Lecture 12. RNAseq, Methyl-seq, & ChipSeq

Michael Schatz

March 8, 2017

JHU 601.749: Applied Comparative Genomics

Assignment 5: Due March 8

Assignment 5: Genome Arithmetic

Assignment Date: Thursday, March 1, 2018
Due Date: Thursday, March 8, 2018 @ 11:59pm

Assignment Overview

In this assignment, you will call structural variants and analyze the properties of variants in the human genome. **Make sure to show your work in your writeup!** As before, any questions about the assignment should be posted to Piazza.

Question 1. Gene Annotation Preliminaries [10 pts]

Download the annotation of build 38 of the human genome from here:
ftp://ftp.ensembl.org/pub/release-87/gtf/homo_sapiens/Homo_sapiens.GRCh38.87.gtf.gz

- Question 1a. How many many GTF data lines are in this file? [Hint: The first few lines in the file beginning with "#" are so-called "header" lines describing thing like the creation date, the genome version (more on that later in the course), etc. Header lines should not be counted as data lines.]
- Question 1b. How many annotated protein coding genes are on each autosome of the human genome? [Hint: Protein coding genes will have "gene" in the 3rd column, and contain the following text: gene_biotype "protein_coding"]
- Question 1c. What is the maximum, minimum, mean, and standard deviation of the span of protein coding genes? [Hint: use the genes identified in 1b]
- Question 1d. What is the maximum, minimum, mean, and standard deviation in the number of exons for protein coding genes? [Hint: you should separately consider each isoform for each protein coding gene]

Project Proposal!
Due March 15

Project Proposal

Assignment Date: March 7, 2018
Due Date: Thursday, March 15, 2017 @ 11:59pm

Review the Project Ideas page

Work solo or form a team for your class project (no more than 3 people to a team).

The proposal should have the following components:

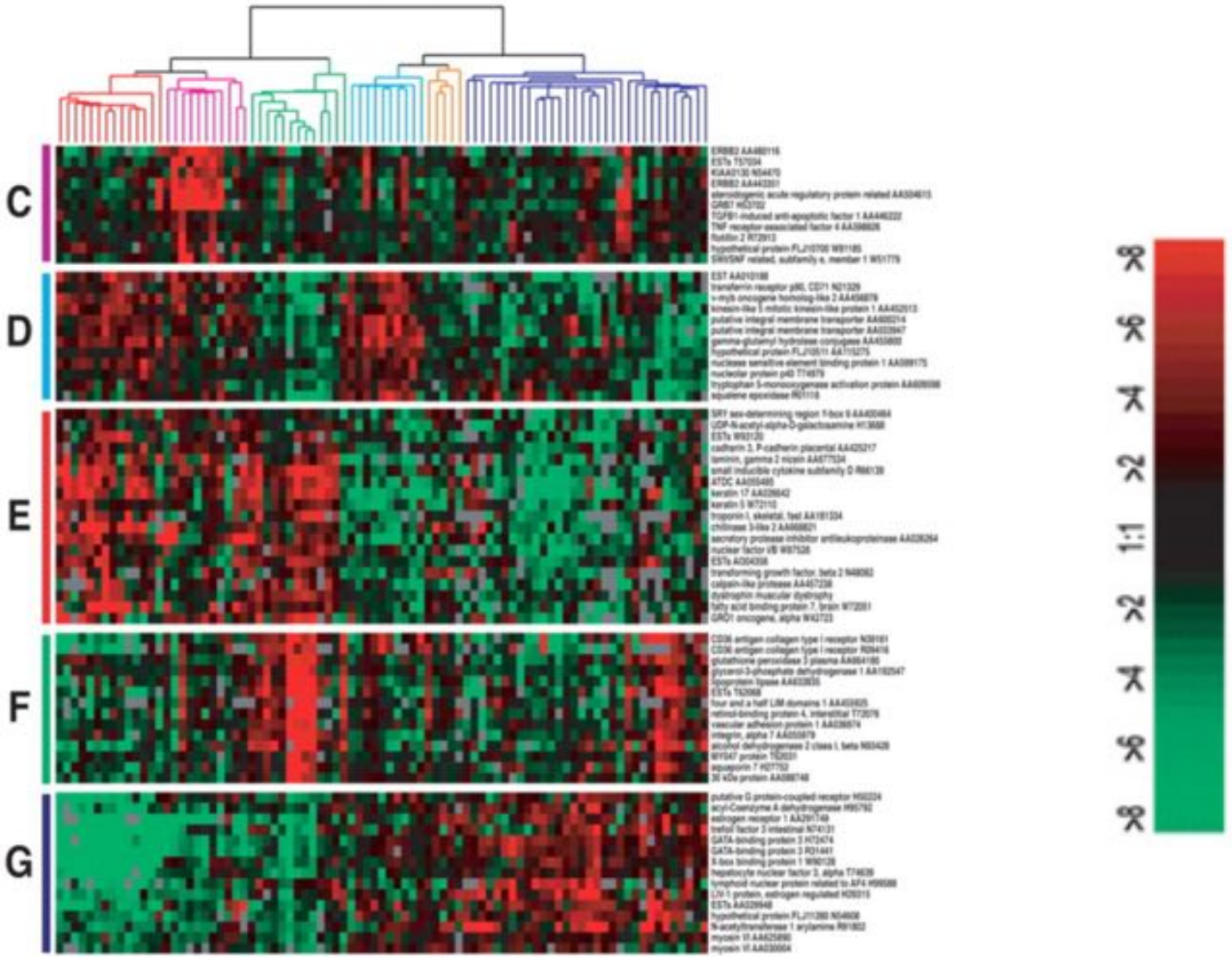
- Name of your team
- List of team members and email addresses
- Short title for your proposal
- 1 paragraph description of what you hope to do and how you will do it
- References to relevant papers
- References/URLs to datasets that you will be studying (Note you can also use simulated data)

Submit the proposal as a single page PDF on blackboard. After submitting your proposal, we will schedule a time to discuss your proposal, especially to ensure you have access to the data that you need. The sooner that you submit your proposal, the sooner we can schedule the meeting. No late days can be used for the project.

Later, you will present your project in class during the last week of class. You will also submit a written report (5-7 pages) of your project, formatting as a Bioinformatics article (Intro, Methods, Results, Discussion, References). Word and LaTeX templates are available at https://academic.oup.com/bioinformatics/pages/submission_online

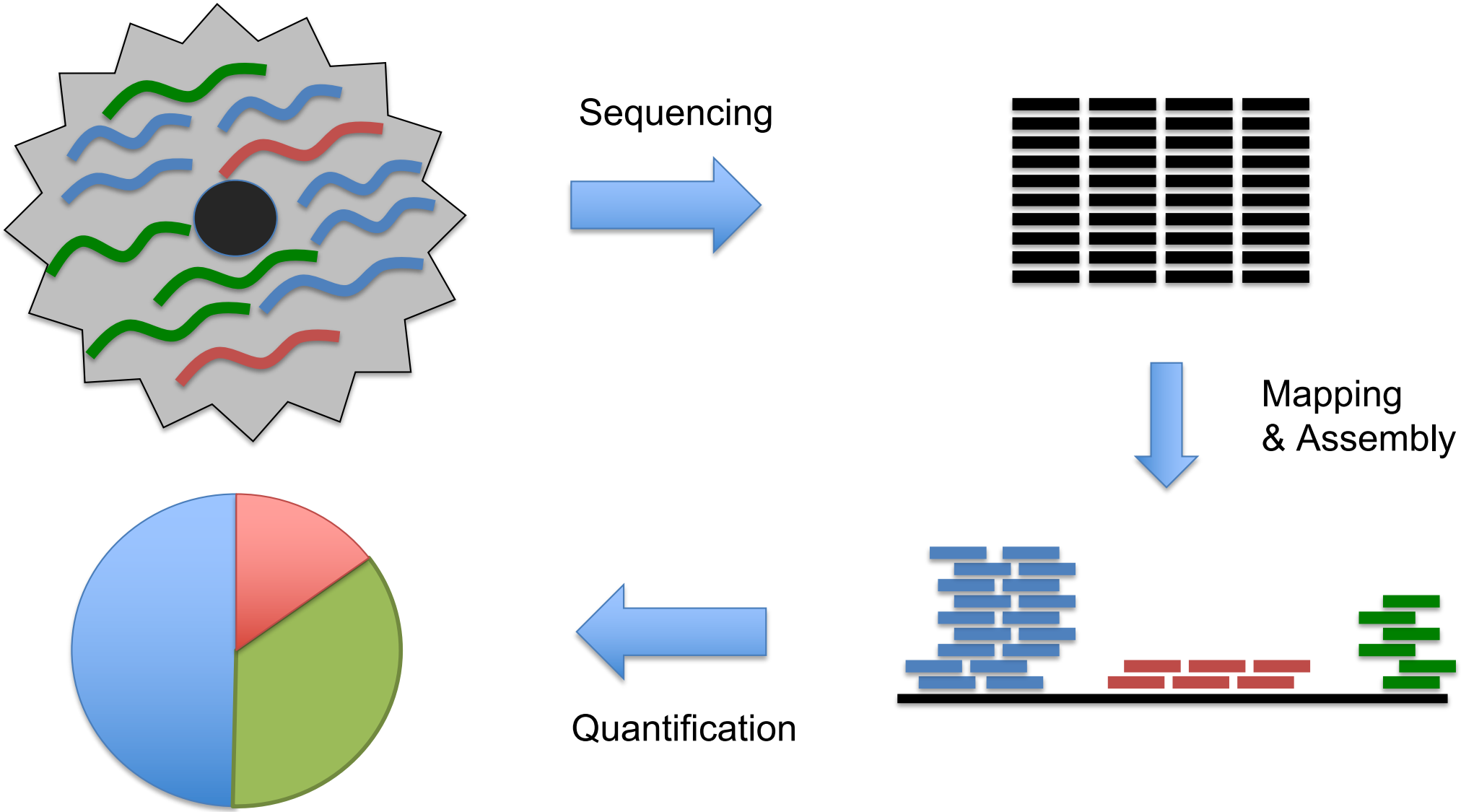
Please use Piazza to coordinate proposal plans!

RNA-seq

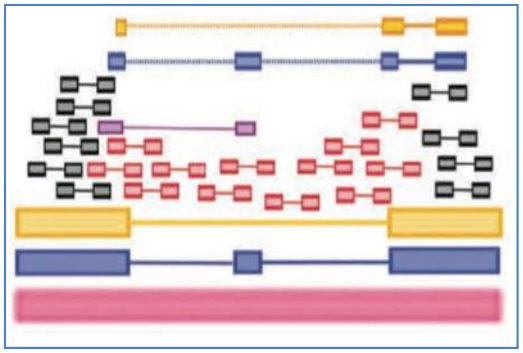


Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.
Sørlie et al (2001) *PNAS*. 98(19):10869-74.

RNA-seq Overview



RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

RNA-Seq Approaches

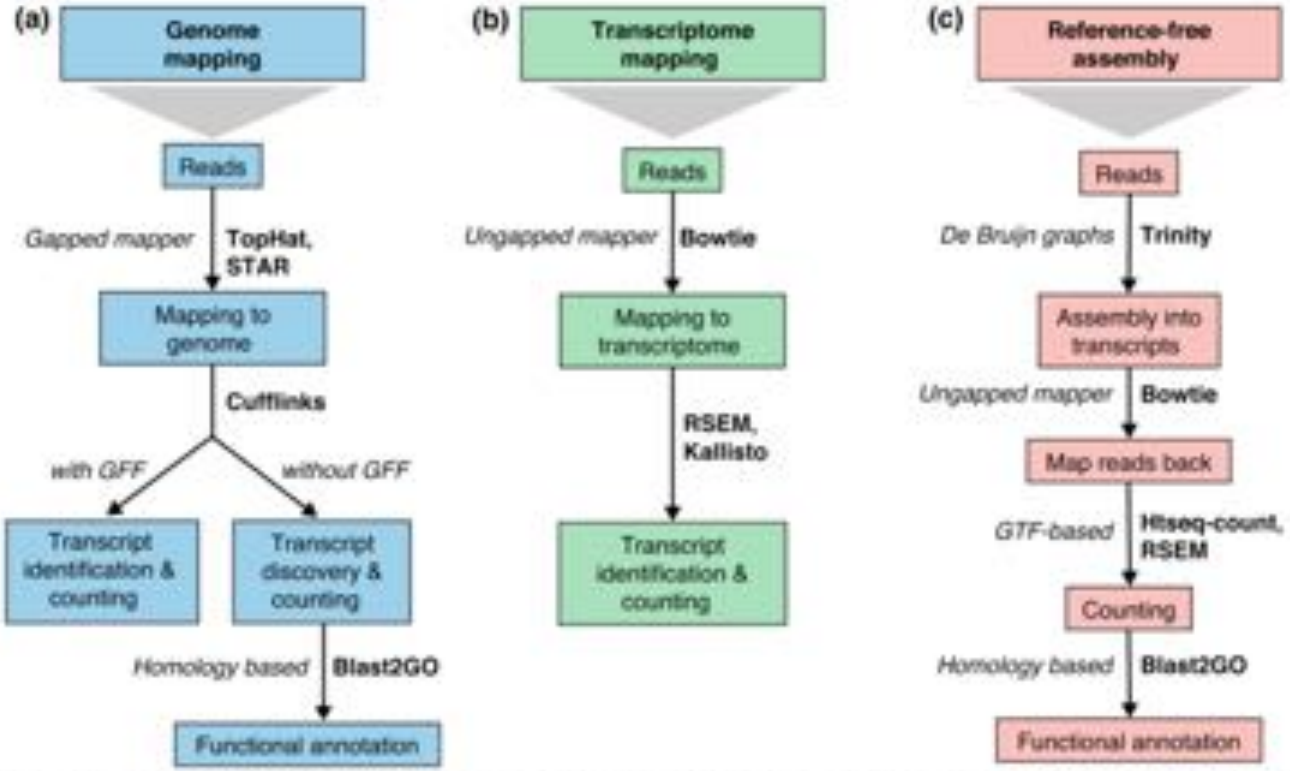


Fig. 2 Read mapping and transcript identification strategies. Three basic strategies for regular RNA-seq analysis. **a** An annotated genome is available and reads are mapped to the genome with a gapped mapper. Next (novel) transcript discovery and quantification can proceed with or without an annotation file. Novel transcripts are then functionally annotated. **b** If no novel transcript discovery is needed, reads can be mapped to the reference transcriptome using an ungapped aligner. Transcript identification and quantification can occur simultaneously. **c** When no genome is available, reads need to be assembled first into contigs or transcripts. For quantification, reads are mapped back to the novel reference transcriptome and further analysis proceeds as in (**b**) followed by the functional annotation of the novel transcripts as in (**a**). Representative software that can be used at each analysis step are indicated in *bold text*. Abbreviations: *GFF* General Feature Format, *GTF* gene transfer format, *RSEM* RNA-Seq by Expectation Maximization

A survey of best practices for RNA-seq data analysis
Conesa et al (2016) Genome Biology. doi 10.1186/s13059-016-0881-8

