Computational Challenges in Microbiome Research

Mihai Pop

UNIVERSITY OF MARYLAND

CBCB
Center for Bioinformatics & Computational Biology

UMIACS
University of Maryland Institute for Advanced Computer Studies

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

DIARRHEAL DISEASE KILLS 800,000 CHILDREN EACH YEAR

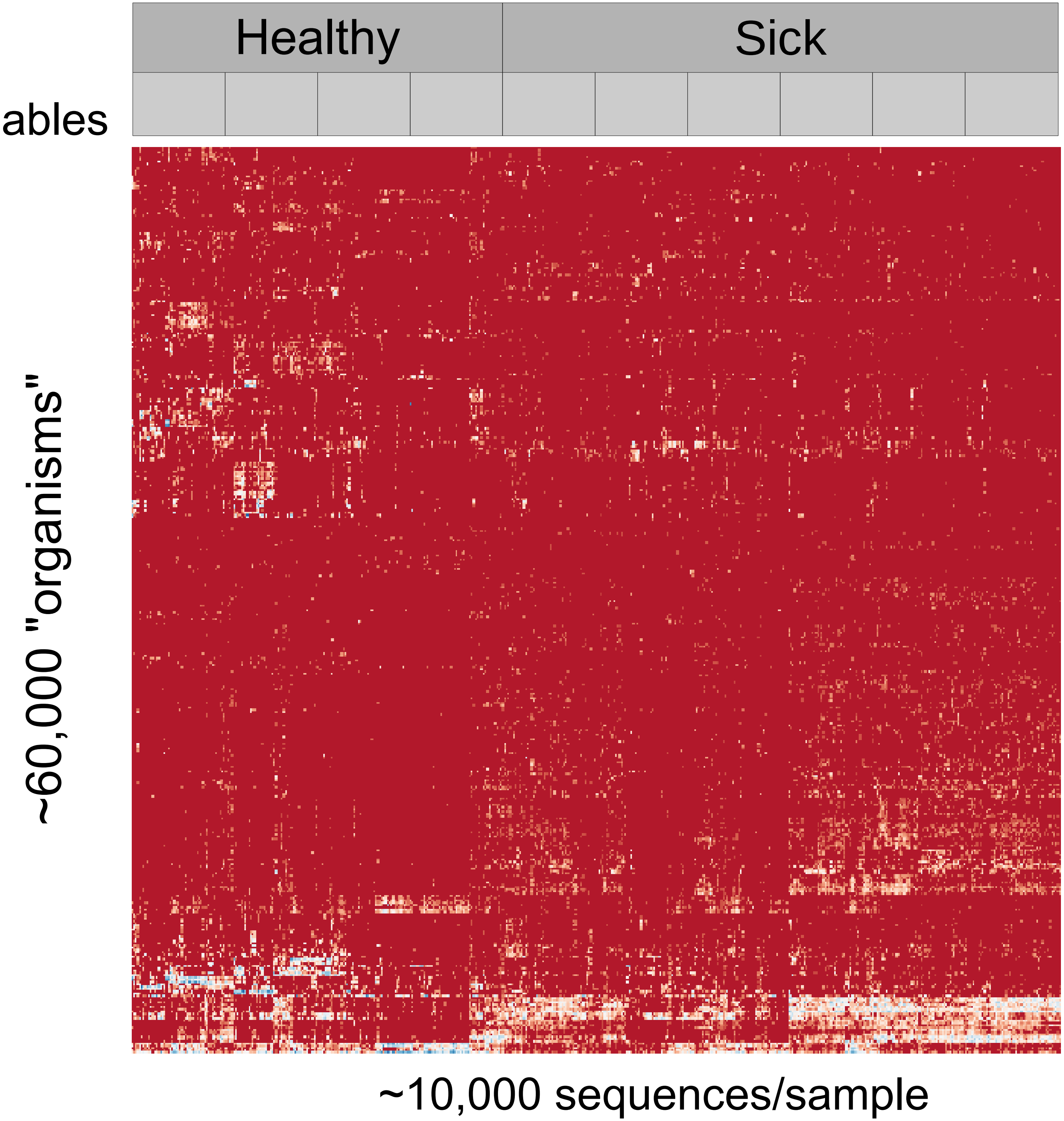
(more than HIV, malaria, and measles combined)

GEMS study: 22,000 children under 5 from 7 African and Asian countries

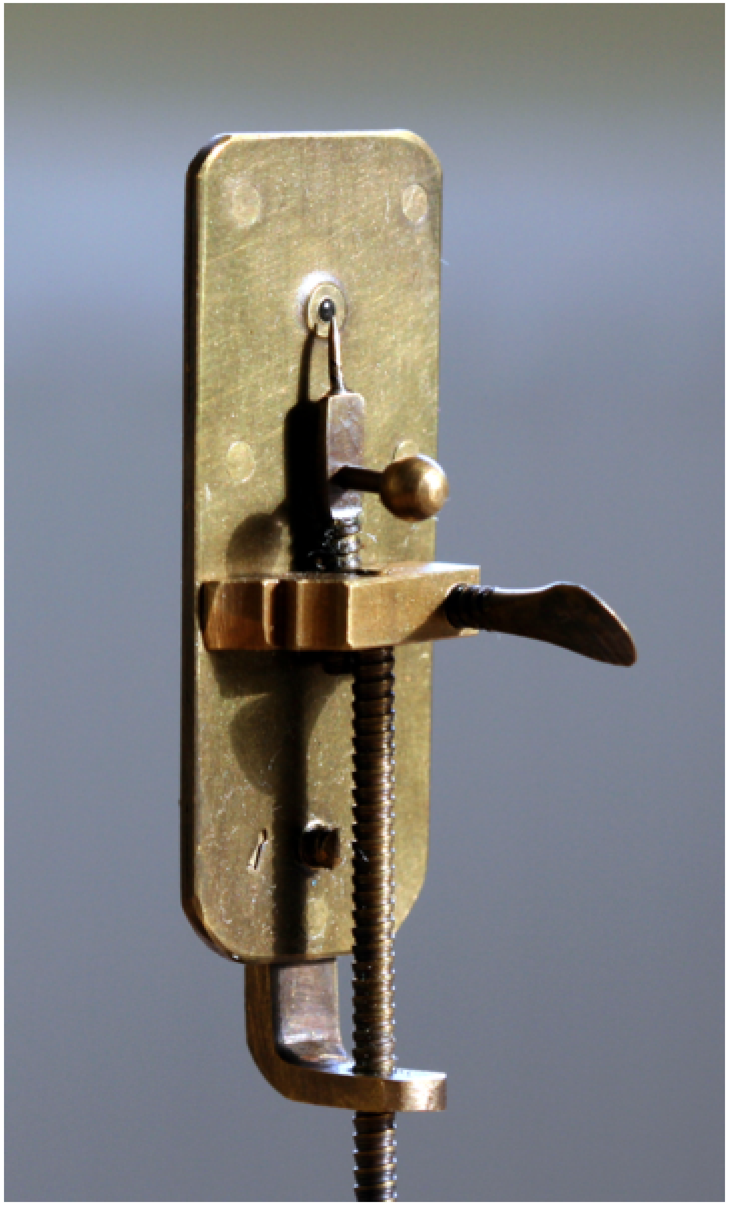
(Lancet, 2013)

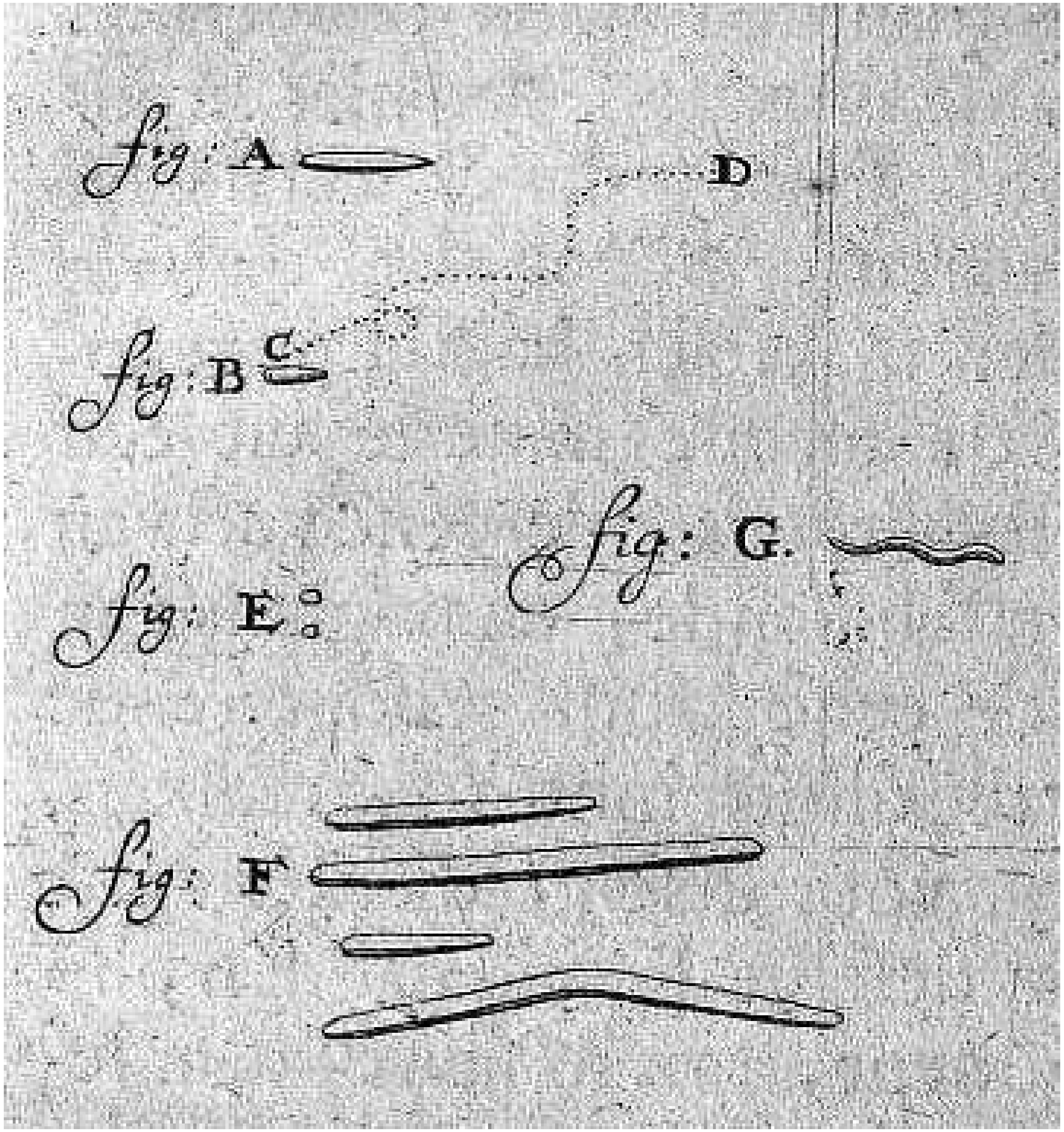
Over half of all cases could not be attributed to any known pathogen

