(Imagine a hip music video playing HERE, at low volume)

Future Thoughts

C. Titus Brown

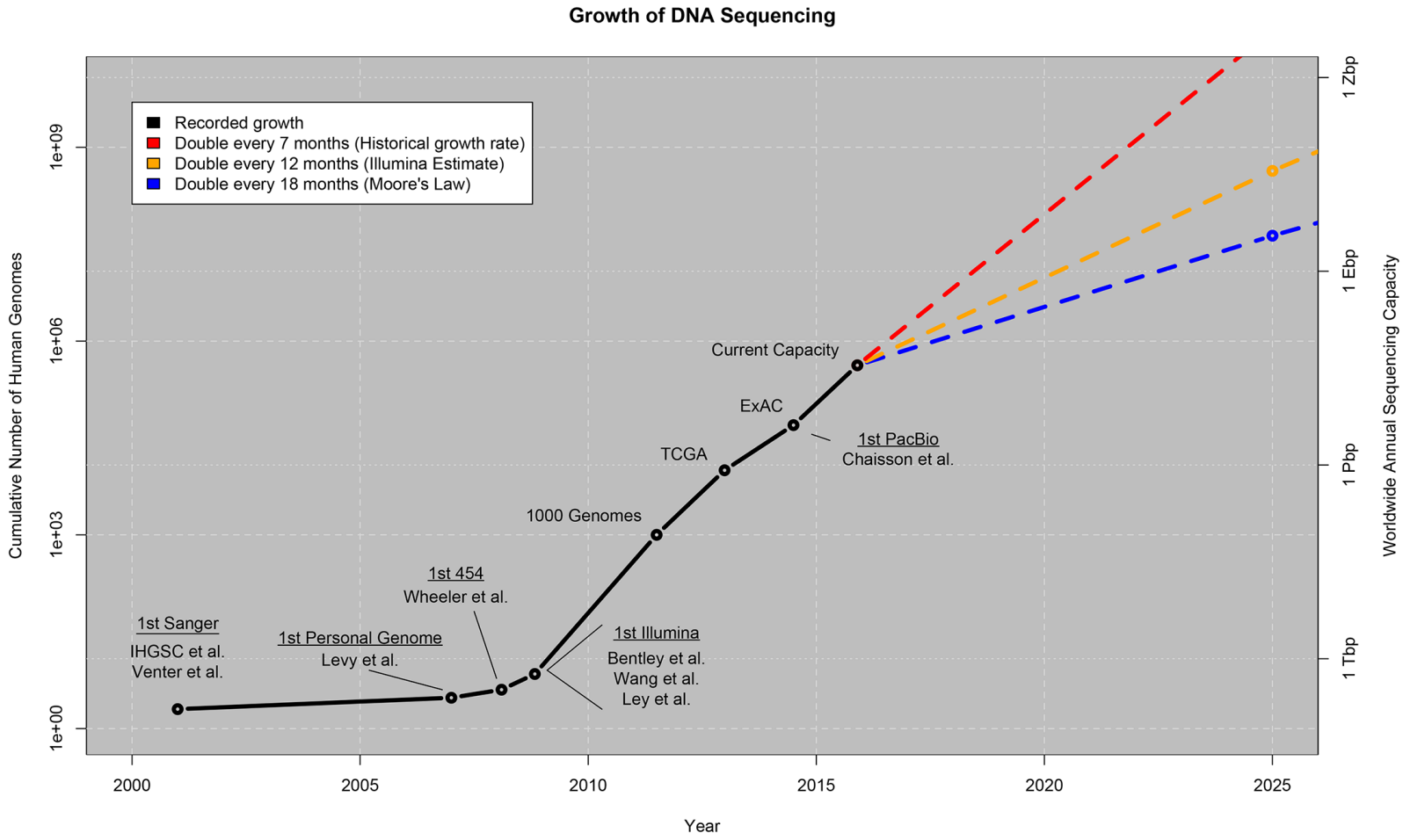
Outline

I. Challenges of data intensive biology research.

II. Some thoughts on dealing with these challenges.

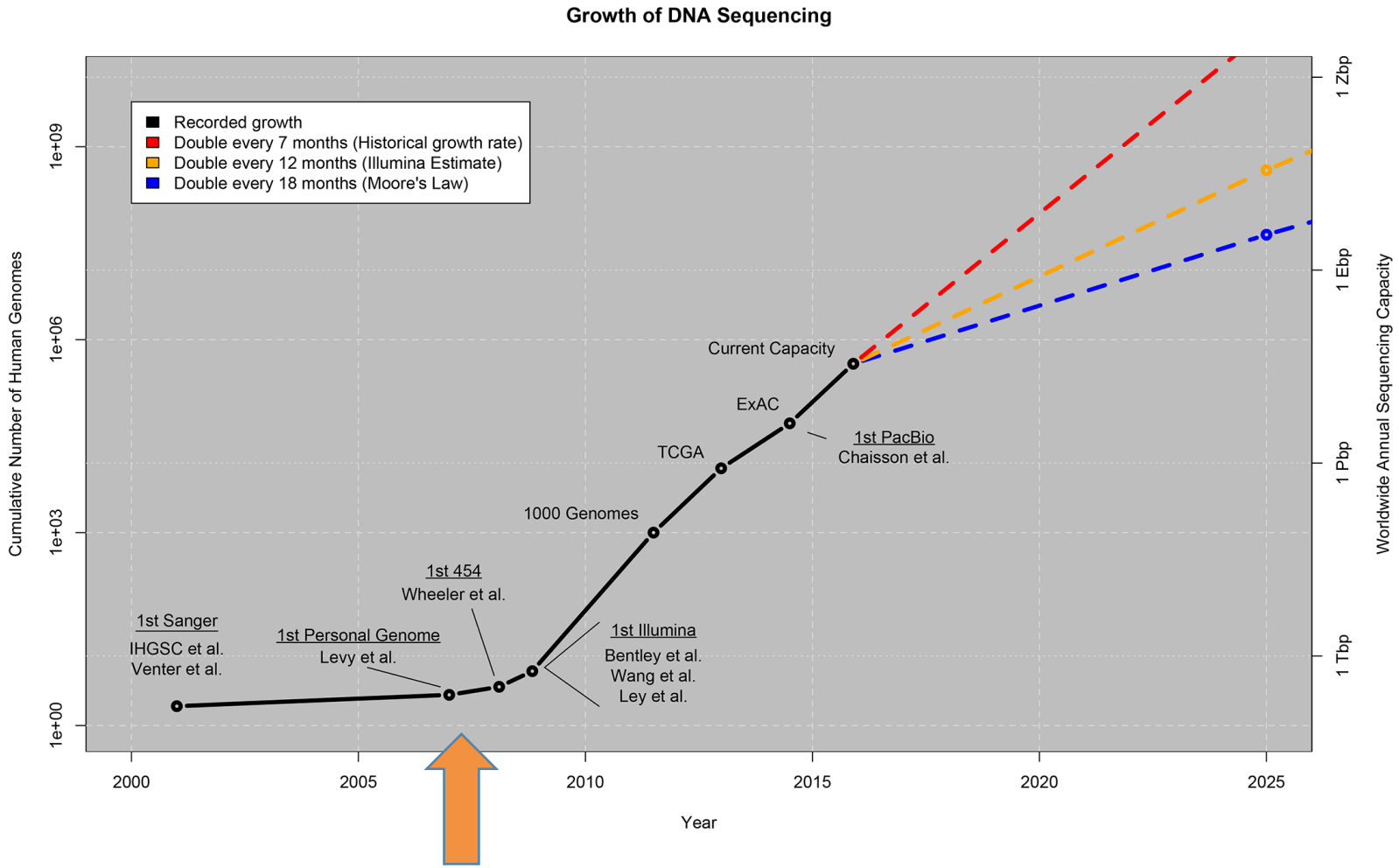
III. Career implications and my advice.

