Thoughts on Publication Rollout Structure for the ENCODE Project

M Gerstein

(Publication analysis done by
D Wang, KK Yan, J Rozowsky, E Pan)

The Human Genome Project

ENCODE Pilot

ENCODE Production

Comparative ENCODE

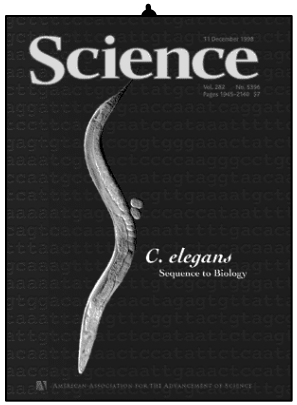
Epigenome Roadmap

2000

2005

2010

2015



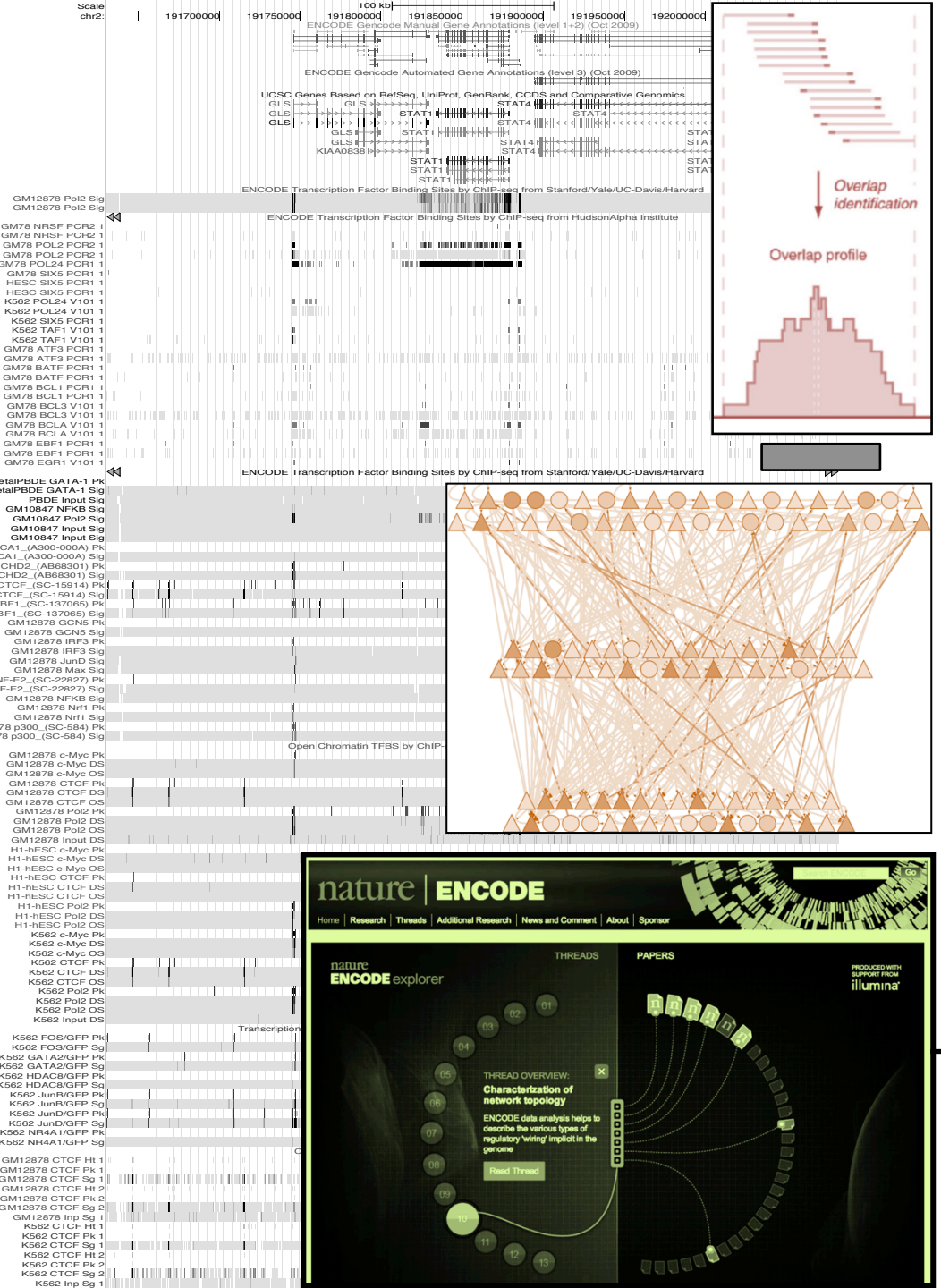
Worm Genome

modENCODE

1000 Genomes Pilot

1000 Genomes Phase 3

GTEx



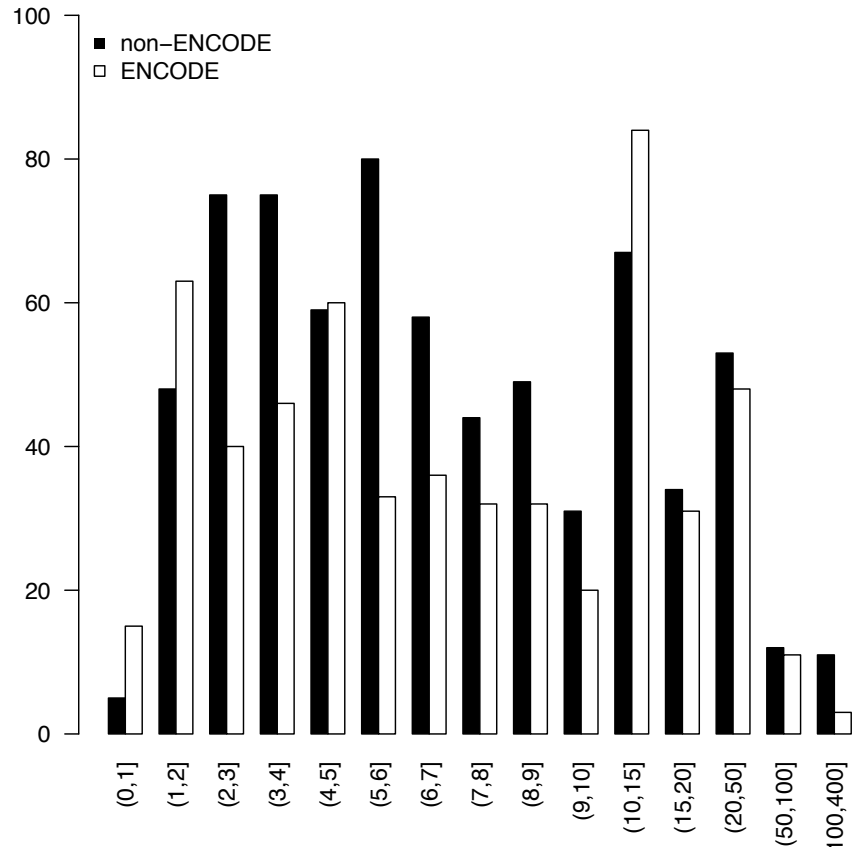
Overall ENCODE presentation as a **Structured Hierarchy** of huge amount of Genomic Data

- **Raw data** (reads) at the bottom
- Progressive Processed **Summaries**
  - Signals (e.g. how much DNA is bound by TFs)
  - Site locations
  - Reg. networks, chromatin states & stat. models
- **Code** & VMs + Result Stats
- **Many linked publications** are near top, documenting everything & forming metadata
- **Abstract of Consortium** paper sits at pinnacle

