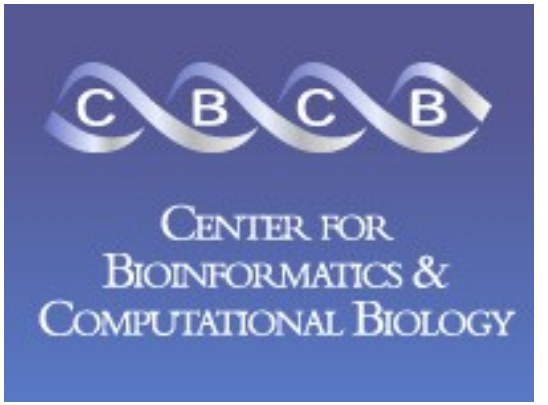
Understanding Mixtures of Organisms

A computational perspective

Mihai Pop
University of Maryland, College Park






(C) ^photo credits: Briana Lindsay, Amy Brown

DIARRHEAL DISEASE KILLS 800,000 CHILDREN EACH YEAR

(more than HIV, malaria, and measles combined)

APPROXIMATELY **1 IN 5**

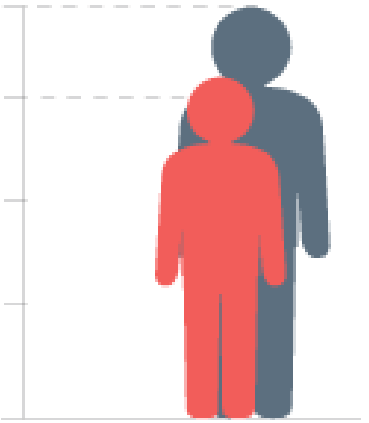
CHILDREN

UNDER THE AGE OF TWO SUFFER FROM AN EPISODE OF MODERATE TO SEVERE (MSD) DIARRHEA EACH YEAR.

THESE CHILDREN WERE

8.5 TIMES MORE LIKELY TO DIE

WITHIN TWO MONTHS OF HAVING DIARRHEAL DISEASE



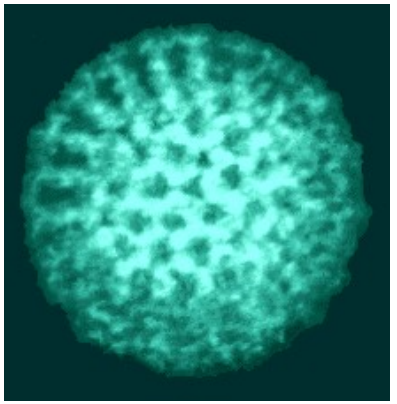
GROWTH IS LIKELY TO BE STUNTED

COMPARED TO PEERS OVER THE SAME TWO MONTH PERIOD

GEMS Study. The Lancet. May 2013.

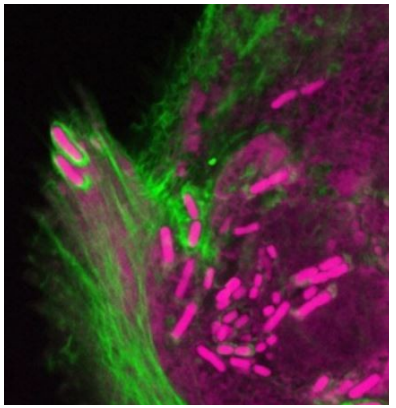
DIARRHEA CAN BE CAUSED BY:

VIRAL



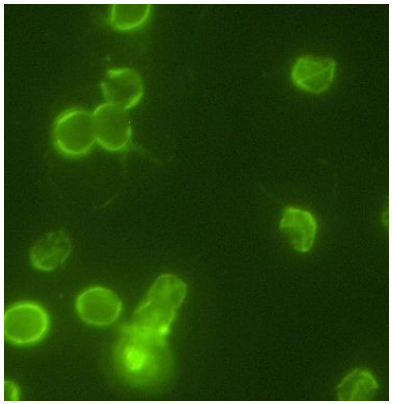
Rotavirus, Norovirus GI, Norovirus GII, Sapovirus, Astrovirus

BACTERIAL



Shigella, Salmonella, Campylobacter, Aeromonas, Vibrio cholerae, Diarrheagenic E.coli

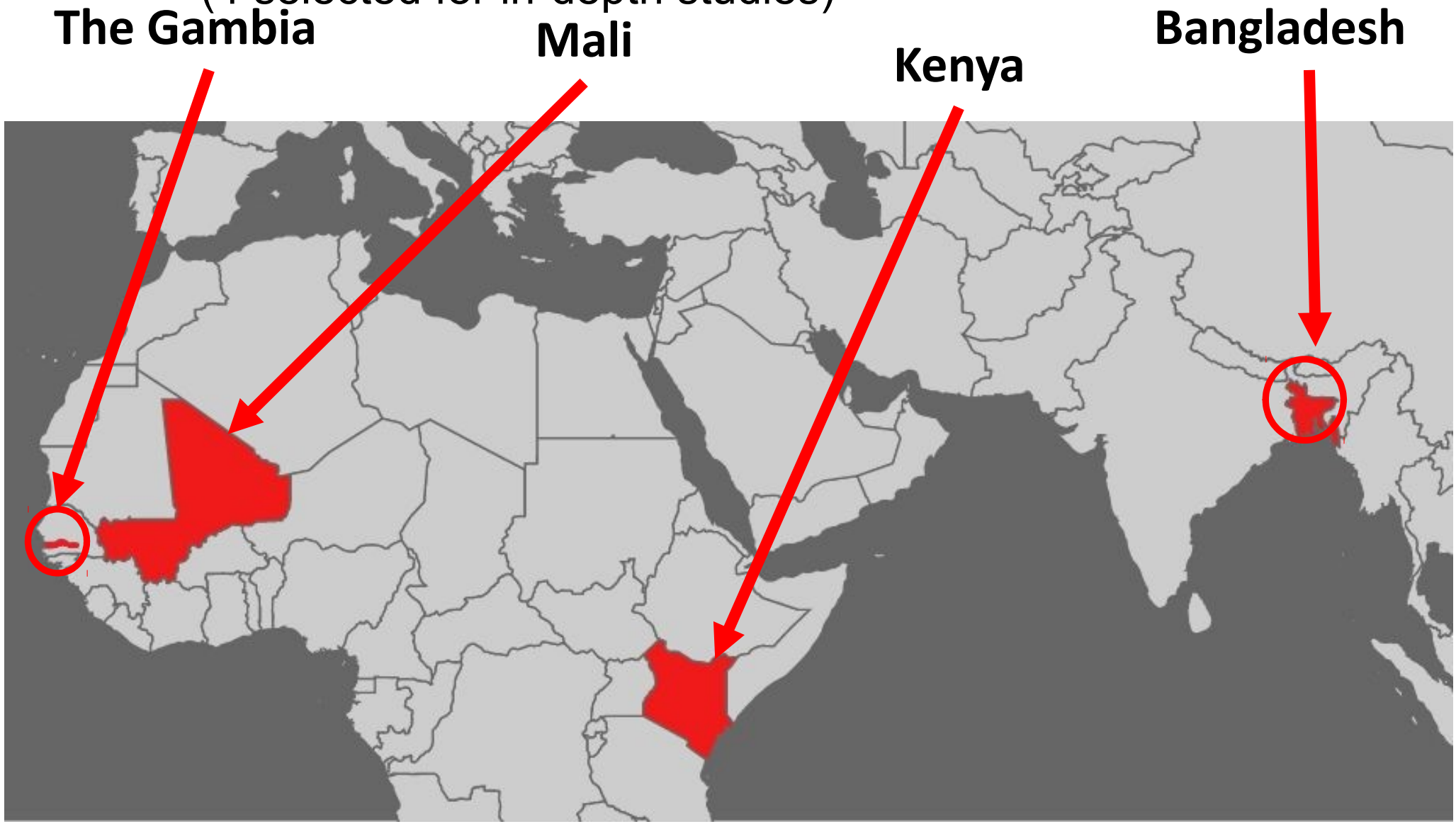
EUKARYOTIC



Cryptosporidium, Giardia, Entamoeba histolytica

GEMS

22,000 children < 5 years of age from 7 countries
(4 selected for in-depth studies)



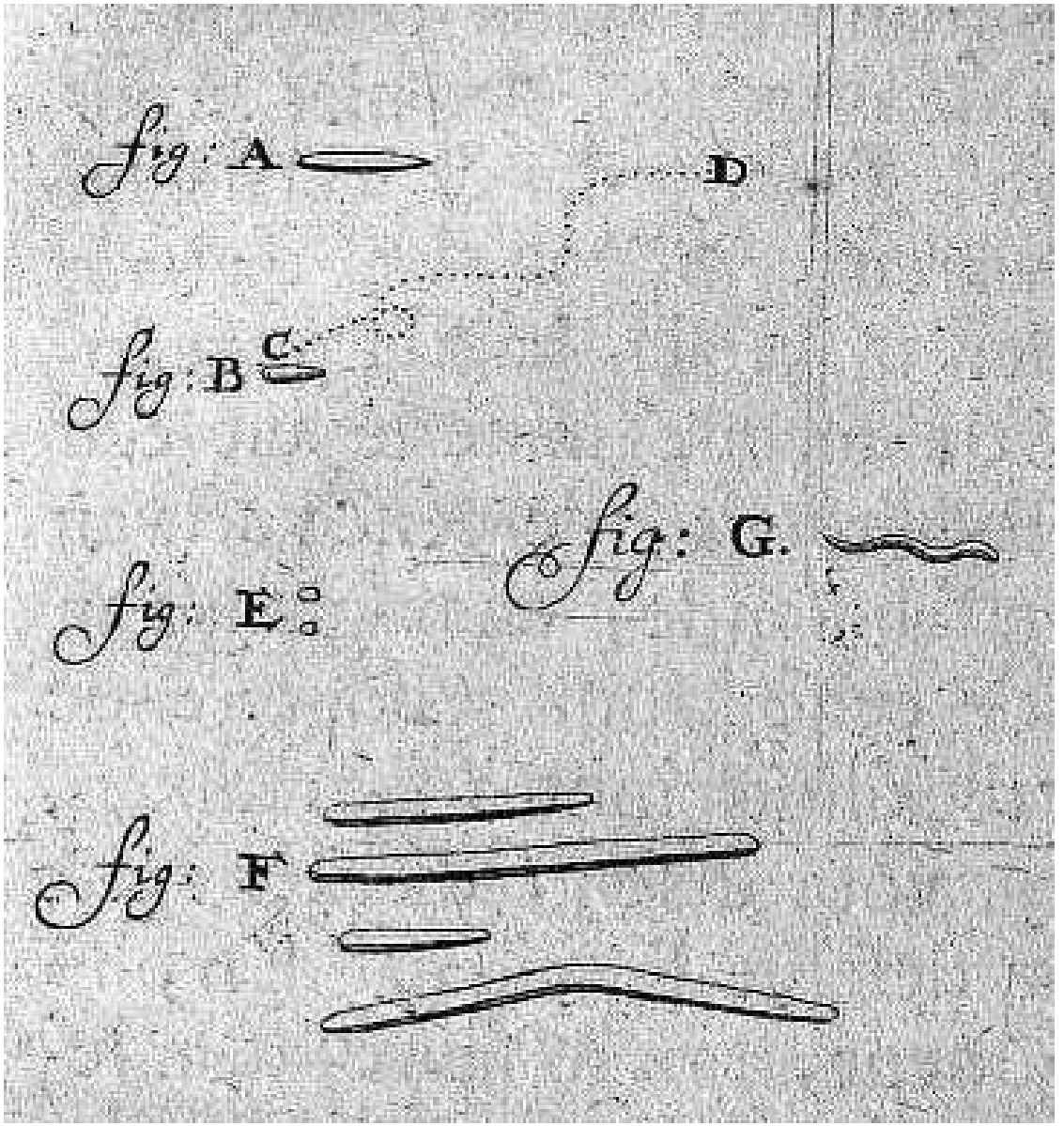
Over half of all cases could not be attributed to any known pathogen

Common core standards: 1st grade

Compare by identifying similarities and differences

Sort and classify into categories

17th century biology

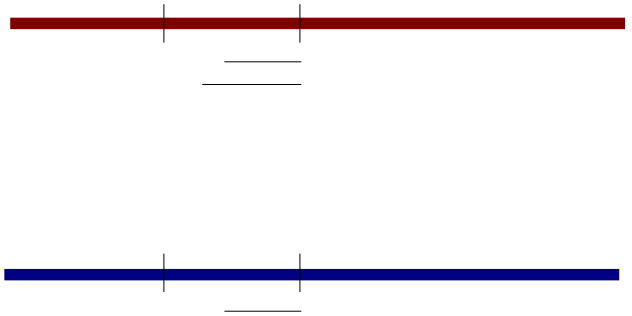


21st century biology

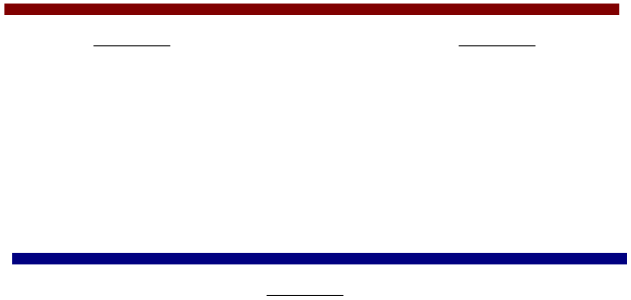
```
>F4BT0V001CZSIM rank=0000138 x=1110.0 y=2700.0 leng
ACTGCTCTCATGCTGCCTCCCGTAGGAGTGCCTCCCTGAGCCAGGATCAAA
>F4BT0V001BBJQS rank=0000155 x=424.0 y=1826.0 lengt
ACTGACTGCATGCTGCCTCCCGTAGGAGTGCCTCCCTGCGCCATCAA
>F4BT0V001EDG35 rank=0000182 x=1676.0 y=2387.0 leng
ACTGACTGCATGCTGCCTCCCGTAGGAGTCGCCGTCCTCGACNC
>F4BT0V001D2HQQ rank=0000196 x=1551.0 y=1984.0 leng
ACTGACTGCATGCTGCCTCCCGTAGGAGTGCCGTCCCTCGAC
>F4BT0V001CM392 rank=0000206 x=966.0 y=1240.0 lengt
AANCAGCTCTCATGCTCGCCCTGACTTGGCATGTGTTAAGCCTGTAGGCTA
>F4BT0V001EIMFX rank=0000250 x=1735.0 y=907.0 length=
ACTGACTGCATGCTGCCTCCCGTAGGAGTGTCGCGCCATCAGACTG
>F4BT0V001ENDKR rank=0000262 x=1789.0 y=1513.0 length=56
GACACTGTCATGCTGCCTCCCGTAGGAGTGCCTCCCTGAGCCAGGATCAAACTCTG
>F4BT0V001D91MI rank=0000288 x=1637.0 y=2088.0 length=56
ACTGCTCTCATGCTGCCTCCCGTAGGAGTGCCTCCCTGAGCCAGGATCAAACTCTG
>F4BT0V001D0Y5G rank=0000341 x=1534.0 y=866.0 length=75
GTCTGTGACATGCTGCCTCCCGTAGGAGTCTACACAAGTTGTGGCCCAGAACCACTGAGCCAGGATCAAACTCTG
>F4BT0V001EMLE1 rank=0000365 x=1780.0 y=1883.0 length=84
ACTGACTGCATGCTGCCTCCCGTAGGAGTGCCTCCCTGCGCCATCAATGCTGCATGCTGCTCCCTGAGCCAGGATCAAACTCTG
```

Same vs. different, 16S vs WGS?

