

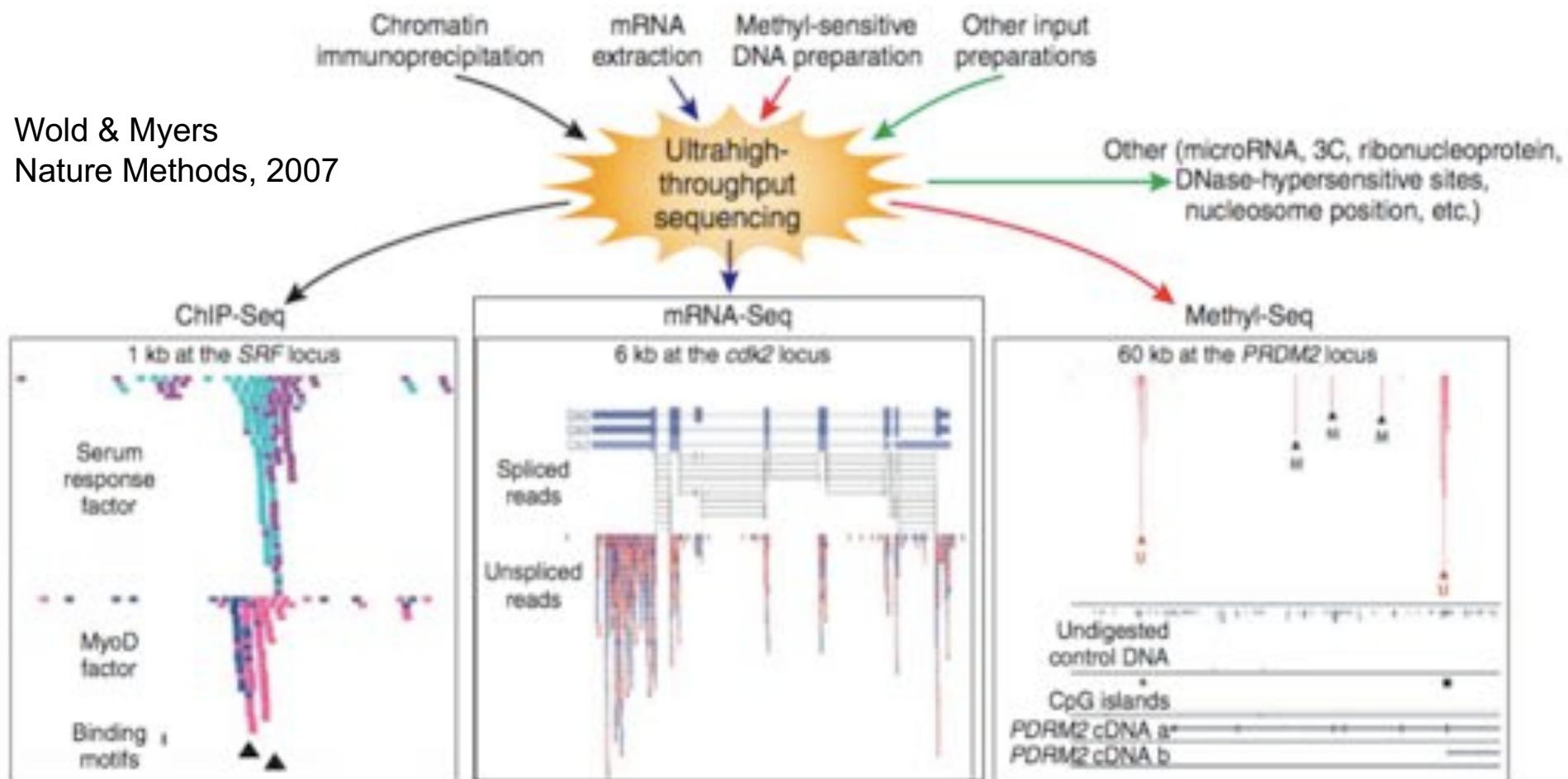
Lecture 3: Transcriptomics and RNA-seq

BioSci D132

Learning Objectives

- Describe RNA-seq workflow
- Calculate gene expression abundances using RNA-seq
- Understand the basics of Hierarchical and K-means clustering
- Describe the structure and applications of Gene Ontology and pathway analysis

Functional genomics using sequencing



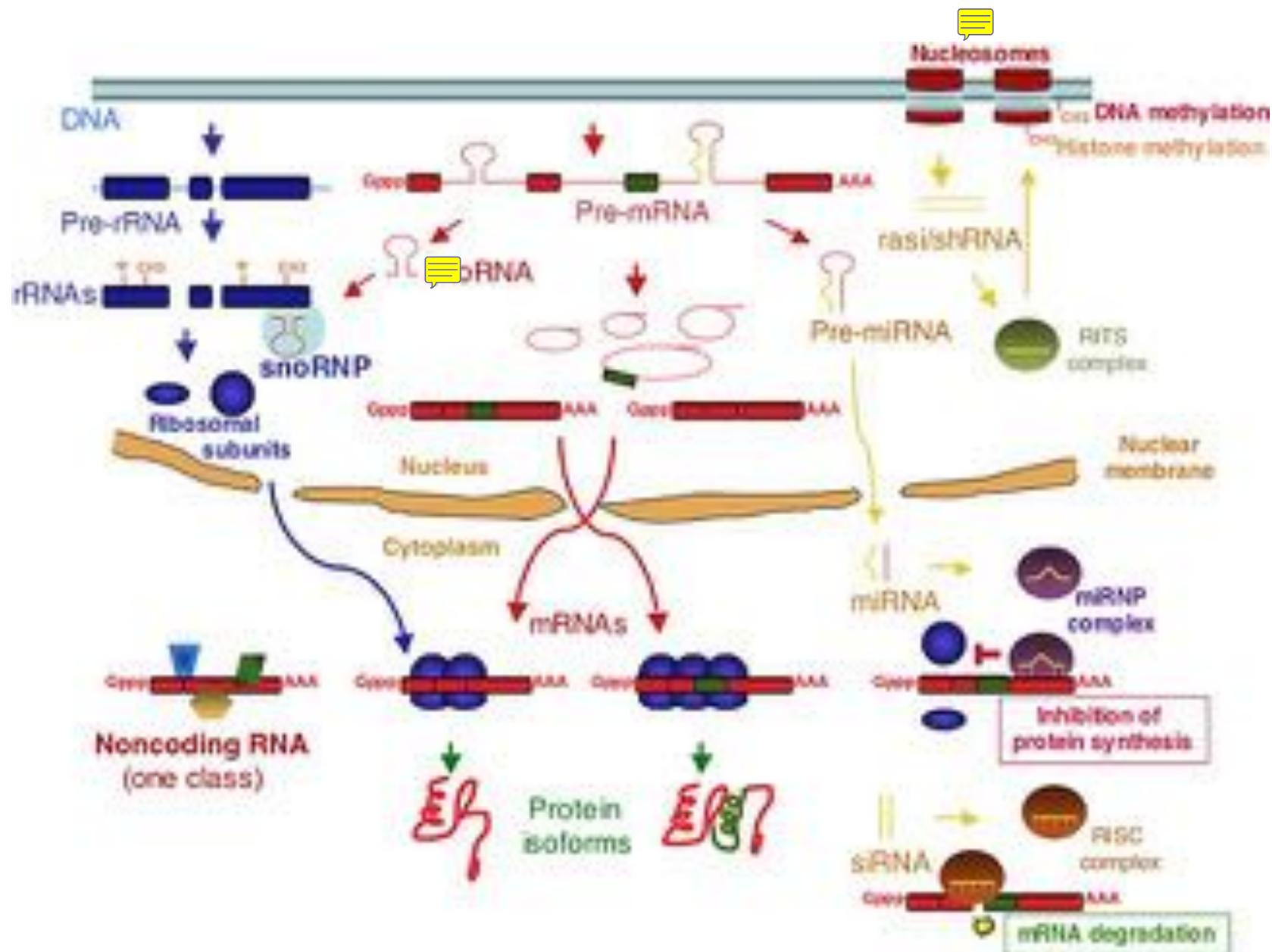
For all sequence-counting assays, the more reads, the better

About half of the worldwide current generation of sequencing capacity is dedicated to these assays.

The transcriptome is a signature of the identity of a cell

- Complete set of transcripts in a cell, tissue or animal.
- Includes:
 - Coding transcripts, mRNA
 - Non-coding Transcripts e.g. lncRNA, siRNA, miRNA
 - rRNA, tRNA (most of the RNA in the cell)
- Dictates the activities of a cell.
- An integral part of functional genomics.
- We will focus primarily on mRNA!

There are many classes of RNA



(Soares, 2006)

Estimate of total RNA content in mammalian cells

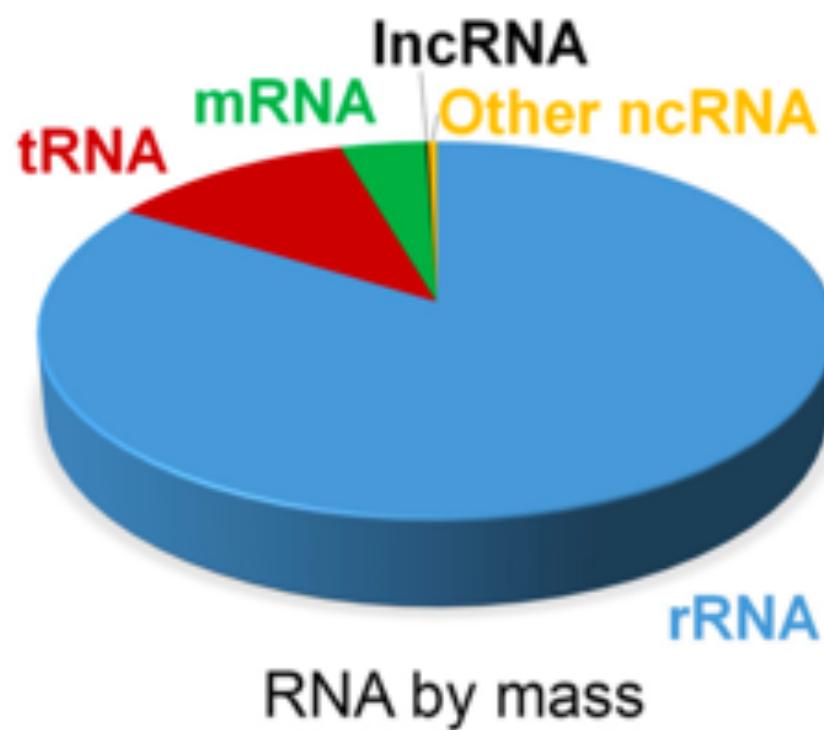
Type	Percent of total RNA by mass	Molecules per cell	Average size (kb)	Total weight picograms/cell	Notes	Reference
rRNAs	80 to 90	$3\text{--}10 \times 10^6$ (ribosomes)	6.9	10 to 30		B Blobel and Potter (1967), Wolf and Schlessinger (1977), Duncan and Hershey (1983)
tRNA	10 to 15	$3\text{--}10 \times 10^7$	<0.1	1.5 to 5	About 10 tRNA molecules /ribosome	Waldron and Lacroute (1975)
mRNA	3 to 7	$3\text{--}10 \times 10^5$	1.7	0.25 to 0.9		Hastie and Bishop (1976), Carter et al. (2005)
hnRNA (pre-mRNA)	0.06 to 0.2	$1\text{--}10 \times 10^3$	10*	0.004 to 0.03	Estimated at 2–4% of mRNA by weight	Mortazavi et al. (2008), Menet et al. (2012)
Circular RNA	0.002 to 0.03	$3\text{--}20 \times 10^3$	~0.5	0.0007 to 0.005	Estimated at 0.1–0.2% of mRNA**	Salzman et al. (2012), Guo et al. (2014)
snRNA	0.02 to 0.3	$1\text{--}5 \times 10^5$	0.1–0.2	0.008 to 0.04		Kiss and Filipowicz (1992), Castle et al. (2010)
snoRNA	0.04 to 0.2	$2\text{--}3 \times 10^5$	0.2	0.02 to 0.03		Kiss and Filipowicz (1992), Cooper (2000), Castle et al. (2010)
miRNA	0.003 to 0.02	$1\text{--}3 \times 10^5$	0.02	0.001 to 0.003	About 10^5 molecules per 10 pg total RNA	Bissels et al. (2009)
7SL	0.01 to 0.2	$3\text{--}20 \times 10^4$	0.3	0.005 to 0.03	About 1–2 SRP molecules/100 ribosomes	Raue et al. (2007), Castle et al. (2010)
Xist	0.0003 to 0.02	$0.1\text{--}2 \times 10^3$	2.8	0.0001 to 0.003		Buzin et al. (1994), Castle et al. (2010)
Other lncRNA	0.03 to 0.2	$3\text{--}50 \times 10^3$	1	0.002 to 0.03	Estimated at 1–4% of mRNA by weight	Mortazavi et al. (2008), Rämsköld et al. (2009), Menet et al. (2012)

*The size for the average unspliced pre-mRNA is 17 kb; however, most pre-mRNAs are partially spliced at any given time, and the average size of hnRNA is estimated at 10 kb (Salditt-Georgieff et al., 1978).

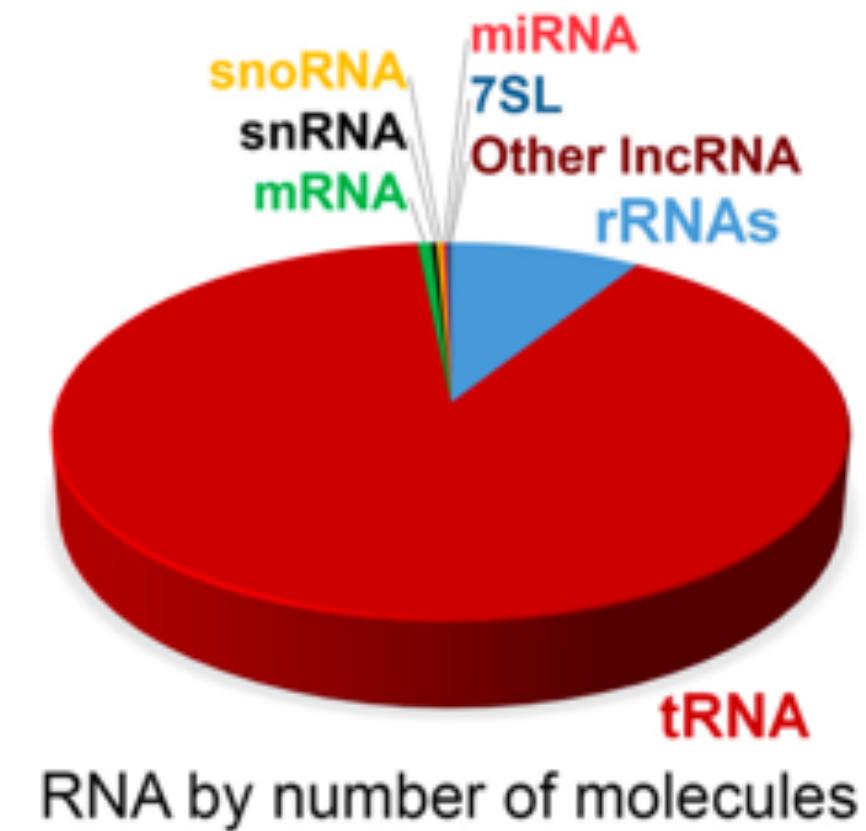
**Based on the finding that 1–2% of all mRNA species generate circular RNA, which is present at 10% of the level of the parental mRNA.

