Decoding our bacterial overlords

Torsten Seemann

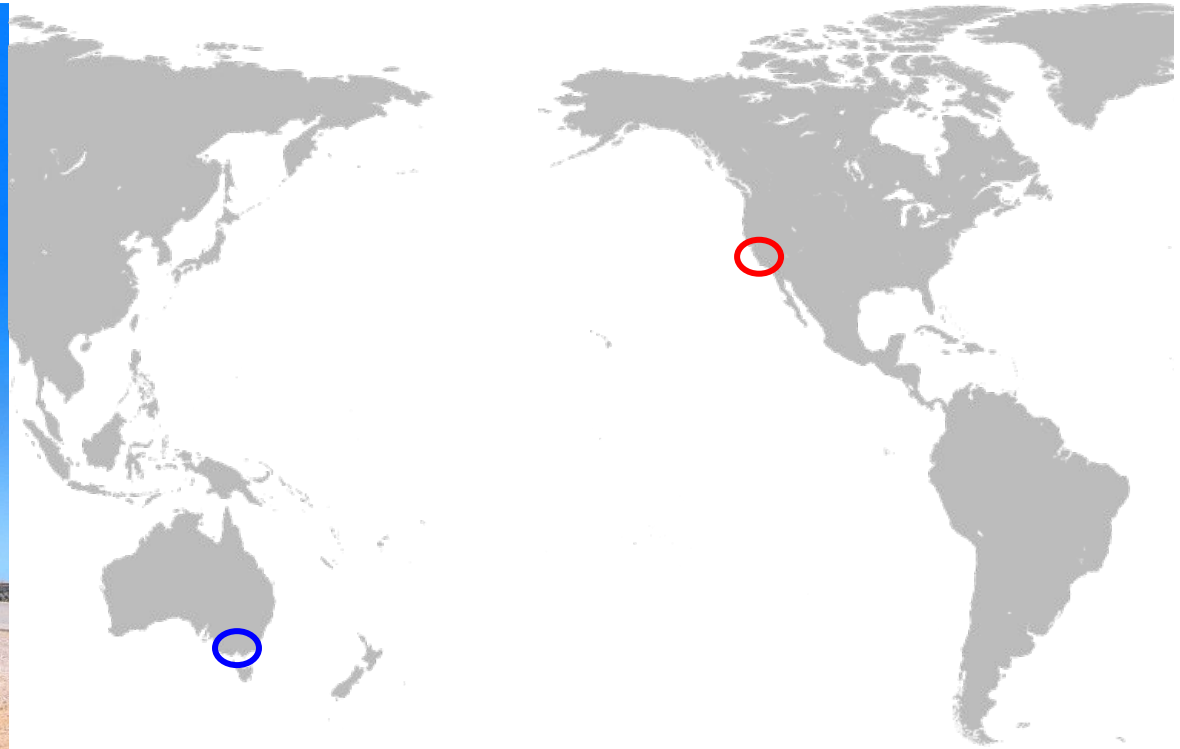
@torstenseemann

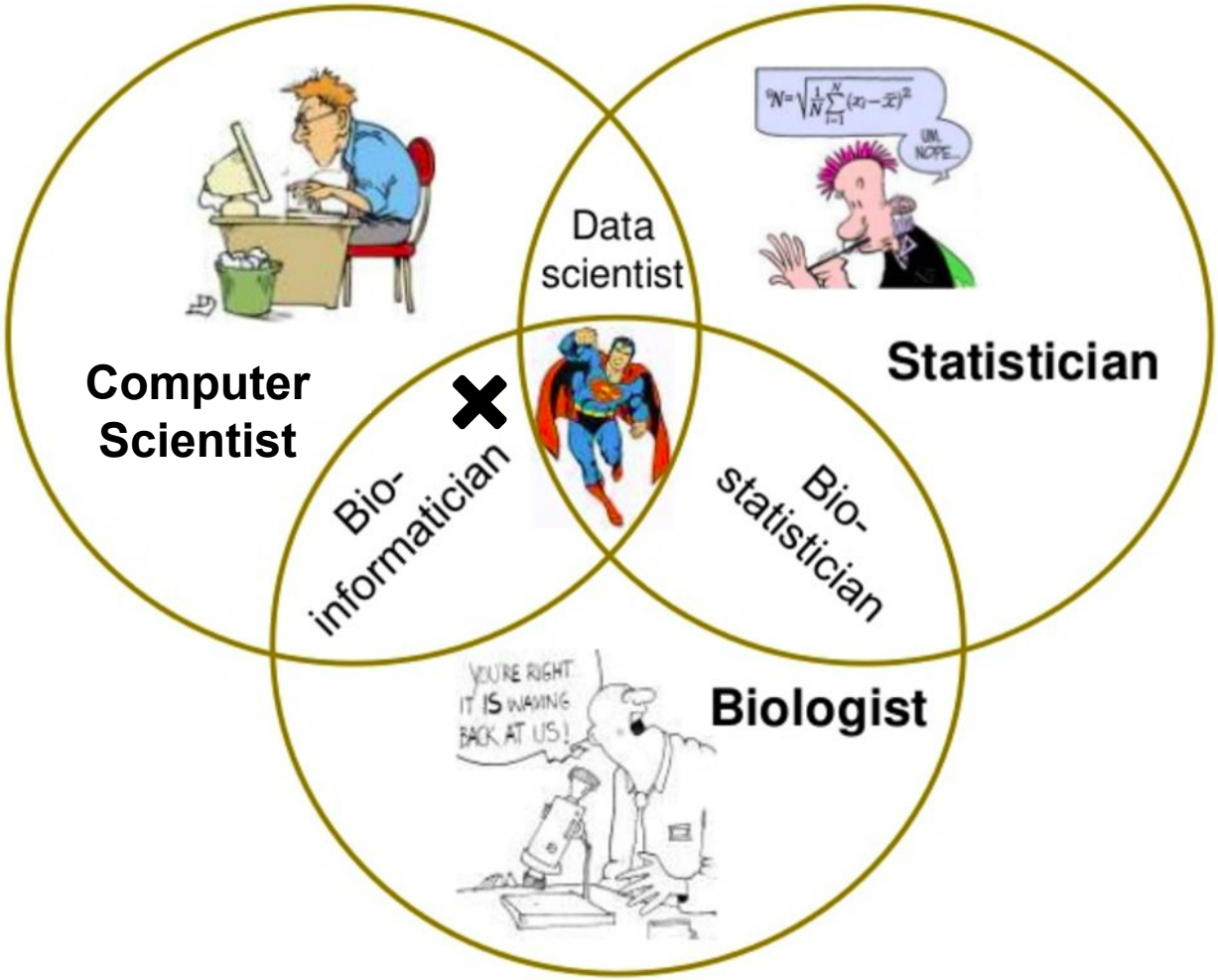
Dedication

This presentation is dedicated to Sabah, whose body is desperately trying to rid itself of some microbes which managed to get somewhere they shouldn't have.

About me

Melbourne, Australia




Data scientist
Computer Scientist
Statistician
Bio-informatician
Bio-statistician
Biologist
UM, NOPE...
YOU'RE RIGHT IT IS WAVING BACK AT US!

Doherty Institute

THE UNIVERSITY OF MELBOURNE

The Royal Melbourne Hospital

"Immunity and infection"

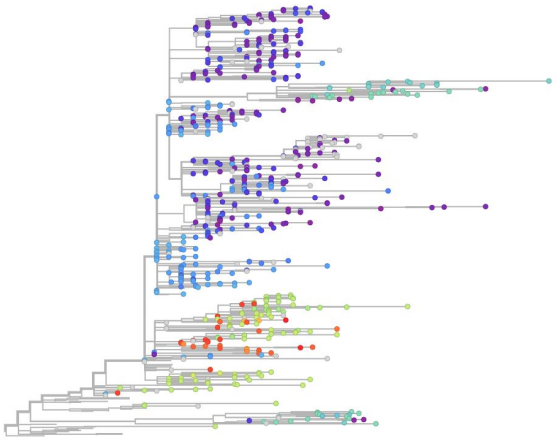
- Research
- Teaching
- Public health and reference labs
- Diagnostic services
- Clinical care in ID and immunity

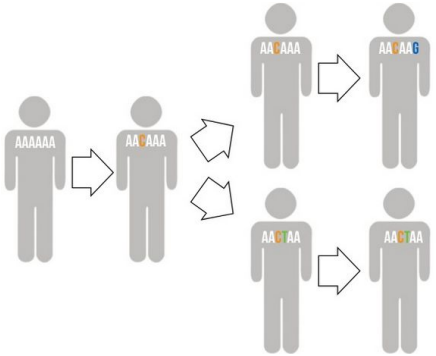
Microbiological Diagnostics Unit

- Public health microbiology lab
- Established in 1897
- Within a University Micro Dept

- Co-locates microbiologists, clinicians, bioinformaticians and epidemiologists
- Strong research links

Mandate: apply WGS *wherever it makes sense*



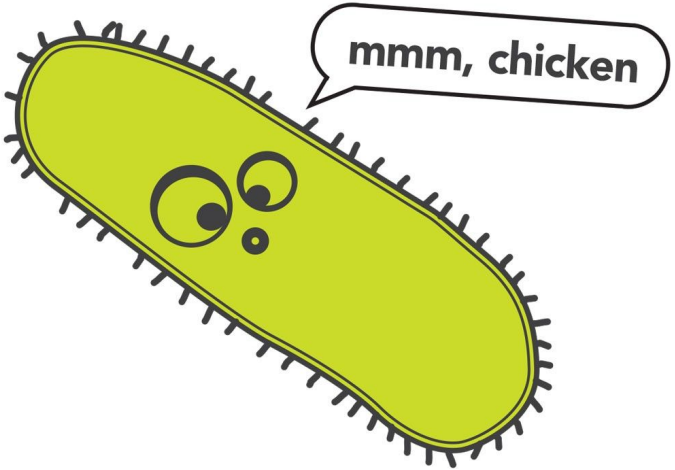


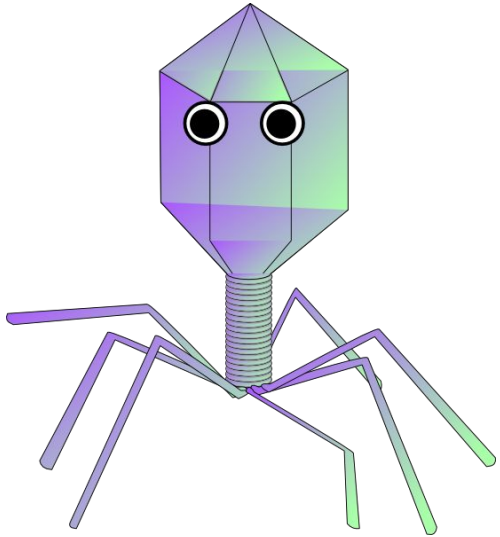
Diagnostics

Surveillance

Outbreak response

Bacteria, Viruses, Archaea, Fungi, Protists





Foodborne, human (clinical), animal and environmental samples

Lots of genomics & transcriptomics

High quality “first” genomes

- Capillary sequences + BACs + primer walking
  - *2006 - Leptospira borgpetersenii* (abortive agent - cows)
  - 2007 - *Mycobacterium ulcerans* (Buruli ulcer - human)
  - 2008 - *Mycobacterium marinum* (Fish granuloma - model for TB)
- Roche 454 + Illumina + BACs + primer walking
  - 2012 - *Enterococcus faecium* (Human pathogen, highly resistant)
- Ion Torrent + Illumina + BACs
- Pacbio RSII + Illumina (~50 done)
- Nanopore + Illumina

Software tools for microbial genomics

BIOINFORMATICS APPLICATIONS NOTE

Vol. 30 no. 14 2014, pages 2068–2069
doi:10.1093/bioinformatics/btu153

Genome analysis

Advance Access publication March 18, 2014

Prokka: rapid prokaryotic genome annotation

Torsten Seemann[1,2]

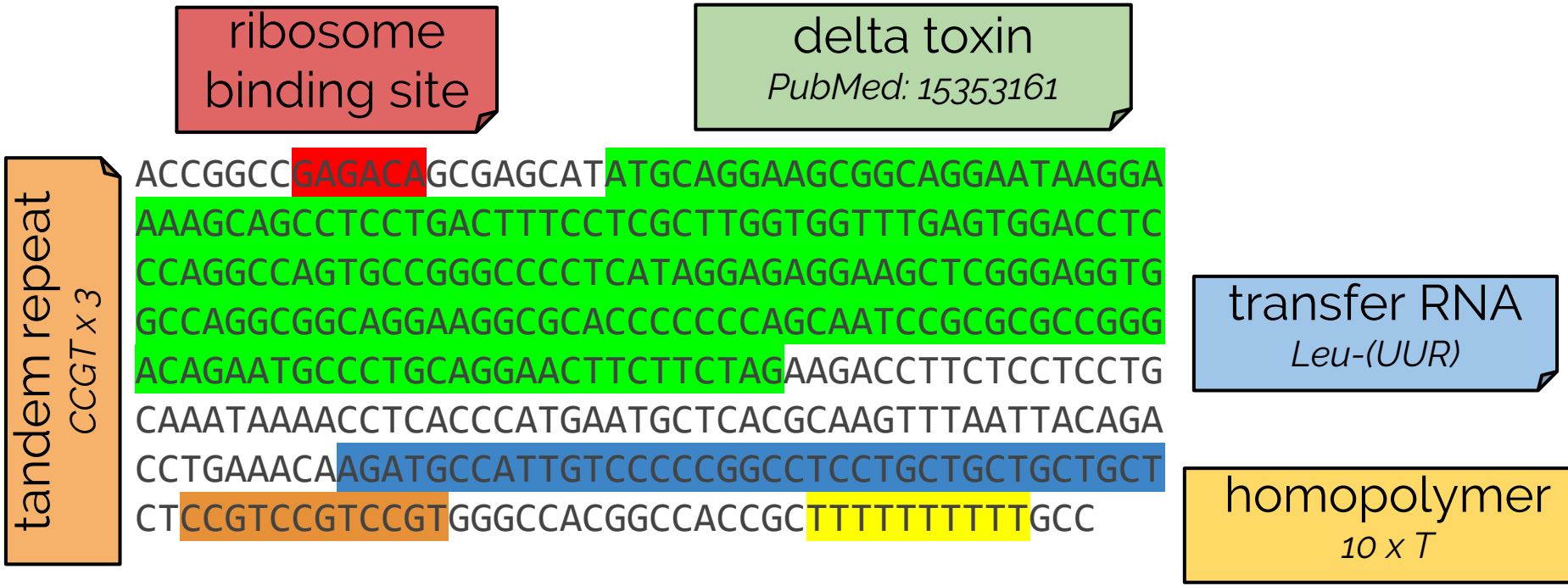
[1]Victorian Bioinformatics Consortium, Monash University, Clayton 3800 and [2]Life Sciences Computation Centre, Victorian Life Sciences Computation Initiative, Carlton 3053, Australia

Associate Editor: Alfonso Valencia

TITLE	CITED BY	YEAR
Prokka: rapid prokaryotic genome annotation T Seemann Bioinformatics 30 (14), 2068-9	1841	2014

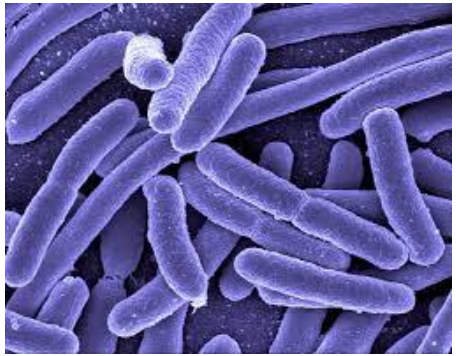
Annotation

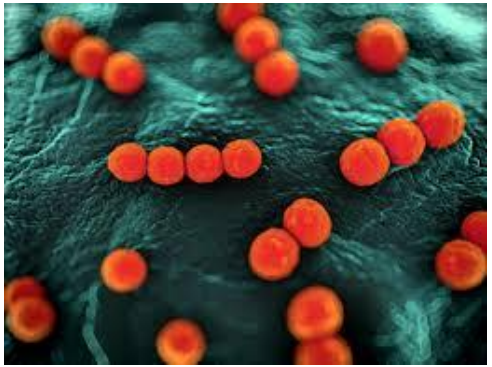
Adding biological information to sequences.

