

Genome Annotation

Paul Gardner

November 9, 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 Institute 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).

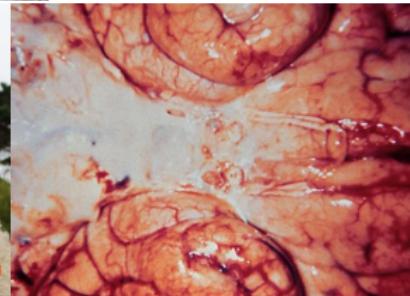
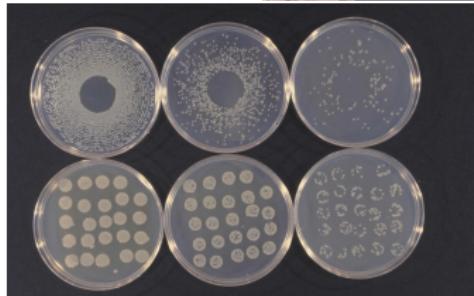


- ▶ Neil Gemmel at Otago University is sequencing the tuatara genome.



Discussion

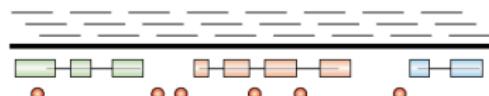
- ▶ How should these researchers annotate their genomes (after they have sequenced and assembled them)?
 - ▶ What are the fast and cheap methods?
 - ▶ What are the most accurate methods?



Nucleotide → Protein/RNA → Process

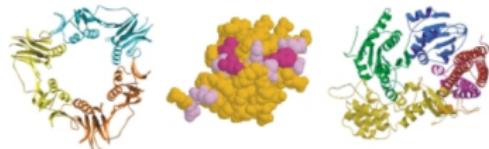
Where?

Nucleotide-level annotation



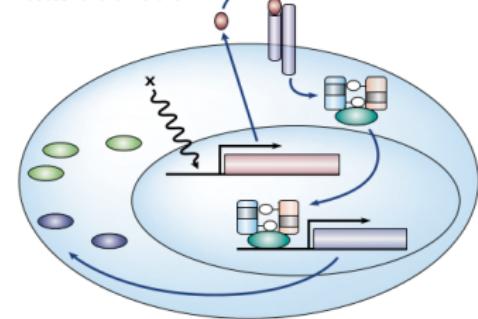
What?

Protein-level annotation



How?

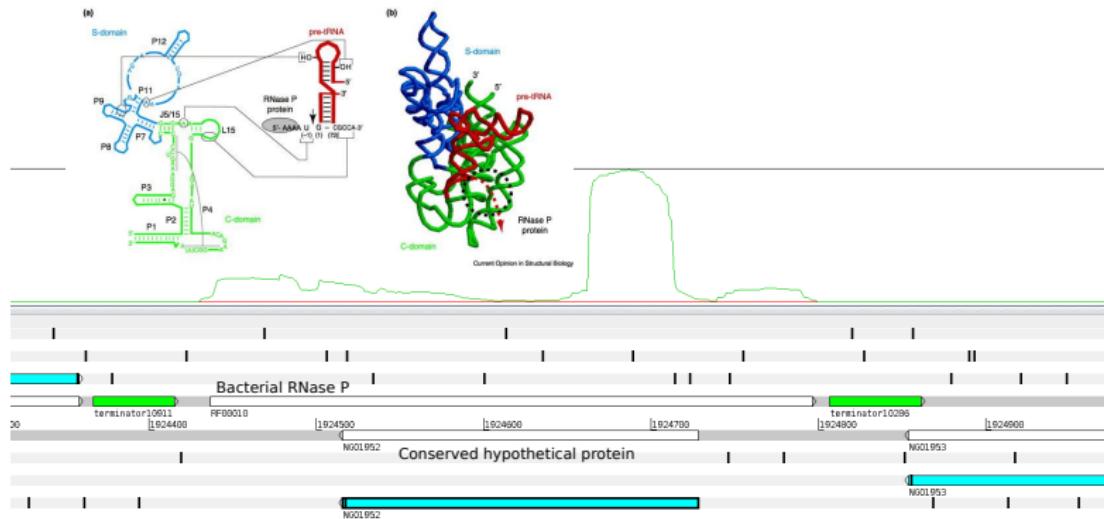
Process-level annotation



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

The screenshot shows the homepage of the journal "DATABASE: The Journal of Biological Databases and Curation". The header features the journal's name in large white letters on a dark blue background. Below the header, there is a navigation bar with links to "ABOUT THIS JOURNAL", "CONTACT THIS JOURNAL", "SUBSCRIPTIONS", and "CURRENT". A red banner at the bottom of the header area displays the text "Institution: University of Canterbury" and "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

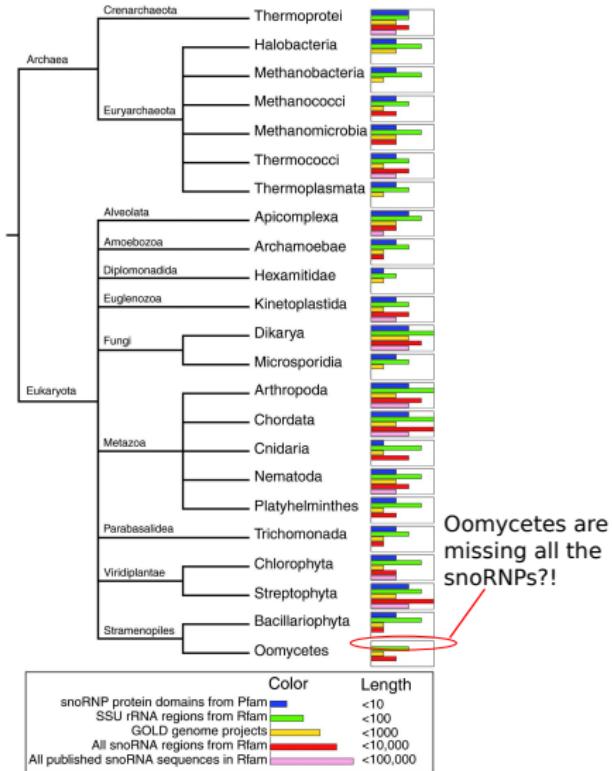
Ruth Y. Eberhardt^{1,*}, Daniel H. Haft², Marco Punta¹, Maria Martin³,

