Applications of Pattern Recognition in Computational Biology

Pattern Recognition Course (2110597)

Chulalongkorn University

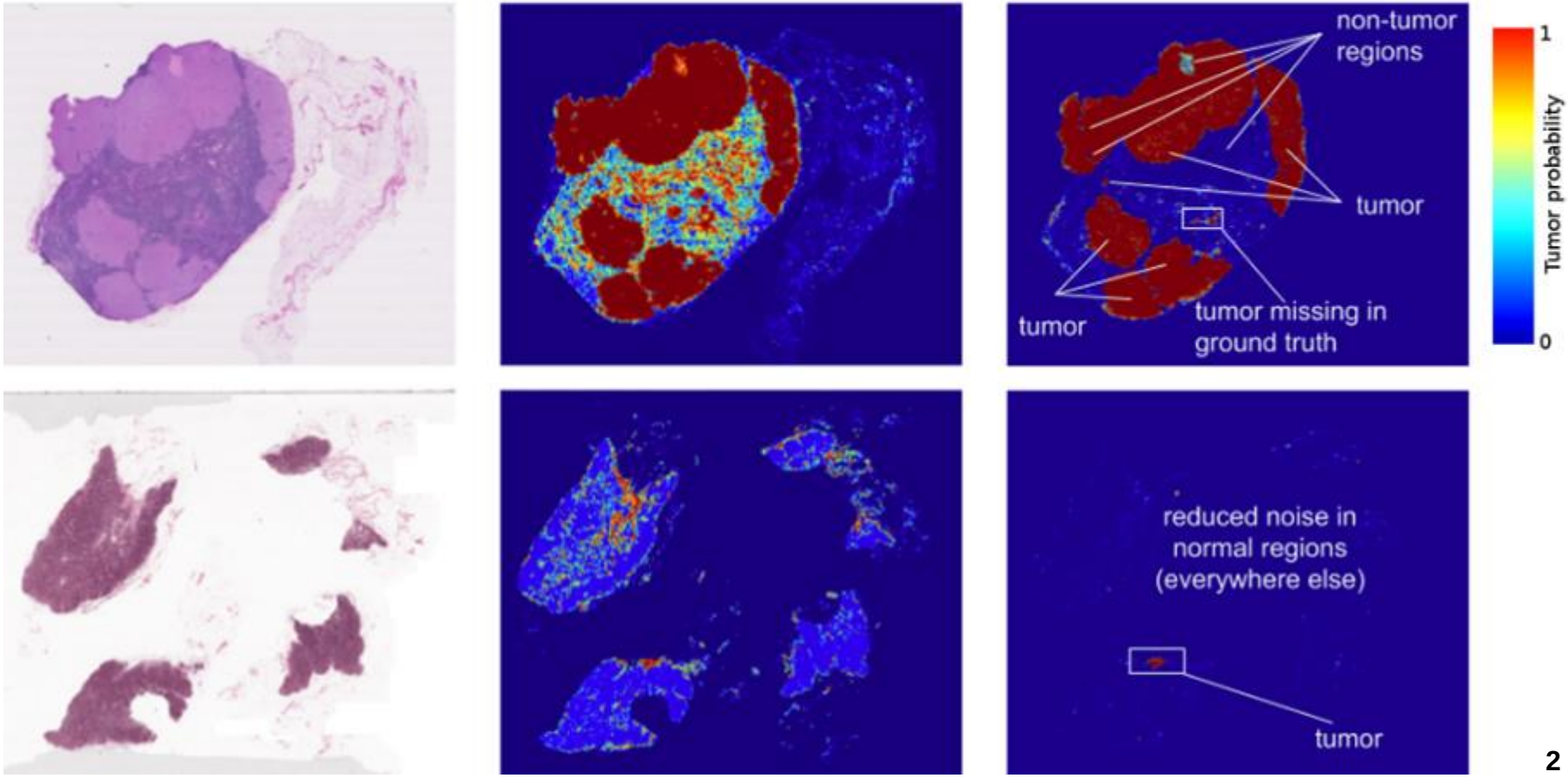
August 22nd, 2017

Instructor: Sira Sriswasdi (สิระ ศรีสวัสดิ์)

Google Deep Learning system diagnoses cancer better than a pathologist with unlimited time

89% accuracy vs 73%

Ben Lovejoy - Mar. 3rd 2017 6:51 am PT @benlovejoy



Biology + Computation

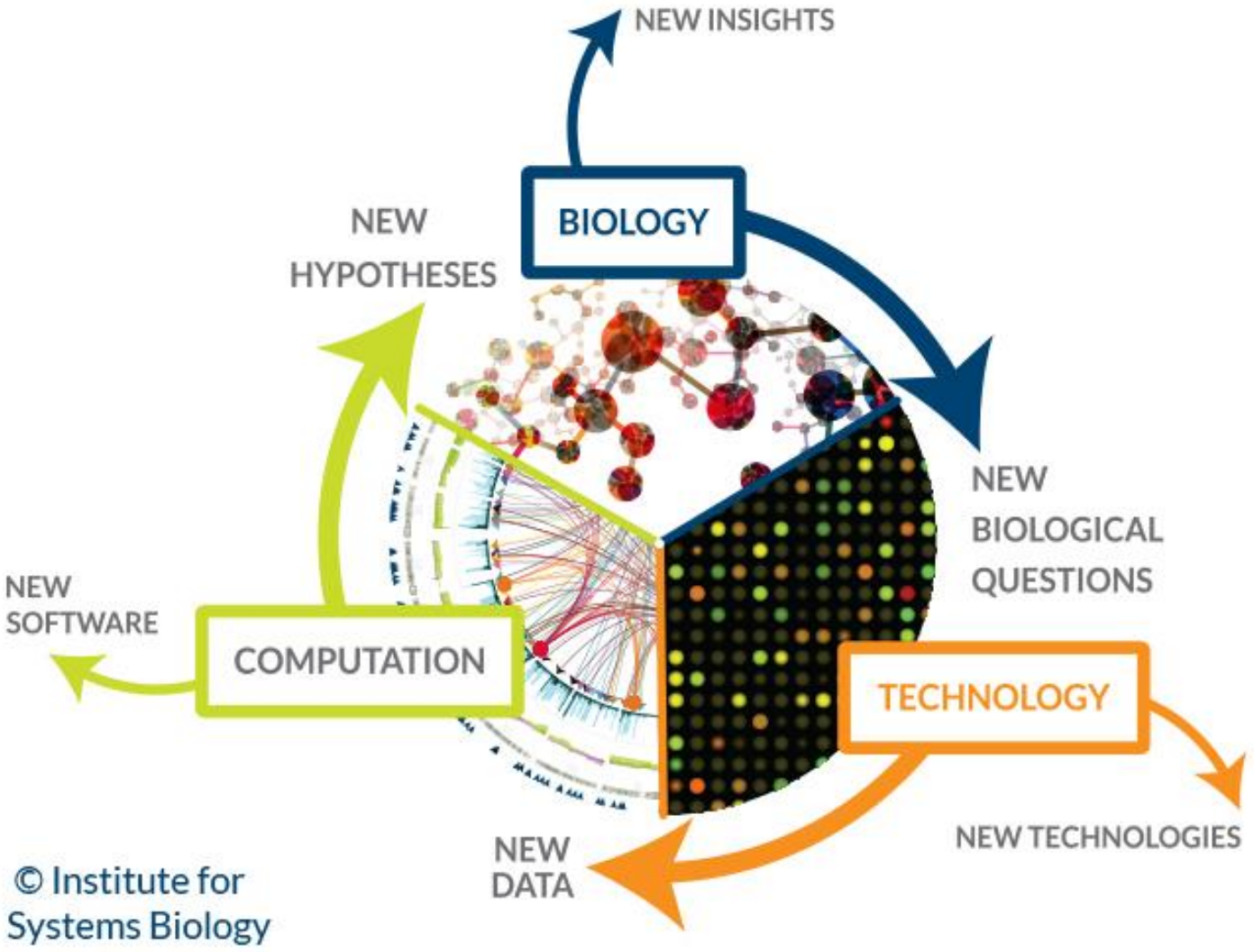


Image from https://www.systemsbiology.org/about/what-is-systems-biology/

Data From High-Throughput Technology

The Central Dogma

"Information Processing in Biology"

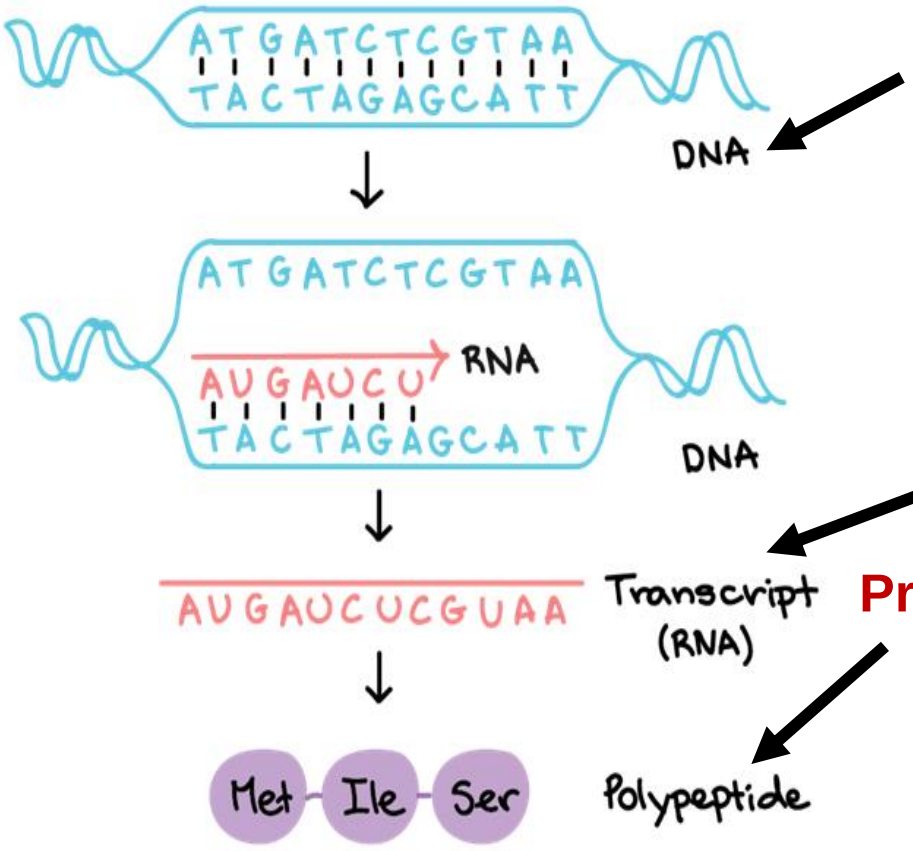
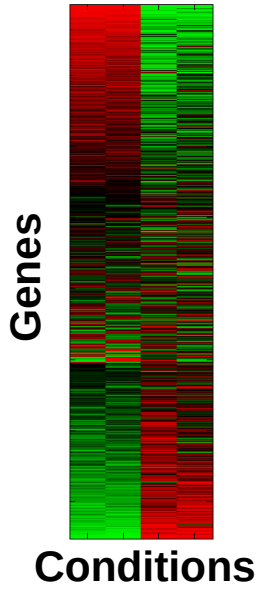


Image from https://www.khanacademy.org/science/biology/gene-expression-central-dogma/transcription-of-dna-into-rna/a/overview-of-transcription

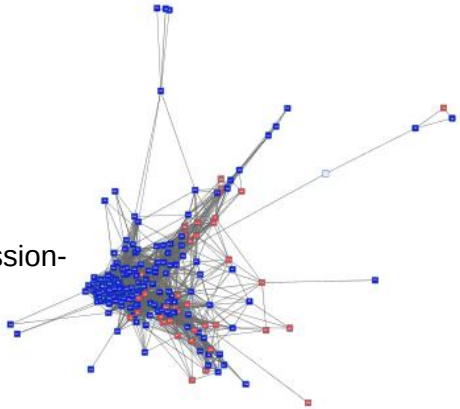
Genome Sequence

ACCAGCGGCGAAGCTCGGGGCGGAGGGGTTGA
GCCACATGAGGCGATGGCGACAATGAGGCGAG
ACATGGCGTGGCTGGCTGTTACATTTTGTTTT
GATGAAAAGCATAACCATGCGGATGATATTTT
TATTATAGACTAGAGATGATTATTGAATAGAC
ATGCTCTTAACCATTTTTAACTCTAATTCCAC

RNA Expression



Protein Interactions



The Omics Era

Protein identification and quantification, post-translational modification

DNA methylation, histone modification

Metabolites, hormones, and other signaling molecules

Image from http://www.cubocube.com/files/images/opengenetics/chapter11/image2.png

RNA sequencing, RNA expression, transcriptional regulation

DNA sequencing, genetic mapping, recombinant DNA

Application I:
Evolutionary Genomics

Genome Sequence As Species Signature

Human 

CCAGCGGCGAAGCTCGGGGCGGAGGGGTTGAGCCACATGAGGCGATGGCGACAATGAGGCGAGACA
TGGCG**TGGCTGGC**TGTTACATTTTGTTTTGATGAAAAGCATAACCATGCGGATGATATTTTTATTA
TAGACTAGAGATGATTATTGAATAGAC**ATTTTAACTCTAATTCCA**...

Zebrafish 

GCGAAGCTCGTTAACCATGCGGATGATATTTTTATTATAGACTAGAGATGATTATTGAATAGAC**AT**
GCTCTTAACCATTTTTAACTCTAA...

Frog 

GCGGCGAAGCTCGGGGCGGAGGGGTTGAGCAAAAGCATAACCATGCGGATGATATTTTTATTATAG
ACTAGAGATGATTATTGAA**CTCTAAA**...

~98%

~60%

Chicken 

CCAGCGGCGAAGCTCGGGGCGGAGGGGTTGAGCATAACCATGCGGATGATATTTTTATTATAGACT
AGAGATGATTATTGAATAGAC**ATTTTAACTCTAATTCCA**...

Chimpanzee 

CCAGCGGCGAAGCGGCGATGGCGACAATGAGGCGAGACATGGCG**TGGCTGGC**TGTTACATTTTGTT
AATCGGGGCGGAGGGGTTGAGCCACATGAGCATAACCATGCGGATGATATTTTTATTATAGACTAG
AGATGATTATTGAATAGAC**ATTTTAACTCTAATTCCA**...

Cow 

~80%

CAGCGGCGAAGCTCGGGGCGGAGGGGTTTATTTTTATTATAGACTAGAGATGATTATTGAATAGAC
ATGCTCTTAACCATTTTTAACTCTAATTCC...

Rat 

GCGGCGAAGCTCGGGGCGGAGGGGTTGAGCCACATGAGGCGATGGCGACAATGAGGCGAGACATGG
CG**TGGCTGGC**TGTTACATTTTGTT**TAACTCTAAT**...

Platypus 

~90%

CGGAGGGGTTGAGCCACATGAGGCGATGGCGACAATGAGGCGAGA
CATGGCG**TGGCTGGC**TGTTAC...

Green spotted pufferfish 

CGGCGAAGCTCGGGGCGGAGGGGTTGA**TTTTTAACTCTAATT**...

Opossum 

AGCGGCGAAGCTCGGGGCGGAGGGGTTGAGCCACATGAGGCGATGGCGACAATGAGGCGAGACATG
GCG**TGGCTGGC**TGTTACATTTTGTTTTGATGAAA**TTTTTAACTCTAATTC**...

Mouse 

CCAGCGGCGAAGCTCGTTGAGCCACATGAGGCGATGGCGACAATGAGGCGAGACATGGCG**TGGCTG**
GCTGTTACATTTTGTTTTGATGAAAAGCAT**TCTAATTCCA**...

Image adapted from http://physrev.physiology.org/content/89/3/921

Evolution Of DNA Sequences

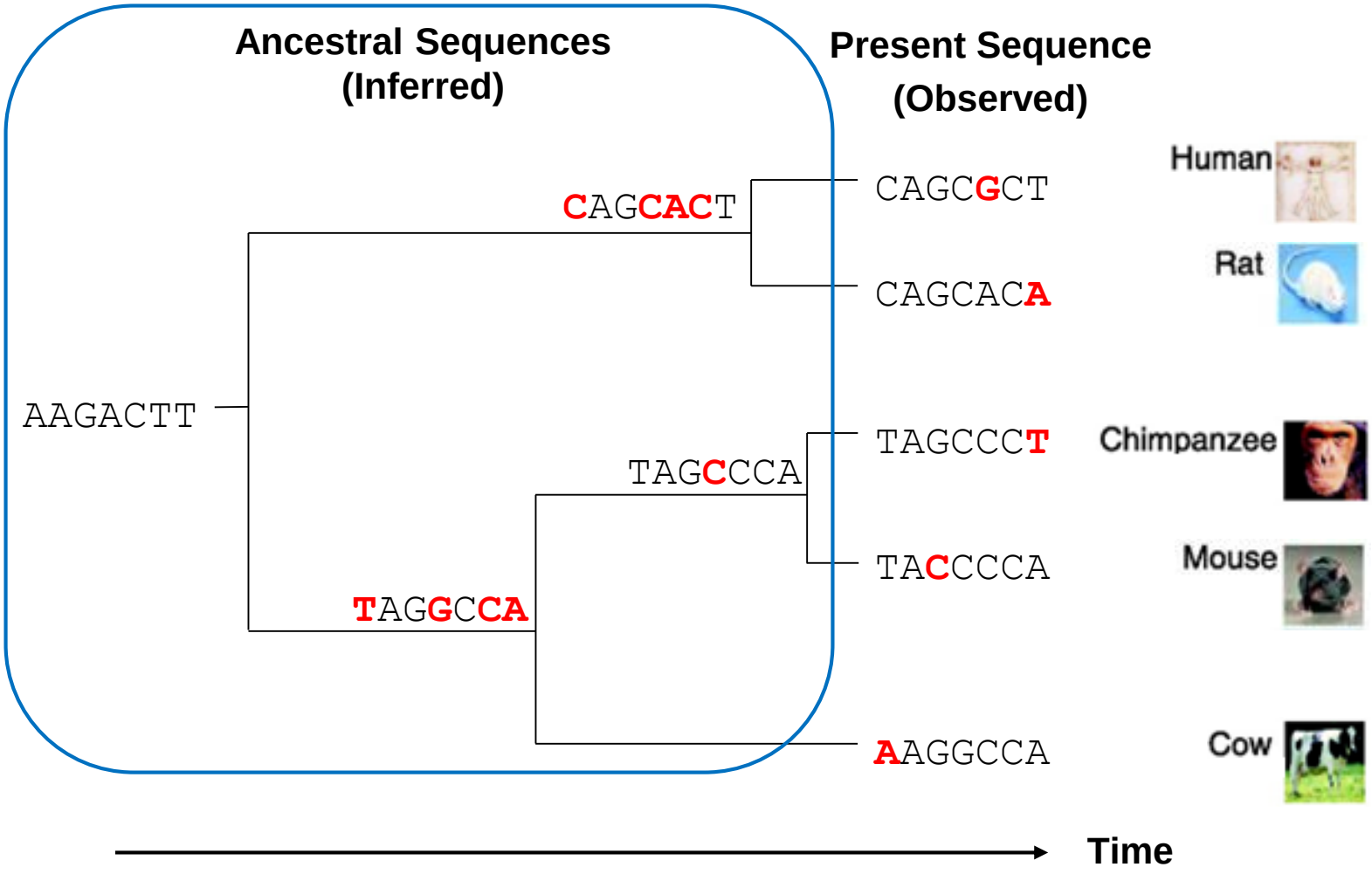
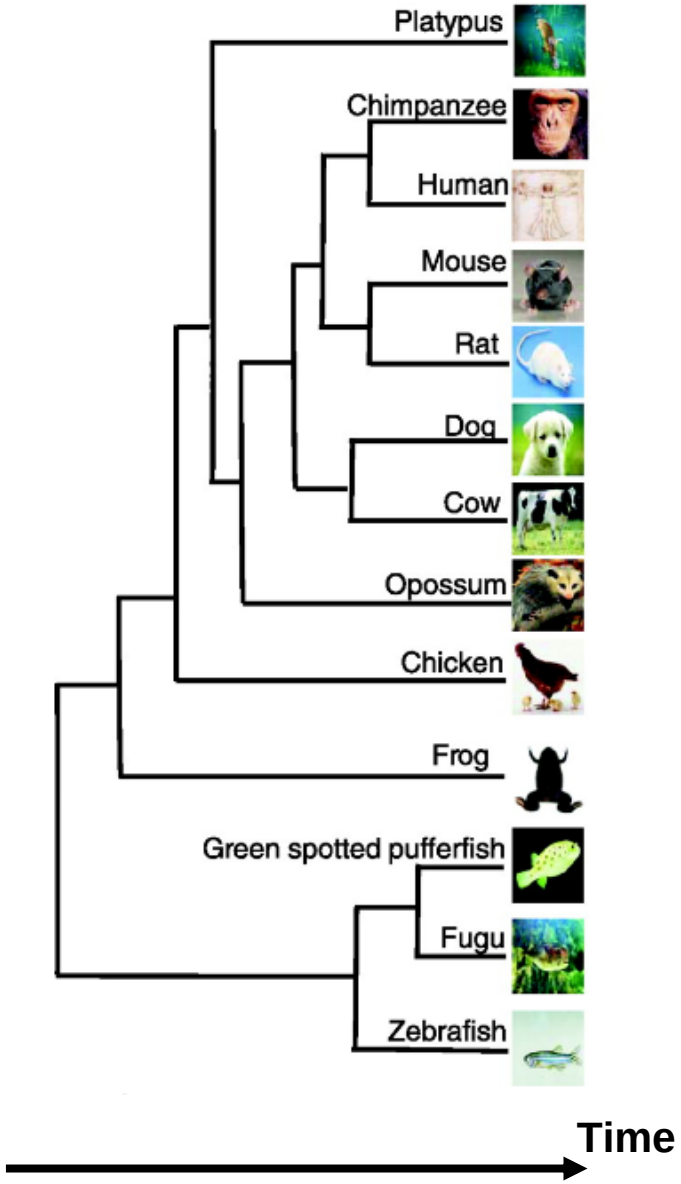


Image adapted from http://physrev.physiology.org/content/89/3/921 and https://www.quora.com/Arent-all-sequences-homologous

