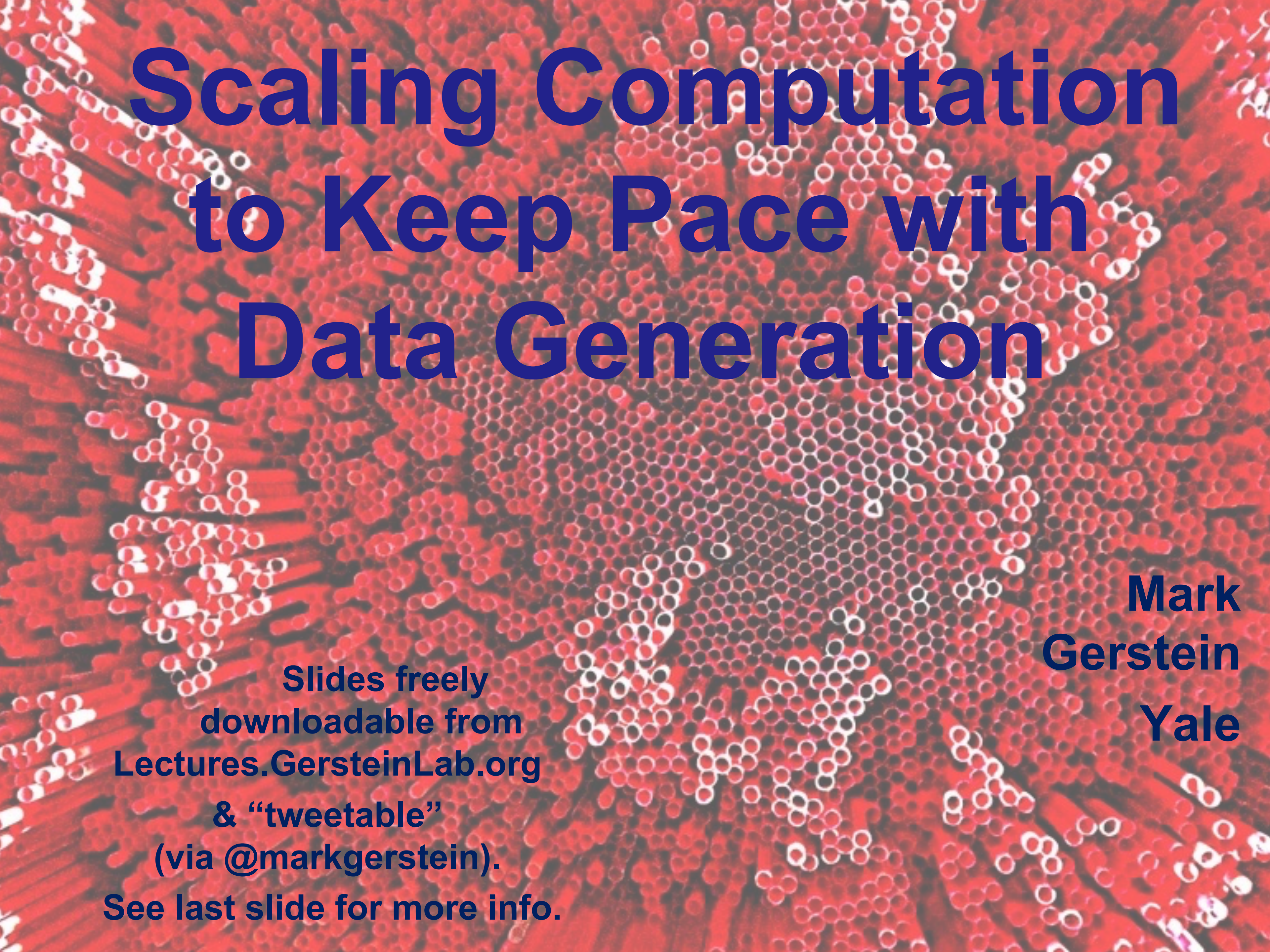


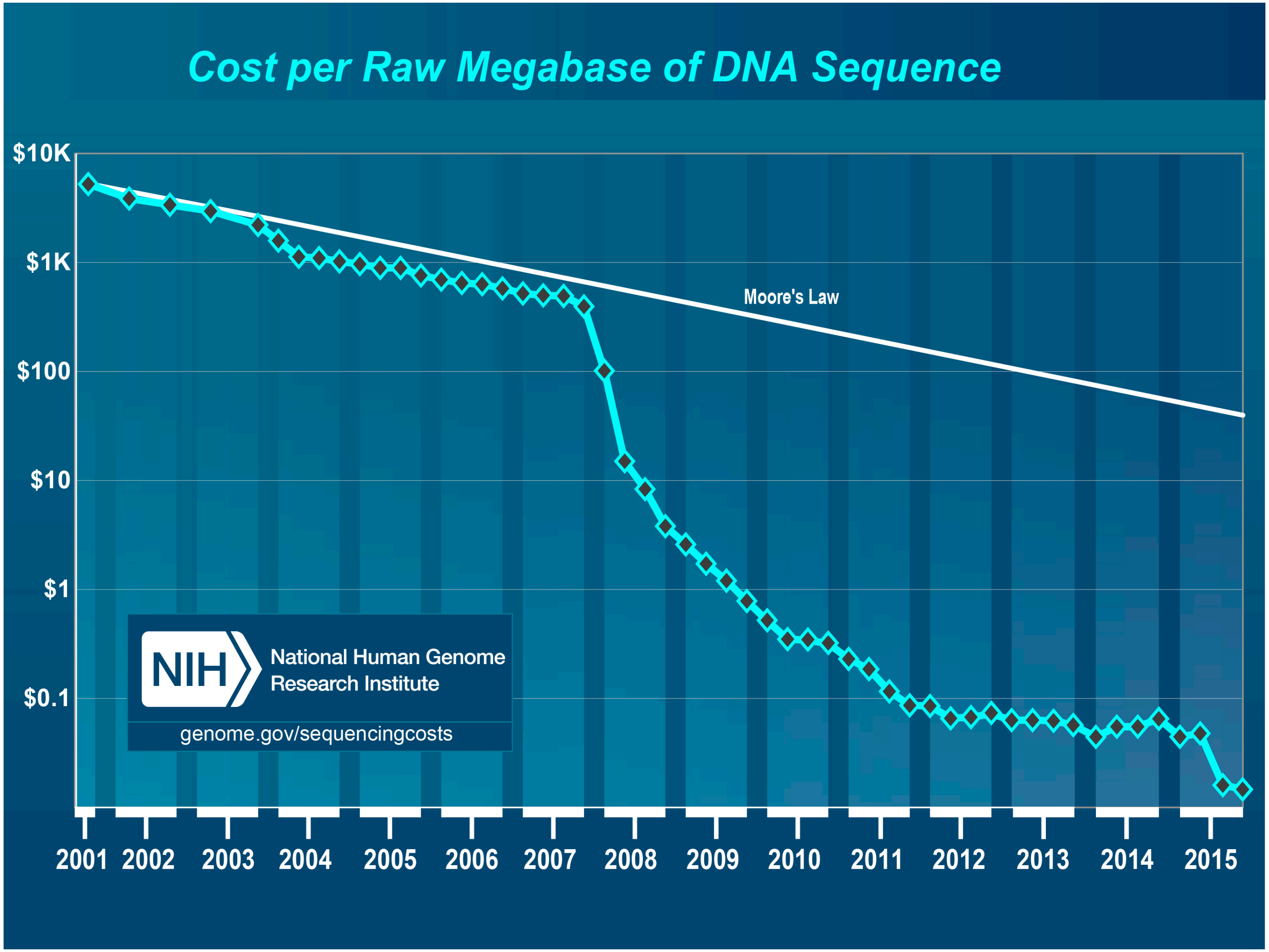
Scaling Computation to Keep Pace with Data Generation