3000 samples
~1000 clinical variables



17th century biology



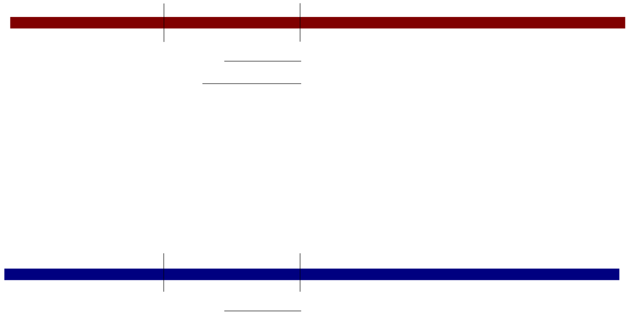


21st century biology

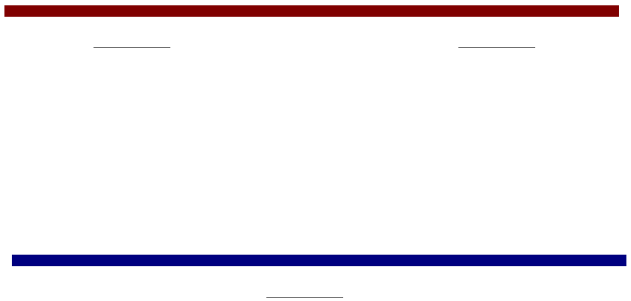
```
>F4BT0V001CZSIM rank=0000138 x=1110.0 y=2700.0 leng
ACTGCTCTCATGCTGCCTCCCGTAGGAGTGCCTCCCTGAGCCAGGATCAAA
>F4BT0V001BBJQS rank=0000155 x=424.0 y=1826.0 lengt
ACTGACTGCATGCTGCCTCCCGTAGGAGTGCCTCCCTGCGCCATCAA
>F4BT0V001EDG35 rank=0000182 x=1676.0 y=2387.0 leng
ACTGACTGCATGCTGCCTCCCGTAGGAGTCGCCGTCCTCGACNC
>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=46
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 versus different

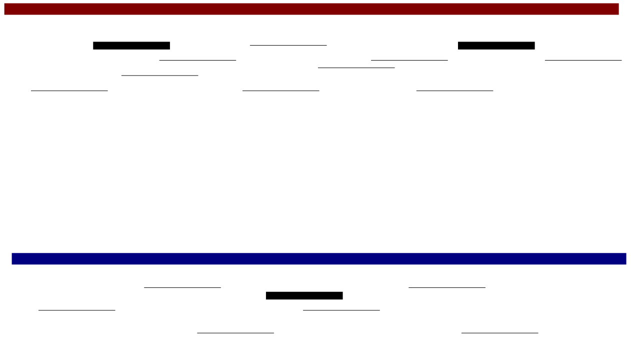
16S



WGS



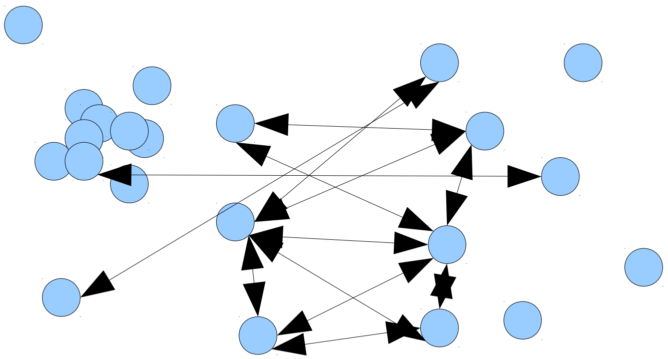
WGS



meta-genome assembly

16S analysis is easy

It's ultimately just clustering...



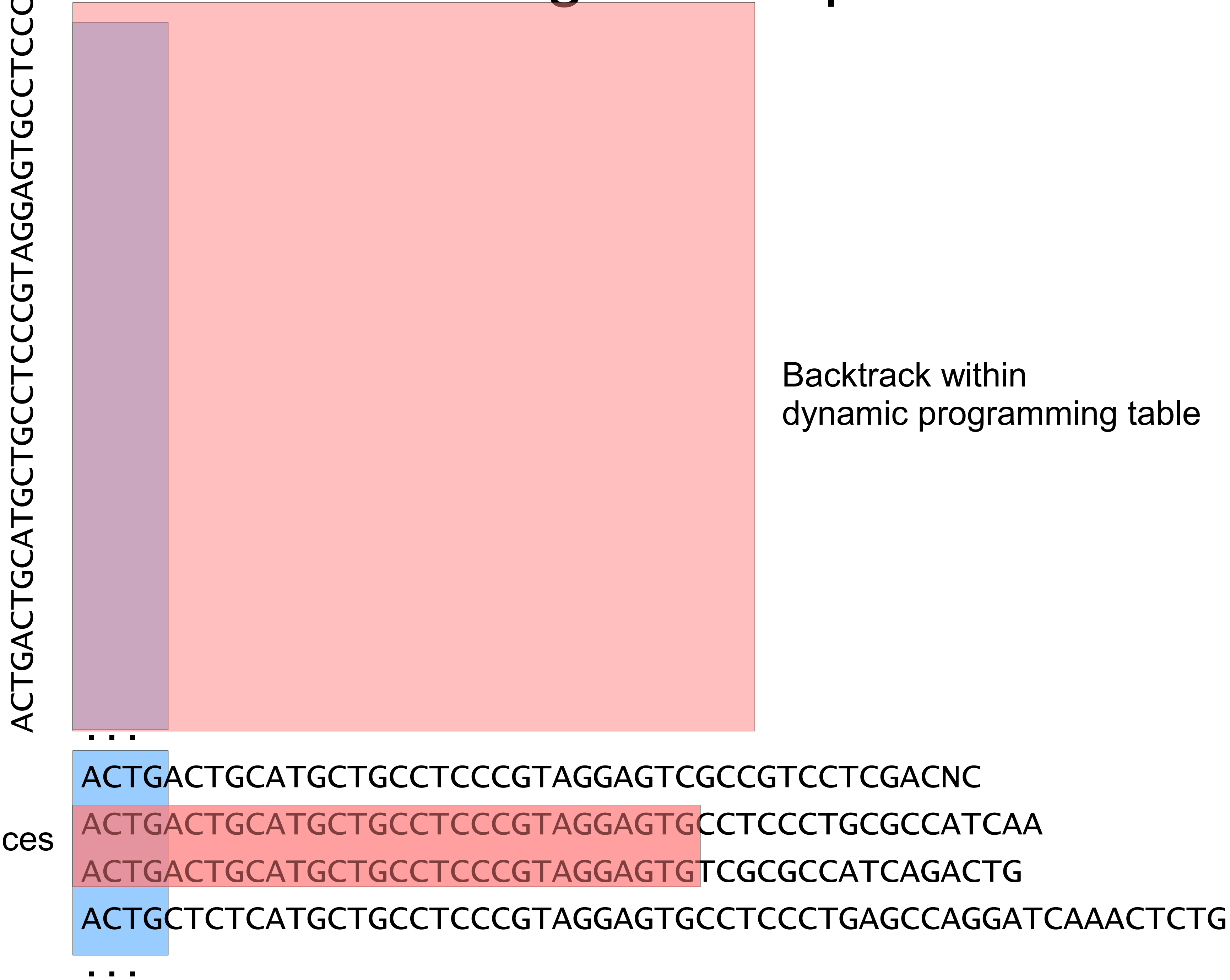
Must compare all versus all
(at least)

30,000,000 X 30,000,000 = 9 X 10^{14} (900 trillion pairs)

```
ACTGCT--CATGCTGCCT--CGTAGGAGTGCCTCCCTGAGCCAGGATCAAACGTCTG
ACTGCTCTCATGGTG-CTCCCGTAGTAGTGCCTCC-TGAGCTAGGATC–ACCTC---
```

(each pair, a full dynamic programming alignment)

Indexing can help



ACTGACTGCATGCTGCCTCCCGTAGGAGTGCCTCCCTGCG

Backtrack within
dynamic programming table

trie
of sequences

...

ACTGACTGCATGCTGCCTCCCGTAGGAGTCGCCGTCCTCGACNC

ACTGACTGCATGCTGCCTCCCGTAGGAGTGCCTCCCTGCGCCATCAA

ACTGACTGCATGCTGCCTCCCGTAGGAGTGTCGCGCCATCAGACTG

ACTGCTCTCATGCTGCCTCCCGTAGGAGTGCCTCCCTGAGCCAGGATCAAACTCTG

...

DNAclust – Ghodsi et al. 2011

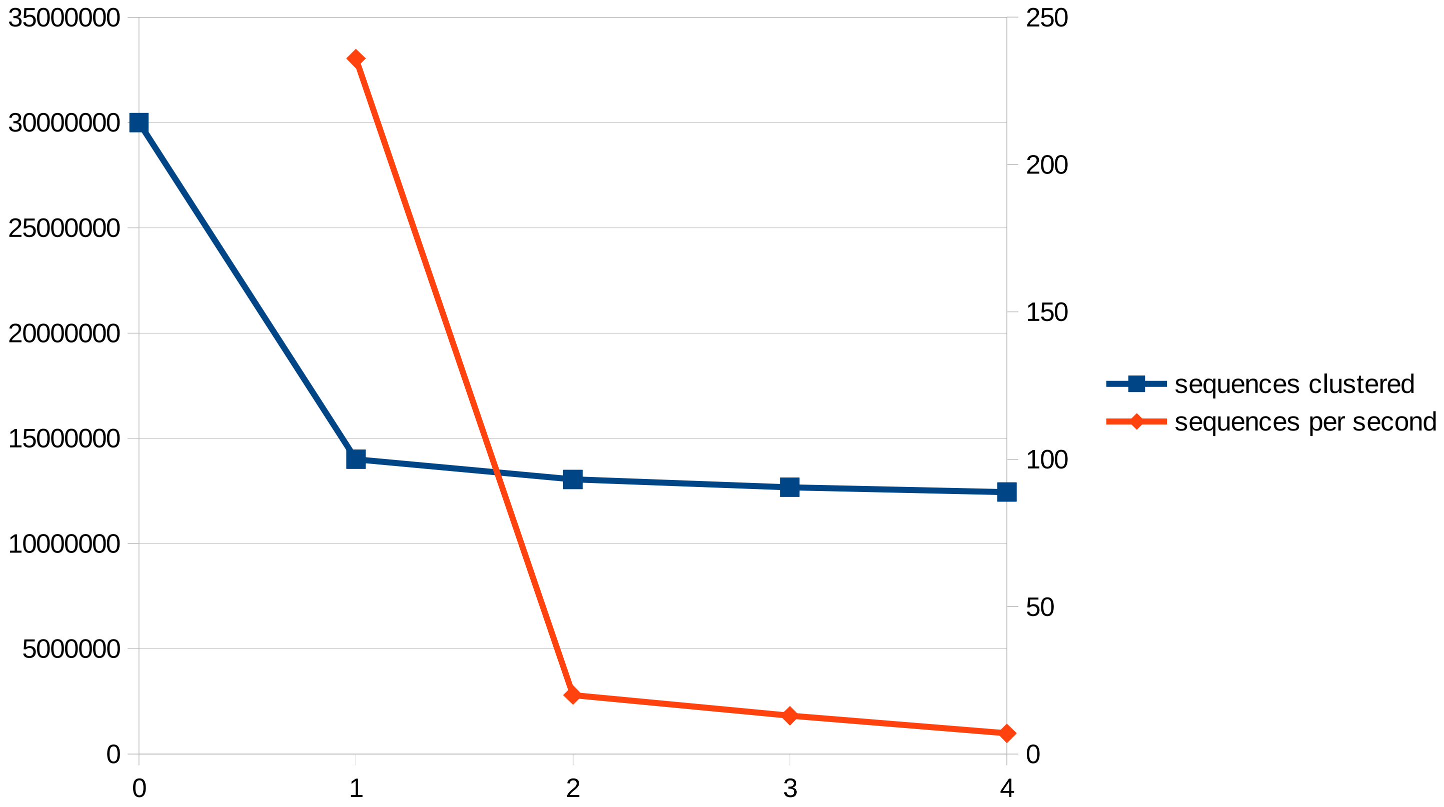
Large clusters can be found quickly

Select a random set of √n sequences
Cluster them
Recruit sequences to the clusters found

=> O(n + c · o(nL))

n sequences of length L
c clusters

... repeat

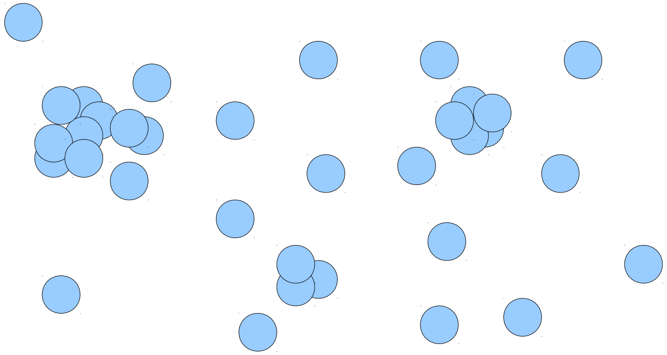


Still too slow - curse of dimensionality

- If we want to find all clusters $O(n^2)$ seems unavoidable

- 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

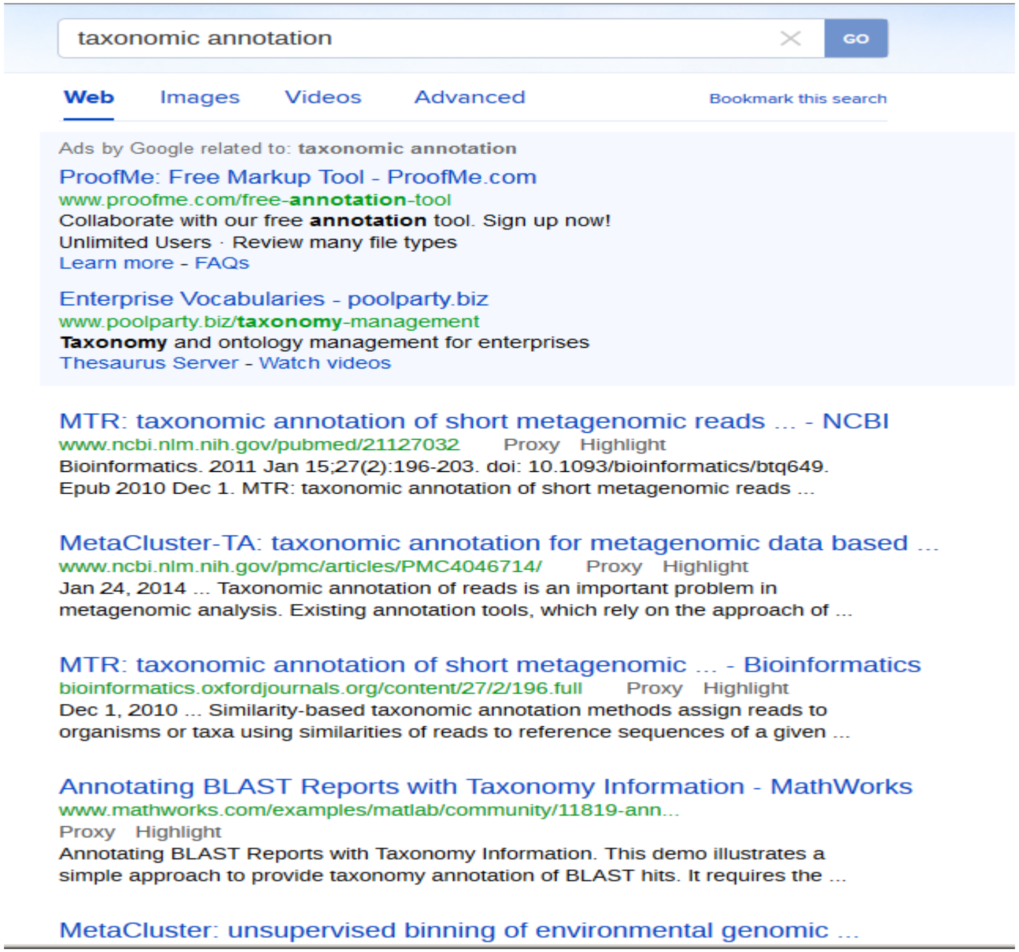
- Simple filtering techniques do not work
- Key issue - error

Annotation

Now that clustering is solved

What do the clusters represent?

Google: "taxonomic annotation"



Ads by Google related to: **taxonomic annotation**

ProofMe: Free Markup Tool - ProofMe.com
www.proofme.com/free-**annotation**-tool
Collaborate with our free **annotation** tool. Sign up now!
Unlimited Users · Review many file types
Learn more - FAQs

Enterprise Vocabularies - poolparty.biz
www.poolparty.biz/**taxonomy**-management
Taxonomy and ontology management for enterprises
Thesaurus Server - Watch videos

MTR: taxonomic annotation of short metagenomic reads ... - NCBI
www.ncbi.nlm.nih.gov/pubmed/21127032 Proxy Highlight
Bioinformatics. 2011 Jan 15;27(2):196-203. doi: 10.1093/bioinformatics/btq649. Epub 2010 Dec 1. MTR: taxonomic annotation of short metagenomic reads ...

MetaCluster-TA: taxonomic annotation for metagenomic data based ...
www.ncbi.nlm.nih.gov/pmc/articles/PMC4046714/ Proxy Highlight
Jan 24, 2014 ... Taxonomic annotation of reads is an important problem in metagenomic analysis. Existing annotation tools, which rely on the approach of ...

MTR: taxonomic annotation of short metagenomic ... - Bioinformatics
bioinformatics.oxfordjournals.org/content/27/2/196.full Proxy Highlight
Dec 1, 2010 ... Similarity-based taxonomic annotation methods assign reads to organisms or taxa using similarities of reads to reference sequences of a given ...

Annotating BLAST Reports with Taxonomy Information - MathWorks
www.mathworks.com/examples/matlab/community/11819-ann...
Proxy Highlight
Annotating BLAST Reports with Taxonomy Information. This demo illustrates a simple approach to provide taxonomy annotation of BLAST hits. It requires the ...

MetaCluster: unsupervised binning of environmental genomic ...

- Database of known pages
- Report all that contain keyword

- Ranking important (which of the thousands is most relevant)

Annotation – as easy as a database search