Claire O'Donovan³ and Alex Bateman¹

Author Affiliations

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

Embarrassing genome annotation screwups



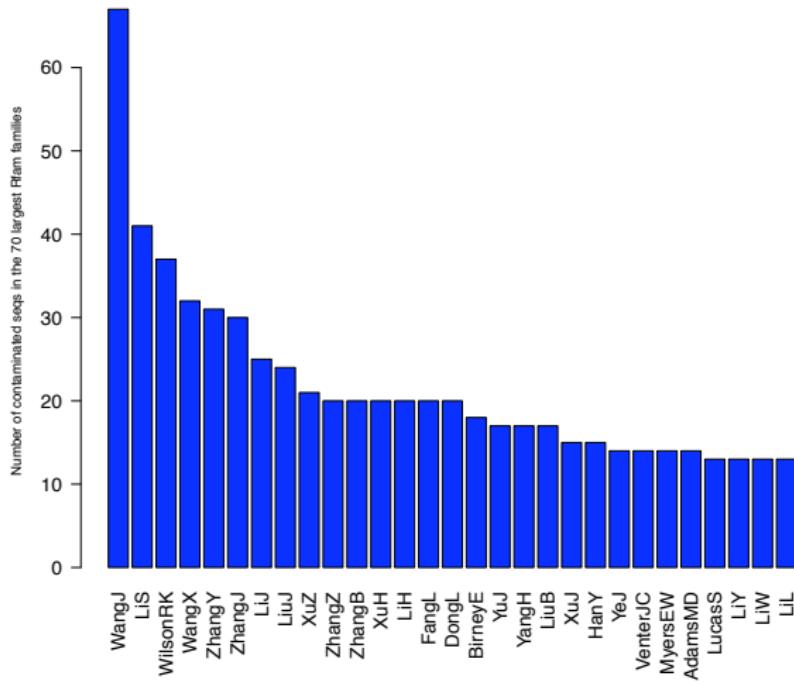
Gardner, Bateman & Poole (2010) SnoPatrol:
how many snoRNA genes are there? *Journal
of Biology*.

Oomycetes are
missing all the
snoRNPs?!

Embarrassing genome annotation screwups

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

The top contaminators of EMBL



Ab initio/De novo genome annotation

- ▶ Genes leave a statistical signal in the genome...
- ▶ Example: identify promoters, 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

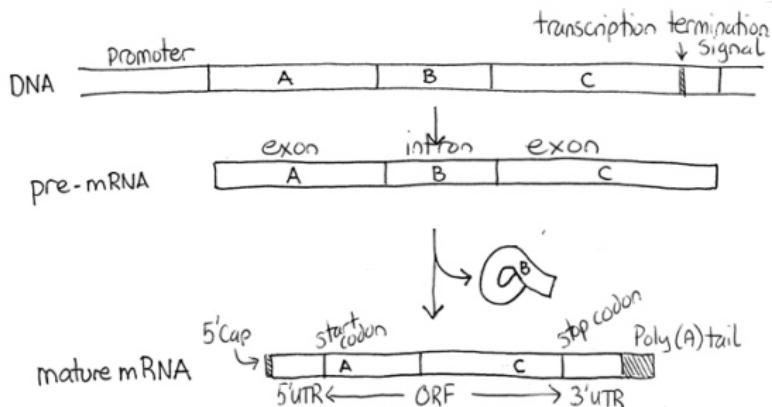


Figure from: <http://zerocool.is-a-geek.net/?p=630>

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



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 occurring by chance)
 - ▶ The log-odds score is $\log_2\left(\frac{\# \text{Observed}}{\# \text{Expected}}\right) = \log_2\left(\frac{p_{AB}}{f_A * f_B}\right)$
 - ▶ 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

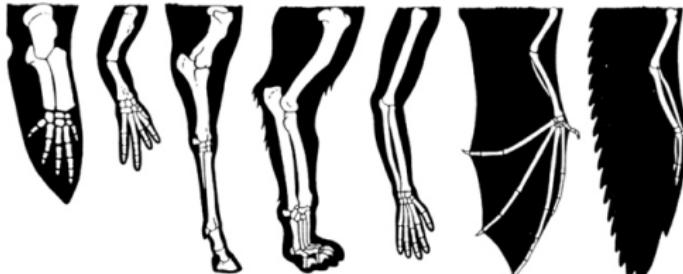
- ▶ Negative controls are at least as important in bioinformatics as they are in the wetlab.
- ▶ Remember, Bioinformaticians lie!

Prodigal practical

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

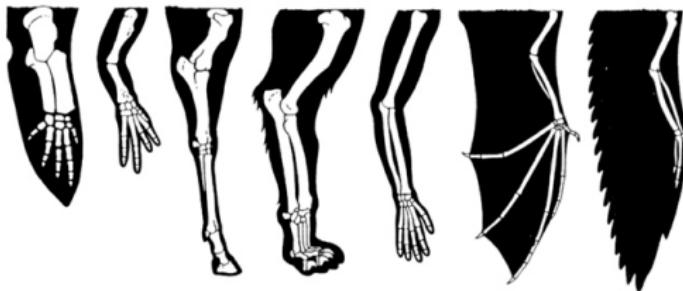
We can use homology...