Slides freely downloadable from Lectures.GersteinLab.org & “tweetable” (via @markgerstein).

See last slide for more info.

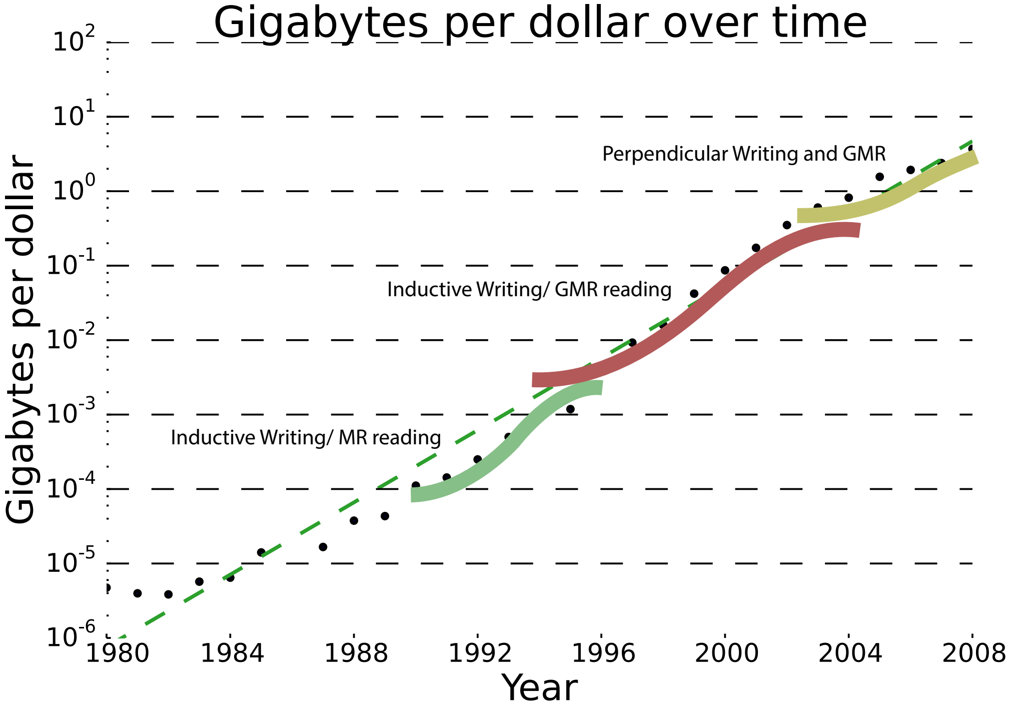
Mark
Gerstein
Yale

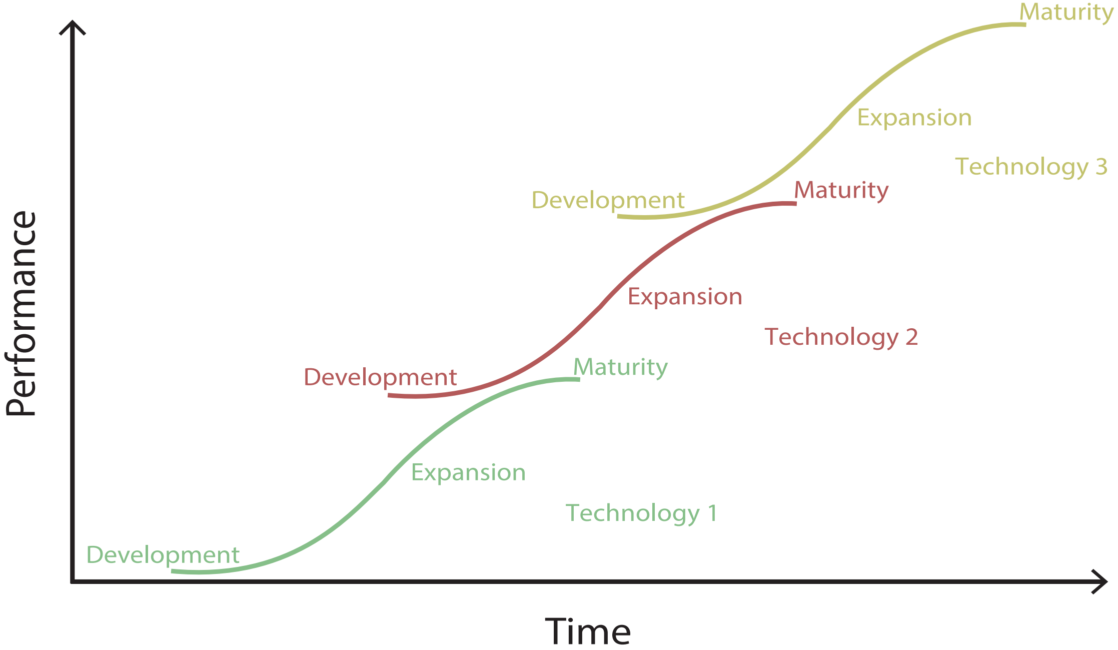
Sequencing Data Explosion: Faster than Moore's Law for a Time



