CS273A

A Computational Tour of The Human Genome

Lecture 6: Genes Enrichment, Gene Regulation I

MW 1:30-2:50pm in Clark **S361*** (behind Peet's)

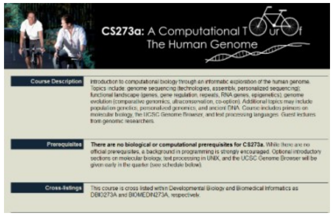
Profs: Serafim Batzoglou & Gill Bejerano

CAs: Karthik Jagadeesh & Johannes Birgmeier

* Mostly: track on website/piazza

Announcements

- http://cs273a.stanford.edu/
    - Lecture slides, problem sets, etc.
- Course communications via Piazza
    - Auditors please sign up too



CS273a: A Computational Tour of The Human Genome

Course Description — Introduction to computational biology through an informatic exploration of the human genome. Topics include: genome sequencing (technologies, assembly, personalized sequencing); functional landscape (genes, gene regulation, repeats, RNA genes, epigenetics); genome evolution (comparative genomics, ultraconservation, co-option). Additional topics may include population genetics, personalized genomics, and ancient DNA. Course includes primers on molecular biology, the UCSC Genome Browser, and text processing languages. Guest lectures from genomic researchers.

Prerequisites — **There are no biological or computational prerequisites for CS273a.** While there are no official prerequisites, a background in programming is strongly encouraged. Optional introductory sections on molecular biology, text processing in UNIX, and the UCSC Genome Browser will be given early in the quarter (see schedule below).

Cross-listings — This course is cross listed within Developmental Biology and Biomedical Informatics as DBIO273A and BIOMEDIN273A, respectively.

piazza Ask. Answer. Explore. Whenever.

Professors & TAs: **Save Time. Teach**

With Piazza, easily answer questions, manage course materials, and track student participation. It'll save you time, and your students will love using it. It's free, and easy to get started.

See why Piazza works

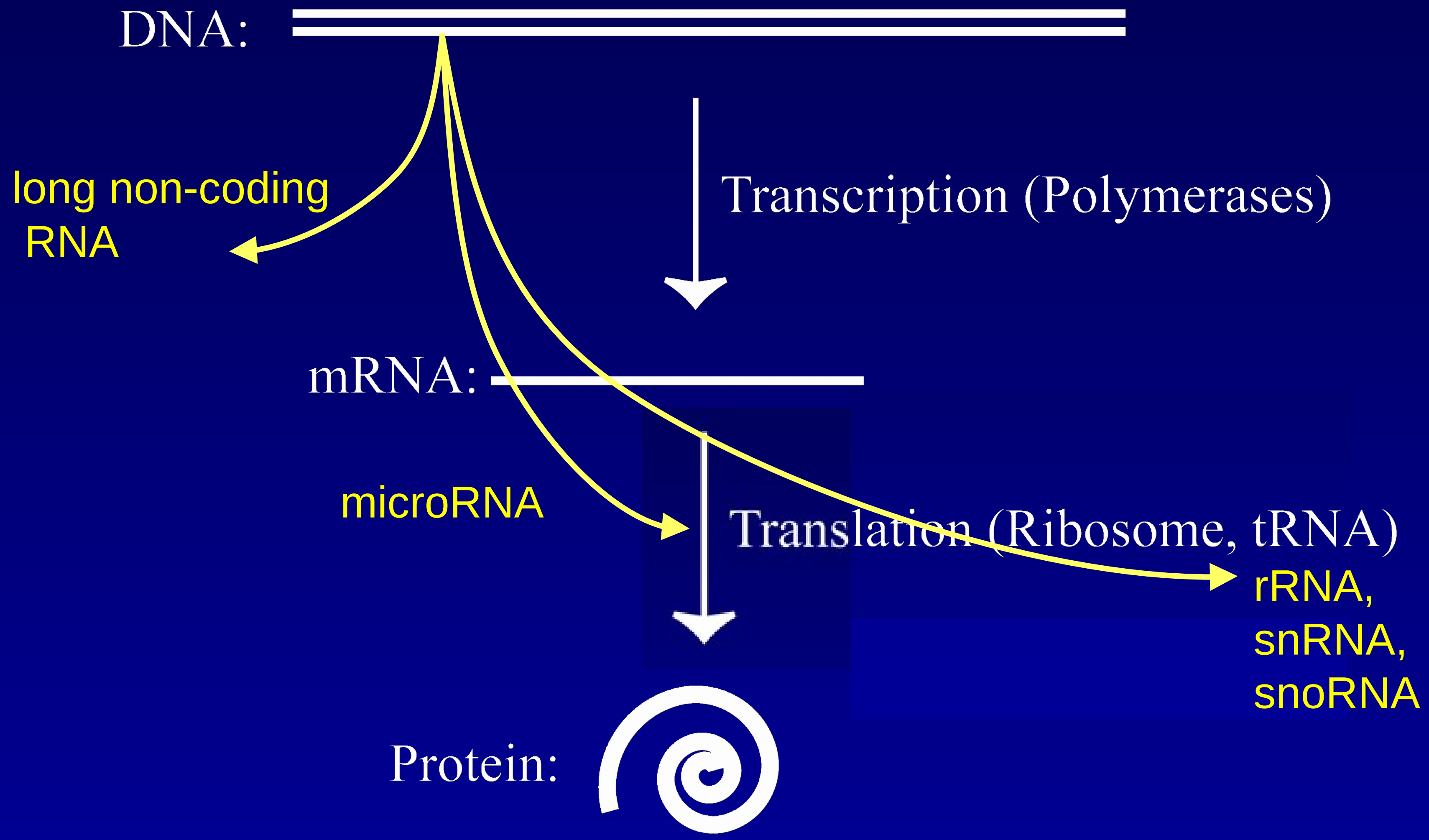
- PS1 is out.

- Last Tutorial this Fri. UCSC Browser. Bring your laptop!

Genome Content

ATATTGAATTTTCAAAAATTCTTACTTTTTTTTTGGATGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATATA
ATCCATATCTAATCTTACTTATATGTTGTGGAAATGTAAAGAGCCCCATTATCTTAGCCTAAAAAAACCTTCTCTTTGGAACTTTC
AATACGCTTAACTGCTCATTGCTATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGACTCTCCTC
GCGTCCTCGTCTTCACCGGTCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTACAATACT
TTTTATGGTTATGAAGAGGAAAAATTGGCAGTAACCTGGCCCCACAAACCTTCAAATTAACGAATCAAATTAACAACCATAGGATG
ATGCGATTAGTTTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCGATGATTTTTGATCTATTAACAGATATATAAATGGAA
CTGCATAACCACTTTAACTAATACTTTCAACATTTTCAGTTTGTATTACTTCTTATTCAAATGTCATAAAAGTATCAACAAAAAAT
TAATATACCTCTATACTTTAACGTCAAGGAGAAAAAACTATAATGACTAAATCTCATTCAGAAGAAGTGATTGTACCTGAGTTCAA
TAGCGCAAAGGAATTACCAAGACCATTGGCCGAAAAGTGCCCGAGCATAATTAAGAAATTTATAAGCGCTTATGATGCTAAACCGG
TTGTTGCTAGATCGCCTGGTAGAGTCAATCTAATTGGTGAACATATTGATTATTGTGACTTCTCGGTTTTACCTTTAGCTATTGAT
GATATGCTTTGCGCCGTCAAAGTTTTGAACGATGAGATTTCAAGTCTTAAAGCTATATCAGAGGGCTAAGCATGTGTATTCTGAAT
TAAGAGTCTTGAAGGCTGTGAAATTAATGACTACAGCGAGCTTTACTGCCGACGAAGACTTTTTCAAGCAATTTGGTGCCTTGATG
GAGTCTCAAGCTTCTTGCGATAAACTTTACGAATGTTCTTGTCCAGAGATTGACAAAATTTGTTCCATTGCTTTGTCAAATGGATC
TGGTTCCCGTTTGACCGGAGCTGGCTGGGGTGGTTGTACTGTTCACTTGGTTCCAGGGGGCCCAAATGGCAACATAGAAAAGGTAA
AAGCCCTTGCCAATGAGTTCTACAAGGTCAAGTACCCTAAGATCACTGATGCTGAGCTAGAAAATGCTATCATCGTCTCTAAACCA
TTGGGCAGCTGTCTATATGAATTAGTCAAGTATACTTCTTTTTTTTTACTTTGTTCAGAACAACTTCTCATTTTTTTCTACTCATAA
TAGCATCACAAAATACGCAATAATAACGAGTAGTAACACTTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGA
TGTTTTCAATGTAAGAGATTTCGATTATCCACAAACTTTAAAACACAGGGACAAAATTCTTGATATGCTTTCAACCGCTGCGTTTT
TACCTATTCTTGACATGATATGACTACCATTTTGTTATTGTACGTGGGGCAGTTGACGTCTTATCATATGTCAAAGTTGCGAAGTT
GGCAAGTTGCCAACTGACGAGATGCAGTAACACTTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATGT
CAATGTAAGAGATTTCGATTATCCACAAACTTTAAAACACAGGGACAAAATTCTTGATATGCTTTCAACCGCTGCGTTTTGGATAC
TTCTTGACATGATATGACTACCATTTTGTTATTGTACGTGGGGCAGTTGACGTCTTATCATATGTCAAAGTCATTTGCGAAGTTCT
CAAGTTGCCAACTGACGAGATGCAGTTTCCTACGCATAATAAGAATAGGAGGGAATATCAAGCCAGACAATCTATCATTACATTTA
GGCTCTTCAAAAAGATTGAACTCTCGCCAACTTATGGAATCTTCCAATGAGACCTTTGCGCCAAATAATGTGGATTTGGAAAAAGA
TAAGTCATCTCAGAGTAATATAACTACCGAAGTTTATGAGGCATCGAGCTTTGAAGAAAAAGTAAGCTCAGAAAAACCTCAATACA
CATTCTGGAAGAAAATCTATTATGAATATGTGGTCGTTGACAAATCAATCTTGGGTGTTTCTATTCTGGATTCATTTATGTACAAC
GACTTGAAGCCCGTCGAAAAAGAAAGGCGGGTTTGGTCCTGGTACAATTATTGTTACTTCTGGCTTGCTGAATGTTTCAATATCAA
TTGGCAAATTGCAGCTACAGGTCTACAACTGGGTCTAAATTGGTGGCAGTGTTGGATAACAATTTGGATTGGGTACGGTTTCGTTG
CTTTTGTTGTTTTGGCCTCTAGAGTTGGATCTGCTTATCATTTGTCATTCCCTATATCATCTAGAGCATCATTCGGTATTTTCTTC
TTATGGCCCGTTATTAACAGAGTCGTCATGGCCATCGTTTGGTATAGTGTCCAAGCTTATATTGCGGCAACTCCCGTATCATTAAT
GAAATCTATCTTTGGAAAAGATTTACAATGATTGTACGTGGGGCAGTTGACGTCTTATCATATGTCAAAGTCATTTGCGAAGTTCT
CAAGTTGCCAACTGACGAGATGCAGTAACACTTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTT
ATGTAAGAGATTTCGATTATCCACAAACTTTAAAACACAGGGACAAAATTCTTGATATGCTTTCAACCGCTGCGTTTTGGATACCT
CTTGACATGATATGACTACCATTTTGTTATTGTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTT
ATGTAAGAGATTTCGATTATCCTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTTTCAATGTAAGA
TTCGATTATCCTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTA
TTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTATCCTTATAGTT
ACAT~~GCTTCAACTACTTAATAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTATCCTTATAGTTCATACA~~TGCTT
CTACTTAATAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTATCCTTATAGTTCATACATGCTTCAACTACTTAA
ATGATTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTATCCTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGT

Genes = coding + “non-coding”



DNA:

Transcription (Polymerases)

long non-coding RNA

mRNA:

microRNA

Translation (Ribosome, tRNA)

rRNA,
snRNA,
snoRNA

Protein:

Genes

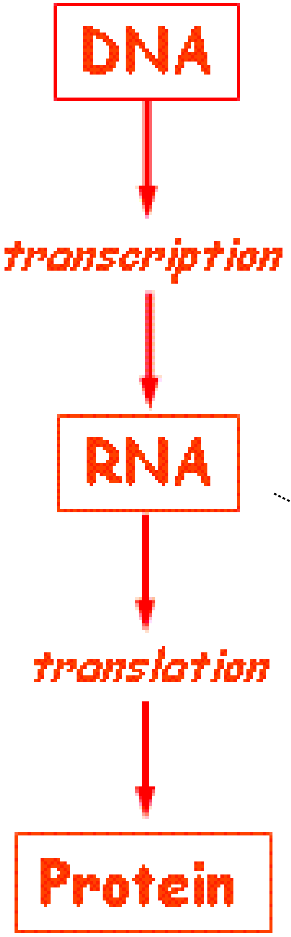
- Gene production is conceptually simple
  - Contiguous stretches of DNA transcribe (1 to 1) into RNA
  - Some (coding or non-coding) RNAs are further spliced
  - Some (m)RNAs are then translated into protein (4^3 to 20+1)
  - Other (nc)RNA stretches just go off to do their thing as RNA

- The devil is in the details, but by and large – this is it.

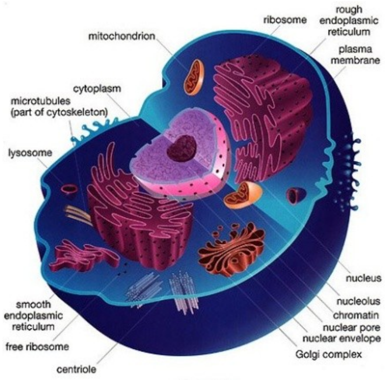
(non/coding) Gene finding - classical computational challenge:

1. Obtain experimental data
2. Find features in the data (eg, genetic code, splice sites)
3. Generalize from features (eg, predict genes yet unseen)
4. Link to biochemical machinery (eg, spliceosome)

Coding and non-coding gene production



To change its behavior a cell can change the repertoire of genes and ncRNAs it makes.



The cell is constantly making new proteins and ncRNAs.

