Biological sequence analysis in the post-data era

Sean R. Eddy
HHMI Janelia Farm
Ashburn, Virginia, USA

1. an embarrassing personal history

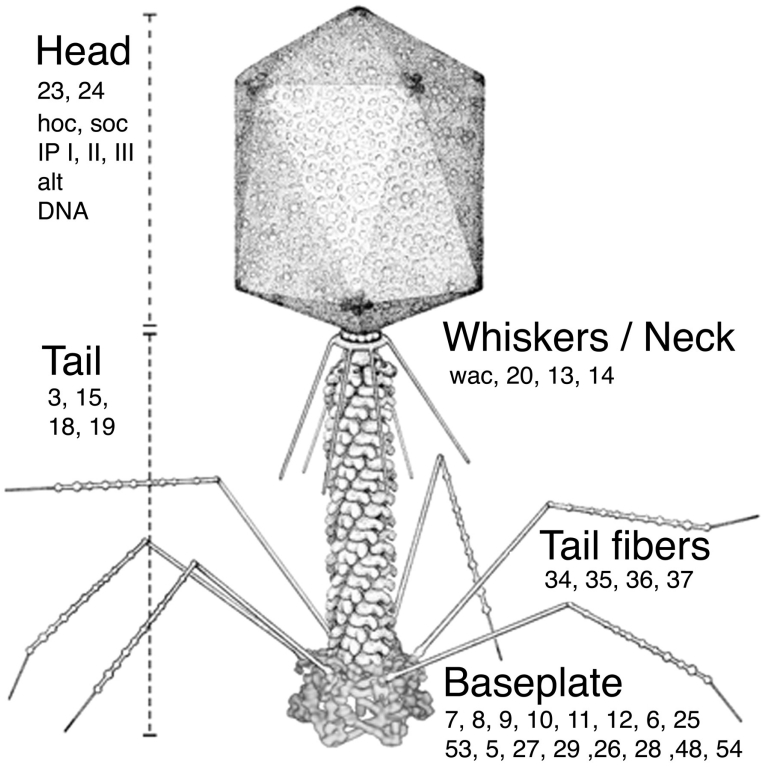
illustrating why I am unqualified to give advice or keynotes

introns in a bacteriophage!? seriously?

Cell, Vol. 47, 81-87, October 10, 1986, Copyright © 1986 by Cell Press

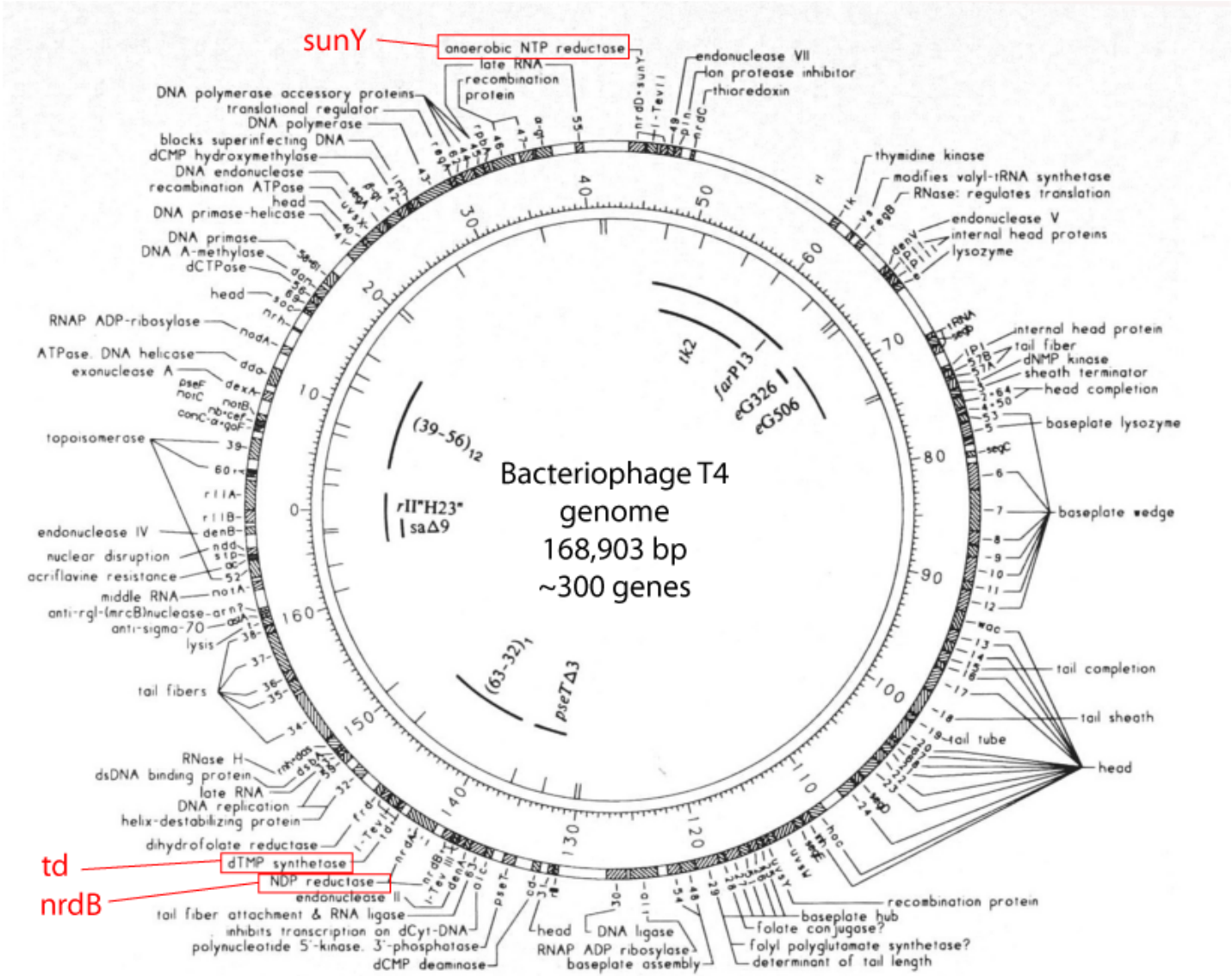
Multiple Self-Splicing Introns in Bacteriophage T4: Evidence from Autocatalytic GTP Labeling of RNA In Vitro

Jonatha M. Gott,* David A. Shub,* and Marlene Belfort †
*Department of Biological Sciences
State University of New York, Albany
Albany, New York 12222
†Wadsworth Center for Laboratories and Research
New York State Department of Health
Albany, New York 12201

