How genomics and bioinformatics is transforming clinical and public health microbiology in Australia

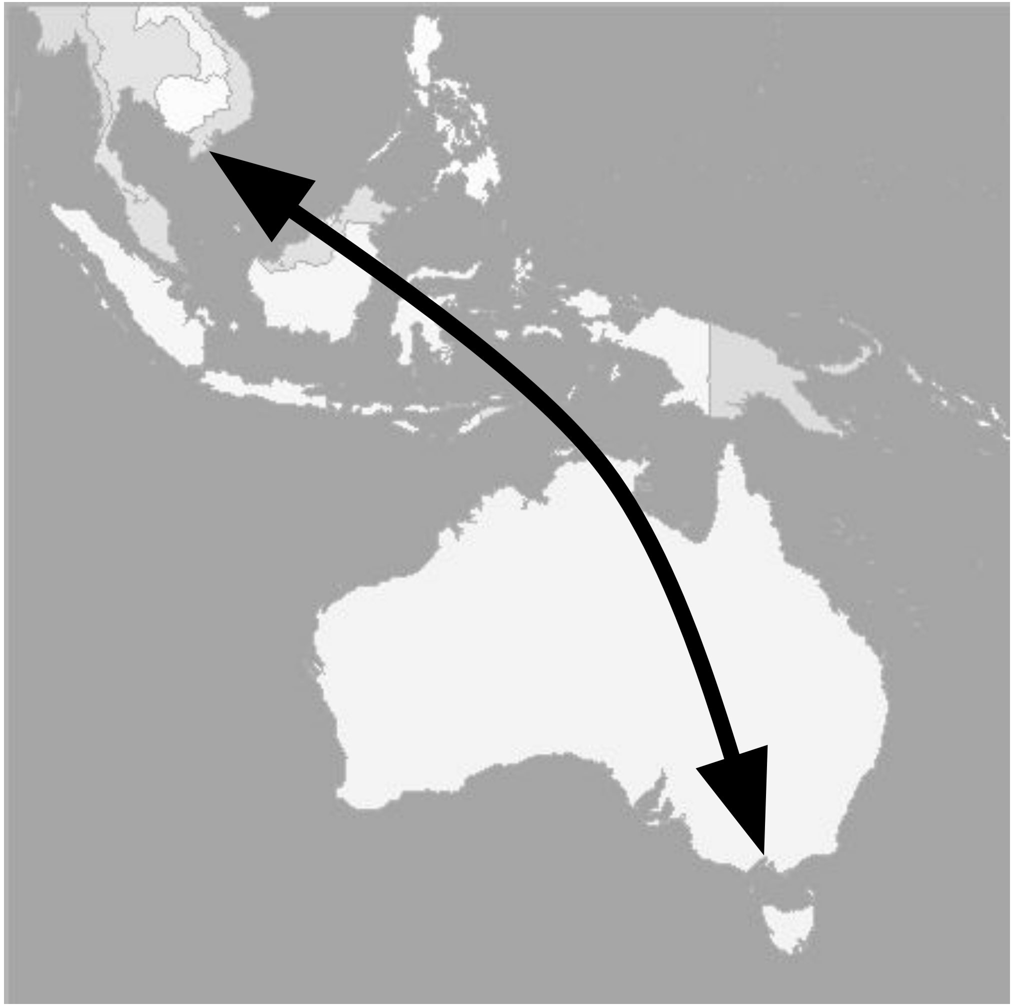
A/Prof Torsten Seemann
@torstenseemann

The University of Melbourne

U

Introduction

Neighbours



Doherty Institute

THE UNIVERSITY OF MELBOURNE + The Royal Melbourne Hospital

"Immunity and infection"

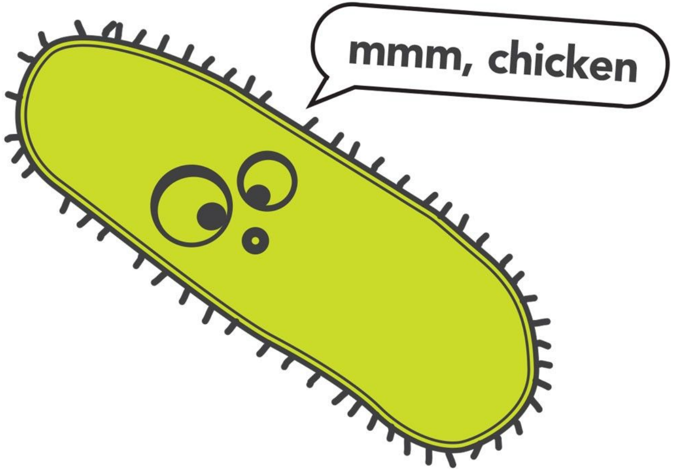
- Research
- Teaching
- Public health and reference labs
- Diagnostic services
- Clinical care in ID and immunity

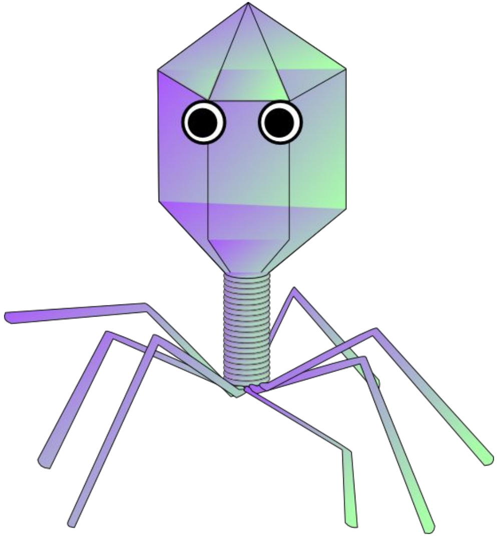
Microbiological Diagnostics Unit

- Oldest public health lab in Australia
    - established 1897 in Melbourne
    - historical ~600,000 isolate collection back to 1950s
- National reference laboratory
    - *Salmonella*, *Listeria*, EHEC
- W.H.O regional reference lab
    - vaccine preventable invasive bacterial pathogens