16S



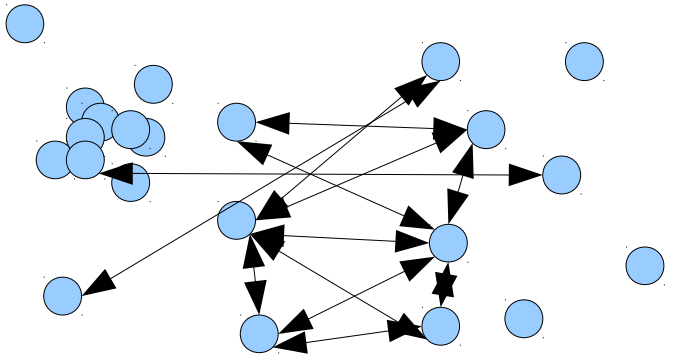
WGS



WGS

meta-genome assembly

Why is clustering difficult?



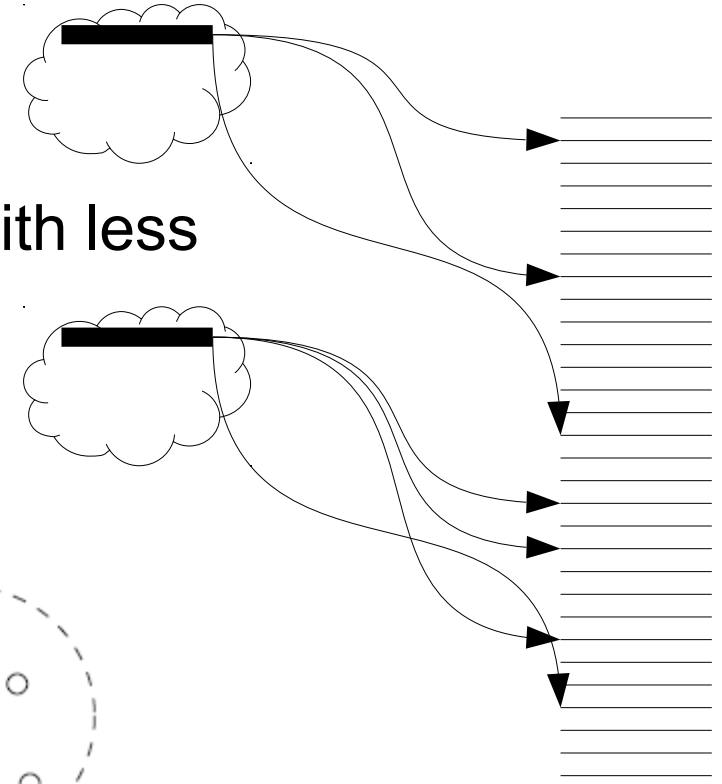
Must compare all versus all
(at least)

3,000,000 X 3,000,000 = 9 X 10^{12} (9 trillion combinations)

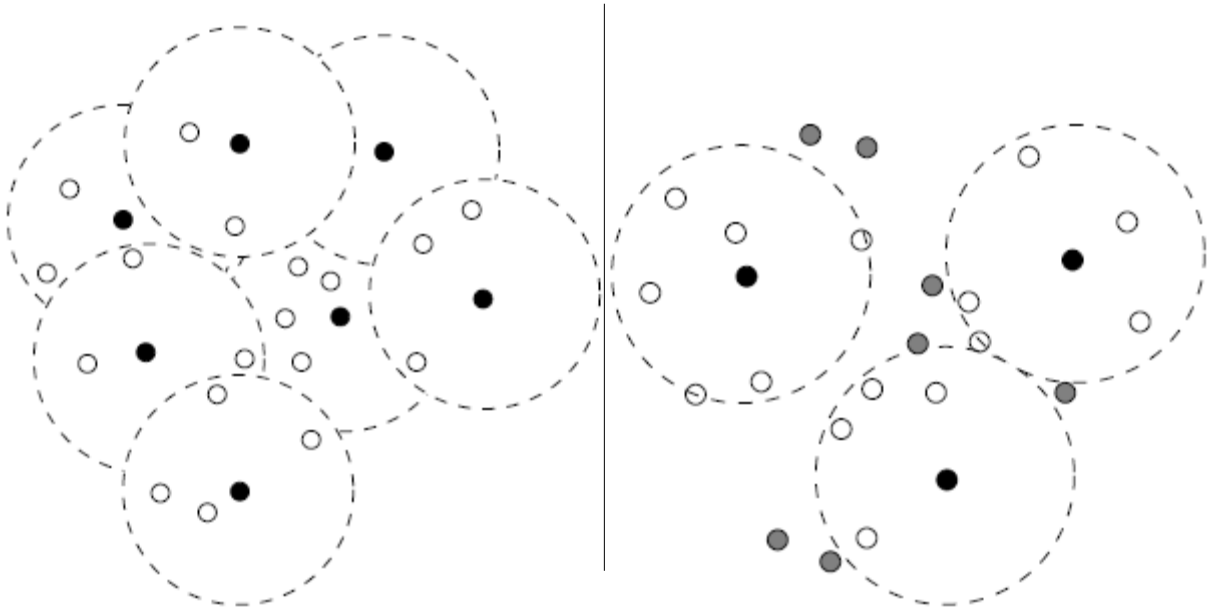
Heuristic clustering

DNAclust

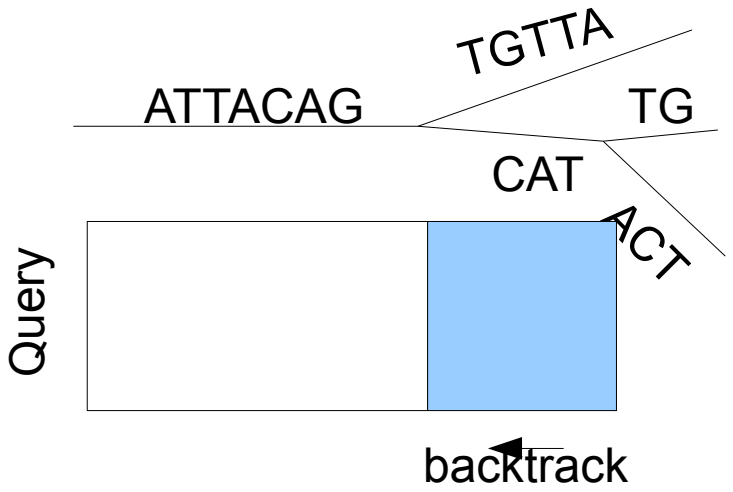
- pick one sequence
- search all other sequences that match it
- repeat



- smart indexing – find N sequences with less than N comparisons
- provides guarantees



Dynamic programming on trie



With Mohammad Ghodsi

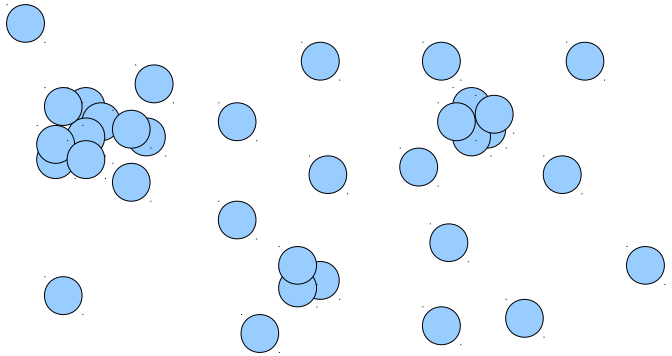
Still too slow - curse of dimensionality

- Iterative clustering – start with most abundant clusters

	1	2	3	4
input seqs	29,129,215	12,523,595	11,567,759	11,191,817
clustered seqs	16,605,620	955,836	375,942	229,282
seqs/sec	236.76	20.68	13.02	7.55

- Curse of dimensionality

$3 \cdot 3^5 \cdot \binom{500}{5} \approx 95 \cdot 10^{12}$ sequences within 5 mismatches in first 500bp and one mismatch in last position

