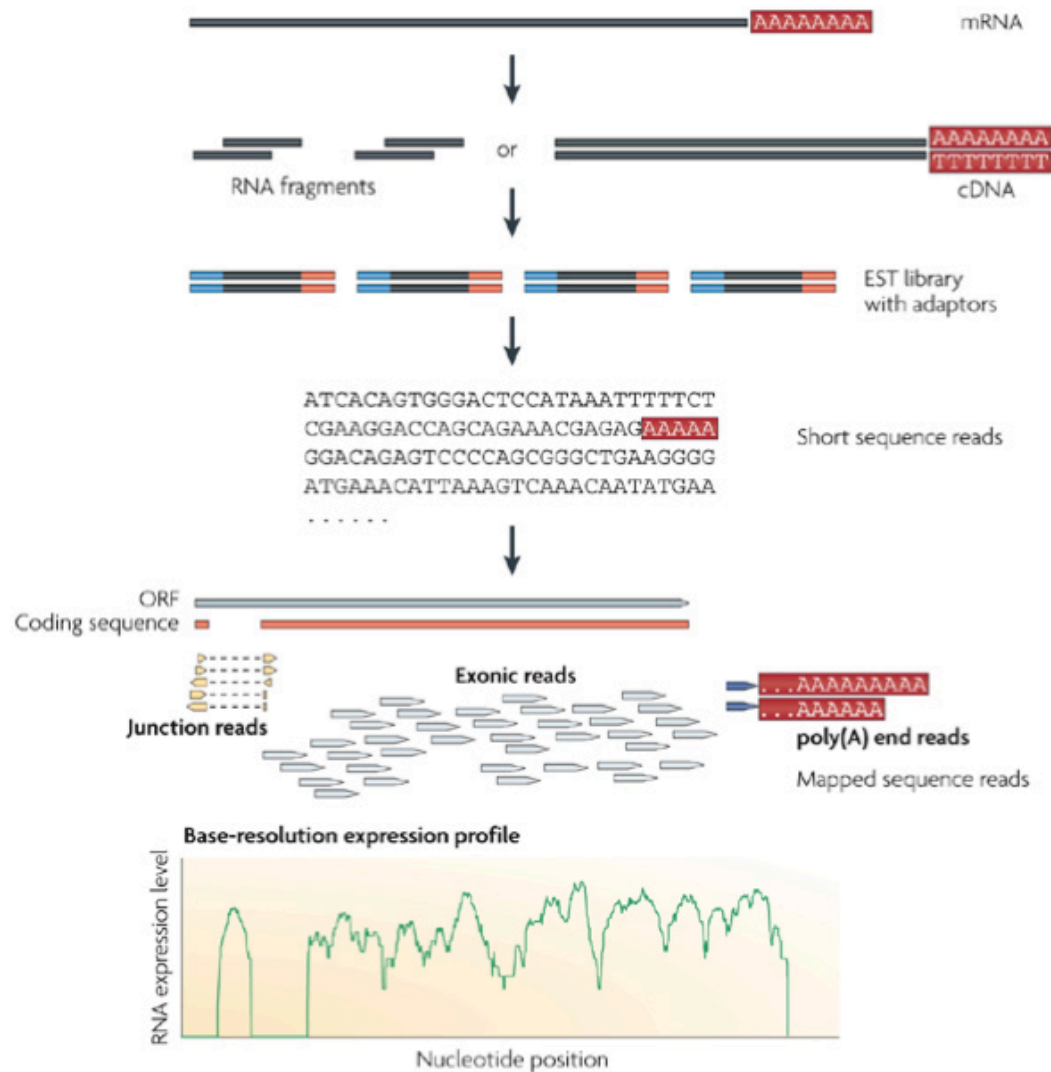
Source: Affymetrix¹² Inc.

Gene expression data matrix

Condition	Mutant			Wild Type		
Replicate	1	2	3	1	2	3
Gene 1	132.724	112.445	128.478	154.888	122.215	138.303
Gene 2	161.825	163.304	210.121	159.003	172.366	163.199
...						
Gene I	1988.66	2063.48	1899.91	1997.77	2156.19	1977.75



RNA-Seq data



Wang, Gerstein, & Snyder (2009) *Nature Reviews Genetics* 10, 57-63.