DNA sequencing rates continue to grow.



Stephens et al., 2015 - 10.1371/journal.pbio.1002195

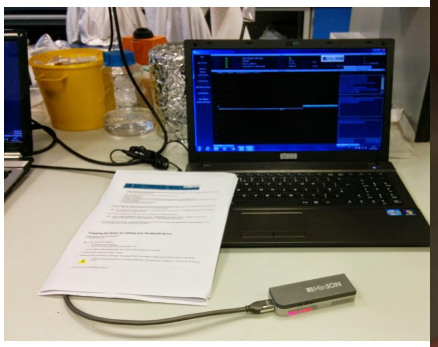
DNA sequencing rates continue to grow.



Stephens et al., 2015 - 10.1371/journal.pbio.1002195

Oxford Nanopore sequencing

MinION - the device



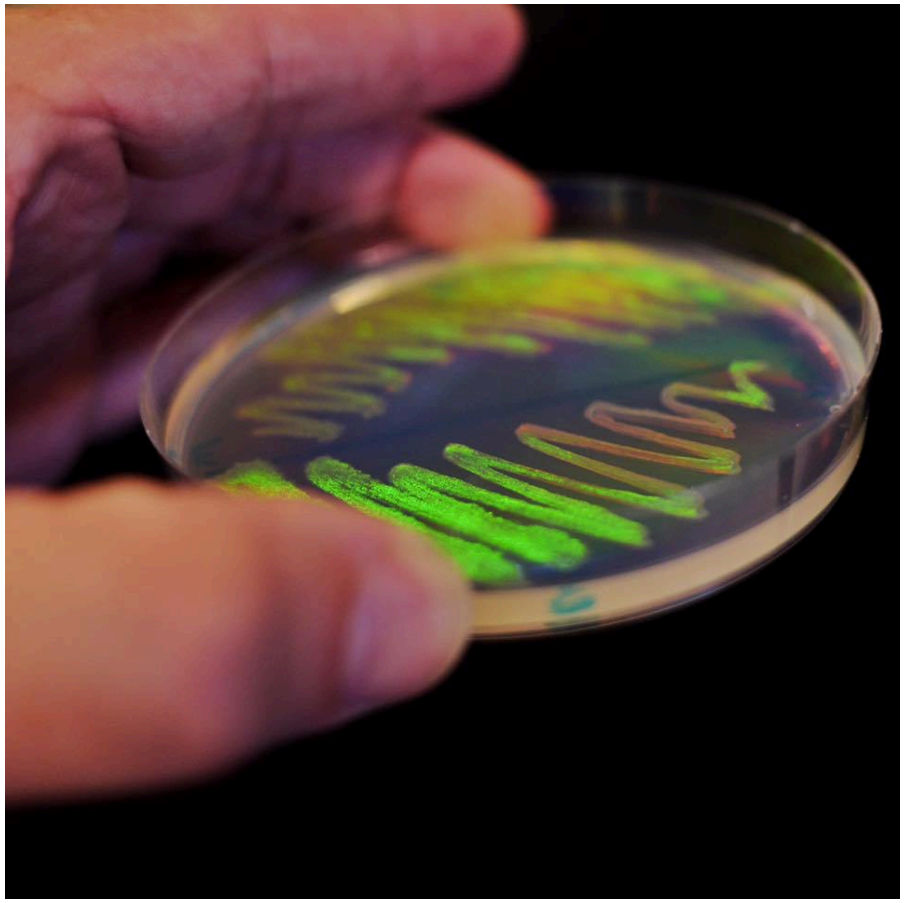
Images via Torsten Seemann & Lisa Cohen

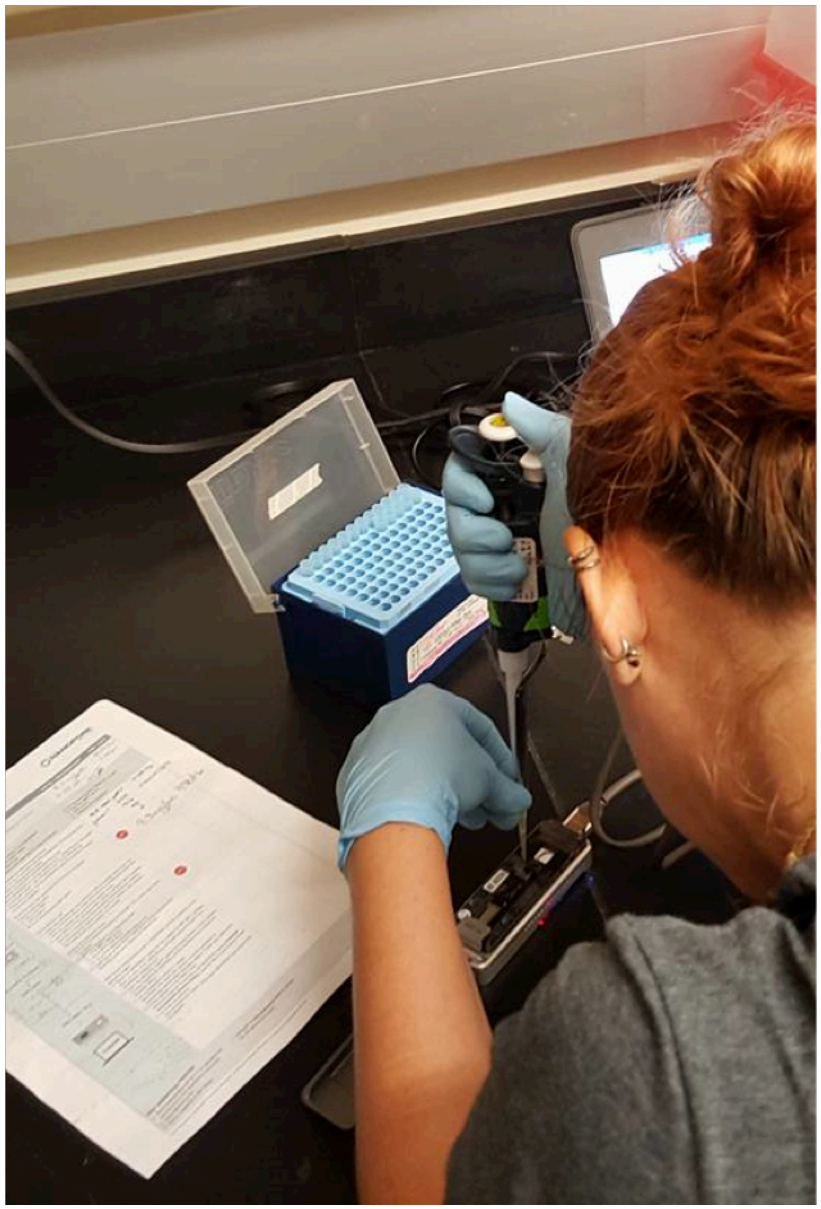
MinION - applications

:: Same as PacBio plus....