RNA-Seq Approaches

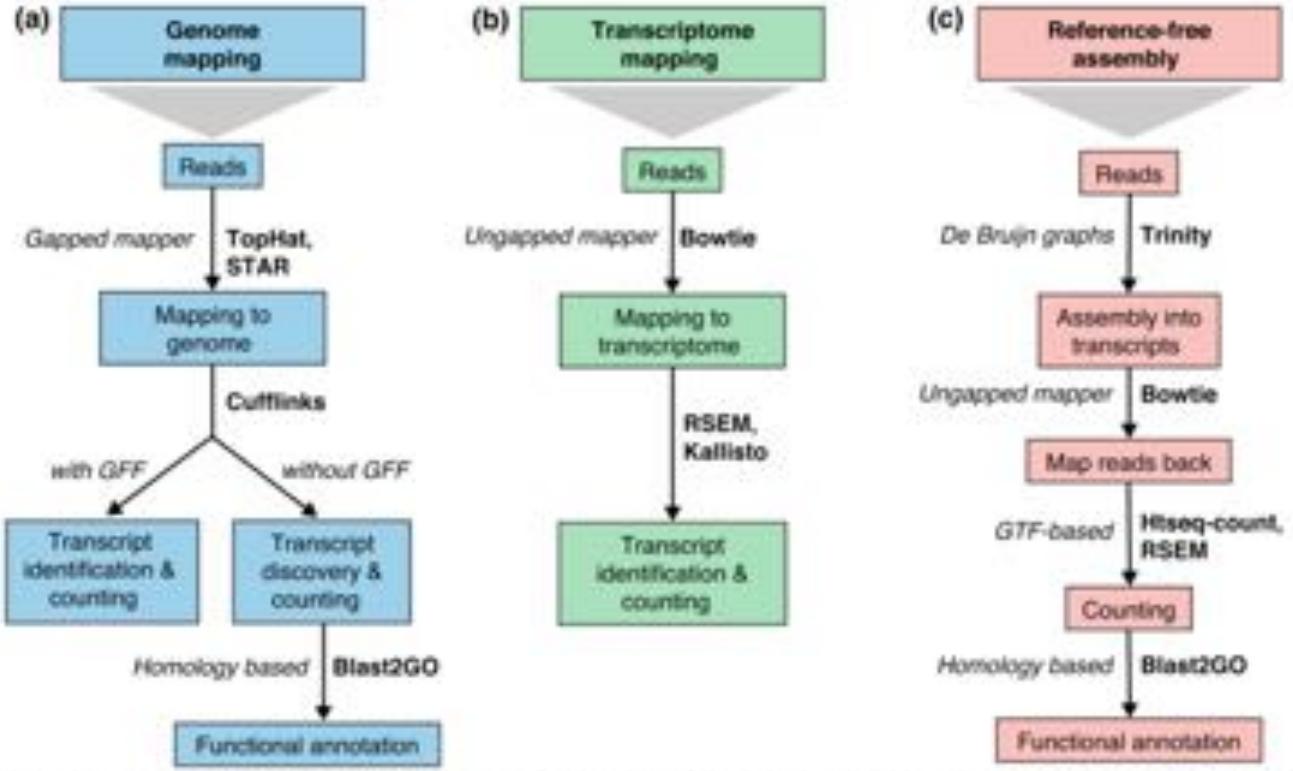


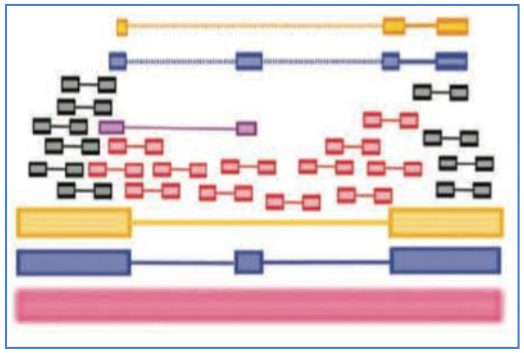
Fig. 2 Read mapping and transcript identification strategies. Three basic strategies for regular RNA-seq analysis. **a** An annotated genome is available and re…ovel) transcript discovery and quantification can proceed with or without an ann… no novel transcript discovery is needed, reads can be mapped to the reference transcriptome using an ungapped aligner. Transcript identification and quantification can occur simultaneously. **c** When no genome is available, reads …o contigs or transcripts. For quantification, reads are mapped back to the novel reference transcriptome and further a…owed by the functional annotation of the novel transcripts as in (**a**). Representative software that can be used a…ated in *bold text*. Abbreviations: *GFF* General Feature Format, *GTF* gene transfer format, *RSEM* RNA-Seq by Expectation Maximization

Which approach should we use?

It depends….

A survey of best practices for RNA-seq data analysis
Conesa et al (2016) Genome Biology. doi 10.1186/s13059-016-0881-8

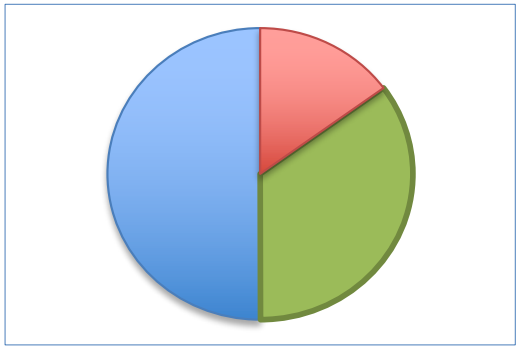
RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

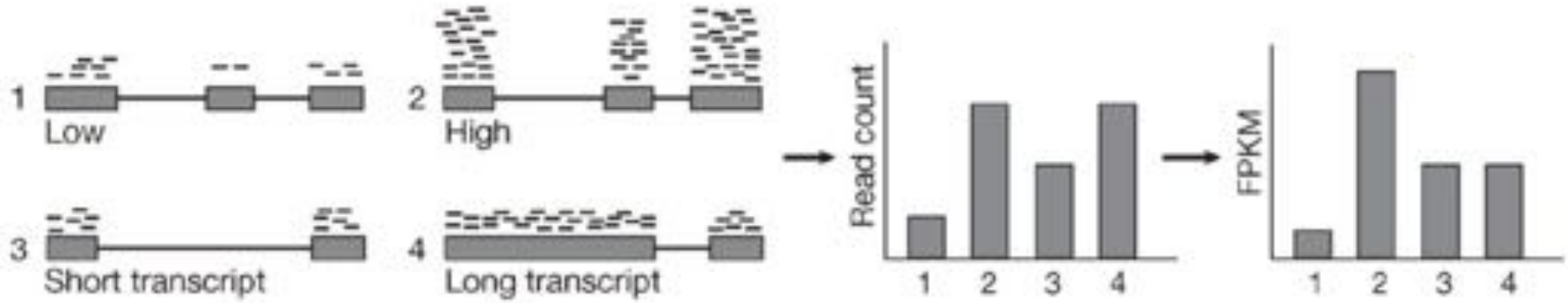
Solution: Use a spliced aligner, and assemble isoforms

TopHat: discovering spliced junctions with RNA-Seq.
Trapnell et al (2009) *Bioinformatics*. 25:0 1105-1111



Challenge 2: Read Count != Transcript abundance

RPKM, FPKM, TPM



Counting Reads that align to a gene DOESN'T work!

- Overall Coverage: 1M reads in experiment 1 vs 10M reads in experiment 2
- Gene Length: gene 3 is 10kbp, gene 4 is 100kbp

1. RPKM: Reads Per Kilobase of Exon Per Million Reads Mapped (Mortazavi et al, 2008)

=> Wait a second, reads in a pair arent independent!

2. FPKM: Fragments Per Kilobase of Exon Per Million Reads Mapped (Trapnell et al, 2010)

=> Wait a second, FPKM depends on the average transcript length!

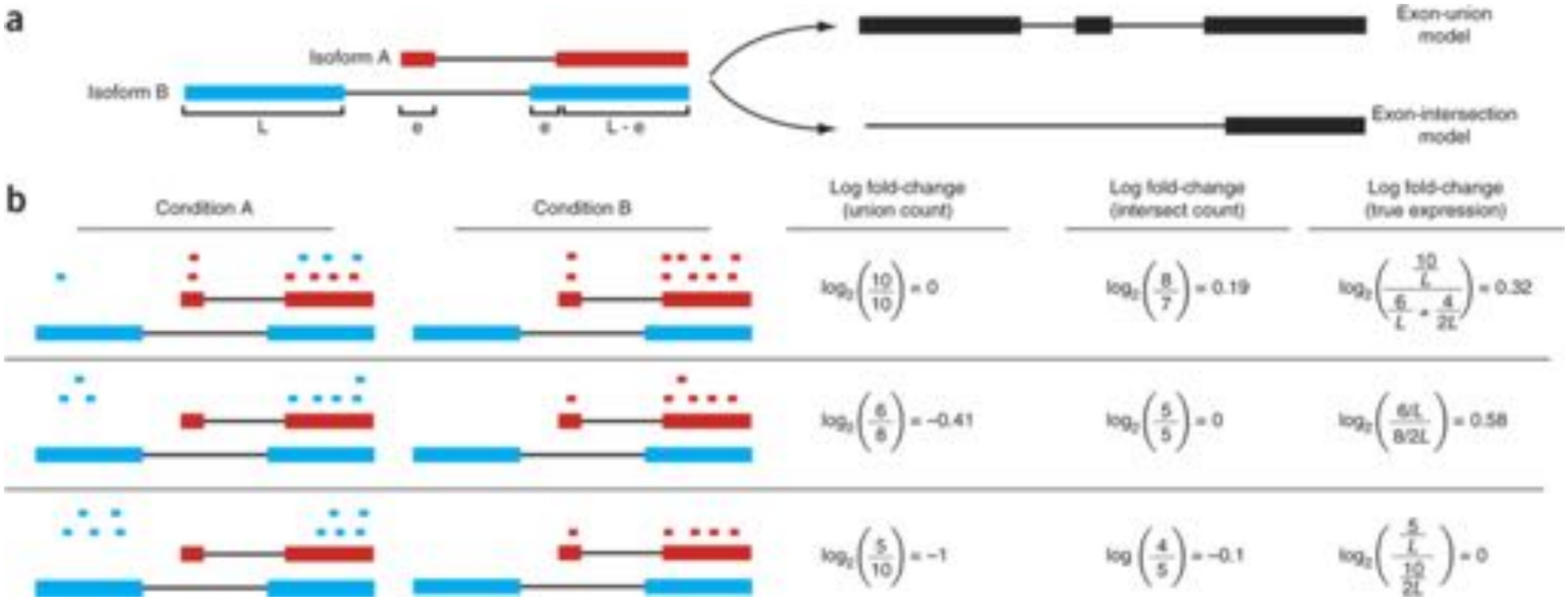
3. TPM: Transcripts Per Million (Li et al, 2011)

⇒ If you were to sequence one million full length transcripts, TPM is the number of transcripts you would have seen of type i, given the abundances of the other transcripts in your sample

=> Recommend you use TPM for all analysis, easy to compute given FPKM

$$\mathrm{TPM}_i = \left(\frac{\mathrm{FPKM}_i}{\sum_j \mathrm{FPKM}_j} \right) \cdot 10^6$$

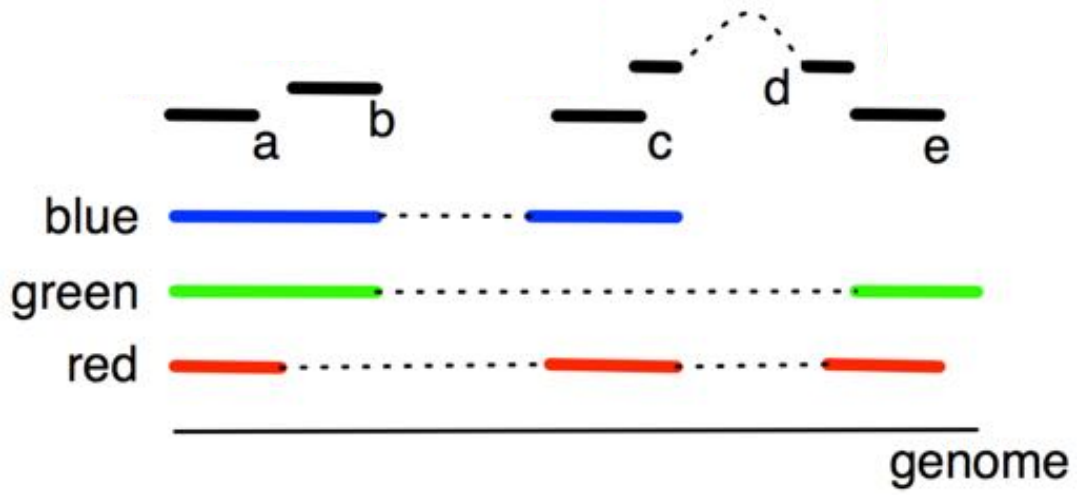
Gene or Isoform Quantification?



Key point : The length of the actual molecule from which the fragments derive is crucially important to obtaining accurate abundance estimates.

Differential analysis of gene regulation at transcript resolution with RNA-seq
Trapnell et al (2013) Nature Biotechnology 31, 46–53. doi:10.1038/nbt.2450

Multi-mapping? Isoform ambiguity? Expectation Maximization to the Rescue



The gene has three isoforms (red, green, blue) of the same length.
Our initial expectation is all 3 isoforms are equally expressed

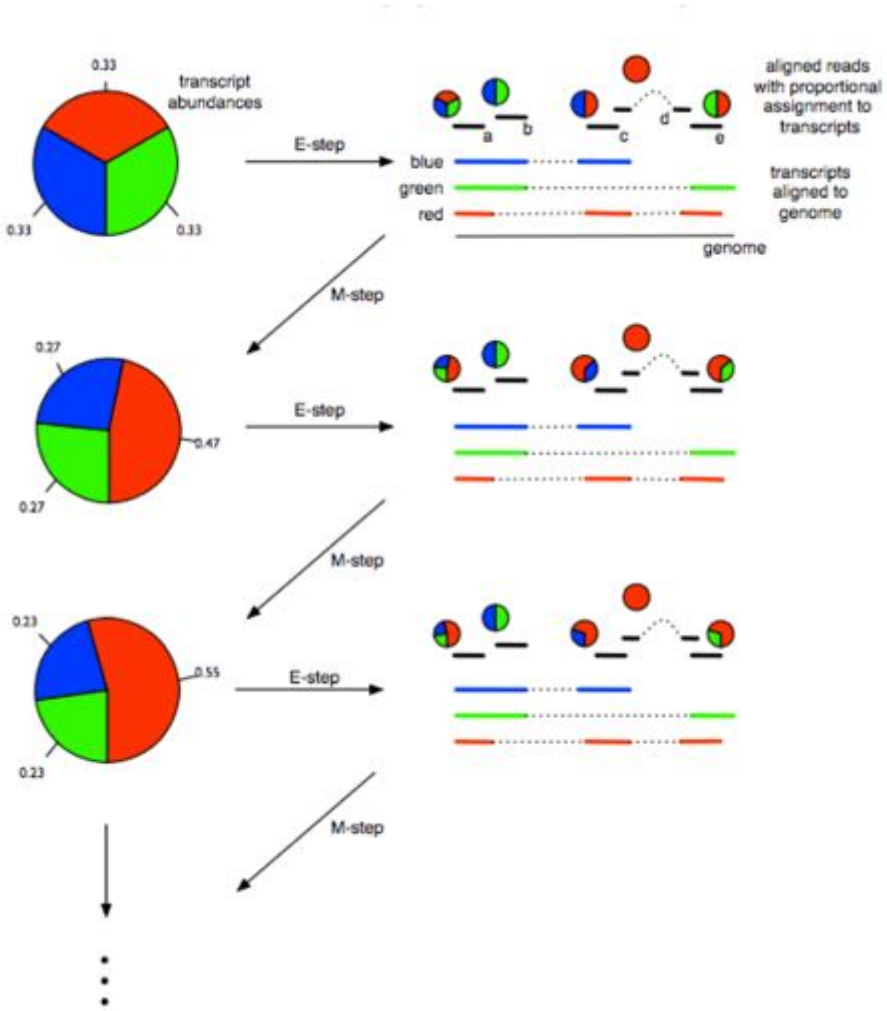
There are five reads (a,b,c,d,e) mapping to the gene.