Kryder's Law and S-curves underlying exponential growth

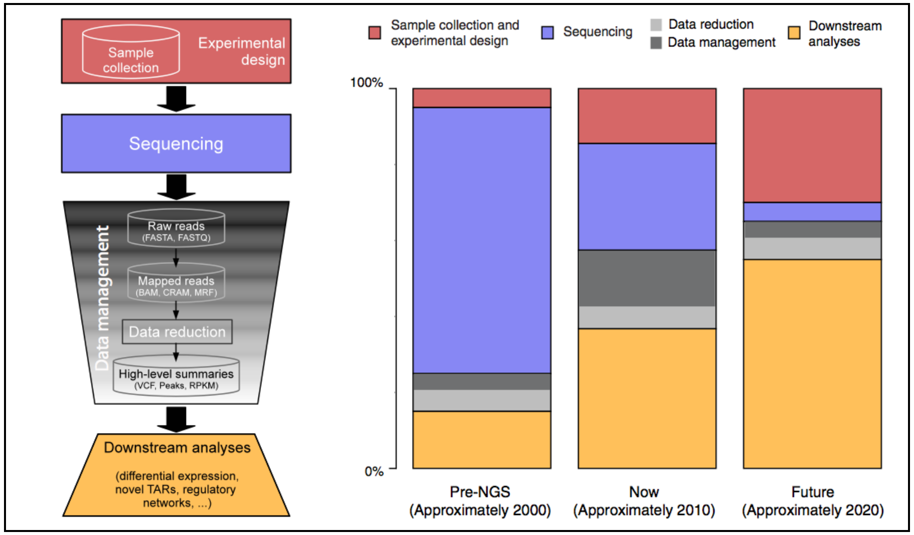
- Moore's & Kryder's Laws
  - As important as the increase in computer speed has been, the ability to store large amounts of information on computers is even more crucial
- Exponential increase seen in Kryder's law is a superposition of S-curves for different technologies





[Muir et al. ('16) GenomeBiol.]

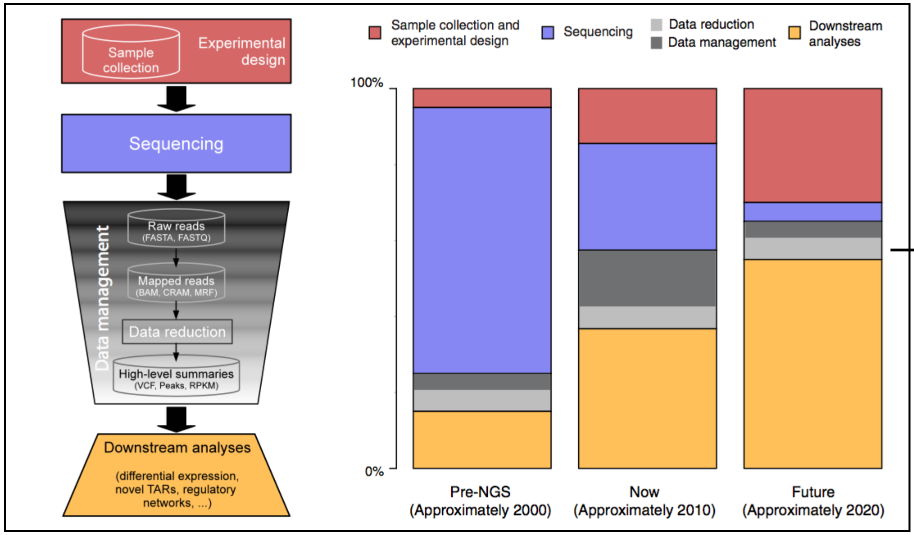
The changing costs of a sequencing pipeline

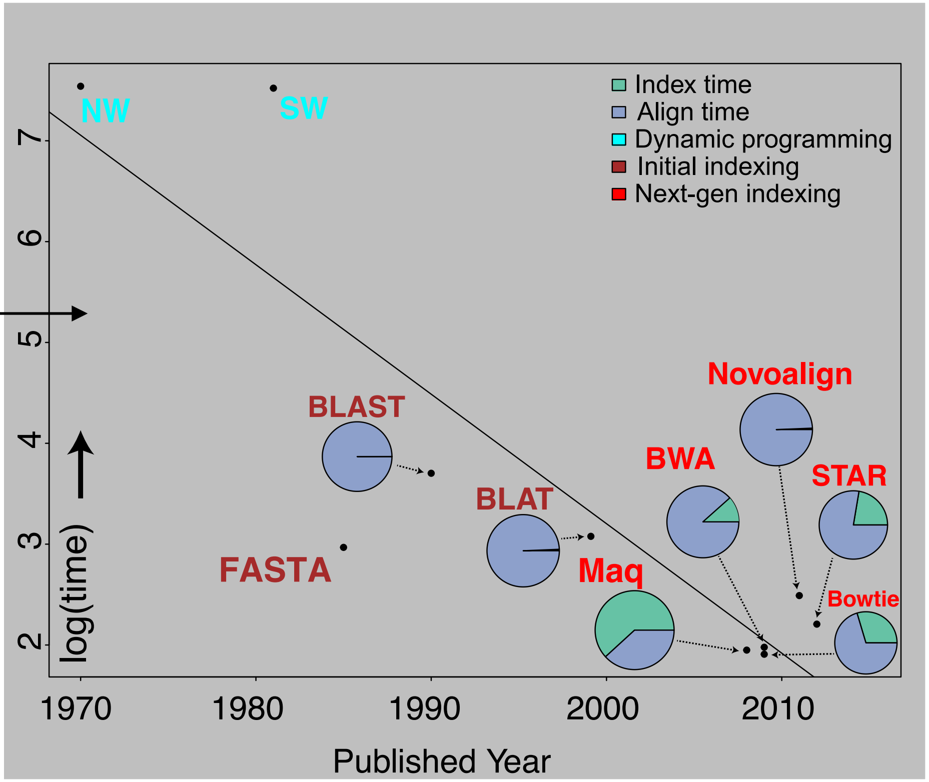


From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

[Sboner et al. (‘11), Muir et al. (‘16) Genome Biology]