RNA-Seq data

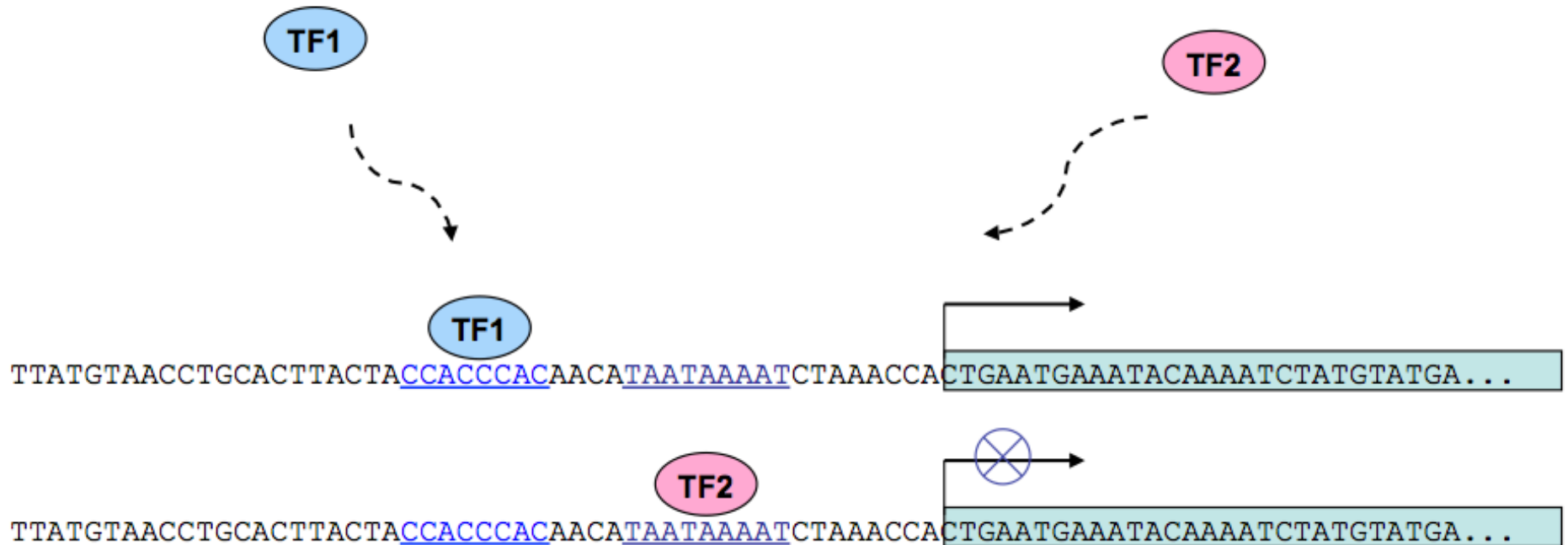
- 1) Long RNAs are first converted into a library of cDNA fragments through either RNA fragmentation or DNA fragmentation.
- 2) Sequencing adaptors (blue) are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using high-throughput sequencing technology.
- 3) The resulting sequence reads are aligned with the reference genome or transcriptome, and classified as three types: exonic reads, junction reads and poly(A) end-reads.
- 4) These three types are used to generate a base-resolution expression profile for each gene, as illustrated at the bottom; a yeast ORF with one intron is shown.