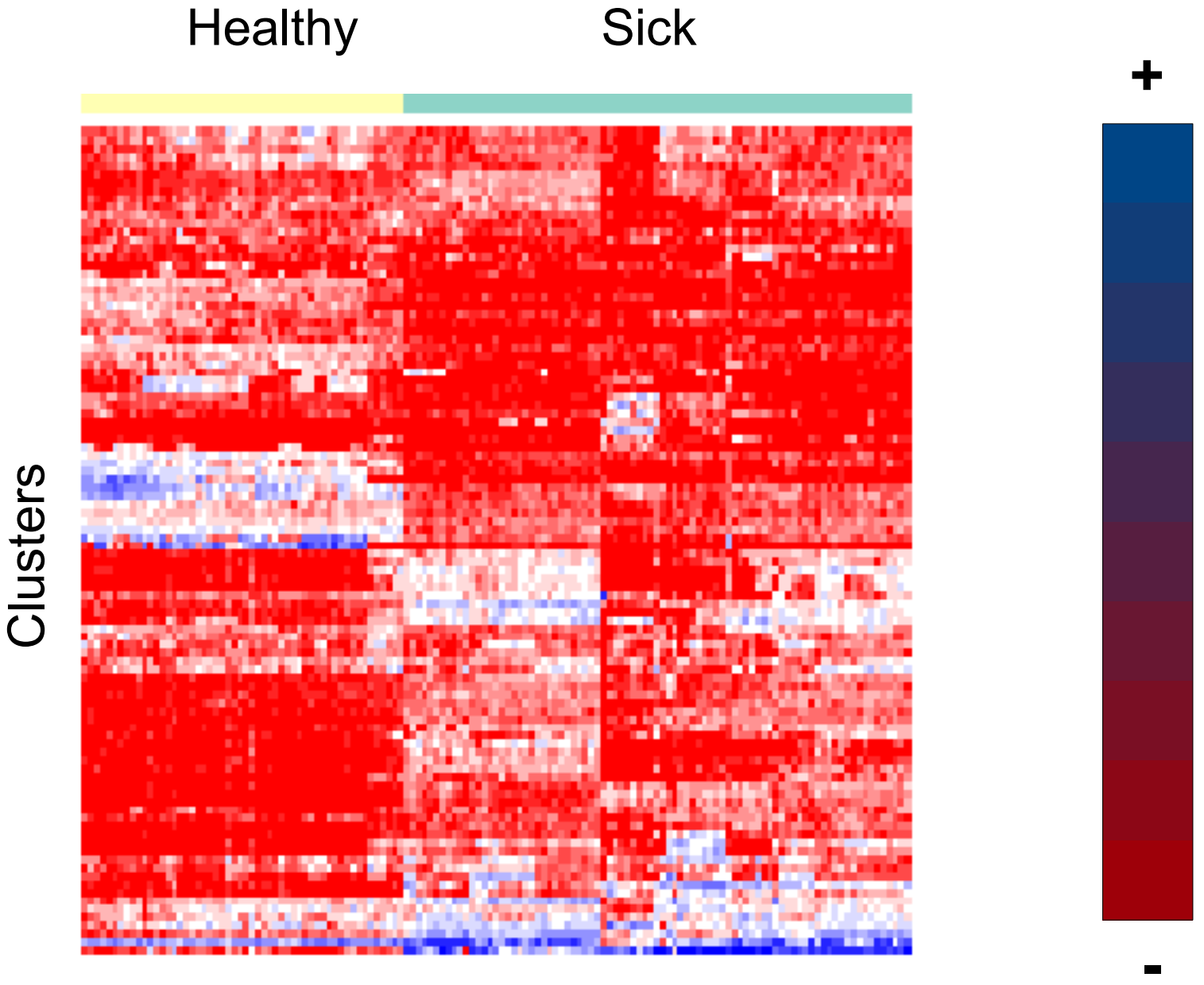


$O(n^2)$ time required to find unclusterable sequences

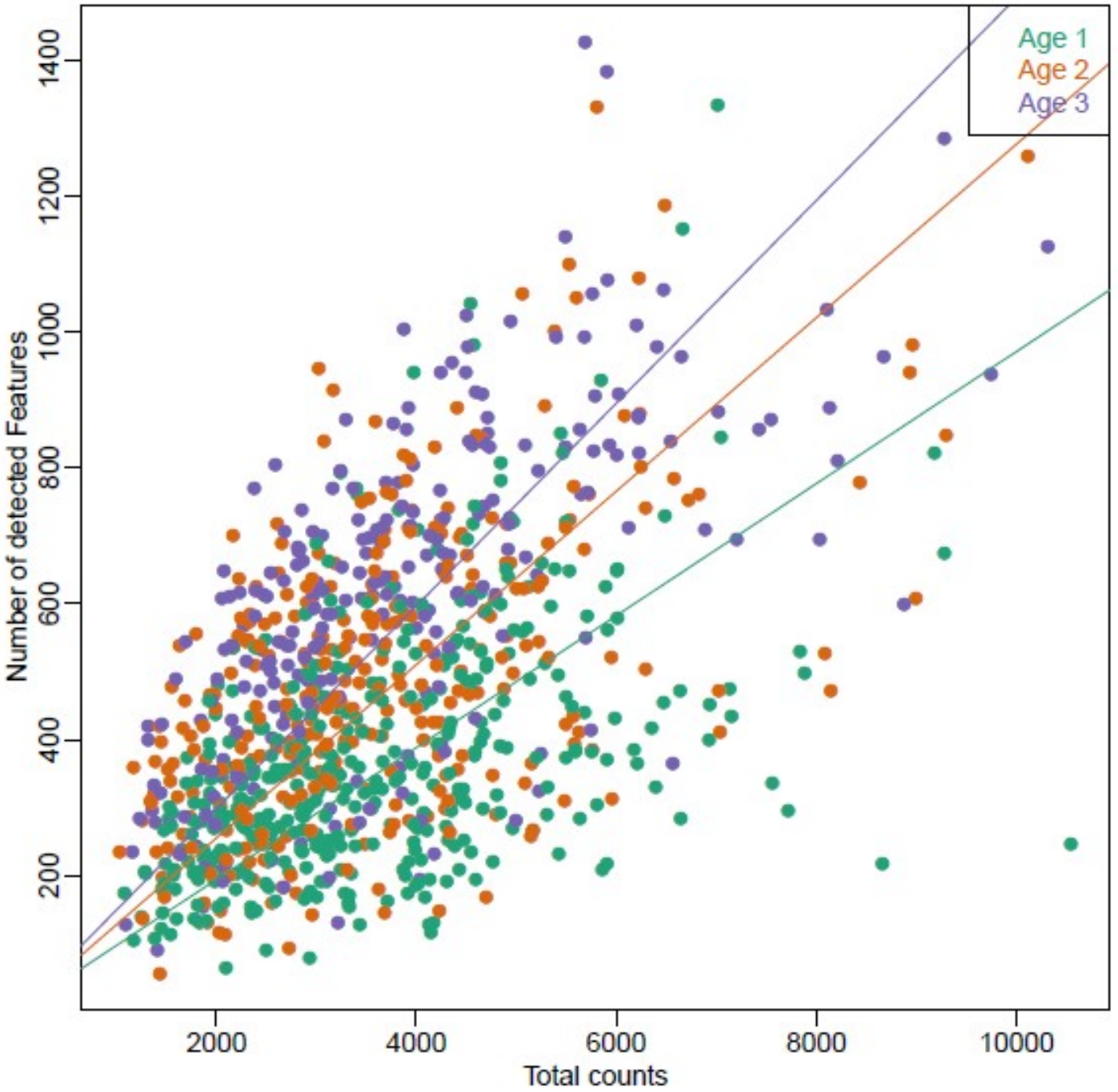
Common core standards: 1st grade

Identify and describe patterns and the relationships within patterns

Abundance for disease associations



Abundance and # of observations dependent on sampling depth



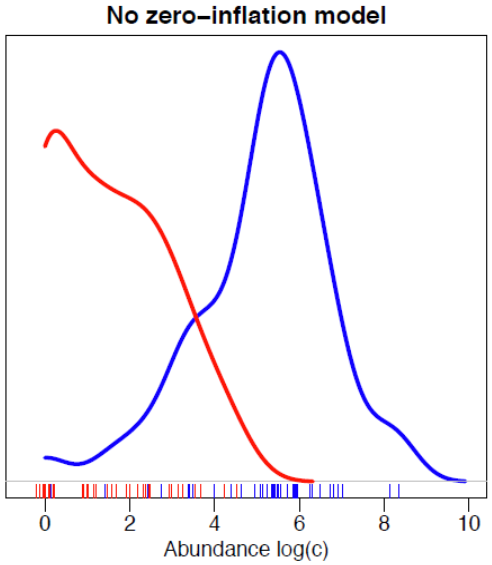
The right statistics also matter

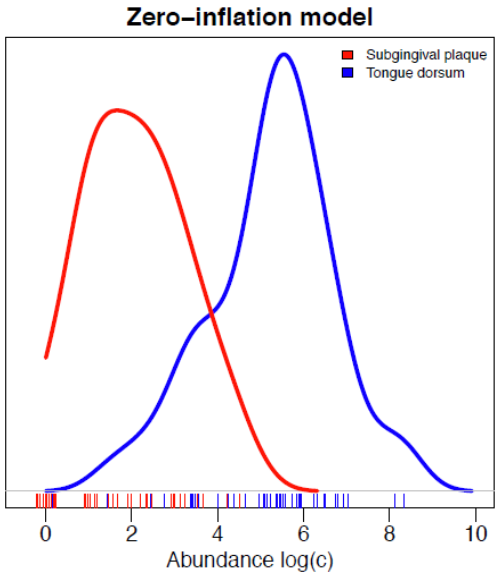
- Note: most counts (# reads in OTU i in sample j) are 0
- Most of the 0s are due to undersampling

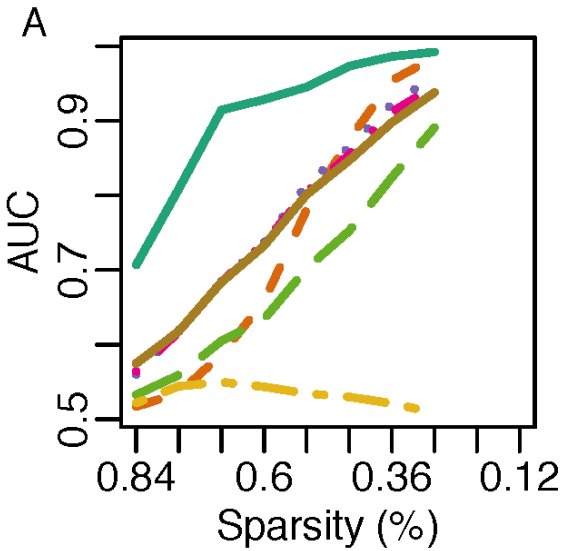
ZIG: Zero-inflated Gaussian model

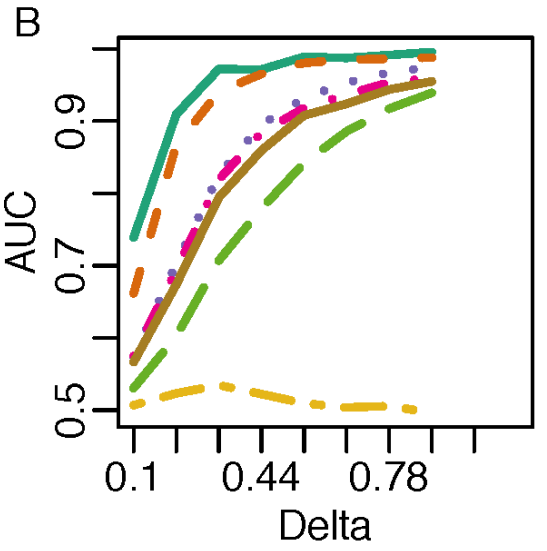
- Mixed model:
  - Model of OTU abundance vs. depth of sequencing in each sample
  - Model of overall abundance in cases/controls for an OTU (the only one in the traditional t-test)

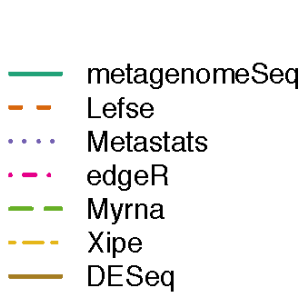
ZIG works as well

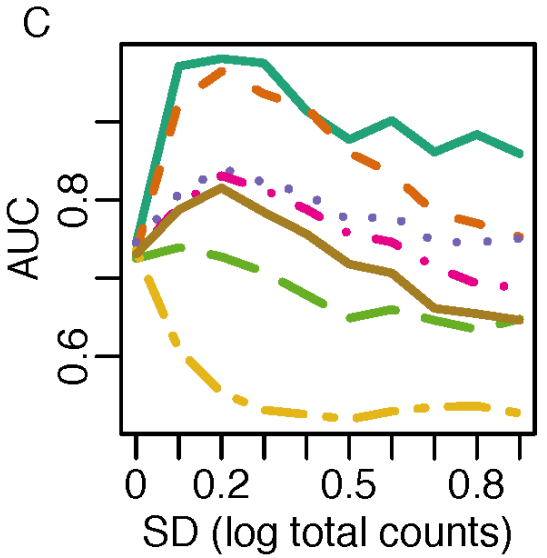


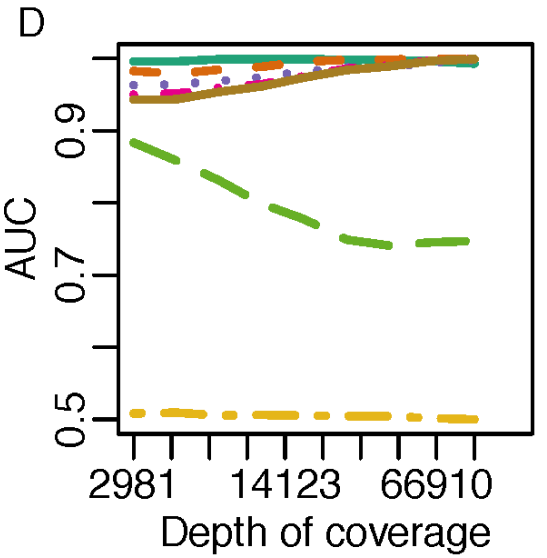










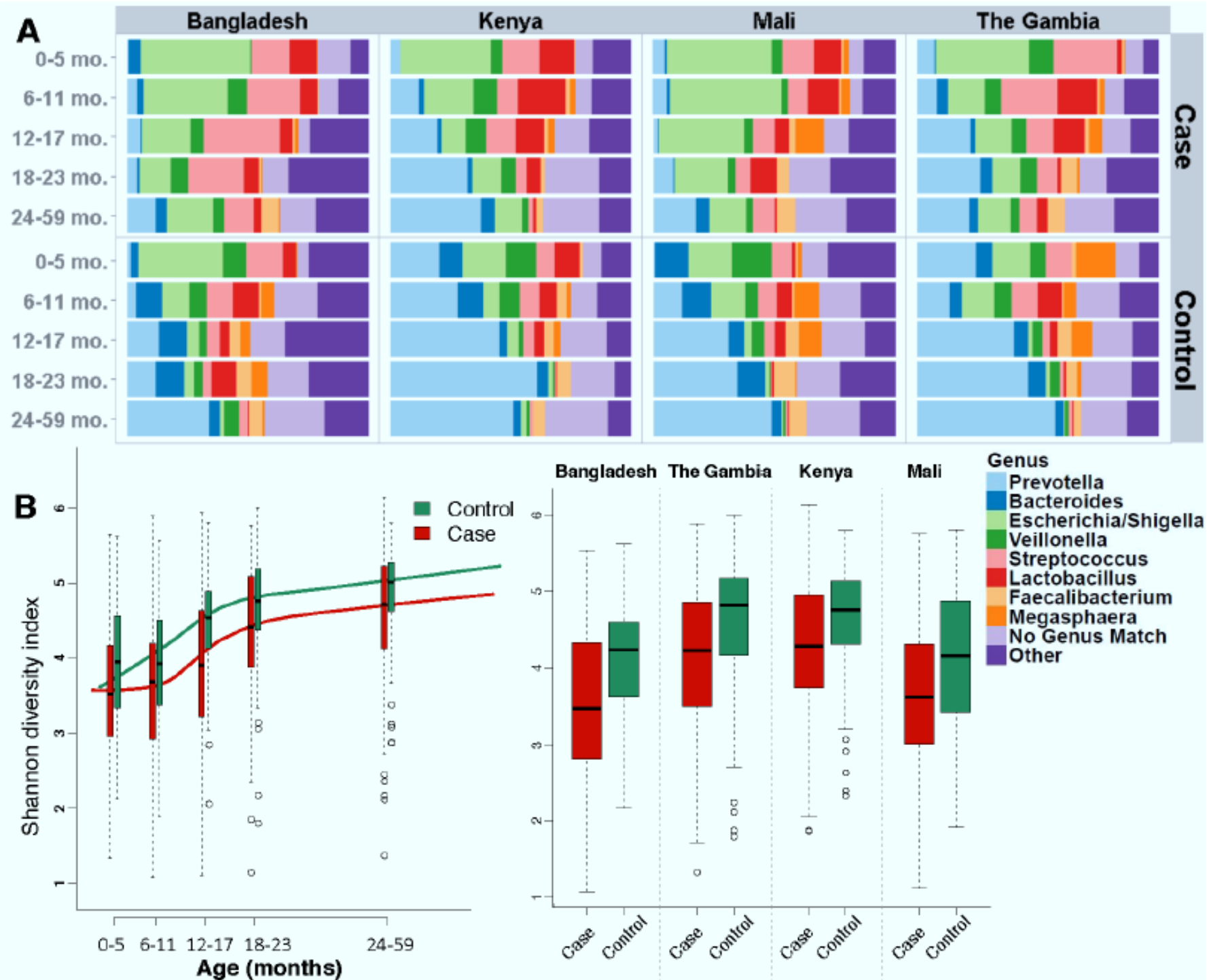


Paulson, J. N., O. C. Stine, H. C. Bravo and M. Pop (2013). Nature Methods **10**(12).

