Will Whole Genome Sequencing Pathogens Revolutionise Infectious Diseases and Public Health?

Derrick Crook

Public Health England

University of Oxford

Oxford University Hospitals FT Trust

UNIVERSITY OF OXFORD

Public Health England

What do we need?

- Accurate species identification

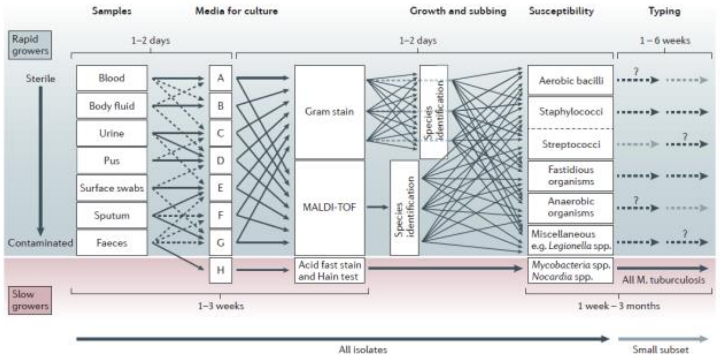
- Feature identification (e.g. resistance prediction; toxin and virulence prediction)

- High resolution typing to identify and characterise outbreaks e.g. time scaled phylogenies/genealogies (family trees)

- Fast, cheap, accurate outputs and on all specimens/isolates

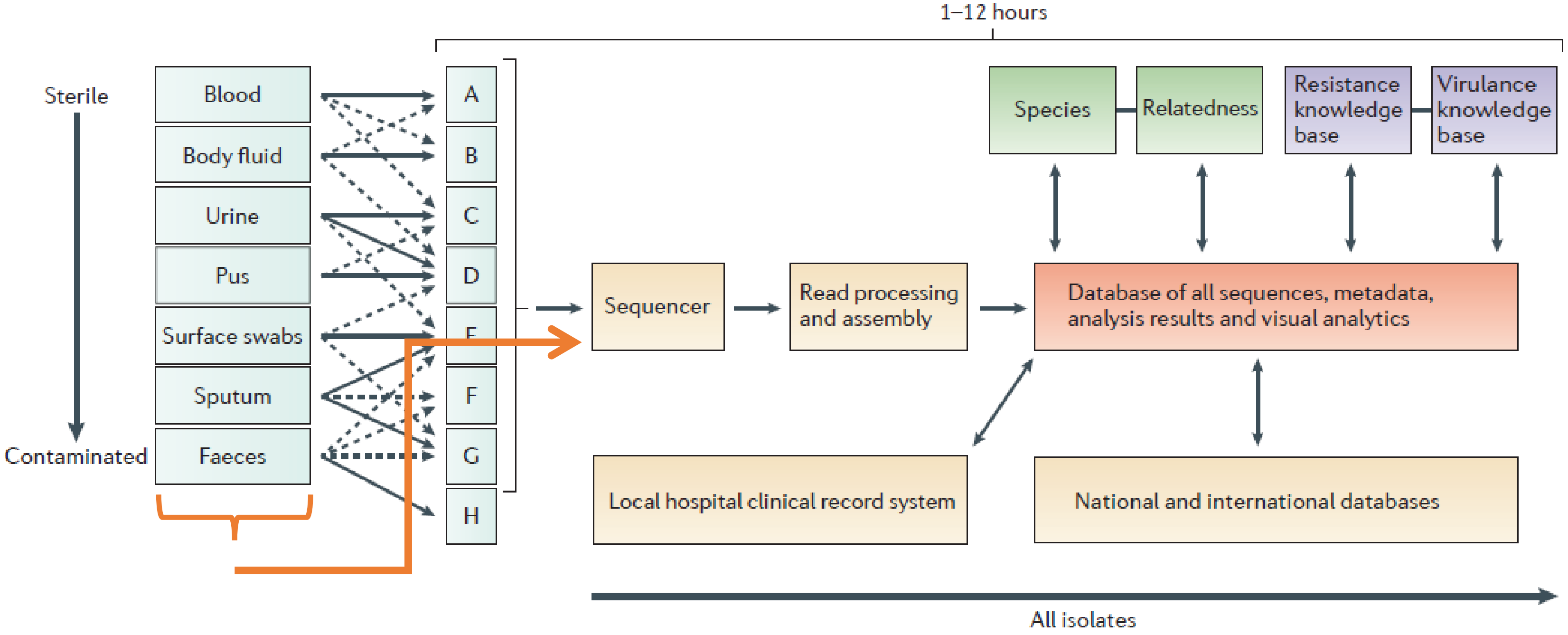
- Linkage to pathogen phenotype and patient epidemiological/clinical record data as an enduring encyclopaedic store of information

Concept for ideal whole genome sequencing solution



Nature Reviews | Genetics

In one step generate the complete diagnostic, typing and surveillance information


1–12 hours
Sterile
Contaminated
Blood
Body fluid
Urine
Pus
Surface swabs
Sputum
Faeces
A
B
C
D
E
F
G
H
Sequencer
Read processing and assembly
Database of all sequences, metadata, analysis results and visual analytics
Species
Relatedness
Resistance knowledge base
Virulance knowledge base
Local hospital clinical record system
National and international databases
All isolates

Nature Reviews Genetics 13, 601-612 (September 2012)

Nature Reviews | Genetics

UNIVERSITY OF OXFORD

Public Health England

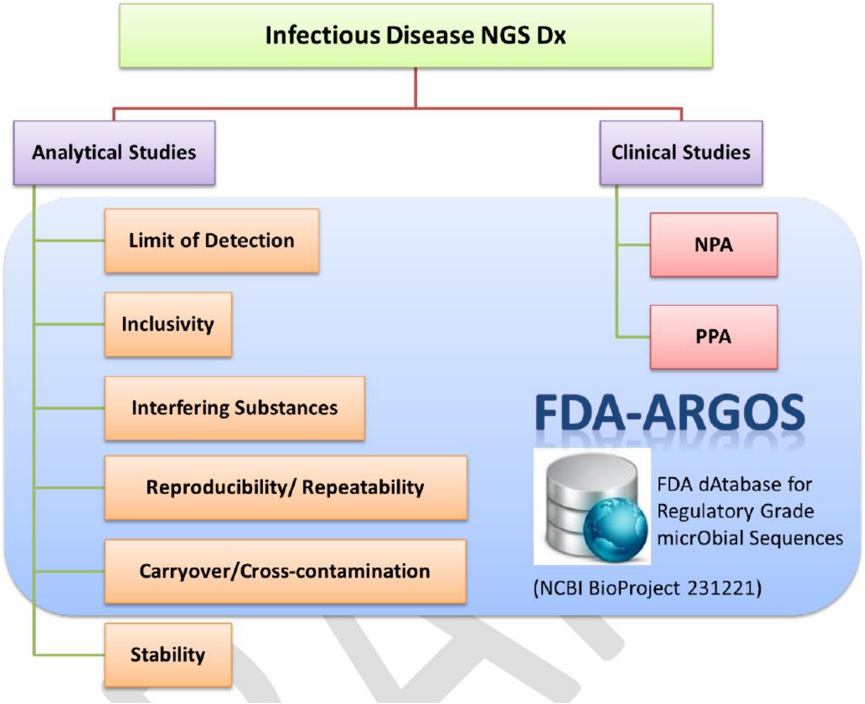
What are the challenges

- To go from research proof-of-principle to a fully accredited service
    - Systematic well validated method for extracting and purifying nucleic acids
    - Sequencing platform which is stable and produces reproducible results
    - Software for processing the data yielding:
        - Species identification
        - Feature prediction - curated knowledge bases
            - Resistance prediction
            - Pathotype
        - Transmission cluster identification
    - Linkage to epidemiological and clinical record data – data protection compliant
    - Software for reporting and presentation/visualisation of data
    - Persistent storage and sharing to benefit from a complete landscape within a species
    - Clinical validation
    - Accreditation

Seven pillars of wisdom needed if each pathogen

re-sequencing & re-calling
an individualat one timepoint
an individual over time
a point source outbreak
the community
PREDICTION (species, pathotype, susceptibility)
VALIDATION
DIVERSITY- Public Health
DIAGNOSTIC UTILITY

Regulatory approval e.g. FDA or UKAS



Infectious Disease NGS Dx

Analytical Studies

- Limit of Detection
- Inclusivity
- Interfering Substances
- Reproducibility/ Repeatability
- Carryover/Cross-contamination
- Stability

Clinical Studies

- NPA
- PPA

FDA-ARGOS

FDA dAtabase for Regulatory Grade micrObial Sequences

(NCBI BioProject 231221)

Will give 3 exemplars

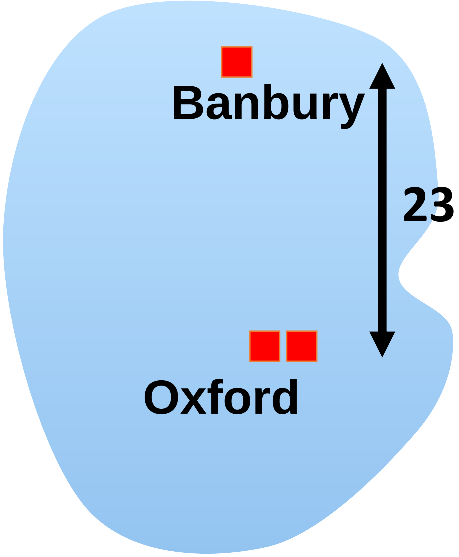
- *Clostridium difficile*
- Enterobacterial carbapenemase resistance
- Mycobacterium tuberculosis TB

Clostridium difficile

Role of symptomatic patients in C. *difficile* transmission

- We sequenced 1223 of all 1251 hospital and community CDI cases (98%) in Oxfordshire, September 2007 – March 2011

- Hospital admission and ward movement data, and home postcode district and GP location available for each case



- 3 Hospitals
  - Typical CDI incidence
  - Infection control in line with published guidelines

Eyre: N Engl J Med 2013; 369:1195-1205

Applying sequencing

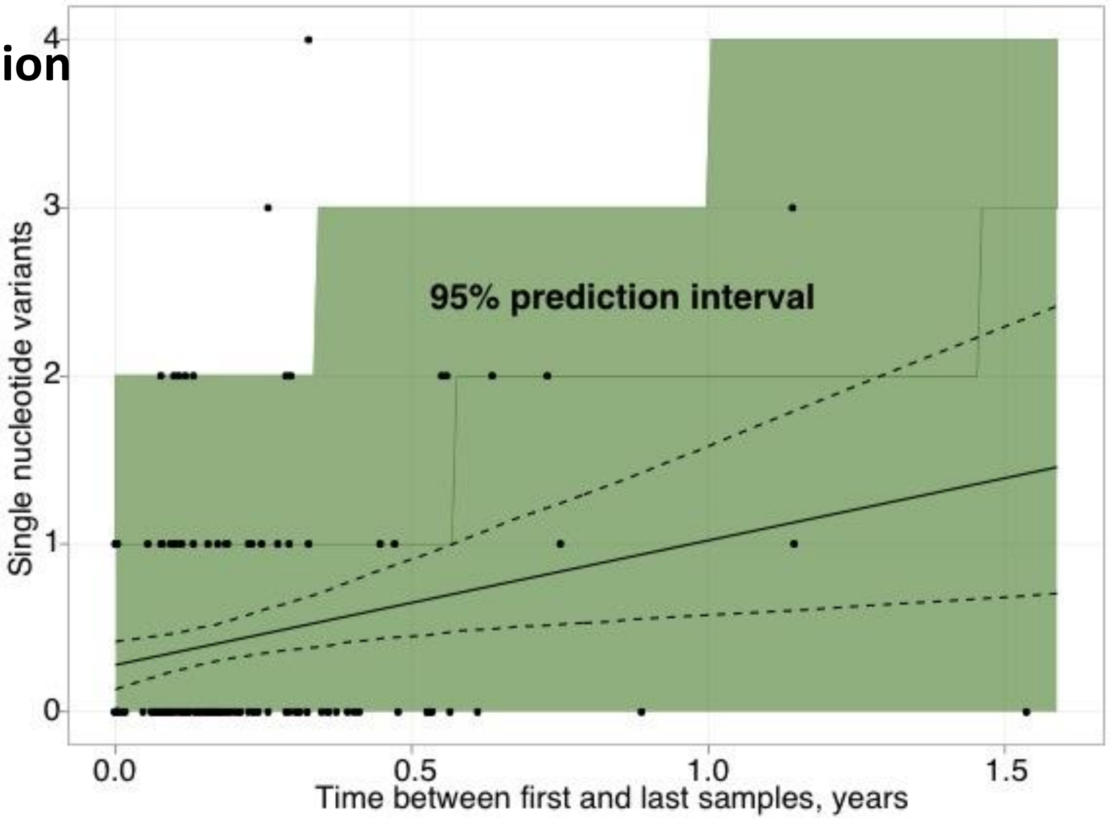
Reproducible sequencing

- 180 genomes sequenced more than once, 1 false SNV per 90 genomes

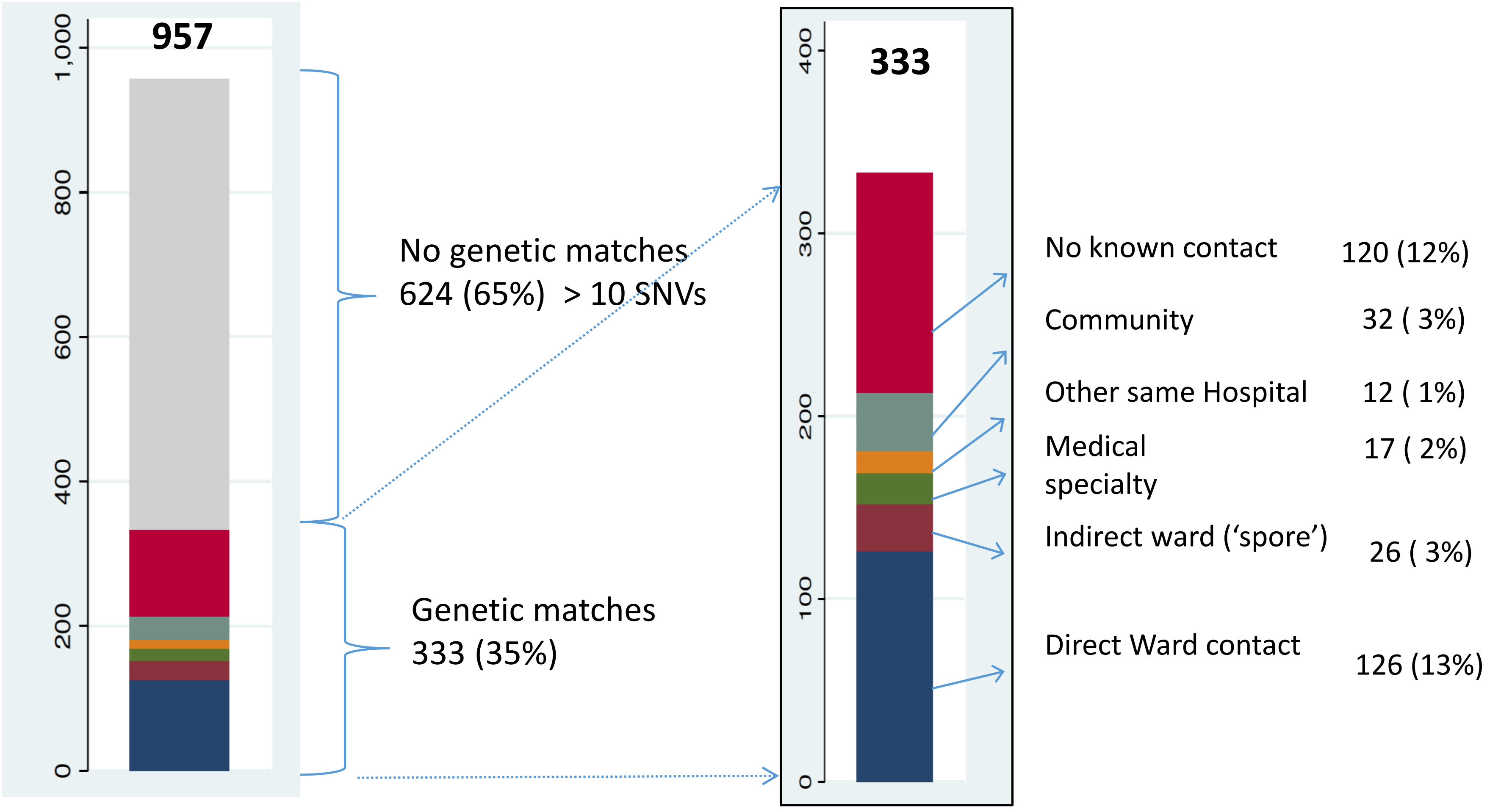
Within host diversity and evolution

- 0-2 SNVs expected between transmitted isolates up to 123 days apart

  - > 10 SNVs likely to be unrelated with a time to most recent common ancestor of ~ 5 years



