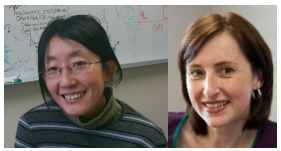
Genome Annotation

Paul Gardner

November 10, 2014

Medical genomics



- Una Ren at ESR & Collette Bromhead at Aotea Pathology are sequencing *Neisseria* genomes. In order to identify signatures associated with meningitis and gonorrhea.

- Vicky Cameron & Anna Pilbrow at Otago are identifying genetic variation and genes associated with an increased risk of heart disease.

- Marcel Dinger at the Garvan Insitute and Mike Stratton at the Sanger Institute are developing methods for identifying genes associated with an increased risk of cancer.

- Rob Knight at UC Boulder is sequencing the microbes that live on us. Finding associations between our health and microbial communities.

Agricultural genomics

- Graeme Attwood at AgResearch is trying to stop cows & sheep from emitting greenhouse gases by studying their gut microbes. He has two methanogenic Archaeal genomes of *Methanobrevibacter* sp.

- Ashley Lu & Andy Pitman at PFR are trying to determine how Pseudomonas syringae pv. actinidiae (PSA) is killing kiwifruit.

- Rebecca Ganley at SCION is investigating how Phytophthora Taxon Agathis (PTA) is causing kauri die-back disease and killing kauri trees.

Academic interest genomics

- Tom Gilbert at the University of Copenhagen is sequencing bird and giant squid genomes.

- Elizabeth Murchison is sequencing tasmanian devils (and their transmissible cancers).

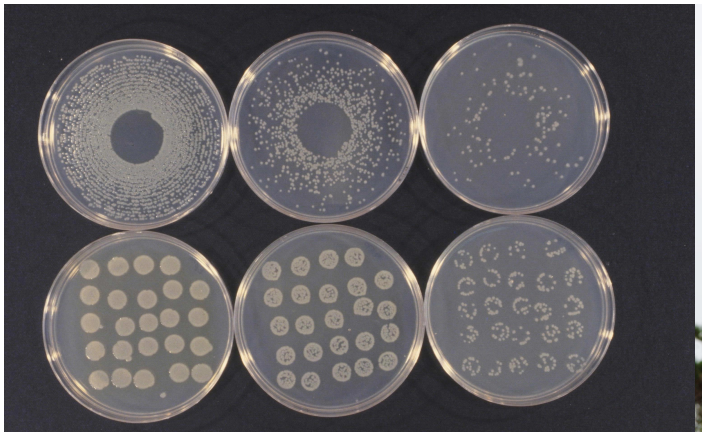
- Neil Gemmel at Otago University is sequencing the tuatara genome.



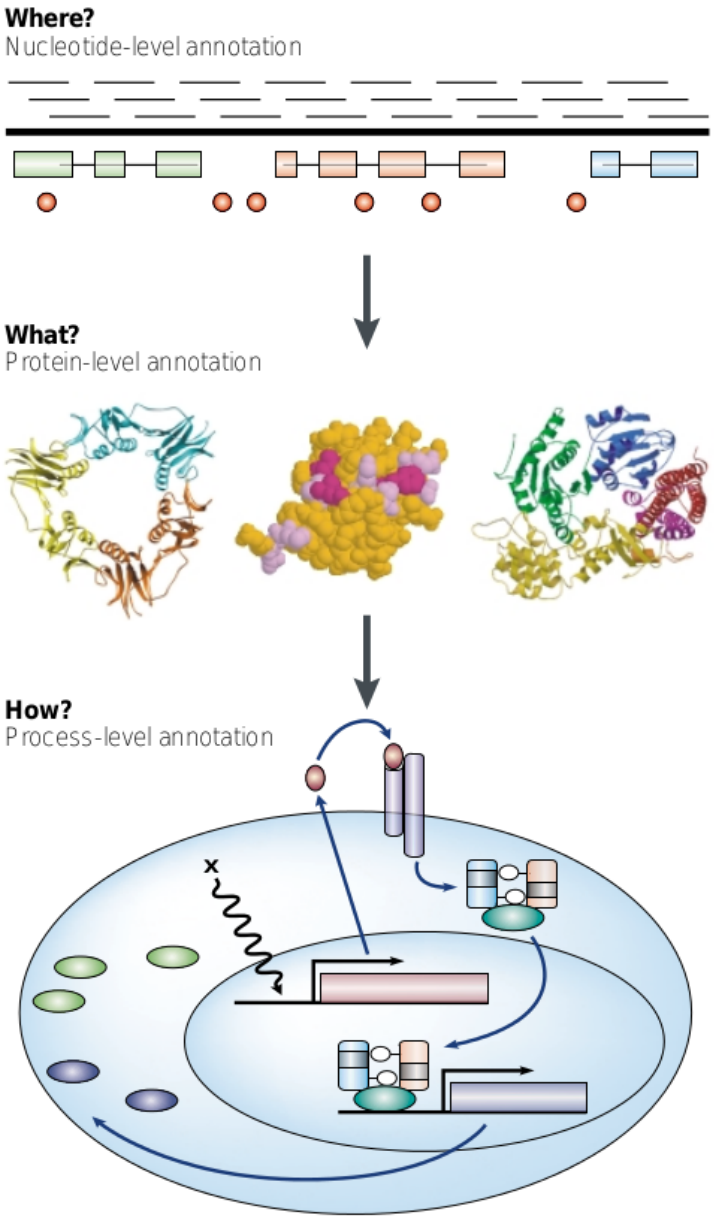
Discussion

- How should these researchers annotate their genomes (after they have sequenced and assembled them)?
  - What are the fast and cheap methods?
  - What are the most accurate methods?





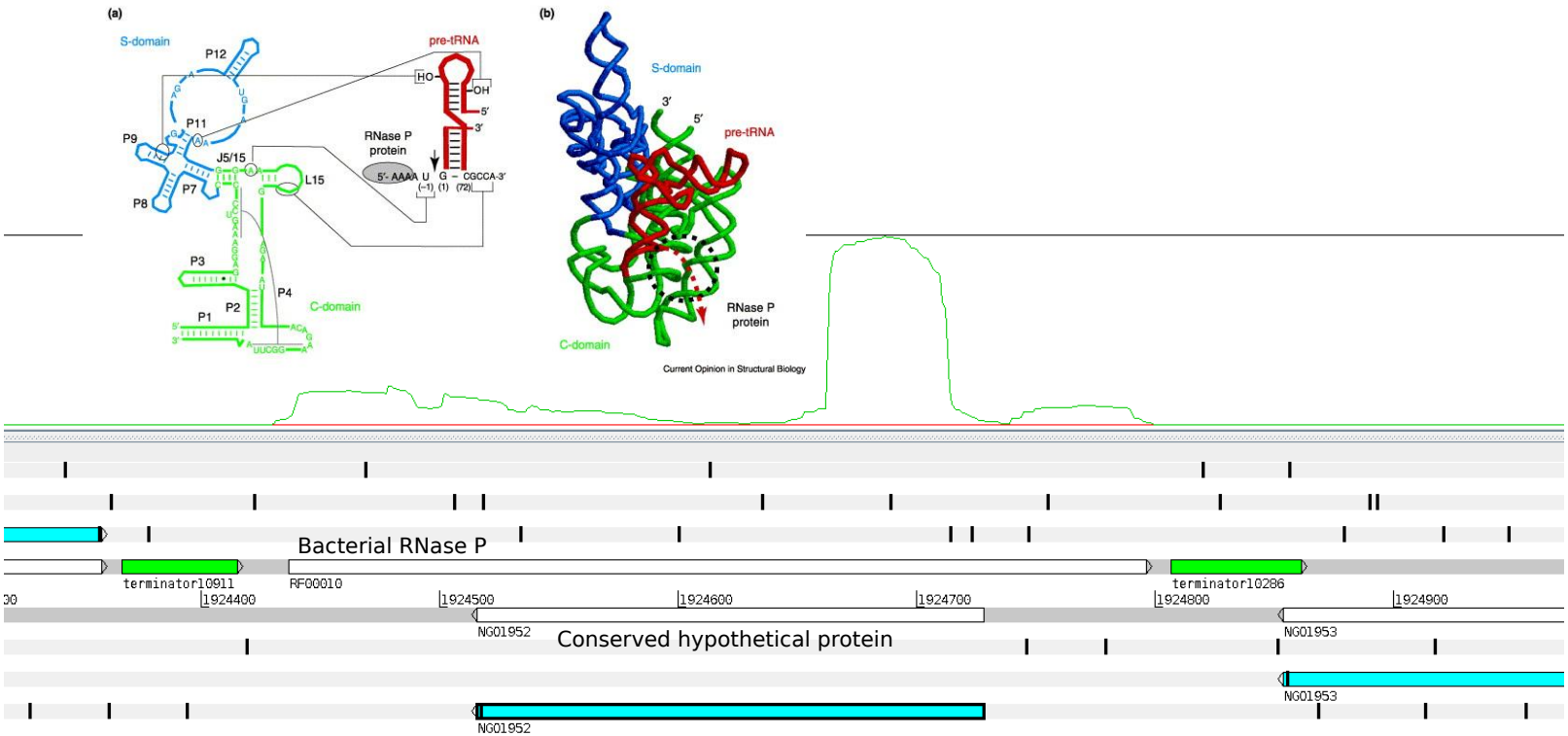

Nucleotide → Protein/RNA → Process



Stein (2001) Genome annotation: from sequence to biology. *Nature Reviews Genetics.*

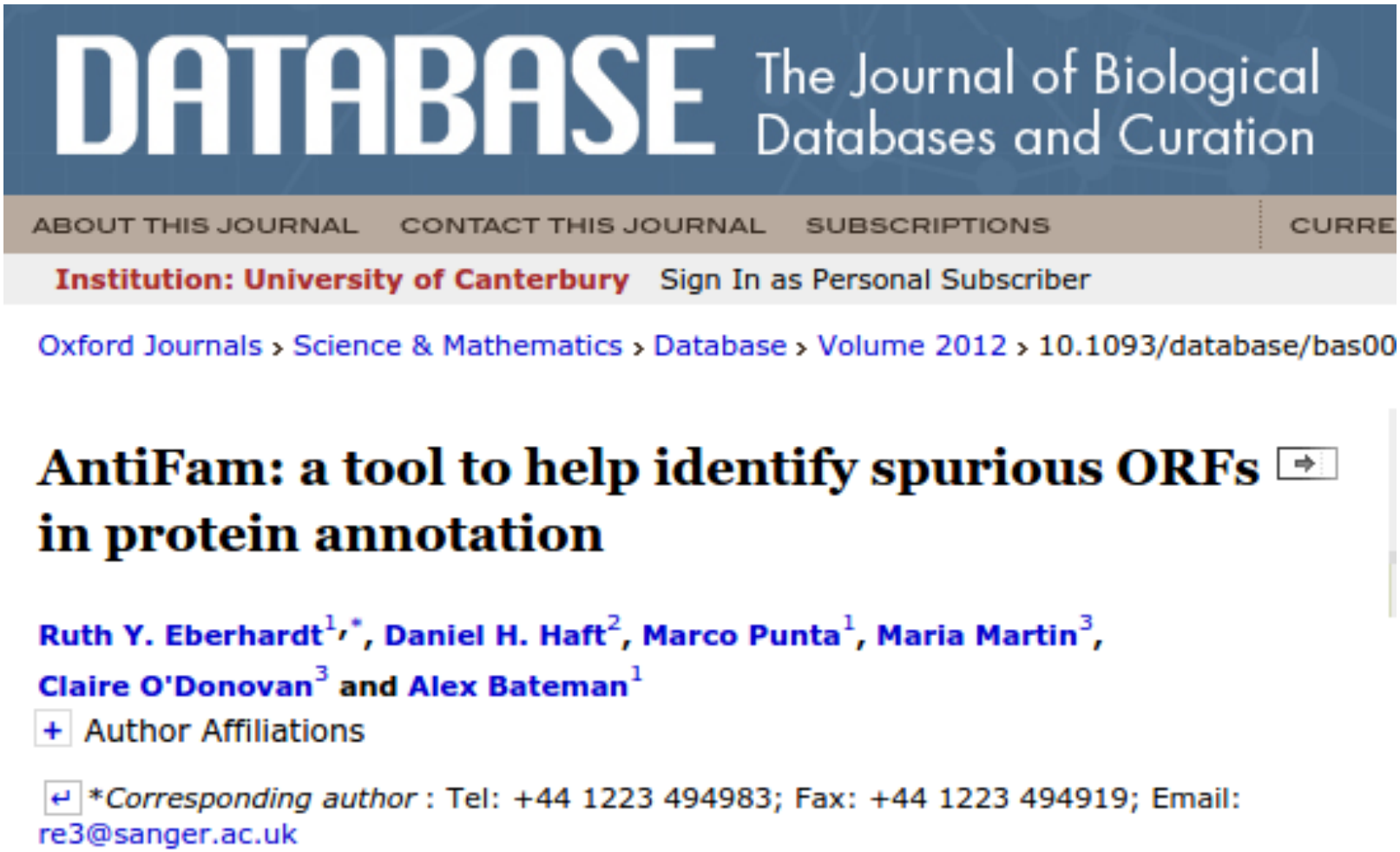
Embarrassing genome annotation screwups

- *Neisseria gonorrhoeae*



Embarrassing genome annotation screwups

- Propagation errors
- Data reuse



DATABASE The Journal of Biological Databases and Curation

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS CURRE

Institution: University of Canterbury Sign In as Personal Subscriber

Oxford Journals › Science & Mathematics › Database › Volume 2012 › 10.1093/database/bas00

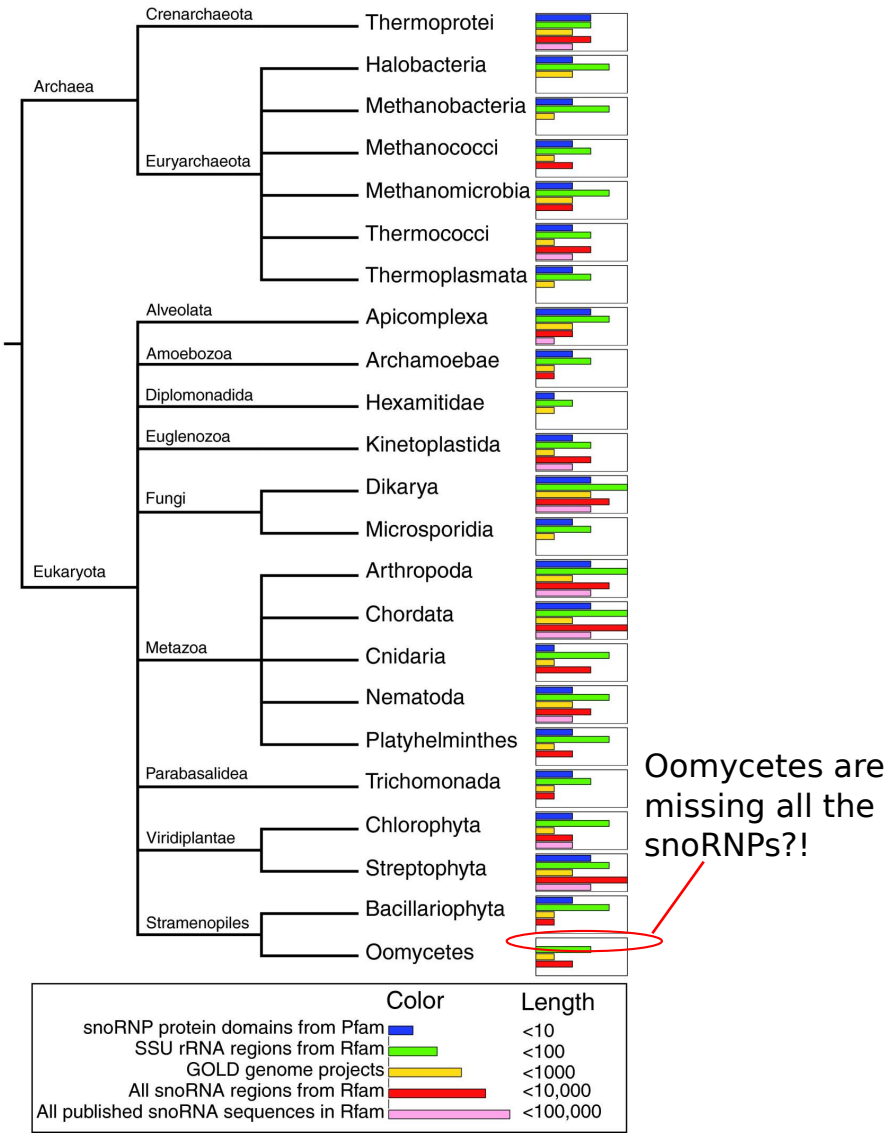
AntiFam: a tool to help identify spurious ORFs in protein annotation