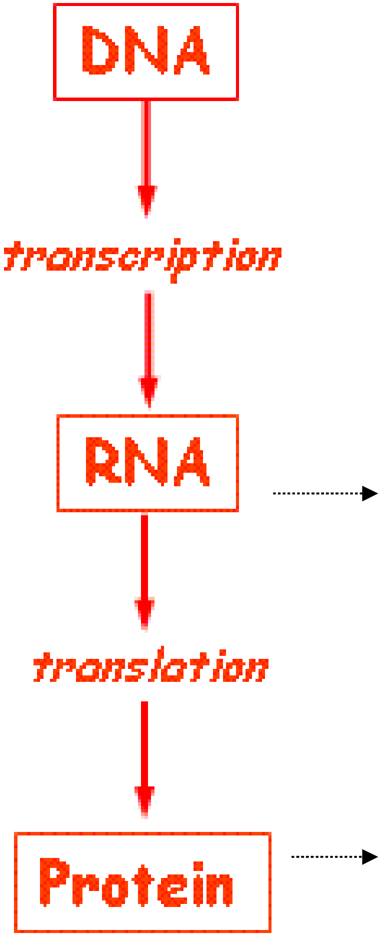
These perform their function for a while,

And are then degraded.

Newly made coding and non coding gene products take their place.

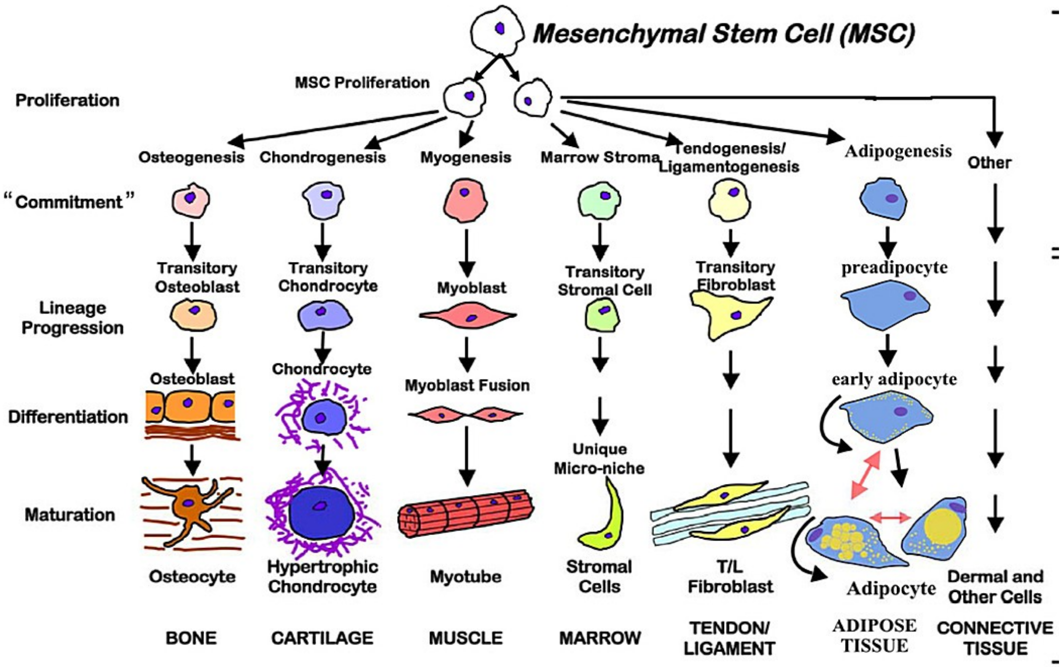
The picture within a cell is constantly “refreshing”.

Cell differentiation



To change its behavior a cell can change the repertoire of genes and ncRNAs it makes.

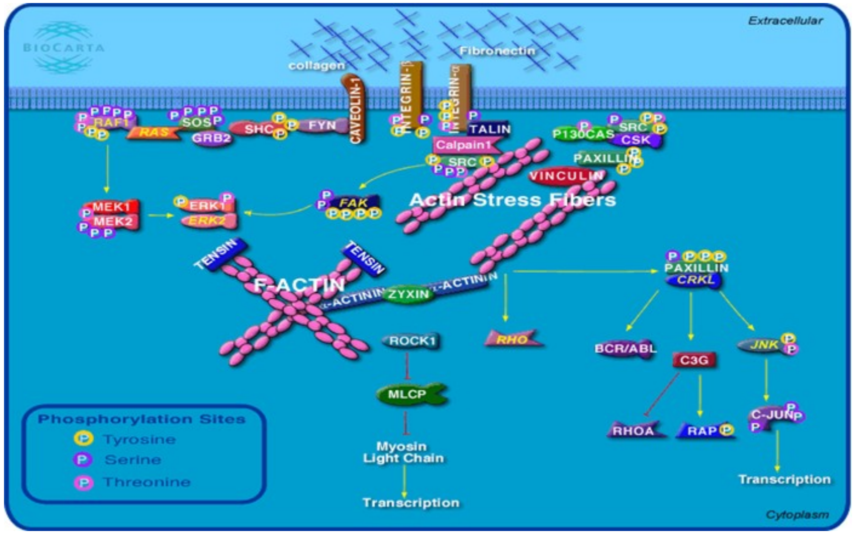
That is exactly what happens when cells differentiate during development from stem cells to their different final fates.

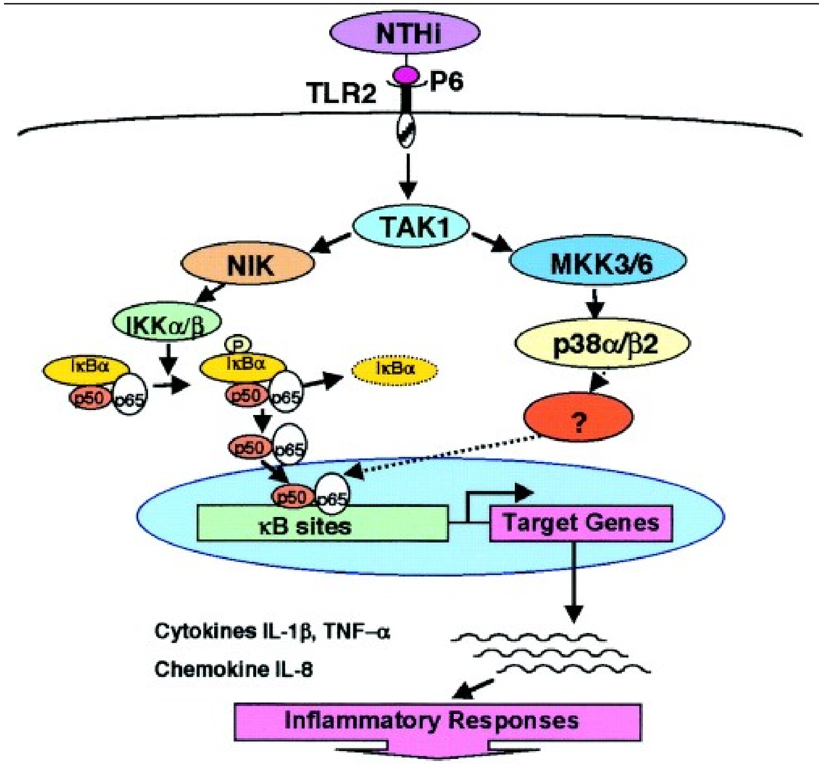


Genes usually work in groups

Biochemical pathways, signaling pathways, etc.

Asking about the expression perturbation of groups of genes is both more appealing biologically, and more powerful statistically (you sum perturbations).





Keyword lists are not enough

Sheer number of terms too much to remember and sort

- Need standardized, stable, carefully defined terms
- Need to describe different levels of detail
- So…defined terms need to be related in a hierarchy

Anatomy keywords

Organ system

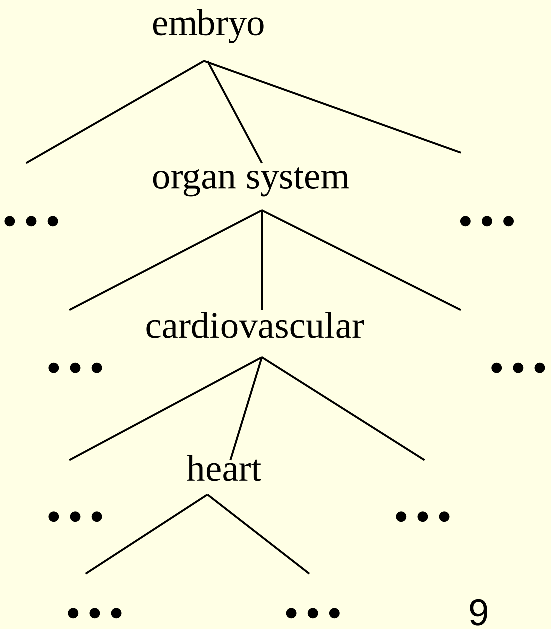
Cardiovascular system

Heart

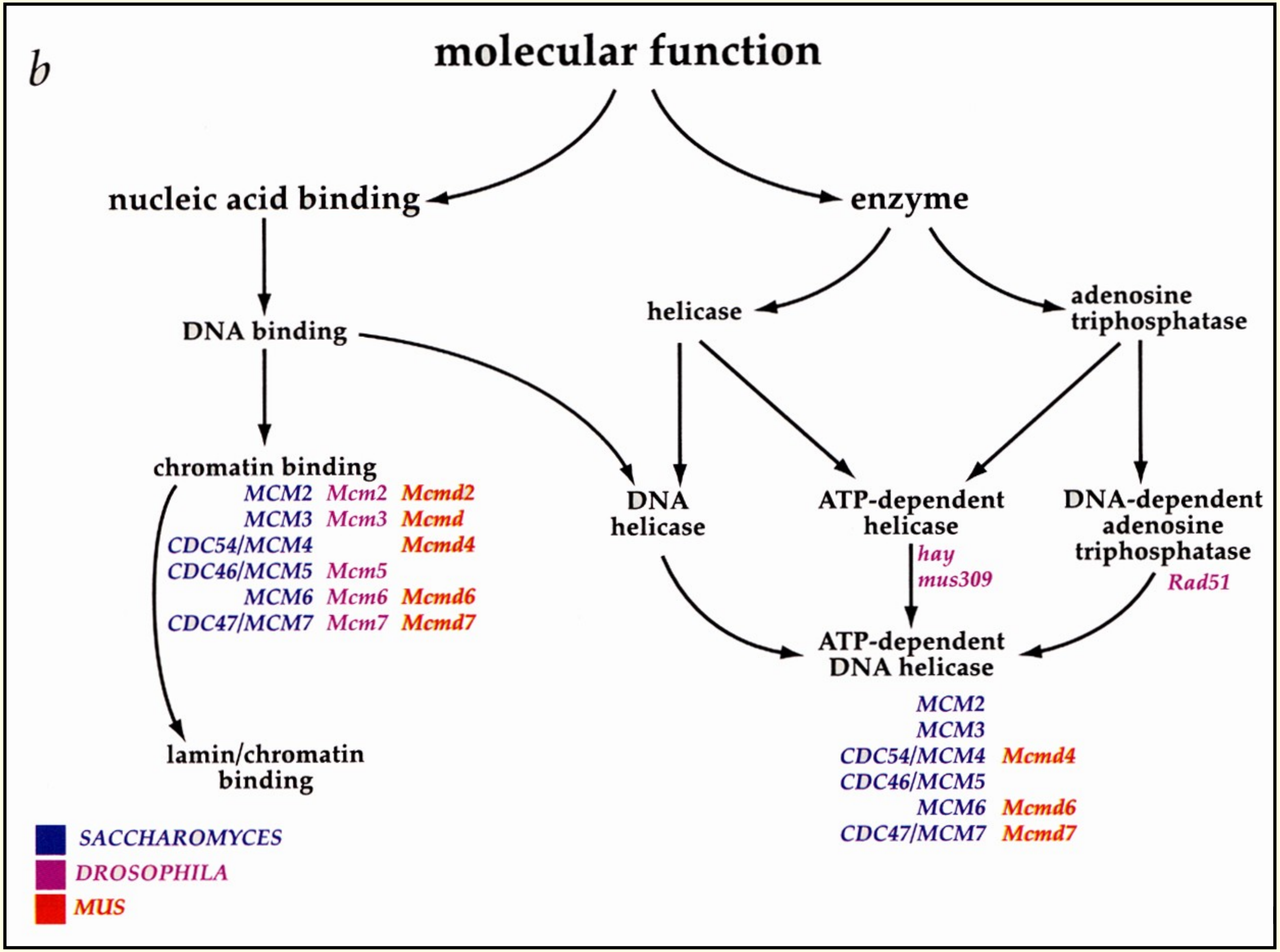
With structured vocabularies/hierarchies

- Parent/child relationships exist between terms
- Increased depth -> Increased resolution
- Can annotate data at appropriate level
- May query at appropriate level