PUBLIC HEALTH
MDU PHL
MICROBIOLOGY

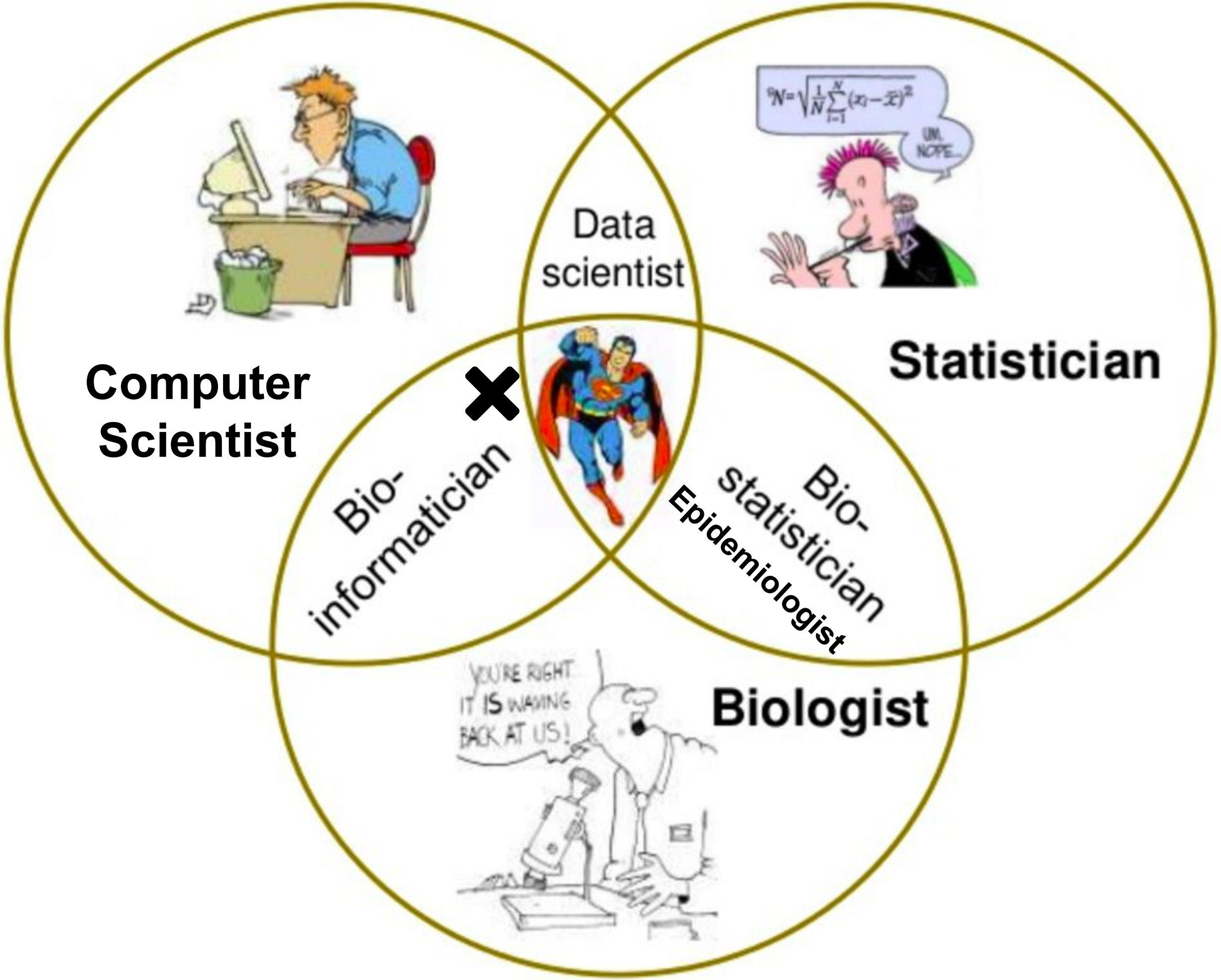
Bacteria + Viruses + Fungi





Foodborne, human (clinical), animal and environmental samples

Lots of genome sequencing



Computer Scientist

Data scientist

Statistician

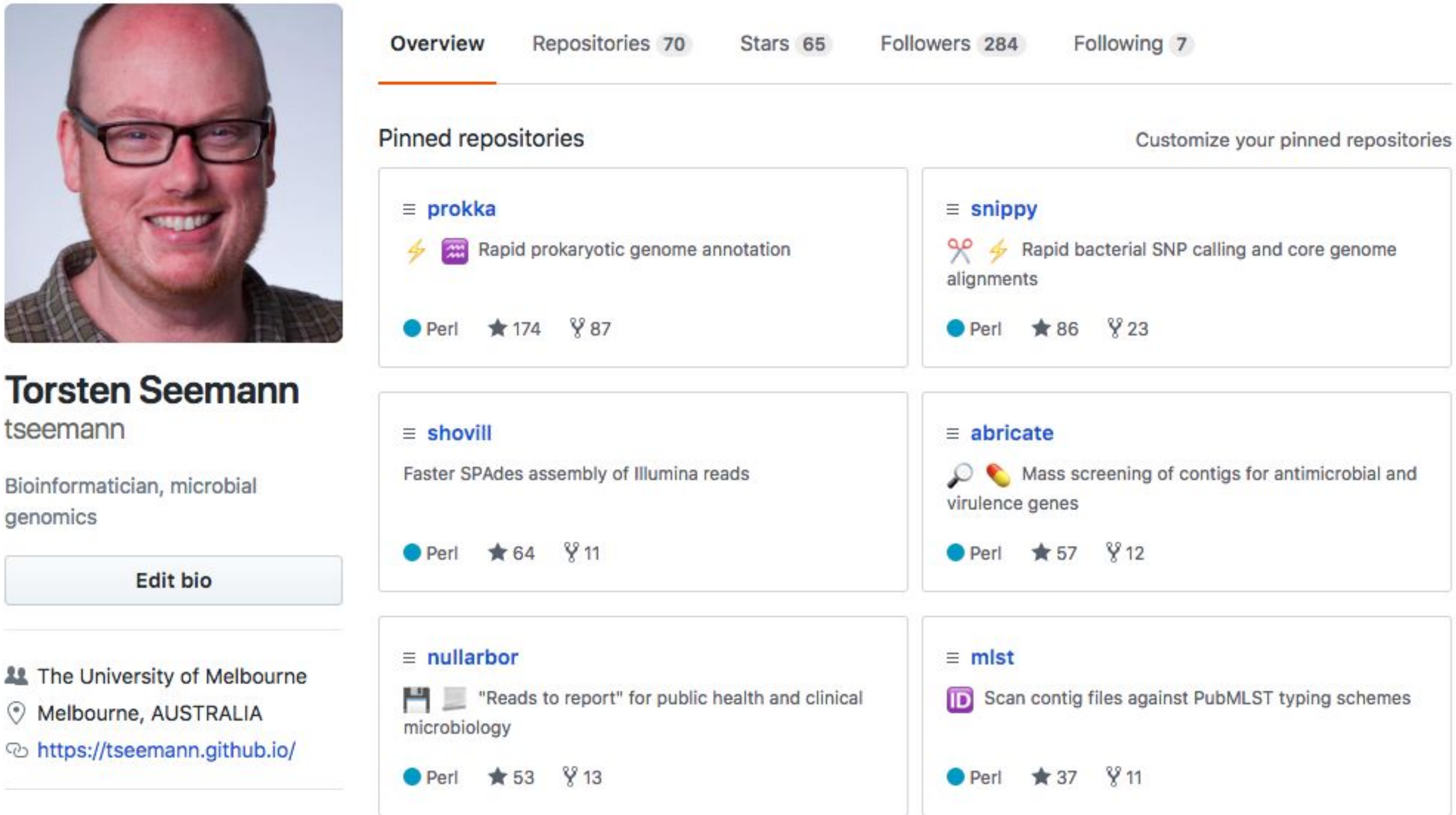
✖ Bio-informatician

Bio-statistician

Epidemiologist

Biologist

Software tools for bacterial genomics



Overview Repositories 70 Stars 65 Followers 284 Following 7

Torsten Seemann
tseemann

Bioinformatician, microbial genomics

Edit bio

The University of Melbourne
Melbourne, AUSTRALIA
https://tseemann.github.io/

Pinned repositories

Customize your pinned repositories

prokka
⚡ ♒ Rapid prokaryotic genome annotation
● Perl ★ 174 ⑂ 87

snippy
✂ ⚡ Rapid bacterial SNP calling and core genome alignments
● Perl ★ 86 ⑂ 23

shovill
Faster SPAdes assembly of Illumina reads
● Perl ★ 64 ⑂ 11

abricate
🔎 💊 Mass screening of contigs for antimicrobial and virulence genes
● Perl ★ 57 ⑂ 12

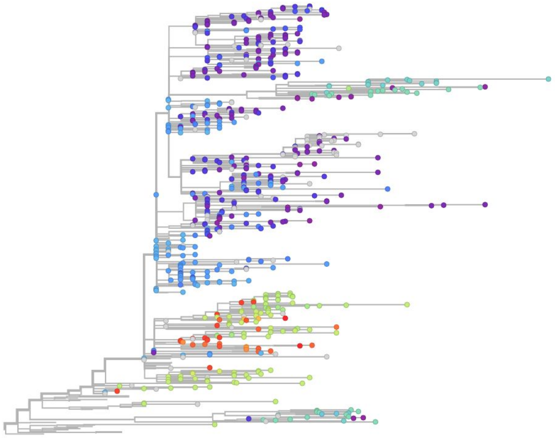
nullarbor
💾 📃 "Reads to report" for public health and clinical microbiology
● Perl ★ 53 ⑂ 13

mlst
🆔 Scan contig files against PubMLST typing schemes
● Perl ★ 37 ⑂ 11

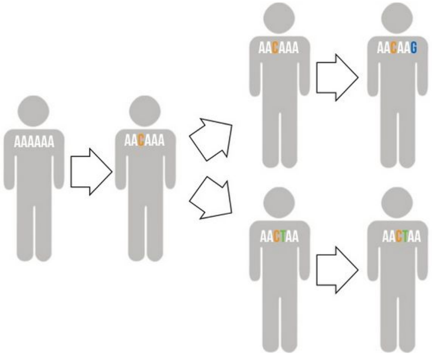
Traditional public health and clinical microbiology

The roles of a public health laboratory

Diagnostics



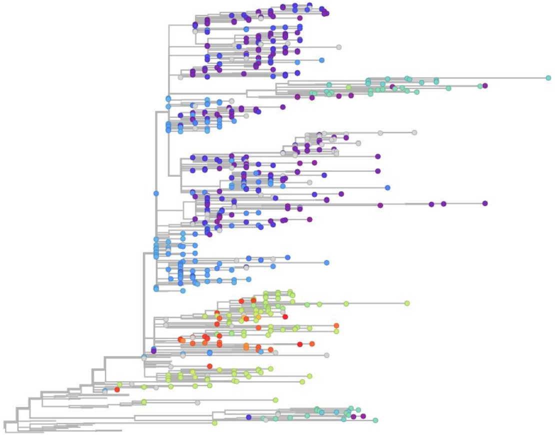
Surveillance



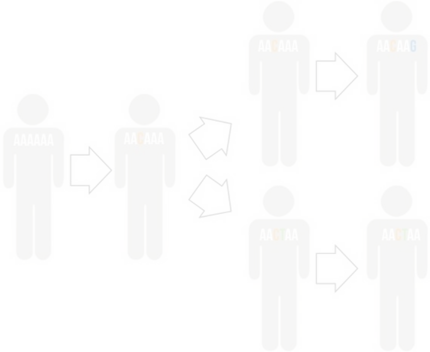
Outbreak response

The roles of a public health laboratory

Diagnostics

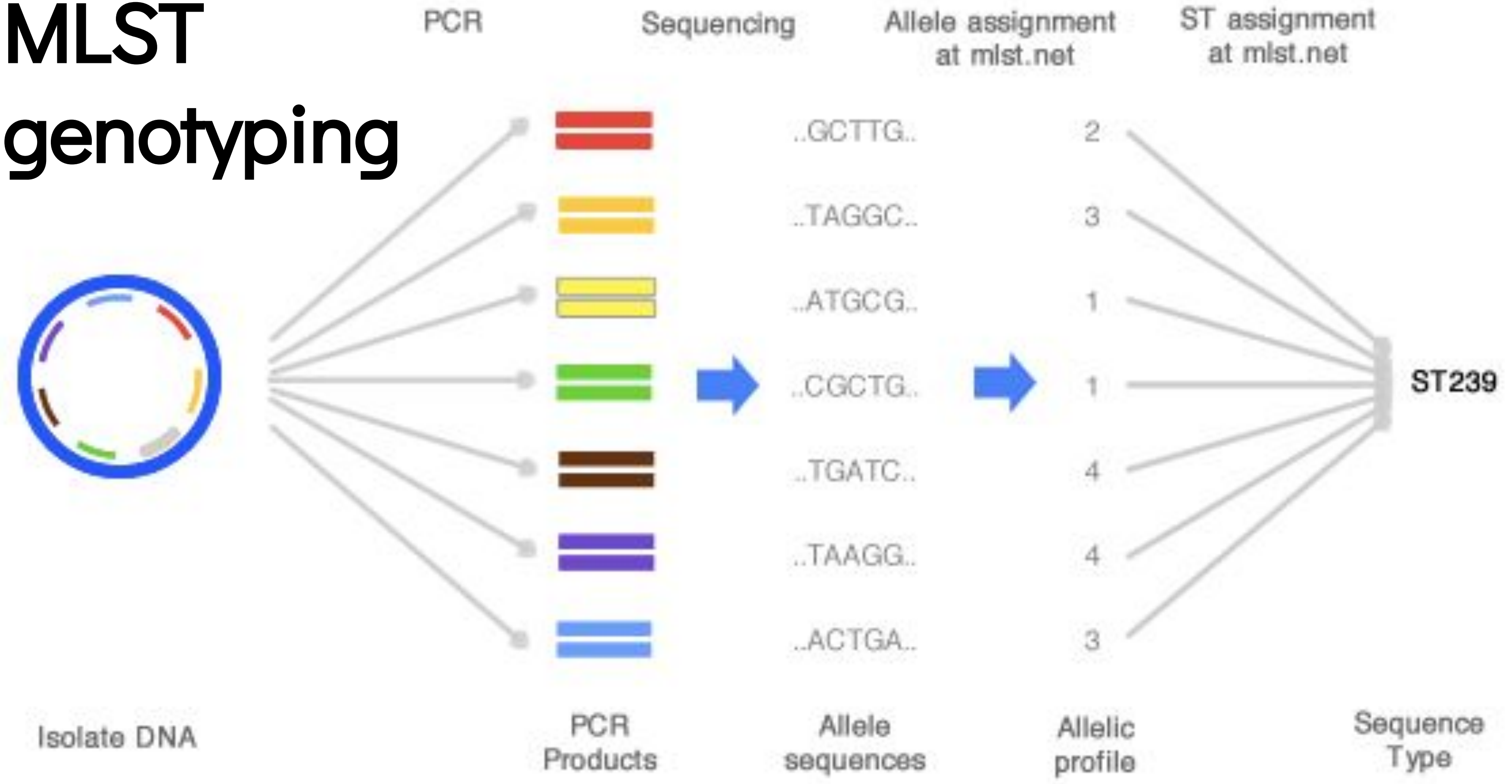


Surveillance



Outbreak response

MLST genotyping

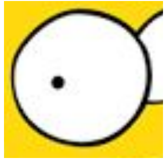


Other schemes

- Spa *S.aureus*
- EMM *S. pyogenes*
- SBT *L.pneumophila*
- NG-MAST *N.gonorrhea*
- MLVA *Salmonella enterica*
- MIRU VNTR *M.tuberculosis*
- Ribotyping *Clostridium sp.*

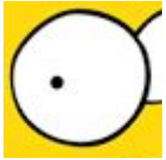
Pathogen: *Homerbacter simpsonii*

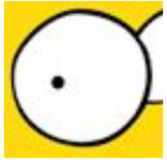
Focus on a small informative section



Genotyping via one or more marker genes.