Ruth Y. Eberhardt[1,*], Daniel H. Haft[2], Marco Punta[1], Maria Martin[3], Claire O'Donovan[3] and Alex Bateman[1]

+ Author Affiliations

*Corresponding author : Tel: +44 1223 494983; Fax: +44 1223 494919; Email: re3@sanger.ac.uk

Embarrassing genome annotation screwups

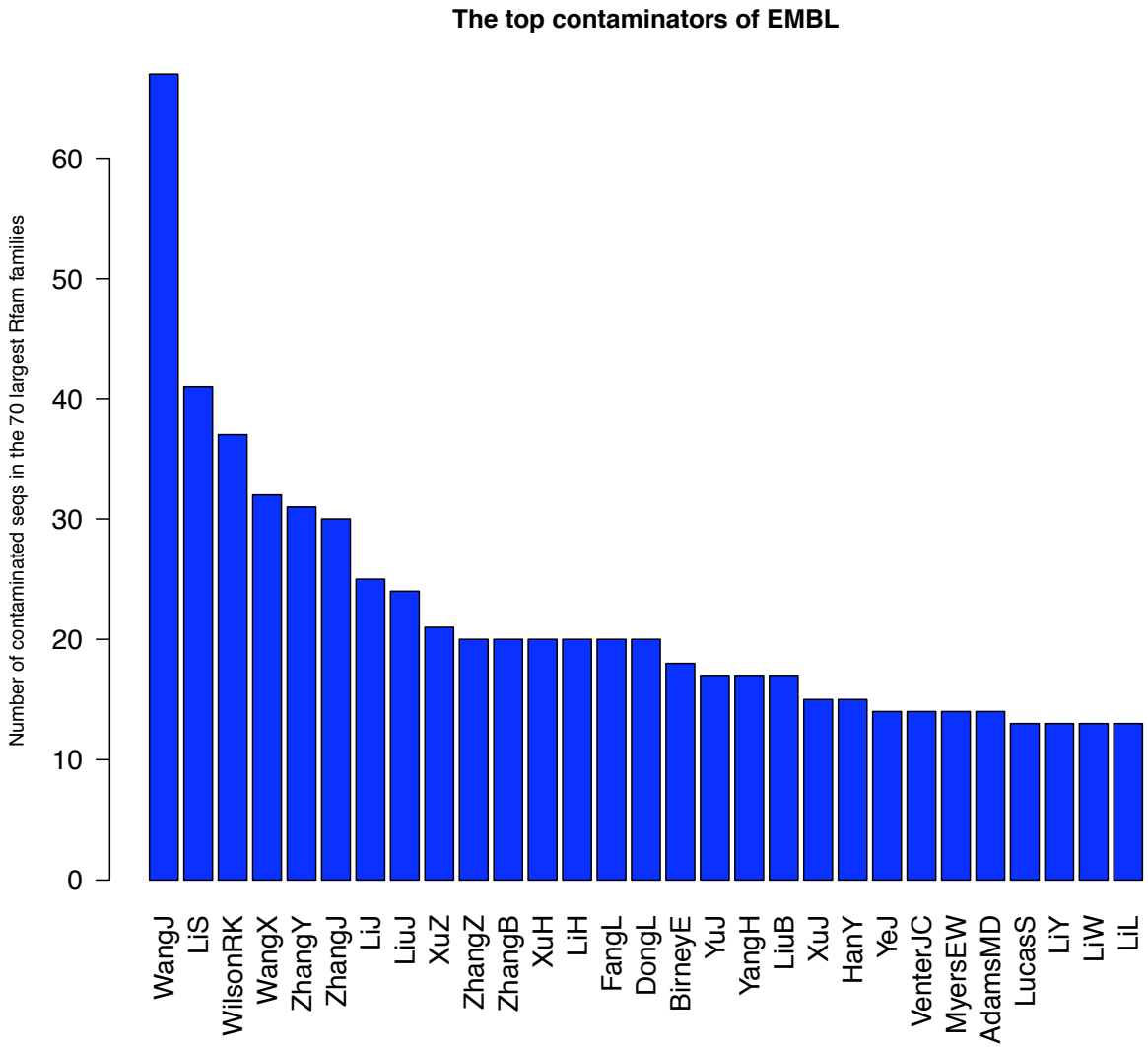


Oomycetes are missing all the snoRNPs?!

Gardner, Bateman & Poole (2010) SnoPatrol: how many snoRNA genes are there? *Journal of Biology.*

Embarrassing genome annotation screwups

- GenBank/ENA/DDBJ is a festering pit of misannotated sequence



Ab initio/*De novo* genome annotation

- Genes leave a statistical signal in the genome...
- Example: identify promotors, ribosome binding sites, open-reading frames (ORFs), terminators
  - In eukaryotes CpG islands, splicing signals and poly-A tails may be incorporated
  - How reliable are these approaches? What are the main weaknesses & strengths?
- Terrible for ncRNAs! (Rivas & Eddy 2000)

**except in A+T-rich hypertherm. genomes

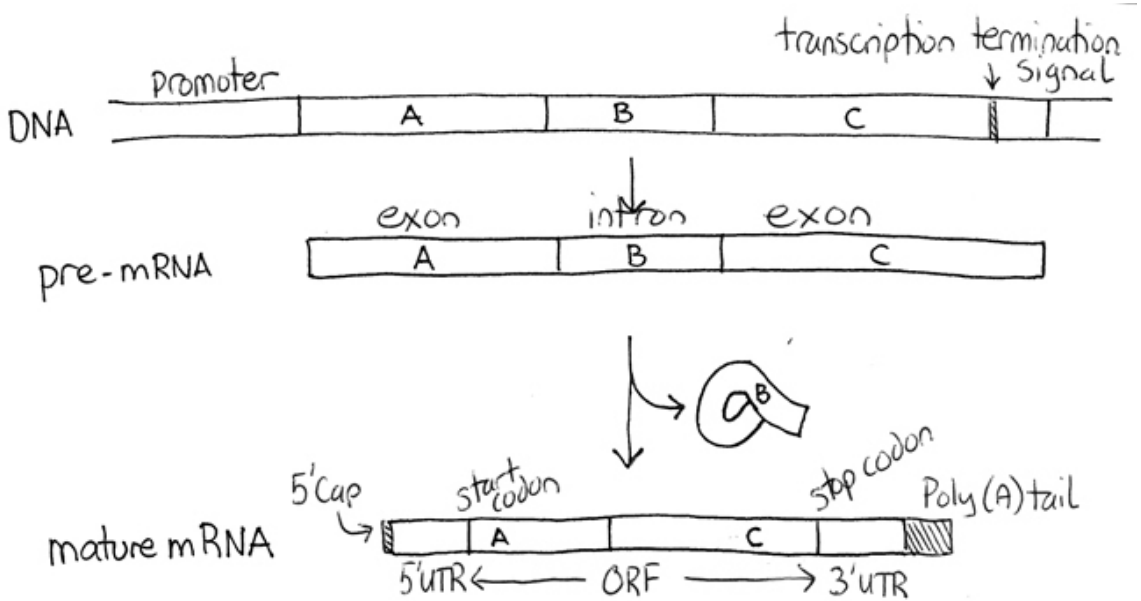


Figure from: http://zerocool.is-a-geek.net/?p=630

Sequence analysis: strengths and weaknesses

- ORF prediction: GLIMMER/PRODIGAL
  - Strengths:
    - very fast
    - cheap
  - Weaknesses:
    - false positives (see AntiFam)
    - misses short peptides (e.g. toxins-antitoxin systems)
    - No ncRNAs

Prodigal pseudocode

```
1.  Read in the sequence
2.  Locate all starts and stops in the genome
3.  Scan all open reading frames and record numbers of G's and C's in each codon position
4.  Build a frame bias model based on ORF length and G/C codon position within each ORF
5.  Record the highest scoring start nodes in each frame that overlap a stop codon by <= 60 bp
6.  Do the first pass dynamic programming, connecting nodes based on frame bias scores
7.  Create a hexamer background of all 6-mers in the entire sequence
8.  FOR each gene model in the dynamic programming output:
      1.  Gather all hexamer statistics
9.  Create log table of hexamer coding scores
10. FOR each gene model in the dynamic programming output:
      1.  Calculate a coding score based on hexamer statistics
      2.  Penalize the score if there is a higher scoring start upstream in the same ORF
      3.  IF the gene is very long but has a negative score, THEN give it a barely positive
          score
11. FOR 10 iterations
      1.  Build a ribosomal binding site and ATG/GTG/TTG background for all nodes
      2.  FOR each gene with a score of > 35.0:
            1.  Gather its Shine-Dalgarno RBS motif data and ATG/GTG/TTG data
      3.  Modify RBS and ATG/GTG/TTG weights by the observations
12. IF organism is not determined to use Shine-Dalgarno THEN run the non-SD finder
13. FOR each gene model:
      1.  Assign a final score of start score + coding score
      2.  Penalize the final score of genes < 250bp
14. Do the second pass dynamic programming, connecting nodes based on hexamer coding
15. FOR each gene model in the final dynamic programming:
      1.  Eliminate negative scoring models
      2.  Resolve very close start pairs (<= 15 bp from each other)
16. Print final output
```

Figure 1 Pseudocode description of the Prodigal algorithm.

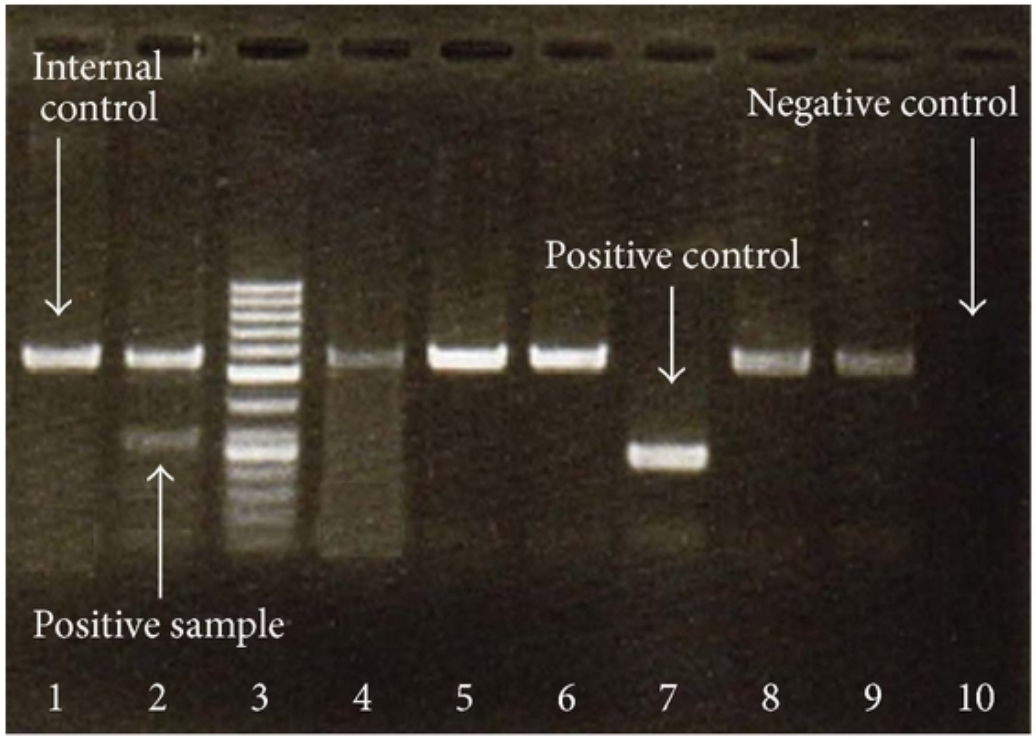
Hyatt *et al.* (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics.*

A quick reminder: Log-odds a.k.a. bit-scores

- Most alignment scores (from BLAST, HMMER, ...) can be interpreted as log-odds or bit-scores
- What is a log-odds score?
  - The observed frequency of two independent events is p_{AB}
  - The expected frequency of two independent events is $f_A * f_B$ (i.e. what is the likelihood of the event occuring by chance)
  - The log-odds score is $log_2(\frac{\#Observed}{\#Expected}) = log_2(\frac{p_{AB}}{f_A * f_B})$
  - What happens when $p_{AB} = f_A * f_B$?
  - What happens when $p_{AB} > f_A * f_B$ or $p_{AB} < f_A * f_B$?
- Often called the bit-score, information theorists like to discuss how many "bits" of information they have, hence "log_2".

Negative controls

- Negative controls are at least as important in bioinformatics as they are in the wetlab.
- Remember, Bioinformaticians lie too!



Prodigal practical

- Visit https://github.com/ppgardne/misc-projects/tree/master/genome_annotation
- Work through Excercise 1-4 in the README, view your predictions in Artemis.