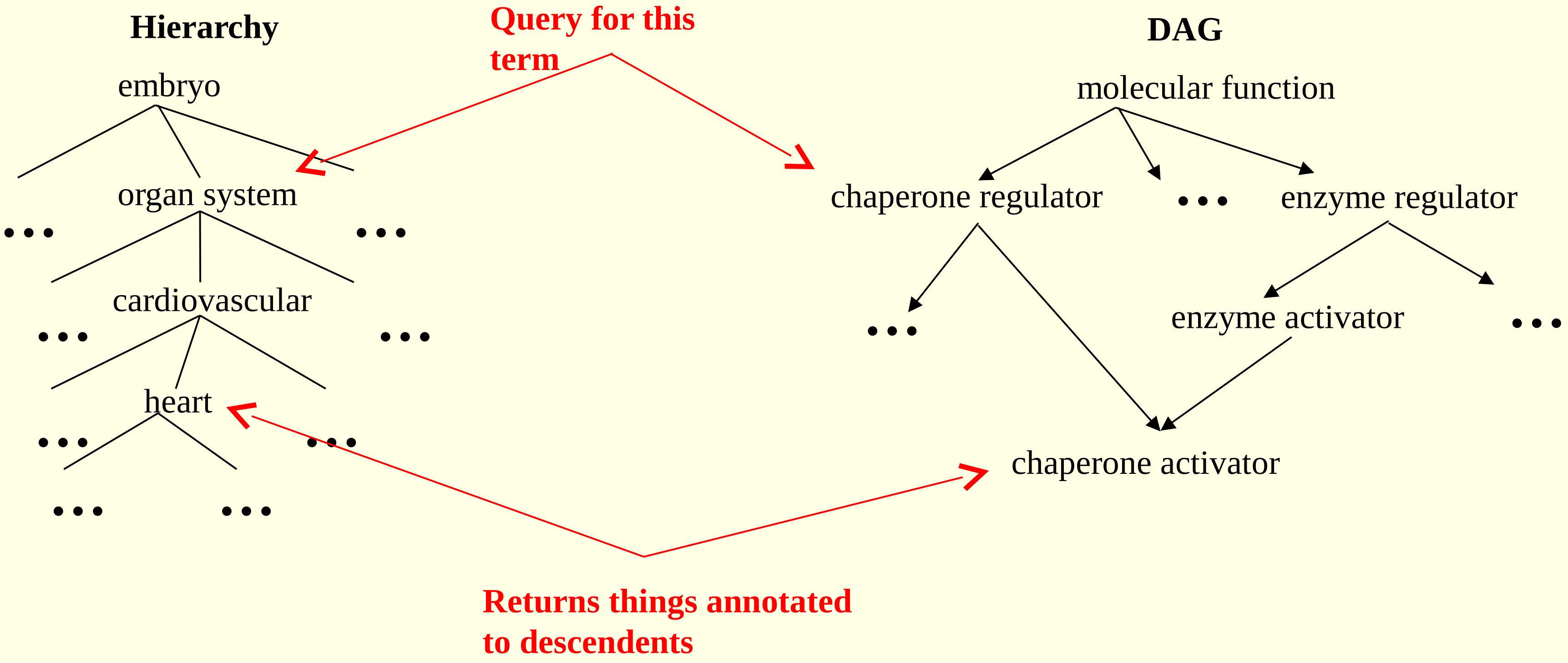
Anatomy Hierarchy



nnotate genes to most specific terms

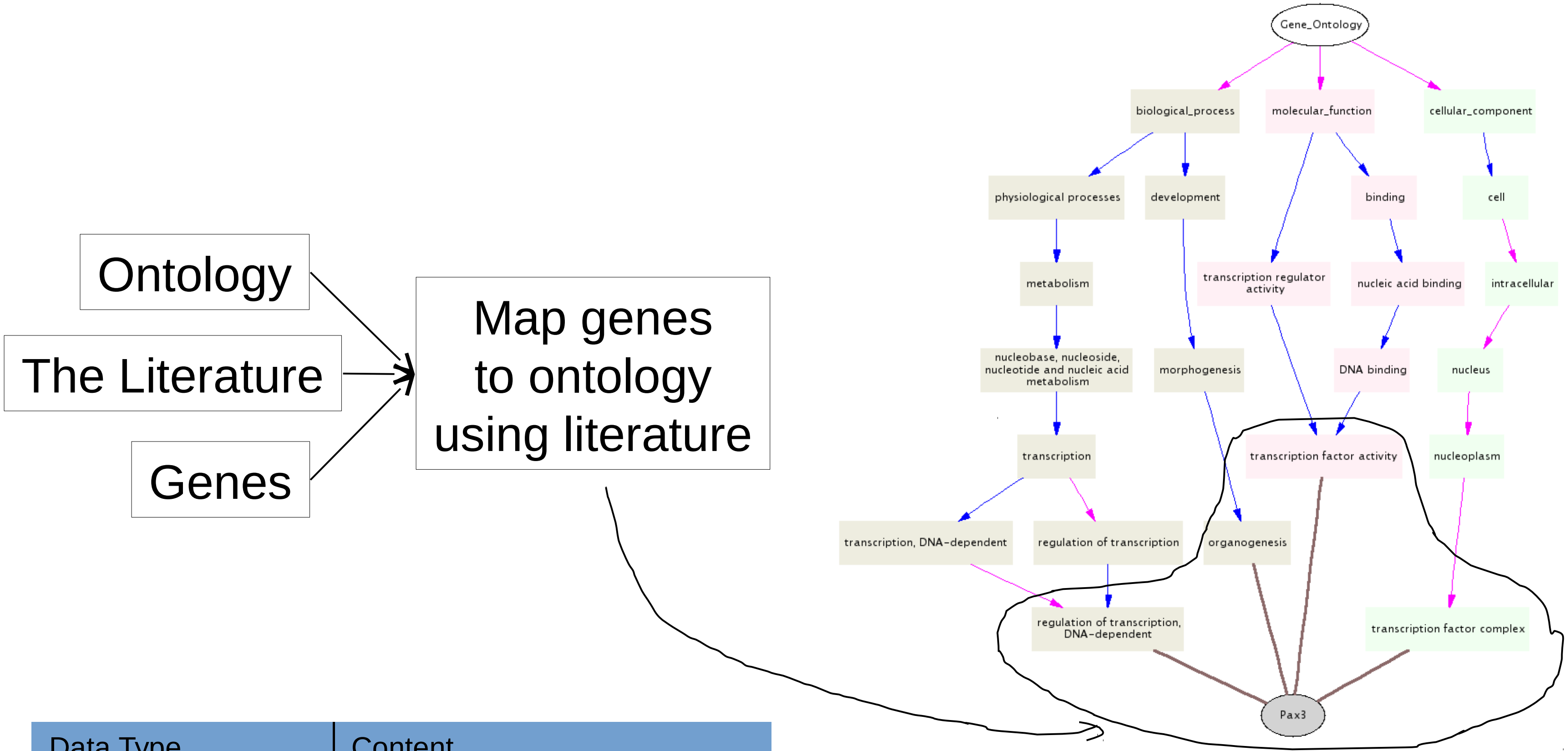


General Implementations for Vocabularies



1. Annotate at appropriate level, query at appropriate level
2. Queries for higher level terms include annotations to lower level terms

Example Research Project (to be revisited)



Ontology

The Literature

Genes

Map genes to ontology using literature

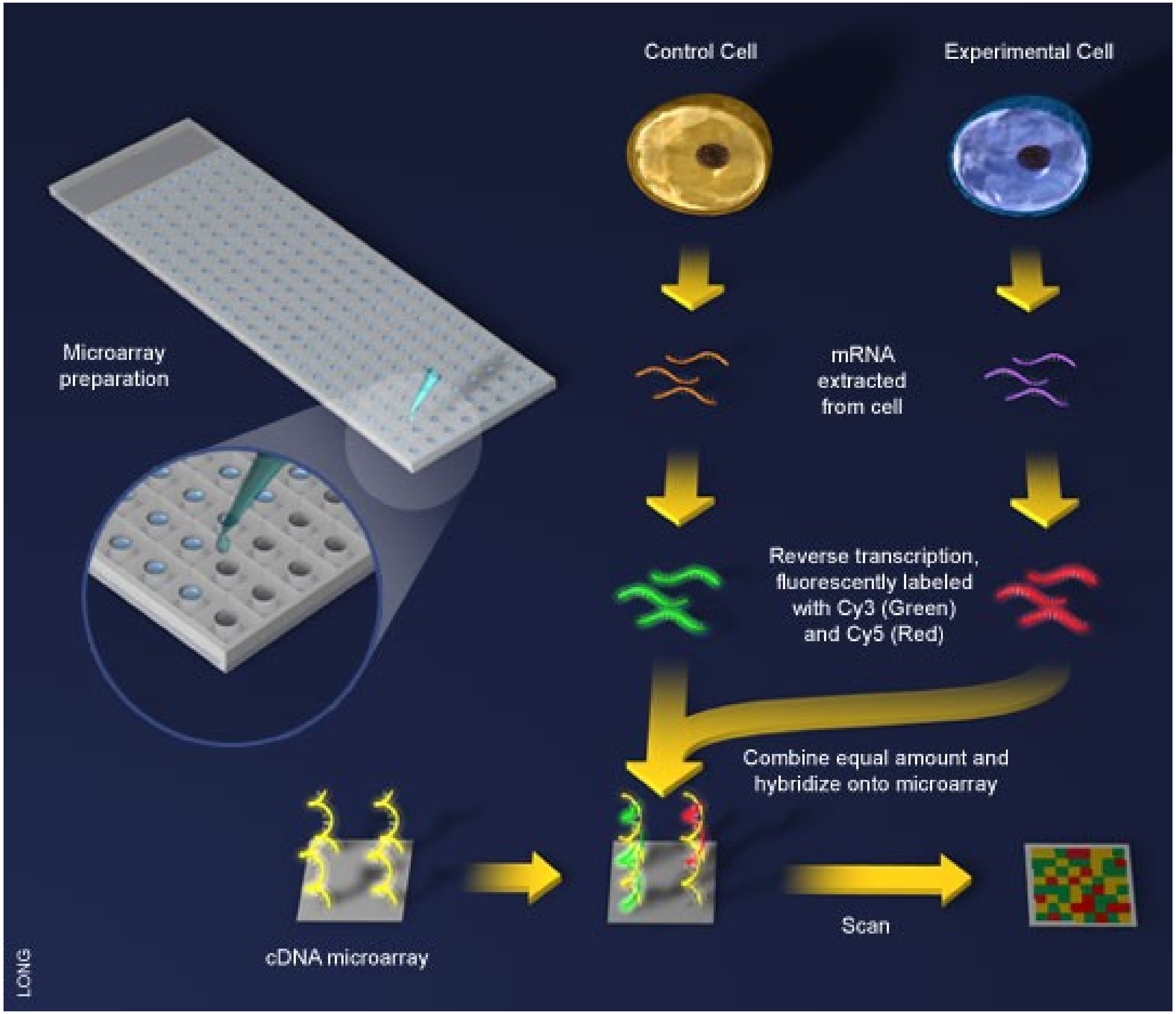
Gene_Ontology
biological_process
molecular_function
cellular_component
physiological processes
development
binding
cell
metabolism
transcription regulator activity
nucleic acid binding
intracellular
nucleobase, nucleoside, nucleotide and nucleic acid metabolism
morphogenesis
DNA binding
nucleus
transcription
transcription factor activity
nucleoplasm
transcription, DNA-dependent
regulation of transcription
organogenesis
regulation of transcription, DNA-dependent
transcription factor complex
Pax3

Data Type	Content
Structured Data	Curated DAGs & mappings
Free Text	**Abstracts, Full Text & Tables**
Diagrams	**Novel models & mappings**
Unstructured Data	**Raw data repos & metadata**

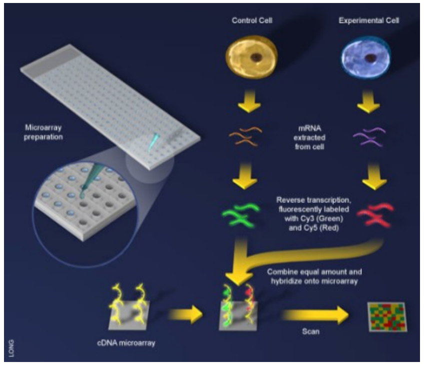
current use

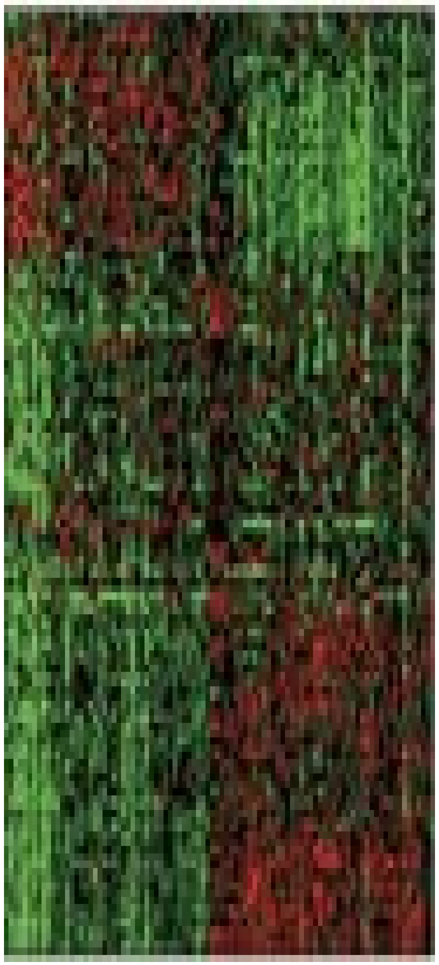
GOAL: attain

Let’s first ask what is changing?


Control Cell
Experimental Cell
Microarray preparation
mRNA extracted from cell
Reverse transcription, fluorescently labeled with Cy3 (Green) and Cy5 (Red)
Combine equal amount and hybridize onto microarray
cDNA microarray
Scan
LONG

Cluster all genes for differential expression


Control Cell
Experimental Cell
Microarray preparation
mRNA extracted from cell
Reverse transcription, fluorescently labeled with Cy3 (Green) and Cy5 (Red)
Combine equal amount and hybridize onto microarray
cDNA microarray
Scan
LONG



Experiment (replicates)

Control (replicates)

genes

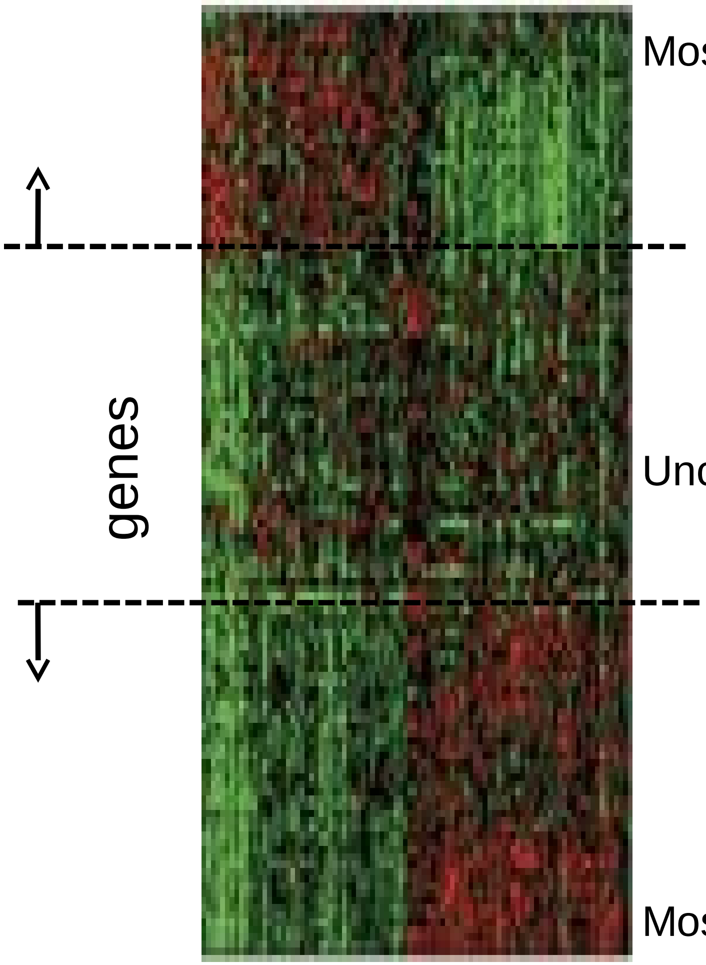
Most significantly up-regulated genes

Unchanged genes

Most significantly down-regulated genes

Determine cut-offs, examine individual genes

Experiment Control
(replicates) (replicates)



