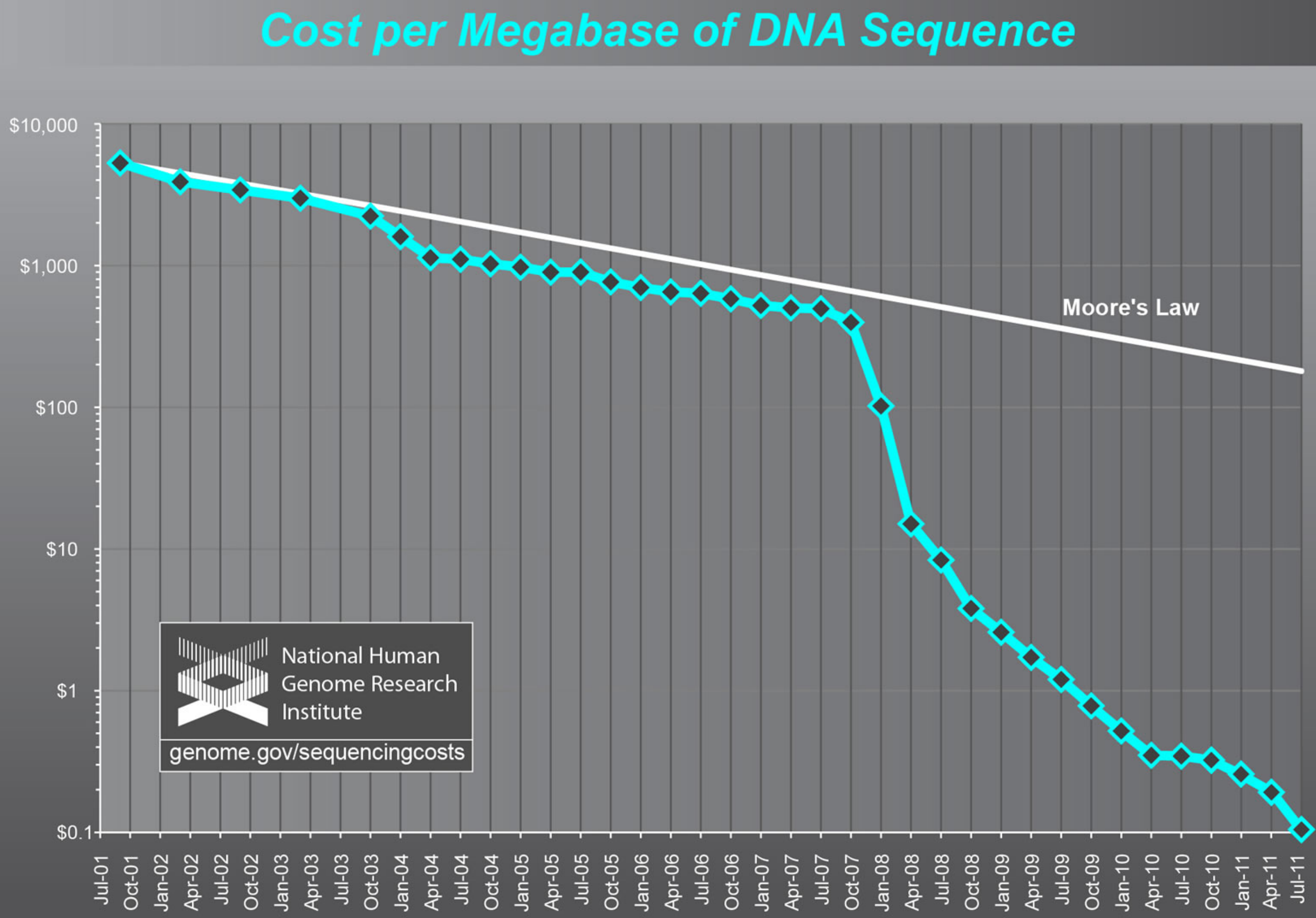
Imperial College London

MRC Medical Research Council

How will the control of infectious disease be improved by genomic analysis?

Christophe Fraser

MRC Centre for Outbreak Analysis
Dept of Infectious Disease Epidemiology



The sequencing revolution is now.

Desktop sequencing in every lab and in every hospital ward

Sequencers are the new microscopes

Why sequence?