messengerRNAs are a small portion of the RNA of a cell

A



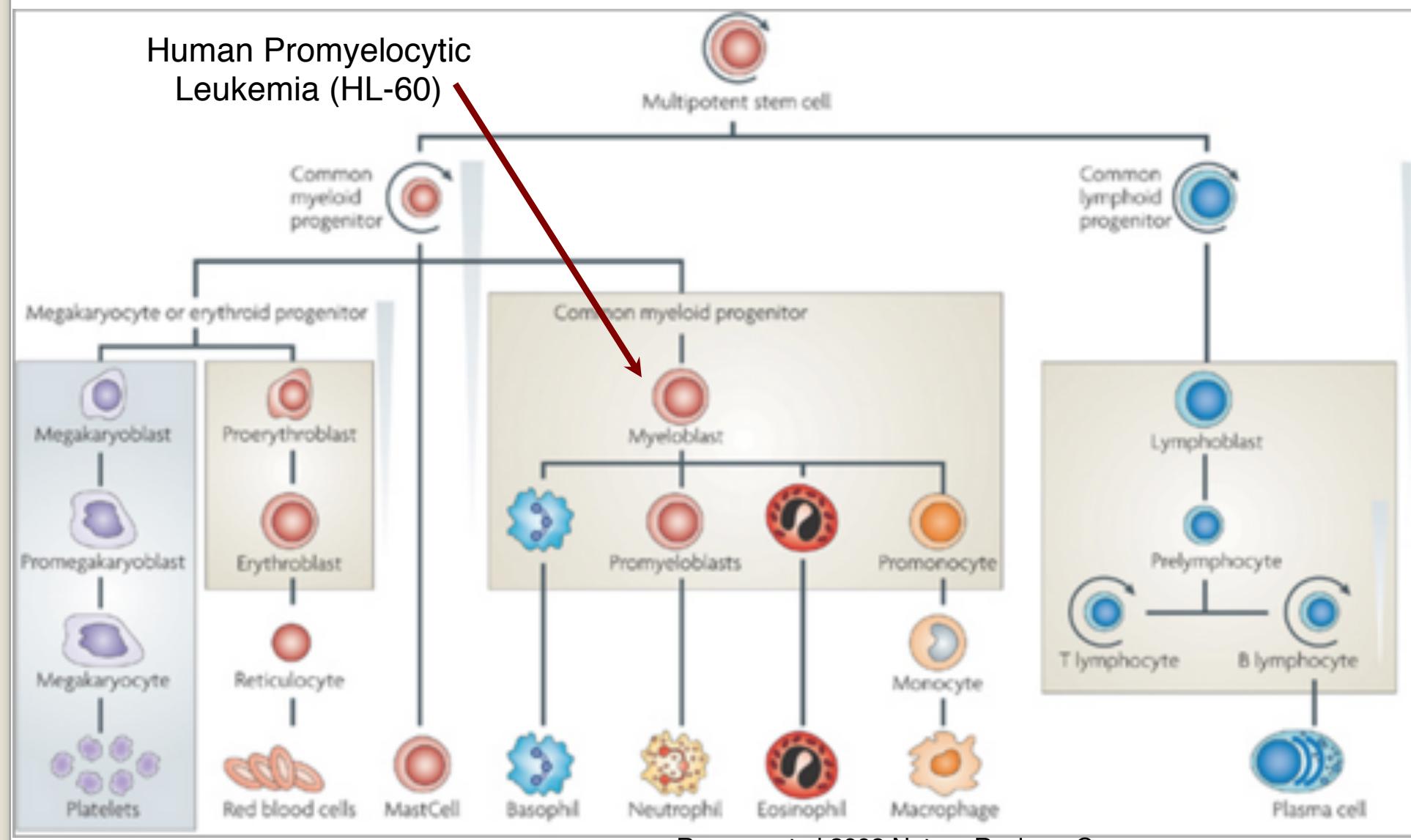
B



Transcriptomics

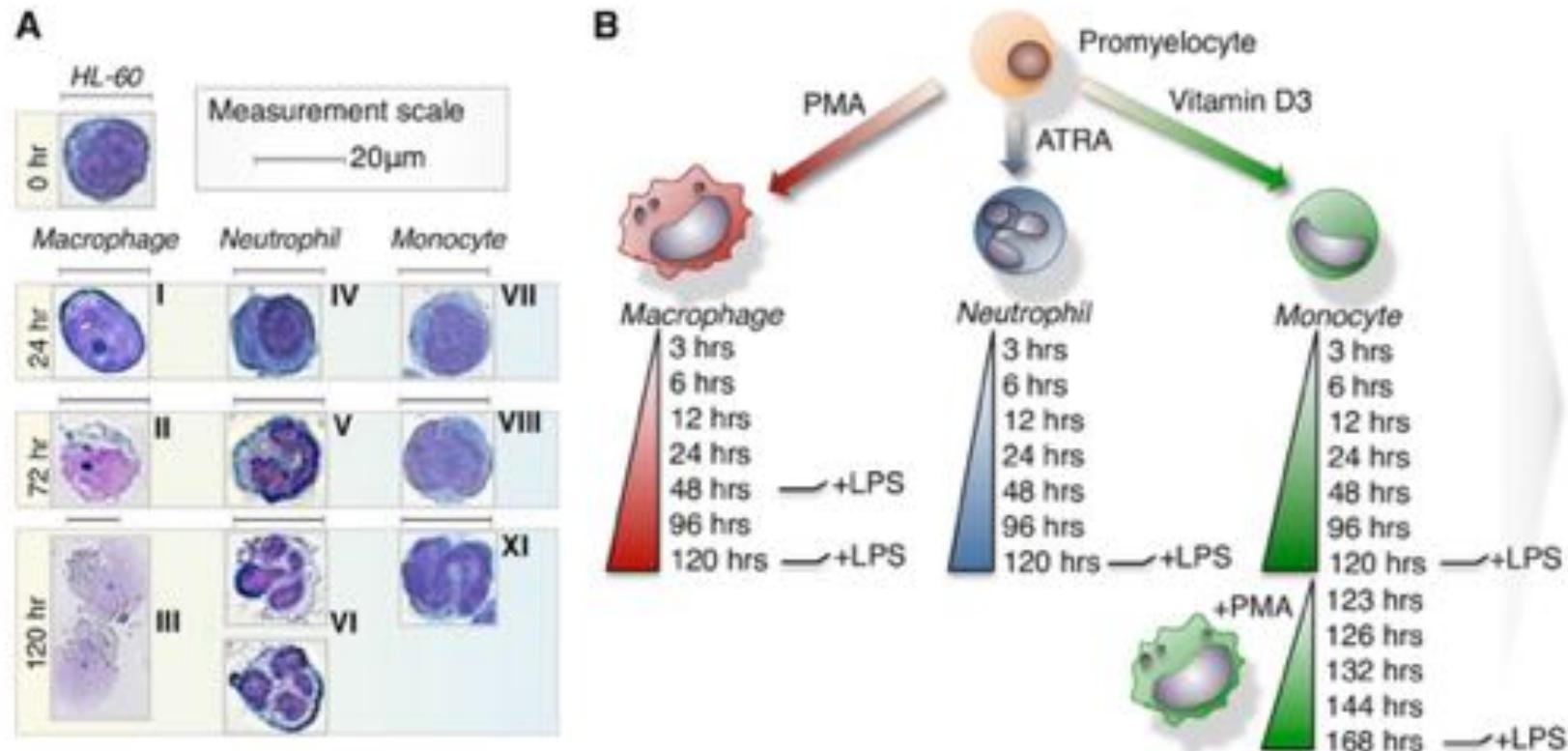
- For a cell or tissue, transcriptomics entails:
 - identification
 - characterization
 - and cataloguing of all transcripts
 - time point and/or treatment
- Applications
 - The big picture:
 - What genes are expressed?
 - Which genes are differentially expressed?

Hematopoiesis – the generation of white blood cells



Ramsay et al 2008 Nature Reviews Cancer

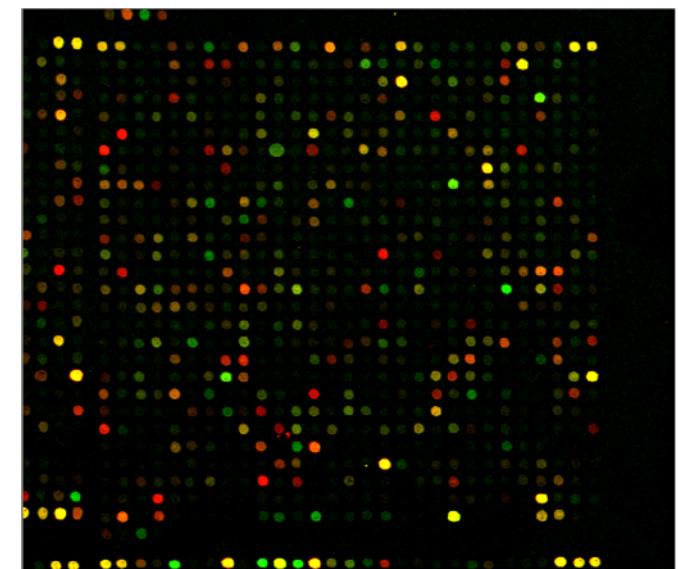
Time courses of gene expression



Measure RNA expression as cells are differentiating

(Obsolete) technologies for transcriptomics

- Microarrays:
 - Was the main technique used for genome-wide functional assays for about 10 years.
 - Based on hybridization to pre-selected probes.
 - Limited dynamic range compared to RNA-seq.
 - High background noise compared to RNA-seq.
 - Lower reproducibility rate than RNA-seq.



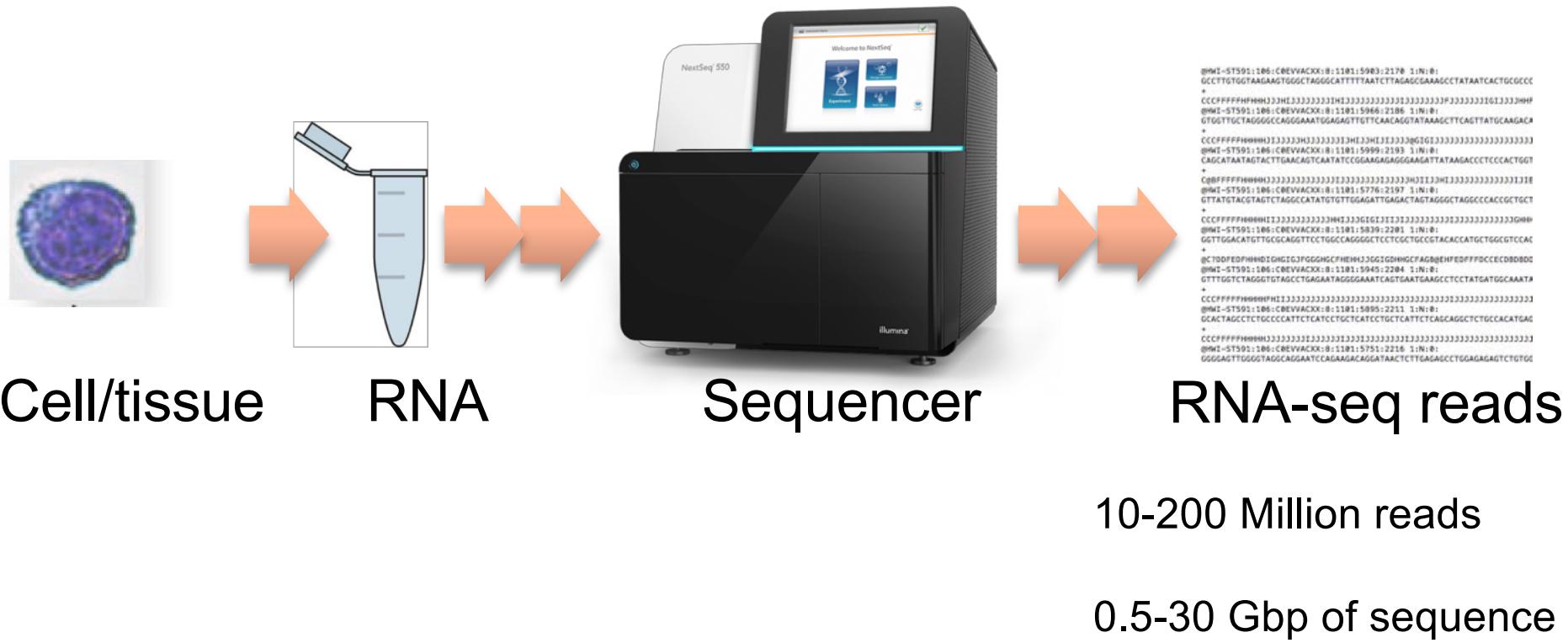
<http://www.microarray.org/sfgf/>

Technologies for transcriptomics

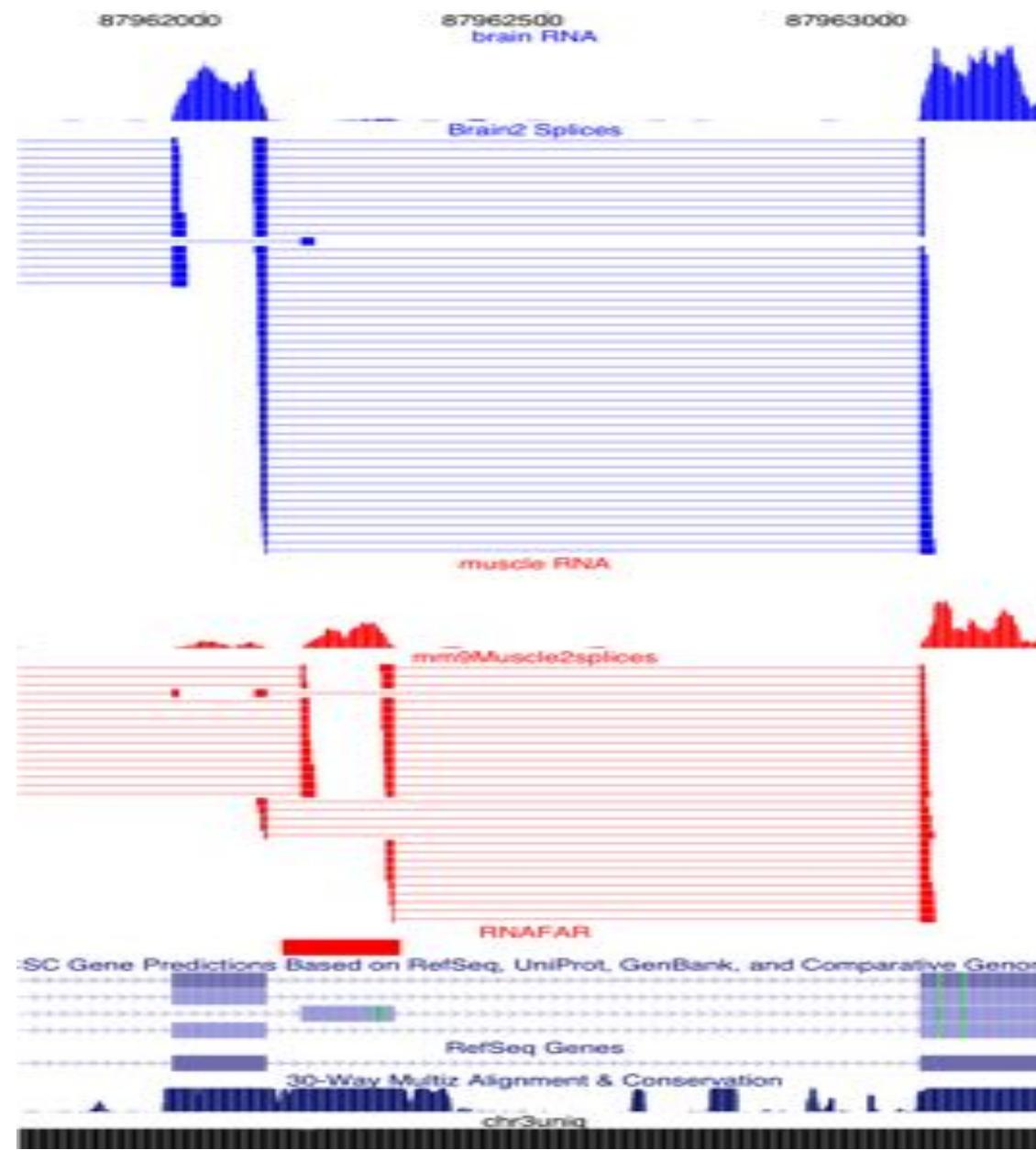
RNA-seq:

- A direct method for ultra-high-throughput sequencing of cDNA using millions of reads.
- Unless noted otherwise, this is really about sequencing mRNA

Overview of RNA-seq

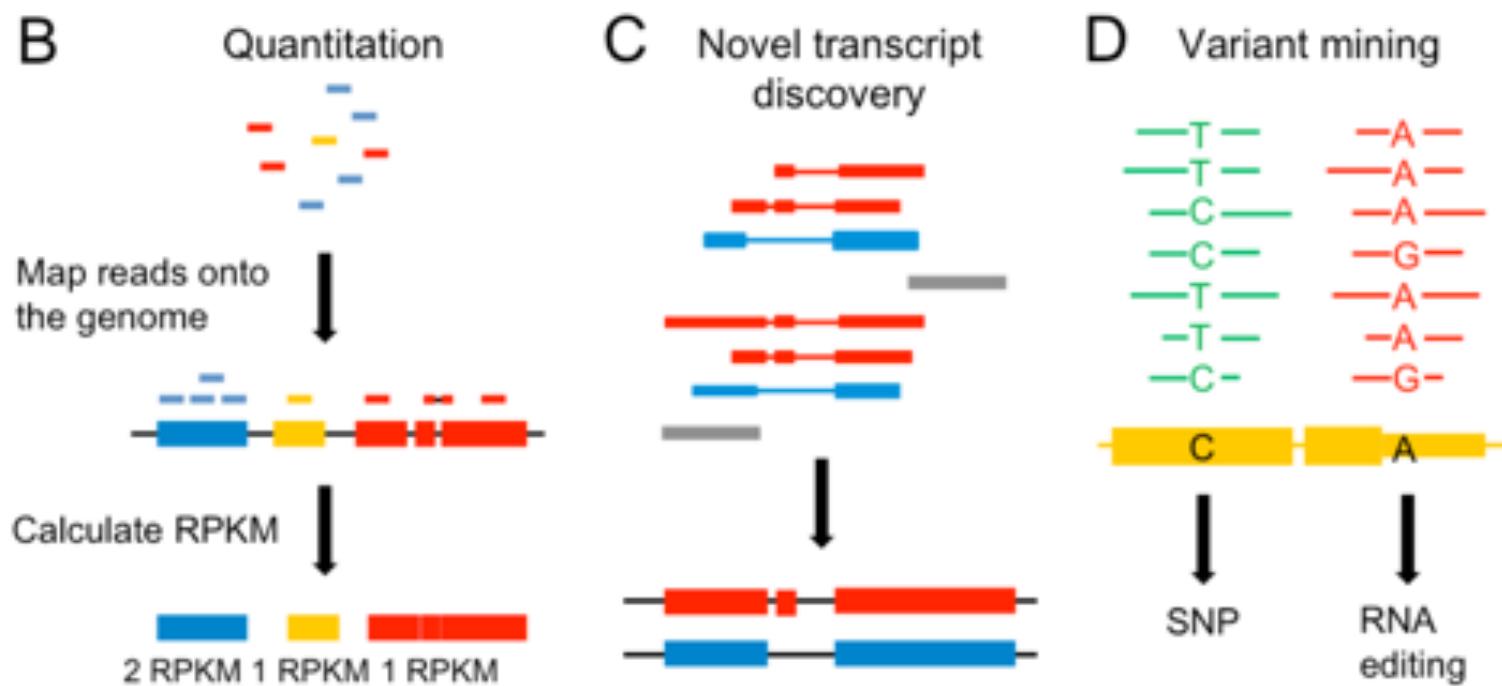


Detecting alternative splicing



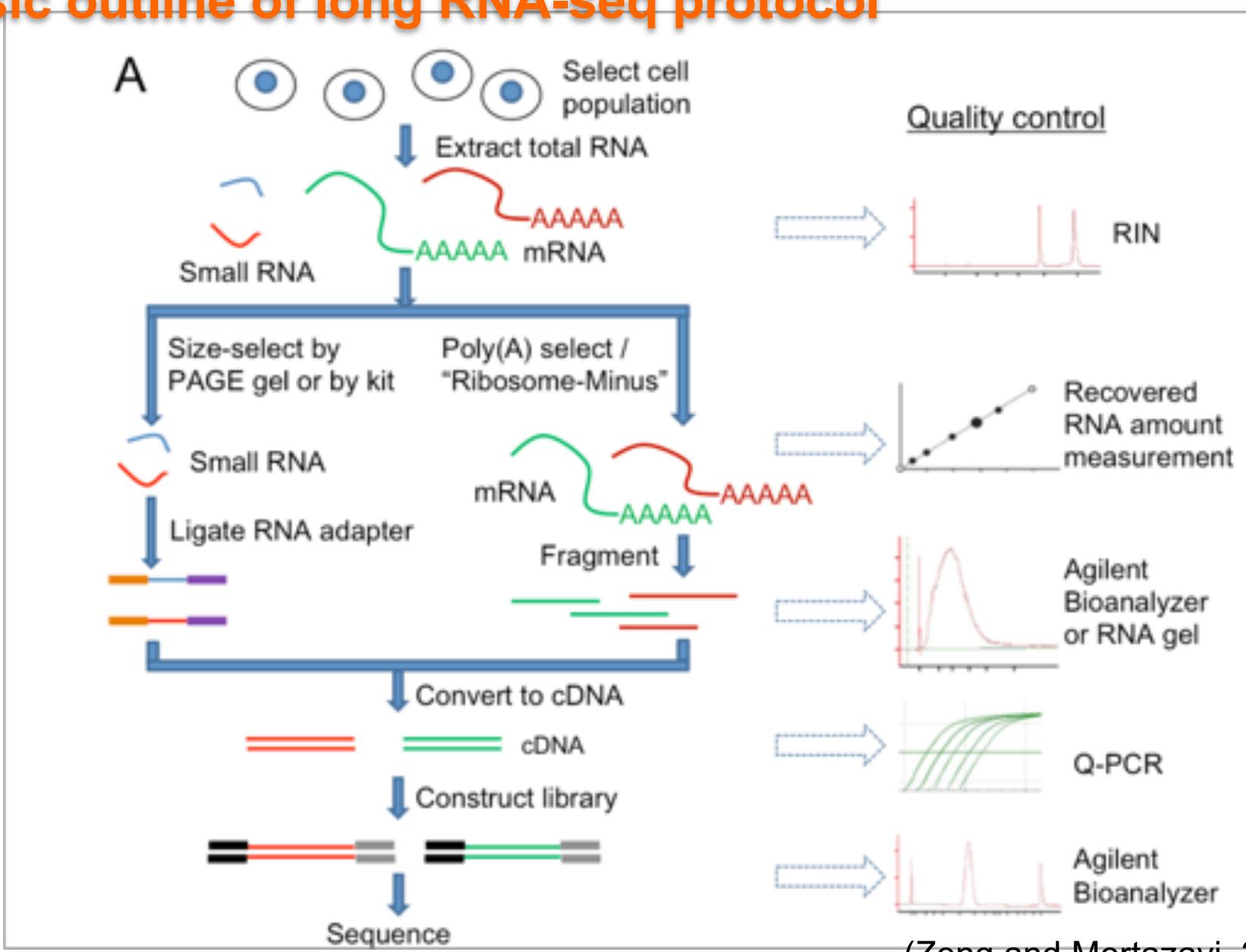
(Mortazavi, 2008)

Applications of RNA-seq



(Zeng and Mortazavi, 2012)

