Clinical metagenomics

Nick Loman
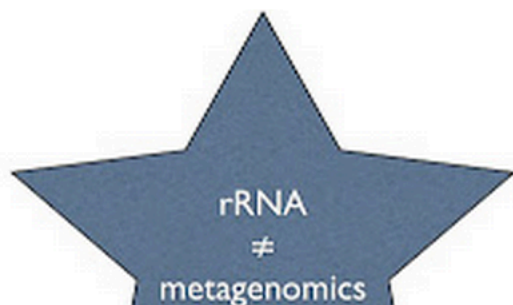
Referring to 16S surveys as "metagenomics" is misleading and annoying #badomics #OmicMimicry

By Jonathan Eisen 8/22/2012 12:12:00 AM

+16   Recommend this on Google

rRNA ≠ metagenomics

Aargh. I am a big fan if ribosomal RNA based surveys of microbial diversity. Been doing them for 20+ years and still continue to - even though I have moved on to more genomic/metagenomic based studies. But it drives me crazy to see rRNA surveys now being called "metagenomics".

You're probably not doing metagenomics

9 Replies

Just to begin, I'd like to say that I'm right about this, and if you think I am wrong, I'm not – you are.

The genome of an organism is the entire complement of genes within an organism's cells, and genomics is therefore the study of entire genomes. Metagenomics refers to the study of all genomes within a particular ecosystem, or group of individuals. Metagenomics therefore refers to studies where entire genomes are assayed.

Jonathan Eisen
Mick Watson

16S vs metagenomics

- Cheap
- Targets single marker gene
- Limited to bacteria
- Relatively easy to analyse
- Lots of known biases
- Taxonomic assignment at species level problematic
- Function can only be inferred, not detected
- Goes deeper

- Expensive
- In theory can detect anything
- Harder to analyse
- Fewer biases (?)
- Function information directly accessible
- Strain-level information
- Shallower

Definition of a **metagenome**

- The collection of genomes and genes from the members of a microbiota

- Microbiota: The assemblage of microorganisms present in a defined environment.

- Microbiome: This term refers to the entire habitat, including the microorganisms, their genomes (i.e., genes) and the surrounding environmental conditions.

# Metagenomics – Your questions

What's the big idea?

- Who is there? (Taxonomic assignments)
- What are they doing? (Functional analysis)
  - What are they capable of doing? (DNA)
- Who is doing what? (Genome reconstruction)

Functional signatures are not the same as taxonomic signatures

Figure 2: Carriage of microbial taxa varies while metabolic pathways remain stable within a healthy population.
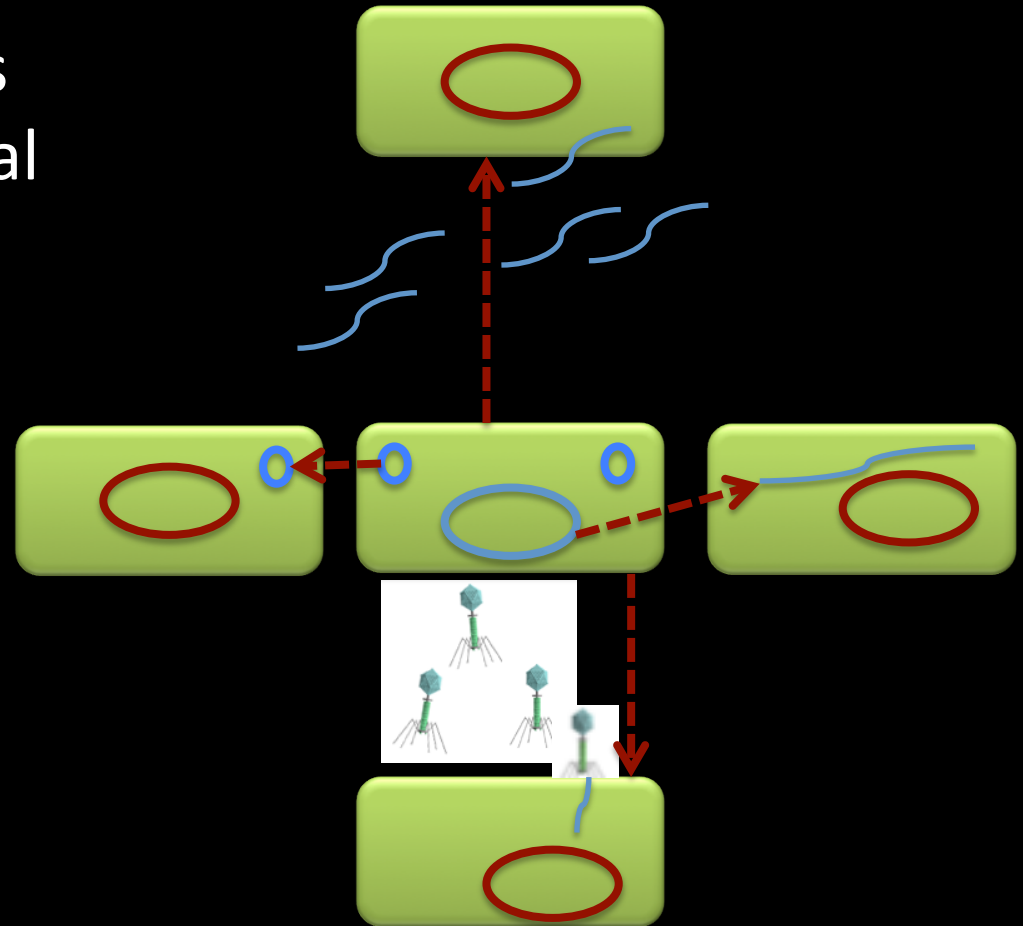

The Human Microbiome Project Consortium, doi:10.1038/nature11234
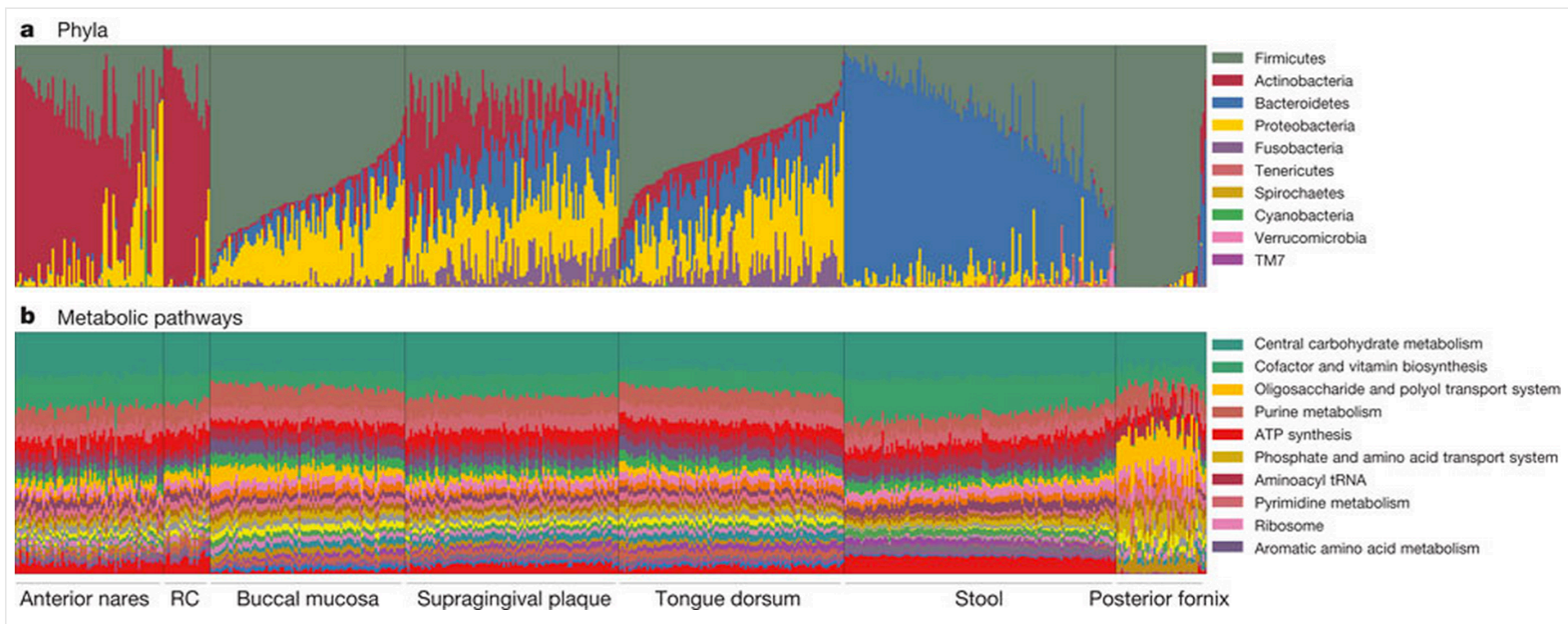
E. coli: more genes than humans?
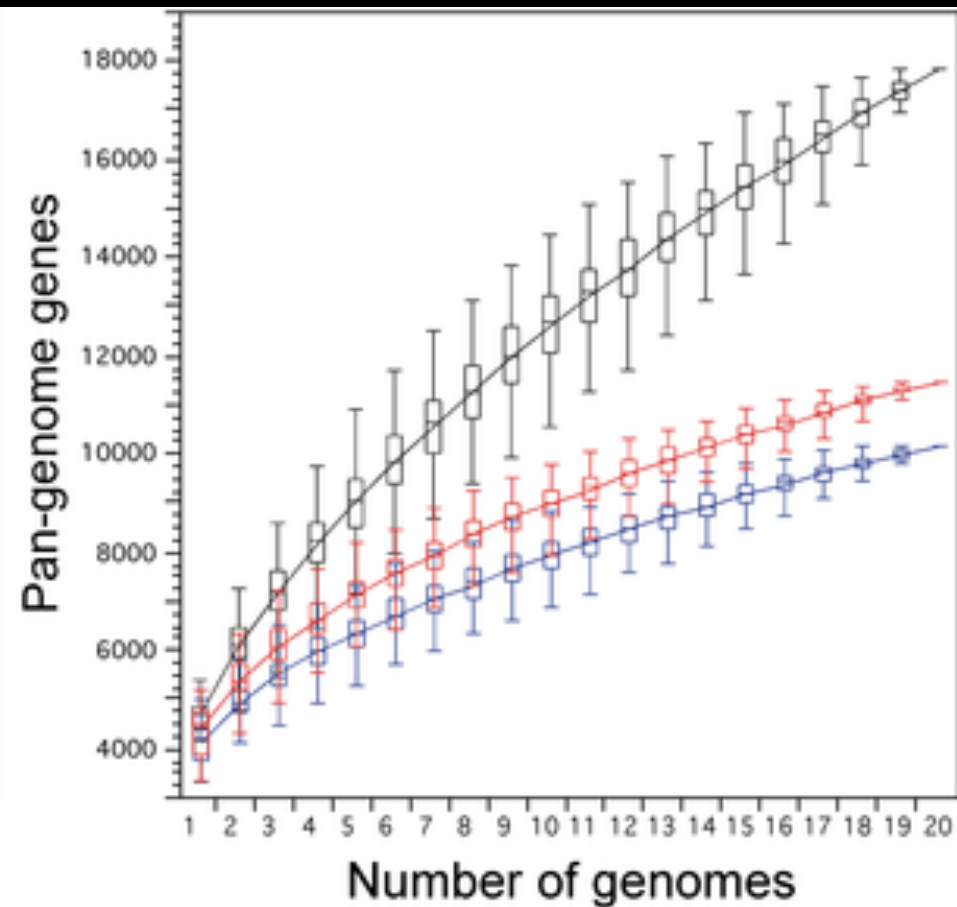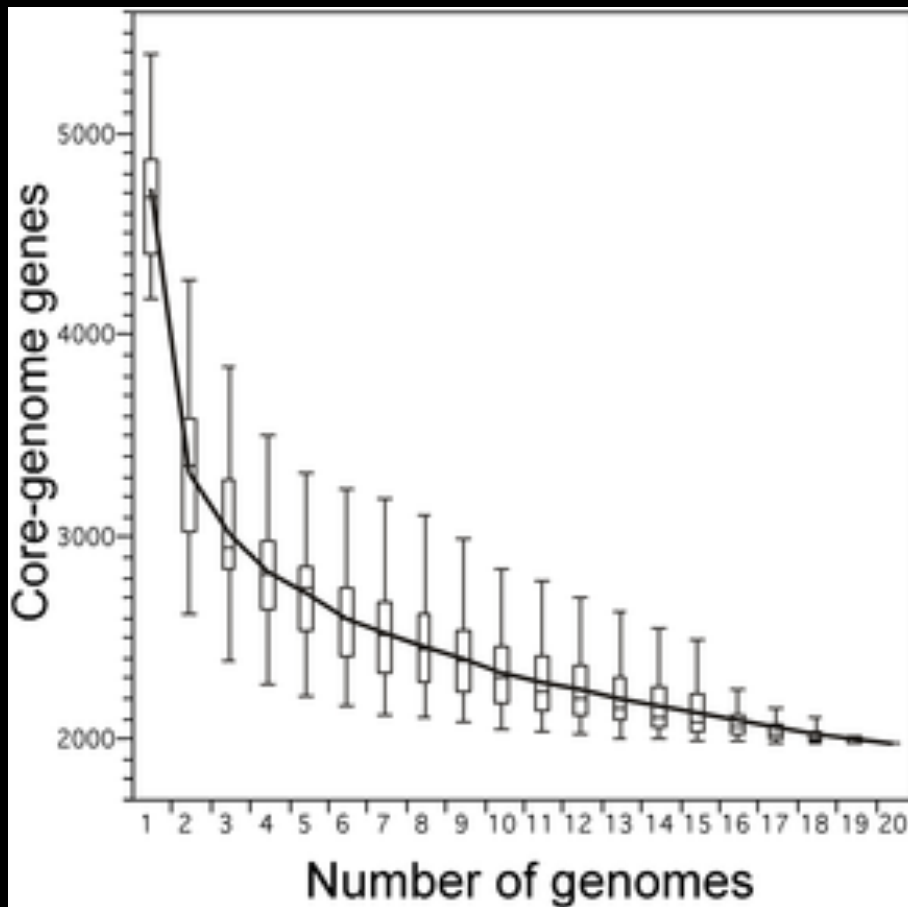
Genetics of virulence

- Many virulence genes acquired via horizontal gene transfer

- On plasmids or chromosome via conjugation

- As naked DNA via transformation

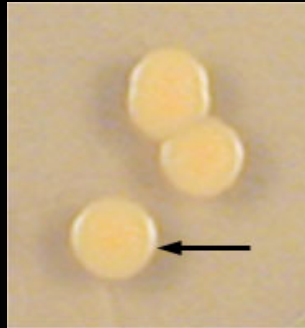- On bacteriophage via transduction

Digital microbiology:

Is high-throughput sequencing a match for Koch and Pasteur?

Diagnostic microbiology
21st Century problem, but 19th Century techniques!
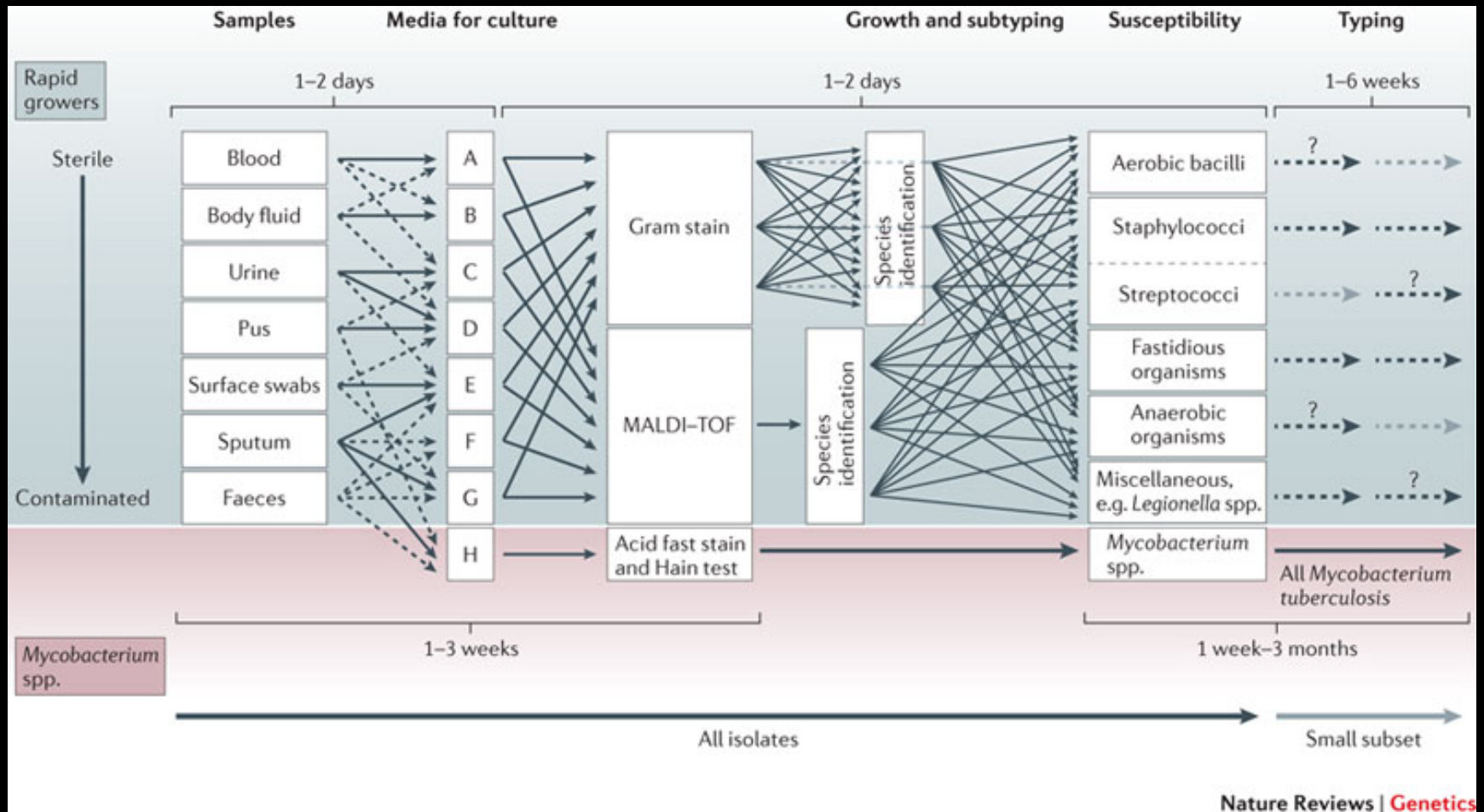
Current Approaches

- Microscopy and culture techniques that date from the time of Koch and Pasteur
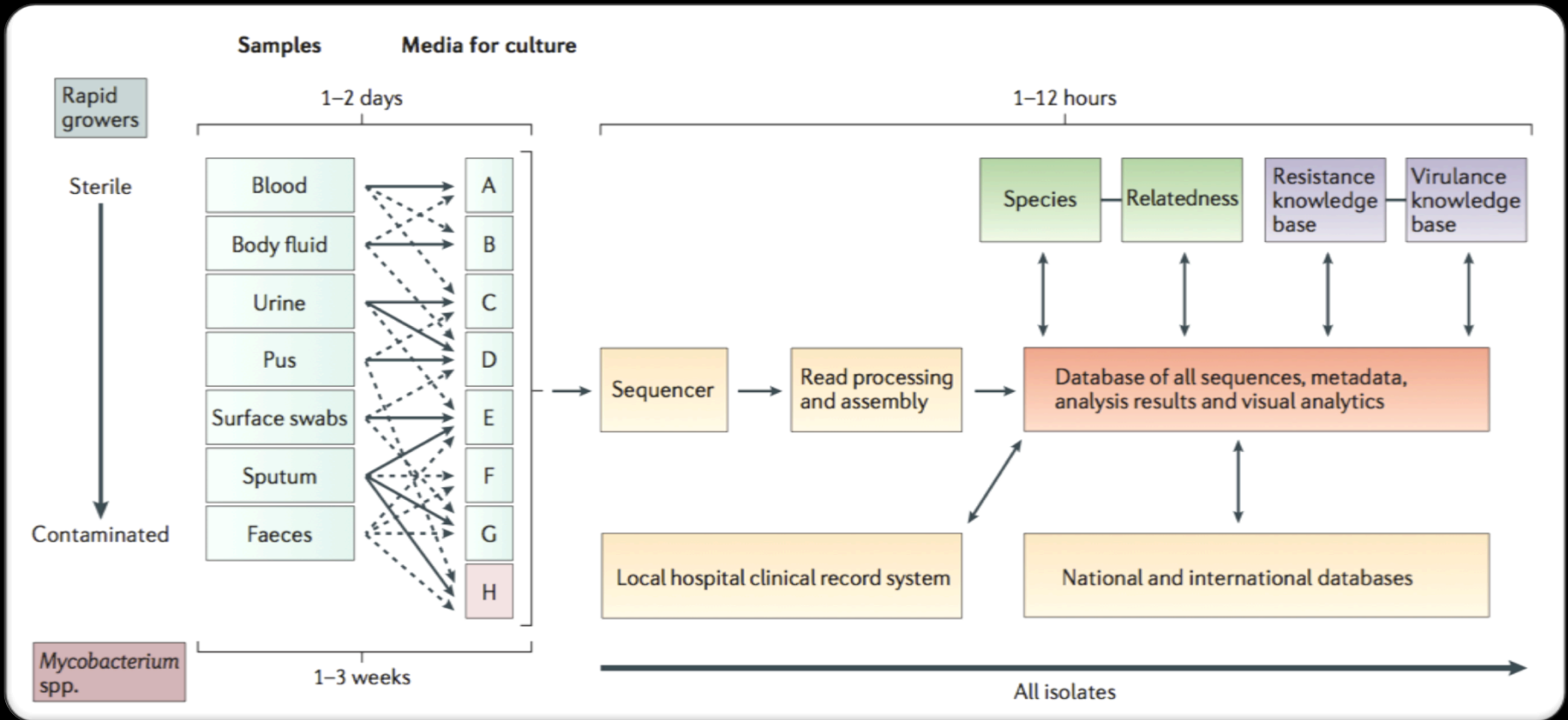

Future Vision

- Digital microbiology
- High-throughput sequencing becomes method of choice in diagnostic microbiology

Clinical microbiology workflow

Didelot, Crook *et al.* PMID 22868263

Digital microbiology?

Didelot, Crook *et al.* PMID 22868263

A New Opportunity: High-Throughput Sequencing

- Brings the advantages of
  - open-endedness (revealing the "unknown unknowns"),
  - universal applicability
  - ultimate in resolution
- Bench-top sequencing platforms now generate data sufficiently quickly and cheaply to have an impact on real-world clinical and epidemiological problems

Mark Pallen
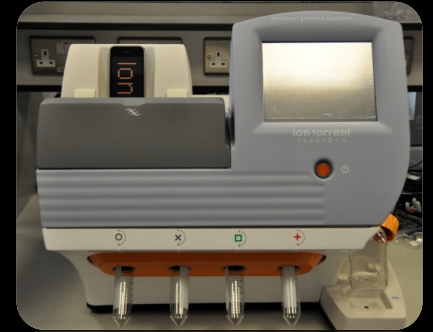
Costs

Application	Library Cost	Sequencing cost
Whole human genome (30x coverage)	£25	£1000 - £5000
Whole bacterial genome (Illumina)	£25	£25
16S phylogenetic profiling	£2	£1 - 10
Metagenome (2Gb per sample)	£20	£750

Whole genome sequencing to track the spread of *Pseudomonas aeruginosa* within a burns unit

- Gram negative bacterium

- Opportunistic pathogen in burns patients

- Infection can lead to graft breakdown and sepsis

- Isolated from 30% of burns patients

Nicola Cumley

Water as a potential source of *Pseudomonas*

- Sources of *Pseudomonas* infection
  - Endogenous
  - Cross infection
  - **WATER**
- Outbreaks linked to contaminated water
- Showering important part of burns care

A DROP OF LONDON WATER.