Source of new *C. difficile* cases



No genetic matches
624 (65%) > 10 SNVs

Genetic matches
333 (35%)

No known contact	120 (12%)
Community	32 (3%)
Other same Hospital	12 (1%)
Medical specialty	17 (2%)
Indirect ward ('spore')	26 (3%)
Direct Ward contact	126 (13%)

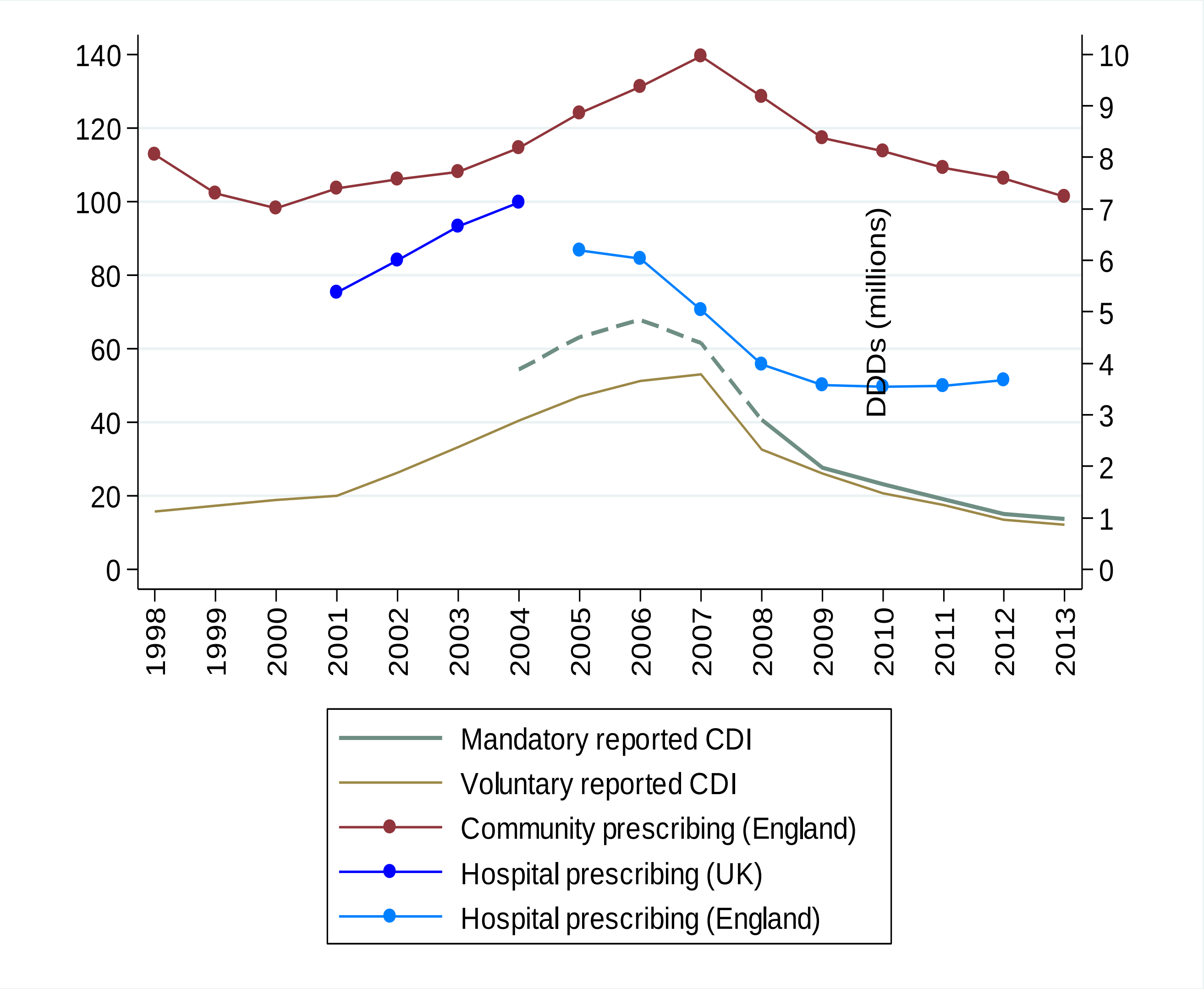
All cases

Genetic Matches (0-2 SNV)

N Engl J Med 2013; 369:1195-1205

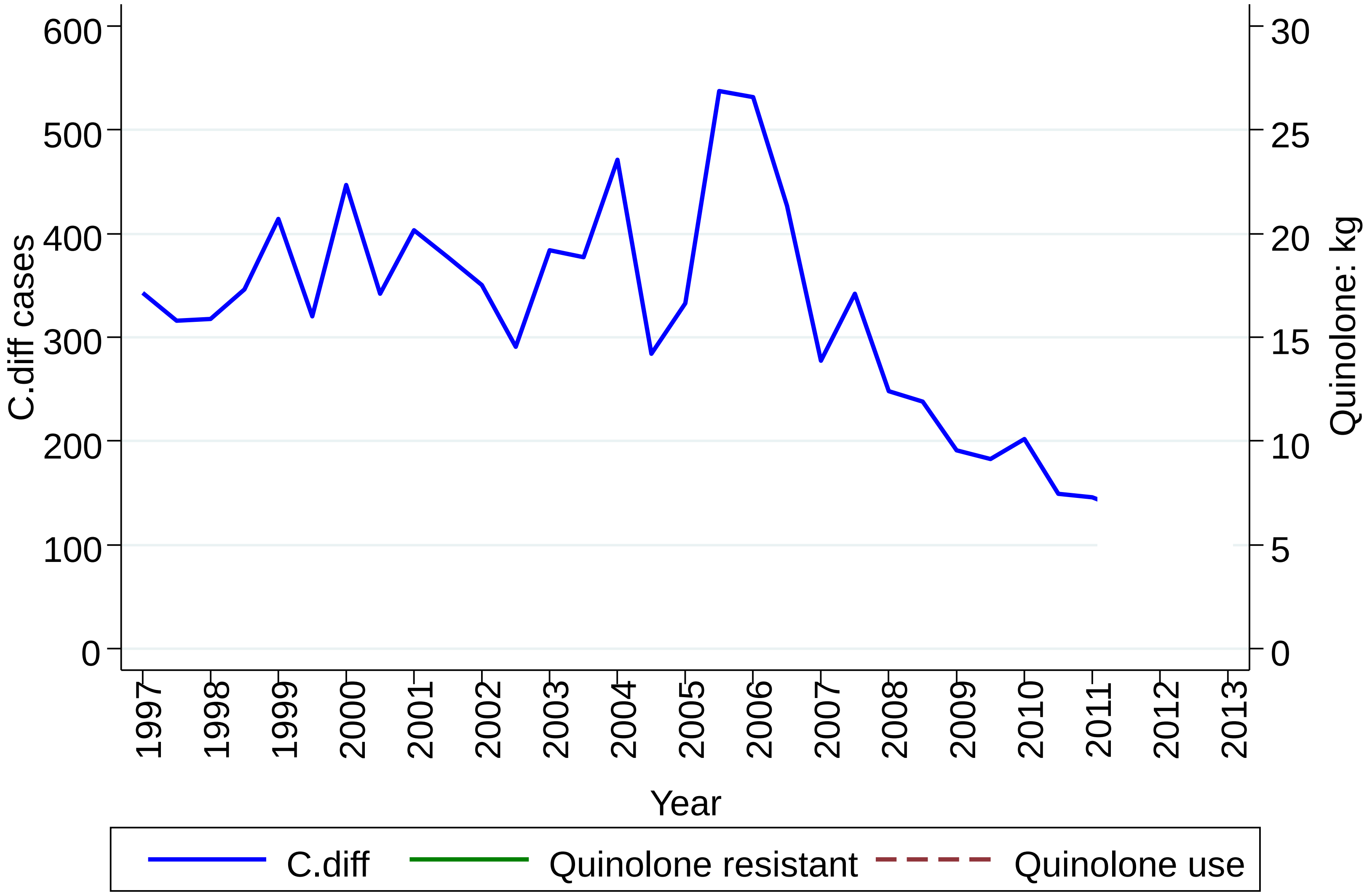
Selection, dispersal and control of C. difficile