Data type 2: Regulatory sequences

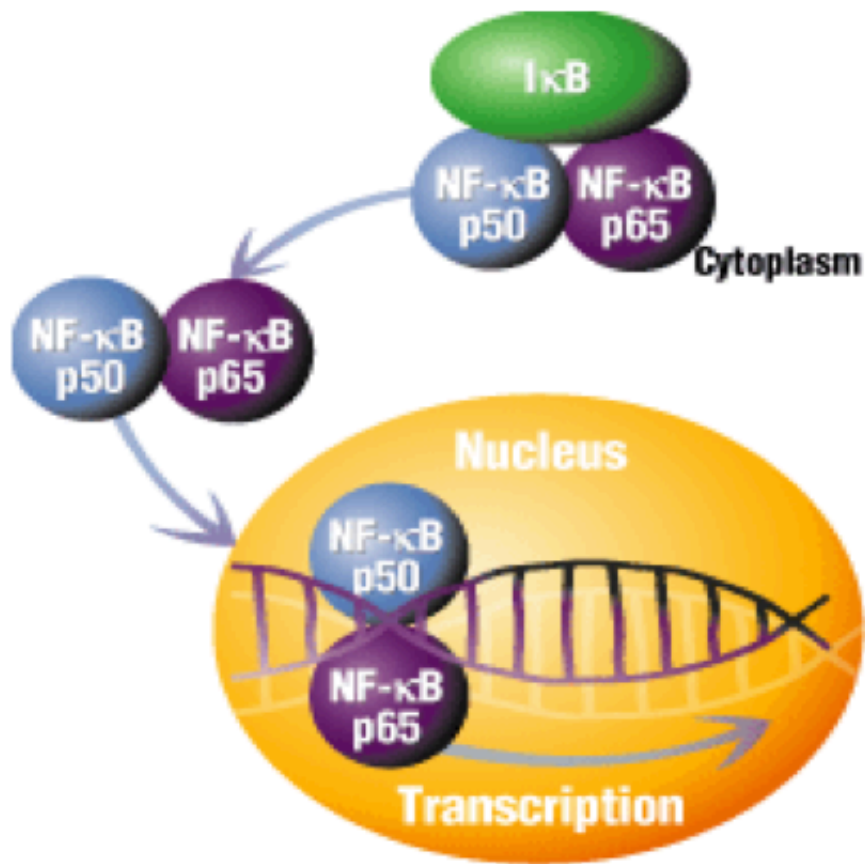
Transcription factors (TF):

TF1

TF2



Transcription factor binding sites & motifs



AATTTC
CATTGCG
ATTGCG
AATTGCA
AATTTCT

A	7	14	0	0	3	0	4
C	6	0	0	0	0	16	2
G	3	0	0	1	8	0	7
T	0	2	16	15	5	0	3



motif

Motifs are regulatory codes

TCAGTTGGAGCTGCTCCCCACGGCCTCTCCTCACATTCCACGTCCTGTAGCTCTATGACCTCCACCTTTGAGTCCCTCCTC
TCACACCACCCATGTTTTGTTTATGAGGATCCTCAAATACCCCGTGATCAGTCTCAGGGTAGCTCTCATAGCCTGGACAGGG
CCCCCTCGGGGGTTGCGCCCAGGTCCAGGCGGGGGATGCACAGCAACAGTCACCGAAGCAGAAGCCGTCACAGTGGTGATG
GGCTGGCAGTAGCTGGGCACAGAGCTGCCCATGGCGGTGGACGTTGGGTTCCGAGGGTTGTGAGAACGGGCCCCACGGGGCC
CTGAGCGGTCCCTATTGCTAGGGCCAGAATGCCCTTCAGTAGAAATTTCAAAGCGTCTCTGCGCGGTCTGTAGGGGGGTGG
CCGCAAGCCTTCTCTAGGGGGATCCCTTCGTTGCTGCTGGCCTTGCCGTCCAGGGGACAAGGAGCCAGAGTCCAGGTGGGGC
TGTTGCCGAGGGGTCAAGGGAGGCTGATGTCTGGAGTCCGGATGGACCACCTGCAGAGGAGAGACATAGGTCAACACAGGGA
GGTAGGATGGTGGTGATGTTCCACCCACAAAAGAAAACCTATTCTTTAGAAACCTCCAGGATGTGAATCCTGCCTGCACCT
GCACAGCTGGCTGGAGGCATATAGCCACTGCCCATAGATCTCAACTTACCCTCACAACCAACTGCCCCCAGGCCTAAGTTCT
CTGCCTCAAACTGCCAAGGCCTGGATAGCCAAGAGCCTGGGTGTCTTGGAATATGCAACCATAAATAGTAGCTTTTAGAA
GTATAAGGCTCCTGTTTCTGGGTCATATTAGTTTTGTTTTCACCTGTCCCCACCCATAAGCCAGGTGTGGCCAGAAGCAAAT
GTACTGTAAGAGCAGAGCAAAAACCTCCACACAGATAGTTCTGTTAGGCAATACATCTCTGCCTGACTATTAGGAATCTGGT
TTCTGGGTCCTCTGTACAAAGCTCGGAGCAACACAGTGGCCACATCAATCAAAGGACCGTGACCAACTTCAAAGTCGGTGA
GCTTGTAACCTATTTTTAGGCTCCTGCTGAACAGAACCAGATTACACTACAGCTCAGCAGGGCATCGTCACGGGTGTGTGTG
TGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGGGGGGGGGGGGTGGACAGAGGACGGGGACACAATTCACTGGCCAGCCCTTC
TCTCCTTCAAGGAAGGCTGCTCTAGCCTGGGACTGGAATACACATTTCTGTAAACATGGTGGGGCCTCAGGCAAGCCAGA
GTTTTGGAGCCTTCCTTAACTCTTCAAGGTGAGCATCTTGACTTGAGGGGTGGGGGTGCGGGTAAGGAAGGAACCTGTGGAC
TCCACCCAACAAGACAGAAAAGGAATAAGCCACGAAGACAATAACGATTTTTGTATCAAGCCTCCTCTCCCATTTTCAGCTTA
CCTGACAATGAAATCAAATTCGGACCTTGCAAGCATCAGTACACCCAGCAGAGTGGACACAGCACCCTCCAGAACGGGAGCA
AACATGTGCTCCAGAGCGAGCATAGCCCTGTGGTTCTTGTCCCCAATGGCTGTCAGAAAGGCCTGAACAAAGGAGAAAATTG
ACACGGTCACATTCTGGGTGTGGTAAAGTGCTCAGCTGTGTCTATACTTGGGTTTTGTAT