- Read a maps to all three isoforms
- Read d only to red
- Reads b,c,e map to each of the three pairs of isoforms.

What is the most likely expression level of each isoform?

Models for transcript quantification from RNA-seq
Pachter, L (2011) arXiv. 1104.3889 [q-bio.GN]

Multi-mapping? Isoform ambiguity? Expectation Maximization to the Rescue



The gene has three isoforms (red, green, blue) of the same length. Initially every isoform is assigned the same abundance (red=1/3, green=1/3, blue=1/3)

There are five reads (a,b,c,d,e) mapping to the gene. Read a maps to all three isoforms, read d only to red, and the other three (reads b,c,e) to each of the three pairs of isoforms.

During the expectation (E) step reads are proportionately assigned to transcripts according to the (current) isoform abundances (RGB): a=(.33,.33,.33), b=(0,.5,.5), c=(.5,.5), d=(1,0,0), e=(.5,.5,0)

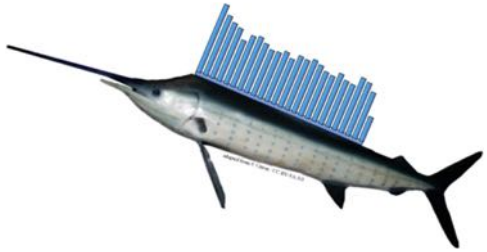
Next, during the maximization (M) step isoform abundances are recalculated from the proportionately assigned read counts:

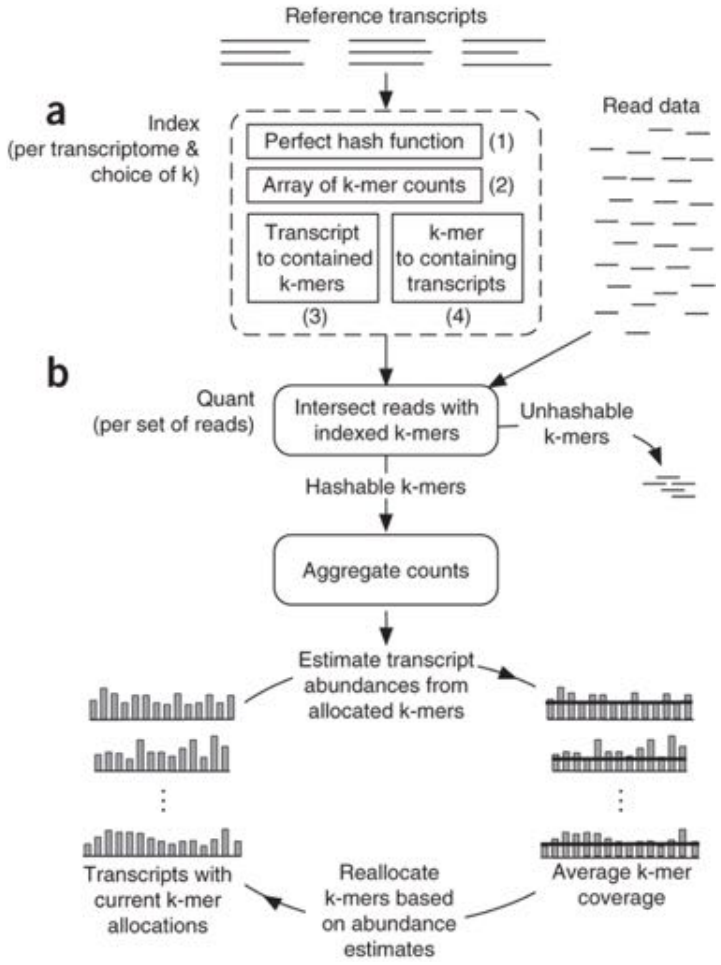
red: 0.47 = (0.33 + 0.5 + 1 + 0.5)/(2.33 + 1.33 + 1.33)
blue: 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)
green: 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)

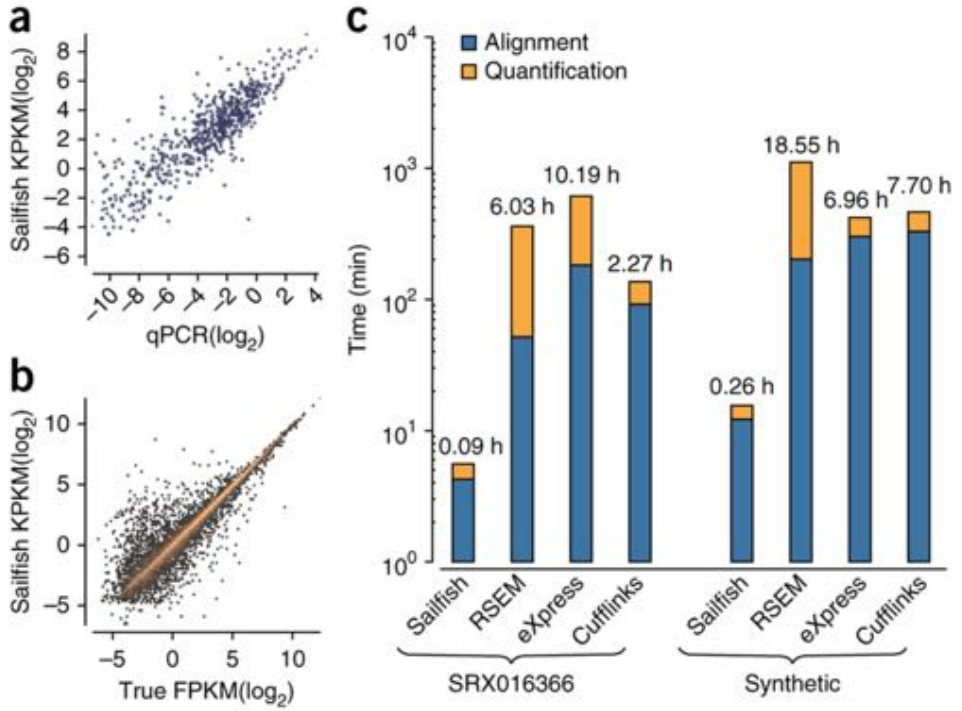
Repeat until convergence!

Models for transcript quantification from RNA-seq
Pachter, L (2011) arXiv. 1104.3889 [q-bio.GN]

Sailfish: Fast & Accurate RNA-seq Quantification





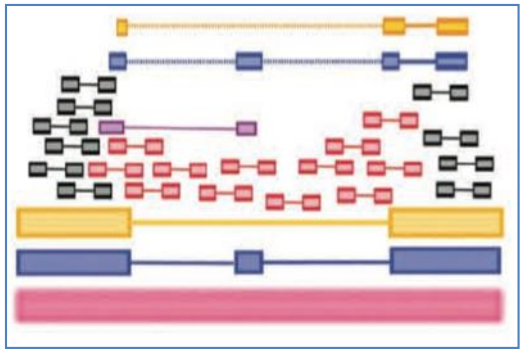


d

	Human brain tissue				Synthetic			
	Sailfish	RSEM	eXpress	Cufflinks	Sailfish	RSEM	eXpress	Cufflinks
Pearson	0.85	0.82	0.85	0.85	0.96	0.96	0.95	0.94
Spearman	0.84	0.80	0.85	0.85	0.76	0.77	0.77	0.75
RMSE	–	–	–	–	6.06	8.90	8.83	10.05
medPE	–	–	–	–	6.50	12.48	14.06	12.42

Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms
Patro et al (2014) Nature Biotechnology 32, 462–464 doi:10.1038/nbt.2862

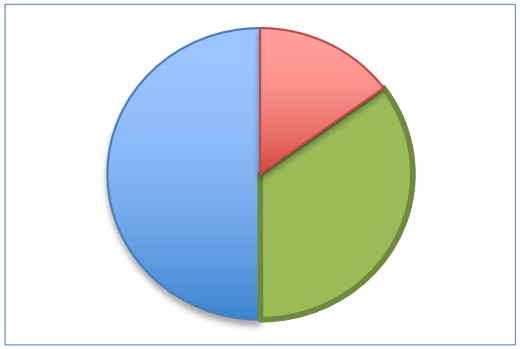
RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

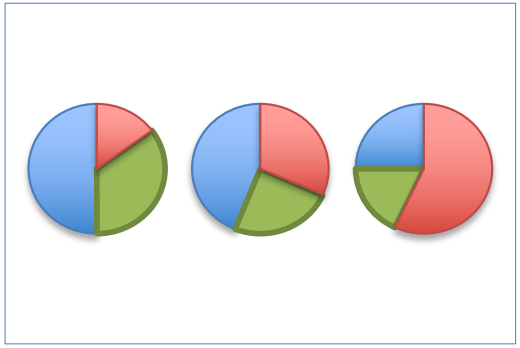
TopHat: discovering spliced junctions with RNA-Seq.
Trapnell et al (2009) *Bioinformatics*. 25:0 1105-1111



Challenge 2: Read Count != Transcript abundance

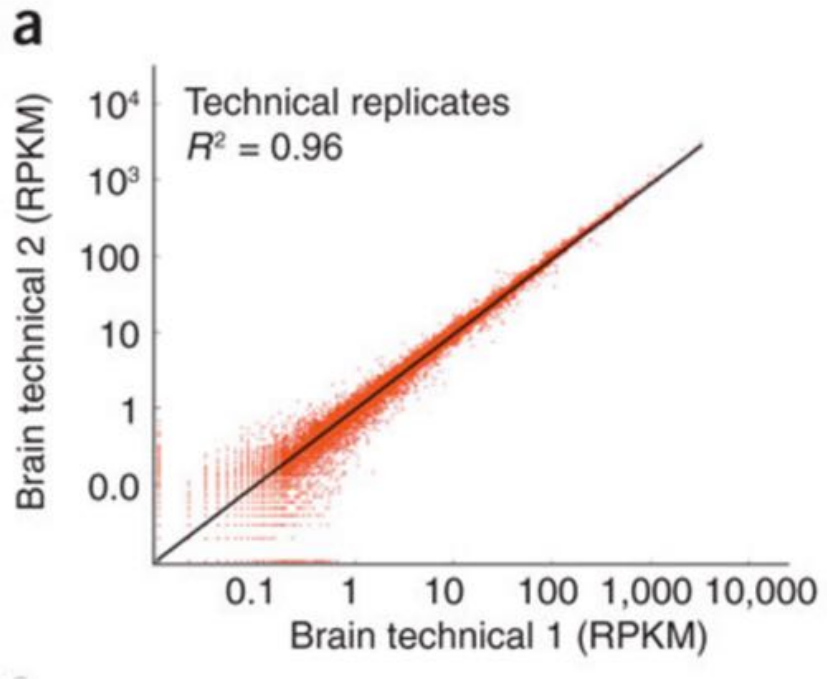
Solution: Infer underlying abundances (e.g. TPM)

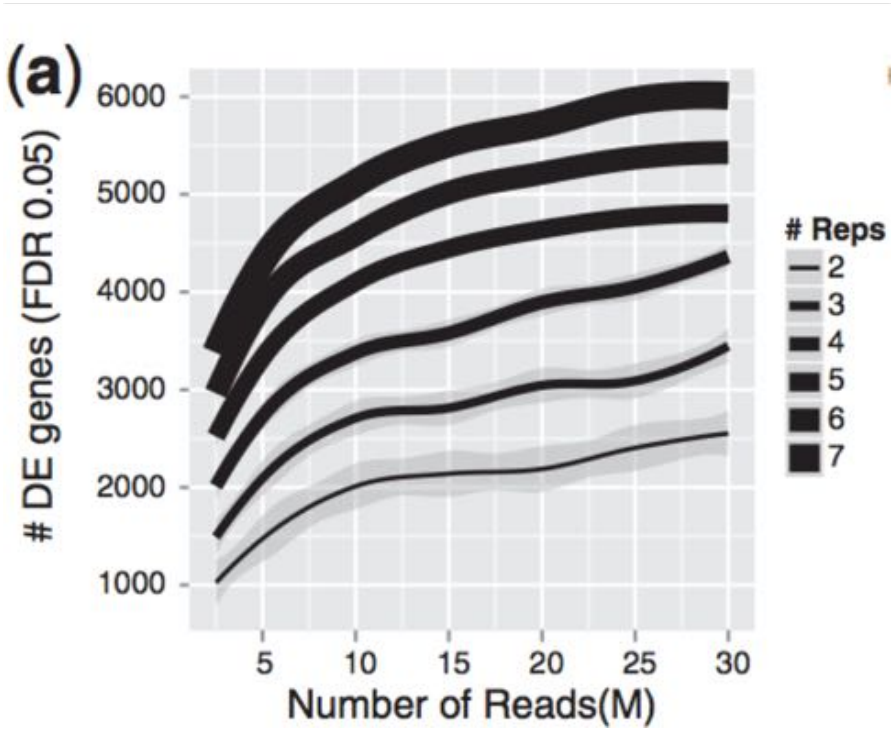
Transcript assembly and quantification by RNA-seq
Trapnell et al (2010) *Nat. Biotech.* 25(5): 511-515



Challenge 3: Transcript abundances are stochastic

How Many Replicates?



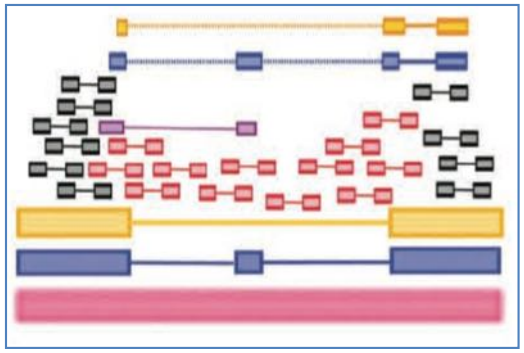


Why don't we have perfect replicates?