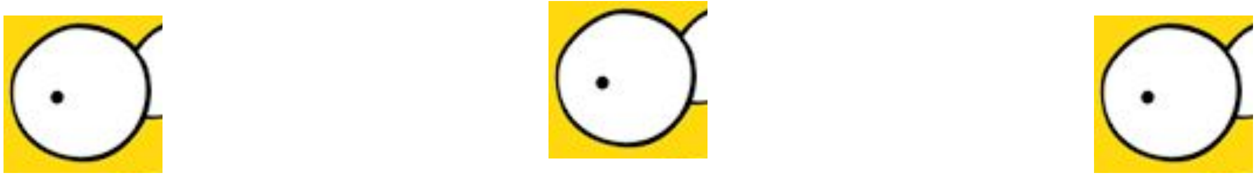
Another sample arrives at the lab





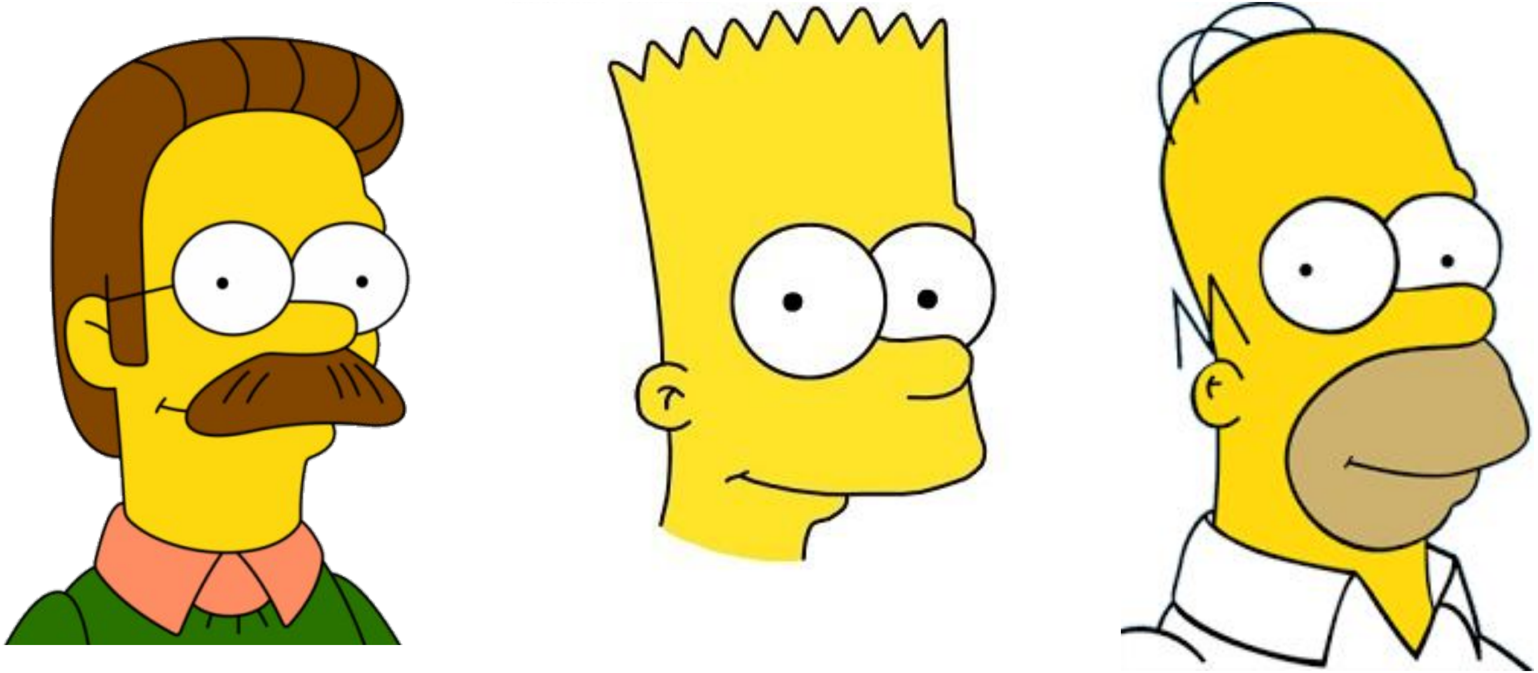
Looks related.

Another sample from same city



Possible outbreak ?

D’oh !



Case study

The Advertiser

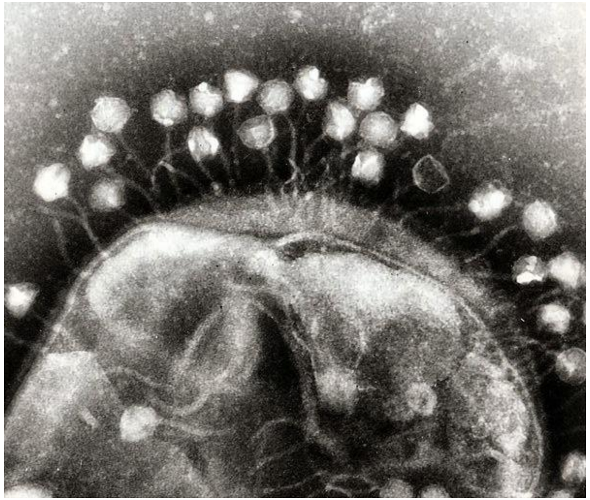
SA Health investigating seven salmonella cases with possible link to toxic salad recalled nationwide

Katrina Stokes, Health Reporter, The Advertiser
February 5, 2016 9:38pm

- 'Toxic salad' victim's absolute agony
- Salmonella outbreak linked to prepacked lettuce

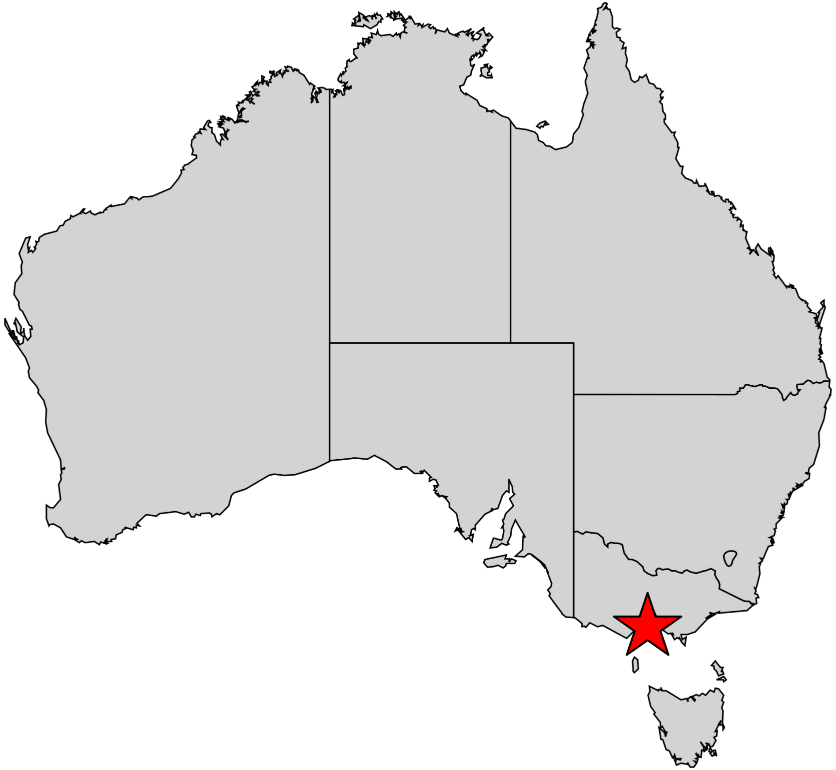
SOUTH Australian health authorities predict more people will be struck down with salmonella in coming days because they have eaten contaminated prepacked lettuce sold in major supermarket chains across the nation.

Salmonella typing



- Serotyping
  - O, H, Vi antigens -> 2500 serovars
- Phage typing
  - Panel of 34 phage -> 200 phage types
- MLVA (multilocus variable tandem repeat)
  - 5 locii -> 100s of MLVA types
- MLST (multilocus sequence typing)
  - 7 locii -> 3000 sequence types

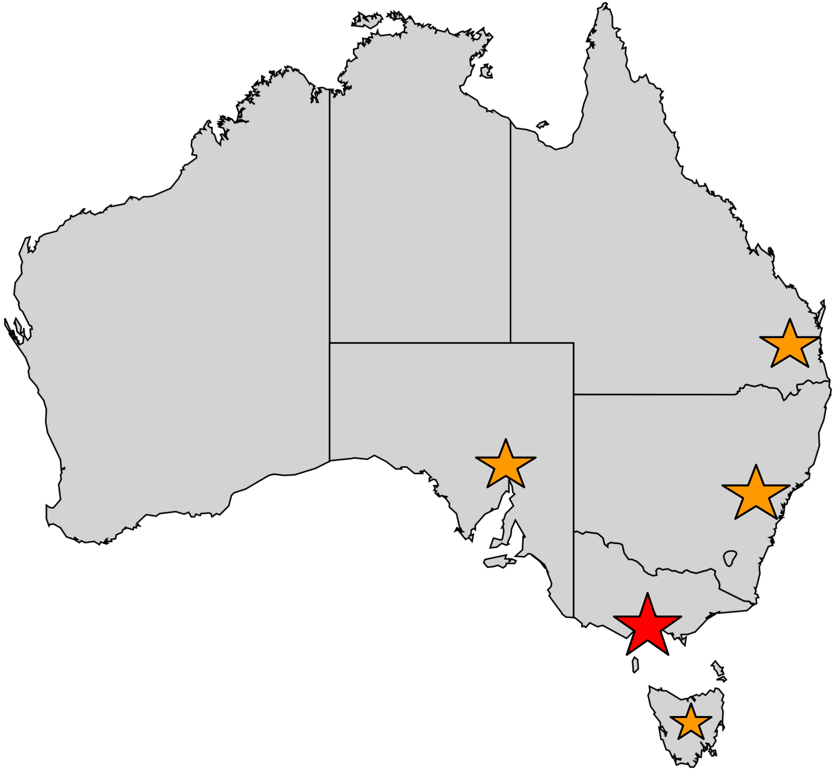
The outbreak begins



- :: Started in Victoria
- :: <100 samples

- :: Sero - **identical**
- :: Phage - **identical**
- :: MLVA - **identical**
- :: MLST - **identical**

The outbreak spreads

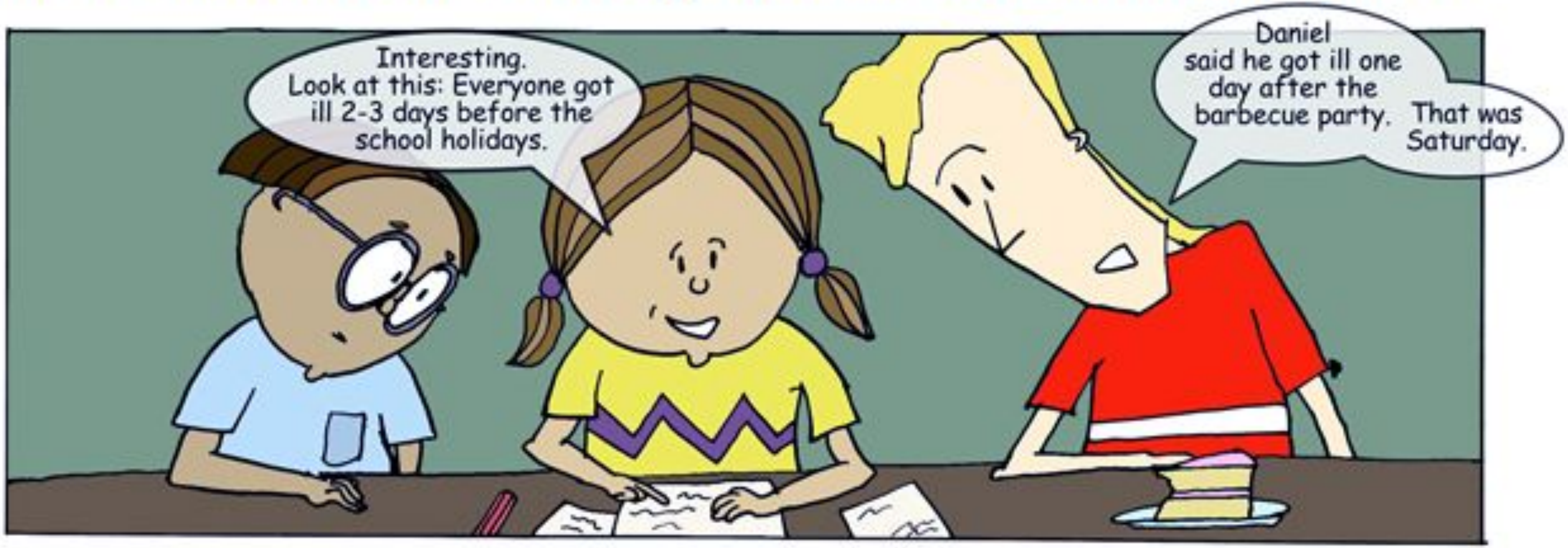


- :: Spreads interstate
- :: 100s of samples

- :: Sero - **identical**
- :: Phage - **identical**
- :: MLVA - **identical**
- :: MLST - **identical**