Clustering and association statistics are well studied problems

Key to our success: understanding the (biological) objective function

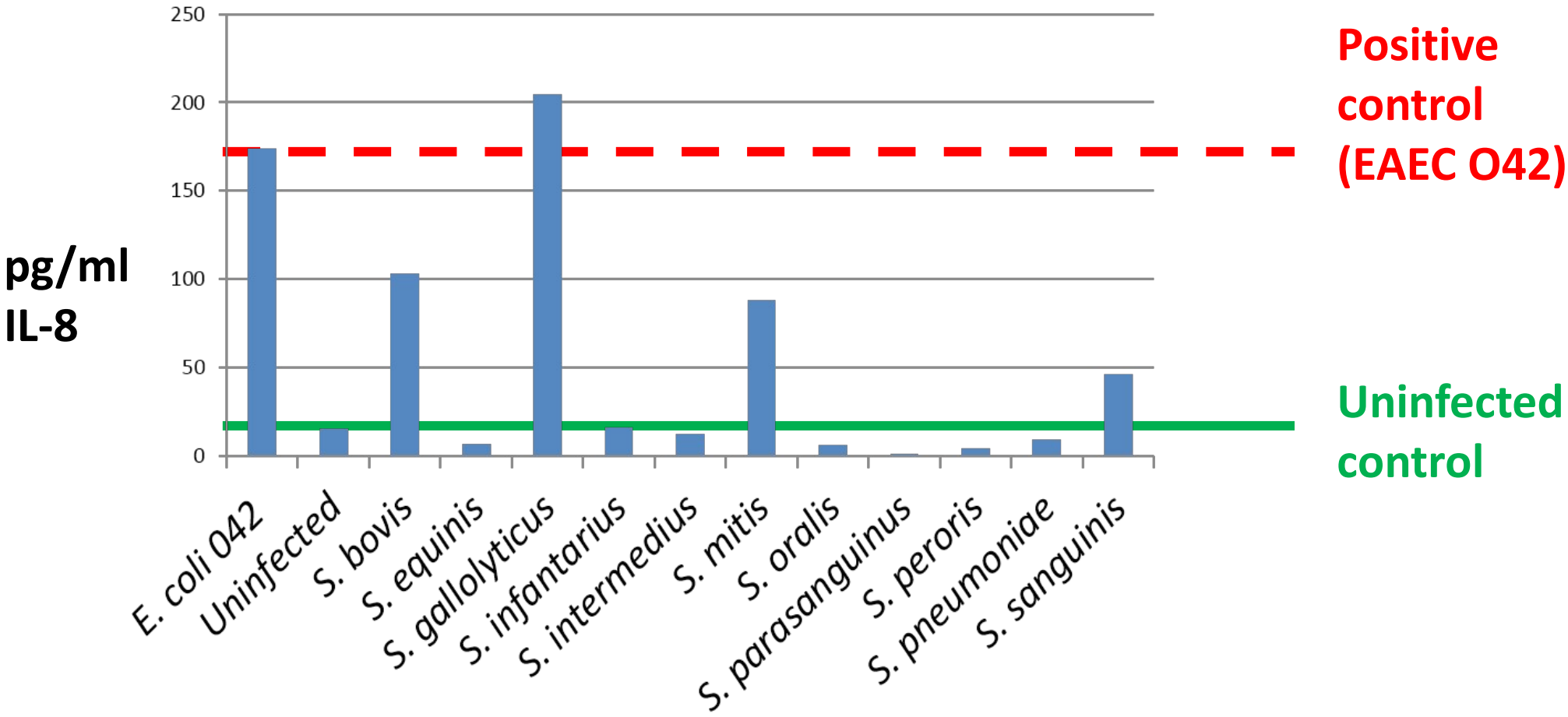
Impact of diarrhea on microbiota



New pathogens

- Streptococci are more common in stools of children with diarrhea than without
  - Regardless of pathogen present
  - No single streptococcal species predominates, but some species are over-represented and others not
  - Chinese CDC reports *S. lutetiensis* associated with diarrhea
  - *S. mutans* recently implicated as injurious to human cells
  - Known correlation between *S. bovis* (*gallolyticus*) and colon cancer

Polarized human colonic (T84) monolayers reveal variation in injurious behavior for streptococcal isolates



Streptococcal isolates incubated with polarized T84 monolayers at 37C for 3 hr; IL-8 release measured by EIA. Results of triplicates

Common core standards: 1st grade

Organize parts to form a new or unique whole

Assembling two cities

it was the best

was the *age of*

best *of times* **it**

it was the age *of times* **it was**

wisdom **it was the**

it was the best

was the best *of*

the worst *of times*

was the worst *of*

was the best *of*

times **it was the**

it was the *age*

times **it was the**

was the *age of*

the best *of times*

worst *of times* **it**

age of wisdom **it**

it was the *age*

of wisdom **it was**

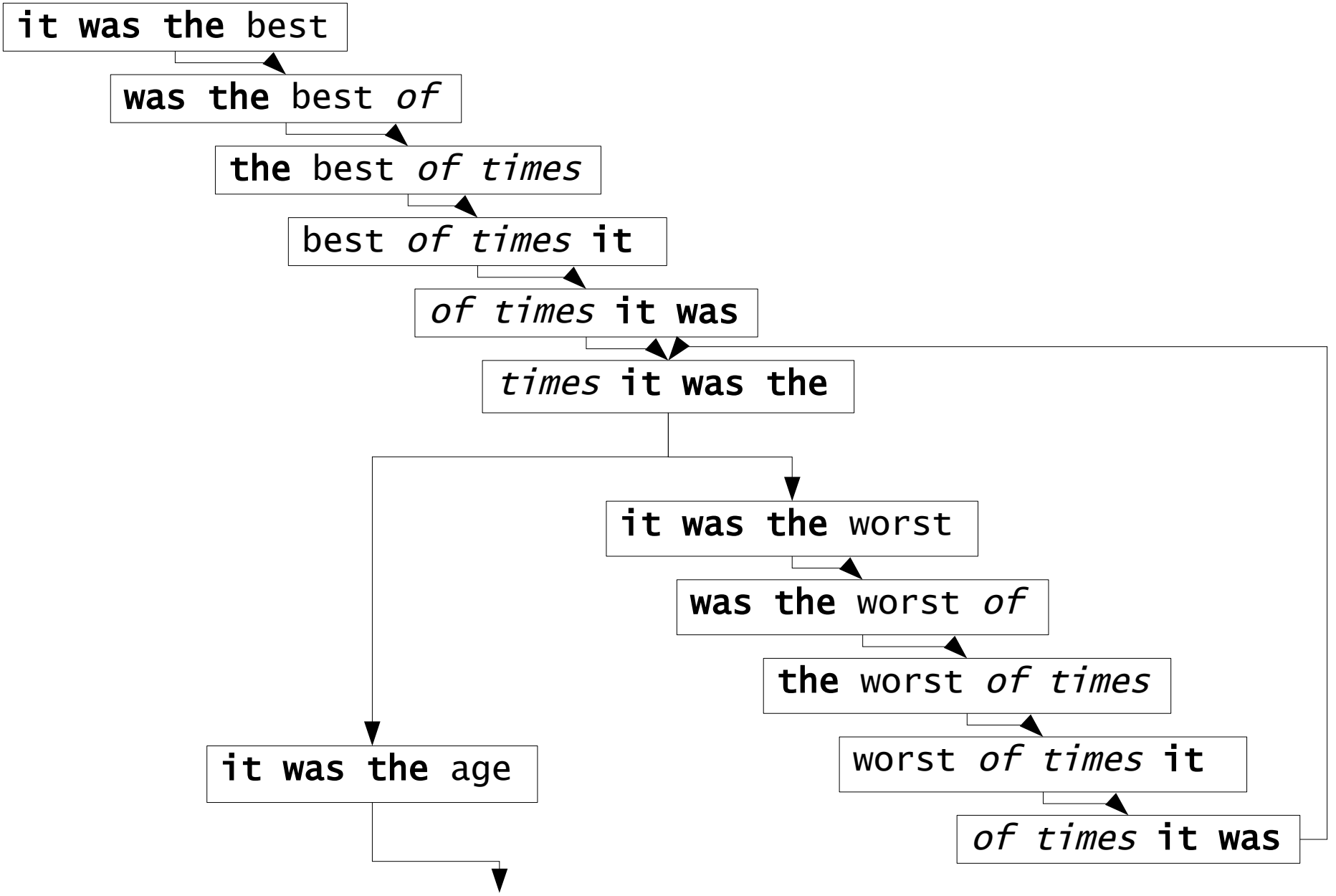
it was the worst

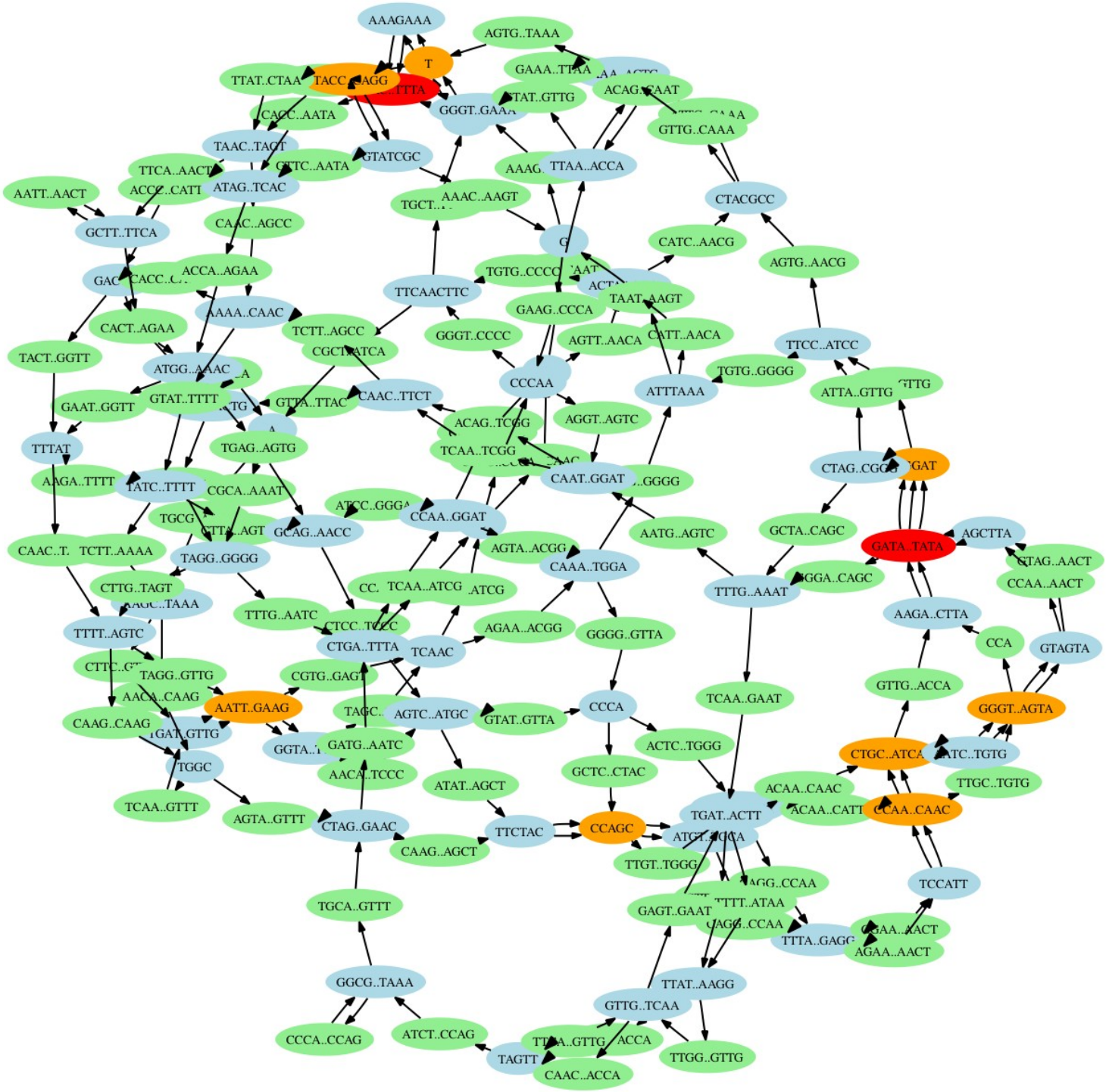
the *age of* wisdom

of times **it was**

the *age of* foolishness

Graph-based approaches



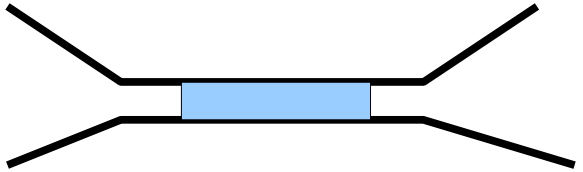


Mycoplasma genitalium, 25 bp reads

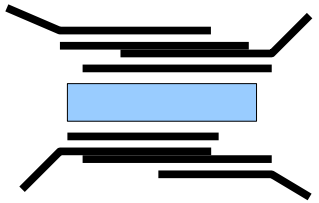
Kingsford et al., BMC Bioinformatics 2010

Read length matters...