Mapping and quantifying mammalian transcriptomes by RNA-Seq
Mortazavi et al (2008) Nature Methods. 5, 62-628

RNA-seq differential expression studies: more sequence or more replication?
Liu et al (2013) Bioinformatics. doi:10.1093/bioinformatics/btt688

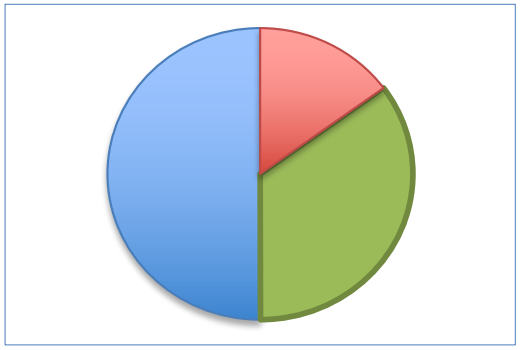
RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

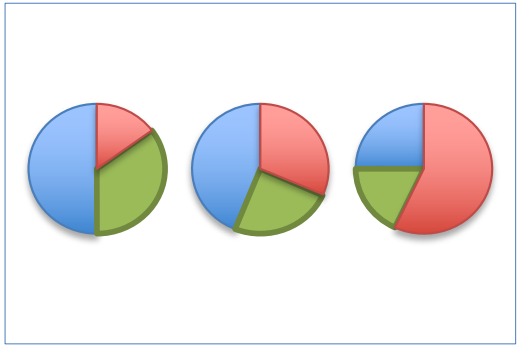
TopHat: discovering spliced junctions with RNA-Seq.
Trapnell et al (2009) *Bioinformatics*. 25:0 1105-1111



Challenge 2: Read Count != Transcript abundance

Solution: Infer underlying abundances (e.g. TPM)

Transcript assembly and quantification by RNA-seq
Trapnell et al (2010) *Nat. Biotech.* 25(5): 511-515

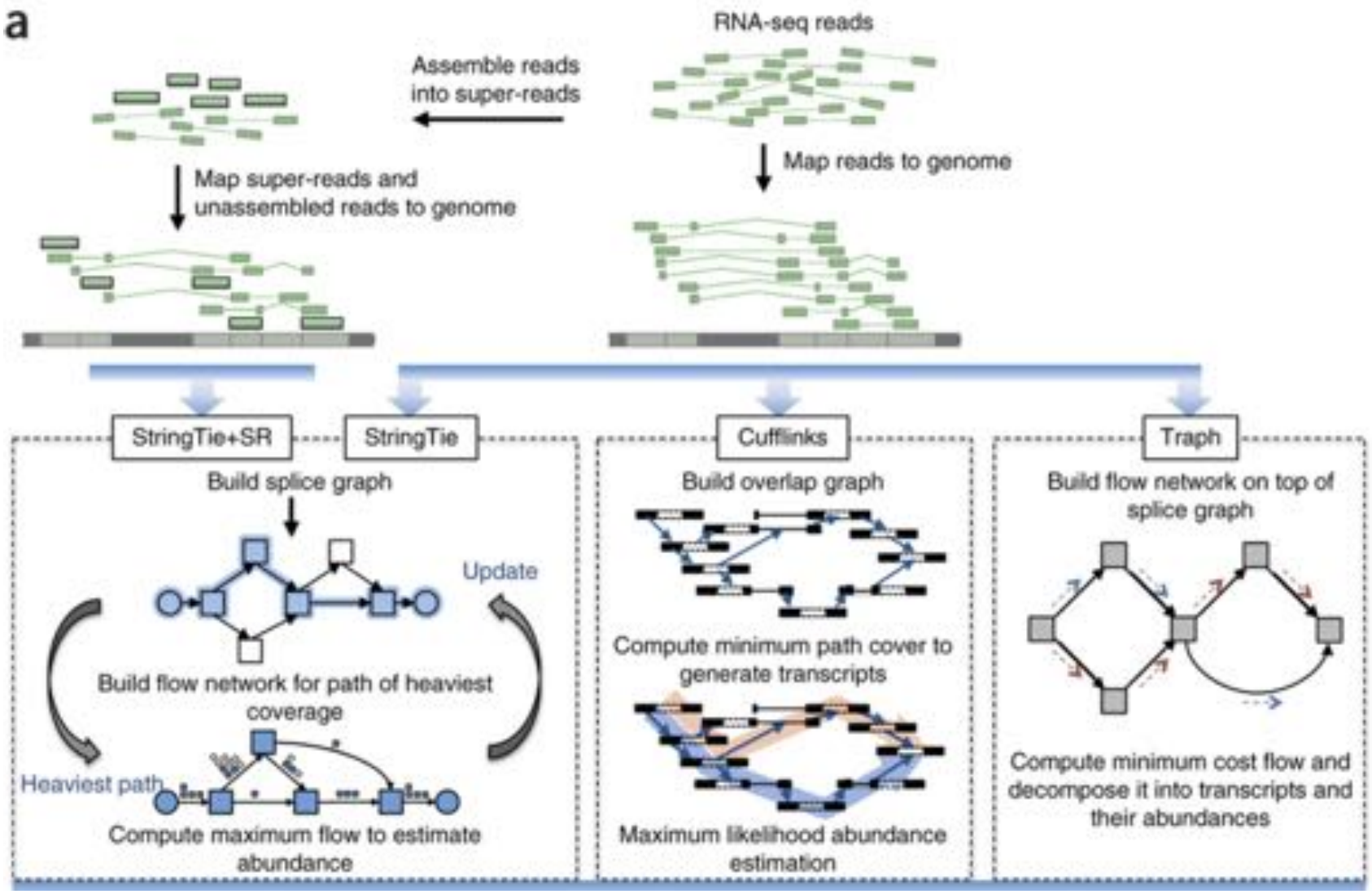


Challenge 3: Transcript abundances are stochastic

Solution: Replicates, replicates, and more replicates

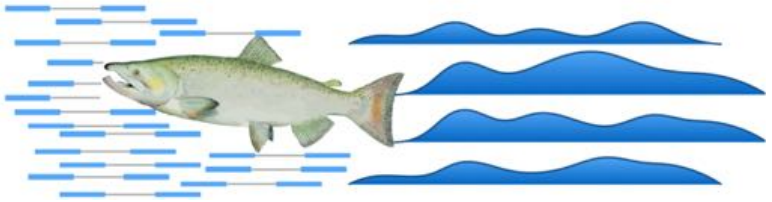
RNA-seq differential expression studies: more sequence or more replication?
Liu et al (2013) *Bioinformatics.* doi:10.1093/bioinformatics/btt688

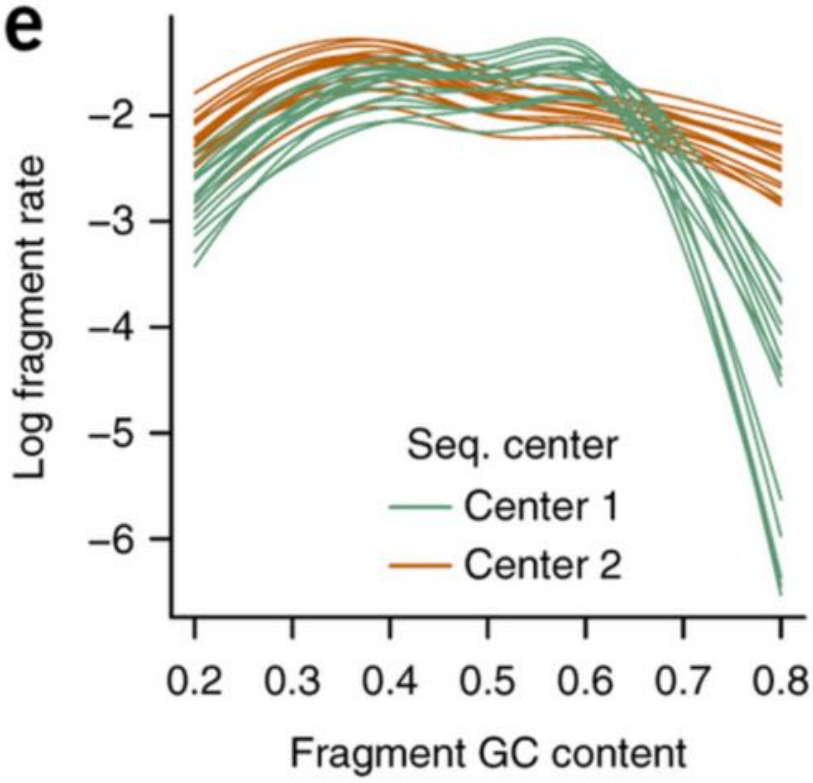
Isoform Quantification Approaches

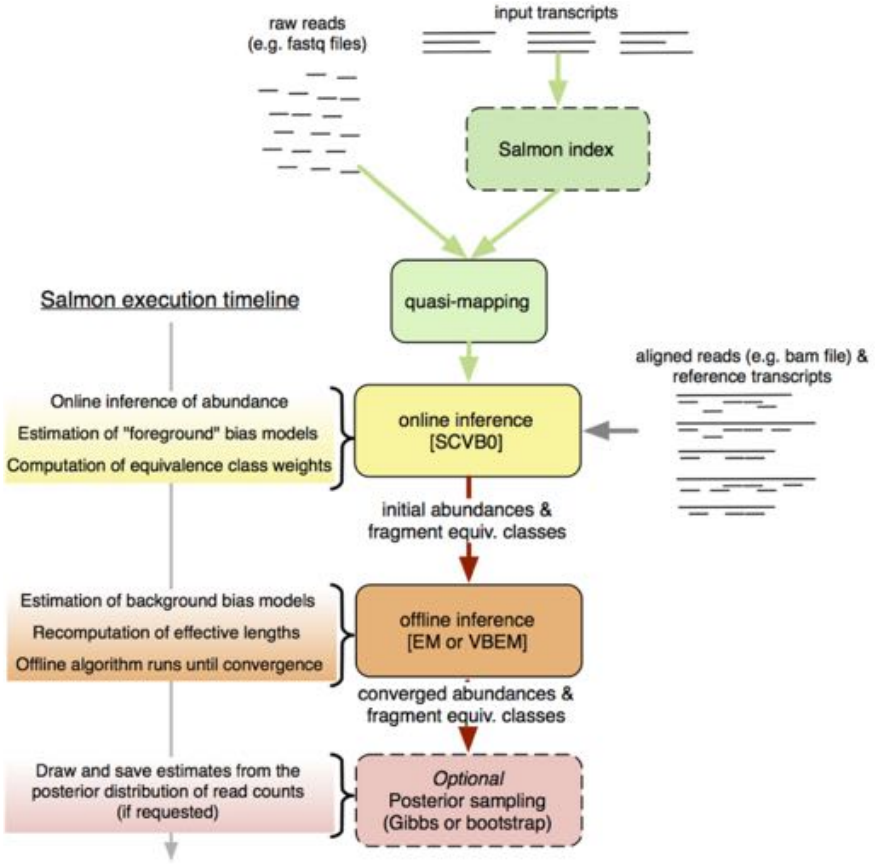


StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.
Pertea M, et al. (2015) Nature Biotechnology. doi: 10.1038/nbt.3122.

Salmon: The ultimate RNA-seq Pipeline?



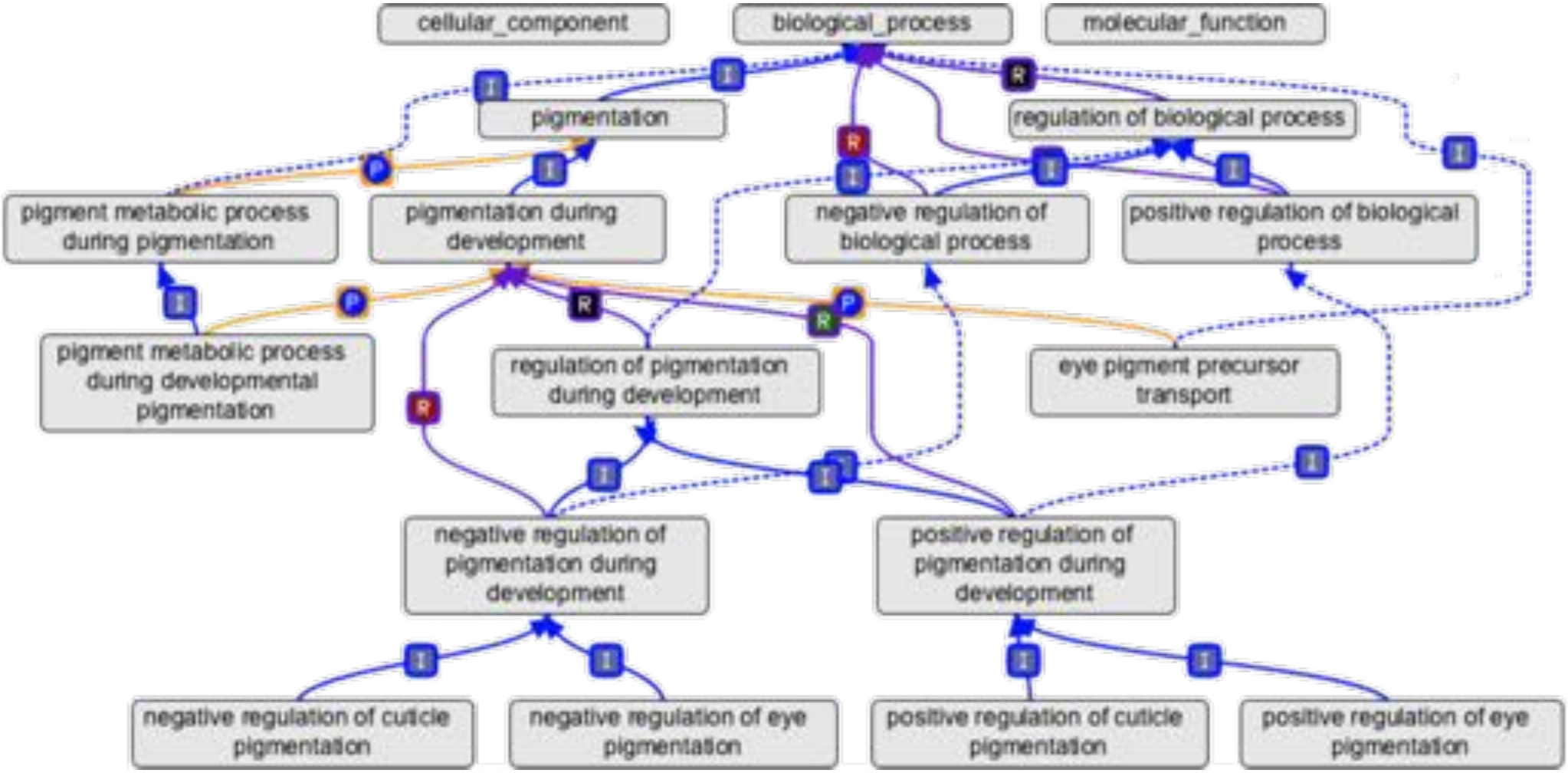




Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation
Love et al (2016) Nature Biotechnology 34, 1287–1291 (2016) doi:10.1038/nbt.3682

