Lecture 12. Functional Genomics I

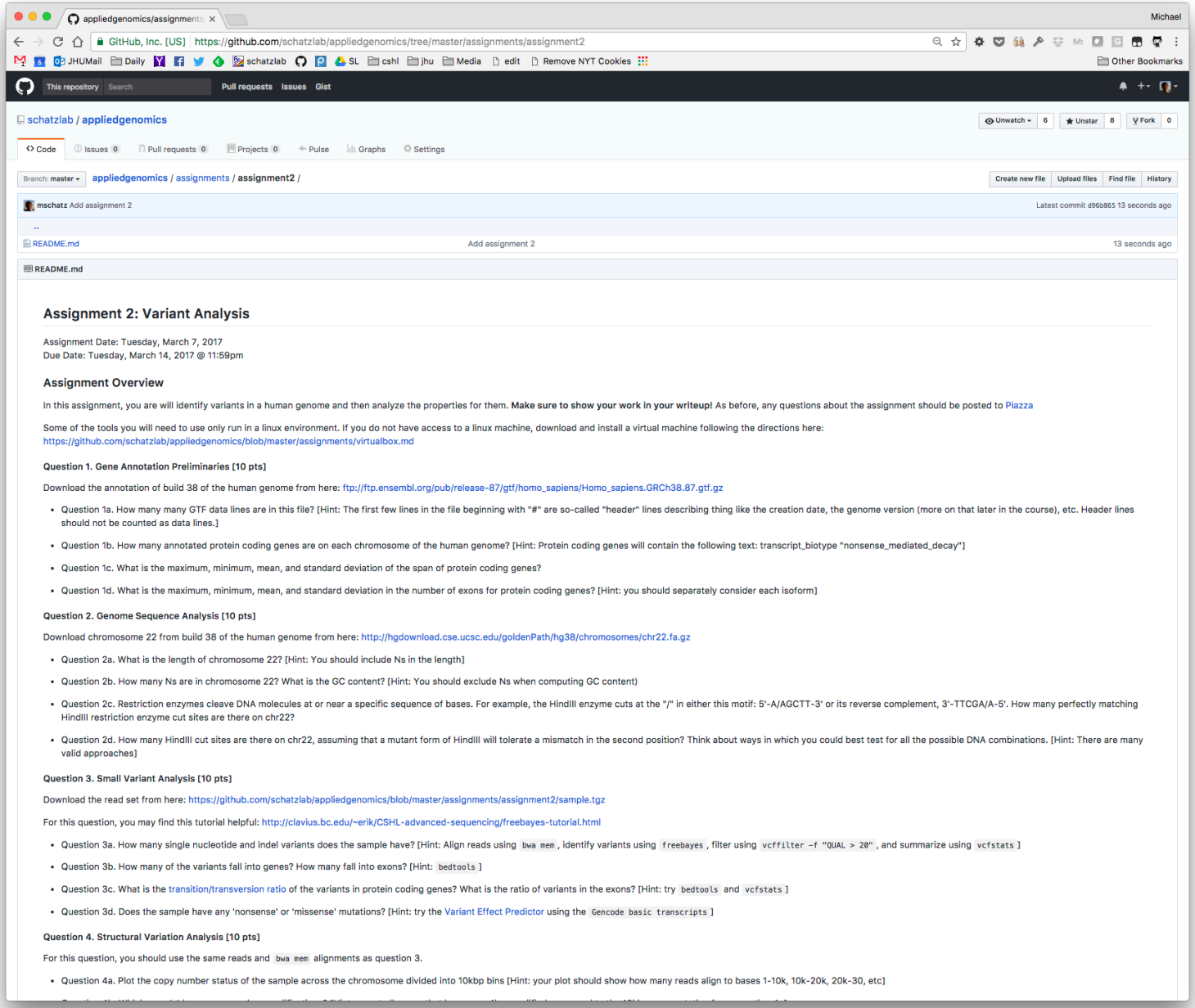
Michael Schatz

March 9, 2017

JHU 600.649: Applied Comparative Genomics

Assignment 2

Due: Tuesday March 14 @ 11:59pm



appliedgenomics/assignments

GitHub, Inc. [US] https://github.com/schatzlab/appliedgenomics/tree/master/assignments/assignment2

schatzlab / appliedgenomics

Code Issues 0 Pull requests 0 Projects 0 Pulse Graphs Settings

Branch: master appliedgenomics / assignments / assignment2 /

mschatz Add assignment 2 — Latest commit d96b865 13 seconds ago

README.md — Add assignment 2 — 13 seconds ago

README.md

Assignment 2: Variant Analysis

Assignment Date: Tuesday, March 7, 2017
Due Date: Tuesday, March 14, 2017 @ 11:59pm

Assignment Overview

In this assignment, you are will identify variants in a human genome and then analyze the properties for them. **Make sure to show your work in your writeup!** As before, any questions about the assignment should be posted to Piazza

Some of the tools you will need to use only run in a linux environment. If you do not have access to a linux machine, download and install a virtual machine following the directions here:
https://github.com/schatzlab/appliedgenomics/blob/master/assignments/virtualbox.md

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 chromosome of the human genome? [Hint: Protein coding genes will contain the following text: transcript_biotype "nonsense_mediated_decay"]
- Question 1c. What is the maximum, minimum, mean, and standard deviation of the span of protein coding genes?
- 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]

Question 2. Genome Sequence Analysis [10 pts]

Download chromosome 22 from build 38 of the human genome from here: http://hgdownload.cse.ucsc.edu/goldenPath/hg38/chromosomes/chr22.fa.gz

- Question 2a. What is the length of chromosome 22? [Hint: You should include Ns in the length]
- Question 2b. How many Ns are in chromosome 22? What is the GC content? [Hint: You should exclude Ns when computing GC content)
- Question 2c. Restriction enzymes cleave DNA molecules at or near a specific sequence of bases. For example, the HindIII enzyme cuts at the "/" in either this motif: 5'-A/AGCTT-3' or its reverse complement, 3'-TTCGA/A-5'. How many perfectly matching HindIII restriction enzyme cut sites are there on chr22?
- Question 2d. How many HindIII cut sites are there on chr22, assuming that a mutant form of HindIII will tolerate a mismatch in the second position? Think about ways in which you could best test for all the possible DNA combinations. [Hint: There are many valid approaches]

Question 3. Small Variant Analysis [10 pts]

Download the read set from here: https://github.com/schatzlab/appliedgenomics/blob/master/assignments/assignment2/sample.tgz

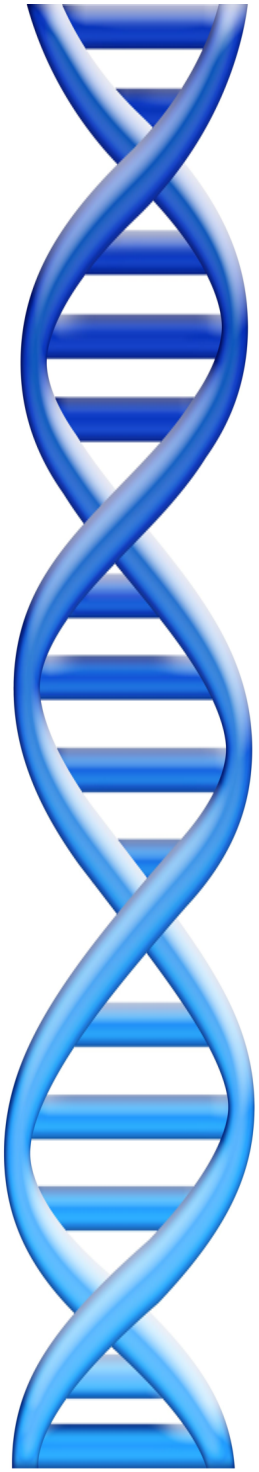
For this question, you may find this tutorial helpful: http://clavius.bc.edu/~erik/CSHL-advanced-sequencing/freebayes-tutorial.html

- Question 3a. How many single nucleotide and indel variants does the sample have? [Hint: Align reads using `bwa mem`, identify variants using `freebayes`, filter using `vcffilter -f "QUAL > 20"`, and summarize using `vcfstats`]
- Question 3b. How many of the variants fall into genes? How many fall into exons? [Hint: `bedtools`]
- Question 3c. What is the transition/transversion ratio of the variants in protein coding genes? What is the ratio of variants in the exons? [Hint: try `bedtools` and `vcfstats`]
- Question 3d. Does the sample have any 'nonsense' or 'missense' mutations? [Hint: try the Variant Effect Predictor using the `Gencode basic transcripts`]

Question 4. Structural Variation Analysis [10 pts]

For this question, you should use the same reads and `bwa mem` alignments as question 3.

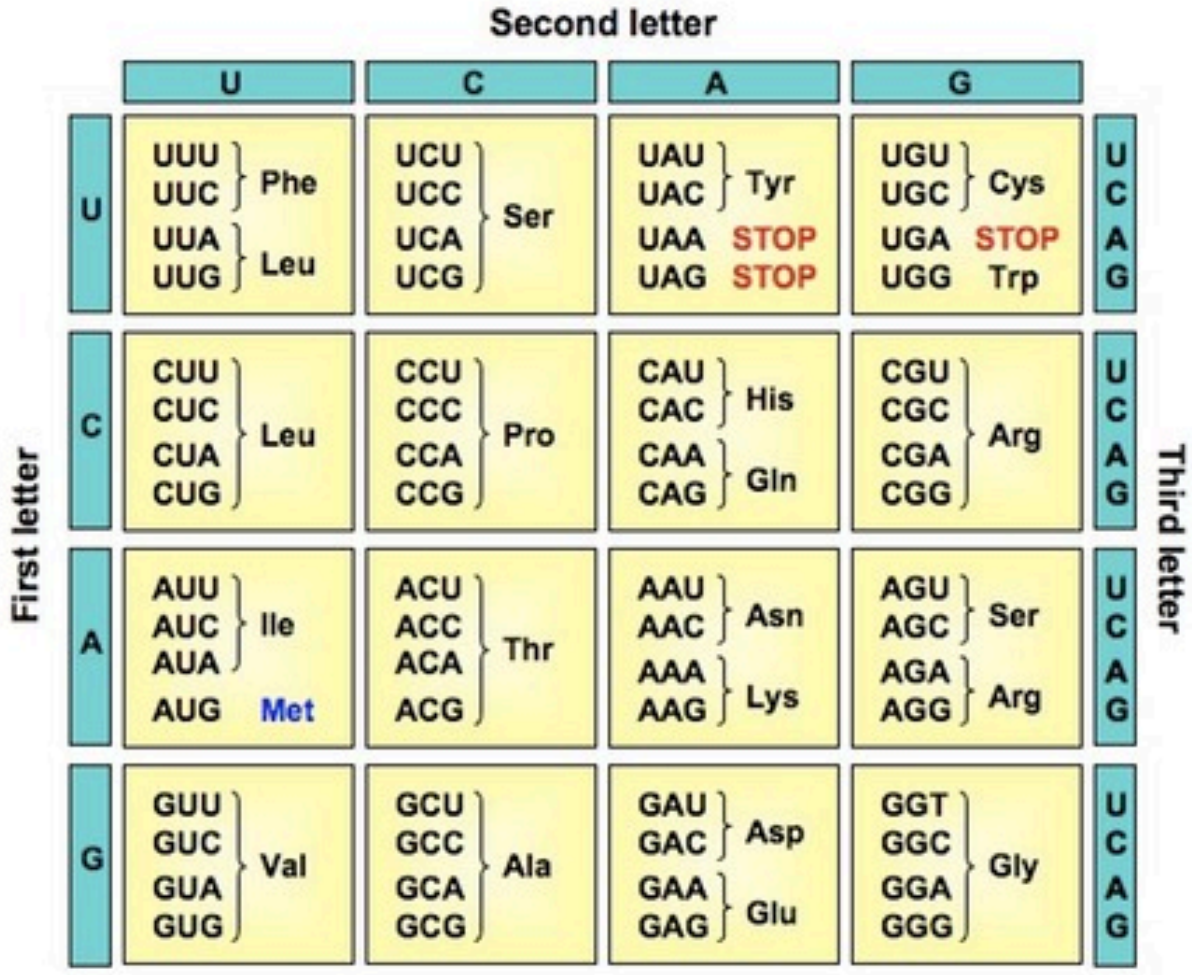
- Question 4a. Plot the copy number status of the sample across the chromosome divided into 10kbp bins [Hint: your plot should show how many reads align to bases 1-10k, 10k-20k, 20k-30, etc]



Genome Annotation

1. Alignment to other genomes
2. Prediction aka "Gene Finding"
3. Experimental & Functional Assays

Genetic Code



Key:

Ala = Alanine (**A**)
Arg = Arginine (**R**)
Asn = Asparagine (**N**)
Asp = Aspartate (**D**)
Cys = Cysteine (**C**)
Gln = Glutamine (**Q**)
Glu = Glutamate (**E**)
Gly = Glycine (**G**)
His = Histidine (**H**)
Ile = Isoleucine (**I**)
Leu = Leucine (**L**)
Lys = Lysine (**K**)
Met = Methionine (**M**)
Phe = Phenylalanine (**F**)
Pro = Proline (**P**)
Ser = Serine (**S**)
Thr = Threonine (**T**)
Trp = Tryptophan (**W**)
Tyr = Tyrosine (**Y**)
Val = Valine (**V**)

Flipping a Biased Coin

P(heads) = 61/64 (95.4%)   P(tails) = 3/64 (4.6%)

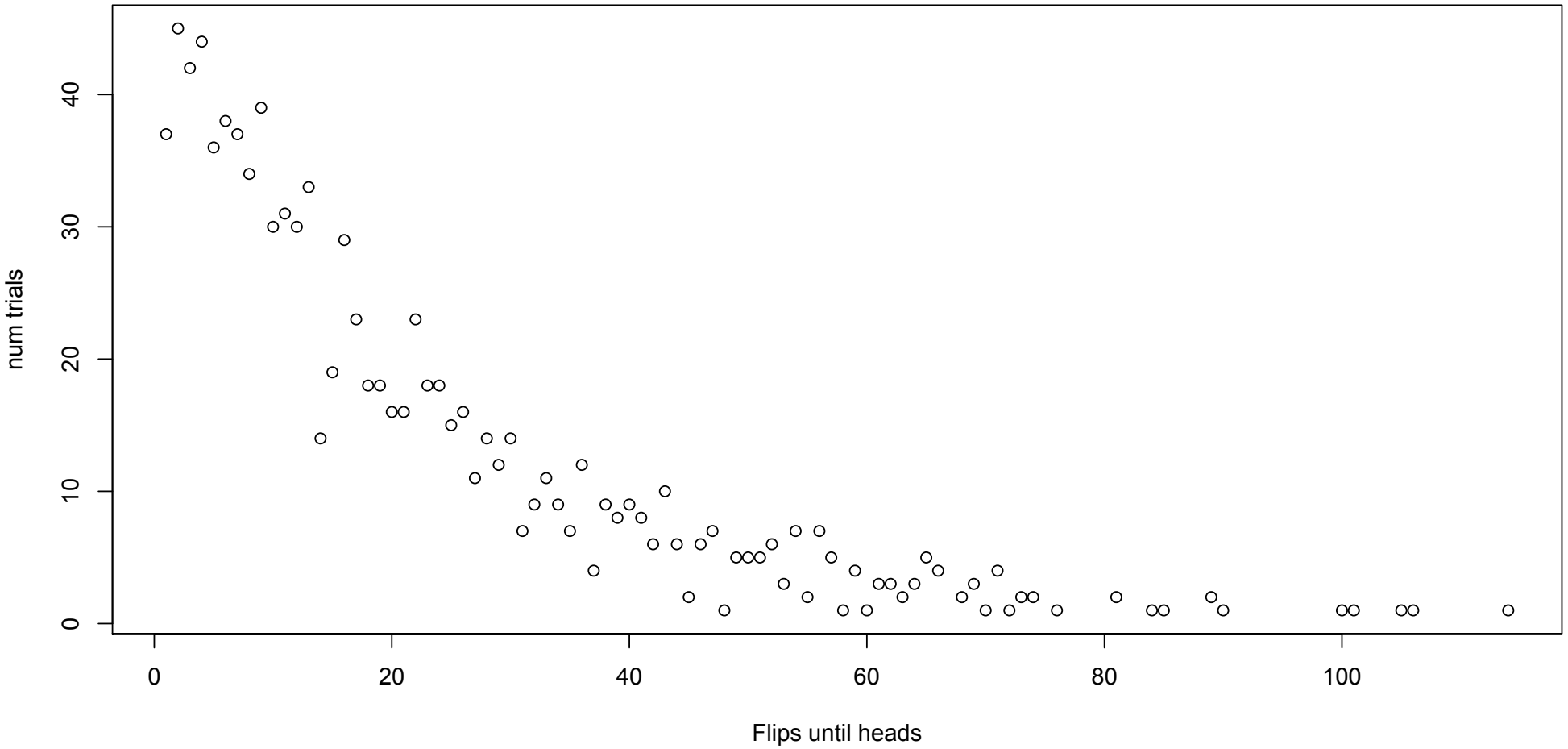
How many flips until my first tail?

```
$ ./coinflip.pl 0.046875 1000
0:	HHHHHHHHHHHHHHT	15
1:	HHHHHHT	7
2:	HHHHHHHHHHHT	12
3:	HHHHHHHHHHHHHHHHHHHHHHHT	24
4:	HT	2
5:	HHHHHHHHHHHHHT	14
6:	HHHHHHHHHT	10
7:	HHHHHHHHHHHHHT	14
8:	HHHHHT	6
9:	HHHHHHHHHHT	11
10:	HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
11:	HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHT	40
12:	HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHT	45
13:	HHHT	4
14:	HHHHHHHHHHHHHHT	15
15:	HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHT	39
16:	HHHHHT	6
17:	HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHT	38
18:	HHHHHHHHHHHHHHHHHHHHHHHHHT	26
19:	HHHHHHHHHHHT	12
```

Flipping a Biased Coin

P(heads) = 61/64 (95.4%) P(tails) = 3/64 (4.6%)

How many flips until my first tail?