- ▶ 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    GGAGCAGACGTCACTCACCCCCCCCAGGCCGGAGAT
opossum     GGAGCAGATGTTACTCACCCCTCCCTGCTGGAGAT
sloth        GGAGCAGACGTCACACACCCCTCCCCGGGGGGAT
armadillo   GGAGCAGACGTCACGCACCCCTCCGGCAAGGGGAT
tenrec       GGGGCCGACGTACGCACCCCCCCTGCGGGGGAT
elephant     GGAGCGGATGTCACACACCCGCCCTGCGGGGGAT
shrew        GGCAGCAGATGTCACGCATCCCTCCAGCAGGGGAC
hedgehog    GGAGCAGATGTCACACACCCCCCCCAGCAGGGAGAT
megabat      GGAGCAGATGTCACACACCCCTCCCTGCAAGGGAGAT
microbat    GGAGCAGATGTCACACACCCCCCCCCTGCAAGGGGAC
dog          GGAGCGGATGTCACACACCCCCCCCAGCAGGGGAC
cat          GGAGCCGATGTCACGCACCCCCCCCAGCAGGGGAT
horse        GGAGCGGATGTCACACACCCCTCCGGCAAGGGGAT
pika         GGAGCAGATGTCACTCACCCCTCCAGCTGGGGAT
rabbit       GGTGAGATGTCACACACCCCCCCCAGCTGGAGAT
squirrel    GGAGCAGATGTCACTCACCCCTCCAGCAGGGAGAT
guinea_pig  GGAGCAGATGTCACACACCCACCAGCAGGGAGAT
mouse        GGAGCAGATGTCACTCATCCACCTGCTGGGGAC
rat          GGAGCAGATGTCACTCATCCACCTGCTGGGGAT