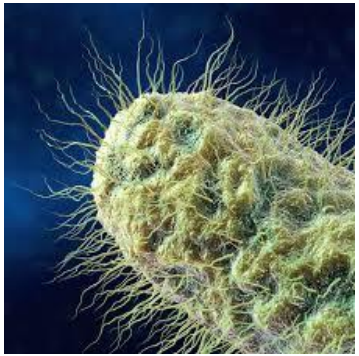


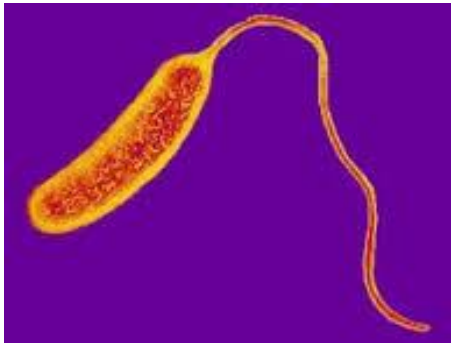
Bacteria

Bacteria are diverse & often super weird









Essential for human life

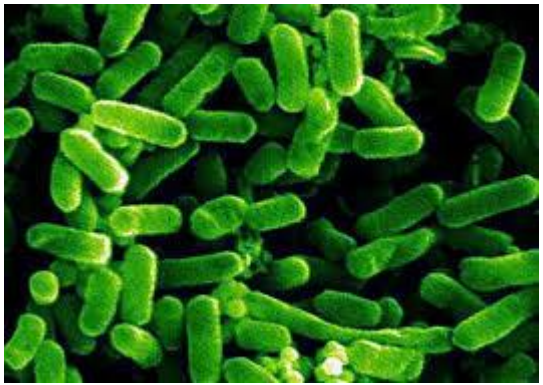
Synthesize vitamins

Help digest our food

Immune system

Bacteria are not malicious

“Good”
(colon)



E.coli

“Bad”
(bladder)

Bacteria run the show

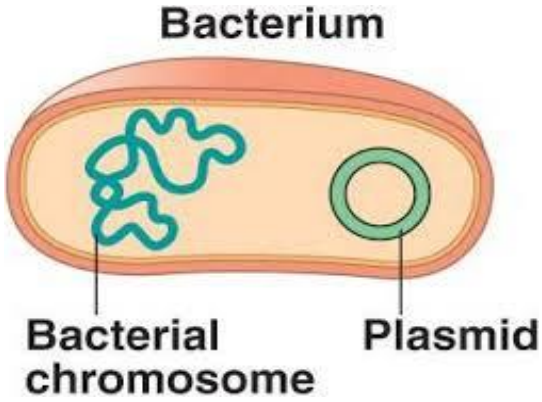
1,000,000

100,000,000,000,000

50-90%
microbial

5,000,000,000,000,000,000,000,000,000,000,000,000
000,000,000,000,000,000,000,000,000,000,000,
000,000,000,000,000,000,000,000.

Replicons



Usually 1 large chromosome (1M to 10M bases)

Sometimes 1-6 “mini” chromosomes (4k - 300k bases)

The circle of life

Bacteria have circular replicons

Bacillus anthracis (Anthrax)

The broken circle of life

Bacteria have circular replicons

Bacillus anthracis (Anthrax)

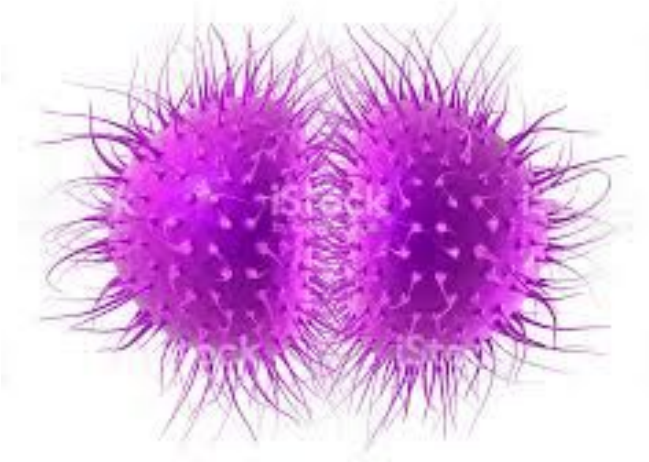
Except when they are linear

Borrelia burgdorferi (Lyme disease)

Bacteria can be polyploid

1

2

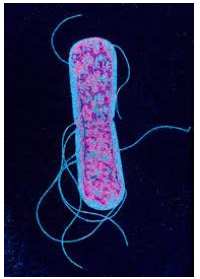


3

4

Neisseria gonorrhoeae has 3-5 copies of its chromosome
Recombination within cell, antigenic variation

Small genome



6,000,000,000 letters

Genome

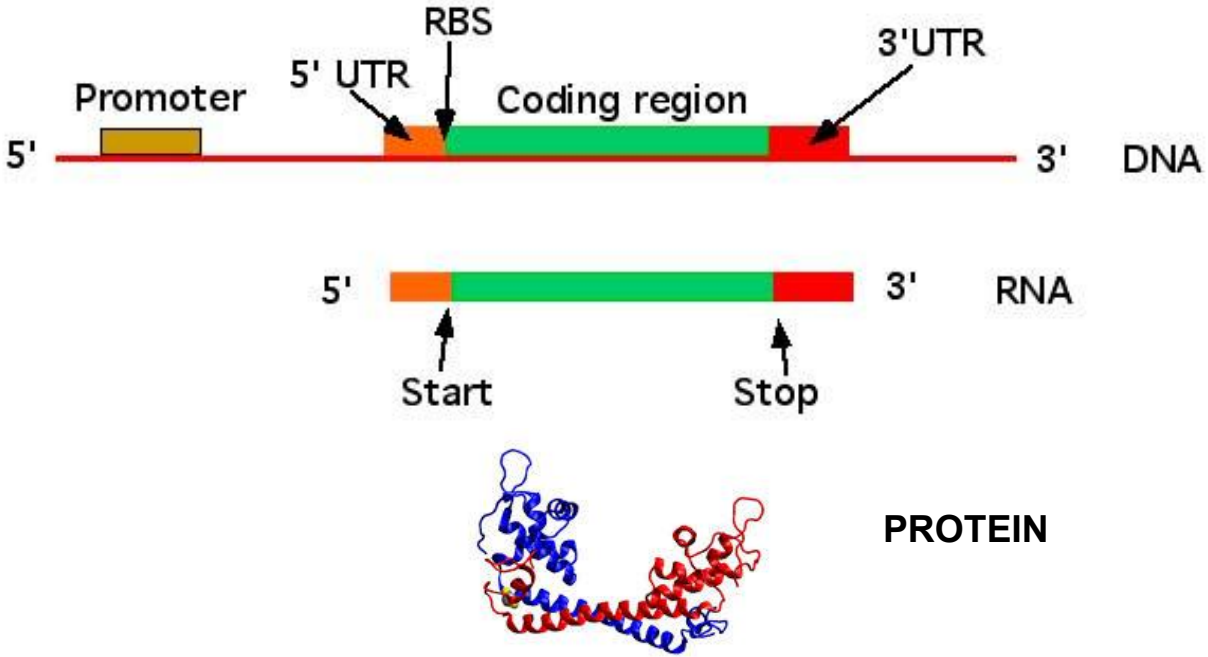
A T G C

3,000,000 letters

30,000 genes

3,000 genes

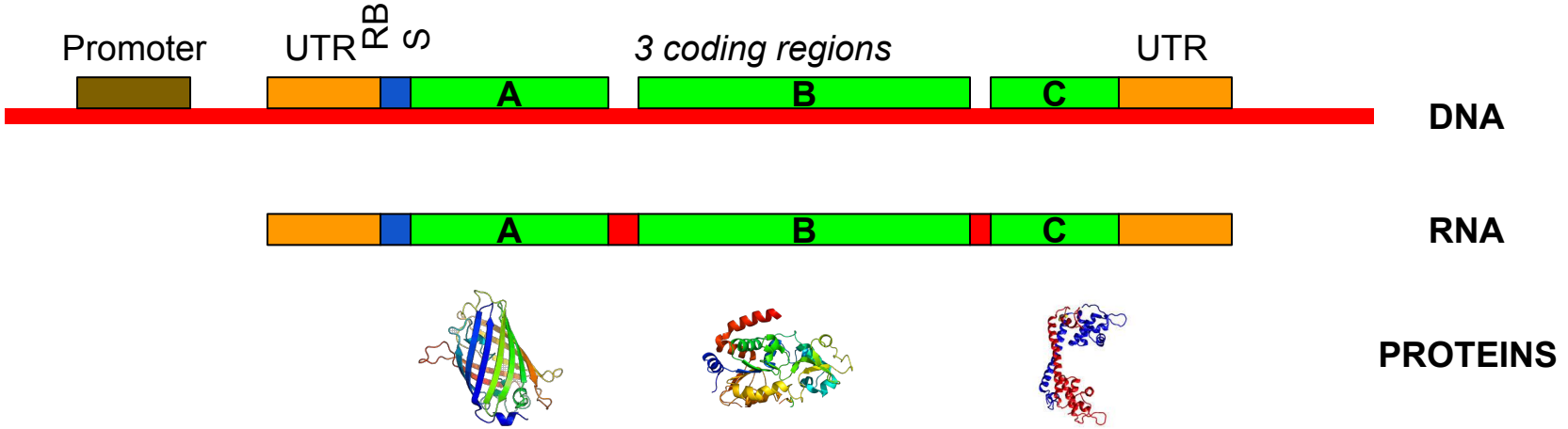
Bacterial genes



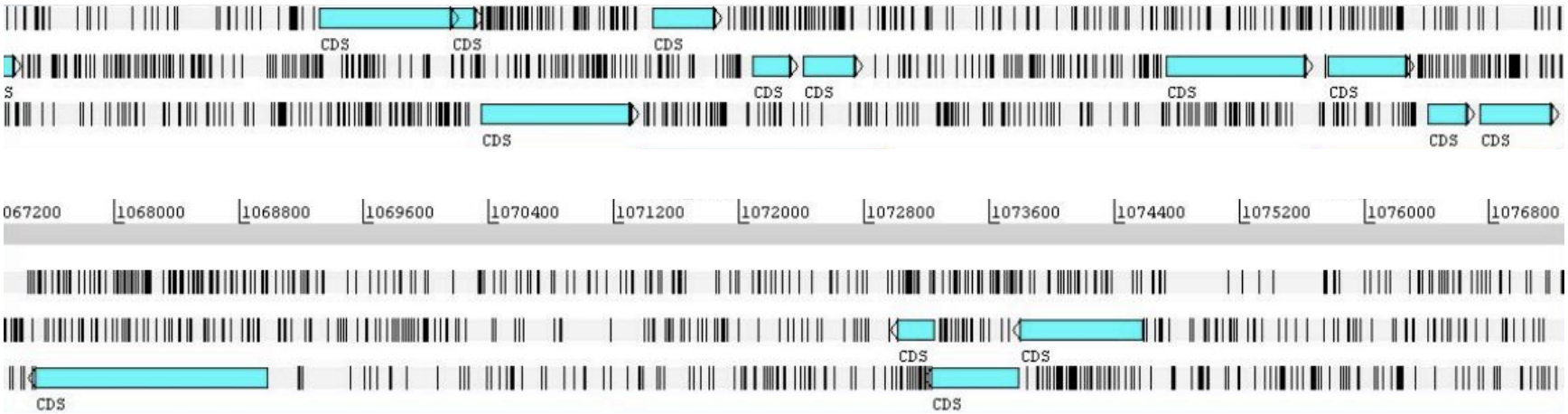
No introns!

Operons

- One RNA transcript, multiple proteins
- Proteins are related: assembly, pathway

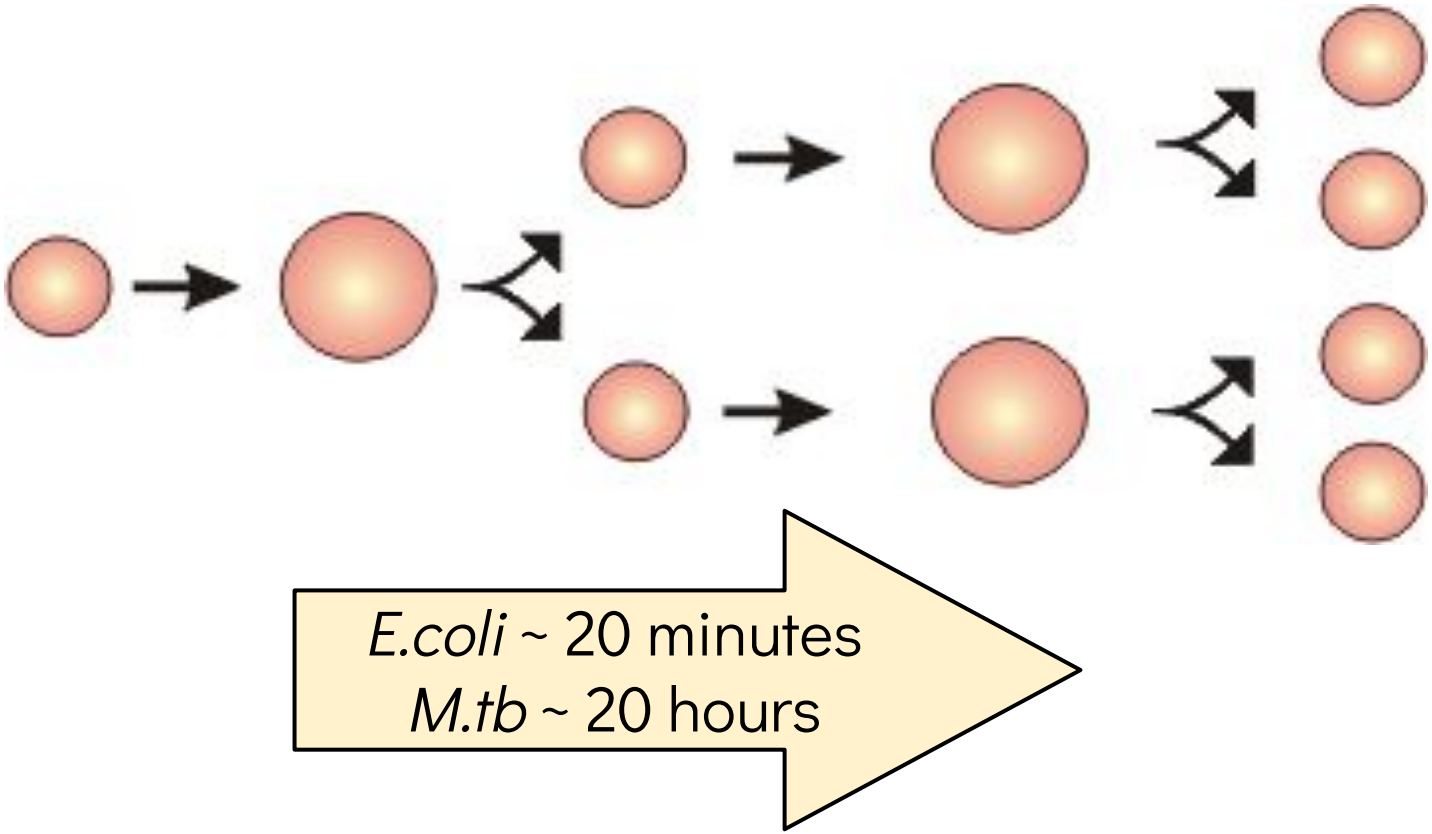


Bacteria are coding dense

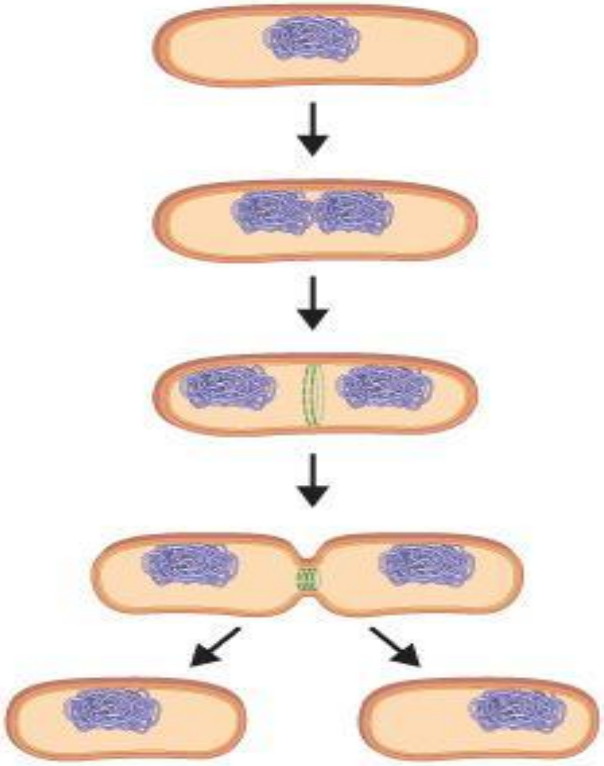


- Overlapping genes
- Very few intergenic regions
- About 1000 genes per 1 Mbp of genome