Change in incidence and quinolone usage nationally

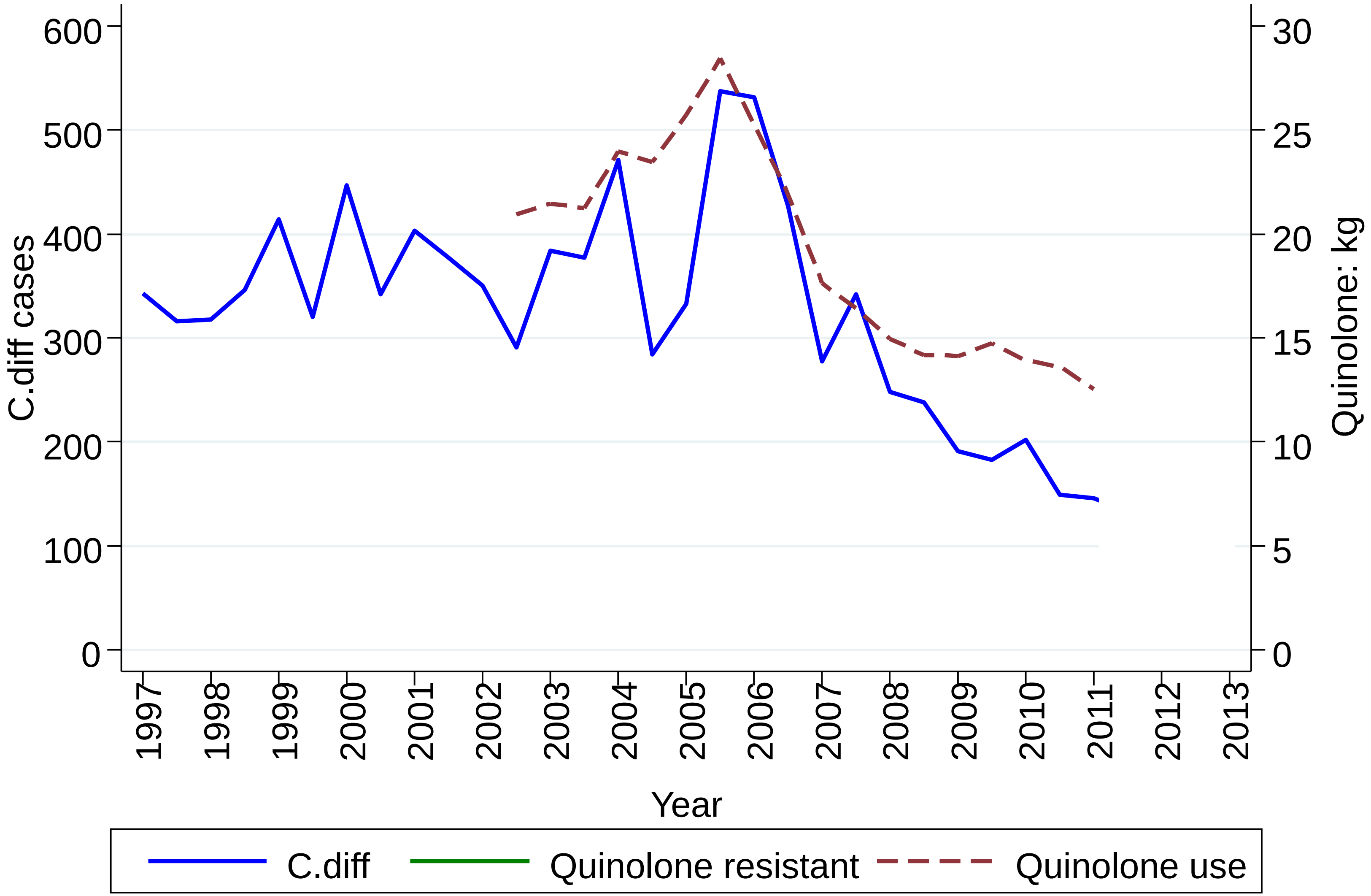


Dingle; Lancet Infect Dis 2017; 17: 411–21

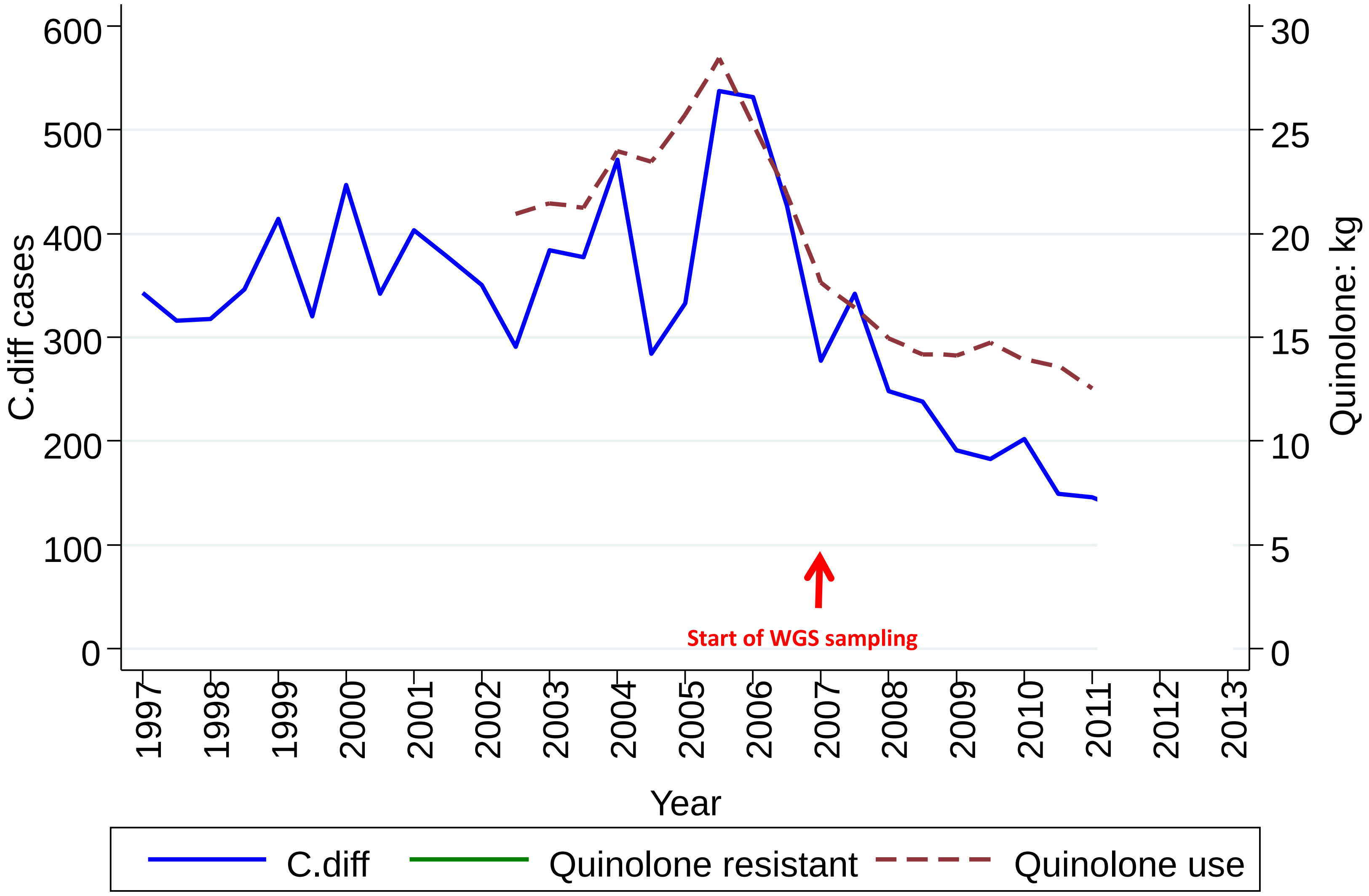
Oxfordshire *C. difficile* cases



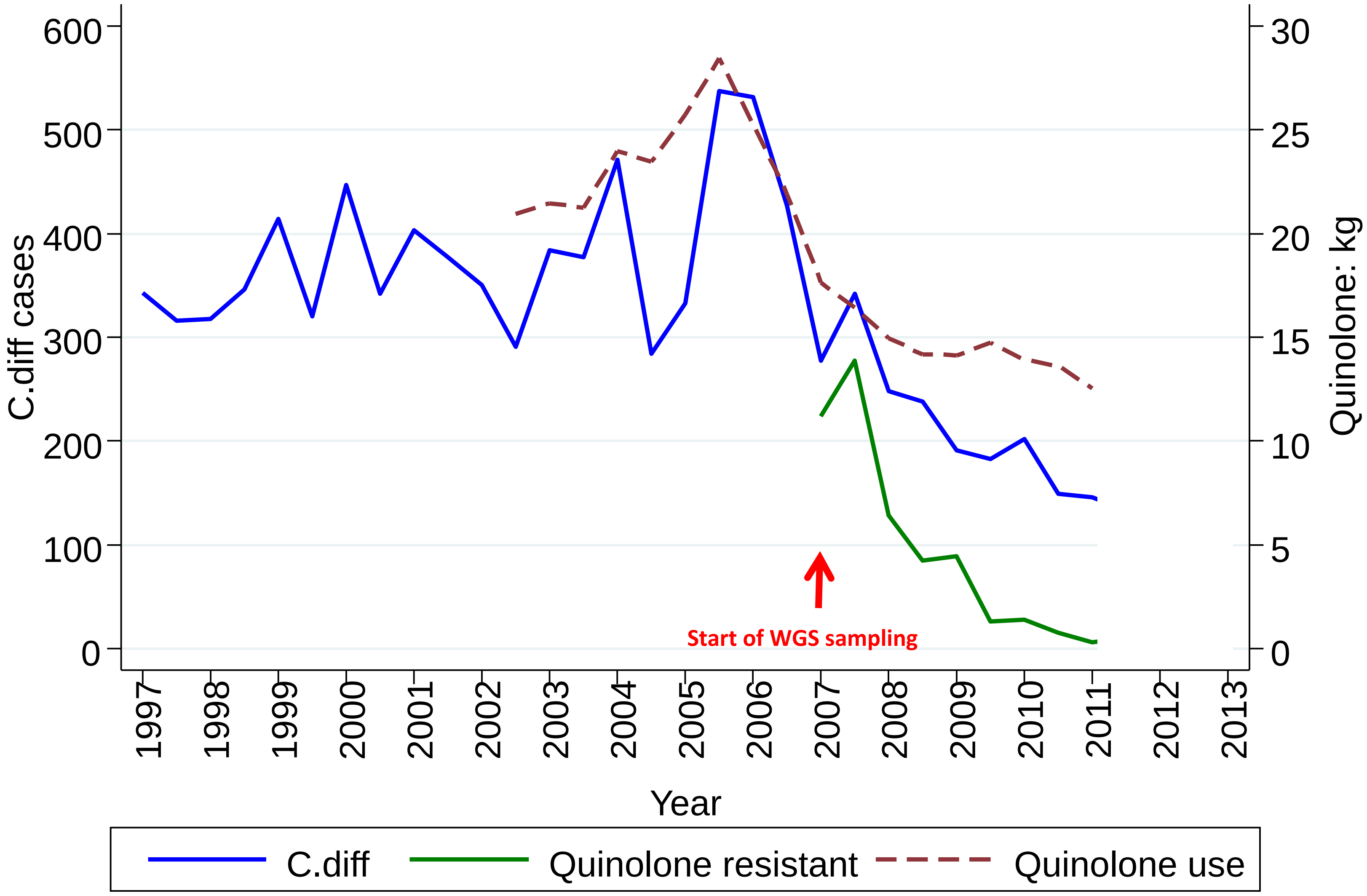
Oxfordshire *C. difficile* cases



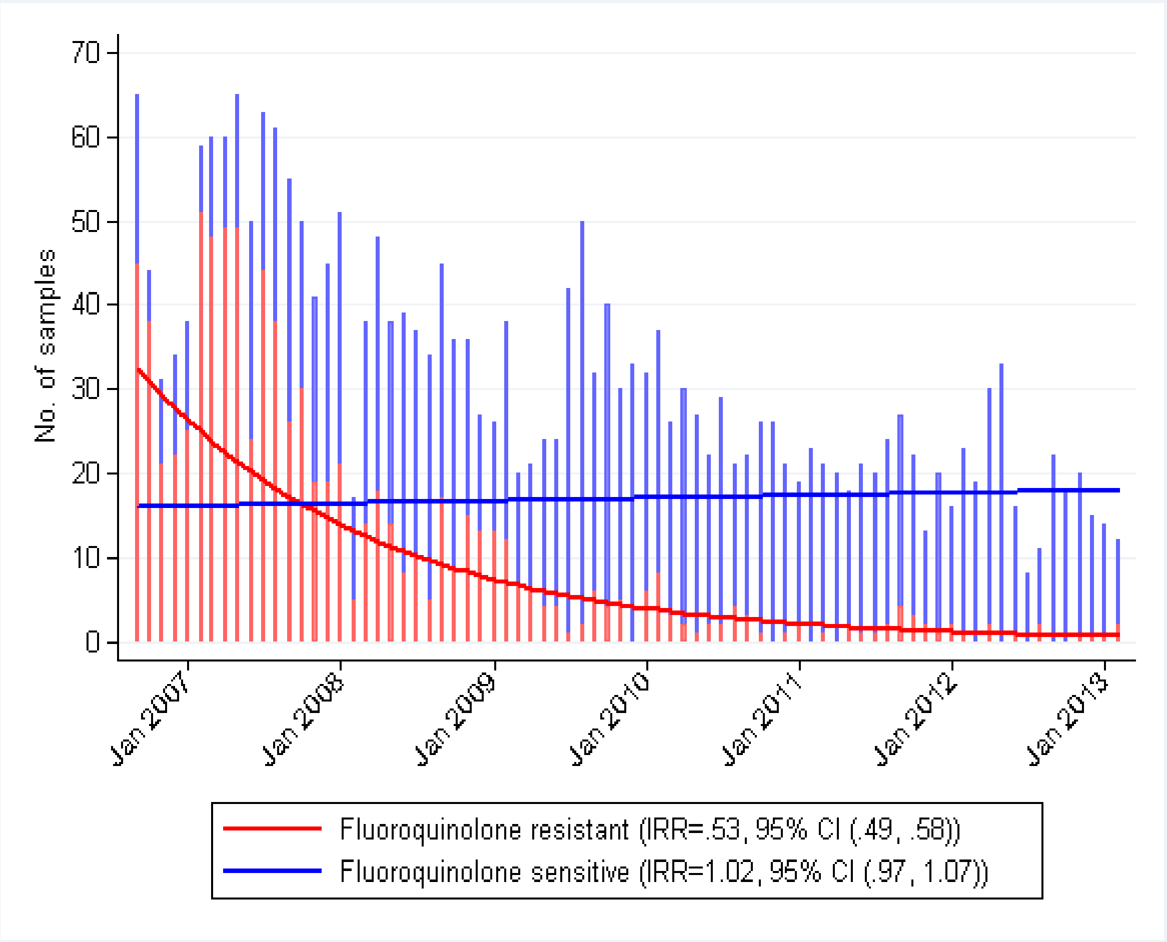
Oxfordshire *C. difficile* cases



Oxfordshire *C. difficile* cases



Fluoroquinolone resistant
Declining ^ CDI in Oxford

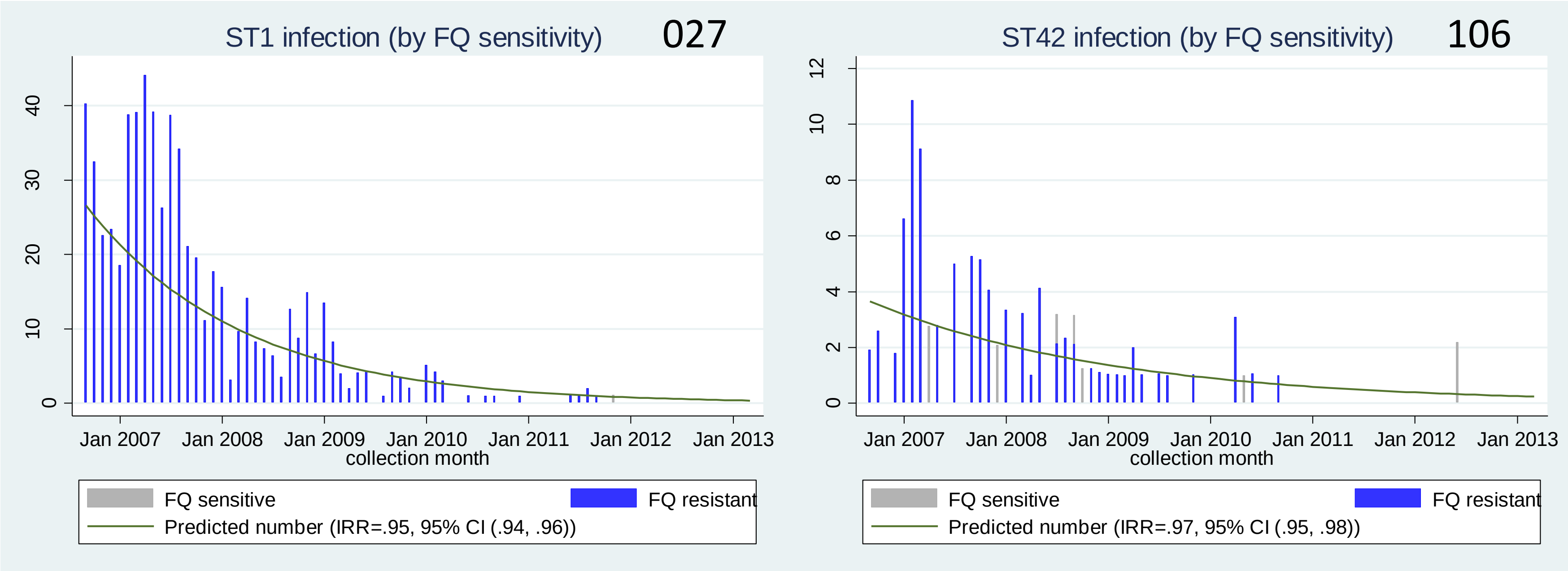


Dingle; Lancet Infect Dis 2017; 17: 411–21

UNIVERSITY OF OXFORD

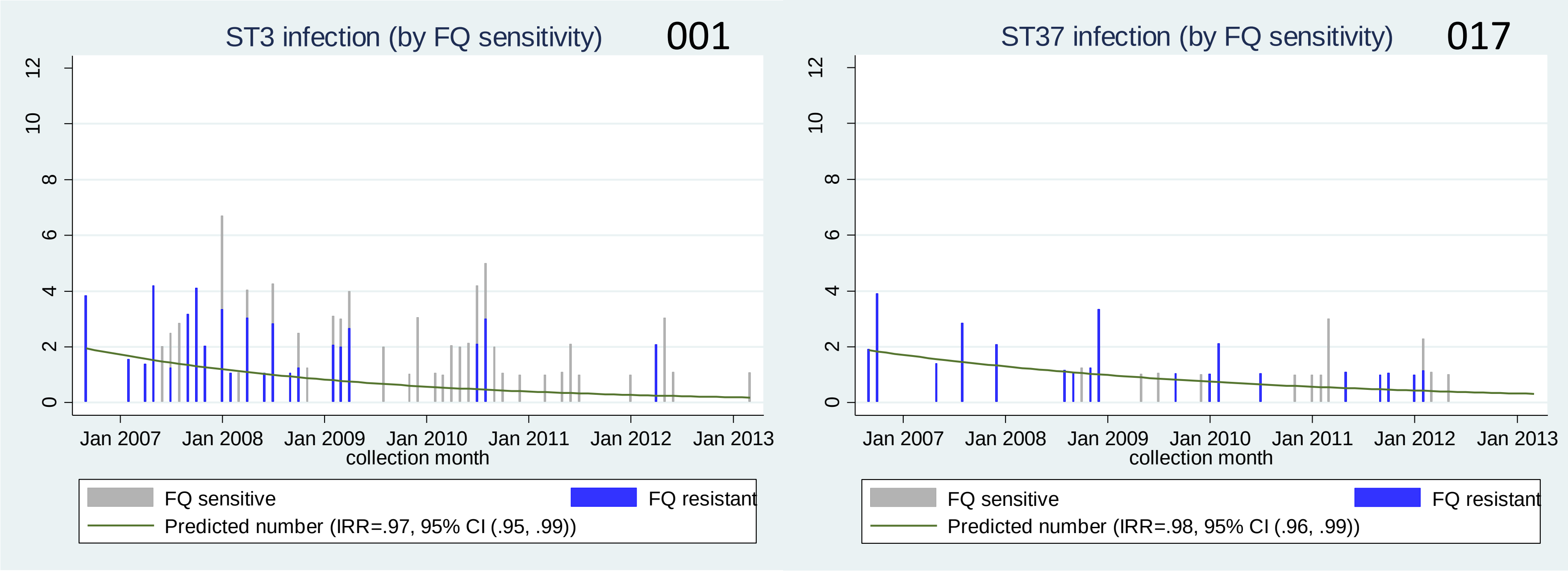
Public Health England

Incidence of FQ resistant genotypes has declined (1)