kangaroo_rat GGAGCAGATGTTACACACCCCTCCAGCAGGGGAT
tree_shrew   GGCAGCAGACGTCACGCACCCCCCCCAGGGGGAT
human        GGAGCGGATGTCACACACCCCCCCCAGCAGGGGAT
tarsier      GGTGCTGATGTCACACACCCCCCCCCTGCAAGGGGAT
marmoset    GGAGCAGATGTCACACACCCACCAGCAGGGGAT
zebrafinch  GGAGCAGATGTCACTCACCCCTCCGGCAAGGGGAT
green_anole  GGGGCAGACGTCACTCACCCGCCAGCAGGGGAC
xenopus      GGAGCAGATGTTACACACCCACCTGCTGGTGAT
pufferfish   GGTGCGGATGTTACTCATCCCTCCCTGCTGGTGAT
fugu         GGGGCTGATGTTACTCACCCCTCCAGCTGGTGAT
stickleback  GGTGCAAGACGTCACACATCCCTCCAGCAGGGTGAT
medaka      GGTGCCGATGTCACTCATCCCTCCGGGGAGAC
zebrafish   GGGGCAGATGTTACACACCCGCCGGCTGGTGAT
lamprey     GGTGCCGATGTCACACACCCCTCCAGCAGGGGAGAC
//
```

Reminder of the degenerate genetic code

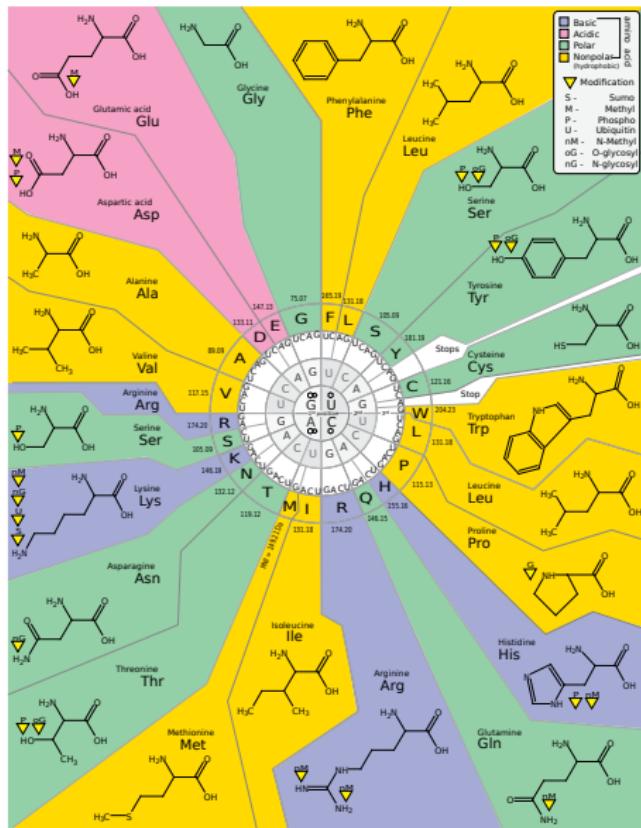
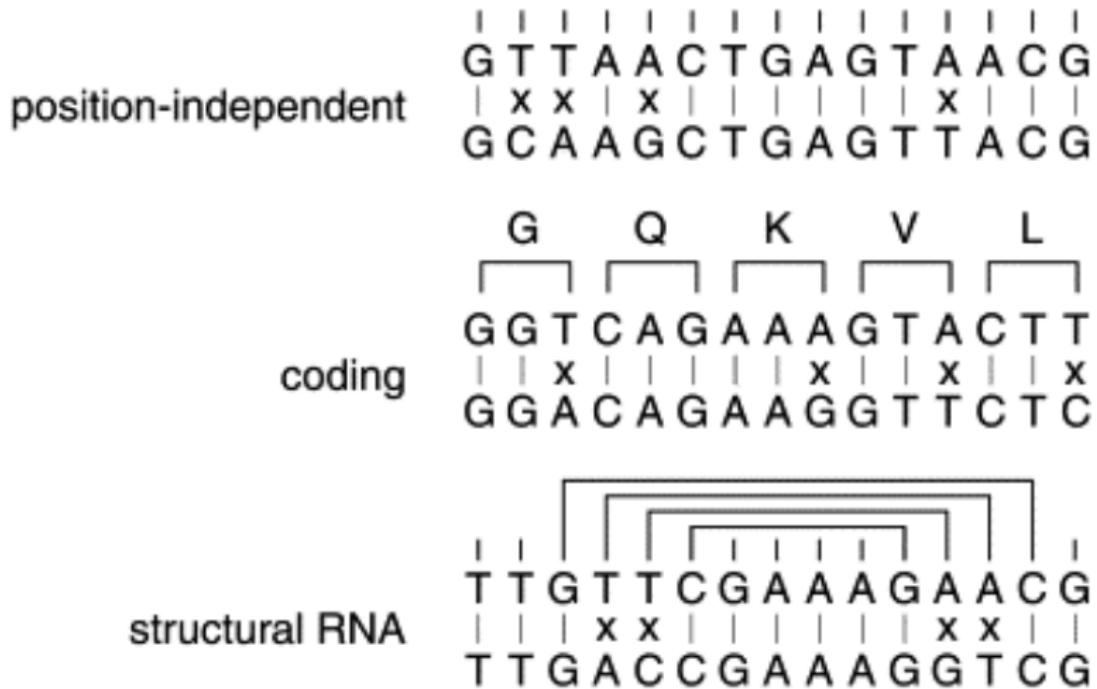


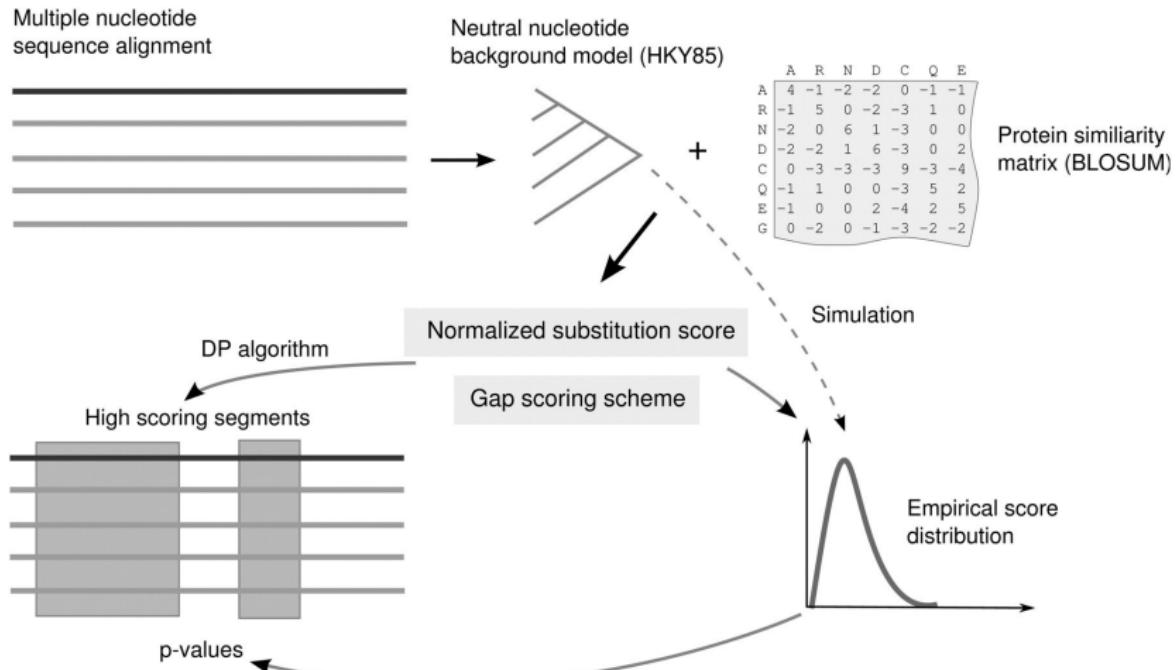
Image source: <http://upload.wikimedia.org/wikipedia/en/d/d6/GeneticCode21-version-2.svg>

The QRNA approach...



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