- To track pathogens and outbreaks
- To identify factors important for disease
- To tailor treatment
- To understand evolution

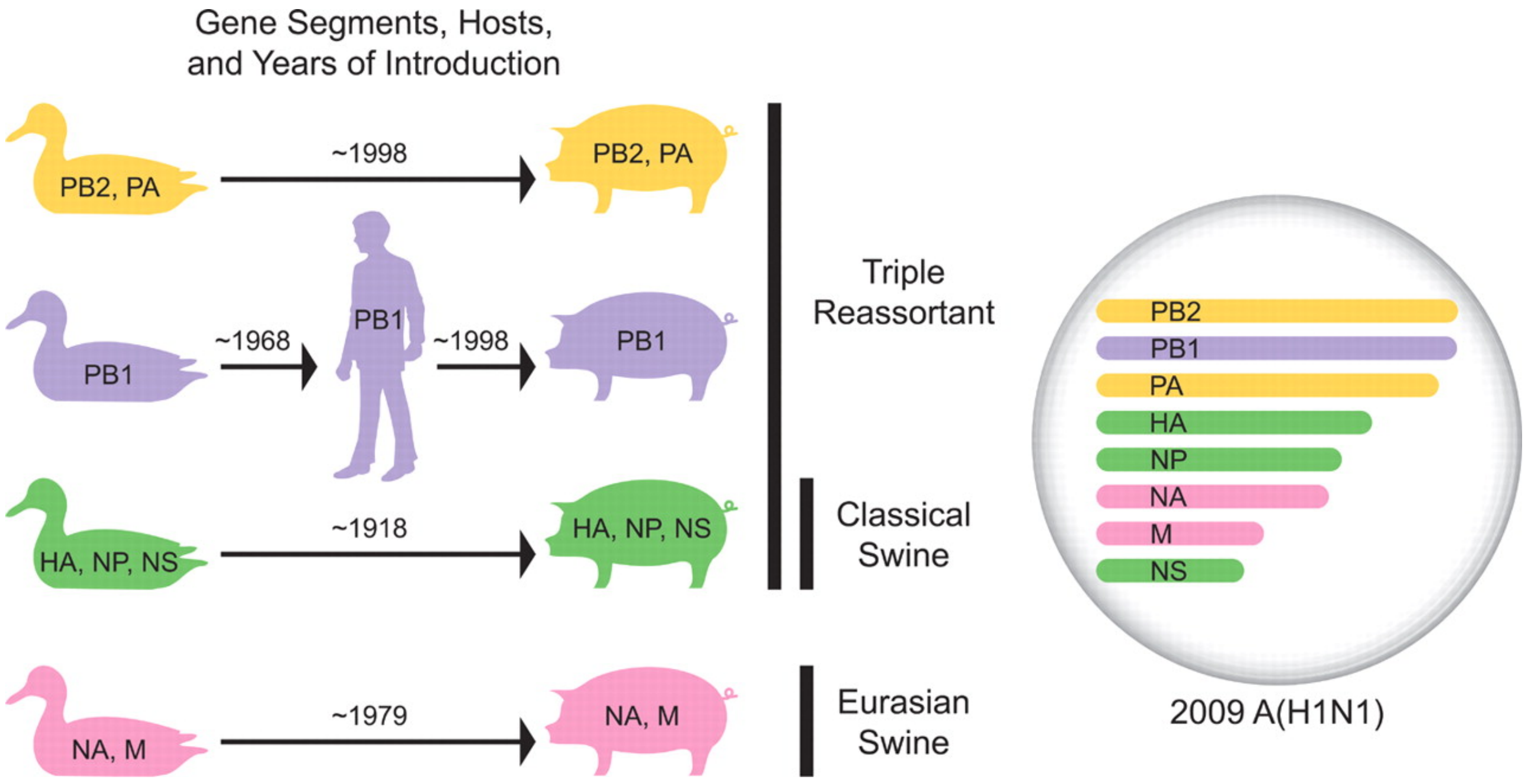
Two Examples:

- The 2009 H1N1 influenza pandemic
- The PMEN-1 lineage of *Streptococcus pneumoniae*