```
5467_464        HM038000.1.1446   E-value: 6e-96  Bit score: 350
```

Bacteria;Cyanobacteria;Melainabacteria;Vampirovibrionales;Vampirovibrio chlorellavorus

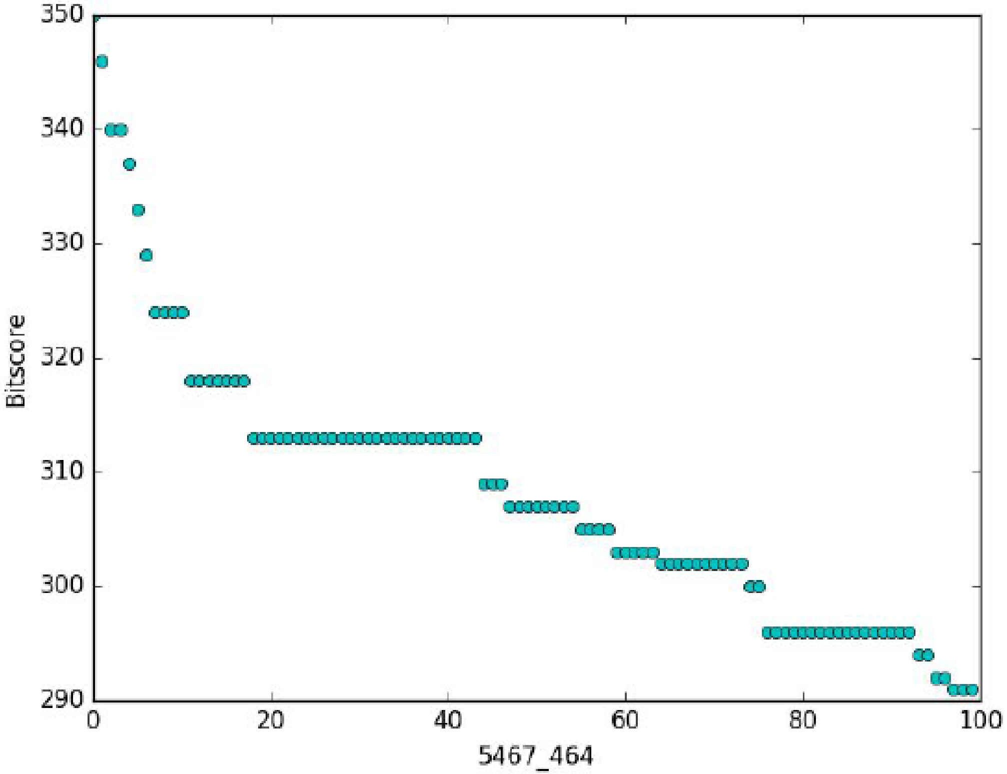
E-value – how many random alignments one expects for the same alignment score/quality

Note: database organized hierarchically to allow one to generalize from inexact matches

Kingdom;Phylum;Class;Order;Family;Genus;Species;