Reader/User goes Top-Down

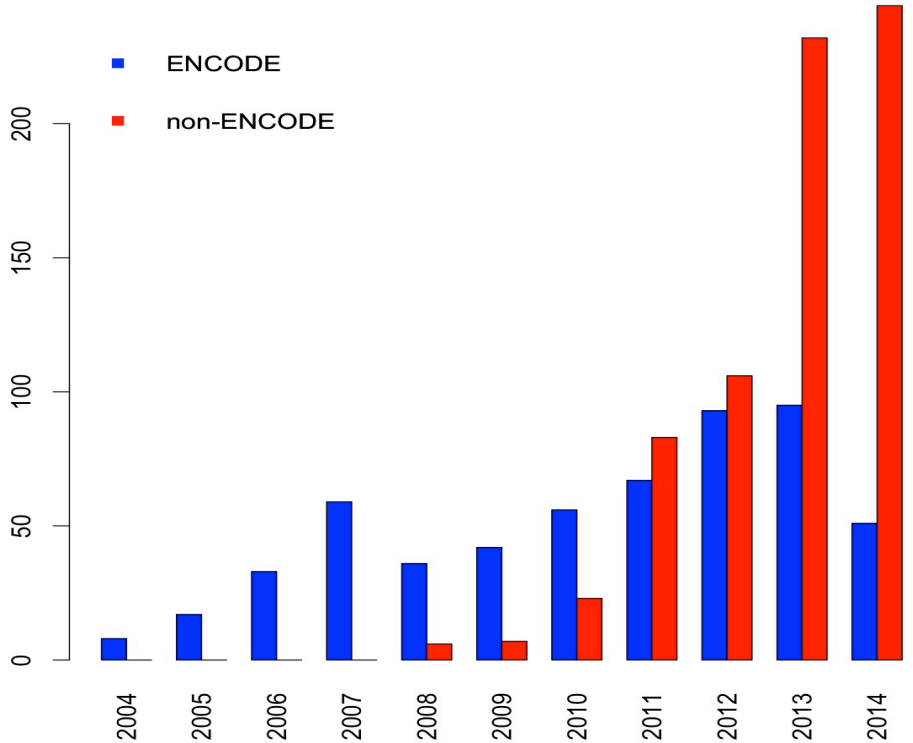
Creator/Author builds Bottom-up

[PLOS CB 4:e1000158;
Nature.com/encode + *Nature* 489: 208]