Transcription Factor Binding Sites (TFBS)

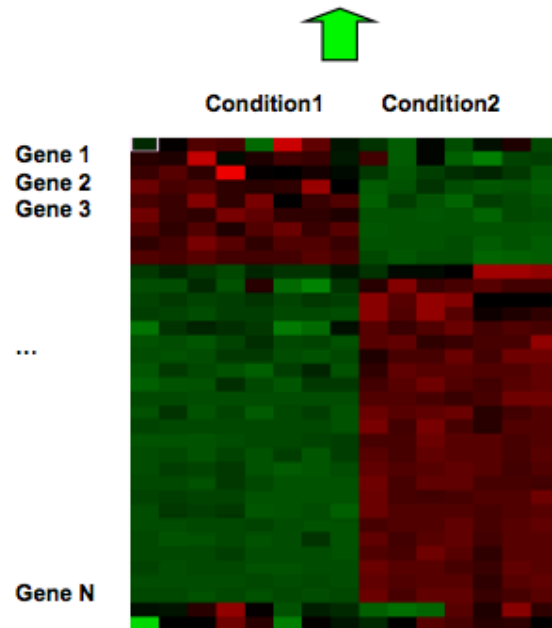
Gene

Finding motifs from co-regulated genes

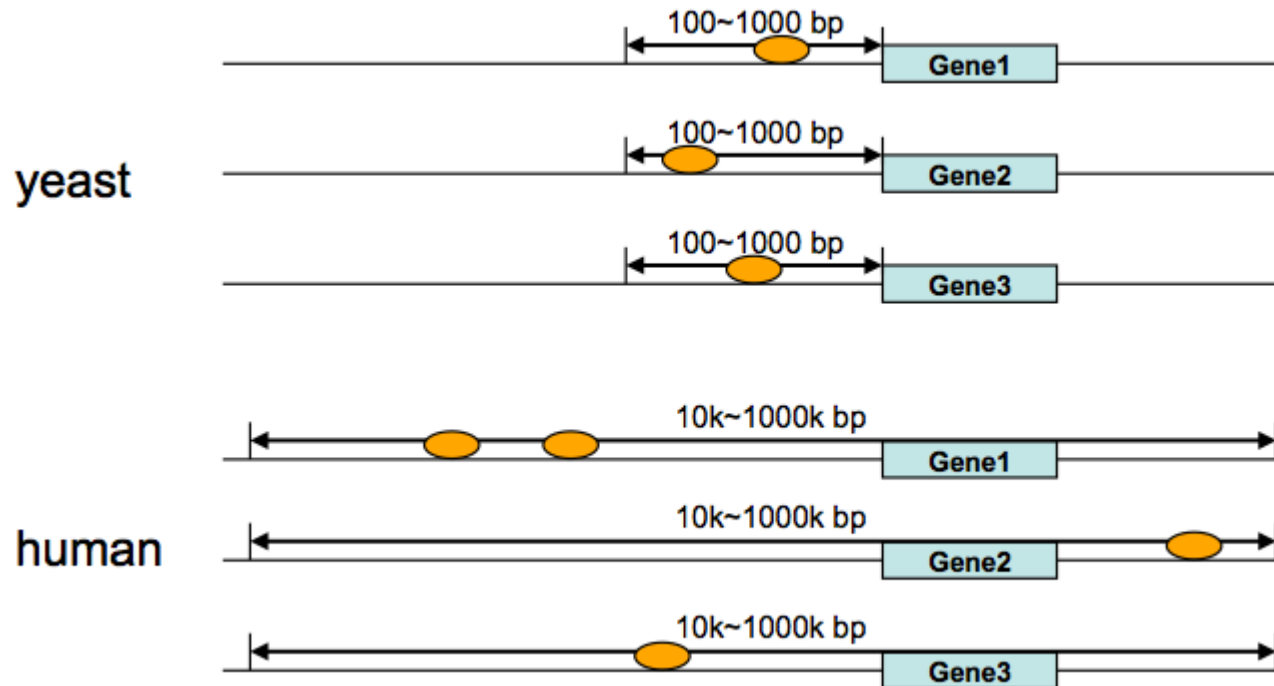
GTATGTACTTACTAAATTGCGAACAAATCTATGTATGAAG **Gene1**
CCATTTCCCTCGGTTTCAGAGTCACAGAGCAGATAATCACC **Gene2**