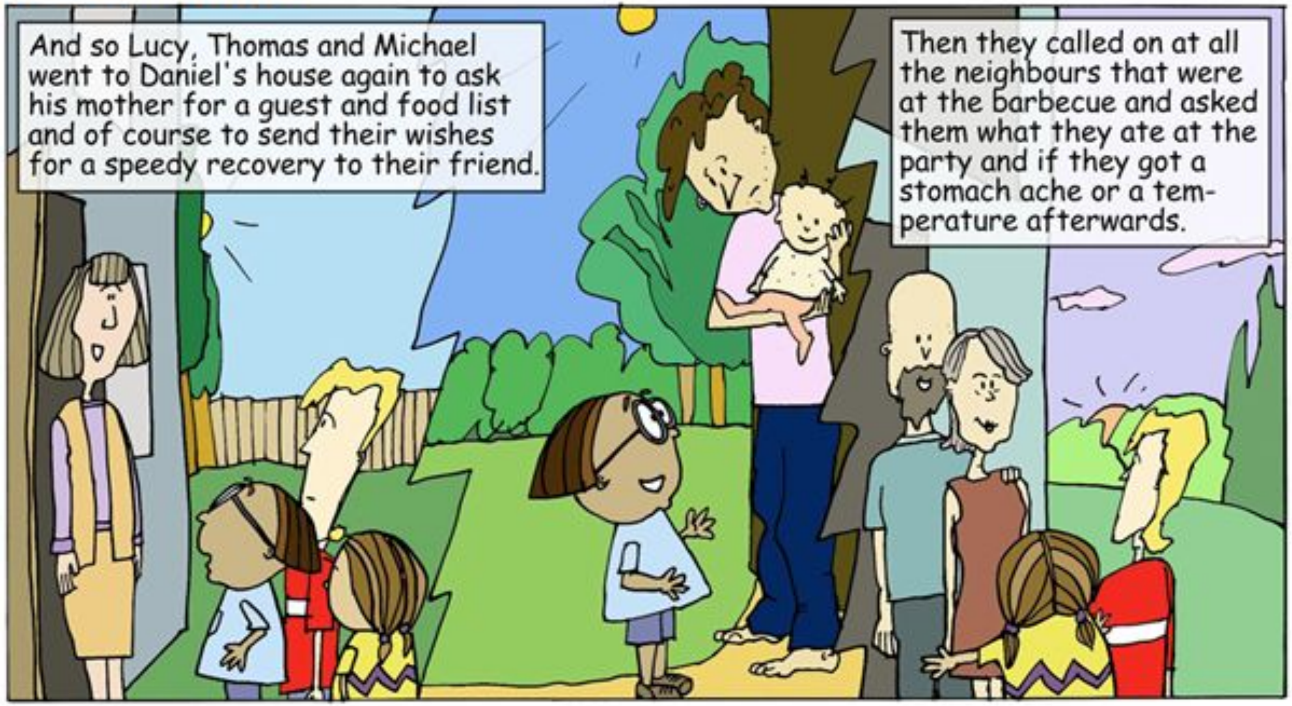
So now what?   Epidemiology!

The Disease Detectives detect patterns of disease in their line list



Good - but includes wasted time & effort

The Disease Detectives conducting a cohort study in their neighbourhood



How can we improve this?

Of course you know the answer.

Whole genome sequencing

- Use 100% of the genome
- Single nucleotide resolution
  - Infer source and transmission of infection
- Full complement of genes
  - Track mobile elements and antimicrobial resistance

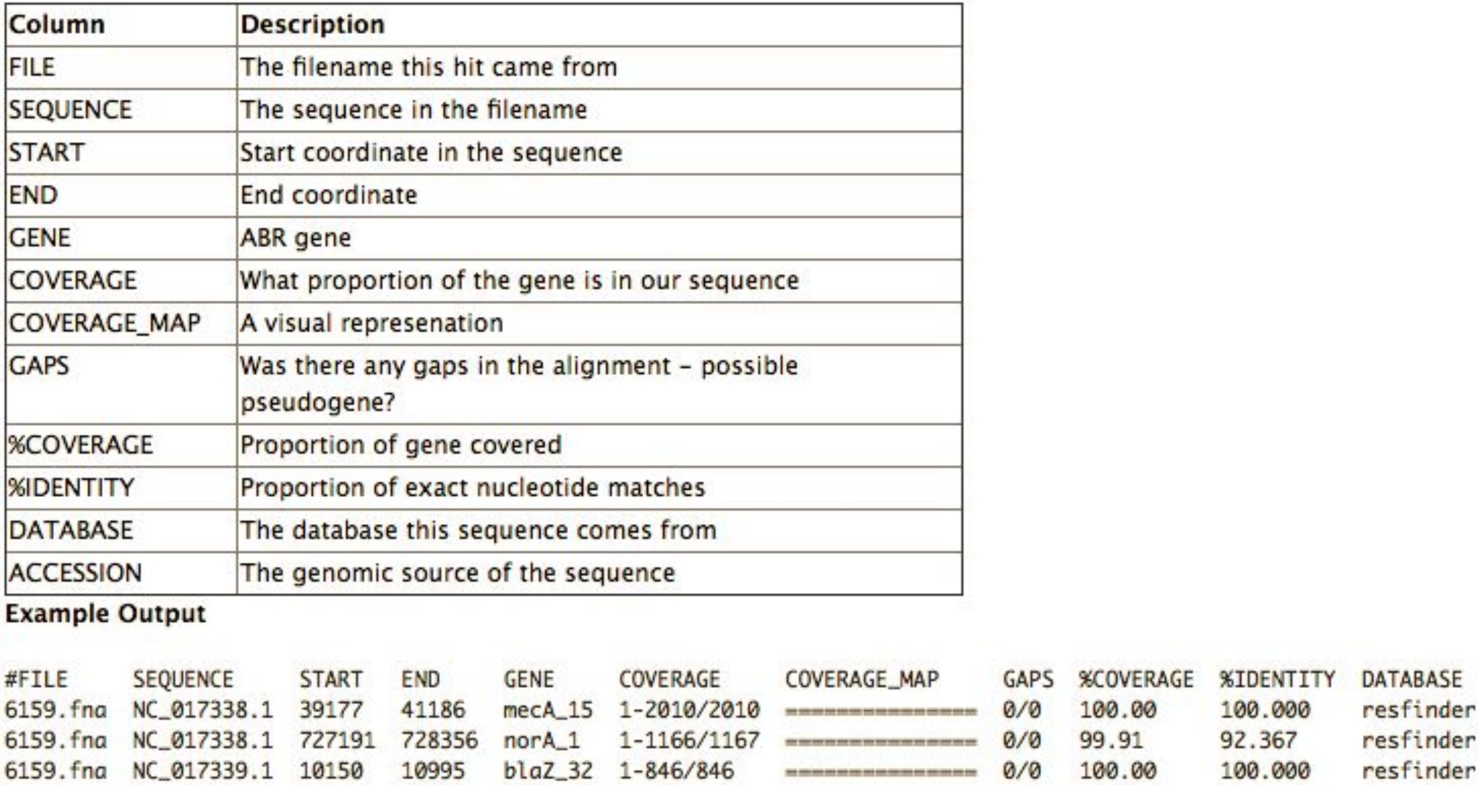
Routine sequencing of isolates

>500 per week
(and increasing+)

Species identification

```
  1.04  1046    1046    U       0       unclassified
 98.96  99624   142     -       1       root
 98.81  99473   1       -       131567    cellular organisms
 98.81  99472   194     D       2           Bacteria
 98.57  99233   111     P       1224          Proteobacteria
 98.45  99110   318     C       1236            Gammaproteobacteria
 98.07  98728   0       O       91347             Enterobacteriales
 98.07  98728   52477   F       543                 Enterobacteriaceae
 44.95  45256   665     G       561                   Escherichia
 44.20  44498   33391   S       562                     Escherichia coli
  8.84  8899    8899    -       1274814                   Escherichia coli APEC O78
  0.29  287     0       -       244319                    Escherichia coli O26:H11
  0.29  287     287     -       573235                      Escherichia coli O26:H11 str 11368
  0.21  216     216     -       316401                    Escherichia coli ETEC H10407
  0.19  193     0       -       168807                    Escherichia coli O127:H6
  0.19  193     193     -       574521                      Escherichia coli O127:H6 str E2348/69
```

Antimicrobial resistance

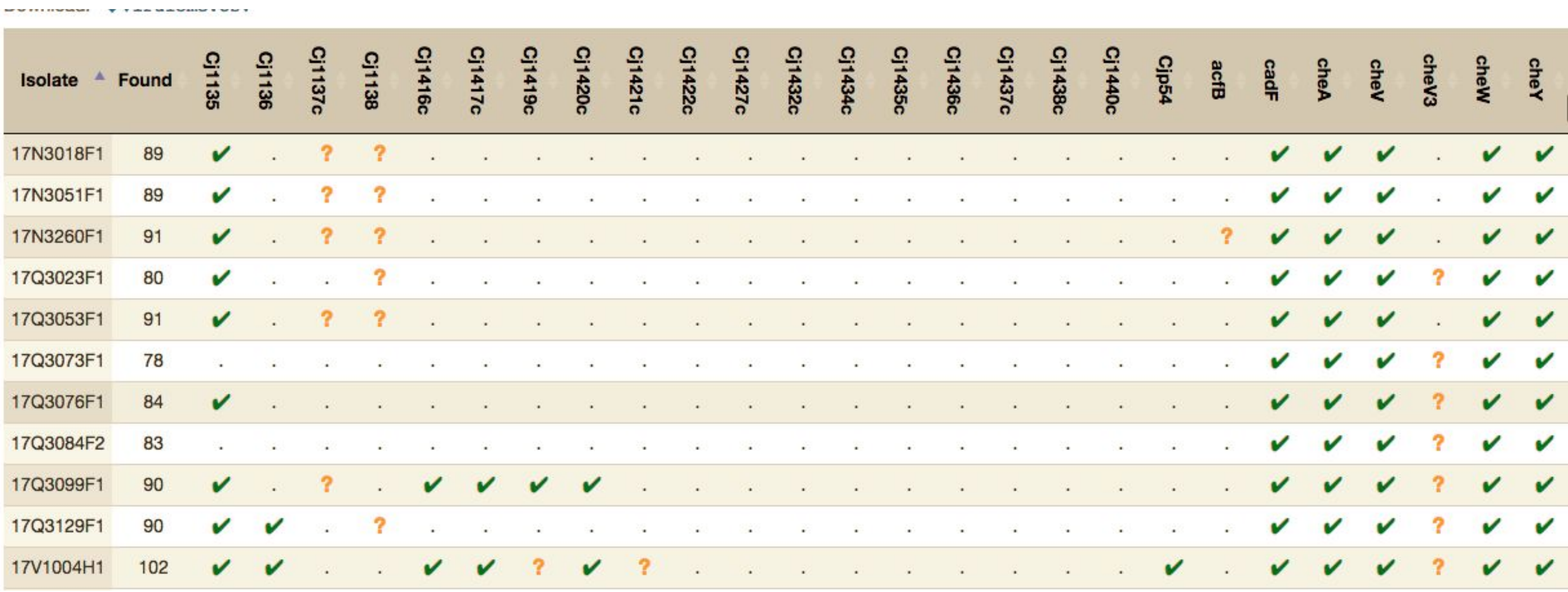


Column	**Description**
FILE	The filename this hit came from
SEQUENCE	The sequence in the filename
START	Start coordinate in the sequence
END	End coordinate
GENE	ABR gene
COVERAGE	What proportion of the gene is in our sequence
COVERAGE_MAP	A visual represenation
GAPS	Was there any gaps in the alignment - possible pseudogene?
%COVERAGE	Proportion of gene covered
%IDENTITY	Proportion of exact nucleotide matches
DATABASE	The database this sequence comes from
ACCESSION	The genomic source of the sequence