Authors

<= # Papers

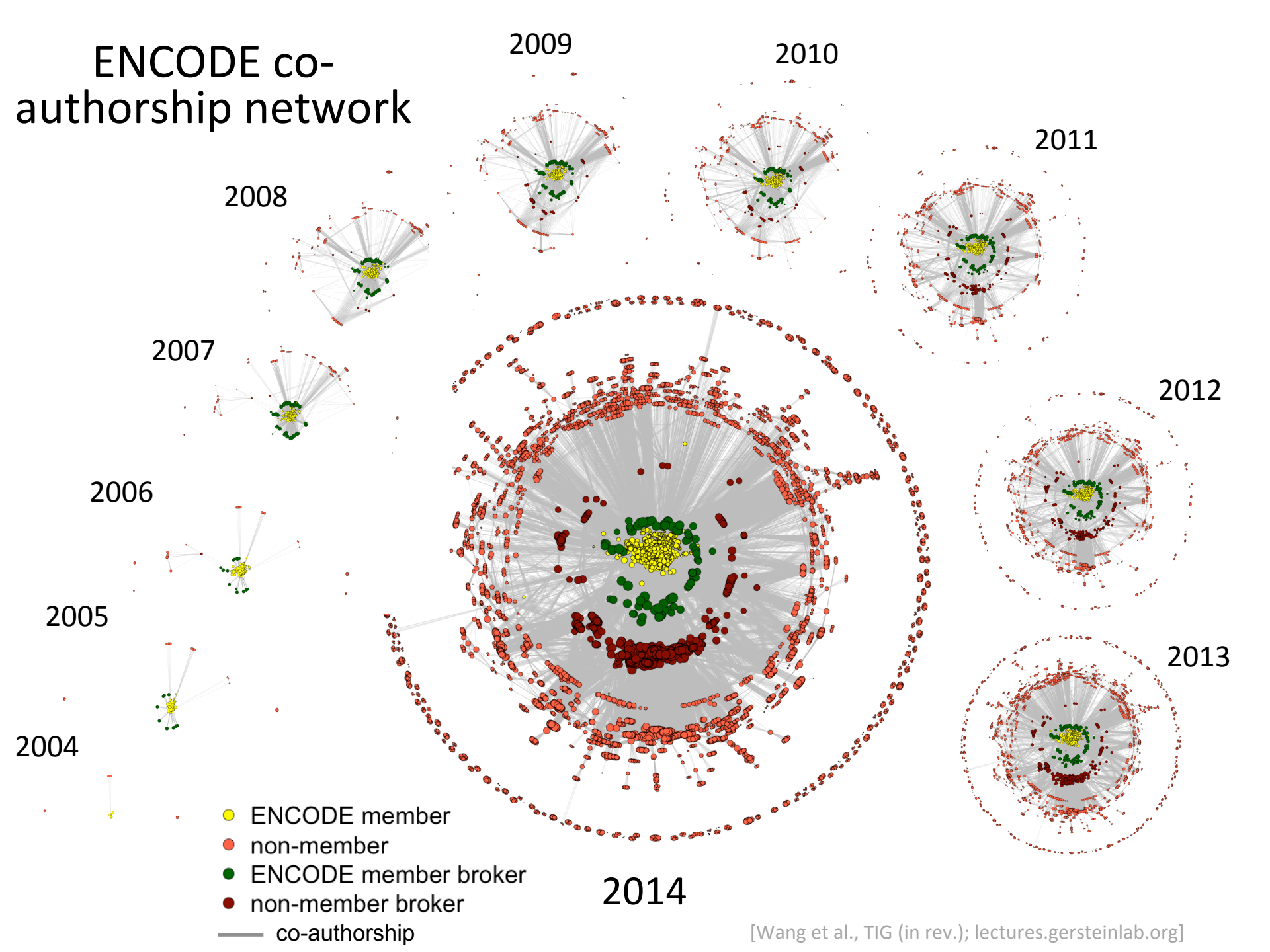


Yr. (‘04 to ‘14)

Papers authored by ENCODE consortium members vs. those that use ENCODE data but were not funded by ENCODE

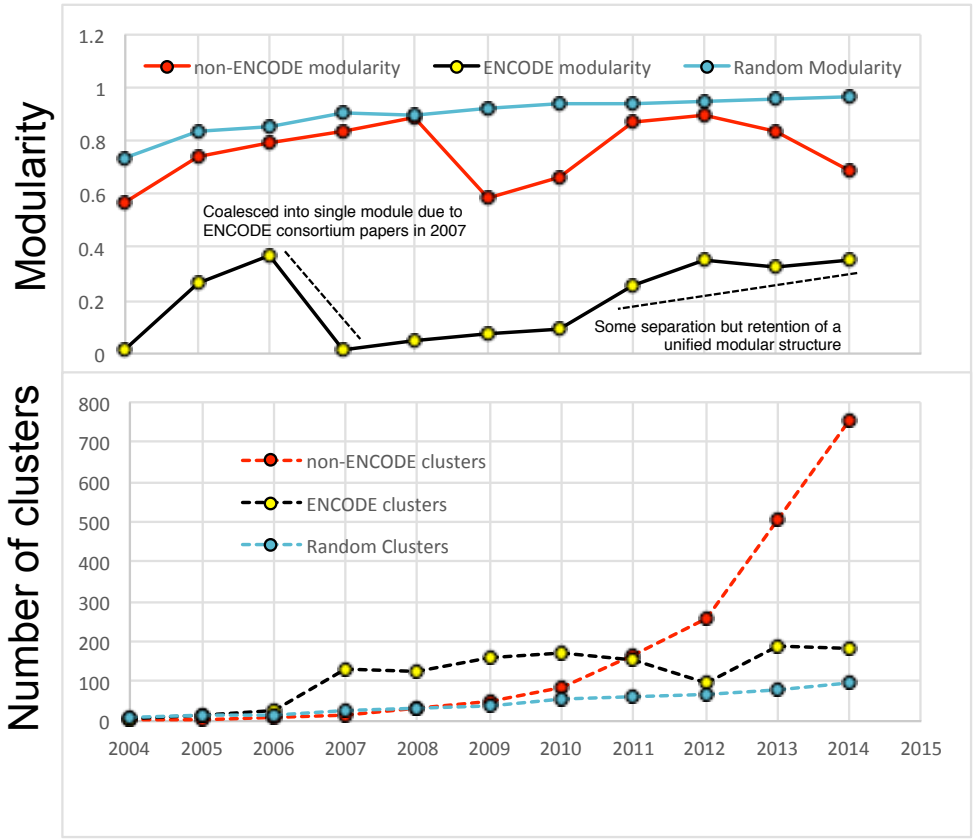
[Wang et al., TIG (in rev.); lectures.gersteinlab.org]