The changing costs of a sequencing pipeline



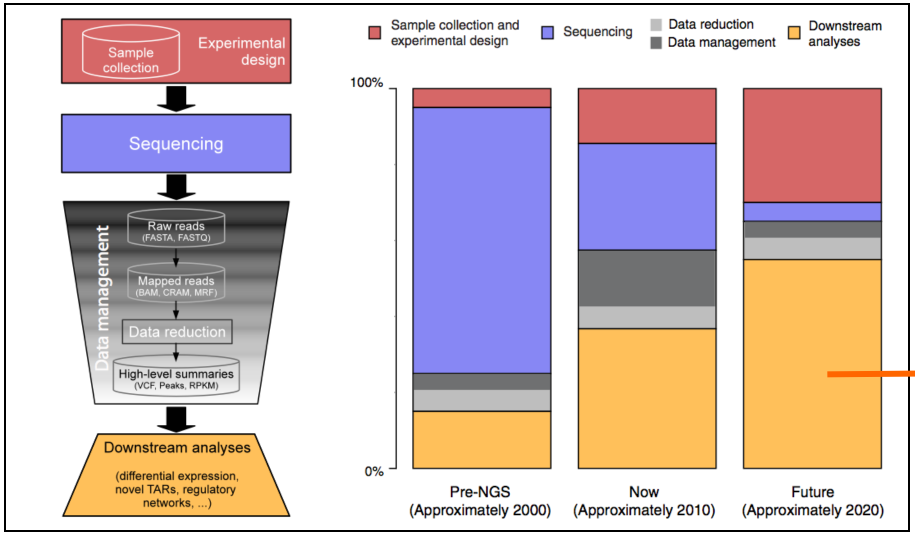


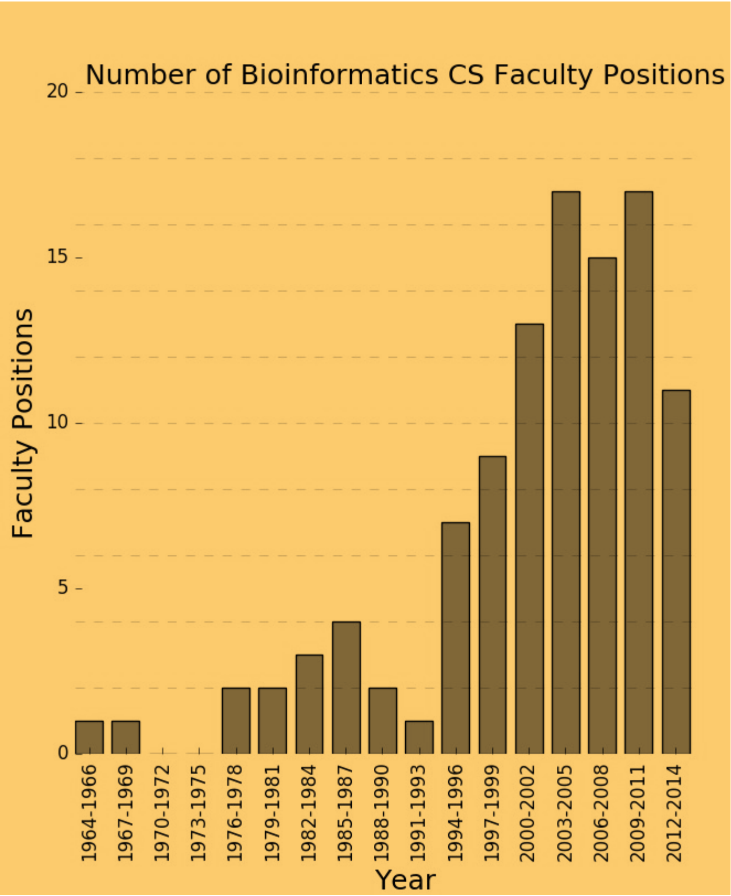
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

Alignment algorithms scaling to keep pace with data generation

[Sboner et al. (‘11), Muir et al. (‘16) Genome Biology]