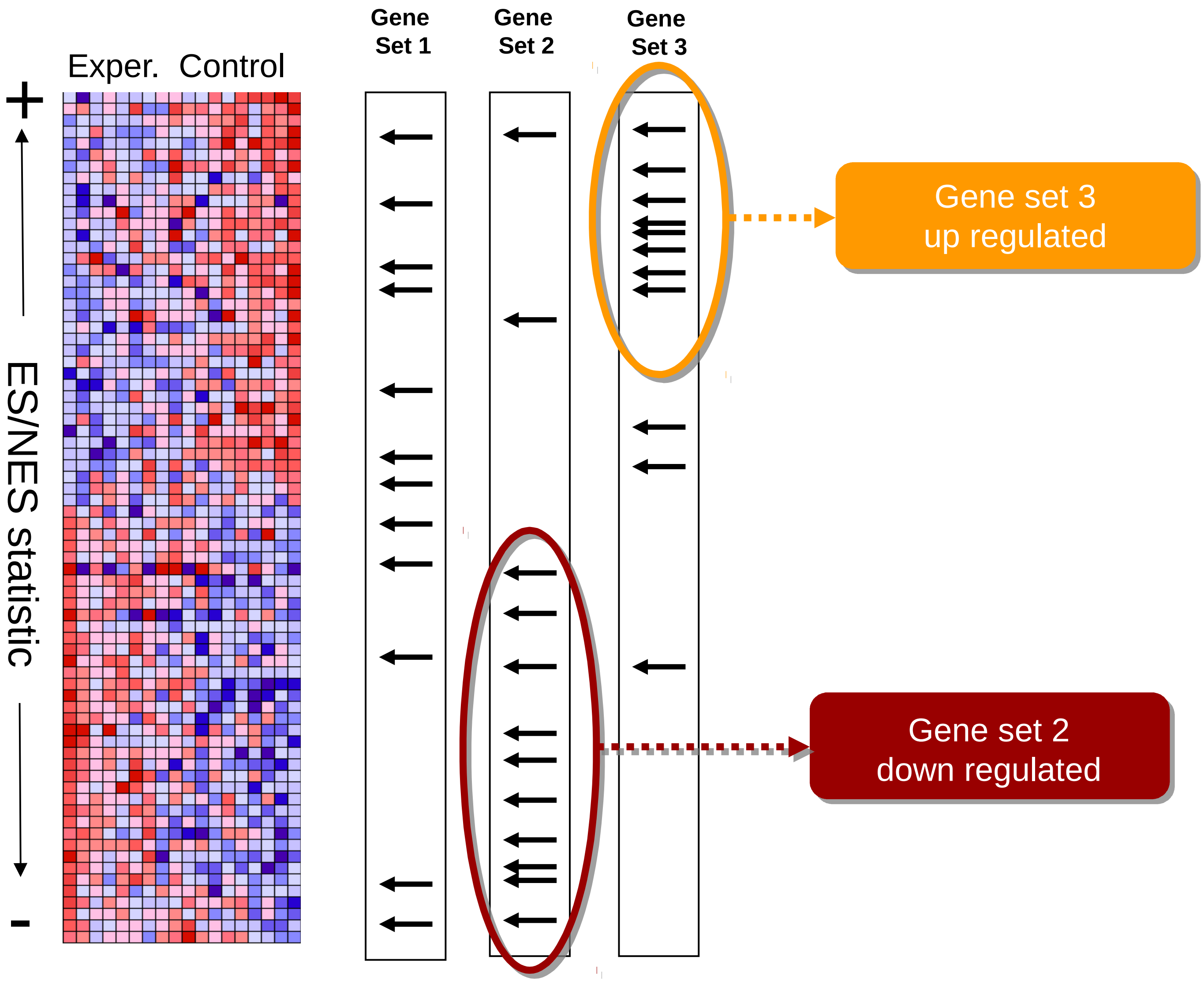
genes

Most significantly up-regulated genes

Unchanged genes

Most significantly down-regulated genes

Ask about whole gene sets



Gene set 3
up regulated

Gene set 2
down regulated

Simplest way to ask: Hypergeometric

Genes measured

N = 20,000

Total genes in set 3

K = 11

I've picked the top

n = 100

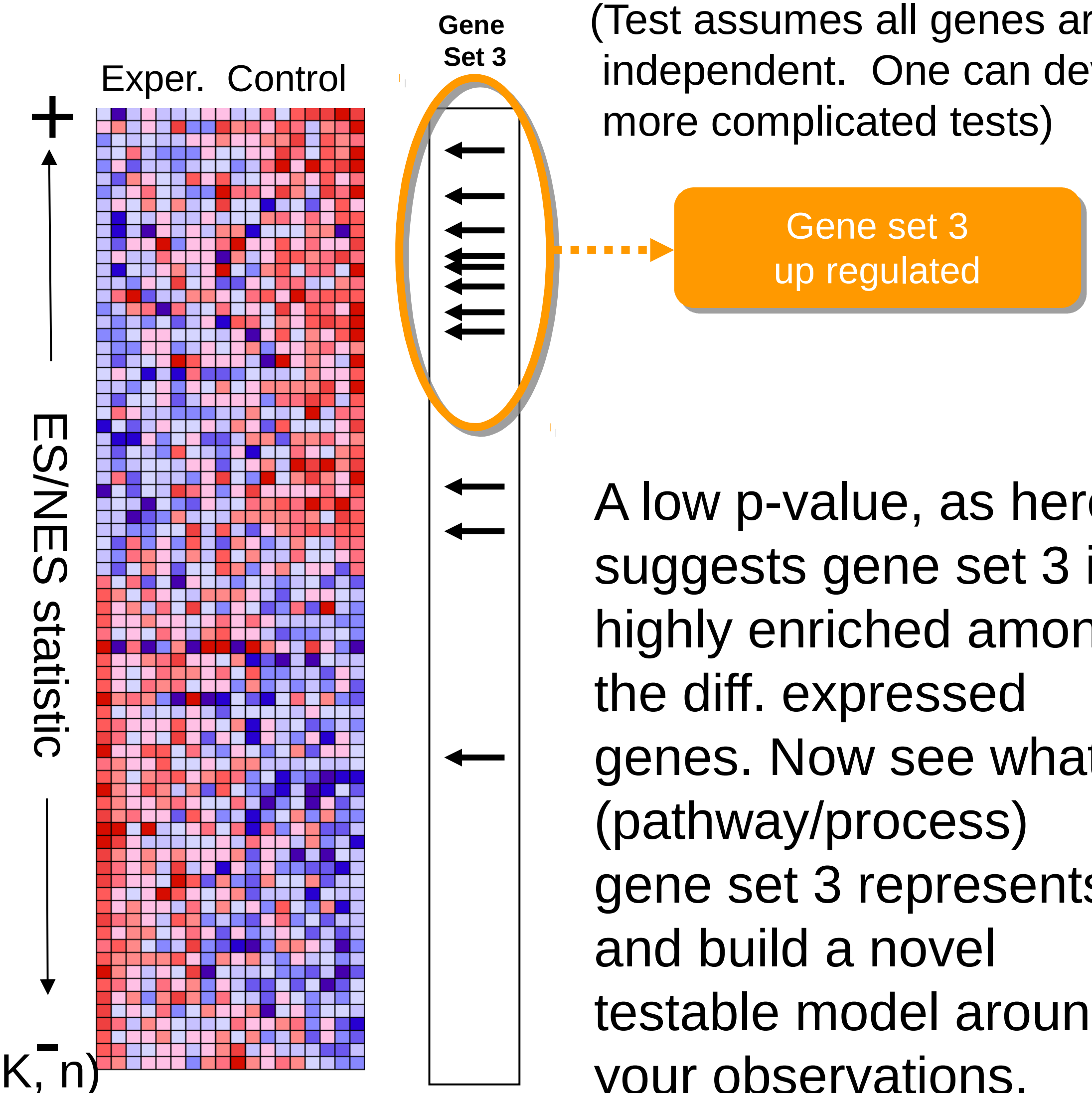
diff. expressed genes.

Of them k = 8

belong to gene set 3.

Under a null of randomly distributed genes, how surprising is it?

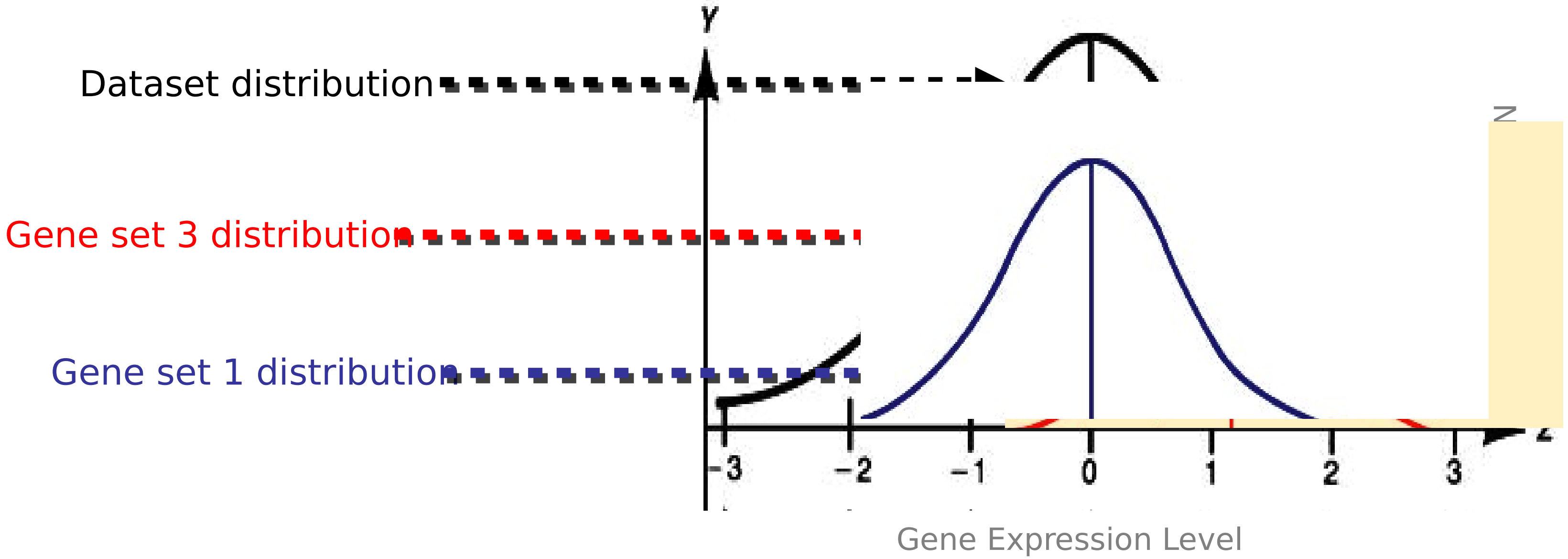
P-value = $\Pr_{\text{hyper}}(k \geq 8 \mid N, K, n)$



(Test assumes all genes are independent. One can devise more complicated tests)

A low p-value, as here, suggests gene set 3 is highly enriched among the diff. expressed genes. Now see what (pathway/process) gene set 3 represents, and build a novel testable model around your observations.

GSEA (Gene Set Enrichment Analysis)