(Relatively) fast growers



Vertical transfer of DNA



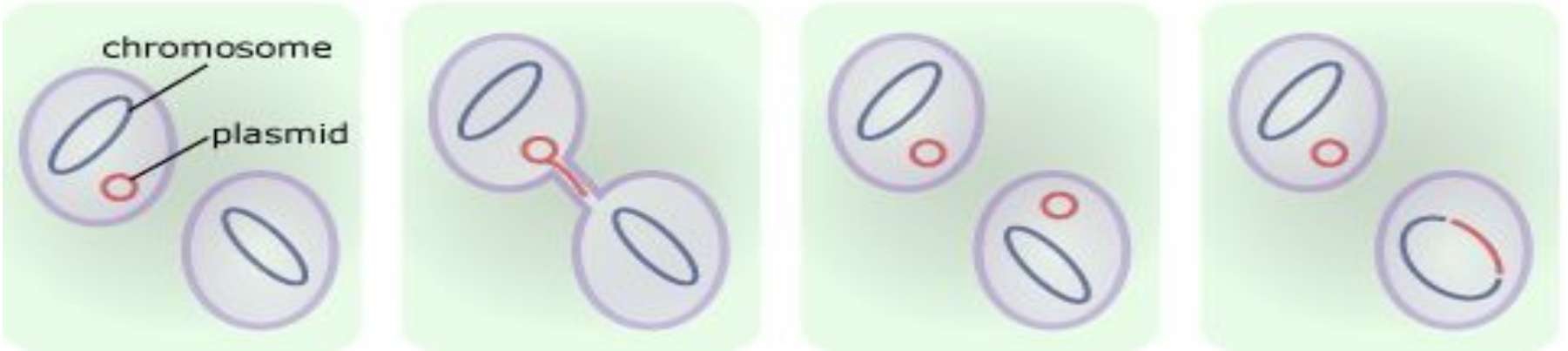
Occurs during cell division

Sometimes it makes an error copying the DNA

eg. A → T

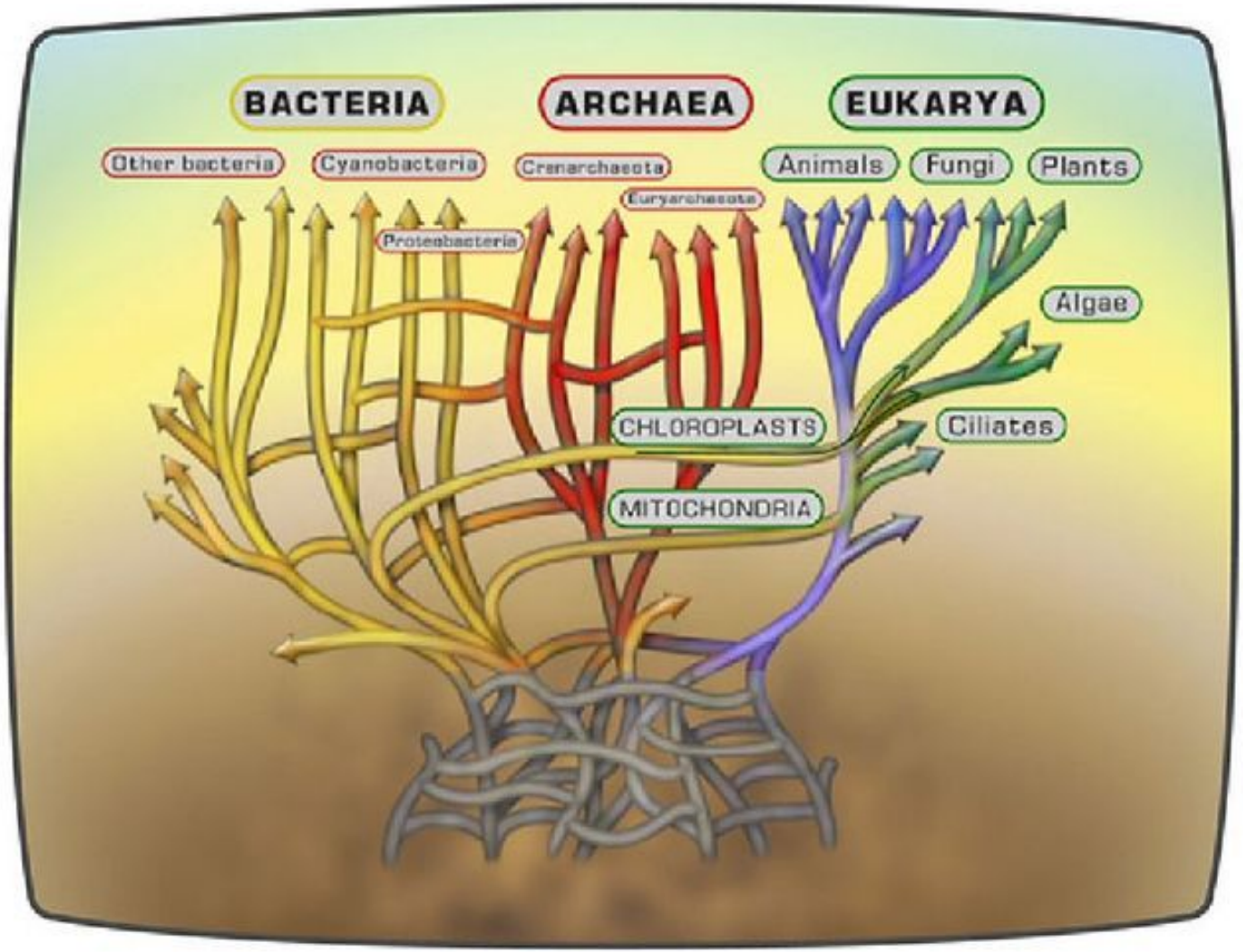
Horizontal/lateral transfer of DNA

Occurs *between* bacterial cells



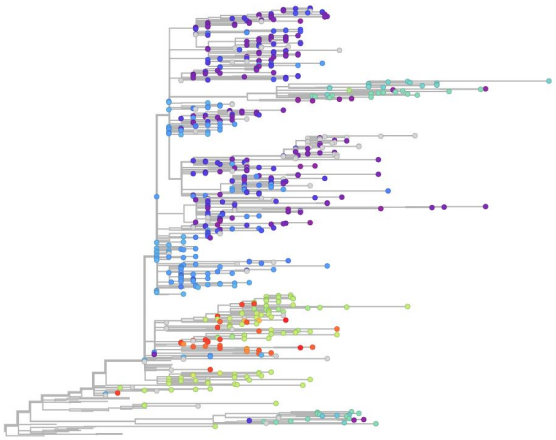
Conjugation and sometimes insertion into chromosome

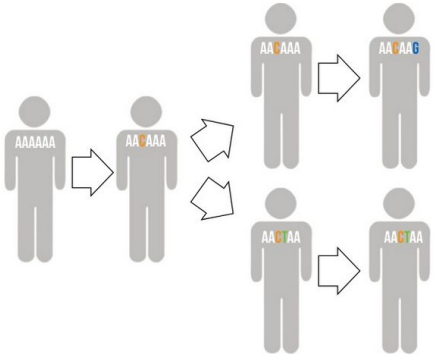
The web of life



Public health and clinical microbiology

Role of a public health laboratory network



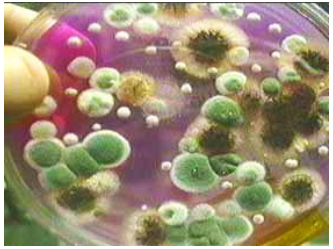


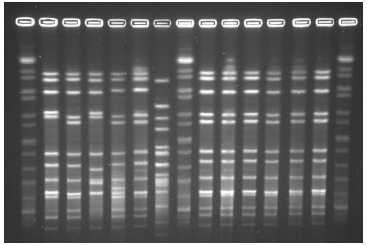
Diagnostics

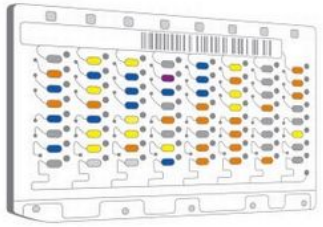
Surveillance

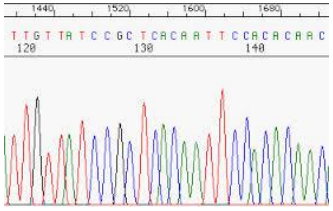
Outbreak response

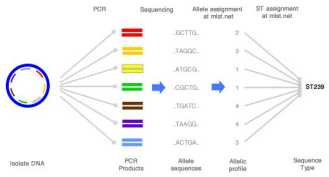
Traditional workflow

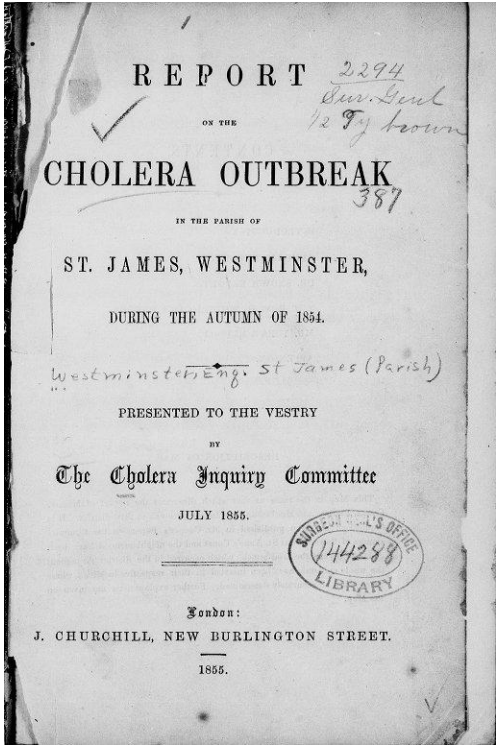






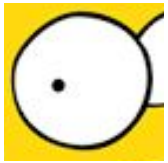






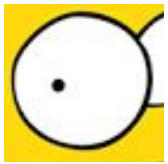
A bacterial isolate

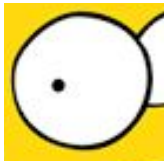
Focus on a small “informative” section

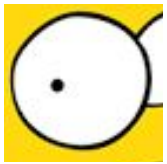


e.g. MLST, VNTR, PFGE, <insert genotyping method here>

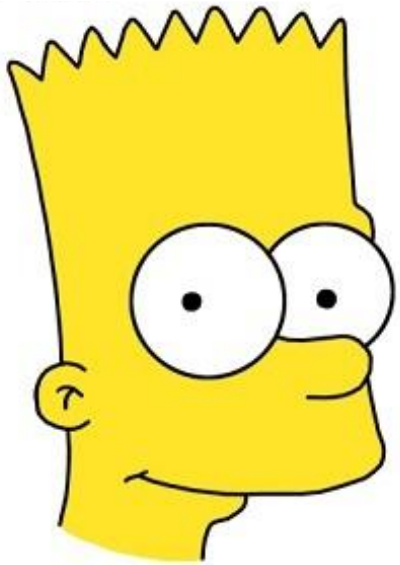
More samples - we have an outbreak!



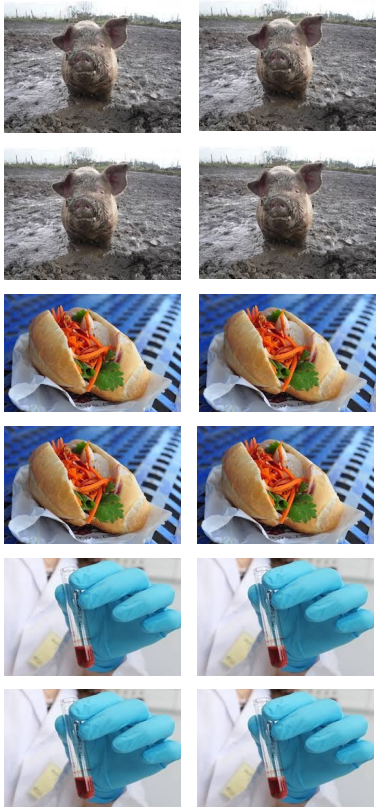


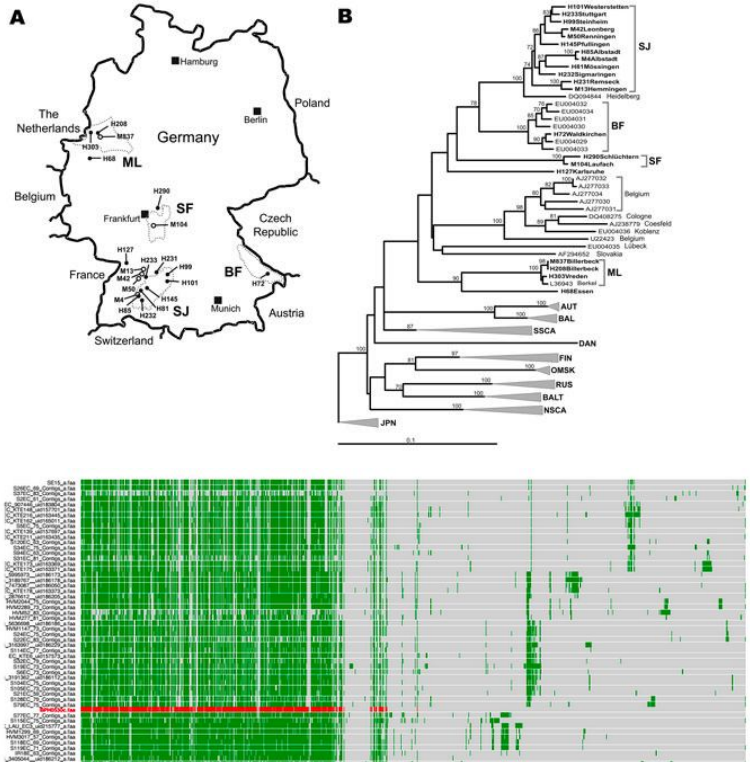


D’oh!



Modern workflow





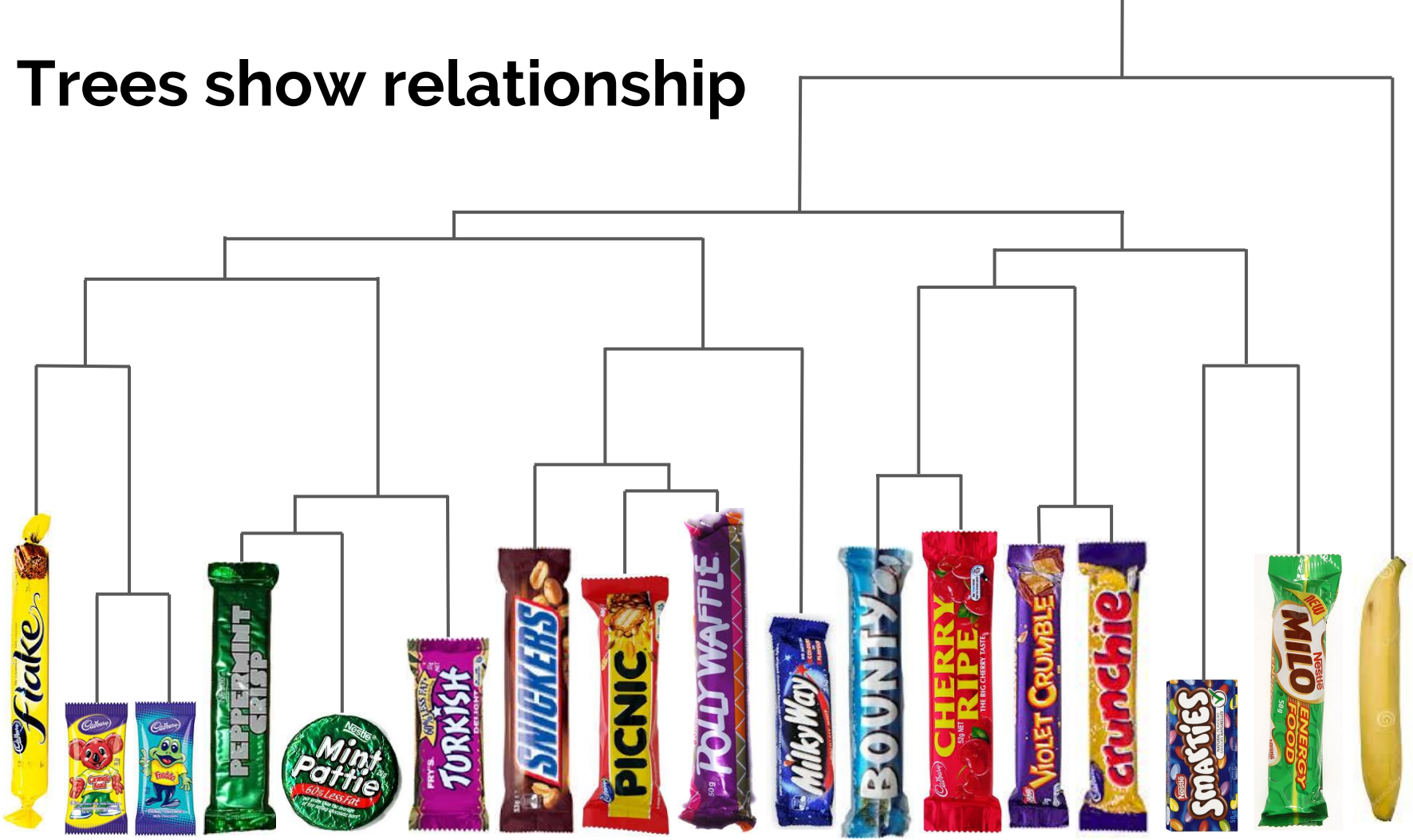
A win for genomics... and bioinformatics!