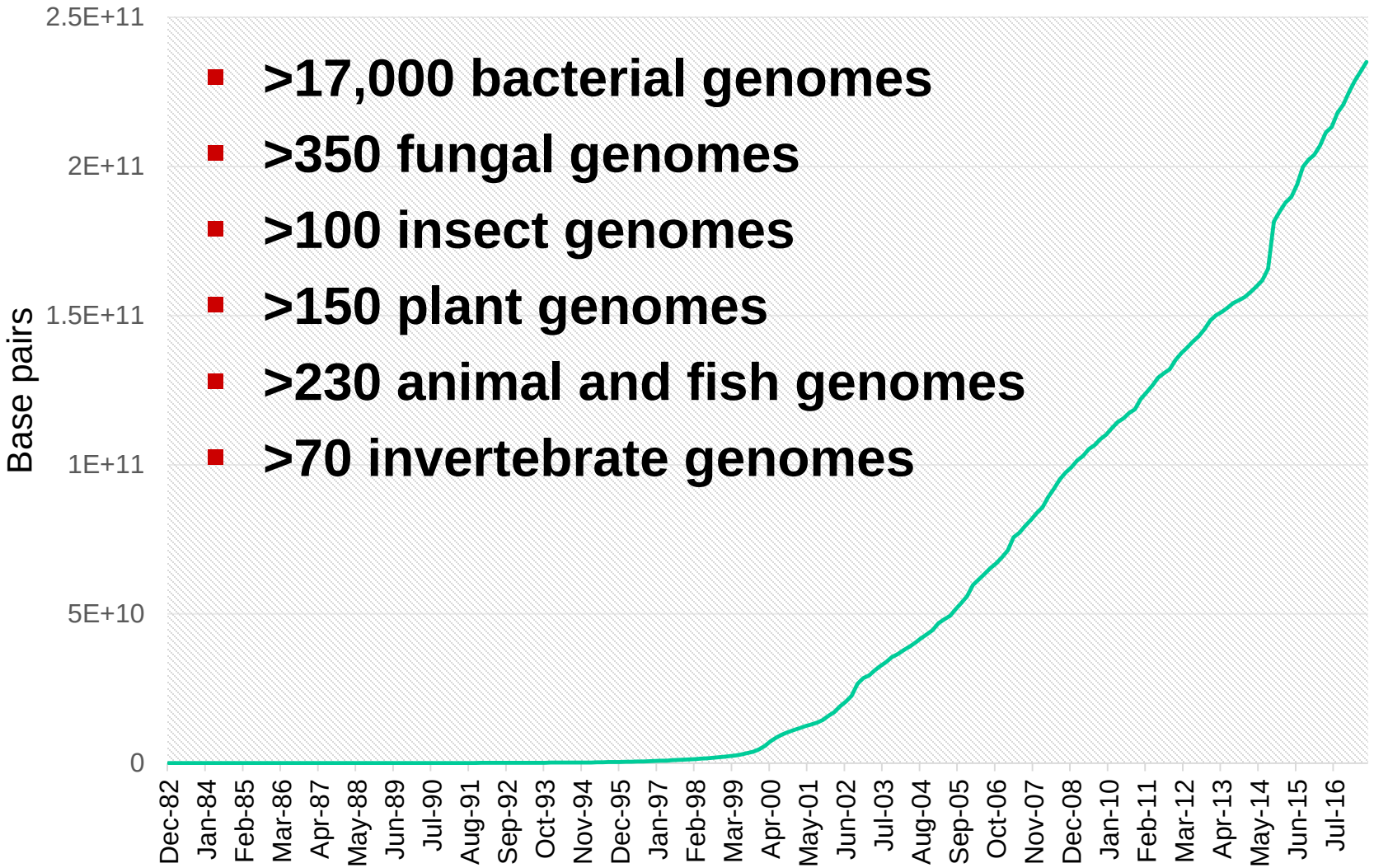
Inferring Evolutionary History (Phylogenetics)



- **Reconstruction of evolutionary events over millions of years**
- **Based on genome sequences of currently existing species**
- **Assume some models of evolution on DNA sequence, e.g. P(A→T), P(G→T)**
- **Output the most likely tree topology and branch lengths**
- **Extremely large number of parameters, search spaces, and number of models to compare**

Image from http://physrev.physiology.org/content/89/3/921

Growing Amount Of Genomic Data



Plotted with data from https://www.ncbi.nlm.nih.gov/genbank/statistics/

Forecasting And Regulating Evolution

- **Epidemiology**
  - **Tracking the spread of disease outbreaks**
  - **Predict the next outbreaks and prepare vaccines in advance**
- **Biotechnology**
  - **Genetic engineering and breeding of new strains with desired characteristics and capabilities**
- **Wildlife Conservation**
  - **Pairing evolutionary history with climate/environmental changes can reveal the factors that drive animal evolution and extinction**

Application II:
Population Genetics

Tracing Population Structure Over Time

Migration

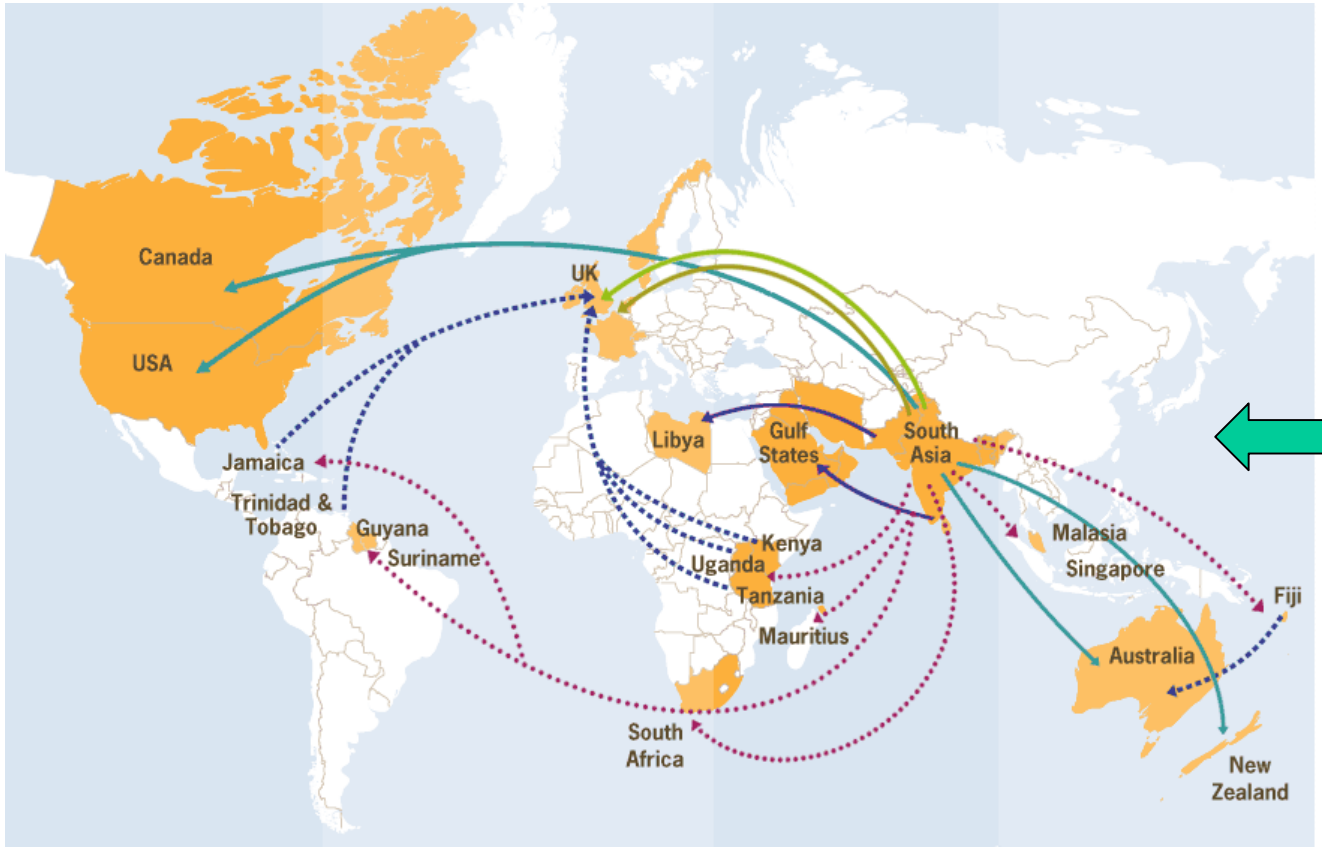


Image adapted from https://www.theodysseyonline.com/why-people-migrate

Genetic Inheritance

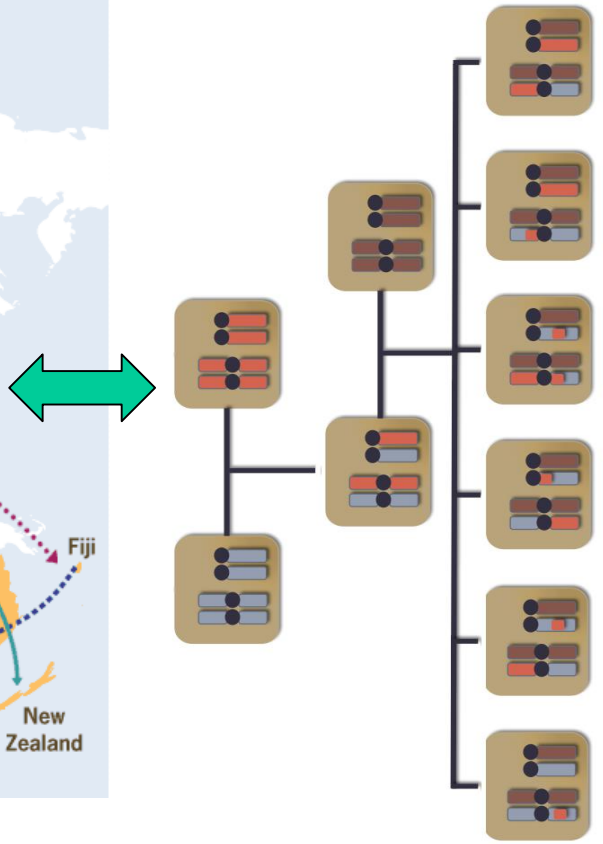


Image adapted from
https://wiki.uiowa.edu/display/2360159/Autosomal+Inheritance

Single Nucleotide Polymorphisms (SNPs) As Individual's Genetic Signature

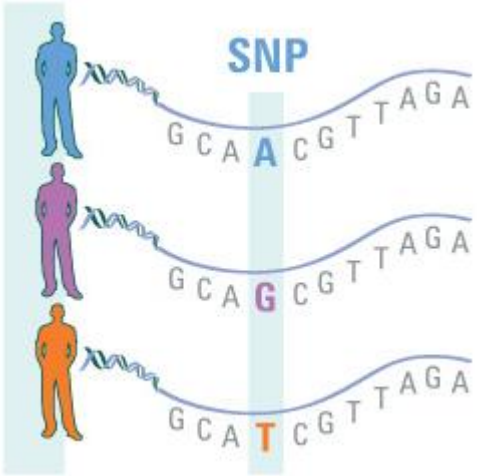
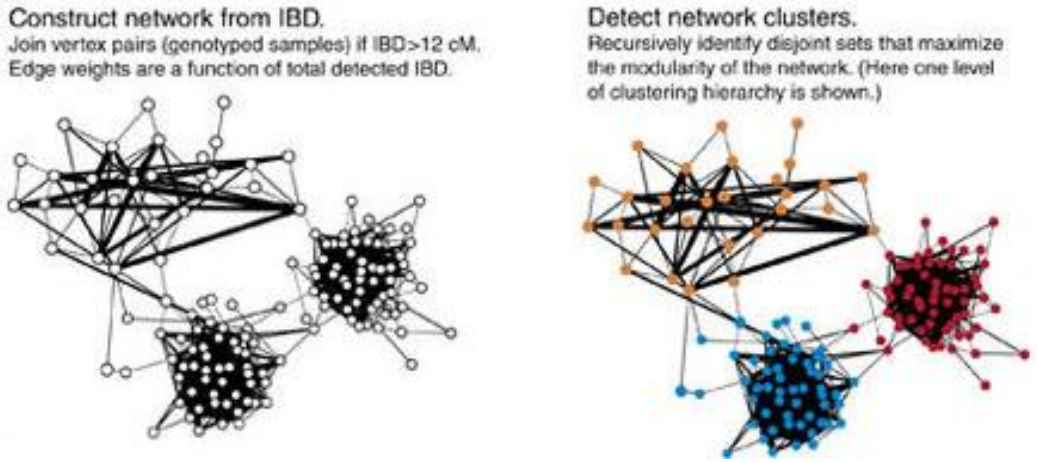
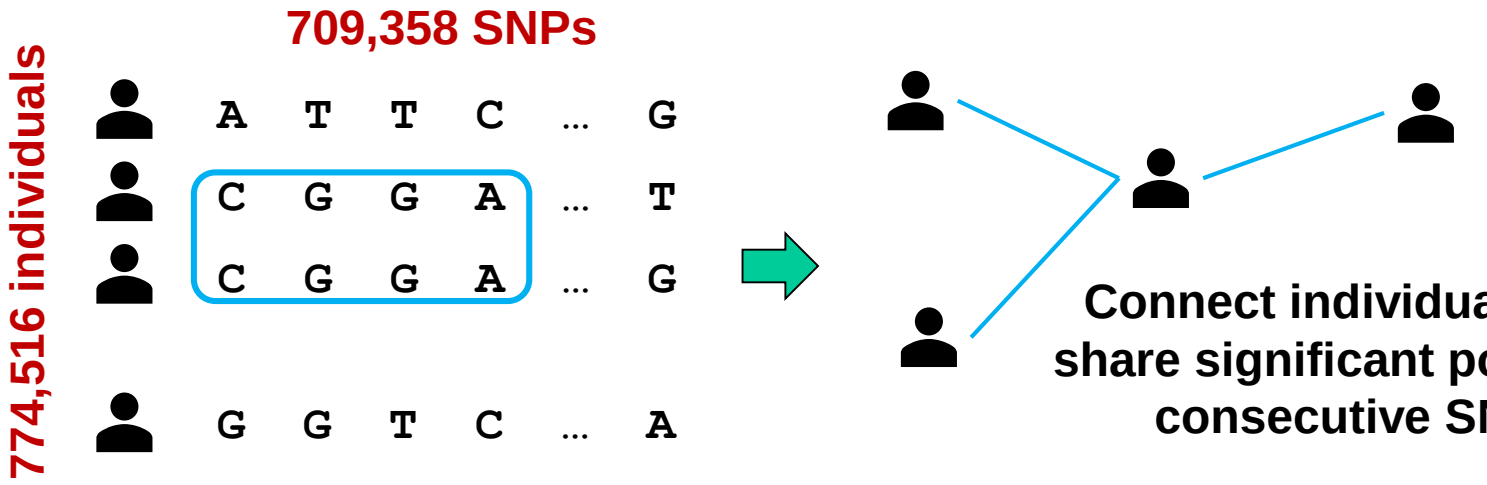


Image from http://uvmgg.wikia.com/wiki/SNP

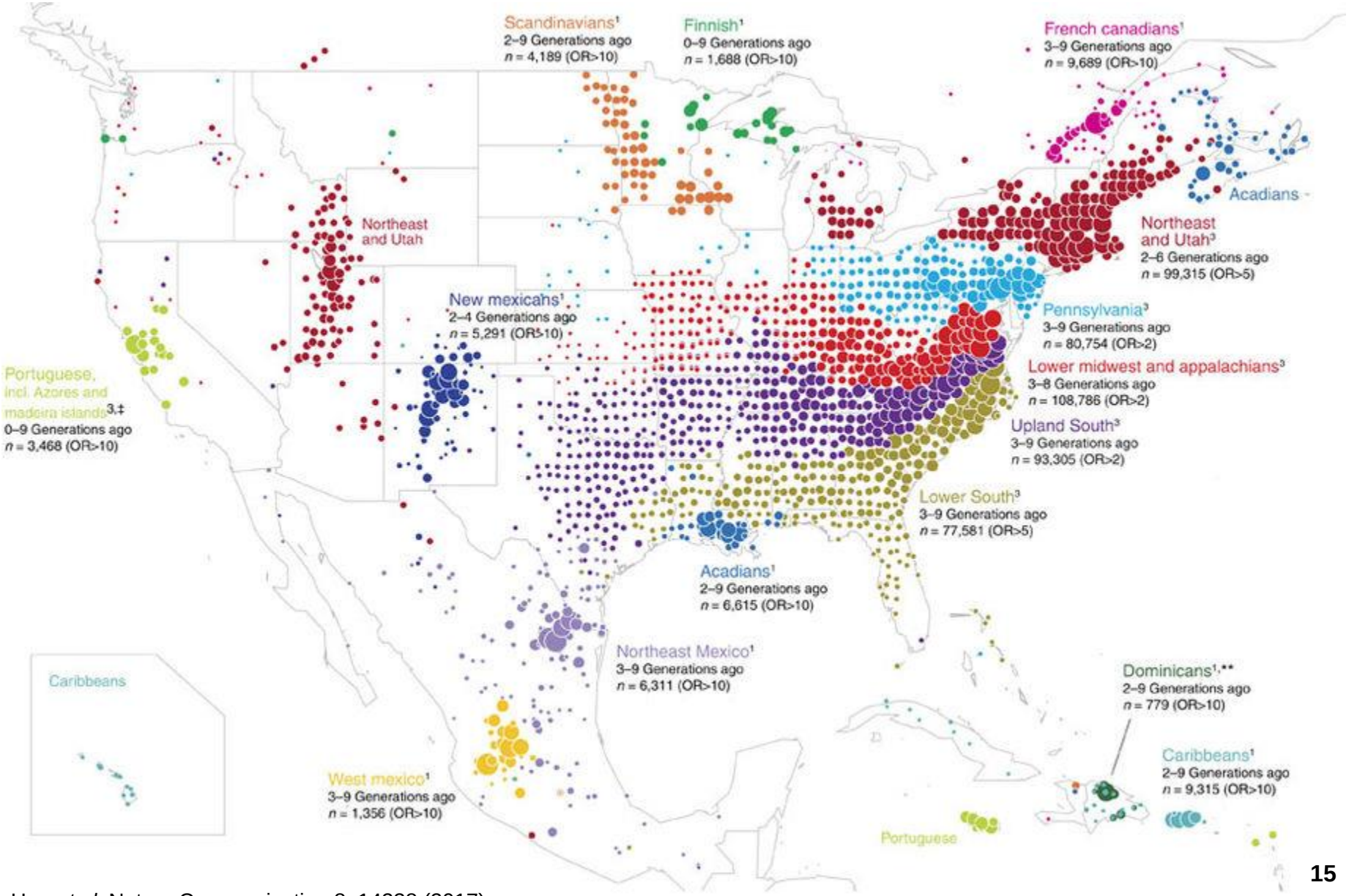
Identity-By-Descent (IBD) Network



Han *et al*. Nature Communication 8, 14238 (2017)



Roots Of North American Population



Han *et al*. Nature Communication 8, 14238 (2017)

From Population To Personalized Medicine

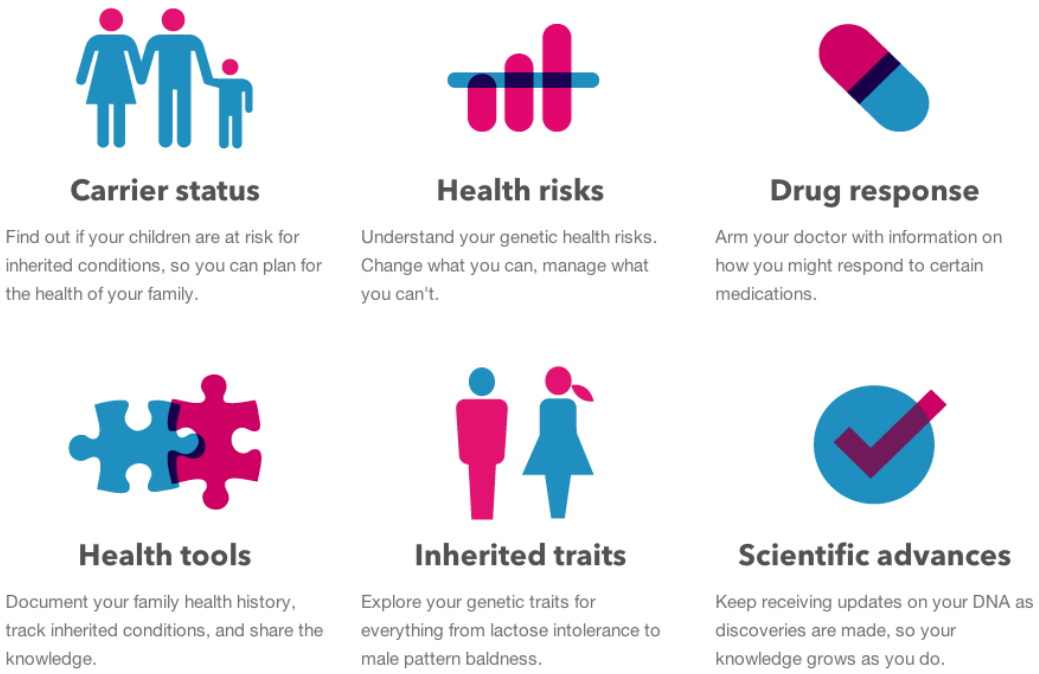


Image from http://poshrx.com/23andme-is-back-on/

- **Social Sciences**
  - **Tracking the dynamics of populations**
  - **Understanding ethnic structures**
- **Medicine**
  - **Identifying common genetic variations within a population that may be associated with drug targets**
  - **Identifying disease risk factors**

Application III:
Gene Expression Analysis

Measuring Gene Expression

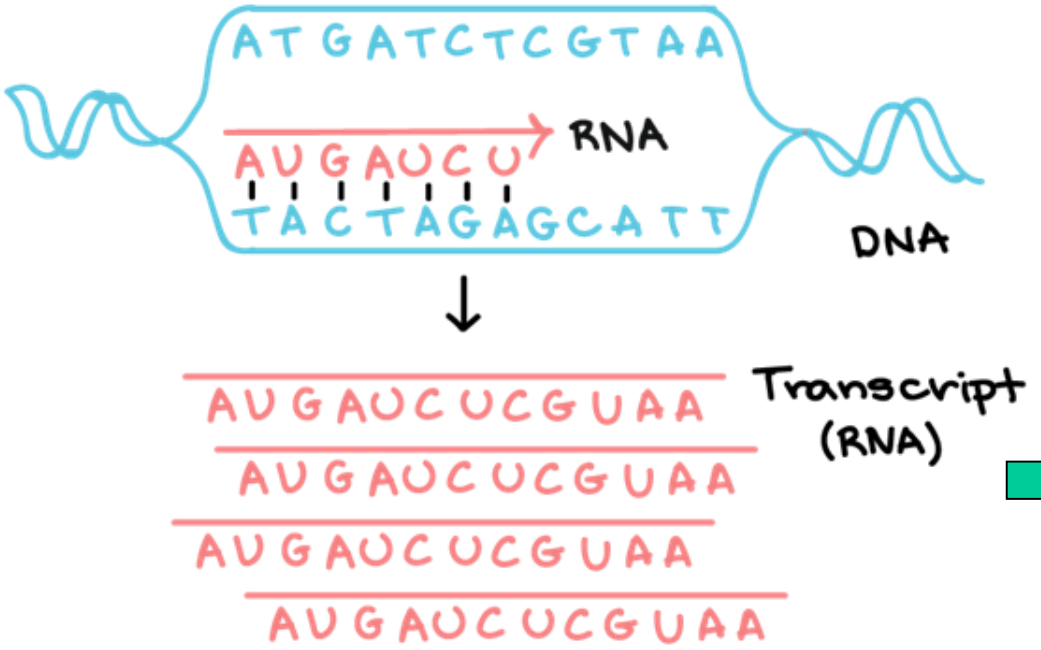


Image adapted from
https://www.khanacademy.org/science/biology/gene-expression-central-dogma/transcription-of-dna-into-rna/a/overview-of-transcription

Sequencing Machine

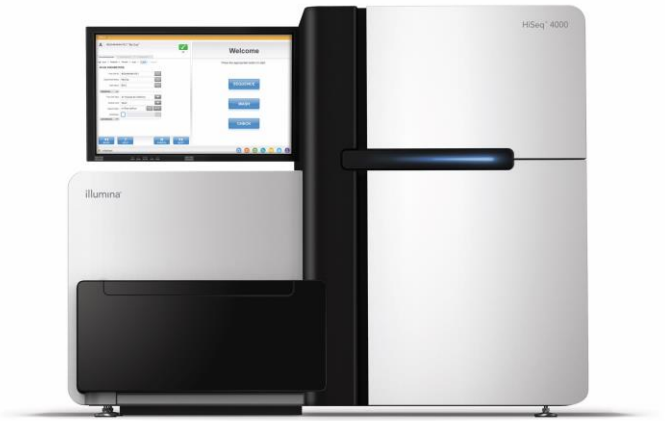


Image from
https://support.illumina.com/sequencing/sequencing_instruments/hiseq-4000.html