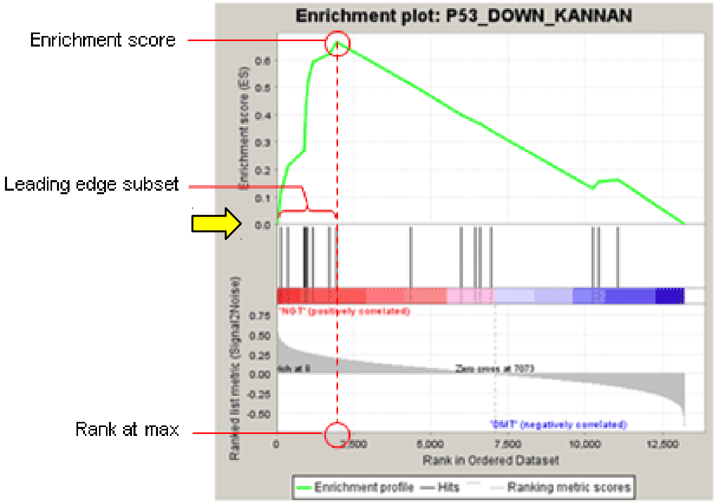


Dataset distribution

Gene set 3 distribution

Gene set 1 distribution

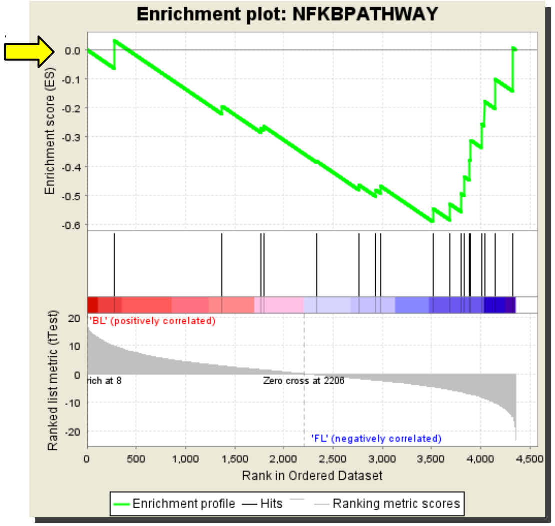
Gene Expression Level

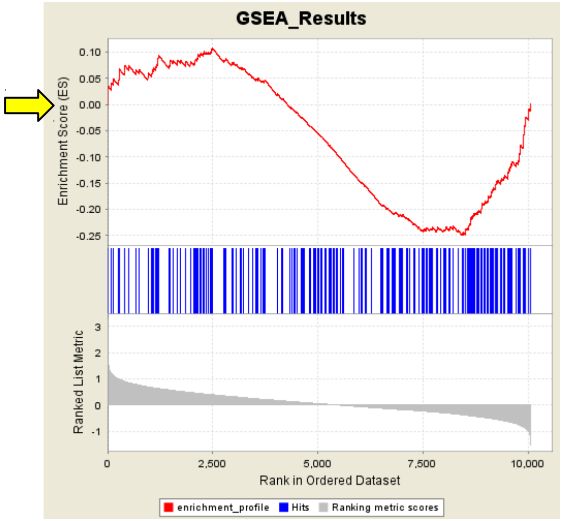


Enrichment score

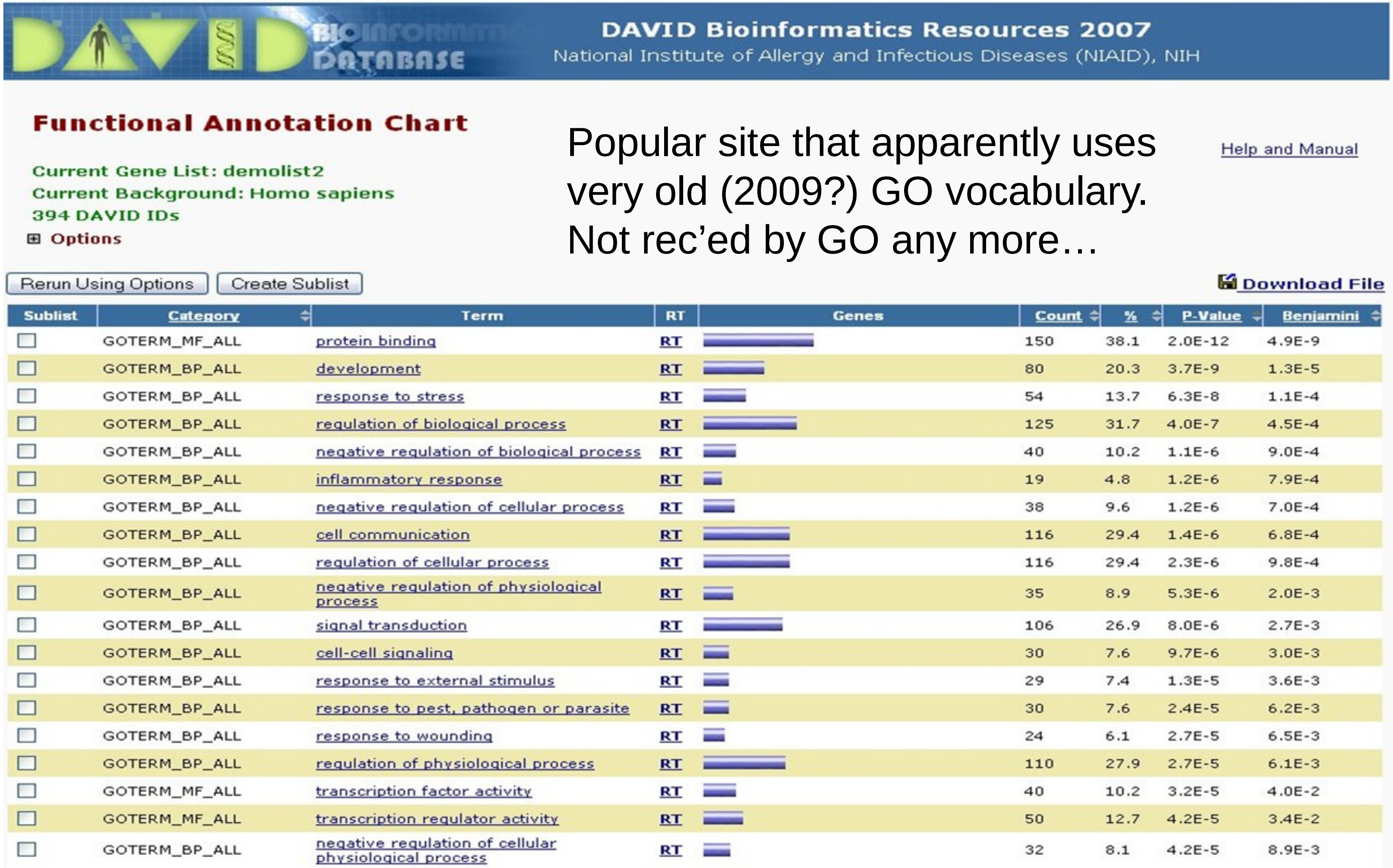
Leading edge subset

Rank at max





Another popular approach: DAVID



DAVID Bioinformatics Resources 2007

National Institute of Allergy and Infectious Diseases (NIAID), NIH

Functional Annotation Chart

Current Gene List: demolist2

Current Background: Homo sapiens

394 DAVID IDs

⊞ Options

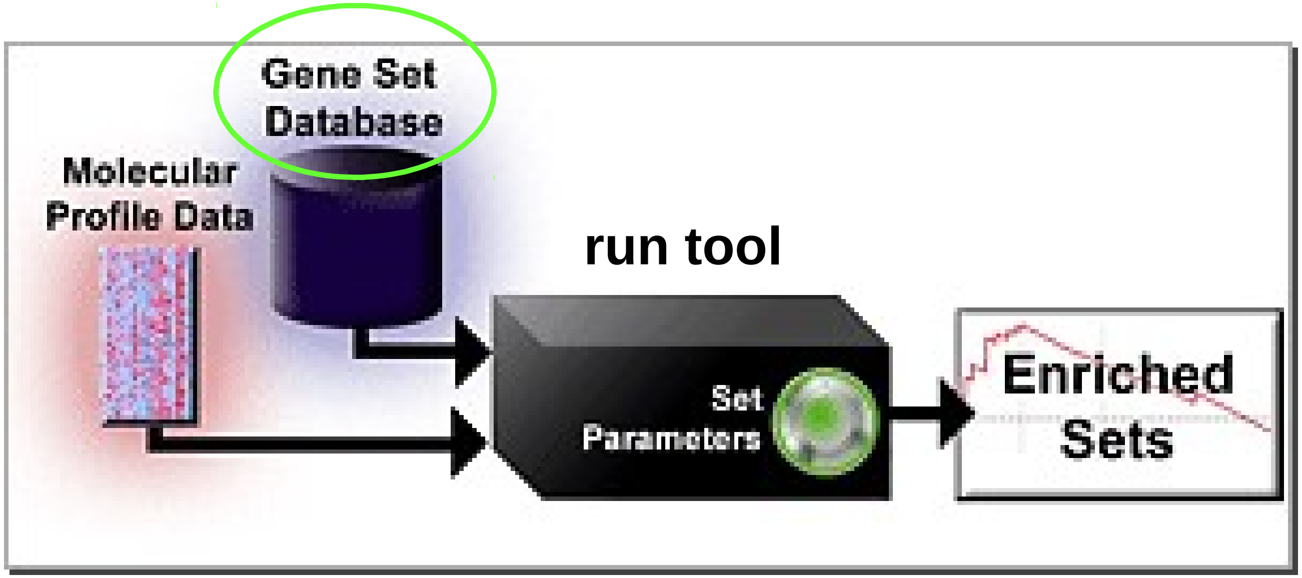
Help and Manual

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	GOTERM_MF_ALL	protein binding	RT	▬	150	38.1	2.0E-12	4.9E-9
□	GOTERM_BP_ALL	development	RT	▬	80	20.3	3.7E-9	1.3E-5
□	GOTERM_BP_ALL	response to stress	RT	▬	54	13.7	6.3E-8	1.1E-4
□	GOTERM_BP_ALL	regulation of biological process	RT	▬	125	31.7	4.0E-7	4.5E-4
□	GOTERM_BP_ALL	negative regulation of biological process	RT	▬	40	10.2	1.1E-6	9.0E-4
□	GOTERM_BP_ALL	inflammatory response	RT	▬	19	4.8	1.2E-6	7.9E-4
□	GOTERM_BP_ALL	negative regulation of cellular process	RT	▬	38	9.6	1.2E-6	7.0E-4
□	GOTERM_BP_ALL	cell communication	RT	▬	116	29.4	1.4E-6	6.8E-4
□	GOTERM_BP_ALL	regulation of cellular process	RT	▬	116	29.4	2.3E-6	9.8E-4
□	GOTERM_BP_ALL	negative regulation of physiological process	RT	▬	35	8.9	5.3E-6	2.0E-3
□	GOTERM_BP_ALL	signal transduction	RT	▬	106	26.9	8.0E-6	2.7E-3
□	GOTERM_BP_ALL	cell-cell signaling	RT	▬	30	7.6	9.7E-6	3.0E-3
□	GOTERM_BP_ALL	response to external stimulus	RT	▬	29	7.4	1.3E-5	3.6E-3
□	GOTERM_BP_ALL	response to pest, pathogen or parasite	RT	▬	30	7.6	2.4E-5	6.2E-3
□	GOTERM_BP_ALL	response to wounding	RT	▬	24	6.1	2.7E-5	6.5E-3
□	GOTERM_BP_ALL	regulation of physiological process	RT	▬	110	27.9	2.7E-5	6.1E-3
□	GOTERM_MF_ALL	transcription factor activity	RT	▬	40	10.2	3.2E-5	4.0E-2
□	GOTERM_MF_ALL	transcription regulator activity	RT	▬	50	12.7	4.2E-5	3.4E-2
□	GOTERM_BP_ALL	negative regulation of cellular physiological process	RT	▬	32	8.1	4.2E-5	8.9E-3

Popular site that apparently uses very old (2009?) GO vocabulary. Not rec'ed by GO any more…

Input: list of genes of interest (without expression values).

Multiple Testing Correction

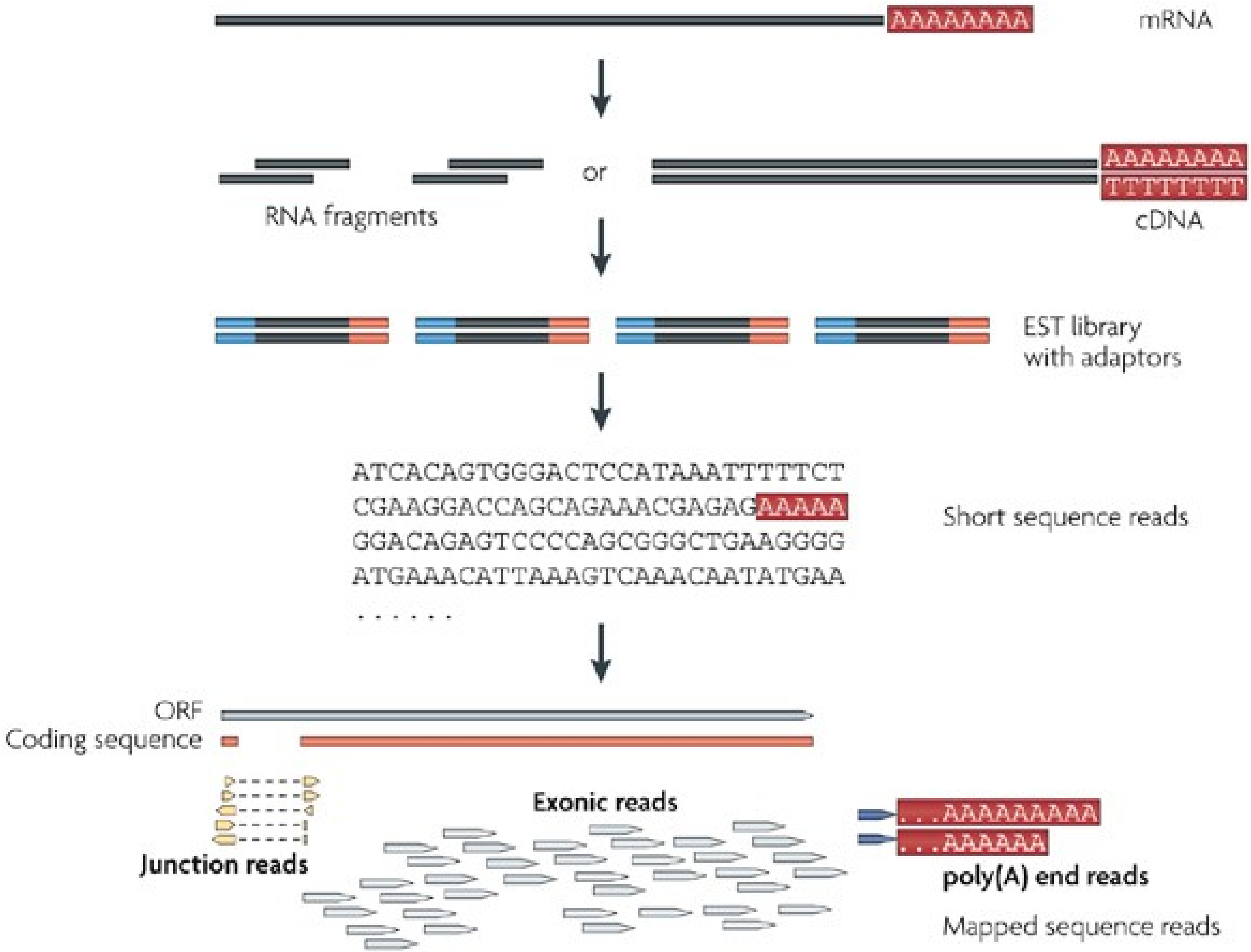


Note that statistically you cannot just run individual tests on 1,000 different gene sets. You have to apply further statistical corrections, to account for the fact that even in 1,000 random experiments a handful may come out good by chance.