Study: Surveillance in burns patients

- Aim: to determine relative contribution of:
  - Endogenous infection
  - Cross-infection
  - The water supply

  - Use whole-genome sequencing for greatest typing resolution and to infer links between positive isolates

Collection of isolates of *Pseudomonas aeruginosa*

- **Screening on admission**
  - >7% burns
  - Wound swabs, stool and Urine for microbiology
    - Stool for molecular testing
  - Environmental sampling of the patients room and shower water

- **Recruited positive patients**
  - Wound swabs at each dressing change
  - Environmental sampling
  - Swabs/tissue for metagenomic analysis
  - Environmental sampling on discharge

# Isolates from the study fall into three main sequence types
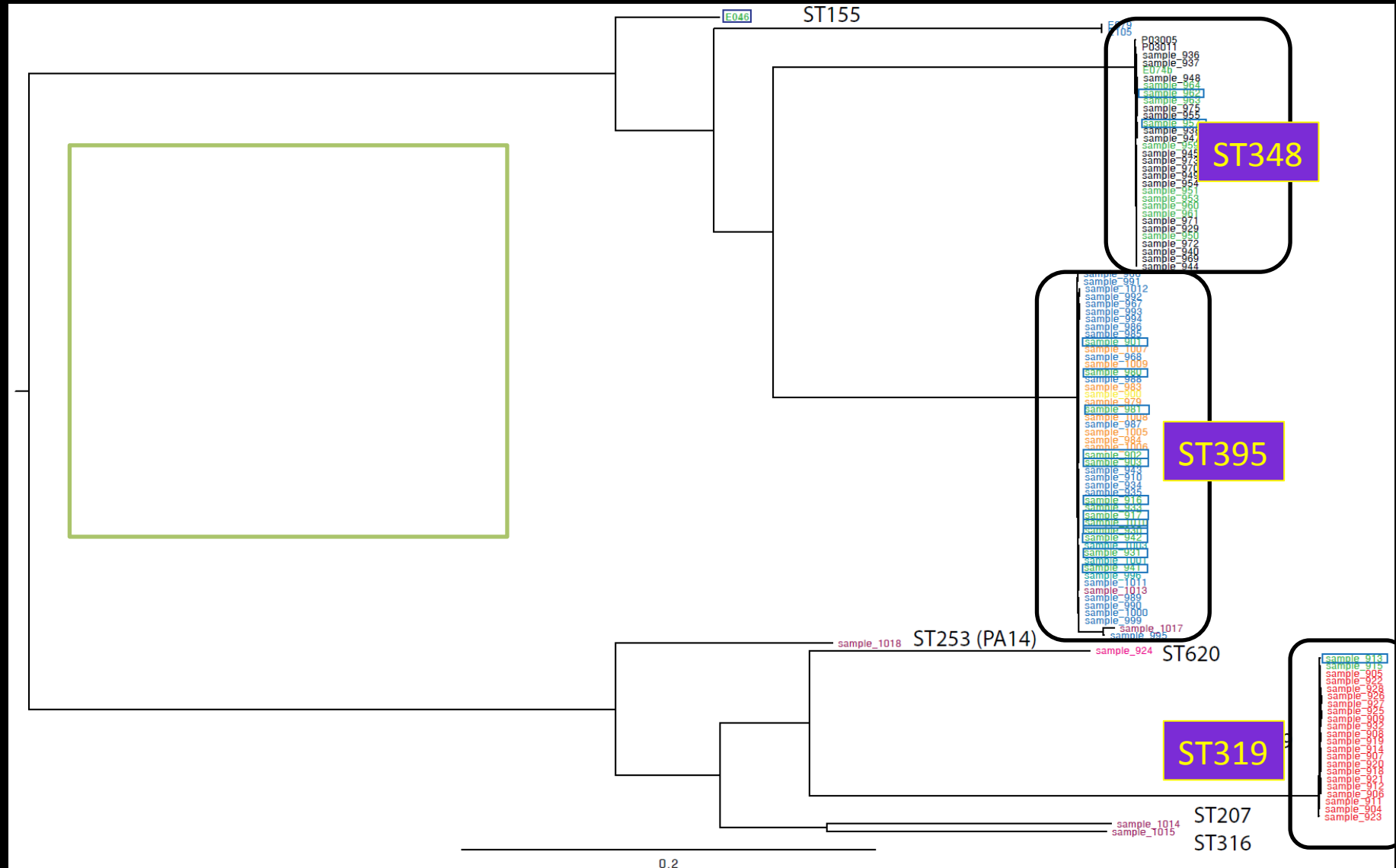
sample_966
sample_991
sample_1012
sample_992
sample_967
sample_993
sample_994
sample_986
sample_985
sample_901
sample_1007
sample_968
sample_1009
sample_980
sample_988
sample_983
sample_900
sample_979
sample_981
sample_1008
sample_987
sample_1005
sample_984
sample_1006
sample_902
sample_903
sample_943
sample_910
sample_934
sample_935
sample_916
sample_933
sample_917
sample_1010
sample_930
sample_942
sample_1003
sample_931
sample_1001
sample_941
sample_996
sample_1011
sample_1013
sample_989
sample_990
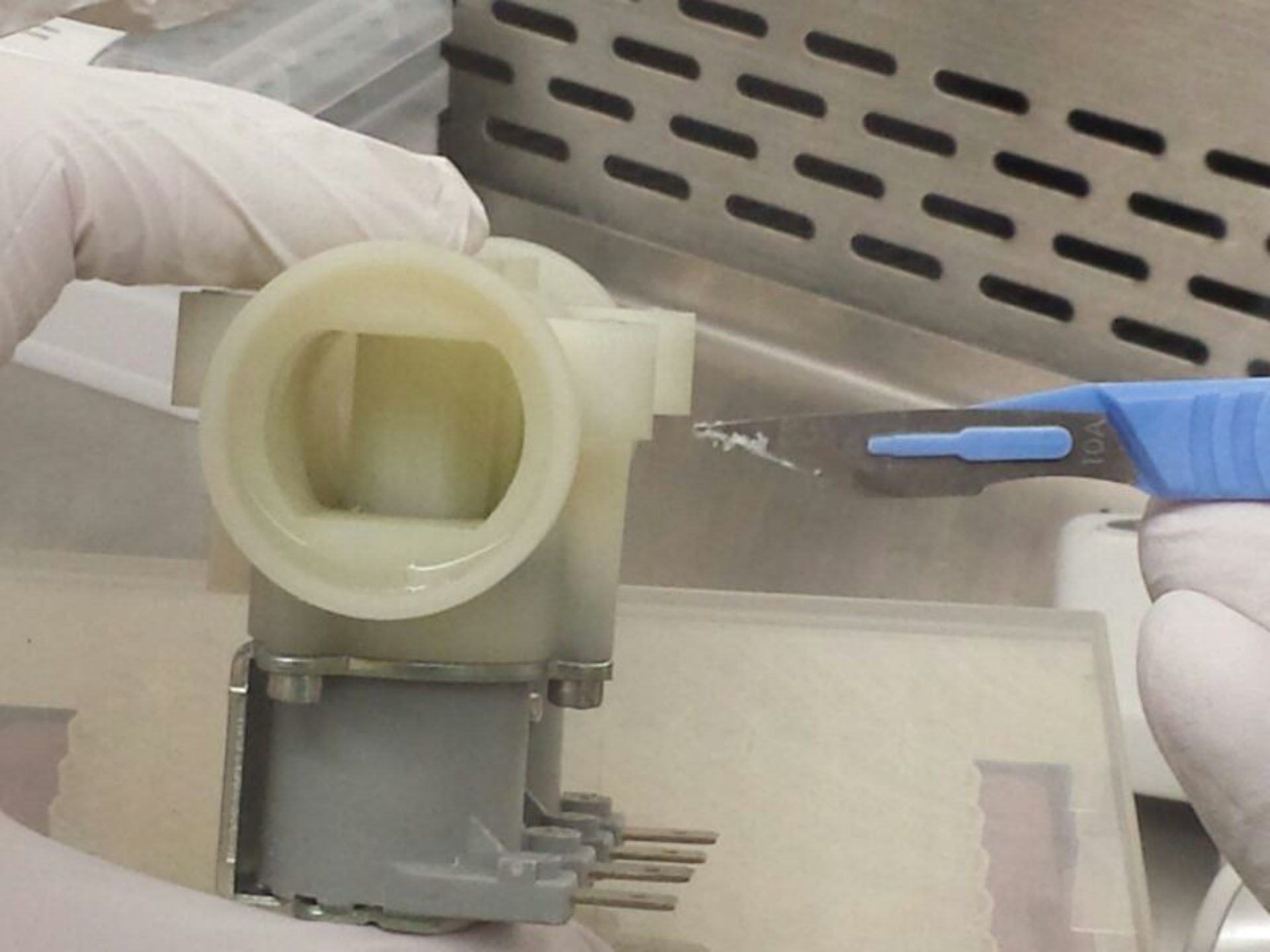sample_1000
sample_999
sample_1017
sample_995

ST395
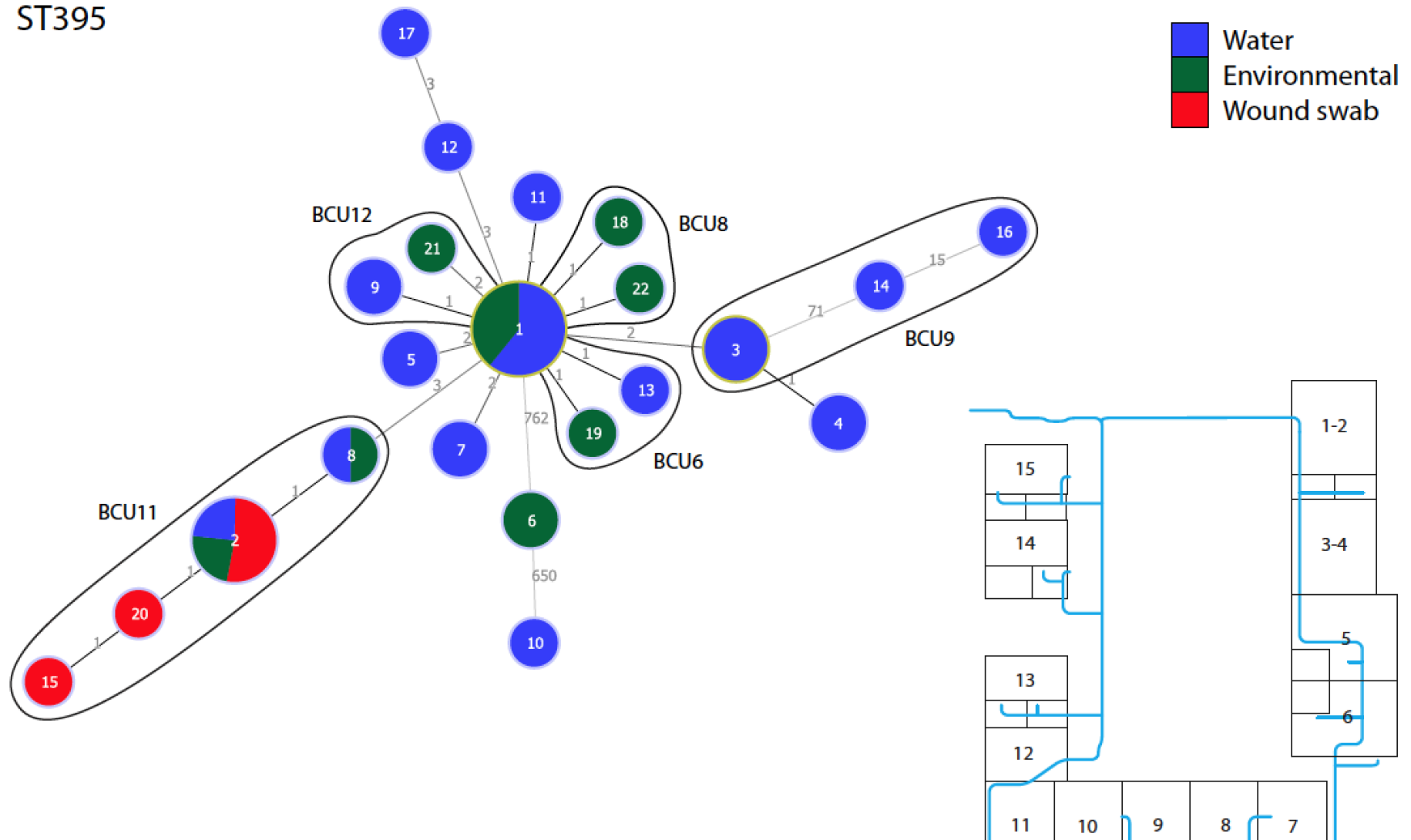
Metagenomics protocol

- Nextera XT protocol NOT according to manufacturer's protocols
- <1ng DNA input
- 1 MiSeq run (V3)
- Short fragments returned
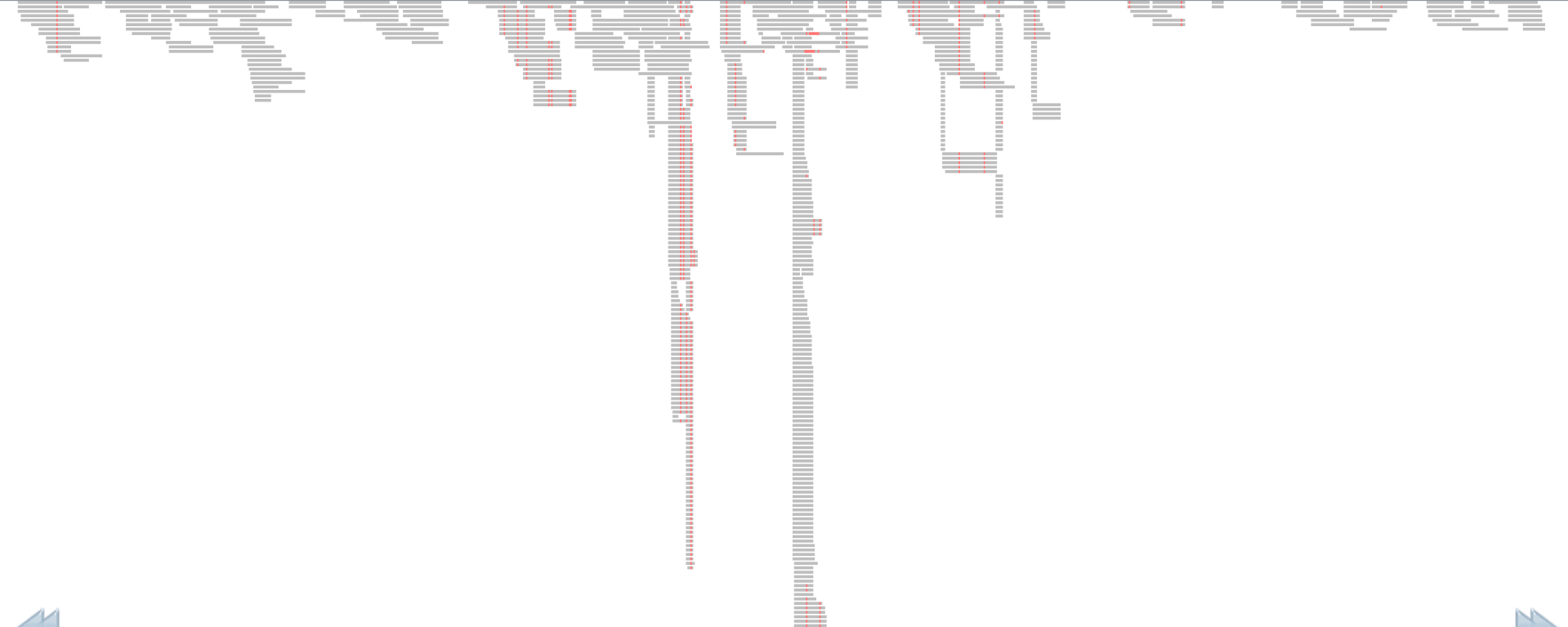- *P. aeruginosa* at ~5x coverage

Alignment to reference genome

70,785 U70,785 ... 79,328 U79,328

SNP typing pipeline

- Alignment of complete genomes to reference
- Trusted set of high quality SNPs
- ML tree (FastTree)
- Call variants from low coverage data
- Place on reference tree with pplacer

SNP typing from metagenomes

- 967 Shower BCU 9
- 1052 Shower BCU 9
- 992 Tap BCU 9
- 993 Shower BCU 9
- 994 Shower BCU 9

German *E. coli* O104:H4 outbreak aka **SPROUTBREAK**

- May-July 2011
- >4000 cases, >50 deaths in Germany
- Link to sprouting seeds
- Increased risk of haemolytic-uraemic syndrome (>25%)
- Females particularly at risk

Crowd-sourcing the genome

EHEC Genome Assembly

By Nick Loman on June 2, 2011

Keep track of the genomic analysis of the EHEC strains on our Github Wiki.

BGI have released 5 runs of Ion Torrent data for the German EHEC/VTEC outbreak strain. I hope it is released with no specific restrictions on use for the benefit of the entire community, but the site doesn't make that entirely clear. Thanks to the BGI for putting it up!

Shall we crowd source some analysis? This comes at a very timely moment as I am currently help organise the Applied Bioinformatics & Public Health conference in Hinxton (#ABPH11), where we are discussing the use of whole-genome sequencing in epidemiology. The problem is I don't have much time to dig into the data.

But I've put a first-pass de novo assembly up using MIRA (3.2.1.17_dev) here. 3,057 contigs, total bases: 5,491,032, N50 3,675. If you want the alignment files etc. get the big file here (282Mb).

37
tweets

retweet

Crowd-sourcing the genom

Search

[   ] [Search]

Profile

Mad rantings about politics, evolution, and microbiology. Comment policy: say what you want, but back it up with an email address. I don't like anonymous trolls.

I Don't Think the German Outbreak *E. coli* Strain Is Novel: Something Very Similar Was Isolated Ten Years Ago...

Category: E. coli • Genomics

Posted on: June 3, 2011 8:10 AM, by Mike

...in Europe. I'll get to that in a moment. You've probably heard of the *E. coli* outbreak sweeping through Germany and now other European countries that has caused over one thousand cases of hemolytic uremic syndrome ('HUS'). What's odd is that the initial reports are calling this a novel hybrid or some new strain of *E. coli*.

BGI has done some sequencing using Ion Torrent of one of these isolates, and Nick Loman assembled the data. Without getting too technical, the genome is actually in about 3,000 pieces, but with those data (and thanks to Nick for assembling them and releasing them) I was able to perform multilocus sequencing typing ('MLST'). Basically, we look at the partial sequences of several genes (in this case, seven) to identify its sequence type--think of it as a molecular barcode (for the scheme and details, see here).

So what did I find?

This EHEC strain is most likely a very close relative of ST678 (details in a bit). In fact, according to the mlst.net strain database, there is a strain "Jan-91", isolated in 2001* from Europe (no further geographic information is provided). That strain belongs to phylogroup D, and is associated with HUS...*just like the outbreak strain*. And the older strain also has the exact same serotype as the outbreak strain, O104:H4.

Analysis

- 2-Jun Nick Loman EHEC genome assembly
- 3-Jun Mike the Mad Biologist I Don't Think the German Outbreak E. coli Strain Is Novel: Something Very Similar Was Isolated Ago...
- 3-Jun Marina Manrique (Era7) Automatic annotation of E. coli TY2482
- 3-Jun Simon Gladman Automatic annotation of E. coli TY2482
- 4-Jun Kat Holt Two strain SNP comparison
- 5-Jun David Studholme Comparisons of E. coli TY2482 against previously sequenced E. coli genomes
- 5-Jun Kat Holt EHEC genomes – plasmid
- 5-Jun Konrad Paszkiewicz TY2482,-LB226692-vs-Genbank-Ecoli
- 5-Jun Phylogeo HUSEC41 German outbreak strains are not that 'new'
- 5-Jun Mariam Rizkallah Automatic annotations with RAST
- 5-Jun Kat Holt Comparative genomics - acquired material
- 5-Jun Raquel Tobes Preliminary functional manual annotation of E. coli TY-2482
- 5-Jun Raquel Tobes Identification of genes involved in colonization, adhesion, pathogenicity and metal resistance
- 5-Jun Raquel Tobes Analysis of TY-2482 genome plasticity
- 6-Jun David Vallenet Explore LB226692 annotations with MicroScope
- 6-Jun Kat Holt EAEC plasmids and aggregative fimbriae
- 7-Jun Kat Holt Gene content and horizontal transfer analysis
- 7-Jun Wolfgang Gerlach Taxonomic analysis of 3057 genes from TY-2482 with CARMA3
- 7-Jun Raquel Tobes & Marina Manrique (Era7) Automatic annotation of E. coli TY2482 BGI V2 assembly
- 8-Jun Raquel Tobes & Marina Manrique (Era7) Automatic annotation of E. coli LB226692 genome
- 8-Jun David Studholme A cluster of E. coli 55989 genes missing from outbreak strain TY2242B. Is this a T6SS?
- 8-Jun Kat Holt Annotation of phage in the latest (6/6; HiSeq data) assembly
- 8-Jun Kat Holt Typing schemes & PCR targets for the outbreak strain
- 9-Jun Scott Weissman & Kat Holt Plasmid MLST analysis of IncI/blaCTX-M plasmid
- 9-Jun Nico Petty via Kat Holt More detailed analysis of prophage
- 9-Jun Konrad Paszkiewicz Pfam domain comparison
- 10-Jun Raquel Tobes Mauve comparison of E. coli H112180280 and TY-2482 strains
- 10-Jun Kwan lab The outbreak strains have similar pathogencity as Ecoli EAEC strain 55989: Alignment of virulence factors from VFDB
- 11-Jun Kat Holt EAEC plasmid comparison with new scaffold assemblies
- 11-Jun Kat Holt Comparison of new BGI and HPA scaffolds
- 11-Jun Kat Holt Resistance genes in the same scaffold as chromosome
- 11-Jun Günter Klambauer, Martin Heusel, Djork-Arné Clevert and Sepp Hochreiter Copy number analysis of the outbreak strain
- 11-Jun Marina Manrique & Raquel Tobes (Era7) Automatic annotation of BGI V3 assembly of E. coli TY-2482 genome
- 11-Jun Marina Manrique & Raquel Tobes (Era7) Automatic annotation of HPA assembly of E. coli H112180280 genome
- 12-Jun Kat Holt Analysis and manual annotation of chromosomal insertion containing adhesin/pathogenicity island plus multiple drug resistance operons
- 13-Jun Kat Holt Biofilm-associated genes and acquired pic mucinase
- 12-Jun Patrik D'haeseleer Alignment of all three assemblies with 55989 and plasmids
- 13-Jun Kwan Lab The outbreak strains harbor the whole set of EAEC secreted proteins
- 14-Jun Peter Slickers Read length matters: identifying the phiStx2 att site
- 14-Jun Lisa Crossman Salmonella matches & annotations on two TY2482 plasmid scaffolds
- 14-Jun Konrad Paszkiewicz & Kat Holt SNP-base phylogeny confirms similarity of E. coli outbreak to EAEC Ec55989
- 14-Jun Peter Slickers MLST and serotyping 55989 in silico
- 15-Jun Lisa Crossman Rearrangements in the plasmids from TY2482 and H112180280
- 16-Jun Kat Holt BRIG visualisation of 4 available genomes
- 16-Jun Marina Manrique & Raquel Tobes (Era7) Automatic annotation of second HPA assembly of H112180280 isolate

Crowd-sourcing the genome

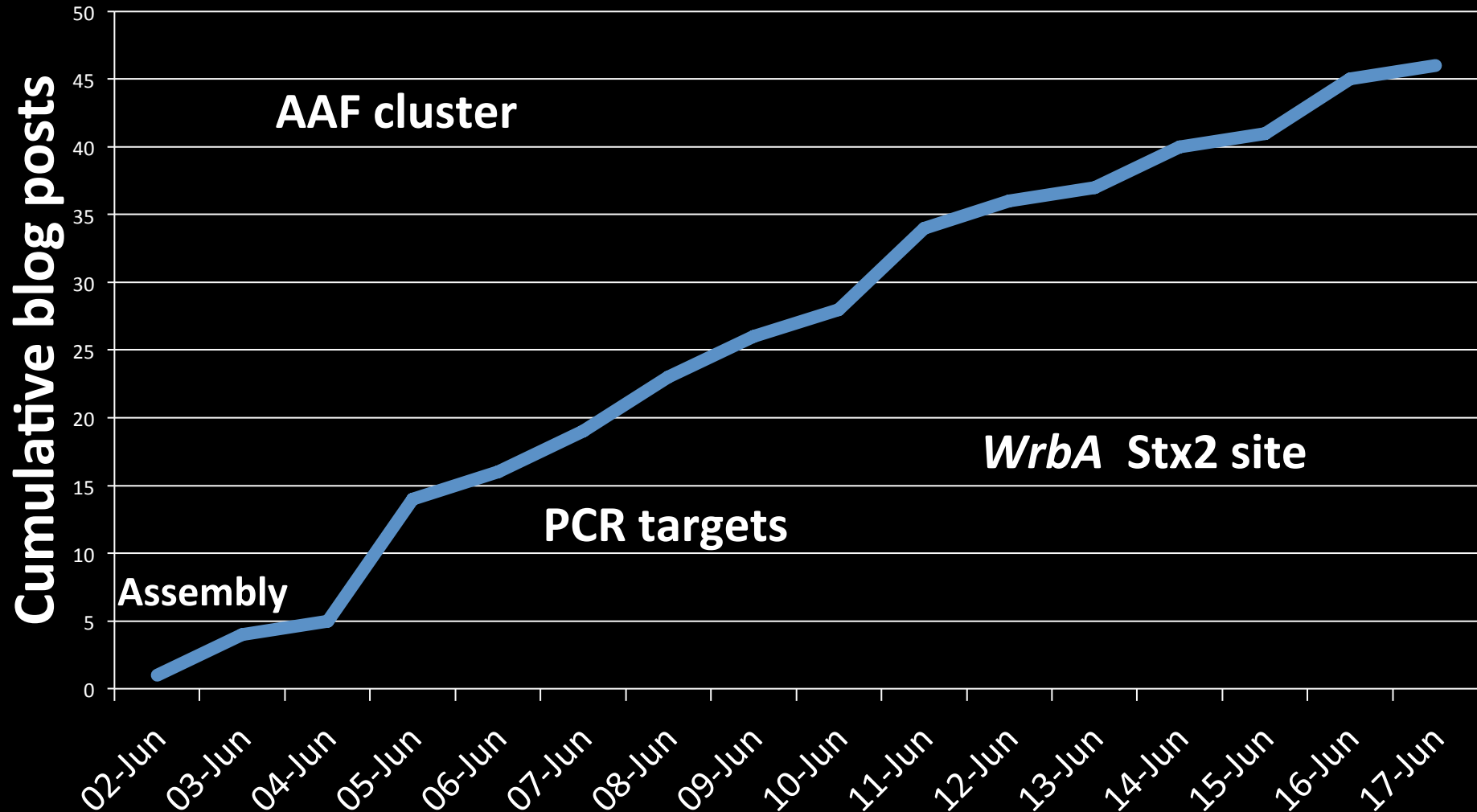https://github.com/ehec-outbreak-crowdsourced/BGI-data-analysis

Table 1. Genetic Elements in Strain TY2482 of Shiga-Toxin–Producing *Escherichia coli* O104:H4.