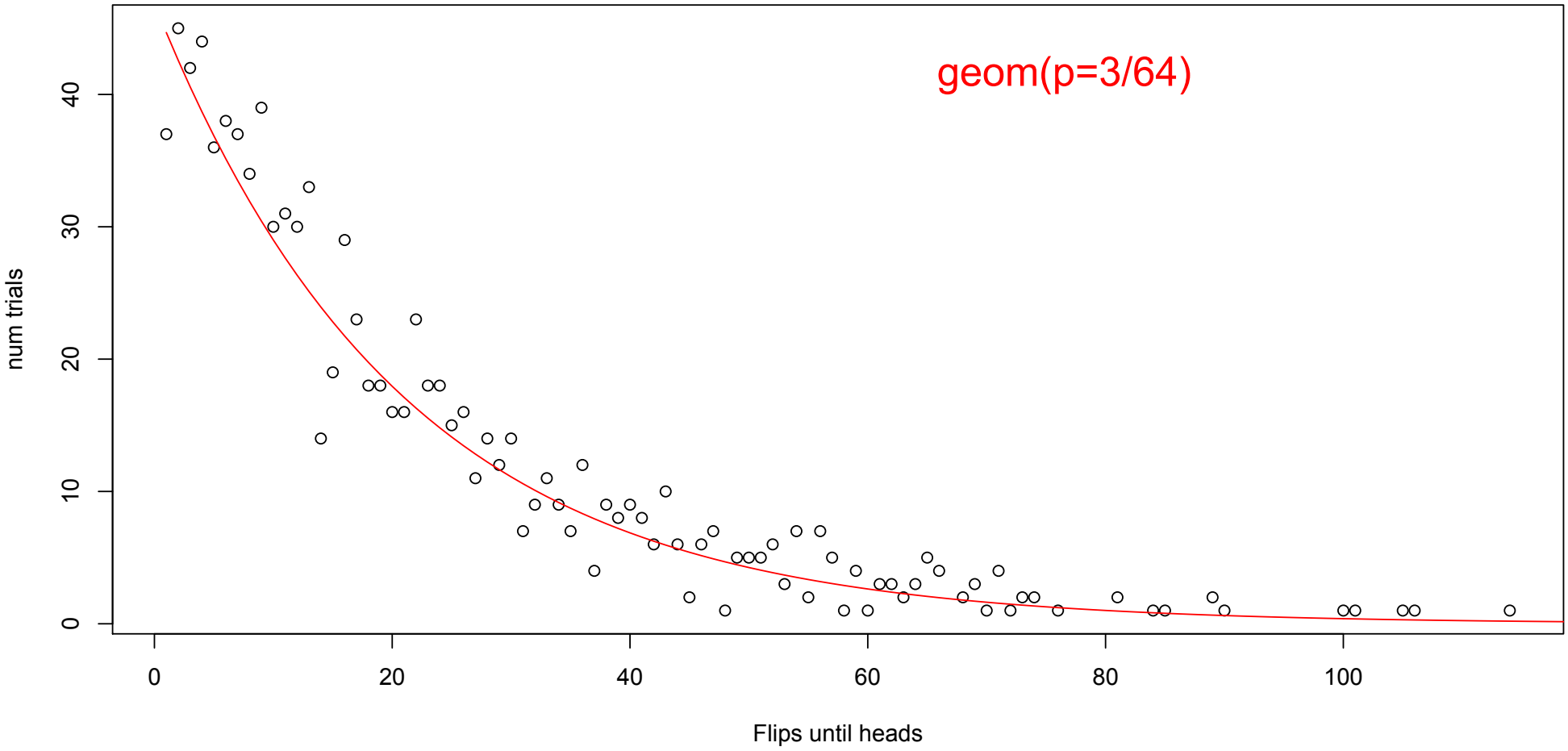


Flipping a Biased Coin

P(heads) = 61/64 (95.4%) P(tails) = 3/64 (4.6%)

How many flips until my first tail?

Geometric Distribution: $P(X=x) = p_{heads}^{x-1} p_{tails}$

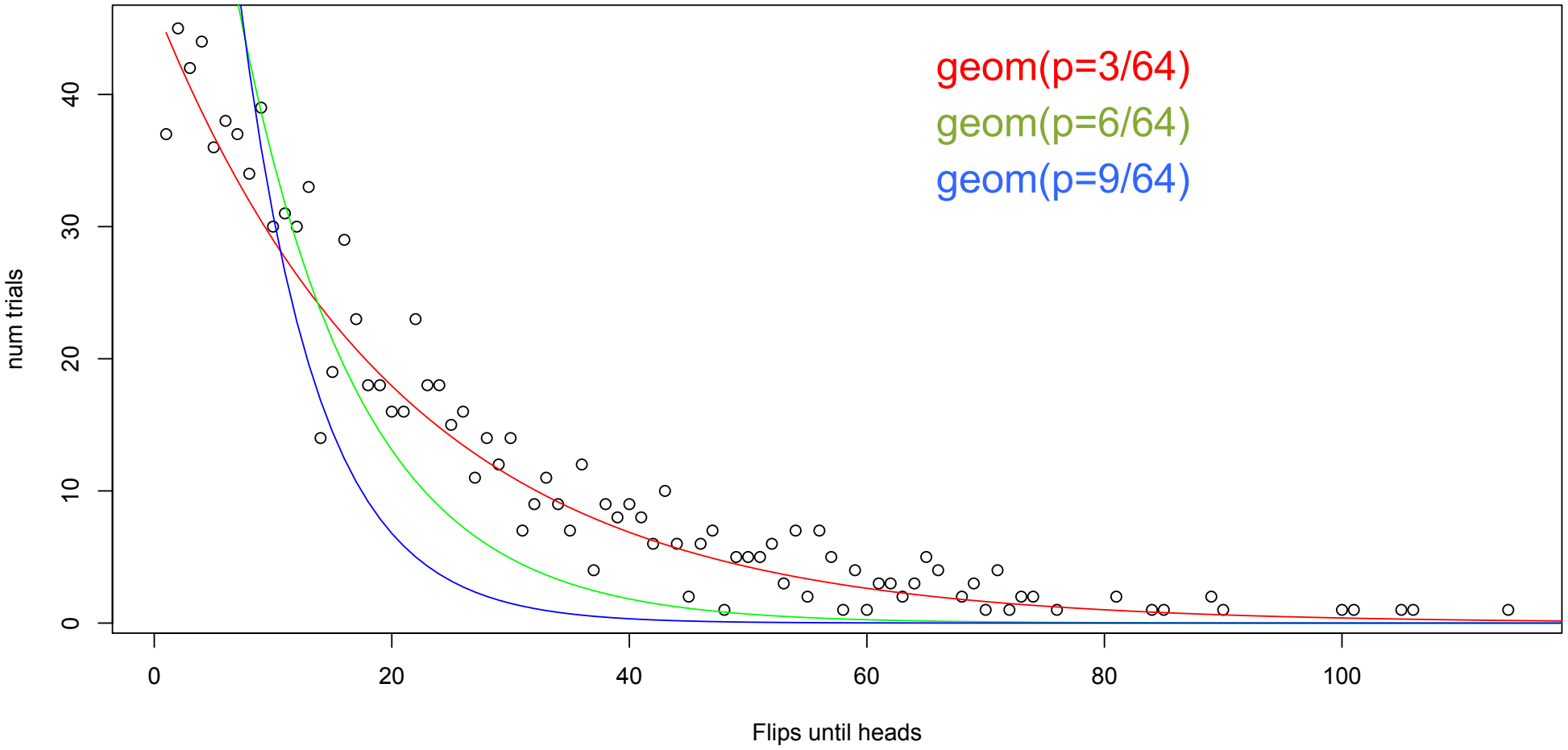


Flipping a Biased Coin

P(heads) = 61/64 (95.4%) P(tails) = 3/64 (4.6%)

How many flips until my first tail?

Geometric Distribution: $P(X=x) = p_{heads}^{x-1} p_{tails}$

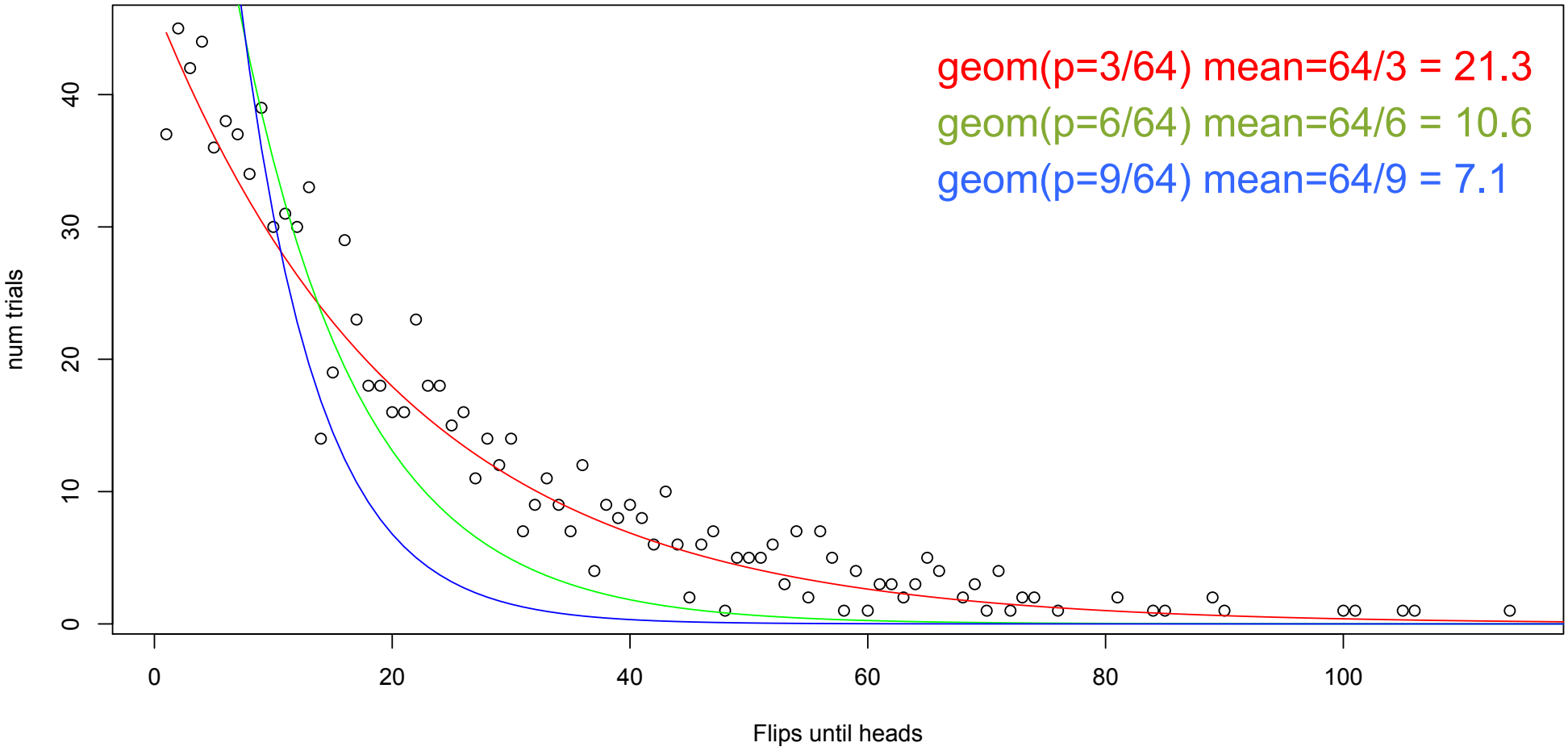


Flipping a Biased Coin

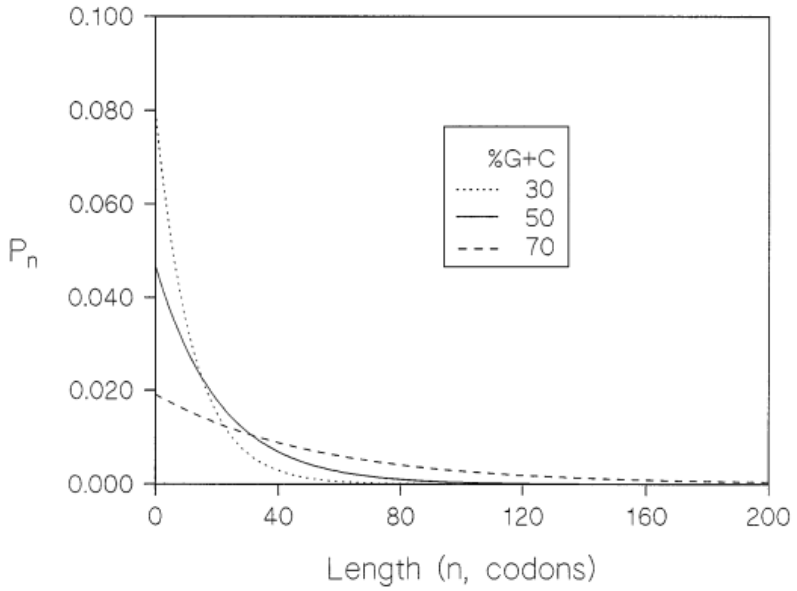
P(heads) = 61/64 (95.4%) P(tails) = 3/64 (4.6%)

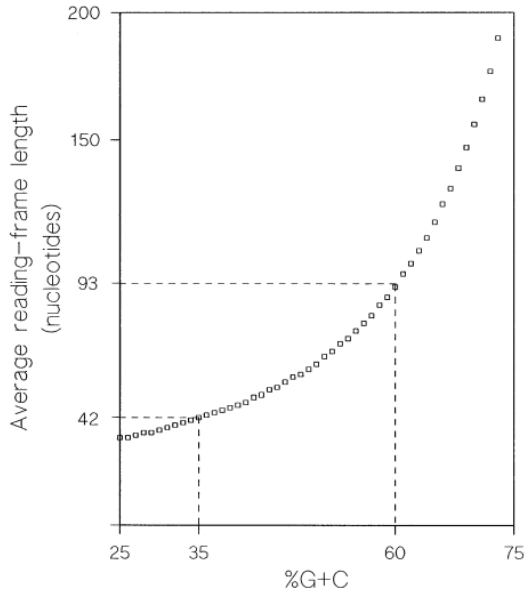
How many flips until my first tail?