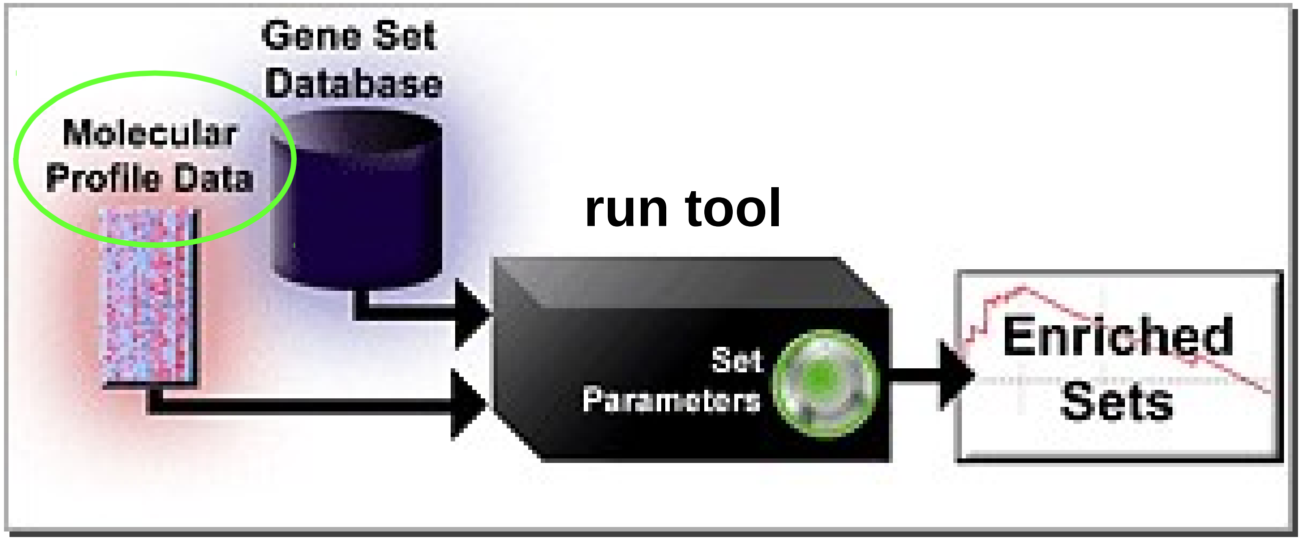
(eg experiment = Throw a coin 10 times. Ask if it is biased.
If you repeat it 1,000 times, you will eventually get an all heads series, from a fair coin. Mustn't deduce that the coin is biased)

RNA-seq

“Next” (2nd) generation sequencing.



What will you test?

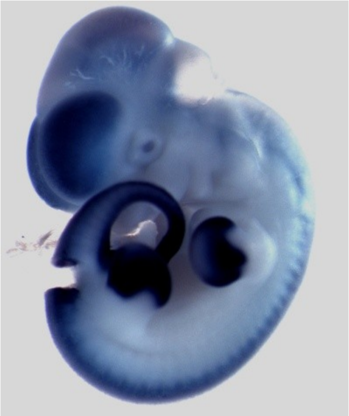


Also note that this is a very general approach to test gene lists.

Instead of a microarray experiment you can do RNA-seq.

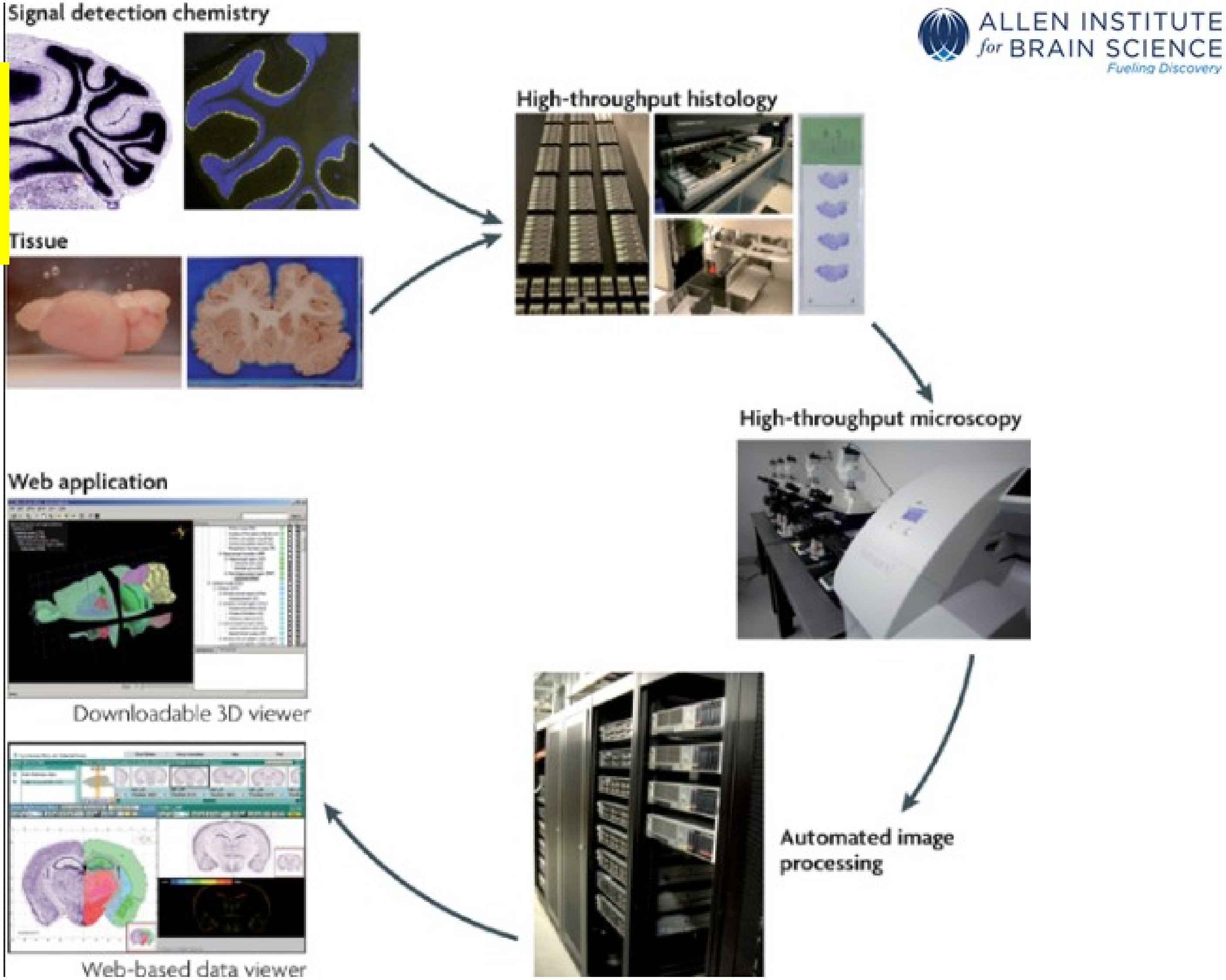
Advantage: RNA-seq measures all genes(up to your ability to correctly reconstruct them). Microarrays only measure the probes you can fit on them. (Some genes, or indeed entire pathways, may be missing from some microarray designs).

Single gene in situ hybridization

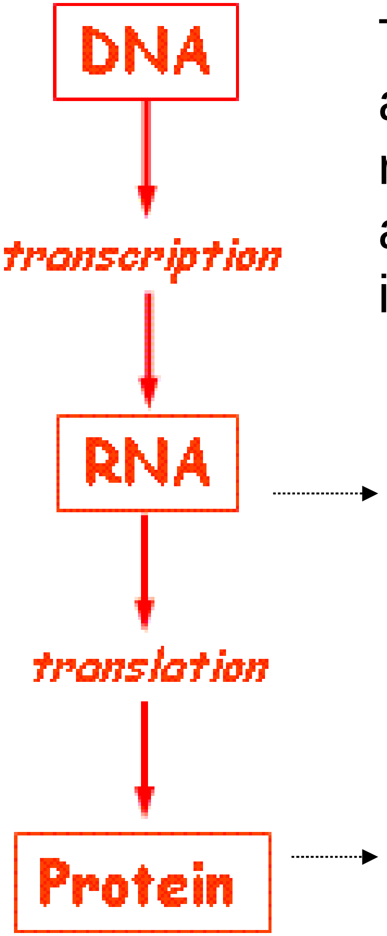


Spatial-temporal maps generation

AI:
Robotics,
Vision

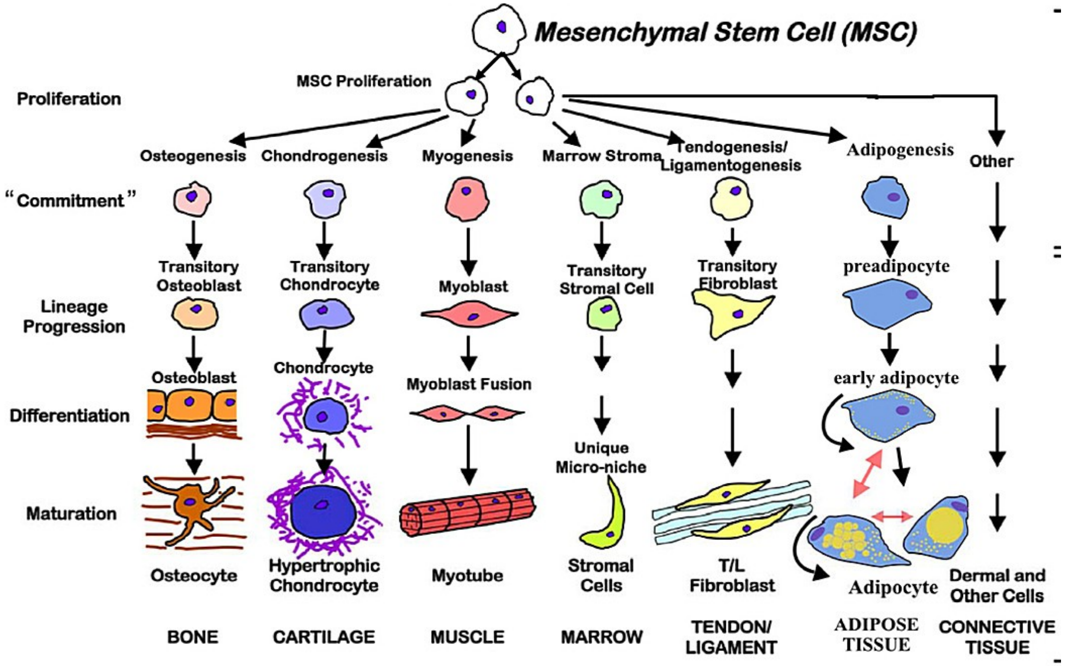

ALLEN INSTITUTE for BRAIN SCIENCE
Fueling Discovery
Signal detection chemistry
Tissue
High-throughput histology
High-throughput microscopy
Automated image processing
Web application
Downloadable 3D viewer
Web-based data viewer

Cell differentiation



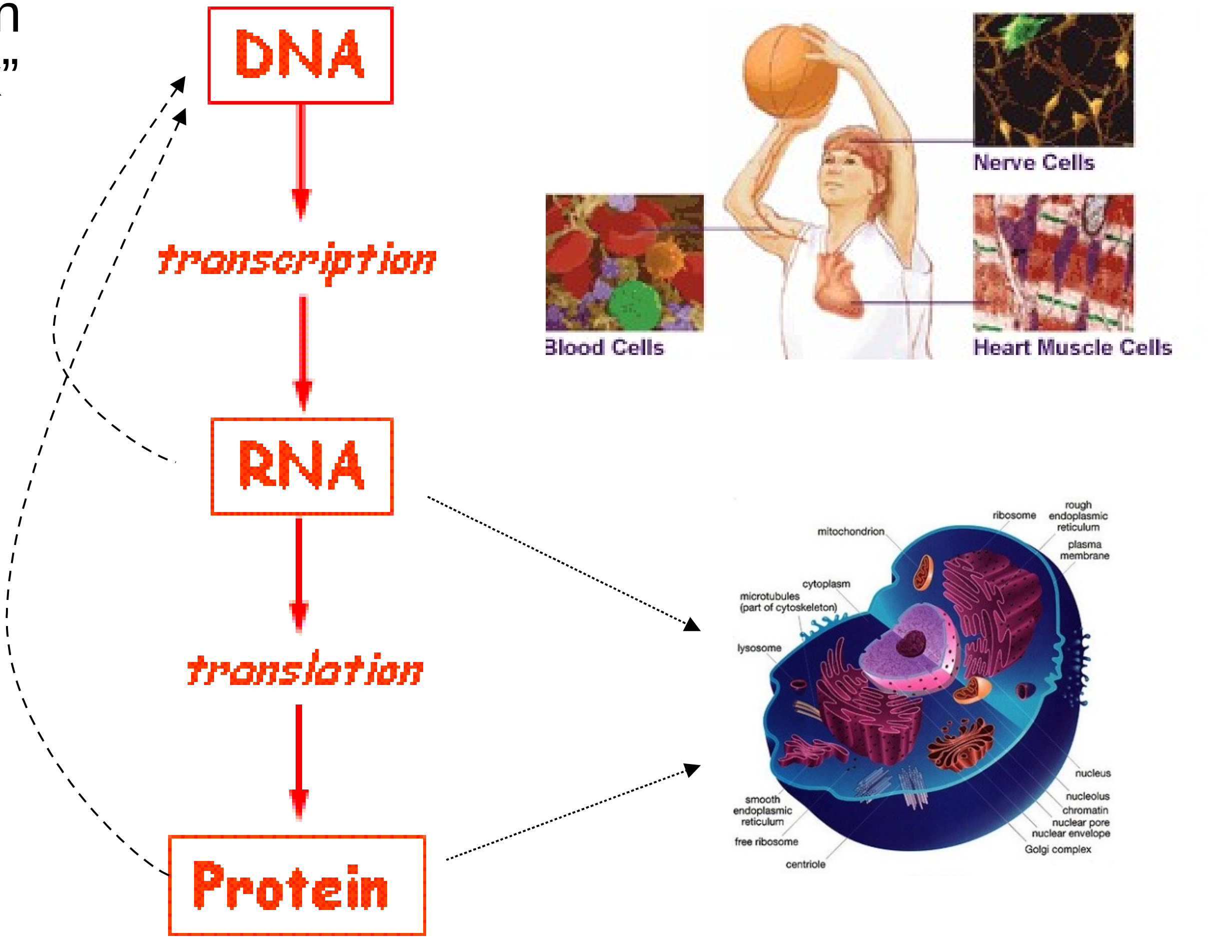
To change its behavior a cell can change the repertoire of coding and non-coding genes it makes.

But how?



Closing the loop

Some proteins and non coding RNAs go “back” to bind DNA near genes, turning these genes on and off.