TAACATGTGACTCCTATAACCTCTTAATTTCGCATGAAGT **Gene3**



GTATGTACTTACTAAATTGCGAACAAATCTATGTATGAAG
CCATTTCCCTCGGTTTCAGAGTCACAGAGCAGATAATCACC
TAACATGTGACTCCTATAACCTCTTAATTTCGCATGAAGT



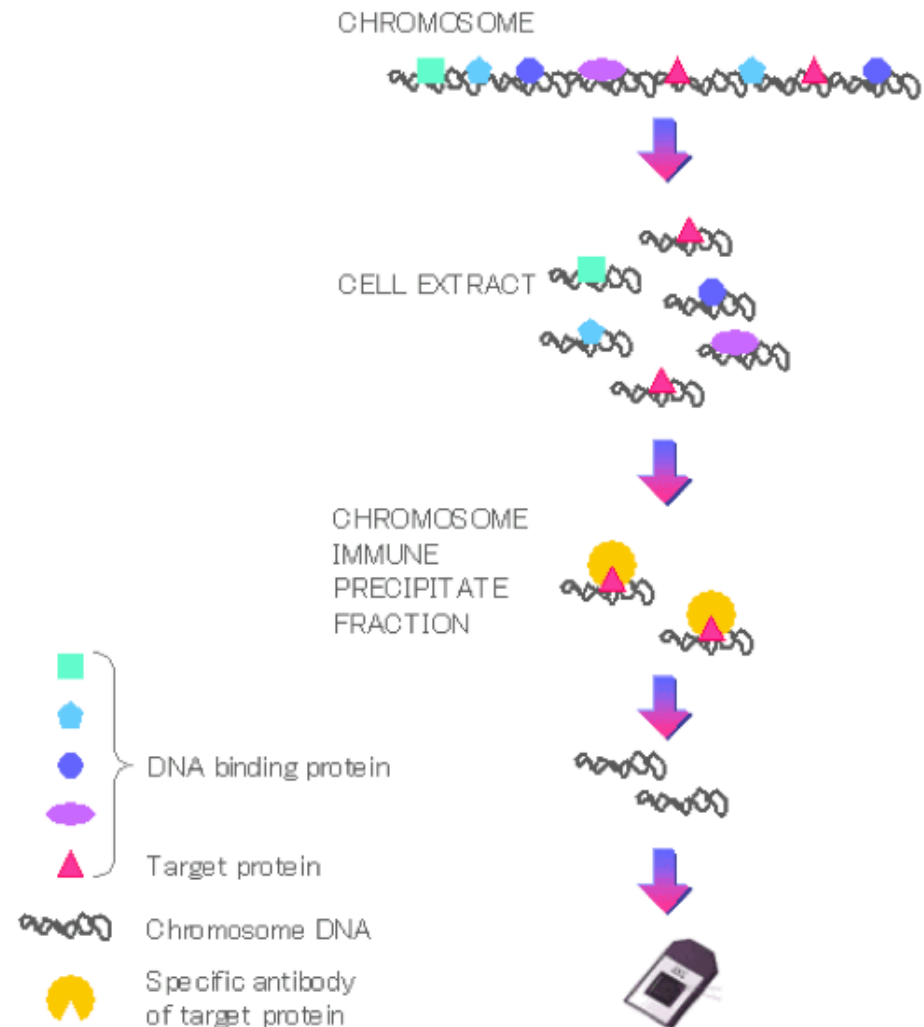
Motif discovery is difficult in mammalian genomes