Basic outline of long RNA-seq protocol



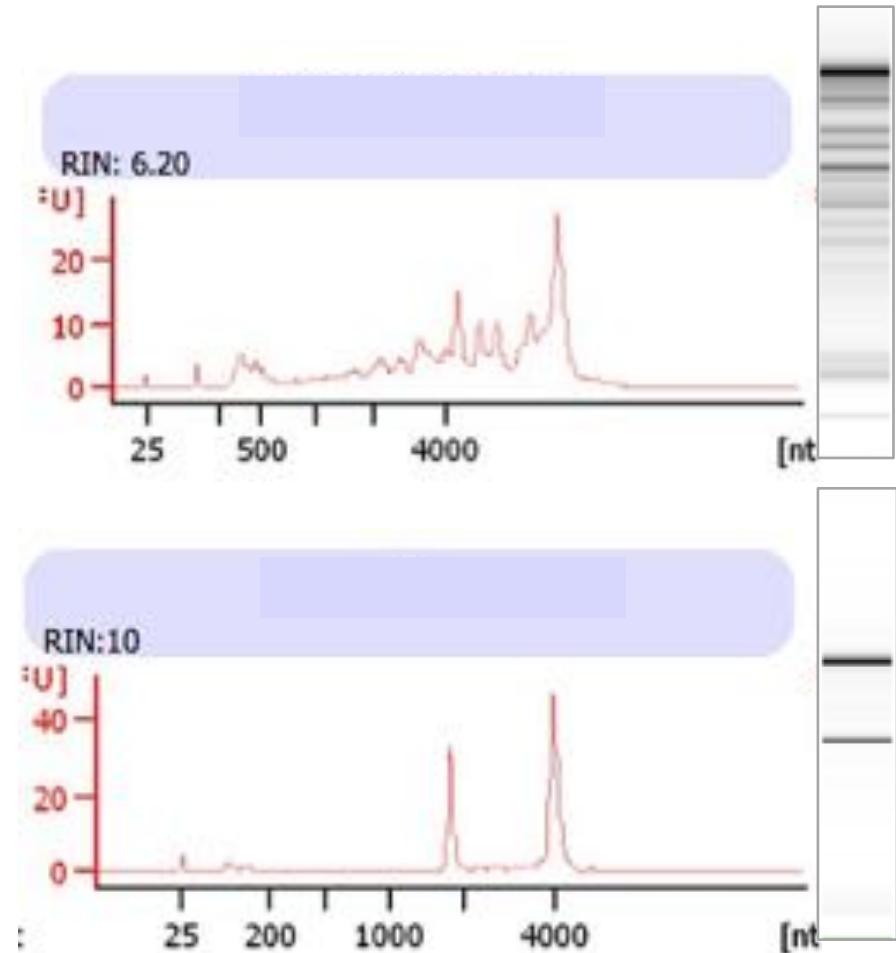
(Zeng and Mortazavi, 2012)

Quality Control (QC) of RNA samples before sequencing

- Best results obtained when RNA isolated from a homogenous cell population or pooled samples.
- In case of extra step with PCR for amplification of low RNA quantities, the quality of RNA should be monitored with Q-PCR for any distortions introduced by PCR amplification.
- The quality of RNA should be quantitatively checked with Bioanalyzer before sequencing.

RNA Integrity number

- Bioanalyzer, a microfluidics platform, can be used for sizing, quantification and quality control of RNA, DNA, proteins and cells.
- RNA integrity number (RIN) is a quantitative measure of RNA integrity on a scale of 1-10 (most-least degraded).
- Any sample with RNA integrity score of 8 or higher is suitable for sequencing.

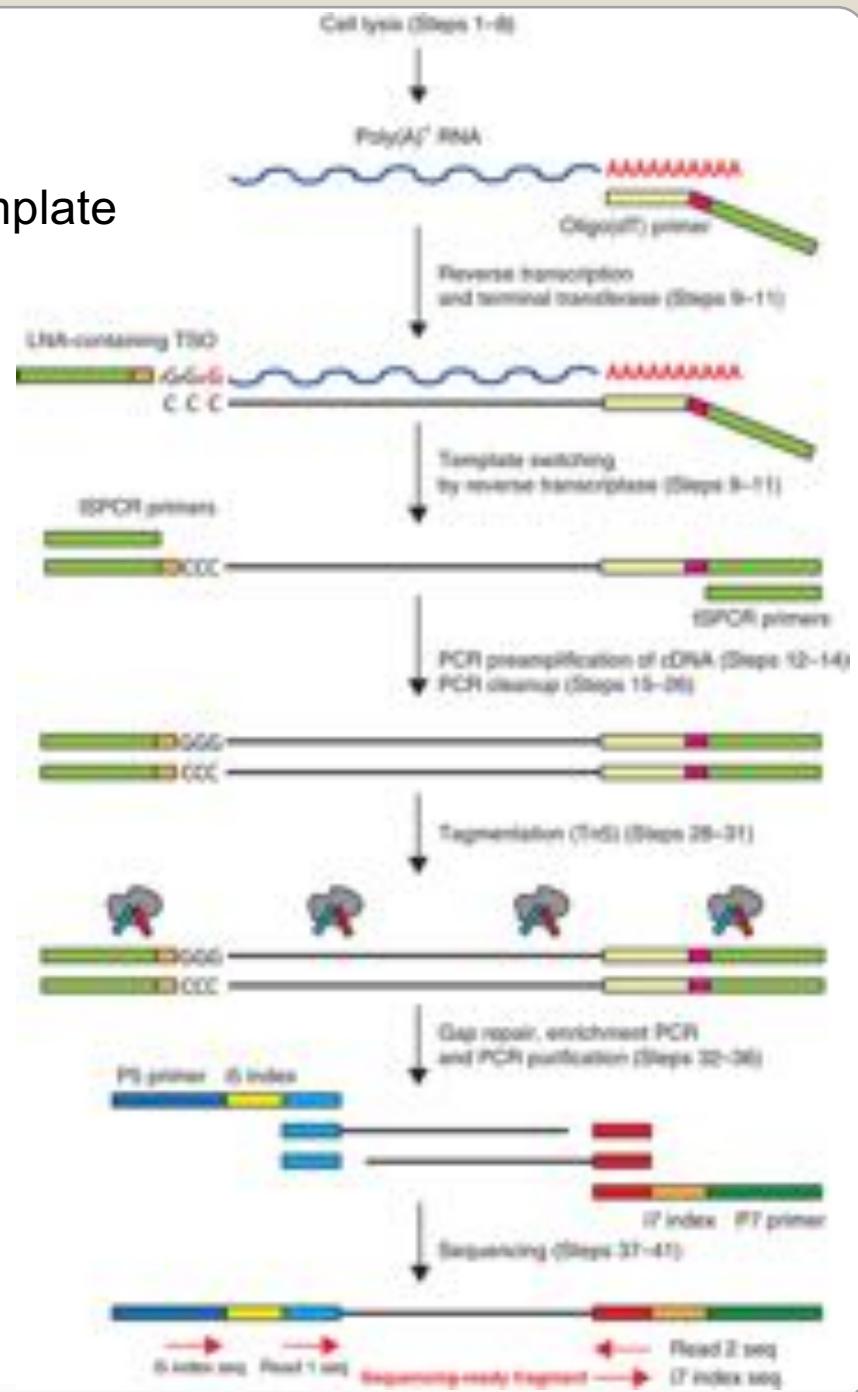


2013: Smart-seq2

SMART: **S**witching **M**echanism **a**t 5' of **R**NA **T**emplate

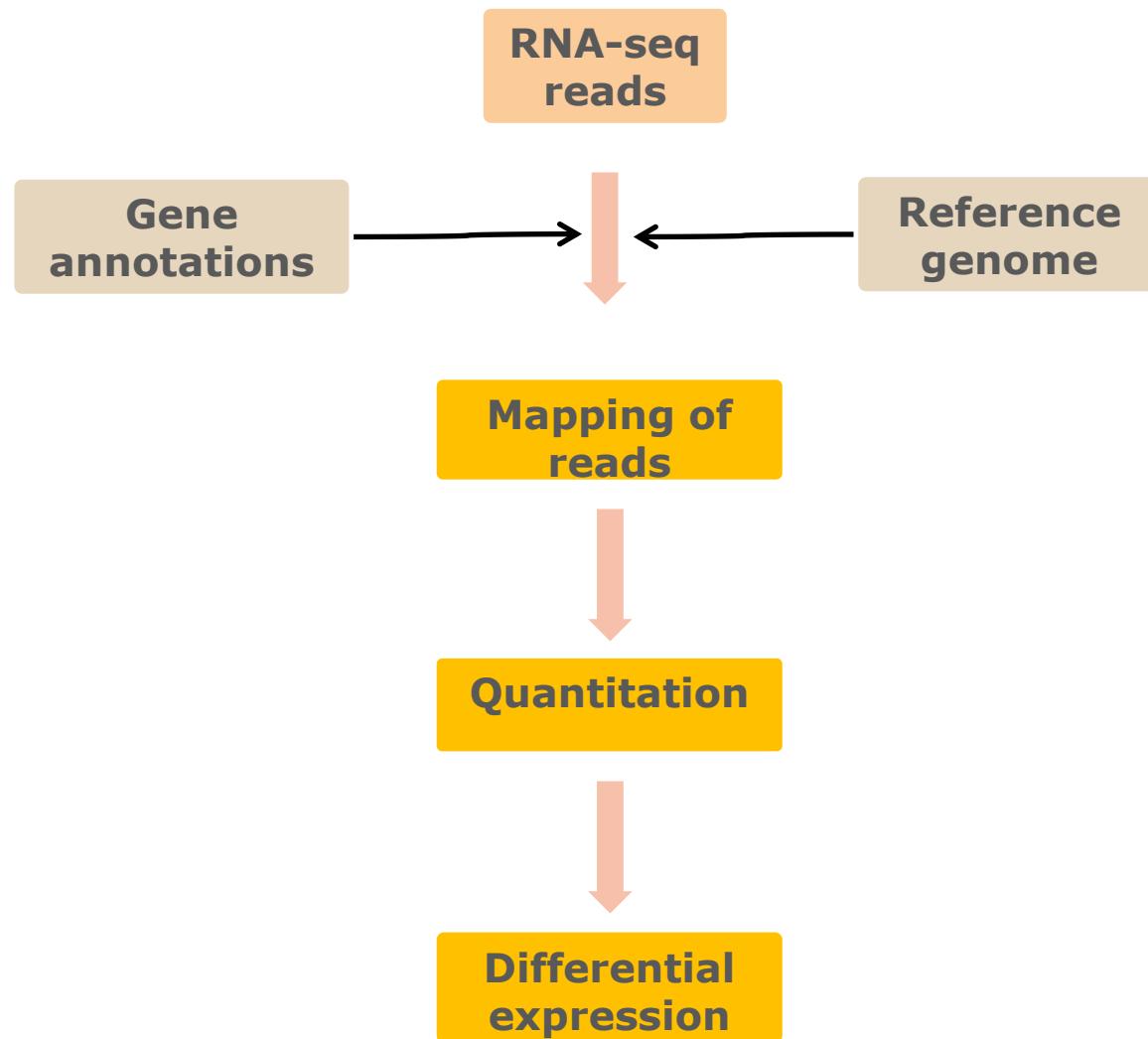
Developed for Full length RNA transcription

Optimized for small amounts of RNA



(Picelli, 2014)

Overview of RNA-seq data analysis

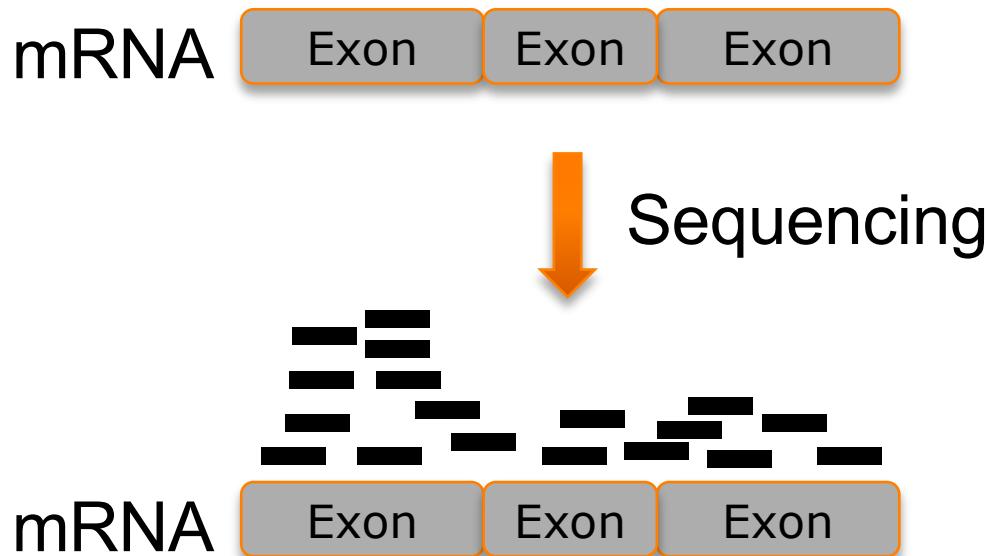


Computational challenges posed by RNA-seq

- Reliable and efficient mapping of short reads to genome.
- Reliable mapping of reads across splice junctions.
- Reliable transcript assembly.
- More difficult in larger, complex genomes due to orthologs, repeats, and pseudo-genes.

Uneven coverage levels

- Transcripts coverage levels differ along their lengths depending on library construction procedure.
 - Fragmentation before or after reverse transcription.



Sequencing bias

- GC-rich and GC-poor regions are under-sampled during sequencing by some chemistries (e.g. Illumina).
- Random priming causes bias against common hexamer e.g. (AAAAAA).
- Not so random priming.



Mapping RNA-seq reads

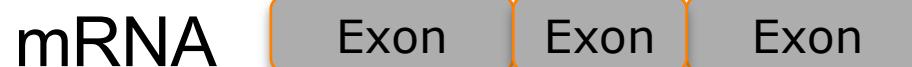
- Mapping reads:
 - determining the position of a read within the reference genome
- “Read mapping” problem:
 - Millions to billions of short reads.
 - Large reference genomes.
 - How to efficiently achieve mapping.
 - Deal with reads from repetitive regions.
 - Deal with sequencing errors .

Spliced and unspliced reads



RNA processing

A large orange arrow points downwards from the pre-mRNA stage to the mRNA stage.



Sequencing,
read mapping

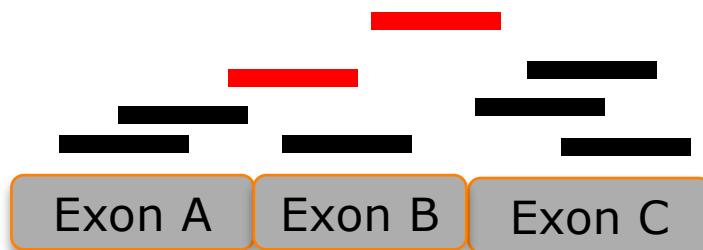
A large orange arrow points downwards from the mRNA stage to the sequencing results stage.



— Spliced read
— Unspliced read

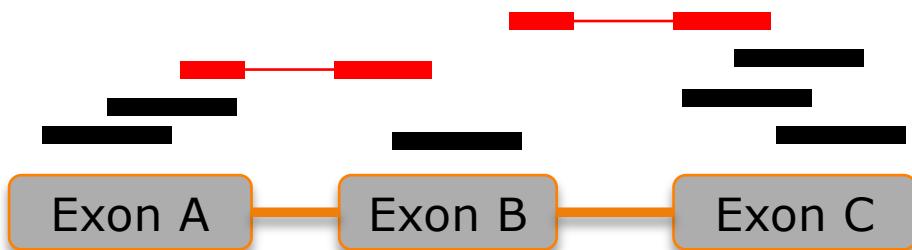
Alternative mapping strategies

- Mapping on the transcriptome



— Junction spanning read
— Aligned read

- Mapping on the genome



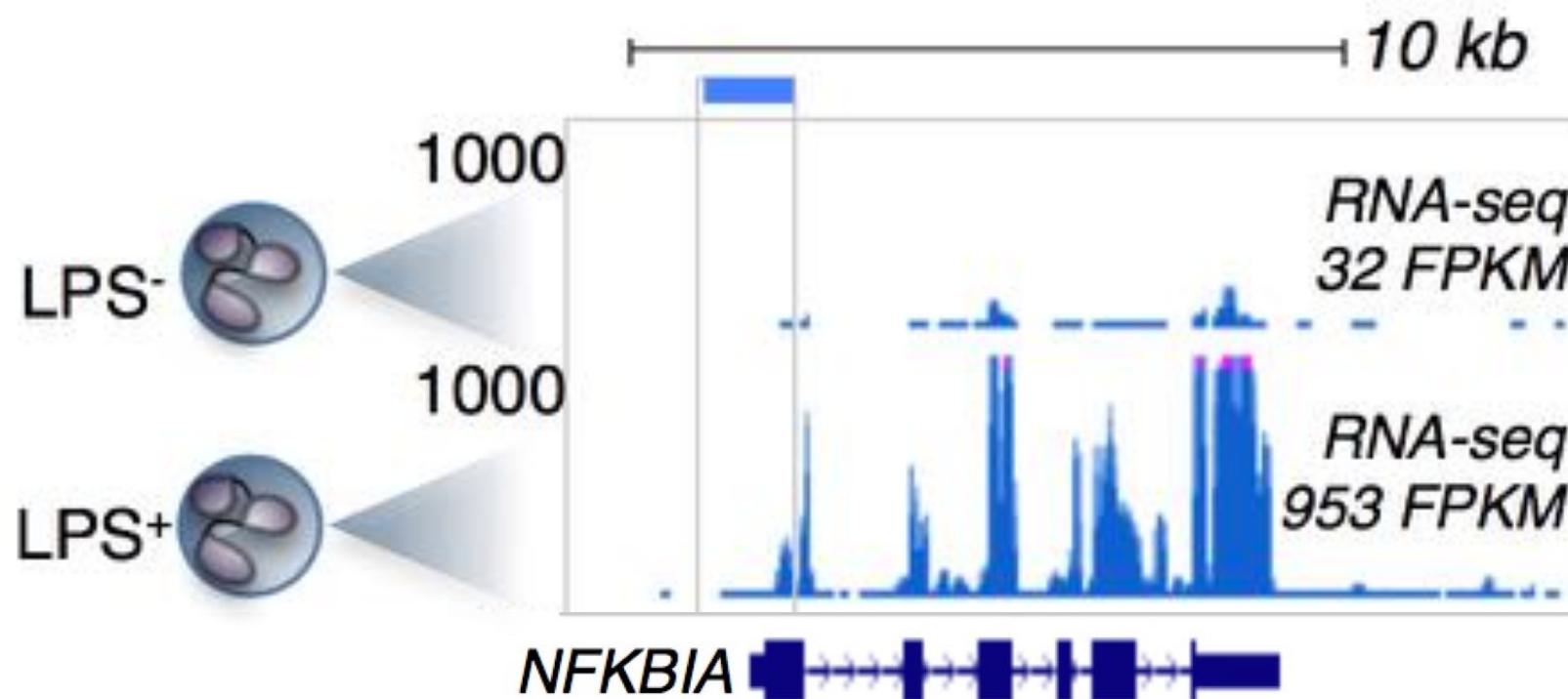
— Intron-spanning read
— Aligned read

- Must provide the mRNAs
- Faster
- No discovery

- Must provide genome
- Slower, more computation
- Can discover new mRNAs

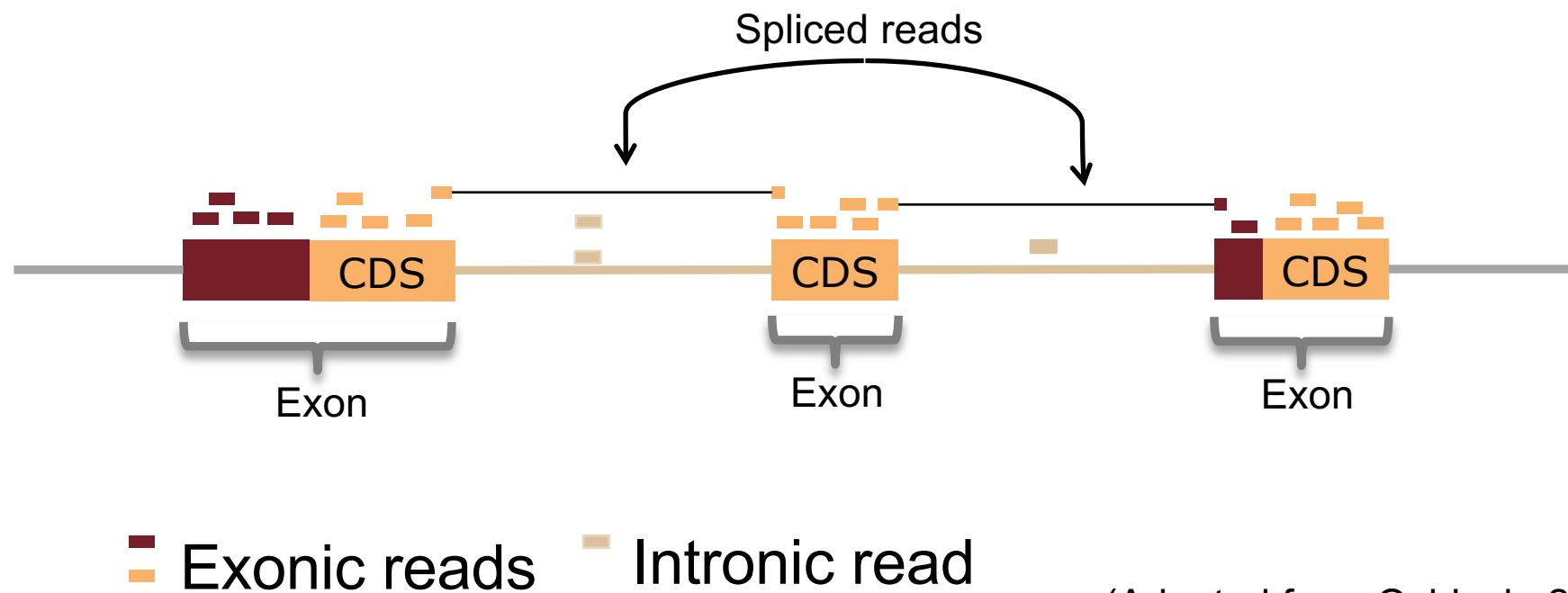
Quantifying gene expression

- Read count or read density can be used to estimate expression levels.
- More “normalized signal” → More expression