Prodigal scores in the *E. coli* K-12 genome

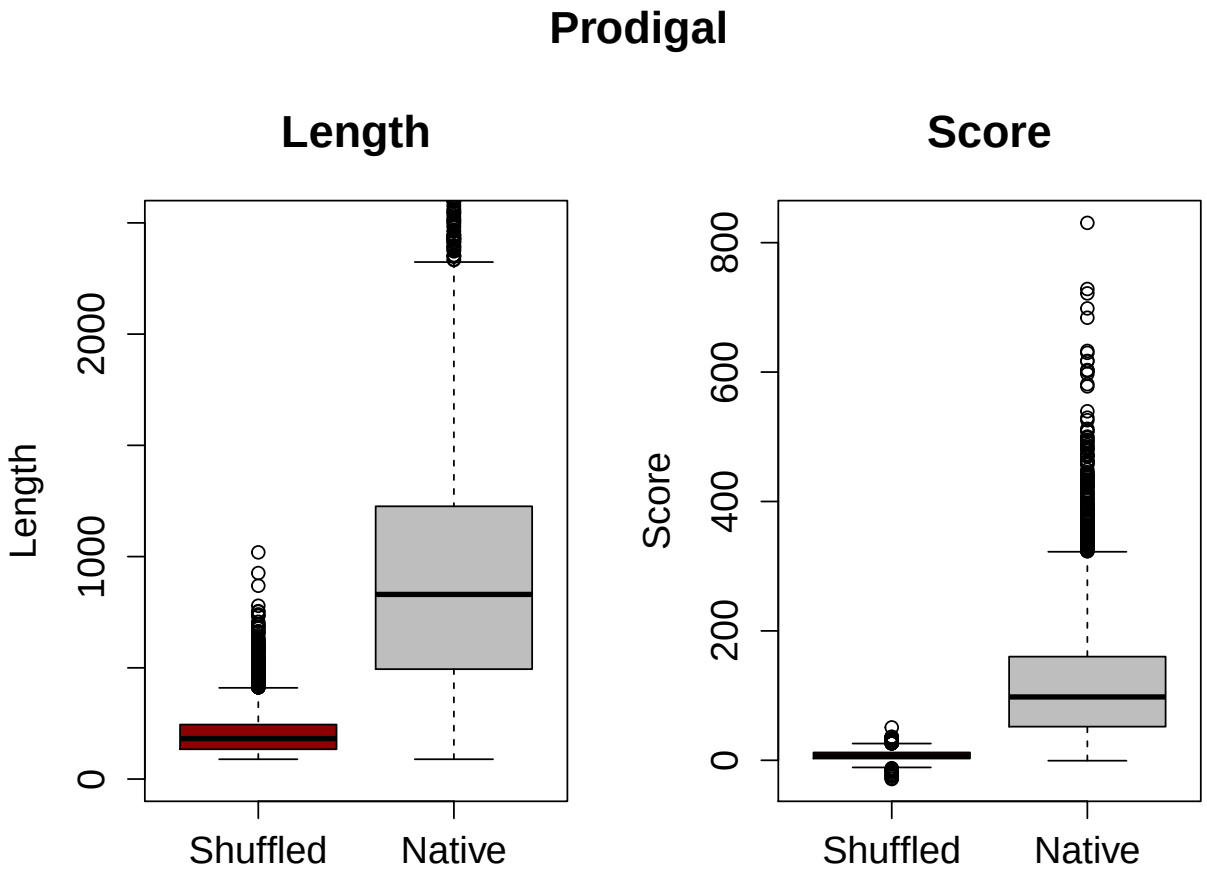
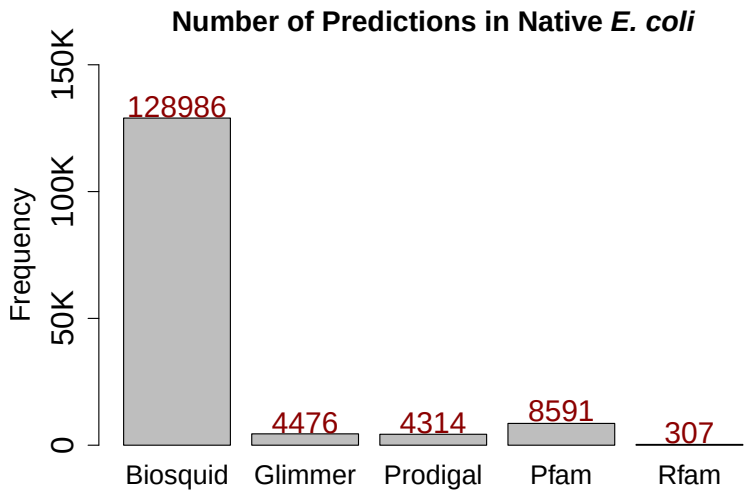


Image provided by Fatemeh Ashari Ghomi.

Number of predictions in the *E. coli* K-12 genome



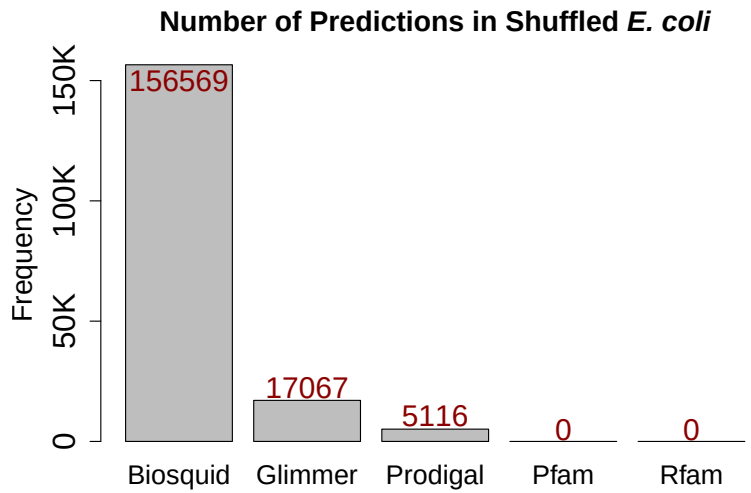
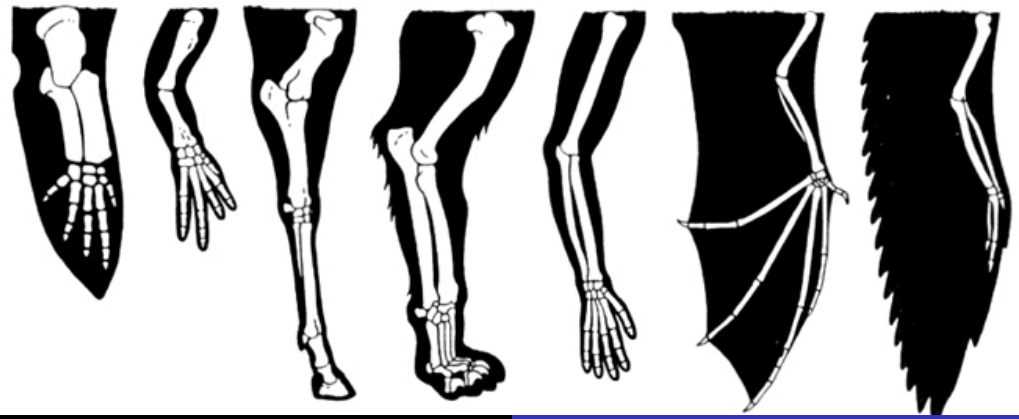


Image provided by Fatemeh Ashari Ghomi.

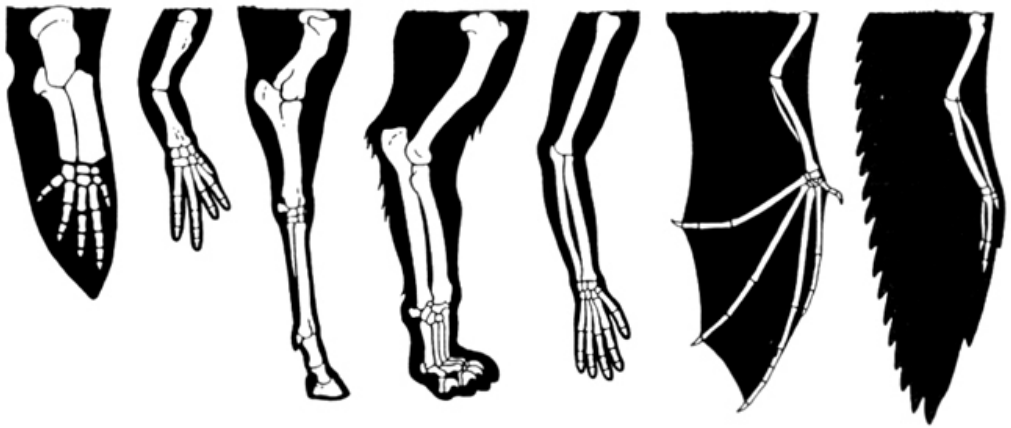
We can use homology...

- Evolution tends to preserve functional genomic regions (see the conservation track in the UCSC browser)...
- Example 1: Use an existing set of genes from related species and map these onto your genome
- Example 2: Align two or more related genomes, look for conserved regions, patterns of variation can be indicative of function
  - How reliable are these approaches? What are the main weaknesses & strengths? Archaeal Methanogen & Tuatara genomes?



We can use homology...

- Example 3: use profile-based models of proteins (e.g. Pfam) on 6-frame translations



Who can spot the pattern?

```
# STOCKHOLM 1.0
platypus        GGAGCAGACGTCACTCACCCCCCAGCCGGAGAT
opossum         GGAGCAGATGTTACTCACCCTCCTGCTGGAGAT
sloth           GGAGCAGACGTCACACACCCTCCCGCGGGGGAT
armadillo       GGAGCAGACGTCACGCACCCTCCGGCAGGGGAT
tenrec          GGGGCCGACGTCACGCACCCCCCTGCGGGCGAT
elephant        GGAGCGGATGTCACACACCCGCCTGCGGGGGAT
shrew           GGCGCAGATGTCACGCATCCTCCAGCAGGGGAC
hedgehog        GGAGCAGATGTCACACACCCCCCAGCAGGAGAT
megabat         GGAGCAGATGTCACACACCCTCCTGCAGGAGAT
microbat        GGAGCAGATGTCACCCACCCCCCTGCAGGGGAC
dog             GGAGCGGATGTCACACACCCCCCAGCCGGGGAC
cat             GGAGCCGATGTCACGCACCCCCCAGCAGGGGAT
horse           GGAGCGGATGTCACACACCCTCCGGCAGGGGAT
pika            GGAGCAGATGTCACTCACCCTCCAGCTGGGGAT
rabbit          GGTGCAGATGTCACACACCCCCCAGCTGGAGAT
squirrel        GGAGCAGATGTCACTCACCCTCCAGCGGGAGAT
guinea_pig      GGAGCAGATGTCACACACCCACCAGCGGGAGAT
mouse           GGAGCAGATGTCACTCATCCGCCTGCTGGGGAC
rat             GGAGCAGATGTCACTCATCCACCTGCTGGGGAT
kangaroo_rat    GGAGCAGATGTTACACACCCTCCAGCAGGGGAT
tree_shrew      GGCGCAGACGTCACGCACCCCCCGGCCGGGGAT
human           GGAGCGGATGTCACACACCCCCCAGCAGGGGAT
tarsier         GGTGCTGATGTCACACACCCCCCTGCAGGGGAT
marmoset        GGAGCAGATGTCACACACCCACCAGCAGGGGAT
zebrafinch      GGAGCAGATGTCACTCACCCTCCCGCCGGGGAT
green_anole     GGGGCAGACGTCACTCACCCGCCAGCCGGGGAC
xenopus         GGAGCAGATGTTACACACCCACCTGCTGGTGAT
pufferfish      GGTGCGGATGTTACTCATCCTCCTGCTGGTGAT
fugu            GGGGCTGATGTTACTCACCCTCCAGCTGGTGAT
stickleback     GGTGCAGACGTCACACATCCTCCAGCGGGTGAT
medaka          GGTGCCGATGTCACTCATCCTCCTGCCGGGGAC
zebrafish       GGGGCAGATGTTACACACCCGCCGGCTGGTGAT
lamprey         GGTGCCGATGTGACACACCCTCCAGCGGGAGAC
//
```

Reminder of the degenerate genetic code

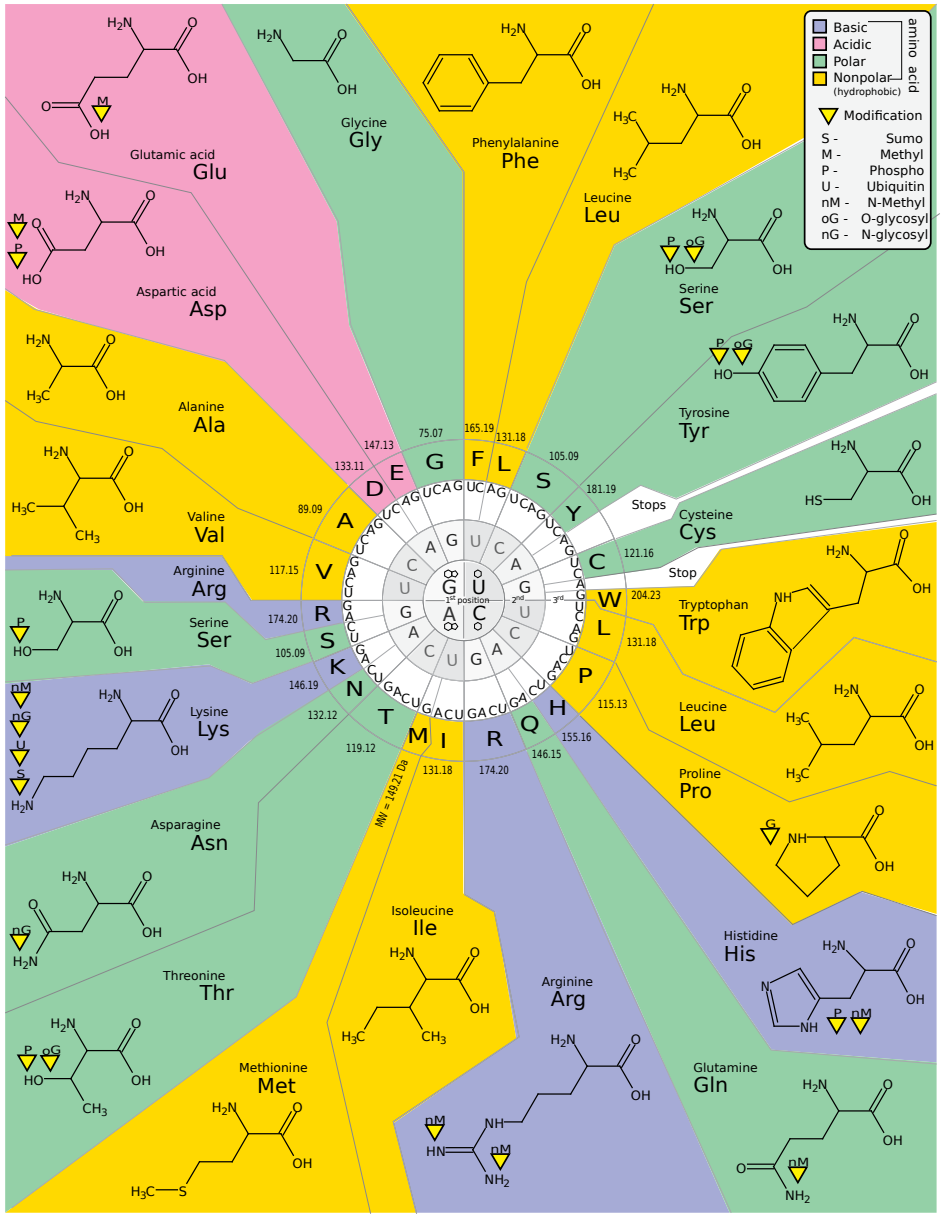
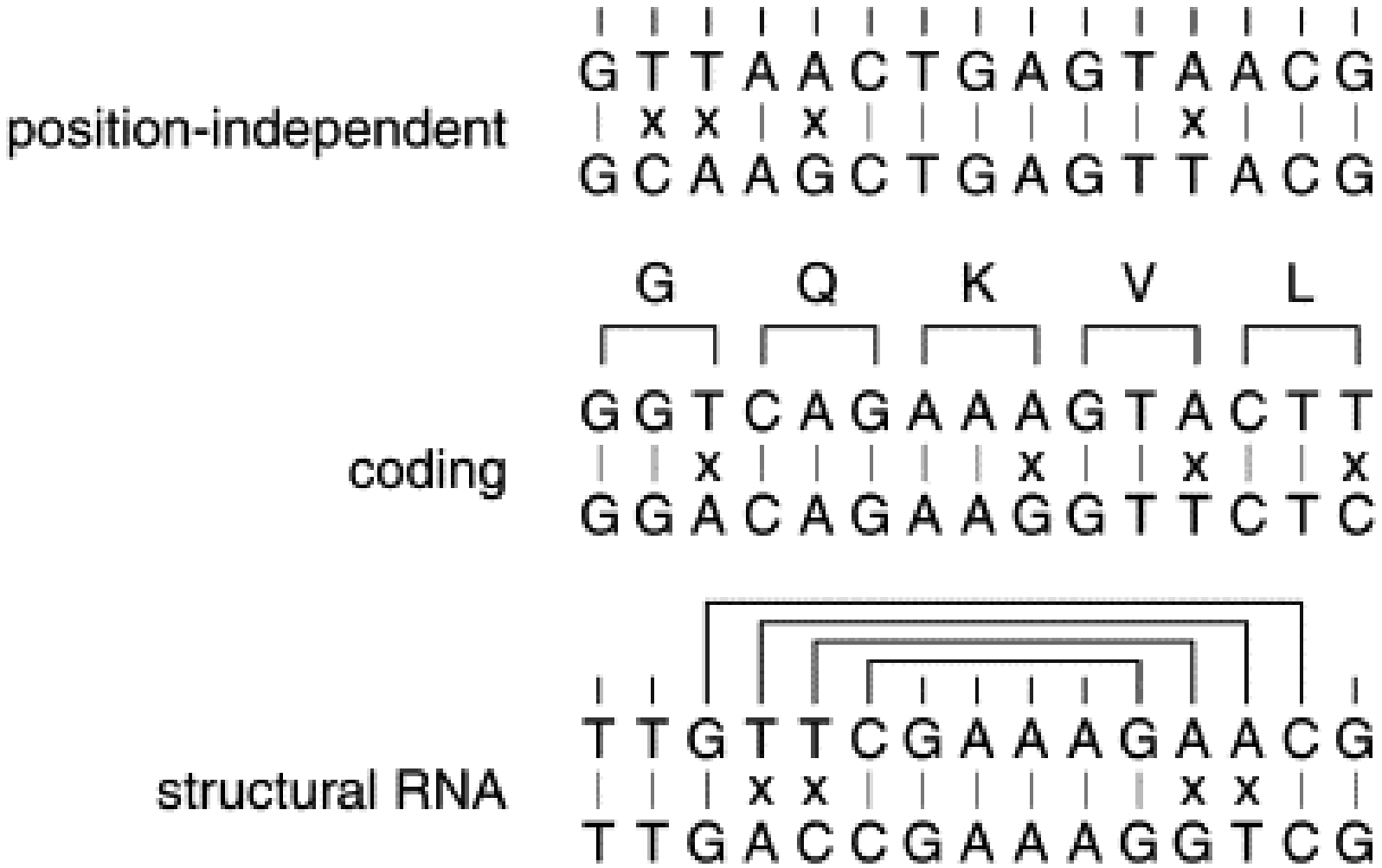