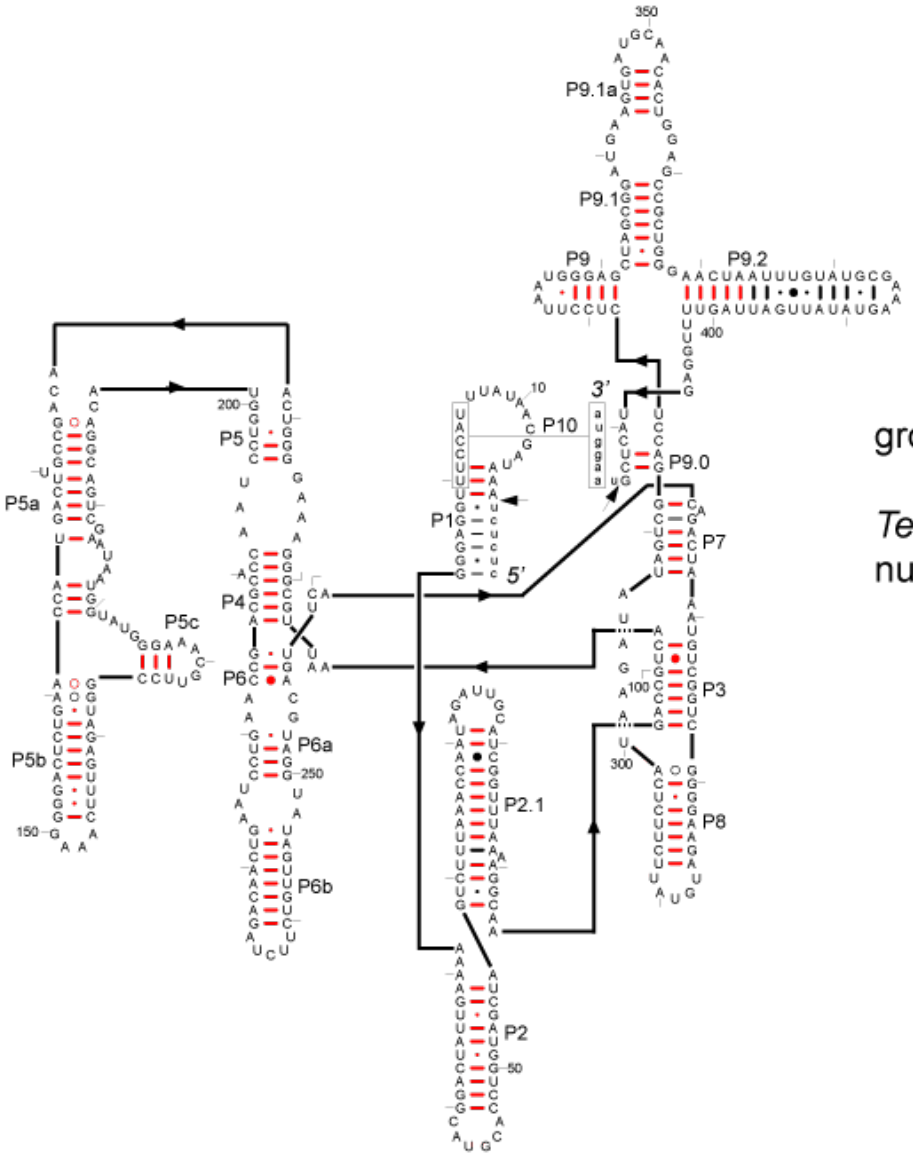


At least three... maybe more?

Miller et al, Microbiol Mol Biol Rev. 67:86, 2003



Group I self-splicing introns

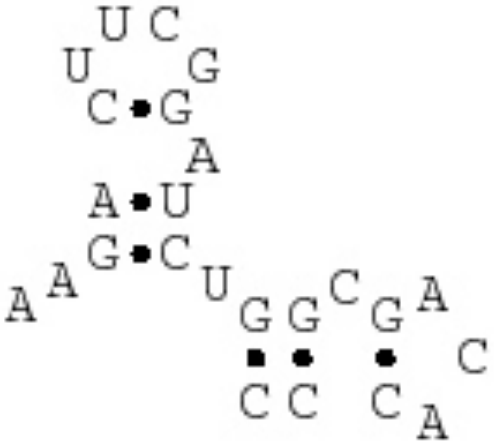


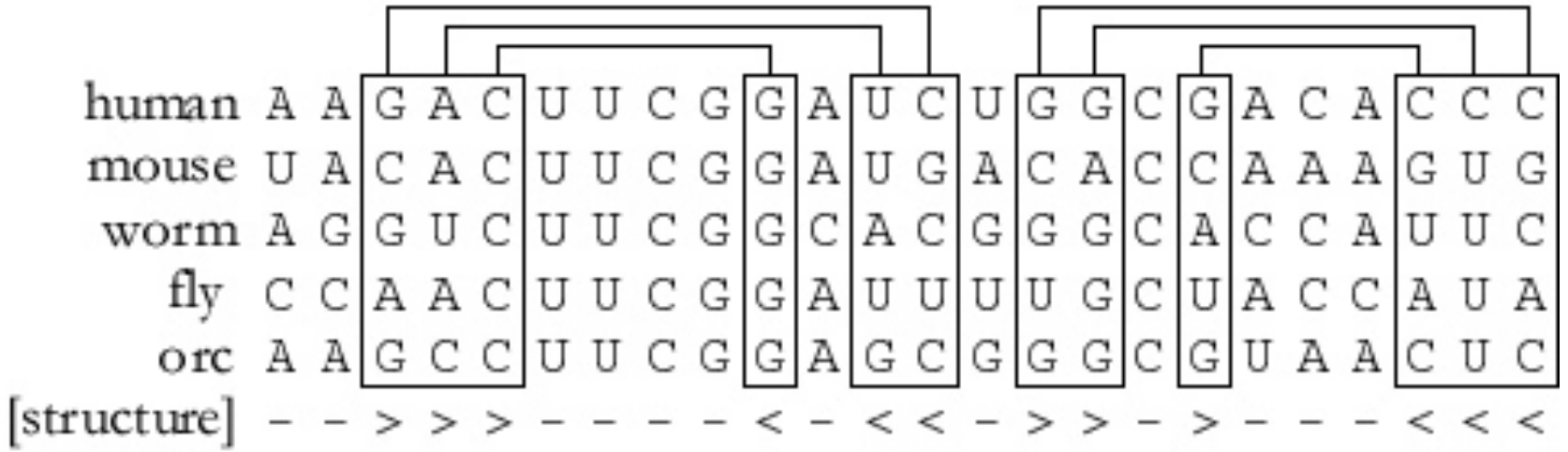
group I self-splicing intron

Tetrahymena thermophila
nuclear LSU rRNA

source: Robin Gutell
Comparative RNA Website
http://www.rna.icmb.utexas.edu/

RNA structure induces pairwise correlation





Pattern matching of RNA consensus

```
H1 s1 H2 s2 H3 s4 H2 s5 H4 s6 H4 s7 H5 s8 H1 s9 H6 s10 H6 s11 H5 s12 H7 s13 H7

H1 5:6 0
H2 3:5 0 GCN:NGC
H3 3:6 0
H4 3:5 0 ARN:NYU
H5 4:5 0 GACU:AGUC
H6 6:9 0
H7 5:8 0
s1 4:8
s2 3:6
s3 40:400
s4 2:6
s5 0:0
s6 30:400
s7 4:5 cann
s8 3:6
s9 0:1
s10 4:1000
s11 5:9 agn
s12 3:6
s13 5:400
```

RNAMOT

Daniel Gautheret, Francois Major, Robert Cedergren
CABIOS 6:325, 1990

RNABOB

a modification of Henry Spencer's regex code
SR Eddy, unpublished, 1991

Hidden Markov Models in Computational Biology: Applications to Protein Modeling UCSC-CRL-93-32

Anders Krogh*†, Michael Brown†, I. Saira Mian§,
Kimmen Sjölander†, David Haussler†