Geometric Distribution: $P(X=x) = p_{heads}^{x-1} p_{tails}$



Stop Codon Frequencies



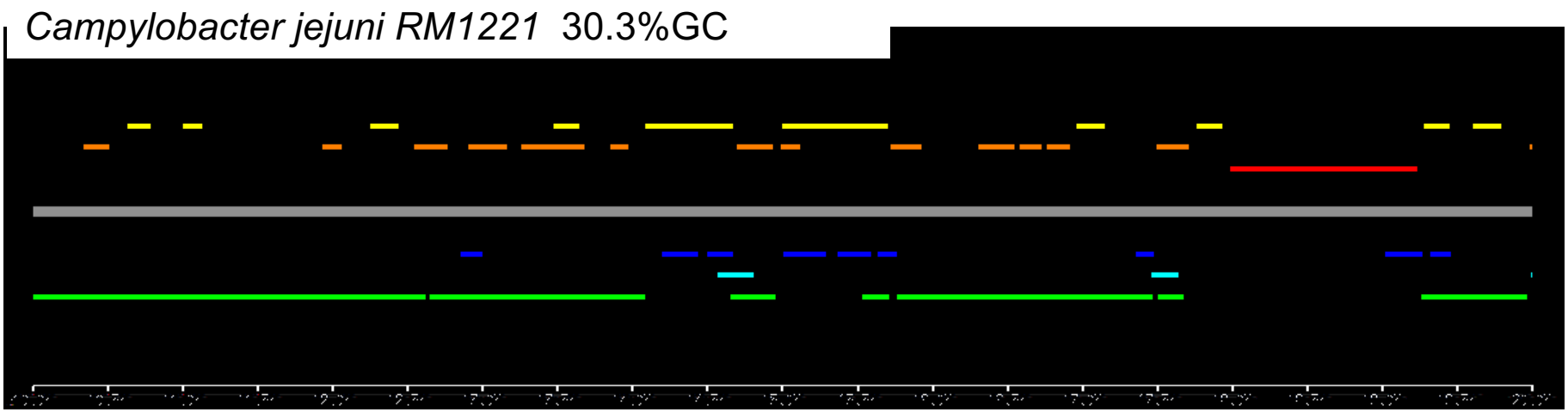


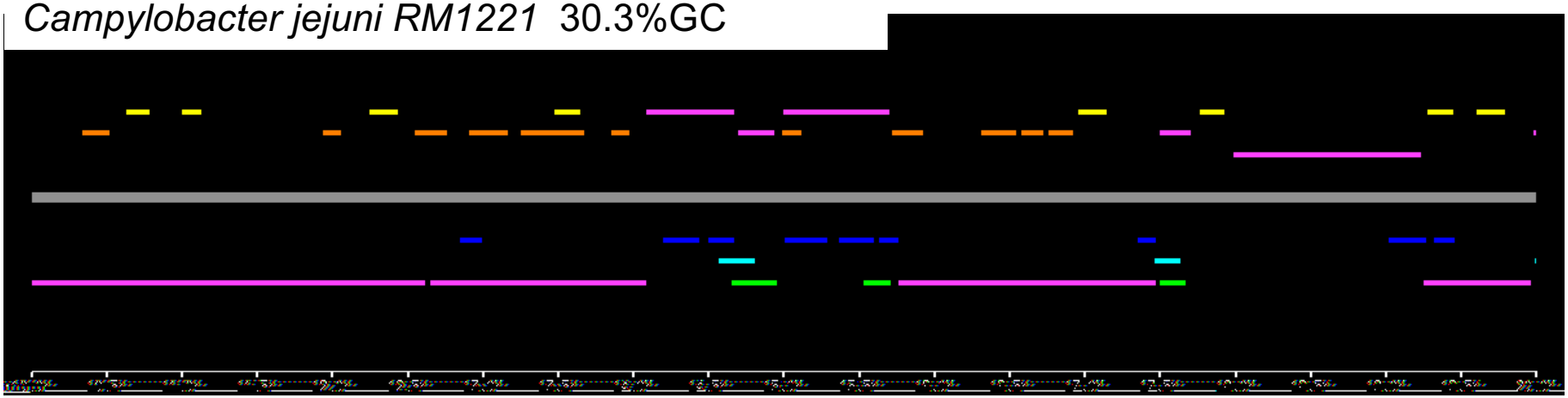
If the sequence is mostly A+T, then likely to form stop codons by chance!

In High A+T (Low G+C):
Random ORFs should be short; long ORFs are more likely to be true genes

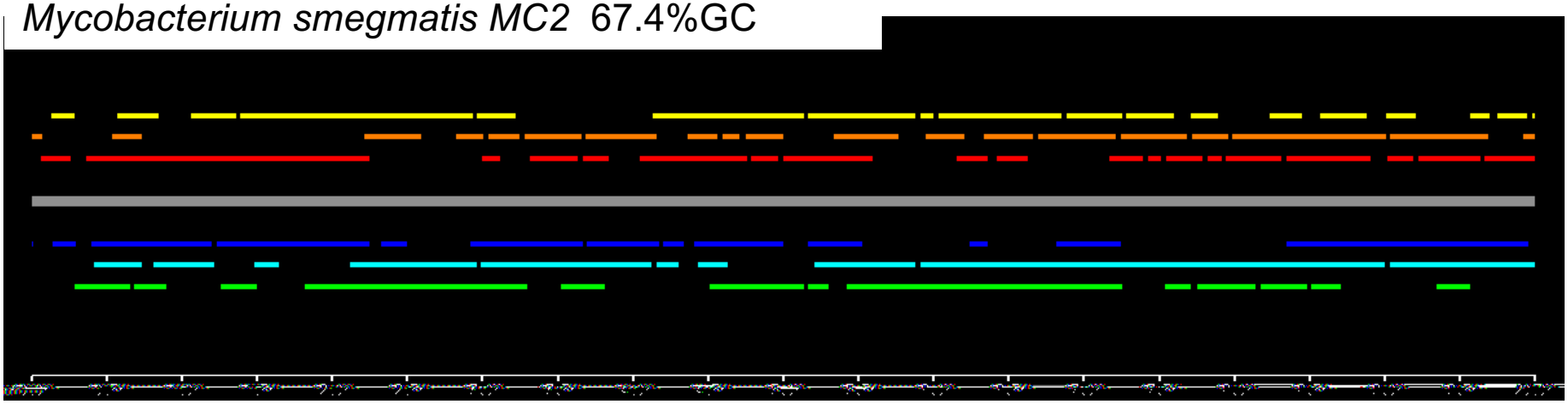
In High G+C (Low A+T):
Random ORFs will be longer; harder to identify true genes

A relationship between GC content and coding-sequence length.
Oliver & Marín (1996) J Mol Evol. 43(3):216-23.

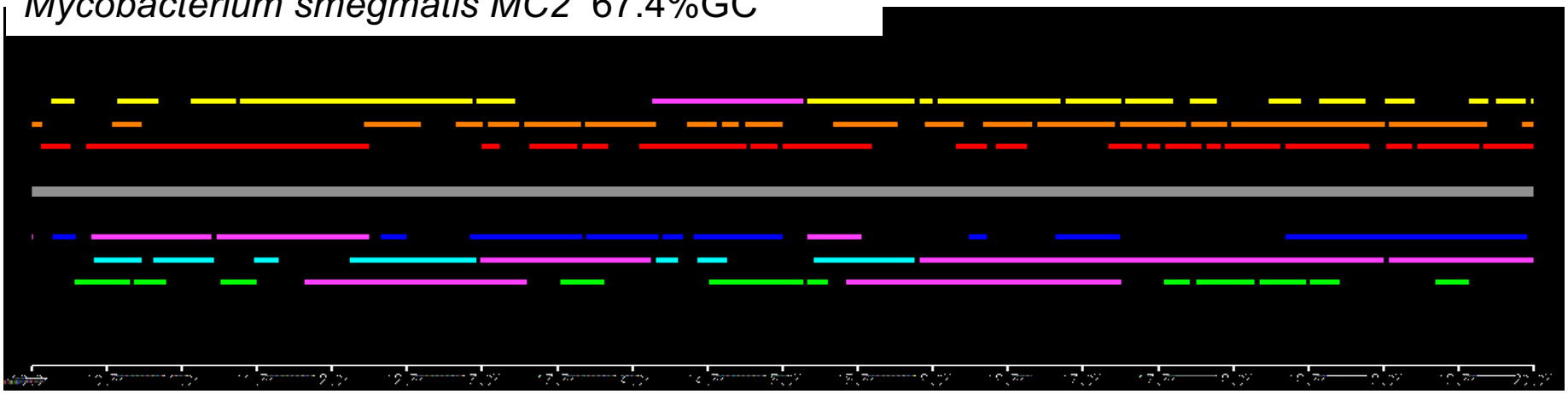

Campylobacter jejuni RM1221 30.3%GC

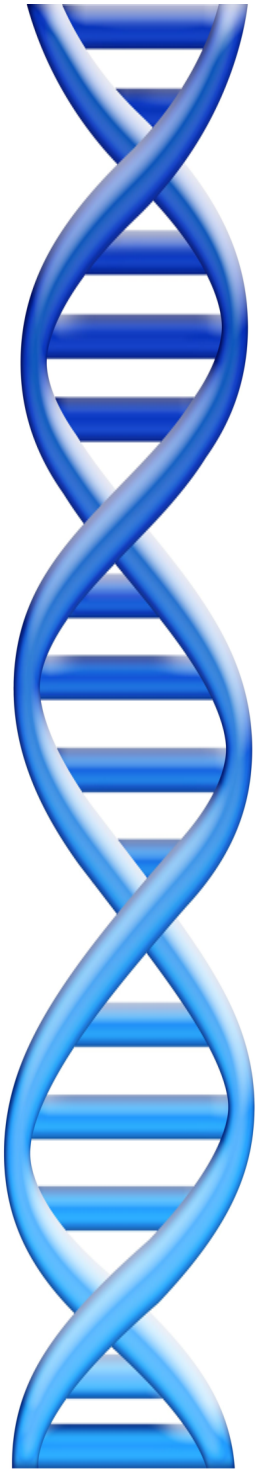

Campylobacter jejuni RM1221 30.3%GC

Mycobacterium smegmatis MC2 67.4%GC



Mycobacterium smegmatis MC2 67.4%GC



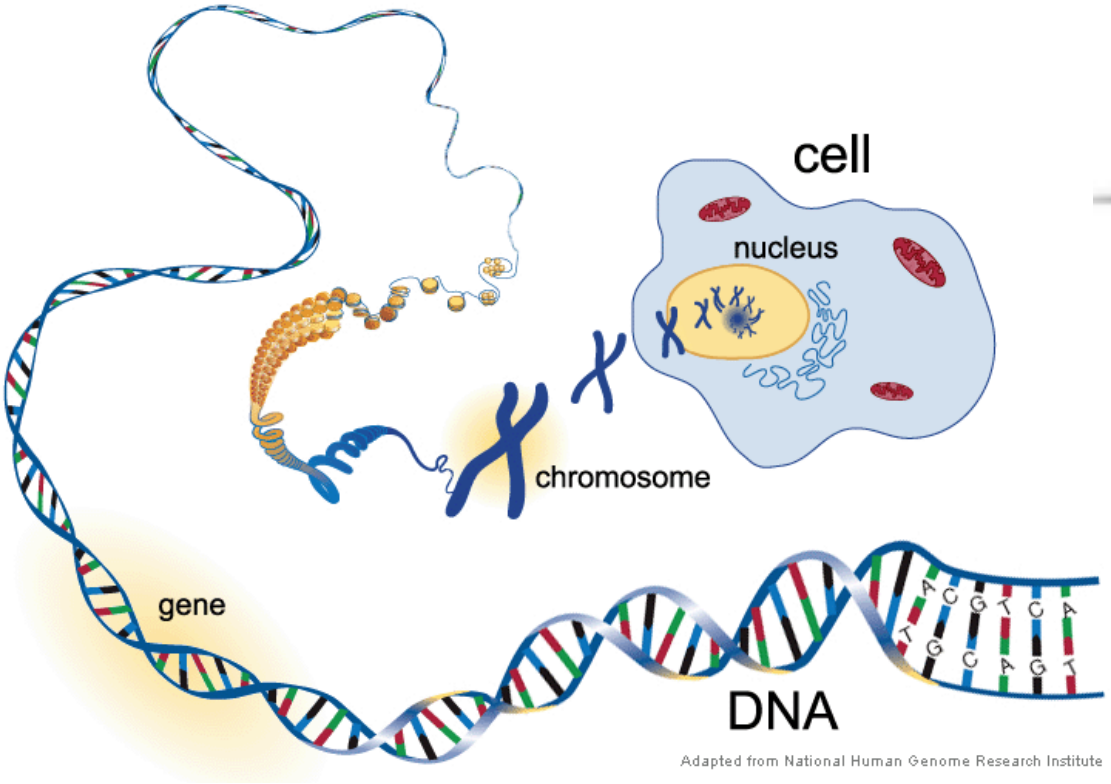


Genome Annotation

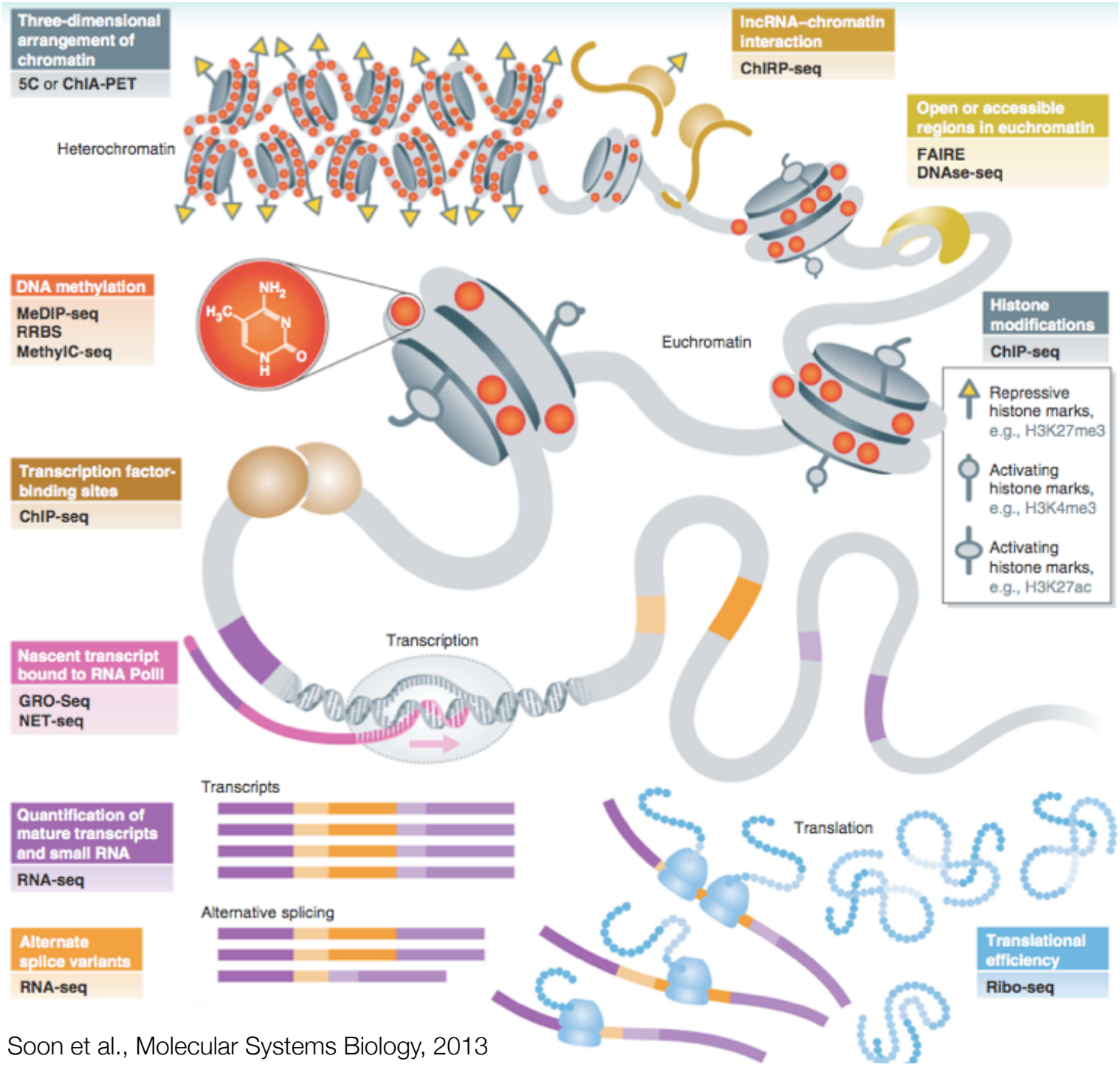
1. Alignment to other genomes
2. Prediction aka "Gene Finding"
3. Experimental & Functional Assays