Genetic Element	Notable Features or Functions	Size or 55989 Coordinates*
Plasmid		
pESBL TY2482	IncI1 plasmid, homologous to pEC_Bactec carrying *bla* CTX-M-15	88 kb
pAA TY2482	Plasmid encoding aggregative adherence fimbriae I	76 kb
pG2011 TY2482	Plasmid with no obvious phenotype	1.5 kb
Region of difference		
I-ROD1	Degenerate prophage	296227 (tRNA-*Thr*)
I-ROD2	*Stx2*-encoding prophage	1176265 (*wrbA*)
I-ROD3	Microcin gene cluster; tellurite resistance gene cluster	1207704 (tRNA-*Ser*)
I-ROD4	Prophage	1811905 (*ynfG*)
I-ROD5	Prophage	2102453 (*yecE*)
I-ROD6	Molybdate metabolism regulator; *yehL*	2426442 (IS1)
I-ROD7	Multidrug-resistant gene cluster (*dfA7, sulI, sulII, strA, strB, tetA*); mercury resistance	4211244 (tRNA-*Sec*)
D-ROD1	Prophage	1094587–1140306
D-ROD2	Prophage	1413924–1446834
D-ROD3	Prophage	1754689–1800354
D-ROD4	Prophage	2688656–2701228
D-ROD5	Type VI secretion genes	3401720–3427357
D-ROD6	Prophage	4944269–5004333

Rohde et al. DOI: 10.1056/NEJMoa1107643
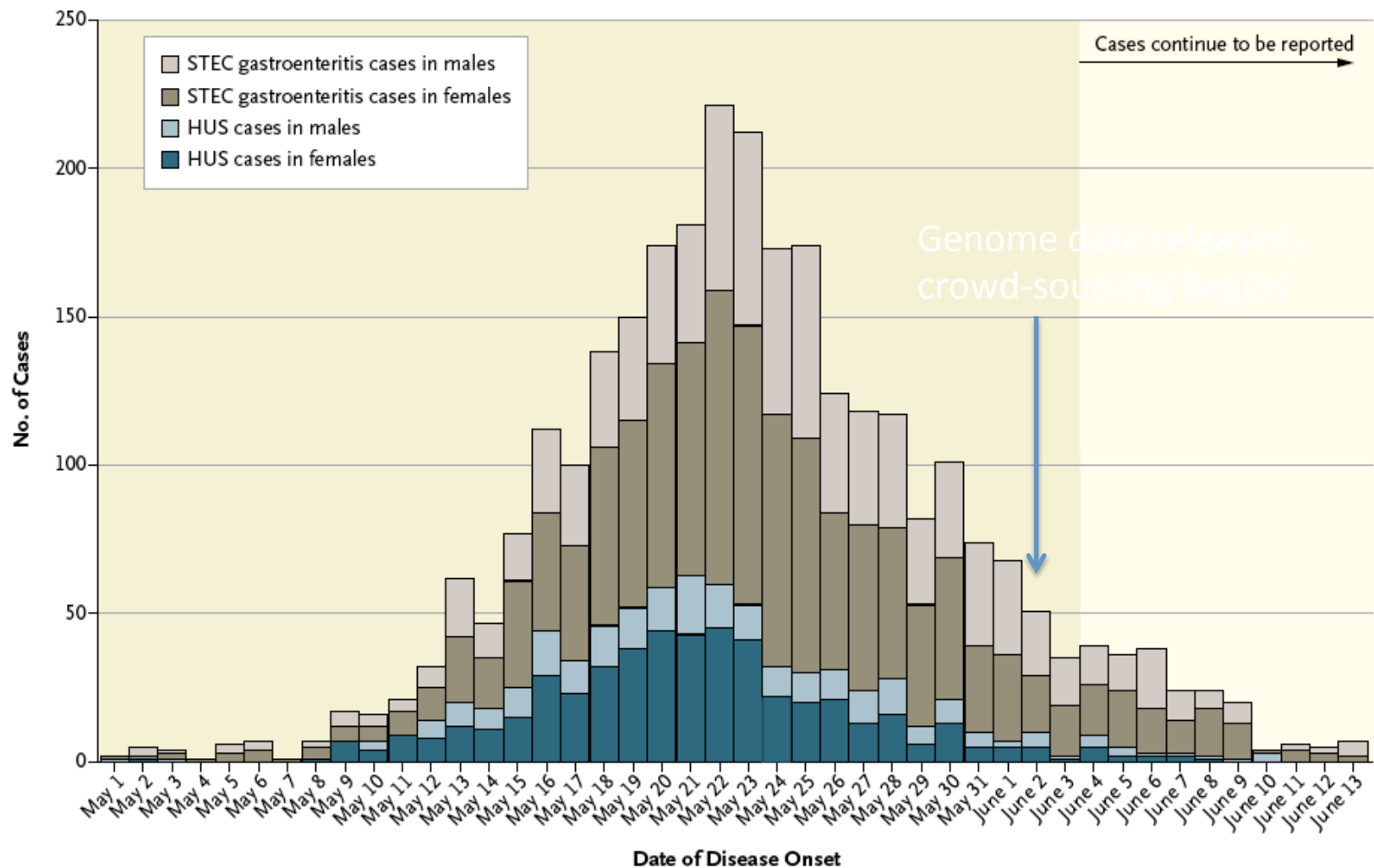
Figure 1. Epidemiologic Curve of the Outbreak.

Shown are the number of cases of the hemolytic–uremic syndrome (HUS) and of Shiga-toxin–producing *E. coli* (STEC) gastroenteritis, according to sex. Only cases with a known date of onset are included here — 748 of 810 cases of the hemolytic–uremic syndrome and 2166 of 2412 cases of Shiga-toxin–producing *E. coli* diarrhea.

Frank *et al* DOI: 10.1056/NEJMoa1106483

Digital microbiology?
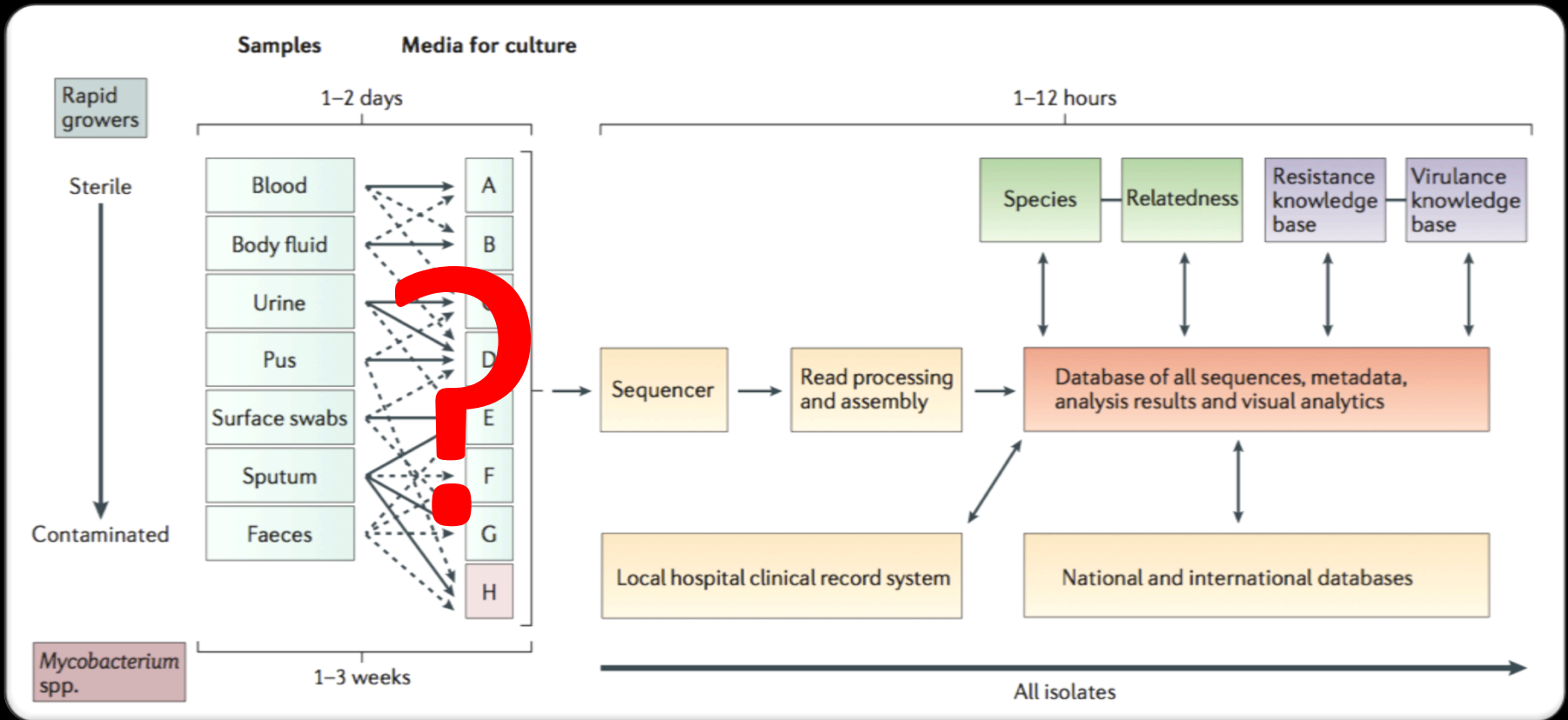
Clinical metagenomics: Pathogen discovery

A New Arenavirus in a Cluster of Fatal Transplant-Associated Diseases

Gustavo Palacios, Ph.D., Julian Druce, Ph.D., Lei Du, Ph.D., Thomas Tran, Ph.D., Chris Birch, Ph.D., Thomas Briese, Ph.D., Sean Conlan, Ph.D., Phenix-Lan Quan, Ph.D., Jeffrey Hui, B.Sc., John Marshall, Ph.D., Jan Fredrik Simons, Ph.D., Michael Egholm, Ph.D., Christopher D. Paddock, M.D., M.P.H.T.M., Wun-Ju Shieh, M.D., Ph.D., M.P.H., Cynthia S. Goldsmith, M.G.S., Sherif R. Zaki, M.D., Ph.D., Mike Catton, M.D., and W. Ian Lipkin, M.D.

ABSTRACT

BACKGROUND
Three patients who received visceral-organ transplants from a single donor on the same day died of a febrile illness 4 to 6 weeks after transplantation. Culture, polymerase-chain-reaction (PCR) and serologic assays, and oligonucleotide microarray analysis for a wide range of infectious agents were not informative.

METHODS
We evaluated RNA obtained from the liver and kidney transplant recipients. Unbiased high-throughput sequencing was used to identify microbial sequences not found by means of other methods. The specificity of sequences for a new candidate pathogen was confirmed by means of culture and by means of PCR, immunohistochemical, and serologic analyses.

RESULTS
High-throughput sequencing yielded 103,632 sequences, of which 14 represented an Old World arenavirus. Additional sequence analysis showed that this new arenavirus was related to lymphocytic choriomeningitis viruses. Specific PCR assays based on a unique sequence confirmed the presence of the virus in the kidneys, liver, blood, and cerebrospinal fluid of the recipients. Immunohistochemical analysis revealed arenavirus antigen in the liver and kidney transplants in the recipients. IgM and IgG antiviral antibodies were detected in the serum of the donor. Seroconversion was evident in serum specimens obtained from one recipient at two time points.

CONCLUSIONS
Unbiased high-throughput sequencing is a powerful tool for the discovery of pathogens. The use of this method during an outbreak of disease facilitated the identification of a new arenavirus transmitted through solid-organ transplantation.

OPEN ACCESS Freely available online — PLoS one

Direct Metagenomic Detection of Viral Pathogens in Nasal and Fecal Specimens Using an Unbiased High-Throughput Sequencing Approach

Shota Nakamura[1], Cheng-Song Yang[2,3], Naomi Sakon[4], Mayo Ueda[2,3], Takahiro Tougan[5], Akifumi Yamashita[1], Naohisa Goto[1], Kazuo Takahashi[4], Teruo Yasunaga[1], Kazuyoshi Ikuta[3], Tetsuya Mizutani[6], Yoshiko Okamoto[7], Michihira Tagami[8], Ryoji Morita[8], Norihiro Maeda[8], Jun Kawai[8], Yoshihide Hayashizaki[8], Yoshiyuki Nagai[7], Toshihiro Horii[2,5], Tetsuya Iida[2], Takaaki Nakaya[2]*

1 Department of Genome Informatics, Research Institute for Microbial Diseases (RIMD), Osaka University, Suita, Osaka, Japan, 2 International Research Center for Infectious Diseases, Research Institute for Microbial Diseases (RIMD), Osaka University, Suita, Osaka, Japan, 3 Department of Virology, Research Institute for Microbial Diseases (RIMD), Osaka University, Suita, Osaka, Japan, 4 Department of Infectious Diseases, Osaka Prefectural Institute of Public Health, Higashinari, Osaka, Japan, 5 Department of Molecular Protozoology, Research Institute for Microbial Diseases (RIMD), Osaka University, Suita, Osaka, Japan, 6 Department of Virology 1, National Institute of Infectious Diseases, Musashimurayama, Tokyo, Japan, 7 Center of Research Network for Infectious Diseases, RIKEN, Chiyoda, Tokyo, Japan, 8 Omics Science Center (OSC), RIKEN, Yokohama, Kanagawa, Japan

Abstract

With the severe acute respiratory syndrome epidemic of 2003 and renewed attention on avian influenza viral pandemics, new surveillance systems are needed for the earlier detection of emerging infectious diseases. We applied a "next-generation" parallel sequencing platform for viral detection in nasopharyngeal and fecal samples collected during seasonal influenza virus (Flu) infections and norovirus outbreaks from 2005 to 2007 in Osaka, Japan. Random RT-PCR was performed to amplify RNA extracted from 0.1–0.25 ml of nasopharyngeal aspirates (N = 3) and fecal specimens (N = 5), and more than 10 µg of cDNA was synthesized. Unbiased high-throughput sequencing of these 8 samples yielded 15,298–32,335 (average 24,738) reads in a single 7.5 h run. In nasopharyngeal samples, although whole genome analysis was not available because the majority (>90%) of reads were host genome–derived, 20–460 Flu-reads were detected, which was sufficient for subtype identification. In fecal samples, bacteria and host cells were removed by centrifugation, resulting in gain of 484–15,260 reads of norovirus sequence (78–98% of the whole genome was covered), except for one specimen that was under-detectable by RT-PCR. These results suggest that our unbiased high-throughput sequencing approach is useful for directly detecting pathogenic viruses without advance genetic information. Although its cost and technological availability make it unlikely that this system will very soon be the diagnostic standard worldwide, this system could be useful for the earlier discovery of novel emerging viruses and bioterrorism, which are difficult to detect with conventional procedures.

Diagnostic metagenomics

STEC

Non-STEC
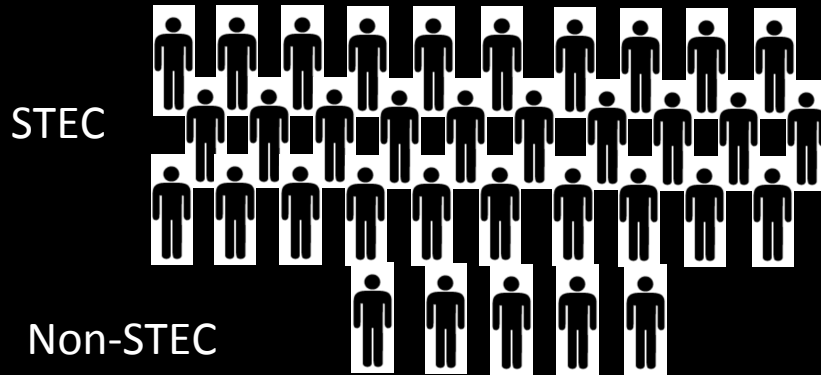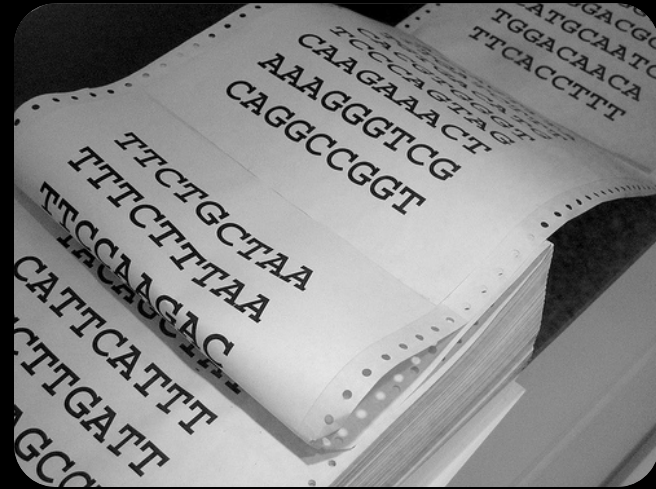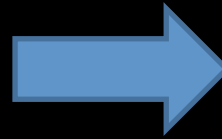
34 patients

C. diff, S. enterica, C. jejuni
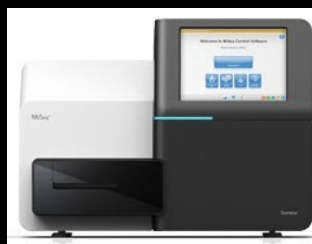
45 samples

17 million reads / >2 gigabases per sample

1) Sample preparation (4.5 hours)

Stool sample → Tissue homogenizer → DNA extraction → DNA fragmentation, tagging and barcoding
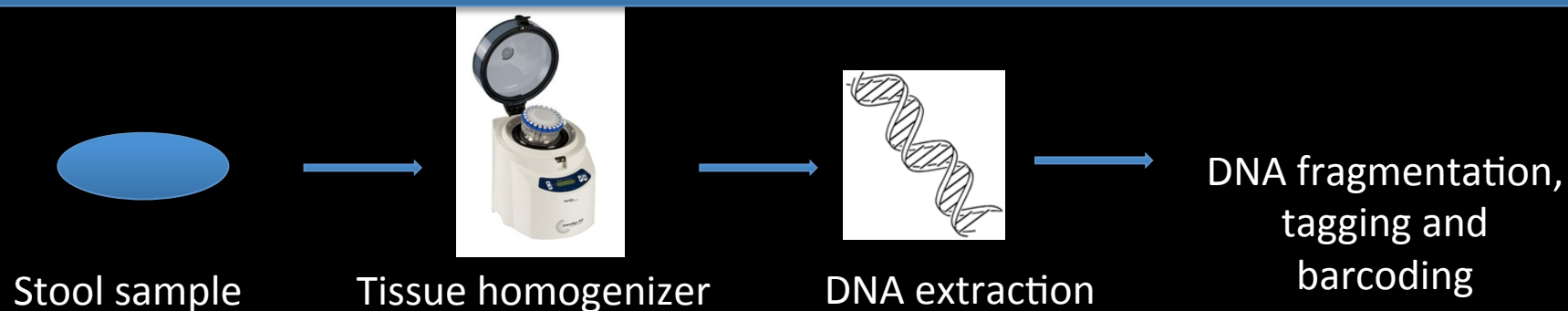
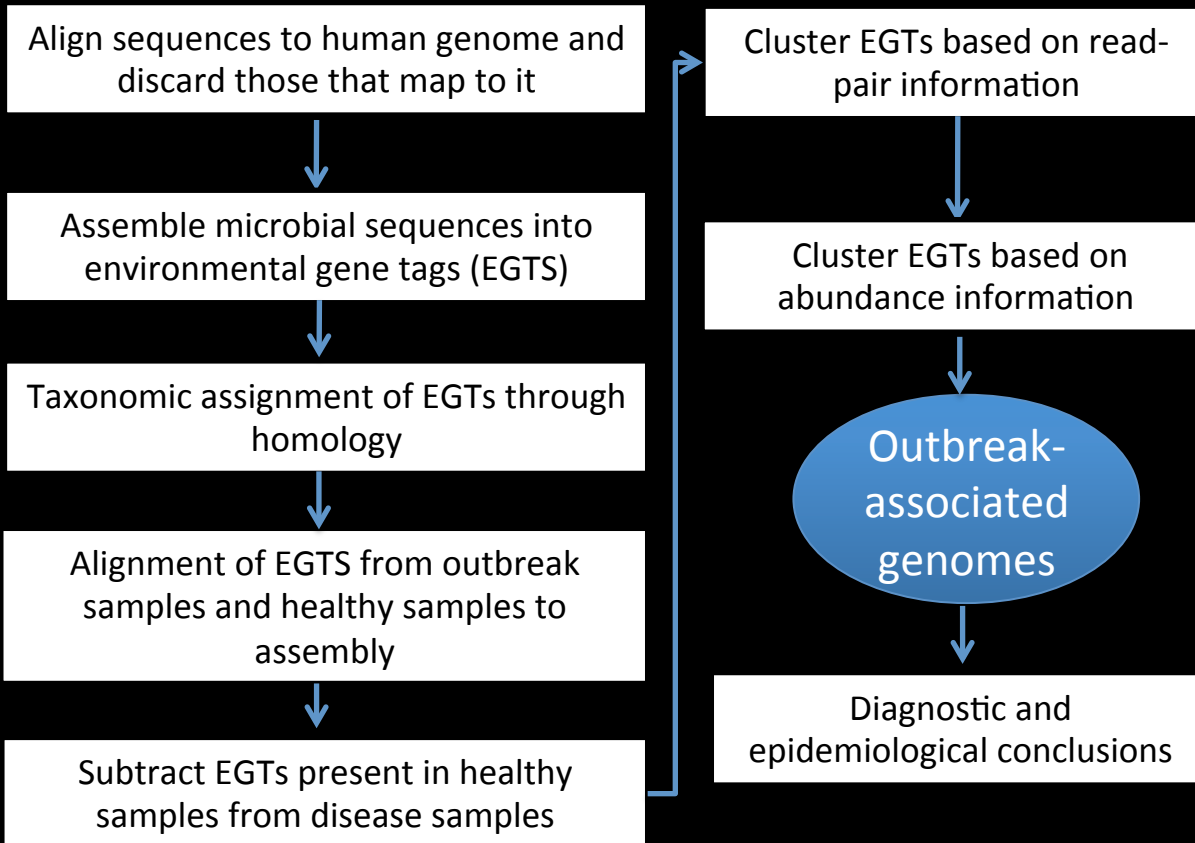# 2) Sequencing (27-40 hours)

HiSeq, 40 hours

MiSeq, 27 hours

Paired 150 base reads

# 3) Bioinformatics analysis (<24 hours of 32-core server time)

Align sequences to human genome and discard those that map to it

↓

Assemble microbial sequences into environmental gene tags (EGTS)

↓

Taxonomic assignment of EGTs through homology

↓

Alignment of EGTS from outbreak samples and healthy samples to assembly

↓

Subtract EGTs present in healthy samples from disease samples

→

Cluster EGTs based on read-pair information

↓

Cluster EGTs based on abundance information

↓

Outbreak-associated genomes

↓

Diagnostic and epidemiological conclusions

Human "contamination"