The changing costs of a sequencing pipeline

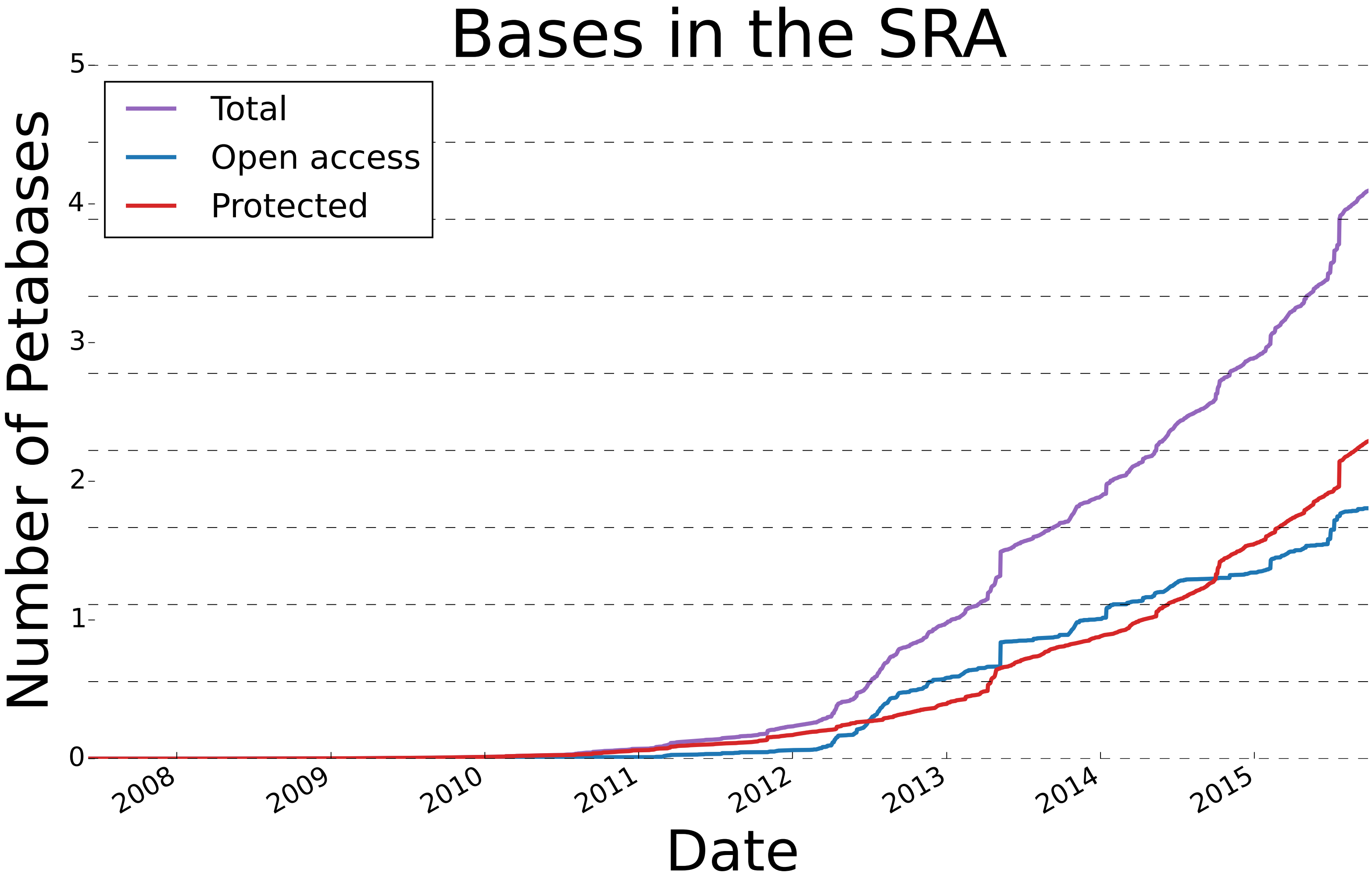




From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

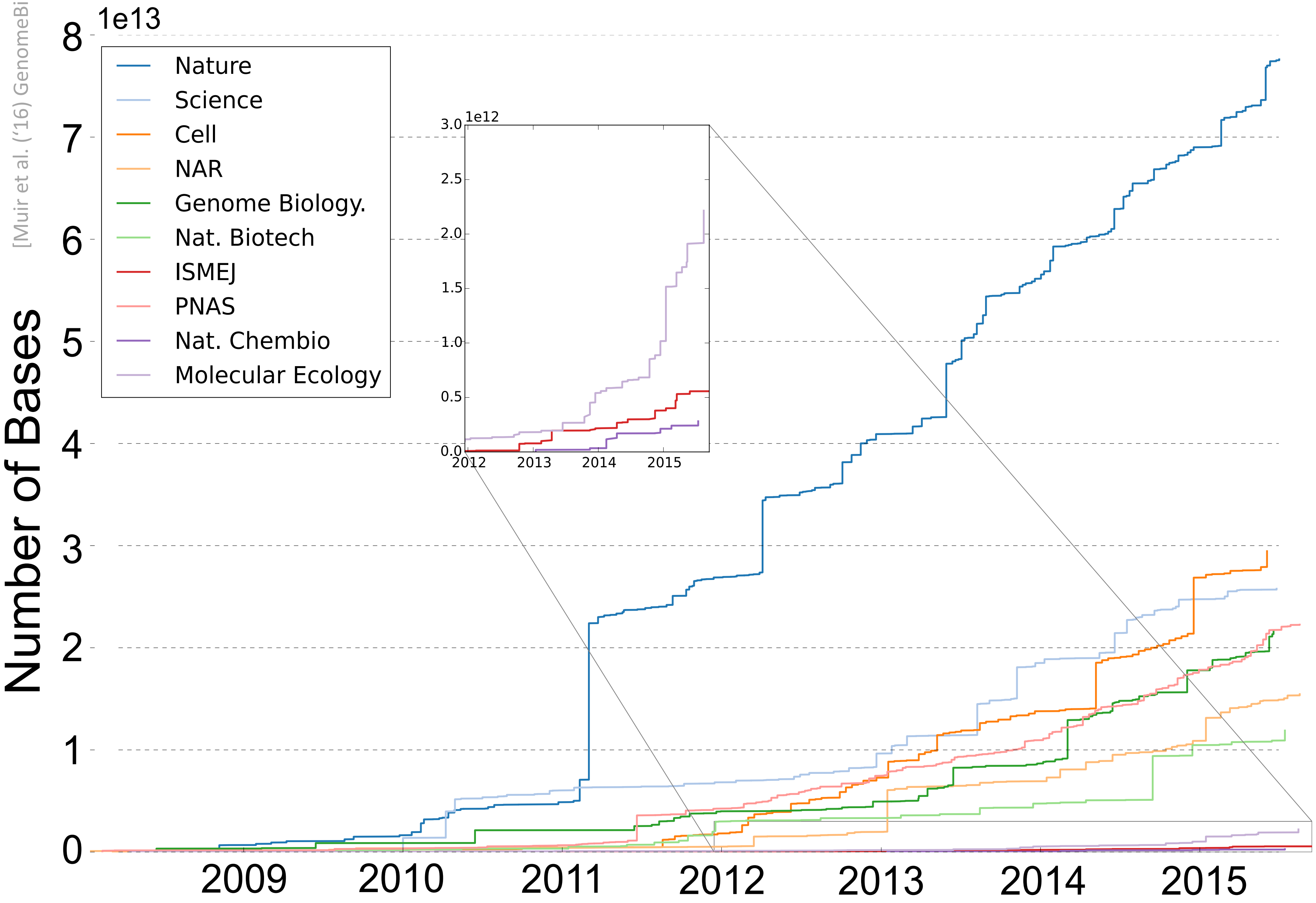
[Sboner et al. (‘11), Muir et al. (‘16) Genome Biology]

Sequencing cost reductions have resulted in a data explosion



[Muir et al. ('16) GenomeBiol.]

Increasing diversity in usage of sequence data

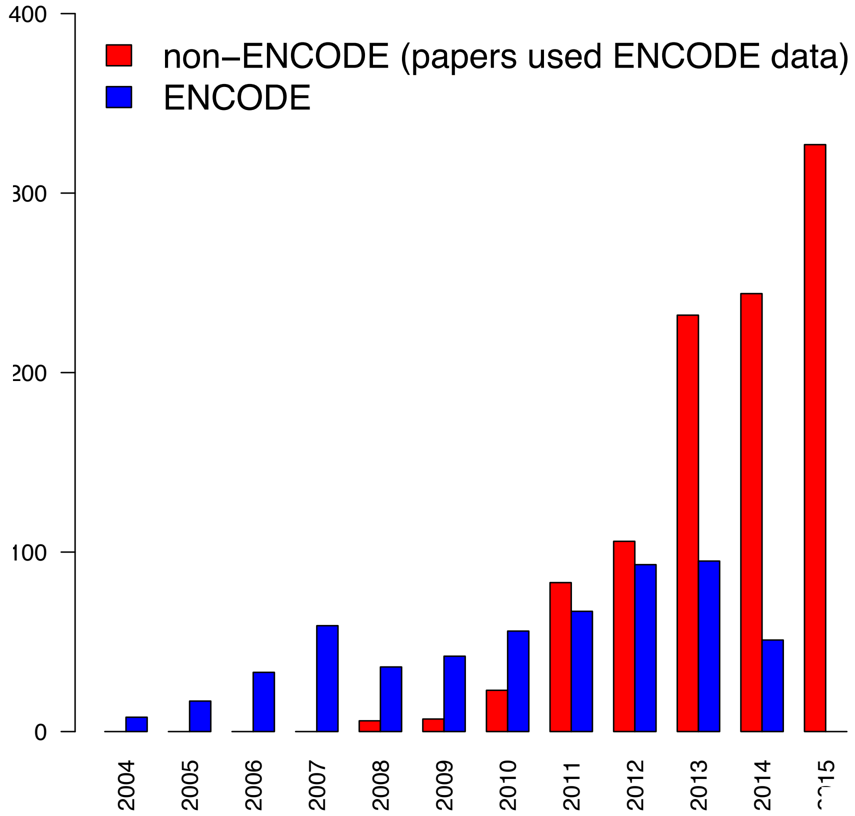


[Muir et al. ('16) GenomeBiol.]