Sequencing techniques

Much of the capacity is used to sequence genomes (or exomes) of individuals…



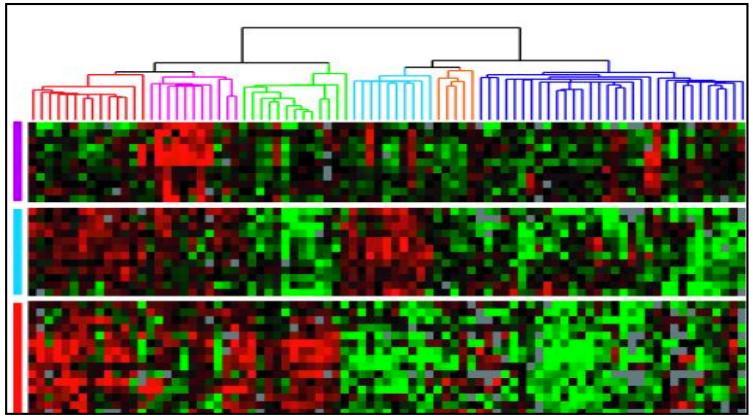
… but biology is much more than just genomes…



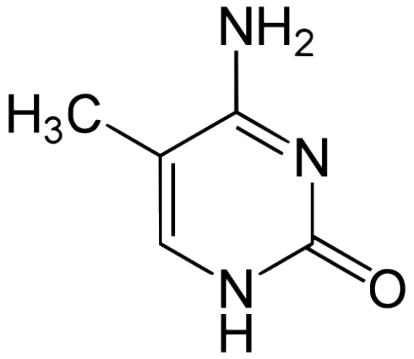
Soon et al., Molecular Systems Biology, 2013

*-seq in 4 short vignettes

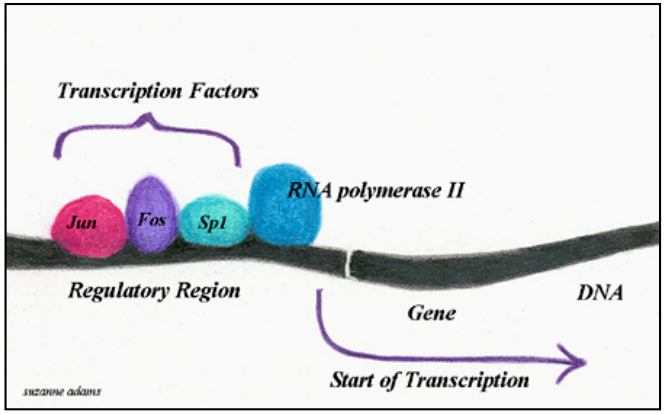
RNA-seq



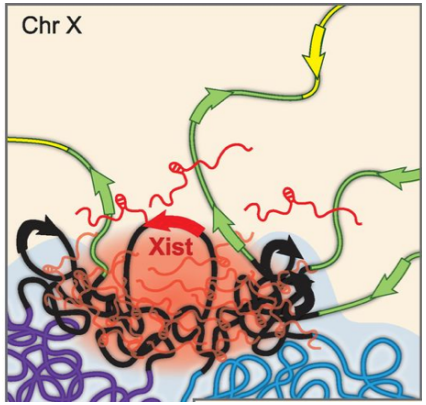
Methyl-seq



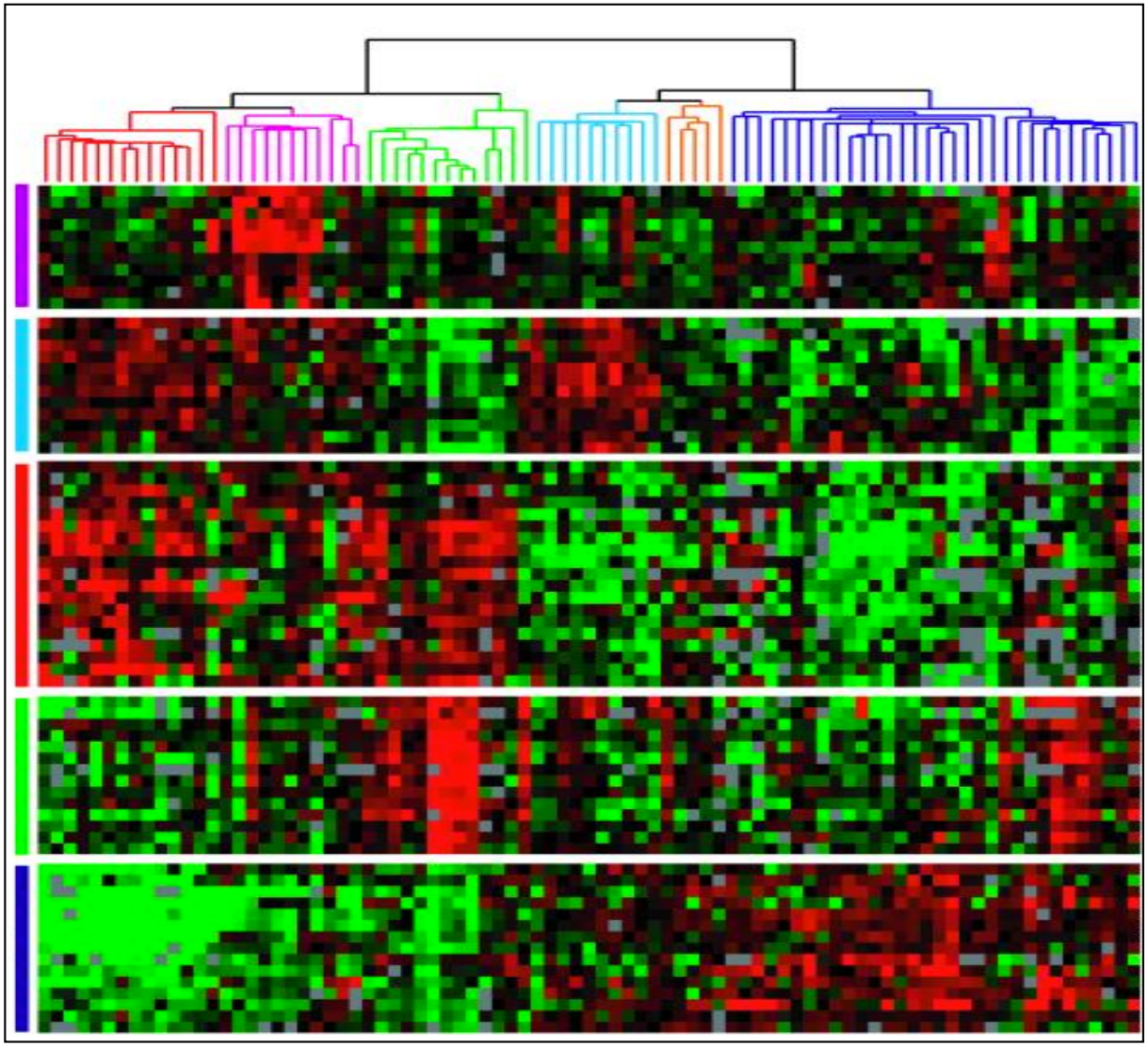
ChIP-seq



Hi-C

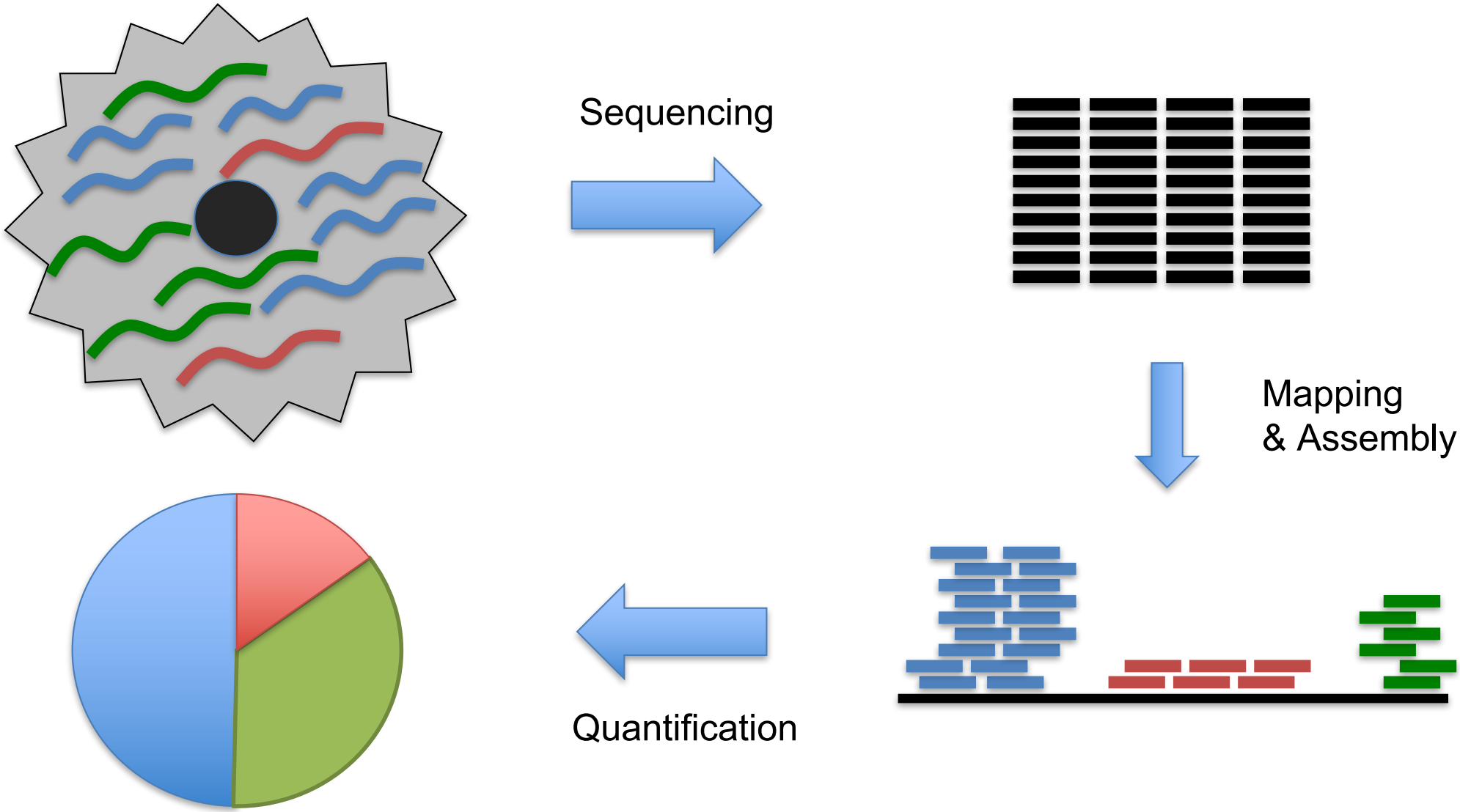


RNA-seq

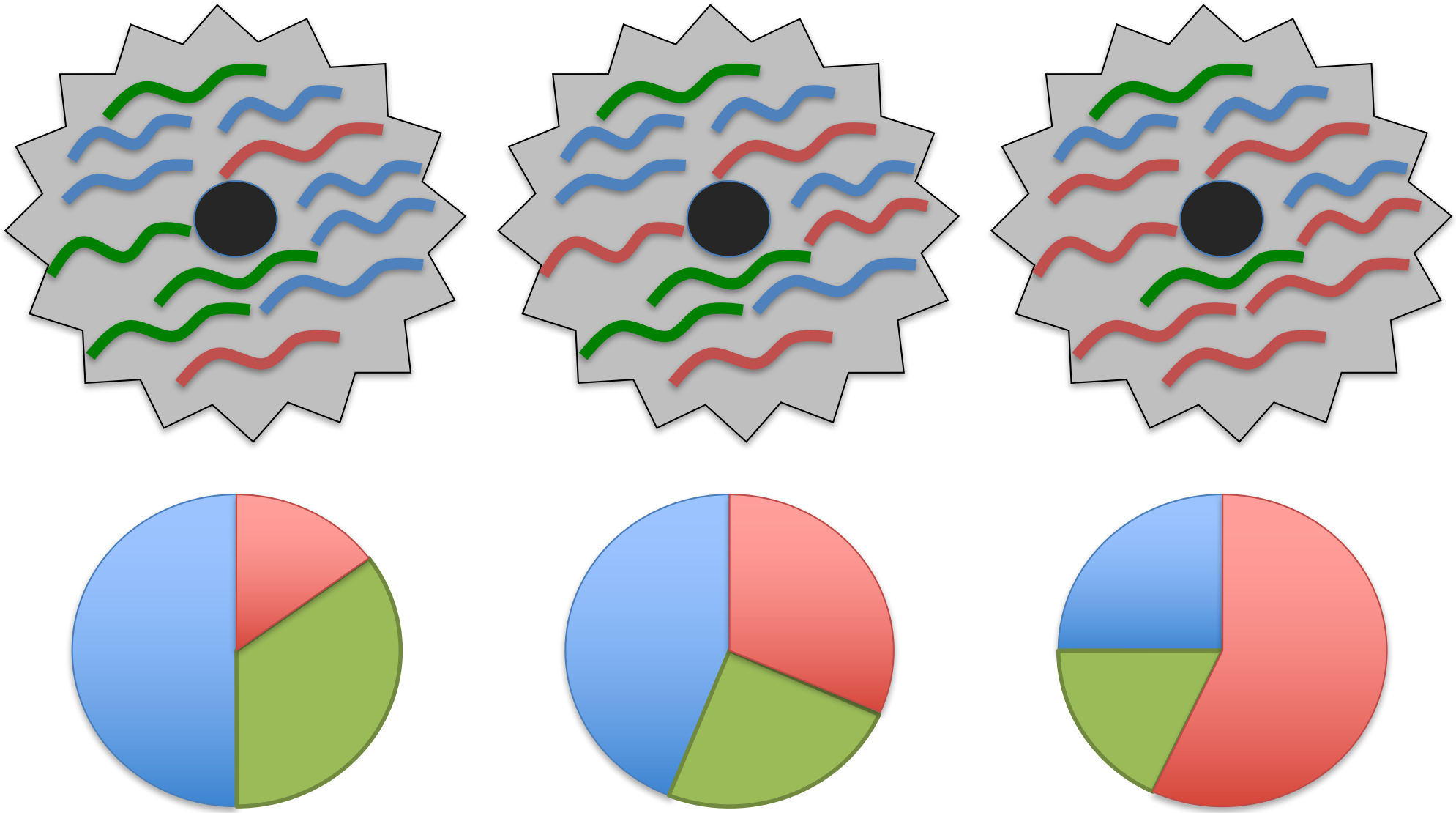


Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.
Sørlie et al (2001) *PNAS*. 98(19):10869-74.