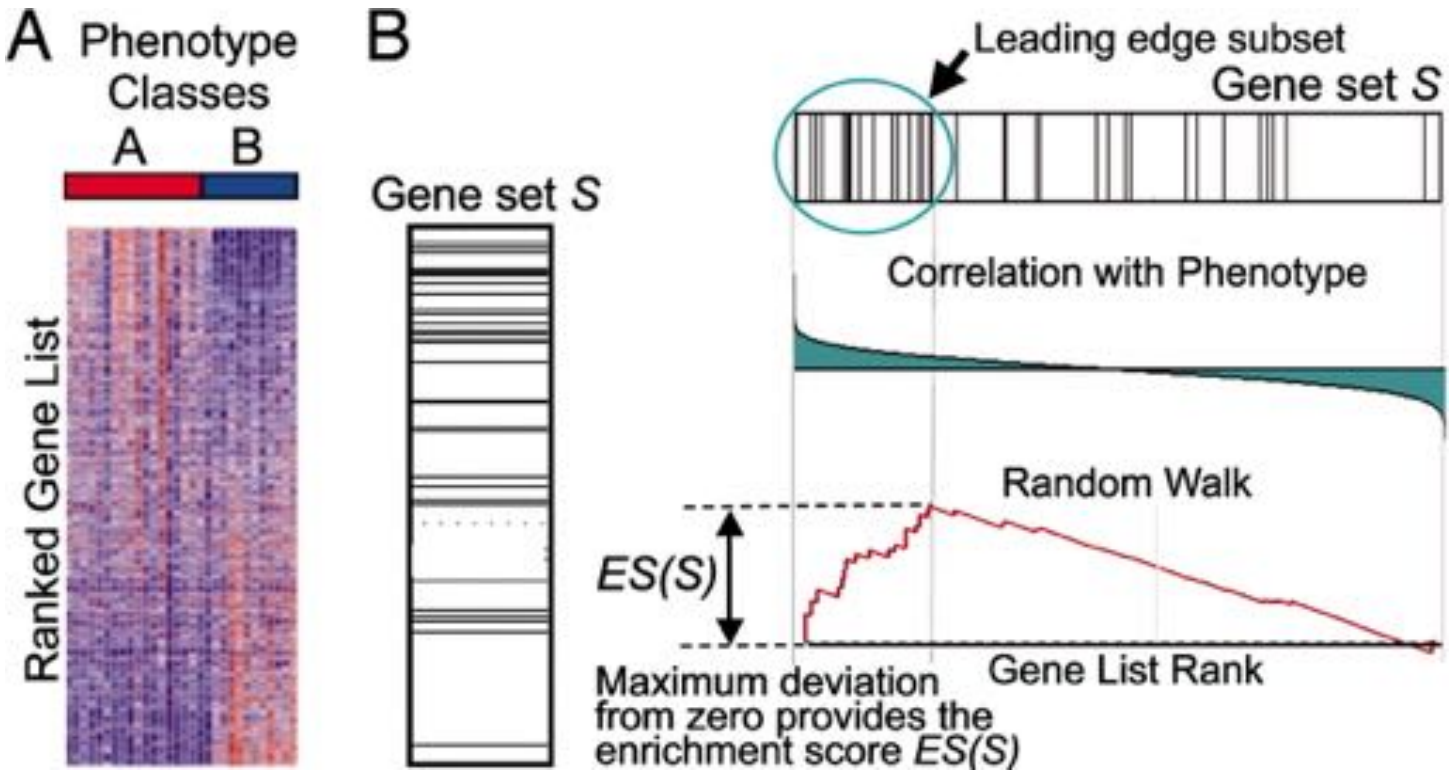
Salmon provides fast and bias-aware quantification of transcript expression
Patro et al (2017) Nature Methods (2017) doi:10.1038/nmeth.4197

Gene Ontology (GO)



AmiGO: online access to ontology and annotation data
Carbon et al (2009) Bioinformatics doi:10.1093/bioinformatics/btn615

GSEA Overview



Collections

The MSigDB gene sets are divided into 8 major collections:

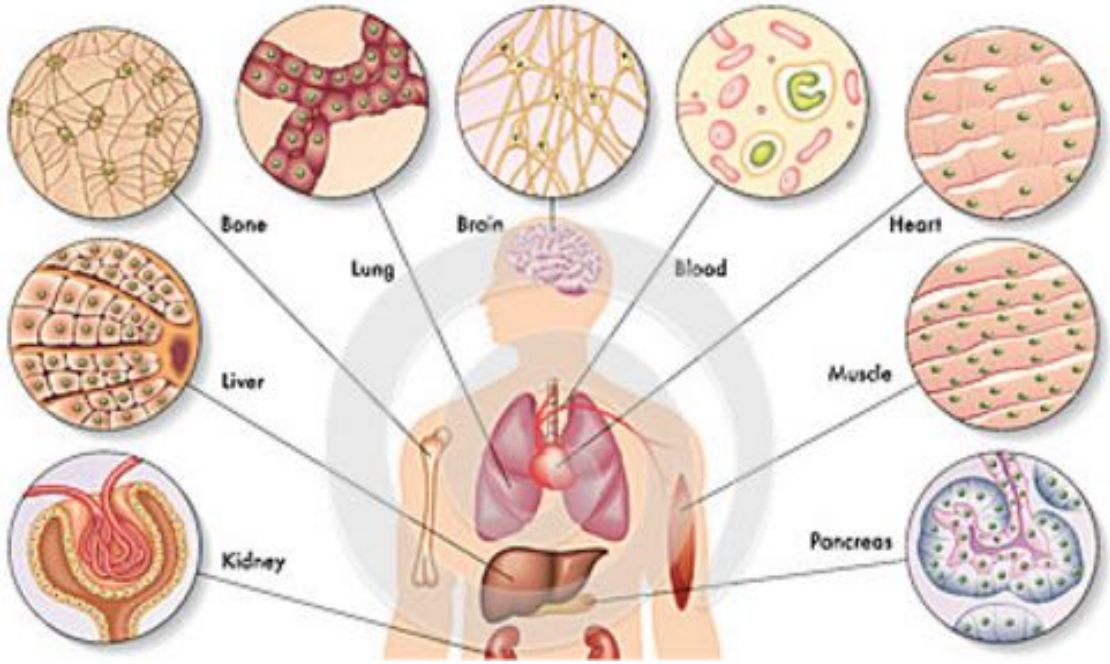
H	**hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.
C1	**positional gene sets** for each human chromosome and cytogenetic band.
C2	**curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.
C3	**motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.
C4	**computational gene sets** defined by mining large collections of cancer-oriented microarray data.
C5	**GO gene sets** consist of genes annotated by the same GO terms.
C6	**oncogenic gene sets** defined directly from microarray gene expression data from cancer gene perturbations.
C7	**immunologic gene sets** defined directly from microarray gene expression data from immunologic studies.

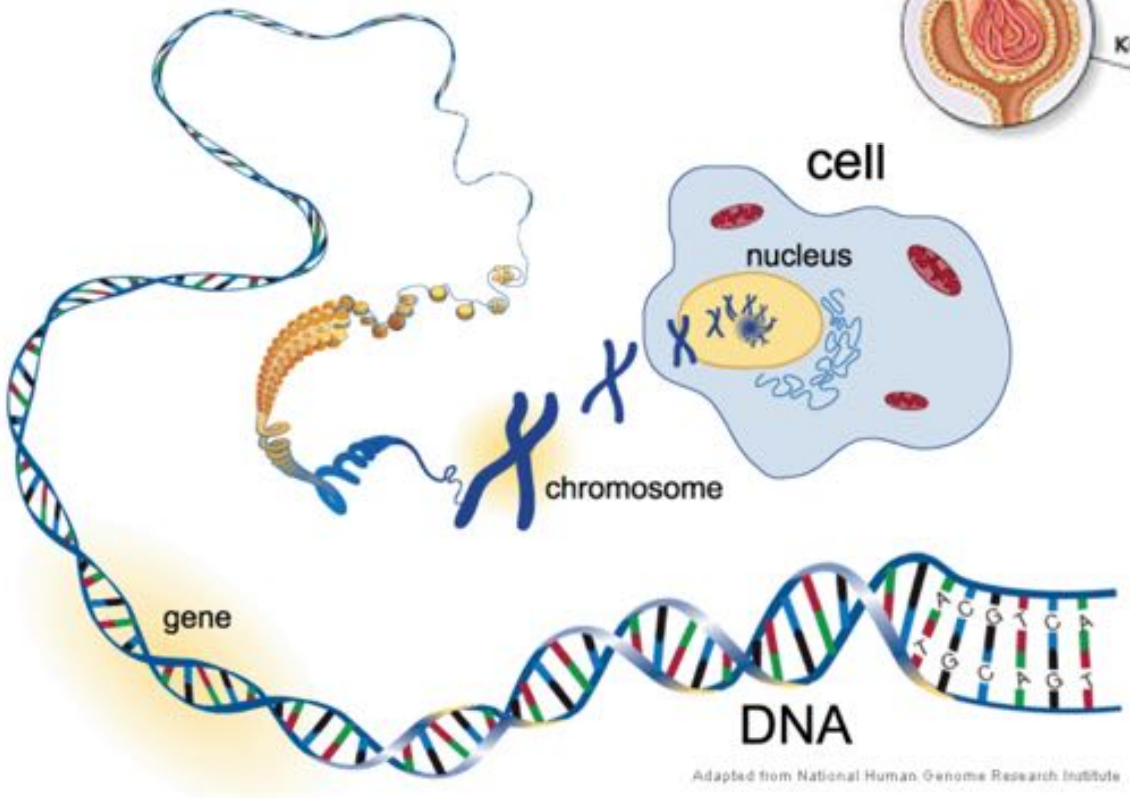
Aravind Subramanian et al. PNAS 2005;102:43:15545-15550

PNAS

Why Genes?

Each cell of your body contains an exact copy of your 3 billion base pair genome.





Your body has a few hundred (thousands?) major cell types, largely defined by the gene expression patterns

Human Evolution

- Humans and chimpanzees shared a common ancestor ~5-7 million years ago (Mya)
- Single-nucleotide substitutions occur at a mean rate of 1.23% but ~4% overall rate of mutation: comprising ~35 million single nucleotide differences and ~90 Mb of insertions and deletions
- Orthologous proteins in human and chimpanzee are extremely similar, with ~29% being identical and the typical orthologue differing by only two amino acids, one per lineage

Initial sequence of the chimpanzee genome and comparison with the human genome (2005) *Nature 437, 69-87 doi:10.1038/nature04072*

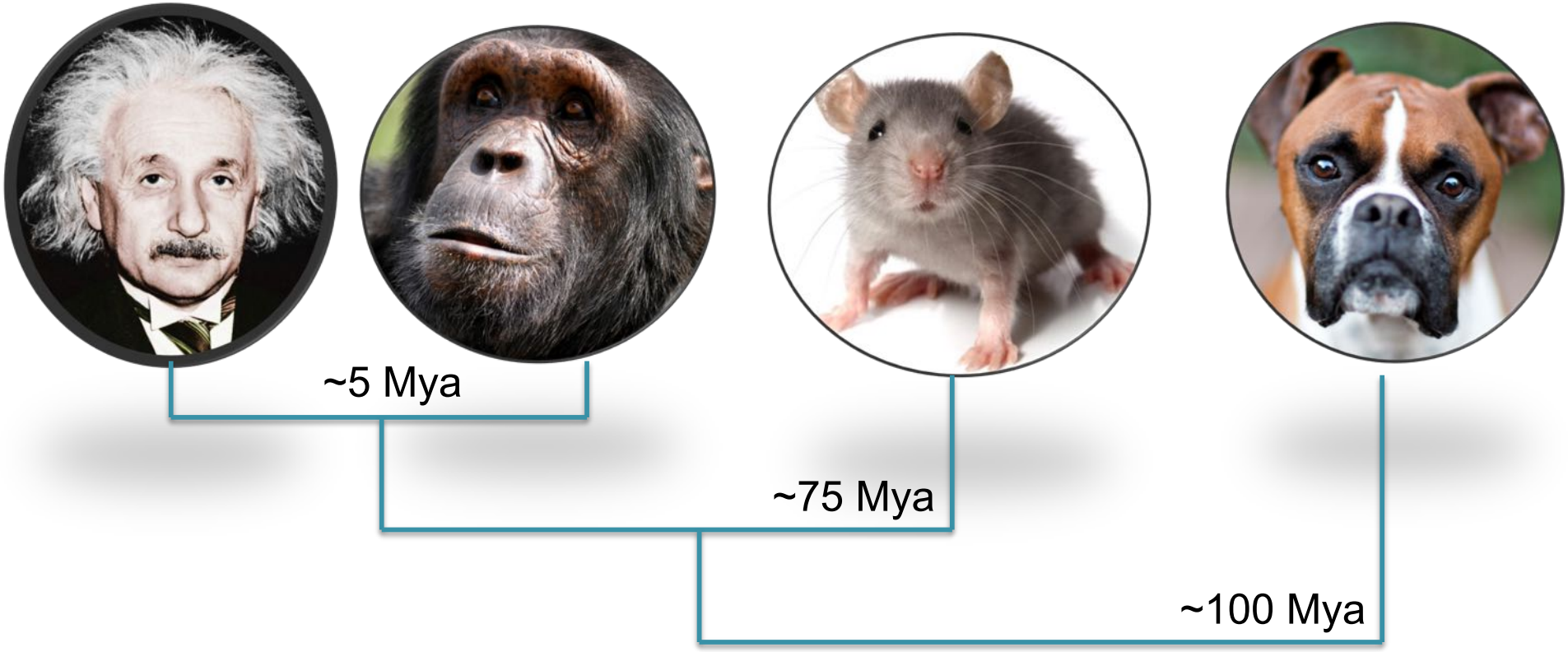
Human Evolution

"In the roughly 75 million years since the divergence of the human and mouse lineages, the process of evolution has altered their genome sequences and caused them to diverge by ***nearly one substitution for every two nucleotides***"

"The mouse and human genomes each seem to contain about 30,000 protein-coding genes. These refined estimates have been derived from both new evidence-based analyses that produce larger and more complete sets of gene predictions, and new de novo gene predictions that do not rely on previous evidence of transcription or homology. The proportion of mouse genes with a single identifiable orthologue in the human genome seems to be approximately 80%. ***The proportion of mouse genes without any homologue currently detectable in the human genome (and vice versa) seems to be less than 1%.***"

Initial sequencing and comparative analysis of the mouse genome
Chinwalla et al (2002) *Nature.* 420, 520-562 doi:10.1038/nature01262

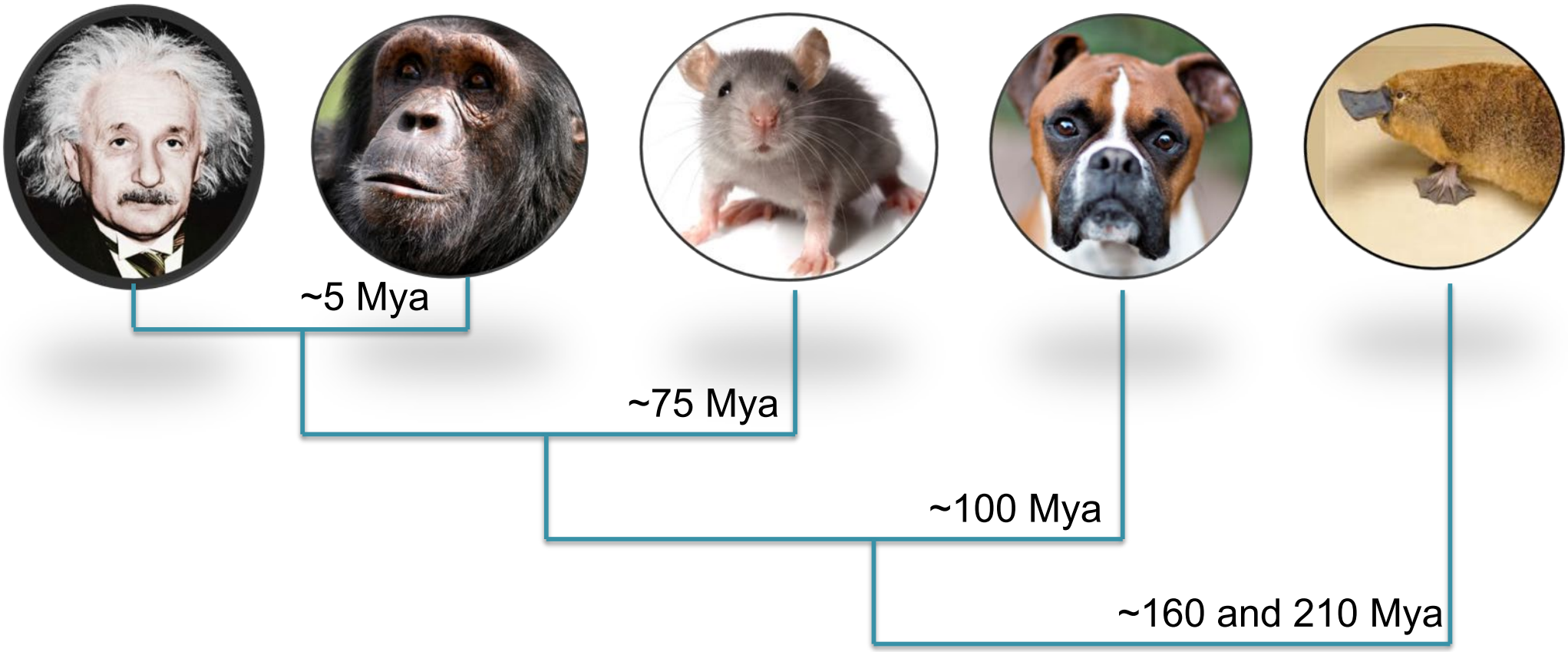
Human Evolution



"We generated gene predictions for the dog genome using an evidence-based method (see Supplementary Information). The resulting collection contains ***19,300 dog gene predictions, with nearly all being clear homologues of known human genes***. The dog gene count is substantially lower than the ~22,000-gene models in the current human gene catalogue (EnsEMBL build 26). For many predicted human genes, we find no convincing evidence of a corresponding dog gene. Much of the excess in the human gene count is attributable ***to spurious gene predictions in the human genome***"