RNAcode (see also PhyloCSF)



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

Profile HMM Background

	A	G	C	U	U	C	G	G	A	A	C
yeast	G	C	U	U	C	G	G	A	G	A	C
fly	G	C	U	U	C	G	G	A	G	A	C
cow	G	C	A	U	U	C	G	U	C	G	C
mouse	G	C	U	U	U	C	G	A	U	G	C
human	G	C	U	U	C	G	C	U	G	C	C
chicken	G	U	A	A	U	C	G	U	A	A	C
snake	G	U	U	U	C	G	C	G	C	A	C
croc	G	U	U	U	C	G	A	G	A	G	C
	1	2	3	4	5	6	7	8	9	10	11

One HMM node per alignment column

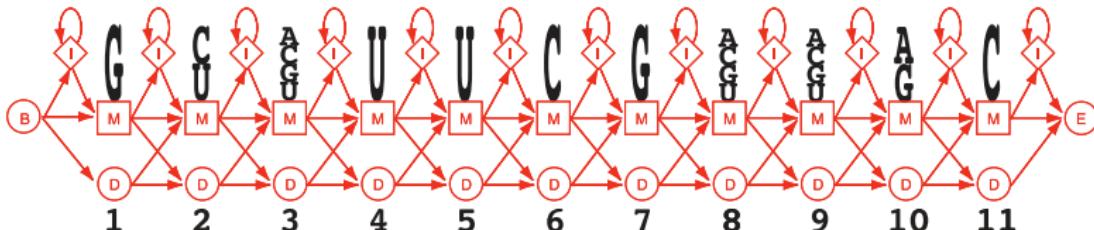
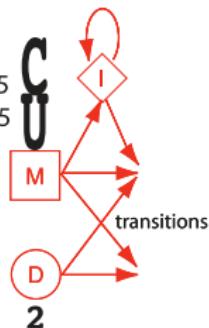
3 states per node:

- (M) Match: emits residues
- (I) Insert: inserts extra residues
- (D) Delete: deletes residues

HMMs generate homologous sequences.

Node for column 2:

$$\begin{aligned} P(C) &= 0.5 \\ P(U) &= 0.5 \end{aligned}$$



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

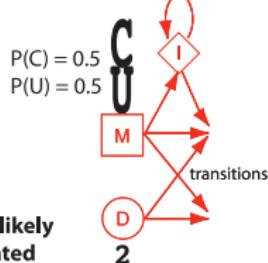
	
yeast	GU.CUUUCGGCAC
fly	GC.CUUUCGGAGC
cow	GC.AUUCGUCGC
mouse	GC.UUUUCGAUGC
human	GC.GUUCGCUGC
chicken	GU.AUUCGUAAC
snake	GU.GUUCGCGAC
croc	GU.UUUUCGAGAC