Example Output

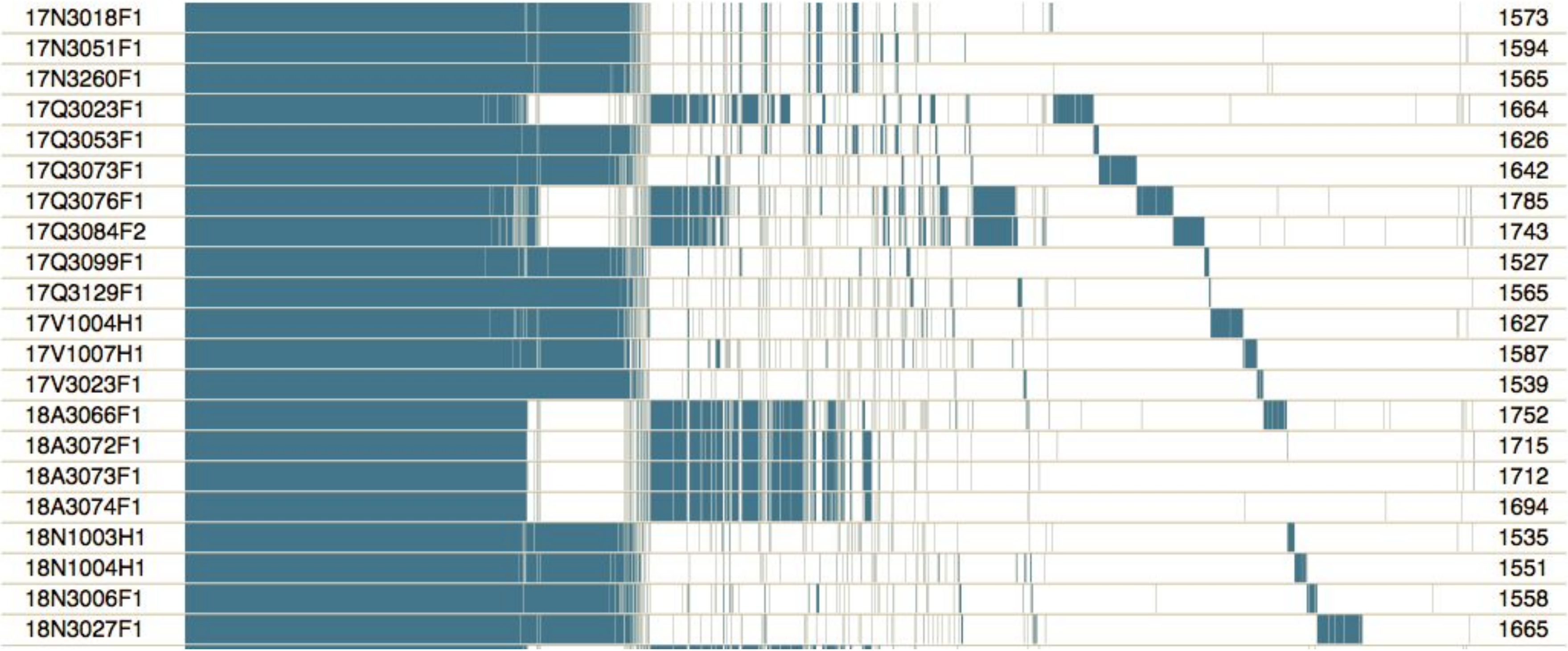
```
#FILE     SEQUENCE     START   END     GENE     COVERAGE     COVERAGE_MAP     GAPS  %COVERAGE  %IDENTITY  DATABASE
6159.fna  NC_017338.1  39177   41186   mecA_15  1-2010/2010  ===============  0/0   100.00     100.000    resfinder
6159.fna  NC_017338.1  727191  728356  norA_1   1-1166/1167  ===============  0/0   99.91      92.367     resfinder
6159.fna  NC_017339.1  10150   10995   blaZ_32  1-846/846    ===============  0/0   100.00     100.000    resfinder
```

Virulome

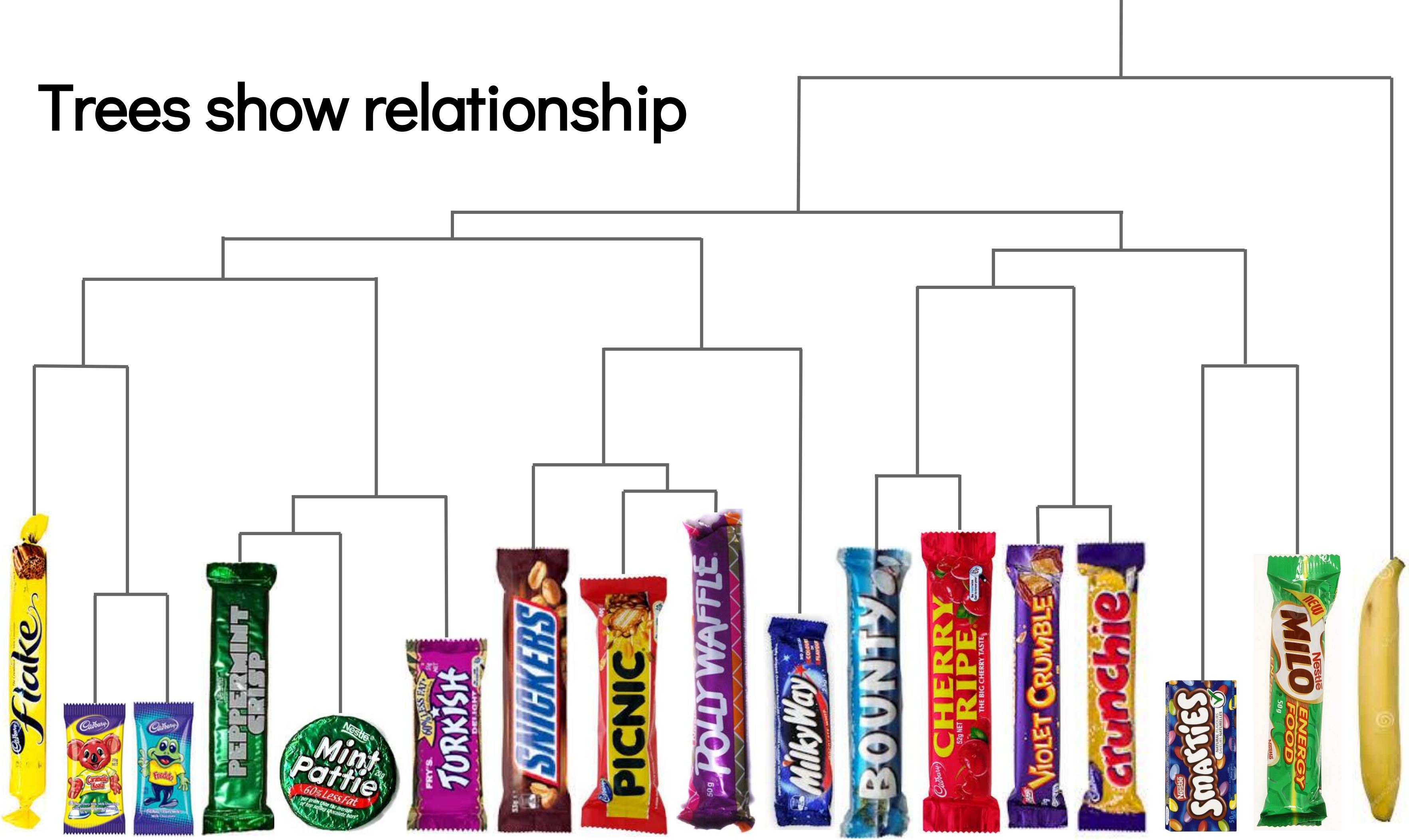


Isolate	Found	Cj1135	Cj1136	Cj1137c	Cj1138	Cj1416c	Cj1417c	Cj1419c	Cj1420c	Cj1421c	Cj1422c	Cj1427c	Cj1432c	Cj1434c	Cj1435c	Cj1436c	Cj1437c	Cj1438c	Cj1440c	Cjp54	acfB	cadF	cheA	cheV	cheV3	cheW	cheY
17N3018F1	89	✓	.	?	?	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	✓	✓	✓	.	✓	✓
17N3051F1	89	✓	.	?	?	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	✓	✓	✓	.	✓	✓
17N3260F1	91	✓	.	?	?	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	?	✓	✓	✓	.	✓	✓
17Q3023F1	80	✓	.	.	?	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	✓	✓	✓	?	✓	✓
17Q3053F1	91	✓	.	?	?	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	✓	✓	✓	.	✓	✓
17Q3073F1	78	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	✓	✓	✓	?	✓	✓
17Q3076F1	84	✓	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	✓	✓	✓	?	✓	✓
17Q3084F2	83	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	✓	✓	✓	?	✓	✓
17Q3099F1	90	✓	.	?	.	✓	✓	✓	✓	.	.	.	.	.	.	.	.	.	.	.	.	✓	✓	✓	?	✓	✓
17Q3129F1	90	✓	✓	.	?	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	✓	✓	✓	?	✓	✓
17V1004H1	102	✓	✓	.	.	✓	✓	?	✓	?	.	.	.	.	.	.	.	.	.	✓	.	✓	✓	✓	?	✓	✓

Pan genome content (genes)



Trees show relationship



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

```
                SNP     Deletion
AGTCTGATTAGCTTAGCTTGTAGCGCTATATTAT    Reference
AGTCTGATTAGCTTAGAT
      ATTAGCTTAGATTGTAG
           CTTAGATTGTAGC-C
    TGATTAGCTTAGATTGTAGC-CTATAT
        TAGCTTAGATTGTAGC-CTATATT      Reads
             TAGATTGTAGC-CTATATTA
             TAGATTGTAGC-CTATATTAT
```

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

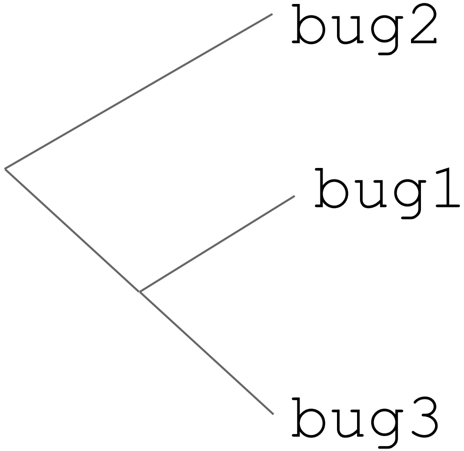
```
bug1  GATTACCAGCATTAAGG-TTCTCCAATC
bug2  GAT---CTGCATTATGGATTCTCCATTC
bug3  G-TTACCAGCACTAA-------CCAGTC
core  | |   |||||||||       ||||||
SNPs         |   |  |          |
```

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

Infer phylogeny from aligned core SNPs

bug1	ATAA
bug2	TTTT
bug3	ACAG

This is the primary information given to to tree building software.