† Computer and Information Sciences

§ Sinsheimer Laboratories

University of California, Santa Cruz, CA 95064, USA.

email: krogh@nordig.ei.dth.dk, haussler@cse.ucsc.edu

August 17, 1993

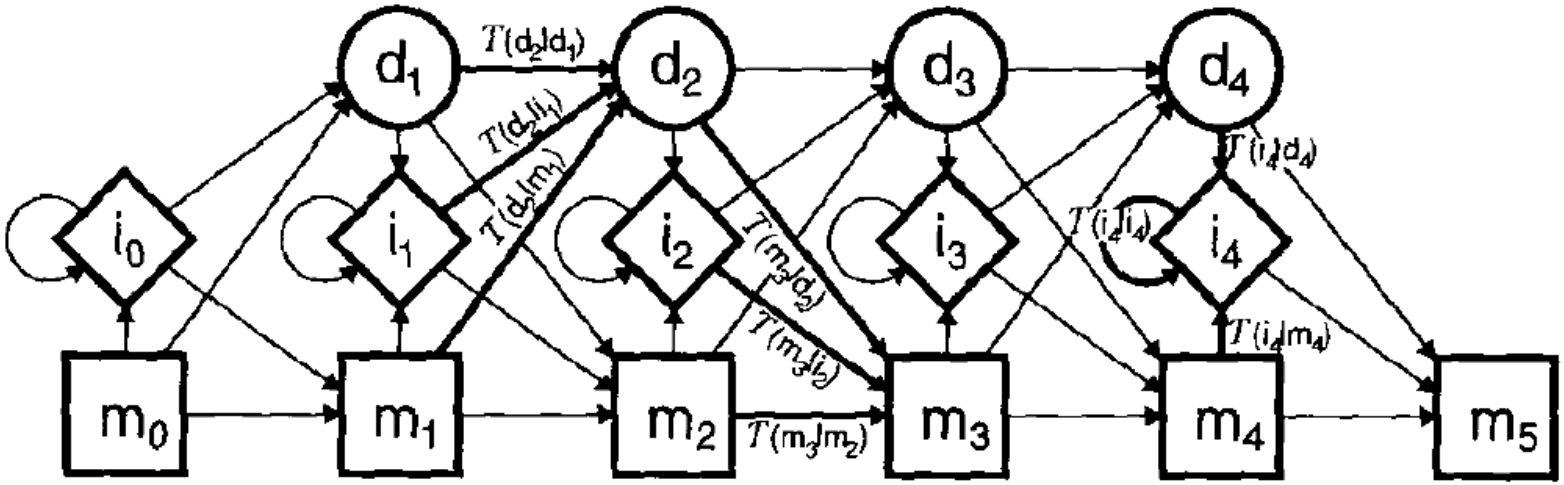
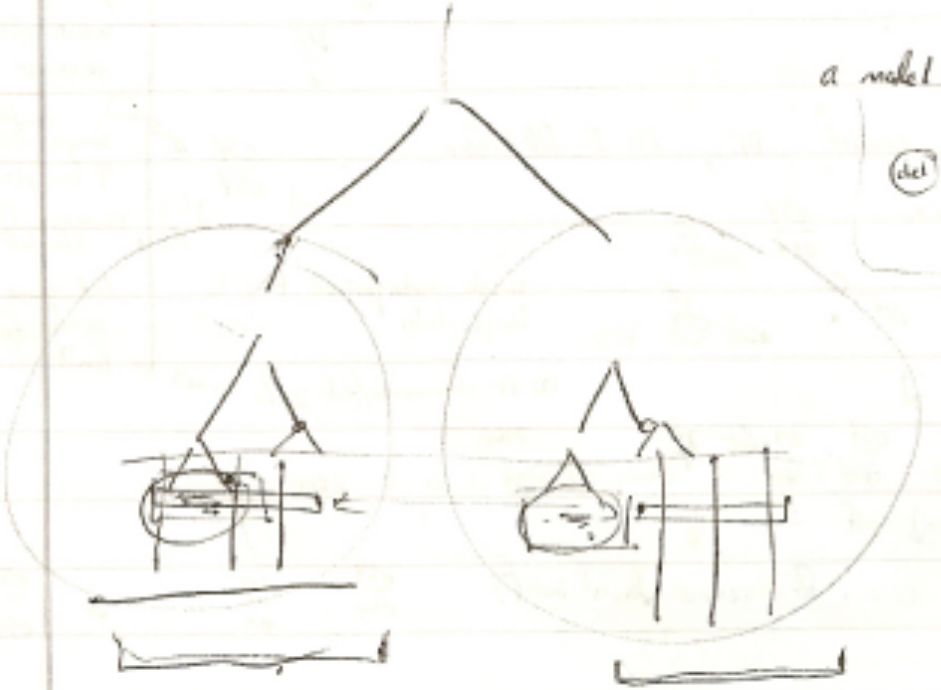


Figure 1. The model.

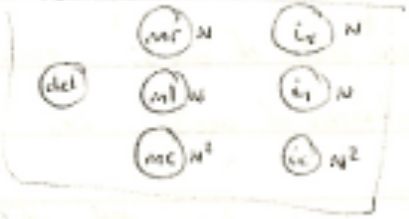
Attempt to repeat some analytic method that is considered unreliable and difficult until patience and hard work yield results similar to those published by the author. Pleasure derived from success, especially if it has come without the supervision of an instructor (that is, working alone), is a clear indication of aptitude for experimental work.

"Advice to a Young Investigator"
Santiago Ramon y Cajal
1916

... HMMER written as a reimplementation of Krogh, Haussler; 1993



a model is:



100 Ps/state

7 substates.

49 transition Ps.

6×8 = 48 emission Ps

tRNA val has 34 ss

21 pairs

∴ needs 55 states

= 5500 Ps.

16s rRNA 1600 nt 600

1000 paired = 500

1100 ~~states~~ states

experiment 1. construct a tRNA model by hand
use it to search sequences.