:: Portable sequencing
- : in the field eg. Josh Quick in Guinea for Ebola
- : in hospitals - infection control
- : monitoring - water/food supply, production facilities
- : at the GP - pathogen test in 10 min from blood prick?
- : spit in a home device every morning?

Slide via Torsten Seemann





Ectocooler

solated & sequenced genomes from two new microbes
n ~2 weeks last year!

- Still costly & requires fine-tuning and pipetting expertise (~$1000-2000 each genome)
- Both genomes assembled into only two contigs (better contiguity than Illumina)
- Revolutionary aspects of this:
  - “Personal” genome sequencing.
  - Inexpensive setup. Currently $5-10k to set up a sequencing station! & decreasing rapidly.

log post –
ttps://monsterbashseq.wordpress.com/2016/08/13/adventures-with-ont-
inion-at-mbls-microbial-diversity-course/

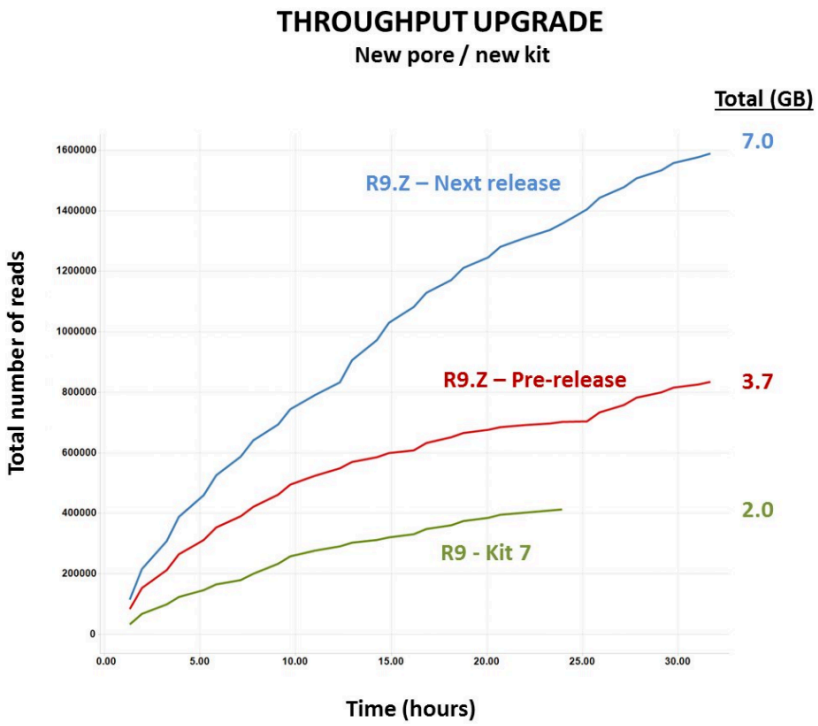
Lisa Cohen, Harriet Alexand
Rebecca Mickol and Kirsten

Scaling up --

Clive G. Brown
@Clive_G_Brown

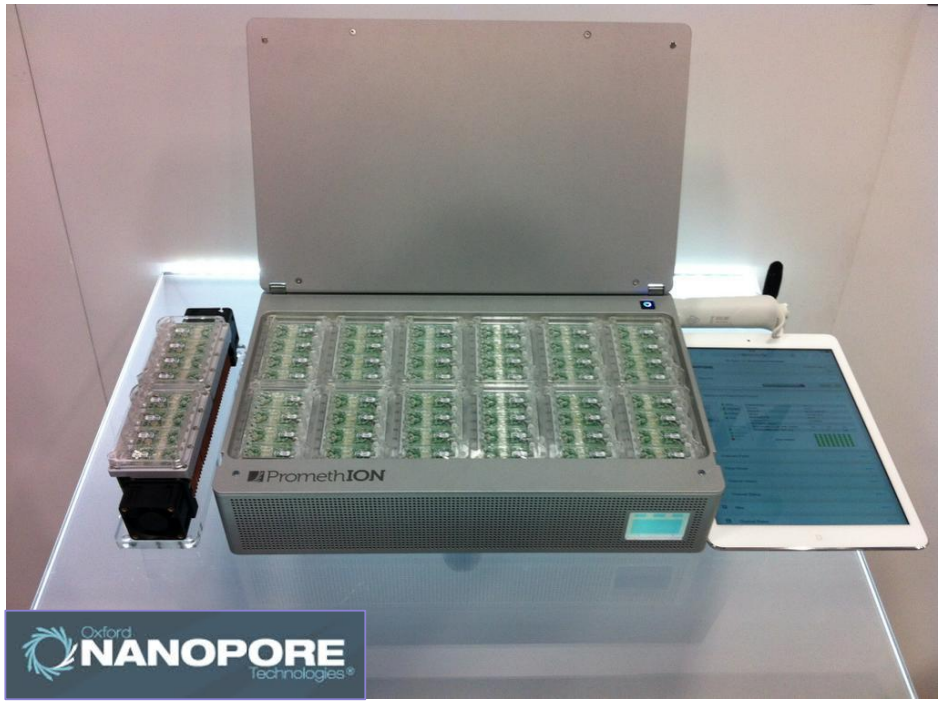
Follow

Further, 7.1G in 30 Hrs on next releases, should get ~10G in 48 hrs with some scripting tweaks. Tech update soon.



Scaling up --

PromethION - large scale device



:: 48 independent flow cells

:: On board ASIC

:: Runs Python

:: Optional compute

“Infinite Data”

Many data sets.

Large data sets.

Data *constantly arriving.*

This results in some interesting conversations

J. Random Student: “Hey, so I’m interested in <X system> and I just sequenced 700 isolates; what now?”

Q. Random Biologist: “What do I do with these terabases of metagenome data?”

Trina McMahon at JGI UM: “Isn’t this a tasty Smörgåsbord of data?”

(Shows literally dozens of data sets on a single plot.)

The *frequency* and *scale* of these conversations is increasing…

Question: how does the field change when:

Data is considerably less precious than **samples** and **analysis expertise?**

Biologists are able to ask targeted and general genomic questions about *their* specific system with ease?

One major limitation becomes *efficiency* and *repeatability of analysis* rather than *availability of data ?*