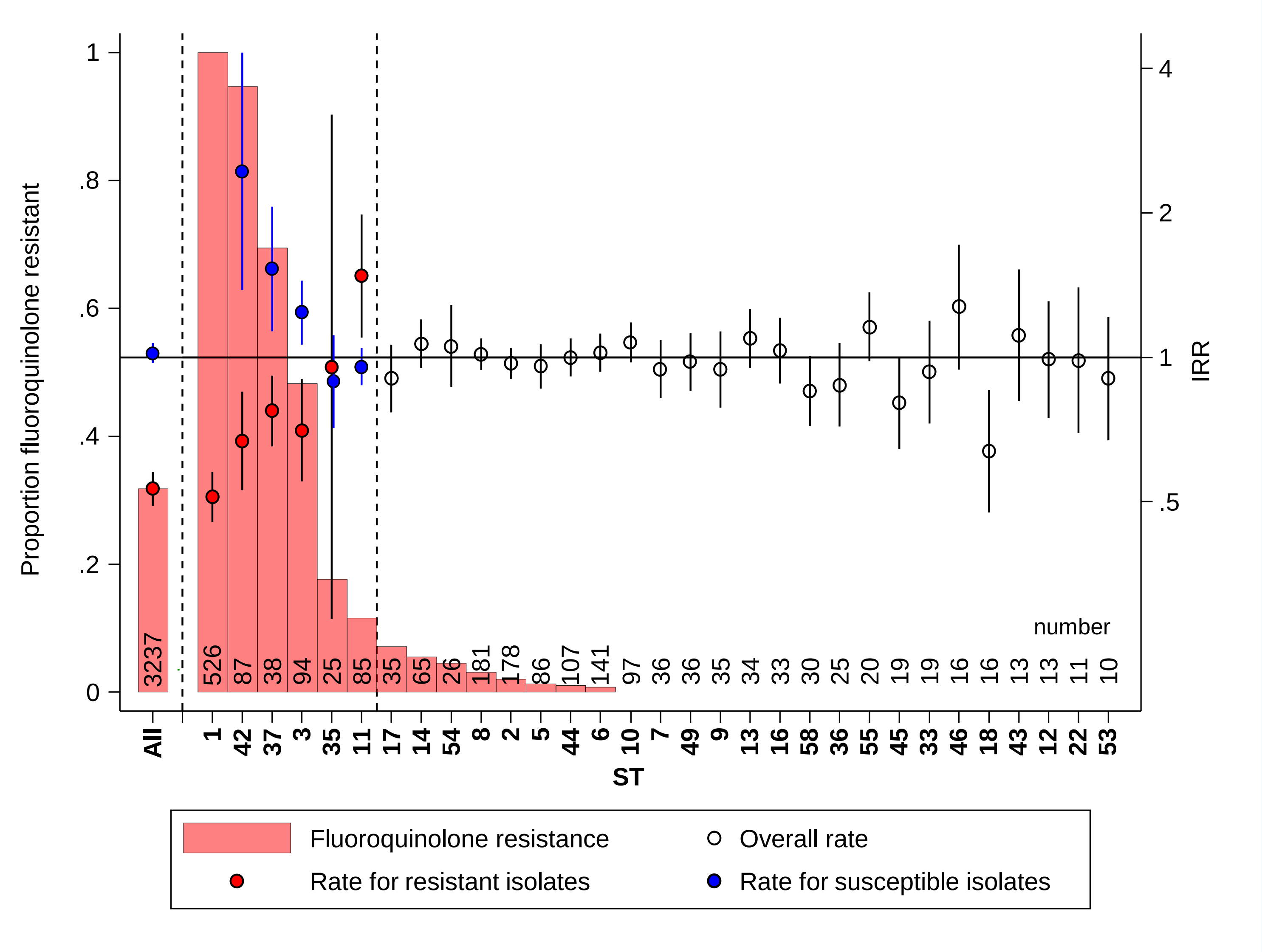
Green line: number of cases (per month) predicted by a Poisson model, (with time as the only covariate), modelling FQ resistant cases (blue) to illustrate declining incidence.

Incidence of FQ resistant genotypes has declined (2)



Green line: number of cases (per month) predicted by a Poisson model, (with time as the only covariate), modelling FQ resistant cases (blue) to illustrate declining incidence.

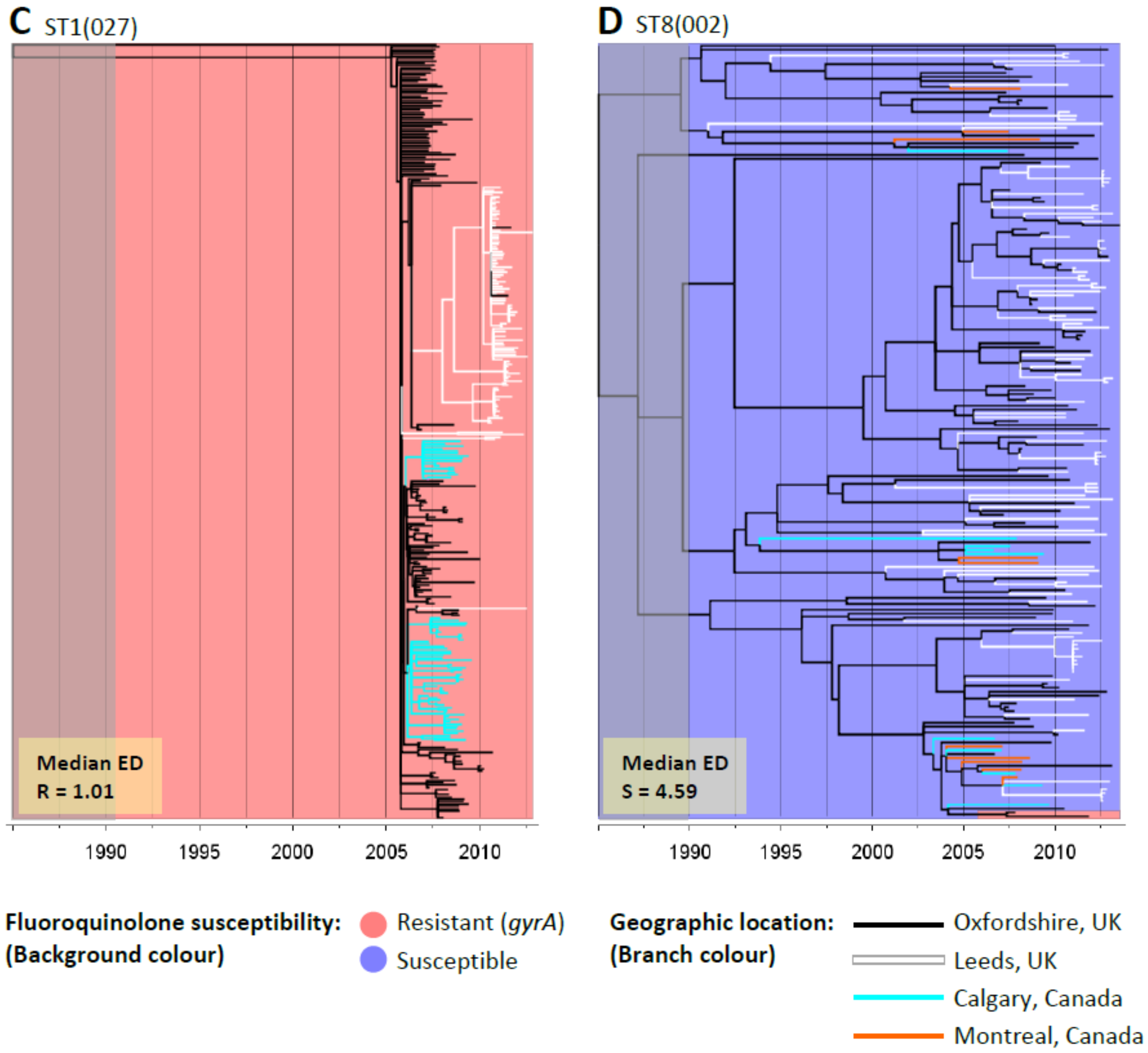
Changes in quinolone resistance over time



UNIVERSITY OF OXFORD

Public Health England

Phylogenetic patterns of quinolone resistant vs susceptible



Dingle; Lancet Infect Dis 2017; 17: 411–21

The decline of *C. difficile* in England

- It has declined by close to 70% since 2006
- Quinolone use declined by ~ 50% preceding the decline in CDI
- The decline is attributable to the simultaneous disappearance of 4 quinolone resistant lineages. The remaining 69 lineages are largely unchanged in incidence
- Resistant lineages had undergone rapid clonal expansion and were geographically structured
- A quinolone effect is a likely explanation for the decline in CDI

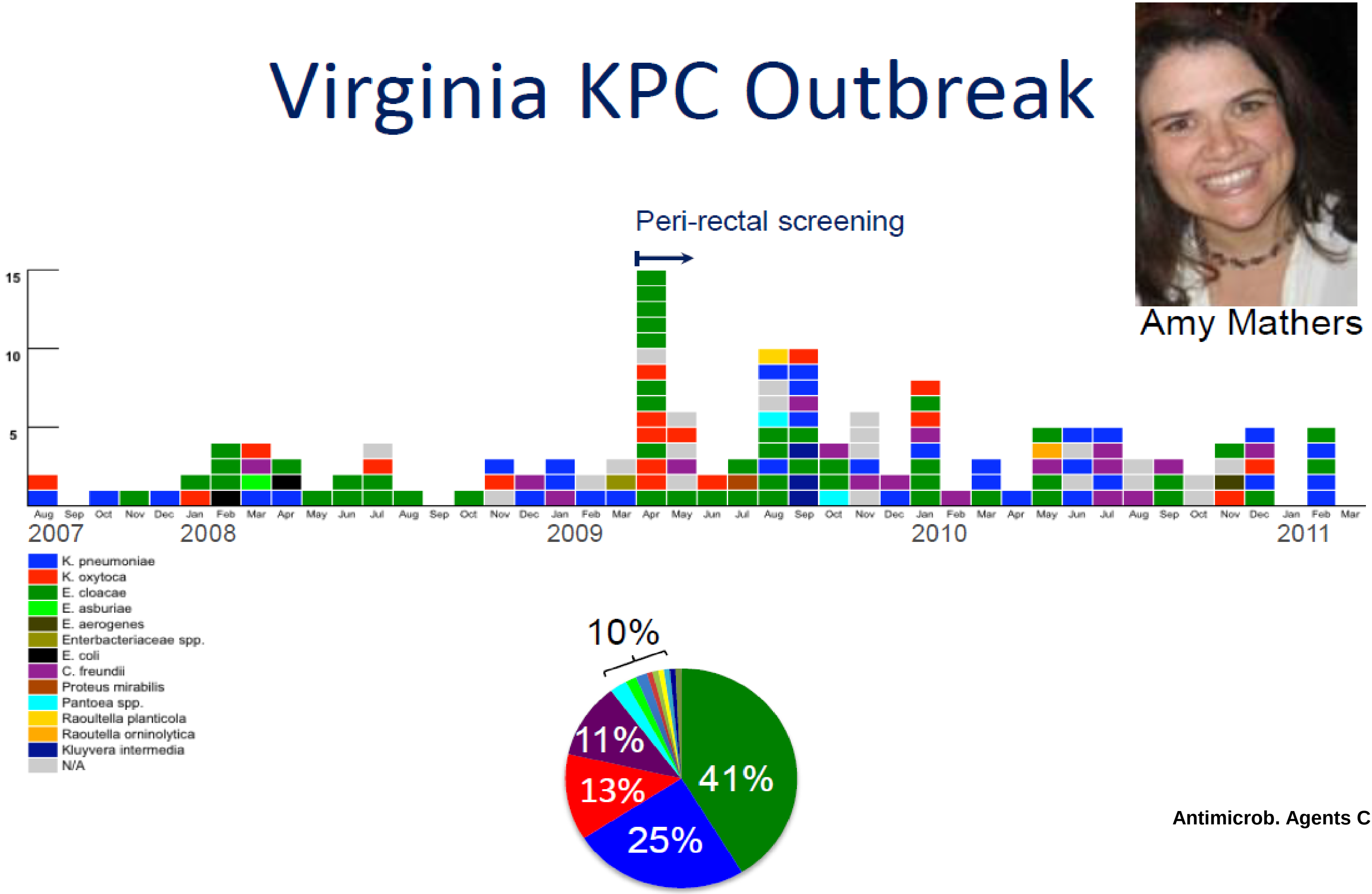
Carbapenemase resistance in Enterobacteriacea