DNA

transcription

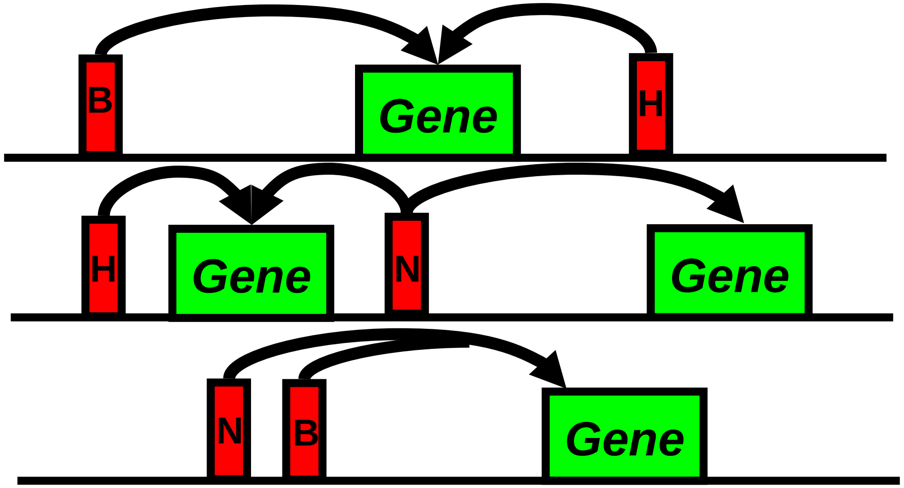
RNA

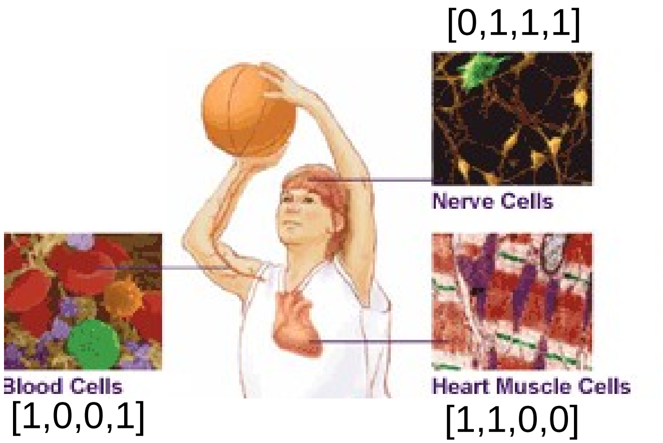
translation

Protein

Genes & Gene Regulation

- Gene = genomic substring that encodes HOW to make a protein (or ncRNA).
- Genomic switch = genomic substring that encodes WHEN, WHERE & HOW MUCH of a protein to make.





Transcription Regulation

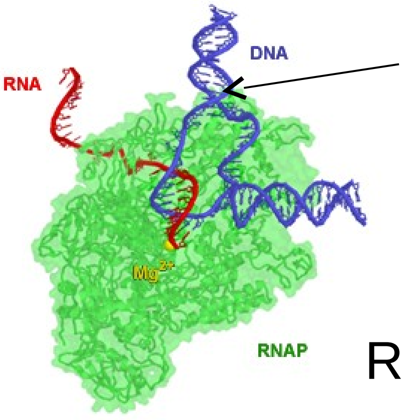
Conceptually simple:

1. The machine that transcribes (“RNA polymerase”)
2. All kinds of proteins and ncRNAs that bind to DNA and to each other to attract or repel the RNA polymerase (“transcription associated factors”).
3. DNA accessibility – making DNA stretches in/accessible to the RNA polymerase and/or transcription associated factors by un/wrapping them around nucleosomes.

(Distinguish DNA patterns from proteins they interact with)

RNA Polymerase

- Transcription = Copying a segment of DNA into (non/coding) RNA
- Gene transcription starts at the (aptly named) TSS, or gene transcription start site
- Transcription is done by RNA polymerase, a complex of 10-12 subunit proteins.
- There are three types of RNA polymerases in human:
  - RNA pol I synthesizes ribosomal RNAs
  - RNA pol II synthesizes pre-mRNAs and most microRNAs
  - RNA pol III synthesizes tRNAs, rRNA and other ssRNAs

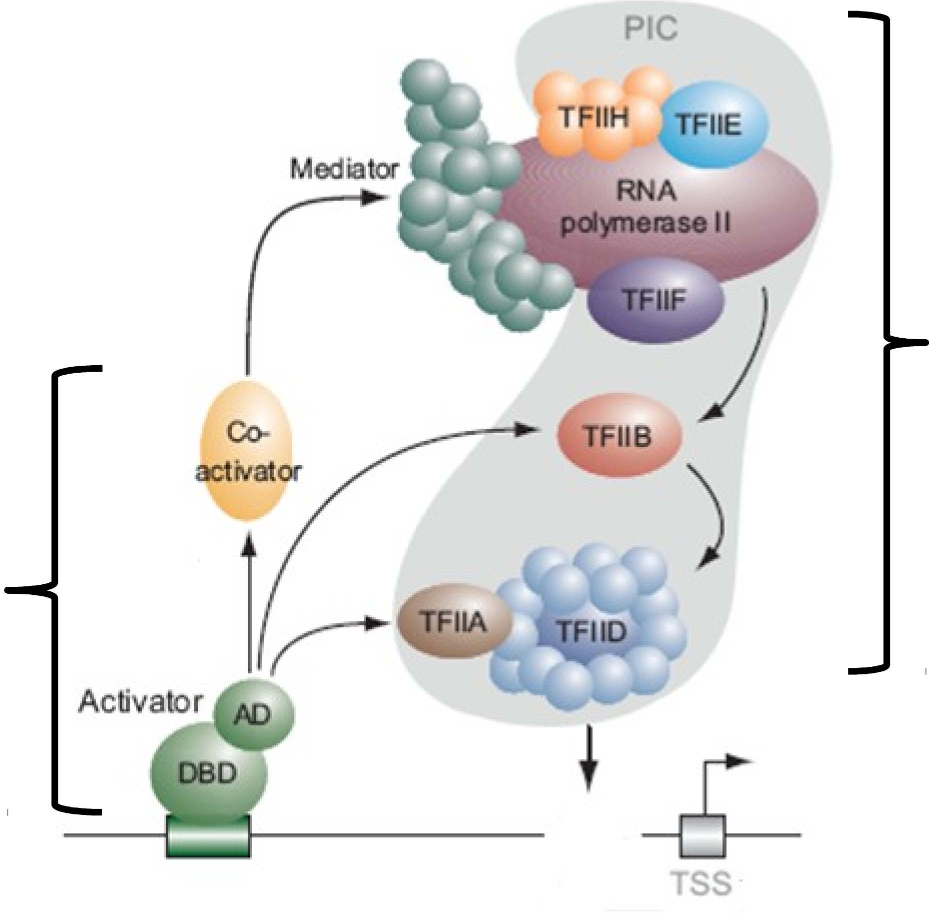


TSS

RNA Polymerase

RNA Polymerase is General Purpose

- RNA Polymerase is the general purpose transcriptional machinery.
- It generally does not recognize gene transcription start sites by itself, and requires interactions with multiple additional proteins.



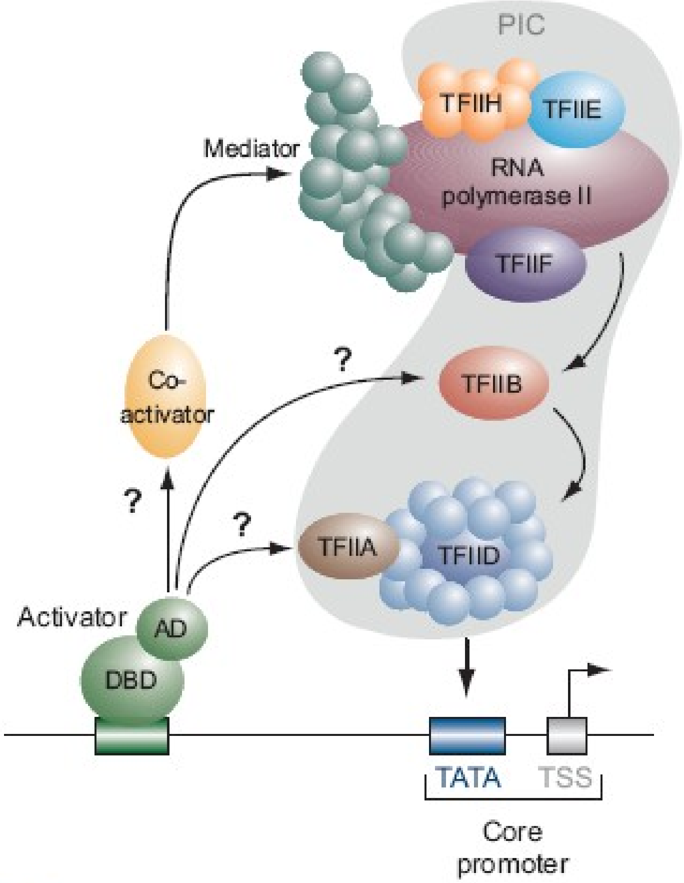
general purpose

context specific

Terminology

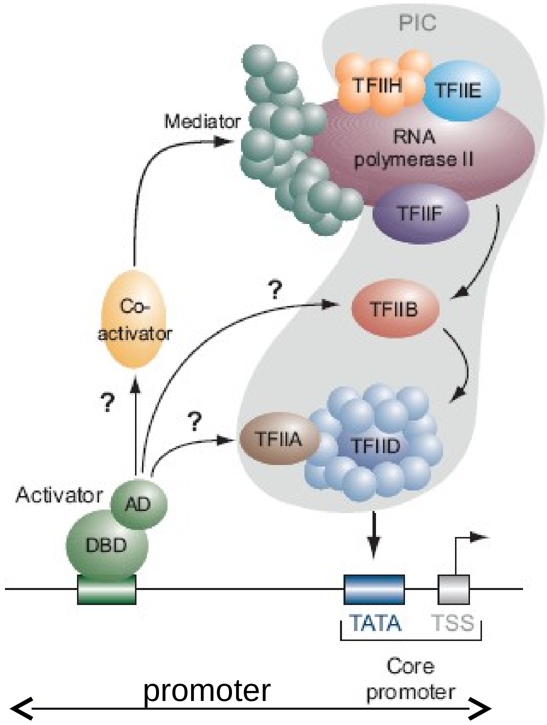
- Transcription Factors (TF): Proteins that return to the nucleus, bind specific DNA sequences there, and affect transcription.
  - There are 1,200-2,000 TFs in the human genome (out of 20-25,000 genes)
  - Only a subset of TFs may be expressed in a given cell at a given point in time.

- Transcription Factor Binding Sites: 4-20bp stretches of DNA where TFs bind.
  - There are millions of TF binding sites in the human genome.
  - In a cell at a given point in time, a site can be either occupied or unoccupied.

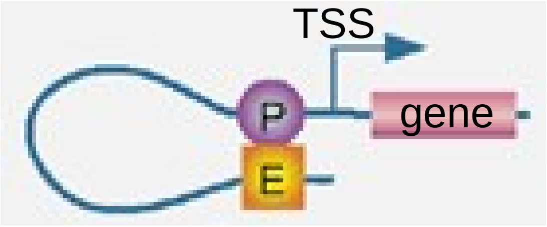


Terminology

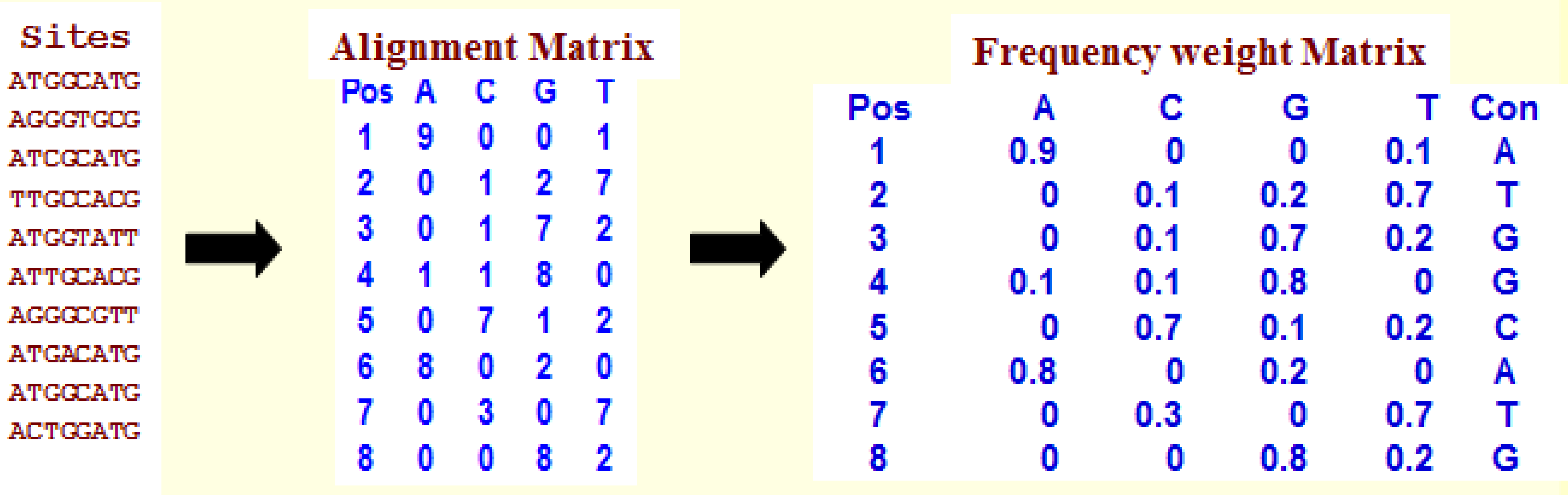
- Promoter: The region of DNA 100-1,000bp immediately "upstream" of the TSS, which encodes binding sites for the general purpose RNA polymerase associated TFs, and at times some context specific sites.
  - There are as many promoters as there are TSS's in the human genome. Many genes have more than one TSS.



- Enhancer: A region of 100-1,000bp, up to 1Mb or more, upstream or downstream from the TSS that includes binding sites for multiple TFs. When bound by (the right) TFs an enhancer turns on/accelerates transcription.
  - Note how an enhancer (E) very far away in sequence (1D) can in fact get very close to the promoter (P) in space (3D).