```
5467_464        HM038000.1.1446 Identity: 80.00% E-value: 6e-96 Bitscore: 350
```

1 in 5 letters is different



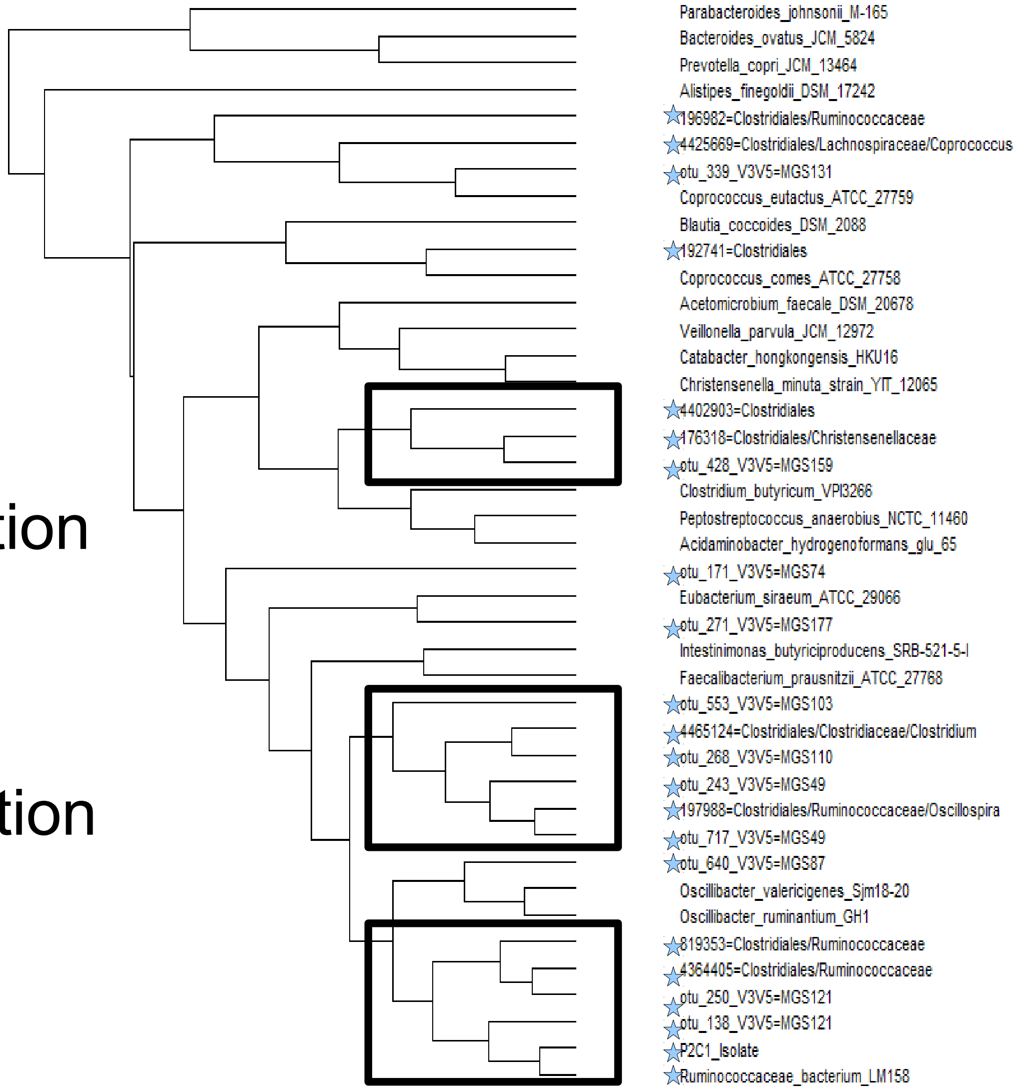
Bacteria;Cyanobacteria;Melainabacteria;Vampirovibrionales;Vampirovibrio chlorellavorus
Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Caulobacteraceae;Brevundimonas;
Brevundimonas mediterranea

Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Caulobacteraceae;Brevundimonas;
Brevundimonas bacteroides

Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Butyricicoccus;Butyricicoccus pullicaecorum

Why biological annotation is hard

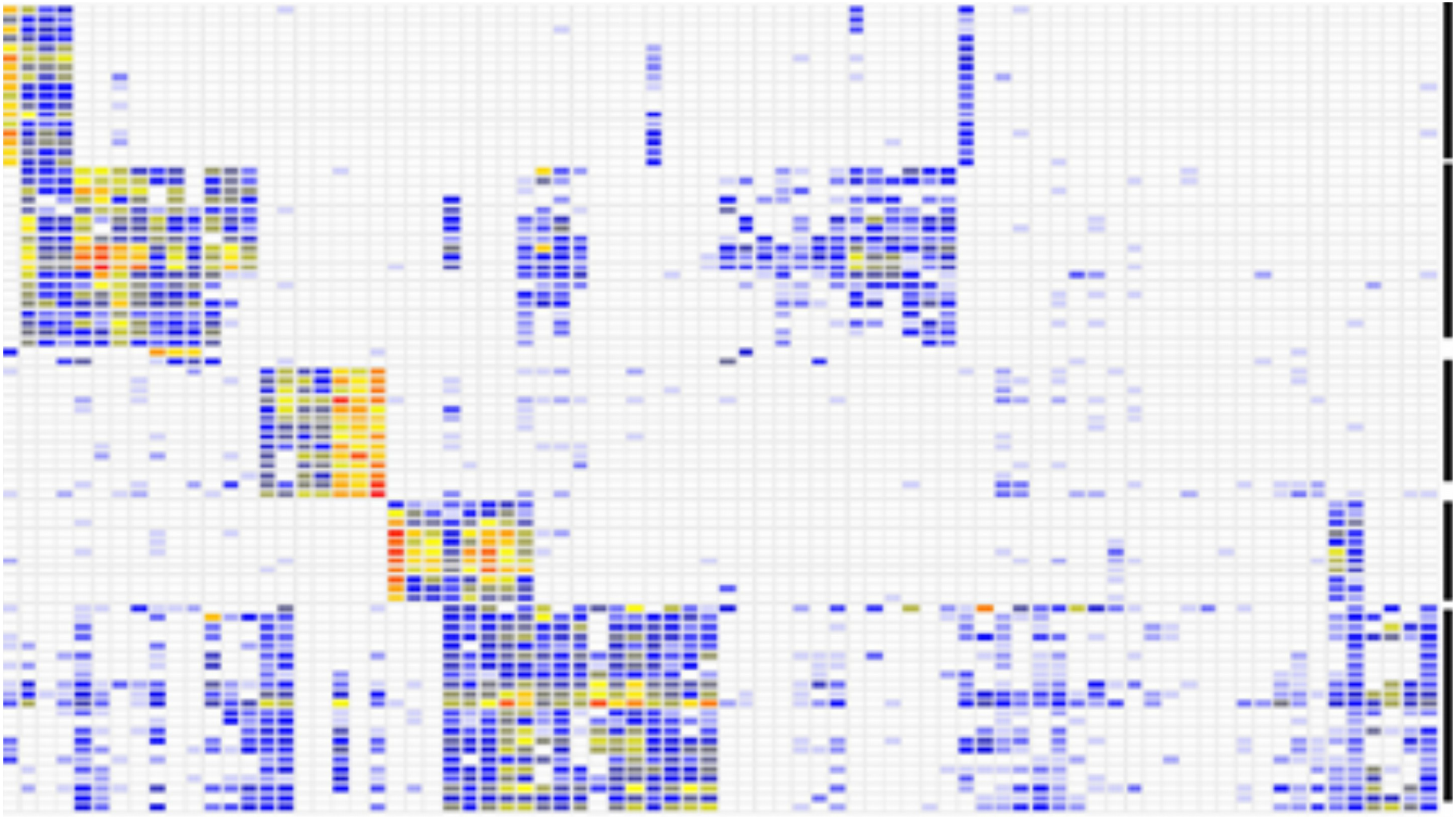
- When sequence is in database – it's a CS problem
- How do we generalize from unknown sequences?
- How do we know we are right?


Parabacteroides_johnsonii_M-165
Bacteroides_ovatus_JCM_5824
Prevotella_copri_JCM_13464
Alistipes_finegoldii_DSM_17242
196982=Clostridiales/Ruminococcaceae
4425669=Clostridiales/Lachnospiraceae/Coprococcus
otu_339_V3V5=MGS131
Coprococcus_eutactus_ATCC_27759
Blautia_coccoides_DSM_2088
192741=Clostridiales
Coprococcus_comes_ATCC_27758
Acetomicrobium_faecale_DSM_20678
Veillonella_parvula_JCM_12972
Catabacter_hongkongensis_HKU16
Christensenella_minuta_strain_YIT_12065
4402903=Clostridiales
176318=Clostridiales/Christensenellaceae
otu_428_V3V5=MGS159
Clostridium_butyricum_VPI3266
Peptostreptococcus_anaerobius_NCTC_11460
Acidaminobacter_hydrogenoformans_glu_65
otu_171_V3V5=MGS74
Eubacterium_siraeum_ATCC_29066
otu_271_V3V5=MGS177
Intestinimonas_butyriciproducens_SRB-521-5-I
Faecalibacterium_prausnitzii_ATCC_27768
otu_553_V3V5=MGS103
4465124=Clostridiales/Clostridiaceae/Clostridium
otu_268_V3V5=MGS110
otu_243_V3V5=MGS49
197988=Clostridiales/Ruminococcaceae/Oscillospira
otu_717_V3V5=MGS49
otu_640_V3V5=MGS87
Oscillibacter_valericigenes_Sjm18-20
Oscillibacter_ruminantium_GH1
819353=Clostridiales/Ruminococcaceae
4364405=Clostridiales/Ruminococcaceae
otu_250_V3V5=MGS121
otu_138_V3V5=MGS121
P2C1_Isolate
Ruminococcaceae_bacterium_LM158

<u>Formally:</u> name equivalent to function

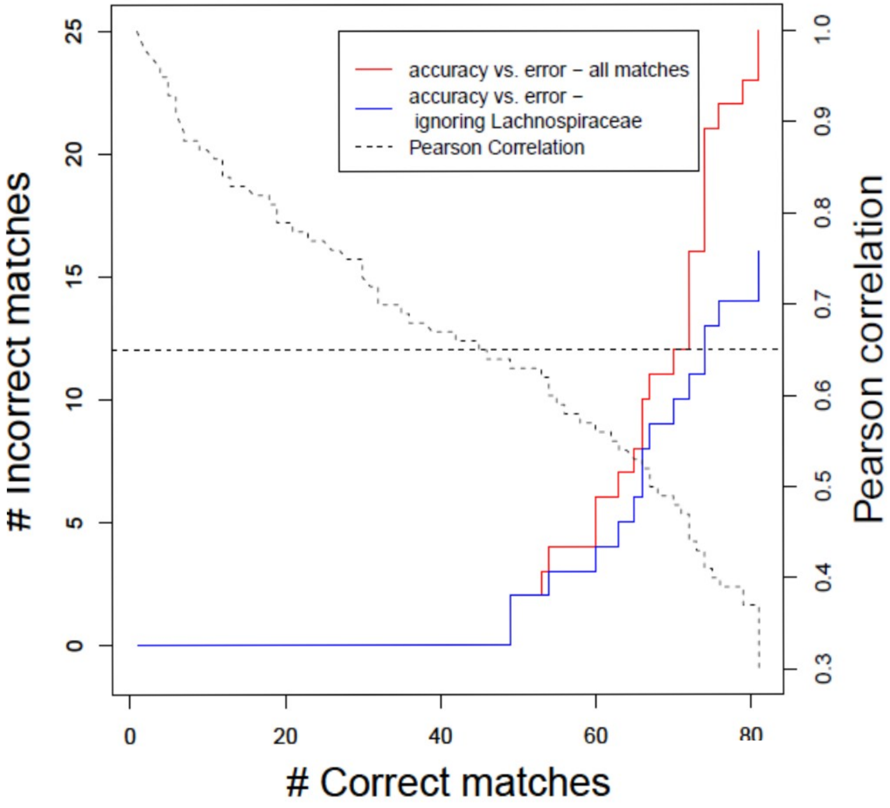
isolate
perform experiments
come up with correct Latin declination