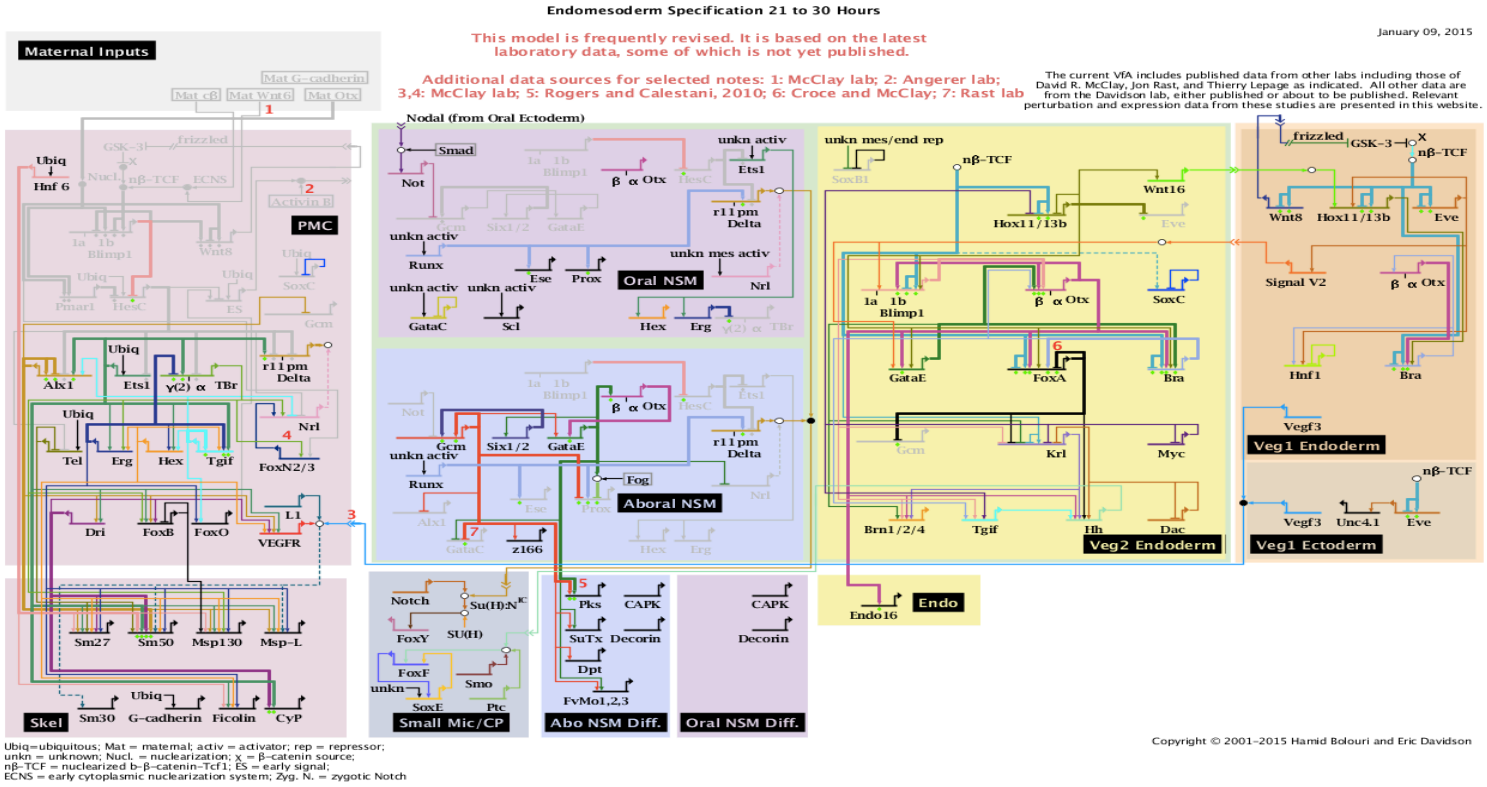
What are the implications for research and career?

Our models for education, training, research, and career are ill-suited for this world in a myriad of ways.

. Some challenges to consider --

1. Biology is very complicated.
2. We know very little about function in biology.
3. Very few people are trained in both data analysis and biology.
4. We are bad at working with multiple data types.
5. Our publishing system is holding back the sharing of knowledge.
6. We don't share data.
7. We are too focused on hypothesis-driven research.
8. Most computational research is not reproducible.

1. Biology is *complicated.*



Sea urchin gene network for early development; http://sugp.caltech.edu/endomes/

=> Biology is underdetermined by the data.

We are very far away from being able to infer *mechanism* from high-throughput data.

This is because there are so many interactions!

2. We know very little, and a lot of what we "know" is wrong.

One recent story that caught my eye – problems with genetic testing & databases. (See URL below for full story.)

- "1/4 of mutations linked to childhood diseases are debatable."
- In a study of 60,000 people, on average each had 53 "pathogenic" variants…

In microbiology, how many of our gene annotations are incomplete, or just plain *wrong?* (We know something specific about only ~50% of genes in *E. coli*!)

http://www.theatlantic.com/science/archive/2015/12/why-human-genetics-research-is-full-of-costly-mistakes/420693/

3. Very few people are trained in both data analysis and biology.

The *practical* and *pragmatic* issues of data analysis and data interpretation are not something that is taught at undergrad or graduate school level.

Most senior faculty do not know how to do this.

Nor do many junior faculty.

But biology increasingly *depends* on skilled interpretation of private + public data.

4. We are bad at working with multiple data types (data *integration*)

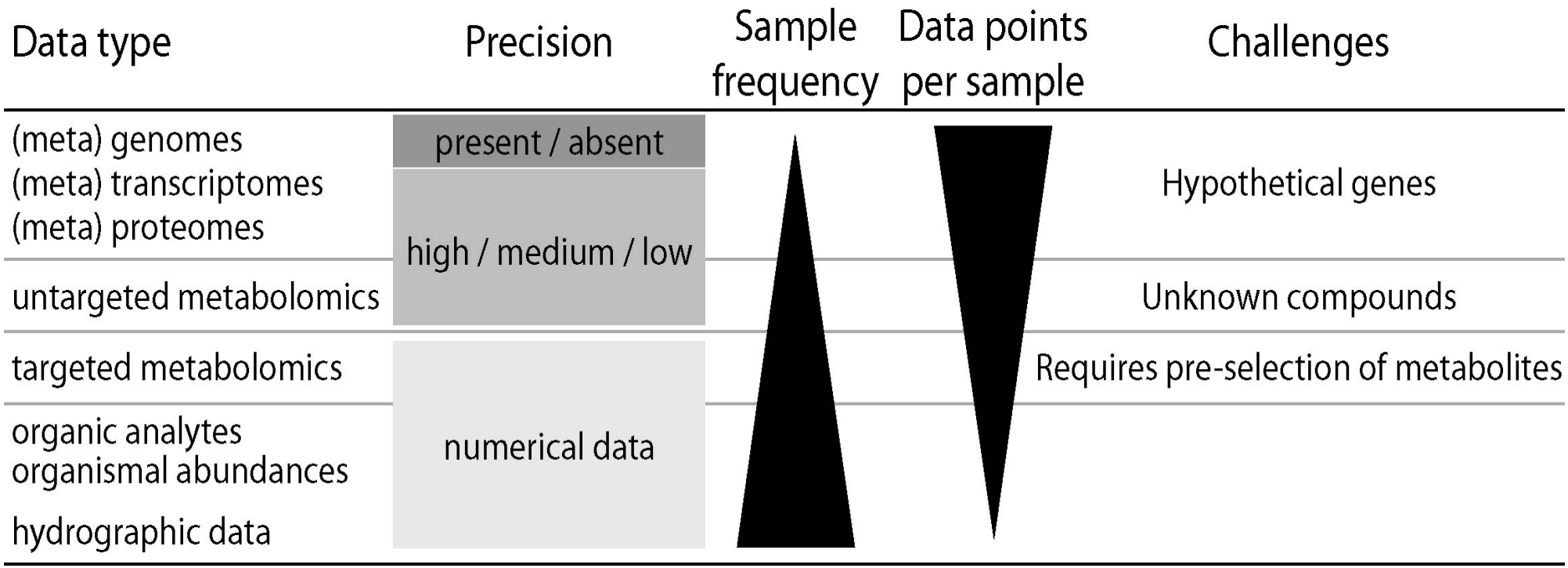


Figure 2. Summary of challenges associated with the data integration in the proposed project.