Healthy human microbiome
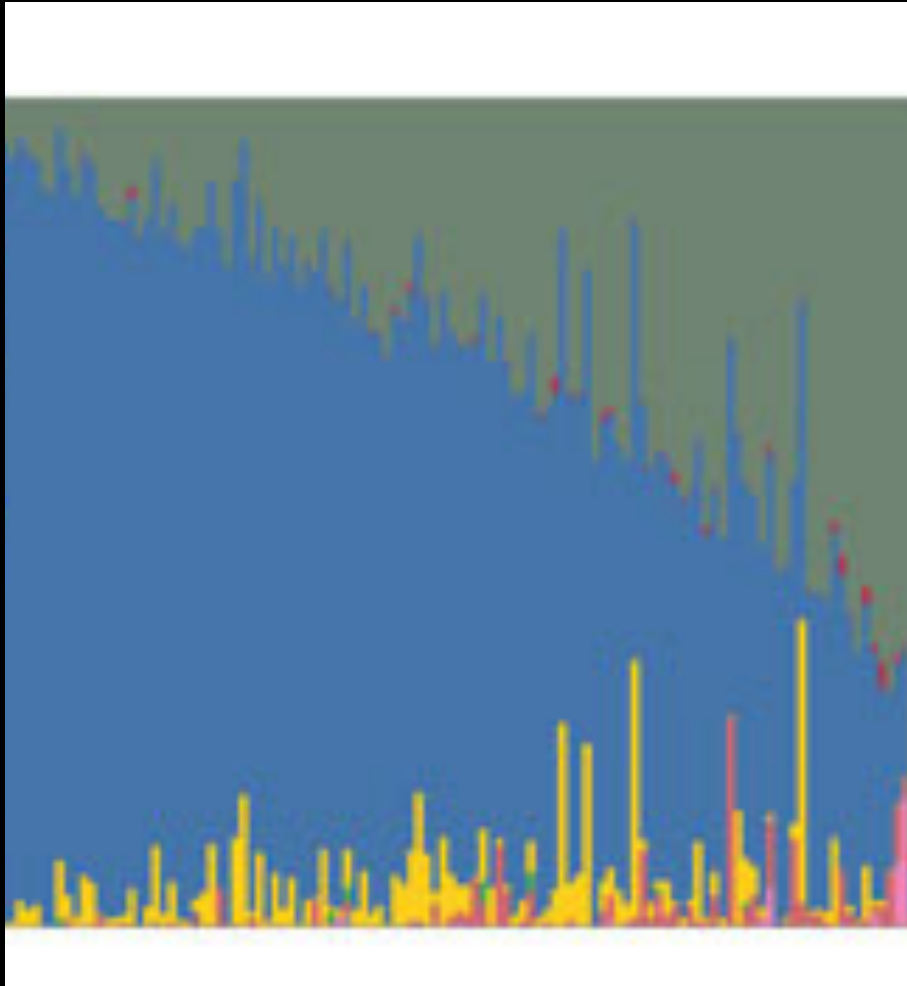
Legend:
- Firmicutes
- Actinobacteria
- Bacteroidetes
- Proteobacteria
- Fusobacteria
- Tenericutes
- Spirochaetes
- Cyanobacteria
- Verrucomicrobia
- TM7

The Human Microbiome Project Consortium
PMID: 22699609

Metagenomics: phylogenetic profiling
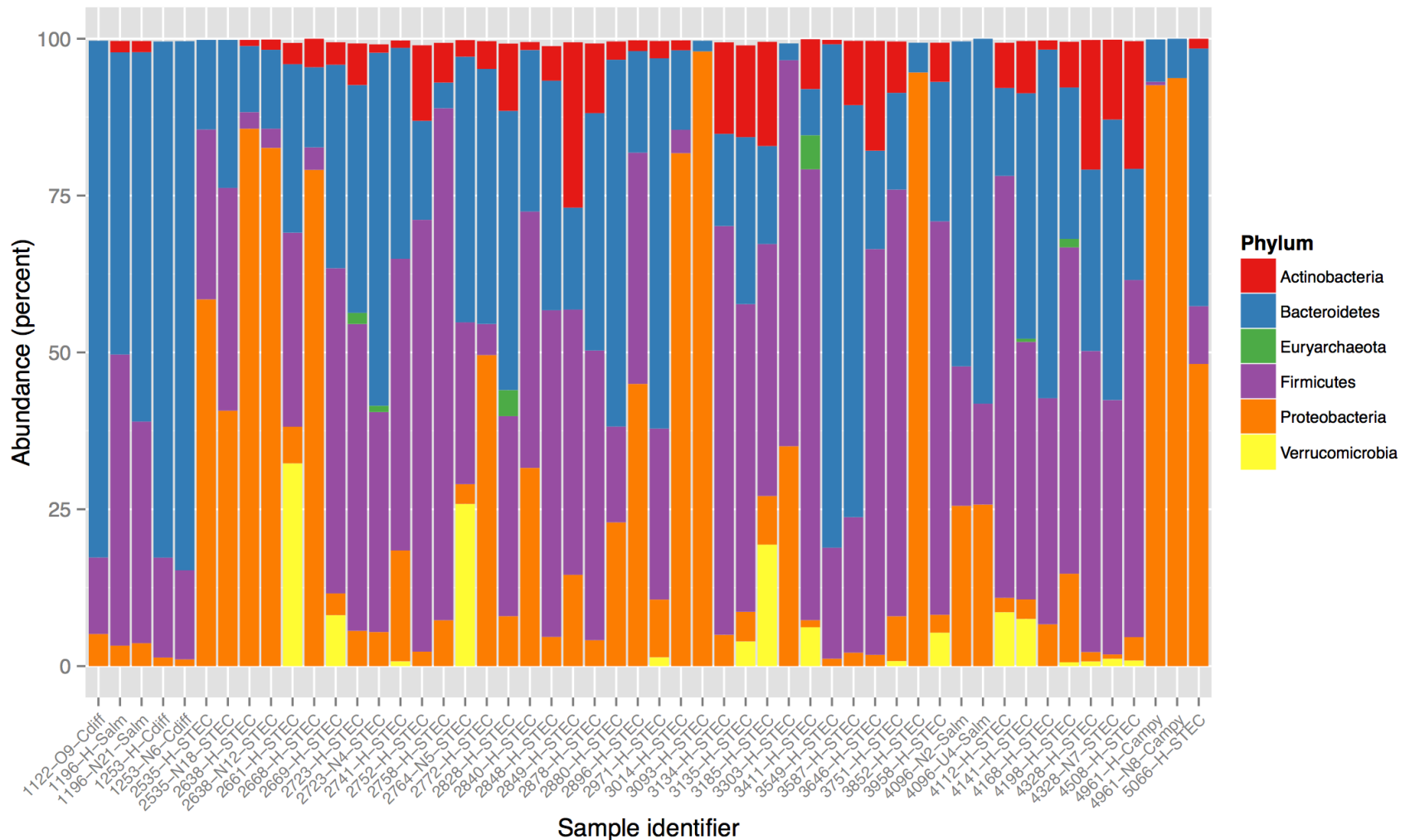
STEC genome coverage plot
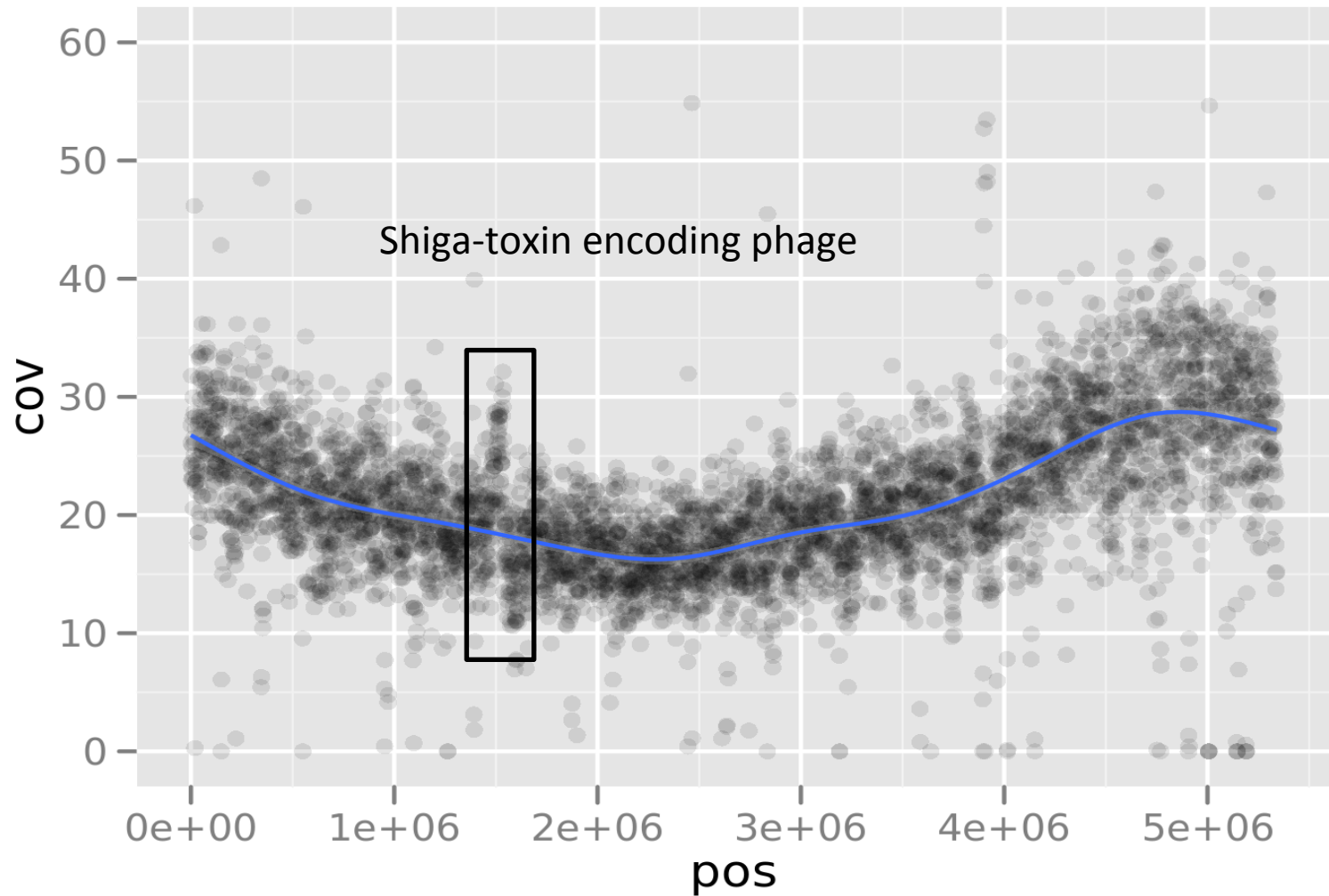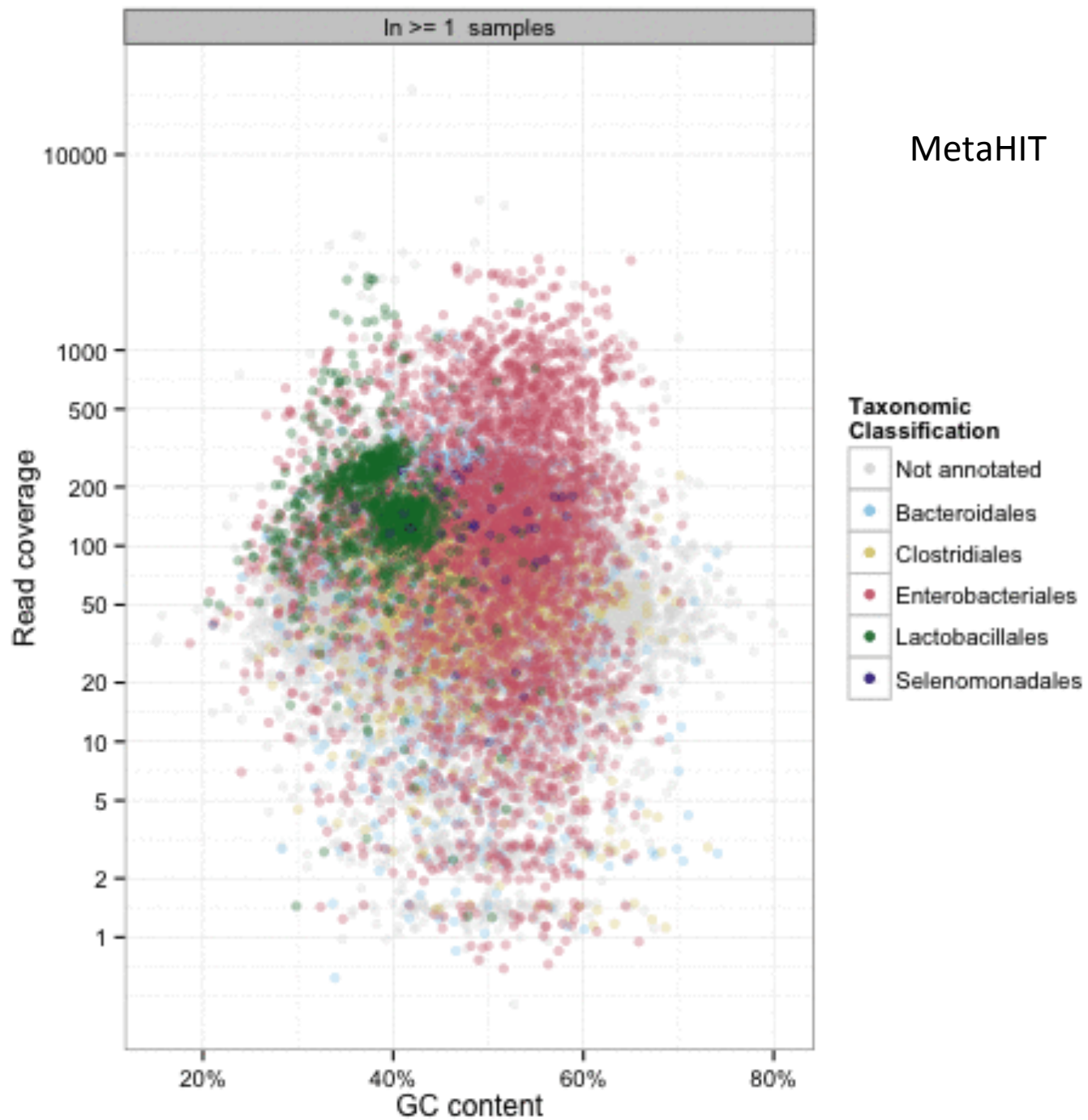
Pathogen discovery
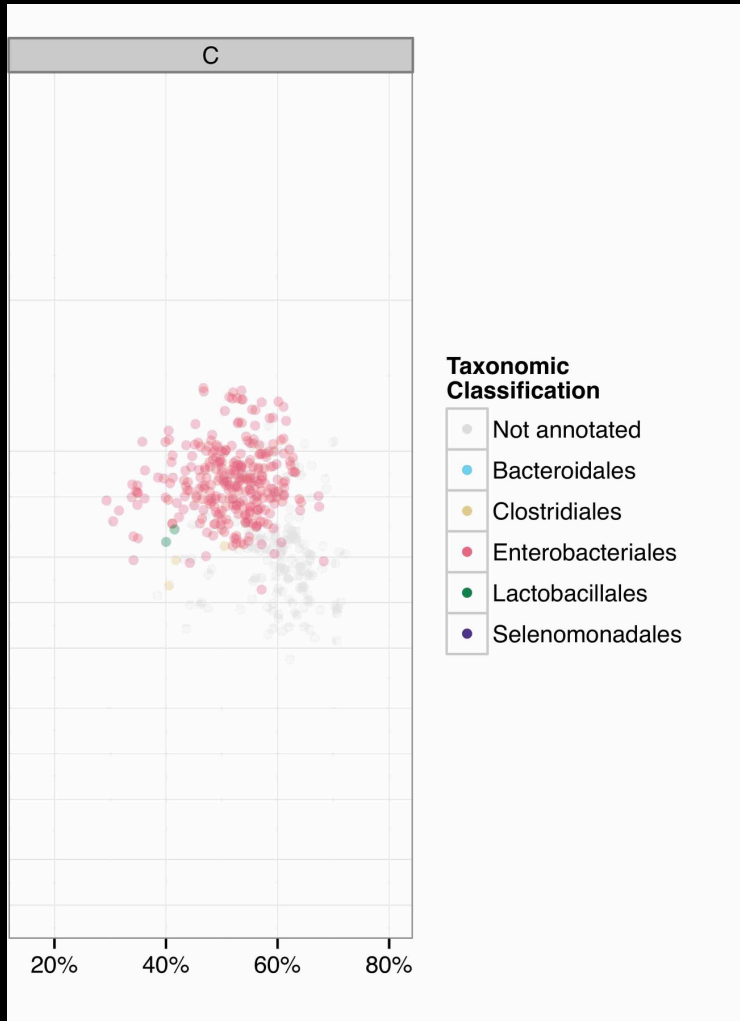
- Shiga-toxin encoding phage genes
- O-antigen determining genes
- pAA and pESBL specific sequences

Motivations for CONCOCT:
- Unsupervised version?
- Can we extract all the genomes?

CONCOCT: Chris Quince, Anders Andersson

E. coli O104:H4 genome reconstruction
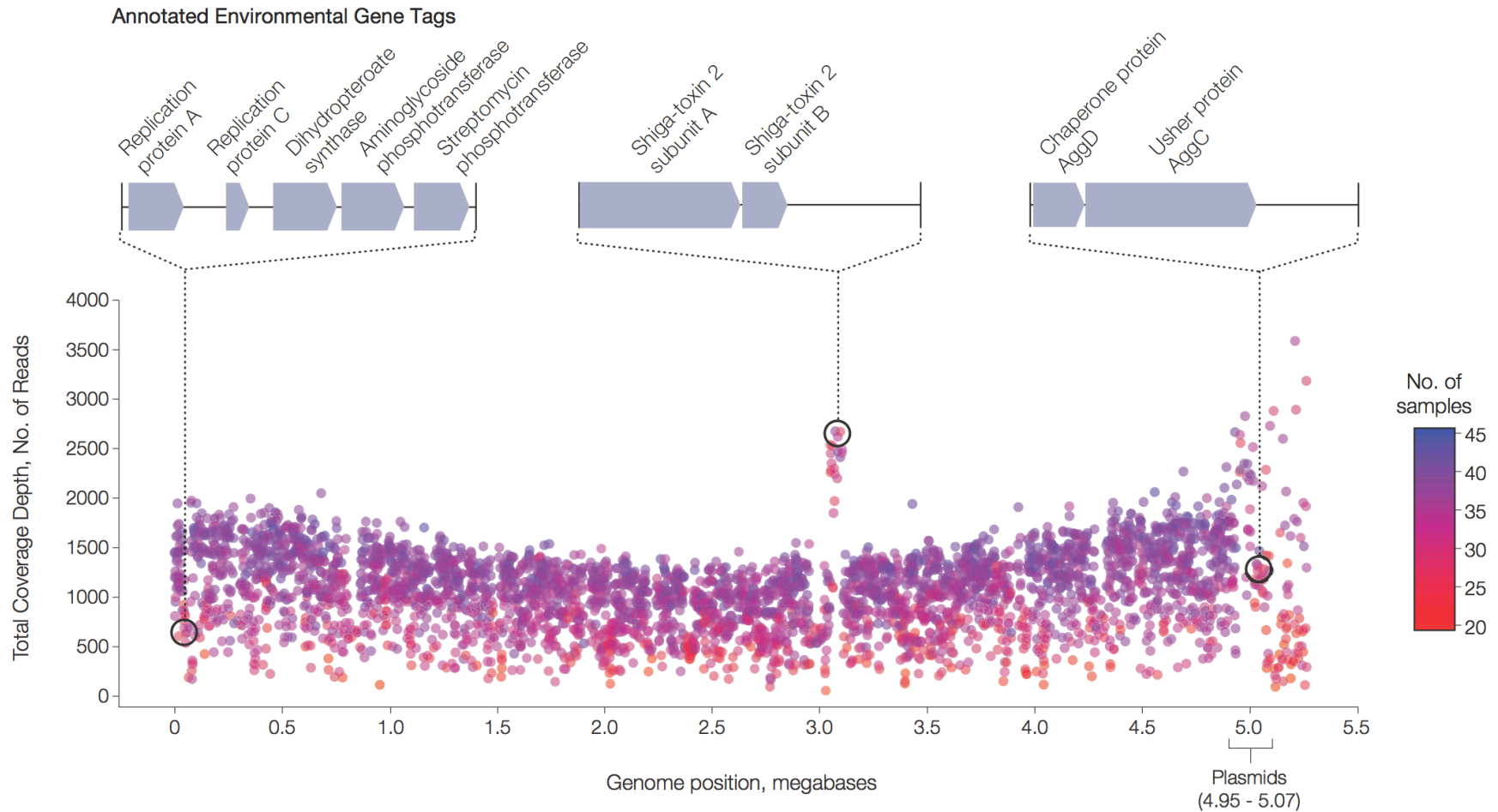
Loman et al. doi: 10.1001/jama.2013.3231.

Getting the diagnosis:
STEC genome reconstruction
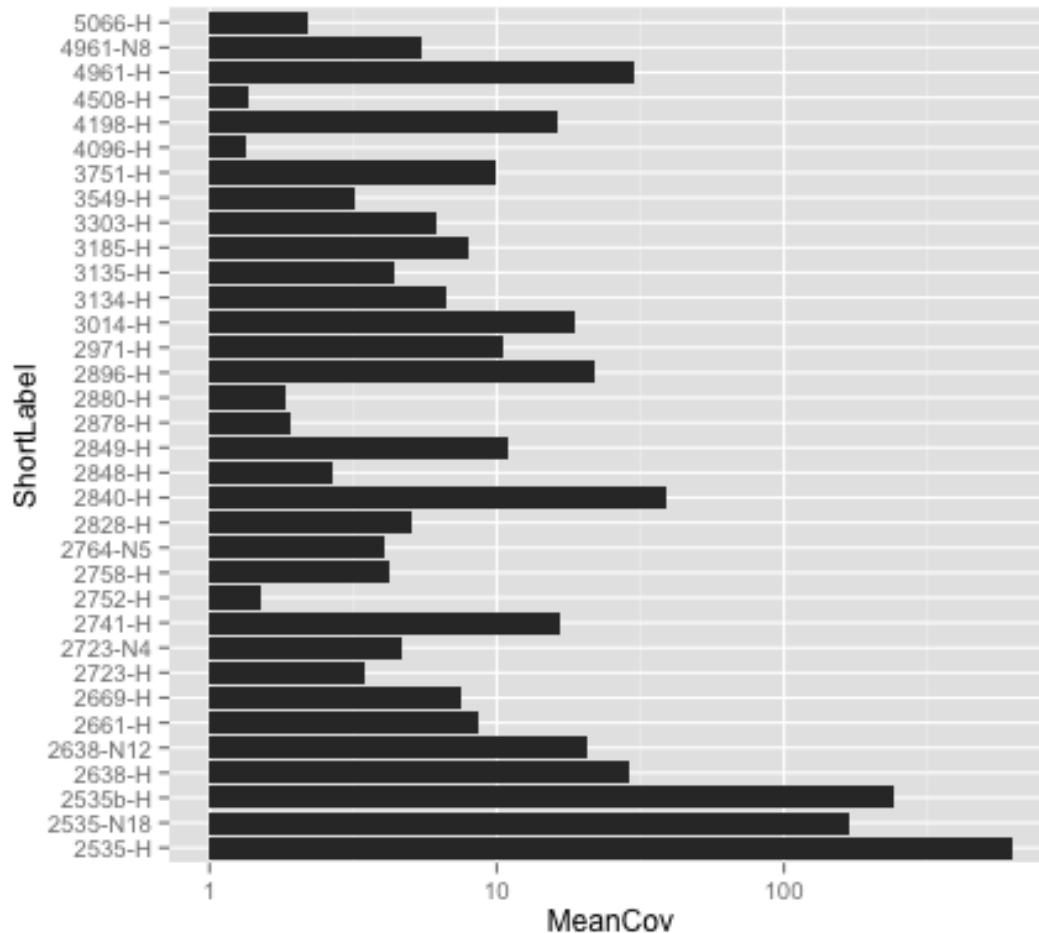
- STEC genome recovered at >10x coverage: 10/40
- STEC genome recovered at >1x coverage: 26/40
- Shiga toxin fragments recovered: 27/40
- Detected STEC genome in 6 samples which were ELISA negative

Shiga-toxin encoding phage:
Copy number variation

Escherichia coli (sample 2669)