Summarizing read counts

- Summarize over a biologically meaningful unit e.g. exon, CDS, transcript, or gene.
- Summarizing over the entire gene length also captures reads mapped to introns.

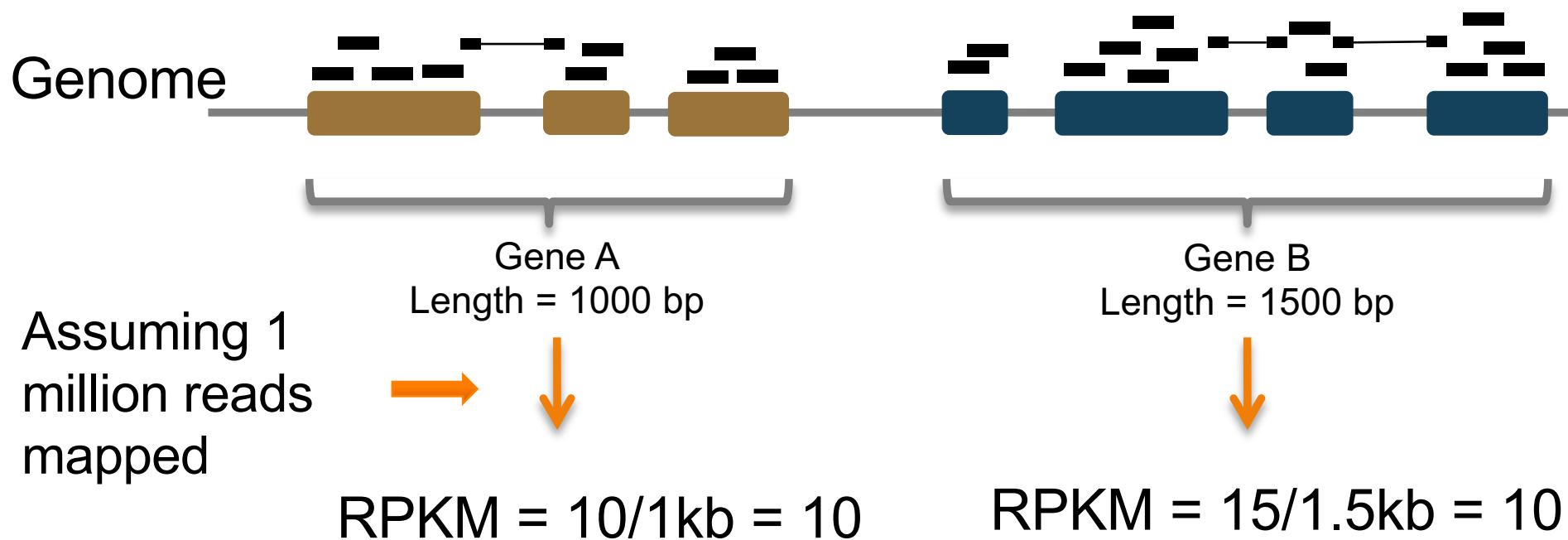


(Adapted from Oshlack, 2010)

RPKM Reads Per Kilobase of exon per Million reads

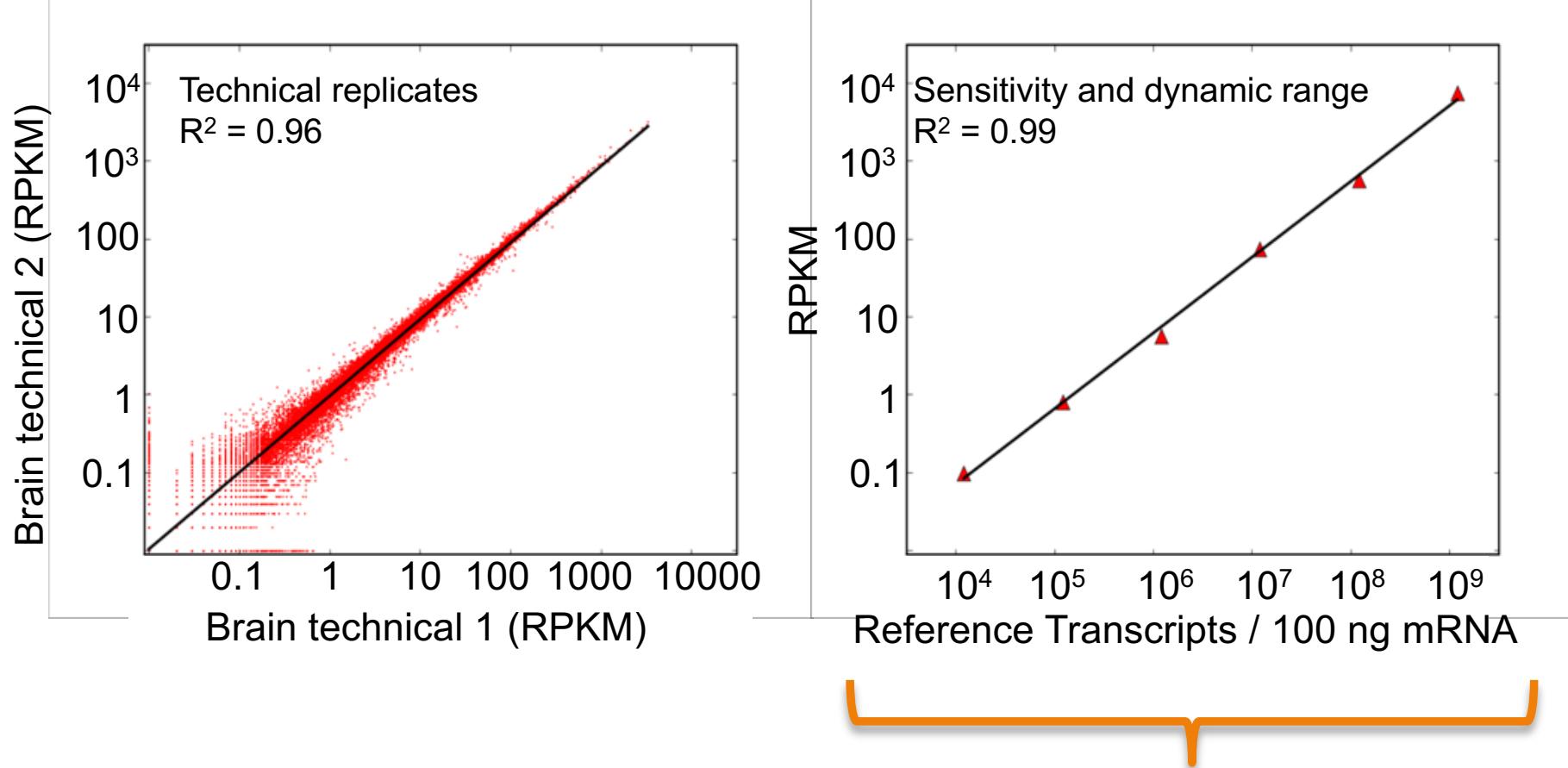
$$RPKM = \frac{\text{Read counts for transcript}}{\frac{N}{1000000} \frac{L}{1000}}$$

Where N = Total reads mapped
L = Length of region (bp)



If paired-end reads, then Fragments Per Kilobase per Million of mapped reads (FPKM). I will use RPKM and FPKM interchangeably, but there are differences – see video.

RPKMs correlate with the amount of transcripts in sample

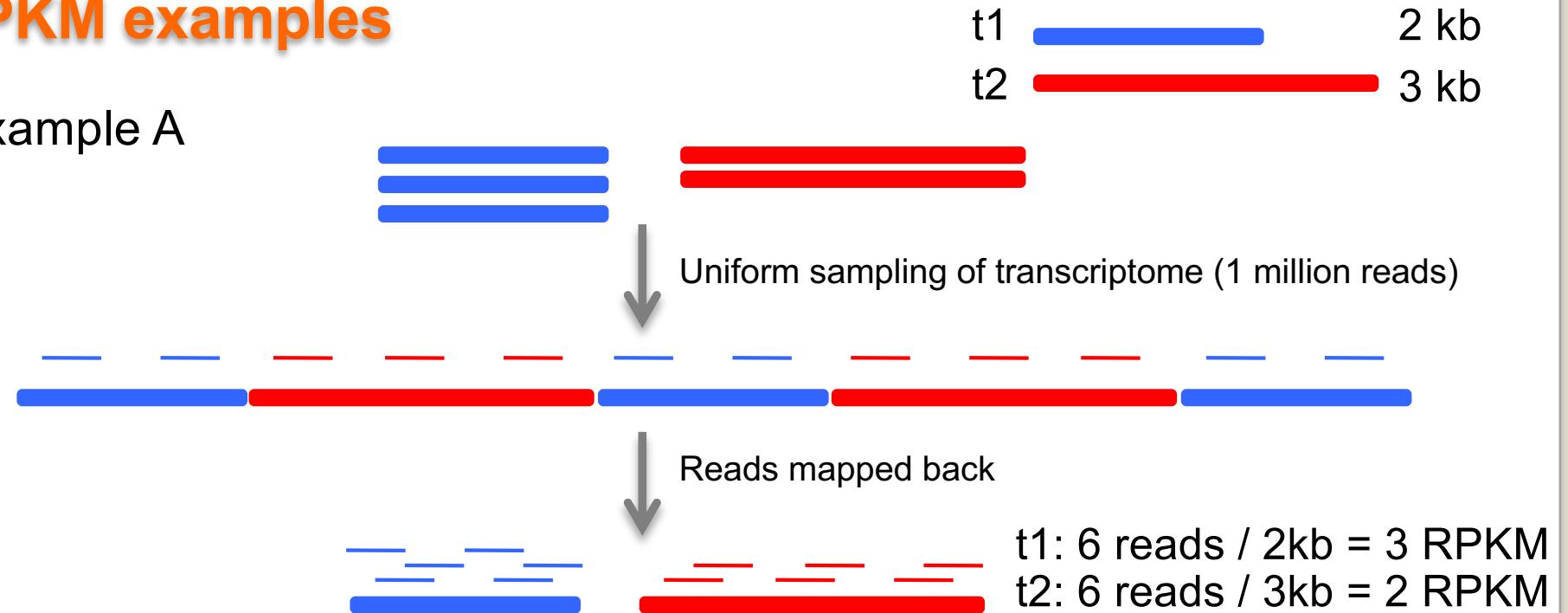


Copies of known transcripts
added at known quantities

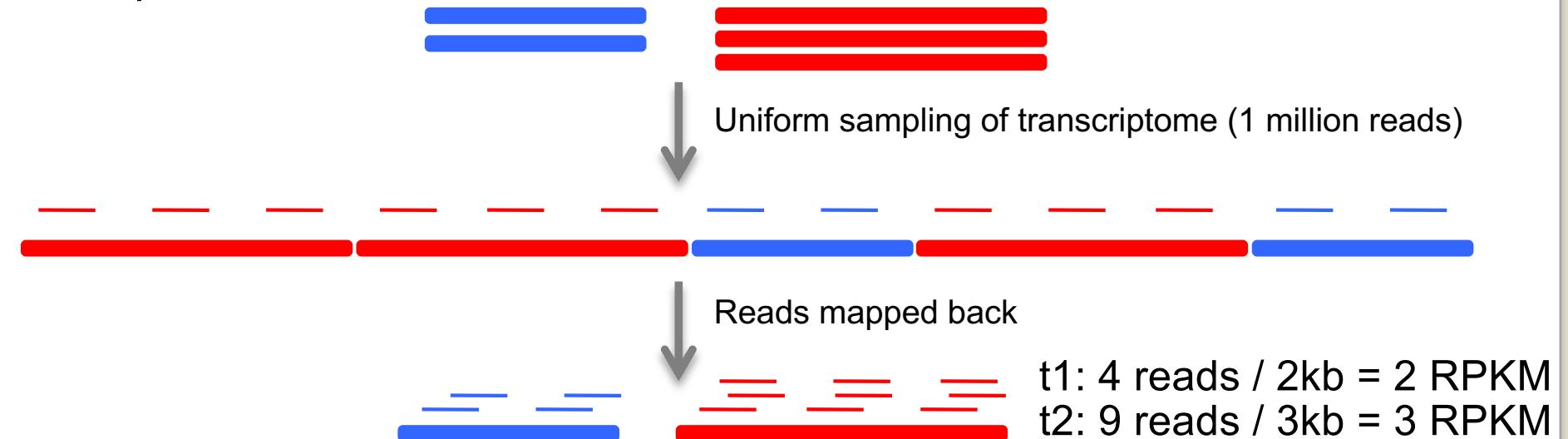
(Mortazavi, 2008)

RPKM examples

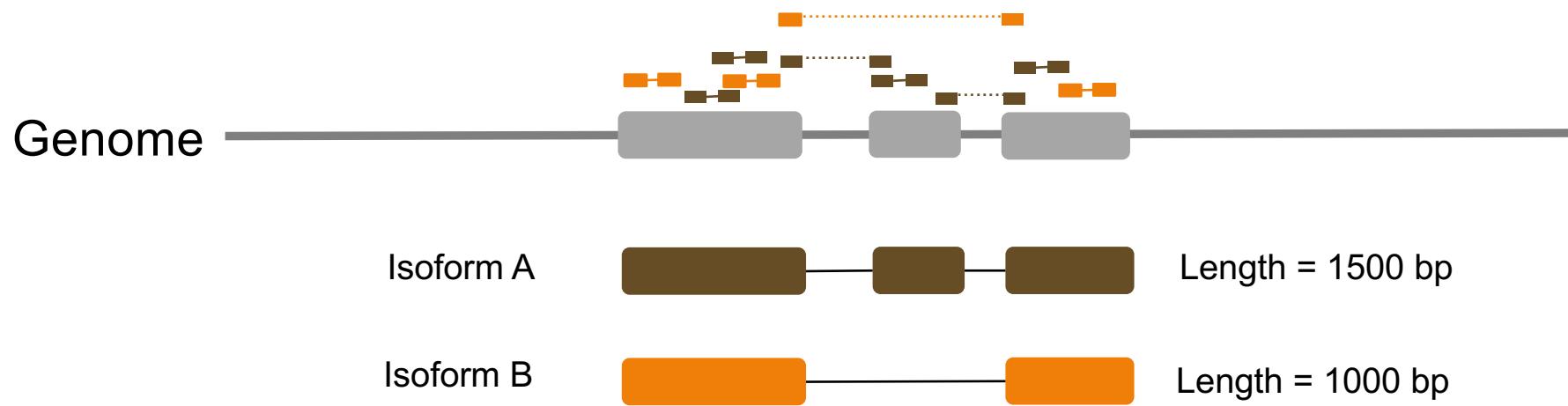
Example A



Example B



Adding transcript FPKMs to get gene FPKMs



Assuming 1 million reads mapped $\rightarrow \text{FPKM} = \text{FPKM}_A + \text{FPKM}_B = 6/1.5\text{kb} + 4/1\text{kb} = 8$

Yet another quantity - TPM: Transcript Per Million

$$TPM_i = \left(\frac{FPKM_i}{\sum_j FPKM_j} \right) * 10^6$$

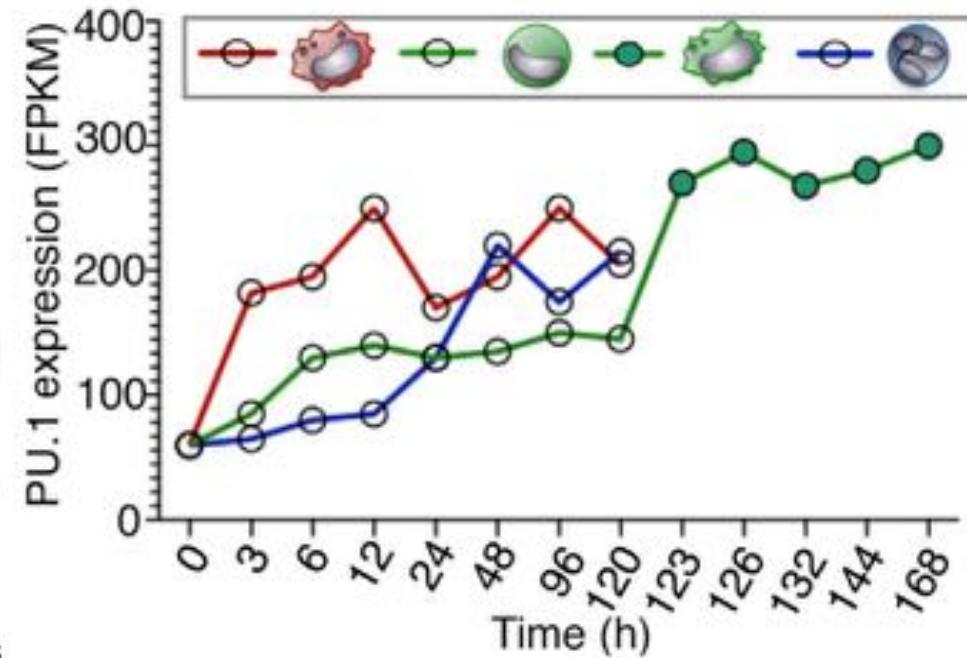
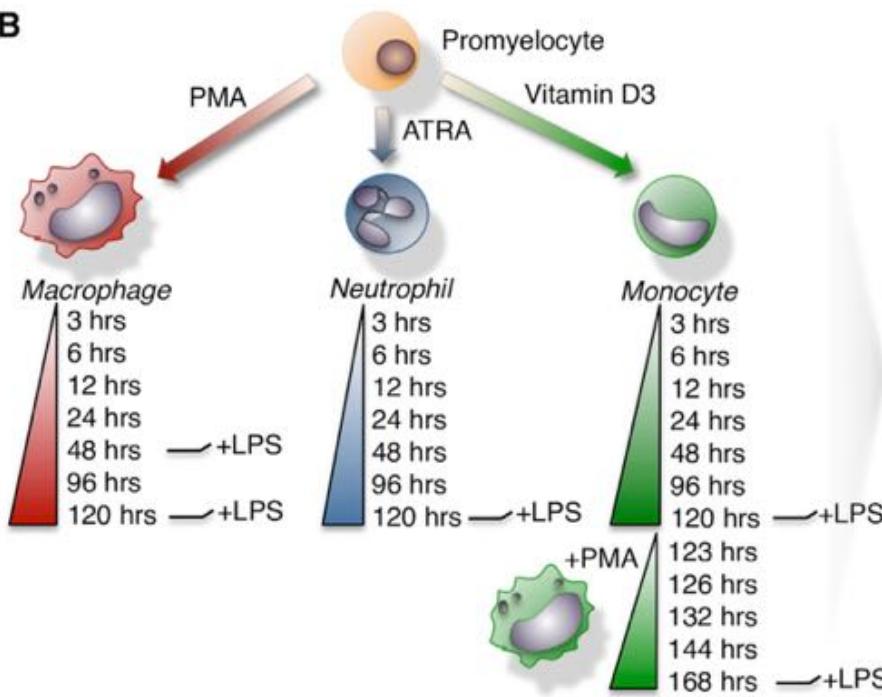
More stable if there is a big difference in distribution of transcripts and transcripts lengths – for example comparing very different cell types

Quantitation of transcript expression

- Some caveats:
 - Uneven read coverage across the length of a transcript.
 - Single-end reads have less information and are harder to assemble. Paired-end reads are better.
 - Unstable solutions that are sensitive to the number of reads mapped.
 - Biological replicates are very important. Count-based analysis tools such as edgeR best used with replicates.

