THE UNIVERSITY OF SYDNEY

The Sydney children's Hospitals Network

care, advocacy, research, education

Molecular epidemiology

Alex Outhred

HK Paediatric Infection Control Workshop 2016

Typing methods

- Phenotypic methods:
  - antimicrobial susceptibility
  - biotype (metabolism or enzyme activity)
  - antibodies (capsule or surface proteins)
  - bacteriophage typing
- Genotypic methods:
  - randomly amplified polymorphic DNA (RAPD)
  - repetitive element PCR (rep-PCR)
  - pulsed-field gel electrophoresis (PFGE) / restriction fragment length polymorphism (RFLP)
  - variable-nucleotide tandem repeats (VNTR) / multi-locus VNTR analysis (MLVA)
  - single locus sequence typing (SLST) / multi-locus sequence typing (MLST)
  - reverse line blot (RLB) binary typing / microarray typing
- Fundamentally, all are based on DNA sequence of the genome
  - whole-genome sequencing (WGS)

Parameters of typing methods: choose your compromise

- Properties of all diagnostic tests:
  - Accuracy
    - true positive rate / specificity - eg. clusters represent a biological truth
    - true negative rate / sensitivity - eg. isolates with differing types less closely related
  - Precision
    - intra-laboratory reproducibility - same people repeat test, get same result
    - inter-laboratory reproducibility - different people repeat test, get same result
- Special parameter for typing:
  - Diversity
    - how likely two isolates selected from unrelated population will show different types
    - Shannon entropy / Simpson index / Hunter-Gaston index
- General parameters:
  - time to receive result
  - costs
    - capital
    - consumable
    - labor

Clonal vs. recombinant organisms

- Clonal organisms:
  - genetic changes accumulate steadily over time
  - changes are primarily point mutations / single nucleotide polymorphisms
  - low rates of genetic exchange between organisms (except “mother/daughter” replication)
  - eg. *M. tuberculosis*, HBV
- Recombinant organisms:
  - genetic changes accumulate at varying rates
  - point mutations, but also large chunks of DNA appearing / disappearing / moving location
  - high rates of exchange of genetic material between organisms (sometimes different species, genera or even kingdom)
  - eg. *Enterobacteriaceae*, influenza
- It’s a continuum:
  - all organisms show some recombination (may be self-only)
  - all organisms show point mutations / SNPs
  - eg. *Staphylococcus aureus* is quite clonal, but shows more recombination than *M. tuberculosis*

What’s a SNP?

- SNP
  - single nucleotide polymorphism
  - point mutation
  - change at a single point from A to C, C to A, C to T, T to G etc.
  - characteristic change of clonal organisms
- MNP / SV / LSP etc.
  - multiple nucleotide polymorphism
  - structural variant
  - large sequence polymorphism
  - string of DNA jumps - disappears, moves from point A to point B, is duplicated, etc.
  - characteristic change of recombinant organisms

Why would you want to do whole-genome sequencing?

In the past, for research purposes:

- Virulence, pathogenicity
- Interactions with immune system
- Transmissibility
- Big picture phylogeny
- Evolution of resistance
  - and of compensatory fitness-restoring changes

Why would you want to do WGS?

Today, for cluster investigation:

- highest resolution typing
  - WGS clusters likely to represent recent transmission
    - target your interventions
  - more accurate rates of recent transmission
  - understand prevalent lineages, sublineages
  - insight into reservoirs, modes of transmission, role of HCW, environment
  - correlate genetic markers of drug resistance with WGS phylogeny

Example: WGS cluster investigation

- (Public health, not infection control)
  - but they're really quite similar!
- Large cluster of tuberculosis cases
- Cases with socioeconomic disadvantage
- Vigorous public health response
- Linked cases still developing after 15 years