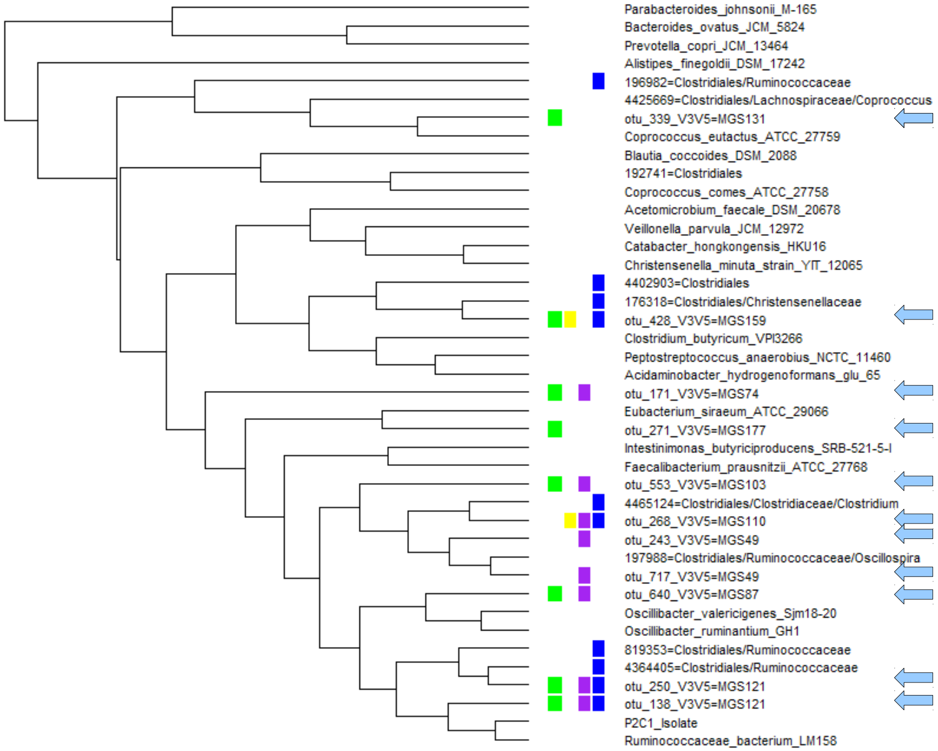
New information: correlation across samples



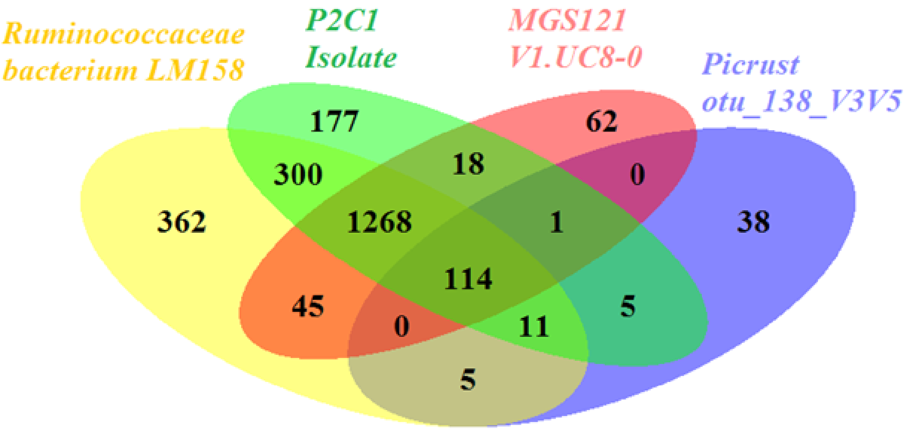
Quince – Concoct
Borenstein – Metagenomic deconvolution

Associating taxonomy markers with genes

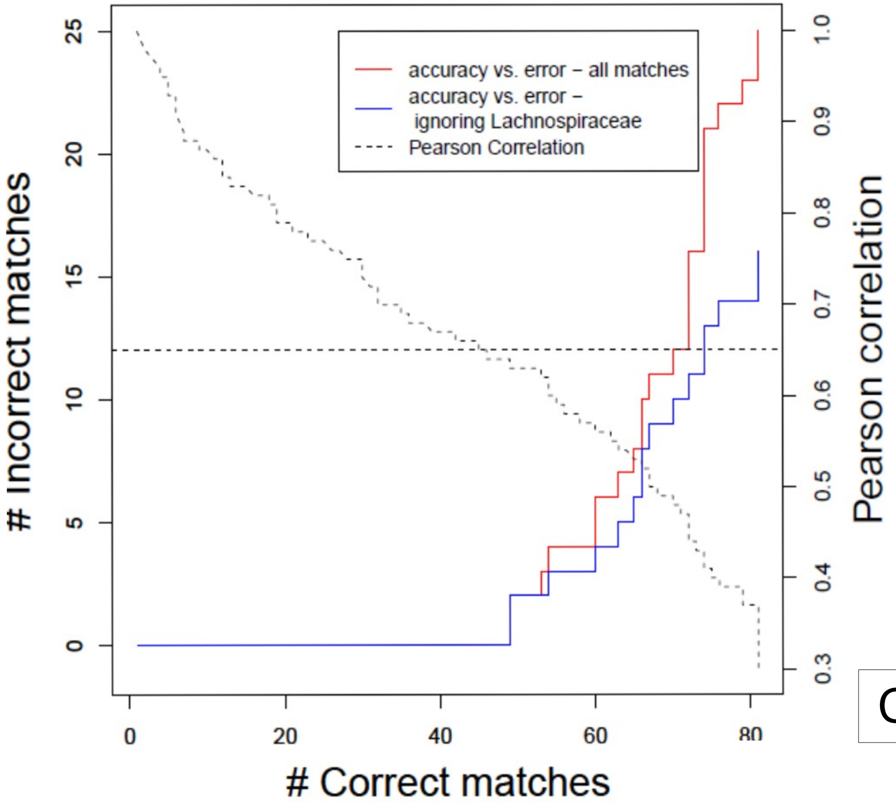


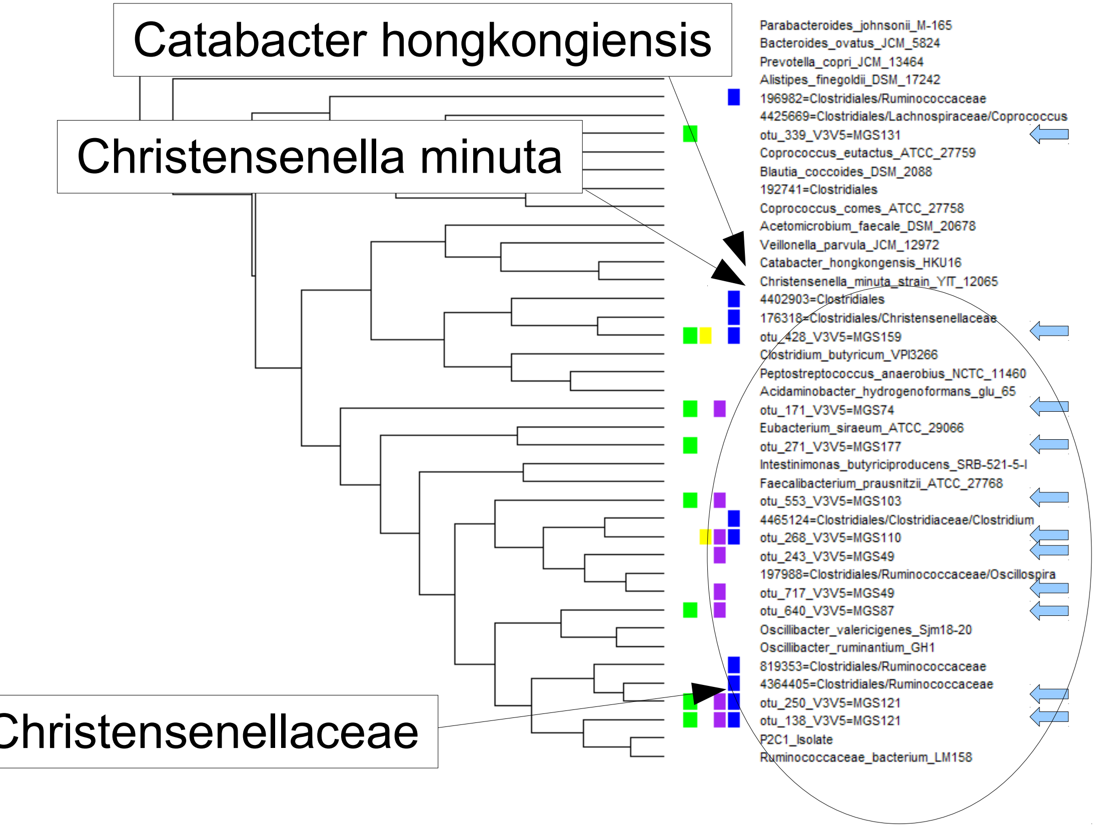


Parabacteroides_johnsonii_M-165
Bacteroides_ovatus_JCM_5824
Prevotella_copri_JCM_13464
Alistipes_finegoldii_DSM_17242
196982=Clostridiales/Ruminococcaceae
4425669=Clostridiales/Lachnospiraceae/Coprococcus
otu_339_V3V5=MGS131
Coprococcus_eutactus_ATCC_27759
Blautia_coccoides_DSM_2088
192741=Clostridiales
Coprococcus_comes_ATCC_27758
Acetomicrobium_faecale_DSM_20678
Veillonella_parvula_JCM_12972
Catabacter_hongkongensis_HKU16
Christensenella_minuta_strain_YIT_12065
4402903=Clostridiales
176318=Clostridiales/Christensenellaceae
otu_428_V3V5=MGS159
Clostridium_butyricum_VPI3266
Peptostreptococcus_anaerobius_NCTC_11460
Acidaminobacter_hydrogenoformans_glu_65
otu_171_V3V5=MGS74
Eubacterium_siraeum_ATCC_29066
otu_271_V3V5=MGS177
Intestinimonas_butyriciproducens_SRB-521-5-I
Faecalibacterium_prausnitzii_ATCC_27768
otu_553_V3V5=MGS103
4465124=Clostridiales/Clostridiaceae/Clostridium
otu_268_V3V5=MGS110
otu_243_V3V5=MGS49
197988=Clostridiales/Ruminococcaceae/Oscillospira
otu_717_V3V5=MGS49
otu_640_V3V5=MGS87
Oscillibacter_valericigenes_Sjm18-20
Oscillibacter_ruminantium_GH1
819353=Clostridiales/Ruminococcaceae
4364405=Clostridiales/Ruminococcaceae
otu_250_V3V5=MGS121
otu_138_V3V5=MGS121
P2C1_Isolate
Ruminococcaceae_bacterium_LM158



Naming is still an issue

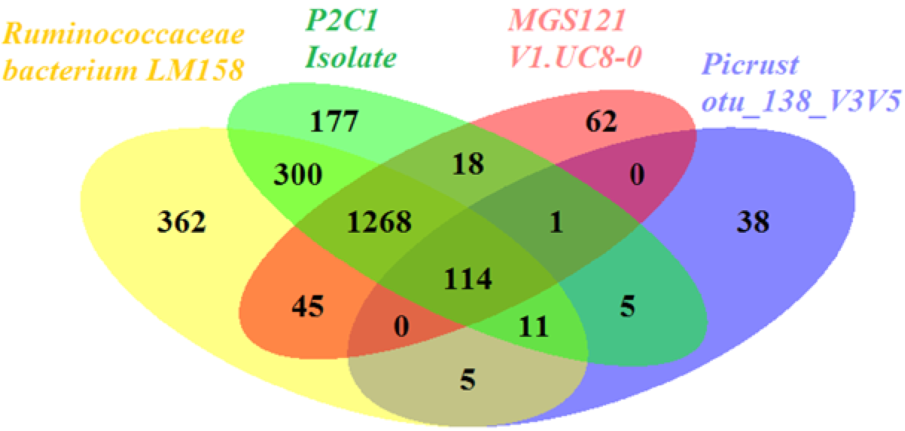



Catabacter hongkongiensis

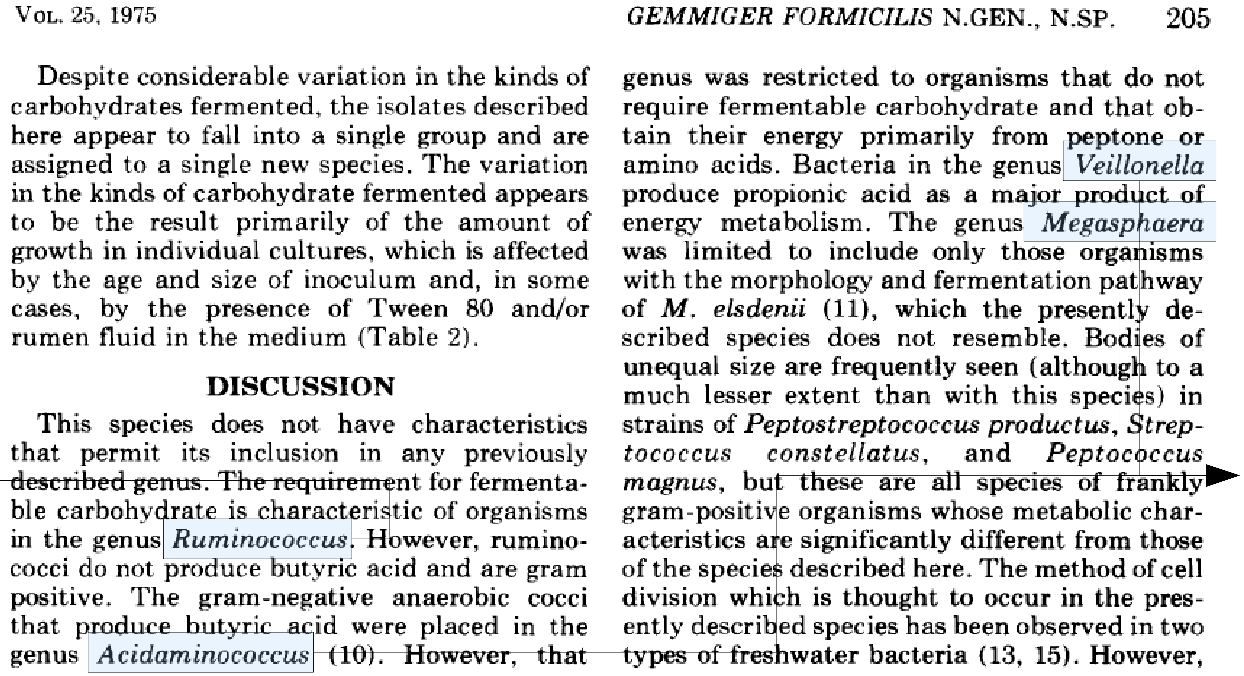
Christensenella minuta

Christensenellaceae

Parabacteroides_johnsonii_M-165
Bacteroides_ovatus_JCM_5824
Prevotella_copri_JCM_13464
Alistipes_finegoldii_DSM_17242
196982=Clostridiales/Ruminococcaceae
4425669=Clostridiales/Lachnospiraceae/Coprococcus
otu_339_V3V5=MGS131
Coprococcus_eutactus_ATCC_27759
Blautia_coccoides_DSM_2088
192741=Clostridiales
Coprococcus_comes_ATCC_27758
Acetomicrobium_faecale_DSM_20678
Veillonella_parvula_JCM_12972
Catabacter_hongkongensis_HKU16
Christensenella_minuta_strain_YIT_12065