Escherichia coli (sample 3134)
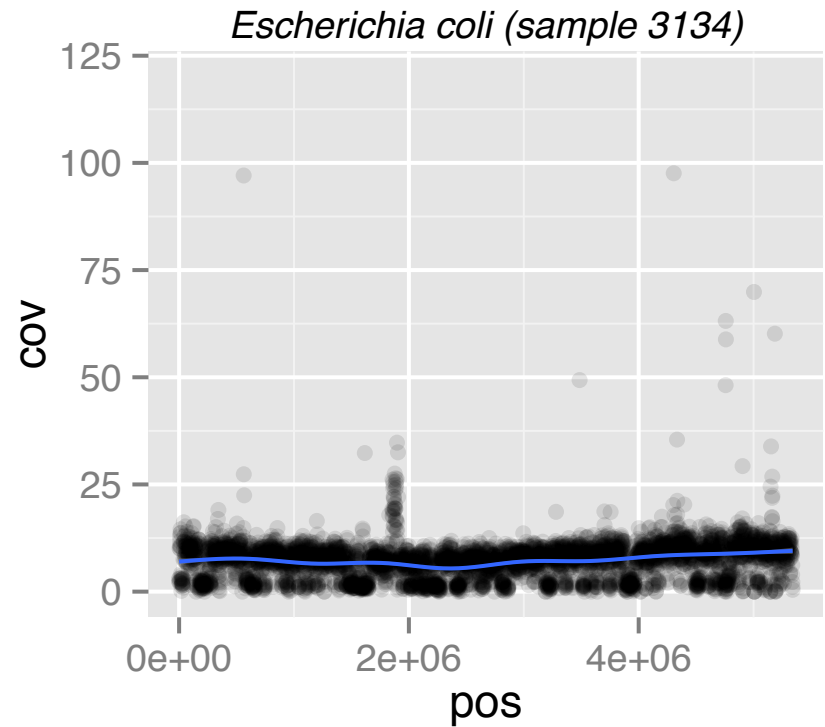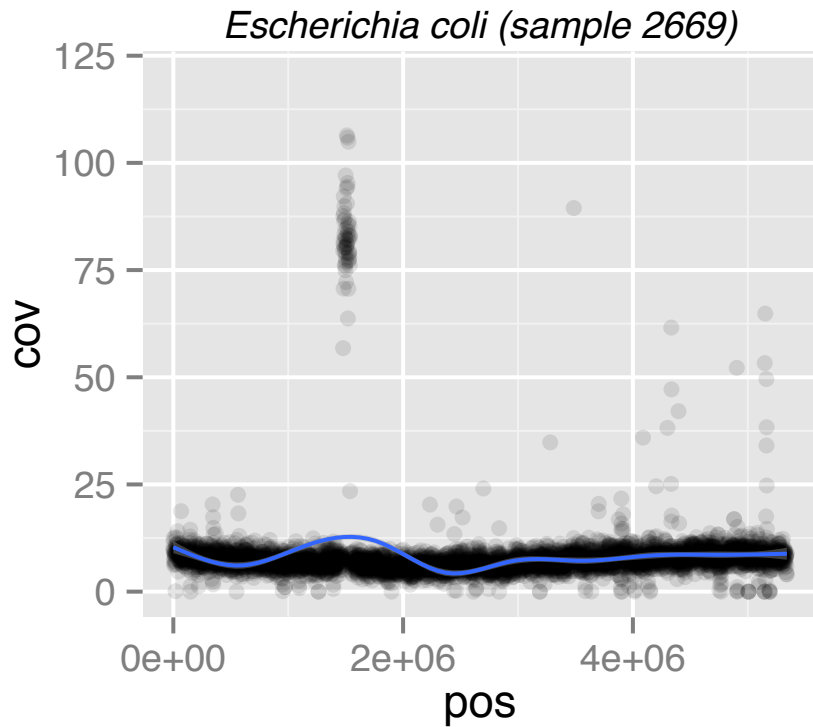
Detection of important virulence genes

– Flagellin (H-antigen)
– MLST

Correct H antigen detected in 17/40

No H antigen detected in 18/40

Mixtures of antigens detected in 3/40

ST678 profile unambiguously detected in 9/40

Mixture of profiles detected in 5/40

The rest either partial or absent profiles

Escherichia coli (sample 2535)
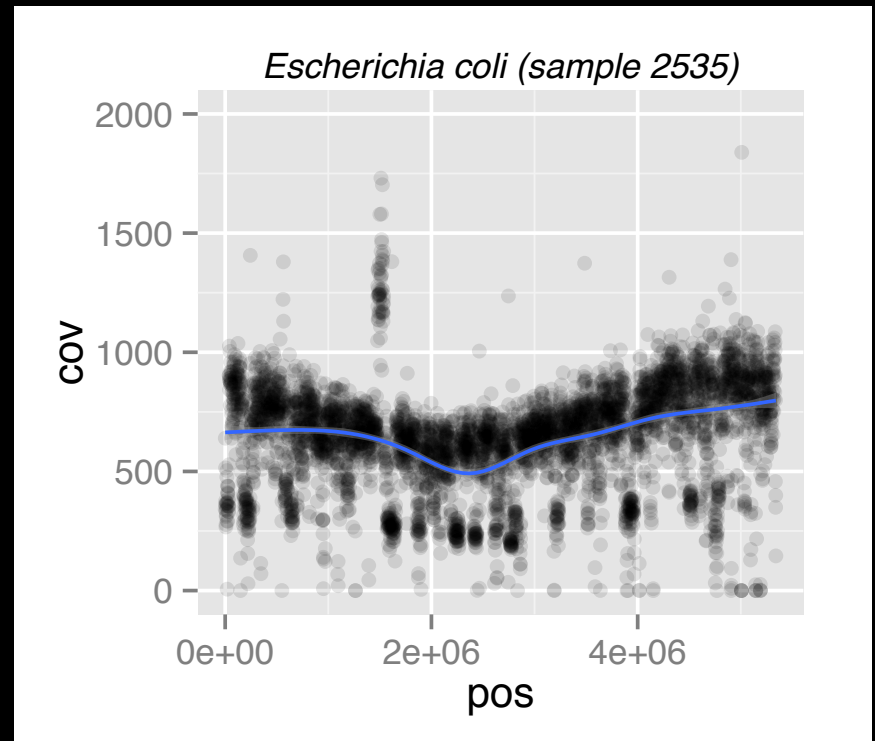
Non-STEC genome reconstructions
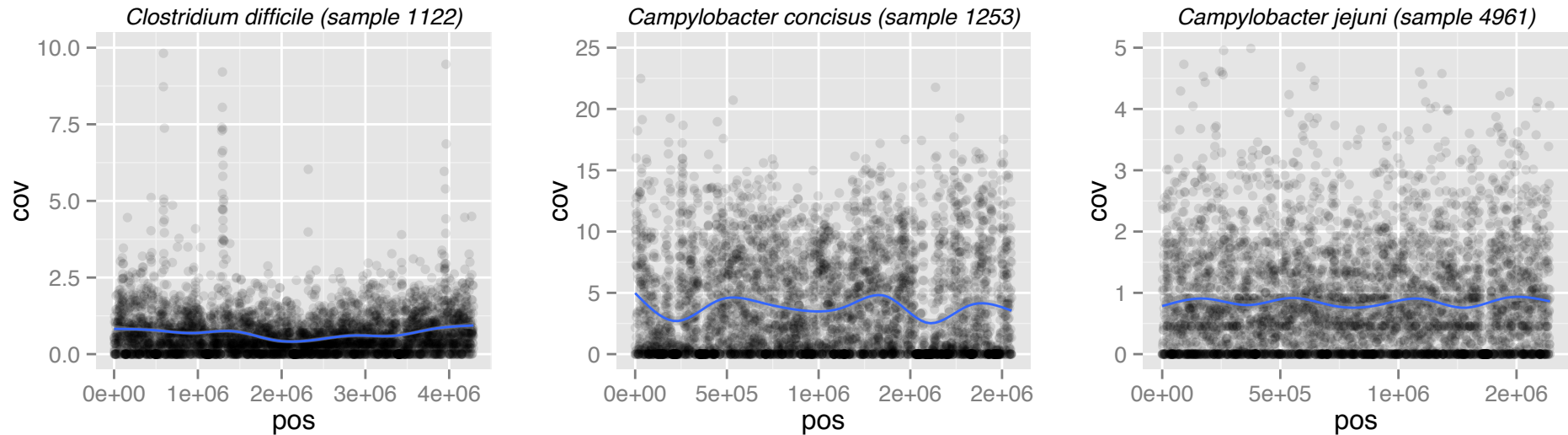
- *Clostridium difficile* toxin sequences detected (2/2)
- *Campylobacter jejuni* toxin sequences detected
- *Salmonella enterica* sequences detected (1/2)
- Emerging human pathogen *Campylobacter concisus* detected

Experimental considerations

Tissue metagenomics: in the dark

- A case of chronic osteomyelitis
- Surgical debridement and reconstruction of a compound fracture wound at the left tibia
- Reconstruction was carried out using a radial forearm free flap
- Initial injury 5 years ago
- 5 tissue samples from operation on a single MiSeq run
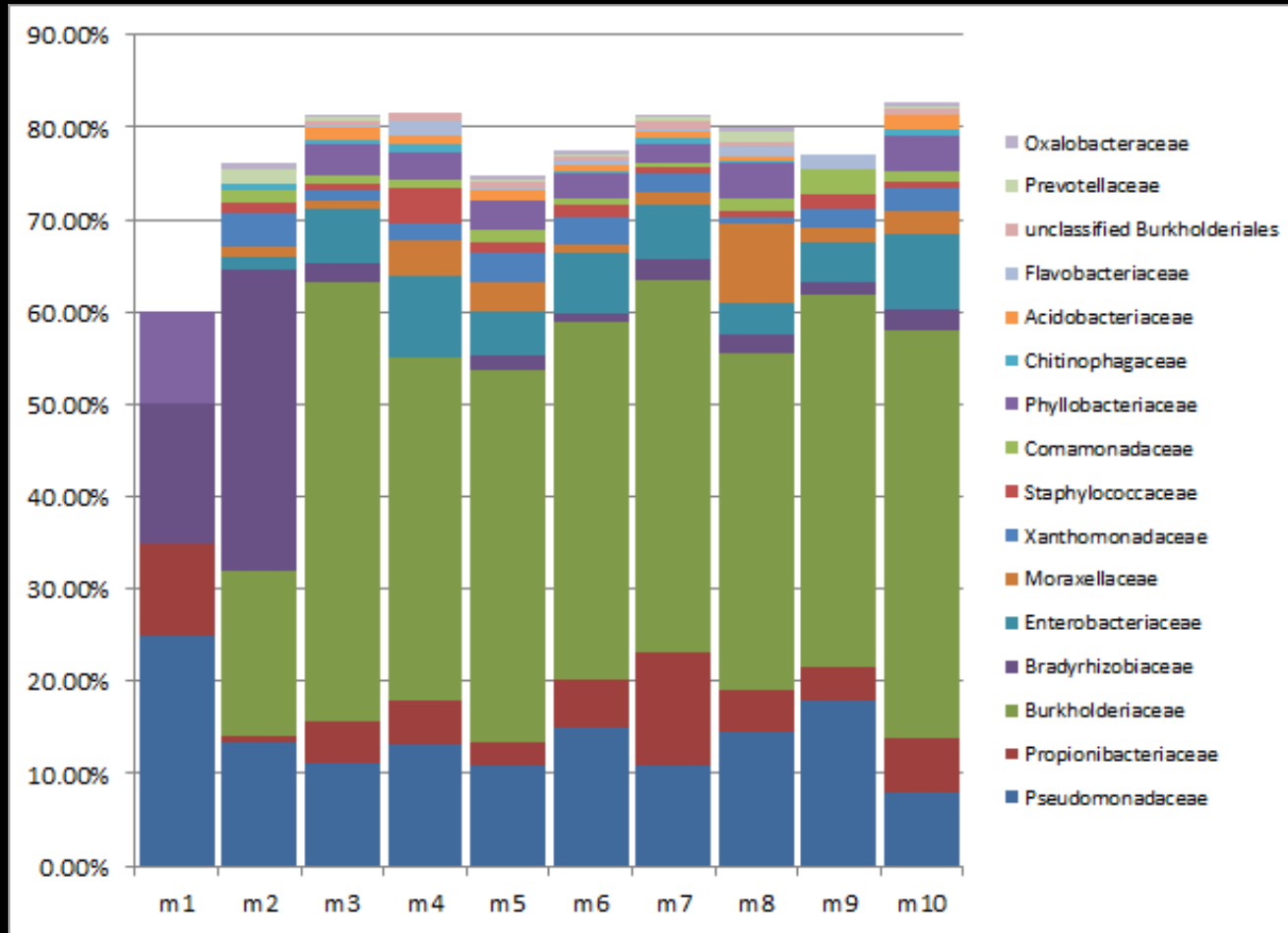- Extracted with Stratec-RTP Bacterial DNA Mini kit.

Chronic osteomyelitis

Negative control!

12358 reads

Taxonomy profile for 7506_S6.rma
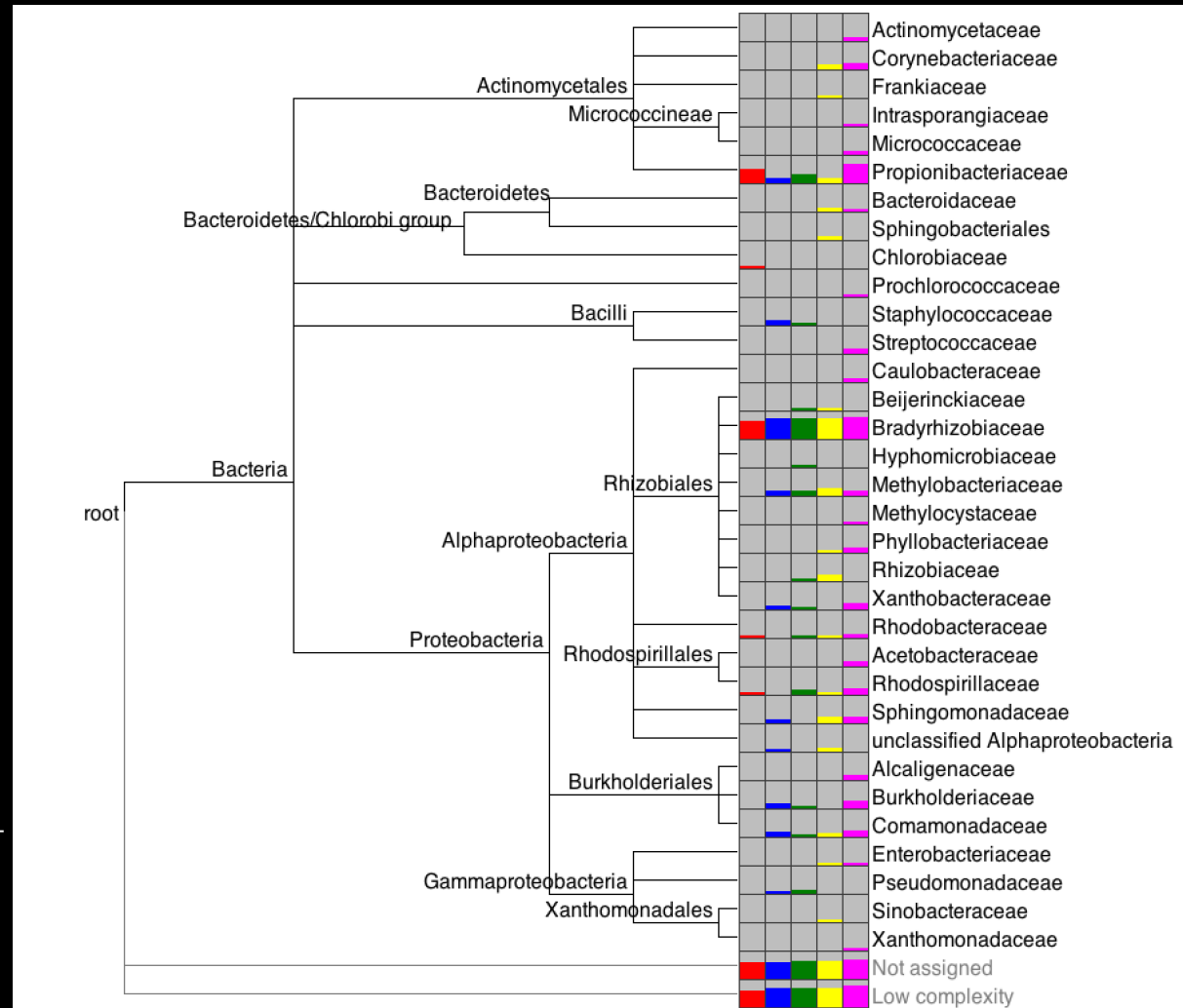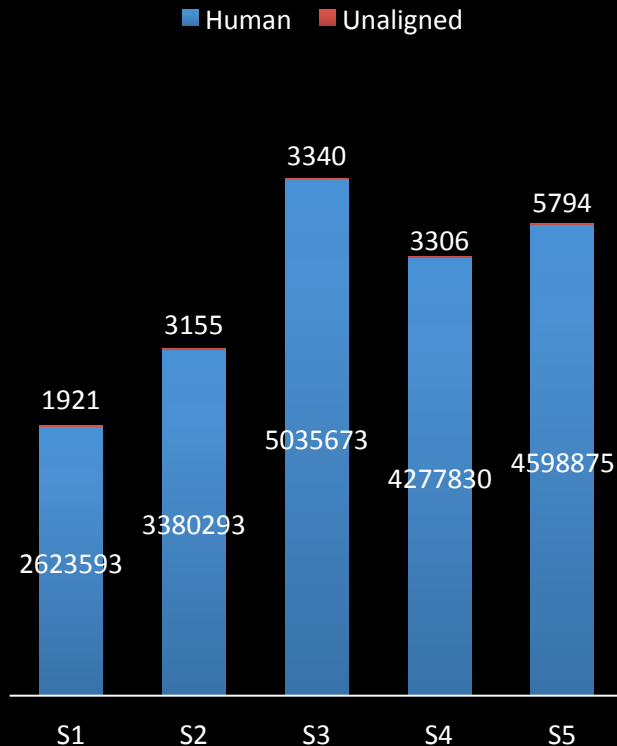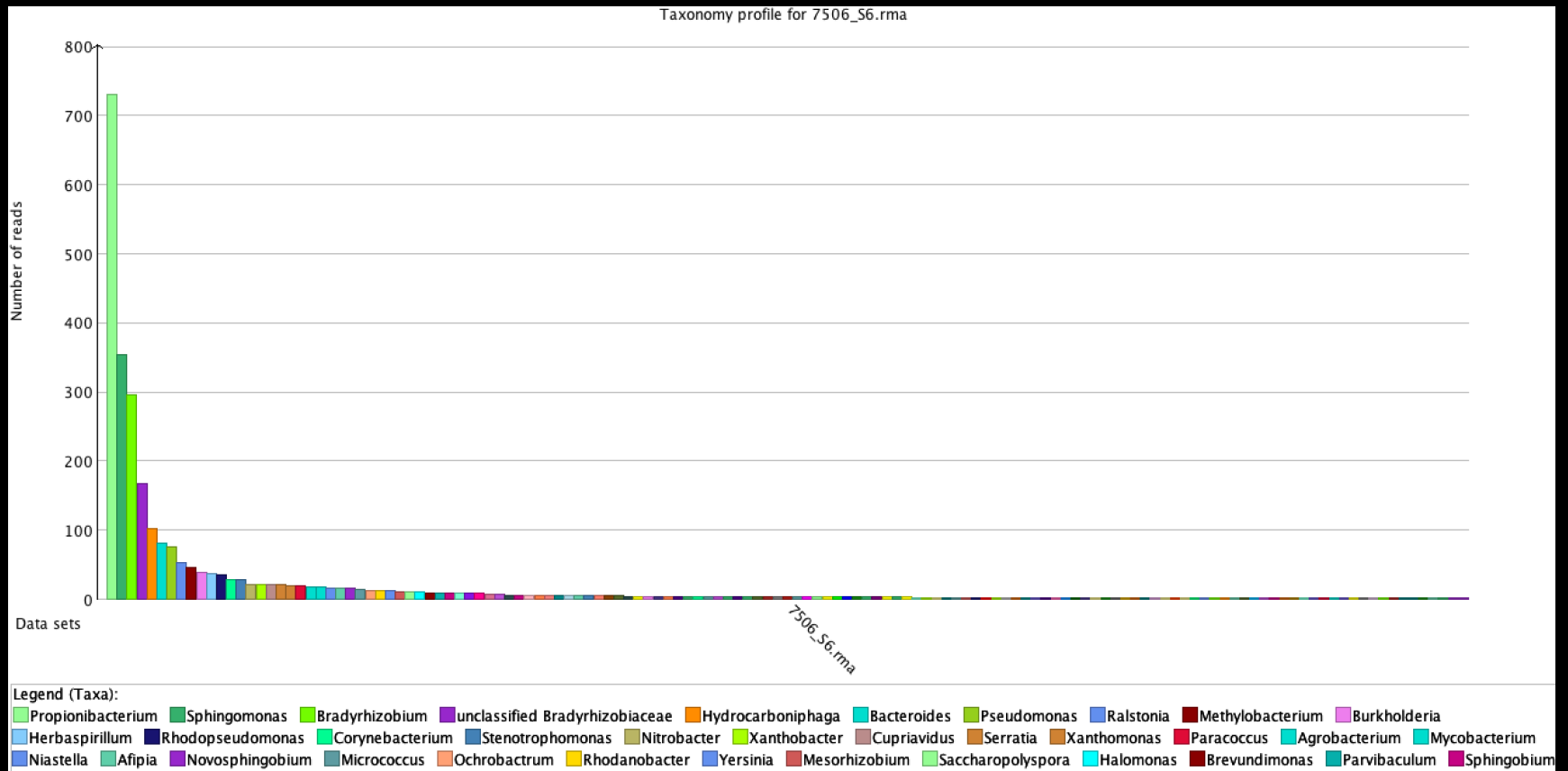
Kitome: QIAGEN Stool Kit
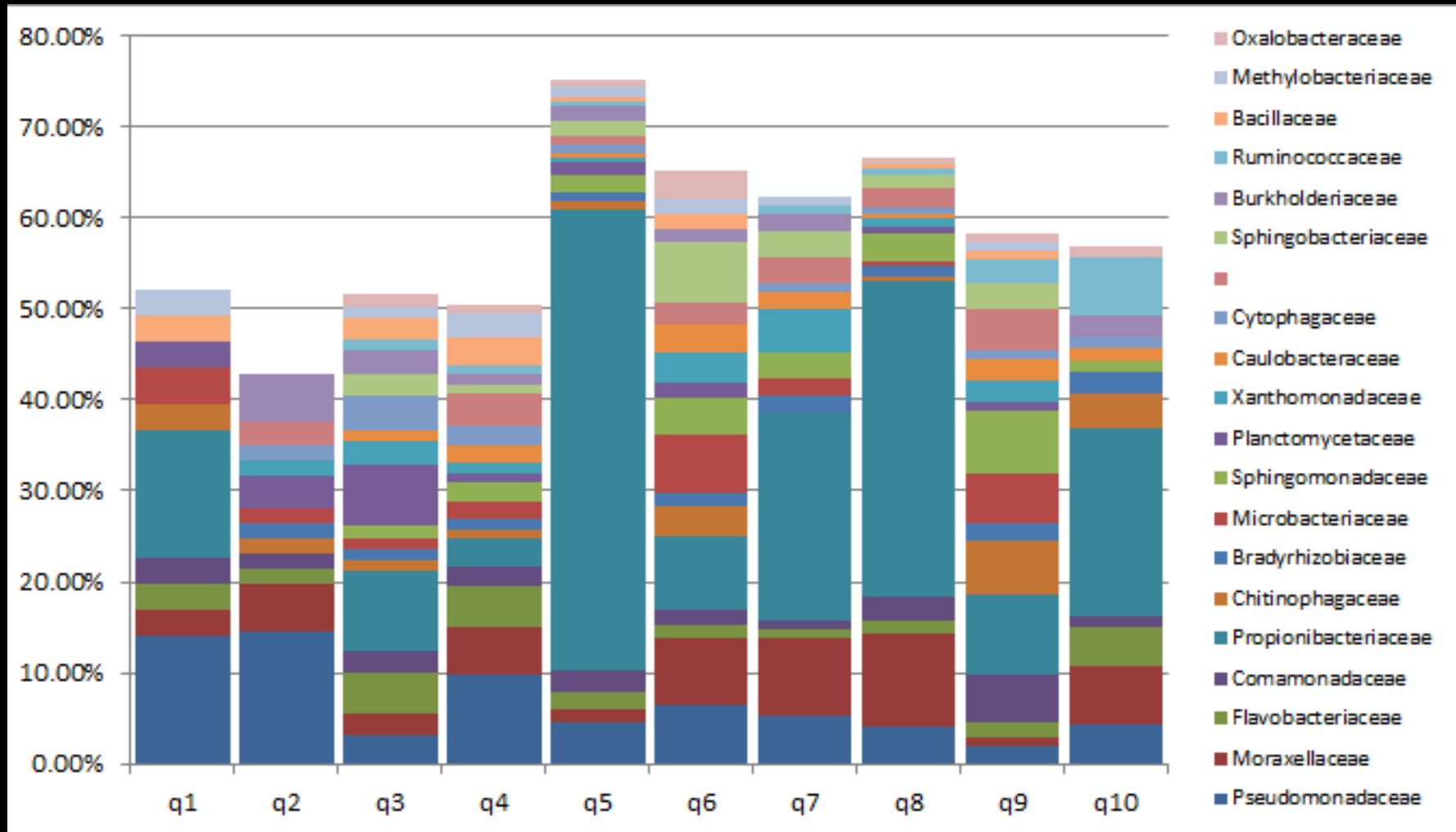
Zannah Salter