UNIVERSITY OF OXFORD

Public Health England

A single hospital

Virginia KPC Outbreak



Antimicrob. Agents Chemother; April 2016

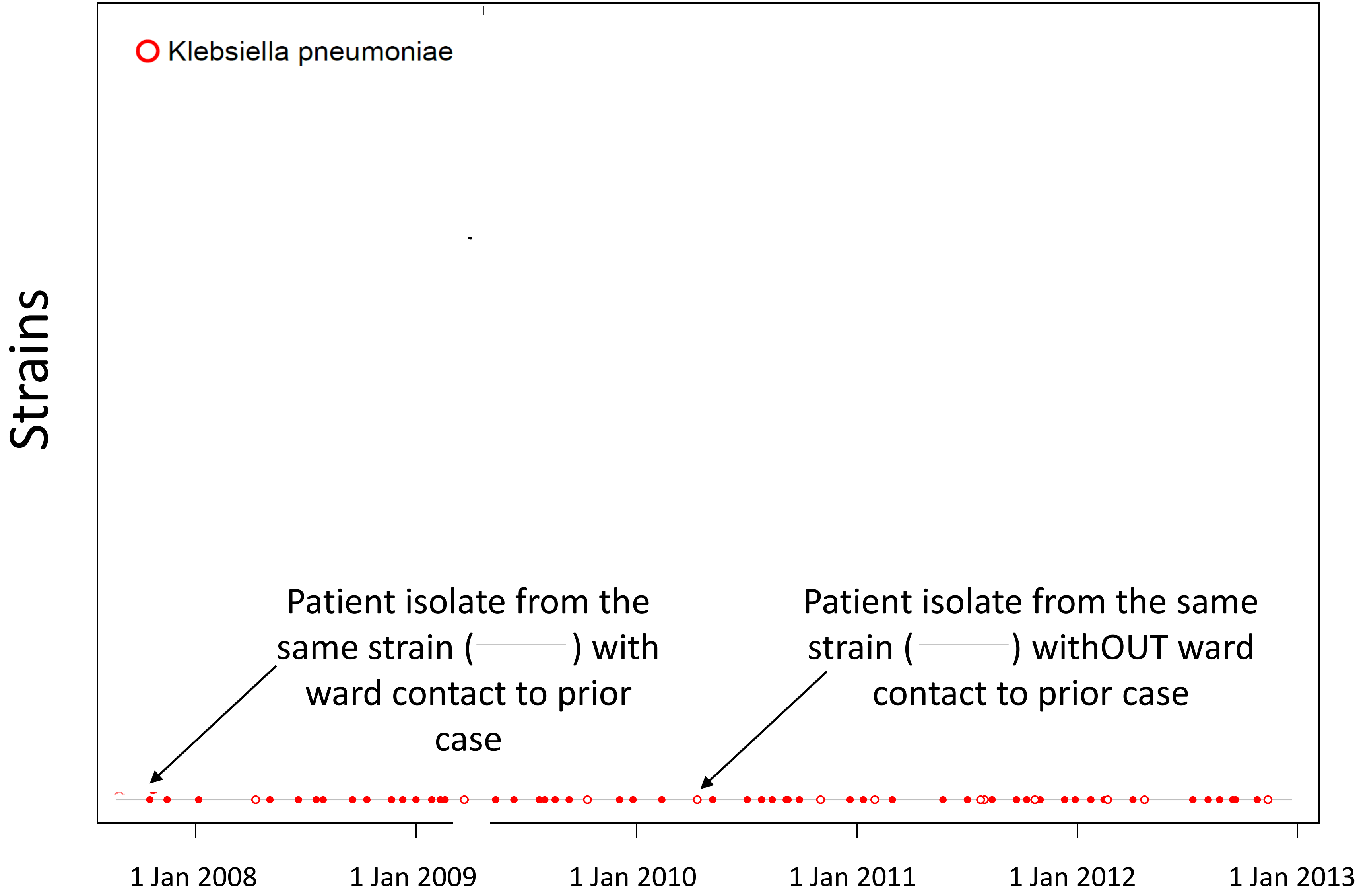
bla$_{\text{KPC}}$ in Virginia

- Virginia “outbreak” – ongoing since August 2007

- 281 *bla*$_{\text{KPC}}$-positive Enterobacteriaceae
  - Isolated August 2007 – December 2012
  - From 182 patients
  - All Illumina sequenced

- Multiple species of *bla*$_{\text{KPC}}$-positive Enterobacteriaceae
  - 9 different genera
  - 13 different species
  - 62 different “strains”
    (defined conservatively as ~500 SNPs variation in “core”)

Idealised outbreak timeline – what we'd like to see



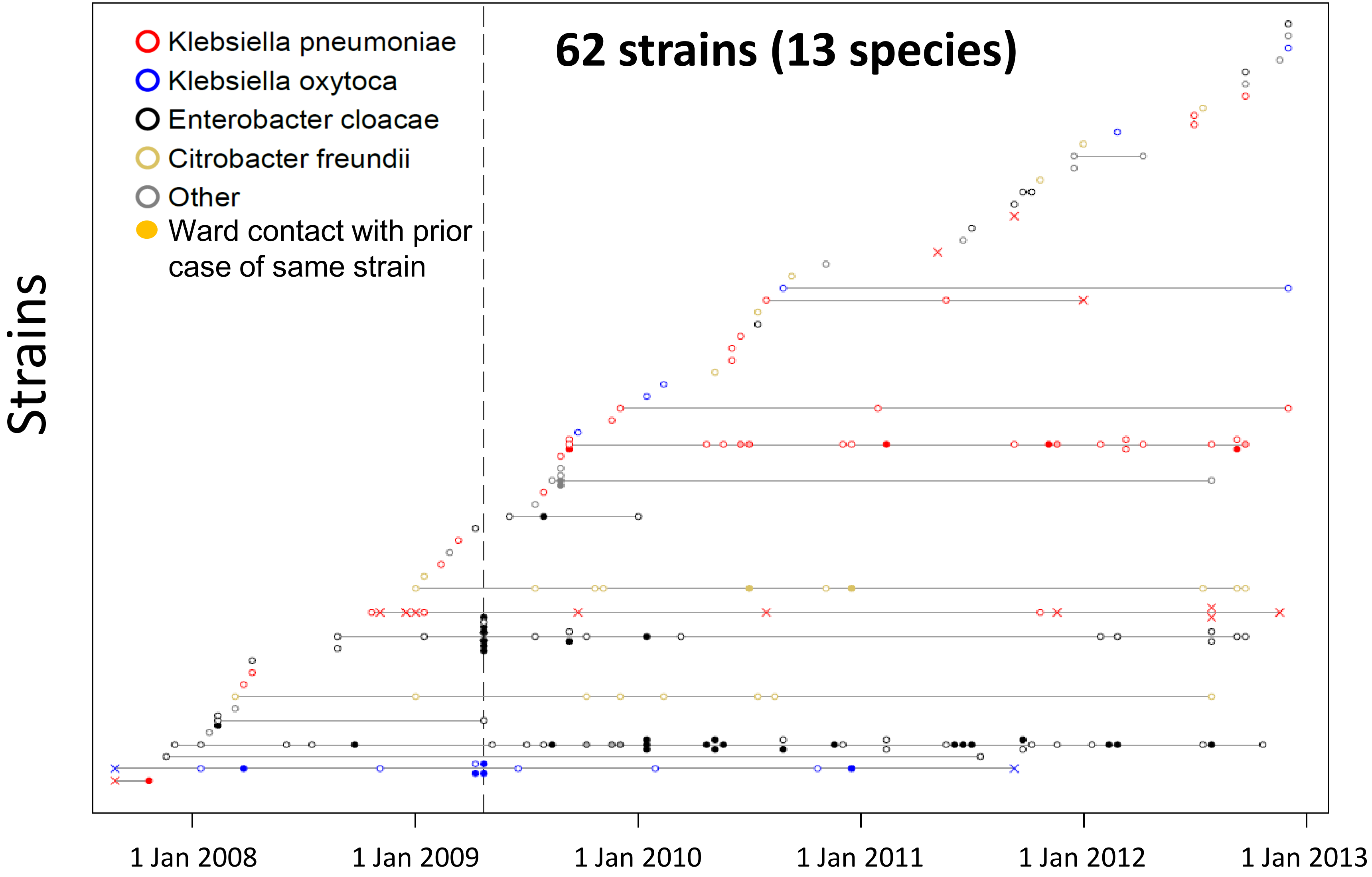
Patient isolate from the same strain (—) with ward contact to prior case

Patient isolate from the same strain (—) withOUT ward contact to prior case

UNIVERSITY OF OXFORD

Public Health England

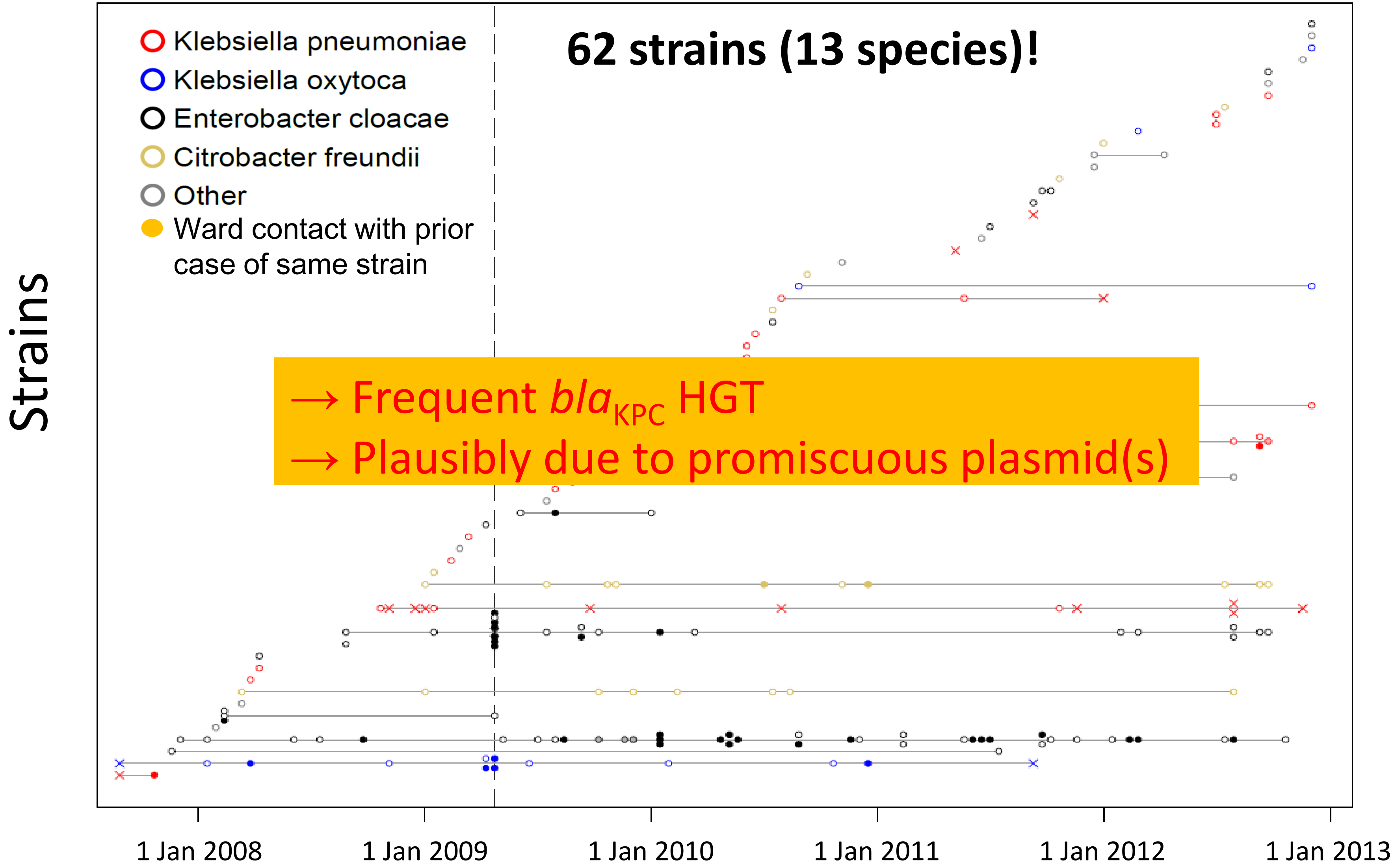
What did we see - enormous host strain diversity



UNIVERSITY OF OXFORD

Public Health England

Enormous host strain diversity



62 strains (13 species)!

→ Frequent *bla*$_{KPC}$ HGT

→ Plausibly due to promiscuous plasmid(s)

Plasmid-mediated outbreak?

- Hypothesis: outbreak is driven by one or a few promiscuous plasmids carrying *bla*$_{\text{KPC}}$
- Assumption: plasmid structures relatively stable within outbreak
- Approach:
  - Generate outbreak-specific plasmid references (index patient)
  - Use these to assess plasmid presence across outbreak isolates
  - **Definition: ≥99% sequence identity over ≥80% reference length**
    - Assessed via BLASTn (reference plasmid vs isolate's *de novo* assembly)
    - Stringent identity threshold: expect few SNP changes
    - Lenient length threshold: single events can affect large regions

    - Note: does not assess structural continuity (since this is impossible in many isolates due to repeat structures)

Spread of index plasmids

- Two *bla*$_{\mathrm{KPC}}$ conjugative plasmids from index patient
    - pKPC_UVA01 (43,621 bp) and pKPC_UVA02 (113,105 bp)

Spread of index plasmids

- Two *bla*$_{\text{KPC}}$ conjugative plasmids from index patient
    - pKPC_UVA01 (43,621 bp) and pKPC_UVA02 (113,105 bp)

Species	Isolates
Citrobacter amalonaticus	2
Citrobacter freundii	30
Enterobacter aerogenes	4
Enterobacter asburiae	1
Enterobacter cloacae	96
Escherichia coli	2
Klebsiella oxytoca	35
Klebsiella pneumoniae	94
Kluyvera intermedia	7
Proteus mirabilis	1
Raoultella ornothinolytica	1
Serratia marcescens	5
Other (unknown)	3
Total	**281**

Spread of index plasmids

- Two *bla*$_{\text{KPC}}$ conjugative plasmids from index patient
  - pKPC_UVA01 (43,621 bp) and pKPC_UVA02 (113,105 bp)

Species	Isolates	pKPC_UVA01
Citrobacter amalonaticus	2	1
Citrobacter freundii	30	29
Enterobacter aerogenes	4	2
Enterobacter asburiae	1	0
Enterobacter cloacae	96	84
Escherichia coli	2	1
Klebsiella oxytoca	35	9
Klebsiella pneumoniae	94	31
Kluyvera intermedia	7	7
Proteus mirabilis	1	1
Raoultella ornothinolytica	1	1
Serratia marcescens	5	0
Other (unknown)	3	0
Total	**281**	**166 (59%)**

Spread of index plasmids

- Two *bla*$_{\text{KPC}}$ conjugative plasmids from index patient
  - pKPC_UVA01 (43,621 bp) and pKPC_UVA02 (113,105 bp)

Species	Isolates	pKPC_UVA01	pKPC_UVA02
Citrobacter amalonaticus	2	1	0
Citrobacter freundii	30	29	7
Enterobacter aerogenes	4	2	0
Enterobacter asburiae	1	0	0
Enterobacter cloacae	96	84	2
Escherichia coli	2	1	0
Klebsiella oxytoca	35	9	25
Klebsiella pneumoniae	94	31	18
Kluyvera intermedia	7	7	0
Proteus mirabilis	1	1	0
Raoultella ornothinolytica	1	1	0
Serratia marcescens	5	0	0
Other (unknown)	3	0	0
Total	**281**	**166 (59%)**	**52 (19%)**

Spread of index plasmids

- Two *bla*$_{\text{KPC}}$ conjugative plasmids from index patient
  - pKPC_UVA01 (43,621 bp) and pKPC_UVA02 (113,105 bp)

Species	Isolates	pKPC_UVA01	pKPC_UVA02	Neither
Citrobacter amalonaticus	2	1	0	1
Citrobacter freundii	30	29	7	1 (3%)
Enterobacter aerogenes	4	2	0	2
Enterobacter asburiae	1	0	0	1
Enterobacter cloacae	96	84	2	10 (10%)
Escherichia coli	2	1	0	1
Klebsiella oxytoca	35	9	25	1 (3%)
Klebsiella pneumoniae	94	31	18	45 (48%)
Kluyvera intermedia	7	7	0	0
Proteus mirabilis	1	1	0	0
Raoultella ornothinolytica	1	1	0	0
Serratia marcescens	5	0	0	5
Other (unknown)	3	0	0	3
Total	**281**	**166 (59%)**	**52 (19%)**	**70 (25%)**

← mostly known endemic clone previously described with other plasmids

→ Consistent with local plasmid-mediated outbreak, plus occasional imports from other healthcare institutions