Neighbour joining
Minimum evolution
Maximum likelihood
Bayesian models

Pathogens without borders

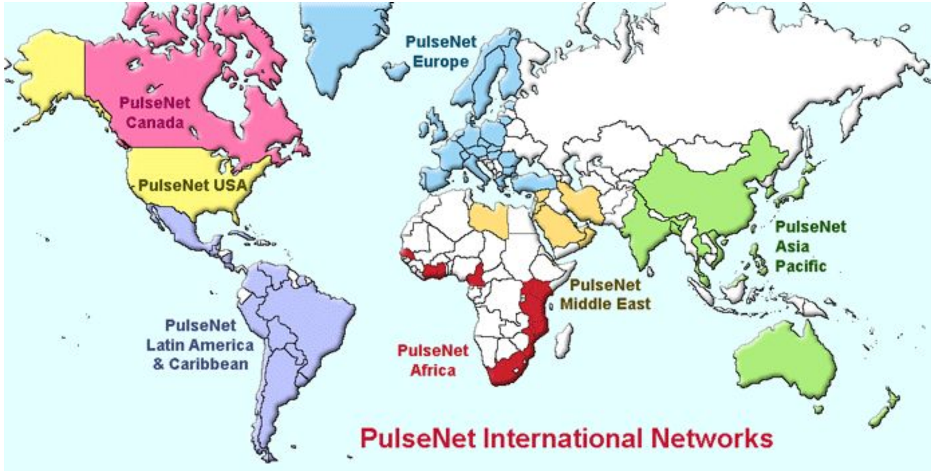
The success of Pulse Net

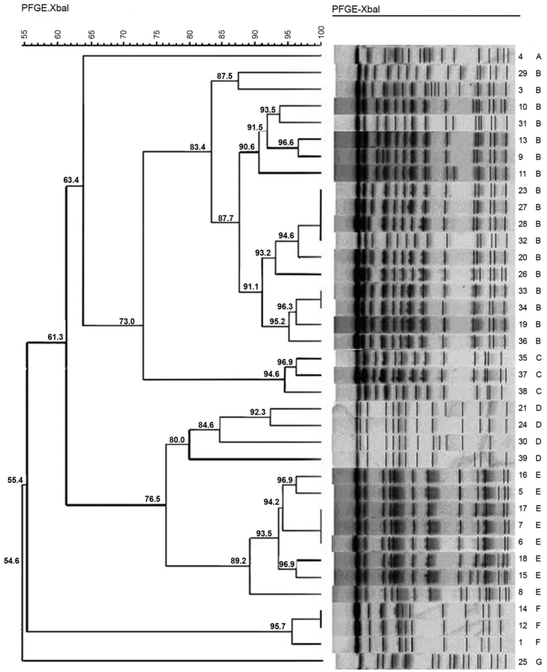
PulseNet™ International

The International Molecular Subtyping Network for Foodborne Disease Surveillance

USA, Canada, Europe, Asia Pacific, Latin America, Middle East

CDC





GenomeTrakr

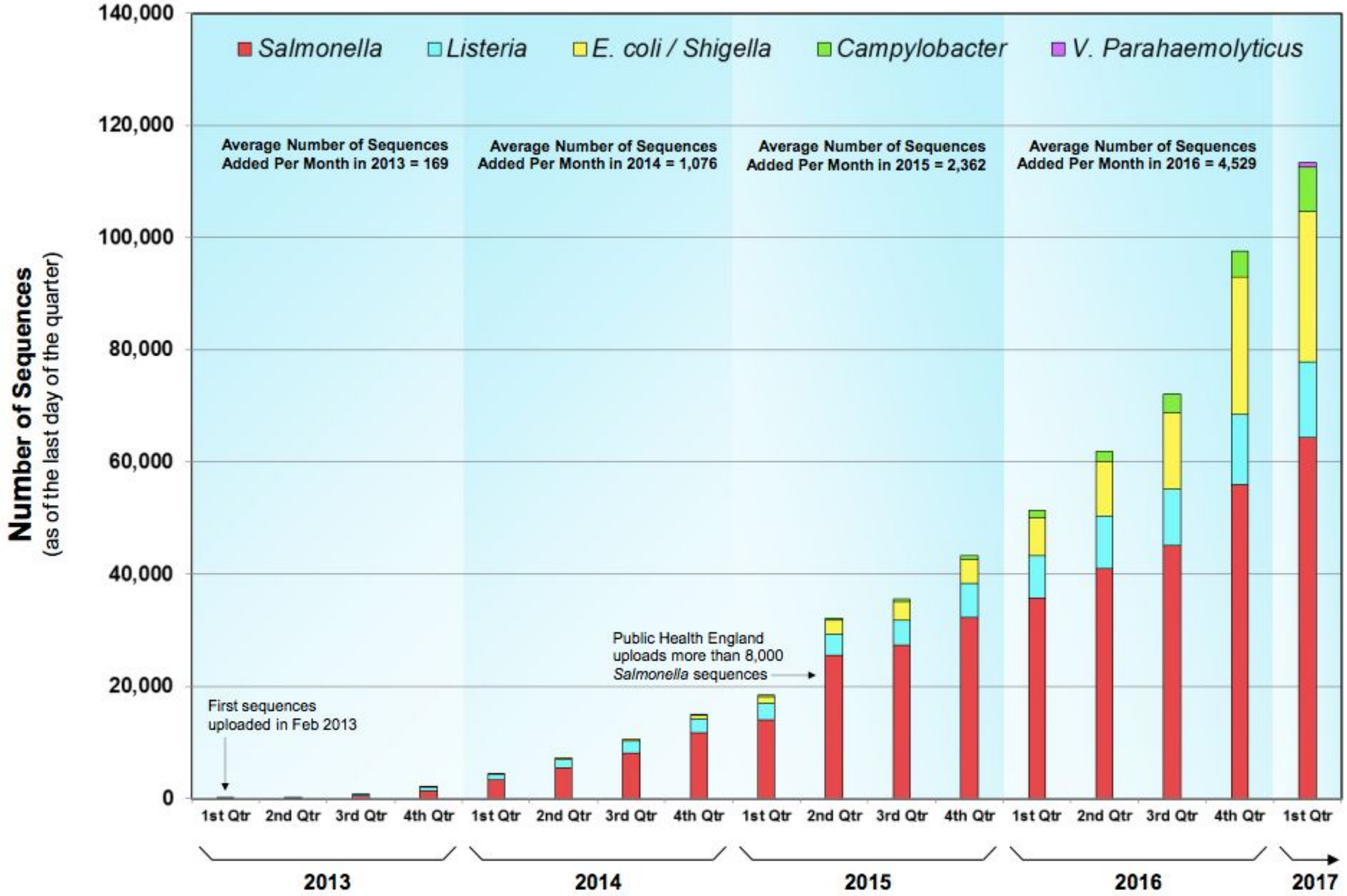
NCBI

FDA U.S. Food and Drug Administration

- International cooperation
  - Led by FDA + NCBI
  - >60 collaborating institutes inc. UK PHE, DK DTU, MX
  - *Salmonella, L.mono, E.coli, Campy, Vibrio*

- Public SRA BioProject #183844
  - Real-time submission of WGS genome reads
  - Nightly updates of phylogenomic trees
  - Contains ~60,000 strains of *Salmonella*

Total Number of Sequences in the GenomeTrakr Database



Average Number of Sequences Added Per Month in 2013 = 169

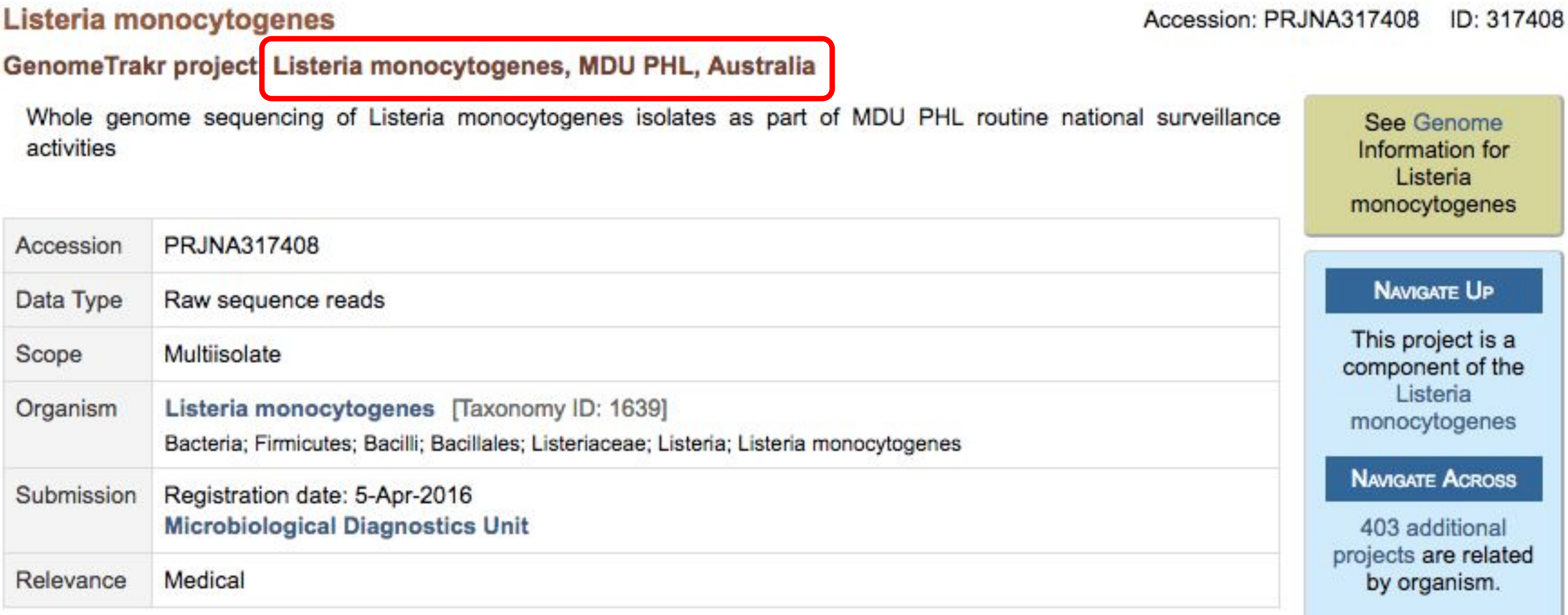
Average Number of Sequences Added Per Month in 2014 = 1,076

Average Number of Sequences Added Per Month in 2015 = 2,362

Average Number of Sequences Added Per Month in 2016 = 4,529

First sequences uploaded in Feb 2013

Public Health England uploads more than 8,000 *Salmonella* sequences



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

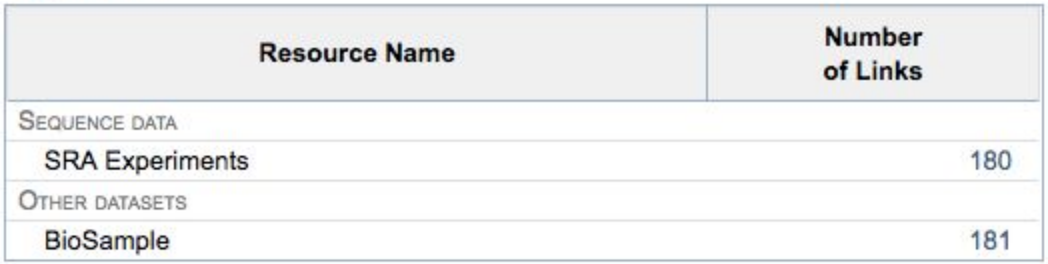
This project is a component of the Listeria monocytogenes

Navigate Across

403 additional projects are related by organism.

33 additional projects are components of the Listeria monocytogenes.

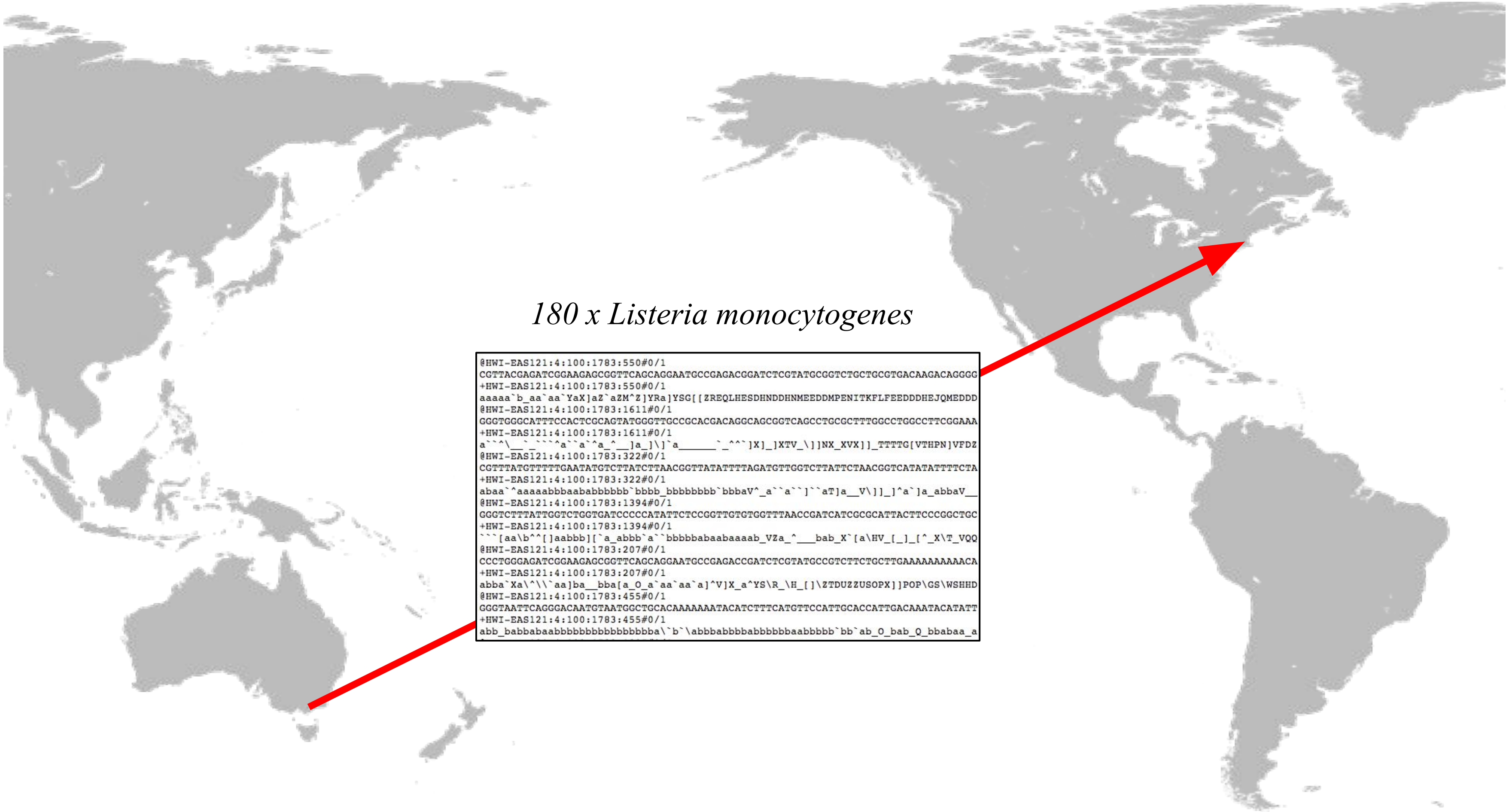
Project Data:



Resource Name	Number of Links
Sequence data	
SRA Experiments	180
Other datasets	
BioSample	181

Bill Klimke

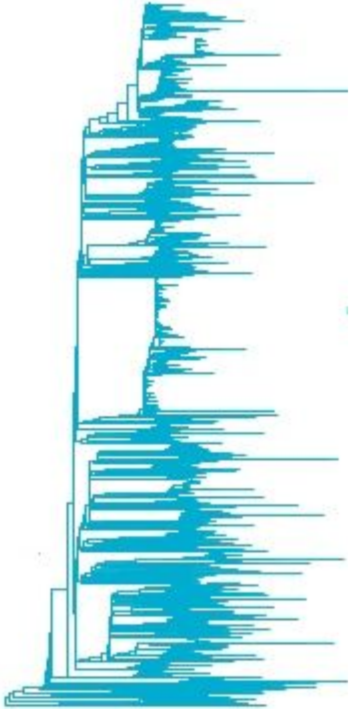
180 x Listeria monocytogenes

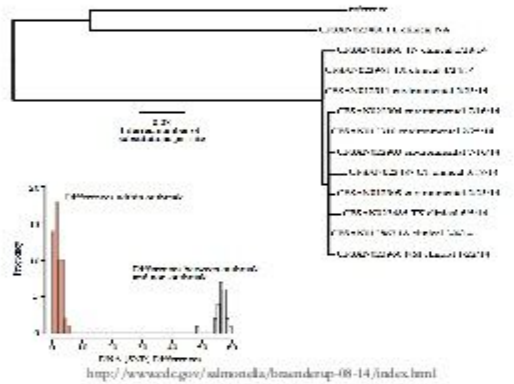


~~~
@HWI-EAS121:4:100:1783:550#0/1
CGTTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACGGATCTCGTATGCGGTCTGCTGCGTGACAAGACAGGGG
+HWI-EAS121:4:100:1783:550#0/1
aaaaa`b_aa`aa`YaX]aZ`aZM^Z]YRa]YSG[[ZREQLHESDHNDDHNMEEDDMPENITKFLFEEDDDHEJQMEDDD
@HWI-EAS121:4:100:1783:1611#0/1
GGGTGGGCATTTCCACTCGCAGTATGGGTTGCCGCACGACAGGCAGCGGTCAGCCTGCGCTTTGGCCTGGCCTTCGGAAA
+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`^aaaaabbbaababbbbbb`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`\abbbabbbbabbbbbbaabbbbb`bb`ab_O_bab_Q_bbabaa_a
~~~

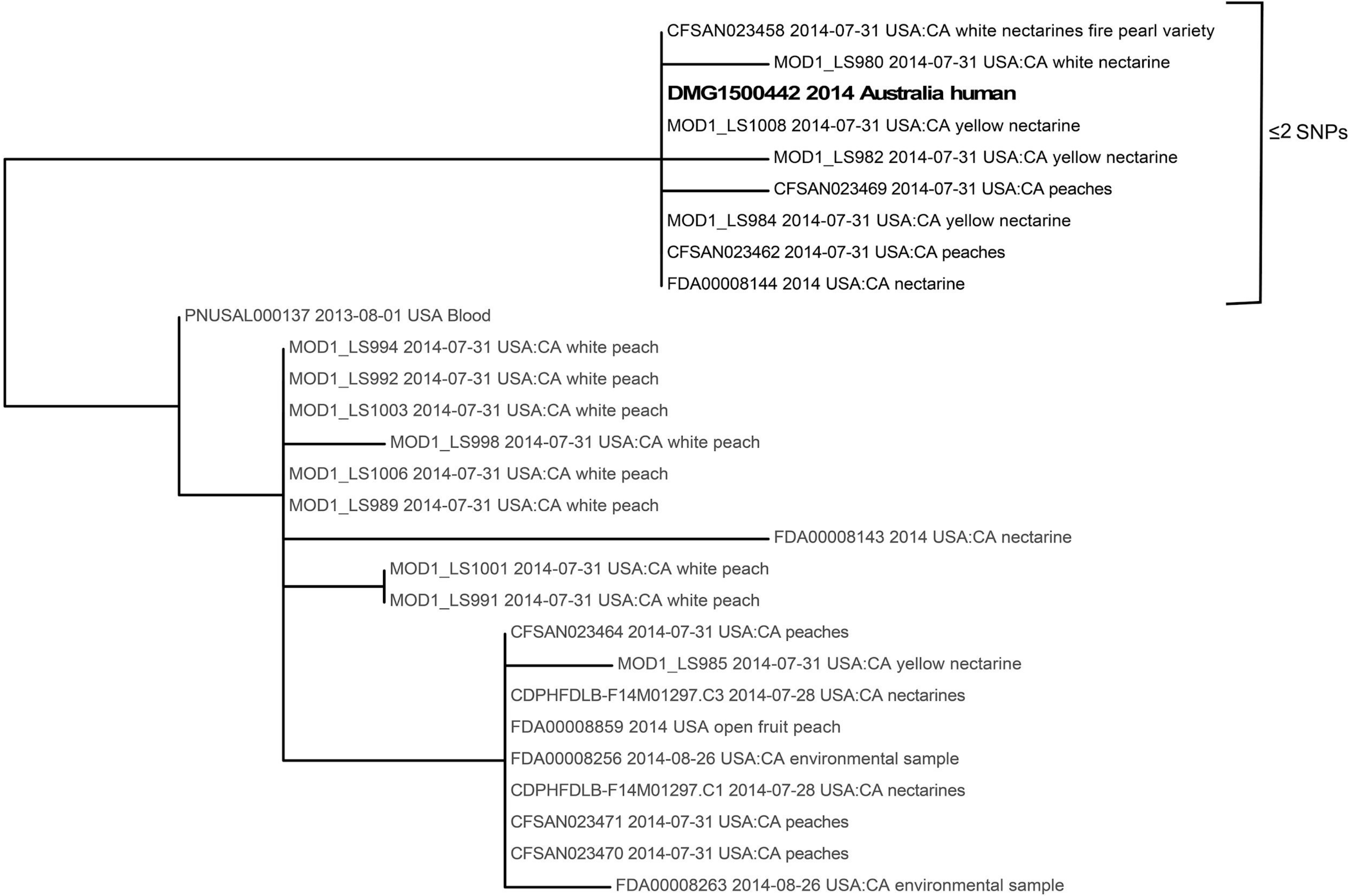
Nightly updates to find new matches

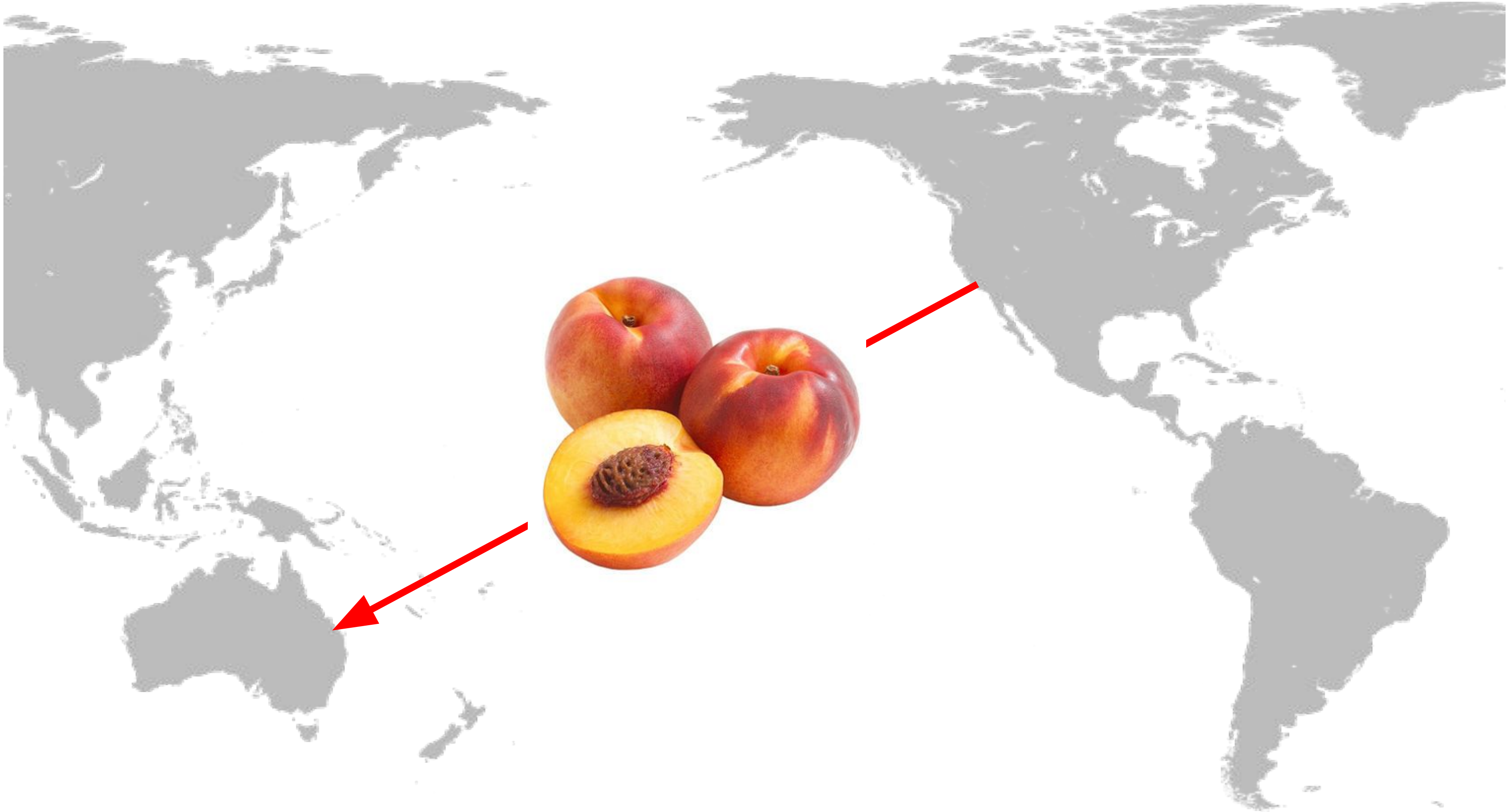
Errol Strain
(FDA)