needs 7 × 55 = 385 states

× ~~70~~

100

× 100

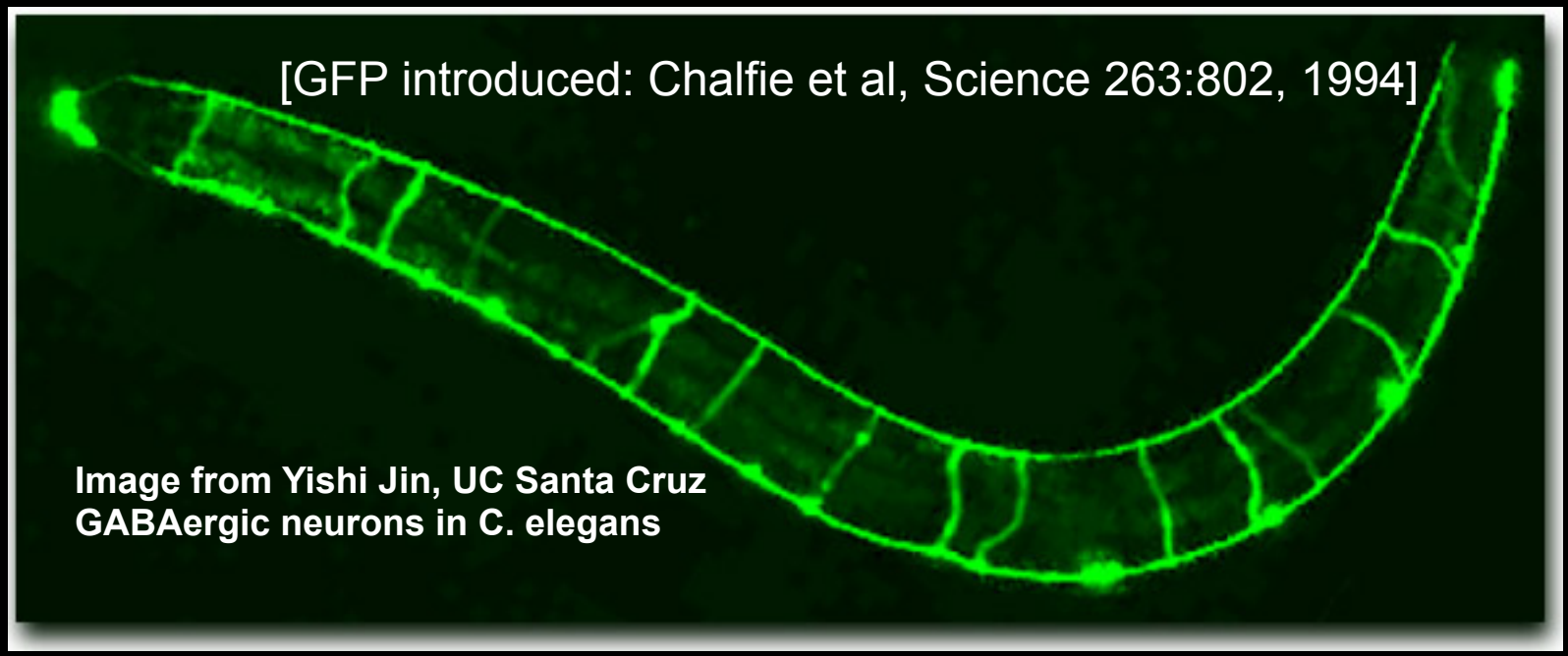
3.8 Mb N^3 in

“covariance HMMs”: April 1993

I had proposed to develop reporter fusions to neural-specific promoters as a tool to visualize axonal processes in live animals, facilitating genetic screens...

I have started to play with... green fluorescent protein (GFP) from the jellyfish...

I found out at the worm meeting in June that Marty Chalfie's lab is onto the same idea. Chalfie has already obtained bright fluorescence in the axons of the six touch neurons...



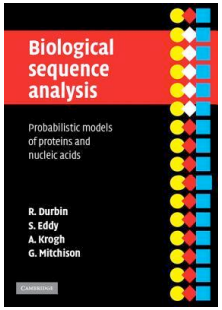
Somewhat embarrassingly, my most productive work has been unrelated to the original proposal, resulting from some moonlighting as a computational biologist... and a lingering interest in RNA structure from my thesis work.... I invented a new kind of statistical model, related to HMMs, which can model the two-dimensional structure consensus of RNAs. **- progress report for my postdoc grant, 1993**

HMMER: profile HMMs of proteins and DNA structure

hmmer.org

Infernal: profile SCFGs of RNA sequence/structure

infernal.janelia.org



Durbin et al., *Biological Sequence Analysis,* 1998

but: 1) I'm a biologist, not a mathematician or computer scientist
2) My biology project was totally, utterly scooped
3) I got into comp bio as a side hobby
4) Tools and methods, not fundamental biology
4) All I really did was learn from other people

2. I will now argue that these are *ideal* conditions.

"My center is giving way, my right is in retreat.
Situation excellent. I shall attack."