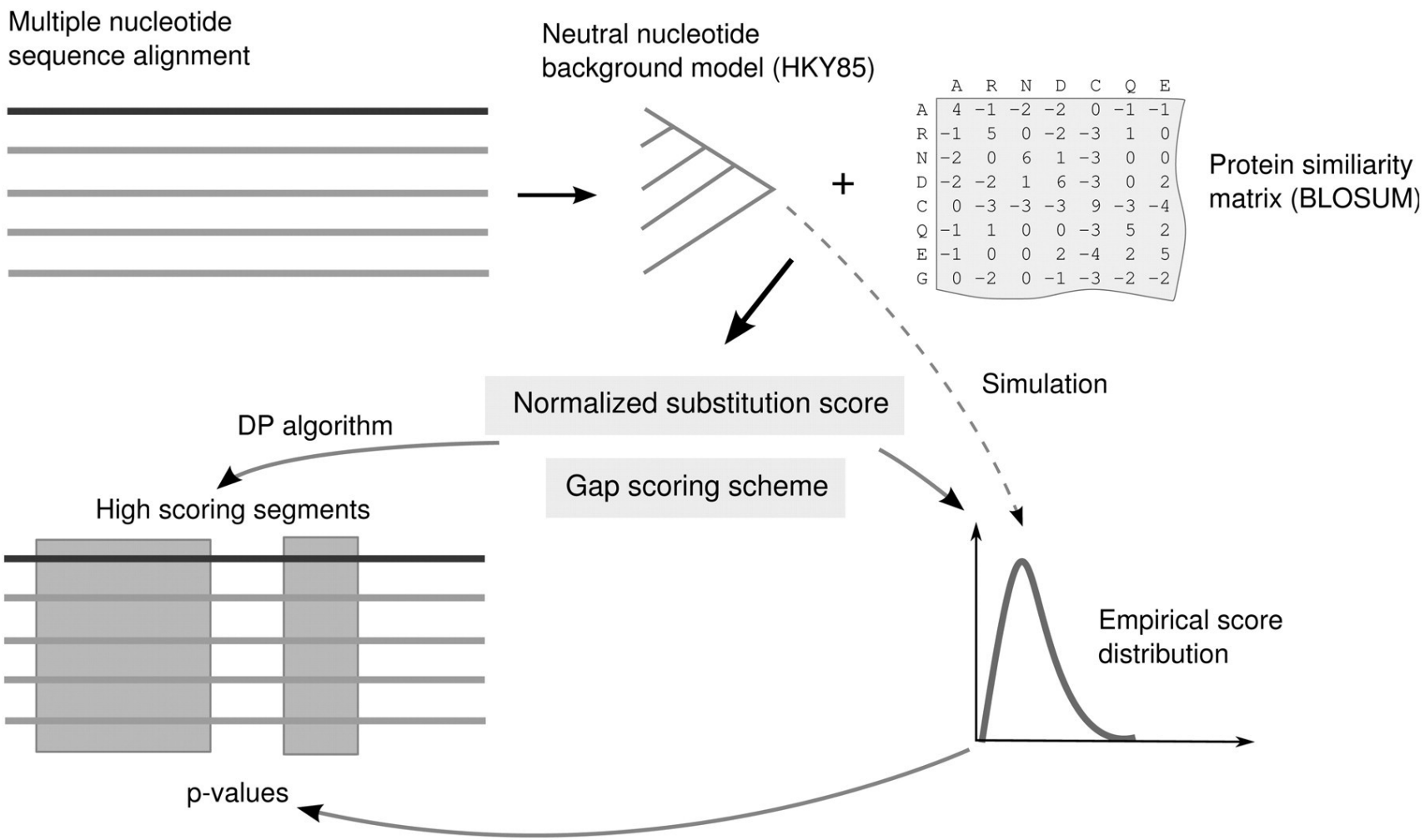
Image source: http://upload.wikimedia.org/wikipedia/en/d/d6/GeneticCode21-version-2.svg

The QRNA approach...



Rivas *et al.* (2001) Computational identification of noncoding RNAs in E. coli by comparative genomics. *Current Biology*.

RNAcode (see also PhyloCSF)



	A	R	N	D	C	Q	E
A	4	-1	-2	-2	0	-1	-1
R	-1	5	0	-2	-3	1	0
N	-2	0	6	1	-3	0	0
D	-2	-2	1	6	-3	0	2
C	0	-3	-3	-3	9	-3	-4
Q	-1	1	0	0	-3	5	2
E	-1	0	0	2	-4	2	5
G	0	-2	0	-1	-3	-2	-2

Washietl *et al.* (2011) RNAcode: Robust discrimination of coding and noncoding regions in comparative sequence data. *RNA.*

Profile HMM Background

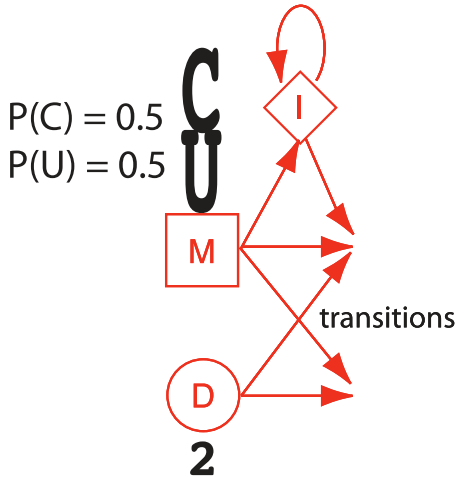
yeast	GUCUUCGGCAC
fly	GCCUUCGGAGC
cow	GCAUUCGUCGC
mouse	GCUUUCGAUGC
human	GCGUUCGCUGC
chicken	GUAUUCGUAAC
snake	GUGUUCGCGAC
croc	GUUUUCGAGAC

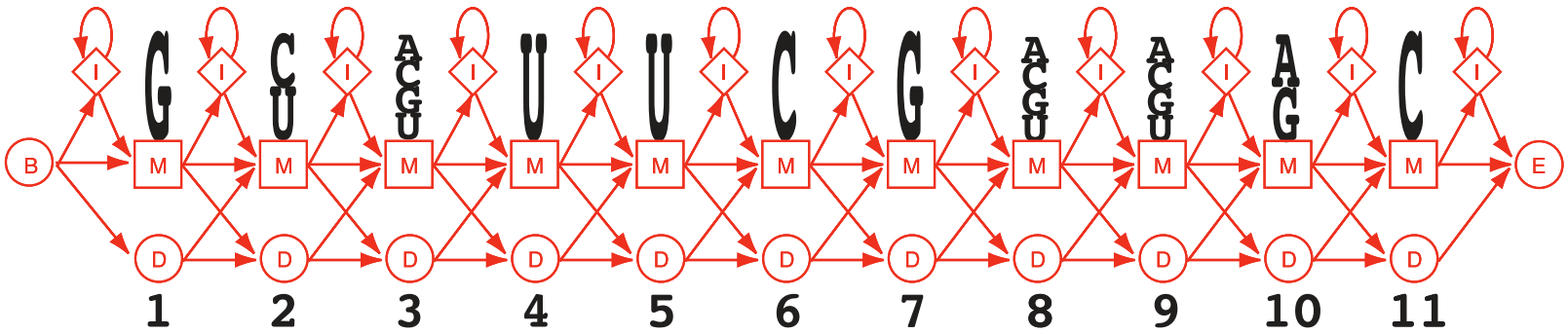
One HMM node per alignment column

3 states per node:
(M) Match: emits residues
(I) Insert: inserts extra residues
(D) Delete: deletes residues

HMMs generate homologous sequences.

Node for column 2:





Krogh, A. *et al.* (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol.*

Image provided by Eric Nawrocki.

Profile HMM Background


```
yeast    GU.CUUCGGCAC
fly      GC.CUUCGGAGC
cow      GC.AUUCGUCGC
mouse    GC.UUUCGAUGC
human    GC.GUUCGCUGC
chicken  GU.AUUCGUAAC
snake    GU.GUUCGCGAC
croc     GU.UUUCGAGAC
worm     GC.GUUCGCGGC
corn     GUGAUUCGU.GC
```

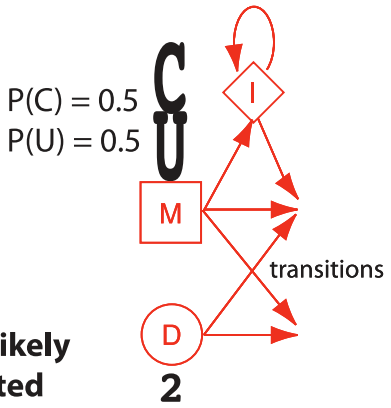
One HMM node per alignment column

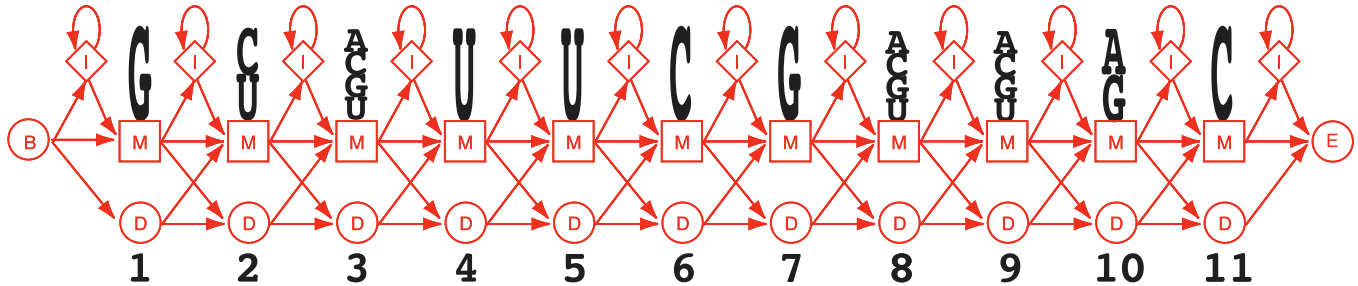
3 states per node:
(M) Match: emits residues
(I) Insert: inserts extra residues
(D) Delete: deletes residues

HMMs generate homologous sequences.

Given a sequence, the most likely path that could have generated that sequence can be computed.

Node for column 2:





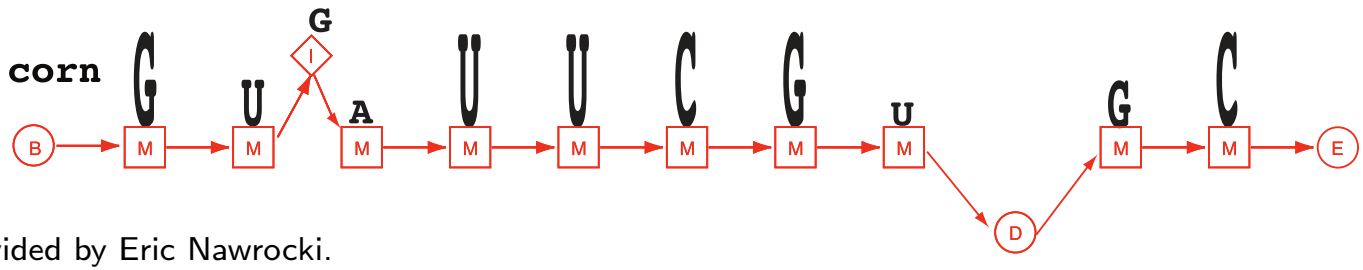
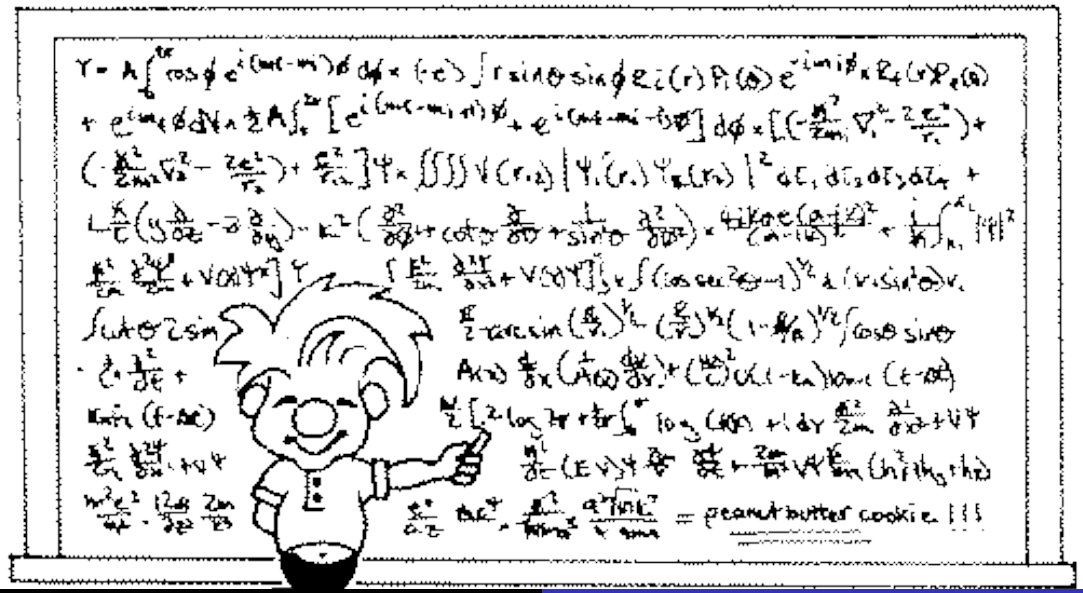


Image provided by Eric Nawrocki.