Amount of RNA product of each gene

gene_symbol	target_id	control 1	control 2	exp 1	exp 2	exp_avera	log10_fold_differen
RANBP2	NM_006267.4	6.48833	12.9884	0.176602	0.117152	0.143837678	1.80497163
CYTH1	XM_011525475.2	15.0945	11.3296	0.249429	0.173519	0.20804007	1.798369664
HNRNPK	NM_031263.3	117.923	184.554	2.67089	2.25173	2.452370922	1.779274833
PSMD2	NM_002808.4	47.5916	97.2485	0.855375	1.50337	1.1339952	1.778095392
NONO	NM_007363.4	91.2359	146.966	2.35517	1.58616	1.932789809	1.77750669
UGP2	NM_001001521.1	45.0717	70.7679	0.982185	0.906358	0.943510059	1.777123589
FDFT1	NM_004462.4	113.559	139.643	3.17652	1.55325	2.221245077	1.753523887
ATP6V1H	NM_015941.3	40.4483	34.6337	0.940035	0.463959	0.660407222	1.753387768
STK10	NM_005990.3	13.8212	25.0476	0.458813	0.246147	0.336058691	1.743240804
DDX3X	NM_001193416.2	59.55	105.791	1.66098	1.25129	1.441654488	1.740804056
KIF5B	NM_004521.2	10.4073	43.989	0.356018	0.437435	0.394632403	1.734148339
TRIB1	XR_428373.2	13.8993	10.355	0.211033	0.232964	0.221727517	1.733251922
SREBF2	NM_004599.3	51.8562	90.6209	1.50785	1.06592	1.267772642	1.732973154

RNA Sequencing (Counting) Biases

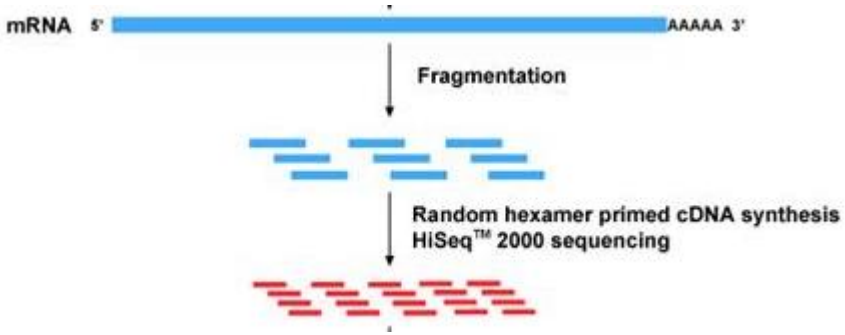


Image adapted from http://bio.lundberg.gu.se/courses/vt13/rnaseq.html/

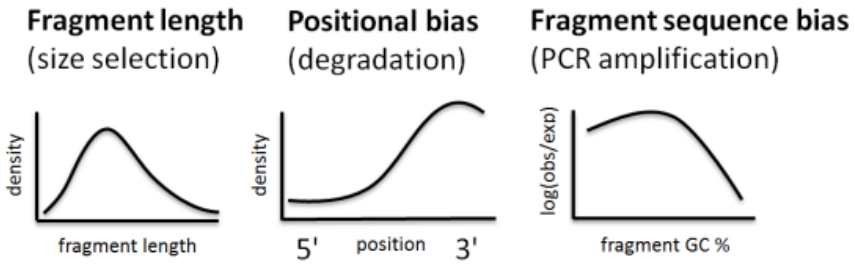


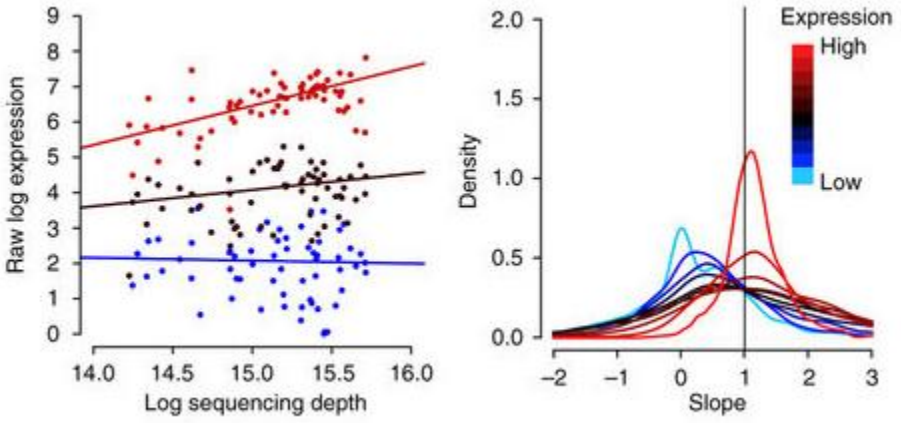
Image adapted from https://mikelove.wordpress.com/2016/09/26/rna-seq-fragment-sequence-bias/

- **Due to technological limitation, the entire length of RNA cannot be sequenced at once**
  - **Full-length RNA has to be fragmented**
  - **Bias in fragment length**
- **To increase sensitivity, fragmented RNA has to be amplified**
  - **Bias in signal amplification**
- **Sequencing is directional**
  - **Bias in head-to-tail read count**

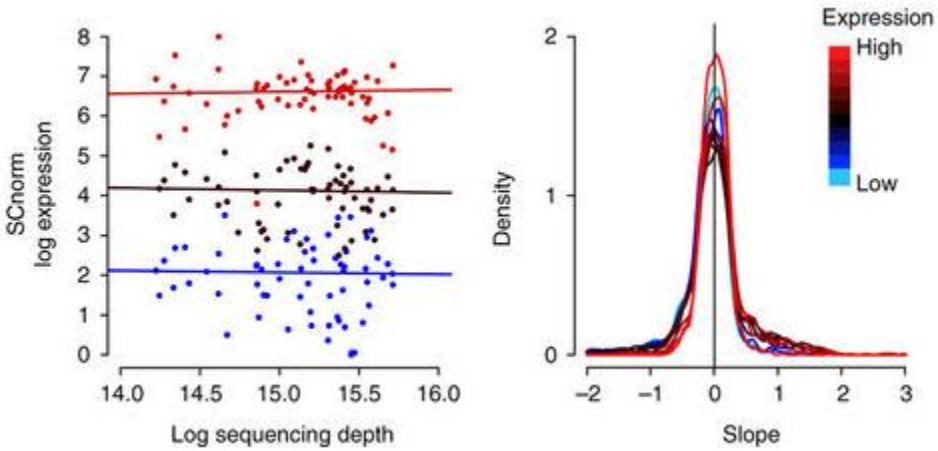
→ **Bias correction**

Bias Normalization via Regression

Before Normalization



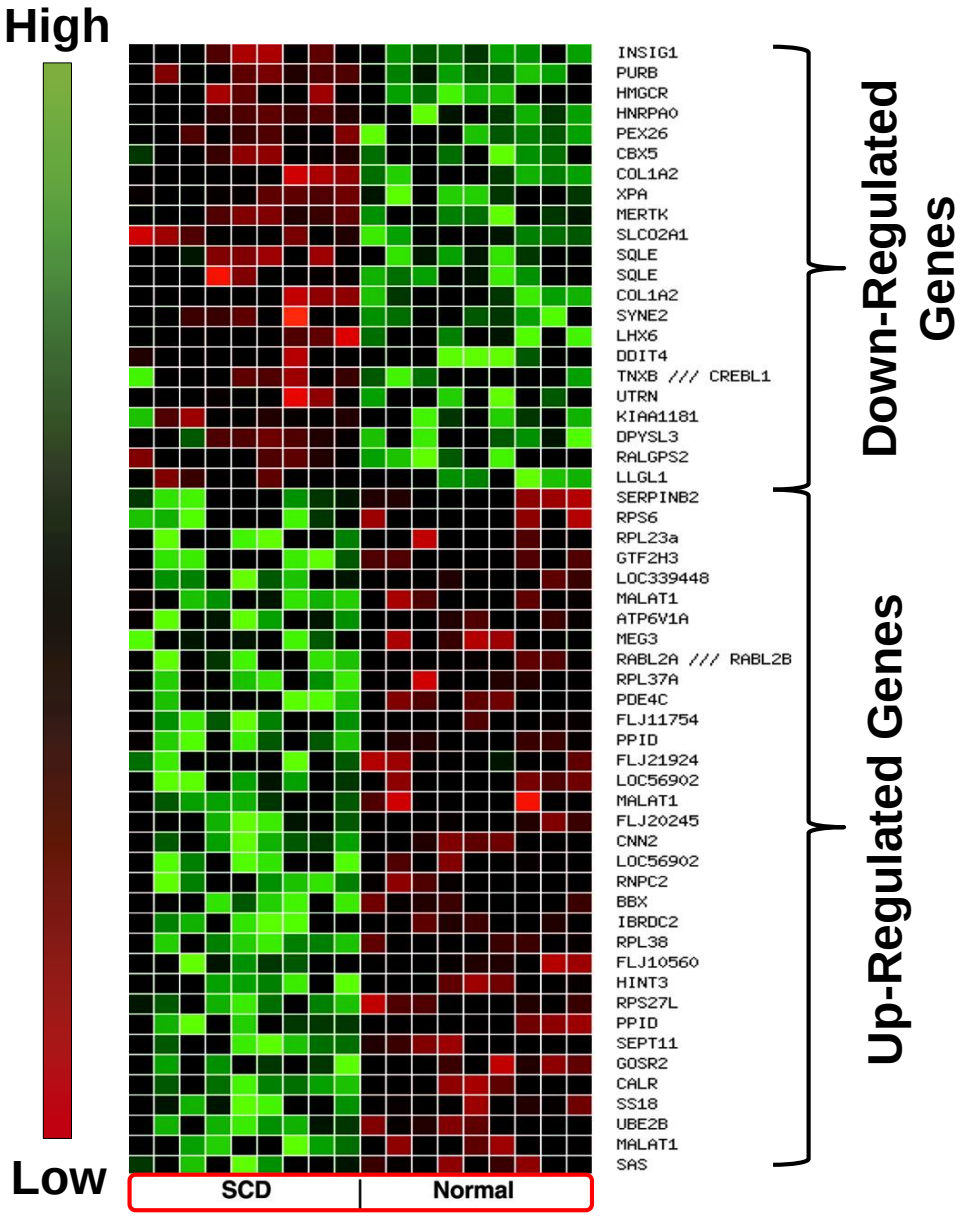
After Normalization



Adapted from Bacher *et al*. Nature Methods 14, 584-586 (2017)

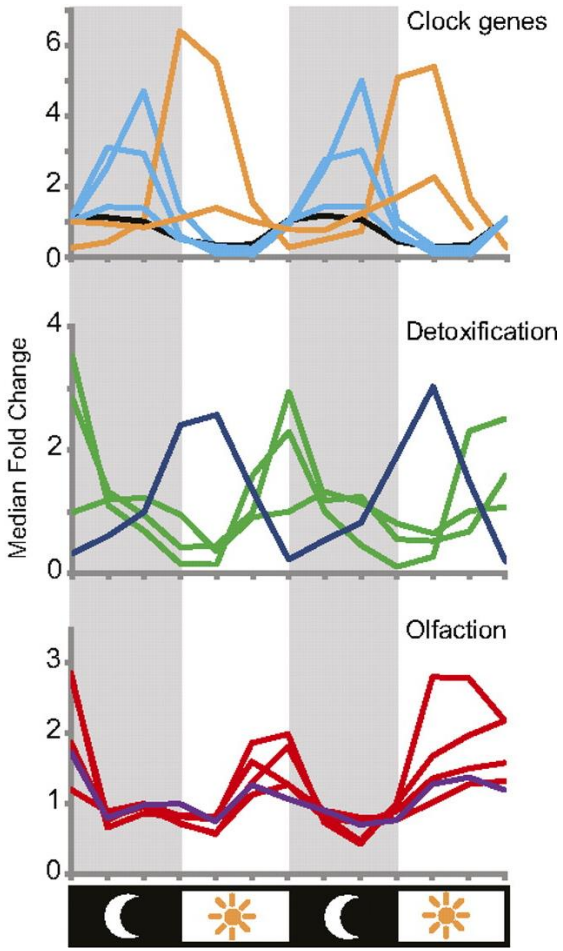
- **Sequencing involves sampling of RNA transcripts**
- **Estimated expression levels of low, medium, and high expression genes are differently affected by the throughput of RNA sequencing experiment**
- **Normalization by regression corrected the biases**

What Can Gene Expression Tell Us?



Klings *et al*. Physiological Genomics 21, 293-298 (2005)

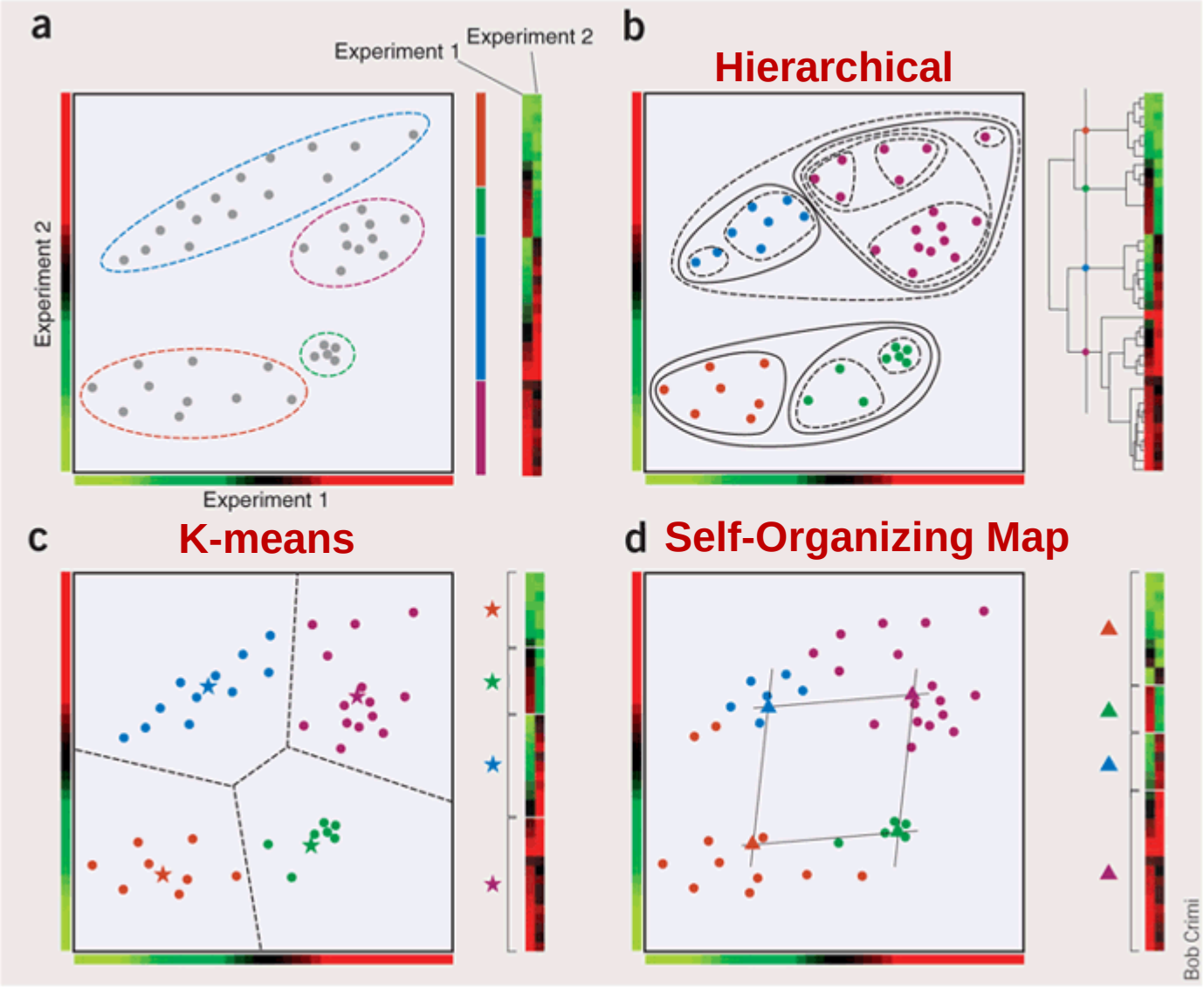
Time Series Analysis



Adapted from Rund *et al*. PNAS 108, E421-430

Structure Behind Gene Expression Profiles

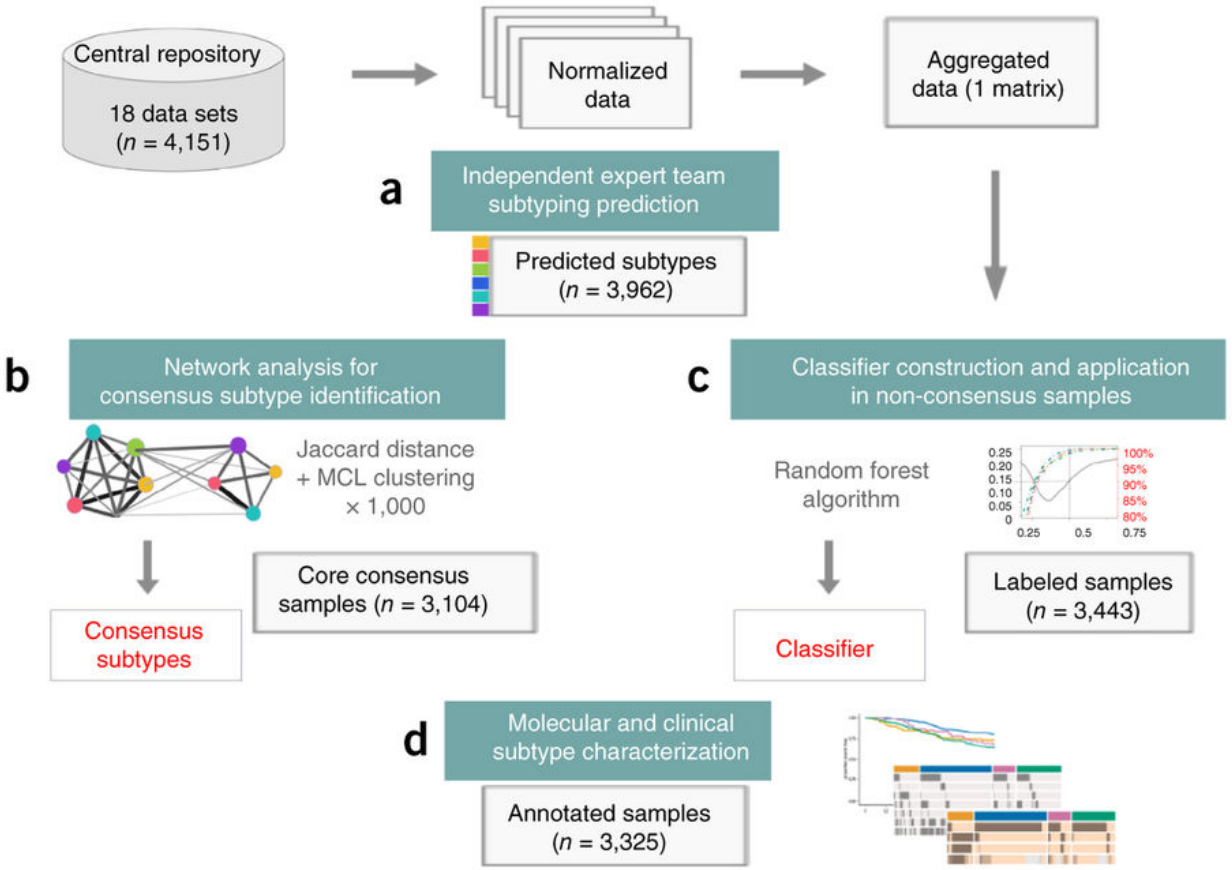
Each cluster represents a group of genes with similar functions

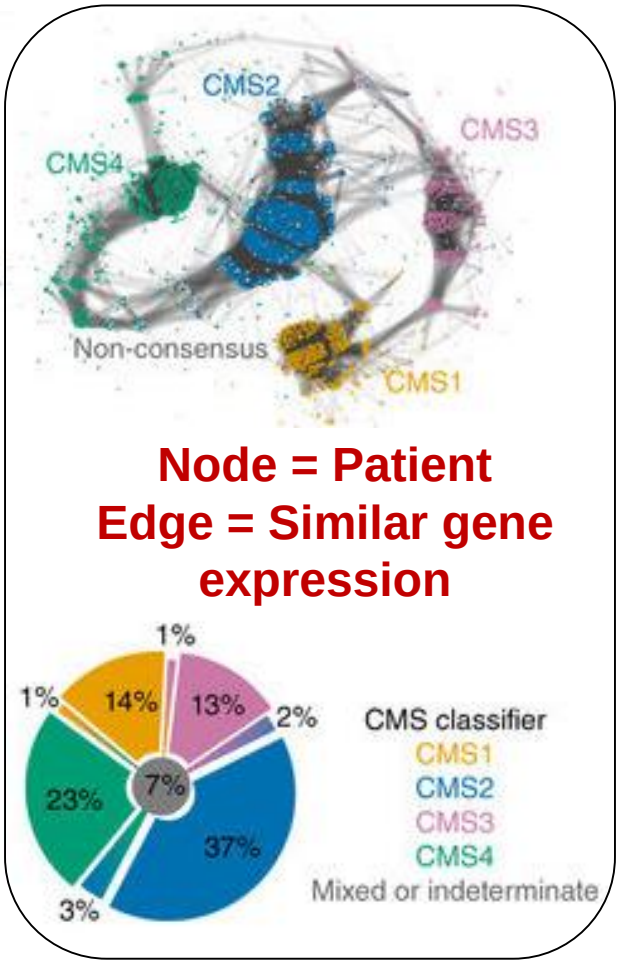


D'haeseleer *et al*. Nature Biotechnology 23, 1499-1501 (2005)

Identifying Disease Subtypes

Gene expression data from >4,000 colorectal cancer patients





Guinney *et al*. Nature Medicine 21, 1350-1356 (2015)