- Many investigations per week
- Dec 2015
  - *Salmonella* Anatum outbreak
  - bagged lettuce recall
  - cases nationally
- Milestone
  - First case definition to include genomics

“You don’t win friends with salad”

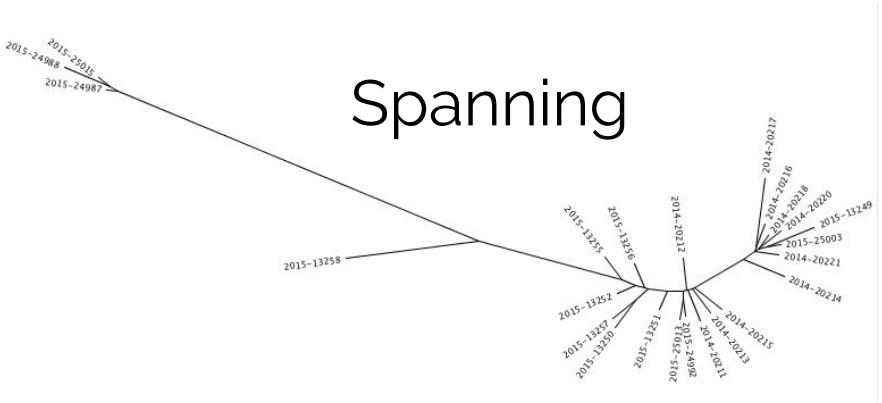
Phylogenomics

Trees show relationship

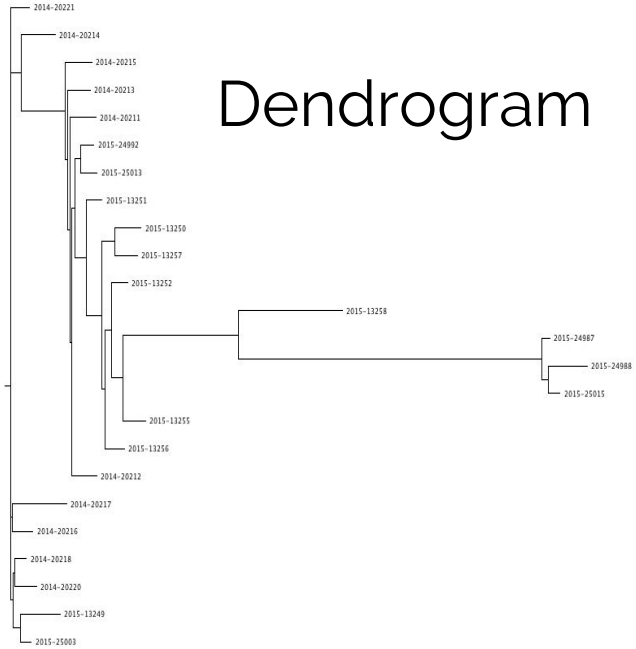


Same tree!

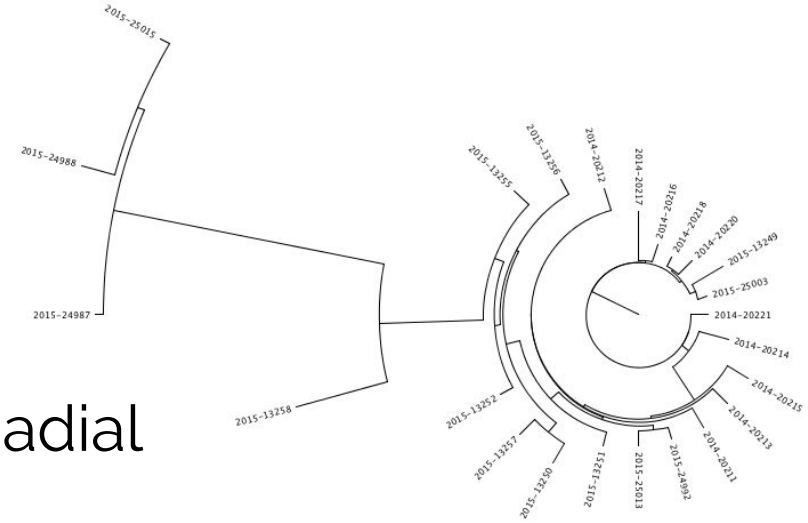
Spanning



Dendrogram



Radial

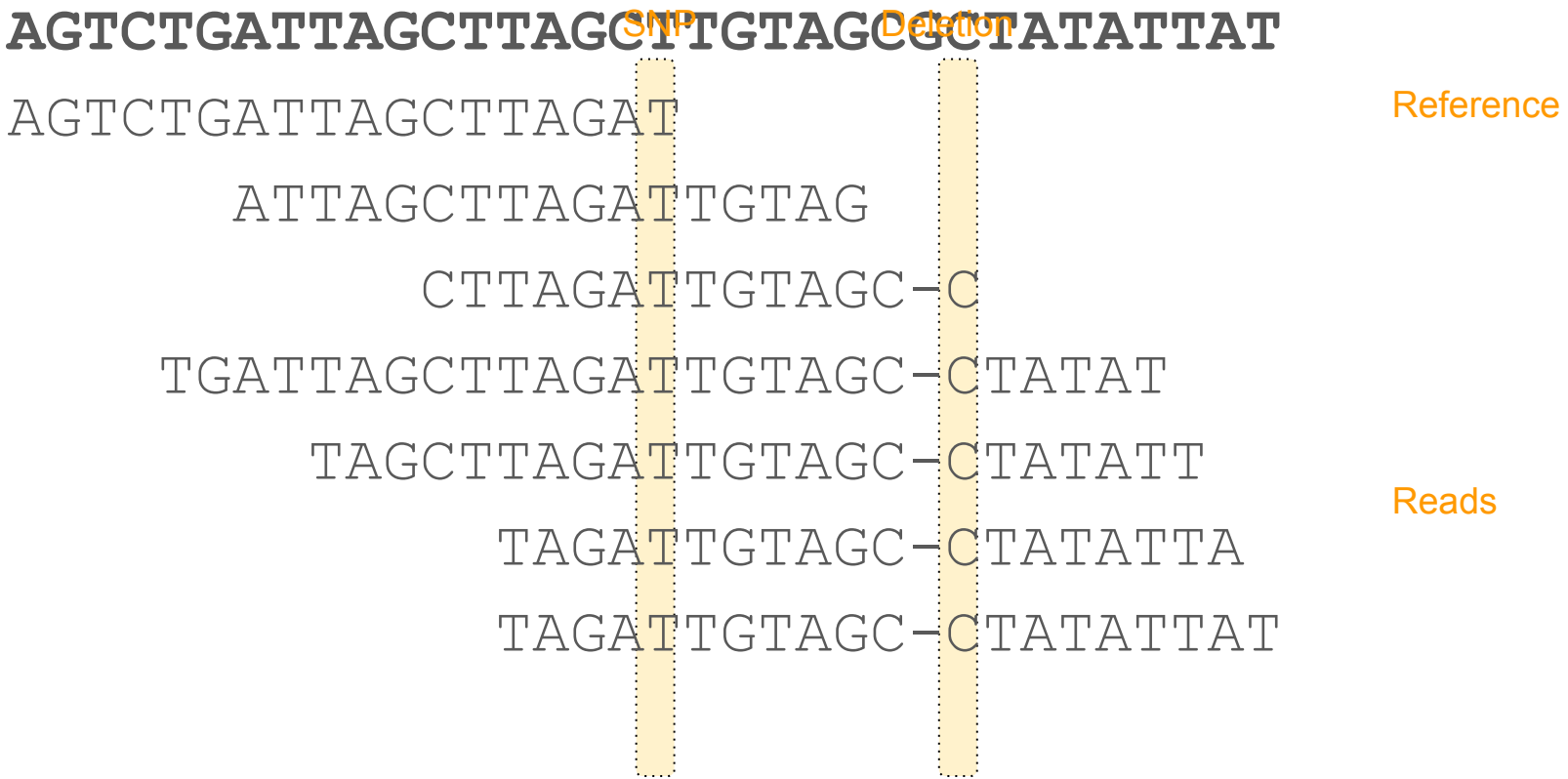


Every SNP is sacred

- Chocolate bar tree
  - branches were based on phenotypic attributes
  - size, colour, filling, texture, ingredients, flavour

- Genomic trees
  - want to use every part of the genome sequence
  - need to find all differences between isolates
  - show me the SNPs!

Finding differences



Collate reference alignments

```
bug1   GATTACCAGCATTAAGG-TTCTCCAATC
bug2   GAT---CTGCATTATGGATTCTCCATTC
bug3   G-TTACCAGCACTAA-------CCAGTC
```

The reference is a “middle man” to generate a “pseudo” whole genome alignment.

Core genome

```
bug1   GATTACCAGCATTAAGG-TTCTCCAATC
bug2   GAT---CTGCATTATGGATTCTCCATTC
bug3   G-TTACCAGCACTAA-------CCAGTC
core   | |   |||||||||       ||||||
```

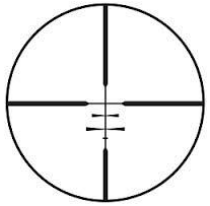
Core sites are present in **all** genomes.

Core SNPs

```
bug1   GATTACCAGCATTAAGG-TTCTCCAATC
bug2   GAT---CTGCATTATGGATTCTCCATTC
bug3   G-TTACCAGCACTAA-------CCAGTC
core   | |    |||||||||        ||||||
SNPs           |    |   |          |
```

Core SNPS = **polymorphic sites in core genome**

Allele sites

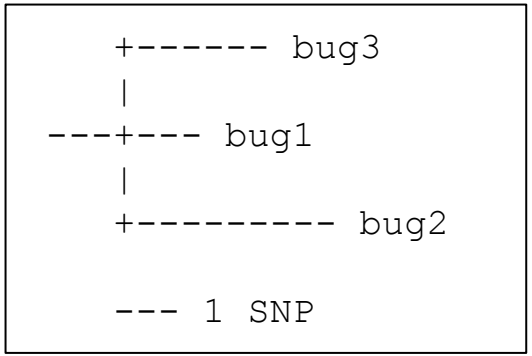


```
bug1   GATTACCAGCATTAAGG-TTCTCCAATC
bug2   GAT---CTGCATTATGGATTCRNCATTC
bug3   G-TTACCAGCACTAA-------CCAGTC
SNPs'         |   |   |          |
            ata ttc ata        atg
             1   2   3          4
```

Alignment ⇢ Distance matrix ⇢Tree

```
>bug1
ATAA
>bug2
TTTT
>bug3
ACAG
```

#SNPs	bug1	bug2	bug3
bug1	-	-	-
bug2	3	-	-
bug3	2	4	-

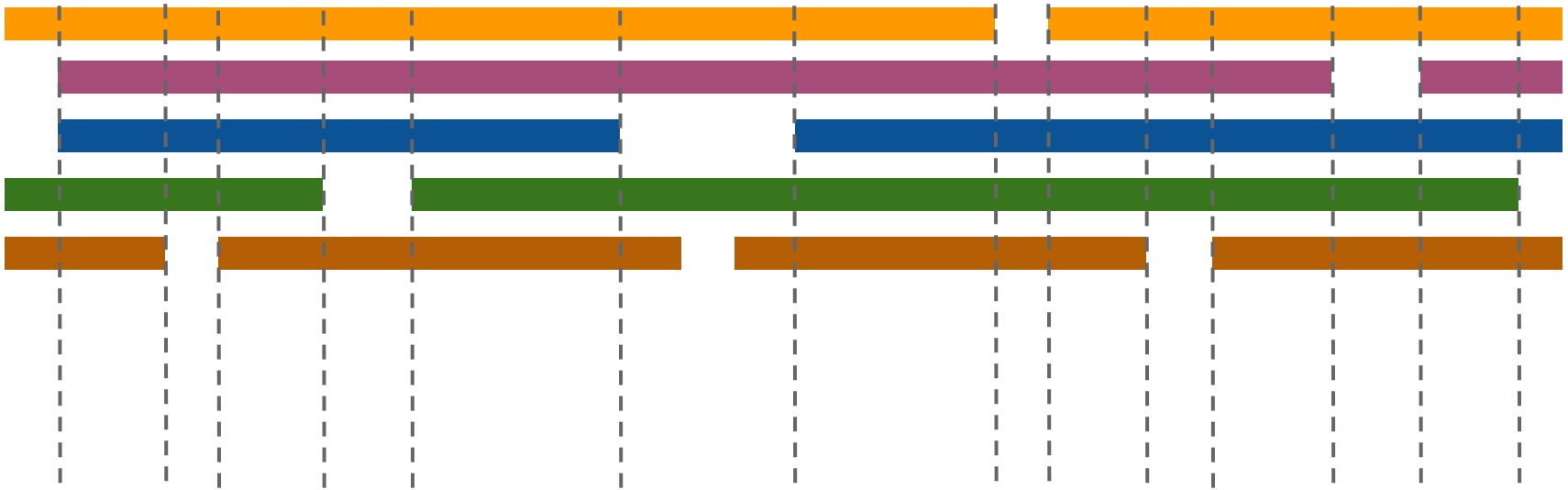


The pan genome

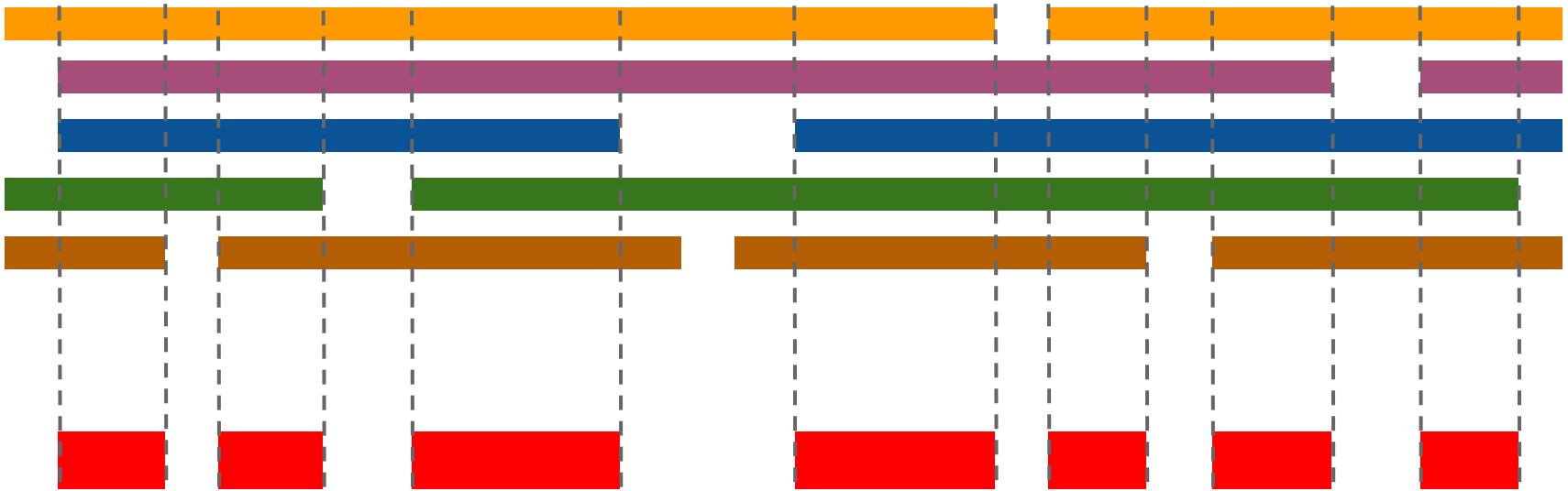
Five genomes

Whole genome multiple alignment

Find “common” segments

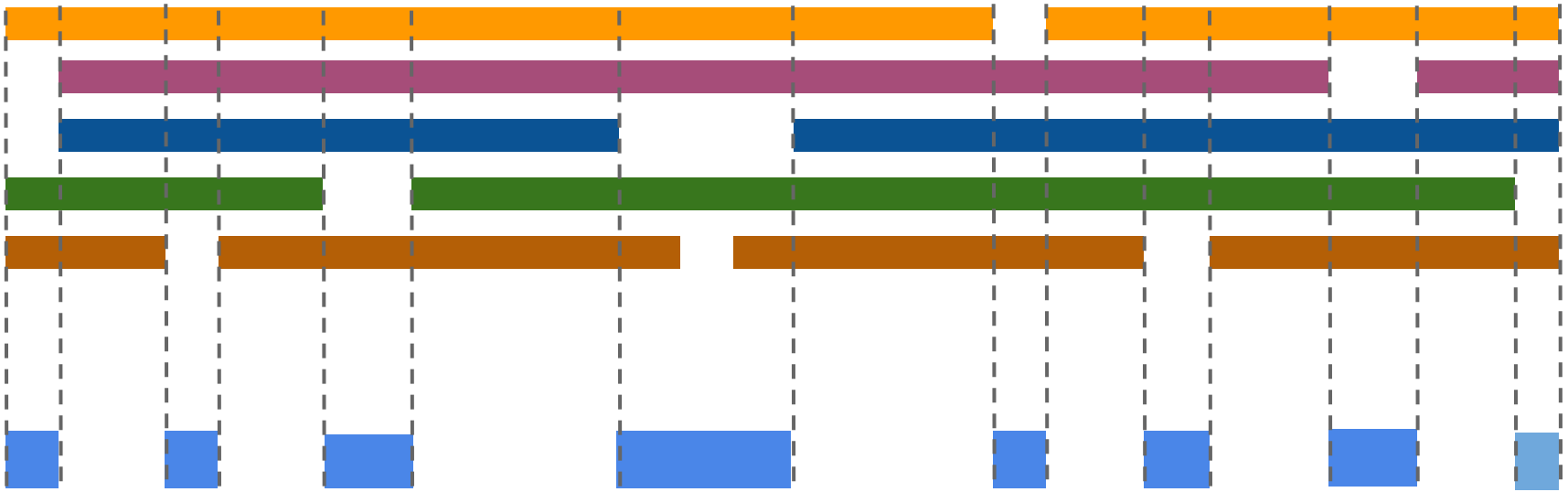


The core genome



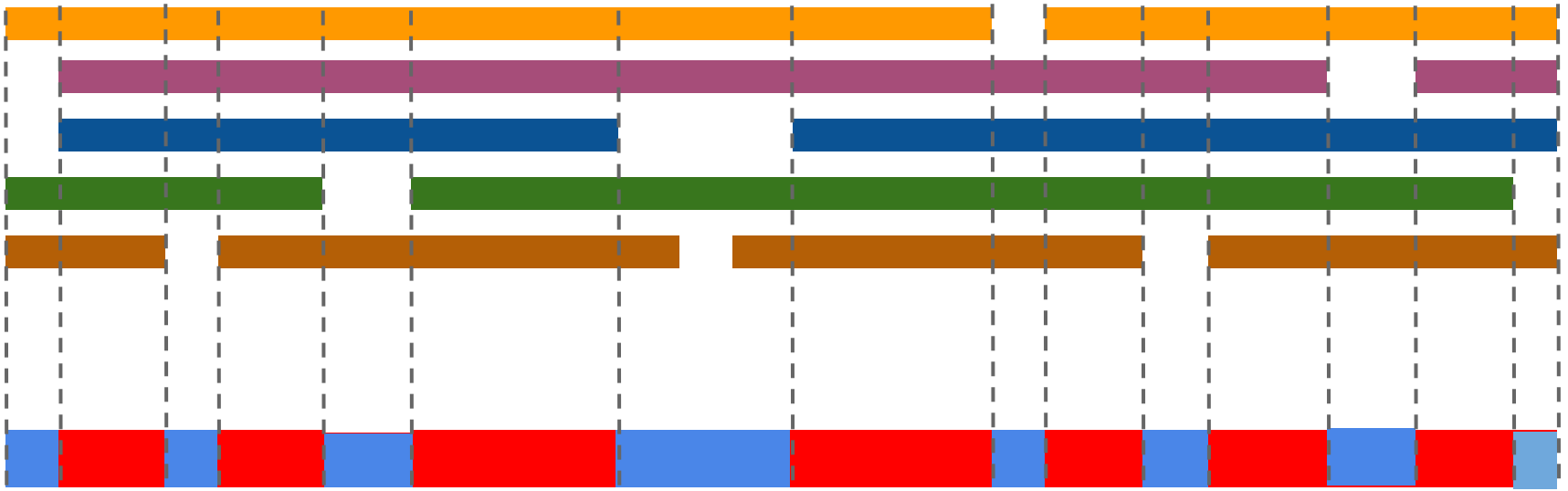
Core is common to all & has similar sequence.