Kitome: MPBio

Zannah Salter

Consider contamination

- Environmental DNA or bacteria themselves can be introduced at many points:

  - Sample collection
  - DNA extraction
  - PCR reagents
  - Lab contamination
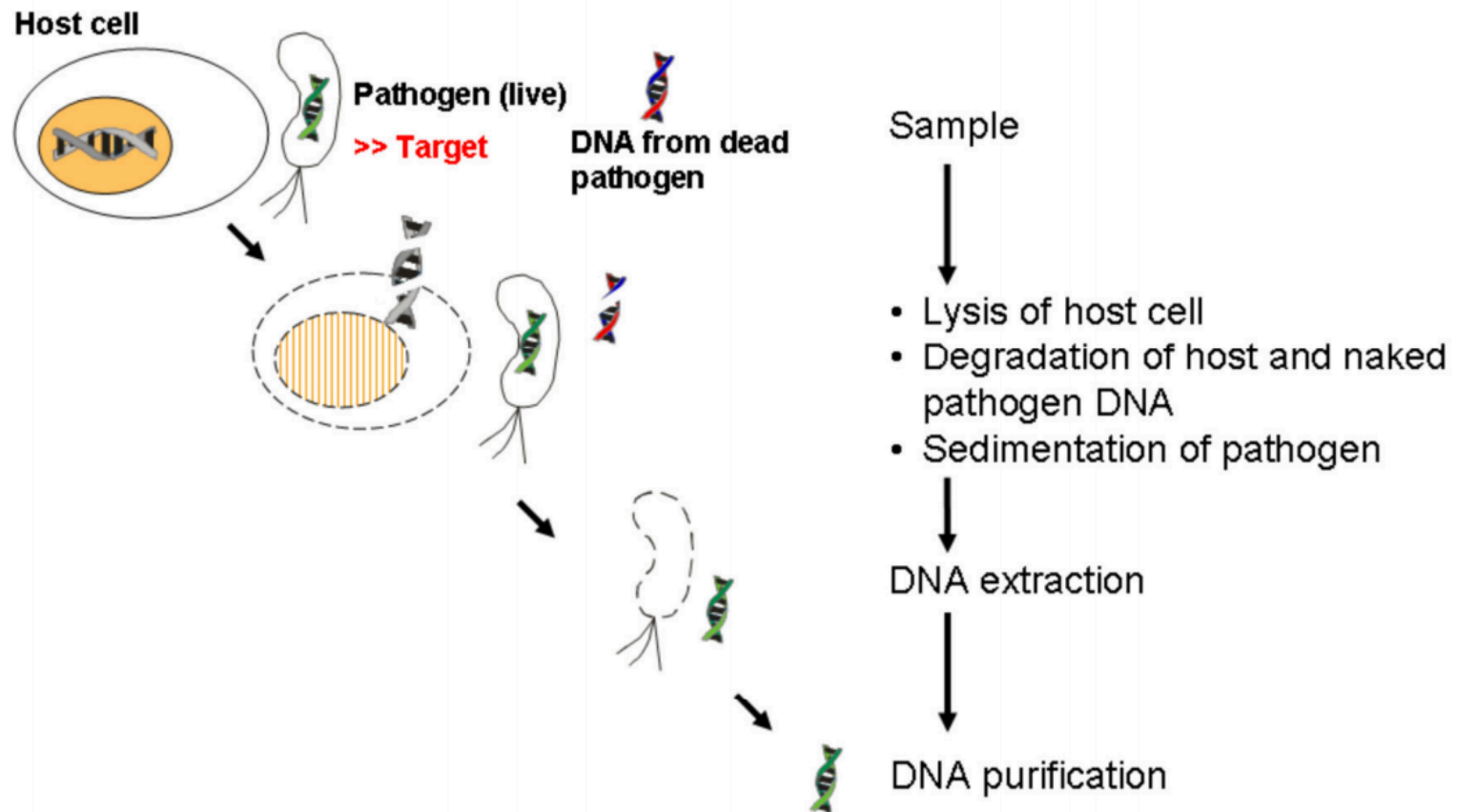
Just throw more reads?

- 5 reads in 5,000,000
  - 1 read in 1,000,000
  - Staph genome ~2Mb
  - 30x coverage required (?)

▸ 240 billion reads required

▸ 6 billion reads per HiSeq 2500 run

   ▸ 40 HiSeq runs

   ▸ Cost of ~£500,000

      Oh dear!

Background depletion

- Differential lysis
  - Human cells fragile
- Eukaryotic DNA is methylated
  - MoLysis
  - NEB Microbiome Kit
  - MEDIP
    - ... what about microbial eukaryotes? ☹

MolYsis™

Fig. 1: The MolYsis™ procedure of human DNA removal and microbial DNA isolation.

NEBNext

https://www.neb.com/products/e2612-nebnext-microbiome-dna-enrichment-kit

Justin O'Grady, Josh Quick, Szymon Calus
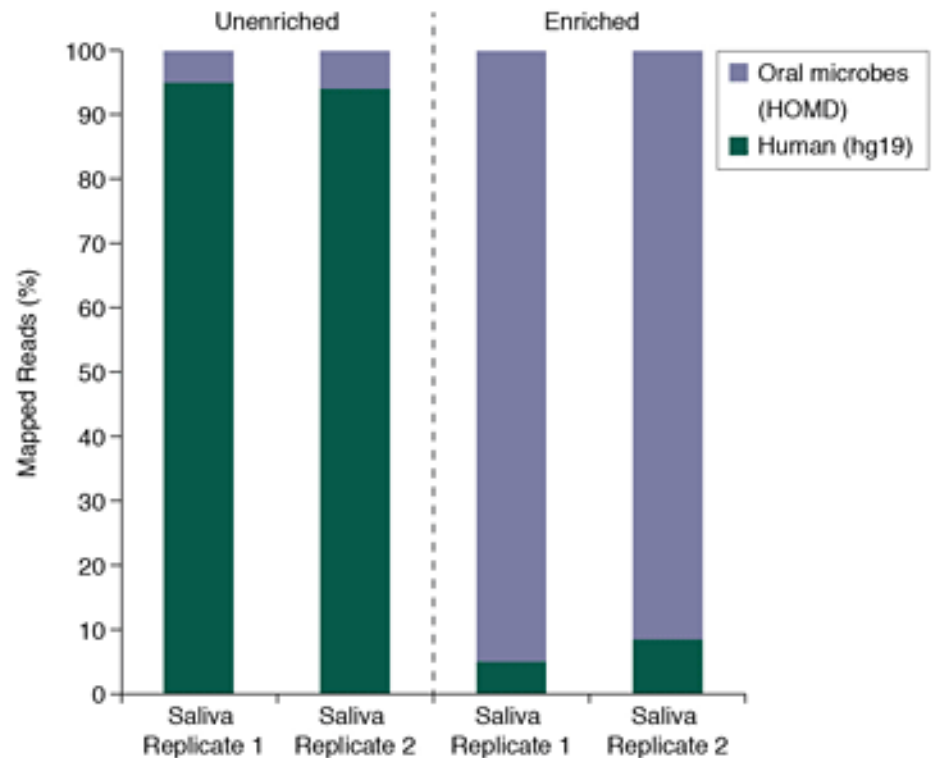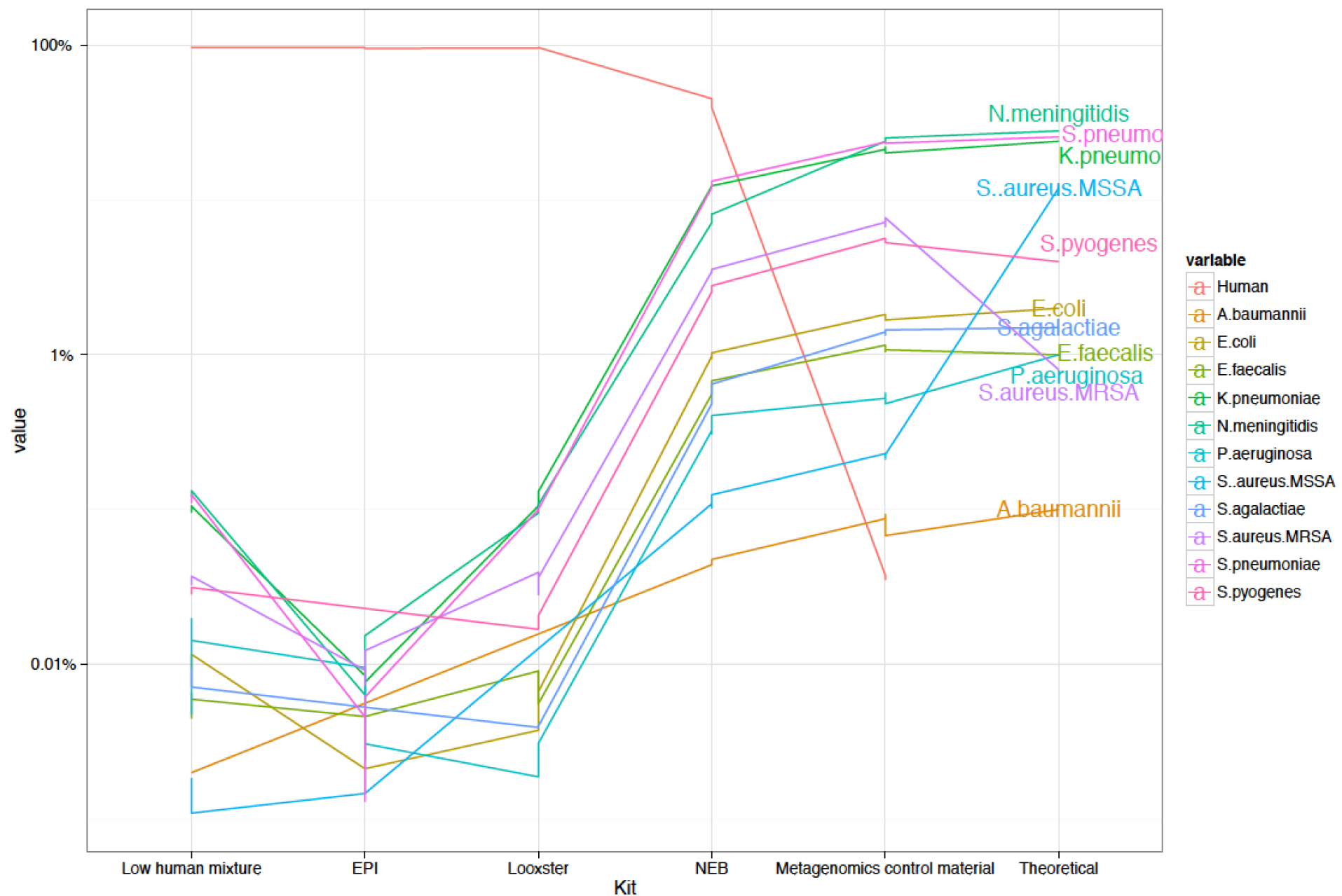
Microbial enrichment

- (Micro)culture(omics)!
  - Also may introduce bias
  - Lose quantitative informatio~~n~~

- PCR: 16S / marker genes
  - Poor species level discrimination, no genome-scale information

- Oligonucleotide probe capture
  - May be practical solution: Restricts to known knowns

- Antibody capture
  - E.g. Immunomagnetic separation (Chlamydia)
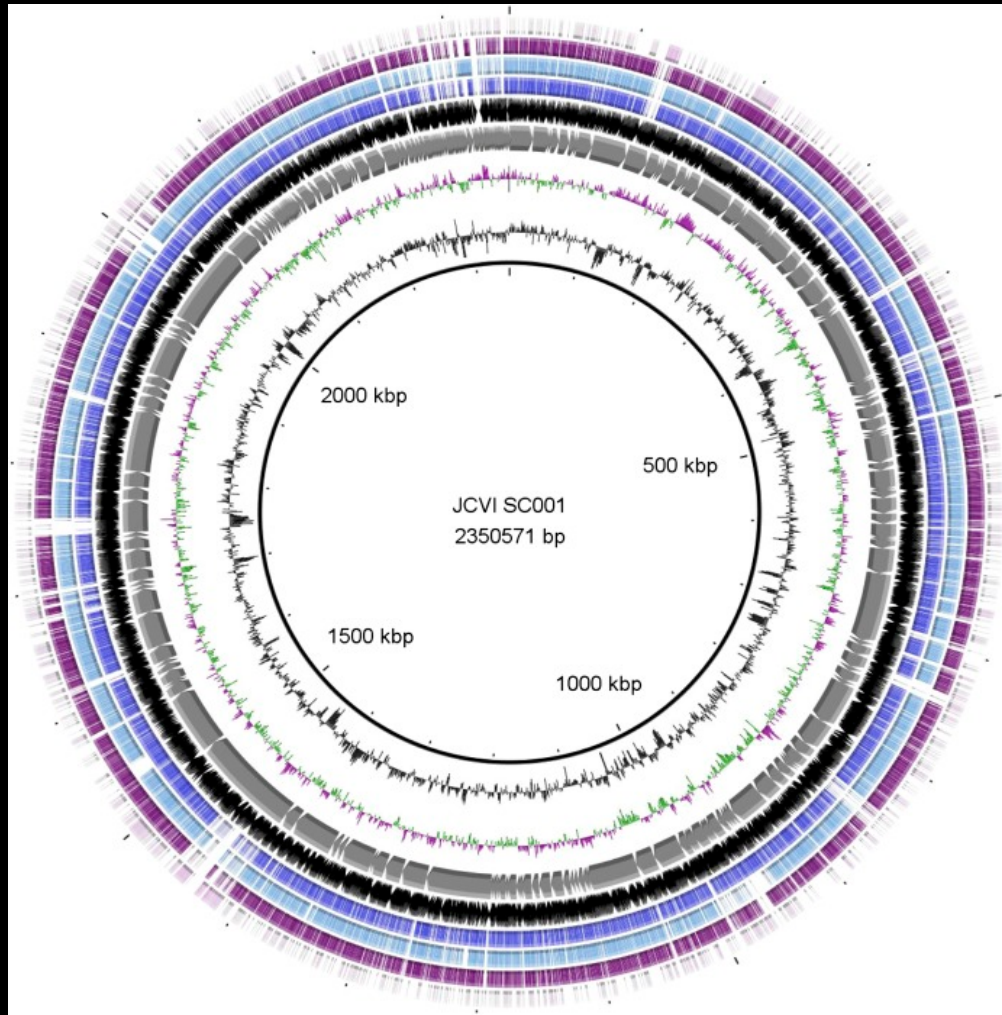
- Cell sorting & whole-genome amplification

Single cell bacterial genomics

McLean et al. PMID:23564253

Experimental design questions

- How many samples for statistical power?
- Replicated design? Biological or technical replicates?
- Extraction method(s) ?
- Level of background contamination?
- Necessary controls
- Longitudinal and/or cross-sectional?
- Sequencing technology: read length
- How many reads?

Taxonomic and functional assignments

- Read-based
- Environmental gene tag or contig-based (from assembly)
- Pathway-based
- Genome-based (from a great assembly)

Adam Bazinet, http://www.biomedcentral.com/1471-2105/13/92
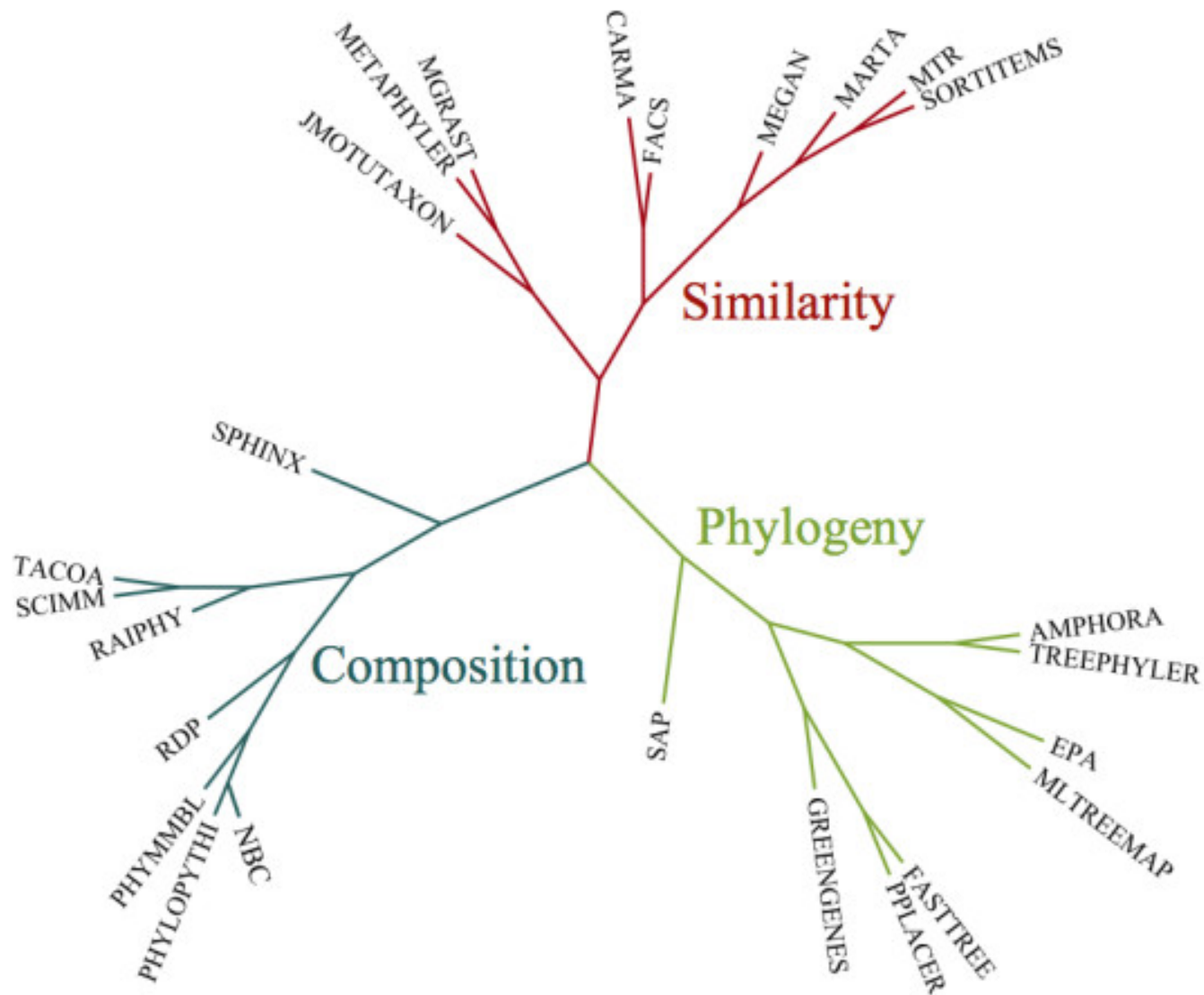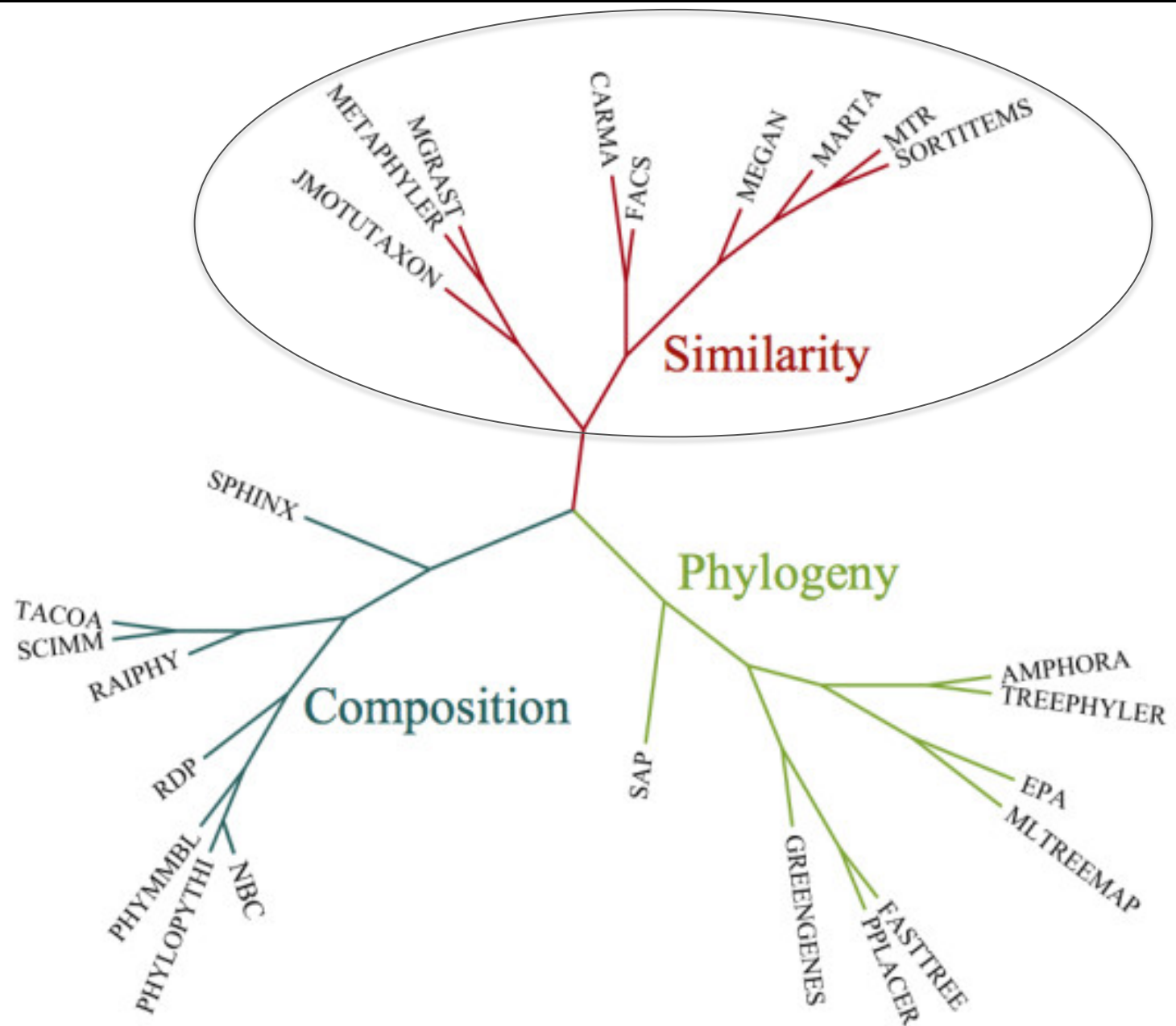
Similarity methods

- Choice of aligner determines speed
- Can make inference from a single read
- Can work in translated BLAST mode
- Very sensitive

- BLAST is very slow!
- Specificity can be low with naïve approach
- Many reads are taxonomically uninformative or taxonomically misleading

Aligners

- BLAST

- RAPSearch2
- LAST

- BWA/Bowtie (in 6-frames with PAUDA)