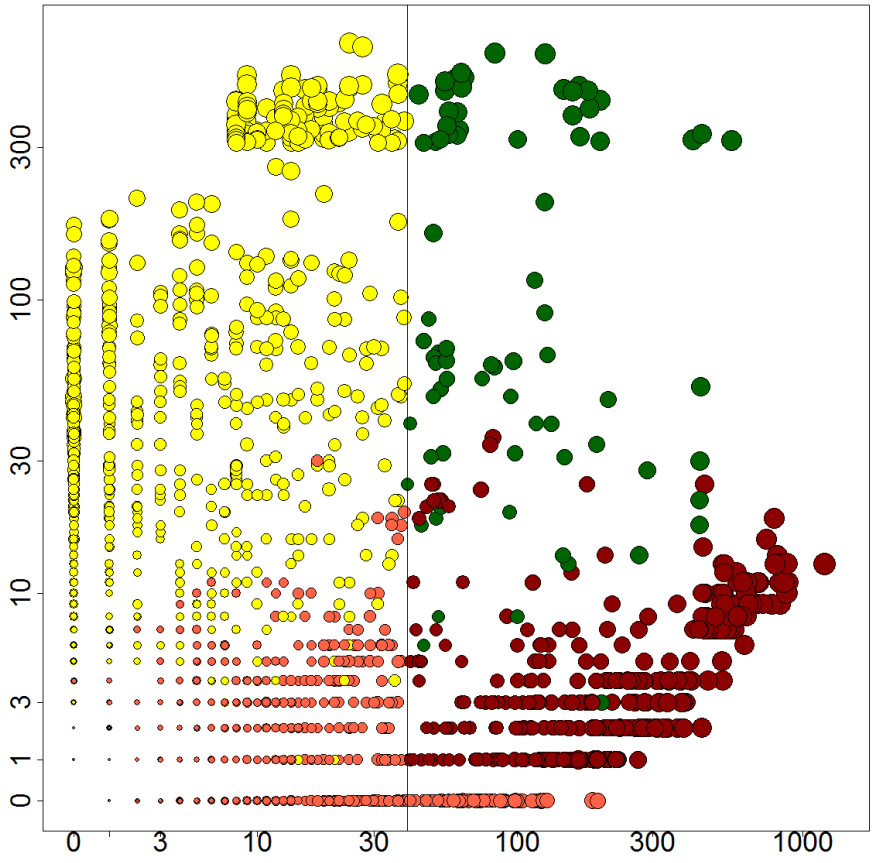
ENCODE co-authorship network



[Wang et al., TIG (in rev.); lectures.gersteinlab.org]

Network statistics highlight change in modularity with consortium rollouts (L) & importance of broker role (R)





[Wang et al., TIG (in rev.); lectures.gersteinlab.org]