Time courses of gene expression

B



PU.1 is an important transcription factor controlling myeloid cell differentiation

Differential expression analysis

- **Goal:**

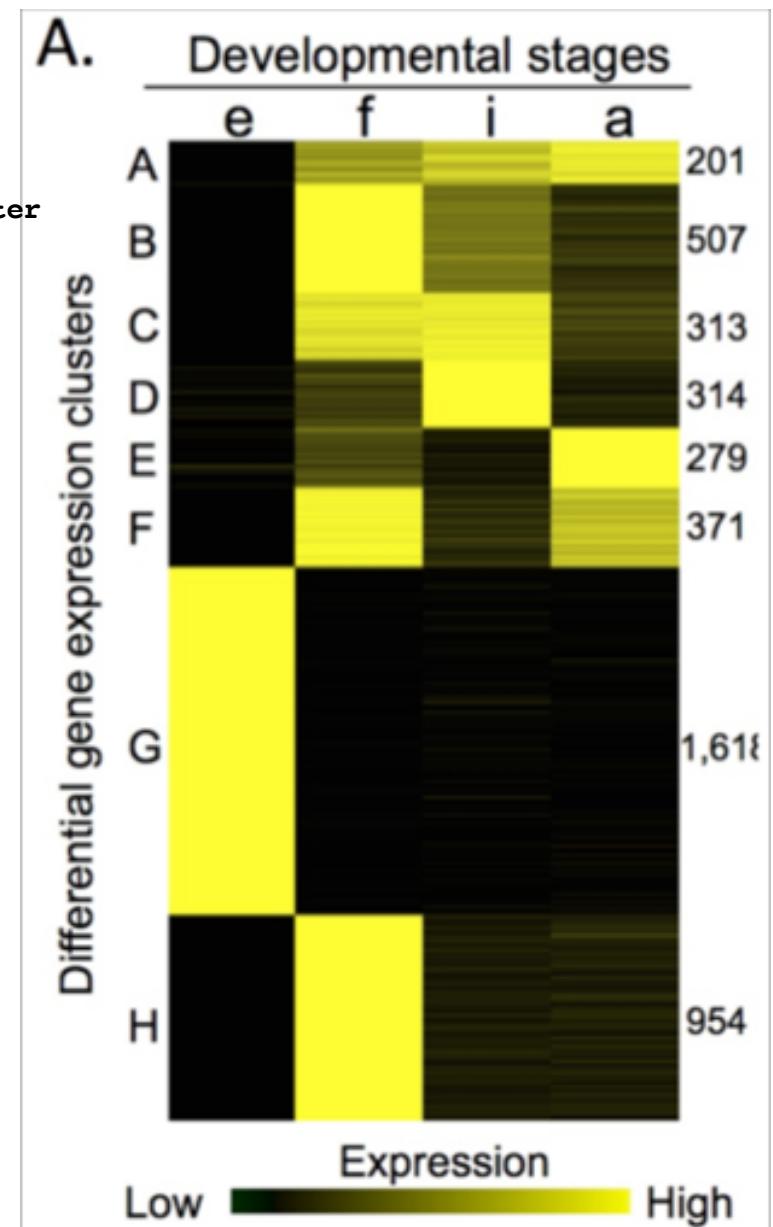
To determine the set of differentially expressed genes, transcripts between two conditions or time points in a time-course experiment.

- **Approach:**

1. Determine gene, transcript abundances.
2. Test for differentially expressed genes and transcripts e.g. test for statistical significance of differences between the expressions of genes from different conditions, time points.

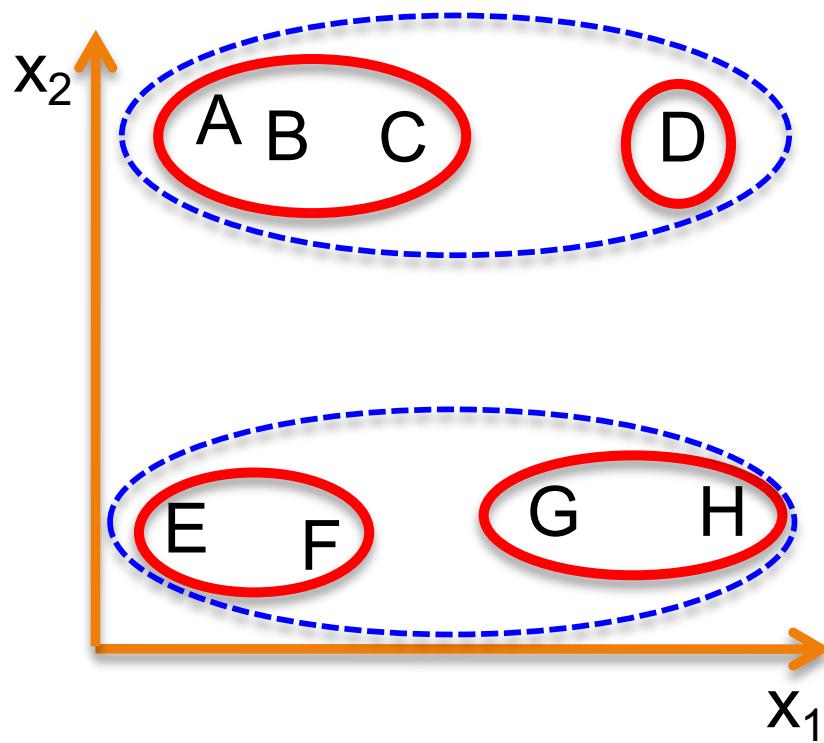
Visualizing differential expression

carpoID	emb	L1	IJ	adult	cluster
L596_g9169	0.0	63.4	291.97	350.93	A
L596_g9943	1.73	66.395	57.695	71.27	A
L596_g23550	3.365	90.53	114.275	88.255	A
L596_g14277	5.515	199.63	388.89	403.54	A
L596_g16663	0.0	11.02	33.11	25.85	A
L596_g12422	4.835	665.535	492.87	463.355	A
L596_g19270	0.765	66.185	66.815	81.23	A
L596_g25810	0.83	47.095	71.08	87.015	A
L596_g17081	1.775	263.21	252.68	399.955	A
L596_g9668	0.065	6.125	8.415	6.34	A
L596_g13068	0.0	14.755	20.515	14.225	A
L596_g26347	11.95	0.63	14.49	19.045	A
L596_g27125	0.0	32.415	22.06	24.805	A
L596_g1905	1.125	44.54	27.73	62.905	A
L596_g6738	1.355	107.825	95.11	91.515	A
L596_g25011	2.6	76.515	51.075	99.53	A
L596_g25018	0.21	13.01	21.8	36.64	A
L596_g27623	1.335	45.795	38.37	61.055	A
L596_g24065	0.055	18.81	37.135	29.335	A
L596_g24064	0.0	23.8	171.315	121.575	A
L596_g10859	2.765	121.065	109.31	147.47	A



We can visualize major groups of genes using clustering

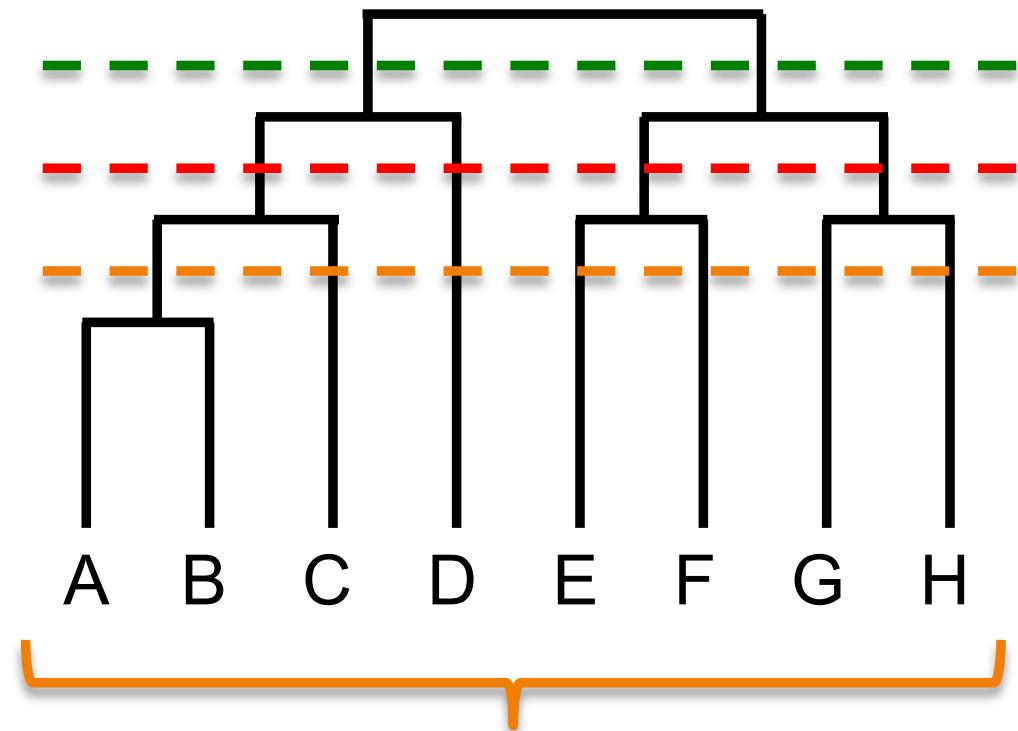
Clustering: Hierarchical & K-means



K-means clustering

K=2 ABCD EFGH

K=4 ABC D EF GH



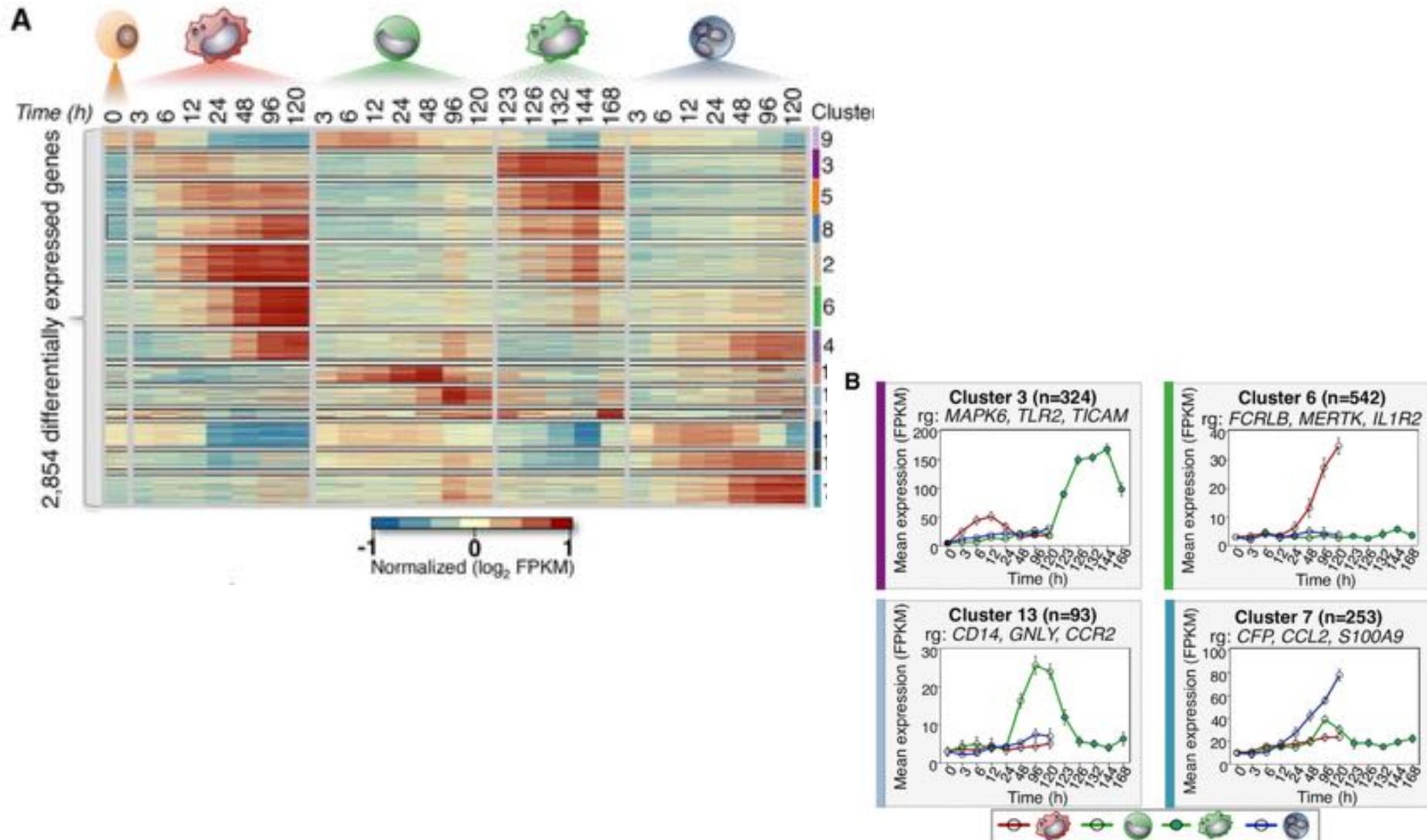
Hierarchical clustering

AB C D E F G H

ABC D EF GH

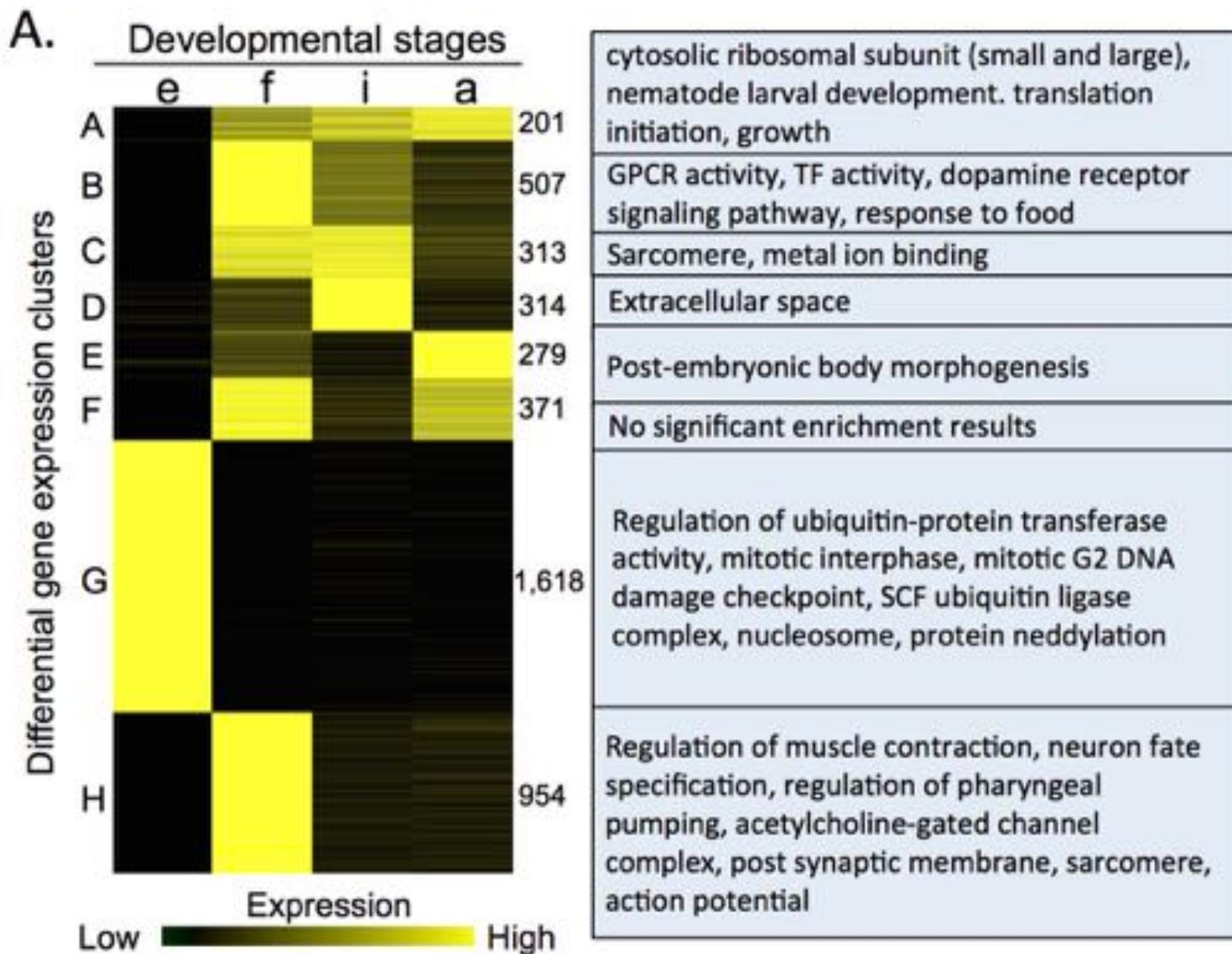
ABCD EFGH

Heatmaps of gene expression



(Ramirez, 2017)

Going from differential expression to function



Using sequence to derive function

Assume that a gene is differentially expressed

carpolD	emb	L1	IJ	adult	DE_cluster
L596_g19086	451.88	0.82	0.0	0.32	G
L596_g28254	294.365	1.125	0.345	0.12	G
L596_g22582	316.345	0.54	0.04	0.805	G