Application Of Gene Expression Analysis

To understand the blueprint of biological systems and diseases

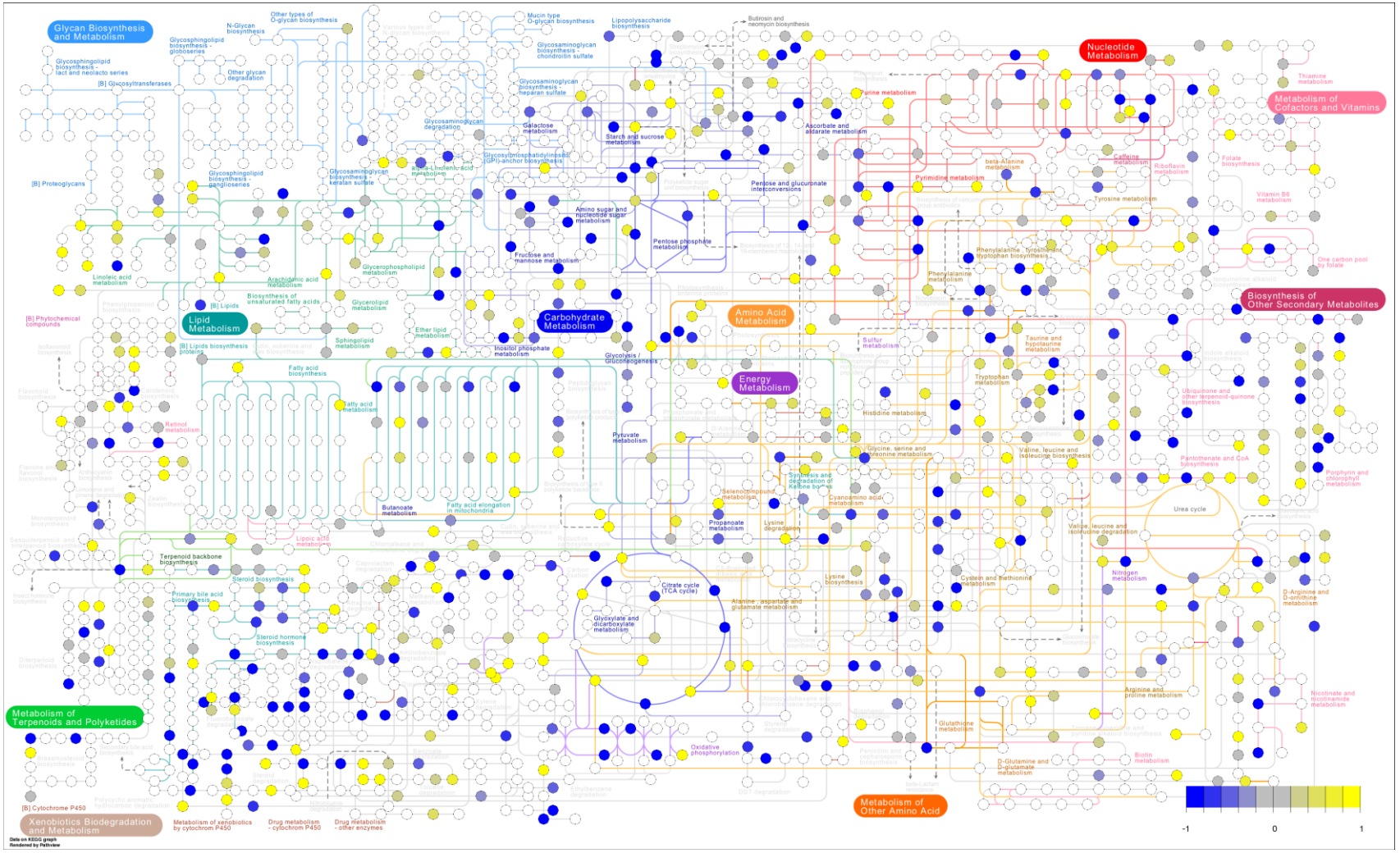
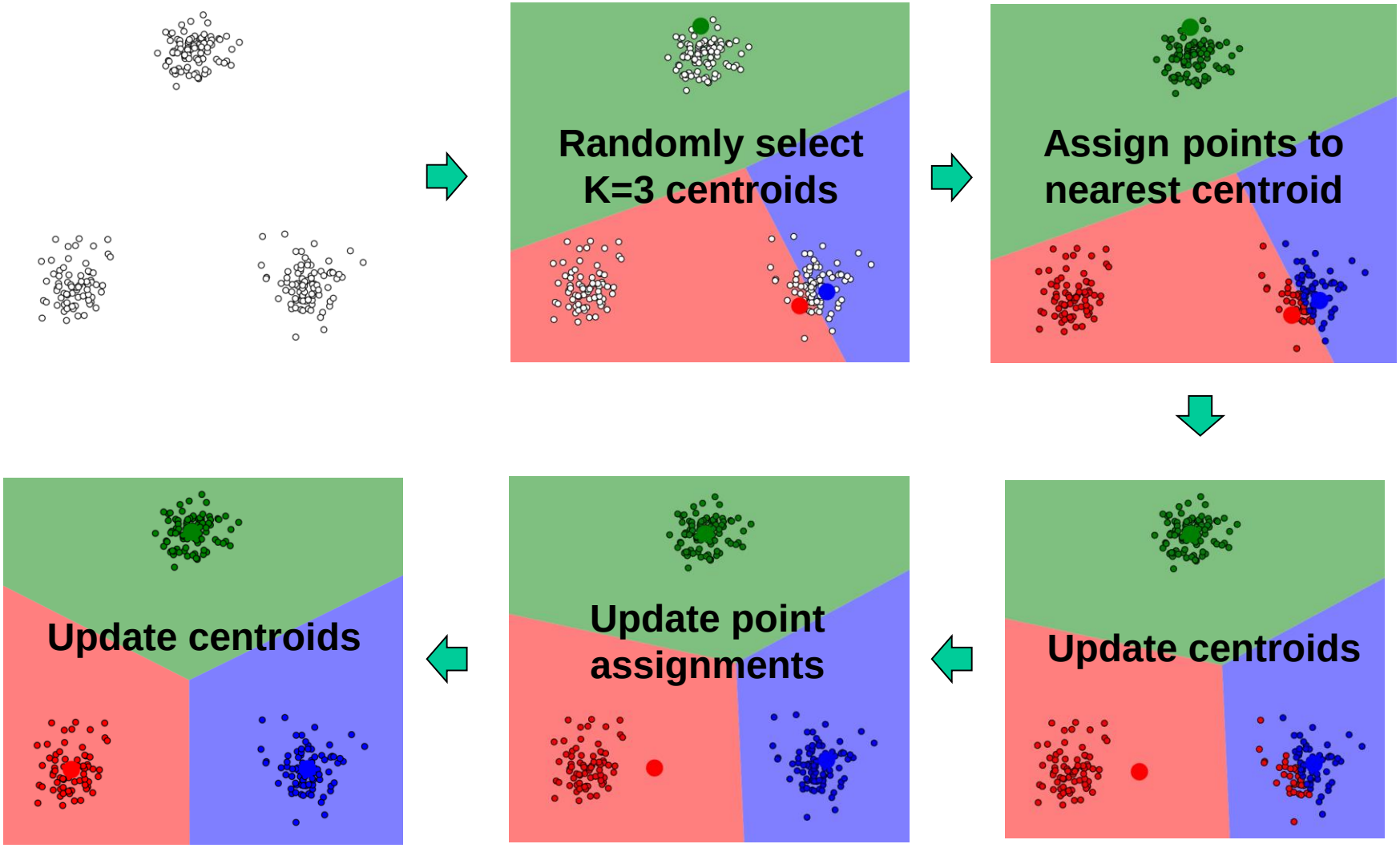


Image from http://pathview.r-forge.r-project.org/

A Break From Biology: Basic Clustering Techniques

An Illustration Of K-Mean Clustering



From https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

Characteristics Of K-Mean Clustering

- **The number of clusters, *K*, is specified in advance.**
- **Euclidean distance**
  - **The nearest centroid minimizes the sum of squares, $||x-m||^2$.**
- **Always converge to a (local) minimum.**
  - **Poor starting centroid locations can lead to incorrect minima.**

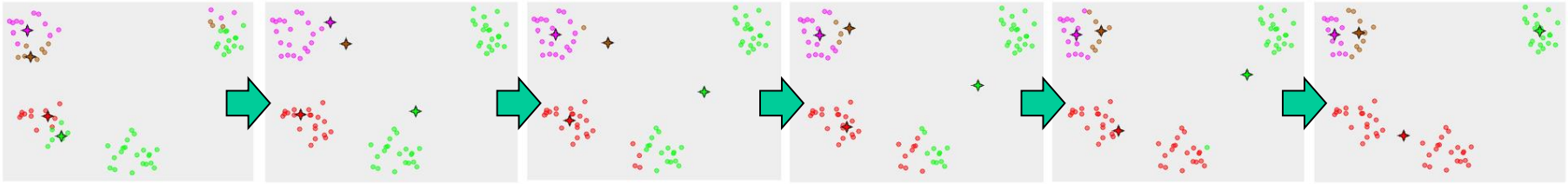
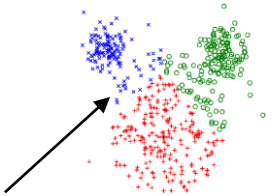
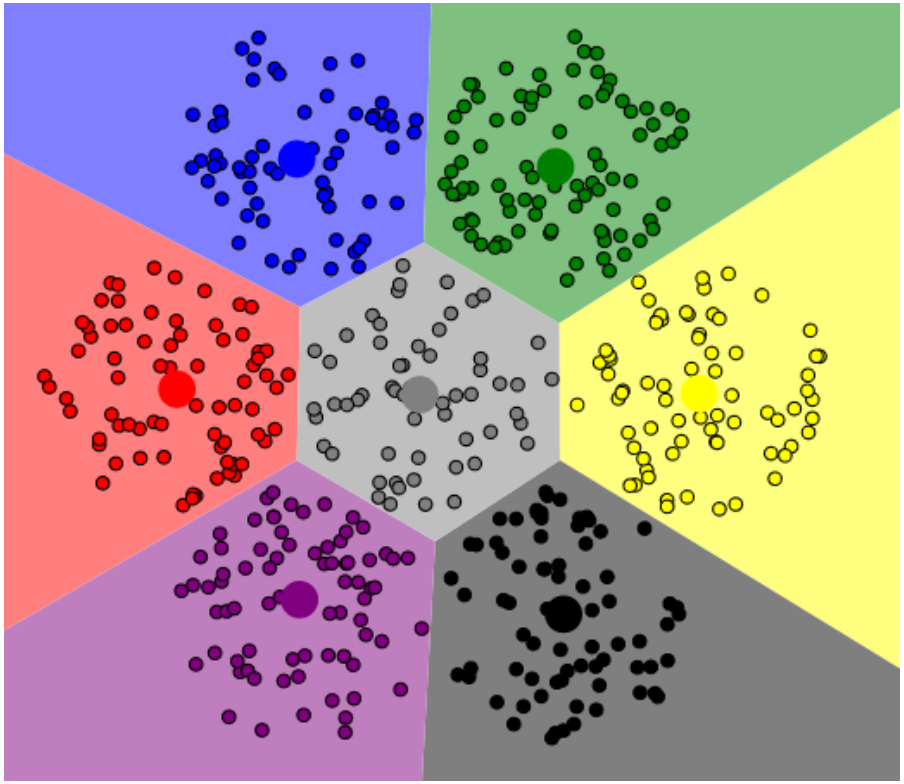


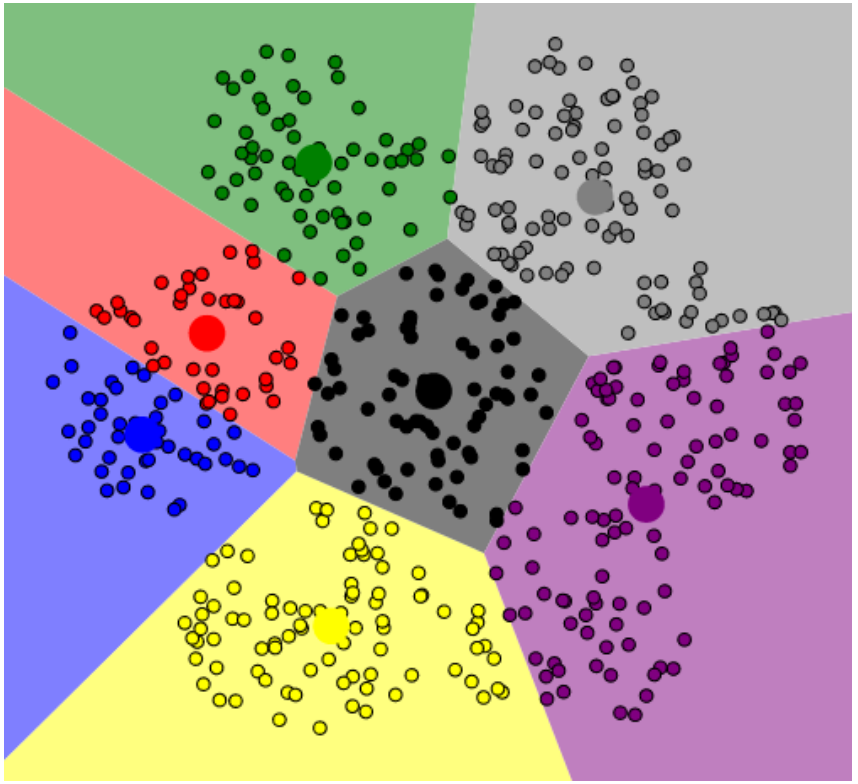
Image from https://en.wikipedia.org/wiki/K-means_clustering

- **The model has several implicit assumptions:**
  - **Data points scatter around cluster's centers.**
  - **Boundary between adjacent clusters is always halfway between the cluster centroids.**



Effect Of Poor Initial Centroid Locations





From https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

Distance Functions

- **Property of distance function**
  - **d(x, y) ≥ 0** — **non-negativity**
  - **d(x, y) = 0 ⇔ x = y** — **identity**
  - **d(x, y) = d(y, x)** — **symmetry**
  - **d(x, z) ≤ d(x, y) + d(y, z)** — **triangle inequality**

- **Example of distance functions**
  - **Euclidean distance** $\mathbf{d}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum(x_i - y_i)^2}$
  - **Squared Euclidean distance** $\mathbf{d}(\mathbf{x}, \mathbf{y}) = \sum(x_i - y_i)^2$
  - **Manhattan distance** $\mathbf{d}(\mathbf{x}, \mathbf{y}) = \sum |x_i - y_i|$
  - **Maximum distance** $\mathbf{d}(\mathbf{x}, \mathbf{y}) = \max |x_i - y_i|$

- $||\mathbf{x}||_p = (\sum_i |\mathbf{x}_i|^p)^{1/p}$ ← ***p*-norm**

More About Distance Functions

Image from https://www.quora.com/What-is-the-difference-between-Manhattan-and-Euclidean-distance-measures

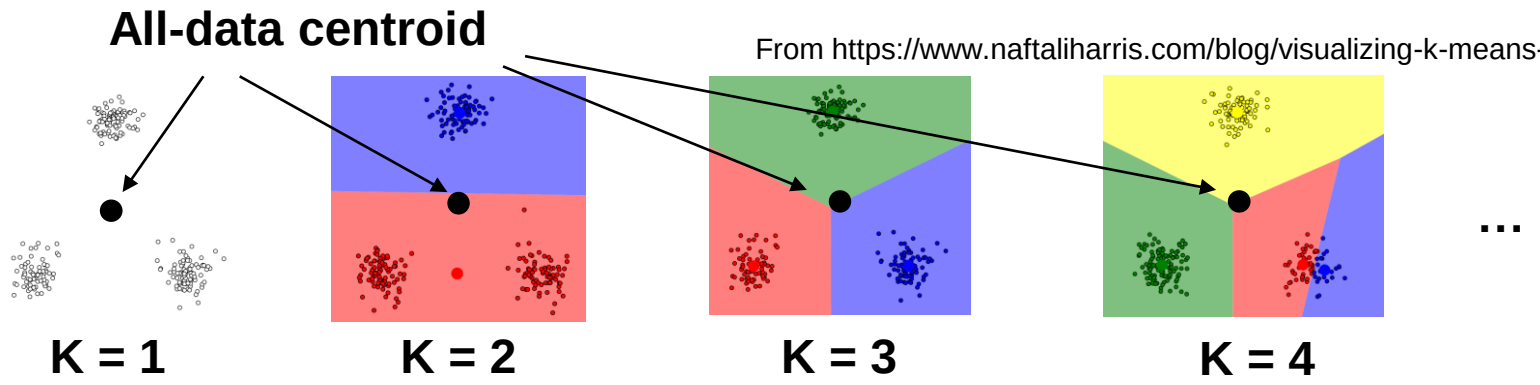
- **Manhattan distance**
  - $\mathbf{d}(\mathbf{x}, \mathbf{y}) = \sum |\boldsymbol{x_i} - \boldsymbol{y_i}|$
  - **Can reflect driving distance**

- **Hamming (Edit) distance**
  - **For two string *s* and *t*, d(*s*, *t*) = # of mismatch positions between the two strings.**
  - **Can reflect the extent of evolution between genes: more changes in sequence ~ more time has passed**

ATGAG**C**ATA**A**CCATGC**G**GAT

ATGAG**G**ATA**C**CCATGC**C**GAT

Elbow Method For Selecting K In K-Mean



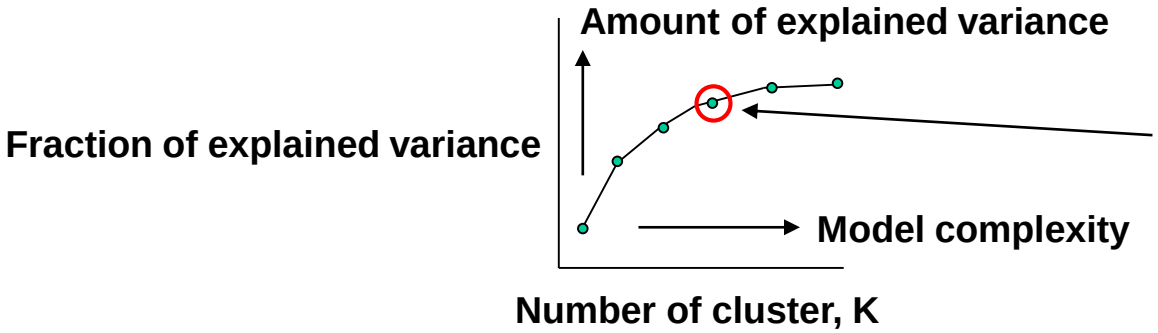
fraction of explained variance = $\frac{\text{between-cluster variance}}{\text{all-data variance}}$

between-cluster variance = $\sum_{i=1}^{K} \frac{n_i(M_i - M)^2}{K-1}$, where n_i = size of i^{th} cluster,

M_i = centroid of i^{th} cluster, and

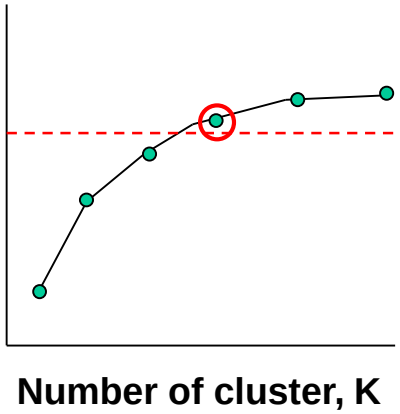
M = all-data centroid.

all-data variance = $\sum_{i=1}^{N} \frac{(x_i - M)^2}{N-1}$, where x_i = i^{th} data point and N = # of data.



The elbow method chooses K where increasing complexity doesn't yield much in return.

Other Ways For Selecting K In K-Mean



Choose minimal K that explains at least 95% of the all-data variance.

K = 2
K = 3
K = 4
⋮

→ **Training K-mean Clustering Model** →

Testing / Cross-validation →

K	Accuracy
2	50%
3	68%
4	83%
⋮	⋮

Choose K that maximizes certain objective (e.g. accuracy on testing data)

K-Mean Clustering Of Gene Expression

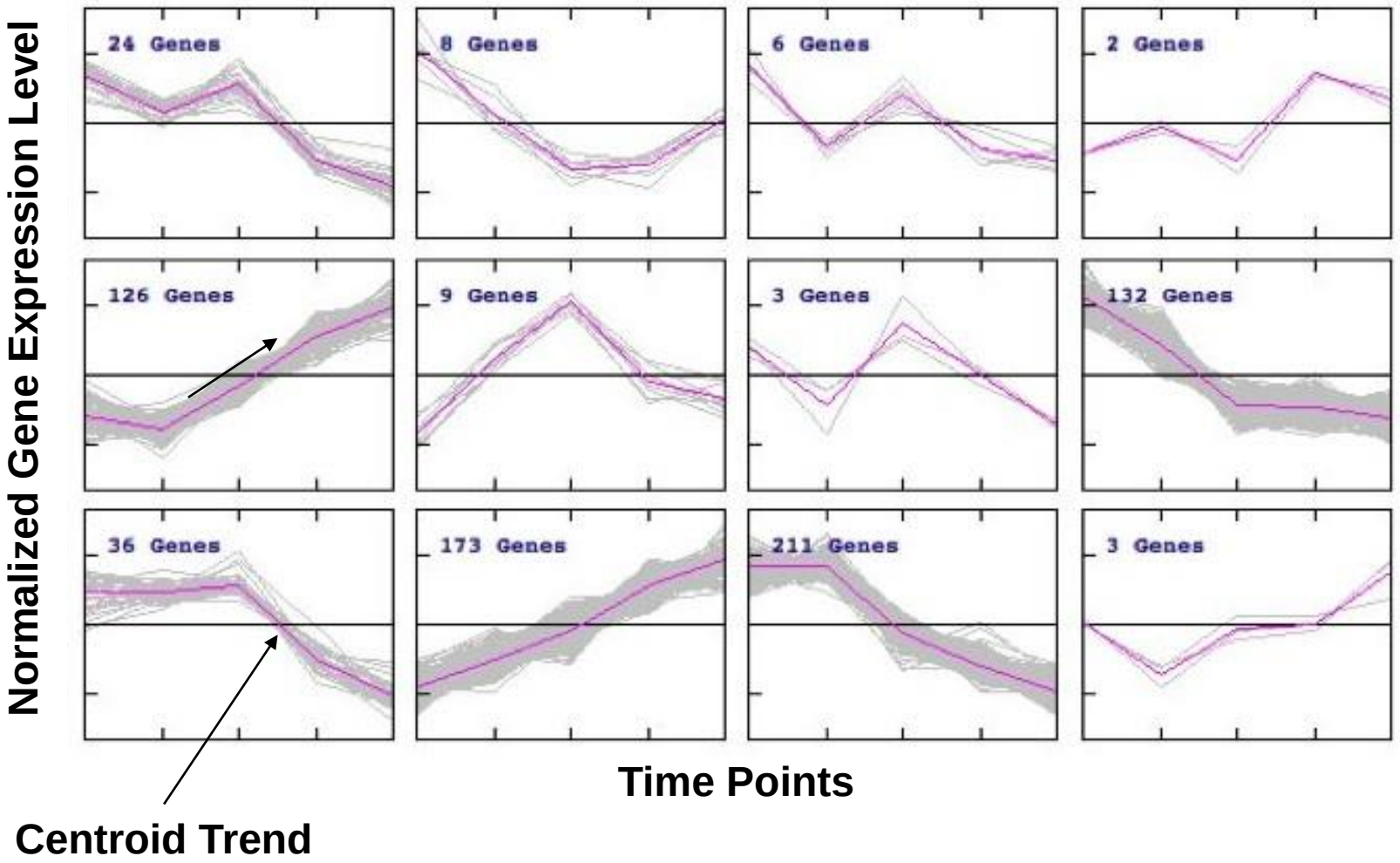


Image from http://www2.warwick.ac.uk/alumni/services/eportfolios/hrrgak/project_overview/systems_biology/