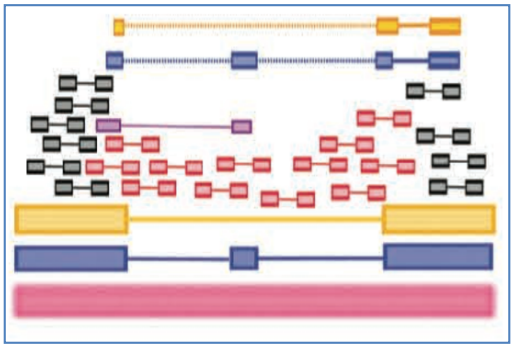
RNA-seq Overview



RNA-seq Overview



RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

RNA-Seq Approaches

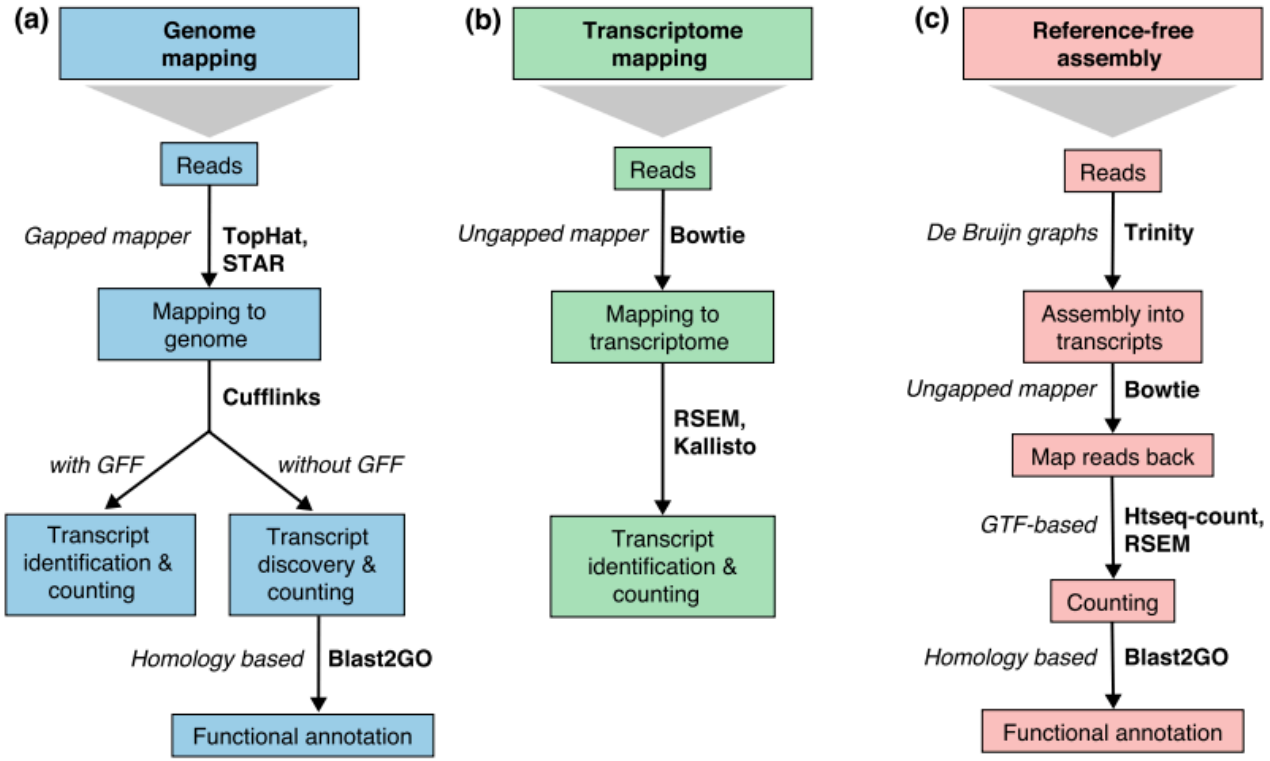


Fig. 2 Read mapping and transcript identification strategies. Three basic strategies for regular RNA-seq analysis. **a** An annotated genome is available and reads are mapped to the genome with a gapped mapper. Next (novel) transcript discovery and quantification can proceed with or without an annotation 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

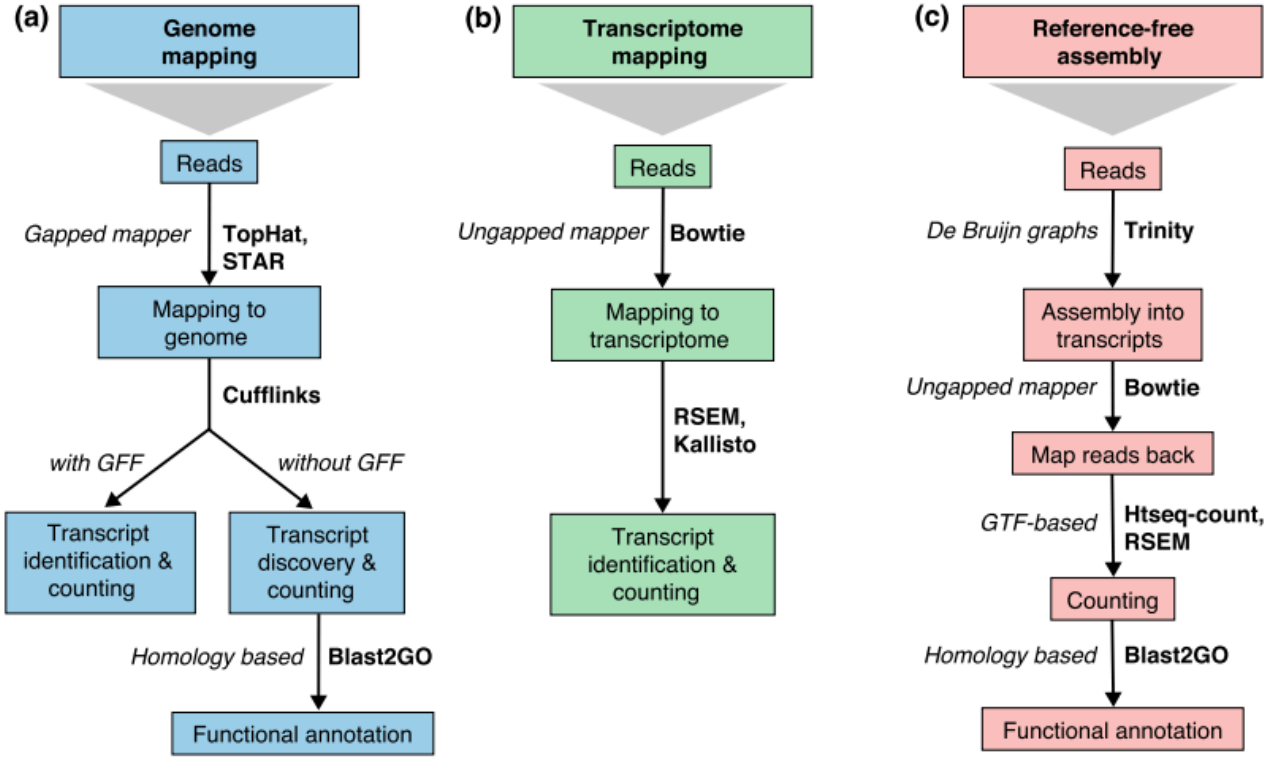


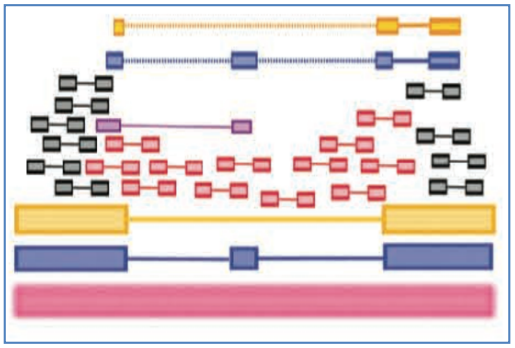
Fig. 2 Read mapping and transcript identification strategies. Three basic strategies for regular RNA-seq analysis. **a** An annotated genome is available and re…ovel) transcript discovery and quantification can proceed with or without an ann… no novel transcript discovery is needed, reads can be mapped to the reference transcriptome using an ungapped aligner. Transcript identification and quantification can occur simultaneously. **c** When no genome is available, reads n…o contigs or transcripts. For quantification, reads are mapped back to the novel reference transcriptome and further ar…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

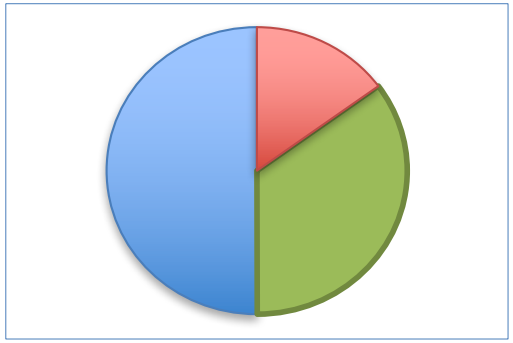
RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

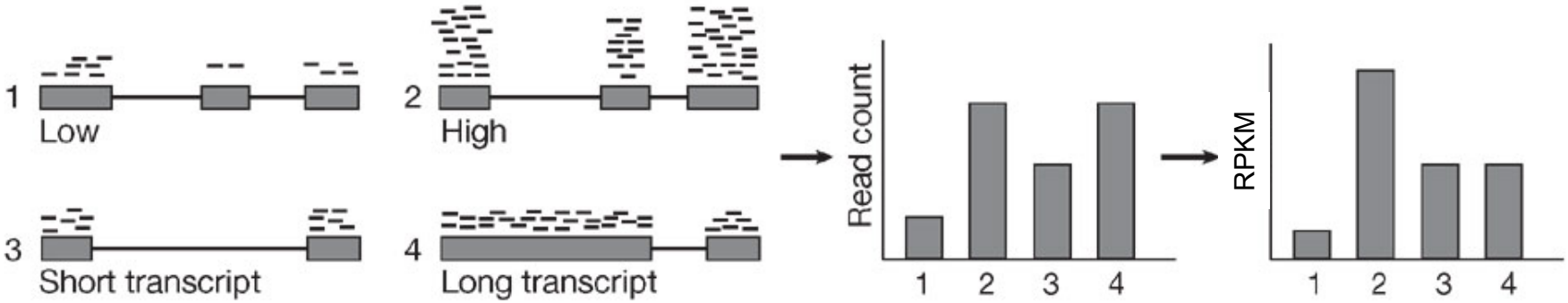
Solution: Use a spliced aligner, and assemble isoforms

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



Challenge 2: Read Count != Transcript abundance

RPLM, FPKM, TPM



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

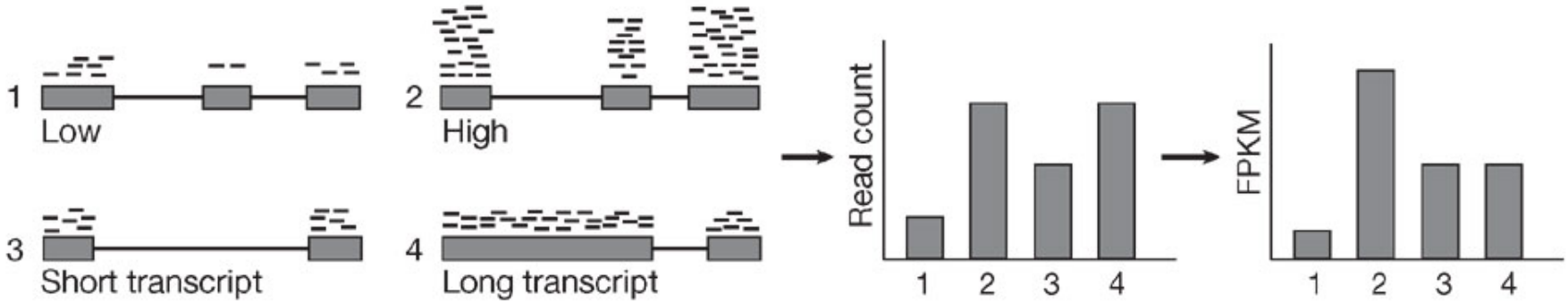
- Overall Coverage: 1M reads in experiment 1 vs 10M reads in experiment 2
- Gene Length: gene 1 is 10kbp, gene 2 is 100kbp

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

(Count reads aligned to gene) / (length of gene in kilobases) / (# millions of read mapped)

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

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 1 is 10kbp, gene 2 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!

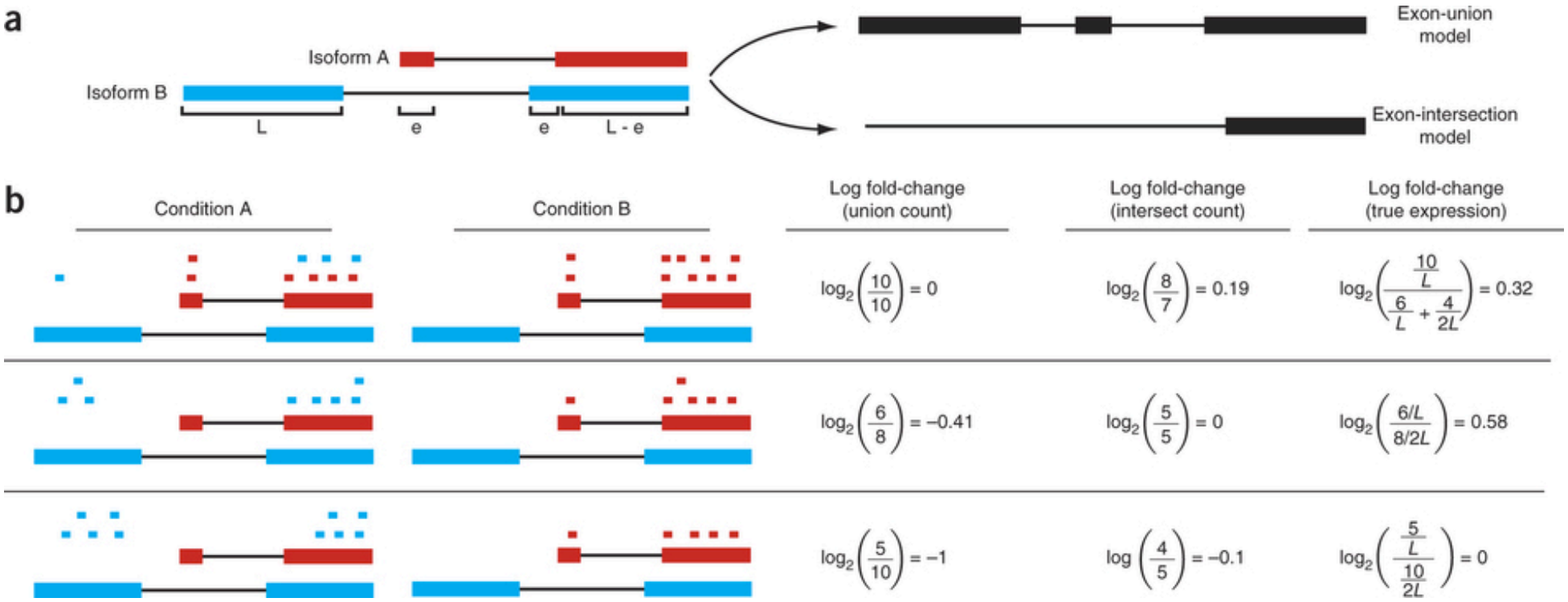
3. TPM: Transcripts Per Million (Li et al, 2011)

$\Rightarrow$ 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

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

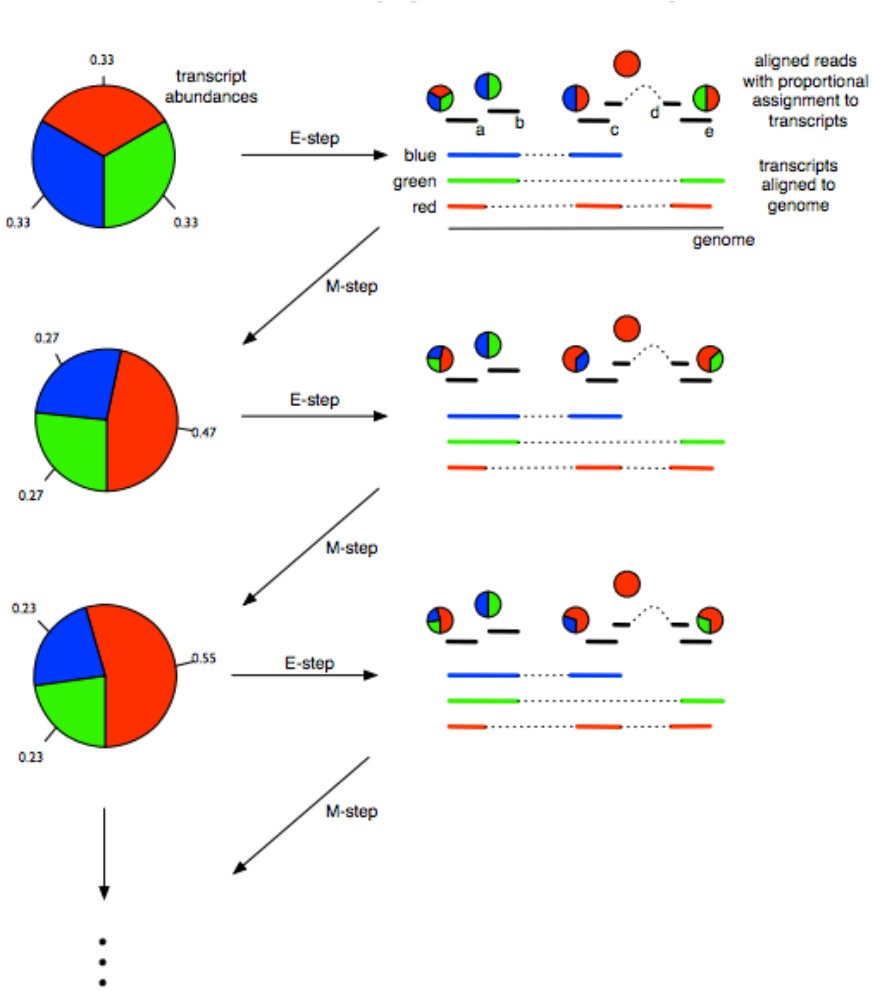
Gene or Isoform Quantification?