TFBS Position Weight Matrix (PWM)



Sites

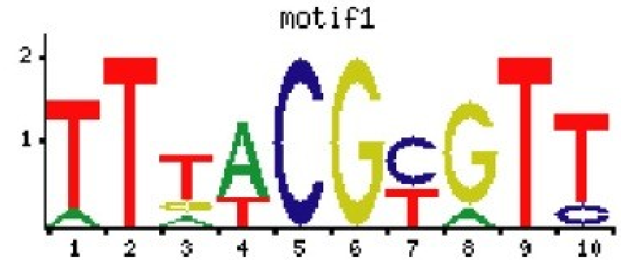
ATGGCATG
AGGGTGCG
ATCGCATG
TTGCCACG
ATGGTATT
ATTGCACG
AGGCCGTT
ATGACATG
ATGGCATG
ACTGGATG

Alignment Matrix

Pos	A	C	G	T
1	9	0	0	1
2	0	1	2	7
3	0	1	7	2
4	1	1	8	0
5	0	7	1	2
6	8	0	2	0
7	0	3	0	7
8	0	0	8	2

Frequency weight Matrix

Pos	A	C	G	T	Con
1	0.9	0	0	0.1	A
2	0	0.1	0.2	0.7	T
3	0	0.1	0.7	0.2	G
4	0.1	0.1	0.8	0	G
5	0	0.7	0.1	0.2	C
6	0.8	0	0.2	0	A
7	0	0.3	0	0.7	T
8	0	0	0.8	0.2	G

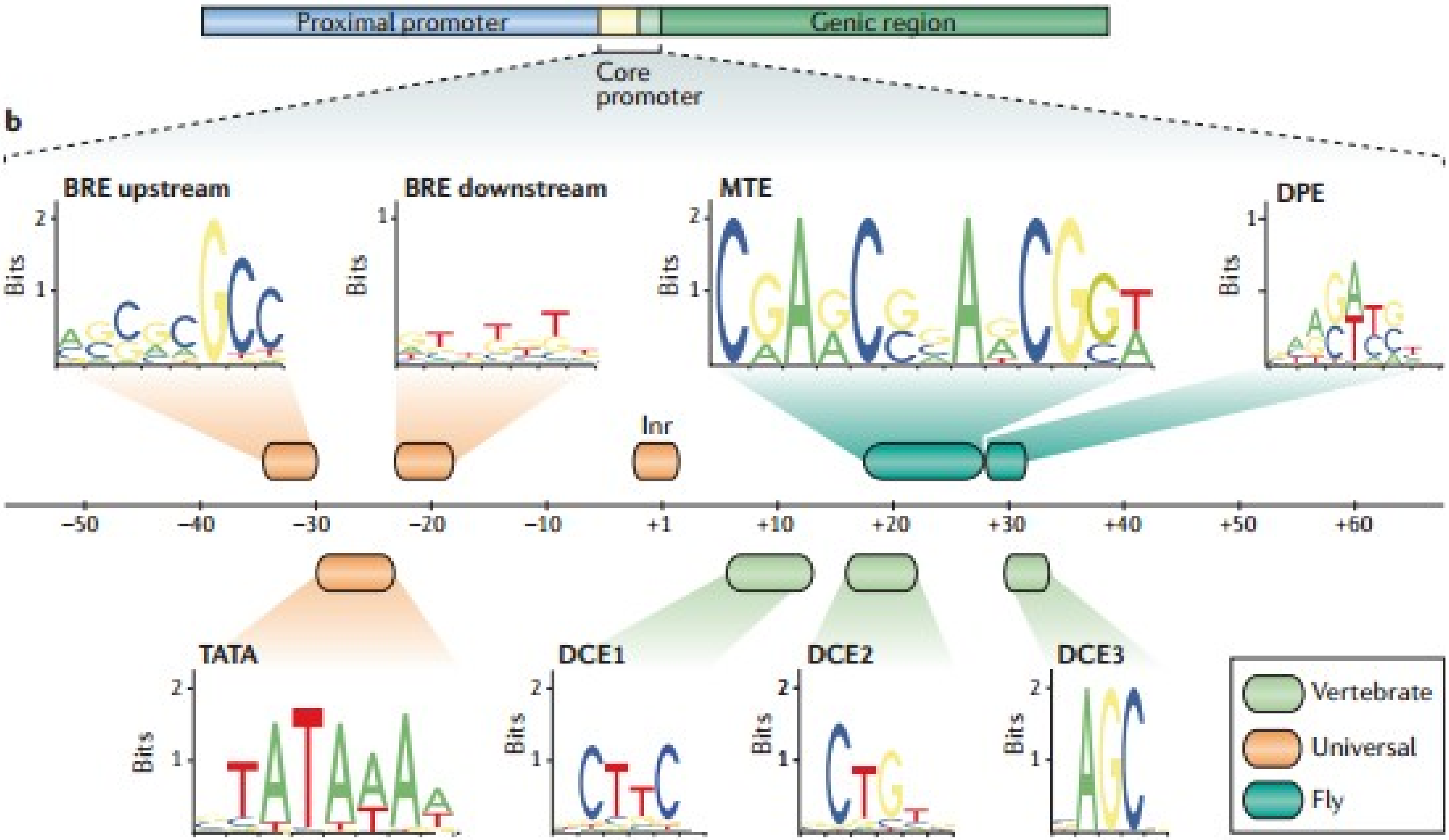


```
A   [ 2   0   2  11   0   0   0   2   0   0 ]
C   [ 0   0   0   0  15   0   9   0   0   3 ]
G   [ 0   0   3   0   0  15   0  13   0   0 ]
T   [13  15  10   4   0   0   6   0  15  12 ]
```

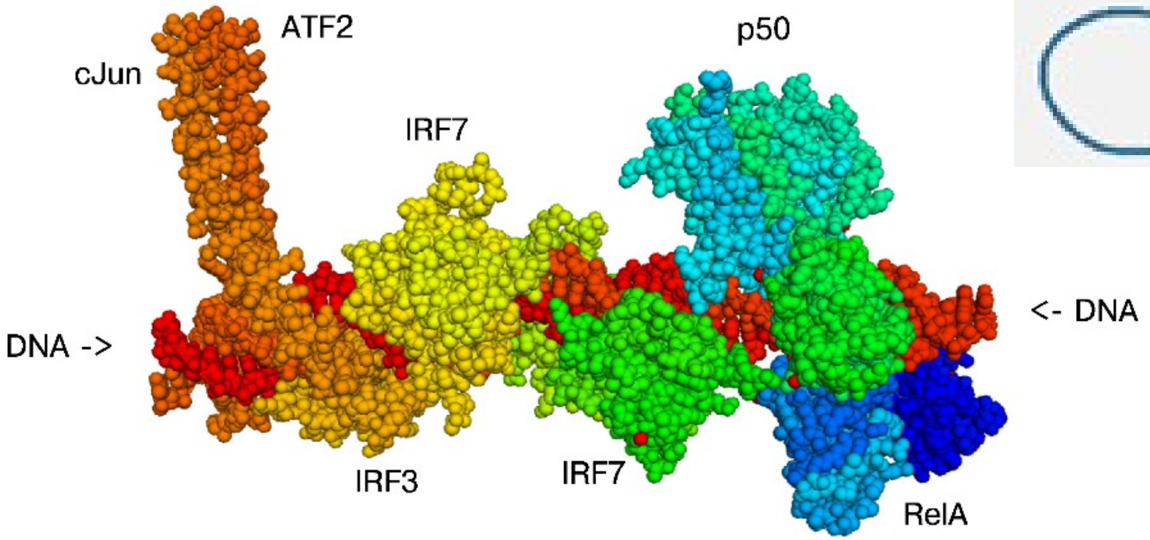
Note the strong independence assumption between positions.

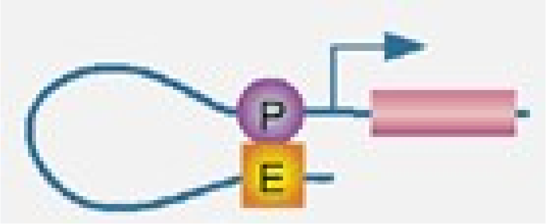
Holds for most transcription binding profiles in the human genome.

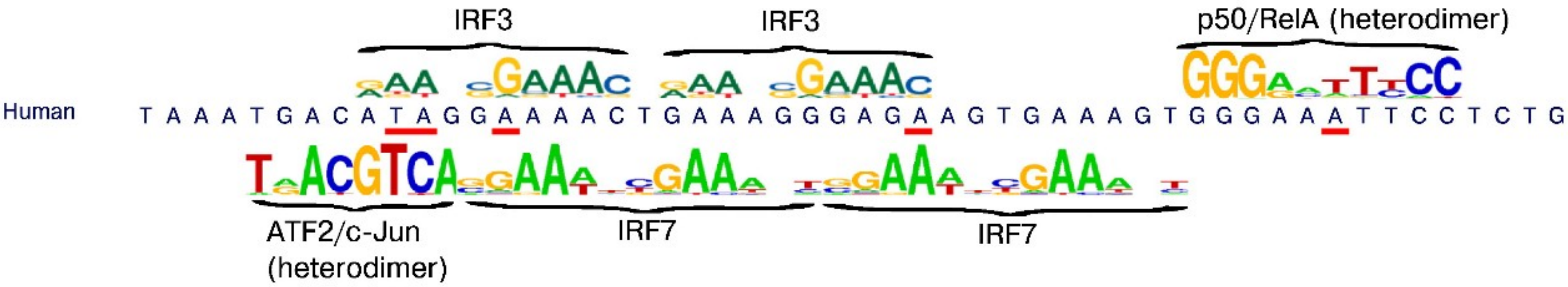
Promoters


Proximal promoter
Genic region
Core promoter
b
BRE upstream
BRE downstream
MTE
DPE
Bits
2
1
Inr
−50
−40
−30
−20
−10
+1
+10
+20
+30
+40
+50
+60
TATA
DCE1
DCE2
DCE3
Vertebrate
Universal
Fly

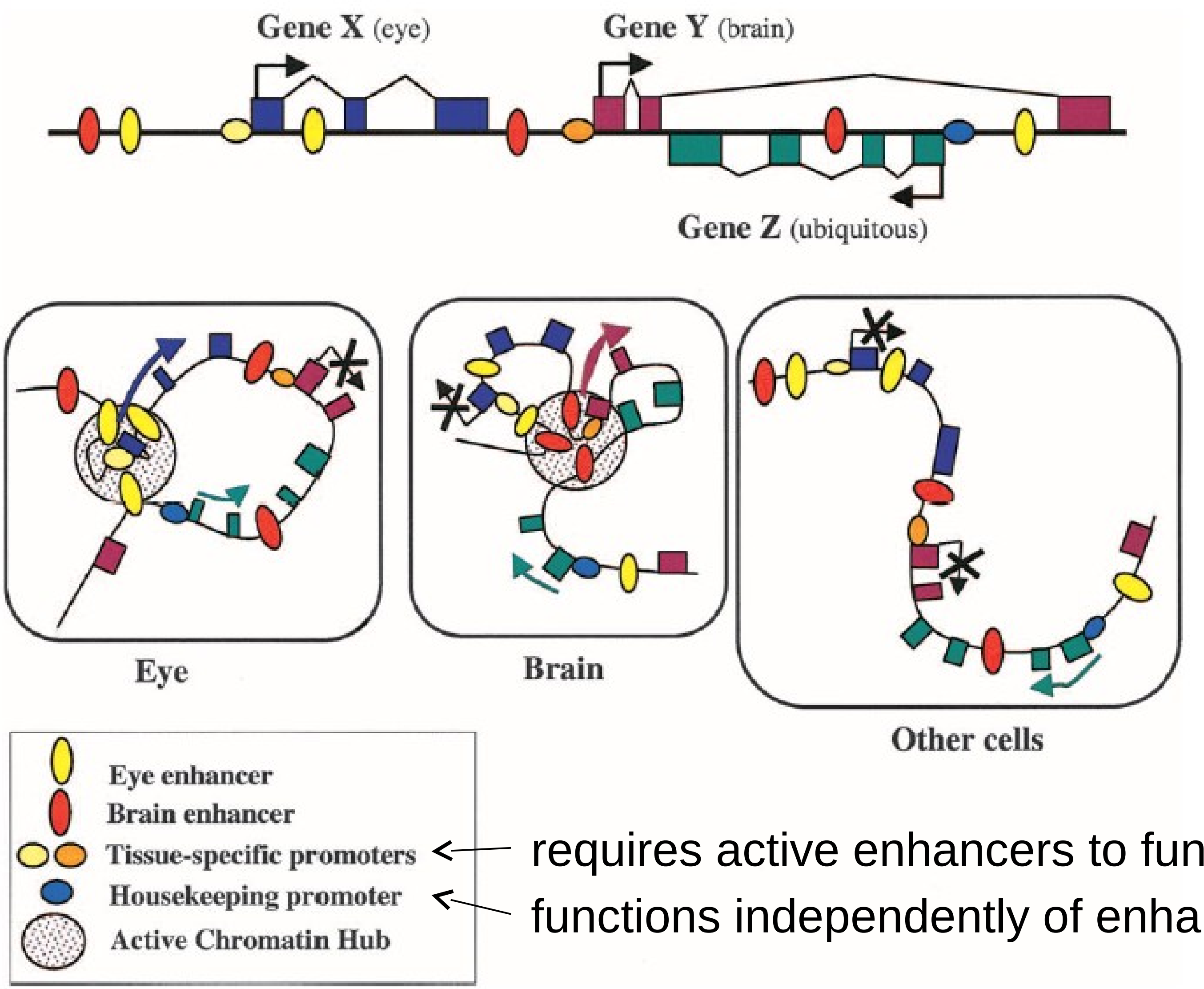
Enhancers


ATF2
p50
cJun
IRF7
DNA ->
<- DNA
IRF3
IRF7
RelA


P
E


IRF3
IRF3
p50/RelA (heterodimer)
Human
T A A A T G A C A T A G G A A A A C T G A A A G G G A G A A G T G A A A G T G G G A A A T T C C T C T G
ATF2/c-Jun (heterodimer)
IRF7
IRF7

One nice hypothetical example



requires active enhancers to function

functions independently of enhancers