worm	GC.GUUCGCGGC
corn	GUGAUUCGU.GC

One HMM node per alignment column

3 states per node:

- (M) Match: emits residues
- (I) Insert: inserts extra residues
- (D) Delete: deletes residues

Node for column 2:



HMMs generate homologous sequences.

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

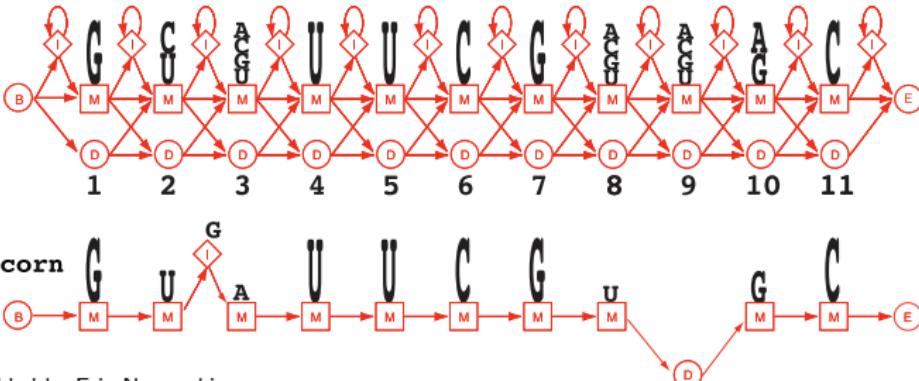
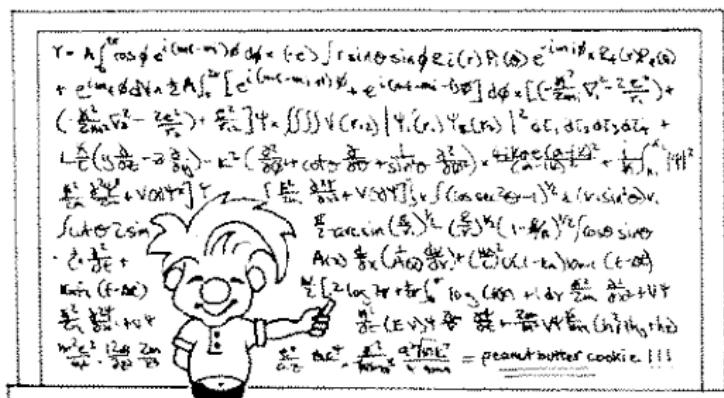


Image provided by Eric Nawrocki.

Profile HMM Background

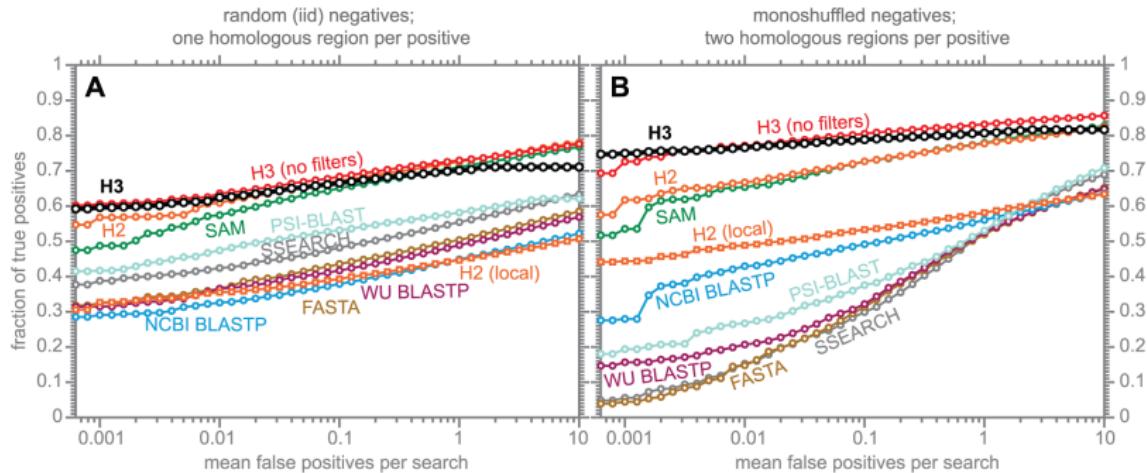
- ▶ A **tree-weighting scheme** takes care of unbalanced alignments
- ▶ **Dirichlet-mixture priors** are used to incorporate information about amino-acid biochemistry
- ▶ **Effective sequence number** is used to down-weight priors when many sequences are available
- ▶ Transition probabilities to Insert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



Why not just use BLAST?

▶ ACCURACY!

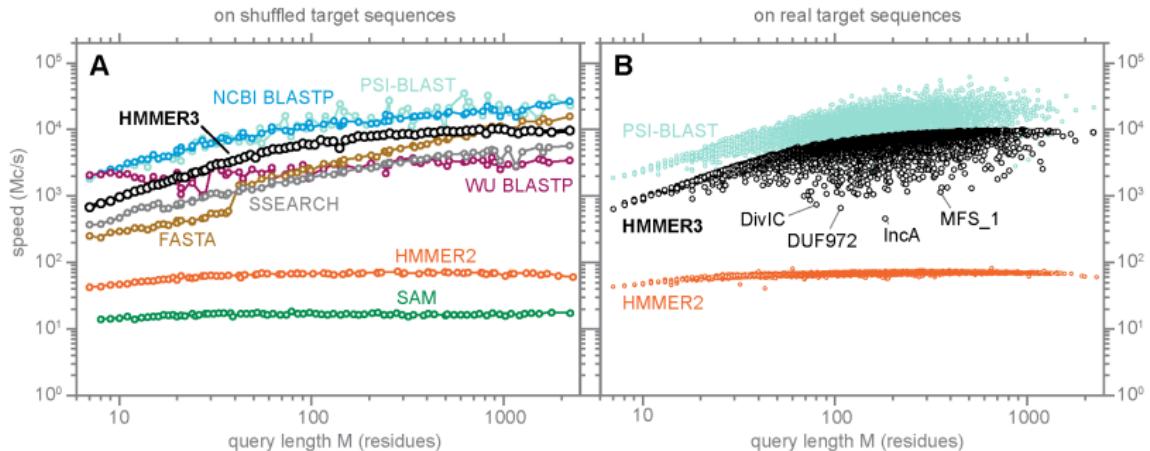