- Ferdinand Foch
Battle of the Marne, 1914

100 years of incremental engineering

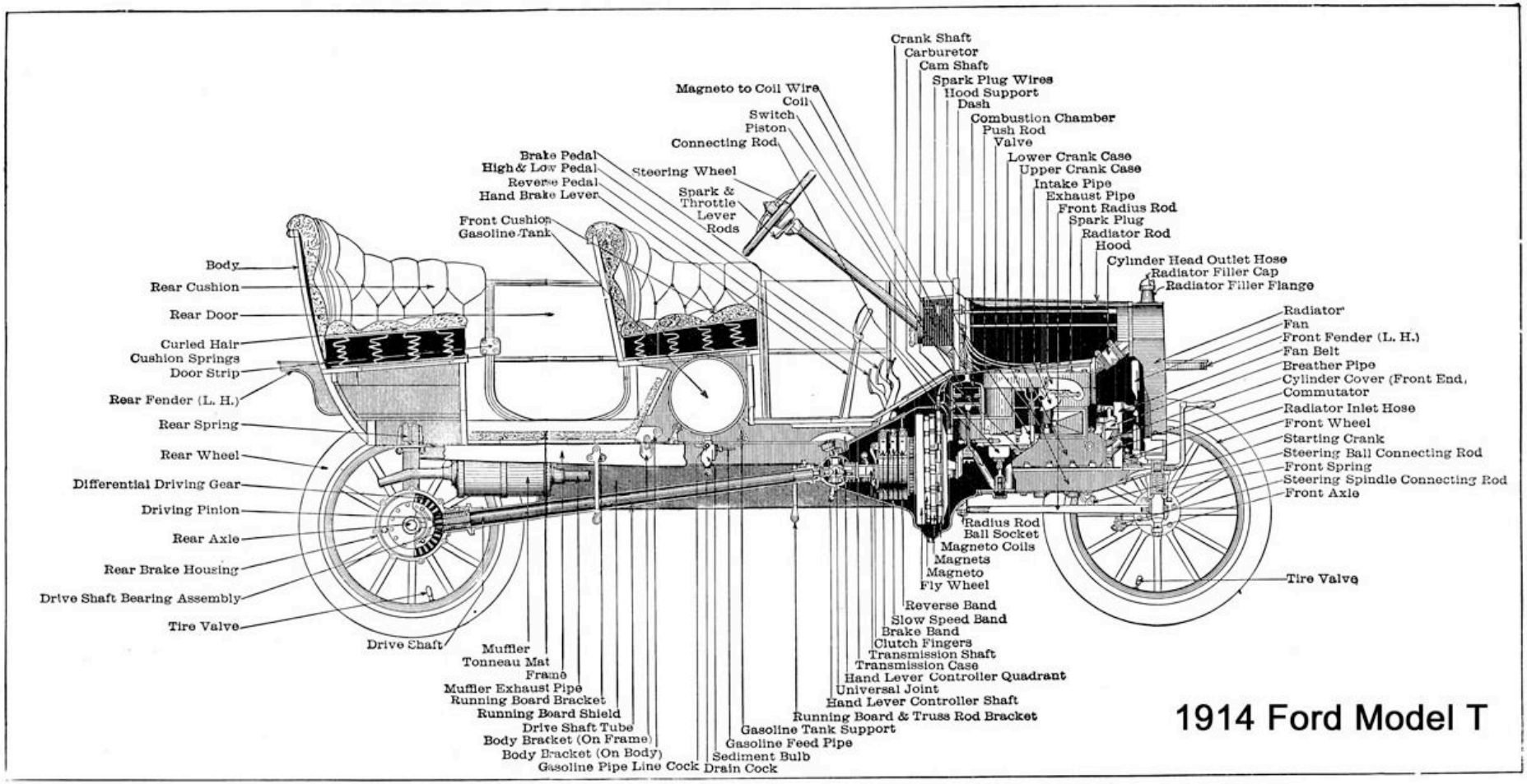
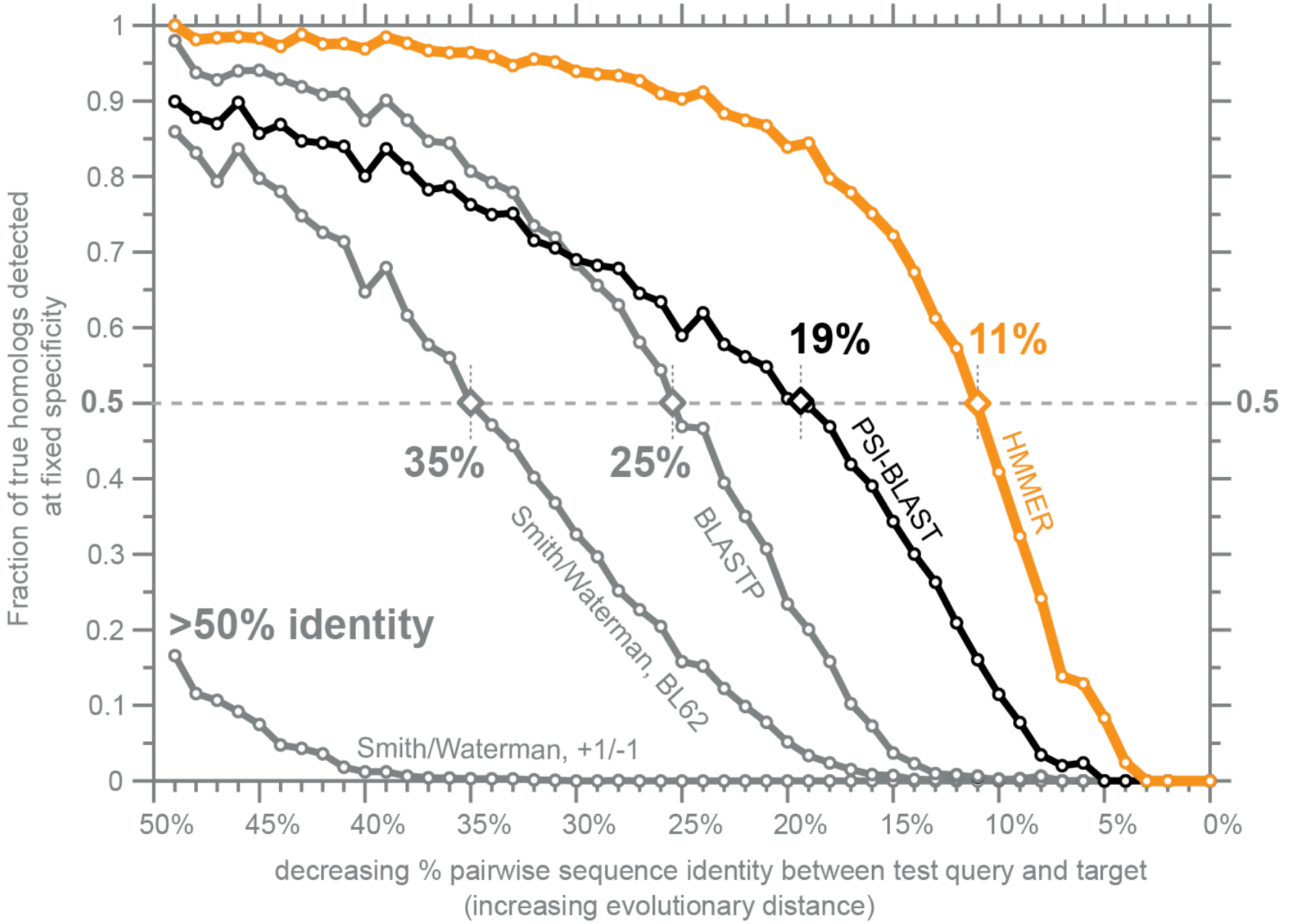


Fig. 468.—Side Sectional View of the Ford Model T Motor Car, the Most Widely Used Automobile in the World.

30 years of incremental engineering



Science vs. engineering in computational biology

Science: an idea, a new algorithm to solve a problem: a paper.
Engineering: a useful tool that many people use: software.

Besides solving a problem, a "useful tool" also:

- Works on real (messy) data, not just exemplars
- Is usable by biologists
  (documented, robust, easy to obtain and learn)
- Outlasts the student or postdoc who developed it
  (perpetually maintained)
- Becomes a building block for even better things
  (simple interface for a complex function)
- Is easier to use than to reinvent.
  (solves the problem thoroughly)

These cost time and money.
They don't result in papers.

The fallacy:
You are rewarded for how many publications you have.

The truth:
You are rewarded for how much impact you have.

The (underappreciated?) consequence:
It *is* worth the time and effort to engineer useful tools.

Arbitrage *(n.)*:
the practice of taking advantage of a value difference between two markets

Academics believe that their career system only values publications.

The biology research community demands strong computational tools.

My comp bio work consciously exploits the middle ground.
To a significant extent, I take other people's ideas and papers,
and engineer the community's advances into robust computational tools.

HMMER accelerated 1000x; Infernal accelerated 10,000x

BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 2 2007, pages 156–161
doi:10.1093/bioinformatics/btl582

Sequence analysis

Striped Smith–Waterman speeds database searches six times over other SIMD implementations

Michael Farrar

Received on June 22, 2006; revised on November 13, 2006; accepted on November 14, 2006
Advance Access publication November 16, 2006
Associate Editor: Nikolaus Rajewsky

SIMD = single instruction multiple data: vector-parallel operations
a market driven in large part by graphics and games

Intel/AMD: SSE (Streaming SIMD Extensions)
AVX (Advanced Vector Extensions)

M Farrar, *Bioinformatics* 23:156 (2007)
T Rognes, *BMC Bioinformatics* 12:221 (2011)
SR Eddy, *PLoS Comp Bio* 7:e1002195 (2011)