Genome sequence, comparative analysis and haplotype structure of the domestic dog
Lindblad-Toh et al (2005) *Nature. 438, 803-819 doi:10.1038/nature04338*

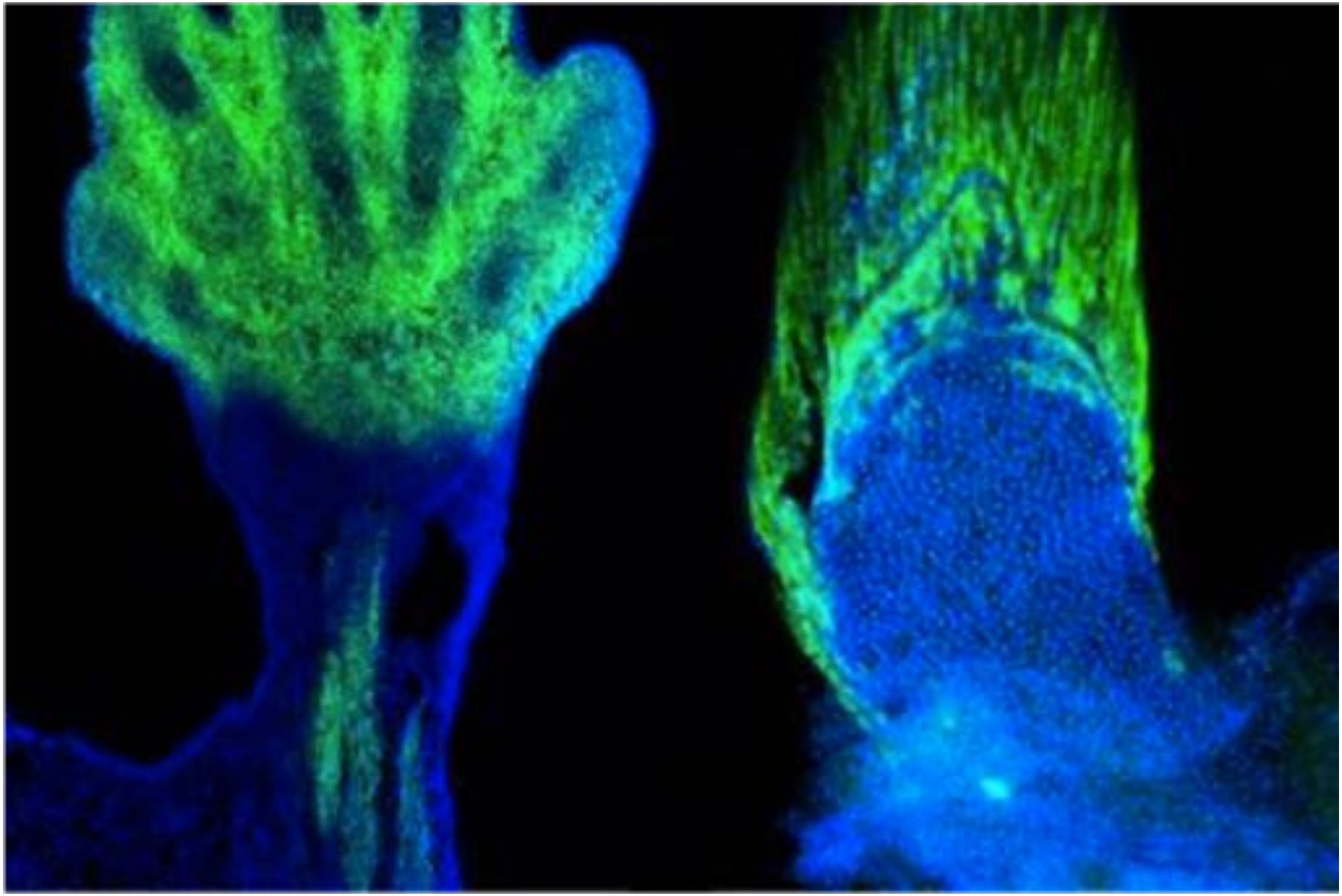
Human Evolution



As expected, the majority of platypus genes (82%; 15,312 out of 18,596) have orthologues in these five other amniotes (Supplementary Table 5). The remaining 'orphan' genes are expected to primarily reflect rapidly evolving genes, for which no other homologues are discernible, erroneous predictions, and true lineage-specific genes that have been lost in each of the other five species under consideration.

Genome analysis of the platypus reveals unique signatures of evolution
(2008) *Nature.* 453, 175-183 doi:10.1038/nature06936

Human Evolution

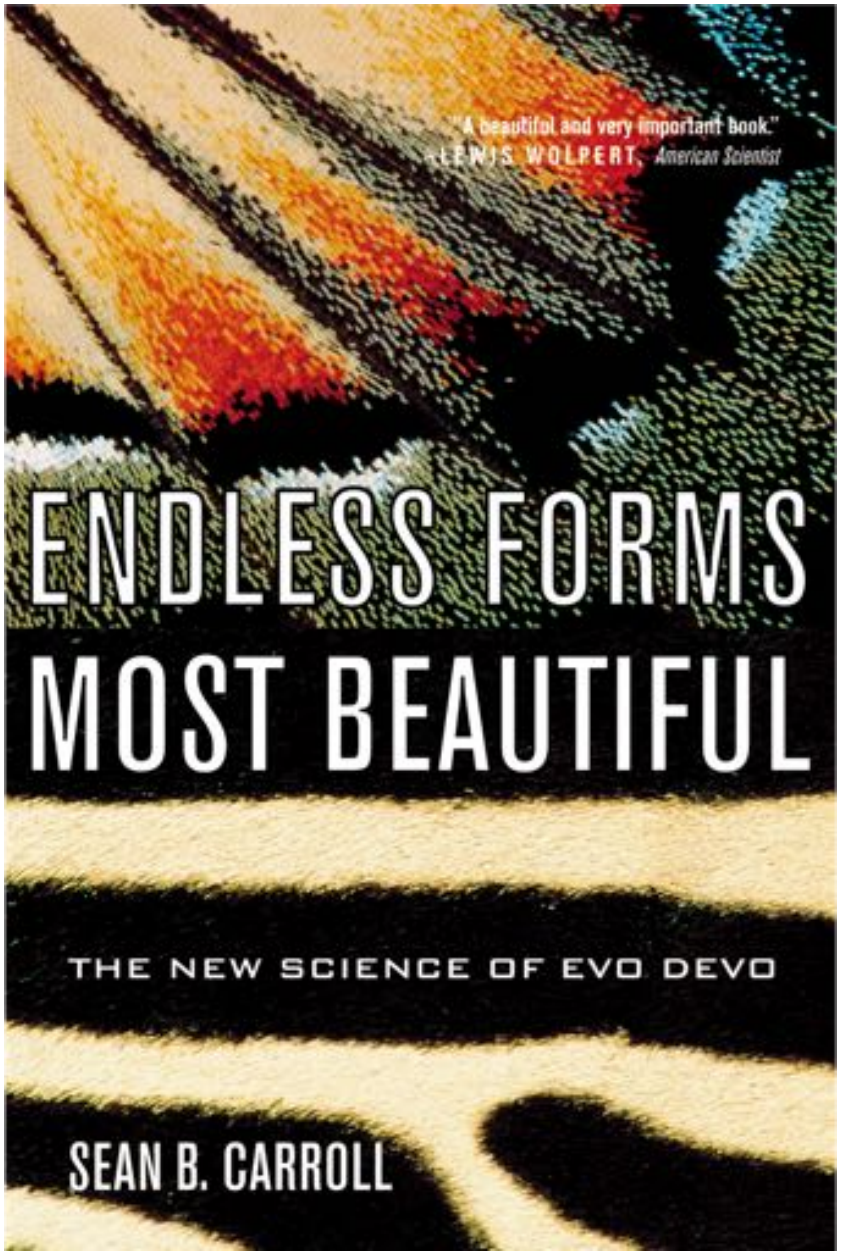


Digits and fin rays share common developmental histories
Nakamura et al (2016) *Nature.* 537, 225–228. doi:10.1038/nature19322

More Information

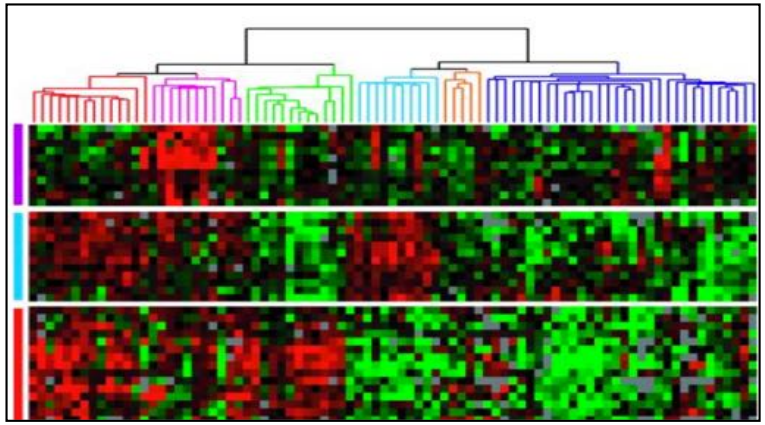
“Anything found to be true of E. coli must also be true of elephants”

-Jacques Monod

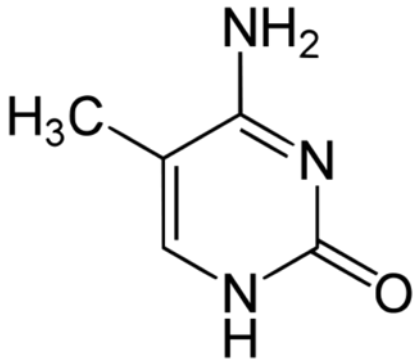


*-seq in 4 short vignettes

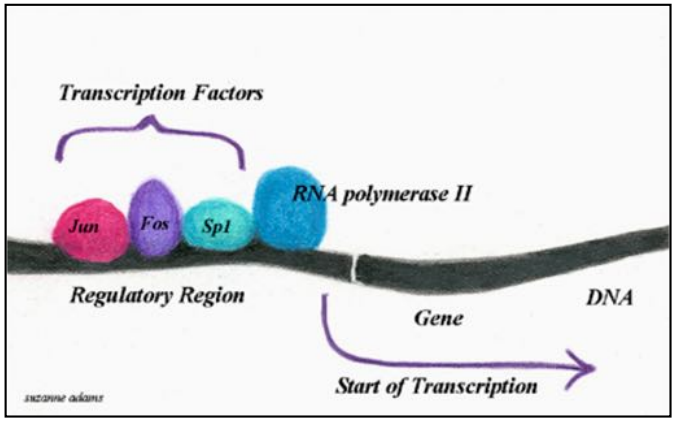
RNA-seq



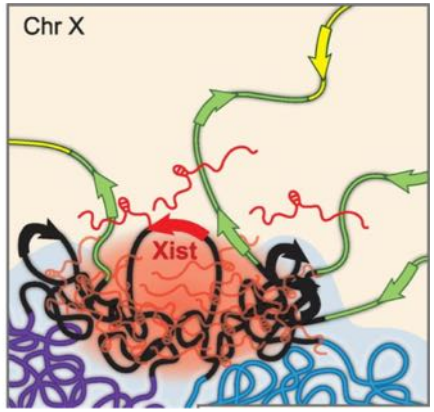
Methyl-seq



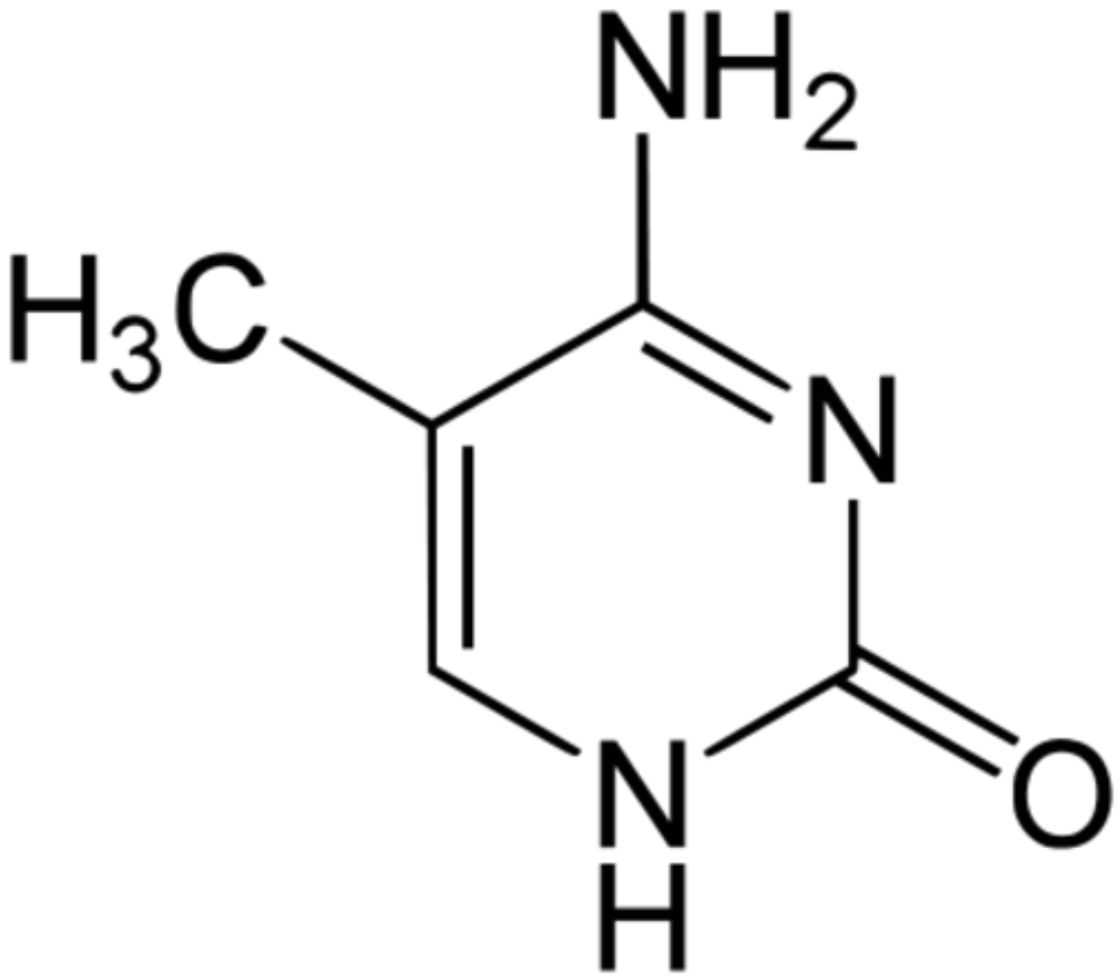
ChIP-seq



Hi-C

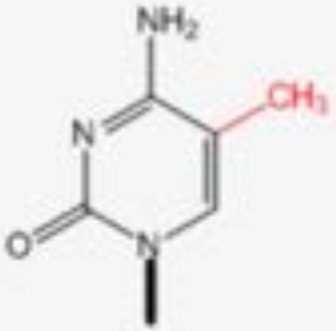


Methyl-seq

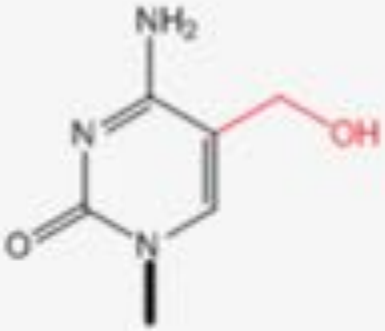


Finding the fifth base: Genome-wide sequencing of cytosine methylation
Lister and Ecker (2009) *Genome Research.* 19: 959-966