- The 2009 H1N1 influenza pandemic

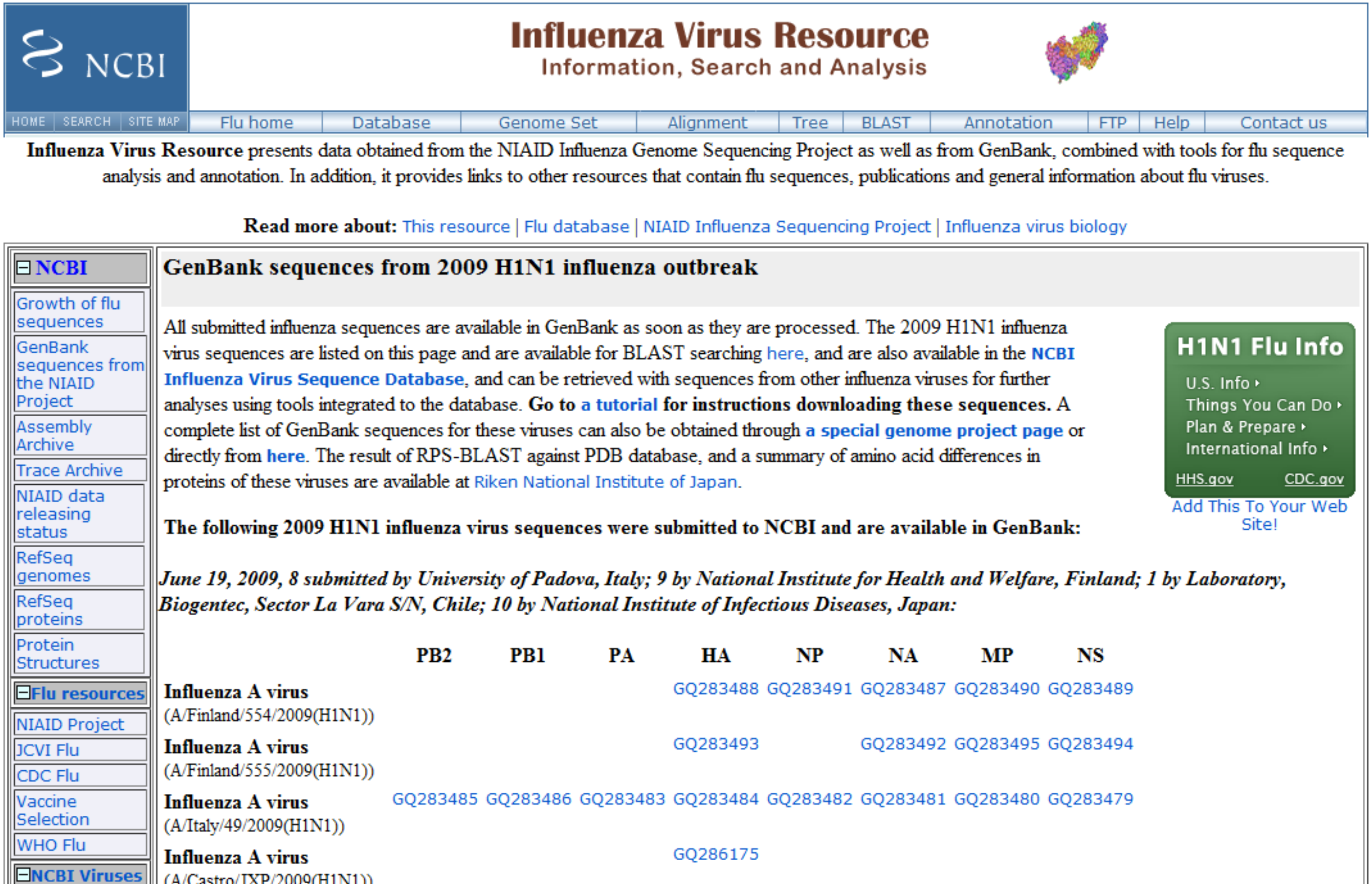
- Tracking bacterial spread and evolution

In April 2009, virological surveillance triggered a pandemic



Garten et al Science 2009, Smith et al Nature 2009

Genetic data was shared in days



Influenza Virus Resource

Information, Search and Analysis

Influenza Virus Resource presents data obtained from the NIAID Influenza Genome Sequencing Project as well as from GenBank, combined with tools for flu sequence analysis and annotation. In addition, it provides links to other resources that contain flu sequences, publications and general information about flu viruses.