And that it's sequence is:

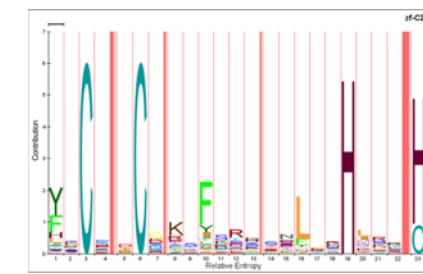
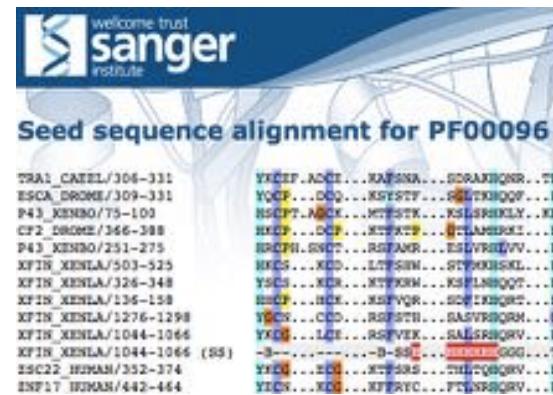
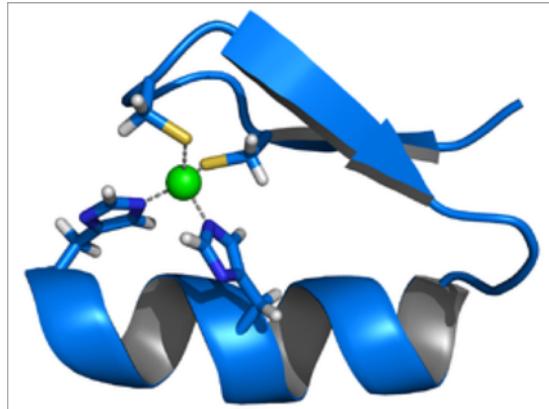
```
>L596_g19086.t1
CCACCTCTACCTACCCAGAAATCCAAAACCCTCCACCATGTACATTCCATCTGTGAGC
ATAAATTCAACCGGAACCTCCCTCTGCTACGGCGACATCGACGGTTACTTGTCTACAGC
CTCAACGACATCAAGCAGAAGAATTCTGTGAGAAAATTGATCAATGGGATCCGTCTACA
GAGCTCTACTTGGCCGAACGTCCTAACACTTCTAAACTTCTGCTGTGCTGTTCCCACAAG
GACAGTCGTGAGGTTCTGATGATCGGCTTAAAGGCTAACGAGGTCTATTGAAGCTCGAC
ATGCCCTCGGATGTTCTAGTGTAGGATTAACAATGAGAGCCTAATCATCTGCATGCAG
CACATGATCCGAATCTTCATCCCCCAGACAAATGCAGGCCATCCACACCATCAACGACATC
CCTTCAAACCTAACGGCGCTGCTGACCTCTCCAAACCCACCTCTATGATCGCATATCCA
ACTTTCGAAAACGGCGGAAAAGTGGTGATATACGACGGAAAGAACCTGAAAGCCGTTGG
GTCATTGACGCCATGATGGACCCGTTGTGGTTCTGAAATTCAACAAGCAAGGAACCCCTC
ATAGCTACTGCCTCAGACAAGGGTACTGTAATCAGAGTCTTCAGCGTGGAAACGGCGTT
CTTGTCCATGAATTCAACGAGGGCGTACACGATTGCGACGATTACTCTATCGCATT
TCTGAAGATTCTCAGTACCTCGCCTGCACAGTAATACGGGTACCGTACACCTCTTCCAT
TTGTCTCCACGTGAAAATGAGCCATGCTTCCCTAAAGATAGTAATCCTATCAGTGACCTC
GTCAGCTATTGTGGAAAAGCGCTGAGGCCTACACACCTGCCGTGGTGAGACCCAAATCT
ACGTCCCTGTGTGCCAGTCGGGCTAACAGTTCGCAGCCTGTGCGCTGAGGATT
ATGAATAACAGGCTGCATTGATTGTGGCTACTCGGAGAAGTATTGTTGTTATGAG
TTGACCCCTATAAGTTAGAGCTGAGTTGAAGAGTCAGTTCGAGTTGGTAGGGAGGAA
GAAGAGAAGGGAGTTGGTTGATGAACTCGTTGAAGGTTGGATAGAAGTGTGTTGAC
TATAAAGATTGTTAAGAGAGATTAAAGTGTGAAAAACACAACCTTAAGGTTAATACATA
AAAGGCTGCAA
```

What is its
Function?

Annotate protein domains from sequence

- Proteins typically have functional domains, such as
 - DNA binding
 - Transmembrane
- We can use profile-HMMs of these domains to annotate novel genes using a combination of :
 - HMMer
 - The Pfam database of protein families

A famous protein domain – C2H2 Zinc fingers



Using PFAM to annotate protein domains



HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

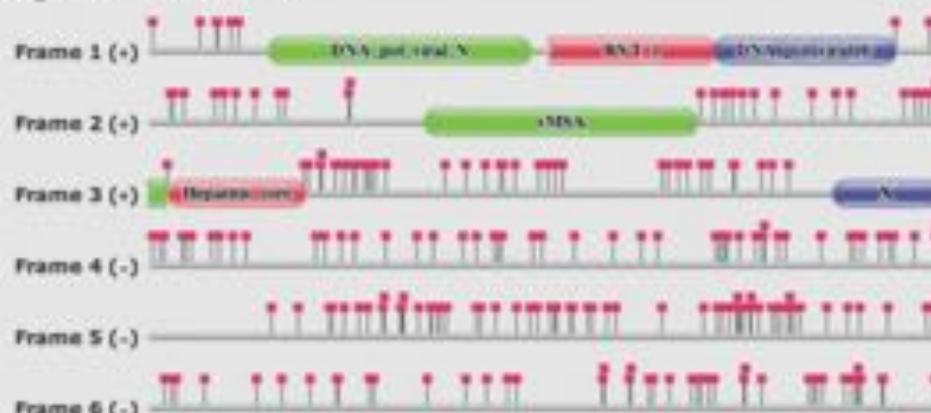


DNA sequence search results

This page shows the results of searching your DNA sequence for Pfam-A matches. To do this we perform a six-frame translation to generate a set of protein sequences, which we then search using the normal Pfam-A HMMs and GA cut-offs.

Show the detailed description of this results page.

We have found 7 significant hits and 0 insignificant hits in 3 frames.



Show the DNA and protein sequences, and the URL for bookmarking these results.

Return to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

Show or hide all alignments. Toogle between amino-acid and DNA sequence coordinates.

Frame (sense)	Family	Description	Entry type	Class	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
					Start	End	Start	End	From	To					
1 (+)	DNA_pol_viral_N	DNA polymerase (viral) N-terminal domain	Family	n/a	163	514	163	514	1	379	379	573.7	1.7e-172	n/a	Show
1 (+)	RT_1	Reverse transcriptase (RNA-dependent DNA)	Family	CL0022	538	762	539	762	2	214	214	206.6	2.7e-61	n/a	Show
1 (+)	DNA_pol_viral_C	DNA polymerase (viral) C-terminal domain	Family	n/a	763	1005	763	1005	1	245	245	477.2	8.1e-144	n/a	Show
2 (+)	xMSA	Major surface antigen from hepadnavirus	Family	n/a	372	737	372	737	1	364	364	563.4	2.2e-169	n/a	Show
3 (+)	Hep_core_N	Hepatitis core protein, putative zinc fi ...	Domain	n/a	1	24	1	24	4	27	27	58.7	2.5e-16	n/a	Show
3 (+)	Hepatits_core	Hepatitis core antigen	Domain	n/a	28	209	28	209	1	187	187	315.3	1.1e-94	n/a	Show
3 (+)	X	Trans-activation protein X	Family	n/a	923	1064	923	1064	1	142	142	268.1	1.4e-80	n/a	Show

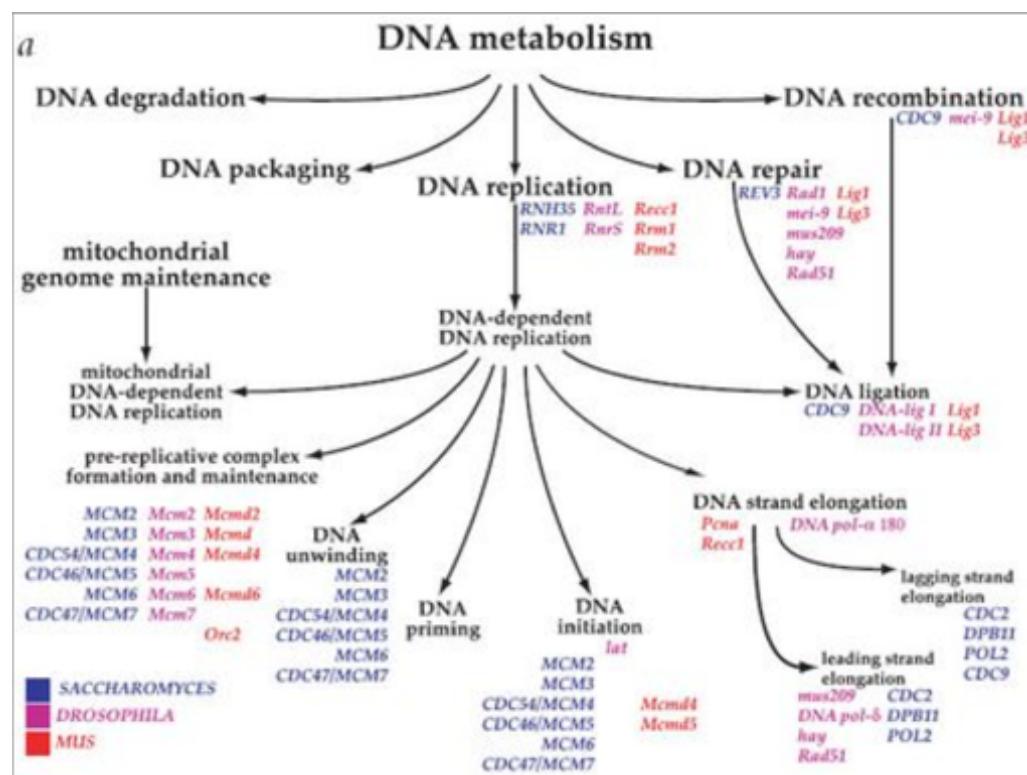
Comments or questions on the site? Send a mail to pfam-help@sanger.ac.uk. Our [privacy policy](#).

The Wellcome Trust

(Finn, 2014)

Annotate transcript function

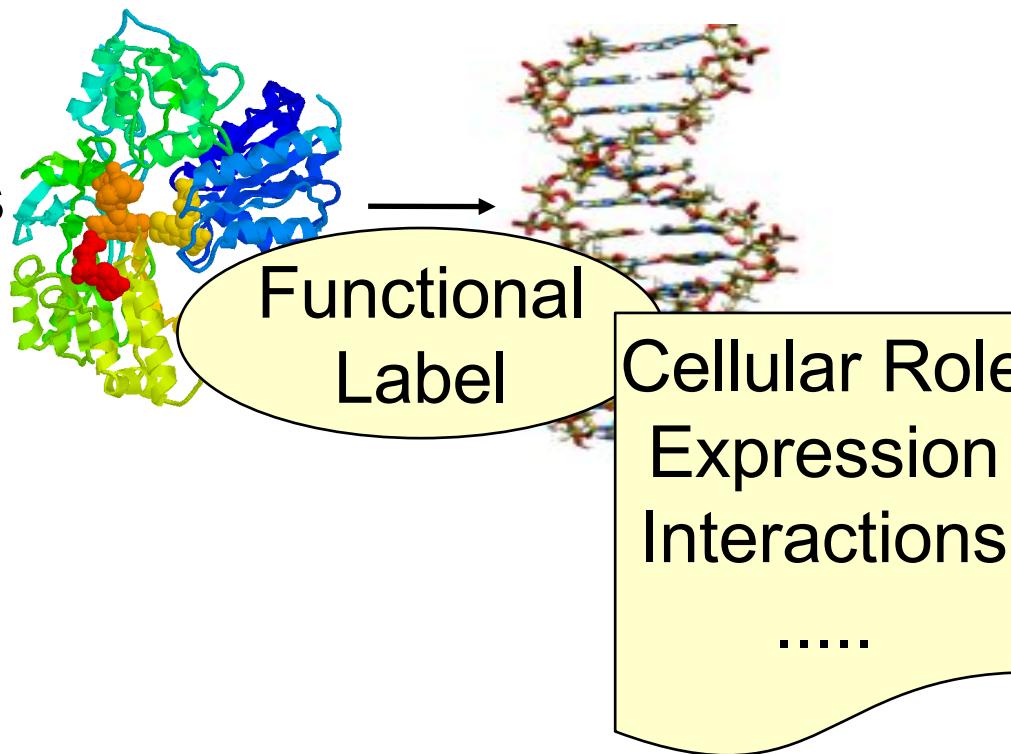
- Once we know functional domains and/or orthologs, what can we say about a gene in relation to others
- To analyze a set of genes, a standardized structured vocabulary is helpful. Why would that be ?



(Ashburner, 2000)

To analyze gene enrichments, must standardize on vocabulary

The function is
on the protein



Controlled
Vocabulary

High
throughput

Accessible

Slides courtesy of Ana Conesa

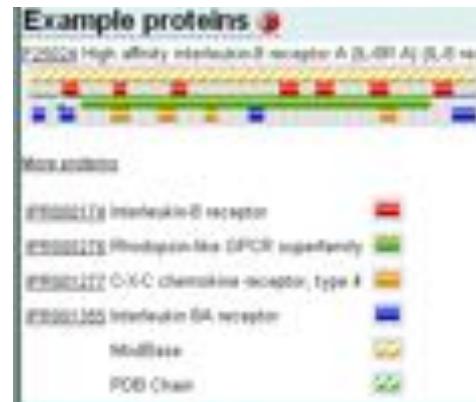
Several databases are designed for data-mining



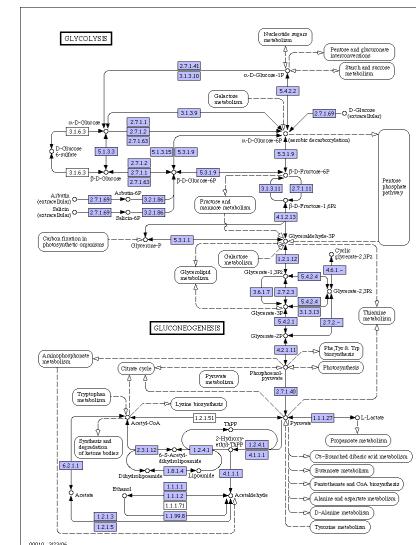
Molecular Function
Biological Process
Cellular Component



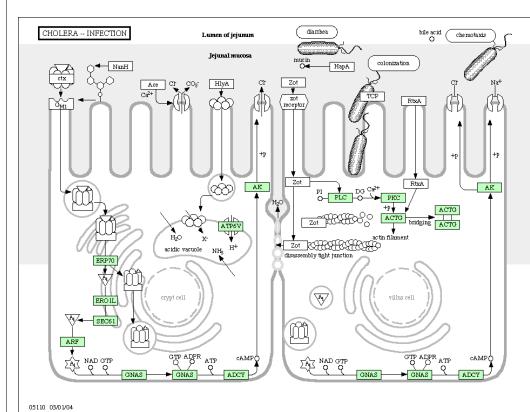
Functional motifs



Metabolic pathways



KEGG orthologues



Slides courtesy of Ana Conesa

Gene Ontology (GO)

- ✓ Project developed by the Gene Ontology Consortium
- ✓ Provides a controlled vocabulary to describe gene and gene product attributes in any organism
- ✓ Latest version more than 33,300 terms
- ✓ Includes both the development of the Ontology and the maintenance of a Database of annotations

<http://www.geneontology.org/>

Slides courtesy of Ana Conesa

The three categories of GO Molecular Function

the tasks performed by individual gene products; examples are *transcription factor* and *DNA helicase*

Biological Process

broad biological goals, such as *mitosis* or *purine metabolism*, that are accomplished by ordered assemblies of molecular functions

Cellular Component

subcellular structures, locations, and macromolecular complexes; examples include *nucleus*, *telomere*, and *origin recognition complex*

- [GO:0003673 : Gene Ontology \(65883\)](#) ●
 - [② GO:0008150 : biological process \(44405\)](#) ●
 - + [① GO:0007610 : behavior \(357\)](#)
 - [① GO:0000004 : biological process unknown \(7877\)](#)
 - [① GO:0009987 : cellular process \(32672\)](#) ●
 - + [① GO:0007154 : cell communication \(5384\)](#)
 - + [① GO:0008219 : cell death \(744\)](#)
 - + [① GO:0030154 : cell differentiation \(464\)](#)
 - + [① GO:0008151 : cell growth and/or maintenance \(28802\)](#)
 - + [① GO:0006928 : cell motility \(911\)](#)
 - + [① GO:0006944 : membrane fusion \(257\)](#)
 - + [① GO:0016265 : death \(793\)](#)
 - + [① GO:0007275 : development \(4615\)](#)
 - + [① GO:0008371 : obsolete \(1581\)](#)
 - + [① GO:0007582 : physiological processes \(31124\)](#)
 - + [① GO:0016032 : viral life cycle \(115\)](#)
 - + [② GO:0005575 : cellular component \(32869\)](#)
 - + [② GO:0003674 : molecular function \(53910\)](#)

The GO hierarchy is a directed graph

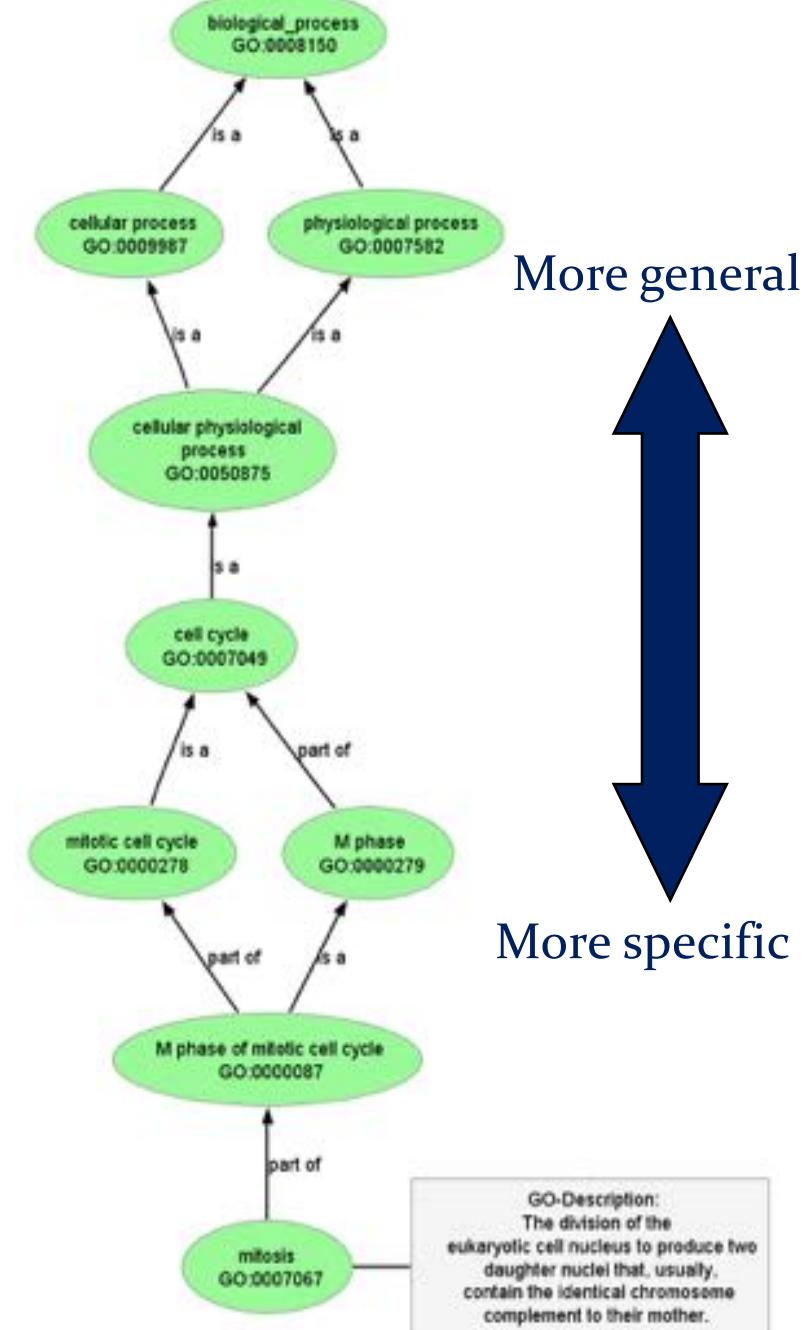


Slides courtesy of Ana Conesa

The GO hierarchical levels

- ✓ Annotations are given to the **most specific** (low) level
- ✓ True path rule: annotation at a given term implies **annotation to all its parent terms**
- ✓ Annotation is given with an **Evidence Code:**
 - **IDA:** inferred by direct assay
 - **TAS:** traceable author statement
 - **ISS:** inferred by sequence similarity
 - **IEA:** electronic annotation
 -