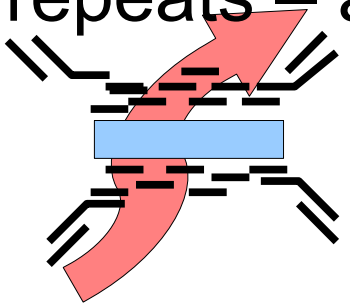
- Reads (much) longer than repeats – assembly trivial



- Reads roughly equal to repeats – assembly computationally difficult (NP-hard)



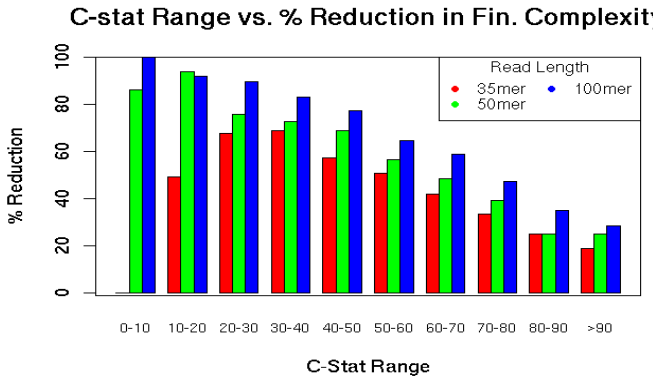
- Reads shorter than repeats – assembly undetermined

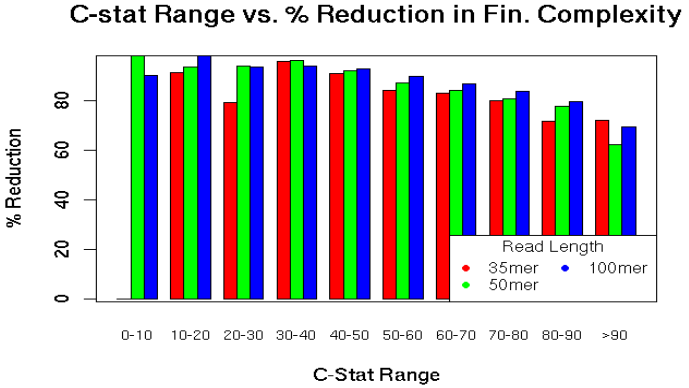


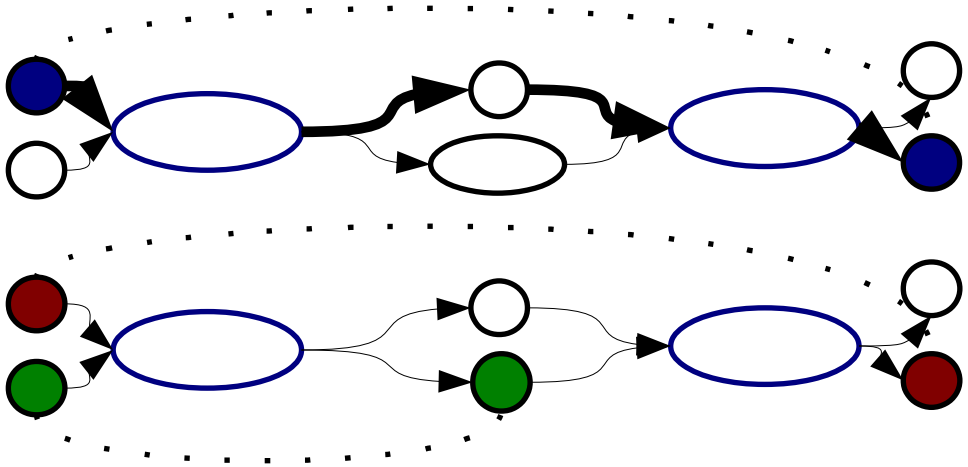
Number of possible reconstructions exponential in # of repeats

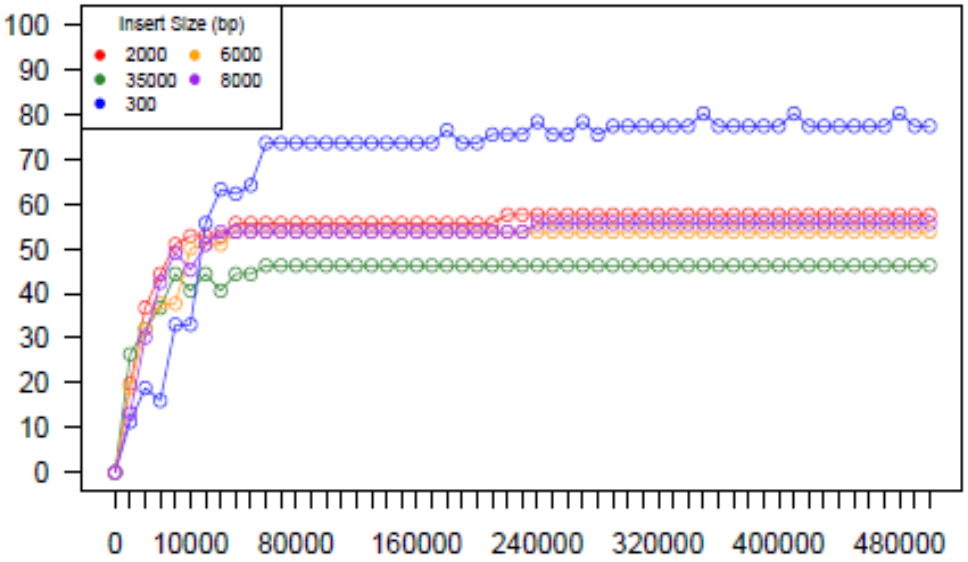
Nagarajan, Pop. J. Comp. Biol. 2009, Kingsford et al., BMC Bioinformatics 2010

Mate-pair information doesn't help much









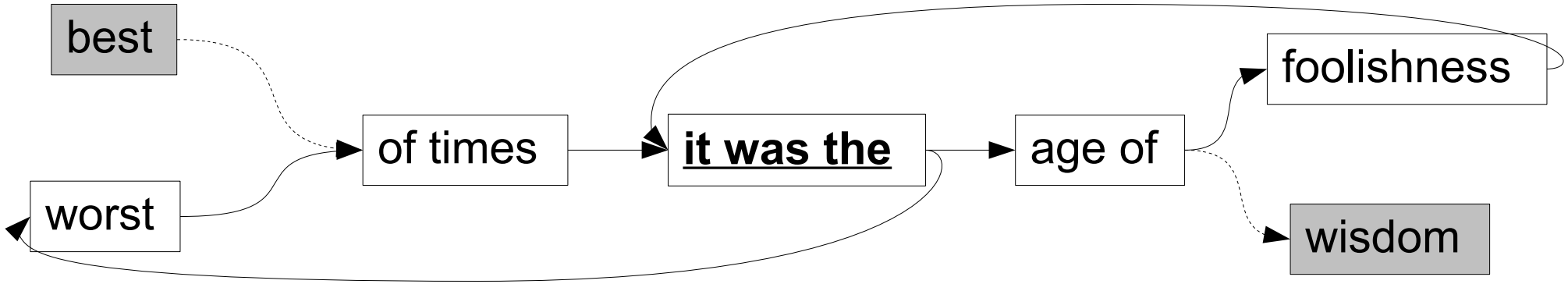
work with Carl Kingsford and Joshua Wetzel

Lack of coverage leads to errors

it was the best *of times* **it was the** worst *of times*
it was the *age of* wisdom **it was the** *age of* foolishness

```
it was the worst of times it, times it was the worst of,
times it was the age of, was the age of foolishness it
```

it was the worst of times it was the age of foolishness



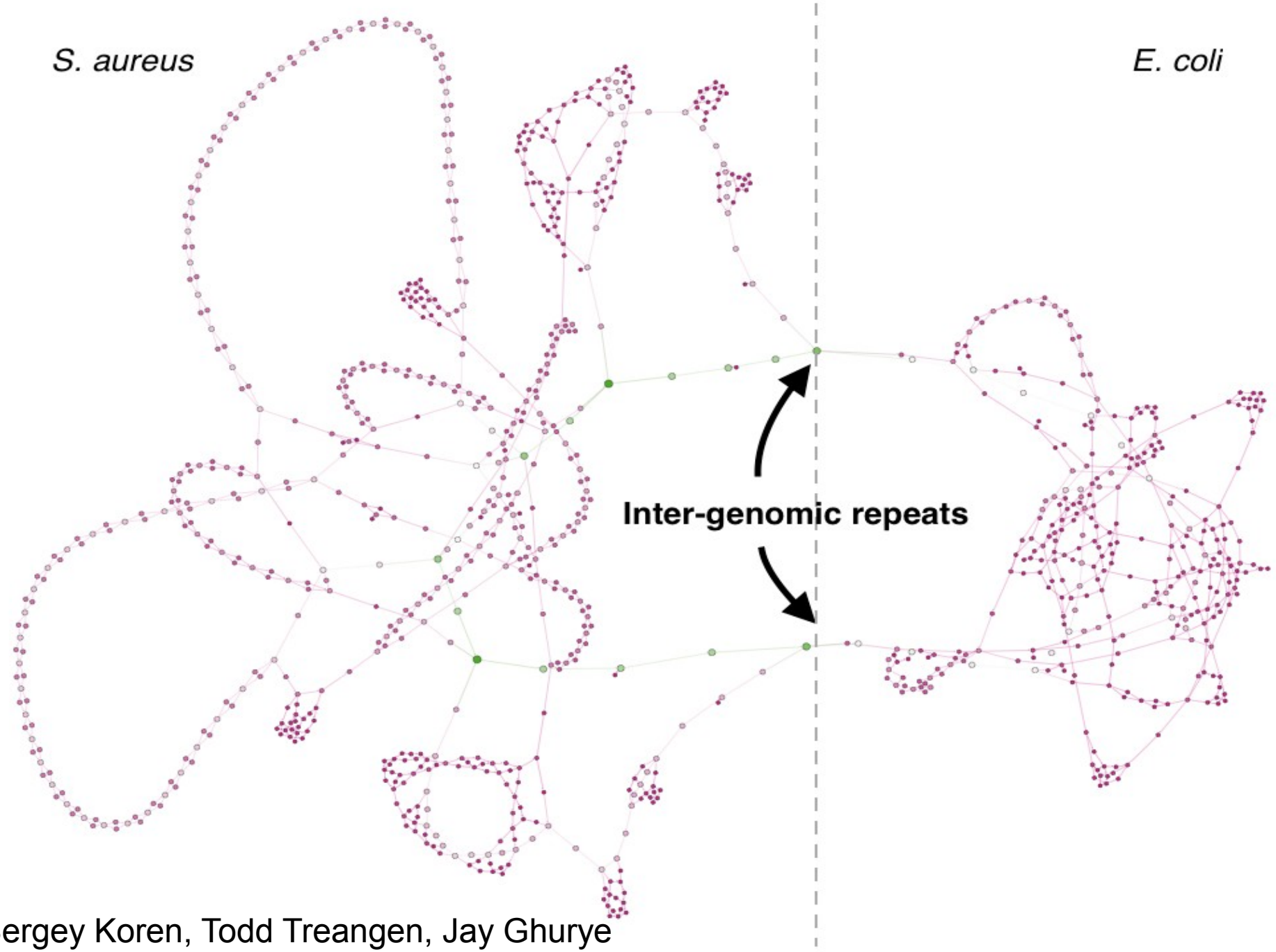
Metagenomic assembly: mixed genomes

- Mathematically not well defined
  - no extensive research in this field (unlike > 30 years of work on isolate genome assembly)

- Biologically not well defined
  - reconstruct organisms
  - reconstruct genes
  - discover genomic variation
  - estimate relative abundances
  - etc...

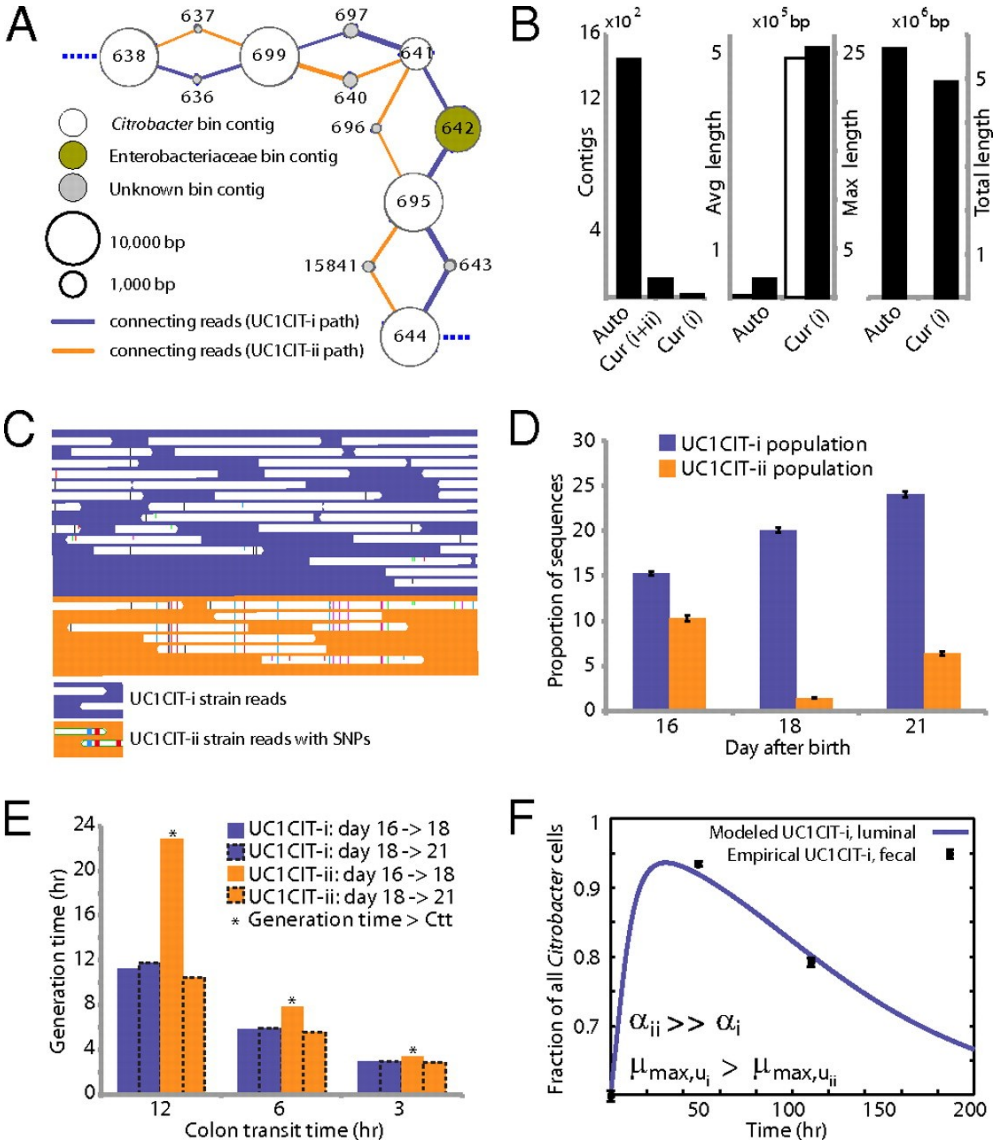
All good problems in life are ill-posed or intractable