Planning for permanence by taming exponential growth

Pfam 27: 14,831 protein families

Wikipedia

Seed alignment
small, representative;
manually curated

Profile
statistical model

Full alignment
comprehensive;
automatically generated

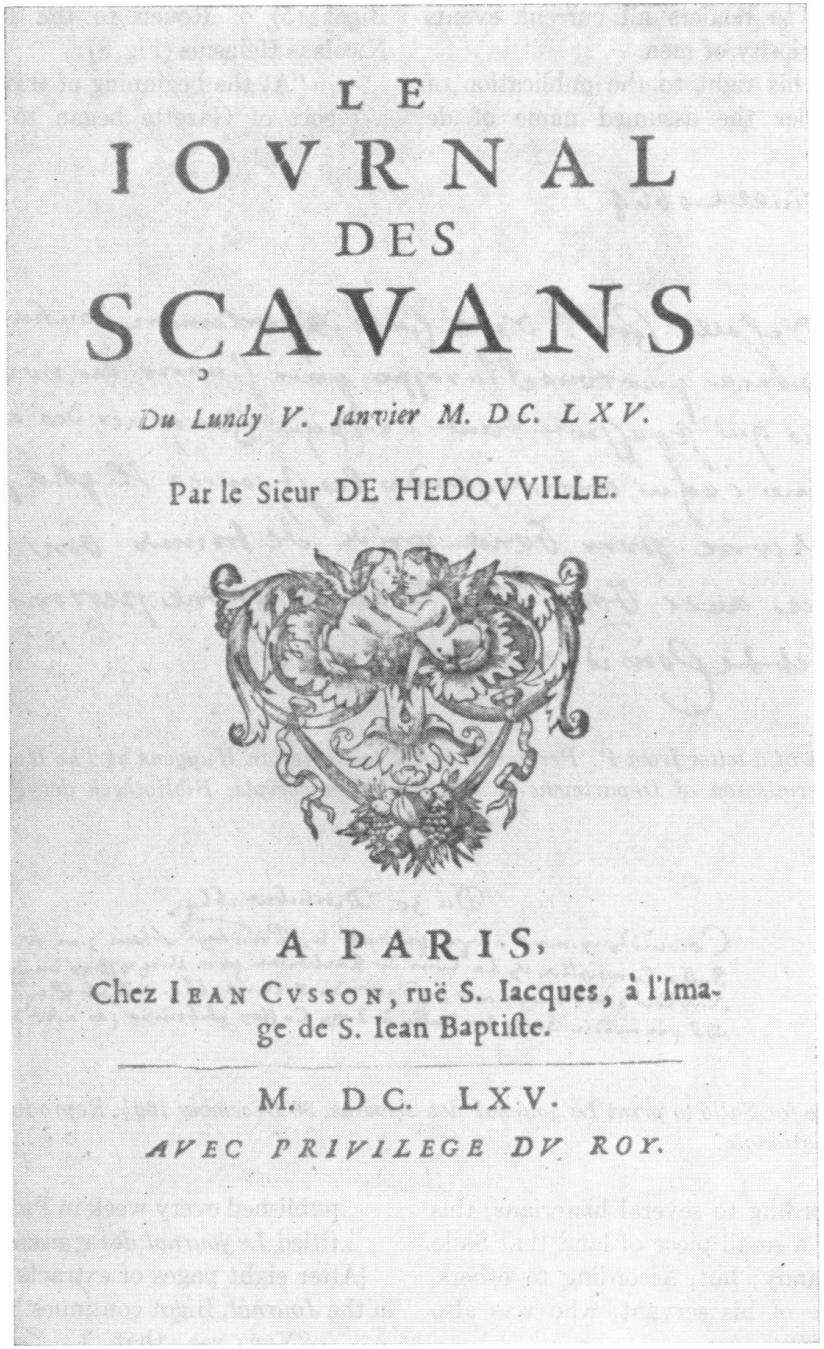
We leverage community annotation

Seed alignments are stable, permanent

We invest in scalable methods (HMMER, Infernal)

Thus the rest of Pfam is automatic

3. Scientific publication: a 350 year old tradition of open science.



LE
IOVRNAL
DES
SCAVANS

Du Lundy V. Ianvier M. DC. LXV.

Par le Sieur DE HEDOVVILLE.

A PARIS,
Chez IEAN CVSSON, ruë S. Iacques, à l'Image de S. Iean Baptiſte.

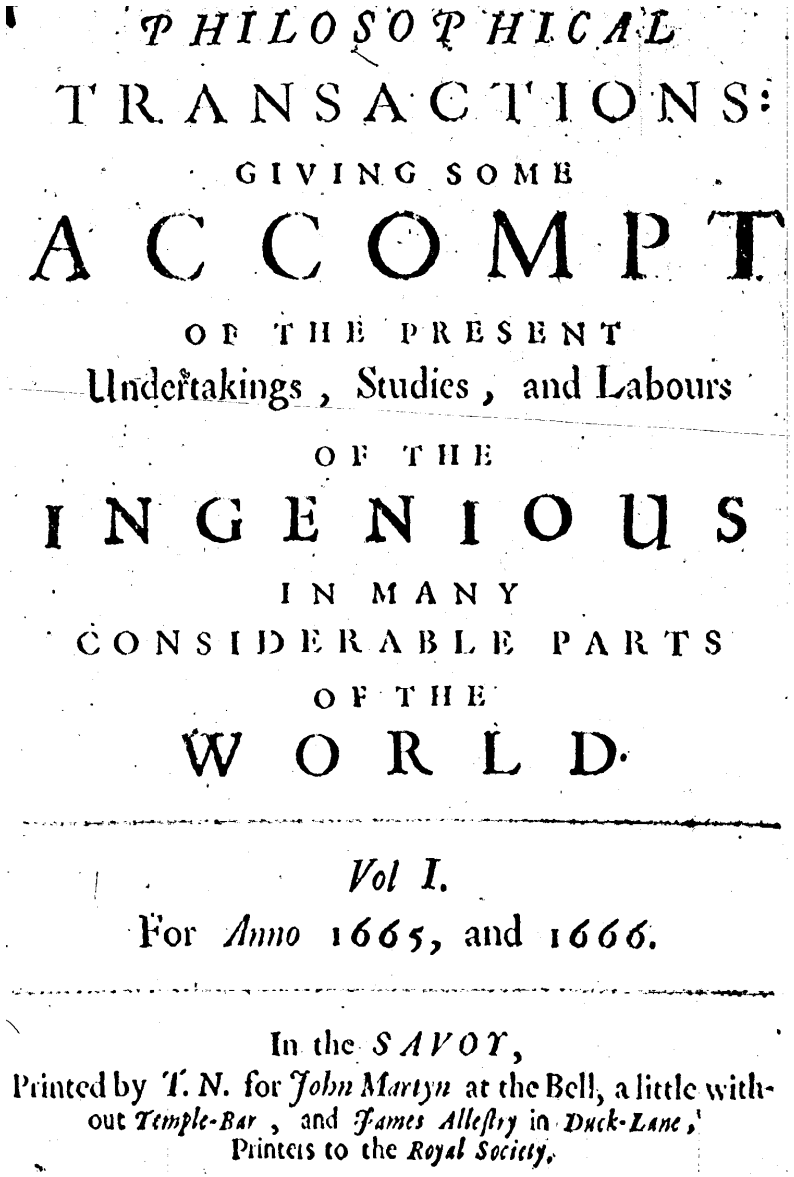
M. DC. LXV.
AVEC PRIVILEGE DV ROY.

The first scientific journal:

Journal des sçavans
(1665)

'for the relief of those either too indolent or too occupied to read whole books'

-Denis de Sallo, founder
quoted in Bernard Houghton, Scientific Periodicals (1975)



PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For Anno 1665, and 1666.

In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane, Printers to the Royal Society.

The oldest extant journal:

Philosophical Transactions of the Royal Society of London
(London, May 1665)

Henry Oldenburg, editor

peer review instituted

famous exchanges with the secretive Isaac Newton

scientific priority (and fame)
in return for publication and disclosure: a *quid pro quo*

The Cech report (2003)

"... the fundamental purpose of publication of scientific information is to move science forward. More specifically, the act of publishing is a *quid pro quo* in which authors receive credit and acknowledgement in exchange for disclosure of their scientific findings.