4402903=Clostridiales
176318=Clostridiales/Christensenellaceae
otu_428_V3V5=MGS159
Clostridium_butyricum_VPI3266
Peptostreptococcus_anaerobius_NCTC_11460
Acidaminobacter_hydrogenoformans_glu_65
otu_171_V3V5=MGS74
Eubacterium_siraeum_ATCC_29066
otu_271_V3V5=MGS177
Intestinimonas_butyriciproducens_SRB-521-5-I
Faecalibacterium_prausnitzii_ATCC_27768
otu_553_V3V5=MGS103
4465124=Clostridiales/Clostridiaceae/Clostridium
otu_268_V3V5=MGS110
otu_243_V3V5=MGS49
197988=Clostridiales/Ruminococcaceae/Oscillospira
otu_717_V3V5=MGS49
otu_640_V3V5=MGS87
Oscillibacter_valericigenes_Sjm18-20
Oscillibacter_ruminantium_GH1
819353=Clostridiales/Ruminococcaceae
4364405=Clostridiales/Ruminococcaceae
otu_250_V3V5=MGS121
otu_138_V3V5=MGS121
P2C1_Isolate
Ruminococcaceae_bacterium_LM158



Database correctness is still an issue



Despite considerable variation in the kinds of carbohydrates fermented, the isolates described here appear to fall into a single group and are assigned to a single new species. The variation in the kinds of carbohydrate fermented appears to be the result primarily of the amount of growth in individual cultures, which is affected by the age and size of inoculum and, in some cases, by the presence of Tween 80 and/or rumen fluid in the medium (Table 2).

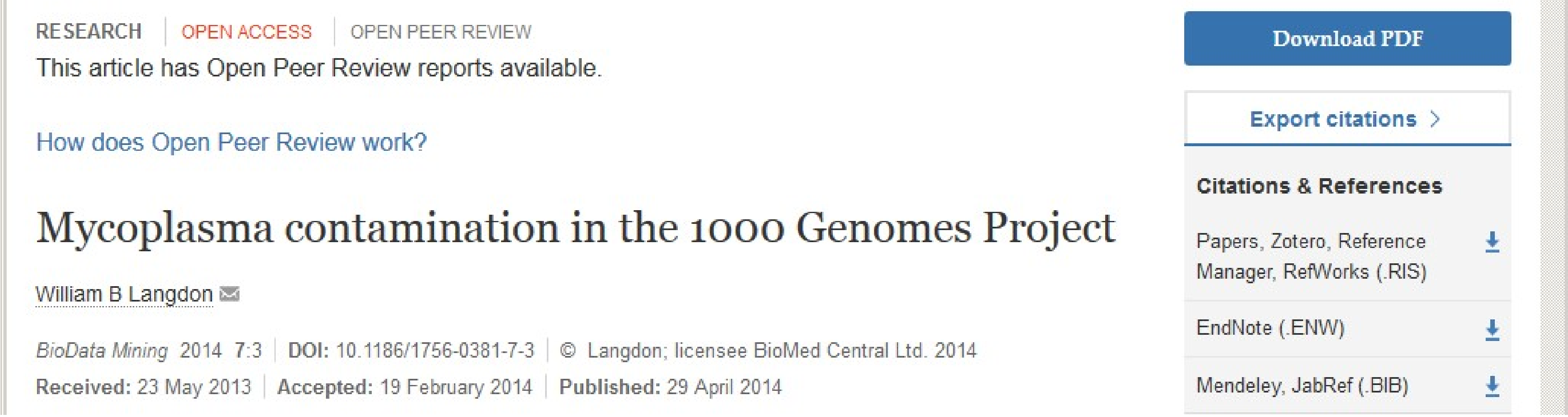
DISCUSSION

This species does not have characteristics that permit its inclusion in any previously described genus. The requirement for fermentable carbohydrate is characteristic of organisms in the genus *Ruminococcus*. However, ruminococci do not produce butyric acid and are gram positive. The gram-negative anaerobic cocci that produce butyric acid were placed in the genus *Acidaminococcus* (10). However, that genus was restricted to organisms that do not require fermentable carbohydrate and that obtain their energy primarily from peptone or amino acids. Bacteria in the genus *Veillonella* produce propionic acid as a major product of energy metabolism. The genus *Megasphaera* was limited to include only those organisms with the morphology and fermentation pathway of *M. elsdenii* (11), which the presently described species does not resemble. Bodies of unequal size are