Graph-based repeat detection



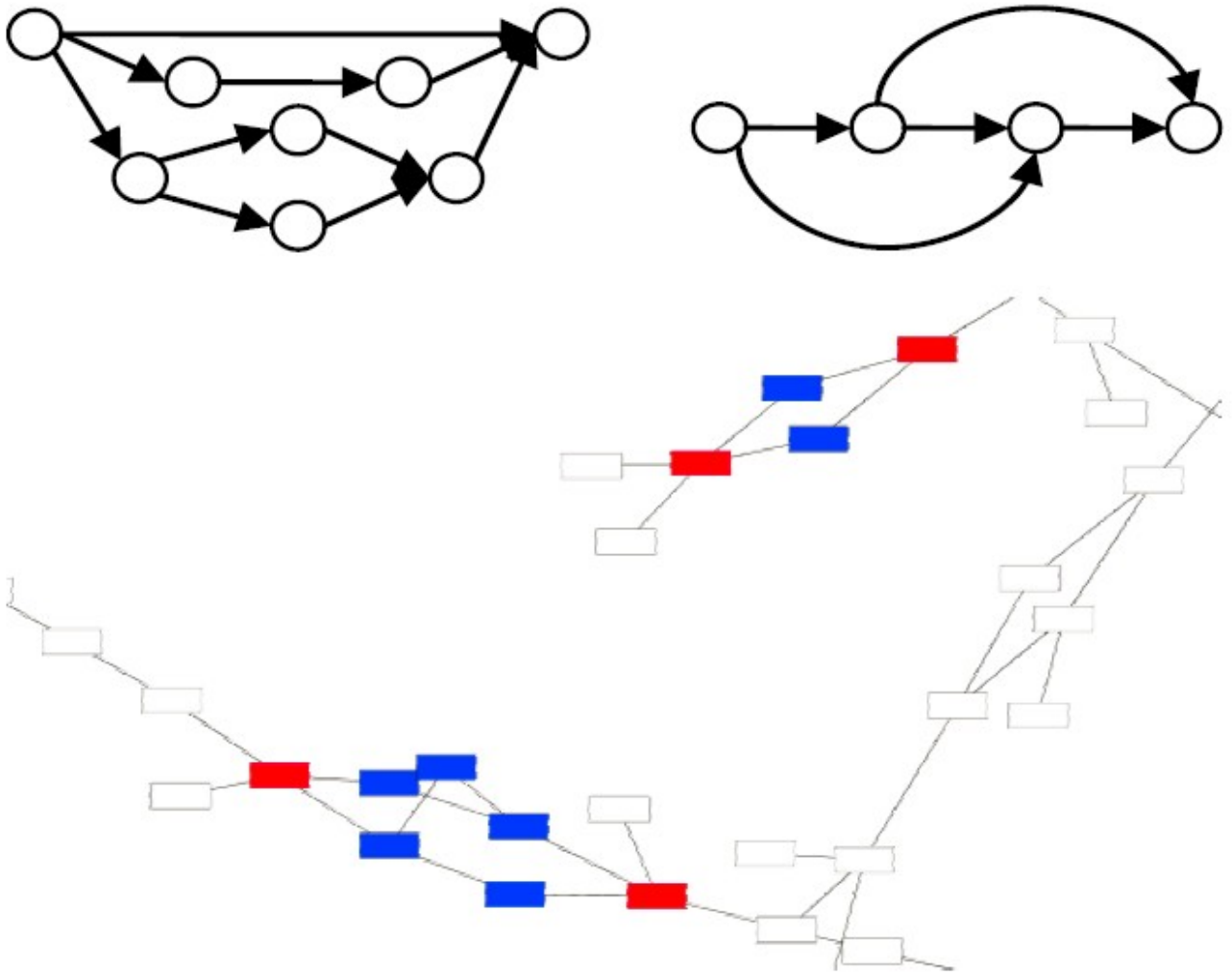
with Sergey Koren, Todd Treangen, Jay Ghurye

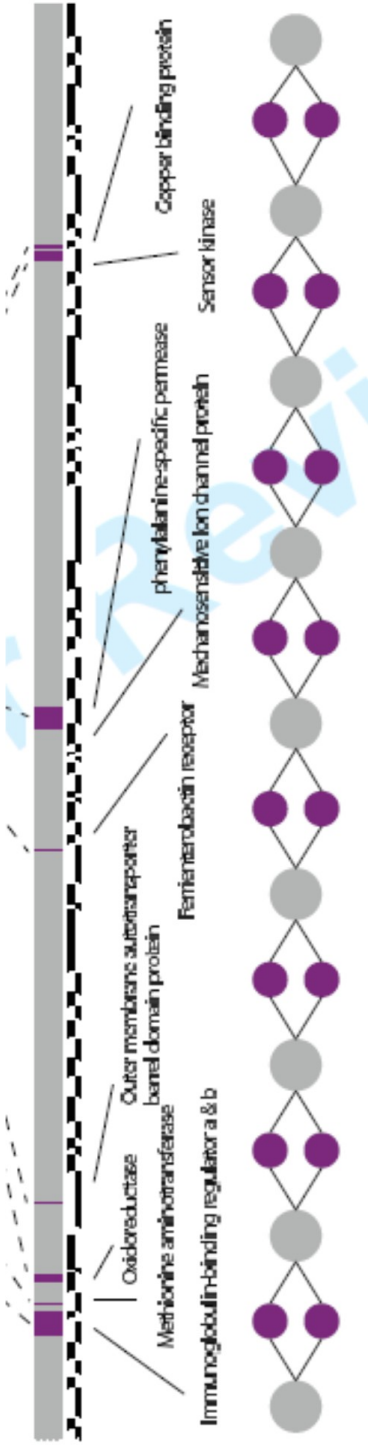
Analyses of two ecologically divergent Citrobacter UC1CIT subpopulations.



Michael J. Morowitz et al. PNAS 2011;108:1128-1133

Finding genomic variants



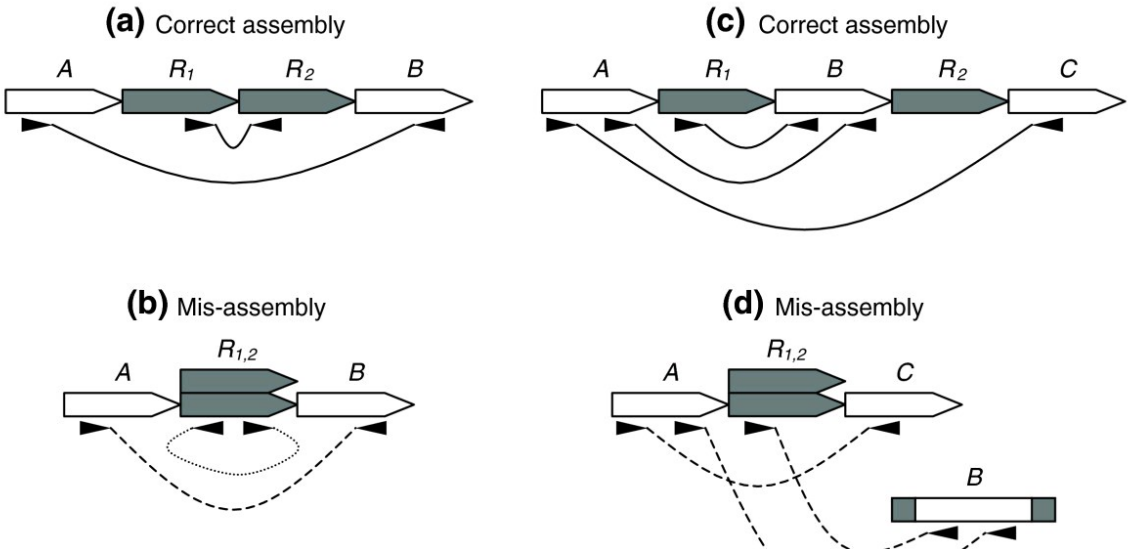


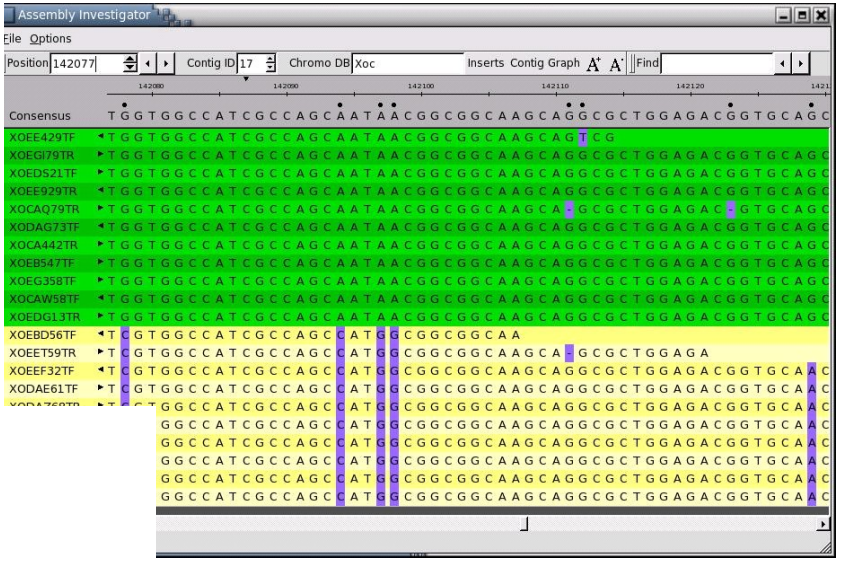
SPQR tree decomposition into tri-connected components
with Jurgen Nijkamp, Matt Myers, Jay Ghurye

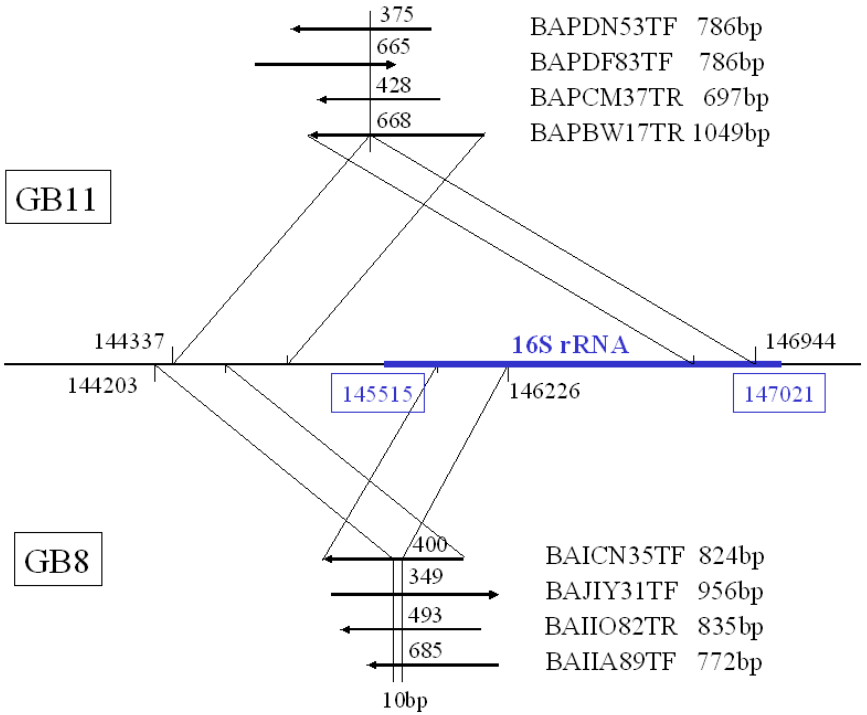
Common core standards: 1st grade

Self-assess effectiveness of strategies

Is my assembly correct?