Hierarchical Clustering

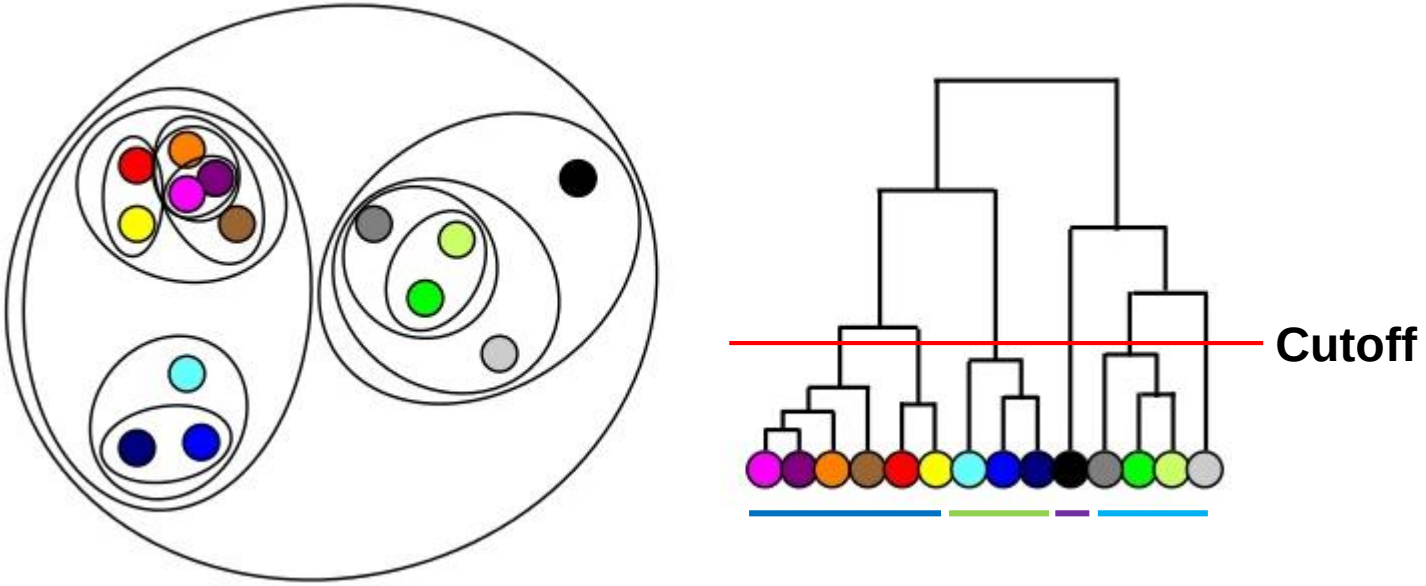
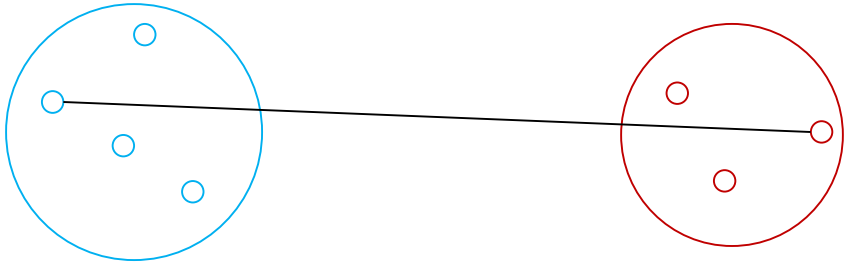


Image from https://www.slideshare.net/ElenaSgis/data-preprocessing-and-unsupervised-learning-methods-in-bioinformatics

- **Each step finds two data points or existing clusters that are closest to each other and group them together.**
- **The number of clusters is defined afterward by setting a cutoff on the distance.**
- **Choices of distance function.**
- **Choices of how to measure distance between clusters (e.g. using cluster centroids or closest members or all members).**

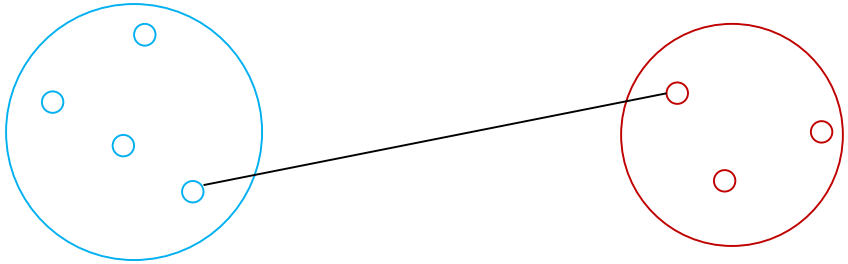
Linkage Criteria: Distance Between Clusters

- **Maximum or complete linkage**



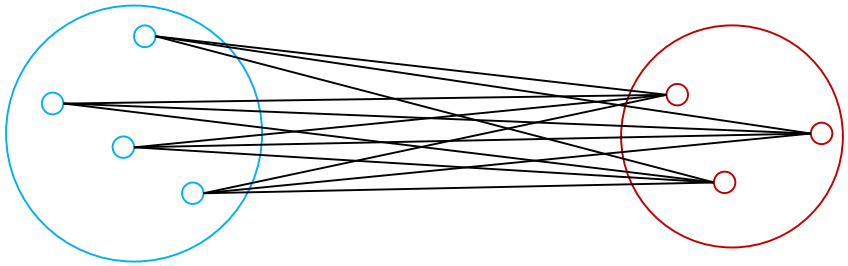
Furthest data points

- **Minimum or single linkage**



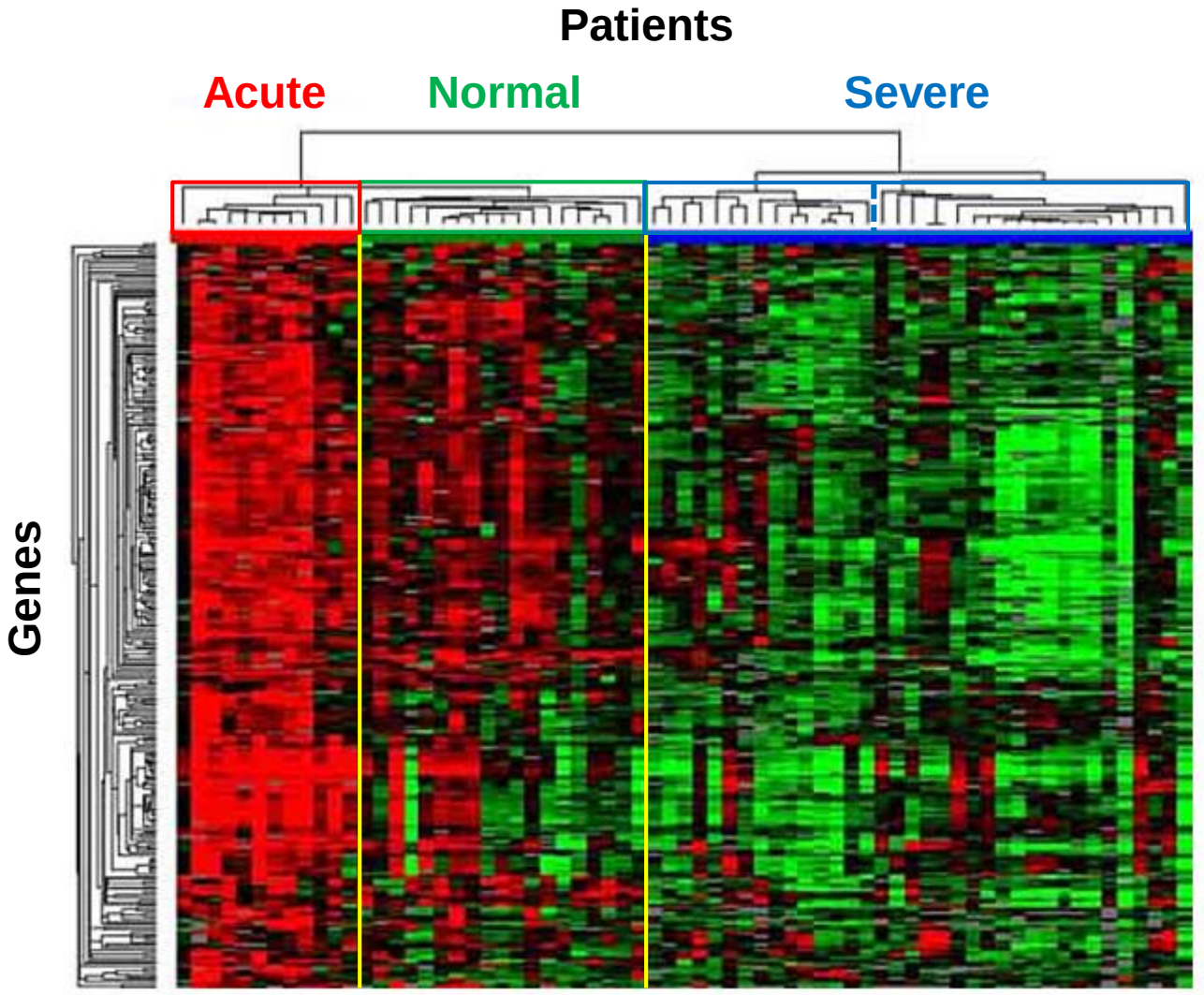
Closest data points

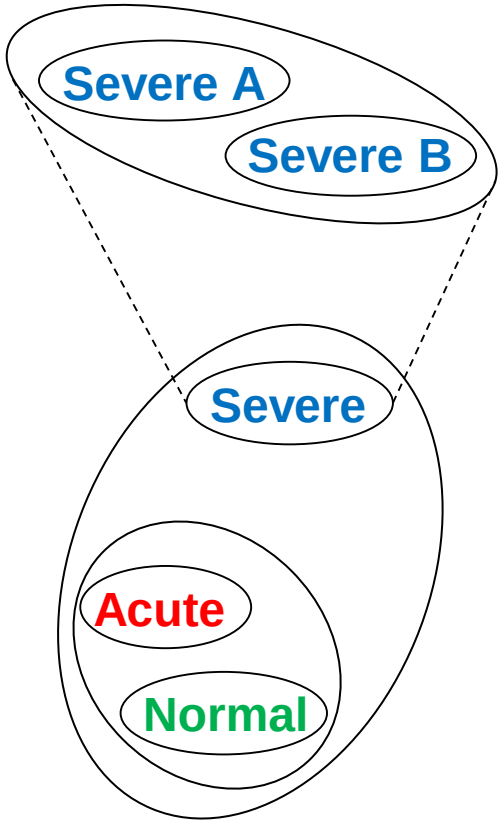
- **Mean or average linkage**



Average

Hierarchical Clustering Of Gene Expression

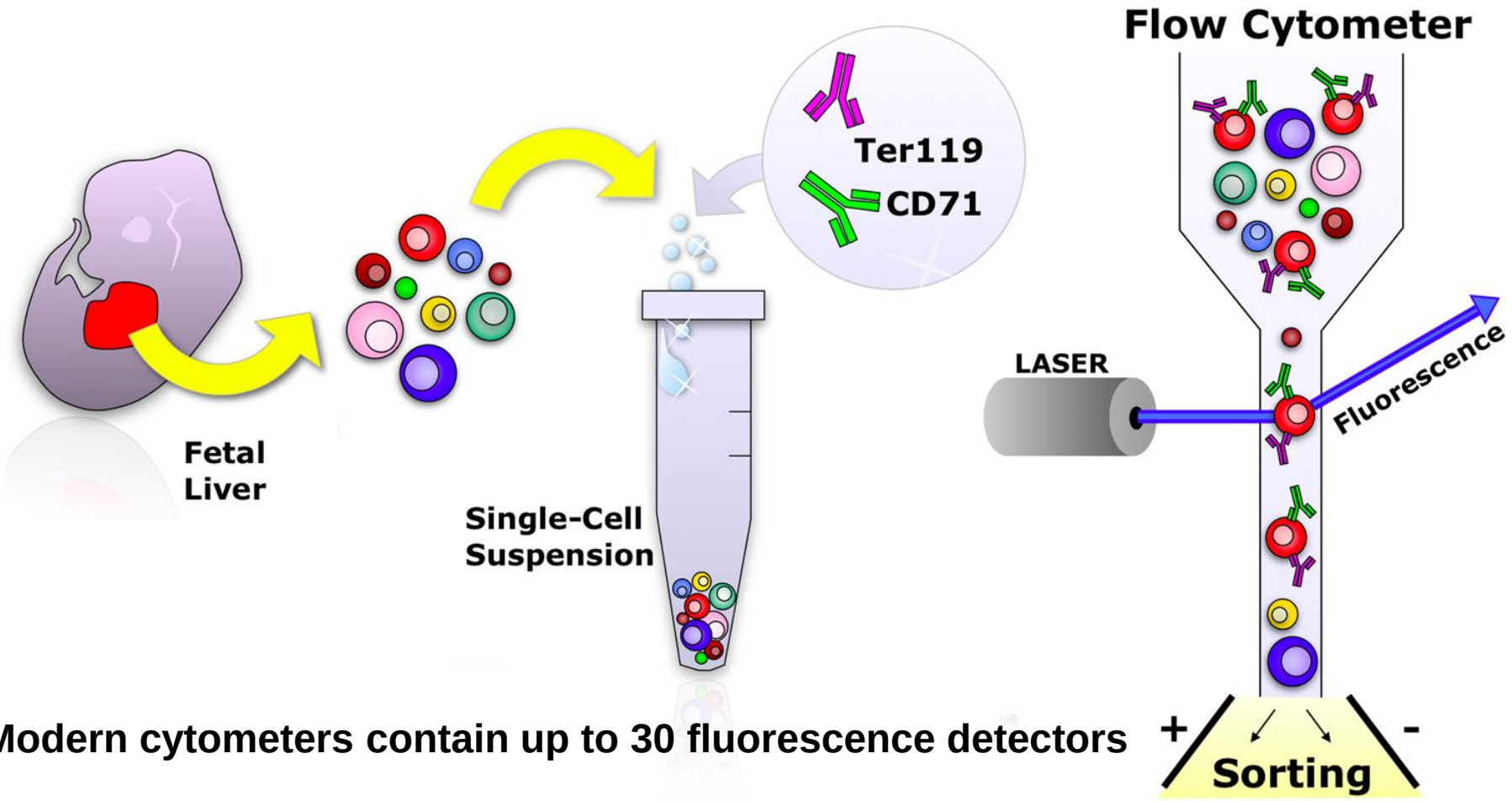




Adapted from Chua *et al*. Frontiers in Bioscience 8, s913-923 (2003)

Application IV:
Annotation of Different Cell Types

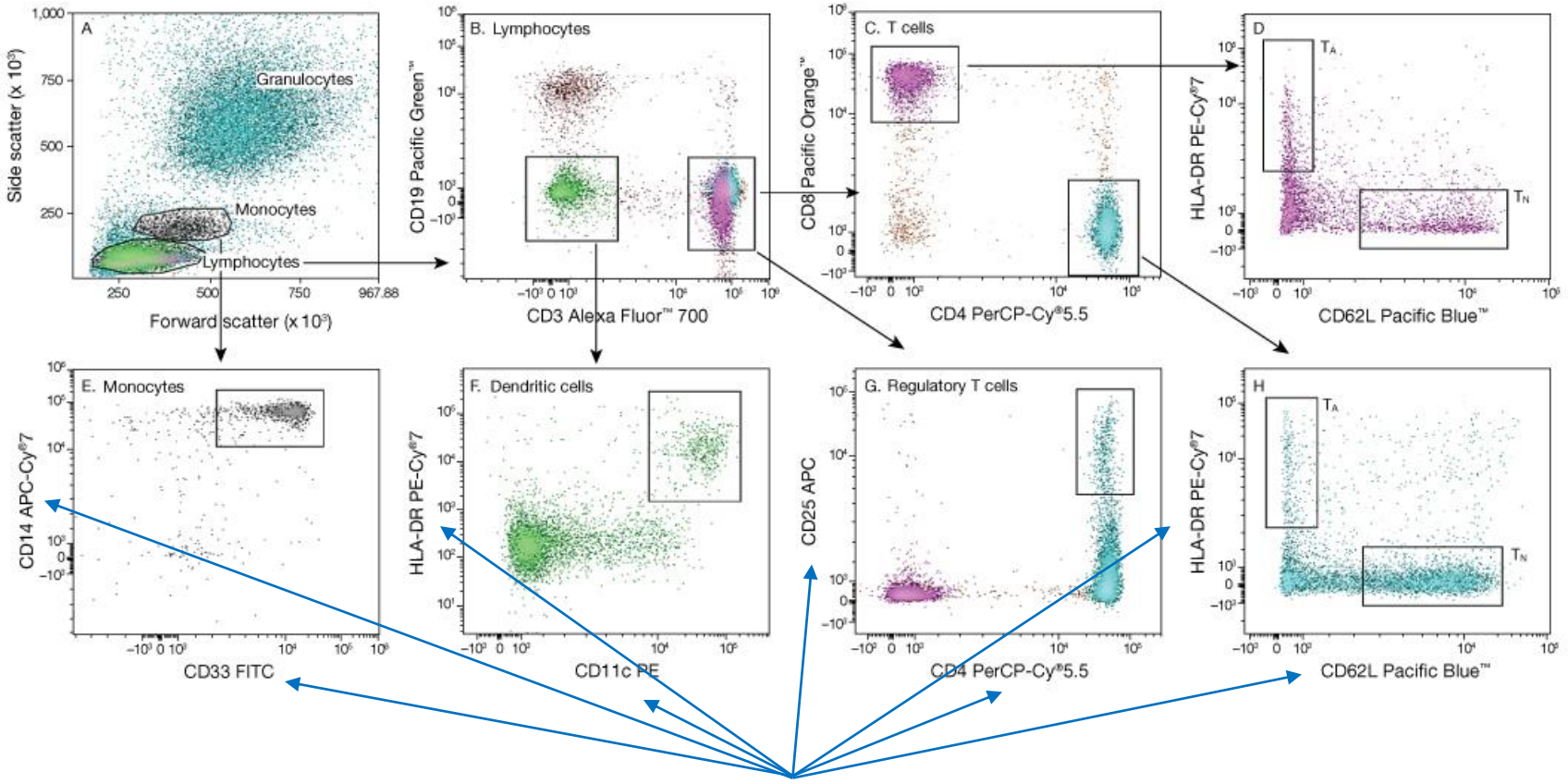
Sorting Different Cell Types



Modern cytometers contain up to 30 fluorescence detectors

Image from http://labs.umassmed.edu/socolovskylab/research~flow_cytometry.html

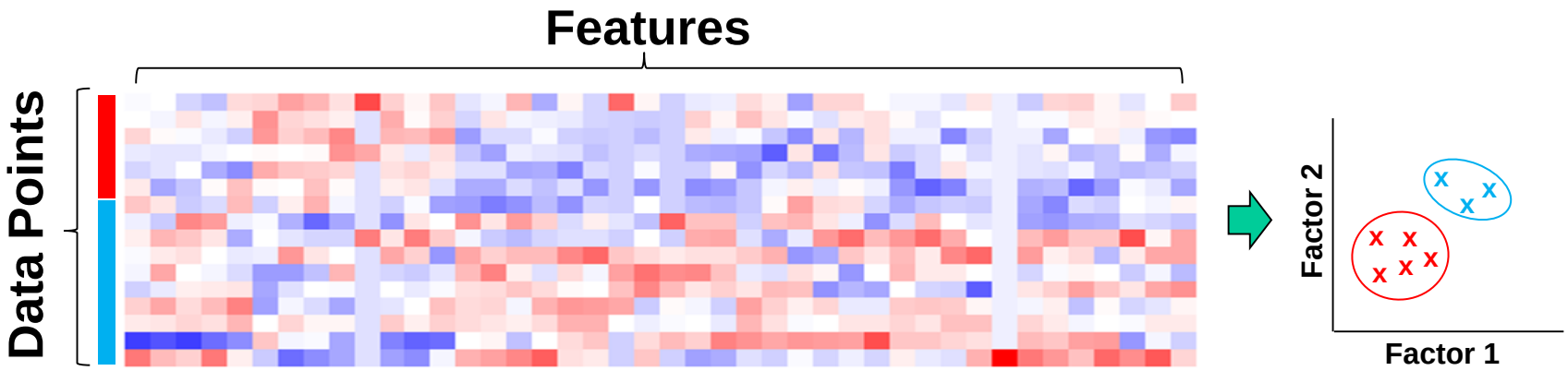
Example Of Flow Cytometry Data



Different fluorescence signals

Image from https://www.thermofisher.com/th/en/home/references/newsletters-and-journals/bioprobes-journal-of-cell-biology-applications/bioprobes-71/bioprobes-71-flow-cytometry-panel-design.html

Simplifying High-Dimensional Data



Dimensionality Reduction

PCA, t-SNE

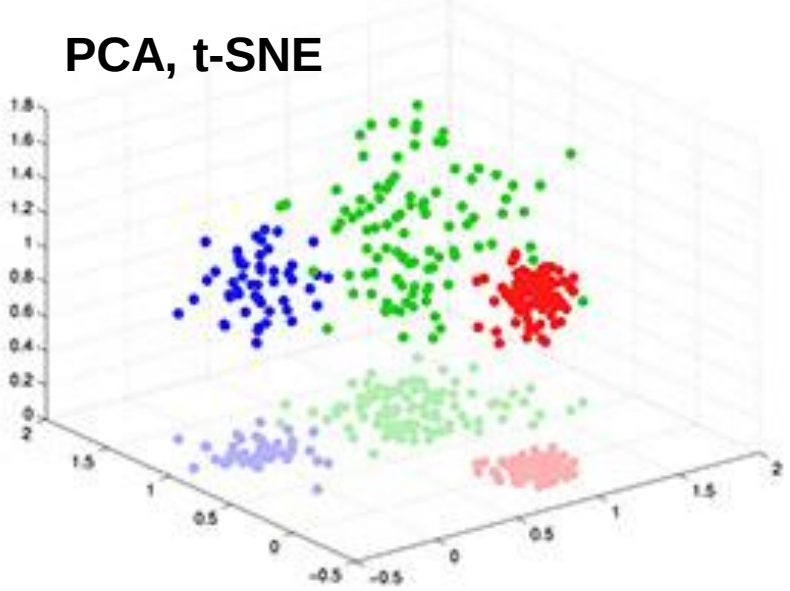


Image from http://bigdata.csail.mit.edu/node/277

Classification

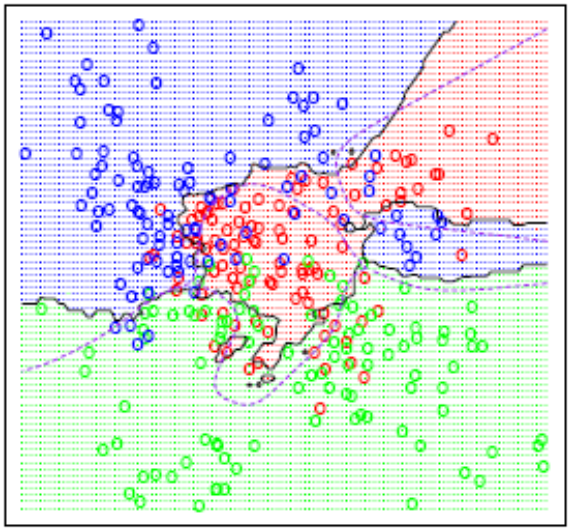
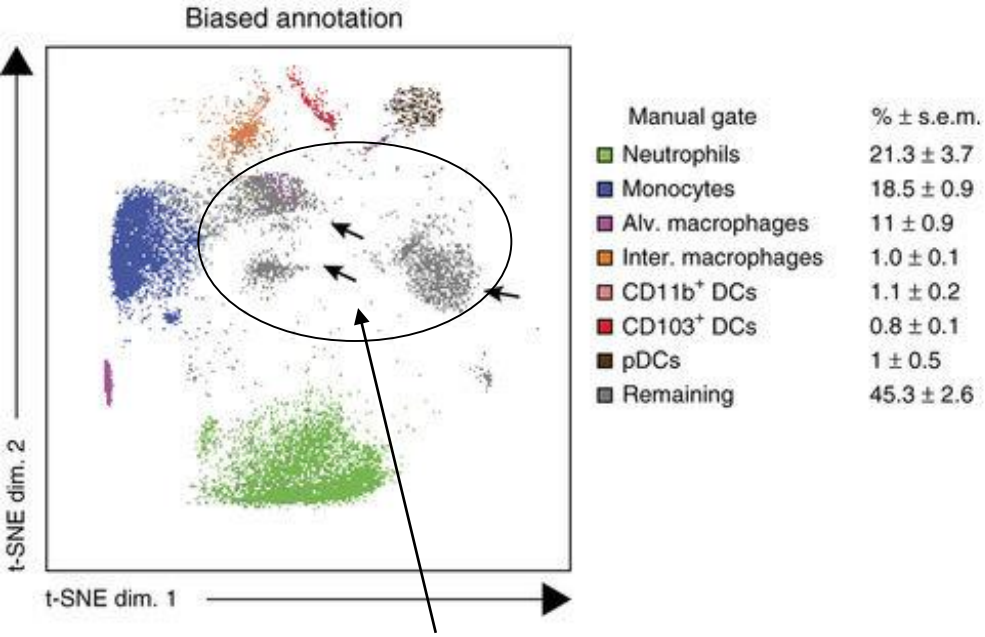
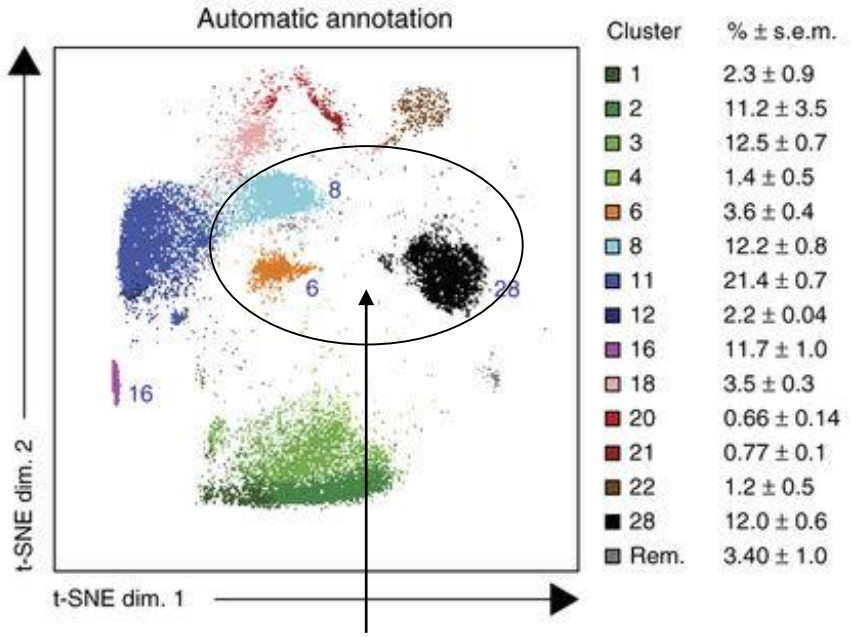