The accessory genome



Accessory = not core (but still similar within)

The pan genome



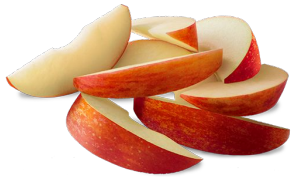
Pan = Core + Accessory

Core

- Common DNA
- Vertical evolution

- Critical genes
- Genotyping
- Phylogenetics

Accessory



- Novel DNA
- Lateral transfer

- Plasmids
- Mobile elements
- Phage

Determining the pan genome

Whole genome alignment is difficult !

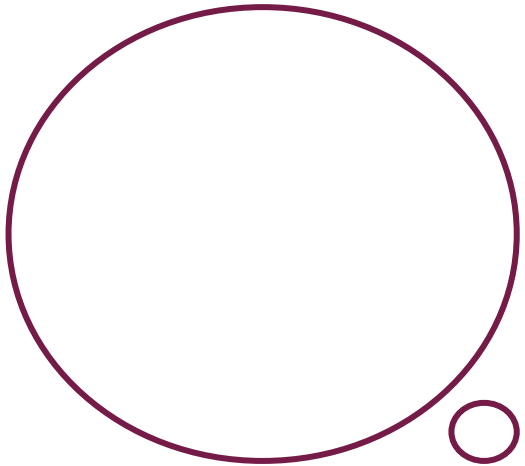
Rearrangements.
Sequence divergence.
Duplications.

Does not scale computationally.

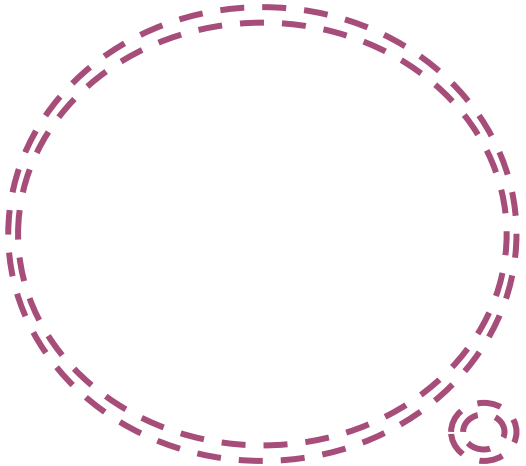
Genome or Gene-ome ?

- The DNA sequence of all replicons
  - Chromosomes, plasmids
- The set of “genes” in an organism
  - “Protein-ome”- just protein coding genes *e.g.* CDS
  - “Gene-ome” - also include non-coding genes *e.g.* RNAs

Genome vs Proteinome



5 Mbp genome



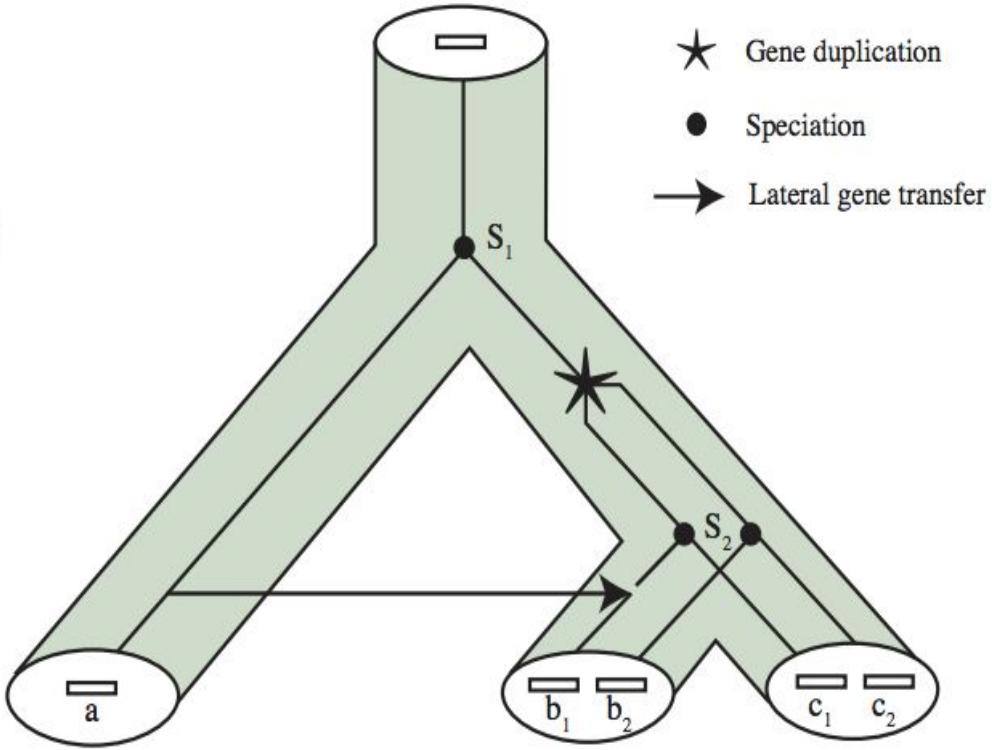
~5000 genes

Reframing the problem

Align whole genomes
(DNA)

Cluster homologous genes
(DNA or AA)

Homologs = common ancestor

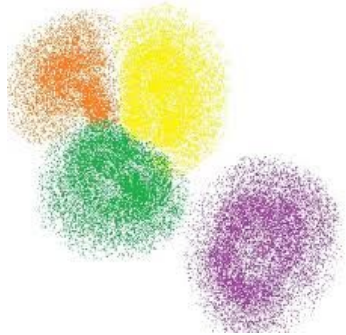


Ortholog
Speciation

Paralog
Duplication

Xenolog
Lateral transfer

Homolog clustering



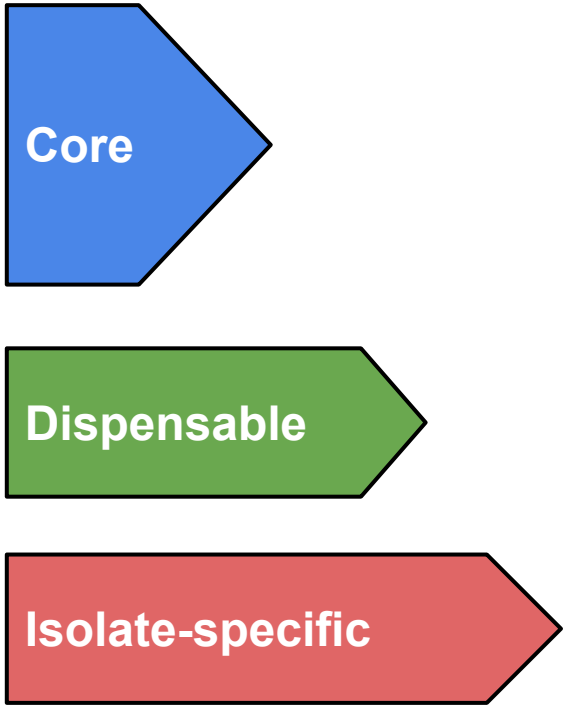
- Group homologous proteins together
  - exploit sequence similarity + synteny + operons
  - all versus all sequence comparison (not scalable)
    - DNA or amino acid (fast heuristics)
  - difficulty increases with taxa distance
- Depends on annotation quality
    - Missing genes
    - False genes

Typical workflow

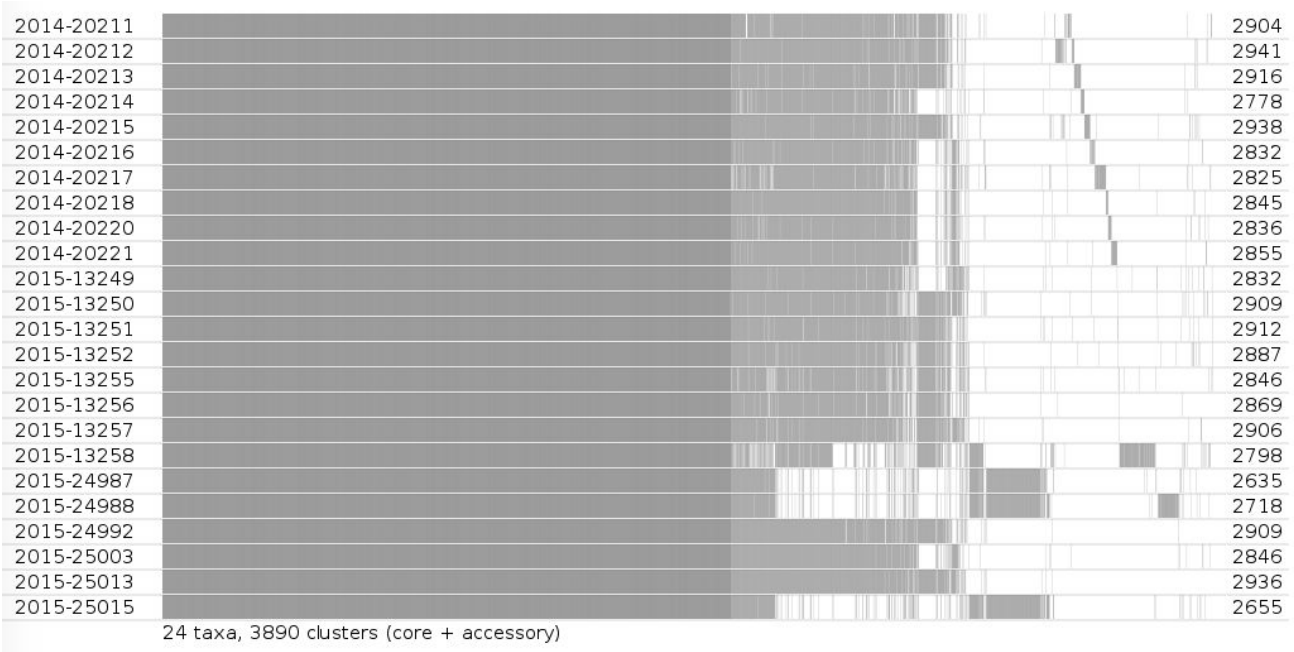
- *De novo* assembly - SPAdes
- Annotation - Prokka
- Pan-genome - Roary
- Visualise - Phandango

Roary → matrix / spreadsheet

CLUSTER	STRAIN1	STRAIN2	STRAIN3
00001	DNO1000	EHEC1000	MRSA_1000
00002	DNO1001	EHEC1002	MRSA_1001
00003	DNO1002	EHEC1003	MRSA_1002
00004	DNO1003	EHEC1004	MRSA_1003
00005	DNO1004	EHEC1005	MRSA_1022
:	:	:	:
02314	DNO1005	na	MRSA_1023
02315	DNO1451	EHEC3215	na
02316	na	EHEC3216	MRSA_1923
:	:	:	:
04197	DNO1456	na	na
04198	na	EHEC3877	na
04199	na	na	MRSA_0533

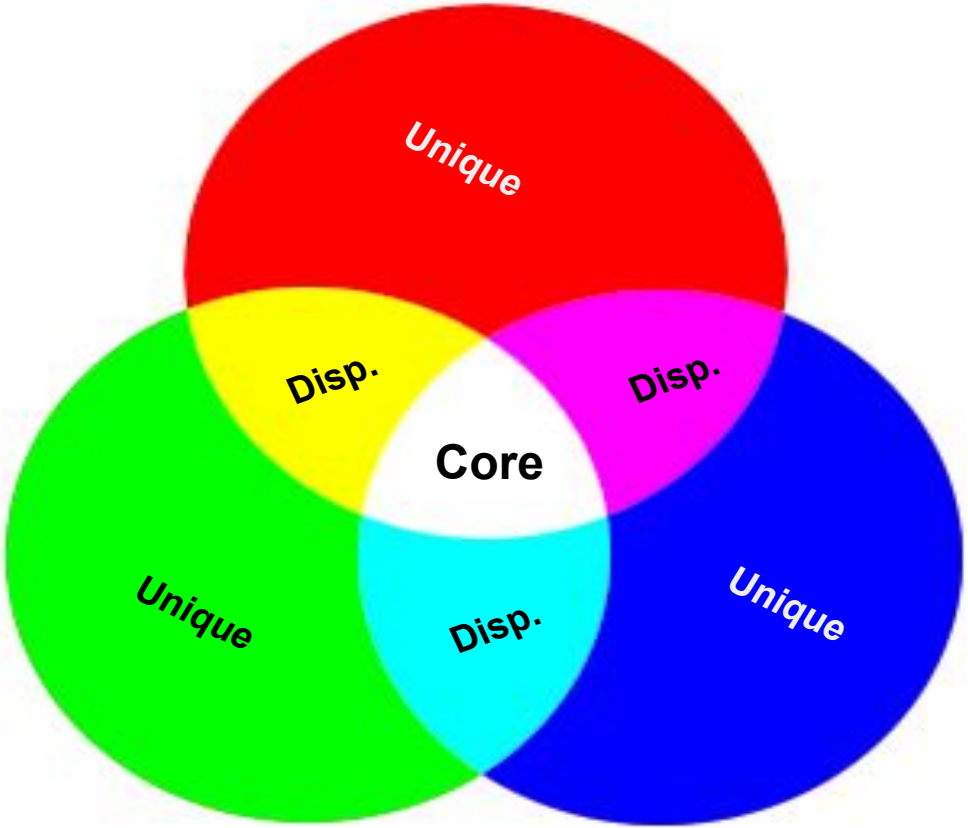


Example pan genome



Rows are genomes, columns are genes.