UNIVERSITY OF OXFORD

Public Health England

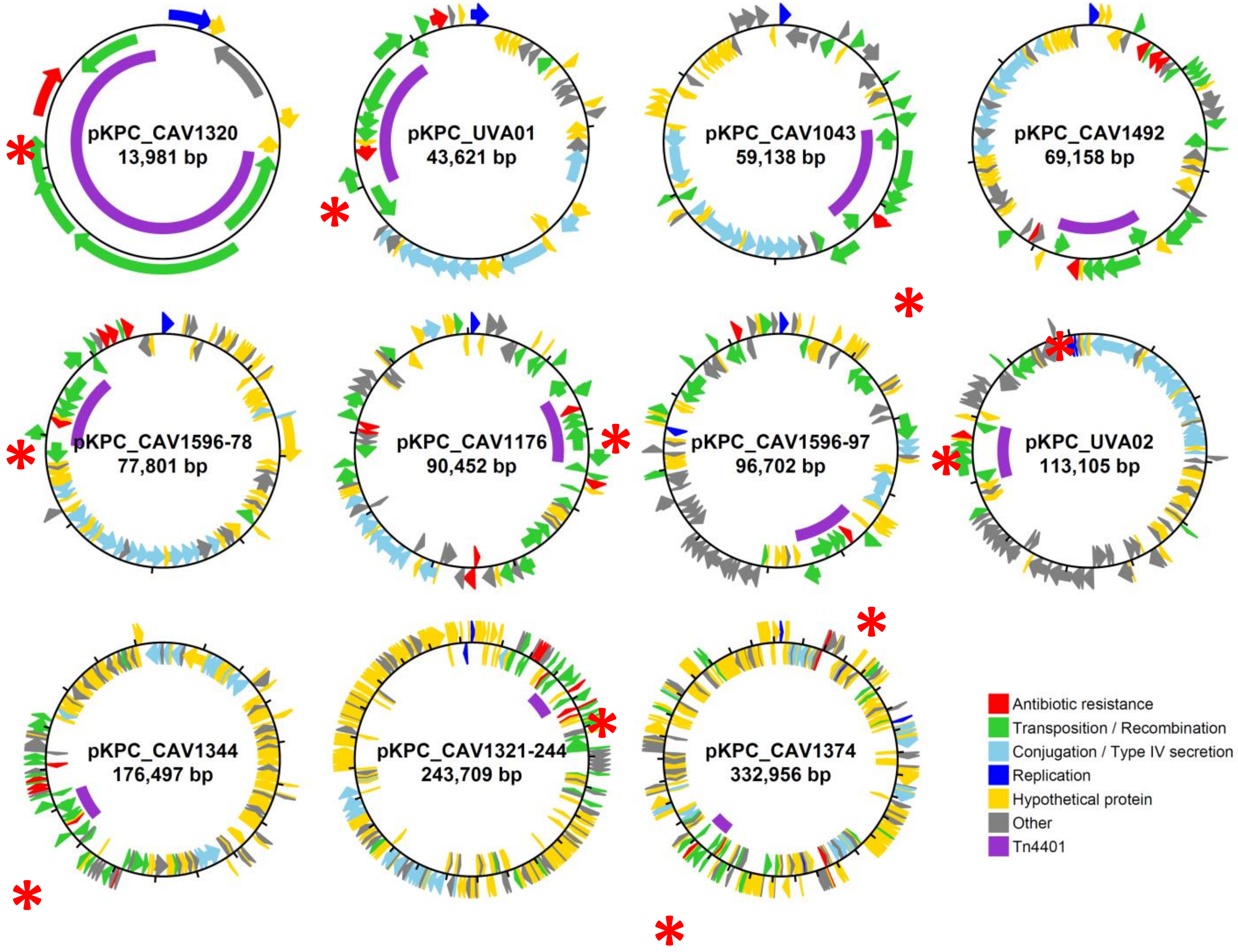
Long-read sequencing

- Needed to validate conclusions, given structural uncertainties of short-read WGS

- PacBio sequencing
  - 17 **randomly chosen** isolates
  - Fully closed plasmid structures

11 different *bla*$_{\text{KPC}}$ (*) plasmids among 80!

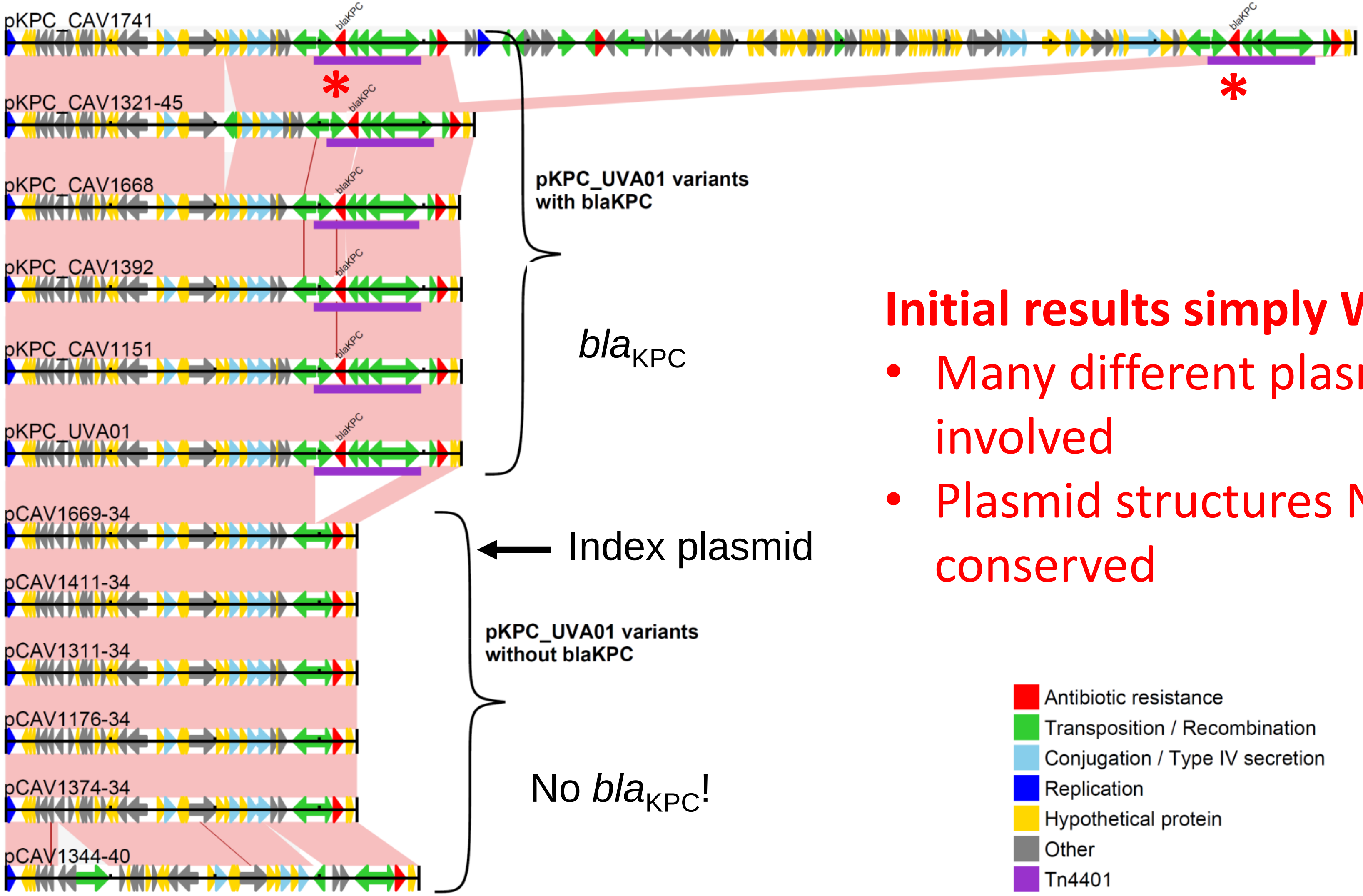
14kb
to
330kb



UNIVERSITY OF OXFORD

Public Health England

Structural diversity of pKPC_UVA01



pKPC_UVA01 variants with blaKPC

bla$_{\mathrm{KPC}}$

Index plasmid

pKPC_UVA01 variants without blaKPC

No *bla*$_{\mathrm{KPC}}$!

Initial results simply WRONG:

- Many different plasmids involved
- Plasmid structures NOT conserved

🟥 Antibiotic resistance
🟩 Transposition / Recombination
🟦 Conjugation / Type IV secretion
🟦 Replication
🟨 Hypothetical protein
⬛ Other
🟪 Tn4401

UNIVERSITY OF OXFORD

Public Health England

A highly dynamic dispersal of KPC within the clinical ecosystem

- KPC dispersing at 3 scales:
  - Isolates spreading KPC between patients
  - Frequent transfer of $\textit{bla}_{\text{KPC}}$ plasmids between strains/species
  - Frequent transfer of $\textit{bla}_{\text{KPC}}$ transposon Tn4401 between plasmids

- Where's the reservoir?

UVa sink study

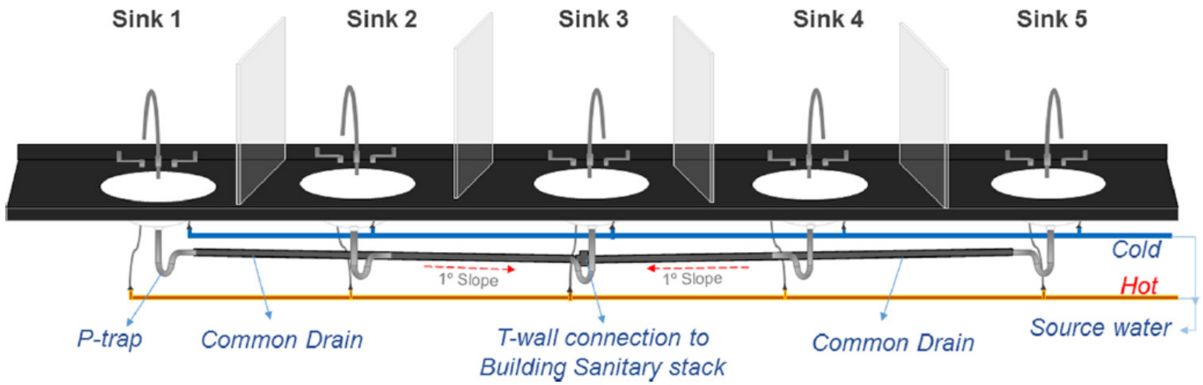


FIG 4 Layout of the sink gallery comprising the 5 sink modules and the associated plumbing.

CPE E. coli were found in > 10 CFU/CM^3 in the basins

Applied and Environmental Microbiology April 2017 Volume 83 Issue 8 e03327-16

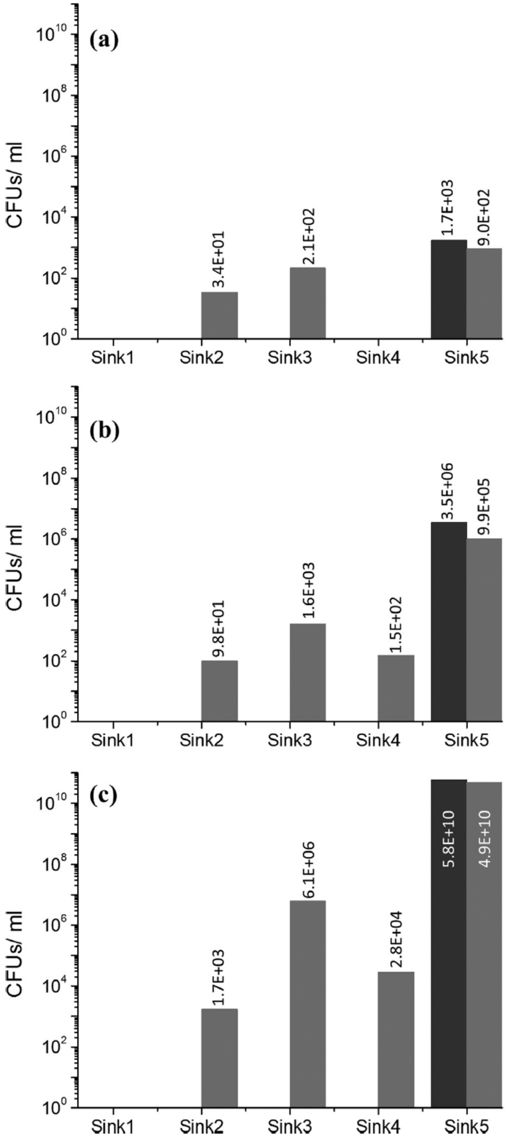
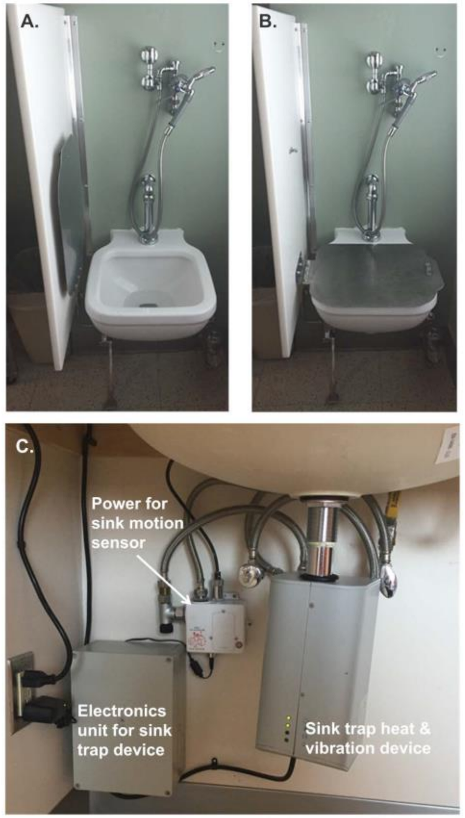