frequently seen (although to a much lesser extent than with this species) in strains of *Peptostreptococcus productus*, *Streptococcus constellatus*, and *Peptococcus magnus*, but these are all species of frankly gram-positive organisms whose metabolic characteristics are significantly different from those of the species described here. The method of cell division which is thought to occur in the presently described species has been observed in two types of freshwater bacteria (13, 15). However,

Bacteria;
Firmicutes;
Clostridia;...

Bacteria;
Firmicutes;
Negativicutes;
Selenomonadales;

Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Hyphomicrobiaceae;Gemmiger; Gemmiger formicilis



RESEARCH | OPEN ACCESS | OPEN PEER REVIEW

This article has Open Peer Review reports available.

How does Open Peer Review work?

Mycoplasma contamination in the 1000 Genomes Project

William B Langdon

BioData Mining 2014 7:3 | DOI: 10.1186/1756-0381-7-3 | © Langdon; licensee BioMed Central Ltd. 2014

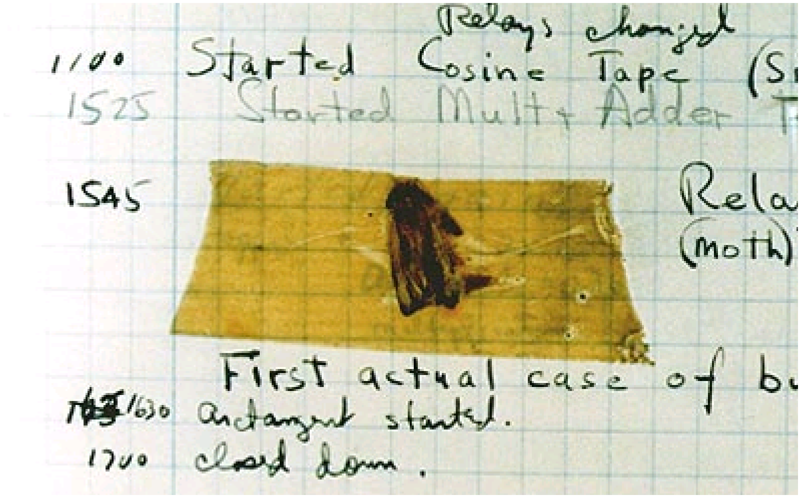
Received: 23 May 2013 | Accepted: 19 February 2014 | Published: 29 April 2014

Important future/continuing challenges

Dealing with errors

- Algorithmic:
  - Incorrect reconstructions/predictions
  - Missing information

- Software errors
  - 15-50 bugs/1000 lines of code
  - Celera Assembler – 300,000 loc



Relays changed

1100 Started Cosine Tape (S

1525 Started Mult+ Adder T

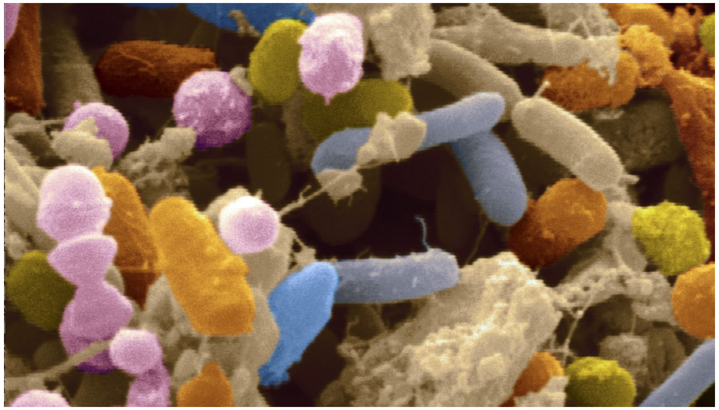
1545 Rela (moth)

First actual case of b

~~143~~/1630 Arctangent started.

1700 closed down.

Computationally modeling biology
... while not ignoring the biology



!= **1011000101000101011011**

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*

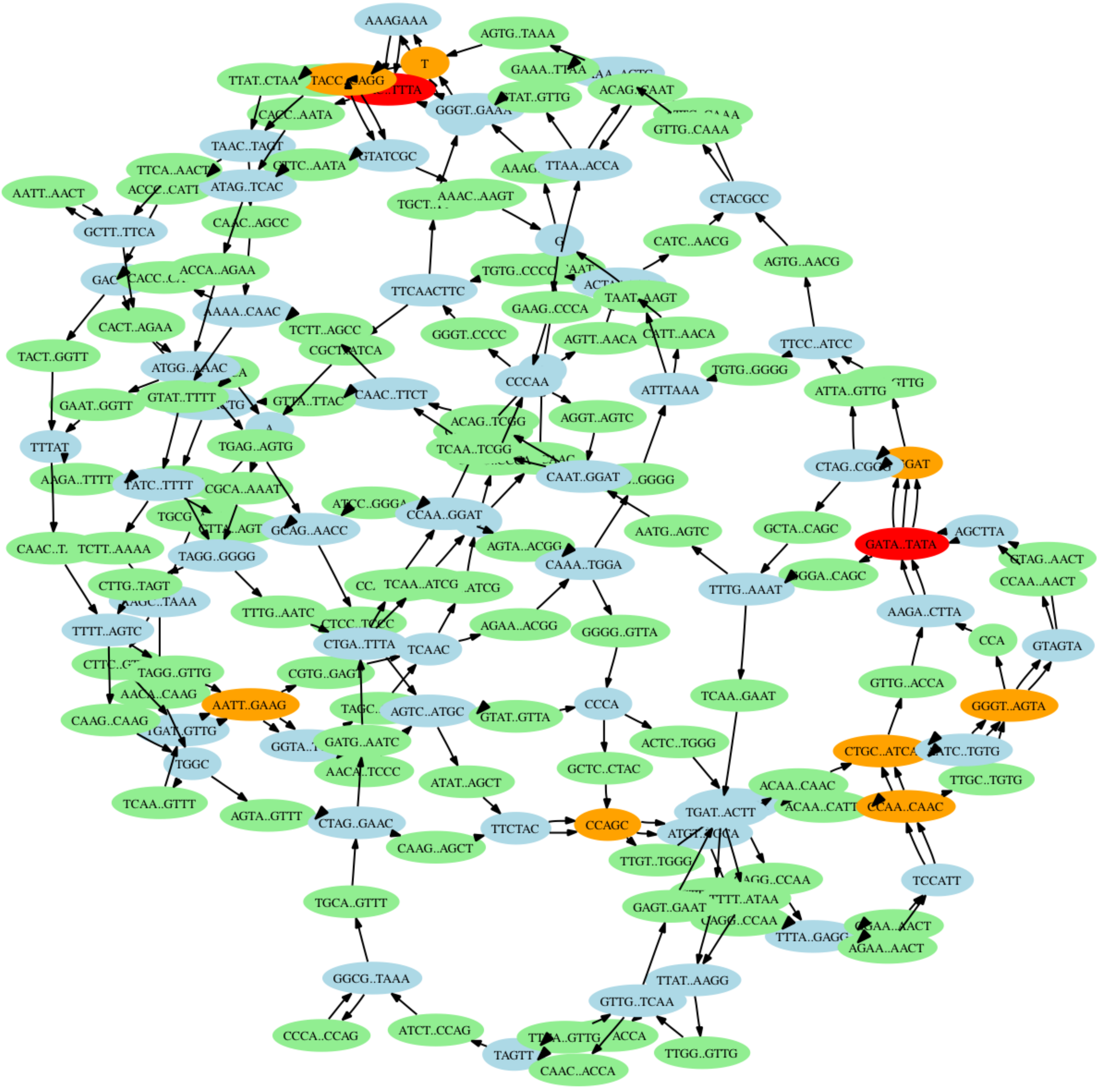
of wisdom **it was**

it was the worst

the *age of* wisdom

of times **it was**

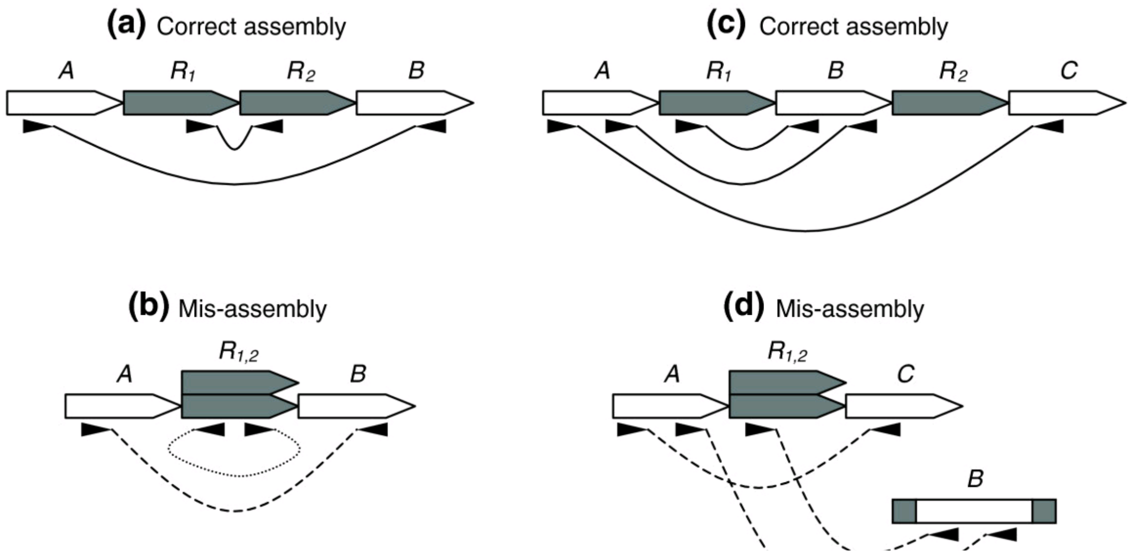
the *age of* foolishness

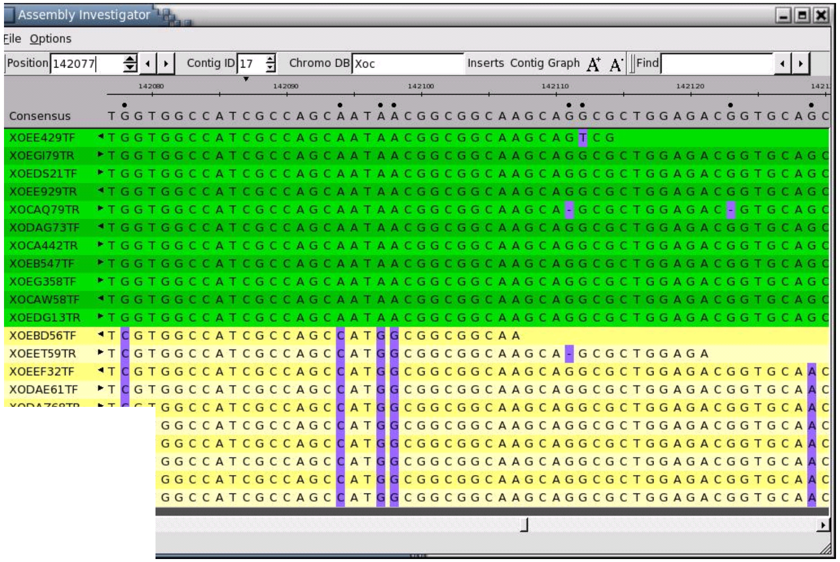


Mycoplasma genitalium, 25 bp reads

Kingsford et al., BMC Bioinformatics 2010

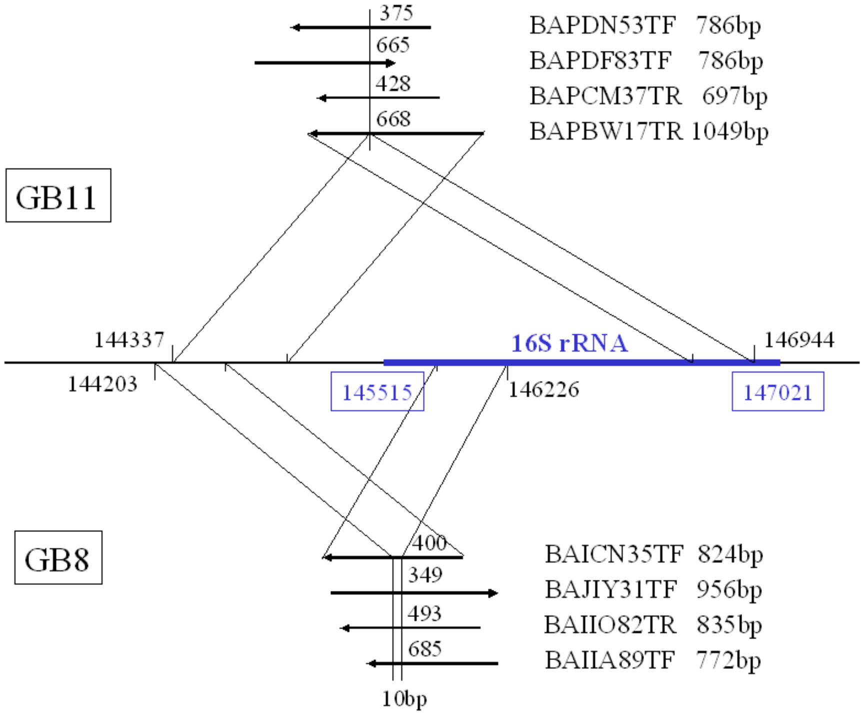
Is my assembly correct?





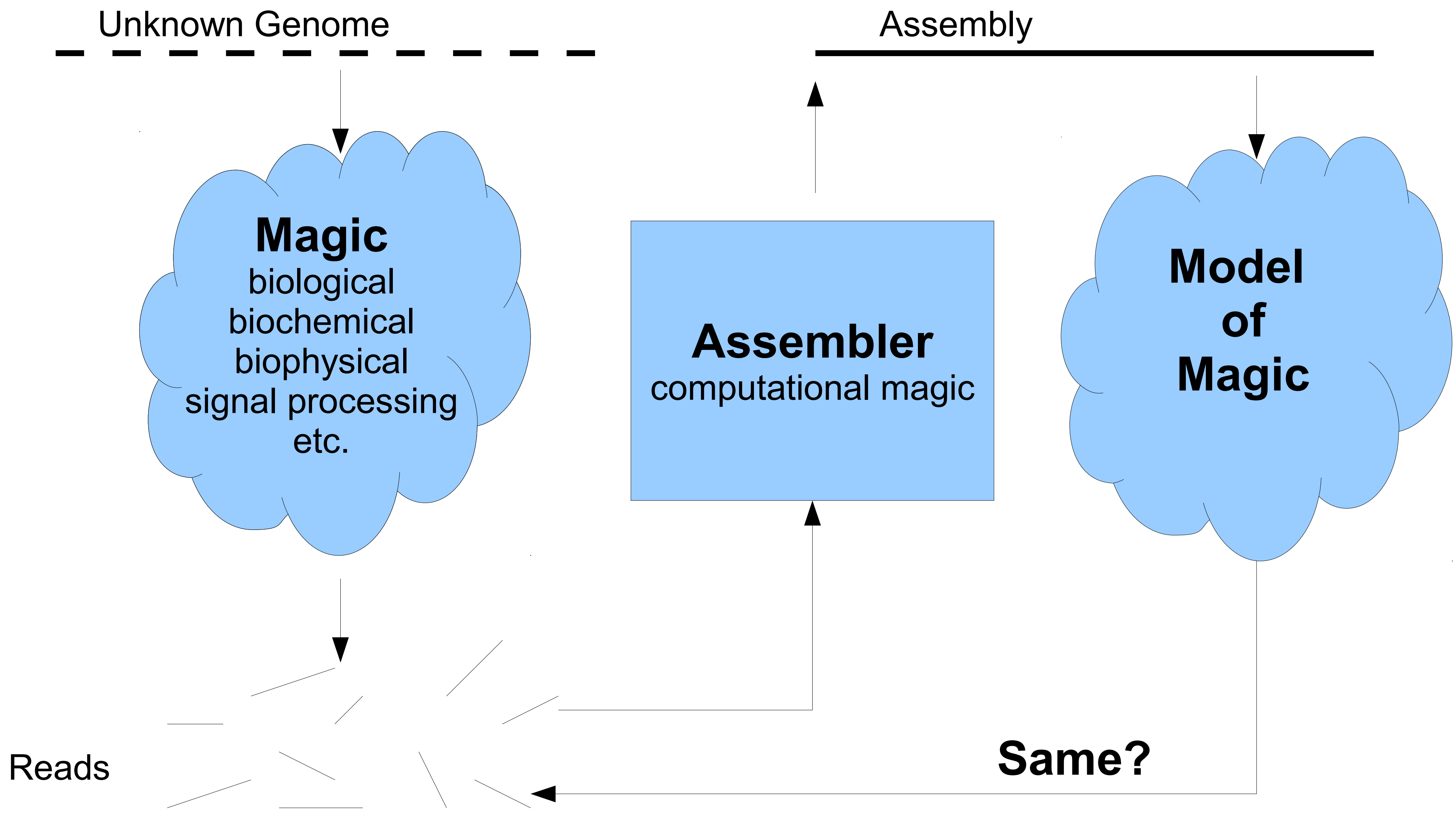
Position 142077 Contig ID 17 Chromo DB Xoc