Three genomes (*N*=3)



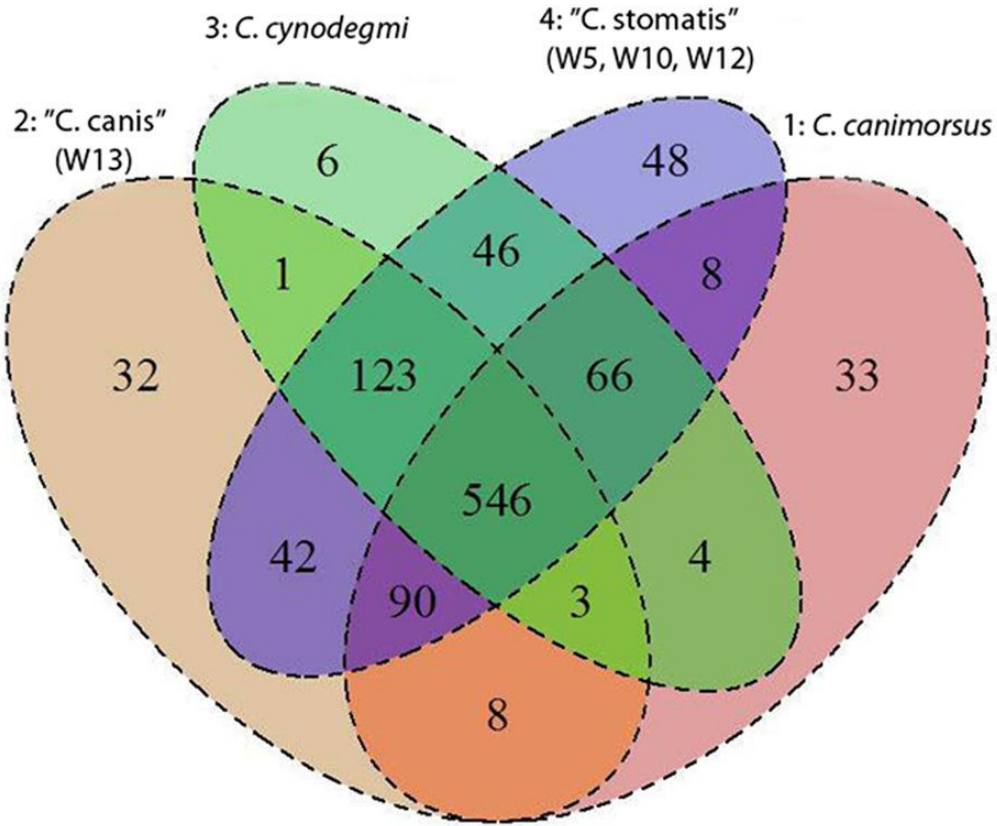
Core
In all 3 strains
(∈ *N* strains)

Dispensable
In 2 strains
(∈ [2,*N*-1] strains)

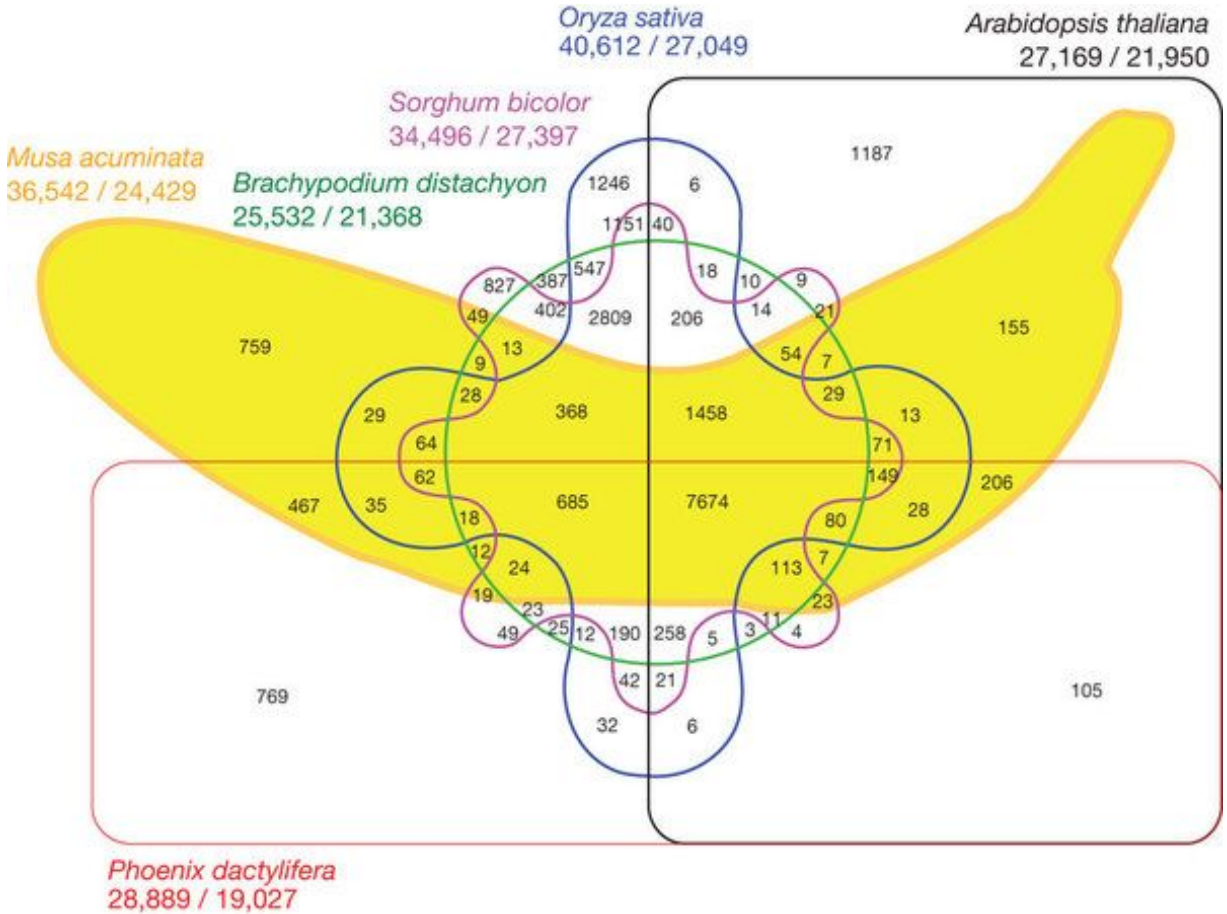
Unique
In only 1 strain
(∈ 1 strain)

Accessory

Flowery Venn (*N*=4)



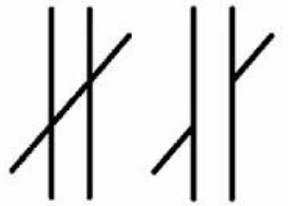
Venn will it end?

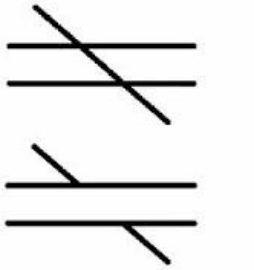


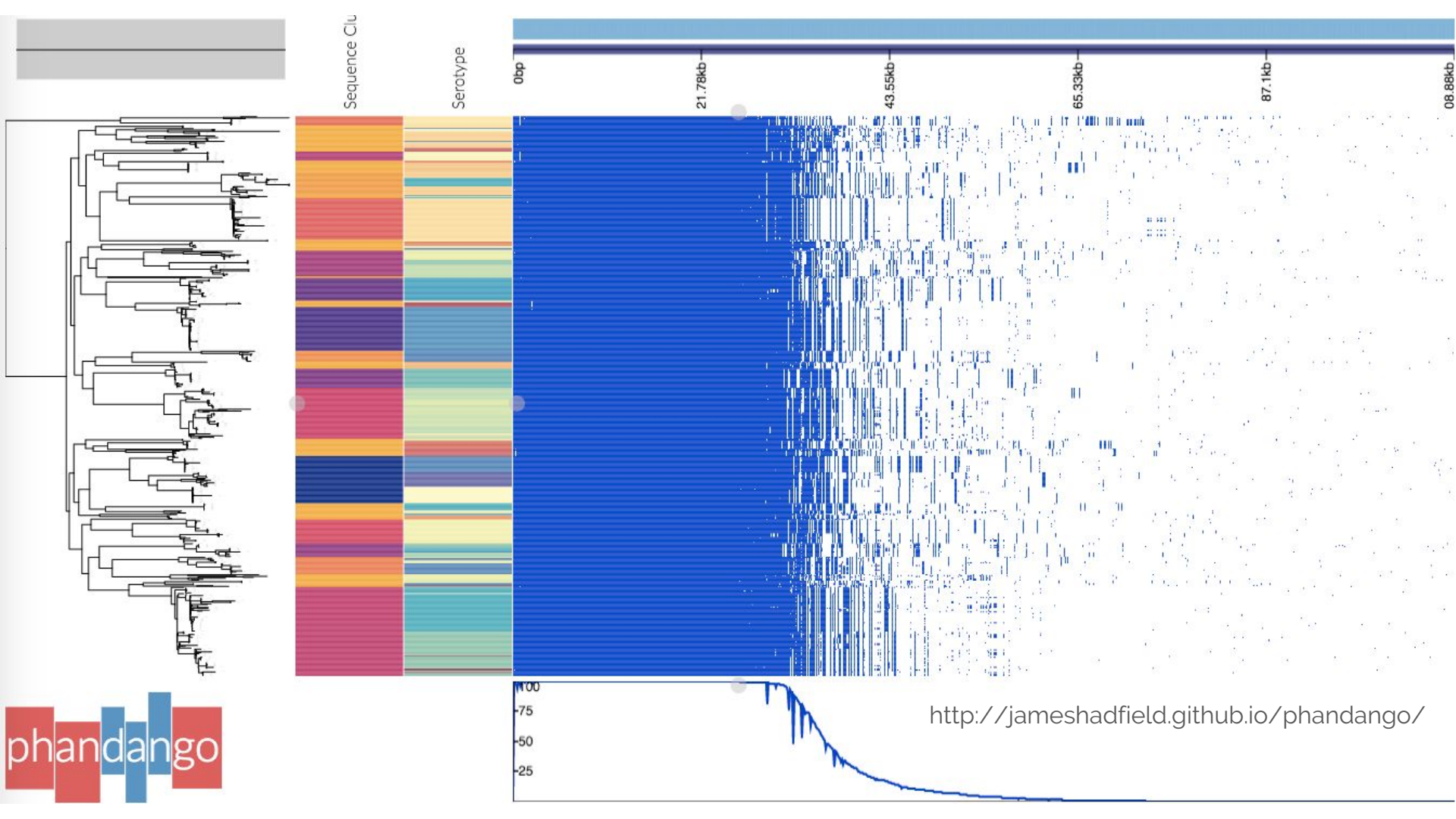
Bringing it together

Diagonal-omics

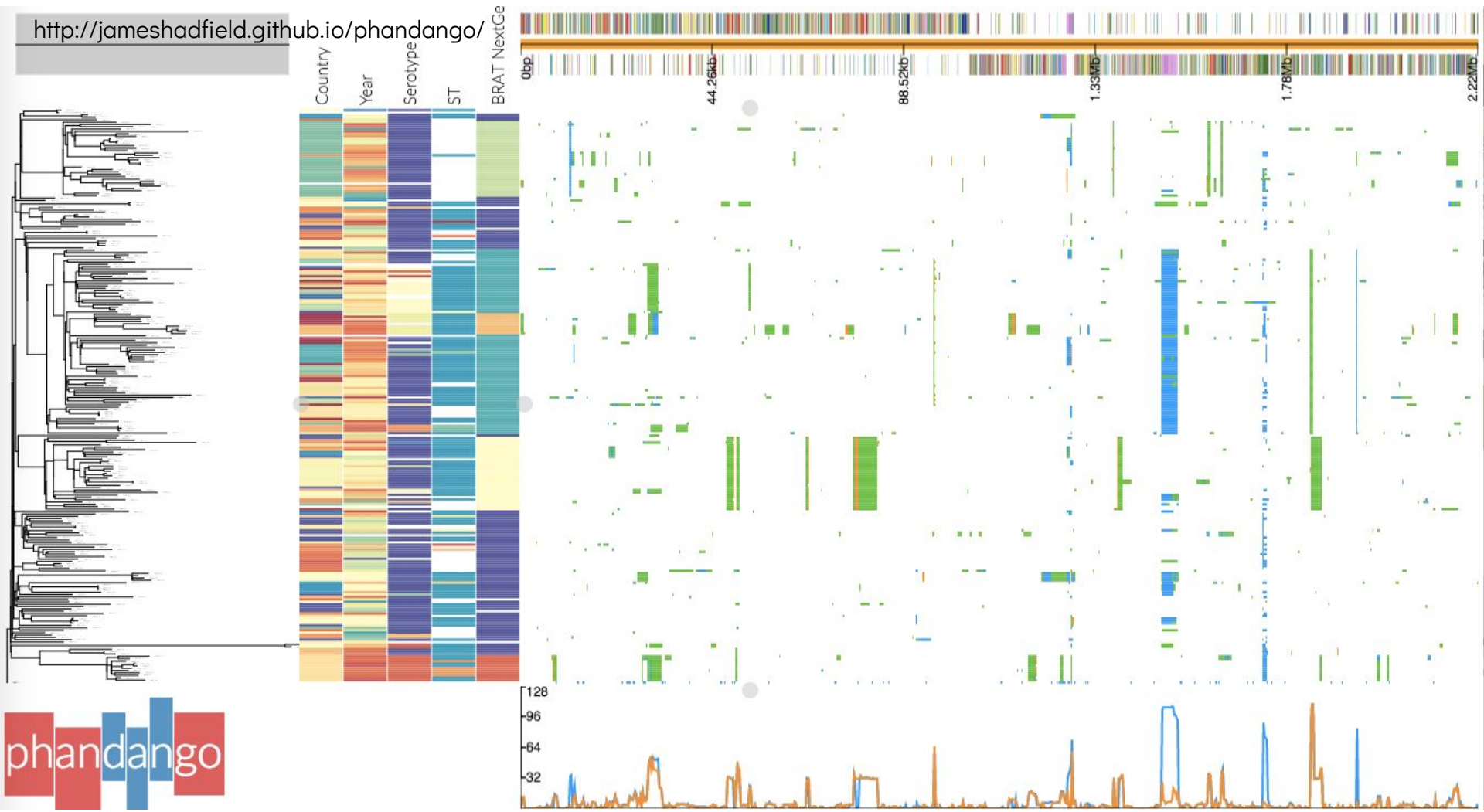
- Phylogenetics
  - Based on core SNPs
  - Vertical transmission
- Pan-genome
  - Looks at accessory genome
  - Horizontal transmission
- Combine for best of both worlds






Sequence Clu
Serotype
0bp
21.78kb
43.55kb
65.33kb
87.1kb
08.88kb
100
75
50
25
phandango
http://jameshadfield.github.io/phandango/

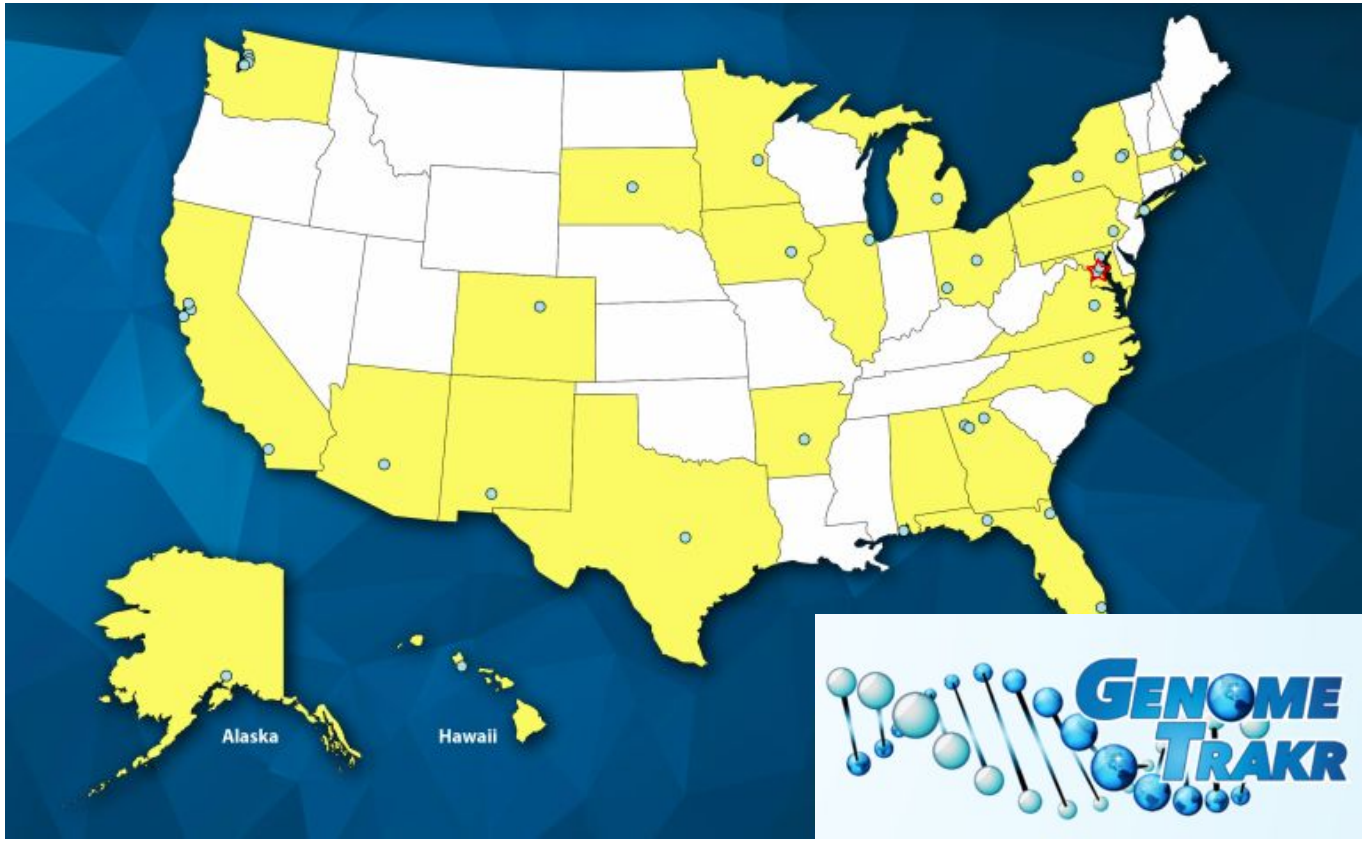
http://jameshadfield.github.io/phandango/



phandango

Data sharing

The GenomeTrakr network



US FDA (CFSAN)