Image from http://www.stat.ucla.edu/~ybzhao/teaching/stat101c/

Identification Of New Cell Types



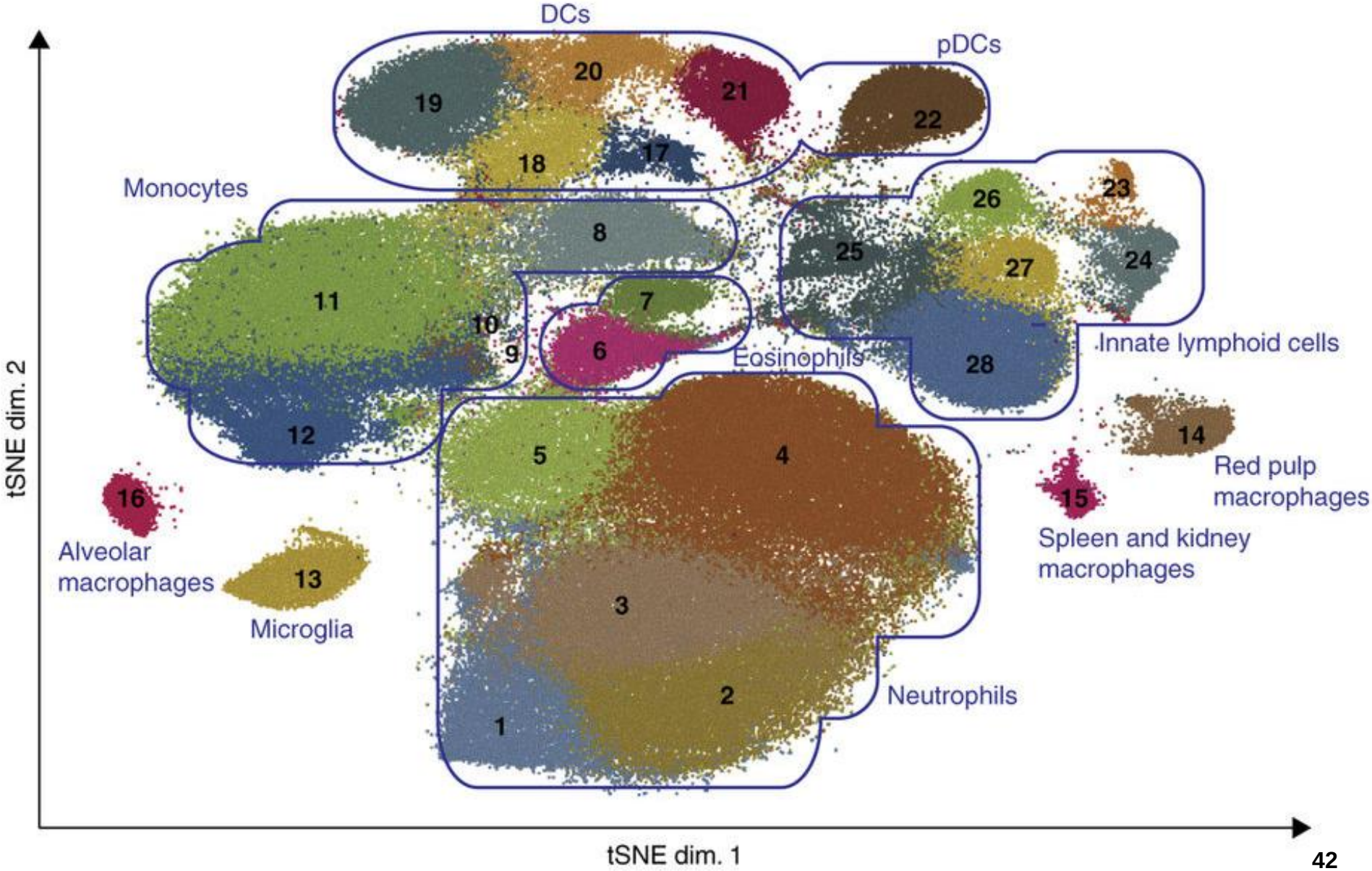
Unannotated cell types according to existing knowledge



Clustering process defines these as new cell types

Becher *et al*. Nature Immunology 15, 1181-1189 (2014)

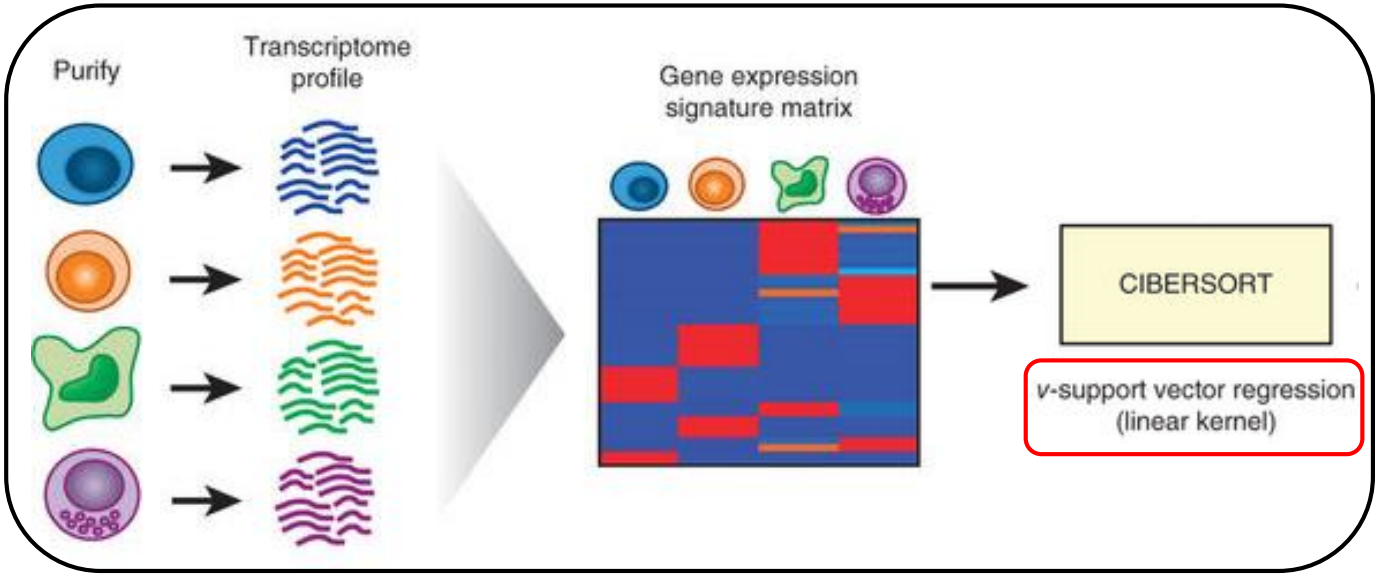
Automatic Cell Type Classification



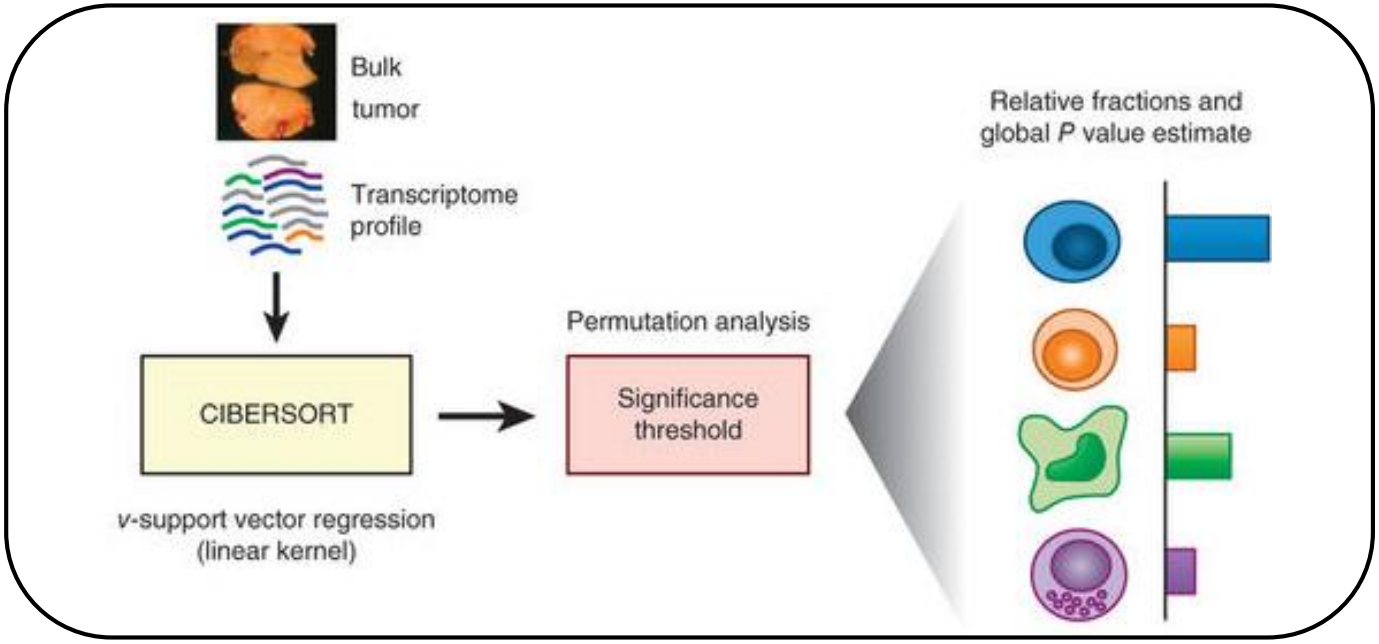
Becher *et al*. Nature Immunology 15, 1181-1189 (2014)

Deconvolution Of Cell Type Composition

Training



Testing



Newman *et al*. Nature Methods 12, 453-457 (2015)

Cell Composition Reflects Disease State

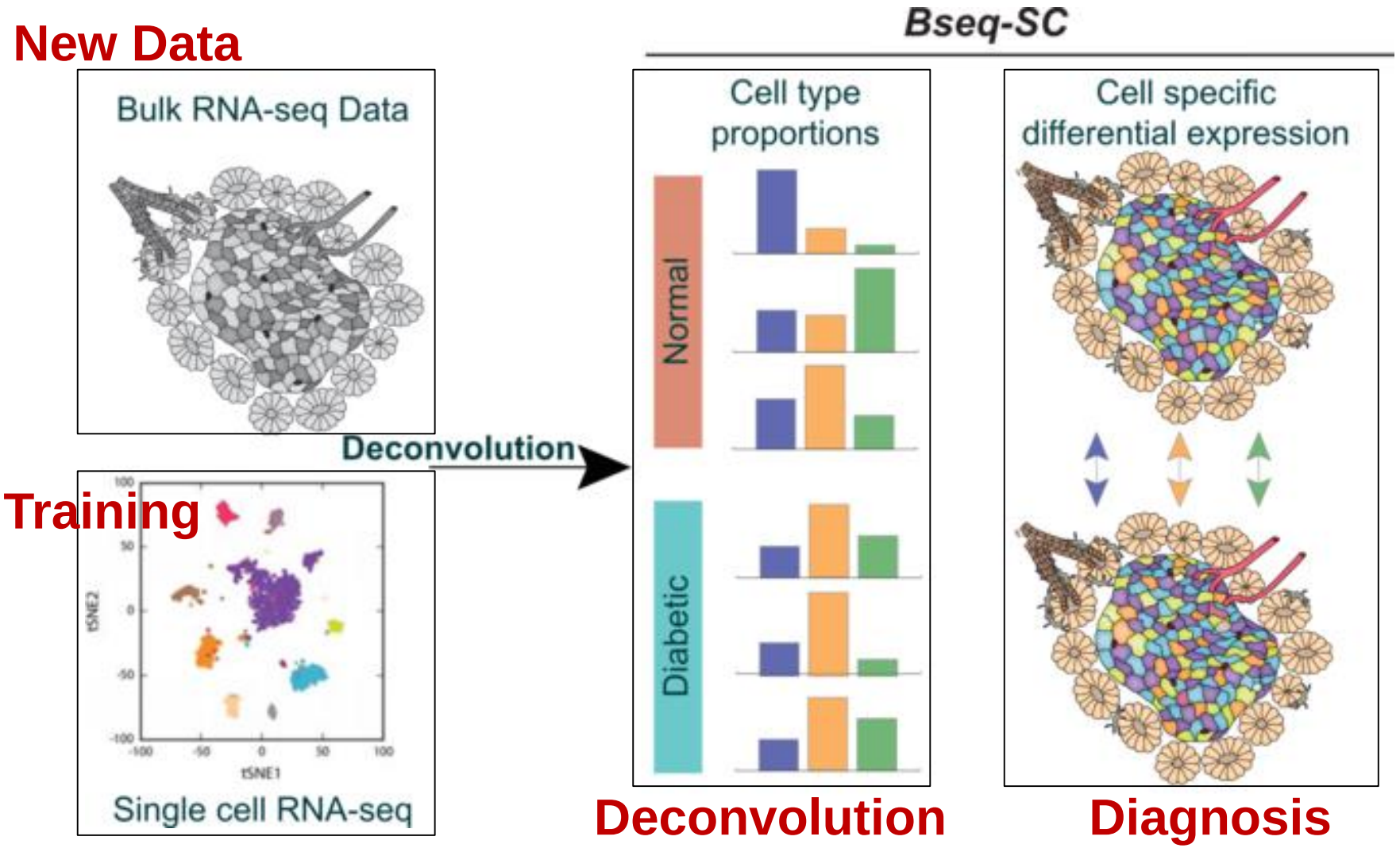


Image from https://shenorrlab.github.io/bseqsc/index.html

Application V:
Protein-To-DNA Binding Prediction

Transcription Factor (TF) Binding

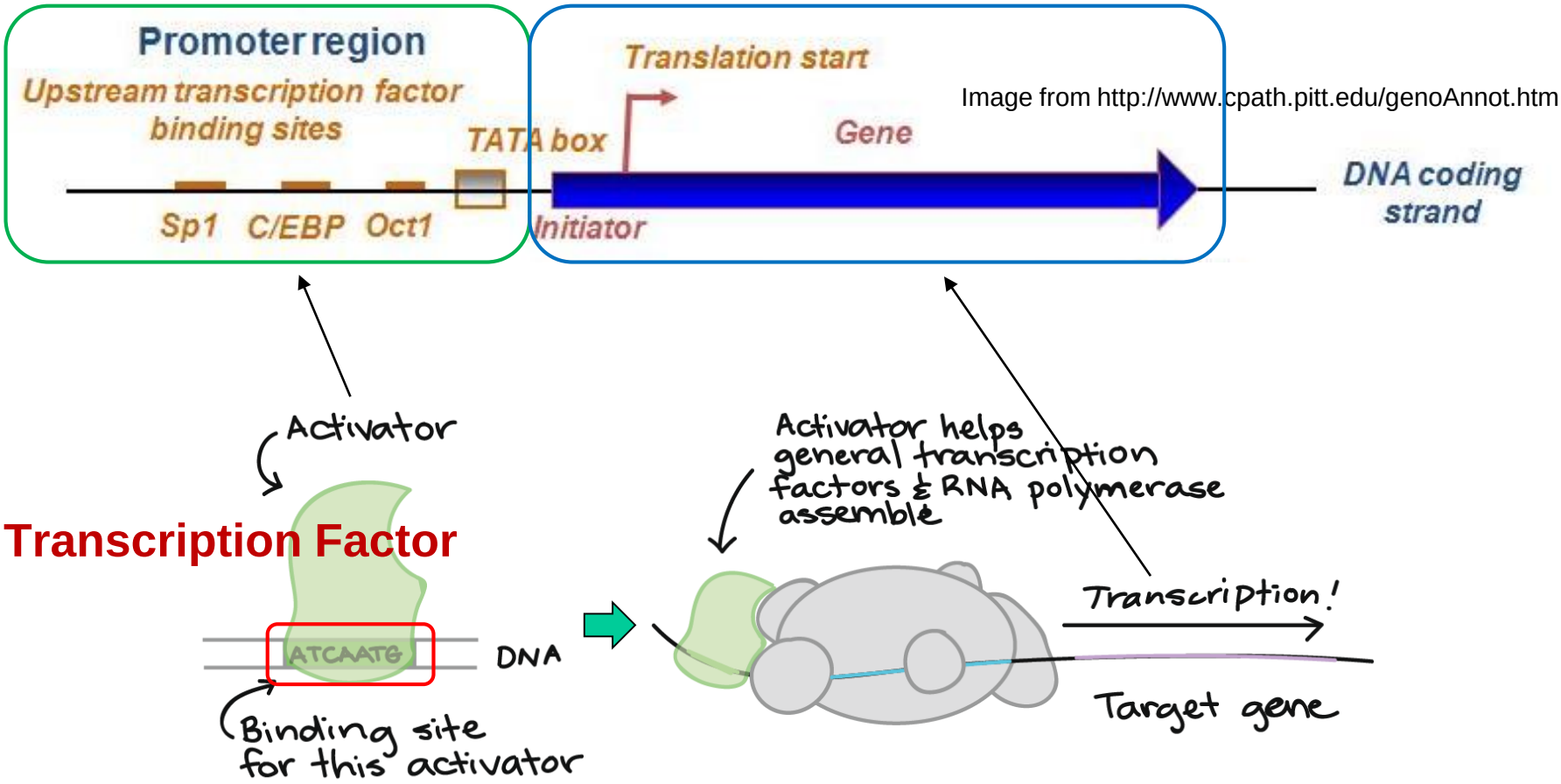
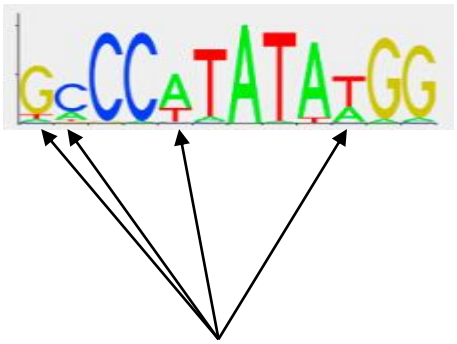


Image from http://www.cpath.pitt.edu/genoAnnot.htm

Image from https://www.khanacademy.org/science/biology/gene-regulation/gene-regulation-in-eukaryotes/a/eukaryotic-transcription-factors

Sequence-Based Prediction Of TF Binding

Expected Motif



High chance of false positives due to flexibility of binding motifs.

Genome Sequence

...ACCAGCGGCGAAGCTCGGGGCGGAGGGGT
TGAGCCACATGAGGCGATGGCGACA**TCCCATA**
TATGGAGACATGGCGTGGCTGGCTGTTACATT
TTGTTTTGATGAAAAGCATAACCATGCGGATG
ATATTTTTATTATAGACTAGAGATGATTATTG
AATAGAC**ATGCTCTTAACCATTTTTAACTCTA**
ATTCCAC...
...ACCAGCGGCGAAGCTCGGGGCGGAGGGGT
TGAGCCACATGAGGCGATGGCGACAGGGACCT
CC**GACCTTATAAGG**AGACATGGCGTGGCTGGC
TGTTACATTTTGTTTTGATGAAAAGCATAACC
ATGCGGATGATATTTTTATTATAGACTAGAGA
TGATTATTGAATAGGCCTACTTTAC**ATGCTCT**
TAACCATTTTTAACTCTAATTCCAC...