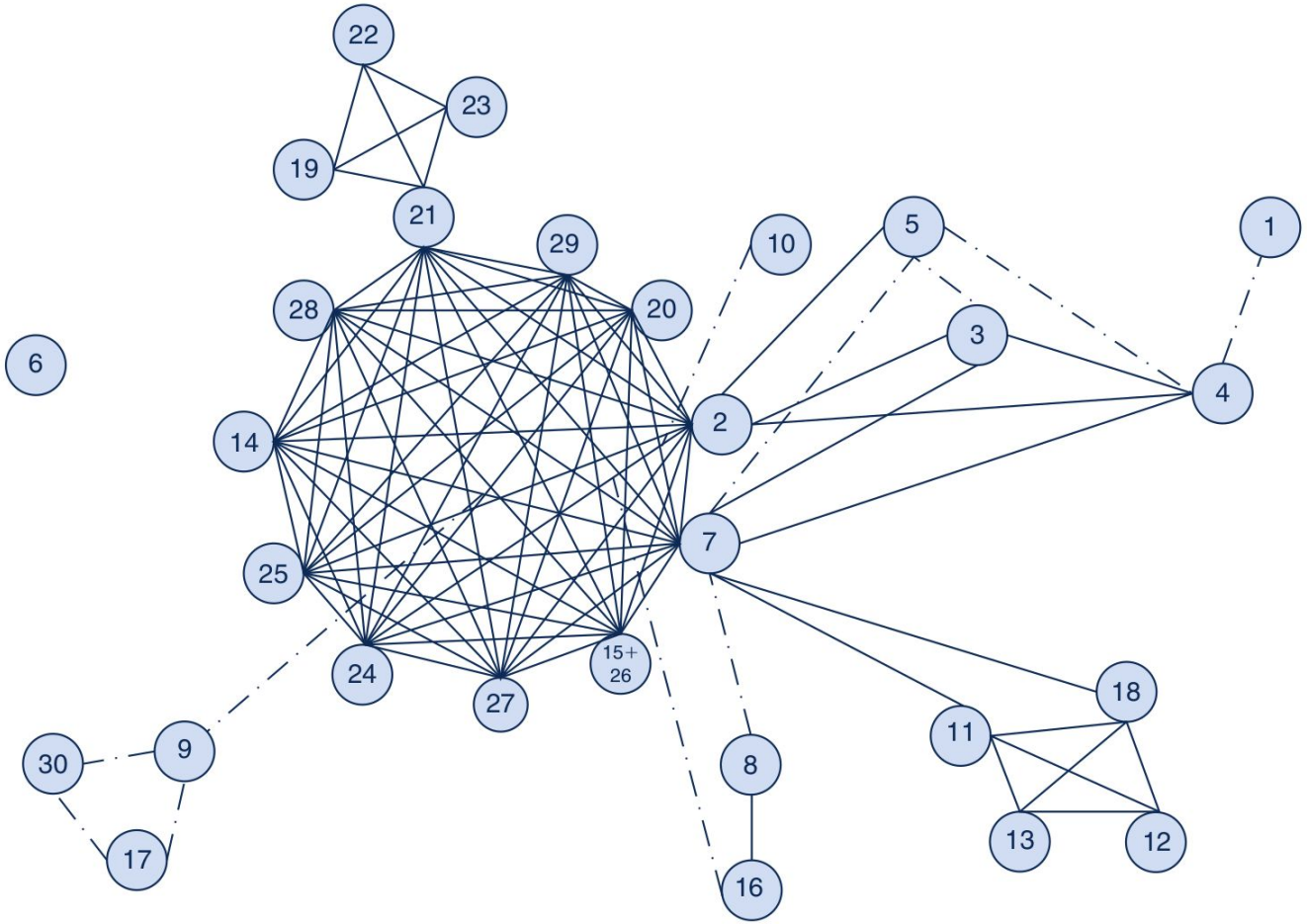


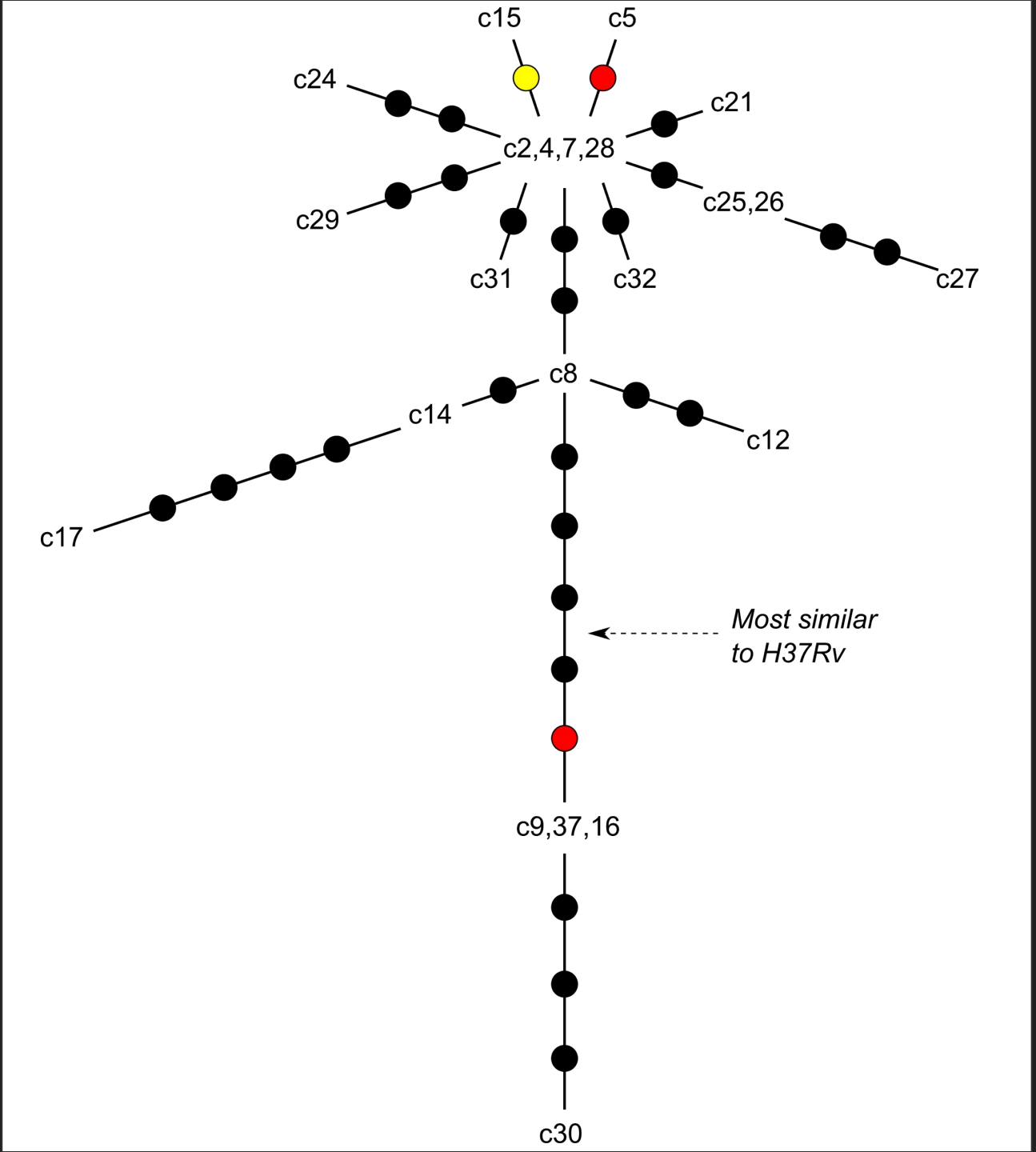
Figure 3. Social networks of tuberculosis cases, NSW/QLD cluster, October 2000–July 2012.

Note: Solid lines denote close social or household contact, dashed lines denote casual or community-level contact.