- ▶ Every benchmark of homology search tools has shown that profile methods are more accurate than single-sequence methods.



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 $\sim 3,000\times$ smaller for profiles
 - ▶ Save Planet Earth, use HMMER3



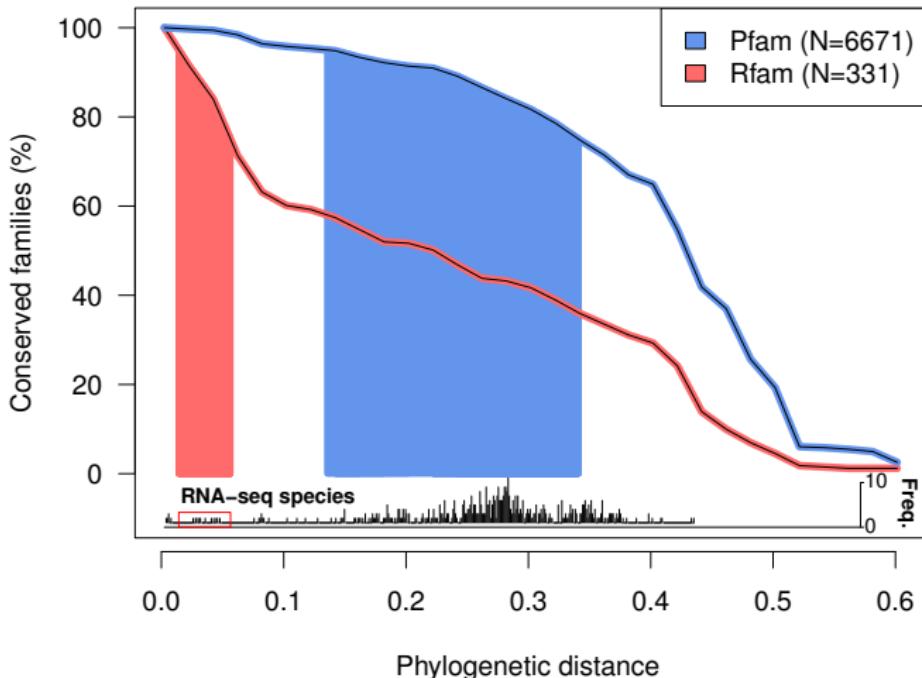
Eddy (2011) Accelerated Profile HMM Searches. *PLoS Computational Biology*.

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

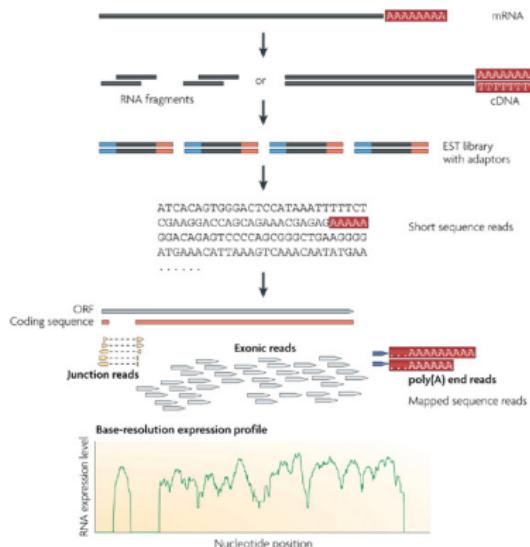
Conservation of Xfam families in bacterial genomes



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

We can use RNA detection methods...

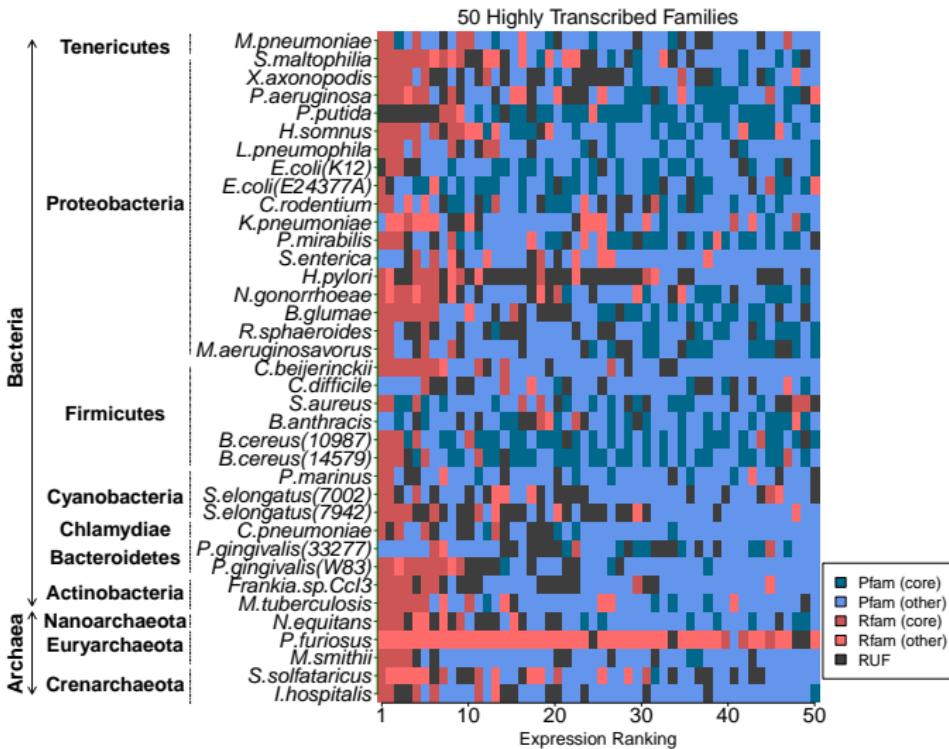
- ▶ Remember the central dogma of molecular biology
- ▶ Example: sequence RNAs from multiple tissues, developmental stages and environmental conditions
 - ▶ How reliable is this approach? What are the main weaknesses & strengths?



Nature Reviews | Genetics

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

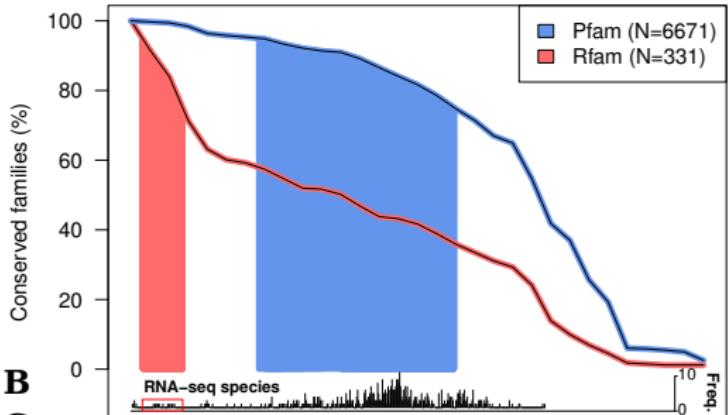
The top 50 most abundant genes