Predicted binding sites

Downstream genes

Adapted from Mathelier *et al*. Cell Systems 3, 278-286 (2016)

DNA Packaging Affects TF Binding

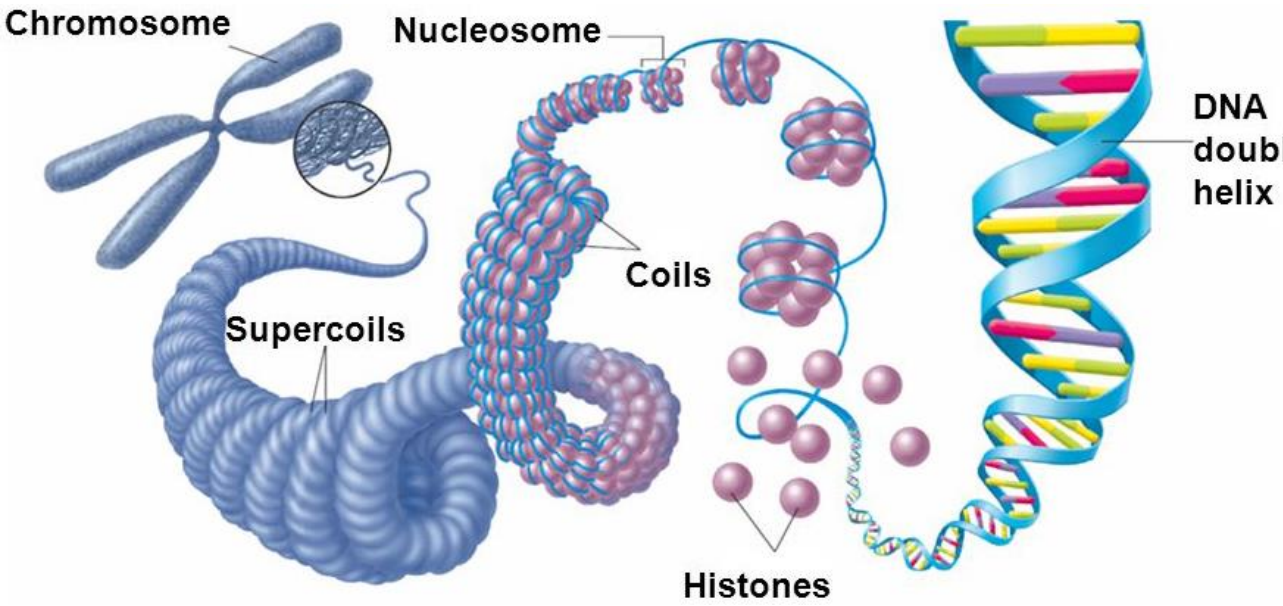
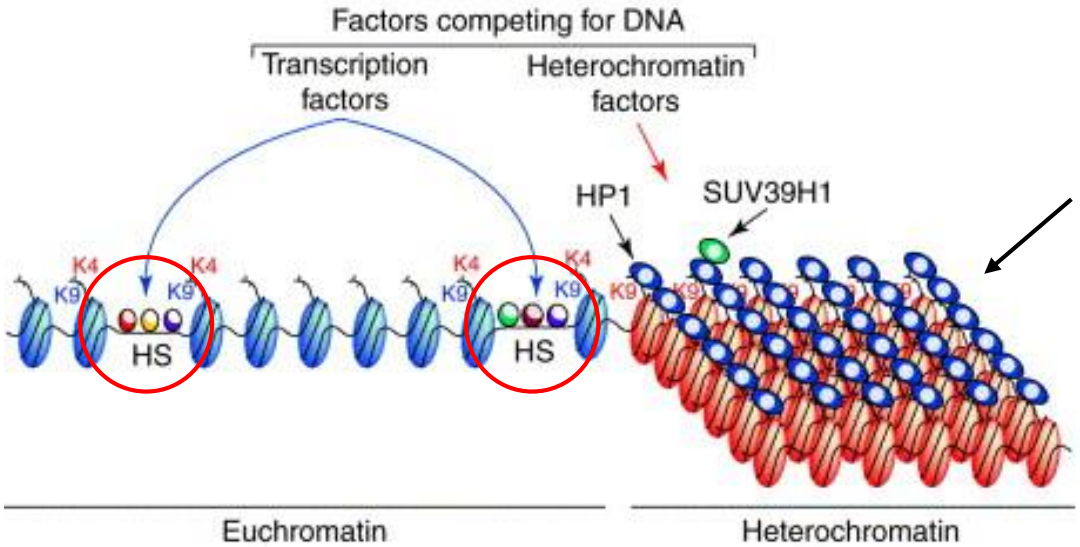


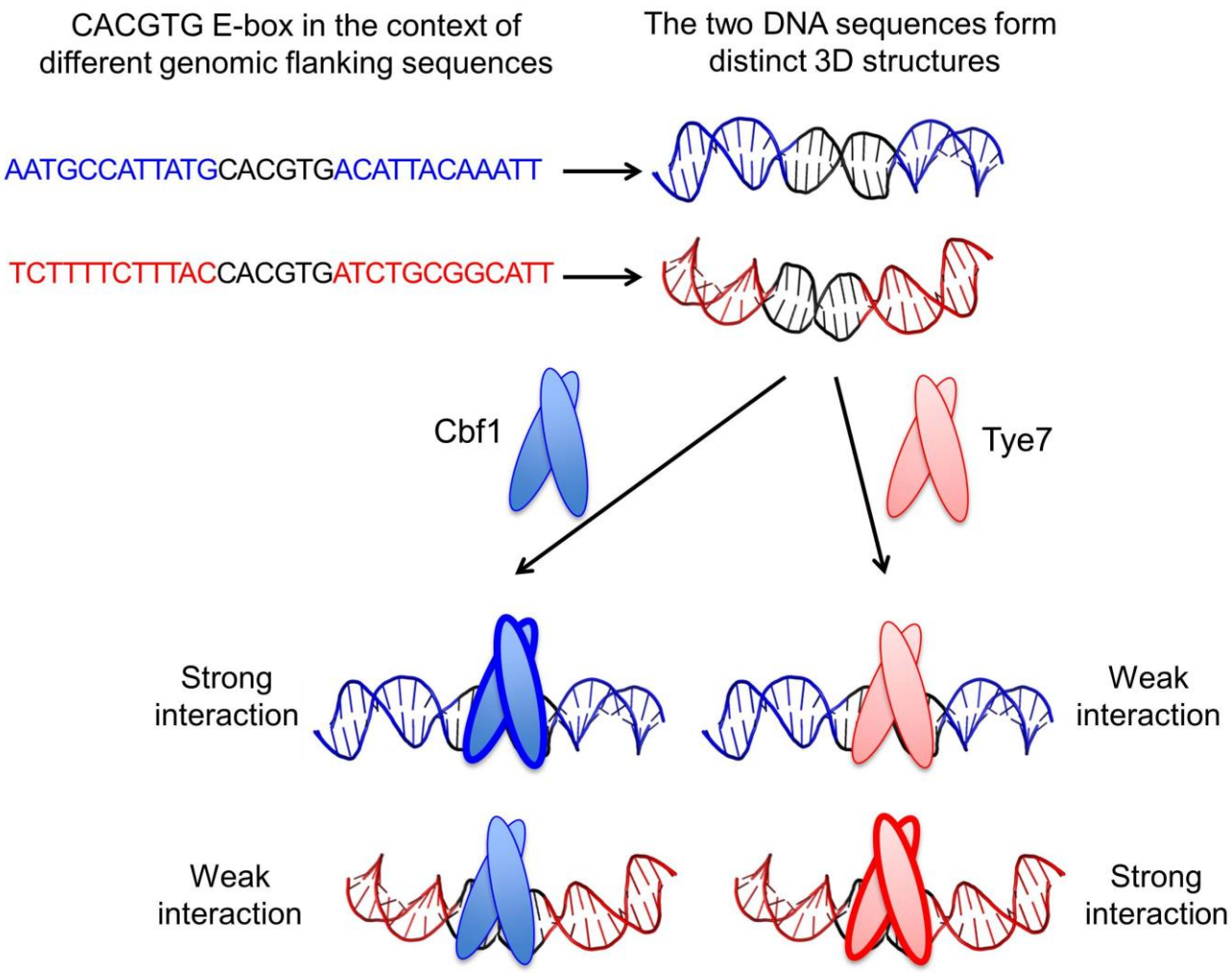
Image from http://images.slideplayer.com/9/2508104/slides/slide_28.jpg



Transcription factors cannot access tightly packed DNA region.

Dillon *et al*. Trends in Genetics 18, 252-258 (2002)

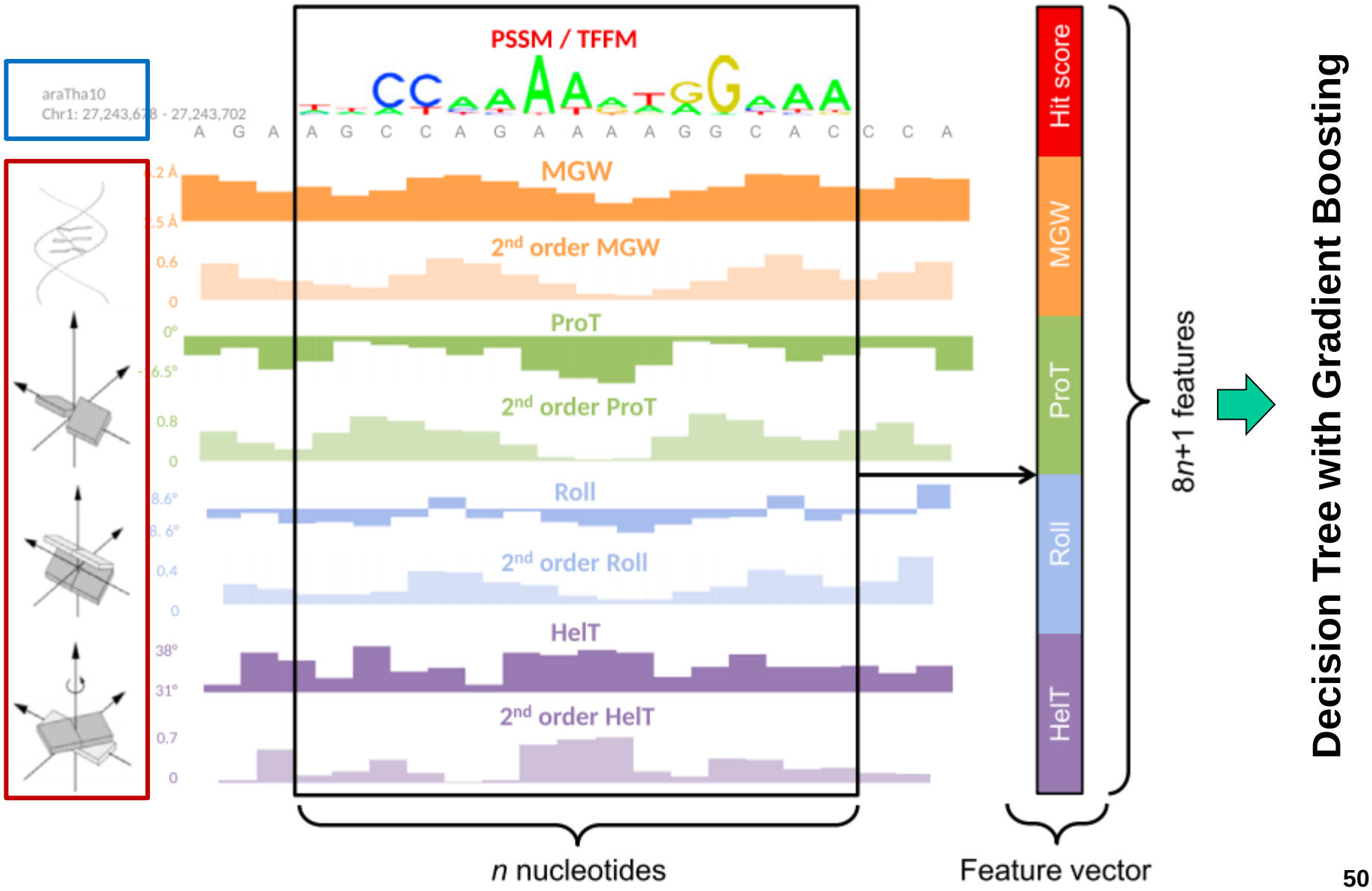
Local DNA 3D Structure Affects TF Binding



Flanking sequences differentially affect binding of Cbf1 and Tye7, through DNA shape

Gordan *et al*. Cell Reports 3, 1093-1104 (2013)

Prediction Of TF Binding Using Physical Data



Mathelier *et al*. Cell Systems 3, 278-286 (2016)

Application VI:
Human Genome Structure

Human Genome

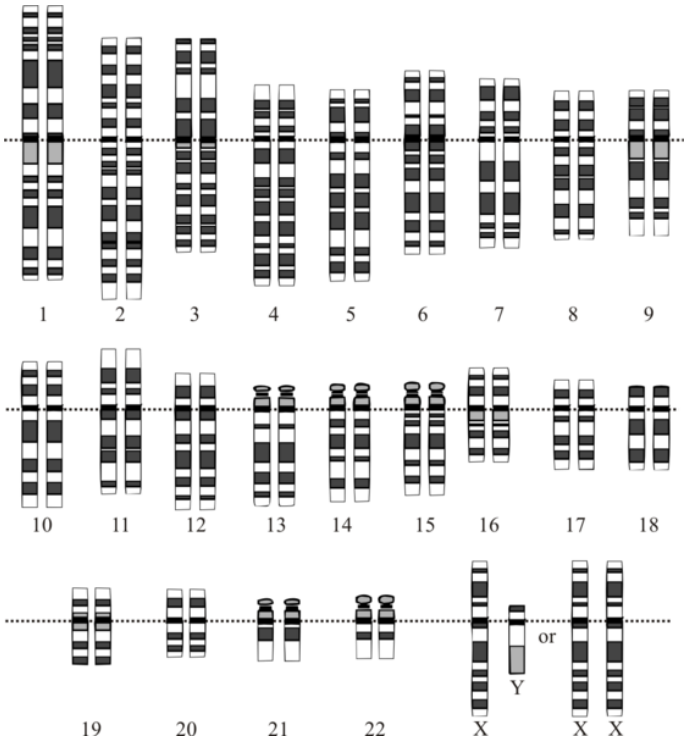


Image from https://en.wikipedia.org/wiki/Human_genome#Coding_vs._noncoding_DNA

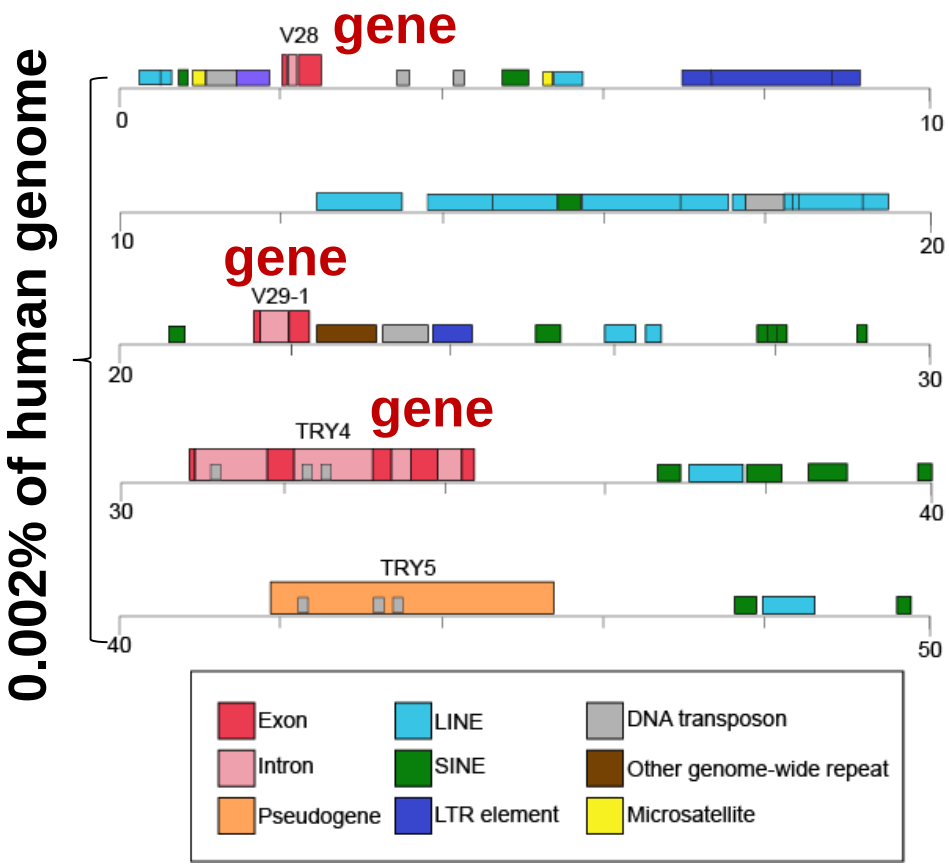
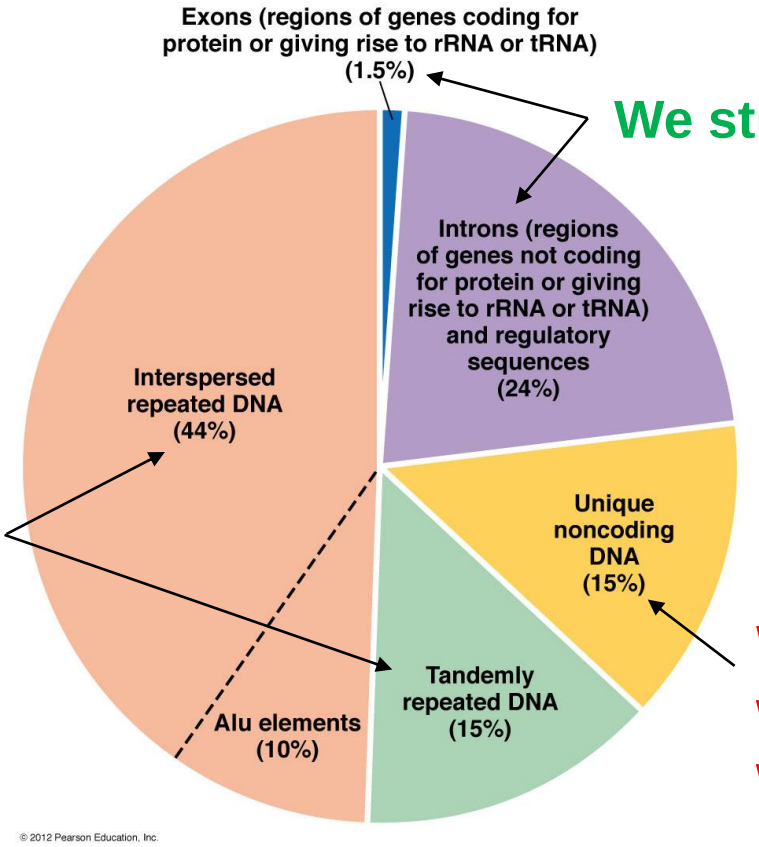


Image from https://bio.libretexts.org/LibreTexts/University_of_California_Davis/BIS_2A%3A_Introductory_Biology_(Easlon)/Readings/26%3A_Genomes%3A_a_Brief_Introduction

- **3.2 Giga-basepairs x2 on 23 pairs of chromosomes.**
- **~0.1% variation between individuals.**
- **<1.5% code for proteins (exons).**

Much Of Genome Remains Unknown



We studied them extensively

We have some clue about what they are but not completely about their impacts on our genome

We don't really know what they are nor what they do

Image from http://www.mun.ca/biology/desmid/brian/BIOL2060/BIOL2060-18/CB18.html /

- **Using data from well-studied genomic regions to predict the structures and functions of other genomic regions in the same or in newly discovered species.**

Dynamic Bayesian Network (DBN)

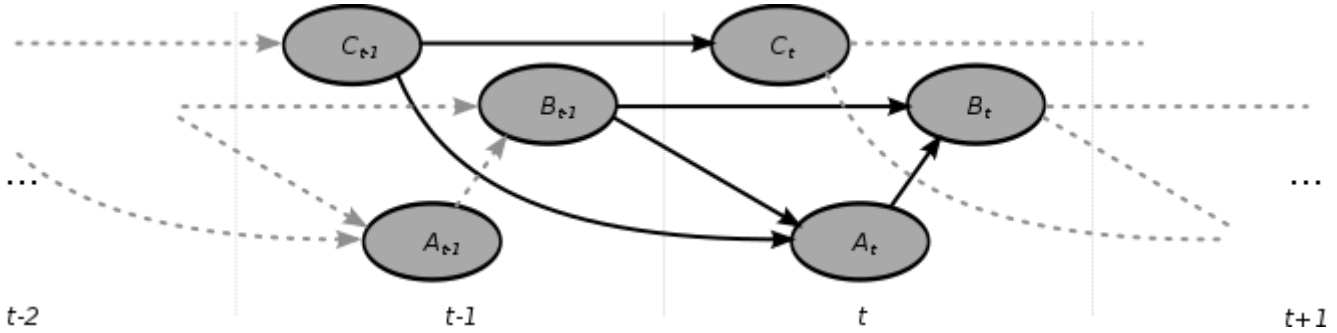


Image from https://en.wikipedia.org/wiki/Dynamic_Bayesian_network

- **The label of a genomic location can be predicted based on the labels and properties of nearby positions.**
- **For example, A_t may indicate label – such as exon or intron – at location *t* while B_t and C_t keep track of the properties of the genome – such as coding-vs-non-coding or loosely-packed-vs-tightly-packed – at location *t*.**
- **The probability of predicting a label depends on B_t and C_t.**

Histone Modifications

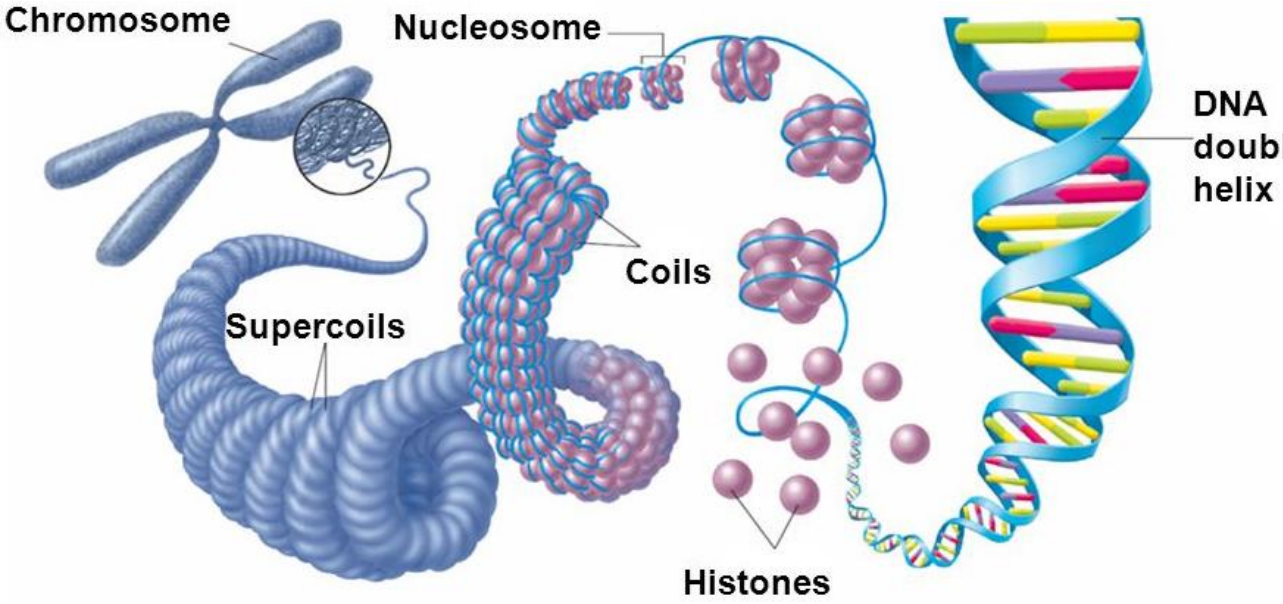
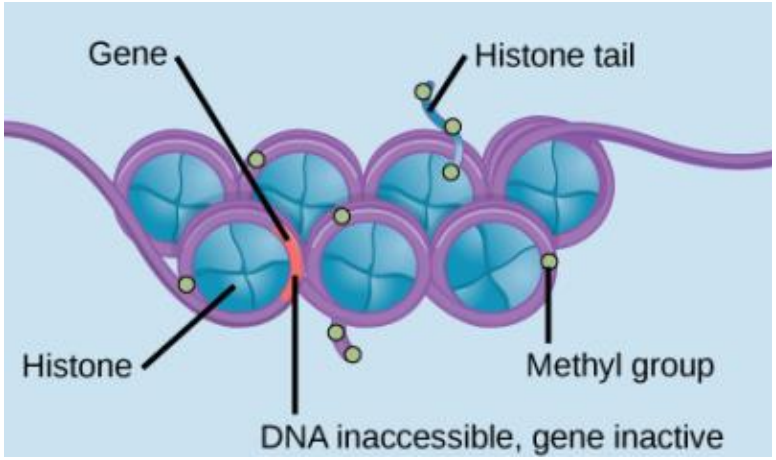