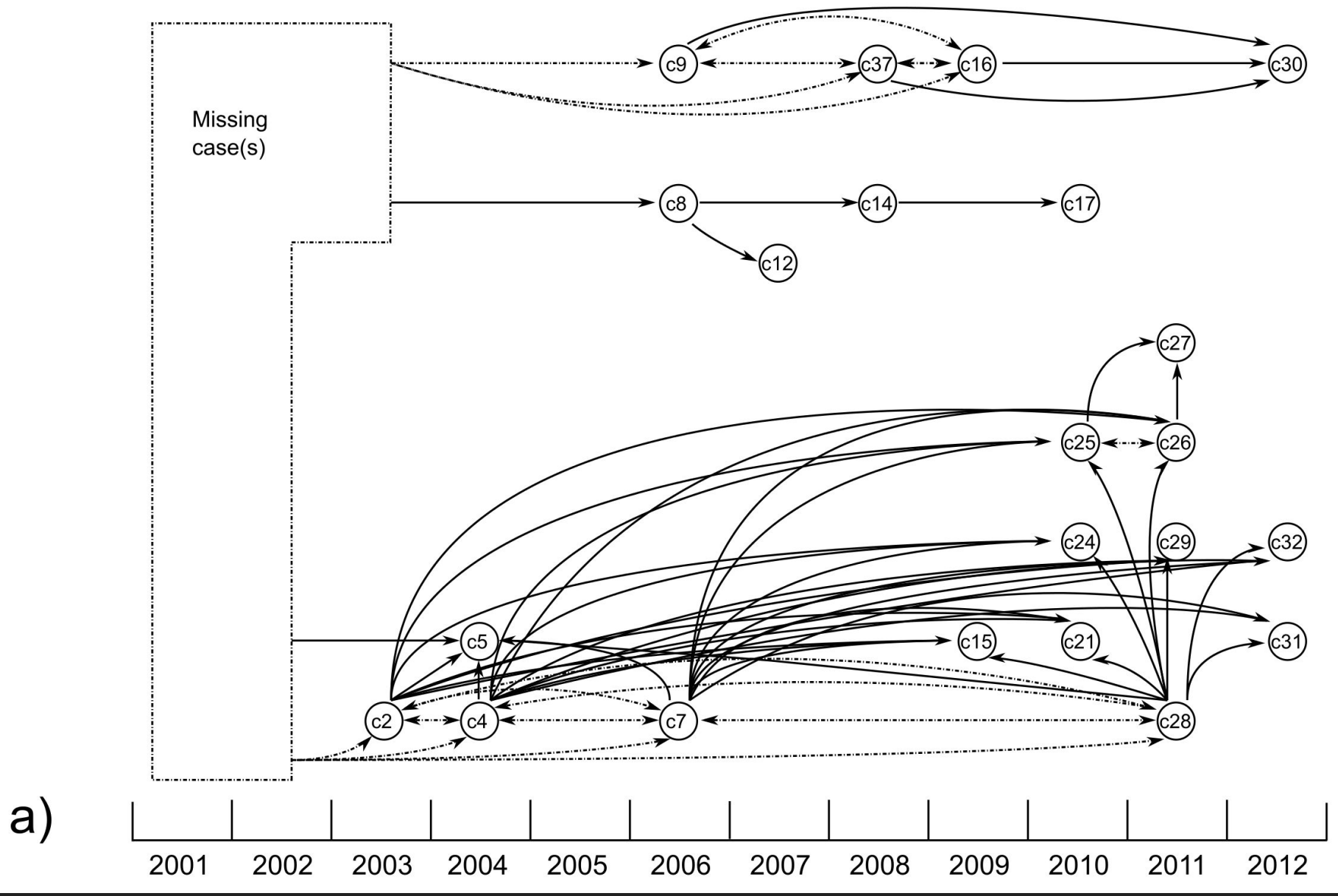
Source: TB Prevention and Control Team, Northern NSW and Mid North Coast Local Health Districts.