Figure via E. Kuja

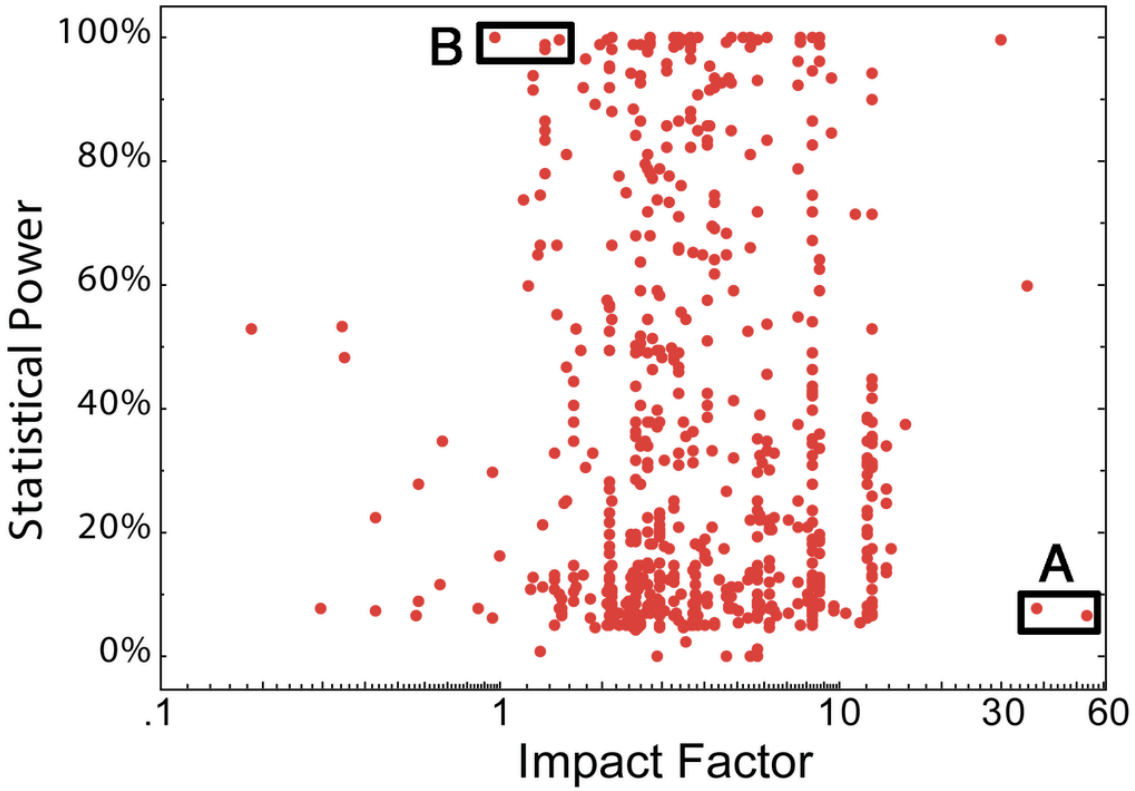
5. Our publishing system has become a real problem.

- The journal system costs more than $10bn/yr, with profit margins estimated at 20-30%*.
- Articles in high impact factor journals have lower statistical power.
- High-IF journals have higher rates of retractions (which cannot solely be attributed to “attention paid”)
- We publish in PDF form, which is computationally opaque.
- Publishing is slow!

* $10bn/year: http://www.stm-assoc.org/2015_02_20_STM_Report_2015.pd

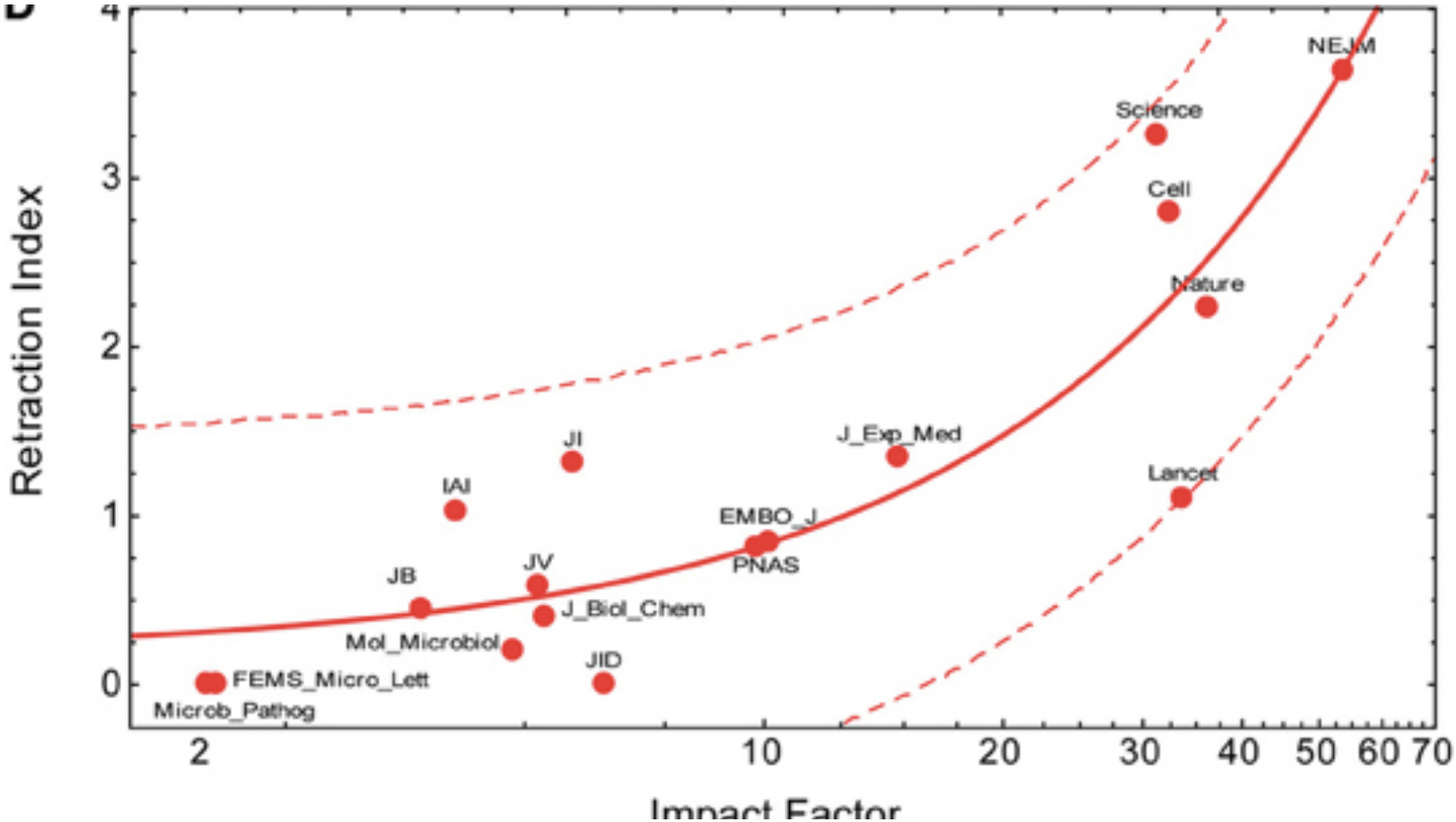
High-impact-factor articles have poor statistical power.