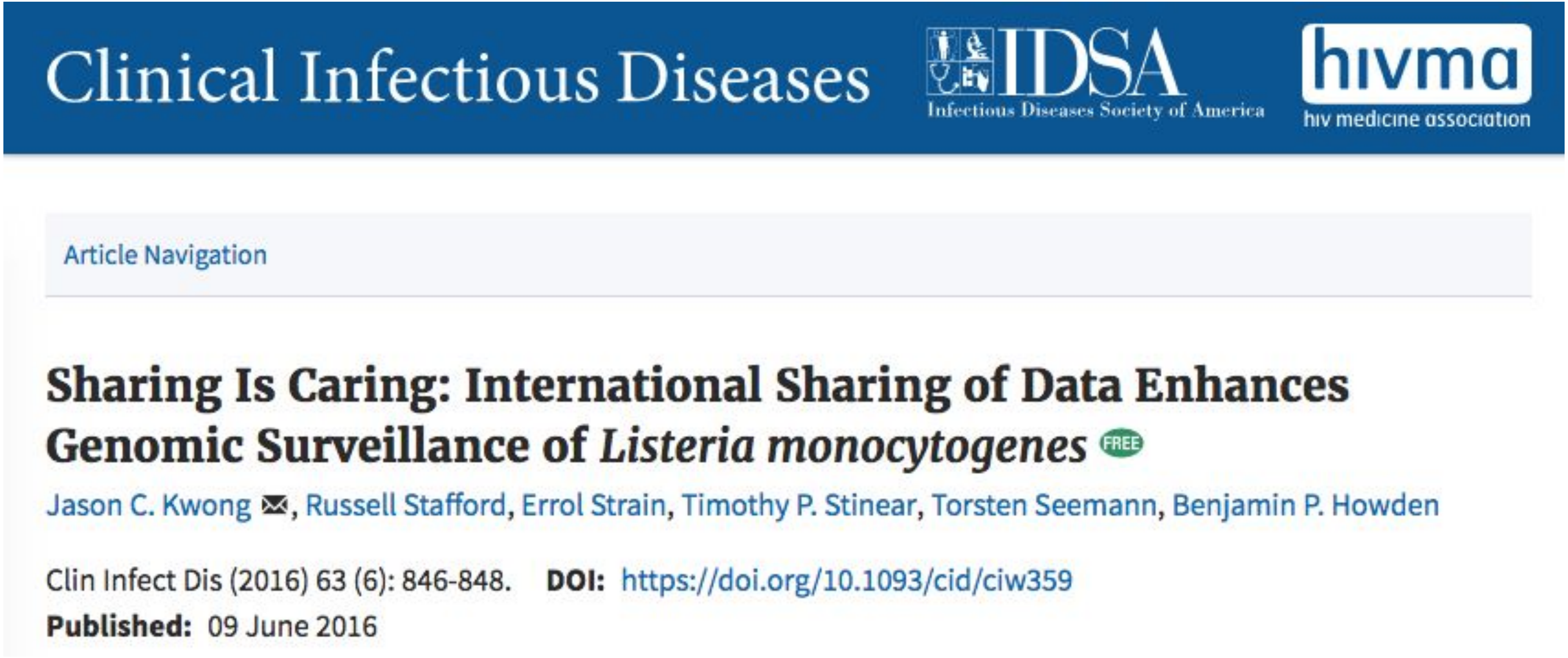


http://www.cdc.gov/salmonella/braenderup-08-14/index.html


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




CASE
CLOSED



Clinical Infectious Diseases

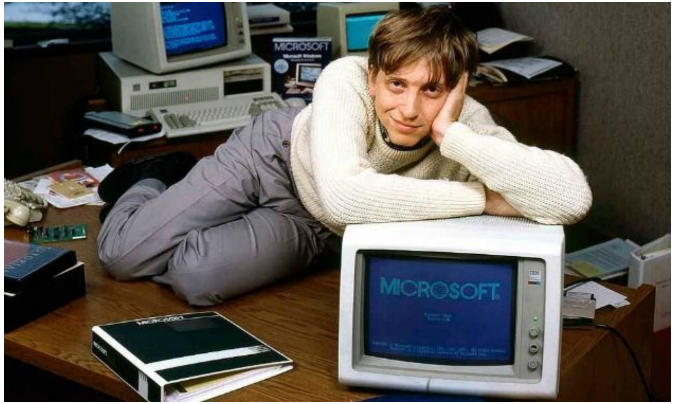
Sharing Is Caring: International Sharing of Data Enhances Genomic Surveillance of *Listeria monocytogenes*

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

What is preventing sharing?



A vision for Australia (and the world)

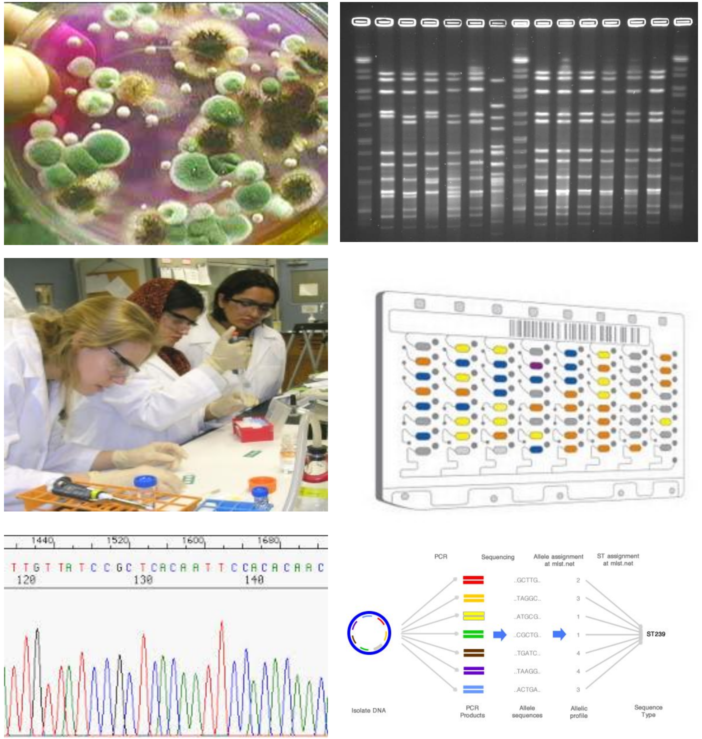
“AusTrakka”

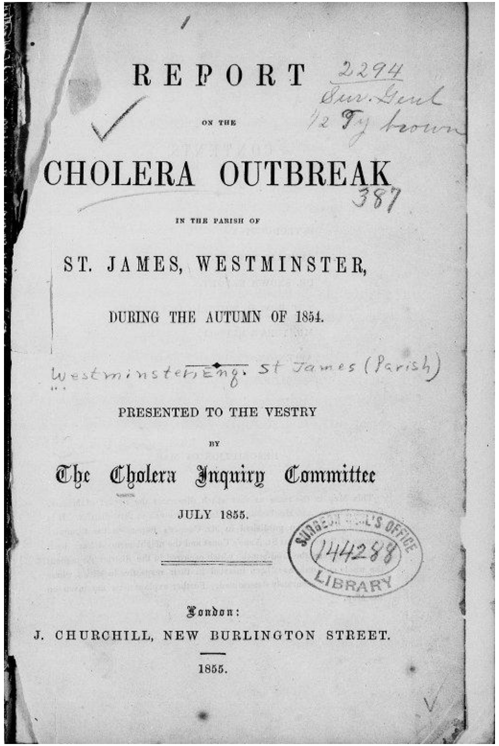
- A shared online system for all Australian labs
  - upload samples
  - automated standard/specific analyses ***in real time***
  - simple reports and visualization
  - easy to submit to international archives (SRA)

- Access control
  - each lab controls their own data
  - jurisdictions can share data in national outbreaks

Conclusions

Traditional workflow





REPORT

2294
Sur. Genl
1/2 Ty brown

ON THE

CHOLERA OUTBREAK

387

IN THE PARISH OF

ST. JAMES, WESTMINSTER,

DURING THE AUTUMN OF 1854.

Westminster, Eng. St James (Parish)

PRESENTED TO THE VESTRY

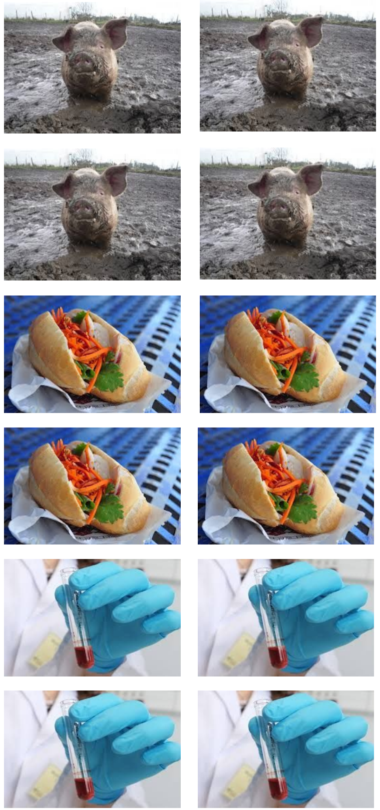
BY

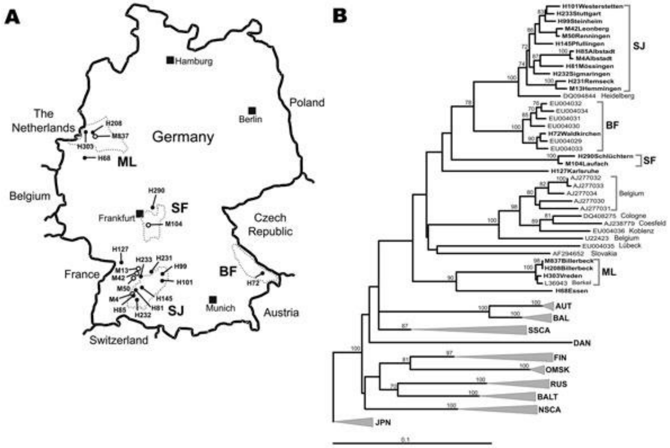
The Cholera Inquiry Committee

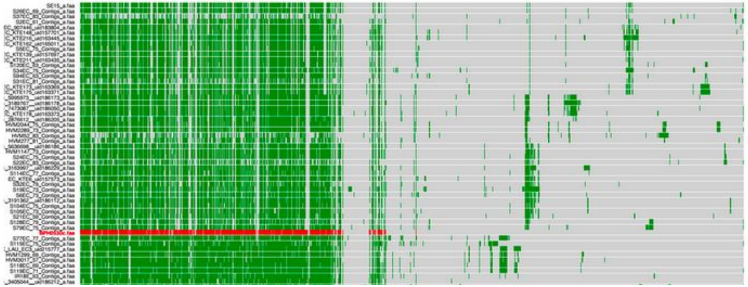
JULY 1855.

London:
J. CHURCHILL, NEW BURLINGTON STREET.
1855.

Modern workflow







Does WGS deliver?

Microbiology

Domain expertise

Bioinformatics

Yes!

Epidemiology

This means ***scientists*** not just software

Technology

Always changing...

Acknowledgements

Phil Ashton

Liz Batty

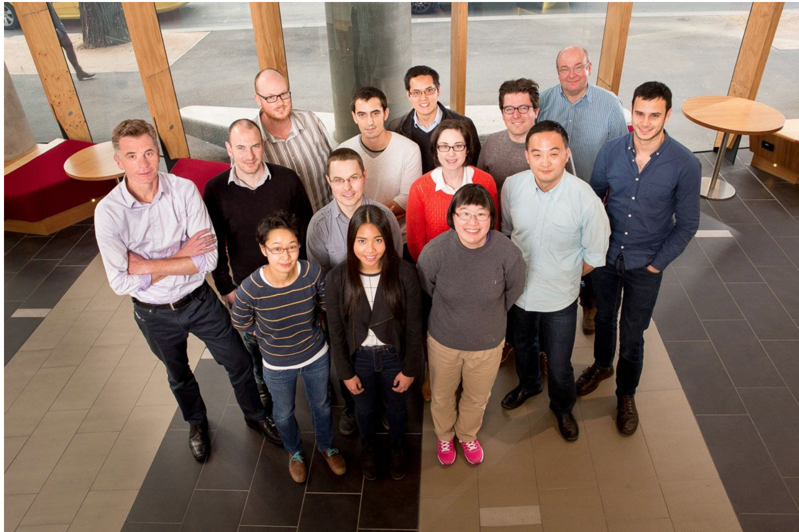
Thuy Vo Thi

Dieter Bulach

Cat Anscombe

Hai Ho

+ my Melbourne colleagues



Thank you for listening

The End