Devlin and Passmore 2013

Position in H37Rv	154900	173422	241392	473306^7	1008303	1669219	1761523	1778869	1891805	2053258	2184243	2320645	2502667	2566036	2662735	2716491	2881037	3138179	3218218	3343411	3387492	3494258	3538087	3719370	3773850	3867072	3952516	4318153	4323153
H37Rv	G	G	T	-	A	C	C	C	G	G	G	G	C	A	C	G	C	C	C	C	C	C	G	G	A	G	G	T	C
c2	G	G	T	-	G	C	C	C	G	G	G	A	C	C	C	C	C	C	T	C	C	C	G	G	A	G	G	T	C
c4	G	G	T	-	G	C	C	C	G	G	G	A	C	C	C	C	C	C	T	C	C	C	G	G	A	G	G	T	C
c7	G	G	T	-	G	C	C	C	G	G	G	A	C	C	C	C	C	C	T	C	C	C	G	G	A	G	G	T	C
c28	G	G	T	-	G	C	C	C	G	G	G	A	C	C	C	C	C	C	T	C	C	C	G	G	A	G	G	T	C
c5	G	G	T	-	G	C	C	C	G	G	G	A	C	C	C	C	C	C	T	C	C	-	G	G	A	G	G	T	C
c15	G	G	T	A	G	C	C	C	G	G	G	A	C	C	C	C	C	C	T	C	C	C	G	G	A	G	G	T	C
c21	G	G	T	-	G	C	C	C	G	G	G	A	C	C	C	C	C	C	T	C	C	C	G	G	A	G	G	T	A
c32	G	G	T	-	G	C	C	C	G	G	G	A	C	C	C	C	C	C	T	C	C	C	G	G	G	G	G	T	C
c31	G	G	C	-	G	C	C	C	G	G	G	A	C	C	C	C	C	C	T	C	C	C	G	G	A	G	G	T	C
c29	G	G	T	-	G	C	C	C	G	G	G	A	C	C	C	C	C	C	T	C	C	C	G	T	A	G	G	C	C
c24	A	G	T	-	G	C	C	C	G	C	G	A	C	C	C	C	C	C	T	C	C	C	G	G	A	G	G	T	C
c25	G	G	T	-	G	C	C	C	G	G	G	A	C	C	C	C	C	C	T	C	A	C	G	G	A	G	G	T	C
c26	G	G	T	-	G	C	C	C	G	G	G	A	C	C	C	C	C	C	T	C	A	C	G	G	A	G	G	T	C
c27	G	G	T	-	G	C	C	C	T	G	G	A	C	C	C	C	C	C	T	T	A	C	G	G	A	G	G	T	C
c8	G	G	T	-	A	C	C	C	G	G	G	A	C	C	C	G	C	C	T	C	C	C	G	G	A	G	G	T	C
c14	G	G	T	-	A	C	C	C	G	G	G	A	C	C	T	G	C	C	T	C	C	C	G	G	A	G	G	T	C
c17	G	G	T	-	A	T	C	A	G	G	C	A	C	C	T	G	C	C	T	C	C	C	G	G	A	G	A	T	C
c12	G	A	T	-	A	C	C	C	G	G	G	A	C	C	C	G	A	C	T	C	C	C	G	G	A	G	G	T	C
c9	G	G	T	-	A	C	C	C	G	G	G	G	T	A	C	G	C	-	C	C	C	C	G	G	A	G	G	T	C
c37	G	G	T	-	A	C	C	C	G	G	G	G	T	A	C	G	C	-	C	C	C	C	G	G	A	G	G	T	C
c16	G	G	T	-	A	C	C	C	G	G	G	G	T	A	C	G	C	-	C	C	C	C	G	G	A	G	G	T	C
c30	G	G	T	-	A	C	G	C	G	G	G	G	T	A	C	G	C	-	C	C	C	C	C	G	A	A	G	T	C