An author's obligation is not only to release data and materials to enable others to verify or replicate published findings (as journals already implicitly or explicitly require) but also **to provide them in a form on which other scientists can build with further research**.

All members of the scientific community – whether working in academia, government, or commercial enterprise – share responsibility for upholding community standards as equal participants in the publication system, and all should be equally able to derive benefits from it."

"Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences"
National Research Council, National Academy of Sciences, 2003

“Open data” has been the scientific community standard since journals began in 1665.

Then why is everyone so exercised about “Open” everything these days?

The problems today are
really more about new *mechanism* than new *principle.*

What happens when it takes more than reading the paper to convey the key result?

- Large data sets — e.g., the Celera human genome
- Software — papers like “FOO: a program to do X”
- Materials — e.g., transgenic mice
- The literature itself — full text indexing/retrieval

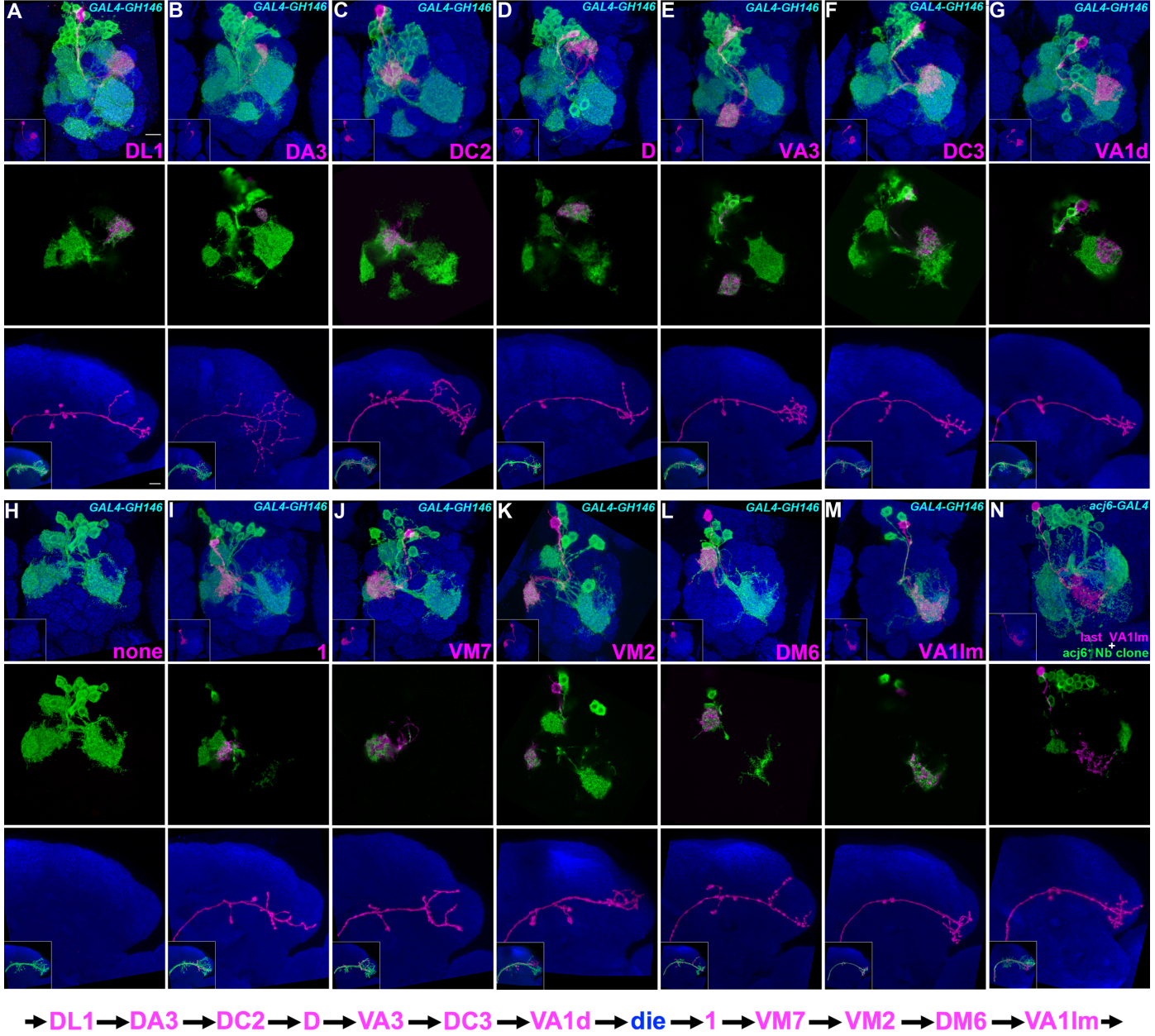
"An article about computational science in a scientific publication is not the scholarship itself, **it is merely advertising of the scholarship**. The actual scholarship is the complete software development environment and that complete set of instructions that generated the figures."

Buckheit and Donoho (1995)
cited in Robert Gentleman, Stat App Genet Mol Biol 4:2 (2005)
emphasis mine

4. Two magic tricks

Any sufficiently primitive technology is indistinguishable from magic.

- little-known bioinformatics corollary to Clarke's Third Law



From one neuroblast:
a precise lineage
of ~40 neuron types

An intrinsic program?

Seriously?

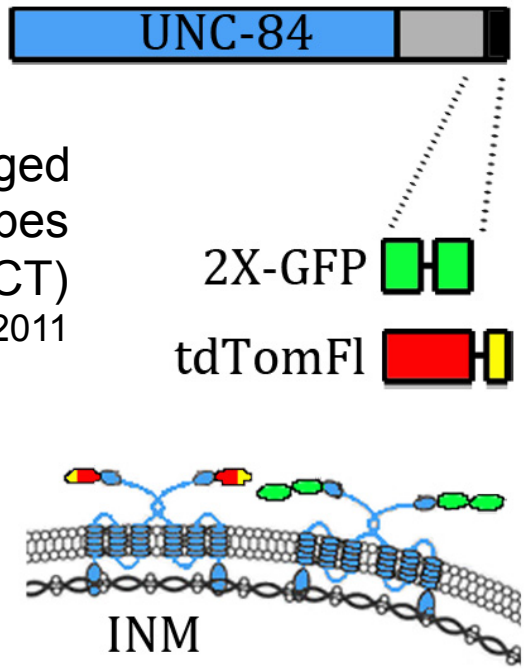
→DL1→DA3→DC2→D→VA3→DC3→VA1d→die→1→VM7→VM2→DM6→VA1lm→

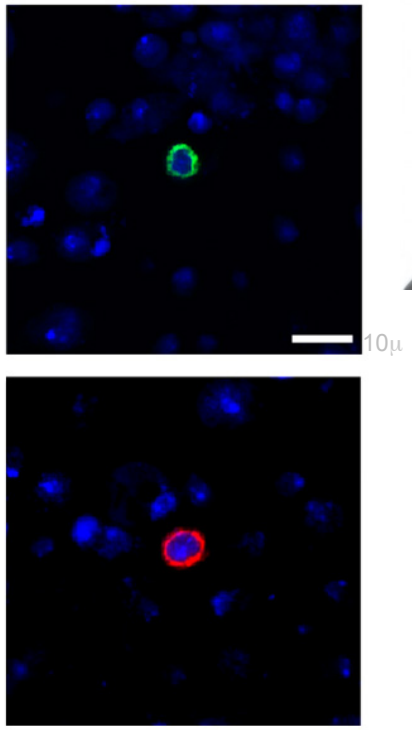
Hung-Hsiang Yu et al (Tzumin Lee lab)
A complete developmental sequence of a Drosophila neuronal lineage...
PLOS Biology (2010)