With help of M Pazin at NHGRI, identified: 702 community papers that used ENCODE data but were not supported by ENCODE funding &
558 consortium papers supported by ENCODE funding
Then identified 1,786 ENCODE members & 8,263 non-members .

🟥 non–ENCODE (papers used ENCODE data) 🟦 ENCODE

<= # Papers



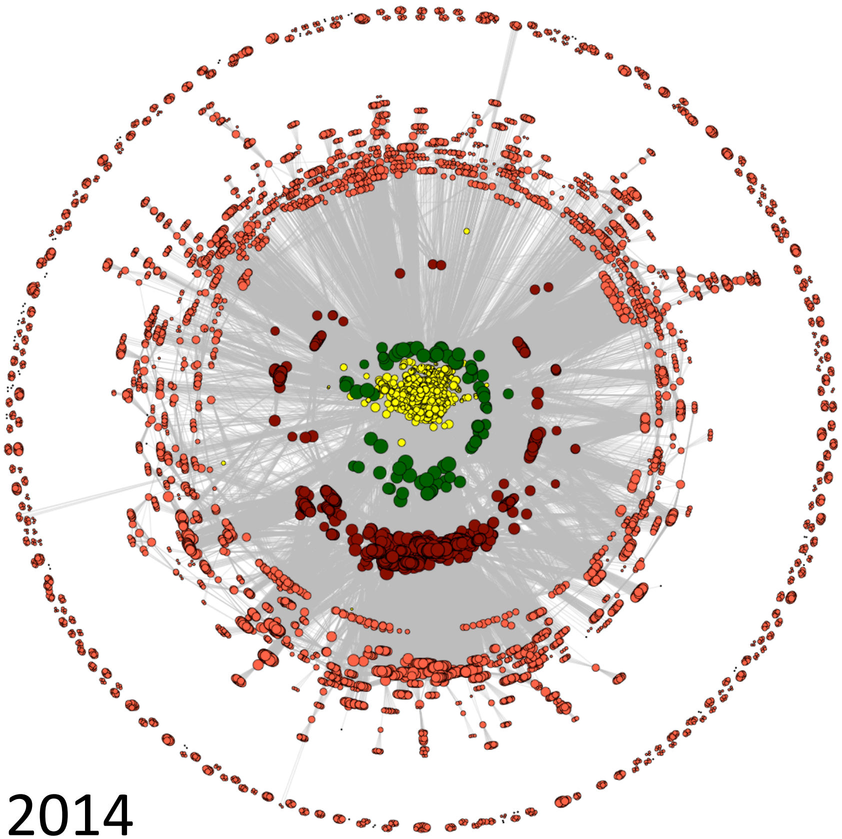
# Authors

[Wang et al., TIG ('16)]

Yr. ('04 to '15)

Co-authorship Network of ENCODE members & Data Users

- ● ENCODE member
- ● non-member
- ● ENCODE member broker
- ● non-member broker
- — co-authorship



2014

[Wang et al., TIG ('16)]