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

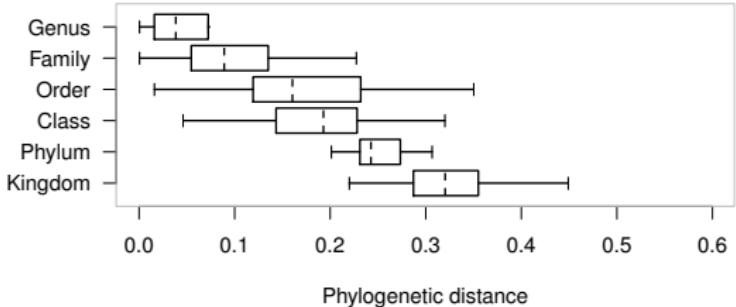
Is there a “Goldilocks Zone” for comparative ncRNA work?

A Conservation of RNAs & Proteins in bacterial genomes



B

C



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?

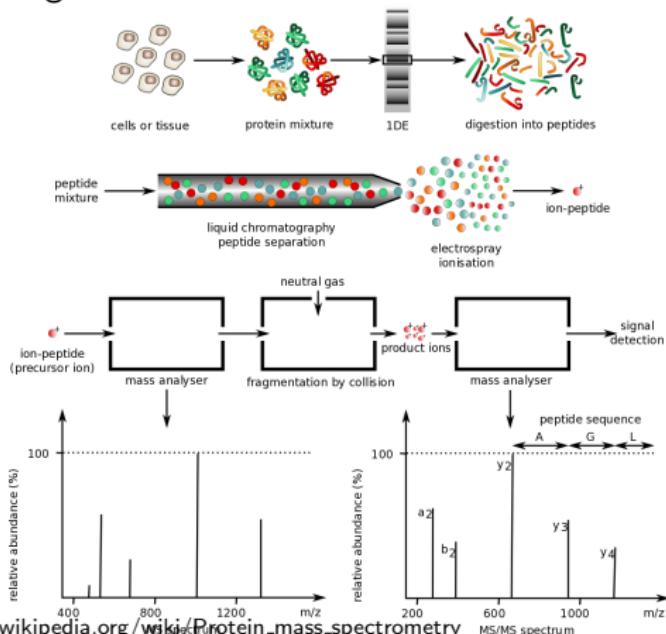


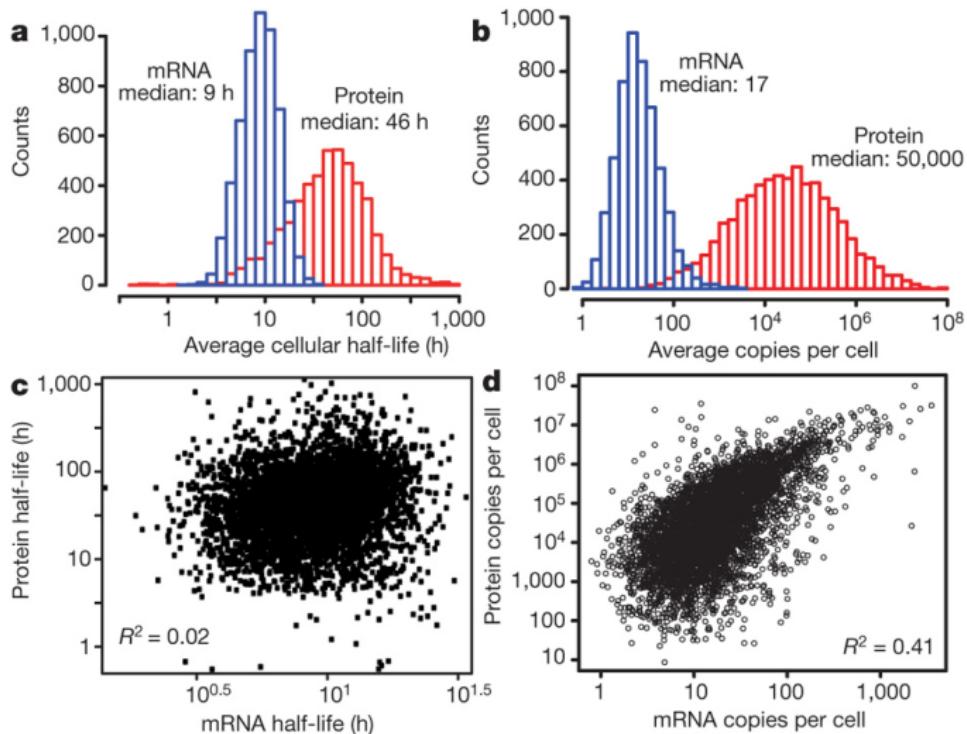
Figure from: http://en.wikipedia.org/wiki/Protein_mass_spectrometry

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

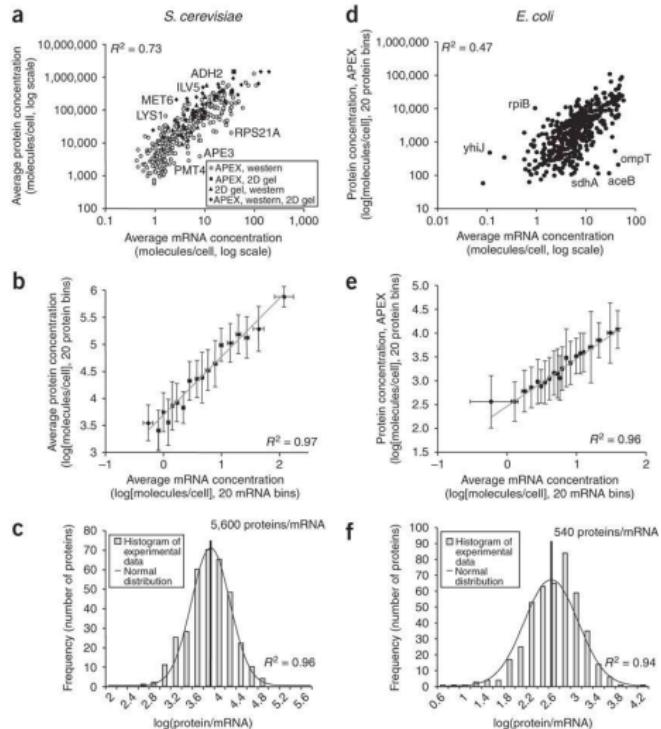
Ezkurdia *et al.* (2012) Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function. *Mol Biol Evol.*

How cool is this?!



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

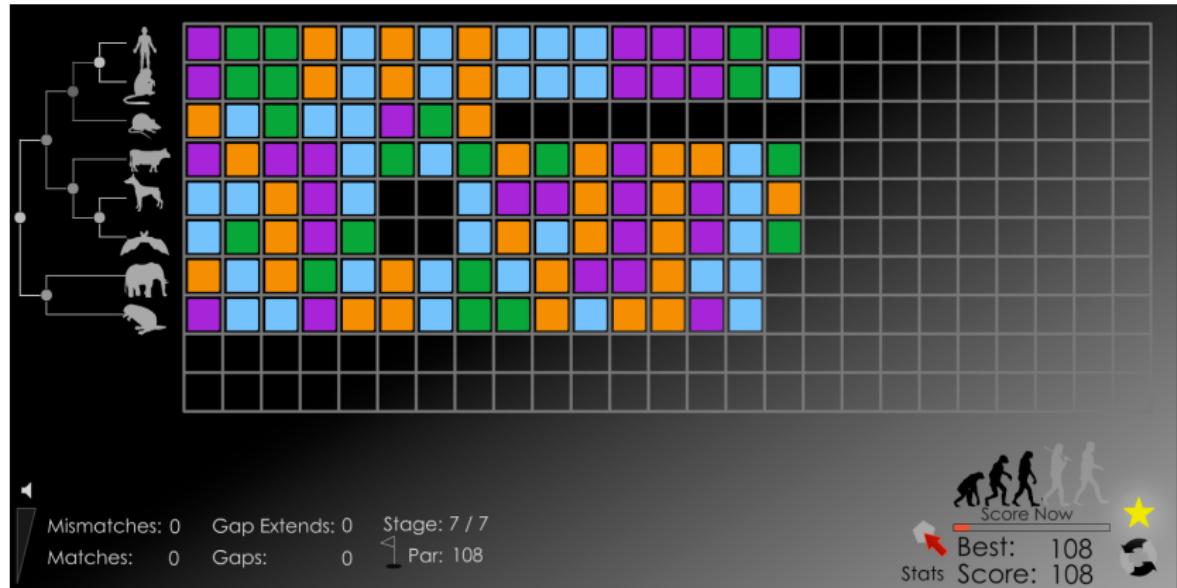
This is also kinda neat...



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

Homework: How to make a sequence alignment?

- ▶ Play: <http://phylo.cs.mcgill.ca>
- ▶ or even better, play SALI: <http://csb.cs.mcgill.ca/sali/>



The End