Our current system rewards A but not B.



Brembs et al., 2013 - http://journal.frontiersin.org/article/10.3389/fnhum.2013.00291/full

High impact factor => high retraction index.



Brembs et al., 2013 - http://journal.frontiersin.org/article/10.3389/fnhum.2013.00291/full

We need to move beyond PDFs

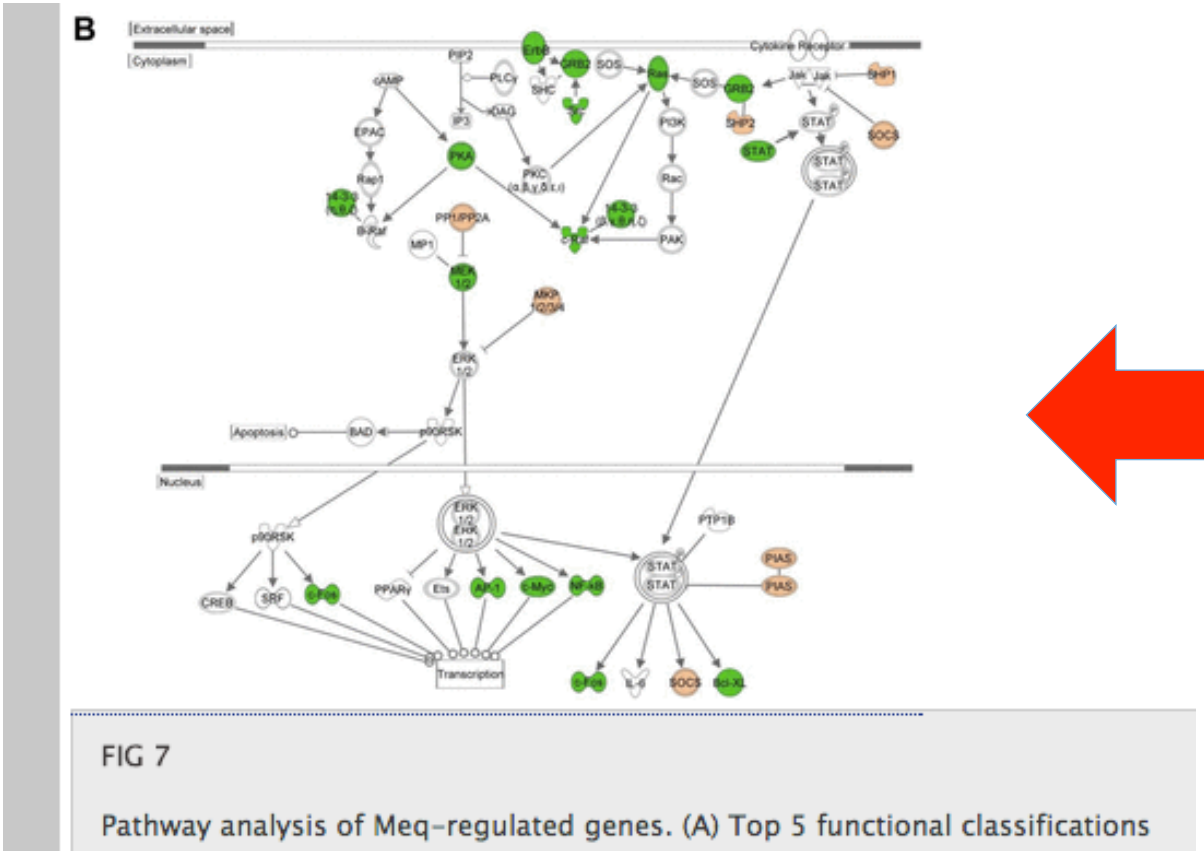


FIG 7

Pathway analysis of Meq-regulated genes. (A) Top 5 functional classifications

This is only part of the story!

Subramanian et al., doi: 10.1128/JVI.01163-13

6. We just don't share our data.

- Progress depends on collaboration and cooperation…
- But researchers have virtually no short-term incentives to share data in useful ways!
- "46% of respondents reported they do not make their data available to others" – study in ecology (Tenopir et al., 2011)
- Some tremendously depressing stories from the rare disease community about how career considerations trump all else.

http://www.newyorker.com/magazine/2014/07/21/one-of-a-kind-2

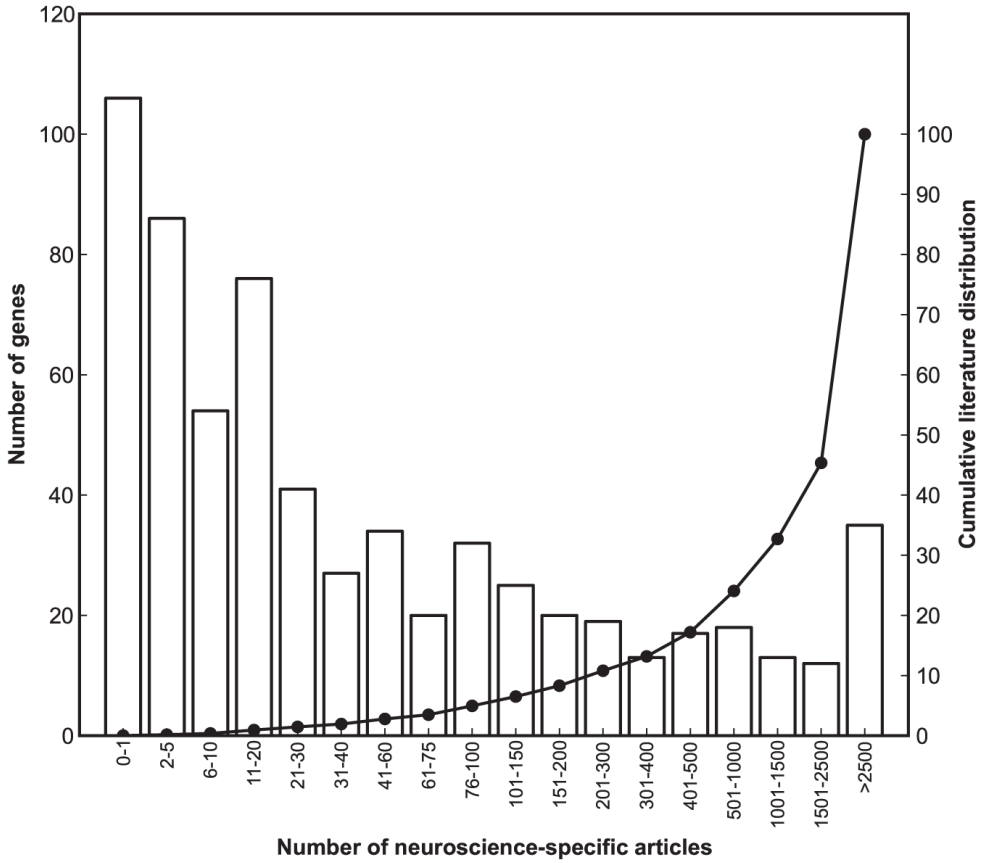
7. We are focused on hypothesis-driven research.

Granting agencies and reviewers typically require specific hypotheses, even when little is known.