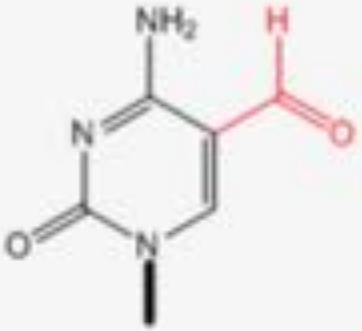
Epigenetic Modifications to DNA



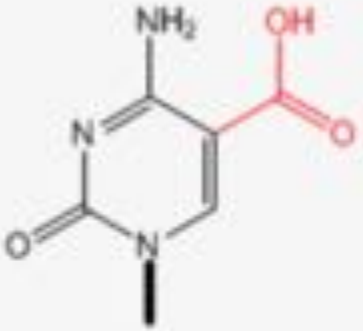
5-mC



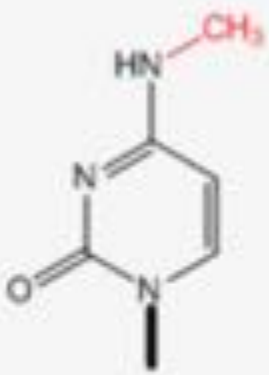
5-hmC



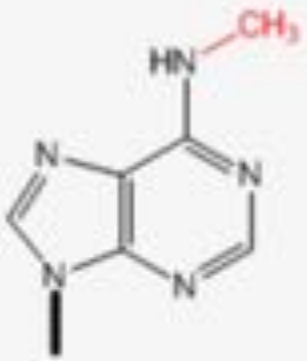
5-fC



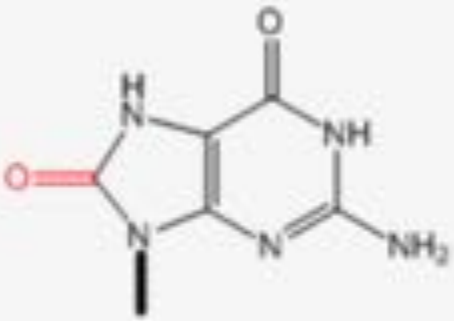
5-caC



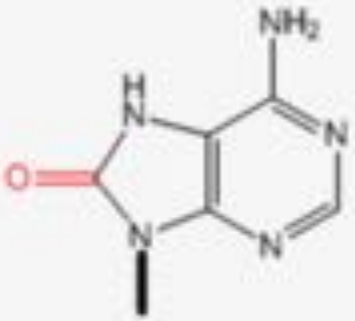
4-mC



6-mA

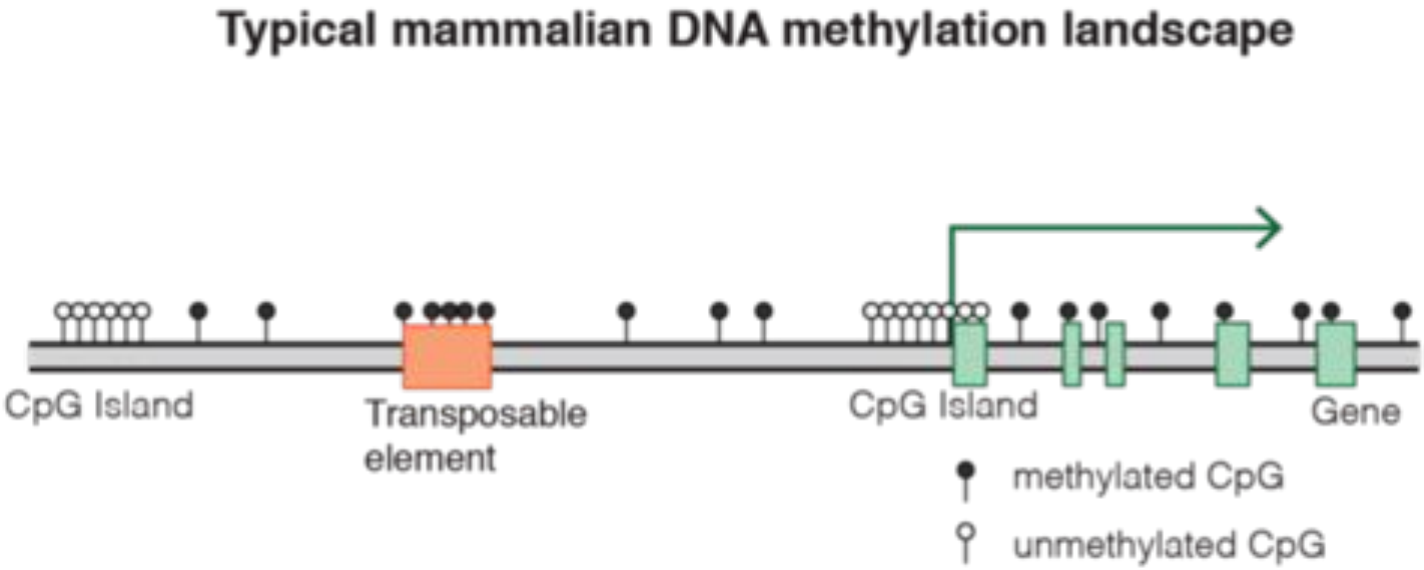


8-oxoG



8-oxoA

Methylation of CpG Islands



CpG islands are (usually) defined as regions with

1) a length greater than 200bp,
2) a G+C content greater than 50%,
3) a ratio of observed to expected CpG greater than 0.6

Methylation in promoter regions correlates negatively with gene expression.

- CpG-dense promoters of actively transcribed genes are never methylated
- In mouse and human, around 60-70% of genes have a CpG island in their promoter region and most of these CpG islands remain unmethylated independently of the transcriptional activity of the gene
- Methylation of DNA itself may physically impede the binding of transcriptional proteins to the gene
- Methylated DNA may be bound by proteins known as methyl-CpG-binding domain proteins (MBDs) that can modify histones, thereby forming compact, inactive chromatin, termed heterochromatin.

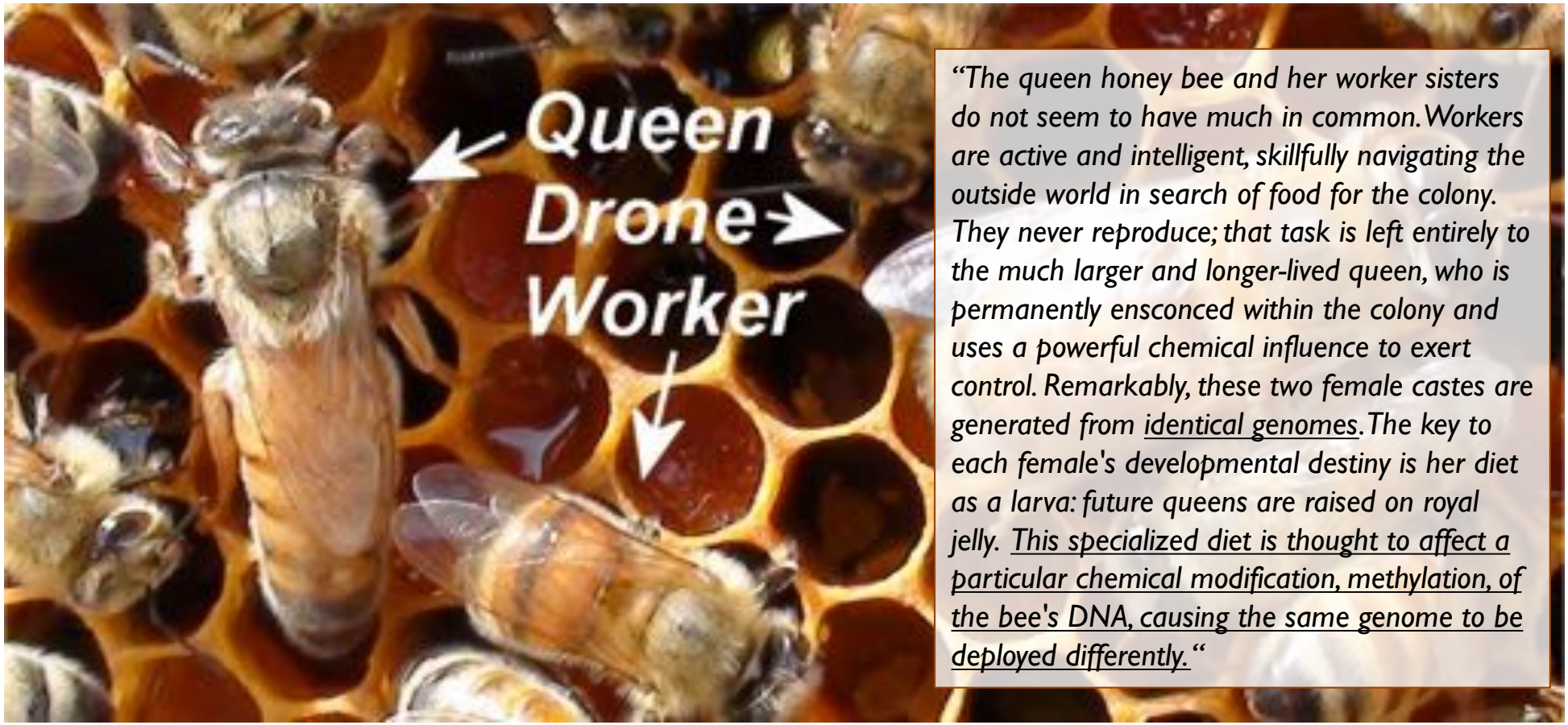
OPEN ACCESS Freely available online

PLoS BIOLOGY

The Honey Bee Epigenomes: Differential Methylation of Brain DNA in Queens and Workers

Frank Lyko[1☯], Sylvain Foret[2☯], Robert Kucharski[3], Stephan Wolf[4], Cassandra Falckenhayn[1], Ryszard Maleszka[3*]

1 Division of Epigenetics, DKFZ-ZMBH Alliance, German Cancer Research Center, Heidelberg, Germany, 2 ARC Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Australia, 3 Research School of Biology, the Australian National University, Canberra, Australia, 4 Genomics and Proteomics Core Facility, German Cancer Research Center, Heidelberg, Germany



"The queen honey bee and her worker sisters do not seem to have much in common. Workers are active and intelligent, skillfully navigating the outside world in search of food for the colony. They never reproduce; that task is left entirely to the much larger and longer-lived queen, who is permanently ensconced within the colony and uses a powerful chemical influence to exert control. Remarkably, these two female castes are generated from identical genomes. The key to each female's developmental destiny is her diet as a larva: future queens are raised on royal jelly. This specialized diet is thought to affect a particular chemical modification, methylation, of the bee's DNA, causing the same genome to be deployed differently."

Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm
Ong-Abdullah, et al (2015) *Nature. doi:10.1038/nature15365*

Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm
Ong-Abdullah, et al (2015) *Nature. doi:10.1038/nature15365*

Somaclonal variation arises in plants and animals when differentiated somatic cells are induced into a pluripotent state, but the resulting clones differ from each other and from their parents. In agriculture, somaclonal variation has hindered the micropropagation of elite hybrids and genetically modified crops, but the mechanism responsible remains unknown. The oil palm fruit 'mantled' abnormality is a somaclonal variant arising from tissue culture that drastically reduces yield, and has largely halted efforts to clone elite hybrids for oil production. Widely regarded as an epigenetic phenomenon, 'mantling' has defied explanation, but here we identify the MANTLED locus using epigenome-wide association studies of the African oil palm Elaeis guineensis. DNA hypomethylation of a LINE retrotransposon related to rice Karma, in the intron of the homeotic gene DEFICIENS, is common to all mantled clones and is associated with alternative splicing and premature termination. **Dense methylation near the Karma splice site (termed the Good Karma epiallele) predicts normal fruit set, whereas hypomethylation (the Bad Karma epiallele) predicts homeotic transformation, parthenocarpy and marked loss of yield.** *Loss of Karma methylation and of small RNA in tissue culture contributes to the origin of mantled, while restoration in spontaneous revertants accounts for non-Mendelian inheritance. The ability to predict and cull mantling at the plantlet stage will facilitate the introduction of higher performing clones and optimize environmentally sensitive land resources.*

Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm
Ong-Abdullah, et al (2015) *Nature. doi:10.1038/nature15365*

Hypomethylation distinguishes genes of some human cancers from their normal counterparts

Andrew P. Feinberg & Bert Vogelstein

Cell Structure and Function Laboratory, The Oncology Center, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

It has been suggested that cancer represents an alteration in DNA, heritable by progeny cells, that leads to abnormally regulated expression of normal cellular genes; DNA alterations such as mutations[1,2], rearrangements[3-5] and changes in methylation[6-8] have been proposed to have such a role. Because of increasing evidence that DNA methylation is important in gene expression (for review see refs 7, 9–11), several investigators have studied DNA methylation in animal tumours, transformed cells and leukaemia cells in culture[8,12-30]. The results of these studies have varied; depending on the techniques and systems used, an increase[12-19], decrease[20-24], or no change[25-29] in the degree of methylation has been reported. To our knowledge, however, primary human tumour tissues have not been used in such studies. We have now examined DNA methylation in human cancer with three considerations in mind: (1) the methylation pattern of specific genes, rather than total levels of methylation, was determined; (2) human cancers and adjacent analogous normal tissues, unconditioned by culture media, were analysed; and (3) the cancers were taken from patients who had received neither radiation nor chemotherapy. In four of five patients studied, representing two histological types of cancer, substantial hypomethylation was found in genes of cancer cells compared with their normal counterparts. This hypomethylation was progressive in a metastasis from one of the patients.

and (3) *Hpa*II and *Hha*I cleavage sites should be present in the regions of the genes.