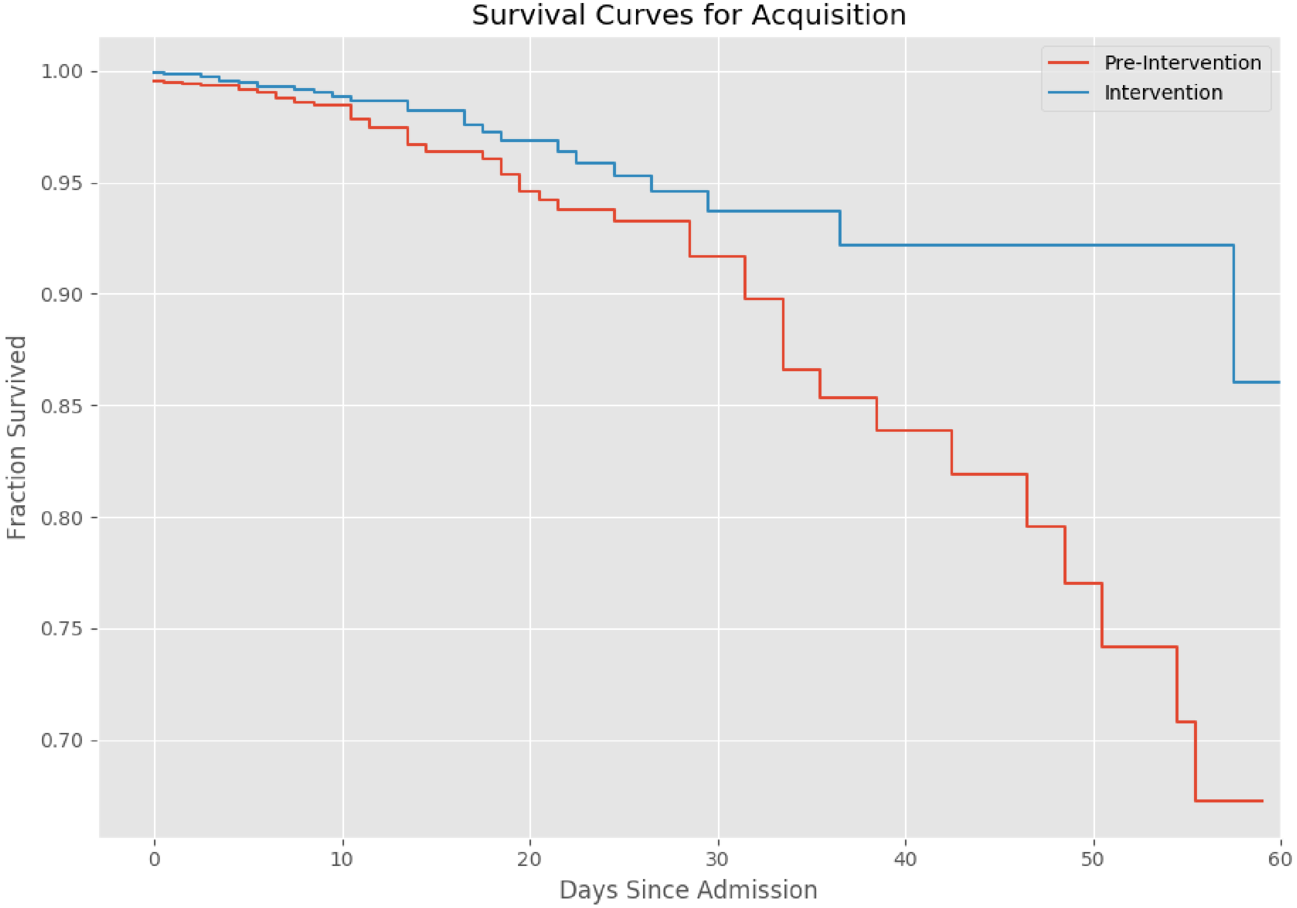
FIG 1 GFP-expressing *E. coli* detected in the P-traps attached to each of the sinks on day 0 (black bars) and day 7 (gray bars) using (a) 10^3, (b) 10^6, and (c) 10^{10} CFU/ml as the starting inoculum concentrations in sink 5.

UNIVERSITY OF OXFORD

Public Health England

University of Virginia Hospital intervention


A.
B.
C.
Power for sink motion sensor
Electronics unit for sink trap device
Sink trap heat & vibration device



Clinical of Infectious Diseases, epublished Feb 2018

Mycobacteria

- Use this as the example of how to implement a WGS solution into **clinical** and **public health** practice
- Give a sense of what the future holds?

The TB problem

- It is a leading infectious disease world-wide
  - In 2014, 1.5 m died; 9.6 m developed TB; 0.5m MDR-TB, and **1/3 undiagnosed**
- Case detection is relatively poor
  - Full microbiological diagnosis is complex, error prone and slow
- Spread is mostly person-to-person with a small zoonotic reservoir
- Can be effectively treated
  - Most treatment is initially empiric; prolonged, and can produce drug resistance
- Can be prevented and even eliminated?
  - Better diagnosis seen as an imperative e.g Cepheid GeneXpert tb/rif

What we can deliver with WGS?

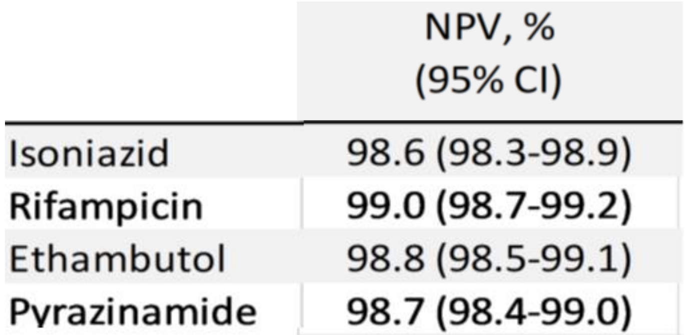
- Developed a MGIT dependent workflow and a software yielding the following:
  - Increasingly fast, cheap and accurate outputs that can be stored and shared **Lancet Respir Med**. 2016 Jan;4(1):49-58**; J Clin Microbiol**. 2018 Jan 24;56(2).
  - Accurate species identification **Lancet Respir Med**. 2016 Jan;4(1):49-58**; J Clin Microbiol**. 2018 Jan 24;56(2).
  - Resistance prediction **Lancet Infect Dis** 2015;15: 1193–1202; **Lancet Respir Med**. 2016 Jan;4(1):49-58; J Clin Microbiol. 2018 Jan 24;56(2).
  - Outbreak detection **Lancet Infect Dis** 2015;15: 1193–1202; **Wyllie. under review**
  - Linkage to pathogen phenotype and patient epidemiological/clinical record data yielding information for treating patients and directing outbreak investigation In pilot deployment.

Full national implementation in England

- Sequencing approximately 30,000 samples/year

- DST will be stopped when predicting susceptibility to the 4 first line drugs
  - Based on:

Analysis of 10,000 isolates from across the world



	NPV, % (95% CI)
Isoniazid	98.6 (98.3-98.9)
Rifampicin	99.0 (98.7-99.2)
Ethambutol	98.8 (98.5-99.1)
Pyrazinamide	98.7 (98.4-99.0)

Diagnostically there is < 2% chance the isolate will be falsely resistant

Where are the gaps?

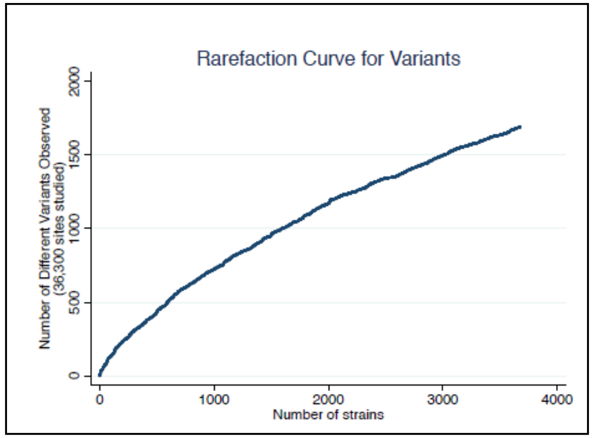
- We need:
  - a comprehensive knowledge base of genomic variants conferring resistance
  - a faster sequencer
  - faster software
  - to process direct from a sample and be equivalent/better than genexpert

Anti-tuberculosis drug resistance prediction

- Arguably 15 drugs are available for treating TB with more new drugs in development
- Is genomic variation which confers resistance limited to somewhere between 20 to 30 genes?
- Current knowledge indicates molecular prediction of INH, rifampicin resistant or pan-susceptible isolates is ~ 95% accurate
- The knowledge base of variation conferring resistance to ‘all drugs’ is incomplete

Filling the resistance gap

Comprehensive Resistance Prediction for Tuberculosis: an International Consortium (CRyPTIC)



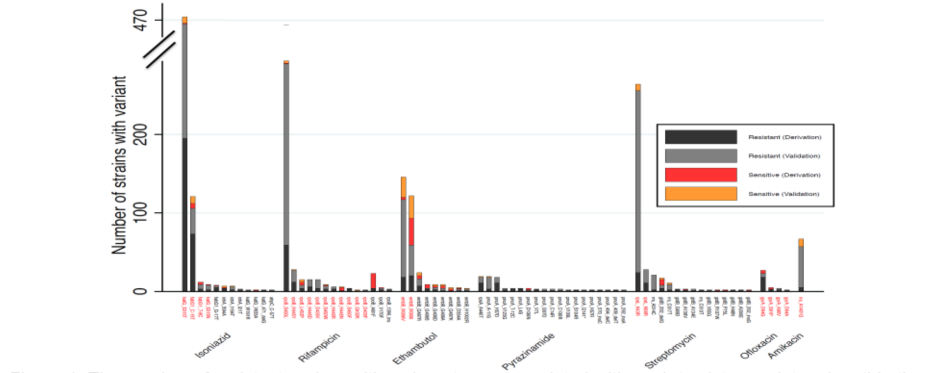
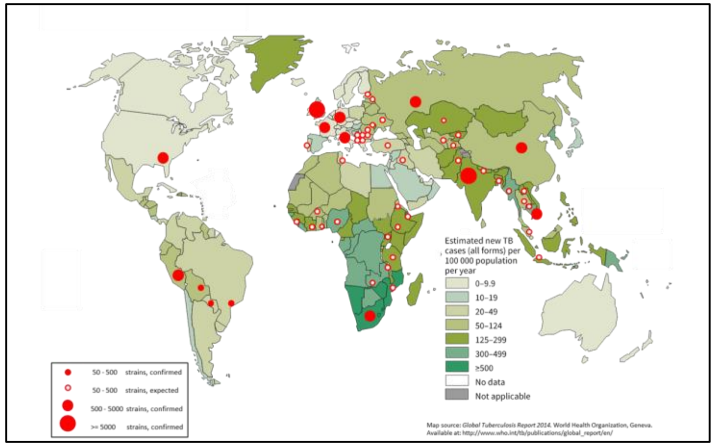


Figure 1: The number of resistant and sensitive phenotypes associated with each 'resistance-determinant' in the derivation- and validation sets: Black and red respectively for the derivation-set, and gray and orange respectively for the validation set. Variants probed by a line-probe assay are highlighted with red labels. Variants that are only seen once in the derivation-set and not in the validation-set (i.e. with no additional information) are not shown.

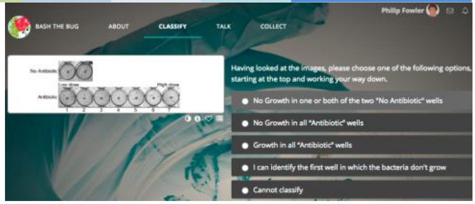

Estimated new TB cases (all forms) per 100 000 population per year
0–9.9
10–19
20–49
50–124
125–299
300–499
≥500
No data
Not applicable
50 - 500 strains, confirmed
50 - 500 strains, expected
500 - 5000 strains, confirmed
>= 5000 strains, confirmed

Map source: *Global Tuberculosis Report 2014*. World Health Organization, Geneva. Available at: http://www.who.int/tb/publications/global_report/en/

Phenotyping

BDQ 2	KAN 16	KAN 8	KAN 4	KAN 2	KAN 1	ETH 8	ETH 4	ETH 2	ETH 1	ETH 0.5	ETH 0.25
BDQ 1	AMI 8	EMB 8	INH 1.6	LEV 8	MXF 4	DLM 1	LZD 2	CFZ 4	RIF 4	RFB 2	PAS 4
BDQ 0.5	AMI 4	EMB 4	INH 0.8	LEV 4	MXF 2	DLM 0.5	LZD 1	CFZ 2	RIF 2	RFB 1	PAS 2
BDQ 0.25	AMI 2	EMB 2	INH 0.4	LEV 2	MXF 1	DLM 0.25	LZD 0.5	CFZ 1	RIF 1	RFB 0.5	PAS 1
BDQ 0.125	AMI 1	EMB 1	INH 0.2	LEV 1	MXF 0.5	DLM 0.125	LZD 0.25	CFZ 0.5	RIF 0.5	RFB 0.25	PAS 0.5
BDQ 0.06	AMI 0.5	EMB 0.50	INH 0.1	LEV 0.5	MXF 0.25	DLM 0.06	LZD 0.125	CFZ 0.25	RIF 0.25	RFB 0.125	PAS 0.25
BDQ 0.03	AMI 0.25	EMB 0.25	INH 0.05	LEV 0.25	MXF 0.125	DLM 0.03	LZD 0.06	CFZ 0.125	RIF 0.125	RFB 0.0625	PAS 0.125
BDQ 0.015	EMB 0.0625	EMB 0.125	INH 0.025	LEV 0.125	MXF 0.0625	DLM 0.015	LZD 0.03	CFZ 0.0625	RIF 0.0625	POS control	POS control

Pyrazinamide will be done by MGIT liquid culture



Having looked at the images, please choose one of the following options, starting at the top and working your way down.

- ○ No Growth in one or both of the two "No Antibiotic" wells
- ○ No Growth in all "Antibiotic" wells
- ○ Growth in all "Antibiotic" wells
- ○ I can identify the first well in which the bacteria don't grow
- ○ Cannot classify

BASH THE BUG

People powered research
zooniverse.org

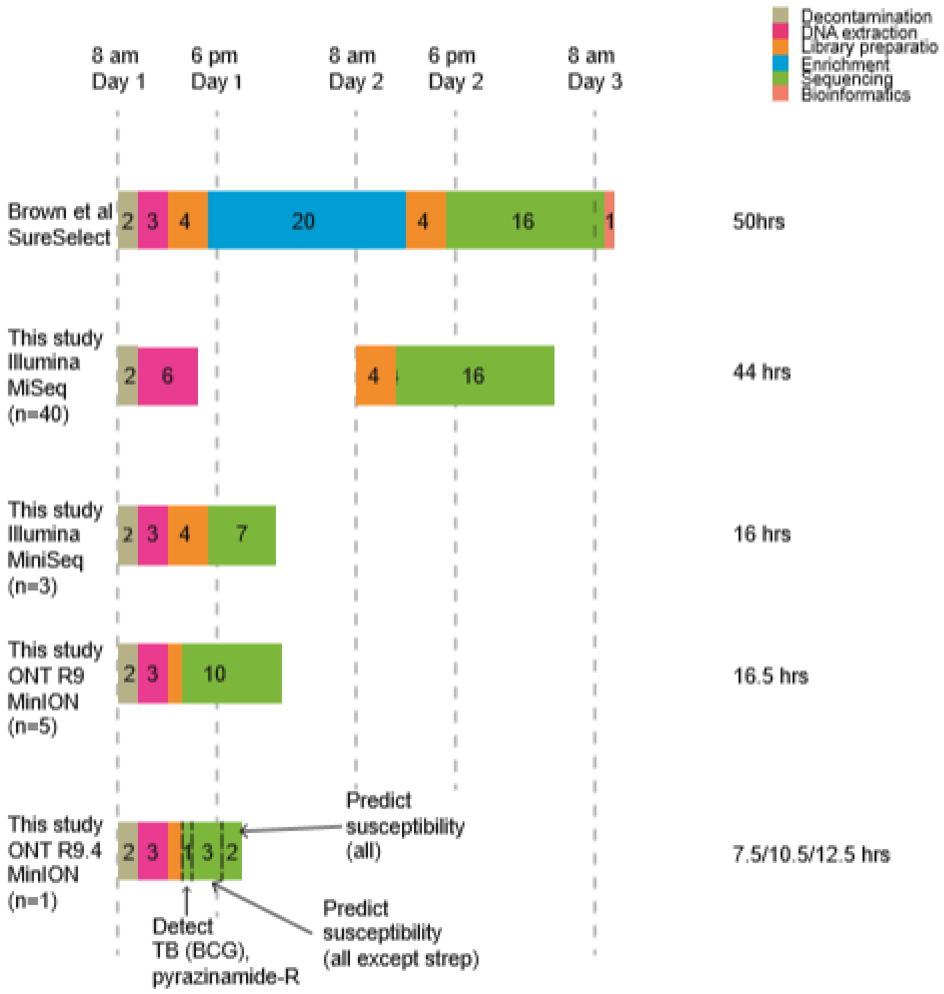
Twitter: @bashthebug

Genotypic characterisation

- 100,000 WGS TB pledged
- ~ 40,000 with extensive DST
- Analysis:
  - Heuristic approach
  - GWAS
  - Machine Learning
  - Thermodynamic modelling of proteins
  - Molecular genetic characterisation

A faster sequencer

How long does it take?