Genomic analysis of individual neuronal cell types

Isolation of Nuclei TAgged in specific Cell Types (INTACT)

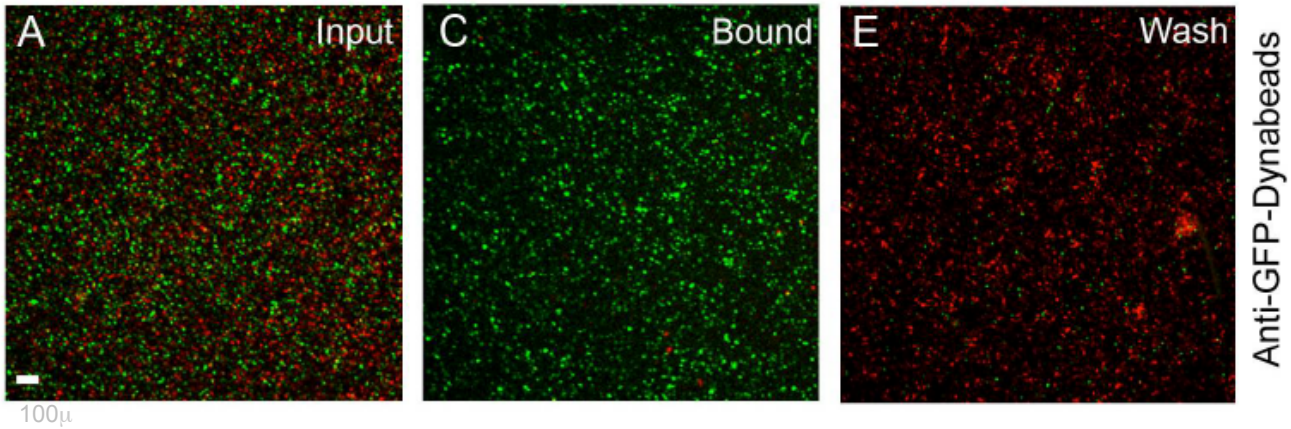
Deal and Henikoff, 2010, 2011





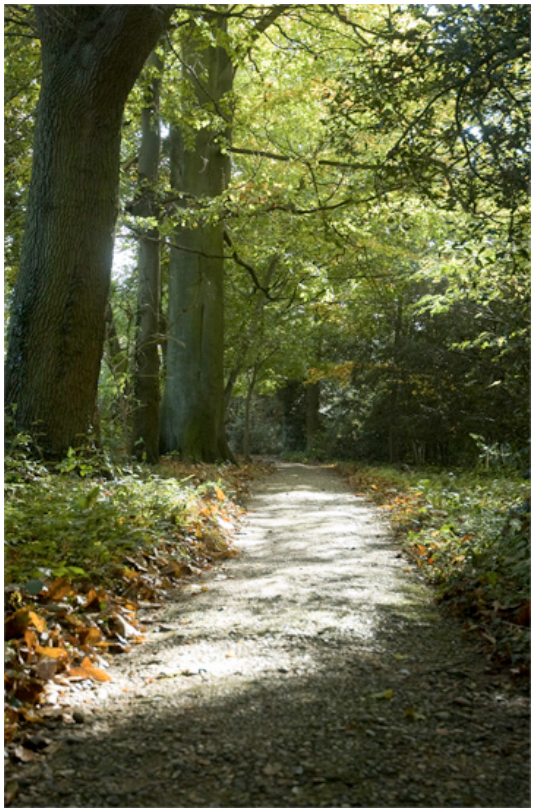
Lee Henry

Fred Davis



Jeremy Nathans, Johns Hopkins

vs.



Darwin's Sandwalk at Down House

Yes, of course, we *do* need sophisticated infrastructure:

- software engineering
- databases
- ontologies...

But more often we're exploring.

For one-off data analysis, premium is on expert biology and tools as simple as possible.

I like control experiments. I don't trust statistical tests.

A statistical test: *See the error?*

The complete sequence analysis of the chromosome reveals 182 ORFs of ≥100 amino acids. ORFs of this length have <0.2% probability of occurring by chance in *S. cerevisiae* DNA[32].

The complete DNA sequence of yeast chromosome III
Oliver et al, Nature 357:42 (1992)

A negative control experiment, using simulation:

```
% esl-shuffle yeastgenome | \
  esl-translate -m -l 100 - | grep "^>" | wc -l

383
```

[1] [2]

```
% esl-shuffle yeastgenome | \
  esl-translate -m -l 100 - | grep "^>" | wc -l
```

[3]

[1] We build a repertoire of trusted methods, like protocols.

[2] Data analysis presumes data availability.

[3] The command line is a powerful data analysis environment.

I look at “big data” by taking small subsamples.

(both random, and outliers)

% cat SupplementaryTable1.txt | randline.pl | head -10

given a bazillion data lines in Supplementary Table1...
randomly sample 10 rows
(...presumes Table S1 in tabular electronic form.)

if 10/10 look good, dataset is in good shape
if 9/10 are artifacts, dataset is in bad shape

If an artifact, design new experiments to fix. Repeat.

a "microexon?"

```
genome: CCCCCTGGTG|agttag----agagcg|CGGG|gtgcca--ttaaggtgcggc|CACT
RNA:    CCCCCTGGTG|                |CGGG|                     |CACT
```

```
genome: CCCCCTG|gtgagttag----agagcgcggggtgcca----ttaag|GTGCGGCCACT
RNA:    CCCCCTG|                                     |GTGCGGGCACT
                                                              x
```

no. an unexpected artifact.
(polymorphisms in strain RNA was from)

"Bioinformatics: Gone in 2012"

Lincoln Stein (2003)

Will bioinformatics be like biostatistics, or like molecular cloning?

Should data analysis be outsourced?

Or should we analyze our own data?

Biologists will be fluent in computational analysis.
This is not "interdisciplinary".
Rather, the discipline of biology will adapt.

Meanwhile, strong computational tools will always be in demand.

The Eddy/Rivas laboratory

Fred Davis
Pat Dennis
Lee Henry
Tom Jones
Seolkyoung Jung
Eric Nawrocki
Elena Rivas
Travis Wheeler

HMMER Web Services

Jody Clements
Rob Finn
Bill Arndt
Ben Miller

The Xfam Consortium

Alex Bateman (EBI Cambridge, UK)
Sarah Burge (Wellcome Trust Sanger Institute, UK)
Paul Gardner (U. Canterbury, New Zealand)
Sam Griffiths-Jones (Manchester University, UK)
Jerzy Jurka (Genetic Information Research Inst.)
Marco Punta (Wellcome Trust Sanger Institute, UK)
Arian Smit (Institute for Systems Biology)
Erik Sonnhammer (Stockholm University, Sweden)
et al.

hmmer.janelia.org

infernal.janelia.org