+

NCBI

+

State reference labs

The GenomeTrakr network is international

Listeria monocytogenes

Accession: PRJNA317408 ID: 317408

GenomeTrakr project: Listeria monocytogenes, MDU PHL, Australia

Whole genome sequencing of Listeria monocytogenes isolates as part of MDU PHL routine national surveillance activities

Accession	PRJNA317408
Data Type	Raw sequence reads
Scope	Multiisolate
Organism	**Listeria monocytogenes** [Taxonomy ID: 1639] Bacteria; Firmicutes; Bacilli; Bacillales; Listeriaceae; Listeria; Listeria monocytogenes
Submission	Registration date: 5-Apr-2016 **Microbiological Diagnostics Unit**
Relevance	Medical

See Genome Information for Listeria monocytogenes

NAVIGATE UP

This project is a component of the Listeria monocytogenes

NAVIGATE ACROSS

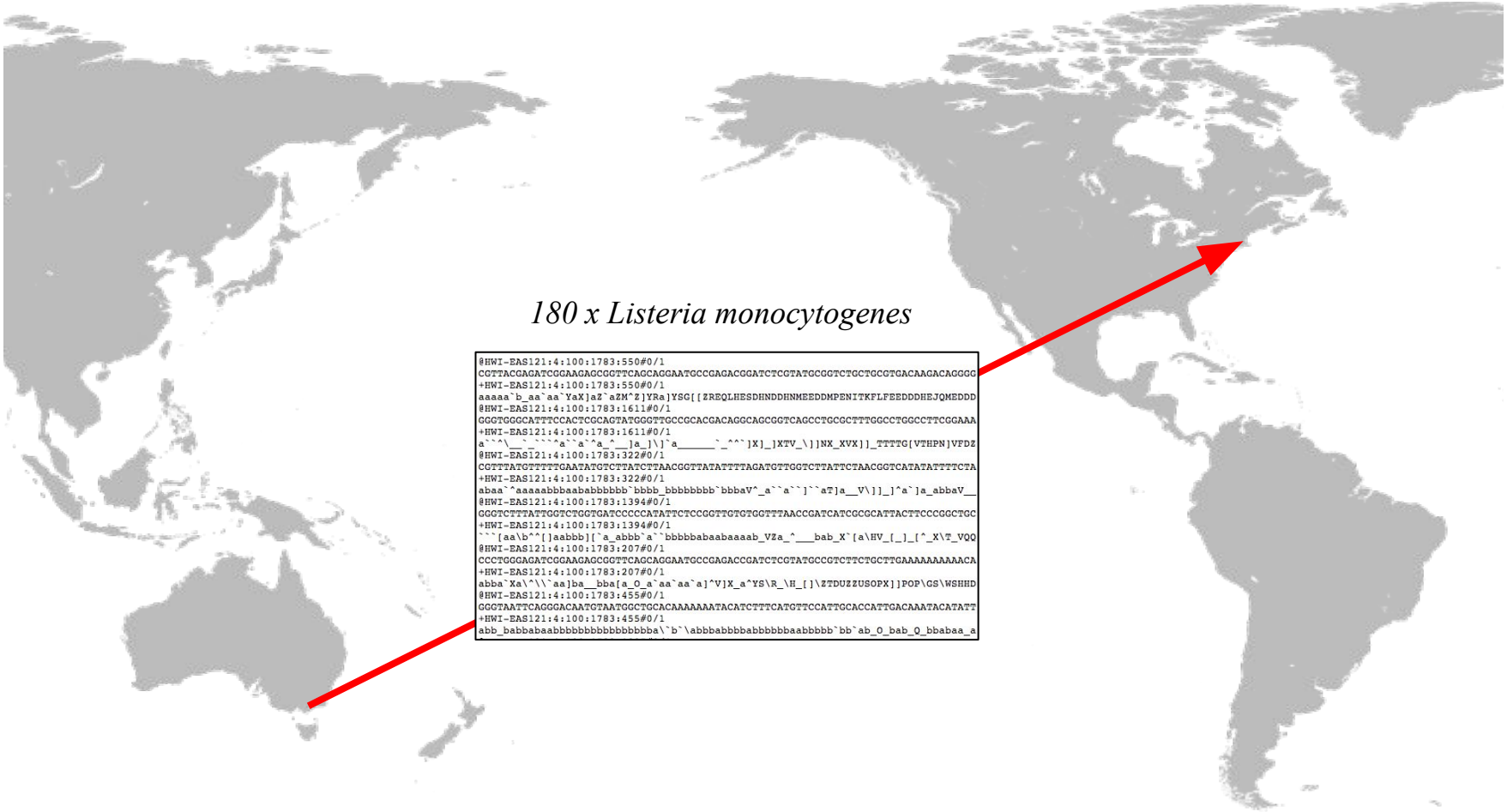
403 additional projects are related by organism.

33 additional projects are components of the Listeria monocytogenes.

Bill Klimke

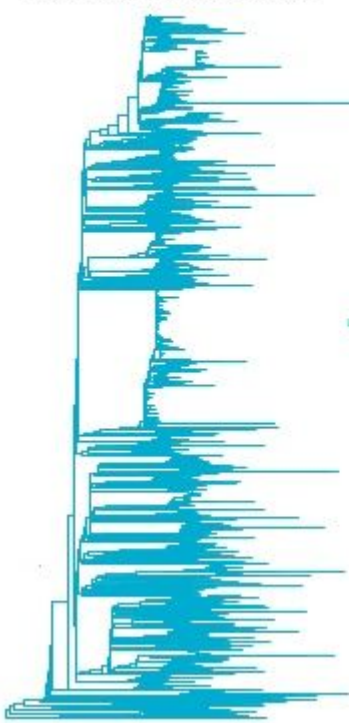
Project Data:

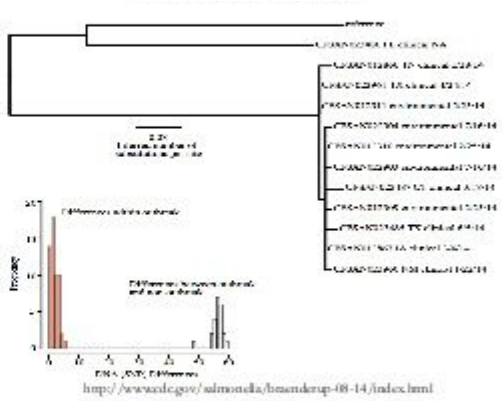
Resource Name	Number of Links
SEQUENCE DATA	
SRA Experiments	180
OTHER DATASETS	
BioSample	181

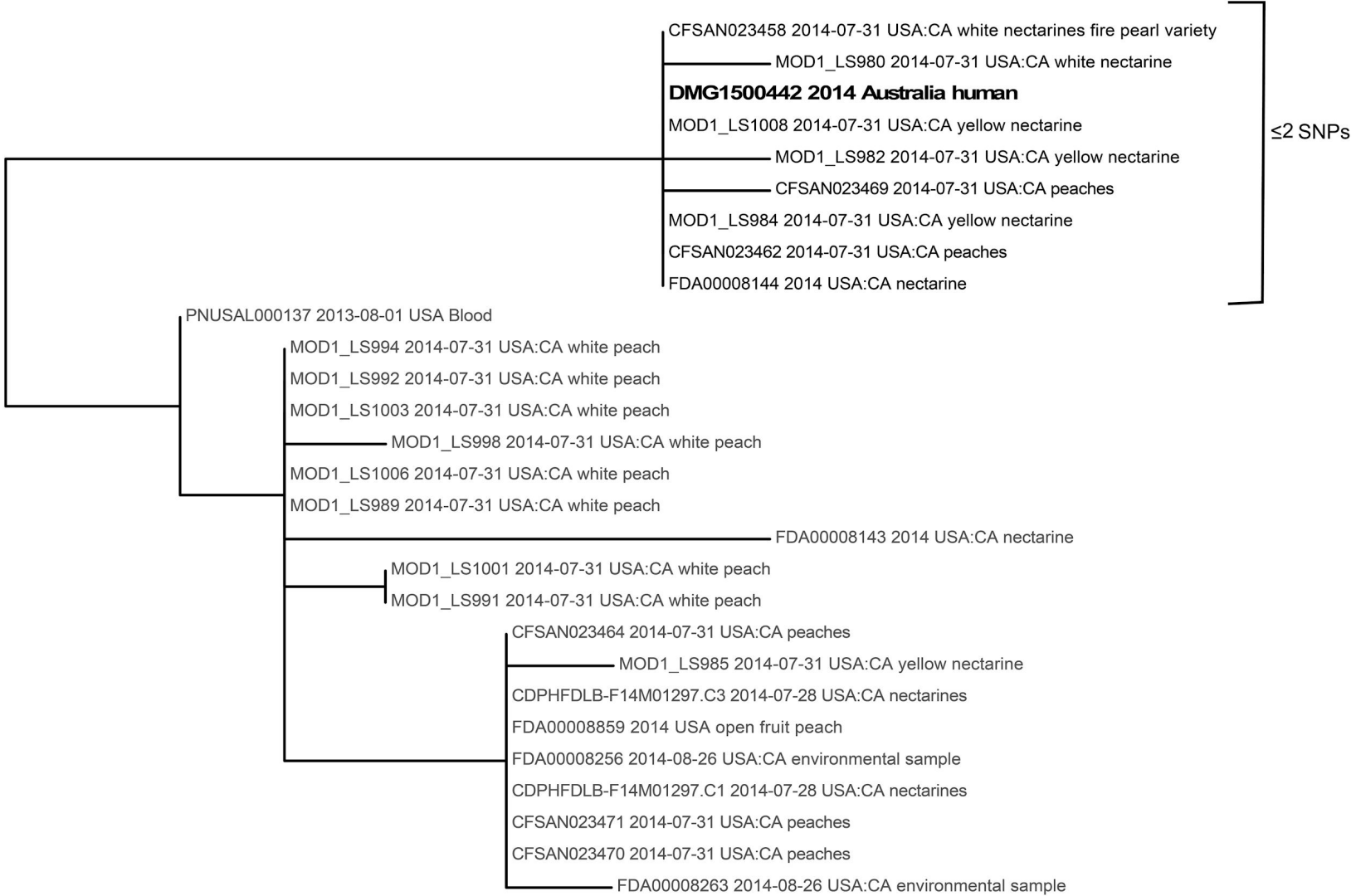

180 x Listeria monocytogenes
@HWI-EAS121:4:100:1783:550#0/1
CGTTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACGGATCTCGTATGCGGTCTGCTGCGTGACAAGACAGGGG
+HWI-EAS121:4:100:1783:550#0/1
aaaaa`b_aa`aa`YaX]aZ`aZM^Z]YRa]YSG[[ZREQLHESDHNDDHNMEEDDMPENITKFLFEEDDDHEJQMEDDD
@HWI-EAS121:4:100:1783:1611#0/1
GGGTGGGCATTTCCACTCGCAGTATGGGTTGCCGCACGACAGGCAGCGGTCAGCCTGCGCTTTGGCCTGGCCTTCGGAAA
+HWI-EAS121:4:100:1783:1611#0/1
a```^__`_```^a``a``^a_^__]a_]\]`a_____`_^^`]X]_]XTV_\]]NX_XVX]]_TTTTG[VTHPN]VFDZ
@HWI-EAS121:4:100:1783:322#0/1
CGTTTATGTTTTTGAATATGTCTTATCTTAACGGTTATATTTTAGATGTTGGTCTTATTCTAACGGTCATATATTTTCTA
+HWI-EAS121:4:100:1783:322#0/1
abaa``^aaaaabbbaabbbbbbb`bbbb_bbbbbbbb`bbbaV^_a``a``]``aT]a__V\]]_]^a`]a_abbaV__
@HWI-EAS121:4:100:1783:1394#0/1
GGGTCTTTATTGGTCTGGTGATCCCCCATATTCTCCGGTTGTGTGGTTTAACCGATCATCGCGCATTACTTCCCGGCTGC
+HWI-EAS121:4:100:1783:1394#0/1
```[aa\b^^[]aabbb][`a_abbb`a``bbbbbabaabaaaab_VZa_^___bab_X`[a\HV_[_]_[^_X\T_VQQ
@HWI-EAS121:4:100:1783:207#0/1
CCCTGGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAACA
+HWI-EAS121:4:100:1783:207#0/1
abba`Xa\^\\`aa]ba__bba[a_O_a`aa`aa`a]^V]X_a^YS\R_\H_[]\ZTDUZZUSOPX]]POP\GS\WSHHD
@HWI-EAS121:4:100:1783:455#0/1
GGGTAATTCAGGGACAATGTAATGGCTGCACAAAAAAATACATCTTTCATGTTCCATTGCACCATTGACAAATACATATT
+HWI-EAS121:4:100:1783:455#0/1
abb_babbabaabbbbbbbbbbbbbbbba\`b`\abbbabbbbbabbbbbbaabbbbb`bb`ab_O_bab_Q_bbabaa_a
```