8 am Day 1
6 pm Day 1
8 am Day 2
6 pm Day 2
8 am Day 3
Decontamination
DNA extraction
Library preparatio
Enrichment
Sequencing
Bioinformatics
Brown et al SureSelect
2
3
4
20
4
16
1
50hrs
This study Illumina MiSeq (n=40)
2
6
4
16
44 hrs
This study Illumina MiniSeq (n=3)
2
3
4
7
16 hrs
This study ONT R9 MinION (n=5)
2
3
10
16.5 hrs
This study ONT R9.4 MinION (n=1)
2
3
1
3
2
7.5/10.5/12.5 hrs
Predict susceptibility (all)
Detect TB (BCG), pyrazinamide-R
Predict susceptibility (all except strep)

J Clin Microbiol. 2017 May;55(5):1285-1298

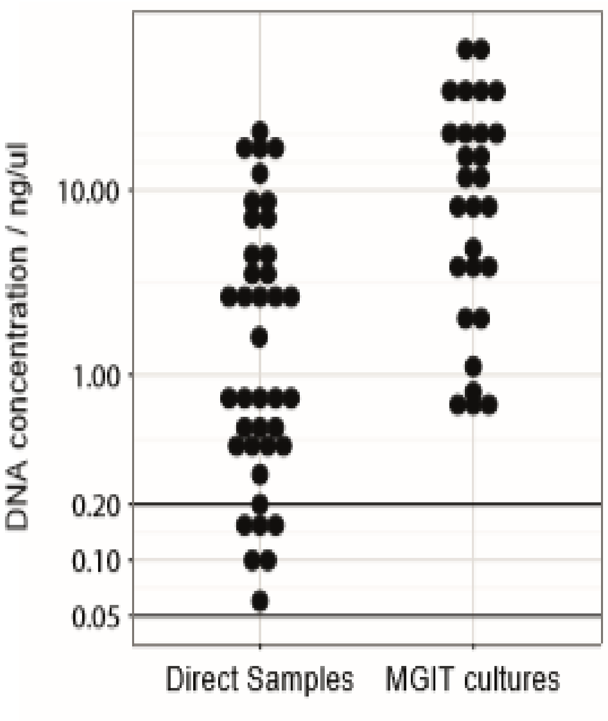
UNIVERSITY OF OXFORD

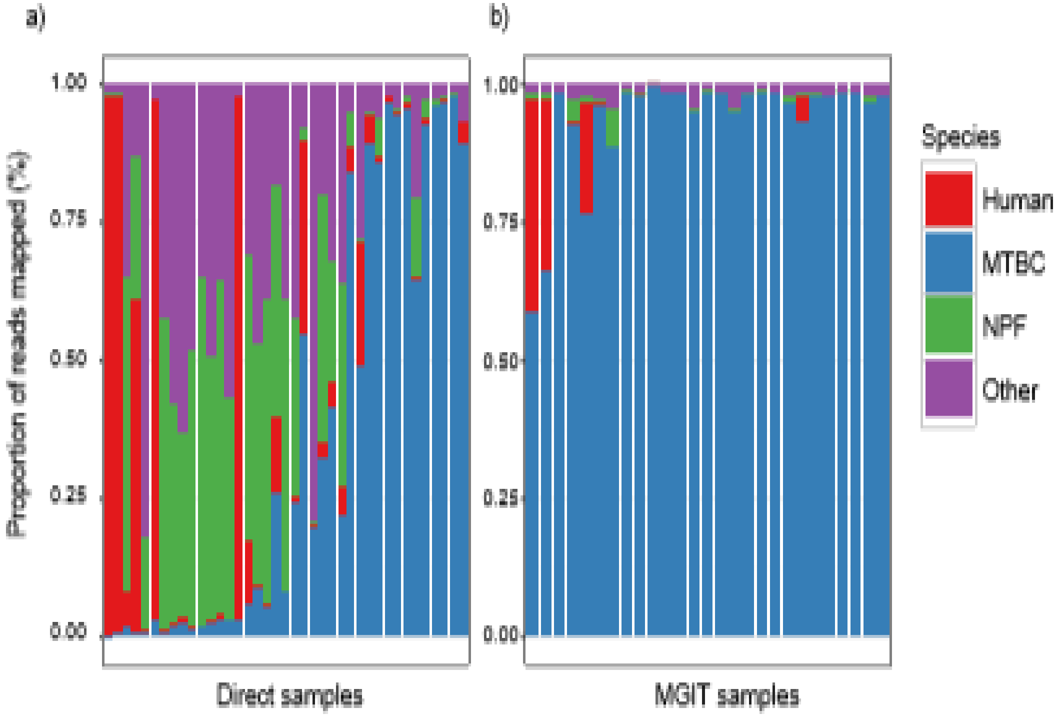
Public Health England

Direct from a sample

Can we do it direct from sputum?

All samples ≥1+ positive for AFB





J Clin Microbiol. 2017 May;55(5):1285-1298

UNIVERSITY OF OXFORD

Public Health England

A faster software

UNIVERSITY OF OXFORD

Public Health England

What limits of detection are we aiming for?

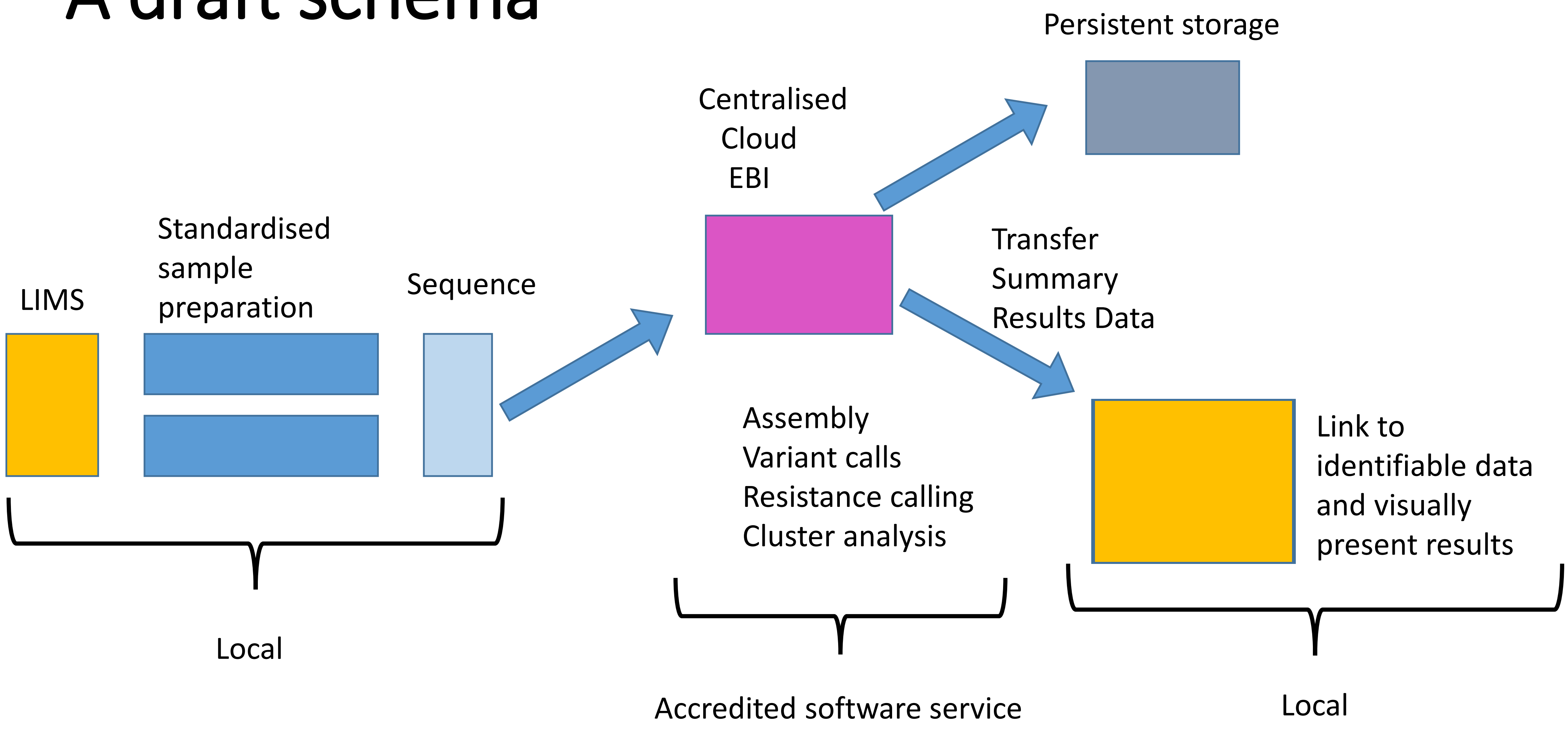
0 – 4+	AFB/ml	HPF/AFB	Genexpert	WGS
4+	10,000,000	10	+	complete
3+	1,000,000	1	+	complete
2+	100,000	0.1	+	complete
1+	10,000	0.01	**+**	In-complete
scanty	3,000	0.003	**+**	In-complete

Establish a WGS software application on the cloud

- Accessible to users anywhere, anytime and will need:
  - reasonable internet bandwidth
  - Simple extraction
  - light-weight sequencing infrastructure

- Partners are setting up field sites in:
  - Mumbai
  - Ho Chi Minh City
  - Madagascar

A draft schema



Persistent storage

Centralised
Cloud
EBI

Transfer
Summary
Results Data

LIMS

Standardised
sample
preparation

Sequence

Assembly
Variant calls
Resistance calling
Cluster analysis

Link to
identifiable data
and visually
present results

Local

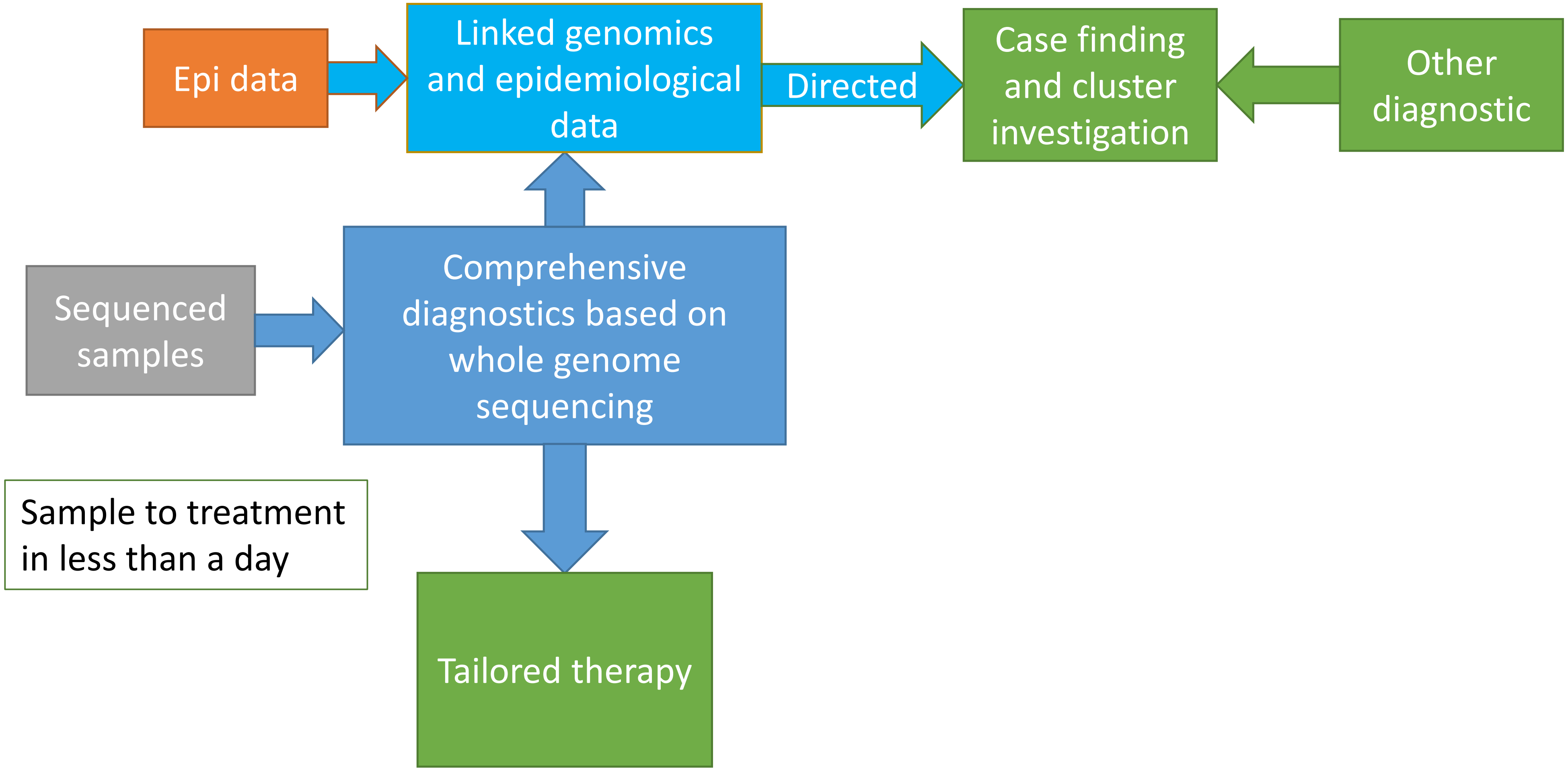
Accredited software service

Local

UNIVERSITY OF
OXFORD

Public Health
England

The schema for diagnostics and prevention



Epi data

Linked genomics and epidemiological data

Directed

Case finding and cluster investigation

Other diagnostic

Sequenced samples

Comprehensive diagnostics based on whole genome sequencing

Sample to treatment in less than a day

Tailored therapy

Acknowledgements

- Sarah Walker
- Zamin Iqbal EBI
- Tim Peto
- Guy Thwaites - Vietnam
- Mark Wilcox – Leeds
- Grace Smith – Birmingham
- Philip Monk - Leicester
- Tim Walker – Oxford
- Esther Robinson – Birmingham
- Research Fellows (6)
- Martin Dedicote - Birmingham
- David Moore - LSHTM and Peru

Microbiology, DNA preparation

- Dai Griffiths
- Kate Dingle
- Nicole Stoesser
- Alison Vaughan
- Bernadette Young
- Claire Gordon

International

Oxford High Throughput Sequencing Hub team

People participating in the studies

Informatics

- David Wyllie
- Fan Turner
- Martin Hunt
- Trien Do
- Jeremy Swann

Bioinformatics and Population Biology

- Danny Wilson
- Carlos del Ojo Elias
- Saheer Gharbia
- Tanya Golubchik
- Anna Sheppard
- Dilrini de Silva
- Xavier Didelot
- Jess Hedge