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. 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.

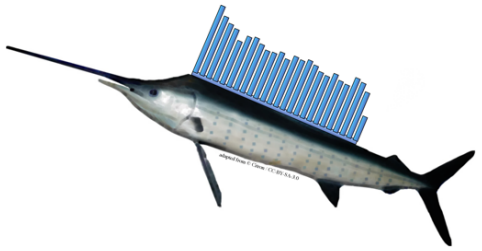
During the expectation (E) step reads are proportionately assigned to transcripts according to the (current) isoform abundances (RGB): a=(.33,.33,.33), b=(0,.5,.5), c=(.5,.5), d=(1,0,0), e=(.5,.5,0)

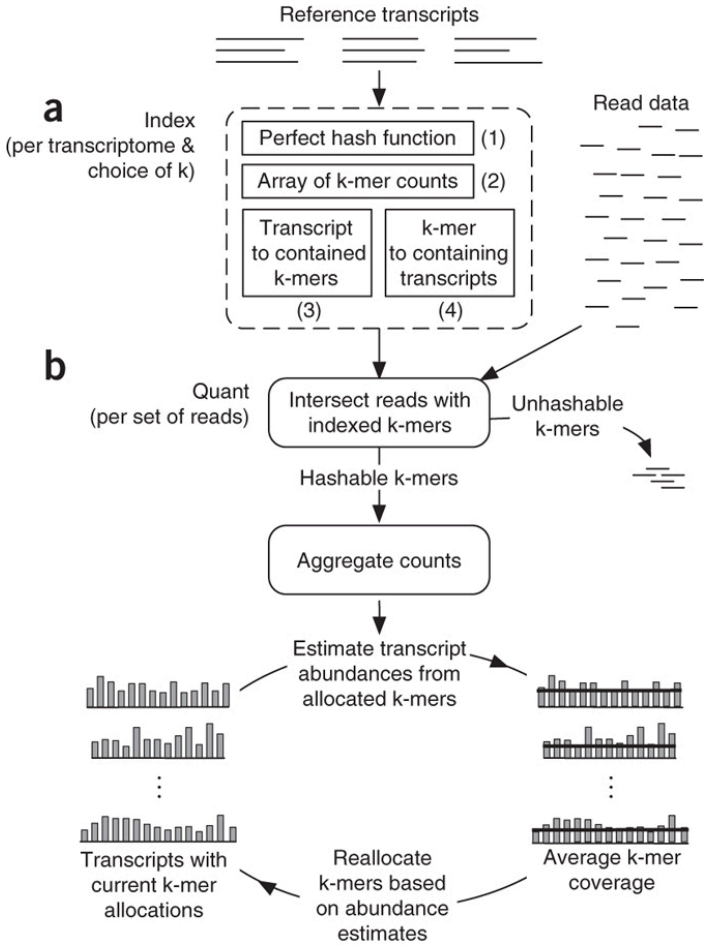
Next, during the maximization (M) step isoform abundances are recalculated from the proportionately assigned read counts:
red: 0.47 = (0.33 + 0.5 + 1 + 0.5)/(2.33 + 1.33 + 1.33)
blue: 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)
green: 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)

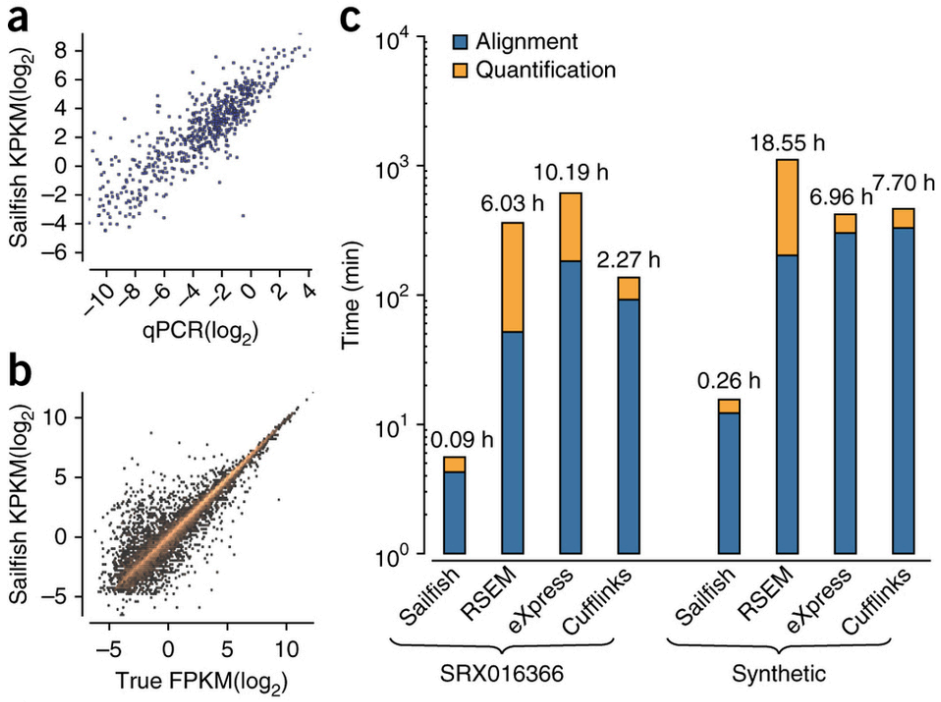
Repeat until convergence!

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

Sailfish: Fast & Accurate RNA-seq Quantification





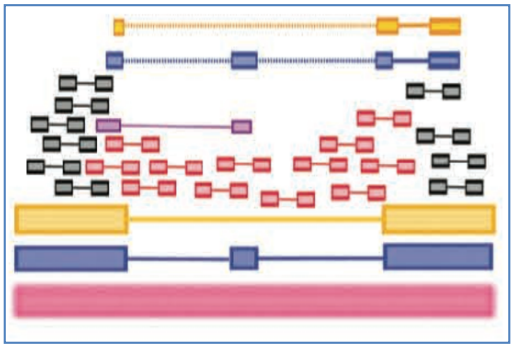


d

	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

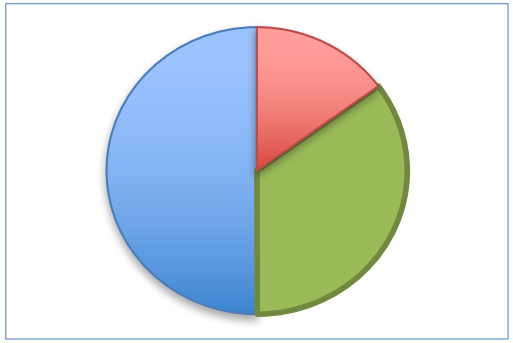
RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

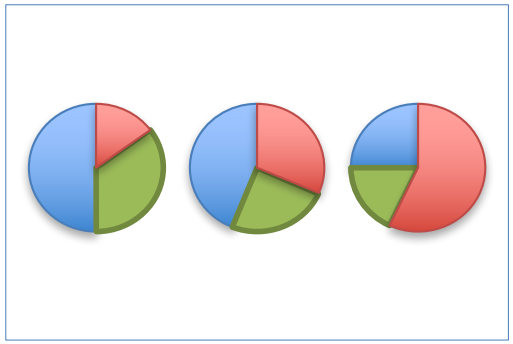
TopHat: discovering spliced junctions with RNA-Seq.
Trapnell et al (2009) *Bioinformatics*. 25:0 1105-1111



Challenge 2: Read Count != Transcript abundance

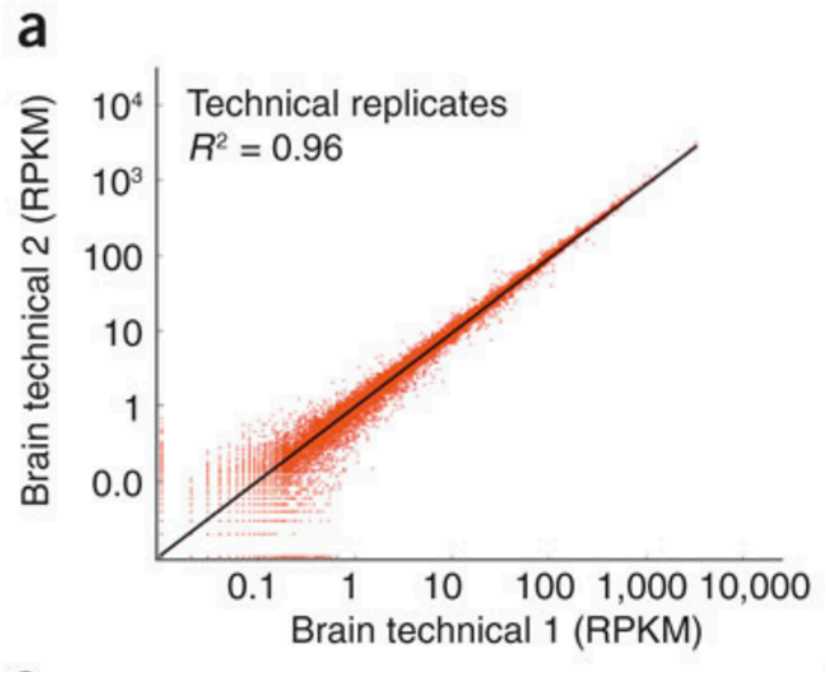
Solution: Infer underlying abundances (e.g. TPM)

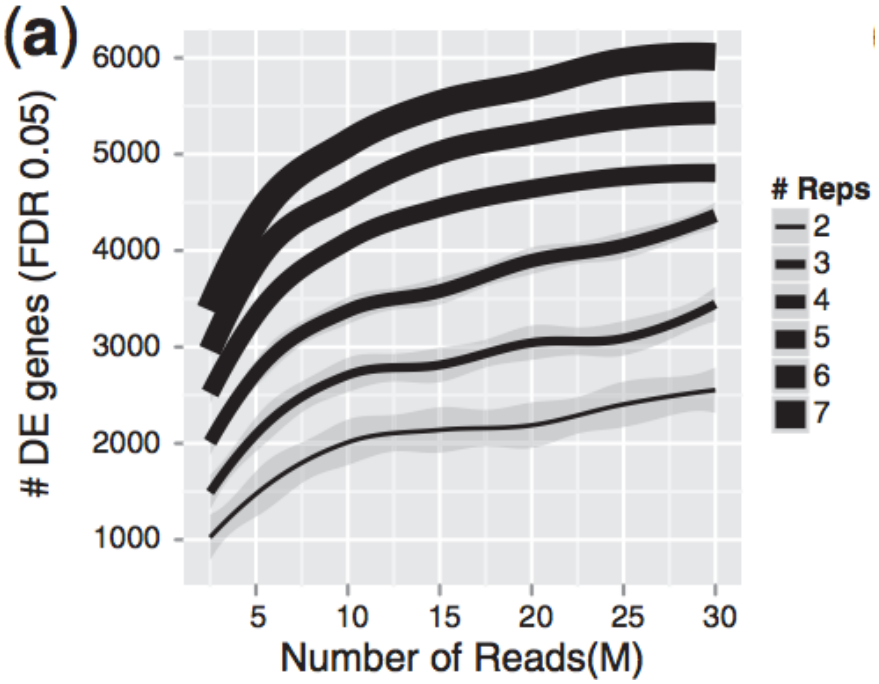
Transcript assembly and quantification by RNA-seq
Trapnell et al (2010) *Nat. Biotech.* 25(5): 511-515



Challenge 3: Transcript abundances are stochastic

How Many Replicates?



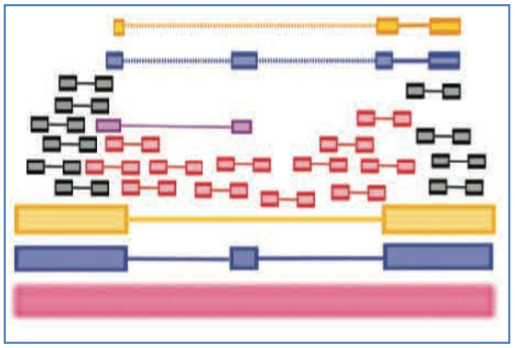


Why don't we have perfect replicates?