Caveat emptor: Similarity methods

MG-RAST
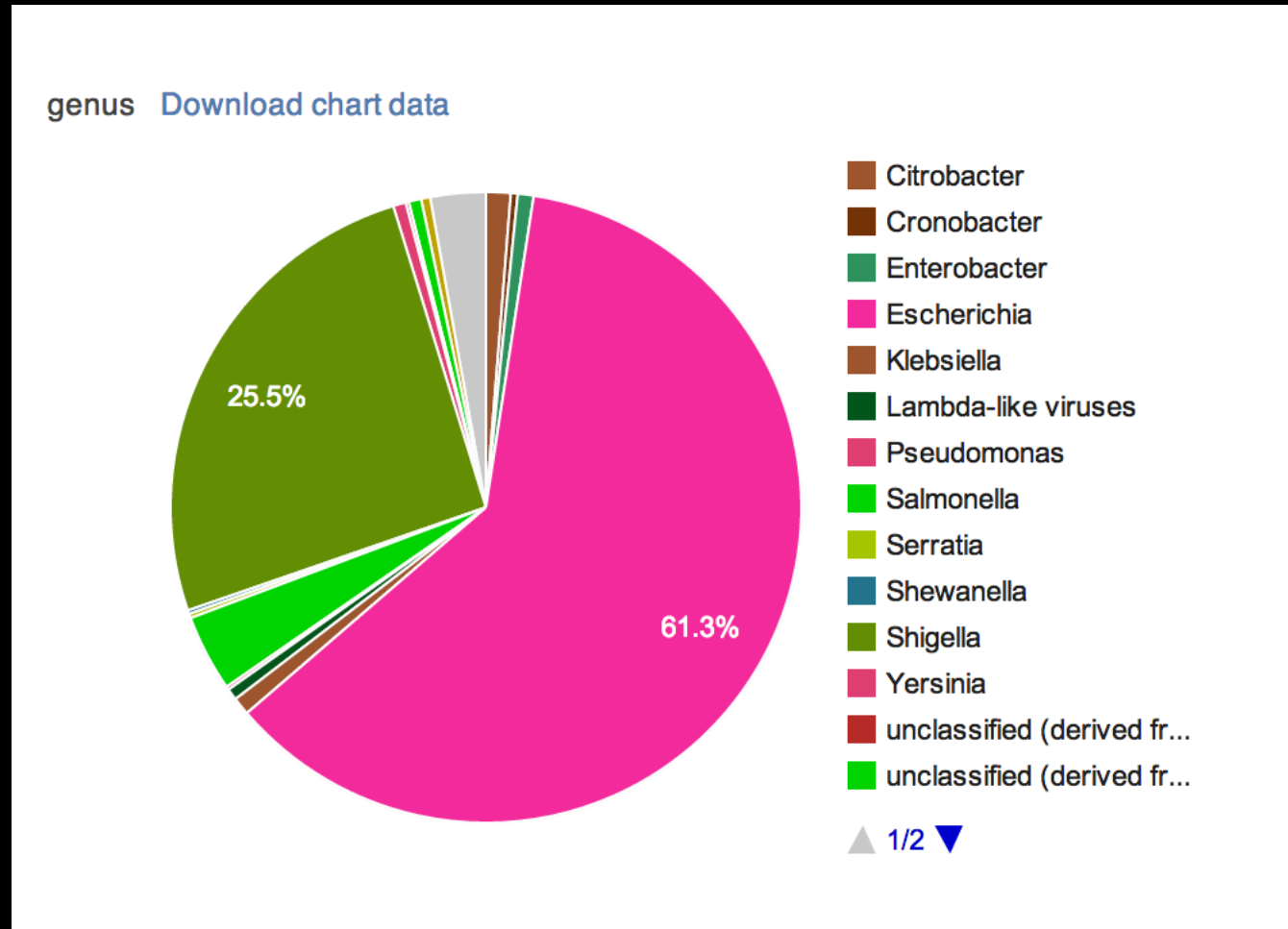
MEGAN

Web   Images   Maps   Shopping   News   More ▾   Search tools

About 1,010,000,000 results (0.42 seconds)

Megan - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Megan ▾
Megan (also spelled Meghan, Meagen, Meagan, Meaghan etc.) is a Welsh female
name, originally a pet form of Meg or Meggie, which is itself a short form of ...
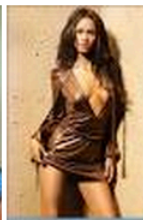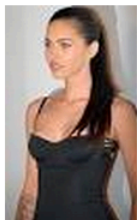Megan Mullally - Megan Fox - Megan Hauserman - Megan Young

Urban Dictionary: Megan
www.urbandictionary.com/define.php?term=Megan ▾
A girl that is very stubborn but at the same time can be the greatest friend on the planet.
She loves pizza and is absolutley gorgeous. She is love...

Images for MEGAN - Report images

Web   Images   Maps   Shopping   Books   More ▾   Search tools

About 2,850,000 results (0.32 seconds)

MEGAN 4 - MEtaGenome ANalyzer — Algorithms in Bioinformatics
ab.inf.uni-tuebingen.de/software/**megan**/ ▾
by AZ Uni - Related articles
3 Dec 2013 - **MEGAN** 4 written by D. H. **Huson**, original design by D. H. **Huson** and
S.C. Schuster, with contributions from S. Mitra, D.C. Richter, P. Rupek, ...
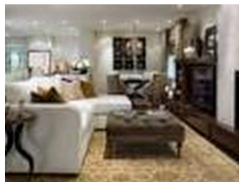MEGAN-download - MEGAN5 - How to use BLAST - Review
You've visited this page many times. Last visit: 21/11/13

MEGAN5 - MEtaGenome ANalyzer — Algorithms in Bioinformatics
ab.inf.uni-tuebingen.de › Software ▾
MEGAN 4 was released at the beginning of 2011 (paper). MEGAN5 will be ...

Images for **MEGAN huson** - Report images

MEGAN - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**MEGAN** ▾
MEGAN ("MEtaGenome ANalyzer") is a computer program that allows optimized
analysis of large ... a b c d **Huson**, H.; A. Auch, Ji Qi, and S. C. Schuster (2007).

MEGAN analysis of metagenomic data
www.ncbi.nlm.nih.gov › Journal List › Genome Res › v.17(3); Mar 2007 ▾
by DH Huson - 2007 - Cited by 737 - Related articles
Daniel H. **Huson**,1,3 Alexander F. Auch,1 Ji Qi,2 and Stephan C. Schuster2,3 ...
MEGAN is then used to compute and explore the taxonomical content of the data ...
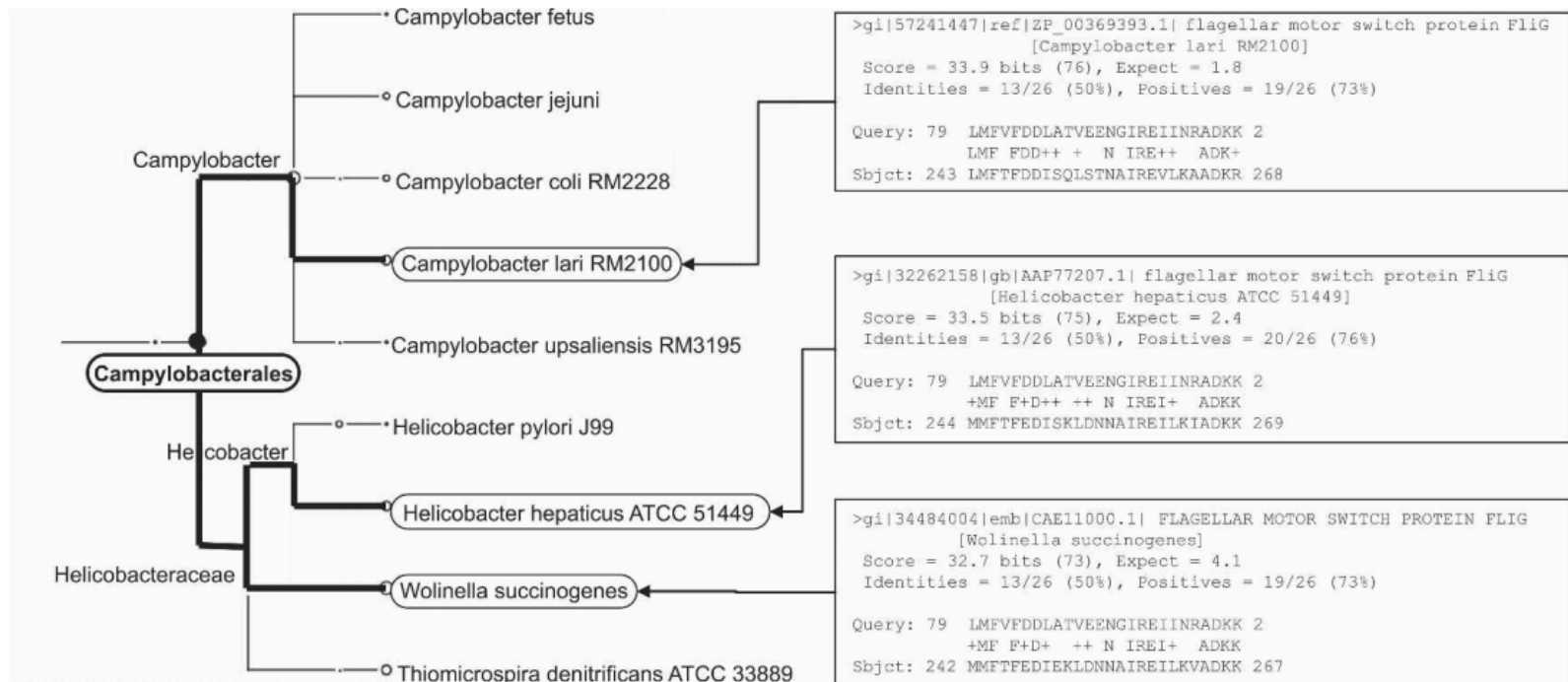
MEGAN

Figure 2. On the *right*, we list the three BLASTX matches obtained for a specific read *r* from the mammoth data set, to sequences representing *Campylobacter lari*, *Helicobacter hepaticus*, and *Wolinella*, respectively. The LCA-assignment algorithm assigns *r* to the taxon *Campylobacterales*, shown on the *left*, as it is the lowest-common taxonomical ancestor of the three matched species.

Huson et al. 10.1101/gr.5969107
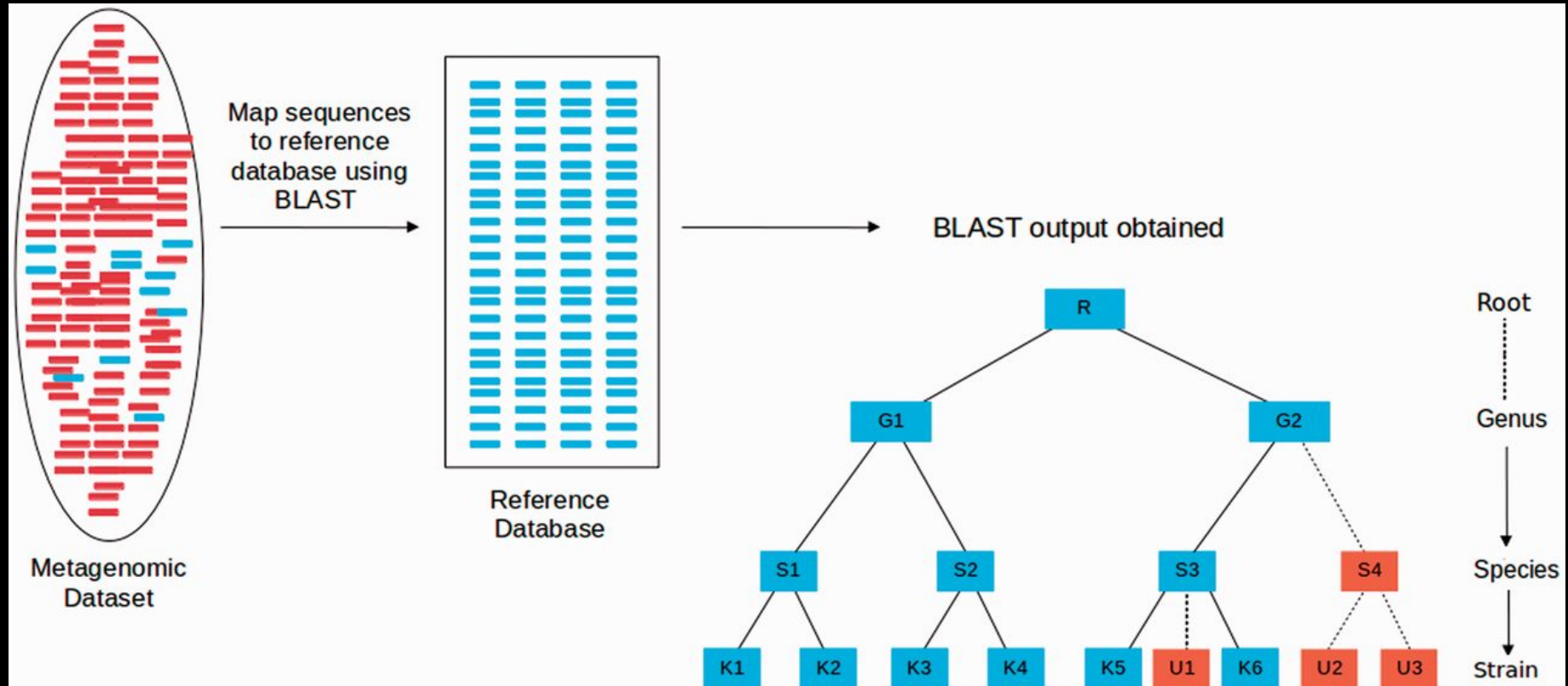
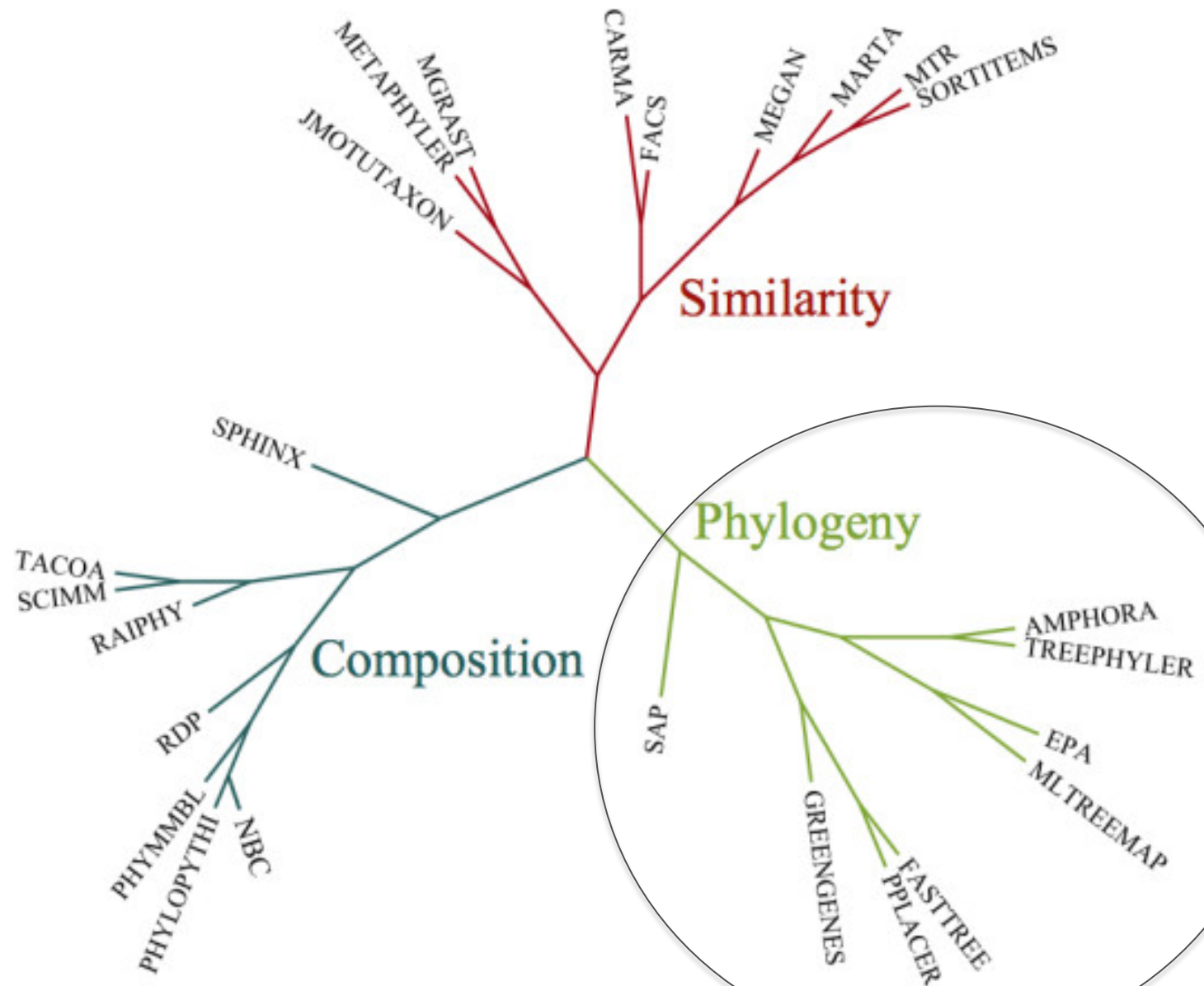| Reads originate from	Significant BLAST Hits	Assingment Strategies	
		Best BLAST Hit Approach	LCA
K1	K1, K2, K3	K1 (√)	G1 (√)
U1	K5, K6	K5 (X)	S3 (√)
U2 and U3	K5, K6	K5 (X)	S3 (X)

Table 1. Results for *E. coli* simulation

	35 bp	100 bp	200 bp	800 bp
Enterobacteriaceae	22%	64%	73%	85%
Gammaproteobacteria	24%	77%	86%	94%
Proteobacteria	25%	83%	89%	96%

For average read lengths of 35, 100, 200, and 800 bp, we sampled 5000 sequence intervals from random locations in the complete genome sequence of *E. coli K12* and then processed the reads with MEGAN. Here we report the percentage of reads classified as Enterobacteriaceae, Gammaproteobacteria, and, even more generally, Proteobacteria. The number of false-positive assignments of reads was ~0%.

Huson et al. 10.1101/gr.5969107

Similarity

Phylogeny

Composition

METAPHYLER
MGRAST
JMOTUTAXON
CARMA
FACS
MEGAN
MARTA
MTR
SORTITEMS

SPHINX

TACOA
SCIMM
RAIPHY

RDP

PHYMMBL
PHYLOPYTHI
NBC

SAP

GREENGENES
FASTTREE
PPLACER

AMPHORA
TREEPHYLER

EPA
MLTREEMAP

Phylogenetic methods

- Accurate
- Give measurement of uncertainty
- Can use multiple conserved marker genes (~40 common)

- Slow
- Dependent on quality and completeness of reference tree
- Prokaryo-centric

A) The MLTreeMap pipeline

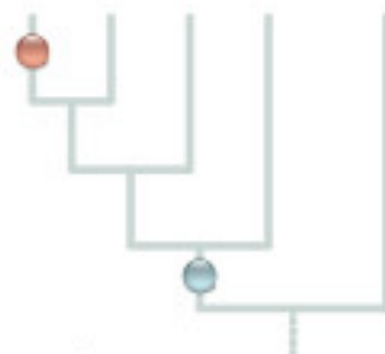1) identify marker genes in raw nucleotide sequence fragments

[BLAST, GeneWise]

2) add detected genes to curated reference alignments
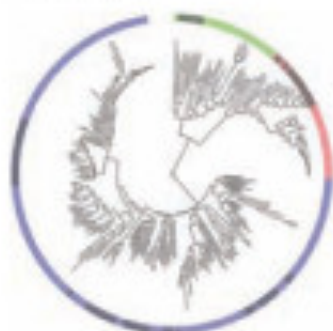
[hmmalign, Gblocks]

3) Maximum Likelihood placement into annotated reference phylogenies

[RAxML]

B) Reference phylogenies in MLTreeMap

Tree of life

Archaea, Bacteria, Eukaryota

RuBisCo

Forms 1, 2, 3, 4a, 4b and unclassified

nifH

Groups 1, 2, 3, 4

nifD

Groups 1, 2, 3 and unclassified

MMO

Methane Monooxygenase

Nick Loman
@pathogenomenick

Pork knee, axe in head, cesky krumlov style #evomics2014 pic.twitter.com/lEp2joTnf4

← Reply   🗑 Delete   ★ Favorite   ••• More

Twitter can be useful?!

S. Strycharz-Glaven @SGlaven   Jan 9

@SahaSurya @subsurface_life @Dr_Bik @pathogenomenick
so how do you choose? Try ur data with all and see which
one is prettiest?

Details   ← Reply   ↻ Retweet   ★ Favorite   ••• More

Nick Loman @pathogenomenick   Jan 9

@SGlaven @SahaSurya @subsurface_life @Dr_Bik that's
the usual way with bioinformatics software, right?

Details   ← Reply   ↻ Retweet   ★ Favorite   ••• More

Surya Saha @SahaSurya   Jan 9

@pathogenomenick @SGlaven @subsurface_life @Dr_Bik
Yes :-) Or come up with test set(s) and get TP/FP/Sp/Sn.
Your mileage *will* vary

Details   ← Reply   ↻ Retweet   ★ Favorite   ••• More
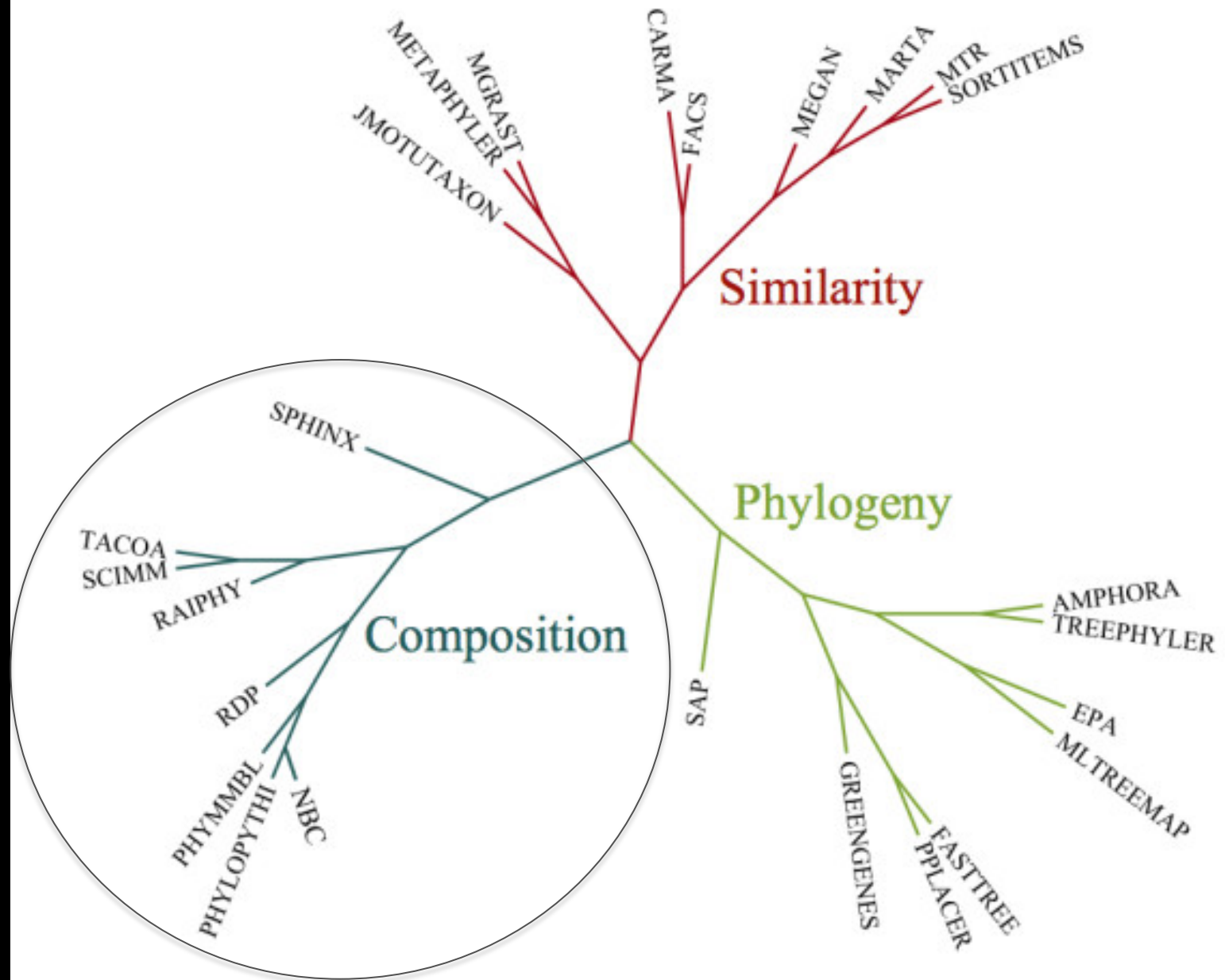

Sam Minot @sminot   Jan 9

@SahaSurya @pathogenomenick @SGlaven
@subsurface_life @Dr_Bik +1 to using a rigorous test set