Slides courtesy of Ana Conesa



GO example: NRSF/REST (part 1)

Gene Ontology (GO): 11 molecular function terms (see first 5): [About this table](#)

GO ID	Qualified GO term	Evidence	PubMed IDs
GO:0001046	core promoter sequence-specific DNA binding	IDA	8568247
GO:0001047	core promoter binding	IDA	17984088
GO:0001078	RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in negative regulation of transcription	IDA	10449787
GO:0003677	DNA binding	--	--
GO:0003682	chromatin binding	ISS	--
GO:0003700	sequence-specific DNA binding transcription factor activity	IDA	19342457
GO:0005515	protein binding	IPI	10449787
GO:0008134	transcription factor binding	IPI	17130167
GO:0015271	outward rectifier potassium channel activity	IMP	18570921
GO:0044212	transcription regulatory region DNA binding	IDA	17555596
GO:0046872	metal ion binding	IEA	--

Gene Ontology (GO): 4 cellular component terms: [About this table](#)

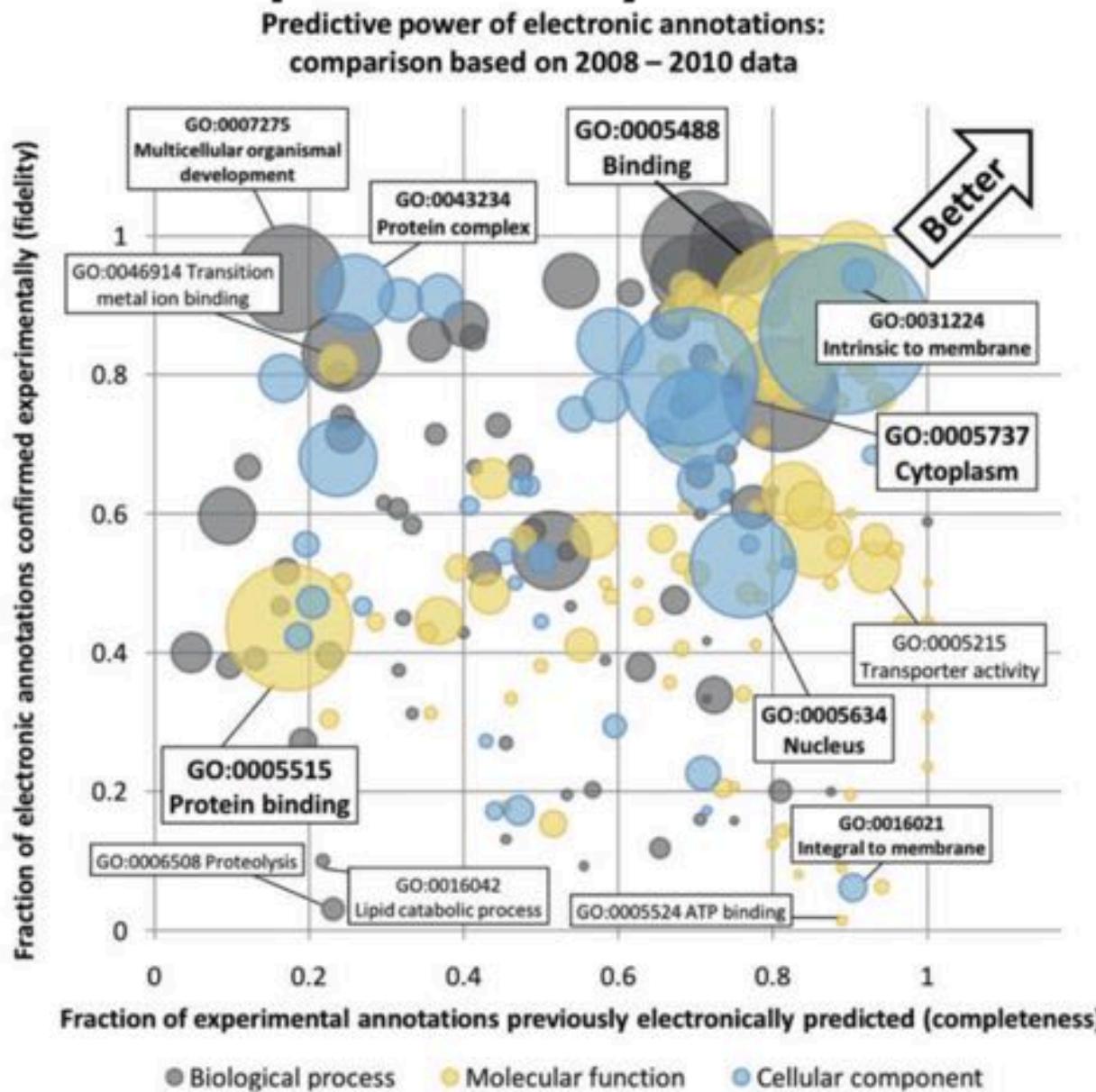
GO ID	Qualified GO term	Evidence	PubMed IDs
GO:0005634	nucleus	NAS	7871435
GO:0005737	cytoplasm	--	--
GO:0005829	cytosol	IEA	--
GO:0017053	transcriptional repressor complex	IDA	10734093

GO example: NRSF/REST (part 2)

Gene Ontology (GO): 24 biological process terms (see first 5): [About this table](#)

GO ID	Qualified GO term	Evidence	PubMed IDs
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	TAS	7697725
GO:0006355	regulation of transcription, DNA-templated	NAS	7871435
GO:0008285	negative regulation of cell proliferation	IMP	--
GO:0010629	negative regulation of gene expression	IMP	--
GO:0032348	negative regulation of aldosterone biosynthetic process	IMP	19342457
GO:0035690	cellular response to drug	IMP	--
GO:0043065	positive regulation of apoptotic process	IMP	--
GO:0043280	positive regulation of cysteine-type endopeptidase activity involved in apoptotic process	IMP	--
GO:0043922	negative regulation by host of viral transcription	IDA	17555596
GO:0045665	negative regulation of neuron differentiation	IMP	18570921
GO:0045892	negative regulation of transcription, DNA-templated	NAS	7871435
GO:0045893	positive regulation of transcription, DNA-templated	IDA	17984088
GO:0045955	negative regulation of calcium ion-dependent exocytosis	ISS	--
GO:0046676	negative regulation of insulin secretion	IMP	--
GO:0050768	negative regulation of neurogenesis	ISS	--
GO:0060379	cardiac muscle cell myoblast differentiation	ISS	--
GO:0070933	histone H4 deacetylation	IDA	17555596
GO:0071257	cellular response to electrical stimulus	IMP	18570921
GO:0071385	cellular response to glucocorticoid stimulus	IDA	17984088
GO:0071805	potassium ion transmembrane transport	IMP	18570921
GO:2000065	negative regulation of cortisol biosynthetic process	IMP	19342457
GO:2000706	negative regulation of dense core granule biogenesis	ISS	--
GO:2000740	negative regulation of mesenchymal stem cell differentiation	IMP	18570921
GO:2000798	negative regulation of amniotic stem cell differentiation	IMP	--

Some computationally predicted GO terms are more likely to be experimentally validated



(du Plessis, 2011)

Applications of GO

Can be applied to analysis of samples from annotated genomes or unannotated genomes

Experimental design

a. Gene of interest known

- Find genes with related functions
- Find genes with same cellular location
- Find genes that interact with each other

b. Function/process of interest known

- Generate list of candidate genes for that function/process

Postexperiment data analyses

a. Genome/transcriptome annotation

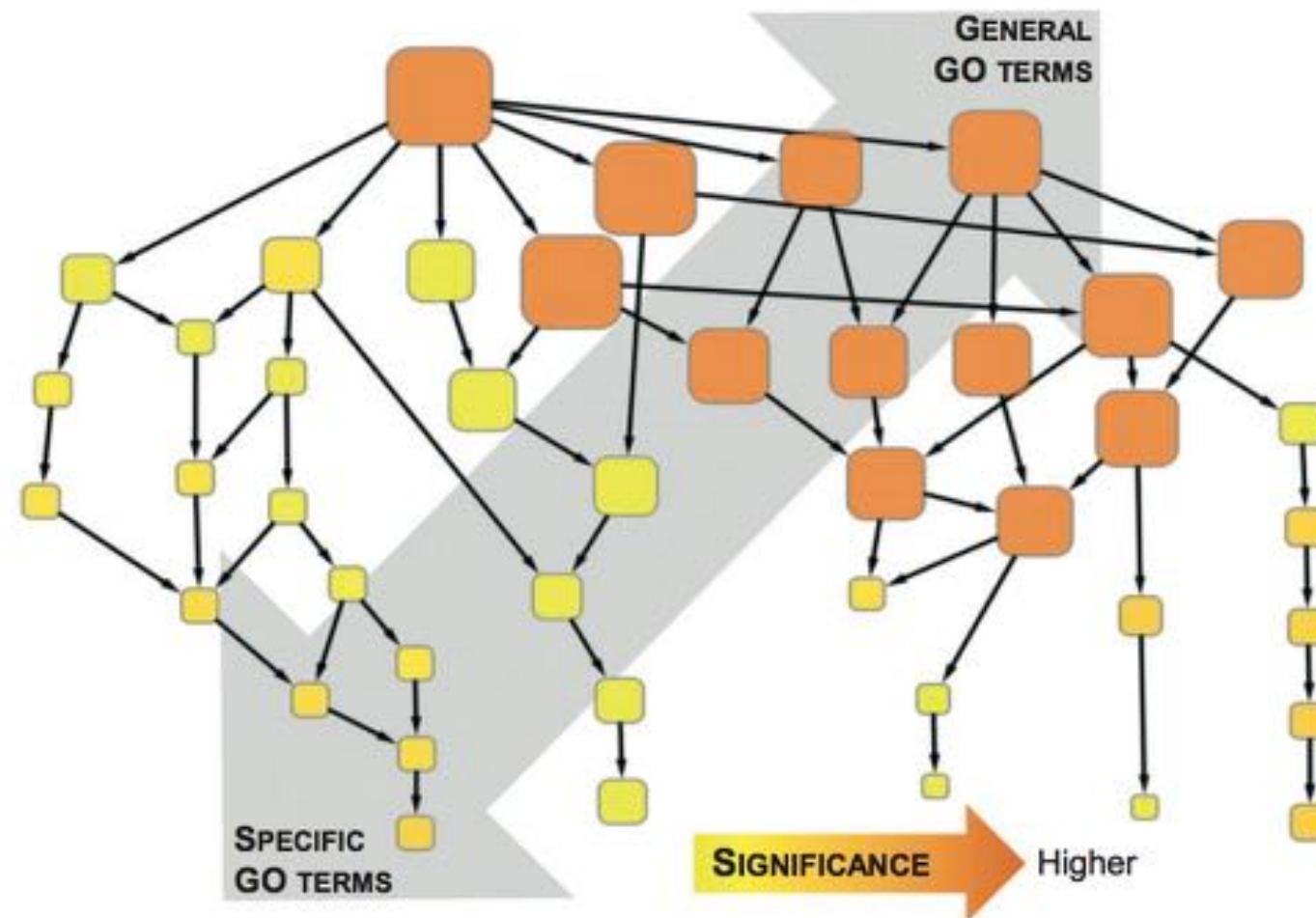
- Characterize a major proportion of gene functions
- Identify classes of genes that are more or less frequent than in related organisms
- Infer pleiotropy at the molecular level

b. Functional enrichment

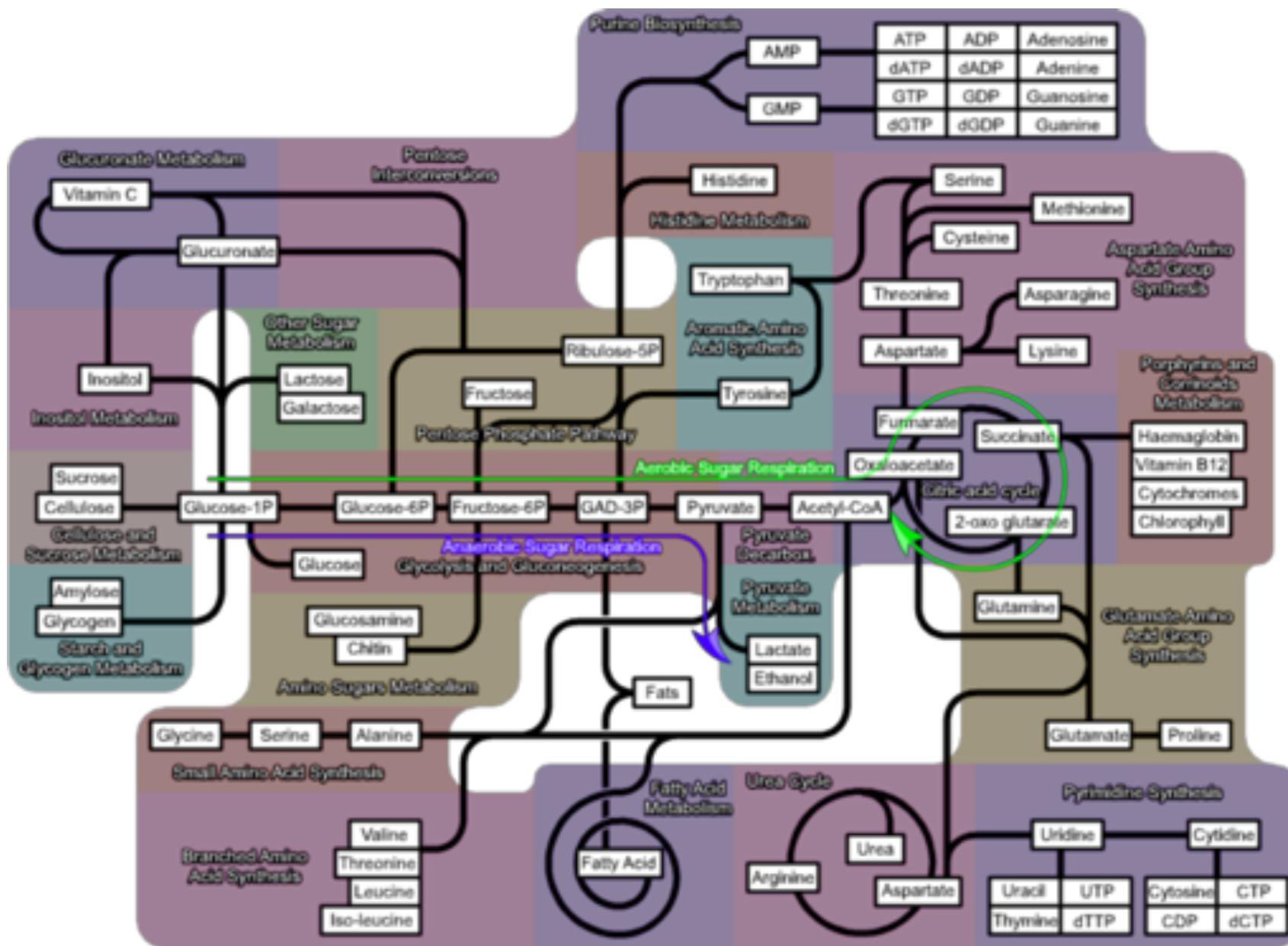
- Between individuals with different
 - Life histories
 - Development stages
 - Experimental treatments
 - Environments
- Between genes with different evolutionary patterns
 - Positive selection vs. neutral
 - Duplicated vs. not



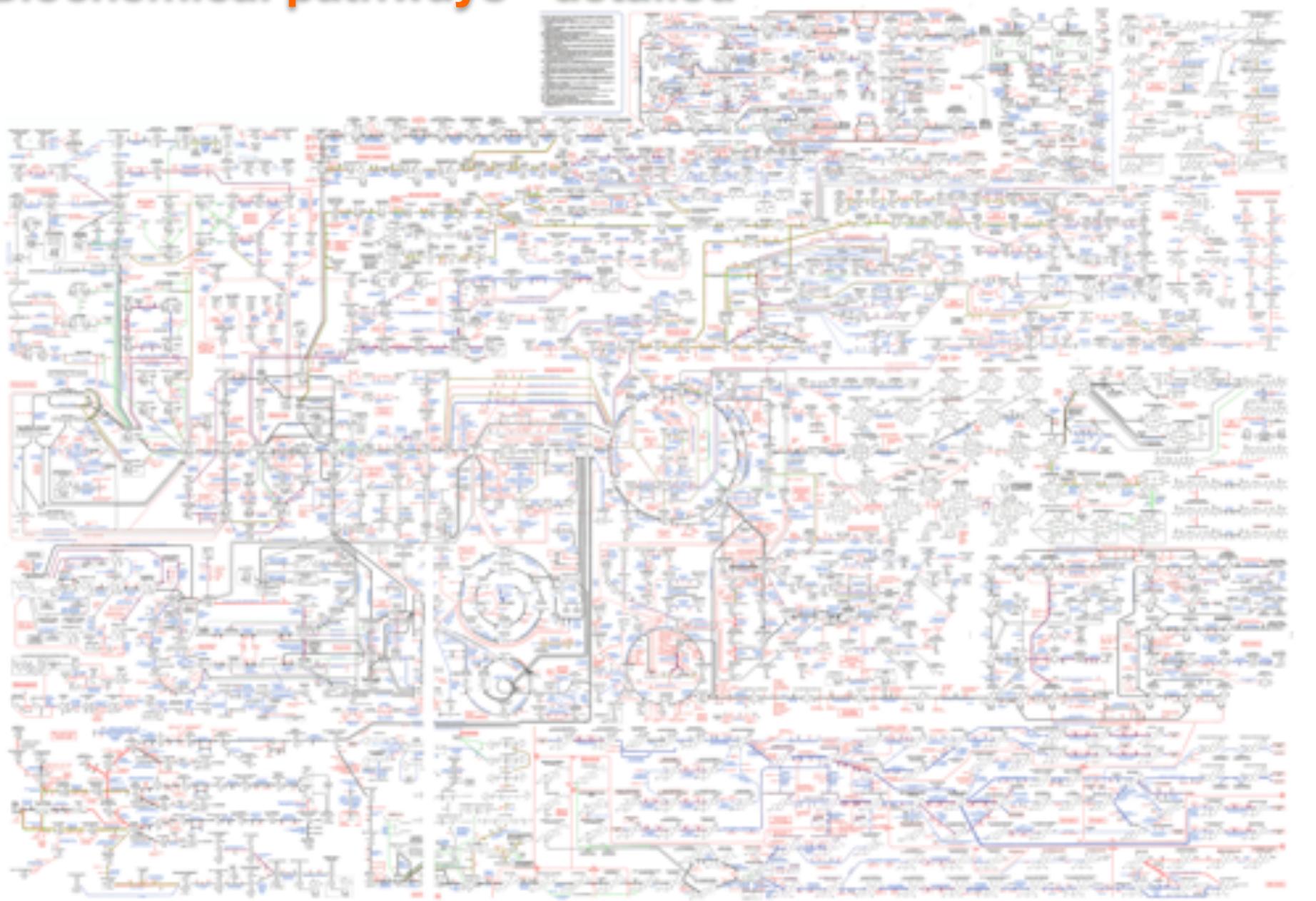
Enrichment analyses of GO terms favors more general terms