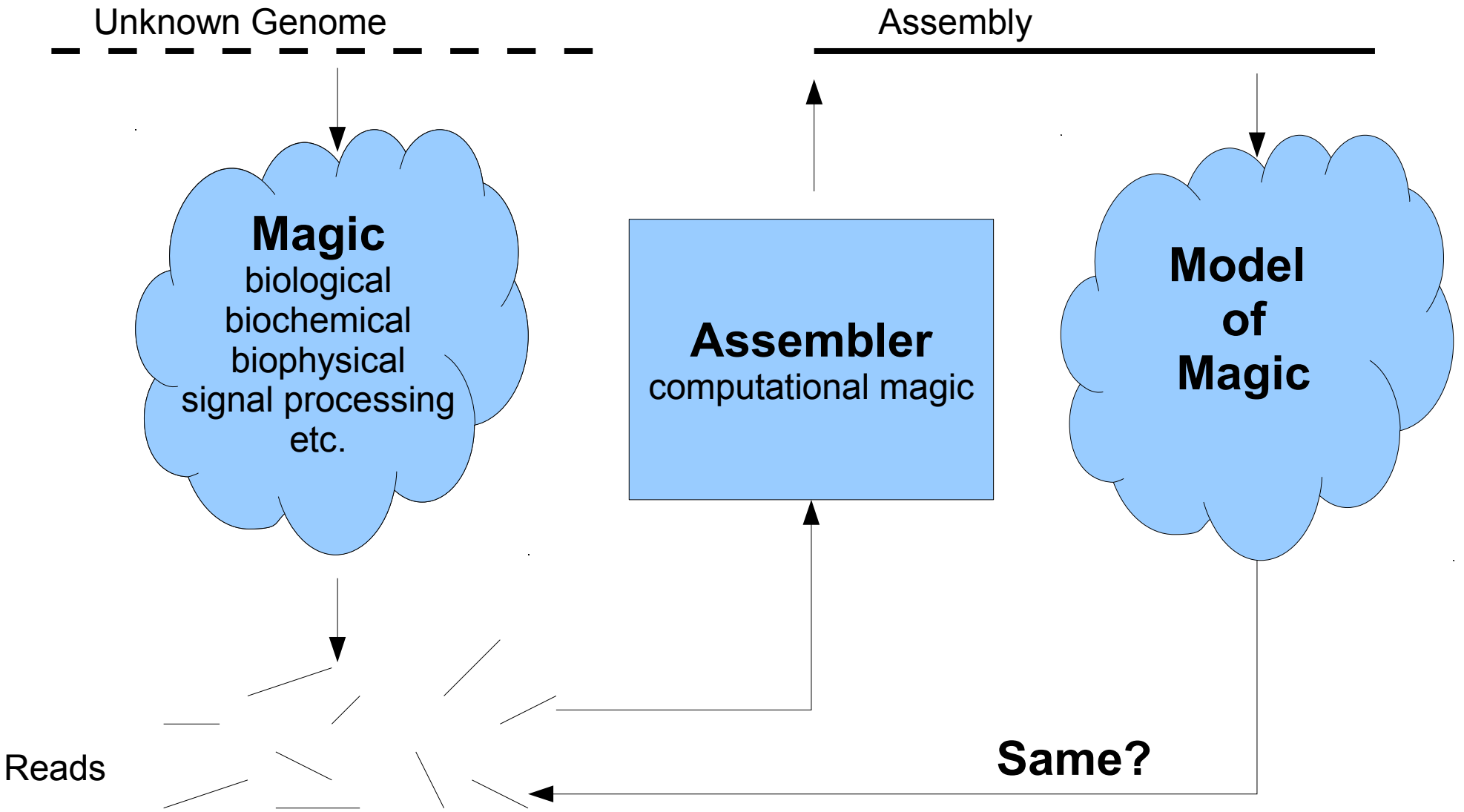






Work with Chris Hill, Atif Memon

Model-based testing

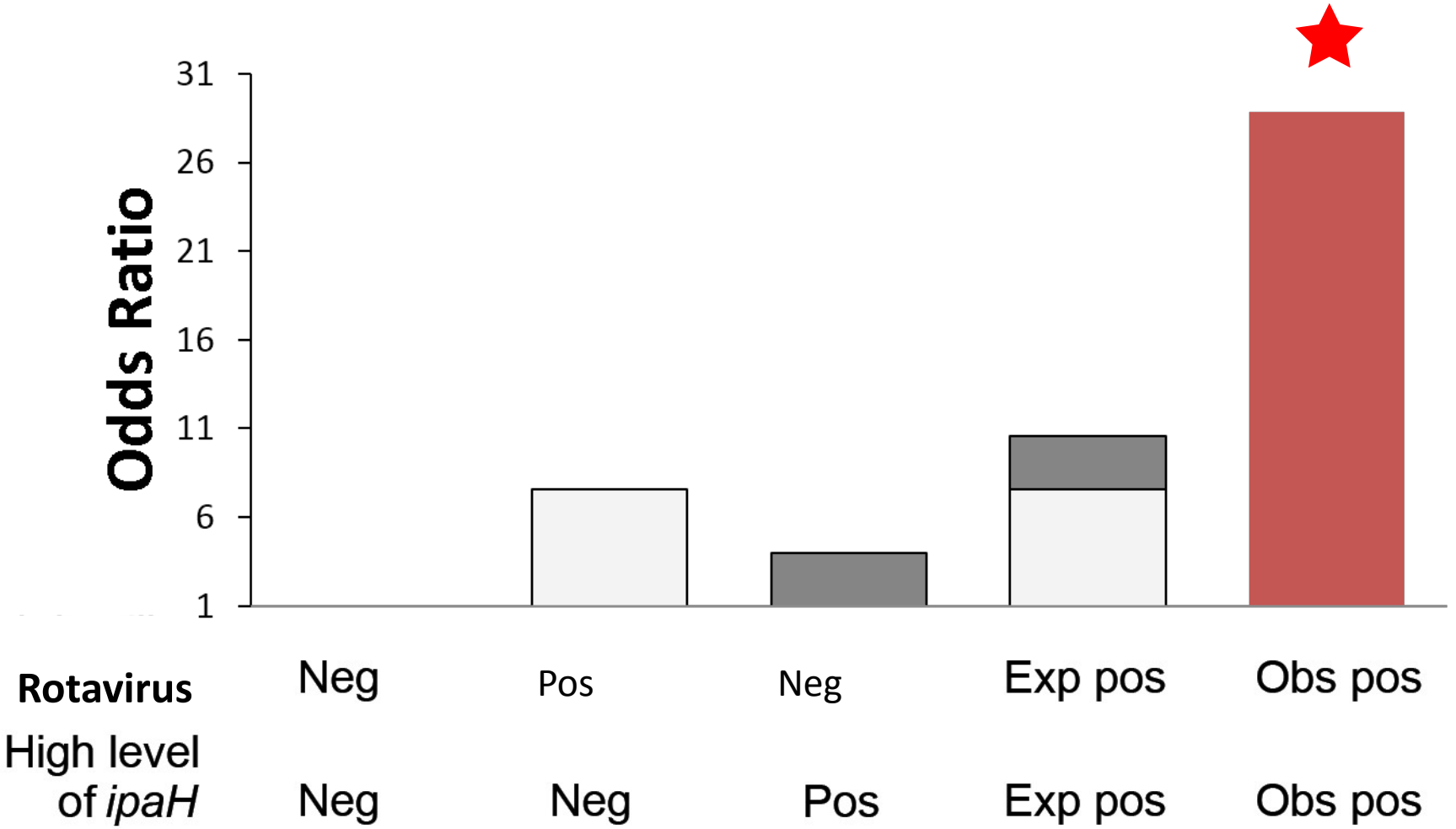


Work with Mohammad Ghodsi, Chris Hill, Bo Liu, Todd Treangen, Irina Astrovskaya

Common core standards: 1st grade

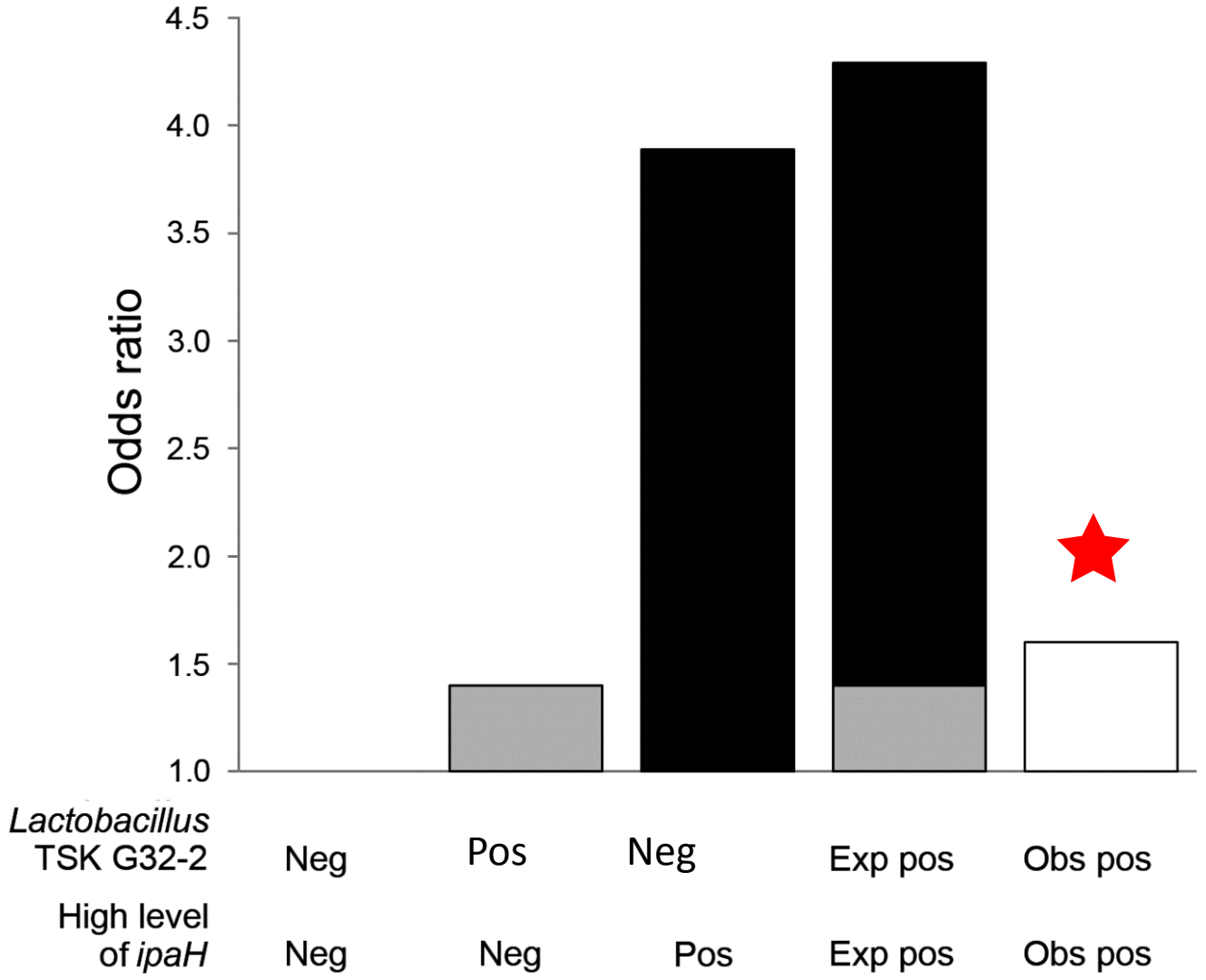
Identify relationships among parts of a whole

Departure from Additivity in Rotavirus/*Shigella* Co-infection



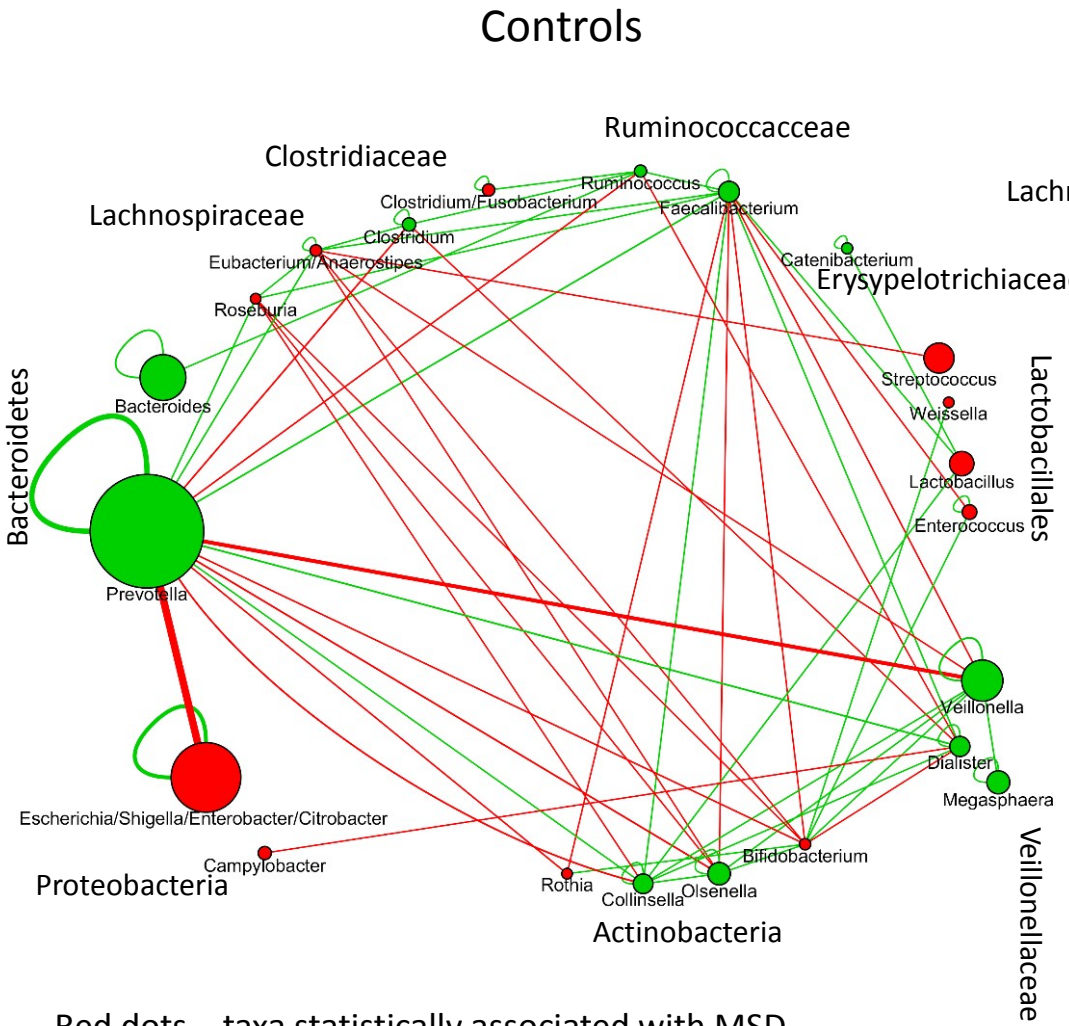
★ Significant increase in OR by factor >2