Image from http://images.slideplayer.com/9/2508104/slides/slide_28.jpg



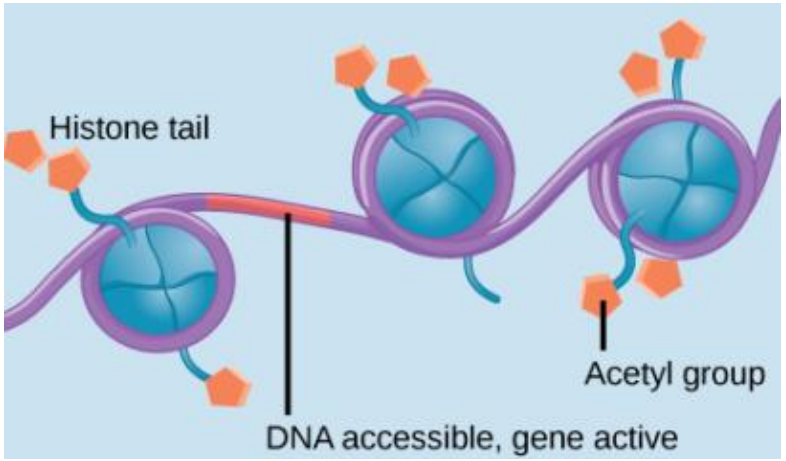
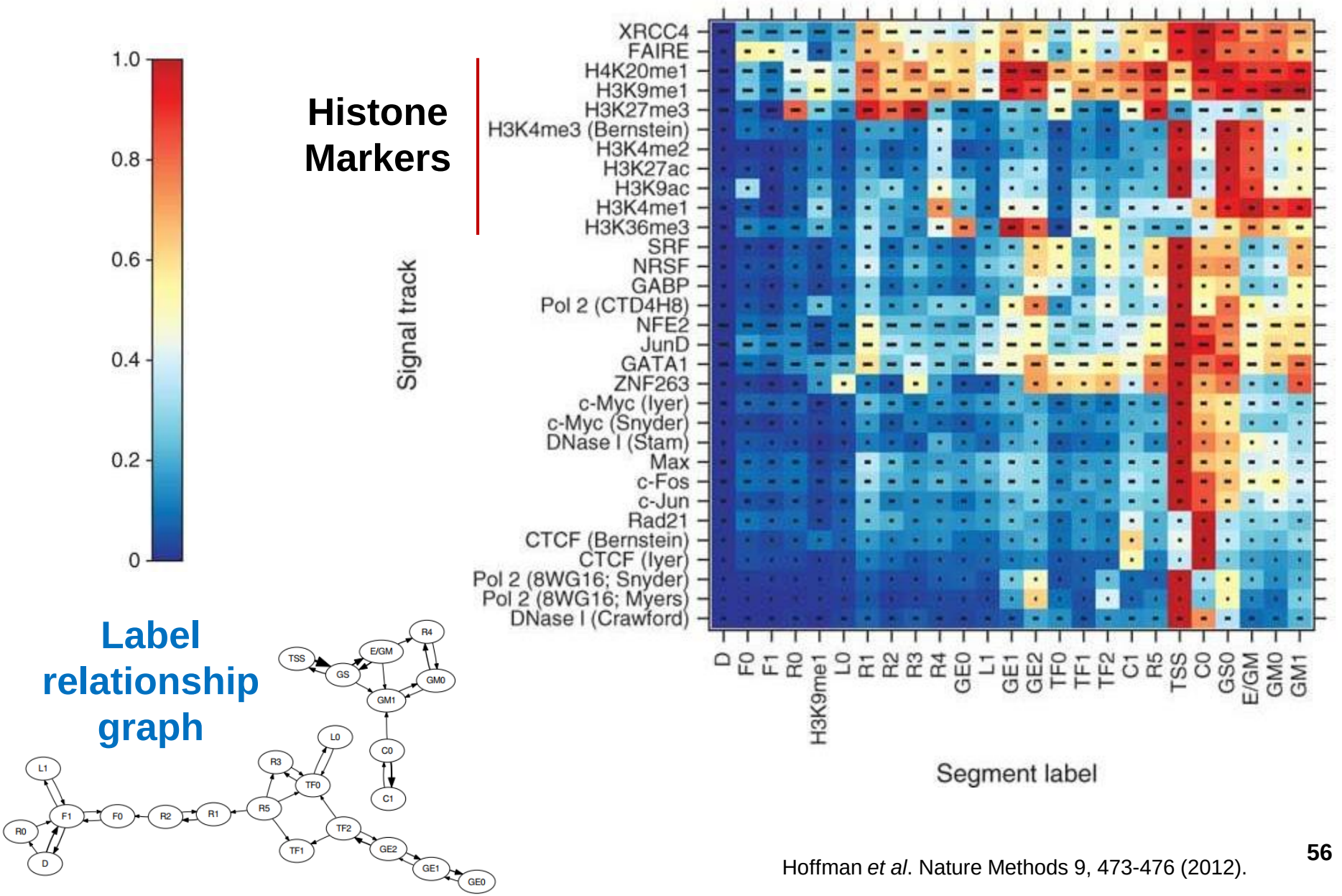


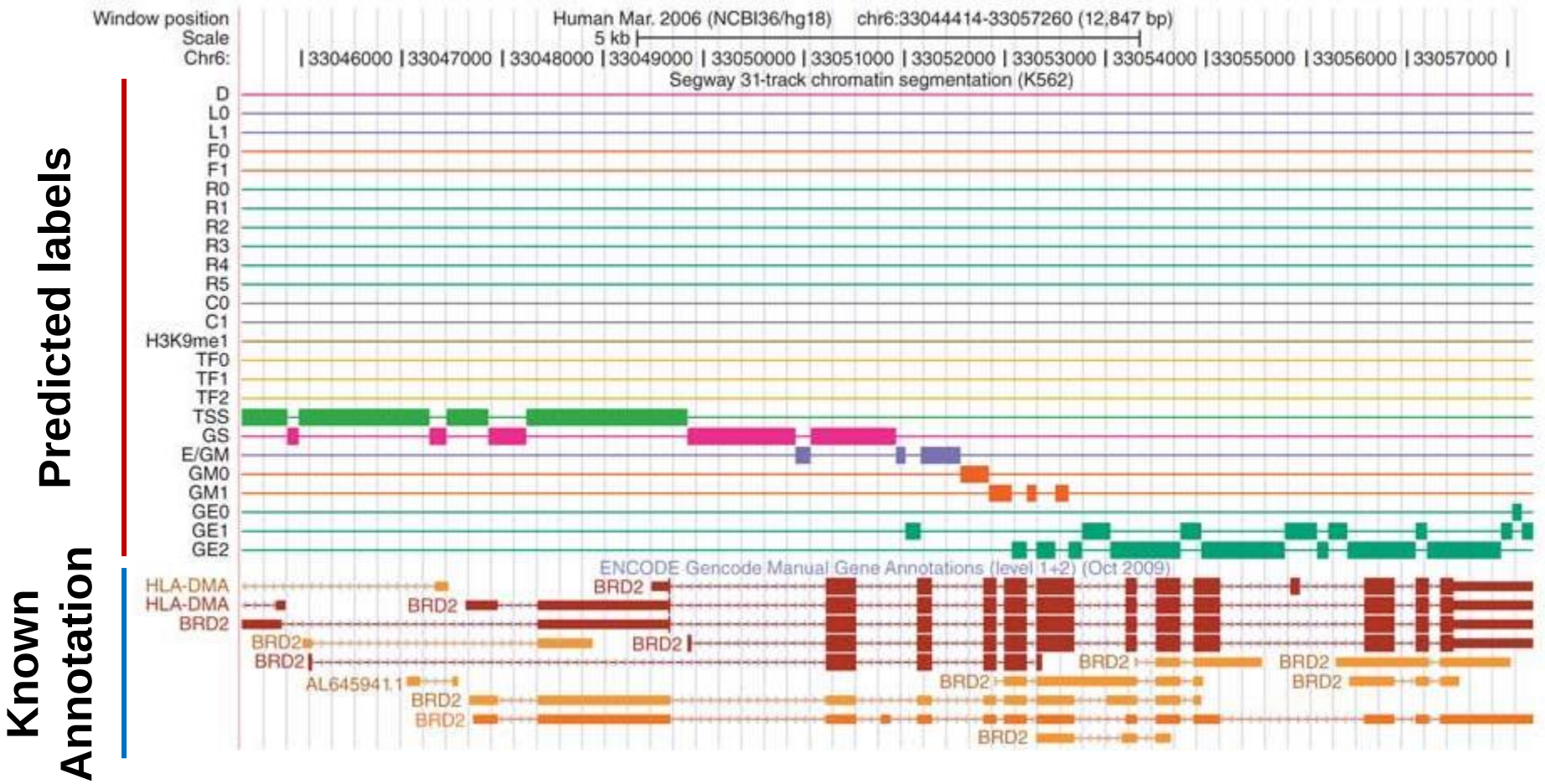
Image from http://www.peirsoncenter.com/articles/dna-methylation-in-down-syndrome

Inferring Labels From Genomic Signals



Hoffman *et al*. Nature Methods 9, 473-476 (2012).

Decent Performance Without Using Sequence

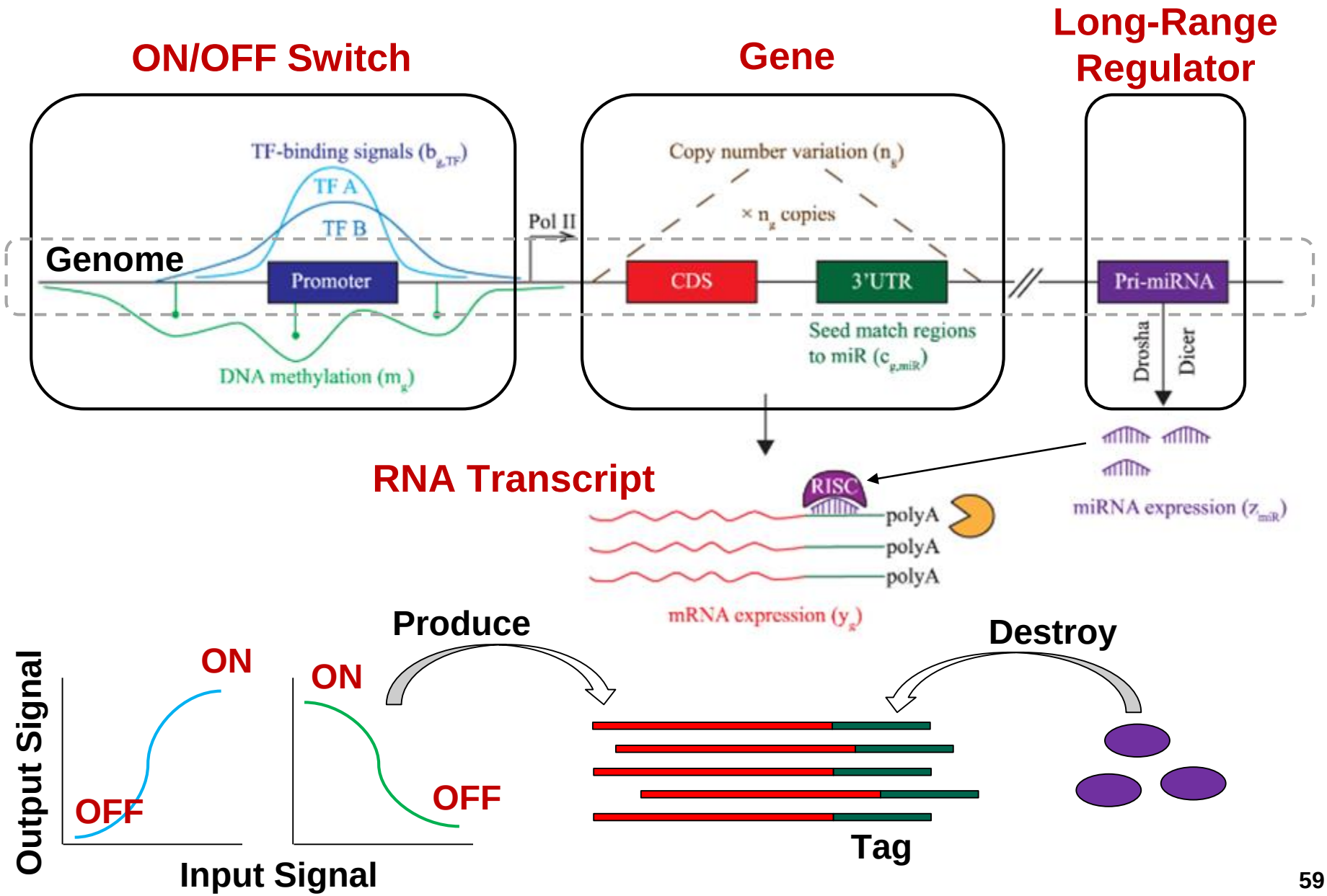


- **Thick bars at the bottom are translated regions (exons).**
- **There are multiple isoform of BRD2 (i.e. different start locations and exon structures)**

Hoffman *et al*. Nature Methods 9, 473-476 (2012).

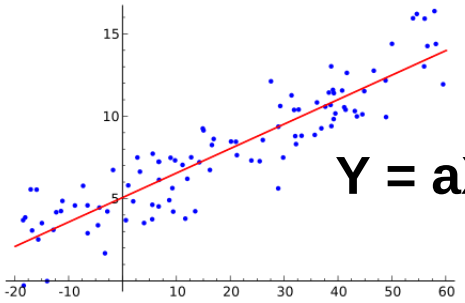
Application X:
Regulation Of Gene Expression

Regulation Of Gene Expression

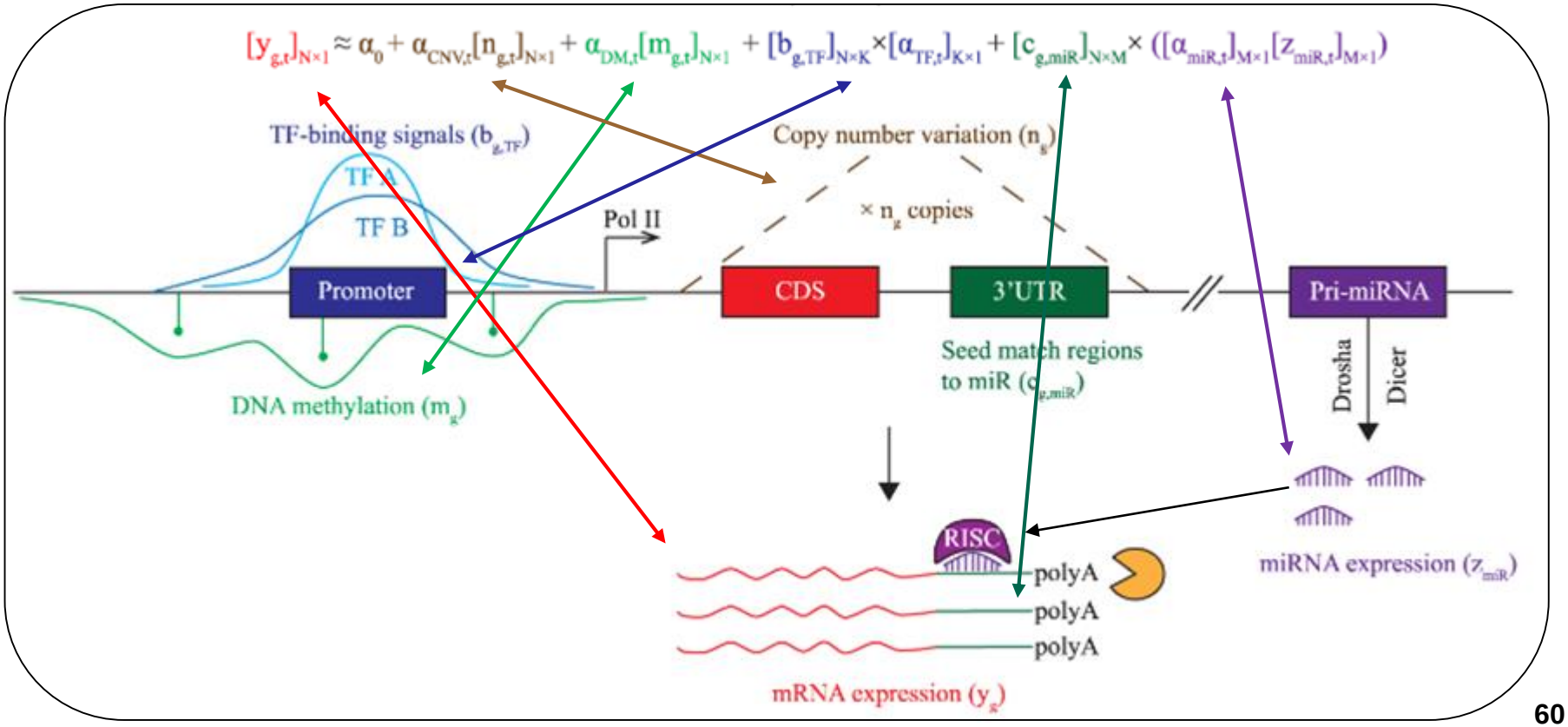


Adapted from Li *et al*. PLoS Comp Biol 10, e1003908 (2014)

Regression Model For Gene Expression I



$Y = aX + b$ ⇒ $Y = a_0 + a_1X_1 + a_2X_2 + \ldots + a_nX_n$



Adapted from Li *et al*. PLoS Comp Biol 10, e1003908 (2014)

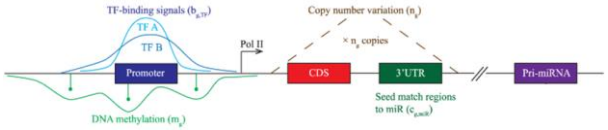
Regression Model For Gene Expression II

Sample #M

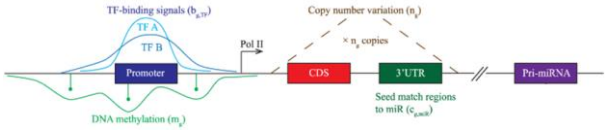
⋮

Sample #2

Sample #1

Gene #1: 

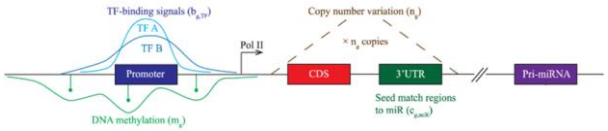
$$Y_1 = a_{01} + a_{11}X_{11} + a_{21}X_{21} + \ldots + a_{n1}X_{n1}$$

Gene #2: 

$$Y_2 = a_{02} + a_{12}X_{12} + a_{22}X_{22} + \ldots + a_{n2}X_{n2}$$

Adapted from Li *et al*. PLoS Comp Biol 10, e1003908 (2014)

⋮ ⋮

Gene #N: 

$$Y_N = a_{0N} + a_{1N}X_{1N} + a_{2N}X_{2N} + \ldots + a_{nN}X_{nN}$$

Same coefficients $a_{i,j}$ across samples

Regression Model For Gene Expression III

Condition A

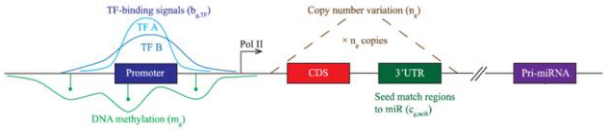
Sample A_M

⋮

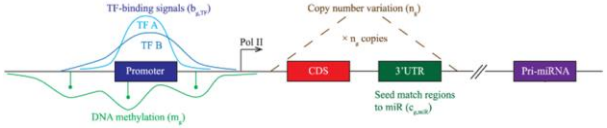
Sample A_2

Sample A_1

Gene #1:



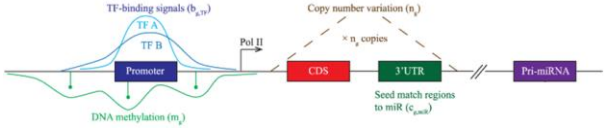
Gene #2:



Adapted from Li *et al*. PLoS Comp Biol 10, e1003908 (2014)

⋮

Gene #N:



Condition B

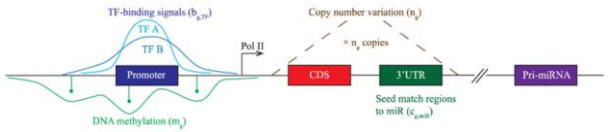
Sample B_P

⋮

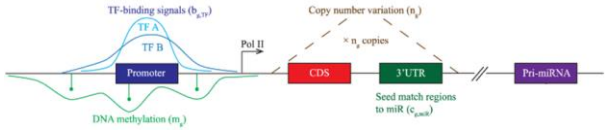
Sample B_2

Sample B_1

Gene #1:



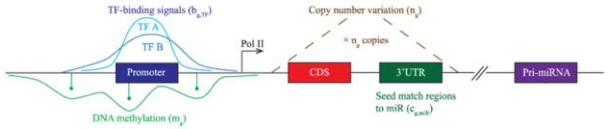
Gene #2:



Adapted from Li *et al*. PLoS Comp Biol 10, e1003908 (2014)

⋮

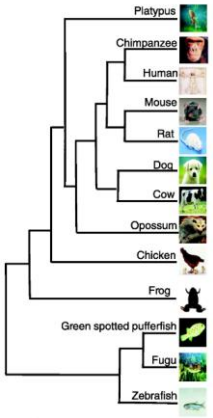
Gene #N:



Same coefficients $a_{i,j}$ across samples, different across conditions

Summary

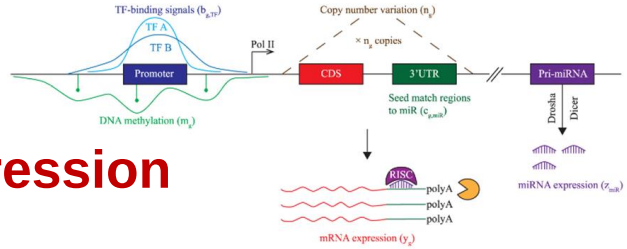
Generative Model



Maximum Likelihood Estimator

Image from http://physrev.physiology.org/content/89/3/921

Regression



Li *et al*. PLoS Comp Biol 10, e1003908 (2014)

Dynamic Bayesian Network

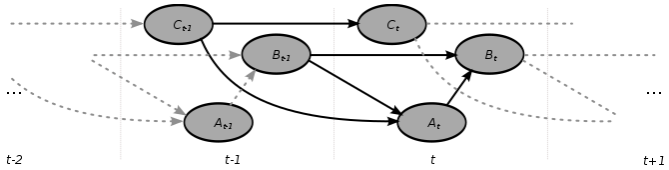
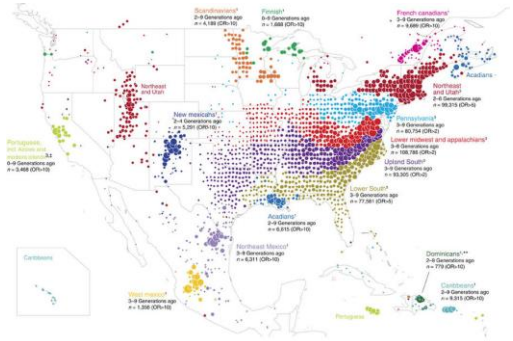
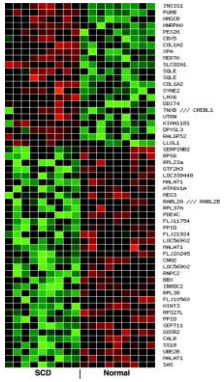


Image from https://en.wikipedia.org/wiki/Dynamic_Bayesian_network

Clustering

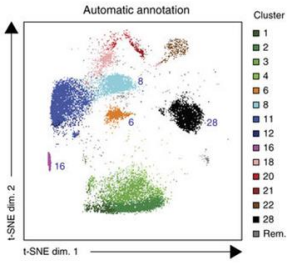




Han *et al*. Nature Communication 8, 14238 (2017)

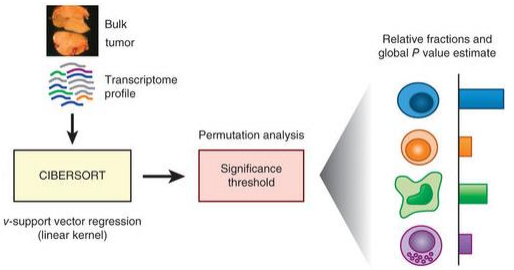
Klings *et al*. Physiological Genomics 21, 293-298 (2005)

Dimensionality Reduction tSNE



Becher *et al*. Nature Immunology 15, 1181-1189 (2014)

Support Vector Machine



Newman *et al*. Nature Methods 12, 453-457 (2015)