The first cancer studied was a grade D (ref. 43), moderately well differentiated adenocarcinoma of the colon from a 67-yr-old male. Tissue was obtained from the cancer itself and also from colonic mucosa stripped from the colon at a site just outside the histologically proven tumour margin. Figure 1 shows the pattern of methylation of the studied genes. Before digestion with restriction enzymes, all DNA samples used in the study had a size >25,000 base pairs (bp). After *Hpa*II cleavage, hybridization with a probe made from a cDNA clone of human growth hormone (HGH) showed that significantly more of the DNA was digested to low-molecular weight fragments in DNA from the cancer (labelled C in Fig. 1) than in DNA from the normal colonic mucosa (labelled N). In the hybridization conditions used, the HGH probe detected the human growth hormone genes as well as the related chorionic somatotropin

Table 1 Quantitation of methylation of specific genes in human cancers and adjacent analogous normal tissues

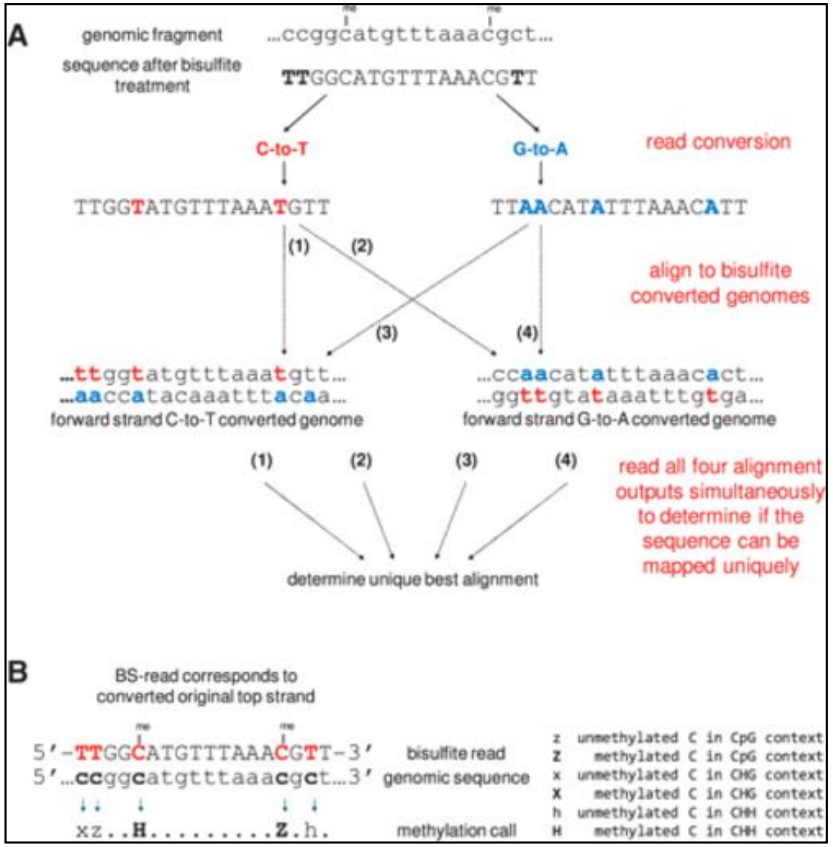
Patient	Carcinoma	Probe	Enzyme	% Hypomethylated fragments		
				N	C	M
1	Colon	HGH	*Hpa*II	<10	35	—
			*Hha*I	<10	39	—
		γ-Globin	*Hpa*II	<10	52	—
			*Hha*I	<10	39	—
		α-Globin	*Hpa*II	<10	<10	—
			*Hha*I	<10	<10	—
2	Colon	HGH	*Hpa*II	<10	76	—
			*Hha*I	<10	85	—
		γ-Globin	*Hpa*II	<10	58	—
			*Hha*I	<10	23	—
		α-Globin	*Hpa*II	<10	<10	—
			*Hha*I	<10	<10	—
3	Colon	HGH	*Hpa*II	<10	41	—
			*Hha*I	<10	38	—
		γ-Globin	*Hpa*II	<10	50	—

Nature 301, 89 - 92 (06 January 1983); doi:10.1038/301089a0

Bisulfite Conversion

Treating DNA with sodium bisulfite will convert <u>un</u>methylated C to T

- 5-MethylC will be protected and not change, so can look for differences when mapping
- Requires great care when analyzing reads, since the complementary strand will also be converted (G to A)
- Typically analyzed by mapping to a "reduced alphabet" where we assume all Cs are converted to Ts once on the forward strand and once on the reverse



Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications
Krueger and Andrews (2010) *Bioinformatics*. 27 (11): 1571-1572.

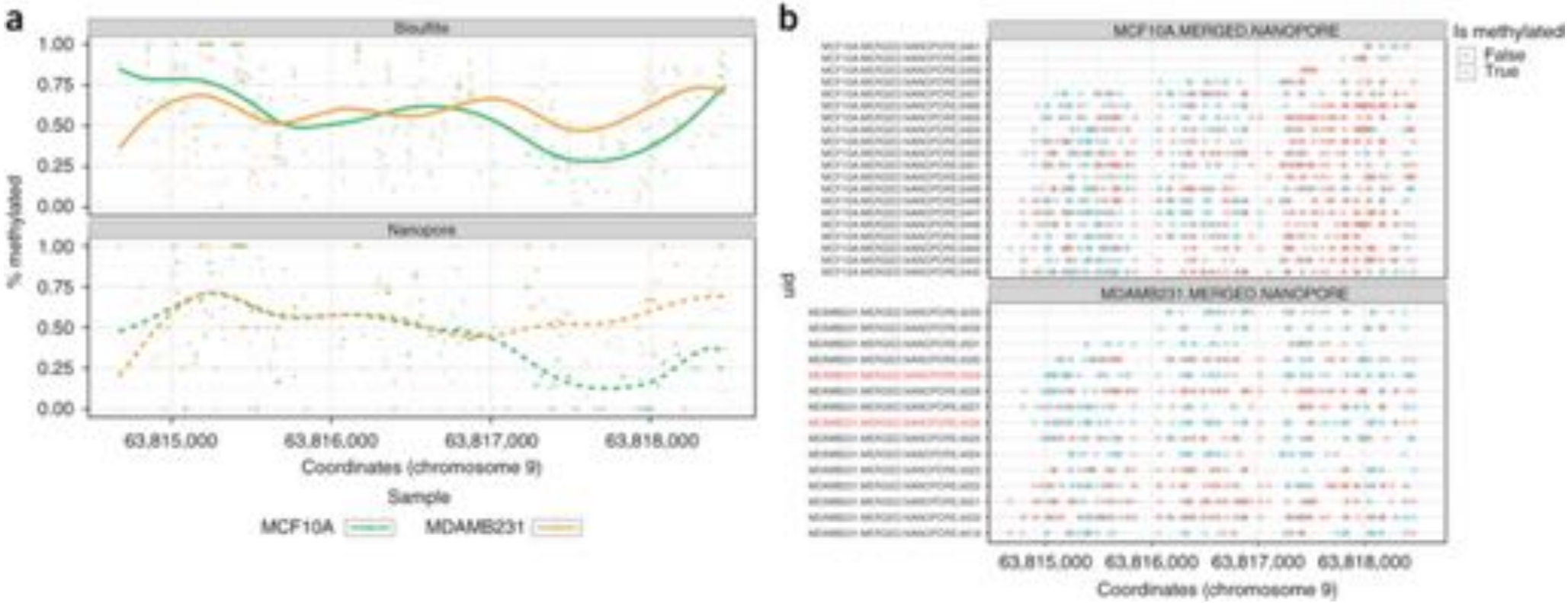
Bisulfite Conversion

Tr
w

-
-
-

Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications
Krueger and Andrews (2010) *Bioinformatics*. 27 (11): 1571-1572.

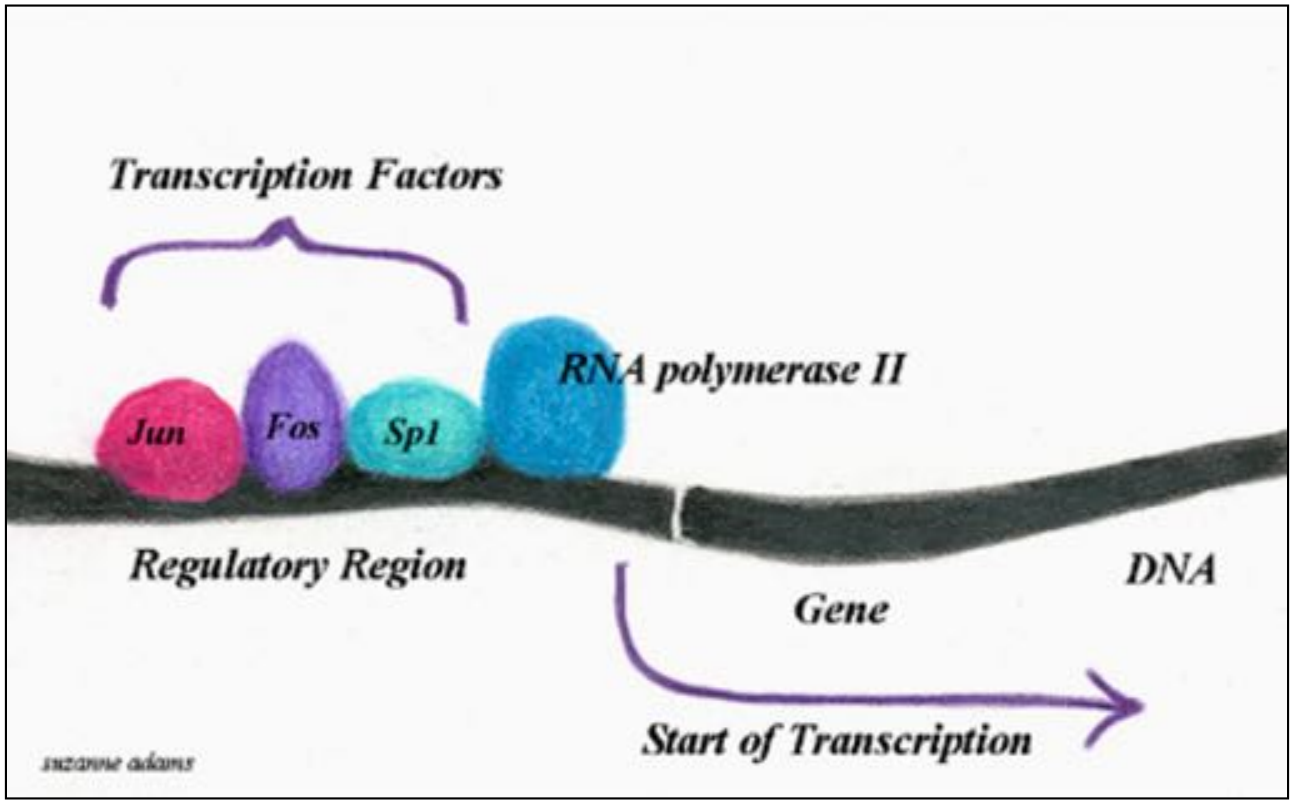
Methylation changes in cancer detected by Nanopore Sequencing



Comparison of bisulfite sequencing and nanopore-based R7.3 data in reduced representation data sets from cancer and normal cells. (a) Raw data (points) and smoothed data (lines) for methylation, as determined by bisulfite sequencing (top) and nanopore-based sequencing using an R7.3 pore (bottom), in a genomic region from the human mammary epithelial cell line MCF10A (green) and metastatic mammary epithelial cell line MDA-MB-231 (orange). (b) Same region as in a but with individual nanopore reads plotted separately. Each CpG that can be called is a point. Blue indicates methylated; red indicates unmethylated.

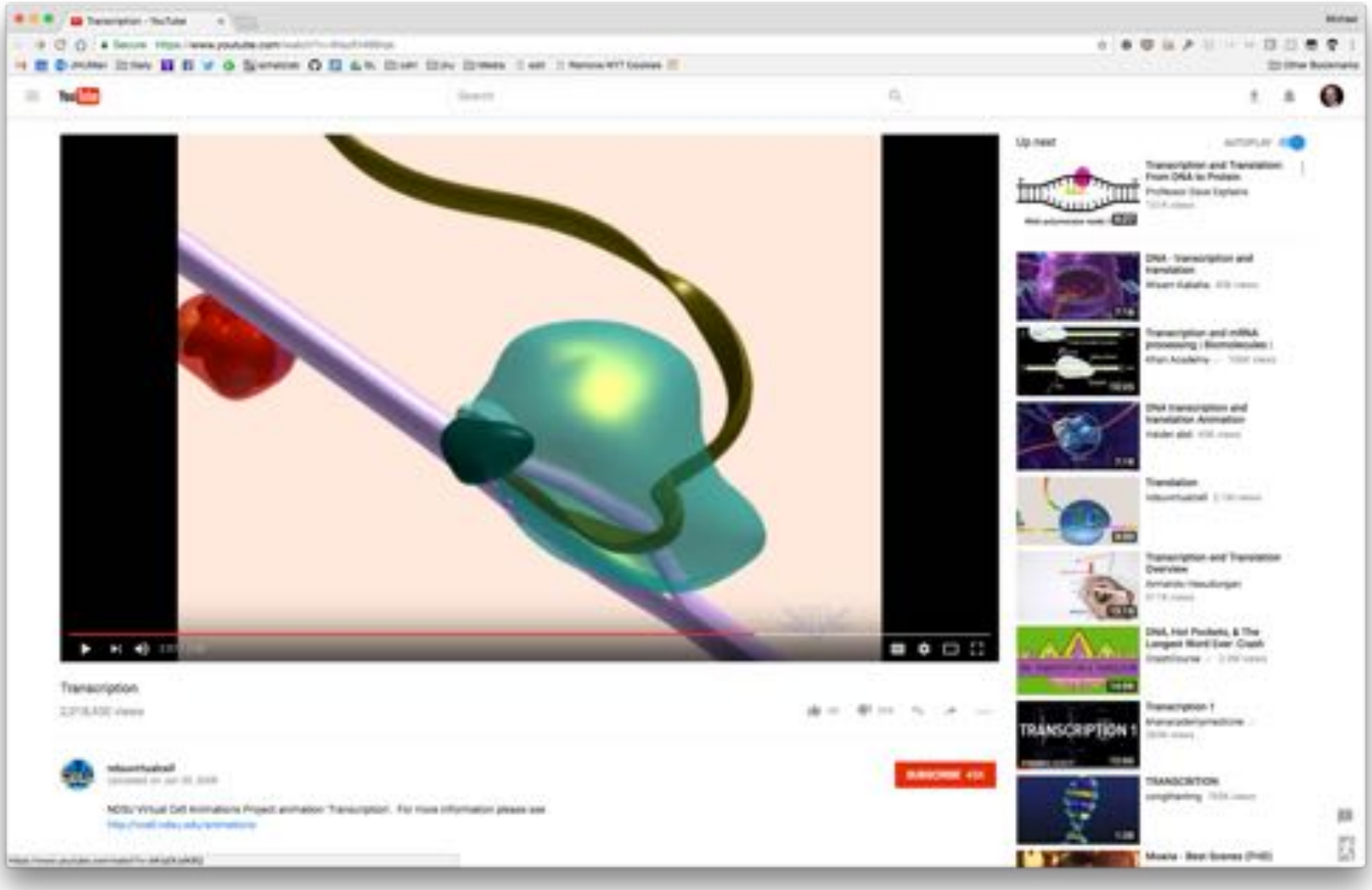
Detecting DNA cytosine methylation using nanopore sequencing
Simpson, Workman, Zuzarte, David, Dursi, Timp (2017) Nature Methods. doi:10.1038/nmeth.4184

ChIP-seq



Genome-wide mapping of in vivo protein-DNA interactions.
Johnson *et al* (2007) *Science.* 316(5830):1497-502

Transcription



https://www.youtube.com/watch?v=WsofH466lqk