Biochemical pathways - simplified



Biochemical pathways - detailed



Known biochemical networks are catalogued by KEGG



1. Metabolism

1.1 Carbohydrate Metabolism

Glycolysis / Gluconeogenesis
Citrate cycle (TCA cycle)
Pentose phosphate pathway
Pentose and glucuronate interconversions
Fructose and mannose metabolism
Galactose metabolism
Ascorbate and aldarate metabolism
Starch and sucrose metabolism
Aminosugars metabolism
Nucleotide sugars metabolism
Pyruvate metabolism
Glyoxylate and dicarboxylate metabolism
Propionate metabolism
Butanoate metabolism
C5-Branched dibasic acid metabolism
Inositol metabolism
Inositol phosphate metabolism

1.2 Energy Metabolism

Oxidative phosphorylation
Photosynthesis
Photosynthesis - antenna proteins
Carbon fixation
Reductive carboxylate cycle (CO₂ fixation)
Methane metabolism
Nitrogen metabolism
Sulfur metabolism

1.3 Lipid Metabolism

Fatty acid biosynthesis
Fatty acid elongation in mitochondria
Fatty acid metabolism
Synthesis and degradation of ketone bodies
Biosynthesis of steroids
DHA acid biosynthesis

KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks for:

1. Metabolism

Global/overview Carbohydrate Energy Lipid Nucleotide Amino acid Other amino Glycan Cofactor/vitamin Terpenoid/PK Other secondary metabolite Xenobiotics Chemical structure

2. Genetic Information Processing

3. Environmental Information Processing

4. Cellular Processes

5. Organismal Systems

6. Human Diseases

and also on the structure relationships (KEGG drug structure maps) in:

7. Drug Development

Current Statistics

KEGG Database as of 2015/7/17

Systems information

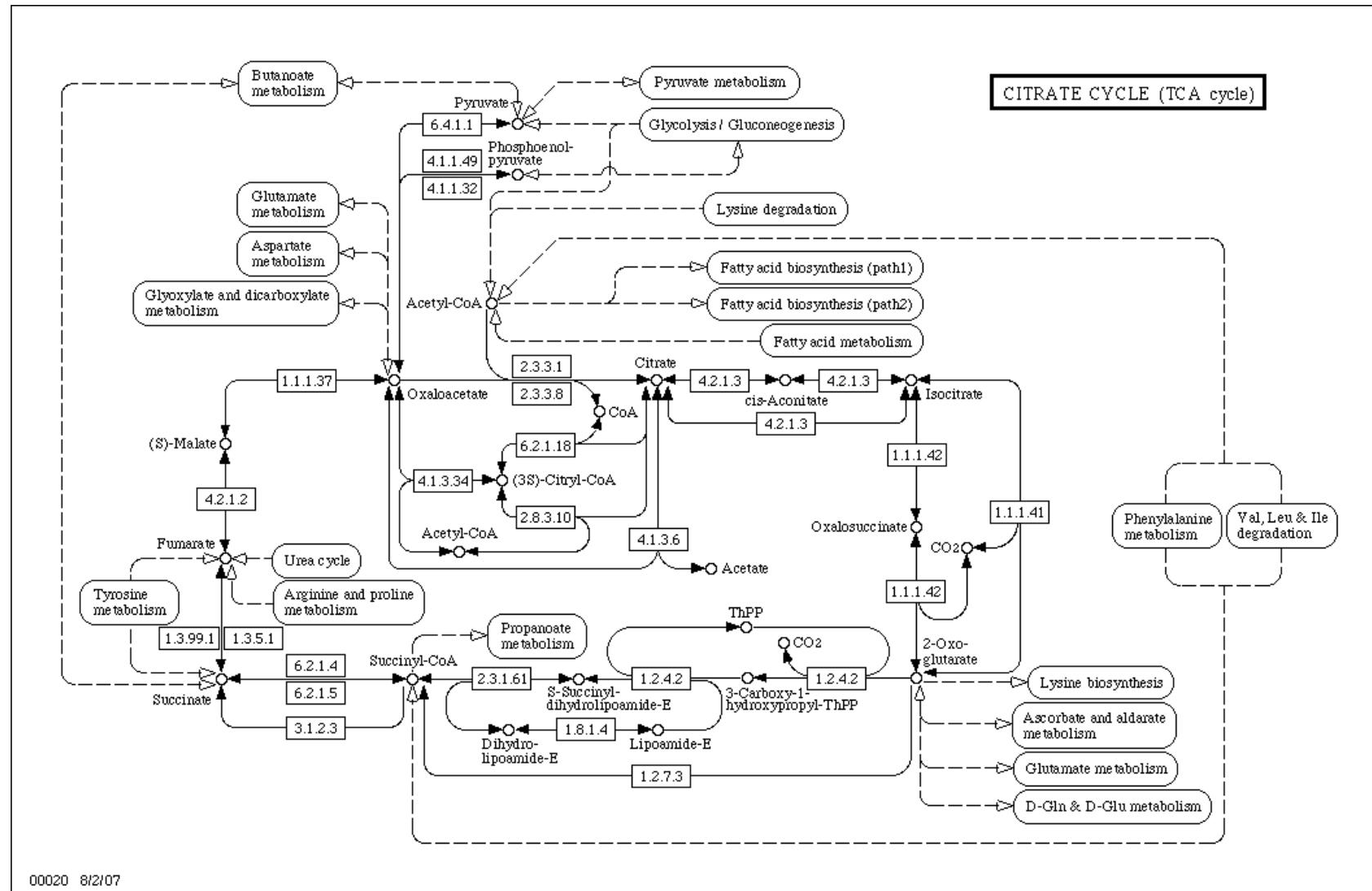
KEGG PATHWAY	Pathway maps, reference (total)	475 (402,444)
KEGG BRITE	Functional hierarchies, reference (total)	205 (138,436)
KEGG MODULE	KEGG modules, reference (total)	708 (325,410)

Genomic information

KEGG ORTHOLOGY	KEGG Orthology (KO) groups	18,879
KEGG GENOME	KEGG Organisms	3,985
KEGG GENES	Genes in high-quality genomes (312 eukaryotes, 3445 bacteria, 211 archaea)	17,556,548

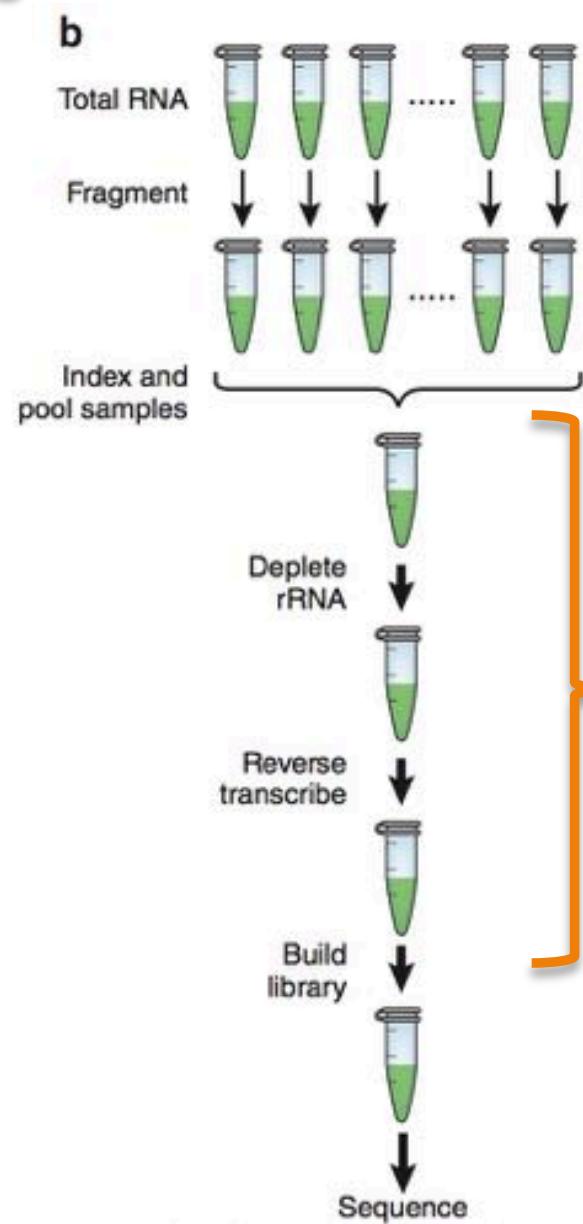
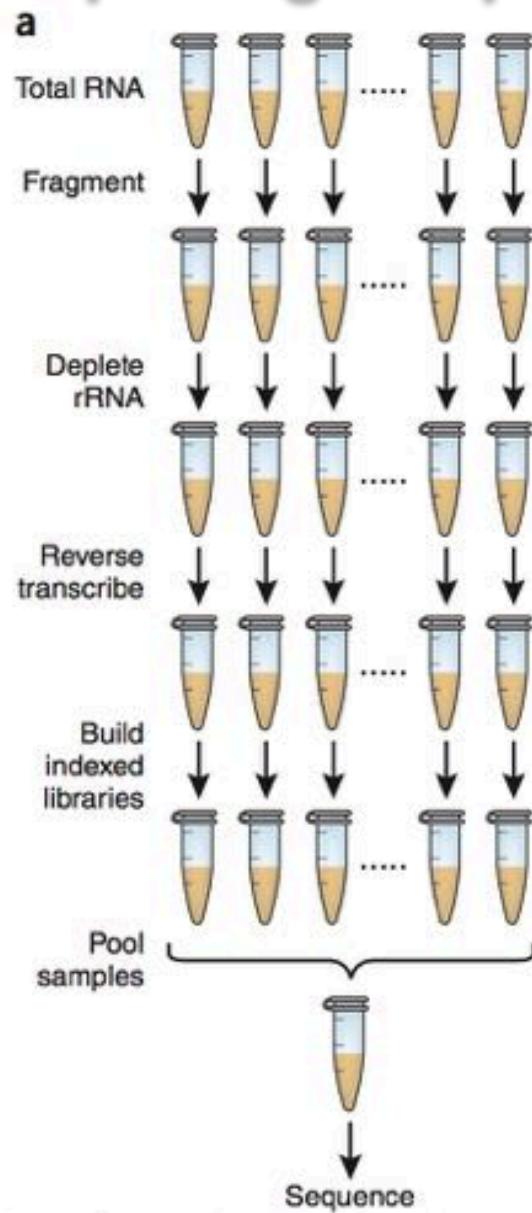
<http://www.genome.jp/kegg/>

A well-known biochemical pathway: the TCA cycle



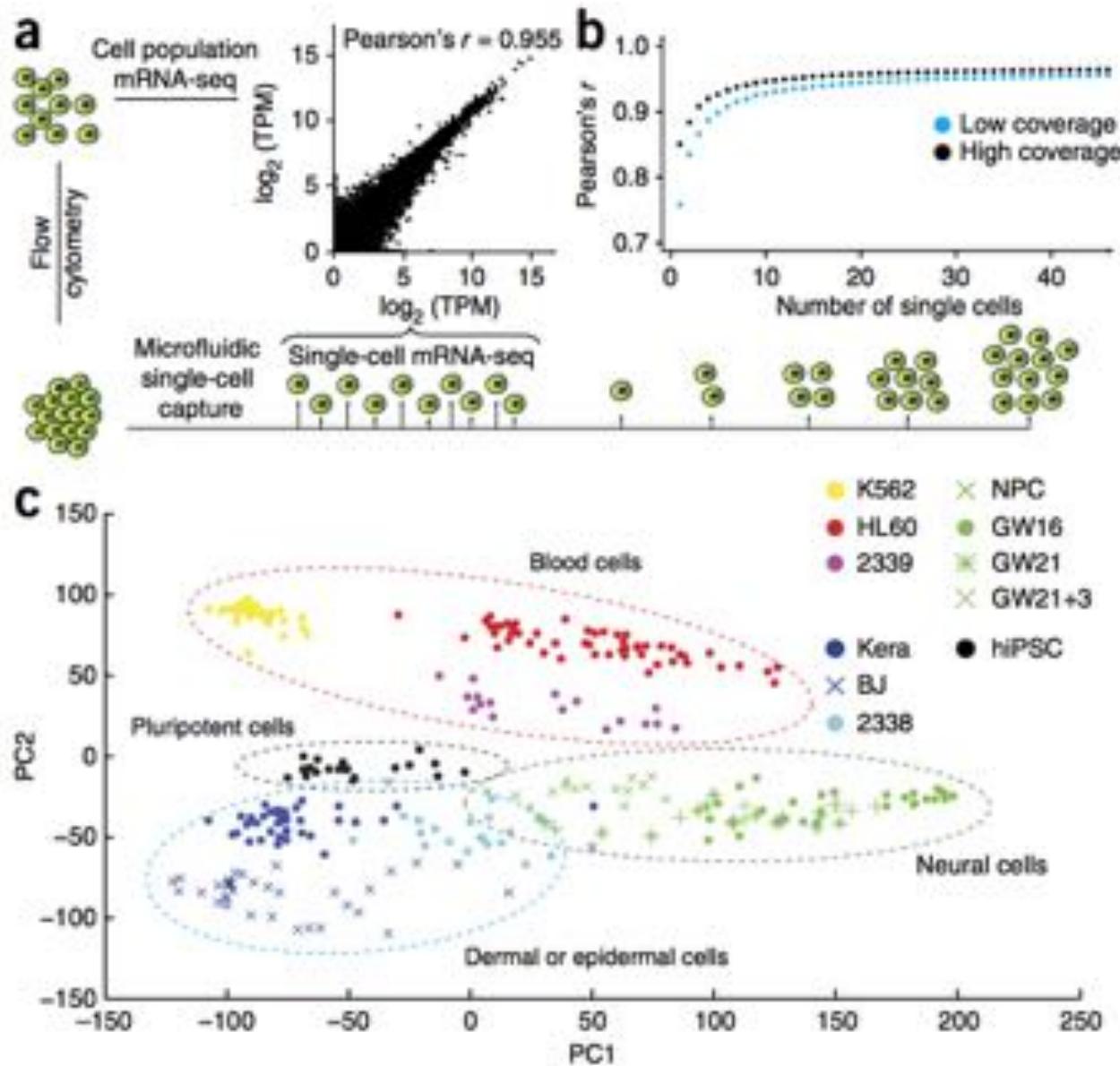
Are several of these genes differentially expressed ?

Multiplexing samples



Less
sample-to-sample
technical variability

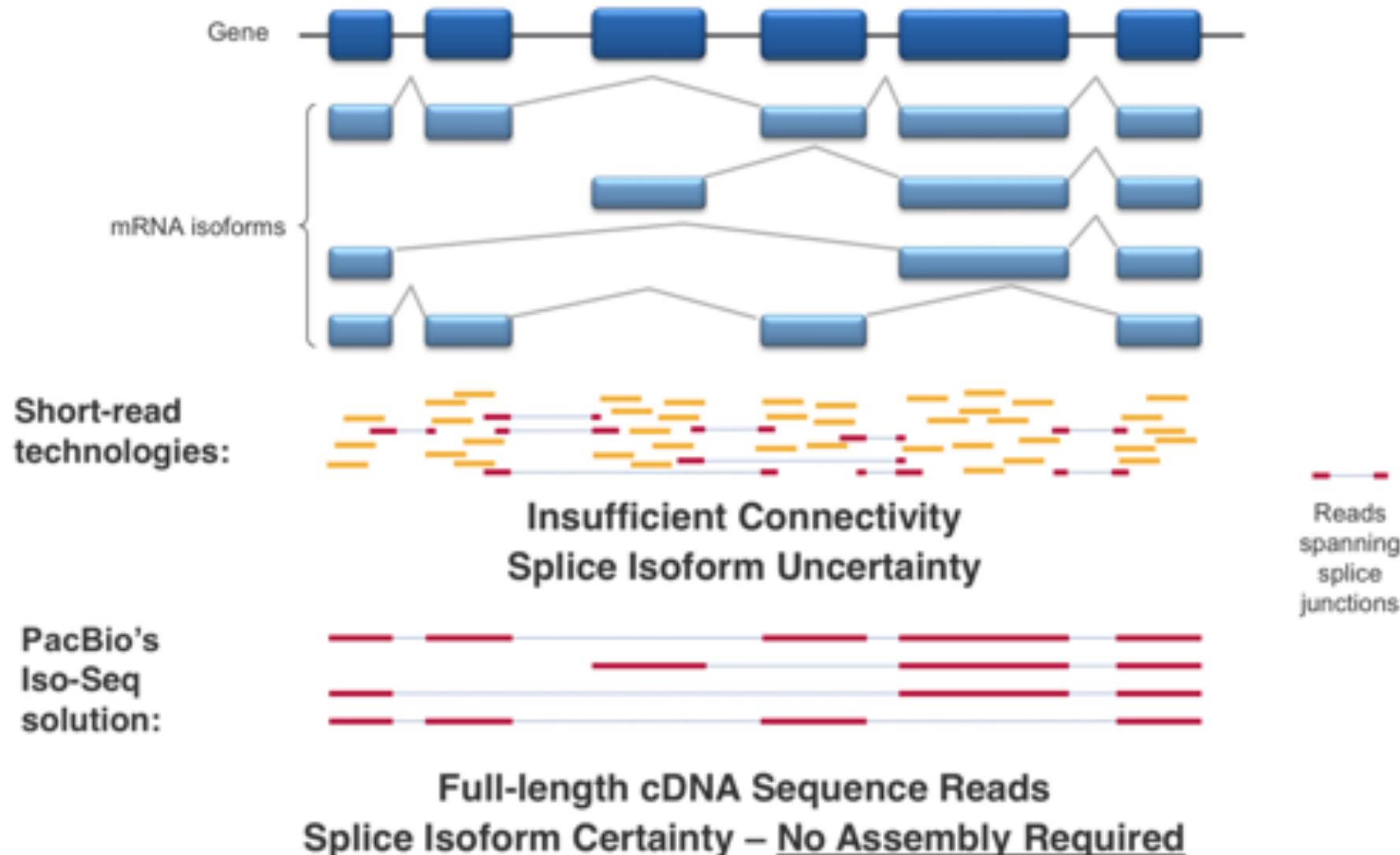
Single-cell RNA-seq captures population behavior



(Pollen, 2014)

Using long-reads to sequence entire mRNAs end-to-end

DETERMINATION OF TRANSCRIPT ISOFORMS



Online videos

RNA-seq:

<https://www.youtube.com/watch?v=tlf6wYJrwKY>

RPKM vs FPKM vs TPM:

<https://www.youtube.com/watch?v=TTUrtCY2k-w>

Gene Ontology:

<https://www.youtube.com/watch?v=3EUaurjK7u8>