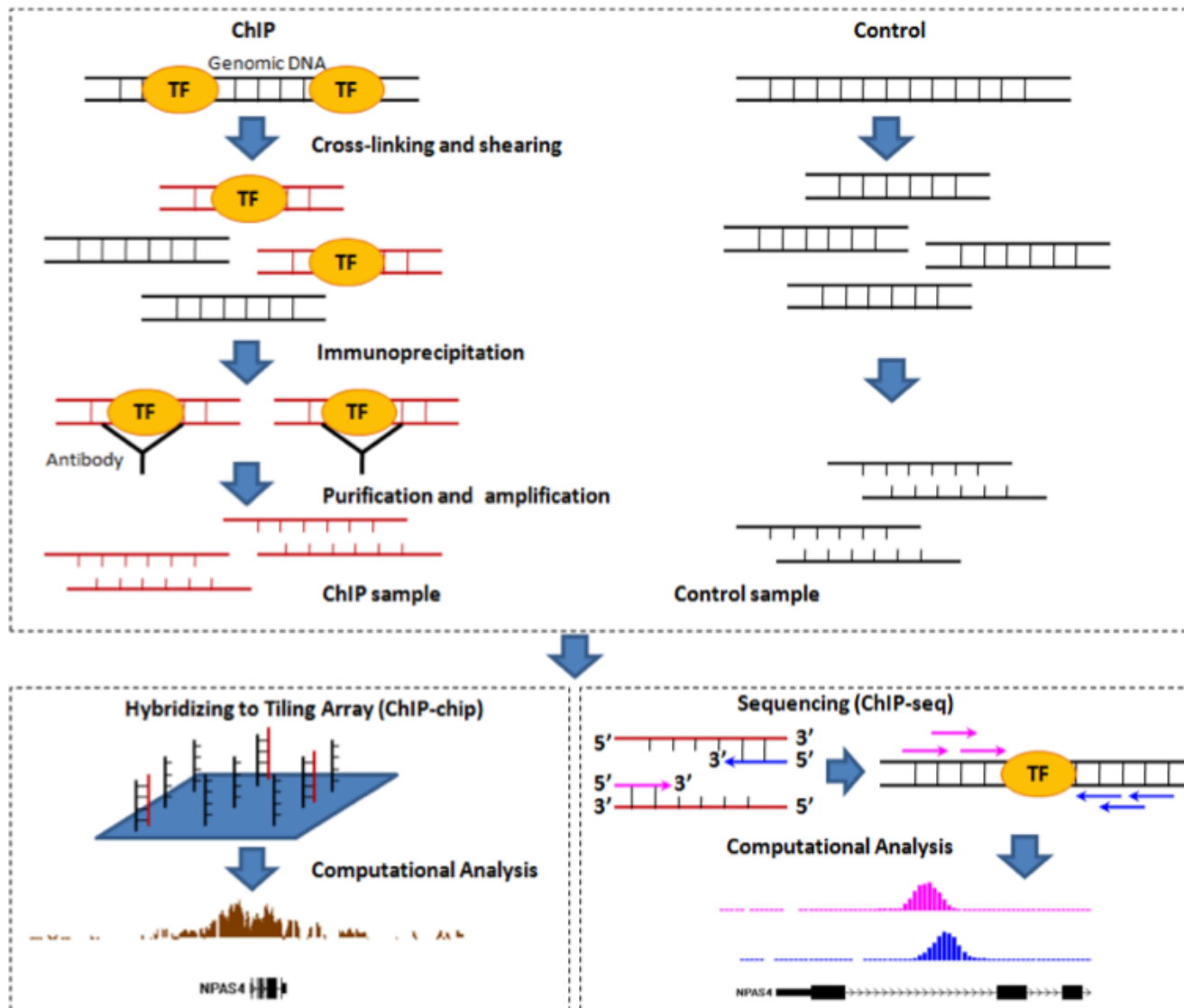
- Advanced methods in regulatory sequence analysis:
 - 1) combinatorial binding pattern
 - 2) multiple species conservation
 - 3) heterogeneity in background
 - 4) predictive modeling