This focuses research on “known unknowns”, and leaves “unknown unknowns” out in the cold.

(This is not true in all fields!)

The problem of lopsided gene characterization is pervasive: e.g., the brain "ignorome"



"...ignorome genes do not differ from well-studied genes in terms of connectivity in coexpression networks. Nor do they differ with respect to numbers of orthologs, paralogs, or protein domains. The major distinguishing characteristic between these sets of genes is date of discovery, early discovery being associated with greater research momentum—a genomic bandwagon effect."

Ref.: Pandey et al. (2014), PLoS One *11*, e88889.

Via Erich Schwarz

3. Most computational research is not reproducible.

I don't know of a systematic study in computational science specifically, but of papers that I read, approximately 95% fail to include details necessary for replication.

It's very hard to build off of research like this.

(There's a lot more to say about repeatability, reproducibility and replicability than I can fit in here…)

I. Some thoughts on dealing with these challenges.

Lots of possible bandaids, etc.

…but what is a long term direction, or goal?

Think about implications of CRISPR, genome editing, and bioengineering…

The challenge with genome editing is fast becoming *what to edit* rather than *how to do it.*

A Major Challenge in biology:

No strongly predictive theory

There is an **amazing** correspondence between theory and experiment in many parts of physics.

As a result, theory and modeling play a central role in physics research.

This simply does not work in biology, because there is no strong predictive theory - neither ecology nor evolution provides such.

(The Central Dogma is about as close as we get.)

A Major Challenge in biology:

No strongly predictive theory

Theory is unlikely to play a direct and central role in understanding biological systems.

Data integration is probably where strong theory would have been most useful… but alas.

Modeling may be our best (only?) hope for integrating, interpreting, and understanding large volumes of data.

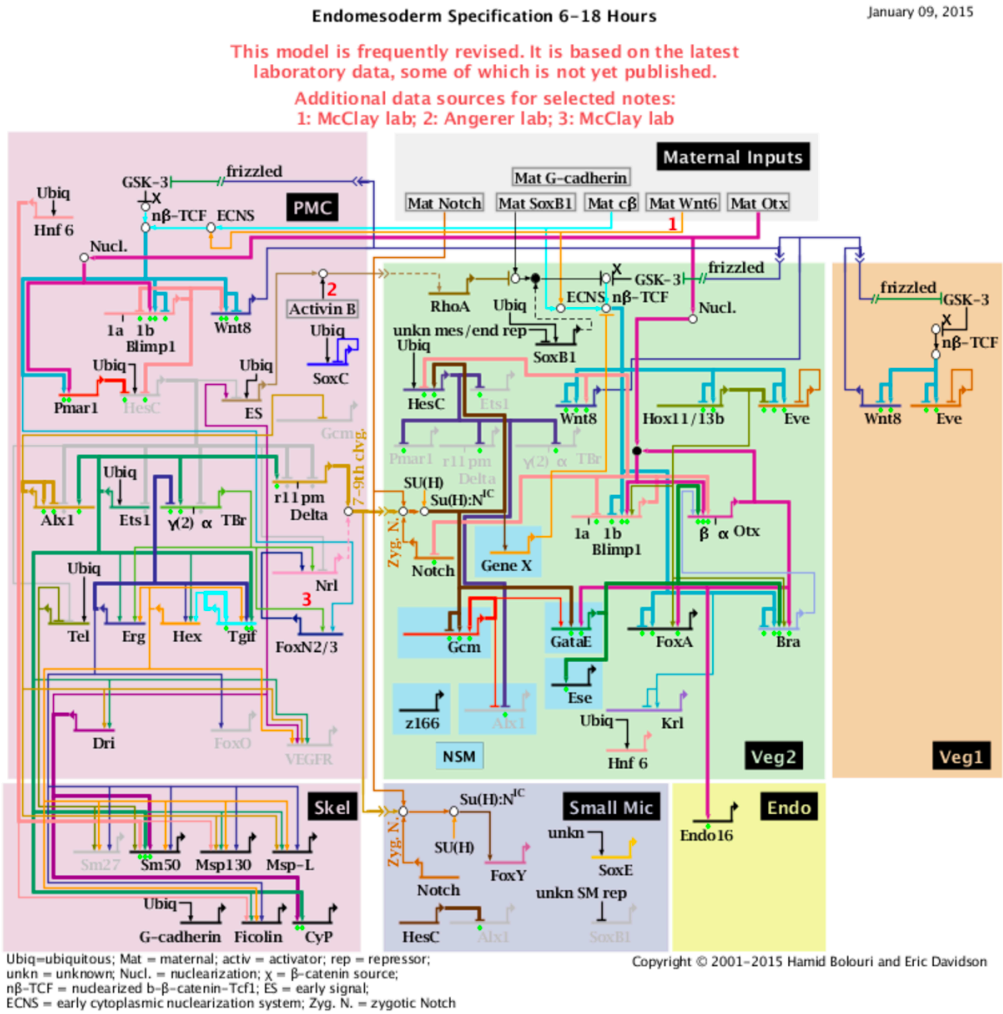
Modeling for data integration

Can we build a model to separate data into "boring" and "interesting" bins?

If our data is explained by the model at a useful level of granularity, then the model is sufficient to explain the data. (Boring! But good!)

If our data does not fit the model, the model is insufficient and must be adjusted! (Interesting!! But *hard*!)

My experience with Sea urchin Gene regulatory networks



http://sugp.caltech.ed

Key features of this model

- Served as common reference point for disagreements, even about things *not* in the model.
- Became a kind of community resource.
- Provided robust tool for in silico thought experiments; led directly to testable hypotheses.
- Illuminated unknowns and uncertainties in a direct and productive way.

Modeling for data integration

Can we build a model to separate data into "boring" and "interesting" bins?

If our data is explained by the model at a useful level of granularity, then the model is sufficient to explain the data. (Boring! But good!)

If our data does not fit the model, the model is insufficient and must be adjusted! (Interesting!! But *hard*!)

Obstacles to building models

- Researchers center their world view around what they are studying!
  (MicDiv 2016 Anecdote here)
  => models must permit ”black boxes”, or a hierarchy of abstractions.
- Funding needs very much distract from focused effort
- Academia disincentivizes deep sharing and clear communication.
- Experimental biologists generally do not seem to pay much attention to models.

n 5 years I would like to be doing a lot more *modeling* and a lot less *bioinformatics.*

(10 years might be more realistic.)

All models are wrong…

All models are wrong…

…but some models are useful.

What would be ***useful*** types of models for the microbiome ***community*** to invest in?

III. What are the implications of all this for research and career?

You don't actually *need* to do anything about any of these challenges.

…but I think they do represent an *opportunity*.

“Chance favors the prepared mind”

- Louis Pasteur

The rapidly decreasing cost of data represents many opportunities that you can deploy in favor of your science. But will you recognize those opportunities, be able to distinguish them from distractions, and be able to take advantage of them?

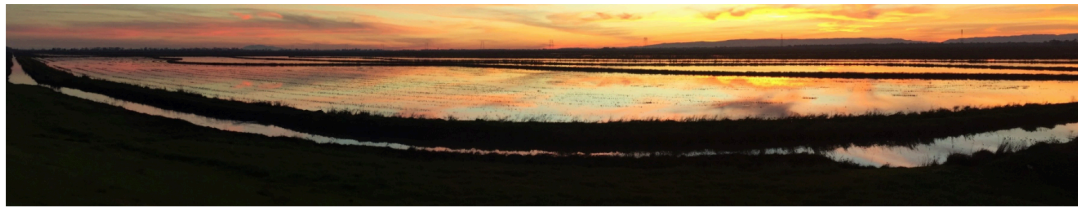
Embrace interdisciplinarity! Explore outside your comfort bubble.