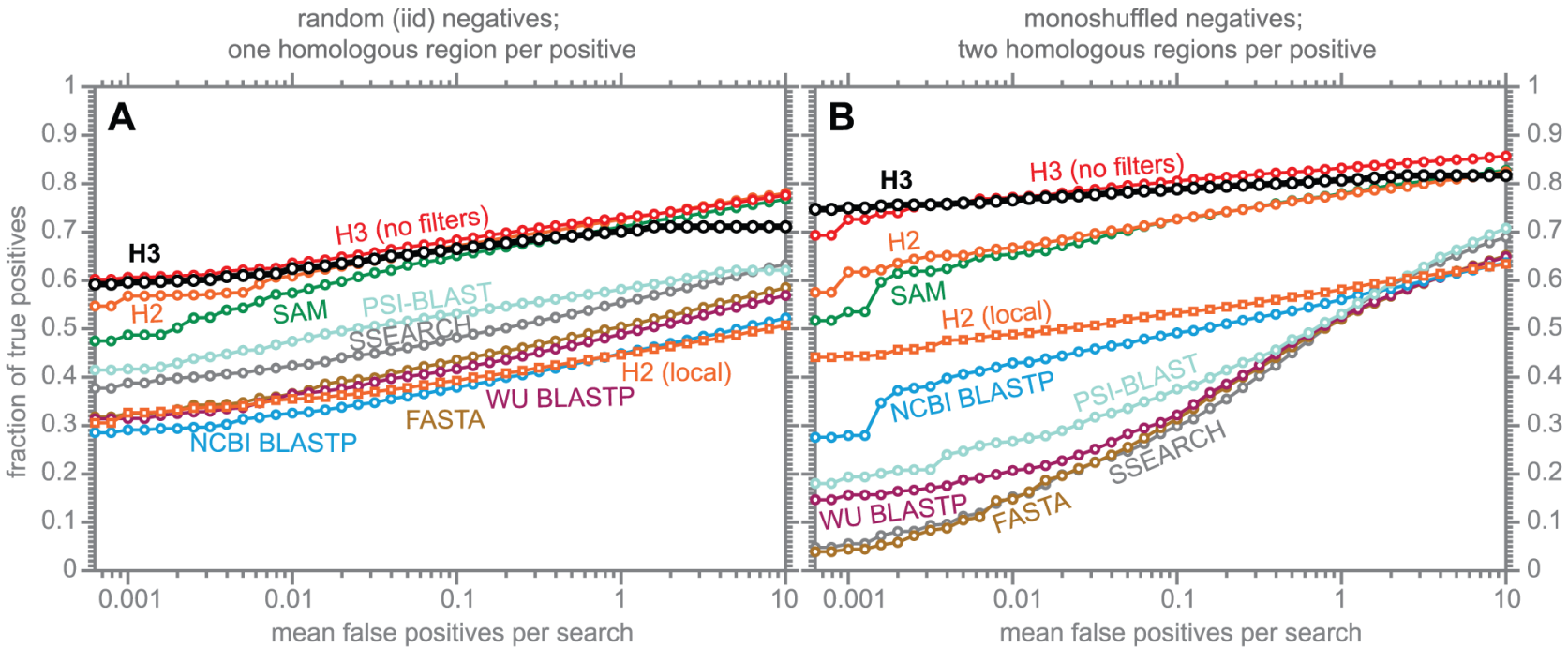
Profile HMM Background

- A **tree-weighting scheme** takes care of unbalanced alignments
- **Dirichlet-mixture priors** are used to incorporate information about amino-acid biochemistry
- **Effective sequence number** is used to down-weight priors when many sequences are available
- Transition probabilities to Insert & Delete states are estimated from the alignment



Why not just use BLAST?

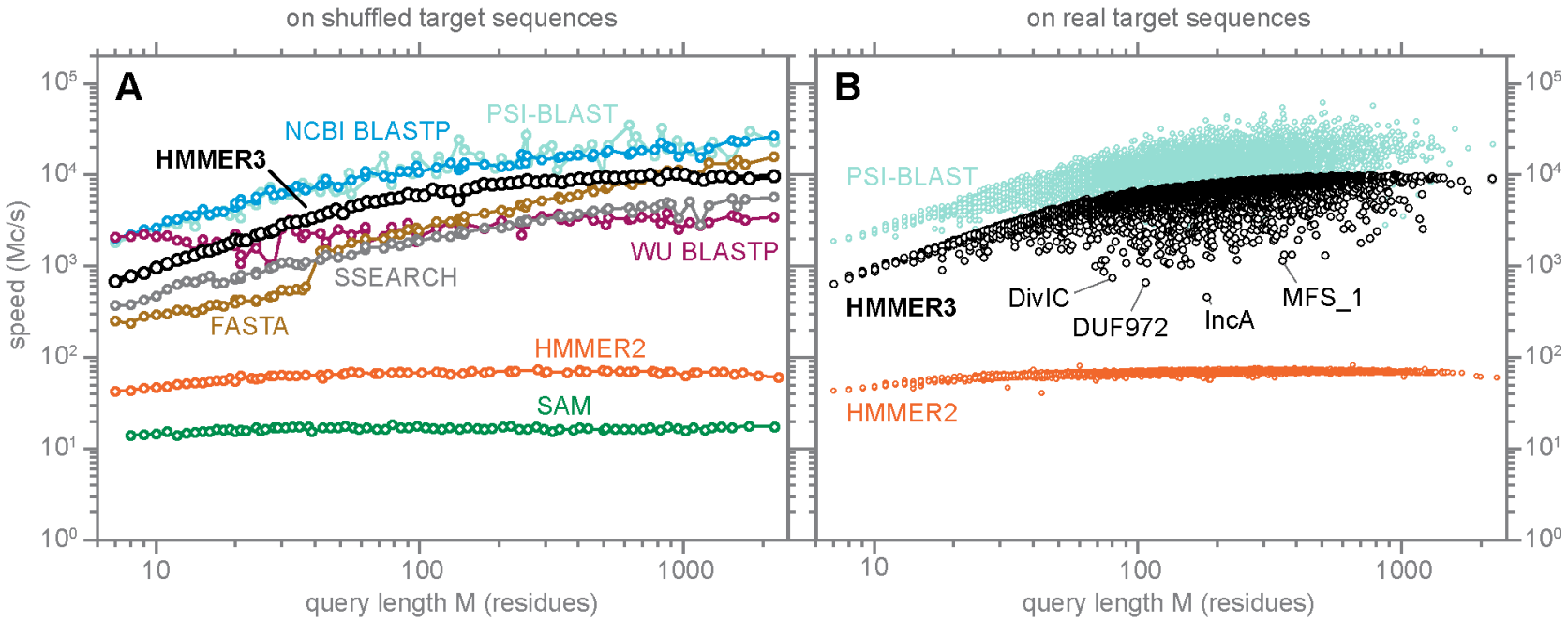
- ACCURACY!
  - Every benchmark of homology search tools has shown that profile methods are more accurate than single-sequence methods.



Eddy (2011) Accelerated Profile HMM Searches. *PLoS Computational Biology.*

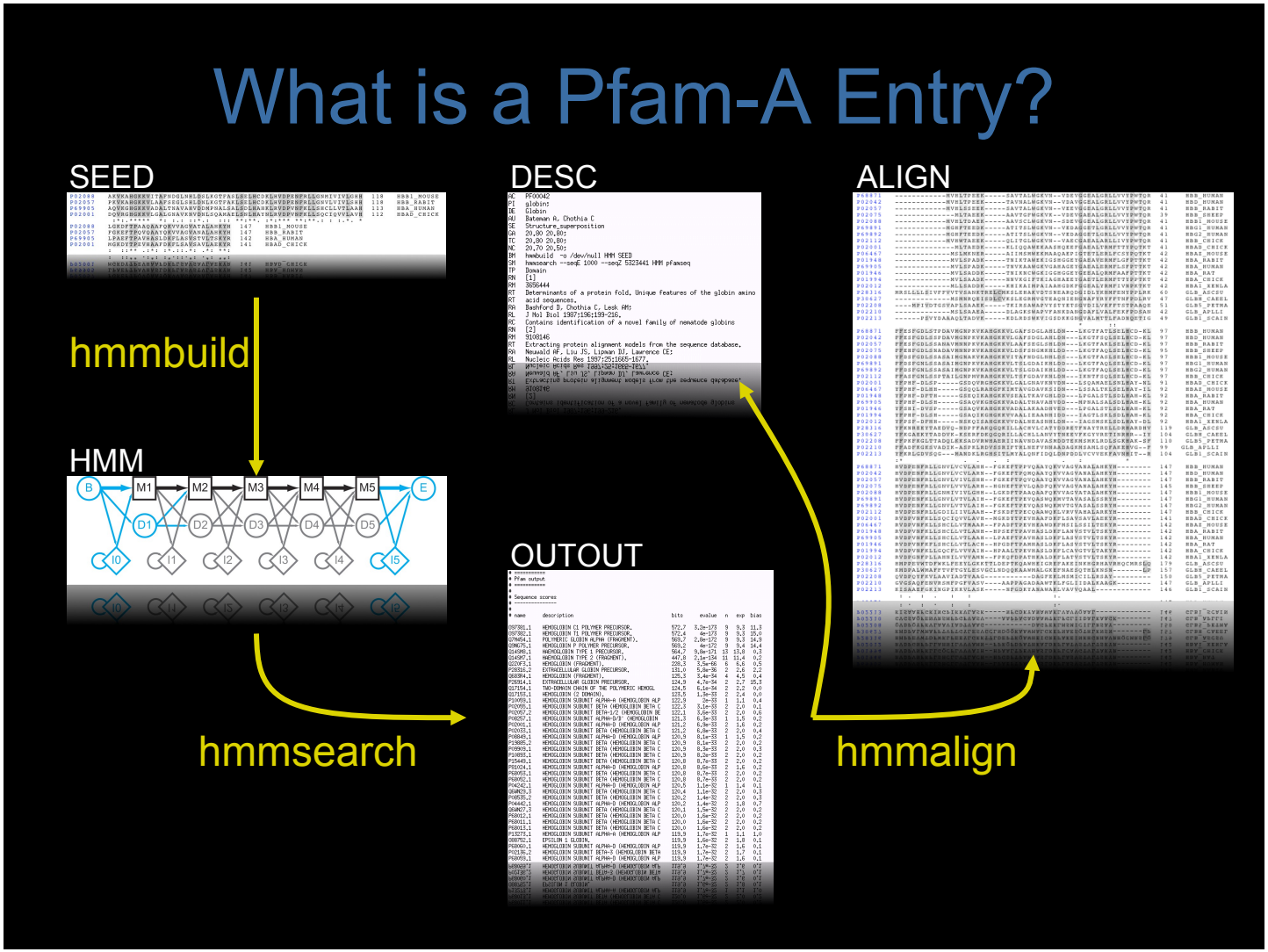
Why not just use BLAST?

- SPEED! To search a single query vs a database of all proteins:
  - BLAST: searches 42 million UniProt sequences
  - HMMER: searches 15,000 Pfam profiles
- The search space is $\sim 3,000x$ smaller for profiles
  - Save Planet Earth, use HMMER3



Eddy (2011) Accelerated Profile HMM Searches. *PLoS Computational Biology.*

What is a Pfam-A Entry?

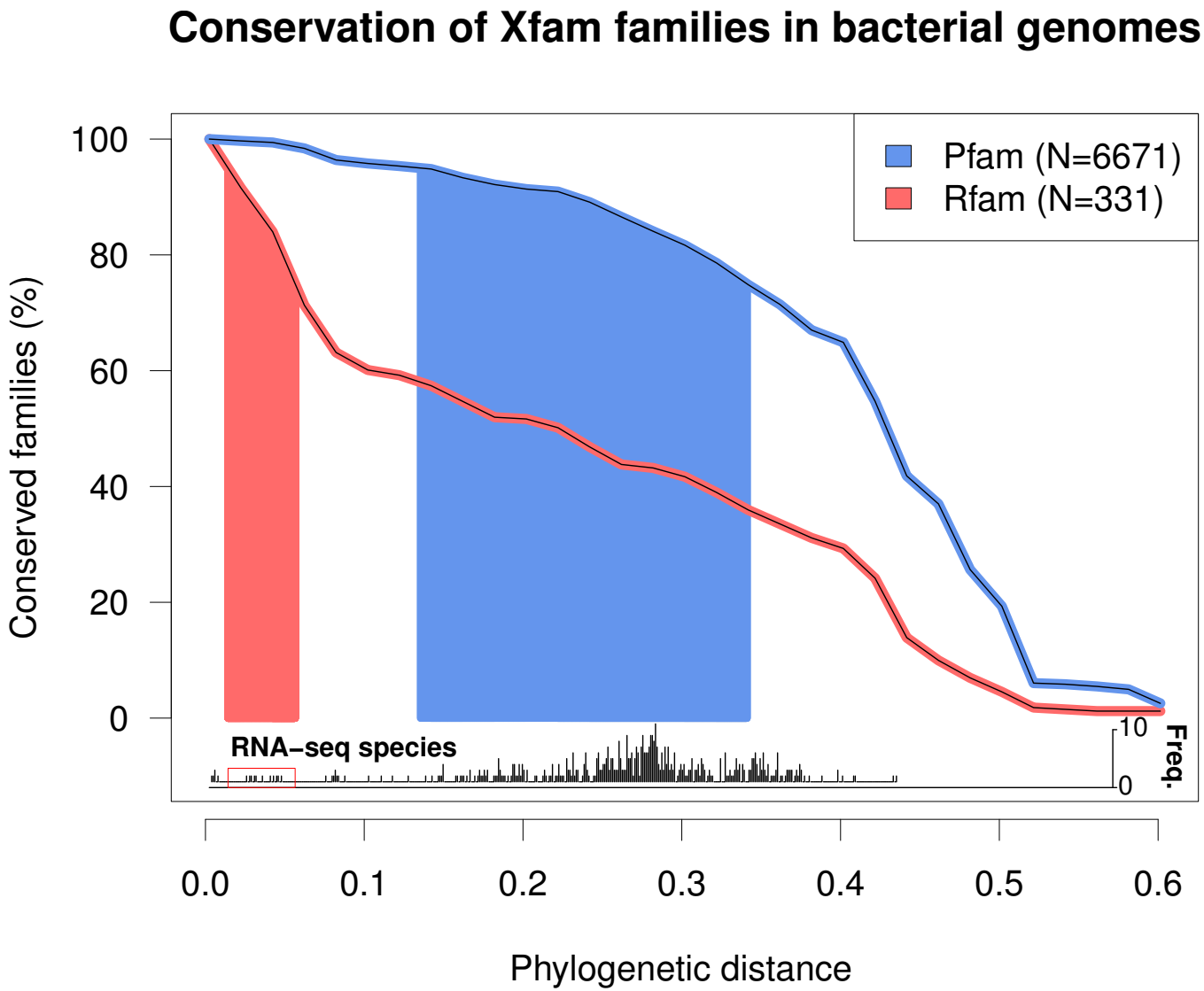


Slide borrowed from Rob Finn.