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

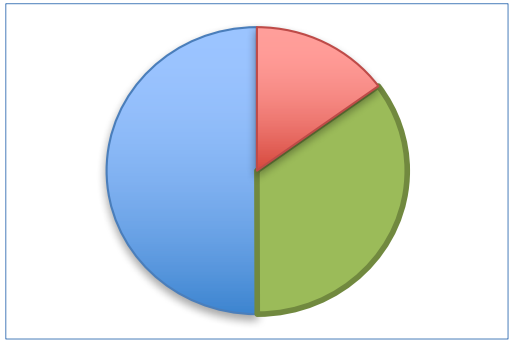
RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

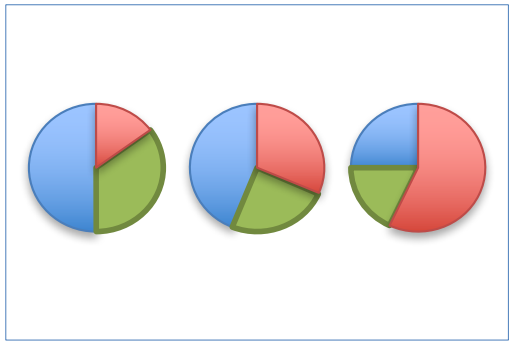
TopHat: discovering spliced junctions with RNA-Seq.
Trapnell et al (2009) *Bioinformatics*. 25:0 1105-1111



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

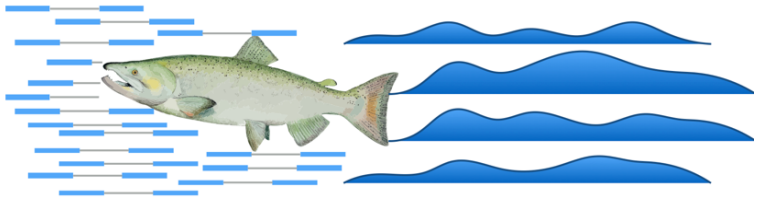


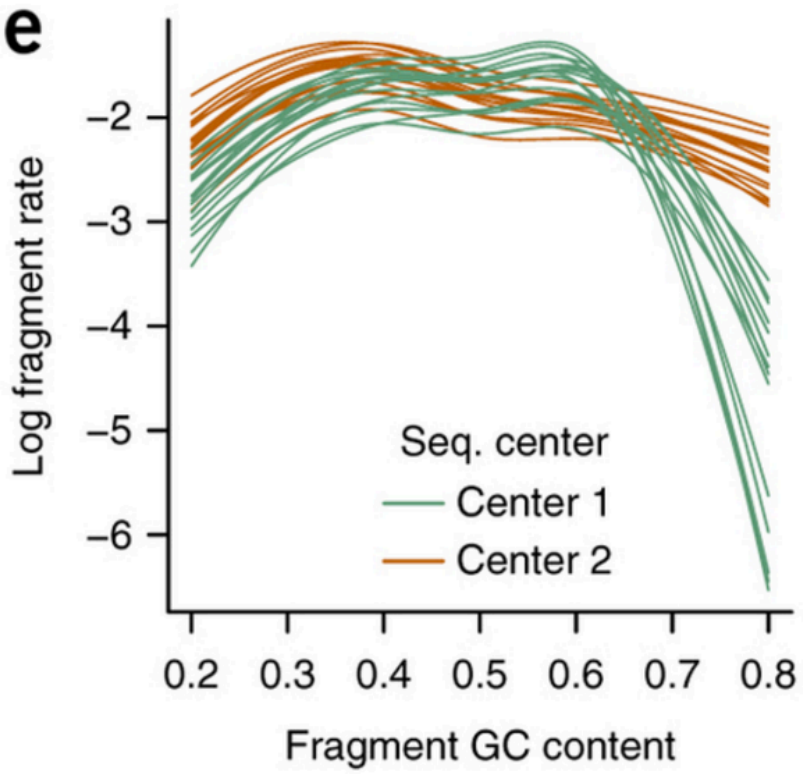
Challenge 3: Transcript abundances are stochastic

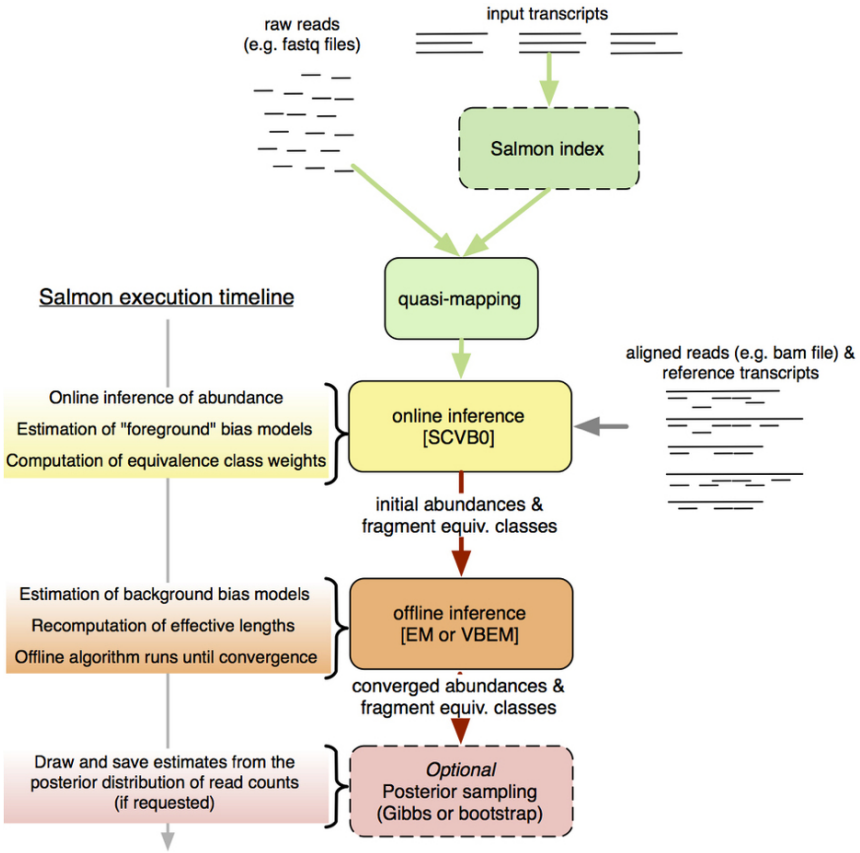
Solution: Replicates, replicates, and more replicates

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

Salmon: The ultimate RNA-seq Pipeline?



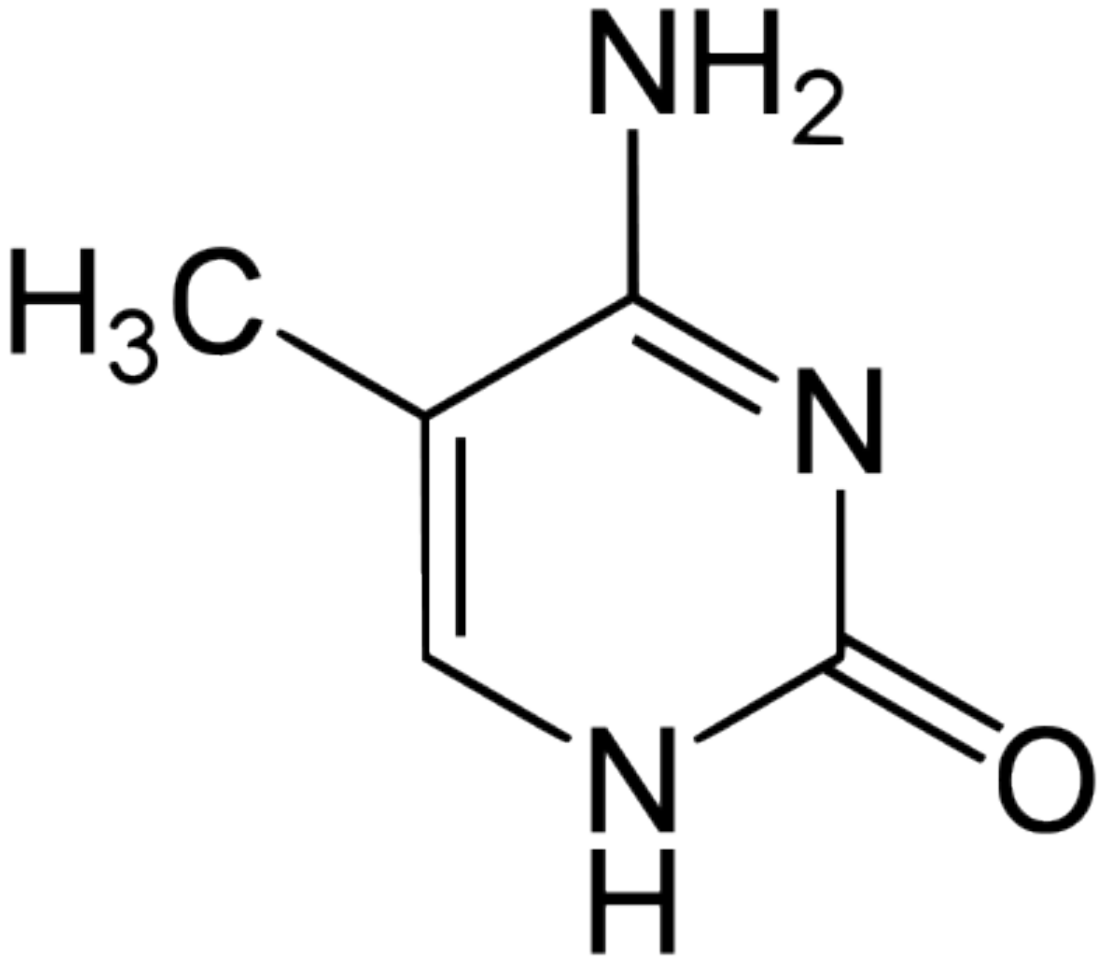




Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation
Love et al (2016) Nature Biotechnology 34, 1287–1291 (2016) doi:10.1038/nbt.3682

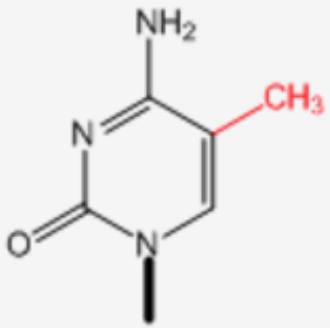
Salmon provides fast and bias-aware quantification of transcript expression
Patro et al (2017) Nature Methods (2017) doi:10.1038/nmeth.4197

Methyl-seq

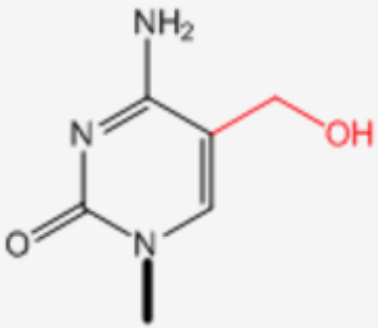


Finding the fifth base: Genome-wide sequencing of cytosine methylation
Lister and Ecker (2009) *Genome Research.* 19: 959-966