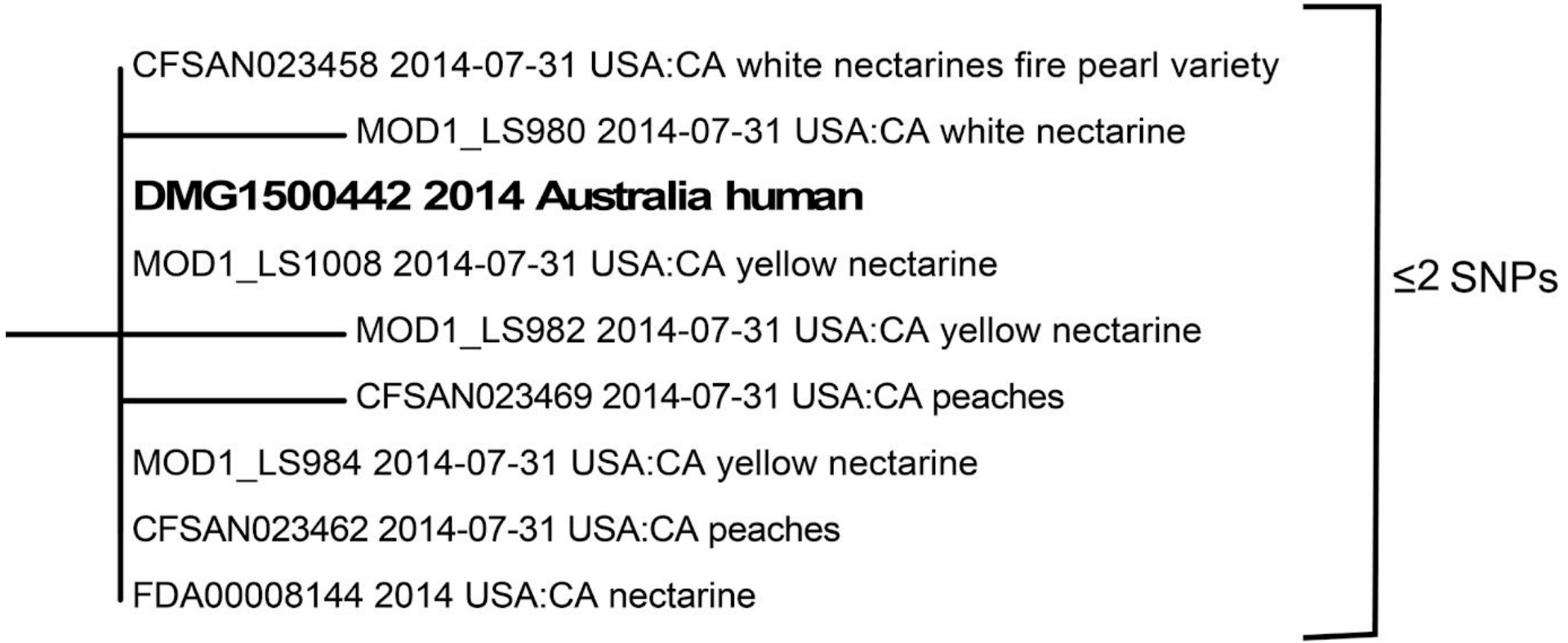
Nightly updates to find new matches

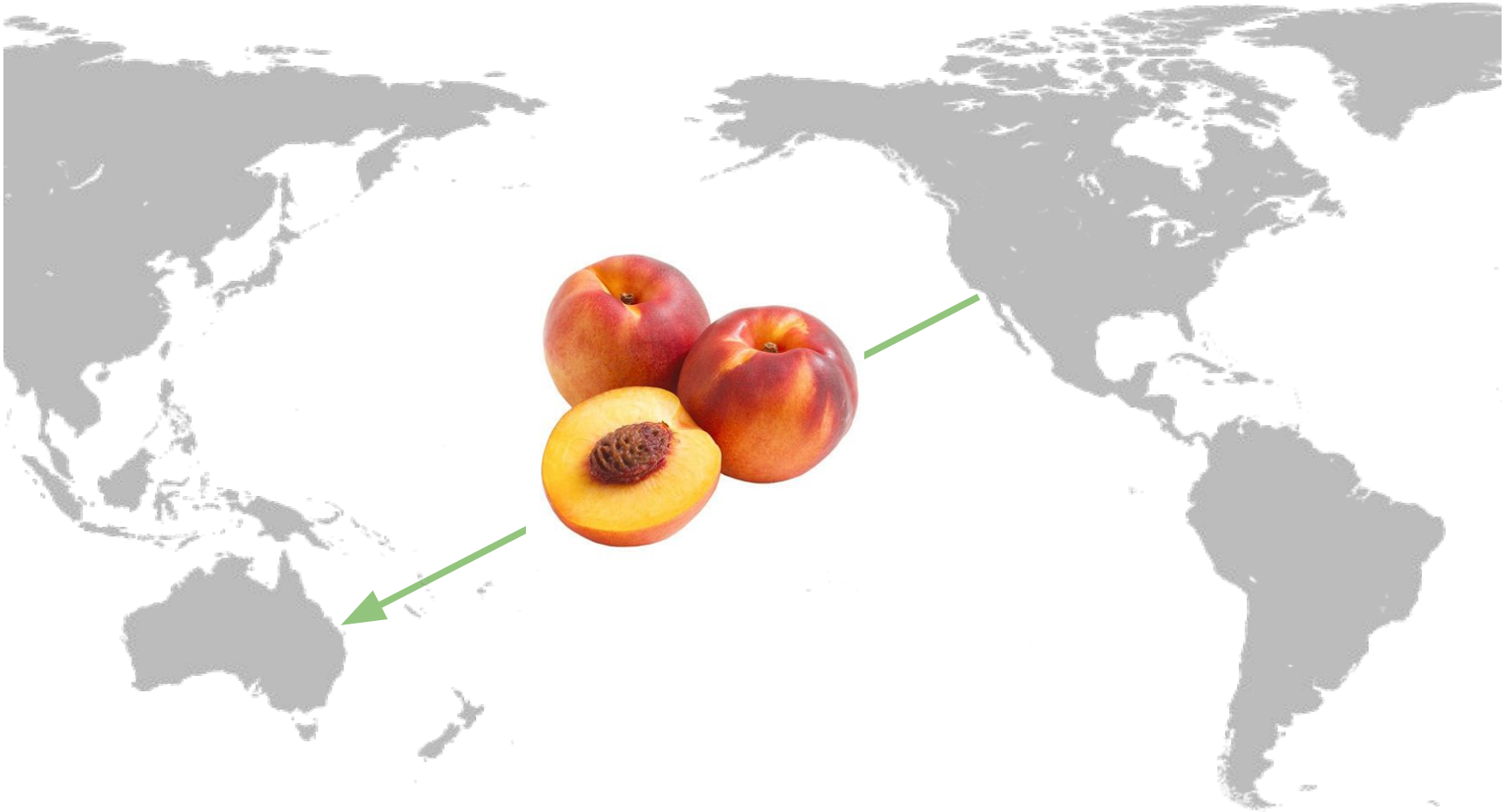
Errol Strain
(CFSAN)






CFSAN023458 2014-07-31 USA:CA white nectarines fire pearl variety
MOD1_LS980 2014-07-31 USA:CA white nectarine
DMG1500442 2014 Australia human
MOD1_LS1008 2014-07-31 USA:CA yellow nectarine
MOD1_LS982 2014-07-31 USA:CA yellow nectarine
CFSAN023469 2014-07-31 USA:CA peaches
MOD1_LS984 2014-07-31 USA:CA yellow nectarine
CFSAN023462 2014-07-31 USA:CA peaches
FDA00008144 2014 USA:CA nectarine
≤2 SNPs
PNUSAL000137 2013-08-01 USA Blood
MOD1_LS994 2014-07-31 USA:CA white peach
MOD1_LS992 2014-07-31 USA:CA white peach
MOD1_LS1003 2014-07-31 USA:CA white peach
MOD1_LS998 2014-07-31 USA:CA white peach
MOD1_LS1006 2014-07-31 USA:CA white peach
MOD1_LS989 2014-07-31 USA:CA white peach
FDA00008143 2014 USA:CA nectarine
MOD1_LS1001 2014-07-31 USA:CA white peach
MOD1_LS991 2014-07-31 USA:CA white peach
CFSAN023464 2014-07-31 USA:CA peaches
MOD1_LS985 2014-07-31 USA:CA yellow nectarine
CDPHFDLB-F14M01297.C3 2014-07-28 USA:CA nectarines
FDA00008859 2014 USA open fruit peach
FDA00008256 2014-08-26 USA:CA environmental sample
CDPHFDLB-F14M01297.C1 2014-07-28 USA:CA nectarines
CFSAN023471 2014-07-31 USA:CA peaches
CFSAN023470 2014-07-31 USA:CA peaches
FDA00008263 2014-08-26 USA:CA environmental sample


CFSAN023458 2014-07-31 USA:CA white nectarines fire pearl variety
MOD1_LS980 2014-07-31 USA:CA white nectarine
DMG1500442 2014 Australia human
MOD1_LS1008 2014-07-31 USA:CA yellow nectarine
MOD1_LS982 2014-07-31 USA:CA yellow nectarine
CFSAN023469 2014-07-31 USA:CA peaches
MOD1_LS984 2014-07-31 USA:CA yellow nectarine
CFSAN023462 2014-07-31 USA:CA peaches
FDA00008144 2014 USA:CA nectarine
≤2 SNPs




CASE
CLOSED

Clinical Infectious Diseases
IDSA Infectious Diseases Society of America
hivma hiv medicine association

Article Navigation

Sharing Is Caring: International Sharing of Data Enhances Genomic Surveillance of *Listeria monocytogenes* FREE

Jason C. Kwong ✉, Russell Stafford, Errol Strain, Timothy P. Stinear, Torsten Seemann, Benjamin P. Howden

Clin Infect Dis (2016) 63 (6): 846-848. **DOI:** https://doi.org/10.1093/cid/ciw359
Published: 09 June 2016

Clinical metagenomics

Infectious disease management

- Integrate genomics into patient care
- Identify pathogen(s)
  - Polymicrobial infections
- Determine antibiotic resistance profile
  - Acquired genes
  - Point mutations
- Determine hospital transmission / sources

Diagnosing the undiagnosable

The NEW ENGLAND JOURNAL of MEDICINE

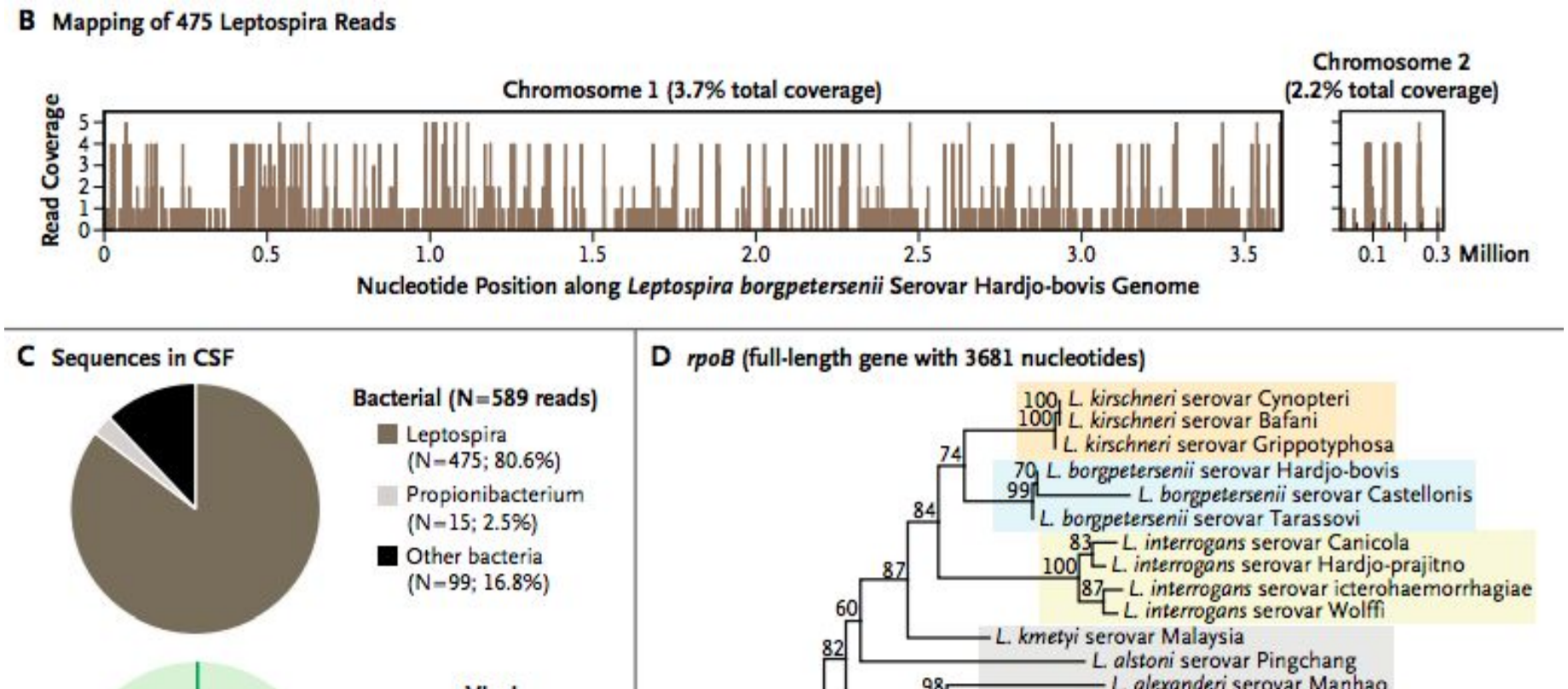
BRIEF REPORT

Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing

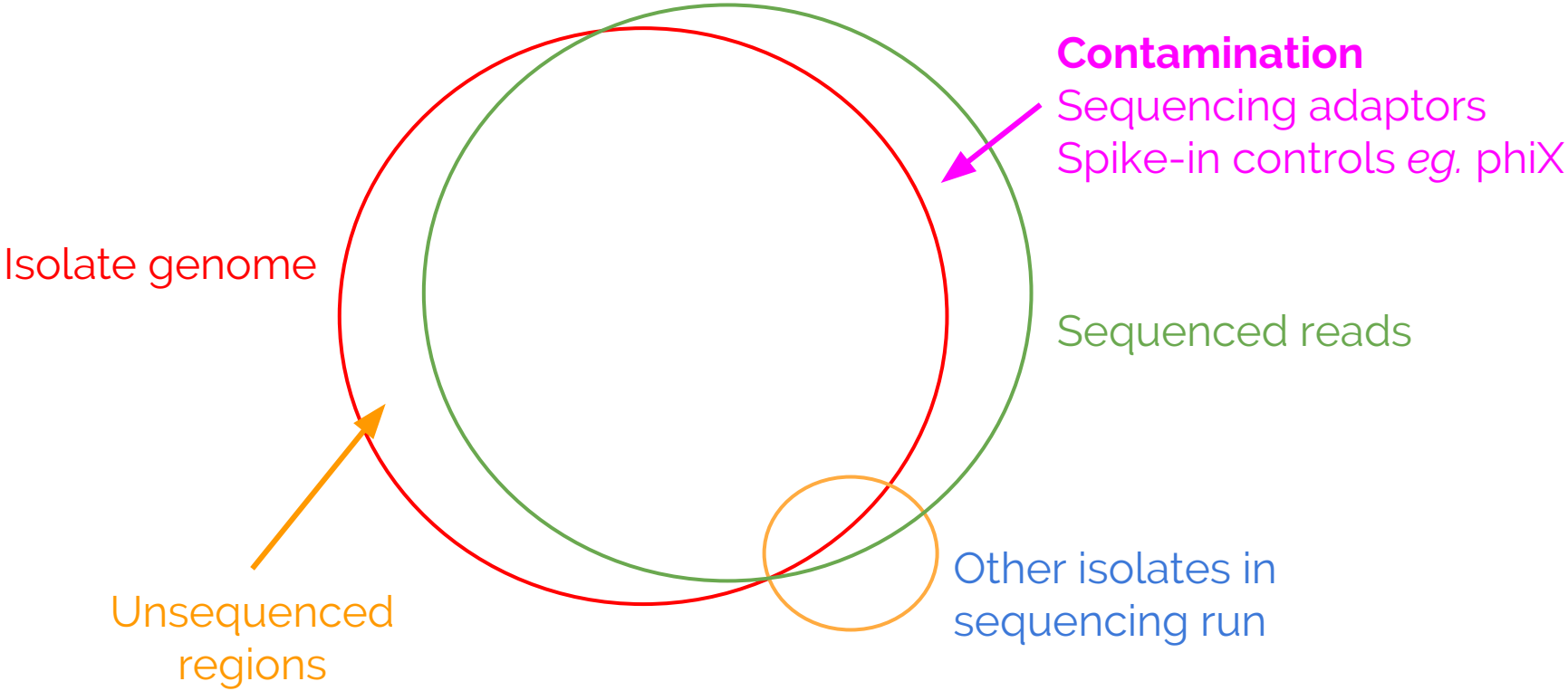
Michael R. Wilson, M.D., Samia N. Naccache, Ph.D., Erik Samayoa, B.S., C.L.S., Mark Biagtan, M.D., Hiba Bashir, M.D., Guixia Yu, B.S., Shahriar M. Salamat, M.D., Ph.D., Sneha Somasekar, B.S., Scot Federman, B.A., Steve Miller, M.D., Ph.D., Robert Sokolic, M.D., Elizabeth Garabedian, R.N., M.S.L.S., Fabio Candotti, M.D., Rebecca H. Buckley, M.D., Kurt D. Reed, M.D., Teresa L. Meyer, R.N., M.S., Christine M. Seroogy, M.D., Renee Galloway, M.P.H., Sheryl L. Henderson, M.D., Ph.D., James E. Gern, M.D., Joseph L. DeRisi, Ph.D., and Charles Y. Chiu, M.D., Ph.D.

This article was published on June 4, 2014, at NEJM.org.

N Engl J Med 2014;370:2408-17.
DOI: 10.1056/NEJMoa1401268


B Mapping of 475 Leptospira Reads
Chromosome 1 (3.7% total coverage)
Chromosome 2
(2.2% total coverage)
Read Coverage
0
1
2
3
4
5
0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.1
0.3 Million
Nucleotide Position along Leptospira borgpetersenii Serovar Hardjo-bovis Genome
C Sequences in CSF
Bacterial (N=589 reads)
Leptospira
(N=475; 80.6%)
Propionibacterium
(N=15; 2.5%)
Other bacteria
(N=99; 16.8%)
Viral
(N=107,016 reads)
Anelloviridae
(N=106,988; 99.97%)
Bacteriophage
(N=28; 0.03%)
D rpoB (full-length gene with 3681 nucleotides)
100 L. kirschneri serovar Cynopteri
100 L. kirschneri serovar Bafani
L. kirschneri serovar Grippotyphosa
74
70 L. borgpetersenii serovar Hardjo-bovis
99 L. borgpetersenii serovar Castellonis
84 L. borgpetersenii serovar Tarassovi
83 L. interrogans serovar Canicola
100 L. interrogans serovar Hardjo-prajitno
87 L. interrogans serovar icterohaemorrhagiae
87 L. interrogans serovar Wolffi
60
L. kmetyi serovar Malaysia
82
L. alstoni serovar Pingchang
98 L. alexanderi serovar Manhao
L. weilii serovar Celledoni
97 L. noguchii strain Caco
100
61 L. noguchii serovar Nicaragua
L. noguchii serovar Panama
L. santarosai strain UW
L. santarosai serovar Shermani
0.04

What data do we really have?



The problem with reference sequences

- Good
  - Biased to pathogens

- Bad
  - Only a fraction of true diversity
  - Protists, fungi poorly represented
  - Contamination
  - Wrong taxonomic assignment

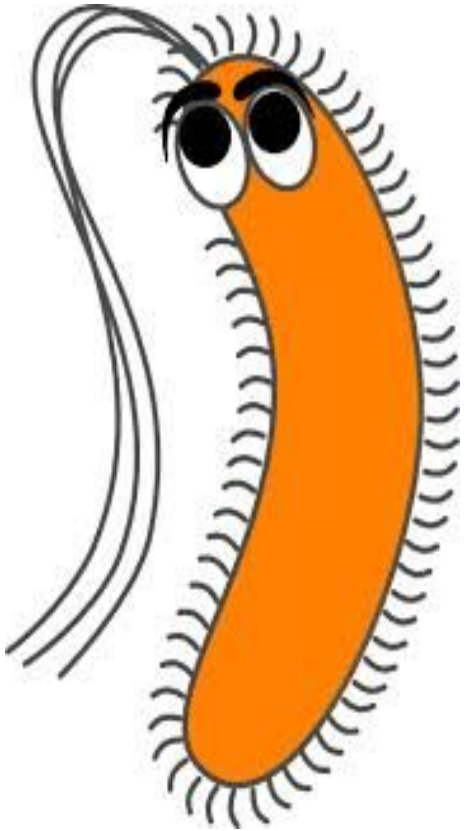
Conclusions

Summary

- Bacteria are cool
- Data wise, they are smaller,

  but we have more of them to deal with
- Small core, huge accessory genome
- Genome wide, SNP resolution has

  transformed public health microbiology
- Data sharing is essential to global health



Acknowledgements

Titus Brown

Lisa Johnson

Amanda Charbonneau

Morgan Price

Karen Word

Erich Schwarz

The end.