Co-authorship Network of ENCODE members & Data Users

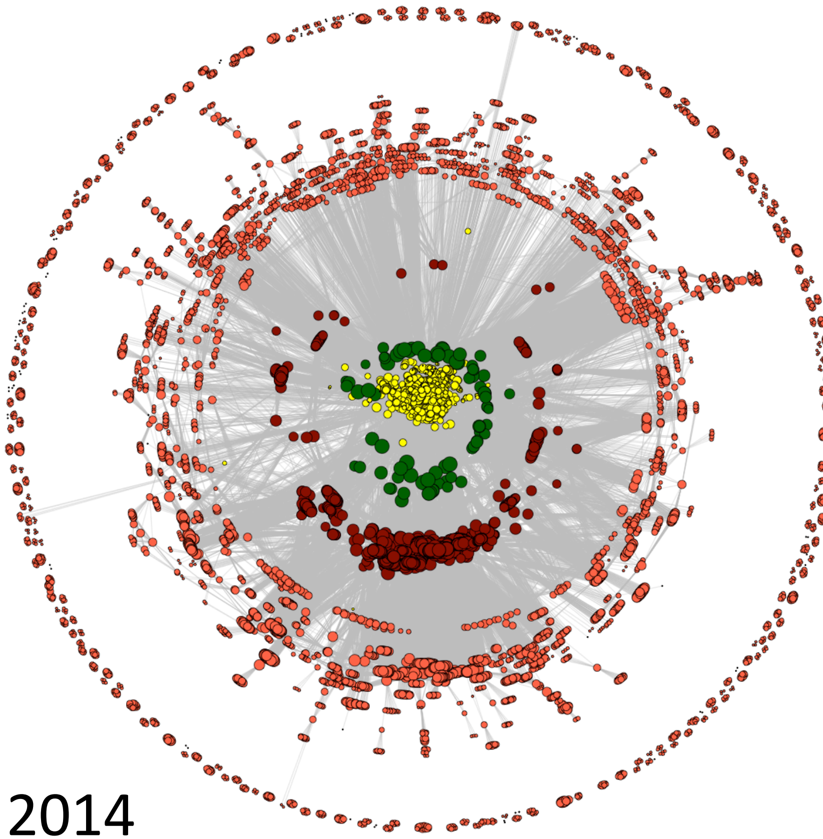
● ENCODE member

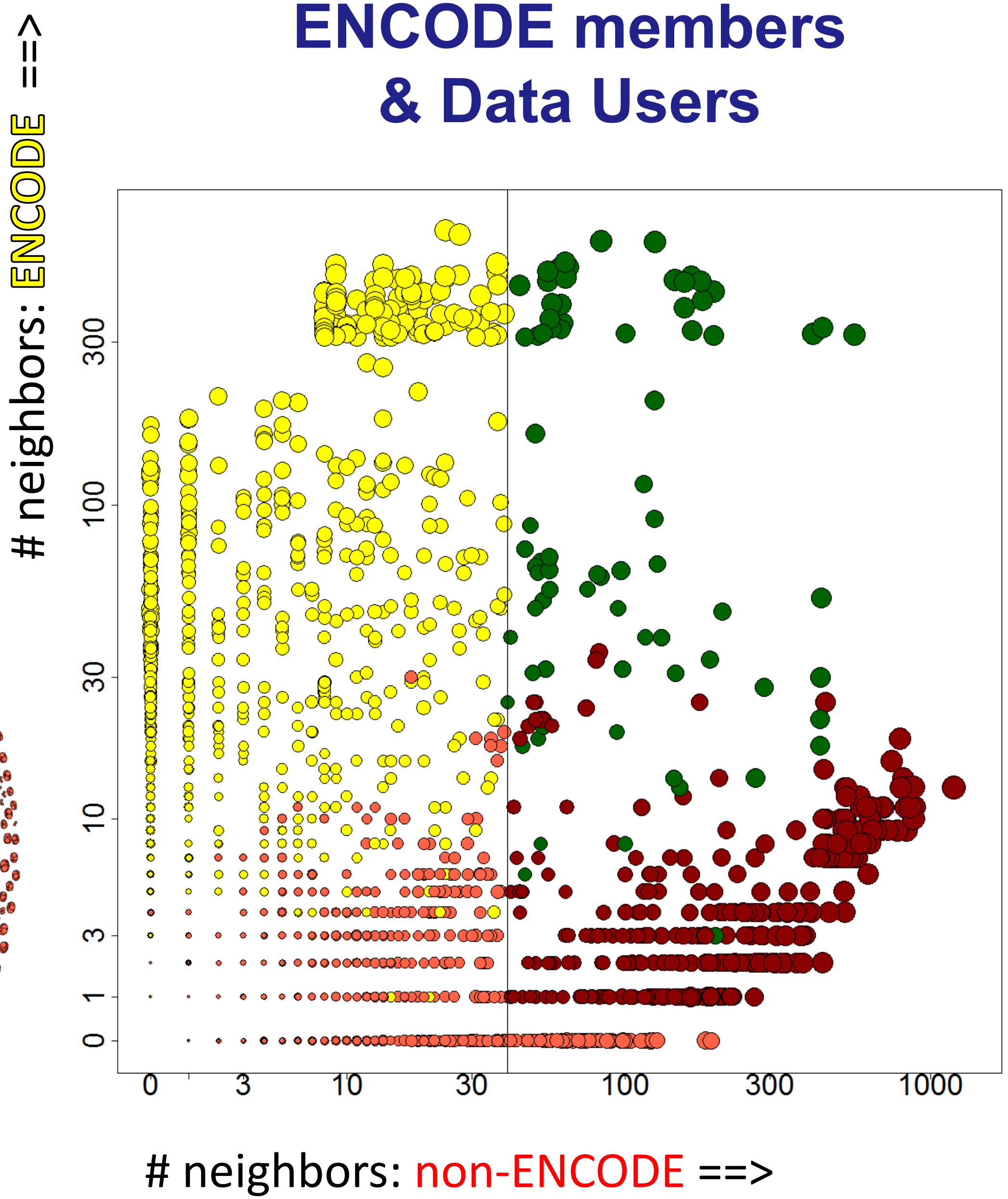
● non-member

● ENCODE member broker

● non-member broker

— co-authorship

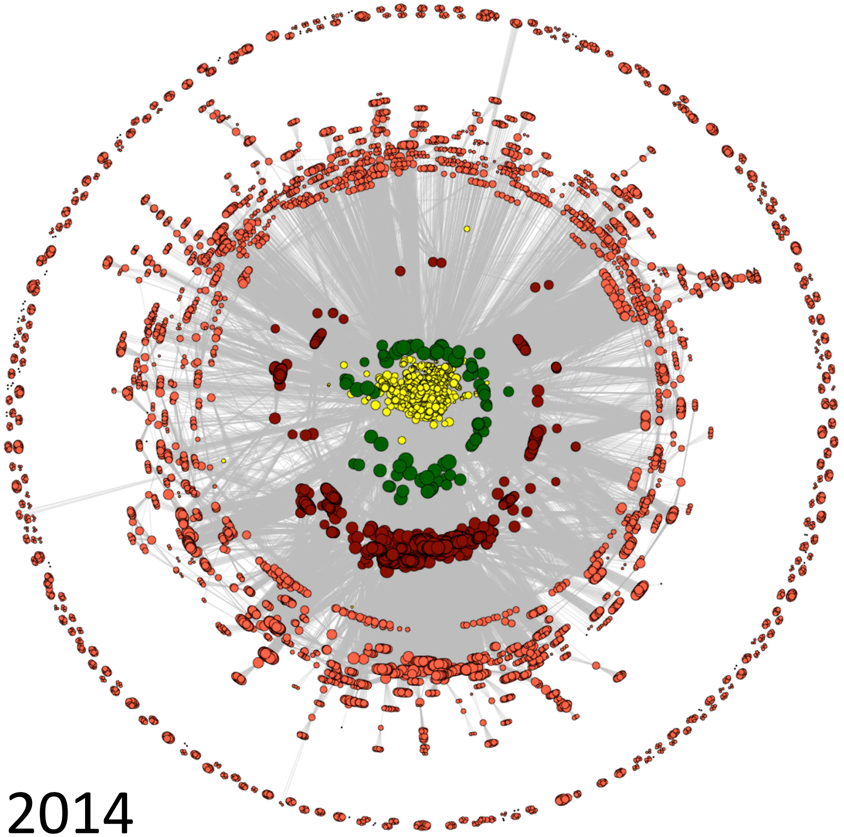




[Wang et al., TIG ('16)]

Co-authorship Network of ENCODE members & Data Users

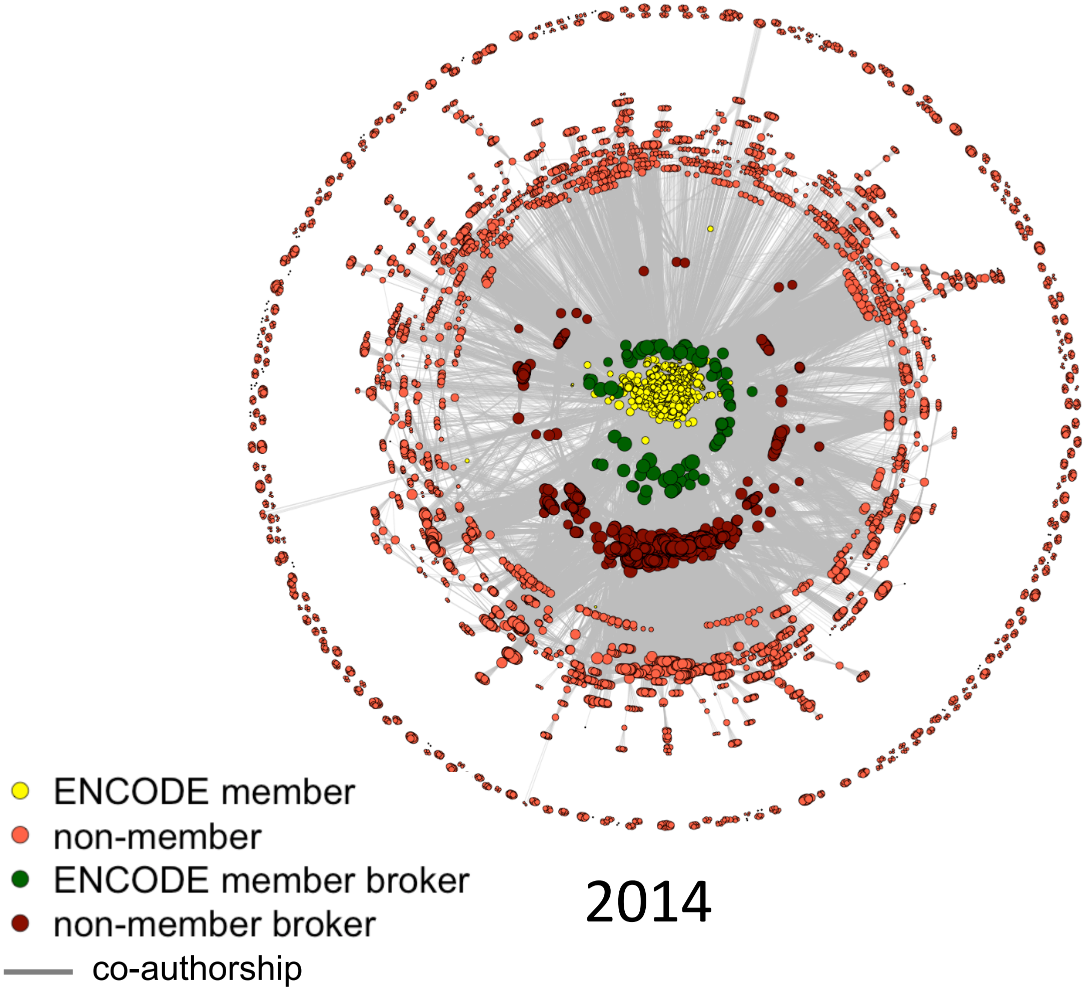
- ● ENCODE member
- ● non-member
- ● ENCODE member broker
- ● non-member broker
- — co-authorship