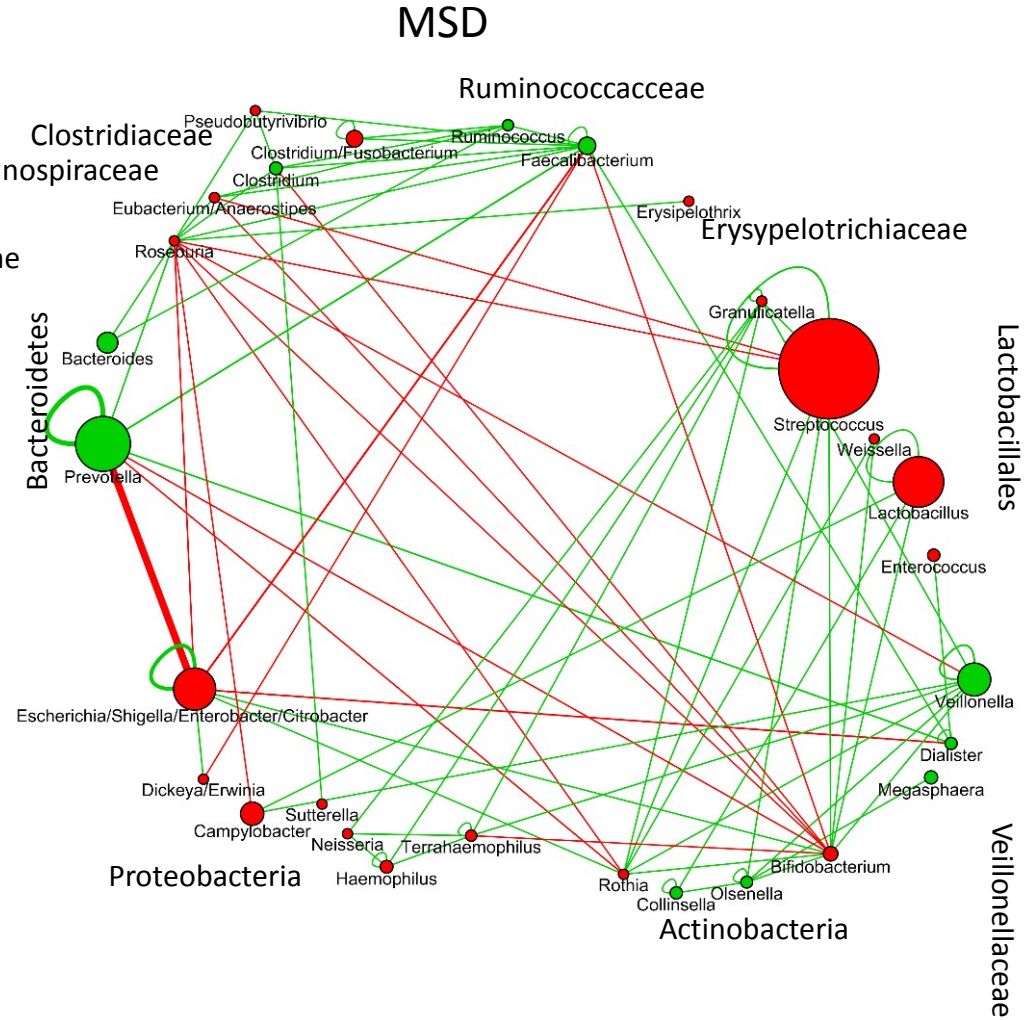
Departure from Additivity in *Lactobacillus/Shigella* Co-infection



★ Significant reduction in OR by factor >2

Network analysis suggests microbial patterns specific to cases and controls





Red dots – taxa statistically associated with MSD
Green dots – taxa statistically associated with controls
Red lines – negative associations between taxa
Green lines – positive associations between taxa

Comparing MSD to controls: Observe similar groups but different connectivity

Sparse Lotka-Volterra Modeling

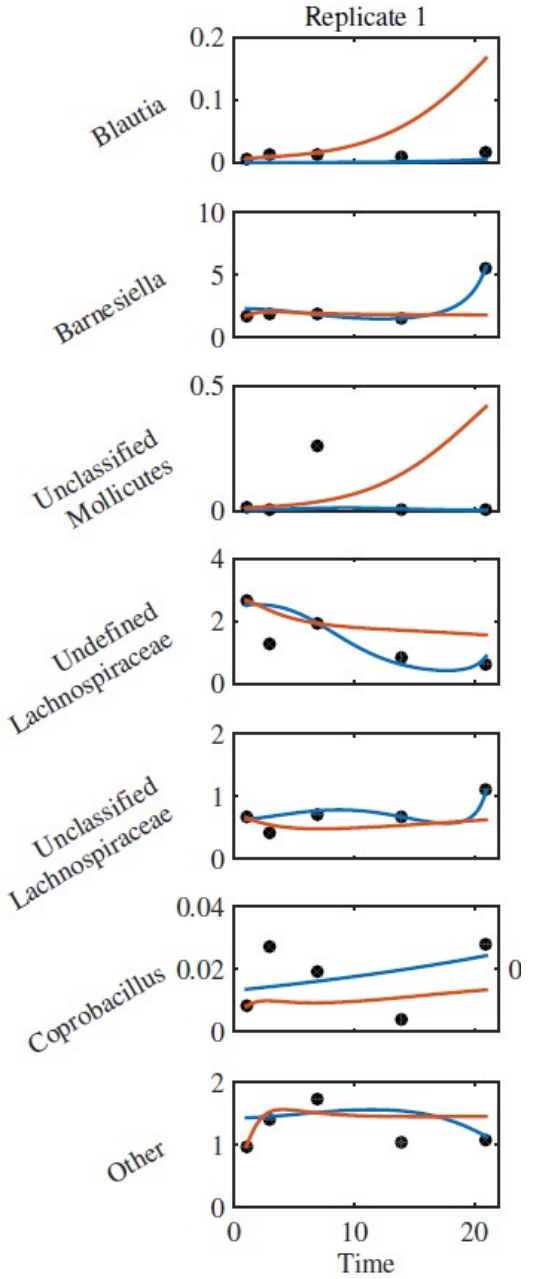
$$y' = diag(y)(b + A\,y)$$

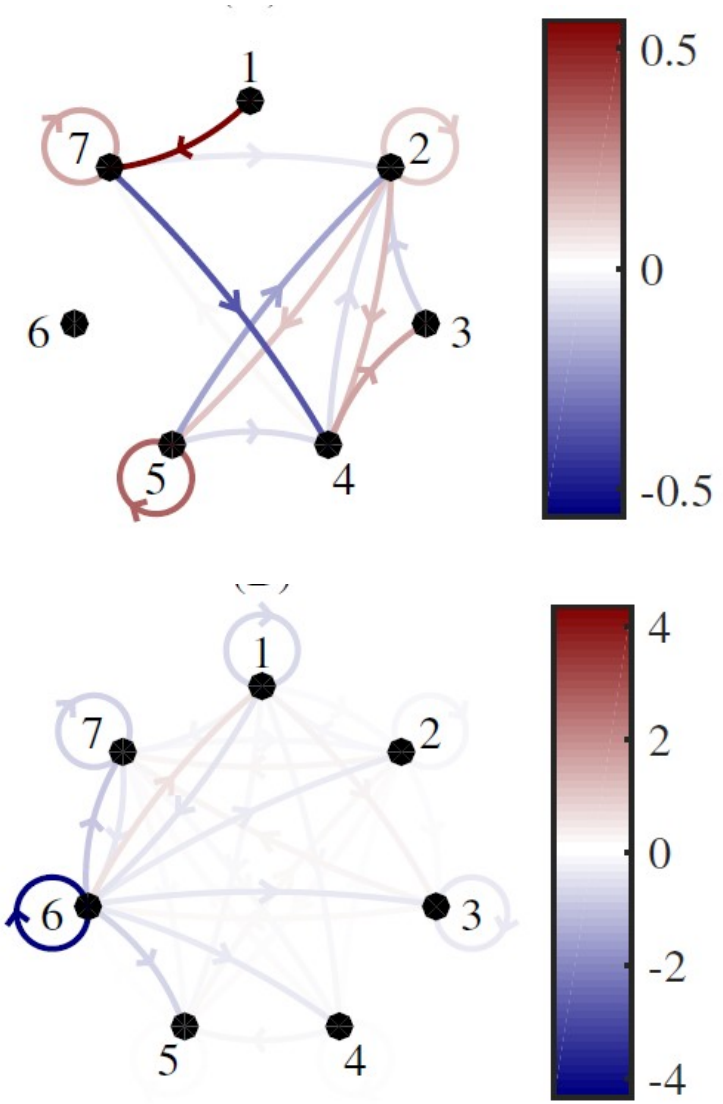
growth rates

$$b = \begin{pmatrix} b_1 \\ \dots \\ b_n \end{pmatrix}$$

interactions

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$



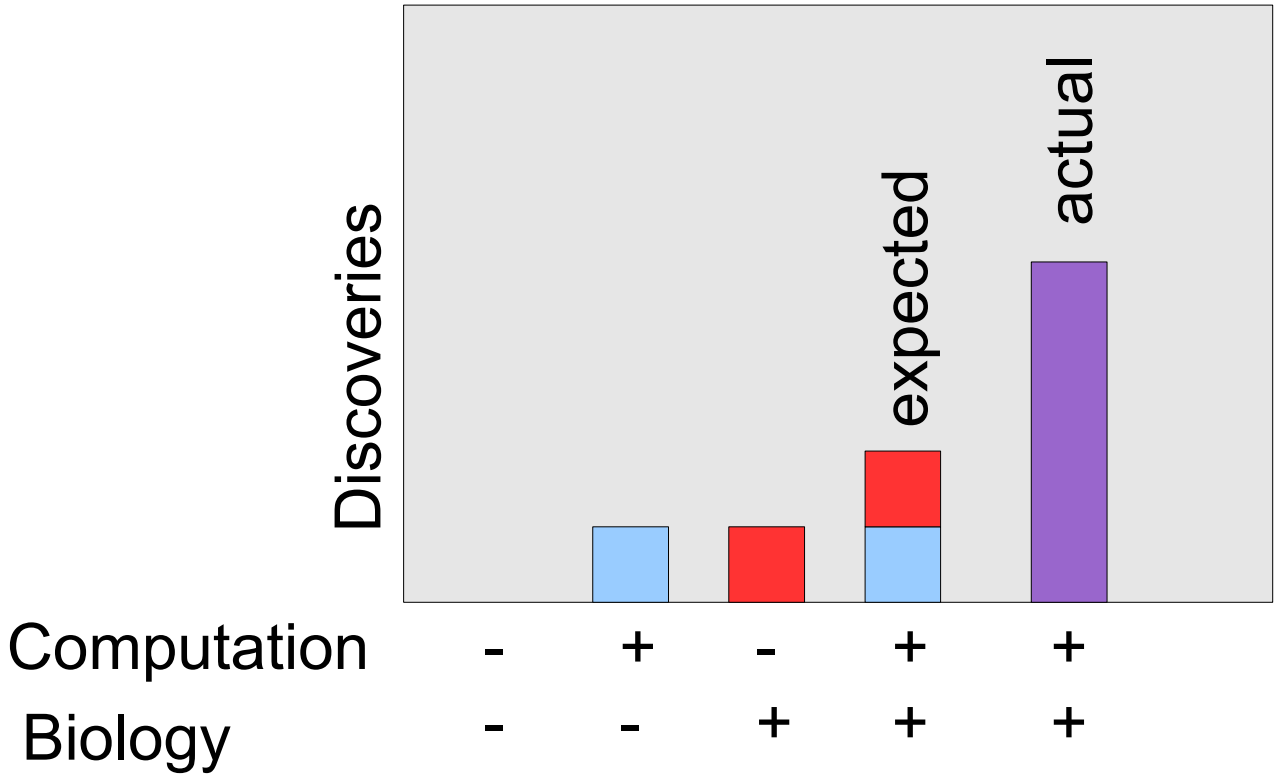


With Matthias Chung and Justin Wagner

What's next?

- Analysis of genome structure across samples
- Models of interactions in microbial communities
- Software testing/evaluation for ill-posed intractable problems

- Education!


Discoveries
expected
actual
Computation - + - + +
Biology - - + + +

Acknowledgments

Too many for a slide:

Pop Lab today
Pop Lab past (now at GIS, JHU, CSHL, Google, Square, Harvard, UW, Nats, etc.)
CS
UMIACS
CBCB
NIH/HMP
INRA (sabbatical host)

Collaborators at:
UMB, UIUC, UVA, VA Tech, BU, TU Delft, U.Wisc.

MY FAMILY

I feel I am nibbling on the edges of this world when I am capable of getting what **Picasso** means when he says to me—perfectly straight-facedly—later of the enormous **new mechanical brains or calculating machines**: "***But they are useless. They can only give you answers***." How easy and comforting to take these things for jokes—boutades!

William Fifield, The Paris Review, 1964

I have been impressed with the urgency of doing. Knowing is not enough; we must apply. Being willing is not enough; we must do.

Leonardo da Vinci