Homology-based annotation: strengths and weaknesses

- Example 1: map known genes onto genomes
  - Strengths: fast, cheap, ...
  - Weaknesses:
    - Inaccurate for divergent species (e.g. Graeme's *Methanobrevibacter* or GEBA genomes)
    - Requires manual correction of border-line results
- Example 2: aligning genomes
  - Strengths:
    - "cheap" if genomes already exist
    - fast for small genomes
    - evolutionary support for all discoveries
  - Weaknesses:
    - Requires lots of powerful computers for large genomes
    - Inaccurate for divergent species (e.g. Neil's tuatara or Graeme's *Methanobrevibacter*)
    - Requires manual correction of border-line results

Homology annotation: nucleotides are difficult to align



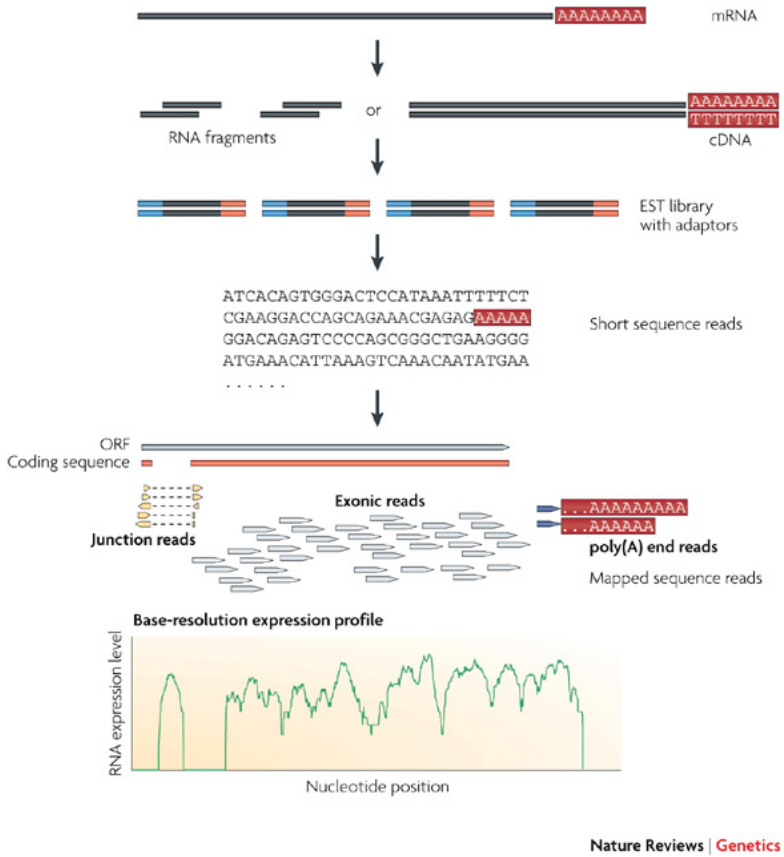
Lindgreen *et al.* (2014) Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. *PLOS Computational Biology.*

HMMER practical

- Visit https://github.com/ppgardne/misc-projects/tree/master/genome_annotation
- Work through Excercise 5 in the README, view your predictions in Artemis.
- Remember, Bioinformaticians still lie!

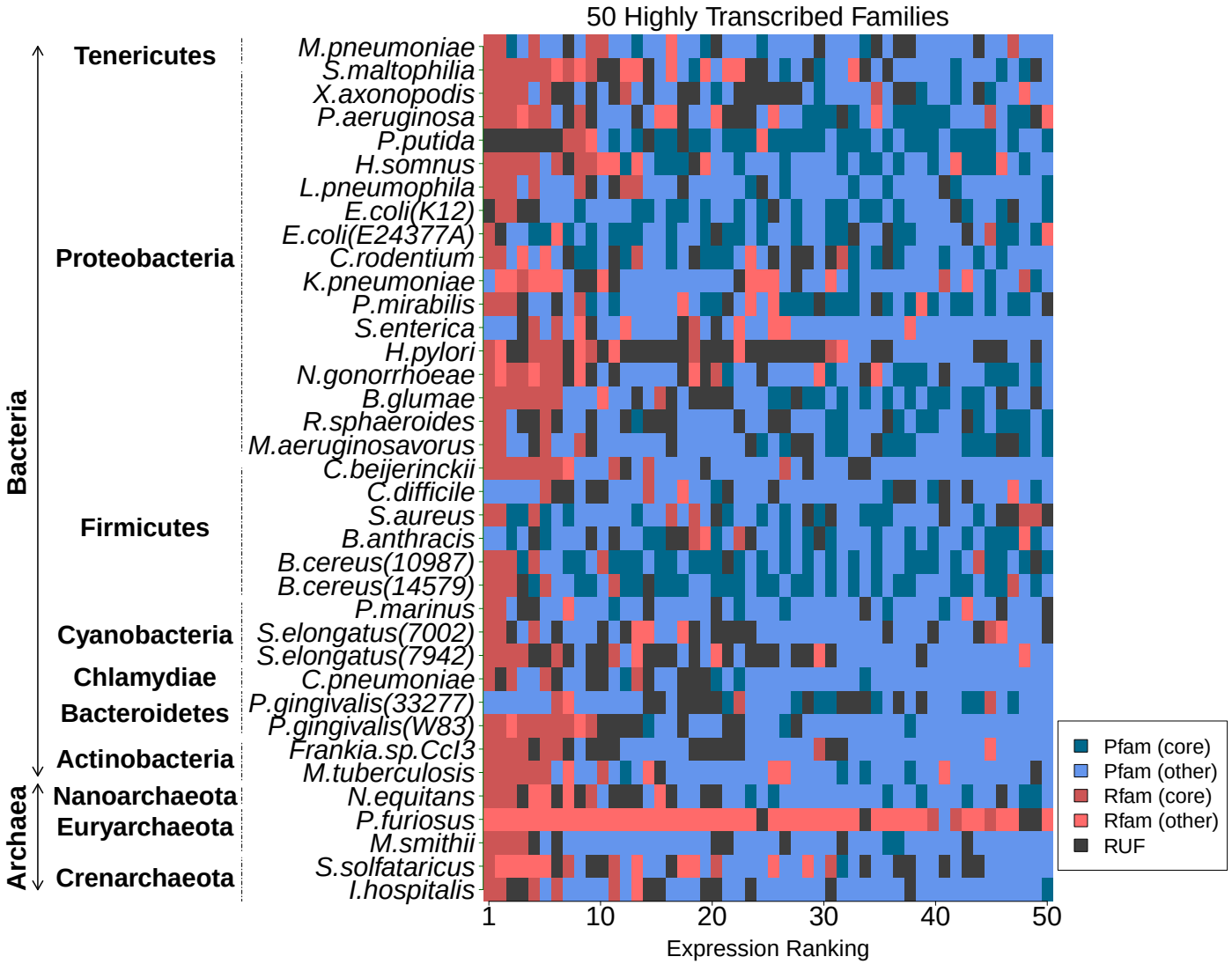
We can use RNA detection methods...

- Remember the central dogma of molecular biology
- Example: sequence RNAs from multiple tissues, developmental stages and environmental conditions
  - How reliable is this approach? What are the main weaknesses & strengths?



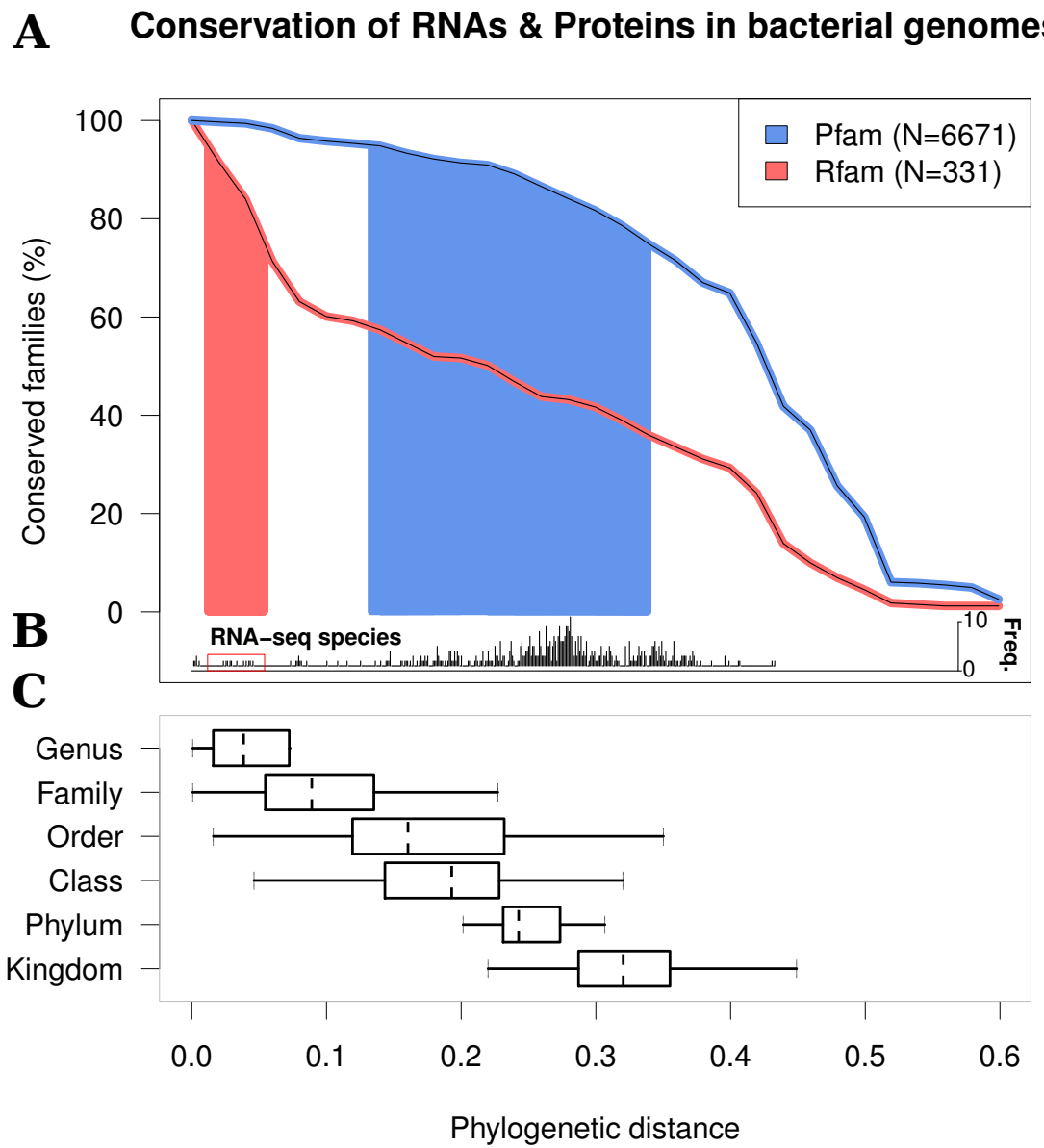
Wang, Gerstein & Snyder (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*.

The top 50 most abundant genes



Lindgreen *et al.* (2014) Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. *PLOS Computational Biology.*

Is there a "Goldilocks Zone" for comparative ncRNA work?



Lindgreen *et al.* (2014) Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. *PLOS Computational Biology.*

RNA-seq: strengths and weaknesses

- RNA-seq
  - Strengths:
    - Experimental support for transcribed regions
    - Identifies untranslated regions (UTRs), ncRNAs, antisense RNAs, ...
    - Identifies alternatively spliced and edited RNAs
  - Weaknesses:
    - Expensive & lots of work
    - RNA degradation and genomic contamination
    - Transcription does not prove translation
    - Will miss genes transcribed in specific developmental stages, tissues & environmental conditions E.g. lsy-6 microRNA

We can use protein detection methods...

- Central dogma of molecular biology
- Example: Protein mass spectrometry
  - How reliable is this approach? What are the main weaknesses & strengths?

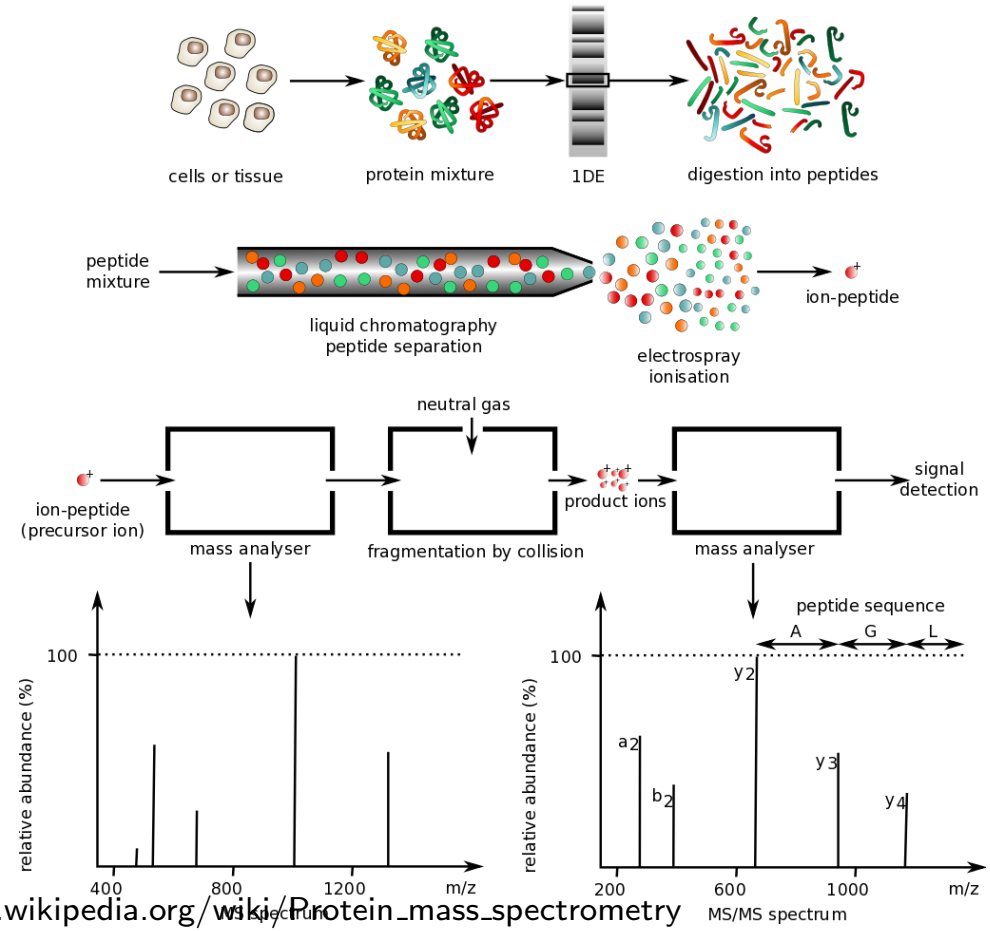


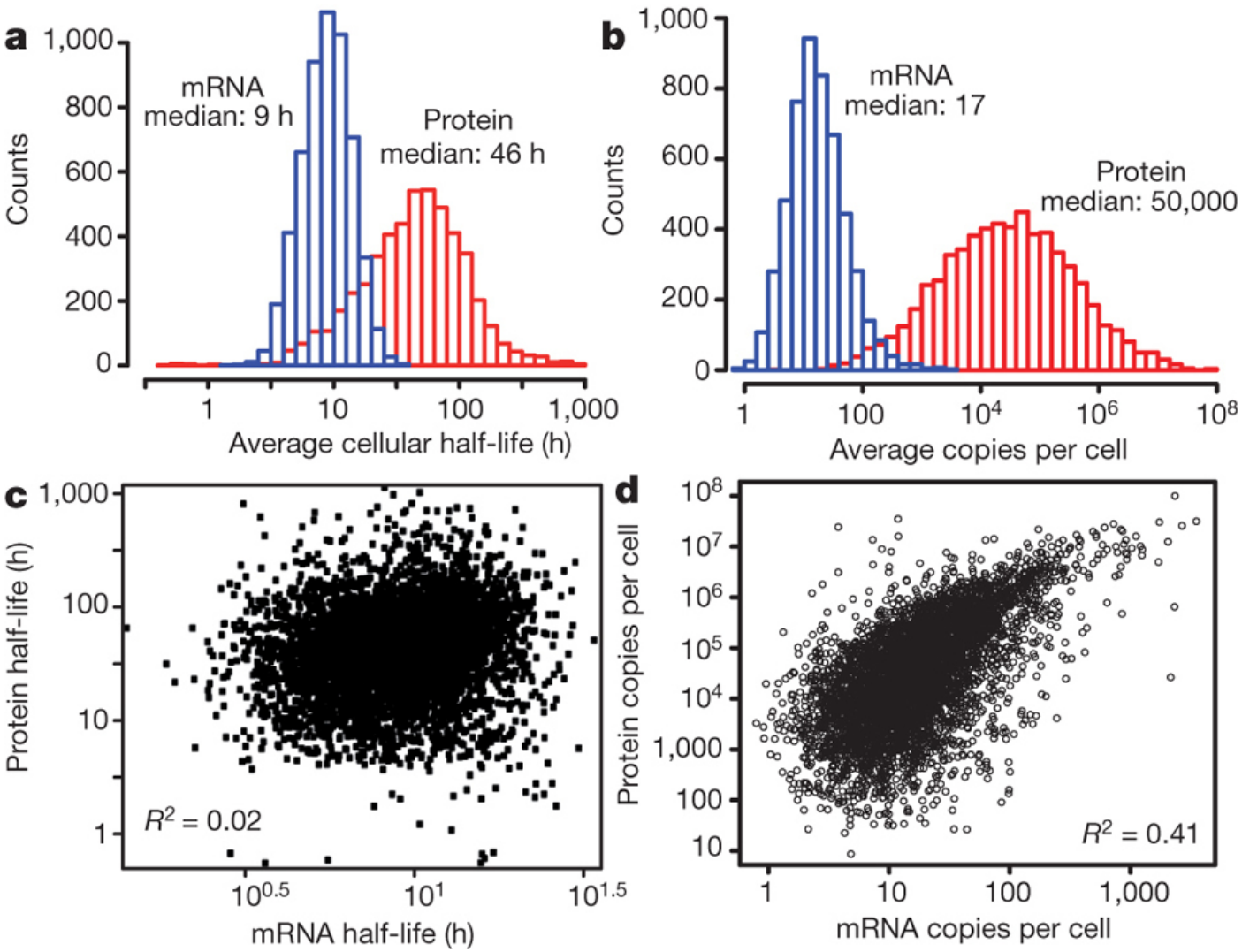
Figure from: http://en.wikipedia.org/wiki/Protein_mass_spectrometry