```
            142080    142090    142100    142110    142120
Consensus     TGGTGGCCATCGCCAGCAATAACGGCGGCAAGCAGGCGCTGGAGACGGTGCAGC
XOEE429TF   ◂ TGGTGGCCATCGCCAGCAATAACGGCGGCAAGCAGTCG
XOEGI79TR   ▸ TGGTGGCCATCGCCAGCAATAACGGCGGCAAGCAGGCGCTGGAGACGGTGCAGC
XOEDS21TF   ▸ TGGTGGCCATCGCCAGCAATAACGGCGGCAAGCAGGCGCTGGAGACGGTGCAGC
XOEE929TR   ◂ TGGTGGCCATCGCCAGCAATAACGGCGGCAAGCAGGCGCTGGAGACGGTGCAGC
XOCAQ79TR   ▸ TGGTGGCCATCGCCAGCAATAACGGCGGCAAGCA-GCGCTGGAGAC-GTGCAGC
XODAG73TF   ◂ TGGTGGCCATCGCCAGCAATAACGGCGGCAAGCAGGCGCTGGAGACGGTGCAGC
XOCA442TR   ▸ TGGTGGCCATCGCCAGCAATAACGGCGGCAAGCAGGCGCTGGAGACGGTGCAGC
XOEB547TF   ▸ TGGTGGCCATCGCCAGCAATAACGGCGGCAAGCAGGCGCTGGAGACGGTGCAGC
XOEG358TF   ▸ TGGTGGCCATCGCCAGCAATAACGGCGGCAAGCAGGCGCTGGAGACGGTGCAGC
XOCAW58TF   ◂ TGGTGGCCATCGCCAGCAATAACGGCGGCAAGCAGGCGCTGGAGACGGTGCAGC
XOEDG13TR   ▸ TGGTGGCCATCGCCAGCAATAACGGCGGCAAGCAGGCGCTGGAGACGGTGCAGC
XOEBD56TF   ◂ TCGTGGCCATCGCCAGCCATGGCGGCGGCAA
XOEET59TR   ▸ TCGTGGCCATCGCCAGCCATGGCGGCGGCAAGCA-GCGCTGGAGA
XOEEF32TF   ◂ TCGTGGCCATCGCCAGCCATGGCGGCGGCAAGCAGGCGCTGGAGACGGTGCAAC
XODAE61TF   ▸ TCGTGGCCATCGCCAGCCATGGCGGCGGCAAGCAGGCGCTGGAGACGGTGCAAC
XODA768TR   ▸ TCGTGGCCATCGCCAGCCATGGCGGCGGCAAGCAGGCGCTGGAGACGGTGCAAC
                 GGCCATCGCCAGCCATGGCGGCGGCAAGCAGGCGCTGGAGACGGTGCAAC
                 GGCCATCGCCAGCCATGGCGGCGGCAAGCAGGCGCTGGAGACGGTGCAAC
                 GGCCATCGCCAGCCATGGCGGCGGCAAGCAGGCGCTGGAGACGGTGCAAC
                 GGCCATCGCCAGCCATGGCGGCGGCAAGCAGGCGCTGGAGACGGTGCAAC
                 GGCCATCGCCAGCCATGGCGGCGGCAAGCAGGCGCTGGAGACGGTGCAAC
```



Work with Chris Hill, Atif Memon

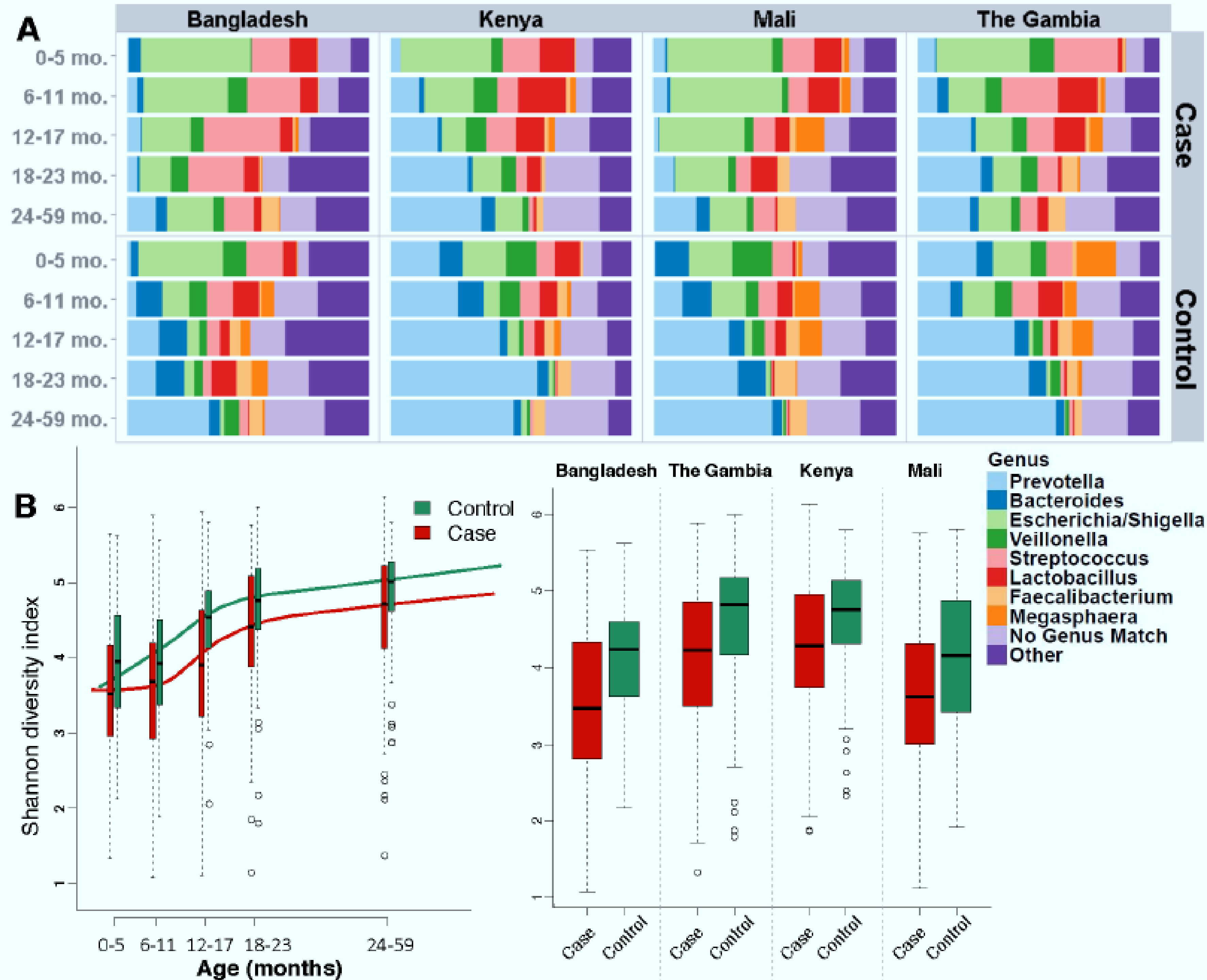
Model-based testing



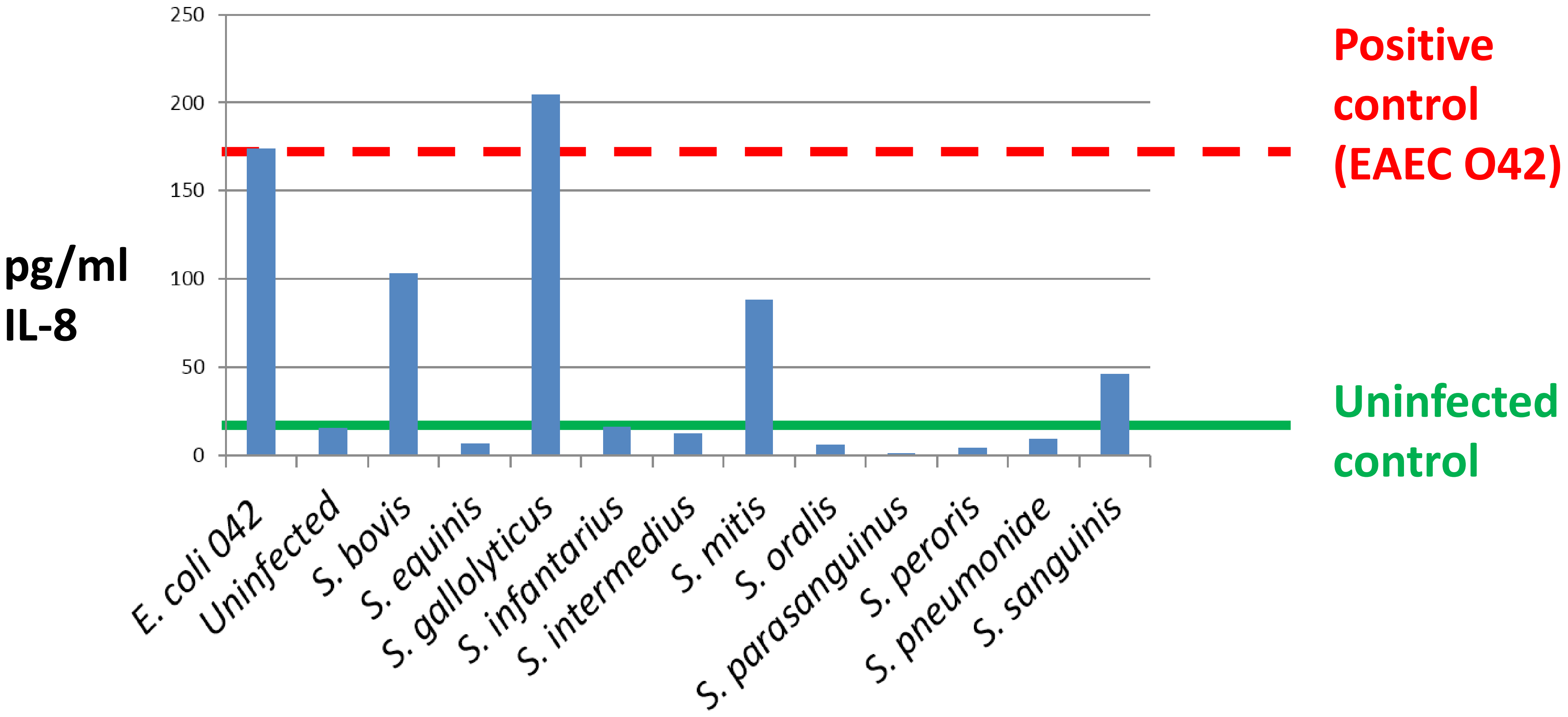
Work with Mohammad Ghodsi, Chris Hill, Bo Liu, Todd Treangen, Irina Astrovskaya

Back to biology

Impact of diarrhea on microbiota

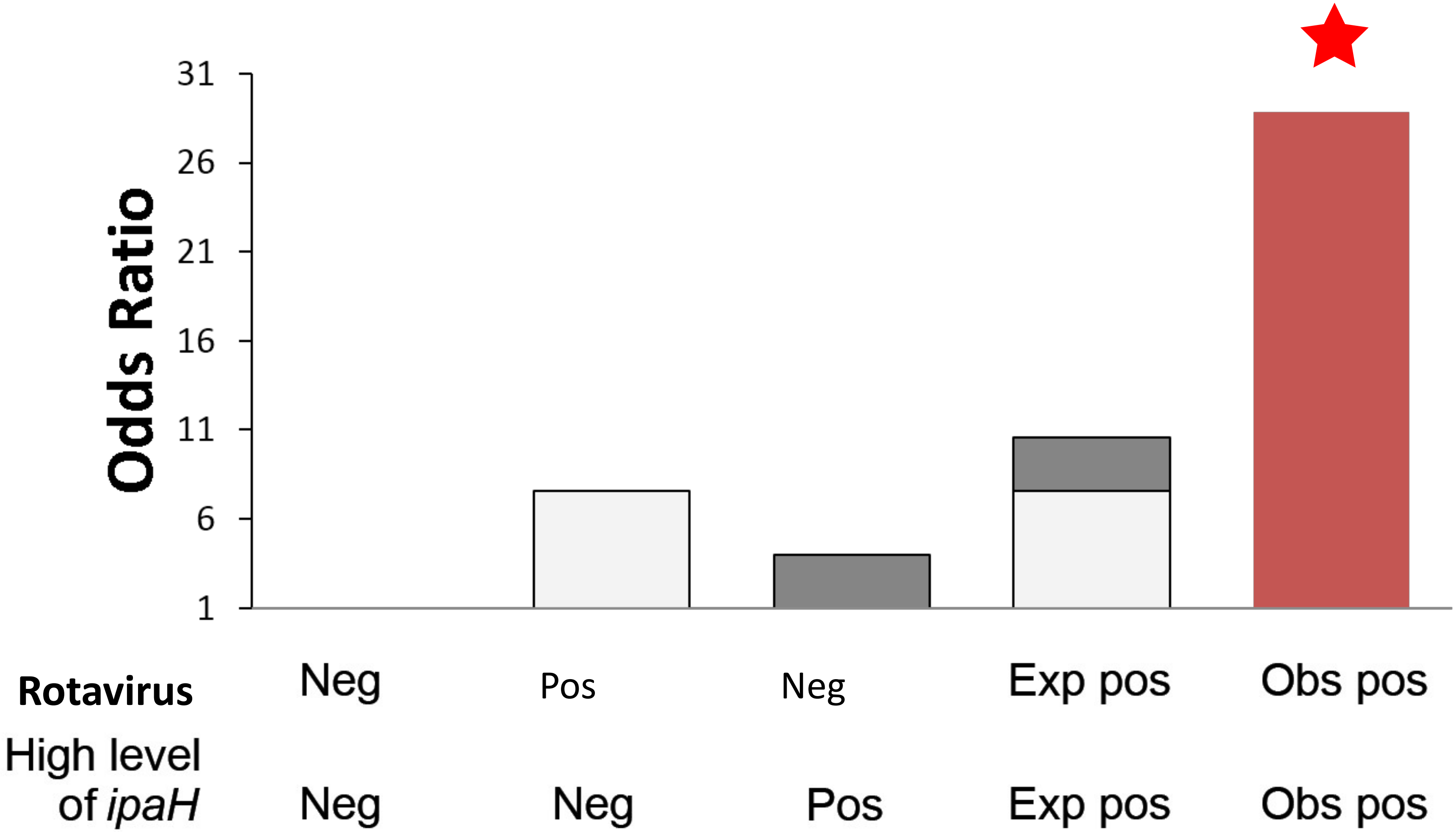


Polarized human colonic (T84) monolayers reveal variation in injurious behavior for streptococcal isolates



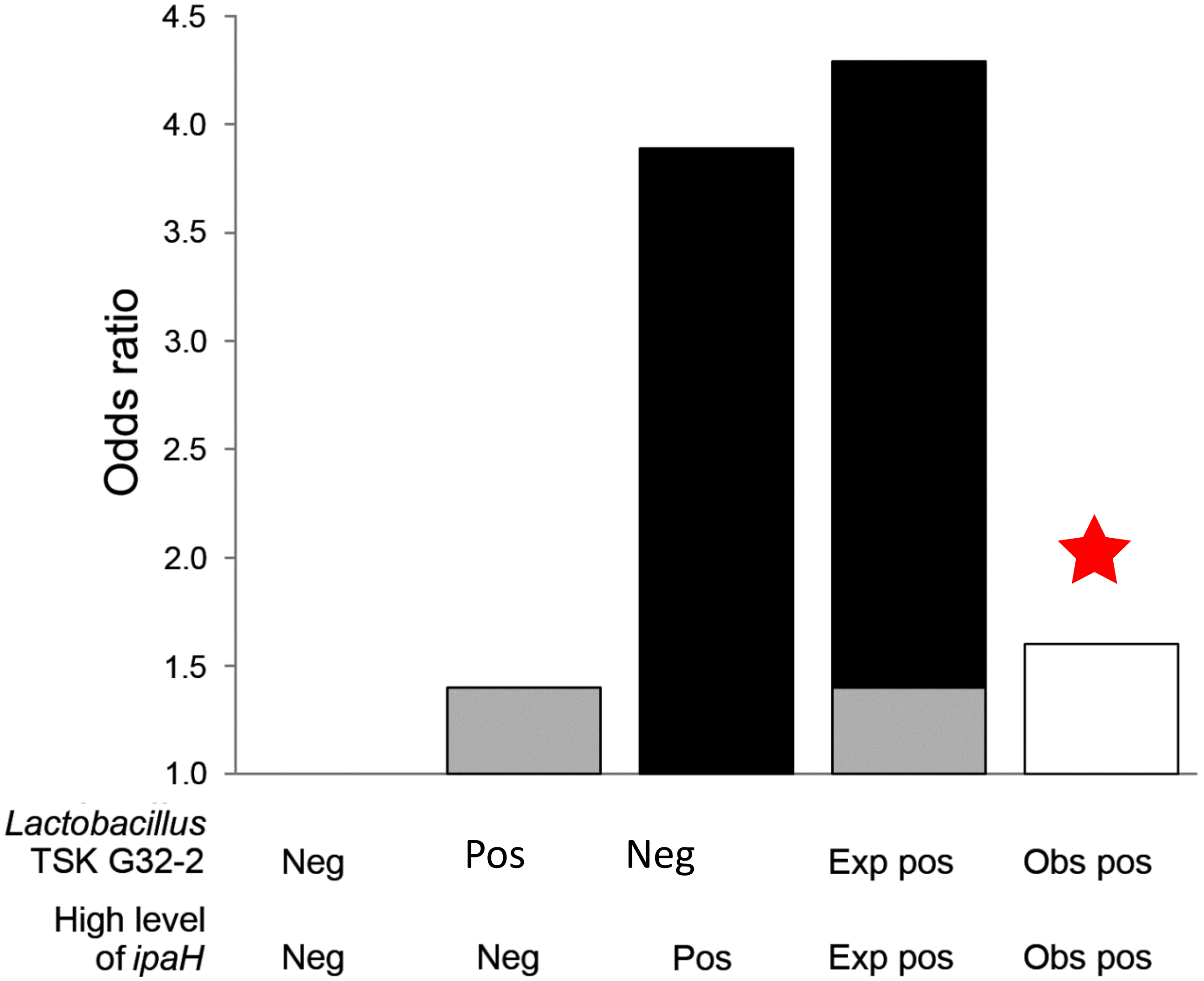
Streptococcal isolates incubated with polarized T84 monolayers at 37C for 3 hr; IL-8 release measured by EIA. Results of triplicates

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

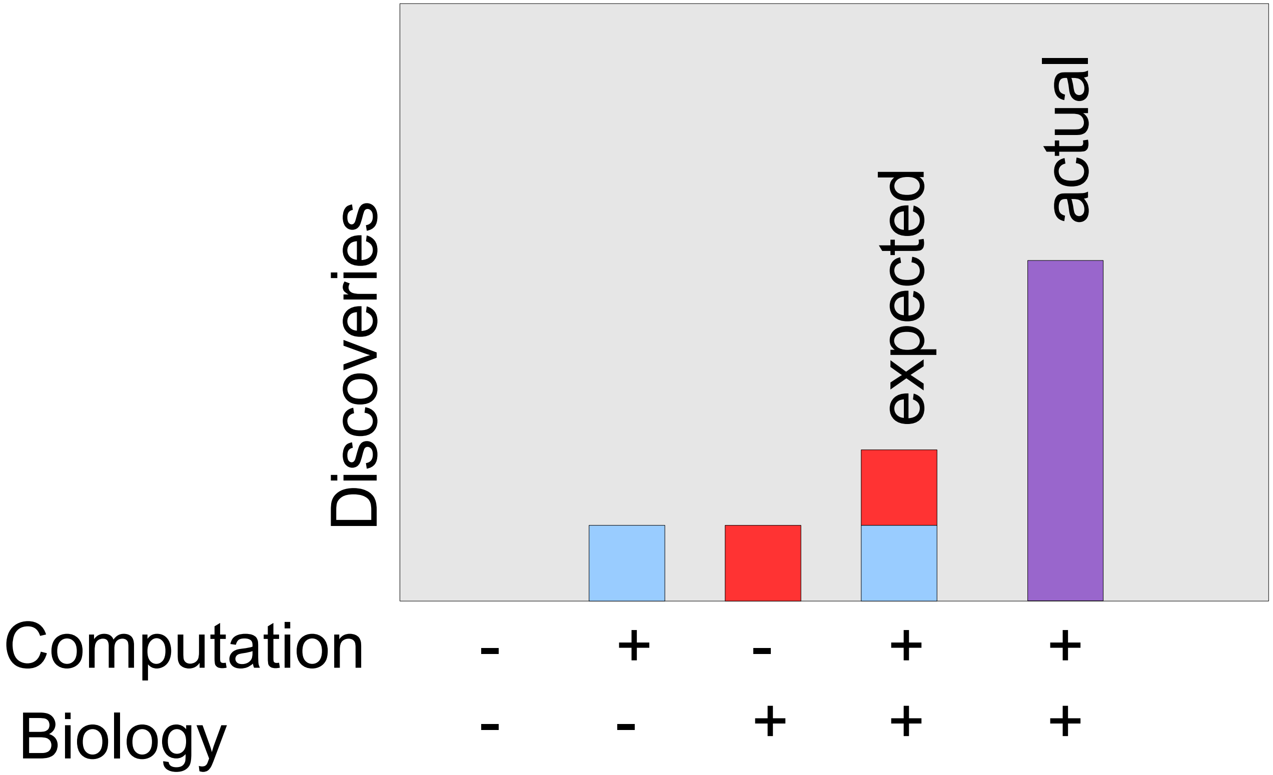


★ Significant increase in OR by factor >2

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



★ Significant reduction in OR by factor >2


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

Acknowledgments

Grainger Initiative
Tandy Warnow

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.

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

Does anyone really believe that data mining could produce the general theory of relativity?

Ed Daugherty, Michael Bittner
Epistemology of the cell, 2011