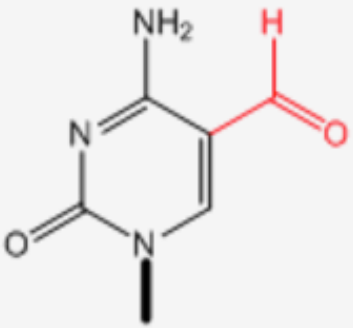
Epigenetic Modifications to DNA



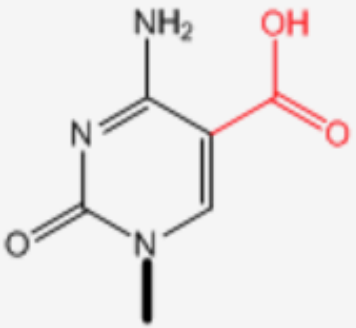
5-mC



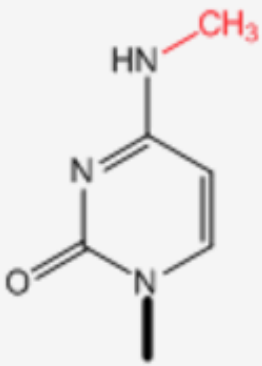
5-hmC



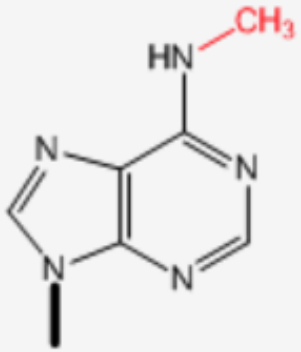
5-fC



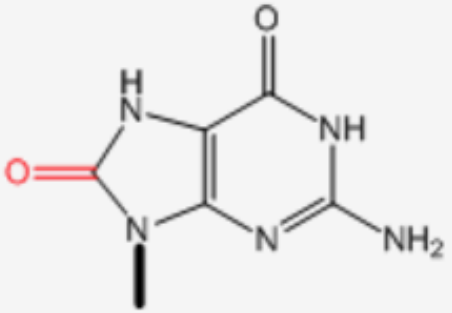
5-caC



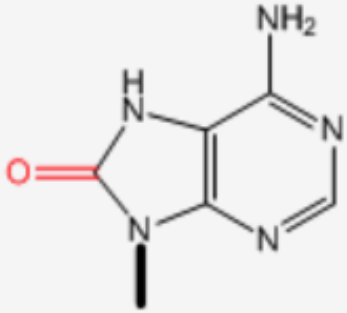
4-mC



6-mA

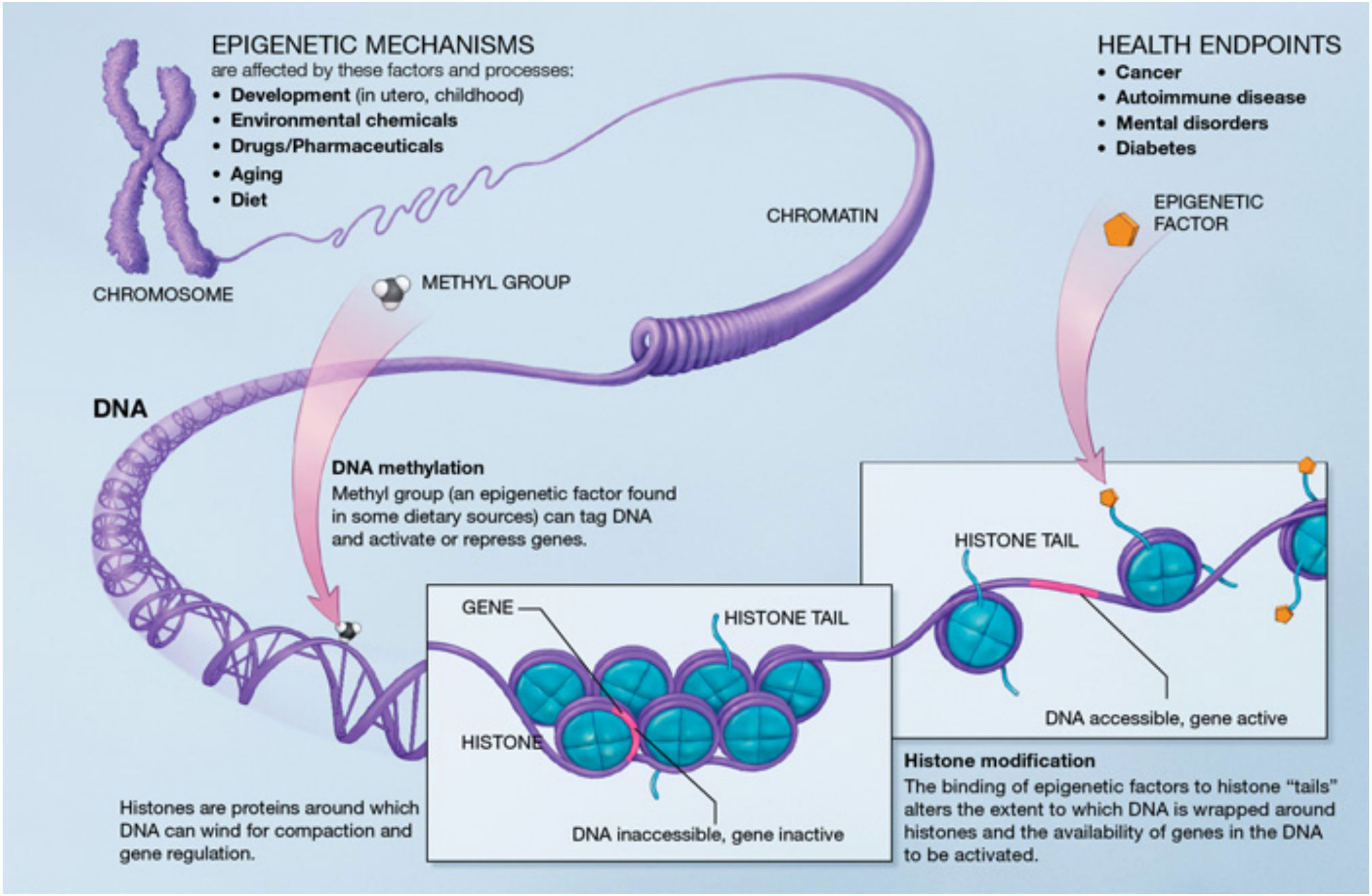


8-oxoG



8-oxoA

Methylation & Epigenetics



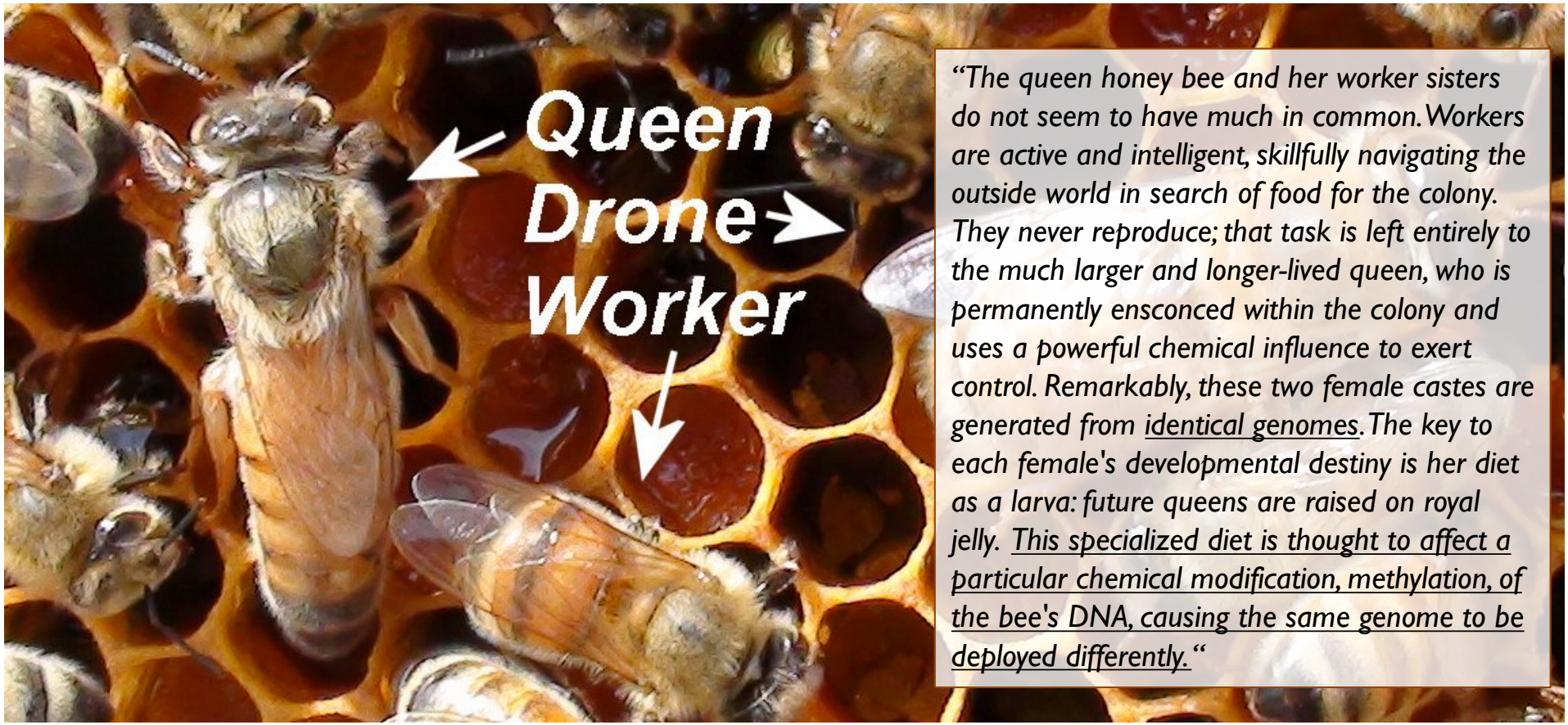
OPEN ACCESS Freely available online

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

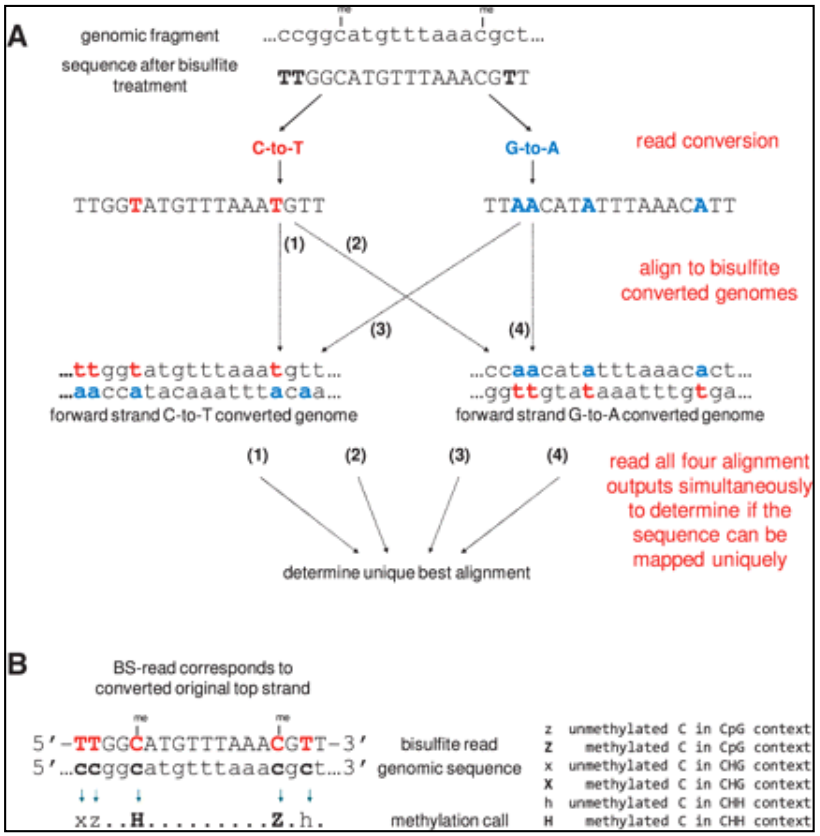


"The queen honey bee and her worker sisters do not seem to have much in common. Workers are active and intelligent, skillfully navigating the outside world in search of food for the colony. They never reproduce; that task is left entirely to the much larger and longer-lived queen, who is permanently 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."

Bisulfite Conversion

Treating DNA with sodium bisulfite will convert unmethylated 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.

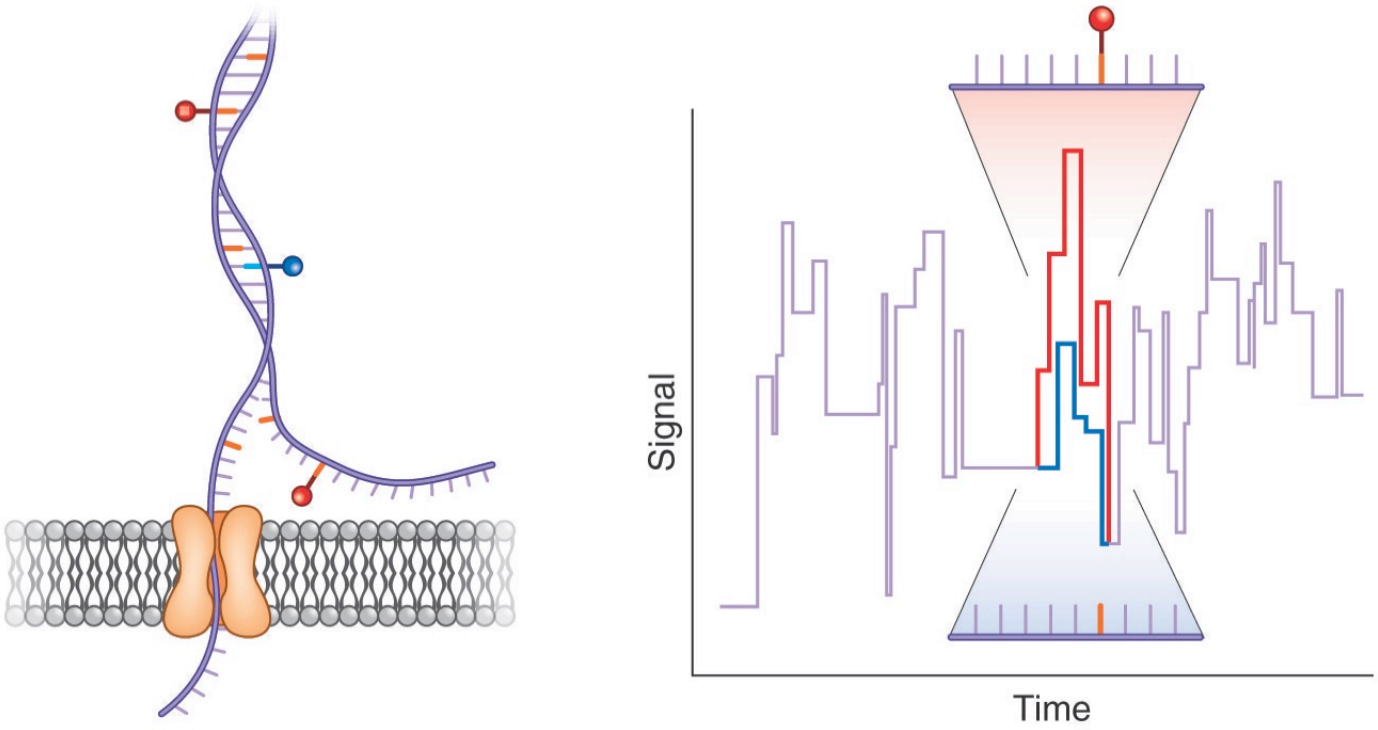
Bisulfite Conversion

Ti
w

-
-
-

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

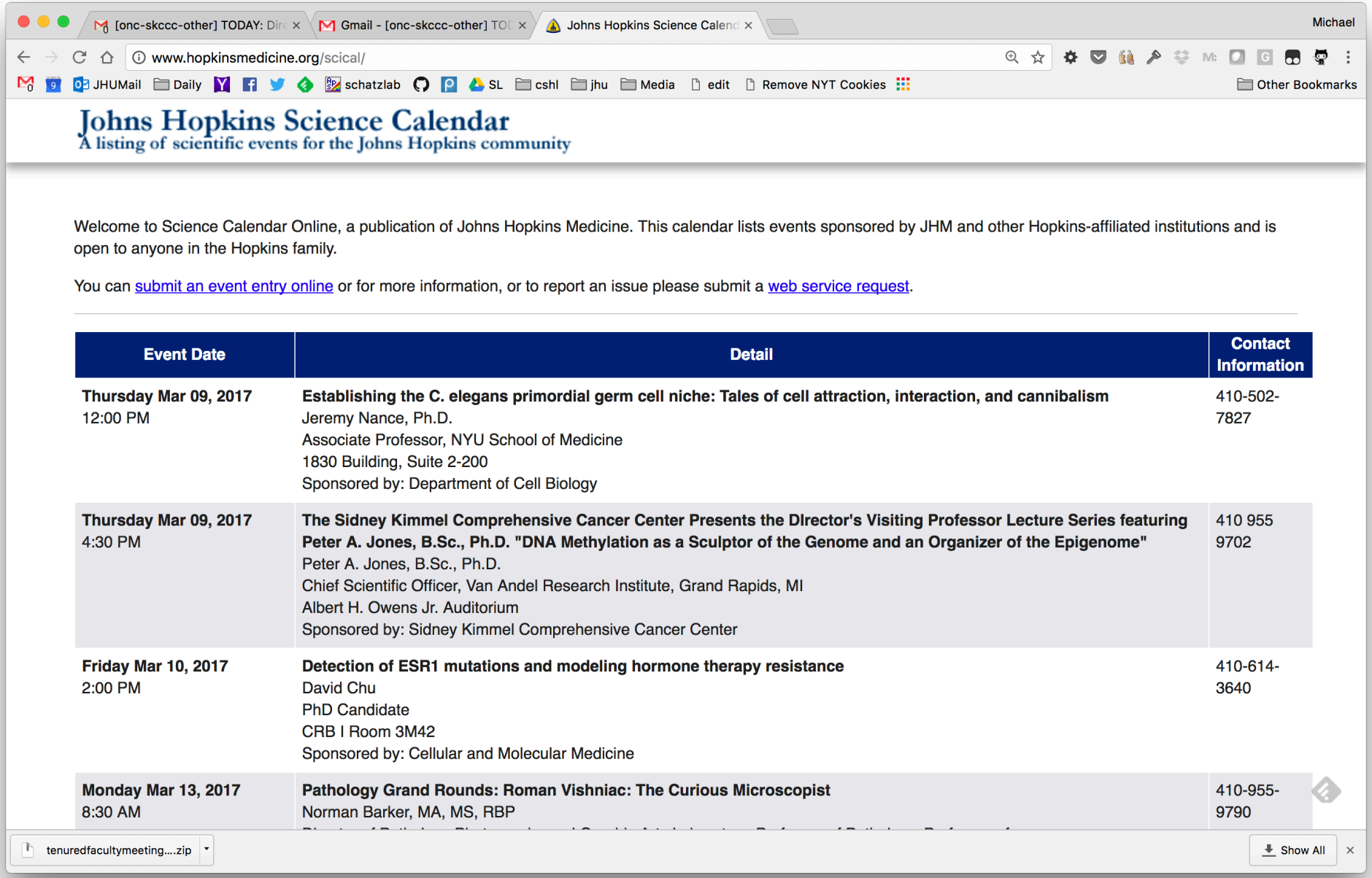
Methylation Detection using Oxford Nanopore Sequencing



Detecting DNA cytosine methylation using nanopore sequencing
Simpson et al (2017) Nature Methods. Doi:10.1038/nmeth.4184

Mapping DNA methylation with high-through nanopore sequencing
Rand et al (2017) Nature Methods. doi:10.1038/nmeth.4189

Today at 4:30!



Johns Hopkins Science Calendar

A listing of scientific events for the Johns Hopkins community

Welcome to Science Calendar Online, a publication of Johns Hopkins Medicine. This calendar lists events sponsored by JHM and other Hopkins-affiliated institutions and is open to anyone in the Hopkins family.

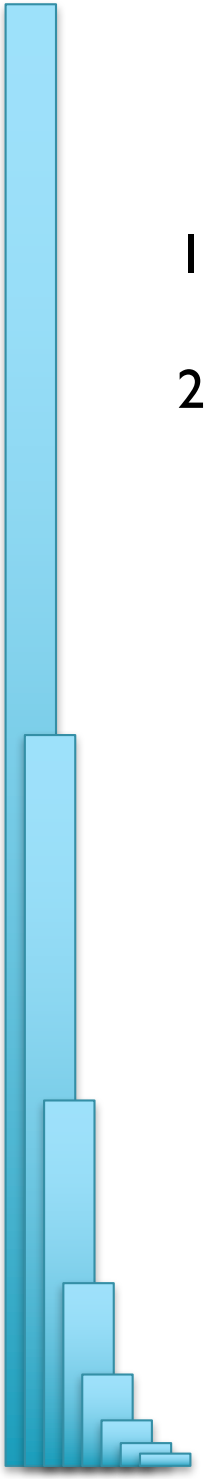
You can submit an event entry online or for more information, or to report an issue please submit a web service request.

Event Date	Detail	Contact Information
Thursday Mar 09, 2017 12:00 PM	**Establishing the C. elegans primordial germ cell niche: Tales of cell attraction, interaction, and cannibalism** Jeremy Nance, Ph.D. Associate Professor, NYU School of Medicine 1830 Building, Suite 2-200 Sponsored by: Department of Cell Biology	410-502-7827
Thursday Mar 09, 2017 4:30 PM	**The Sidney Kimmel Comprehensive Cancer Center Presents the Director's Visiting Professor Lecture Series featuring Peter A. Jones, B.Sc., Ph.D. "DNA Methylation as a Sculptor of the Genome and an Organizer of the Epigenome"** Peter A. Jones, B.Sc., Ph.D. Chief Scientific Officer, Van Andel Research Institute, Grand Rapids, MI Albert H. Owens Jr. Auditorium Sponsored by: Sidney Kimmel Comprehensive Cancer Center	410 955 9702
Friday Mar 10, 2017 2:00 PM	**Detection of ESR1 mutations and modeling hormone therapy resistance** David Chu PhD Candidate CRB I Room 3M42 Sponsored by: Cellular and Molecular Medicine	410-614-3640
Monday Mar 13, 2017 8:30 AM	**Pathology Grand Rounds: Roman Vishniac: The Curious Microscopist** Norman Barker, MA, MS, RBP	410-955-9790

Next Steps

1. Questions on assignment 2?

2. Check out the course webpage




Welcome to Applied Comparative Genomics
https://github.com/schatzlab/appliedgenomics

Questions?