Read broadly! There is lots of of deep thinking on biology and science and engineering; figure out what you agree, disagree with.

nvest in researcher-focused computational training

DATA CARPENTRY

software carpentry



2017 Data Intensive Biology Summer Institute at UC Davis

From June 19 to July 21, 2017, we'll be running several different computational training events at the University o California, Davis – a two-week workshop for biologists to learn basic bioinformatics and high performance computing, a week-long instructor training in how to reuse our materials, and several other focused workshops.

All of this will be taking place at the UC Davis School of Veterinary Medicine, on the main campus of UC Davis (in Davis, CA). Read more...

Everyone is welcome, from everywhere! Grad students, postdocs, faculty, staff, industry, non-profits, teachers, journalists, and grant managers may all find these workshops useful. **We welcome applications from international students and regard the current policies of the United States with frustration and dismay.** Rea more...

Each workshop costs between $350 and $500, and on-campus room & board will be approximately $400/week f non-local attendees.

June 19-23 - Instructor training

June 26-July 8th: Two week introductory bioinformatics workshop (ANGUS)

July 10-15, 2017 - One-week workshops

July 17-21, 2017 - One-week workshops

Lots of online tutorials:

http://www.datacarpentry.org/ - command line, R, etc.

Two week intensive bioinformatics workshop:
https://angus.readthedocs.io/en/2017/

16s metagenomics focused:

https://nceas.github.io/oss-lessons/metagenomics/

Shotgun metagenomics focused:

https://2017-dibsi-metagenomics.readthedocs.io/

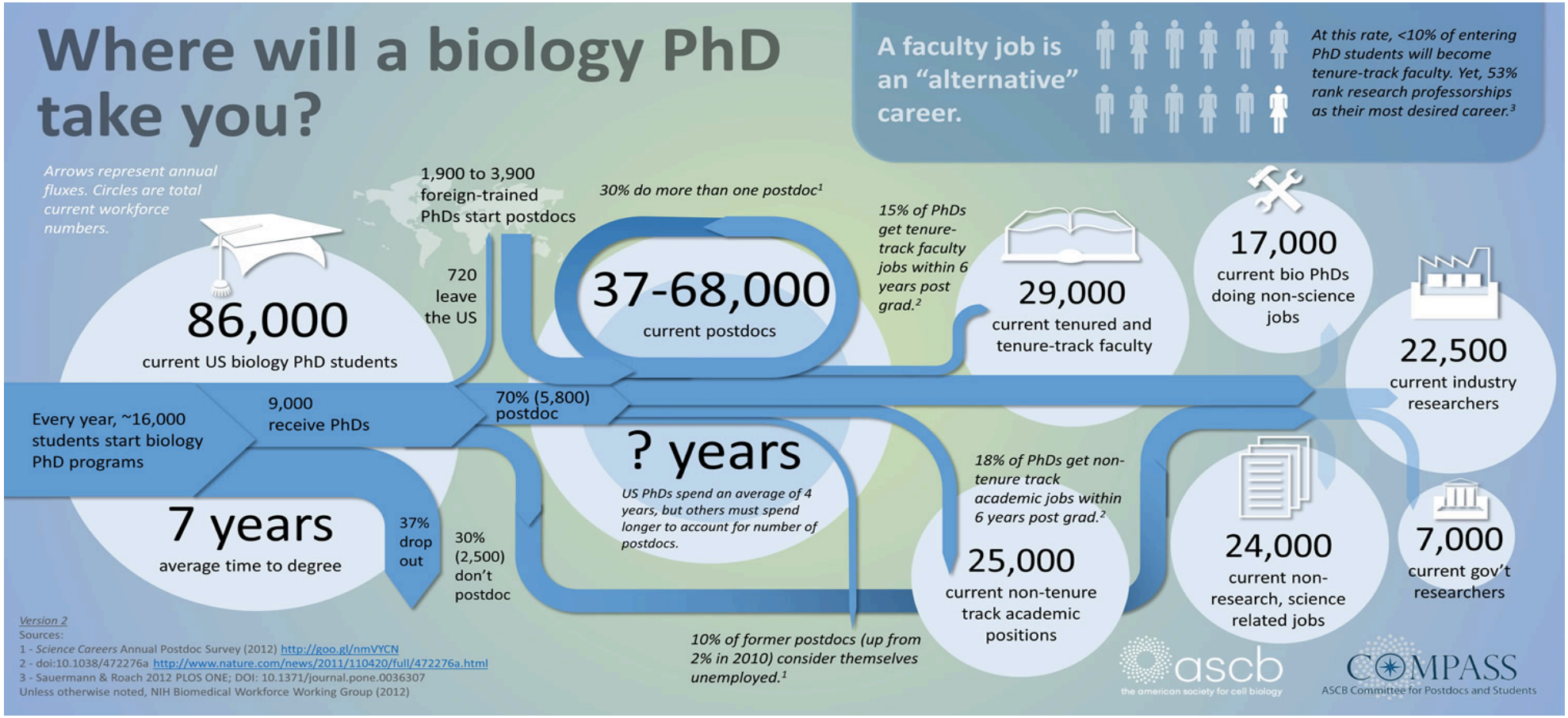
Also, come visit!

Davis is an excellent place to visit (avoid June-August tho)

Scott and I are both there!

Post-course research $$ from MBL can help support!

Think broadly about careers



http://www.ascb.org/ascbpost/index.php/compass-points/item/285-where-will-a-biology-phd-take-you

Tenure track / R1 positions are in the minority.

< 10% of entering PhD students will become tenure track faculty.*

53% rank research professorships as their desired career.*

Universities have increasingly little provision for stable non-tenure-track positions.

* http://www.ascb.org/ascbpost/index.php/compass-points/item/285-where-will-a-biology-phd-take-you

But! Don't despair!

PhD research prepares you *marvelously* for tackling an immense range of problems!!

Biotech, startups, research institutes, teaching, science communication…

(PhD advisors generally do *not* do such a good job of preparing you for non-tenure track positions.)

Papers are *necessary* to graduate but *insufficient* to get you a non-academic job afterwards. You want to have *demonstrable skills.*

Concluding retrospective

1. Data-intensive research presents lots of challenges!

2. These challenges can be turned into opportunities! Keep your eye on the ball (understanding something) but also build your skills up!

3. (I think modeling is maybe a solution for data integration/ communication challenges in biology.)

4. Biologists are living in “interesting times”…

Points for reflection…

There has never been a better time to be in biology, in terms of our ability to get at mechanistic questions! And it's only getting better!

Increasingly, the best guide to the next 10 years of biology is *science fiction* ...

Some reading in the form of blog posts.

"How to analyze, integrate, and model large volumes of biological data - some thoughts." http://ivory.idyll.org/blog/2017-oregon-microbiome-data-integration.html

"What about all those genes of unknown function?" http://ivory.idyll.org/blog/2014-function-of-unknown-genes.html

Thanks for listening!

Please contact me at ctbrown@ucdavis.edu!