Data type 3: ChIP-chip and ChIP-seq

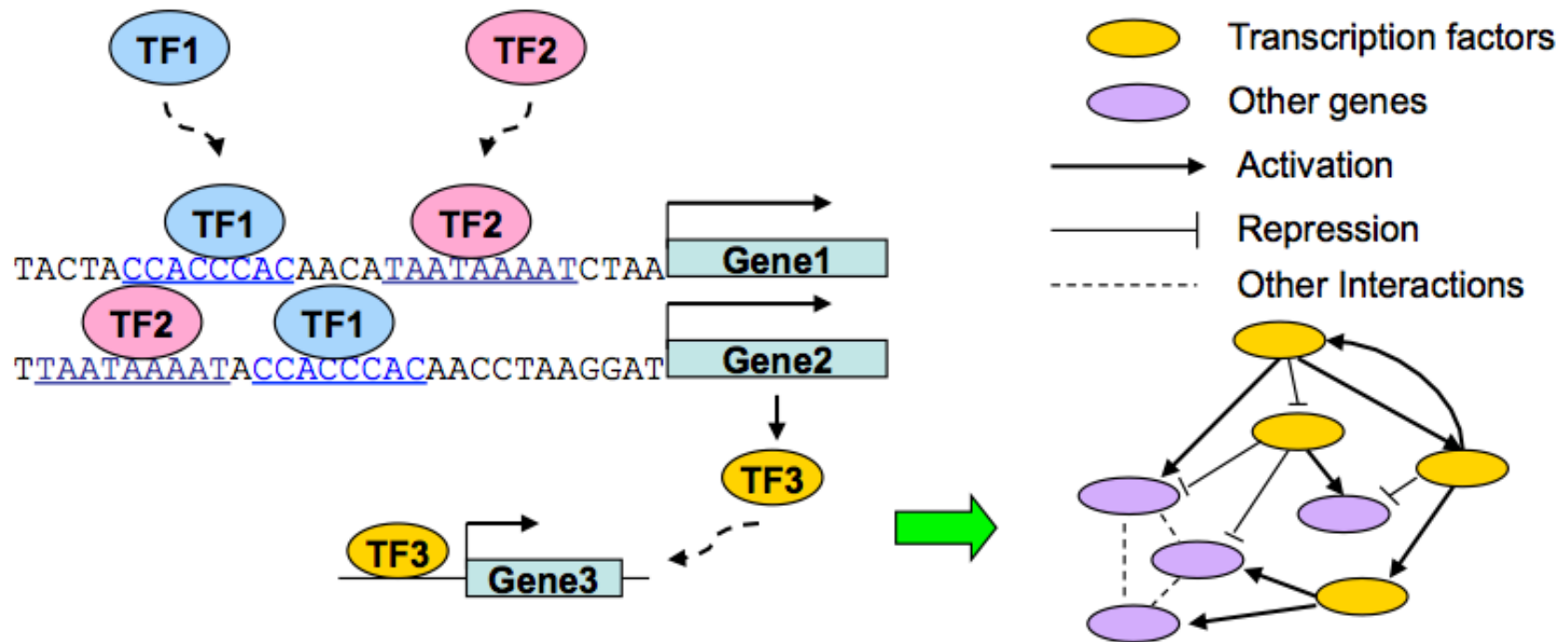
- ChIP: Chromatin ImmunoPrecipitation
- chip: DNA micorarray
- seq: massive sequencing



Array vs. Sequencing



Gene regulatory network



Combine all types of data:

Gene expression, ChIP-chip/seq, regulatory sequences.

Acknowledgments

- For sharing slides on the internet:
 - Dr. Qing Zhou, UCLA
 - Dr. Hongkai Ji, Johns Hopkins University
 - Dr. Cheng Li, Peking University