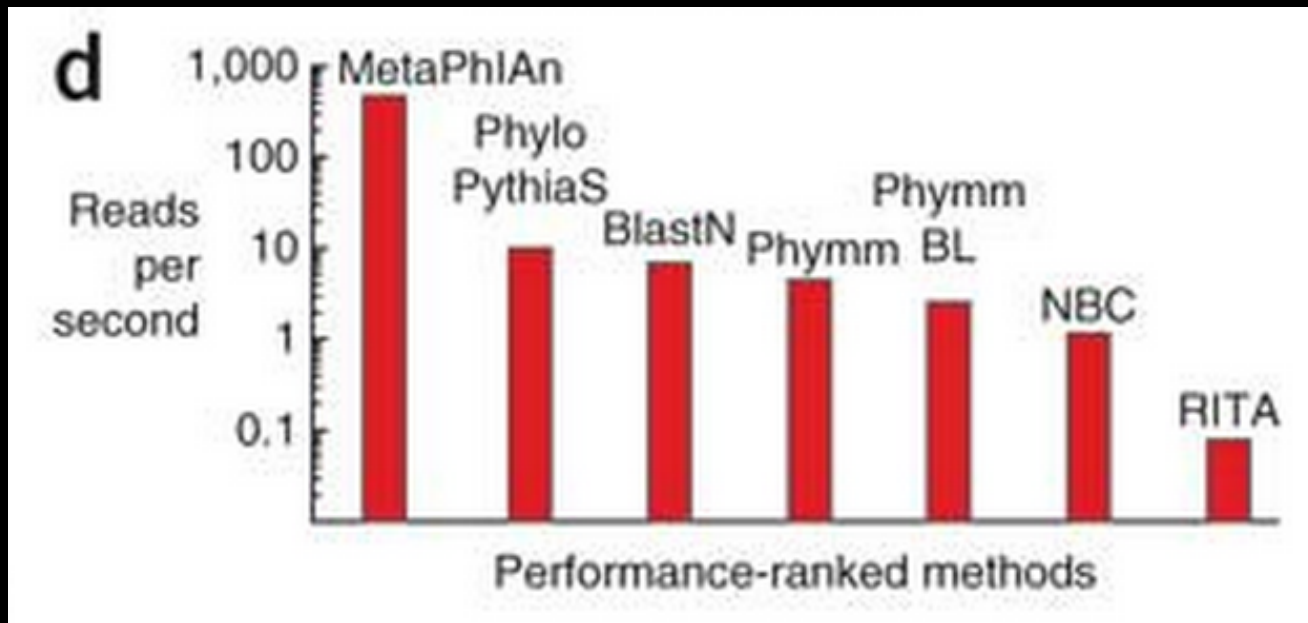Details   ← Reply   ↻ Retweet   ★ Favorite   ••• More
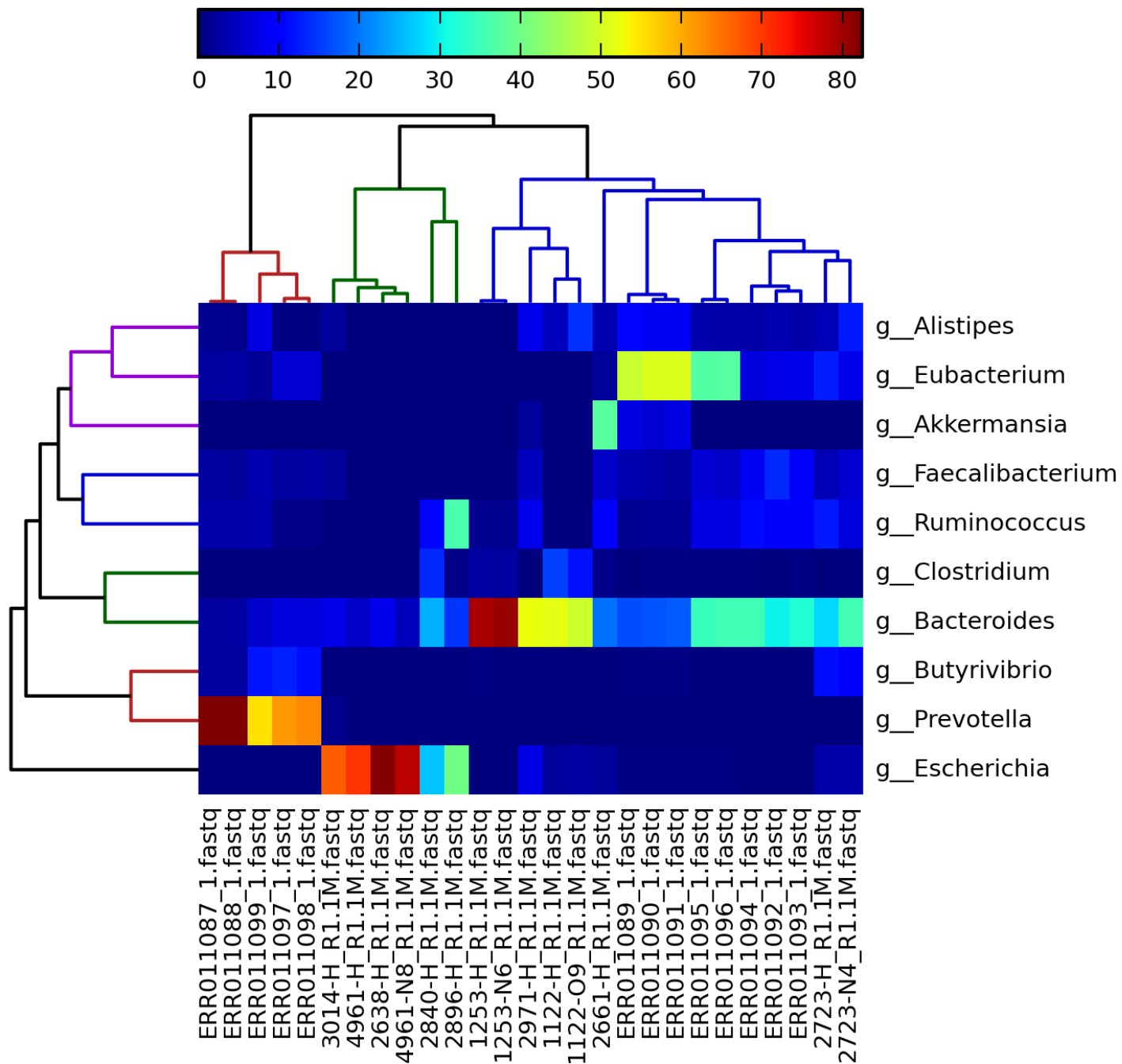
Phylogenetic assignment

- mOTUs
- Phylosift
- MLTreeMap

Metaphlan

- Lineage-specific marker genes
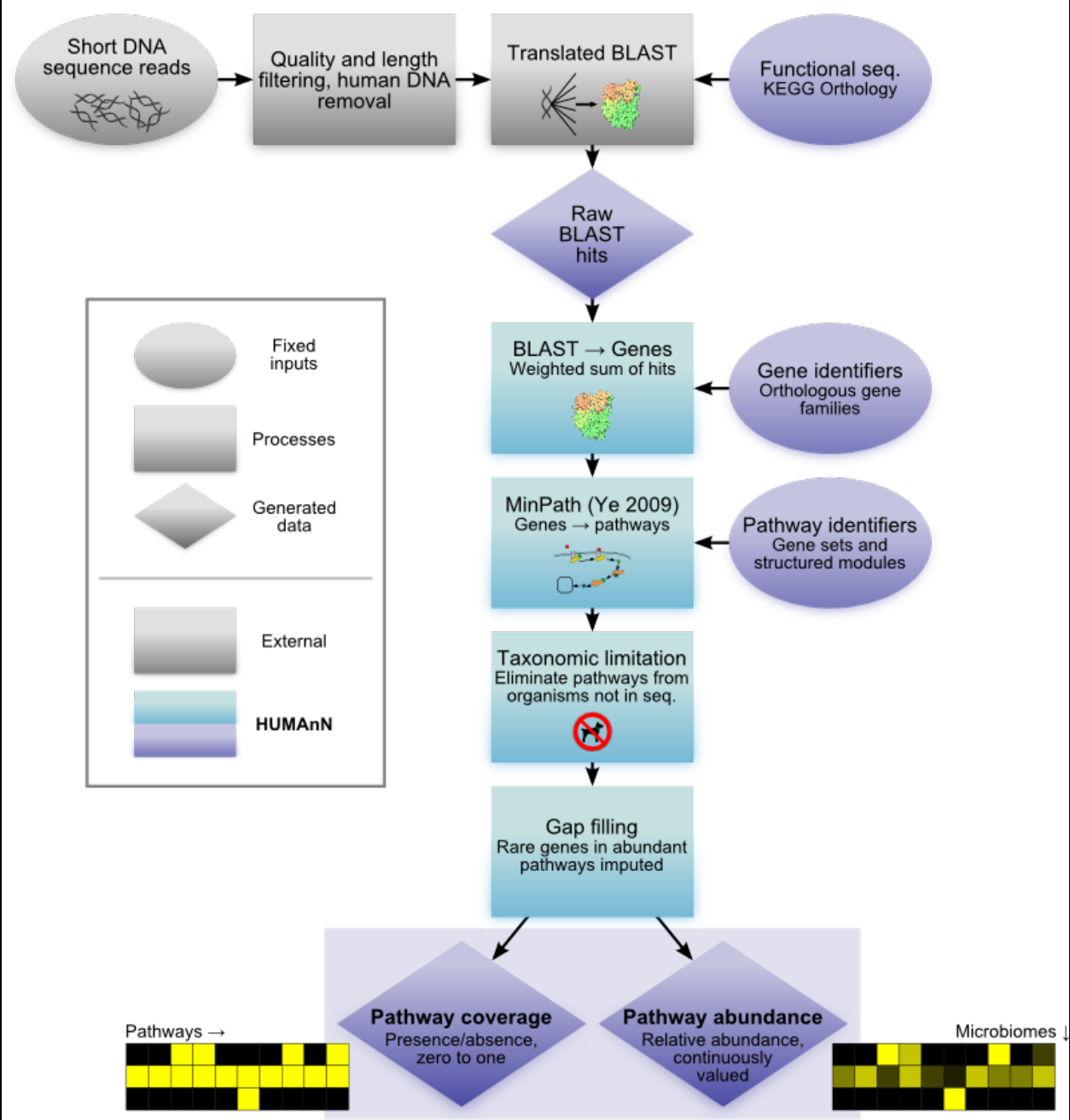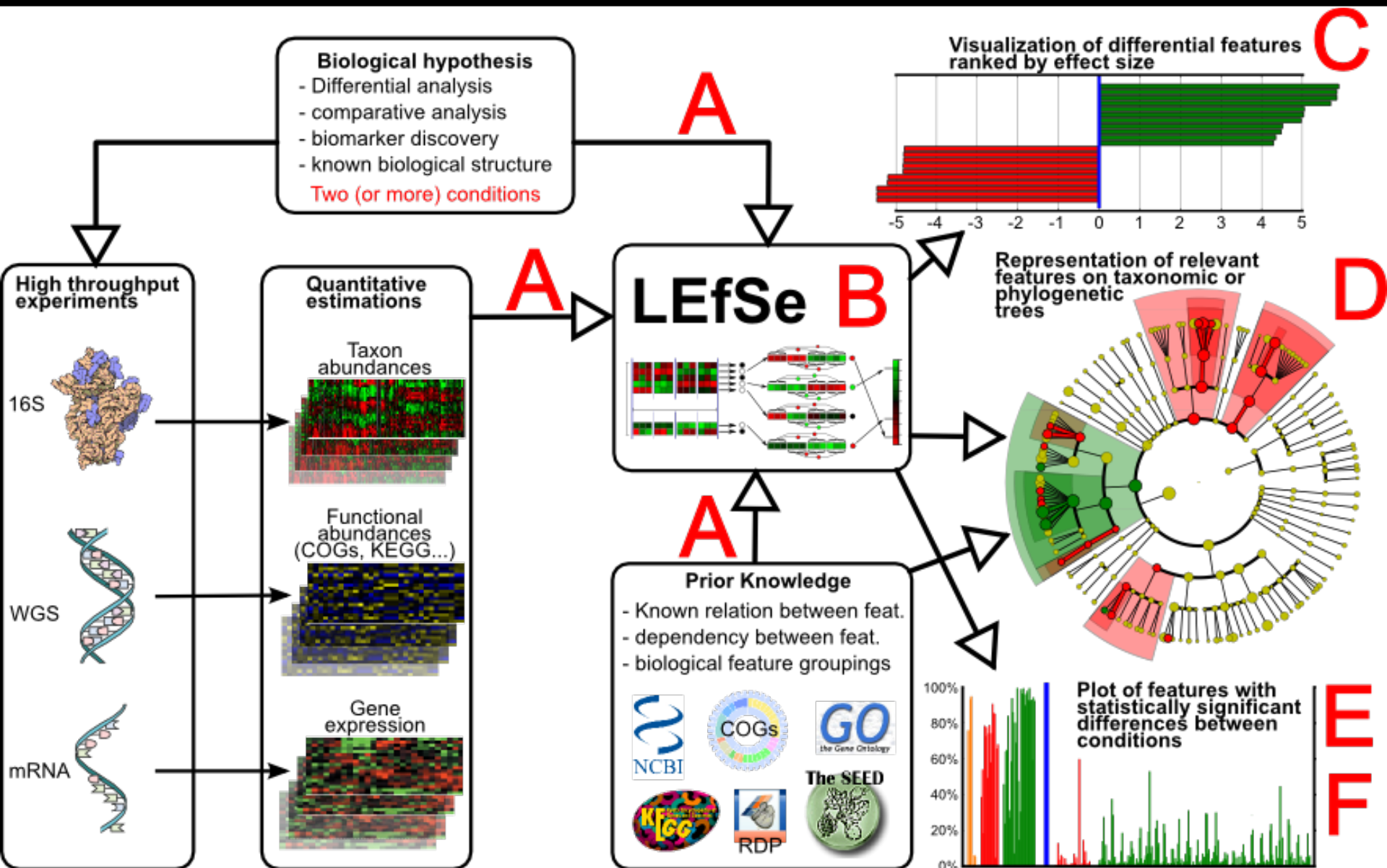- Hybrid similarity and phylogenetic method

Functional assignments

- SEED
- COGs
- EggNOG
- KEGG

HUMANn

Abubaker et al. 10.1371/journal.pcbi.1002358

LEfSe

Metagenomics assembly

- A great way of compressing your dataset
- Fundamental unit of analysis are now contigs (environmental gene tags)
- Assignment methods can be used as before (+/- gene prediction)
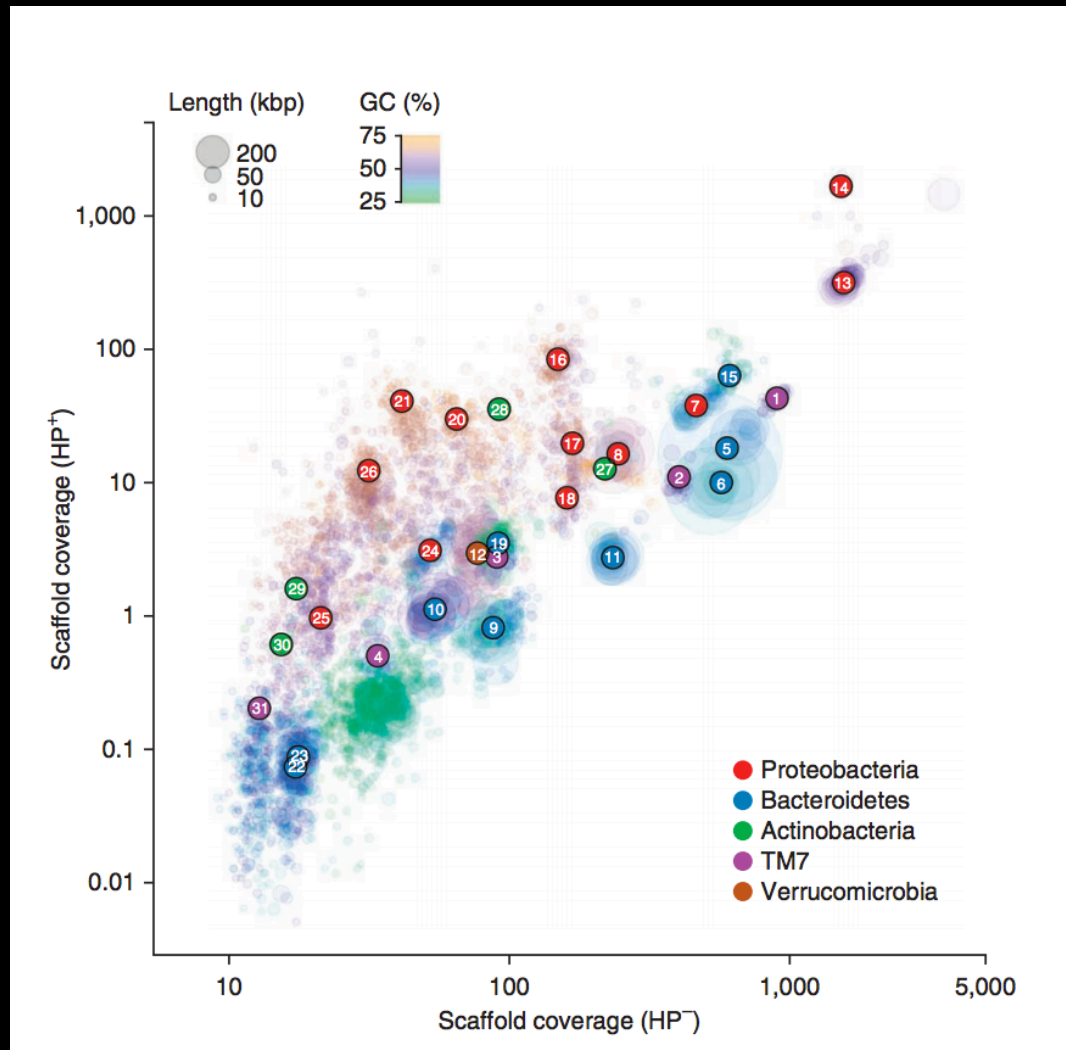- Coverage information can be restored through mapping reads to assembly

Metagenomics assembly: issues

- Large datasets may overwhelm available RAM
- Contigs may represent chimeras of multiple, related species or strains
- Low coverage organisms won't assemble

http://davis-assembly-masterclass-2013.readthedocs.org/en/latest/outputs/opinionated-guide.html

Opinionated guide to assembly

- Metagenome assembly:
  - Ray
  - Velvet
  - IBDA-UD
- (Optional) digital normalization

http://davis-assembly-masterclass-2013.readthedocs.org/en/latest/outputs/
opinionated-guide.html

Genome binning

http://madsalbertsen.github.io/multi-metagenome/

CONCOCT: **C**lustering c**ON**tigs on **CO**verage and **C**omposi**T**ion

- Data pre-processing:
  - Perform coassembly across all samples

  - Map reads back to contigs to get mean coverage of contig in each sample

  - Generate k-mer frequency vector for each contig

  - Join vectors and log transform

  - PCA

Oxford Nanopore: Hope or hype?

Strand sequencing
Biologically modified nanopore
Tri-nucleotides detected
Theoretically unlimited read length
Sequence directly without extensive sample preparation
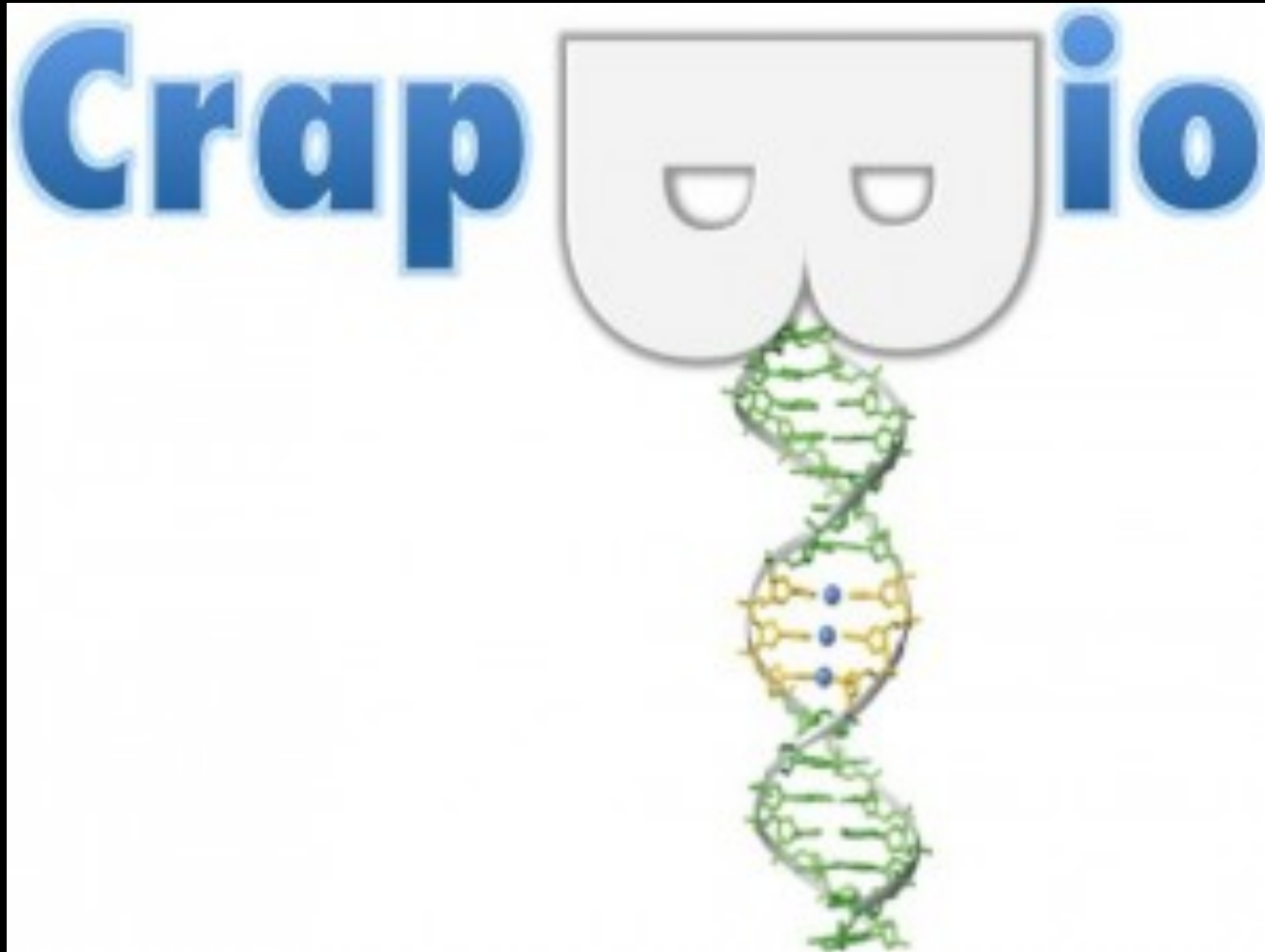
MinIon: 150mb/hour for up to 6 hours for $900 – disposable
GridIon: rack-mounted sequencer with 2000-8000 nanopores

MinION access programme starting end November

http://pathogenomics.bham.ac.uk/blog/2013/10/the-oxford-nanopore-golden-ticket/

SHTSeq™

Metagenomics - Questions

- What are the best ways to address getting representation of bacteria, viruses, fungi and others? Techniques for doing so?
    - Thoughts on the use of physical enrichment techniques to isolate microbe of interest rather than traditional metagenomic sequencing?
- What are the best bioinformatic software packages and pipelines for functional analysis?
    - What are the best analysis pipelines for full viral sequencing to detect whether mutations are true or not? Comparing closely related taxa?
- As an initial approach, should one try 16s sequencing prior to shotgun sequencing if interested in bacteria (or 18s/ITS1 for Fungi)? Which region?
- Shotgun metagenomics versus single cell genomics - for high diversity samples is a shift toward single cell beneficial?
    - Any expertise in microbial or viral single cell genomics? Software suggestions for assembling viral genomes and large scale microbial genome comparison?
- Metatranscriptomics versus metagenomics?
    - Benefits/disadvantages of each?
- Best tools for de novo assembly and annotation
    - Most useful databases for metagenomics
    - Thoughts on combining methodologies - i.e. PacBio sequencing for scaffolding and Illumina/454 for depth/decreased error?

Practical

- All read data

- Subsampling

- Metaphlan assignments (and heatmaps)

- MEGAN
  - Taxonomic assignments with least common ancestor

  - Functional assignments with KEGG

  - Comparison of metagenomics data

Acknowledgements