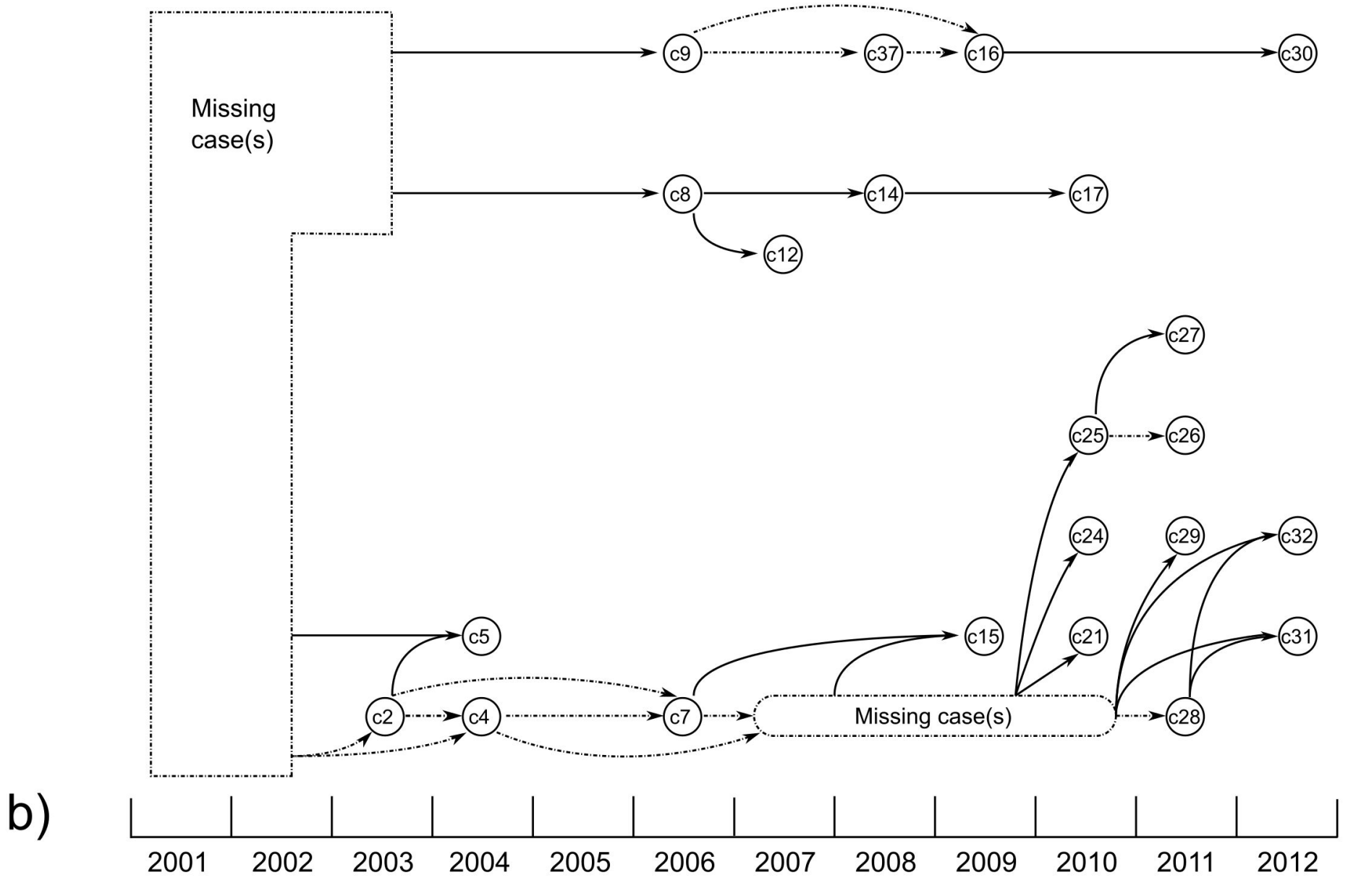


Outhred et al, PLoS ONE, 2016 (in press)


Missing
case(s)
c9
c37
c16
c30
c8
c14
c17
c12
c27
c25
c26
c24
c29
c32
c5
c15
c21
c31
c2
c4
c7
c28
a)
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012

Can we simplify pathways with epidemiological assumptions?

- (No homoplasy / convergent evolution)
  - already applied in previous figure, (a)
- No transmission after diagnosis
- No progression to disease more than 3 years after infection
- Cases diagnosed later cannot be sources for cases diagnosed earlier


Missing case(s)
c9
c37
c16
c30
c8
c14
c17
c12
c27
c25
c26
c24
c29
c32
c5
c15
c21
c31
c2
c4
c7
Missing case(s)
c28
b)
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012

Barriers to WGS: cost

- WGS cost: 100USD/isolate is still expensive
  - Costs for other methods vary greatly between institutions and countries
    - labour can be the dominant determinant
  - Comparison with other methods
    - ~20USD for SLST, RLB binary typing
    - 30-40USD per isolate for PFGE, microarrays
    - 60USD per isolate for MLST
  - For all methods, lower costs if you type more isolates at once (especially true for WGS)
  - WGS cost has fallen rapidly, likely to continue to fall

Barriers to WGS: infrastructure

- Sequencing infrastructure:
  - WGS machines expensive, depreciate quickly
  - Therefore outsource - send extracted genomic DNA to commercial or academic partners for WGS
- Bioinformatic infrastructure:
  - Bacterial genomes mostly modest size
  - Standard desktop (or laptop) computers, extra RAM
  - Hard part is high-speed data network between sequencing facility and analysis location
    - but mail/courier transport of portable hard drives is an alternative

Barriers to WGS: expertise

- Expertise:
  - Bioinformatics can be very complex
  - But open, no secret knowledge or hardware needed
  - For some problems, more user-friendly software is becoming available
  - Collaborate with international partners
    - potential to rapidly develop the skills needed
    - mutual benefits: getting science done and improving public health and infection control programmes

Summary

- WGS not yet the answer to all our typing needs
  - but probably will supplant other methods over the next decade, so start thinking about it
- Keep choosing the typing method that is the best compromise for your setting
- Even with WGS, still need accurate epidemiologic data to understand transmission