2014

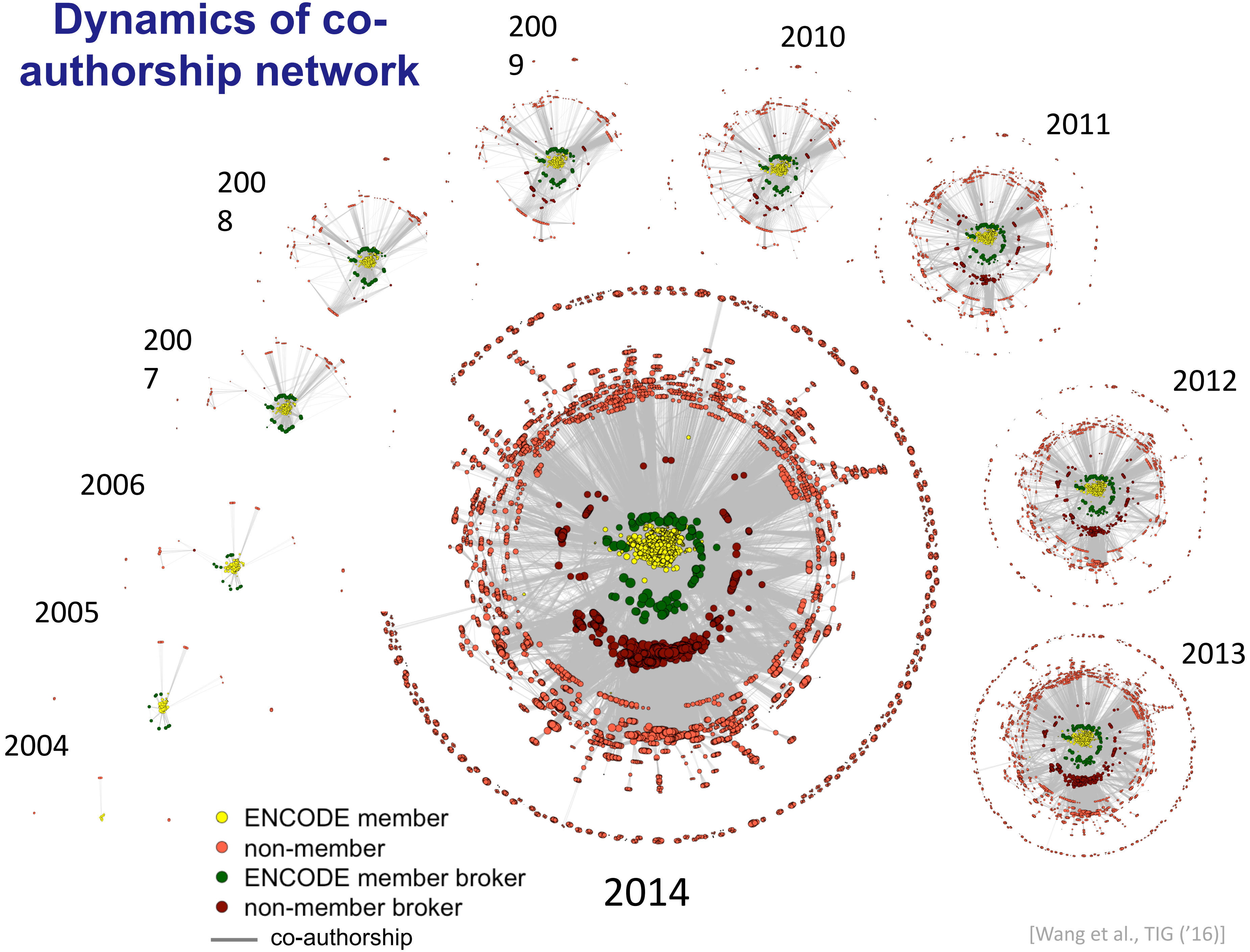
[Wang et al., TIG ('16)]

Dynamics of co-authorship network



[Wang et al., TIG ('16)]

Dynamics of co-authorship network



[Wang et al., TIG ('16)]

Scaling Computation to Keep Pace with Data Generation

- **Sequencing costs are falling exponentially** & the amount of data is increasing exponentially (in accordance with Kryder's law)
- Hence, shift in emphasis to computation. Here, **pipeline processing & data management is keeping pace** with sequencing (roughly), **but downstream analysis work is increasing even faster**
- Seq. data **analysis is diffusing out of genomics** into other disciples (eg ecology). Often this process is **mediated by key connector individuals**

- P **Muir**, S Li, S Lou, D Wang, DJ Spakowicz, L Salichos, J Zhang, F Isaacs, J **Rozowsky** ('16) GenomeBiology
- D **Wang**, KK Yan, J **Rozowsky**, E Pan ('16) TIG

Extra

Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University, 2016.
  - Please read statement at www.**gersteinlab.org/misc/permissions.html** .
  - Feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org). Paper references in the talk are mostly from Papers.GersteinLab.org.

- PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see streams.gerstein.info .