Protein mass spectrometry: strengths and weaknesses

- Protein mass spectrometry
  - Strengths:
    - Experimental support for translated regions
    - Identifies alternative isoforms and post-translational modifications (Ezkurdia *et al.* 2012)
  - Weaknesses:
    - Expensive & lots of work
    - Misses genes transcribed in specific developmental stages, tissues & environmental conditions
    - Currently technology generally only detects the most abundant proteins

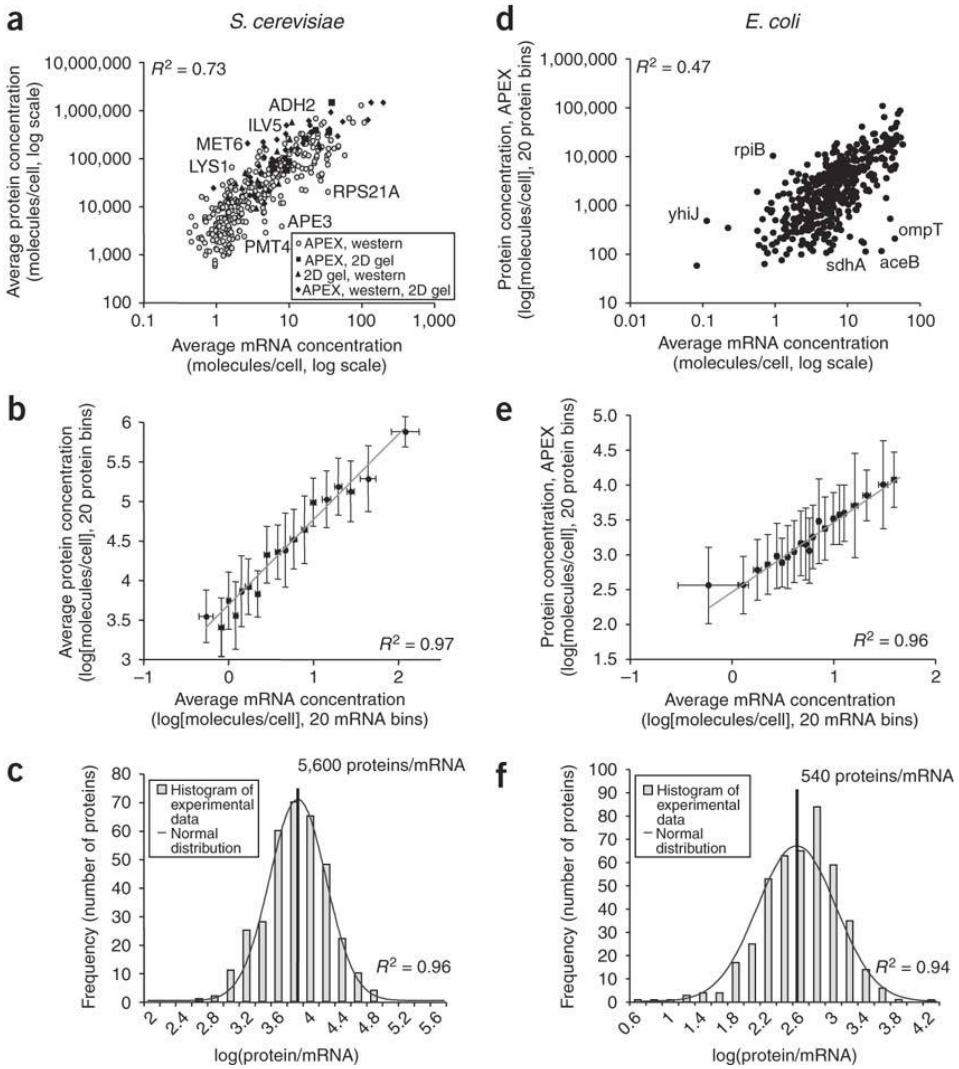
Ezkurdia *et al.* (2012) Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function. *Mol Biol Evol.*

How cool is this?!



Schwanhäusser *et al.* (2011) Global quantification of mammalian gene expression control. *Nature*

This is also kinda neat...



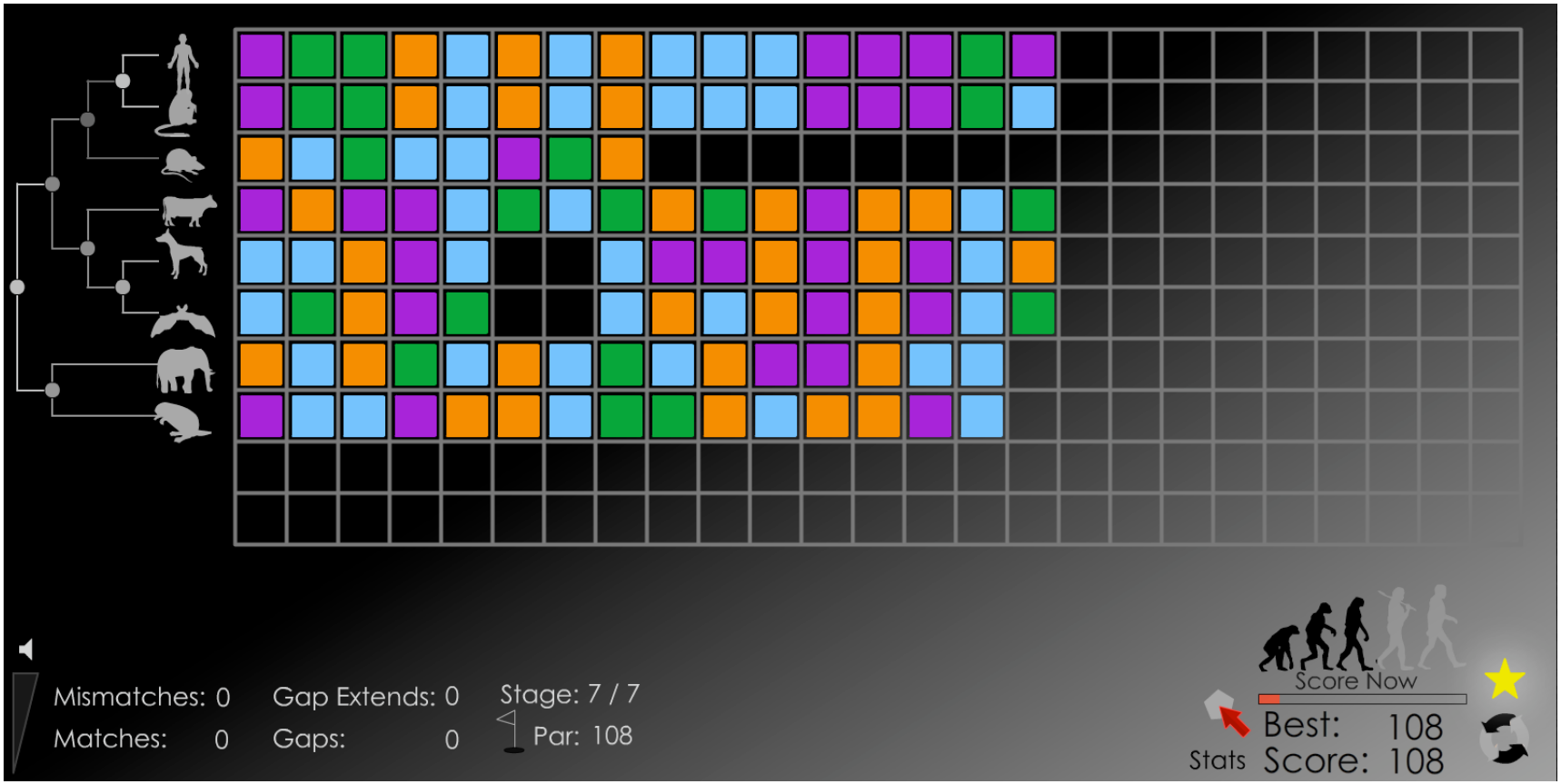
Lu *et al.* (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature Biotechnology*

RNA-seq & Mass-spec practical

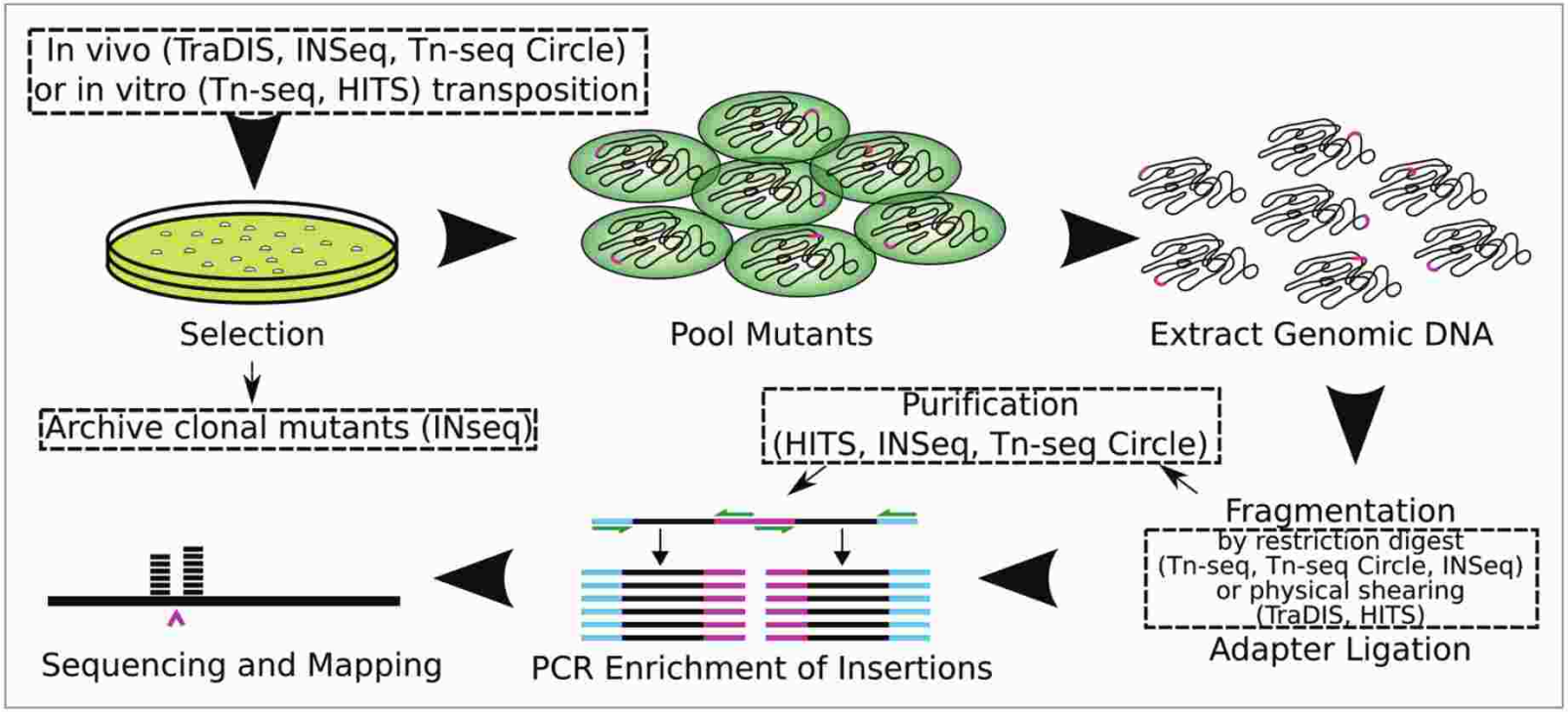
- Visit https://github.com/ppgardne/misc-projects/tree/master/genome_annotation
- Work through Excercise 7-8 in the README, view your predictions in Artemis.

Homework: How to make a sequence alignment?

- Play: http://phylo.cs.mcgill.ca
- or even better, play Ribo: http://ribo.cs.mcgill.ca/



Tn-seq

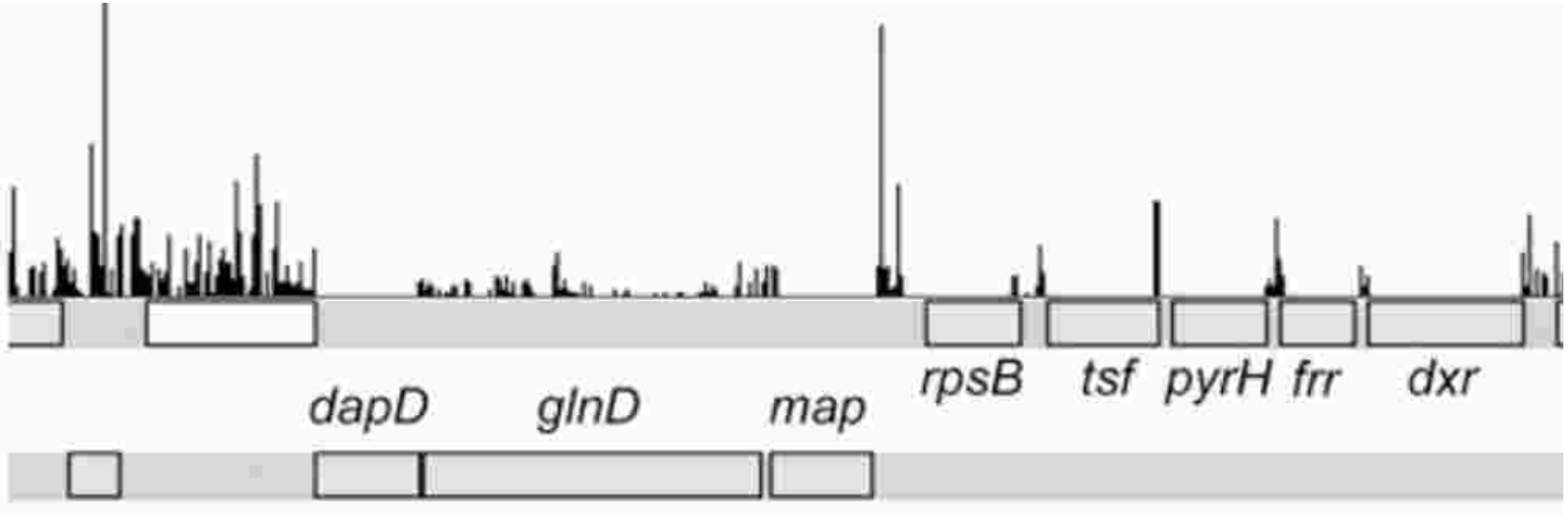


- Transposons are represented by pink lines, sequencing adaptors by blue, genomic DNA by black and PCR primers by green.

Barquist, Boinett & Cain (2013) Approaches to querying bacterial genomes with transposon-insertion sequencing. *RNA Biology*

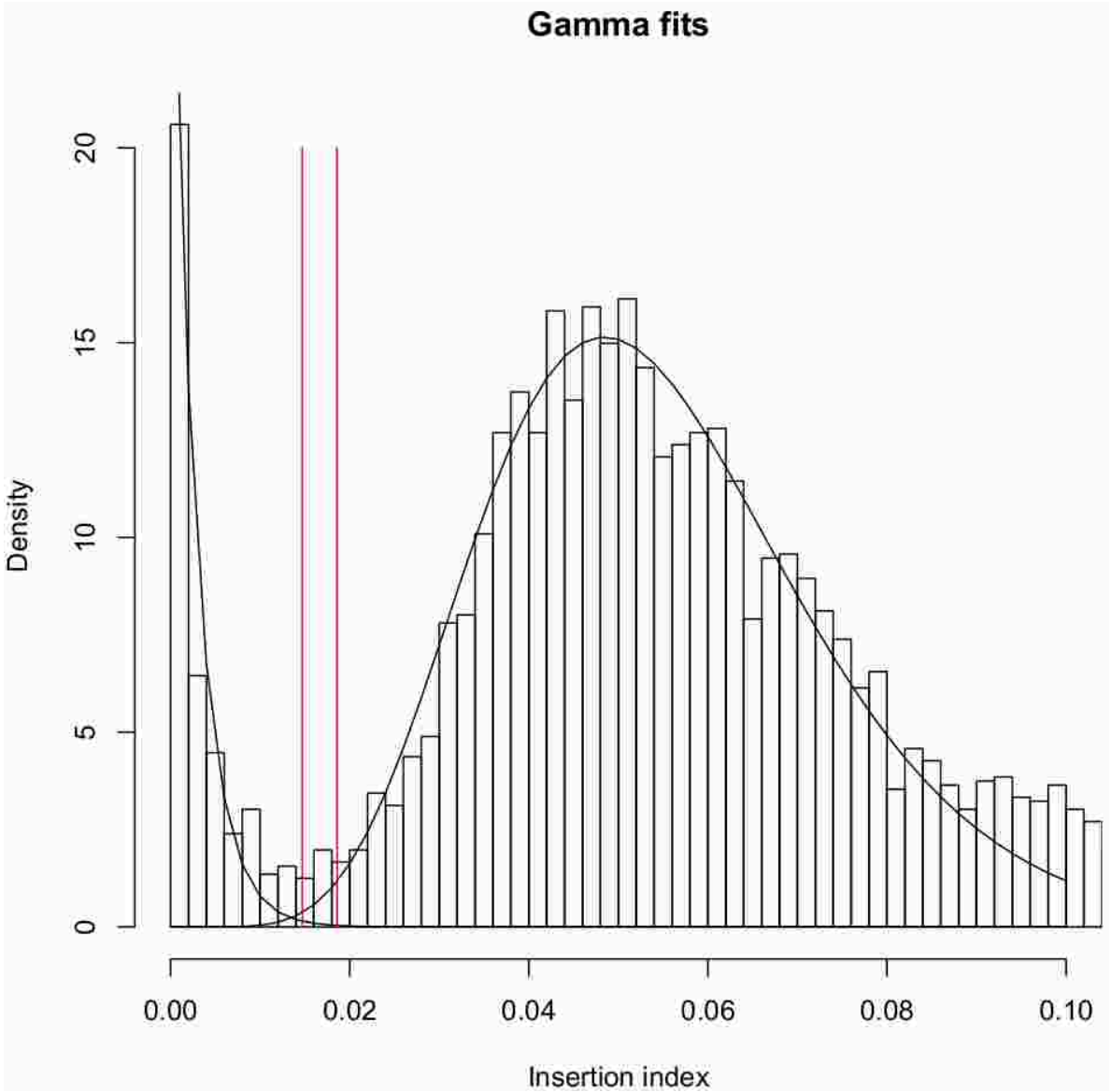
Tn-seq (called TraDIS in the UK)

- Electrotransformation with Tn5-derived transposon/transposase complex containing a kanamycin-resistance gene
- > 10 transformations per batch at $42,000 - 146,000$ mutants per batch; 13 batches $\rightarrow$ 1.1 million mutants
- Short read sequencing with transposon specific primers



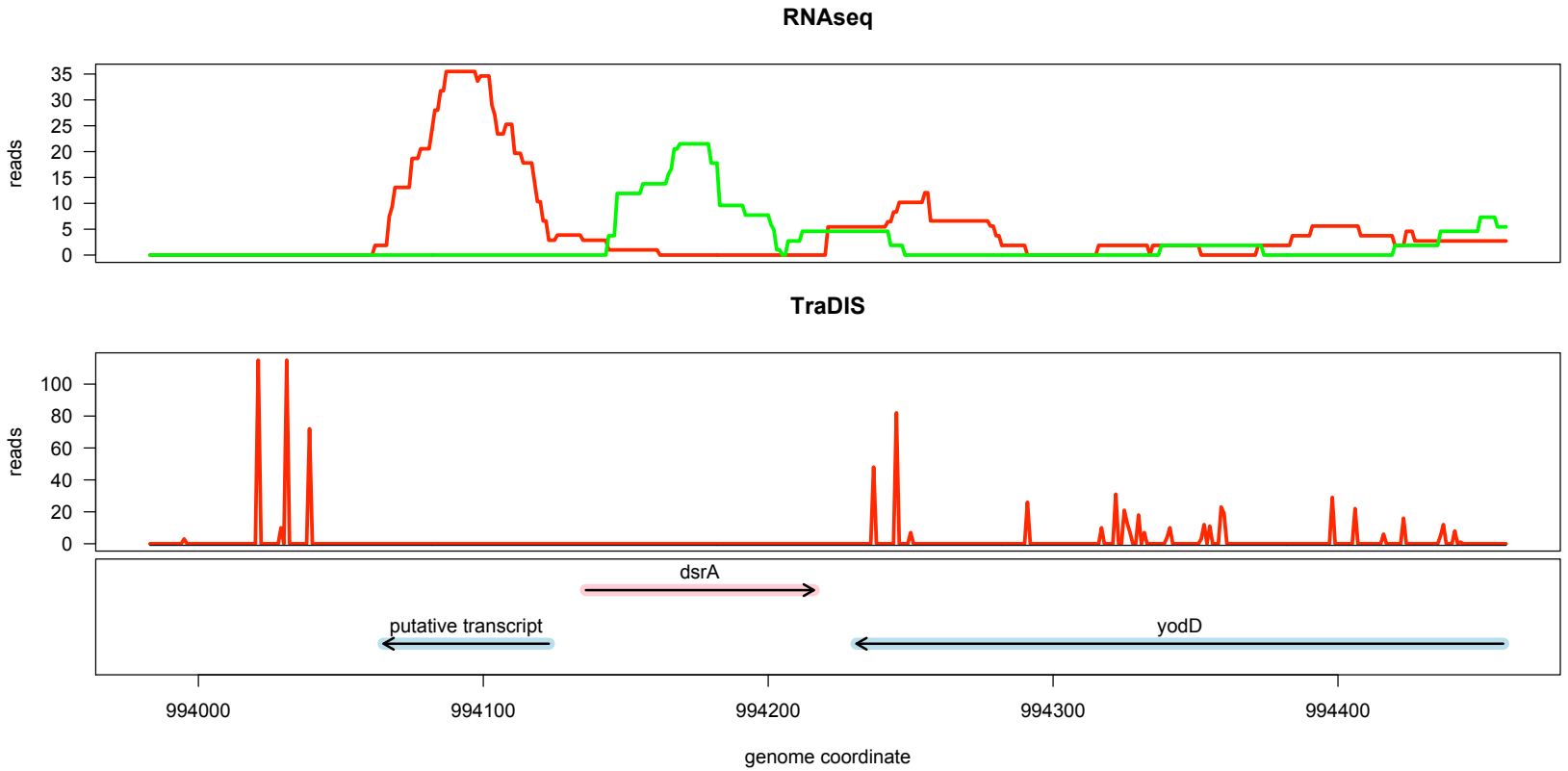
Langridge *et al.* (2009) Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants *Genome research*

Tn-seq

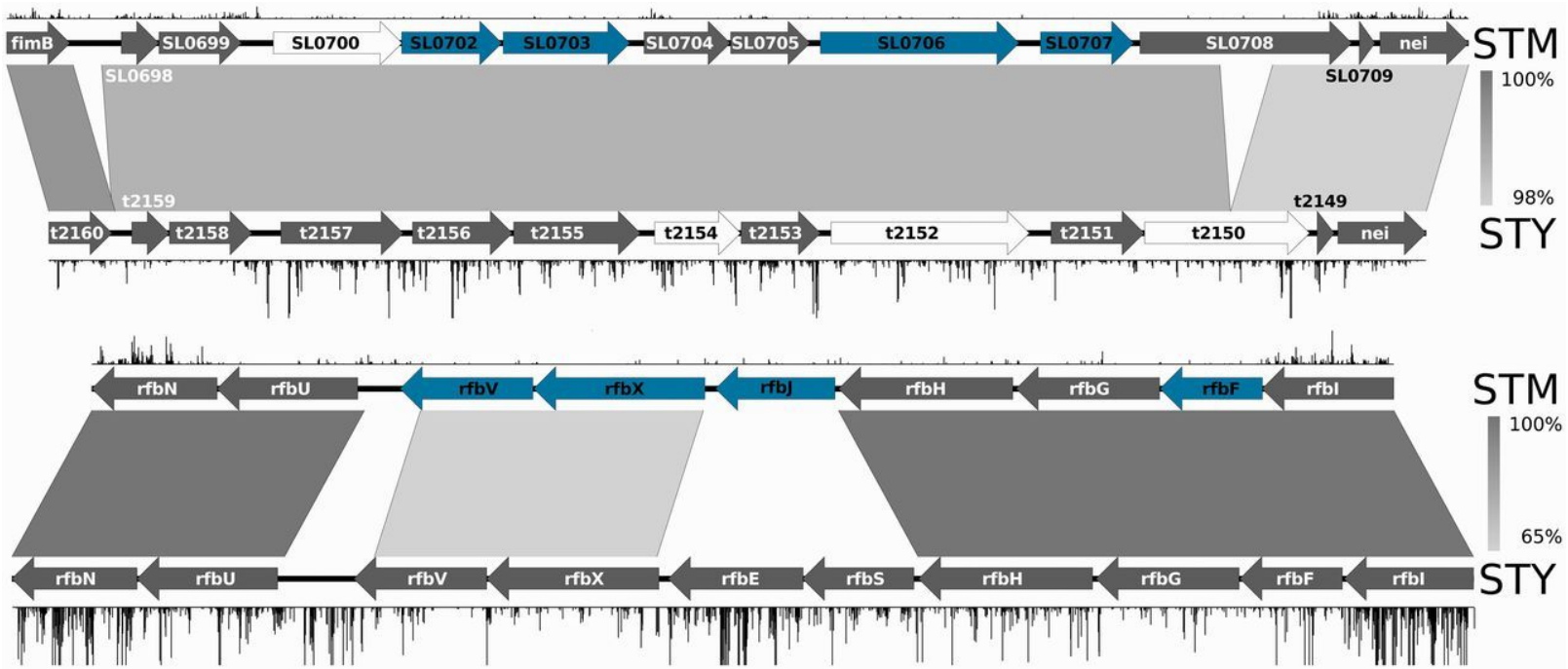


Langridge *et al.* (2009) Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants *Genome research*

Tn-seq on RNA in *Salmonella*

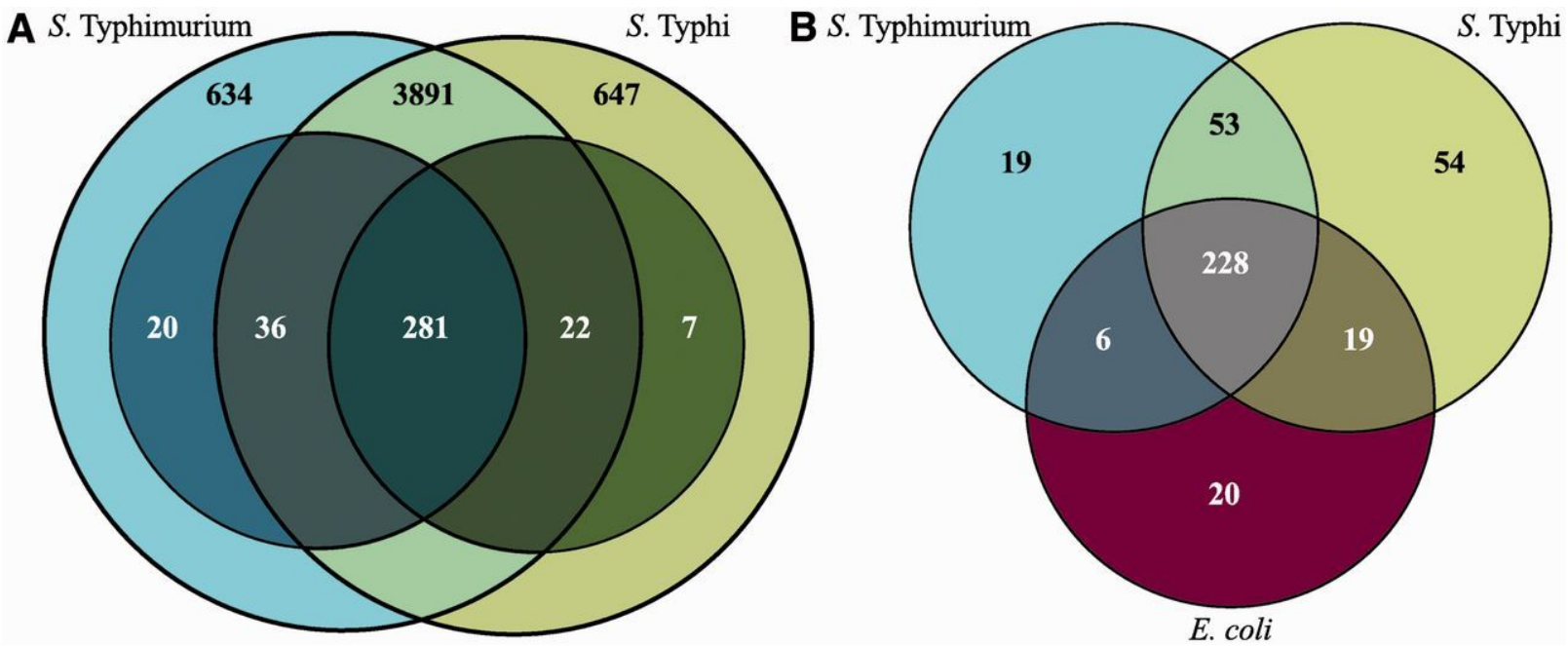


Comparative Tn-seq



Barquist L, Langridge GC, Turner DJ, Phan MD, Turner AK, Bateman A, Parkhill J, Wain J, **Gardner PP** (2013) A comparison of dense transposon insertion libraries in the Salmonella serovars Typhi and Typhimurium. *Nucleic Acids Research.*

Comparative Tn-seq



(A) the overlap of all genes (outer circles) and required genes (inner circles) Black numbers refer to all genes, white numbers to required genes.

(B) the overlap of all required genes between S. Typhimurium, S. Typhi and E. coli.

Barquist L, Langridge GC, Turner DJ, Phan MD, Turner AK, Bateman A, Parkhill J, Wain J, **Gardner PP** (2013) A comparison of dense transposon insertion libraries in the Salmonella serovars Typhi and Typhimurium. *Nucleic Acids Research.*

The End