Read more about: This resource | Flu database | NIAID Influenza Sequencing Project | Influenza virus biology

GenBank sequences from 2009 H1N1 influenza outbreak

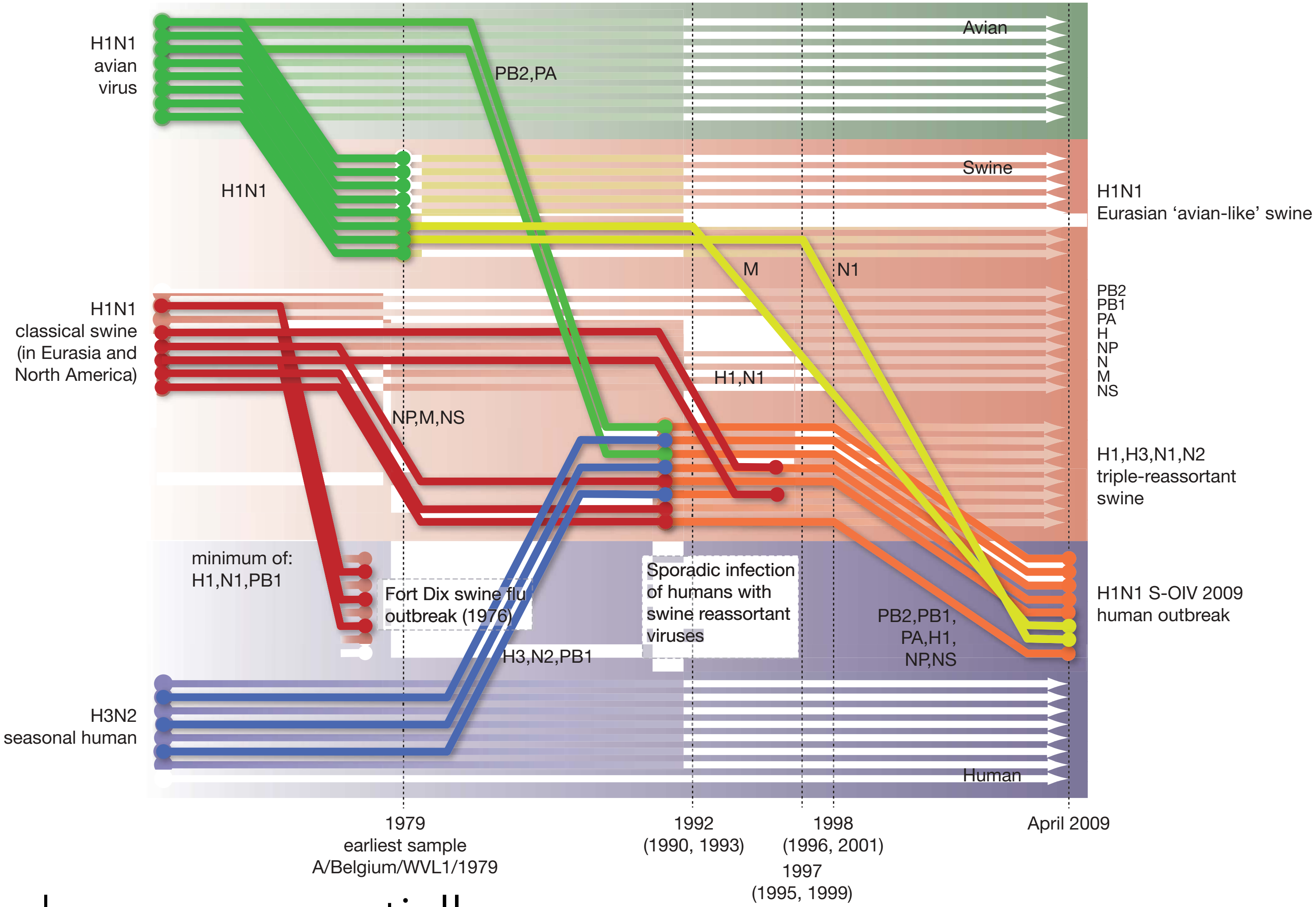
All submitted influenza sequences are available in GenBank as soon as they are processed. The 2009 H1N1 influenza virus sequences are listed on this page and are available for BLAST searching here, and are also available in the **NCBI Influenza Virus Sequence Database**, and can be retrieved with sequences from other influenza viruses for further analyses using tools integrated to the database. **Go to a tutorial for instructions downloading these sequences.** A complete list of GenBank sequences for these viruses can also be obtained through **a special genome project page** or directly from **here**. The result of RPS-BLAST against PDB database, and a summary of amino acid differences in proteins of these viruses are available at Riken National Institute of Japan.

The following 2009 H1N1 influenza virus sequences were submitted to NCBI and are available in GenBank:

June 19, 2009, 8 submitted by University of Padova, Italy; 9 by National Institute for Health and Welfare, Finland; 1 by Laboratory, Biogentec, Sector La Vara S/N, Chile; 10 by National Institute of Infectious Diseases, Japan:

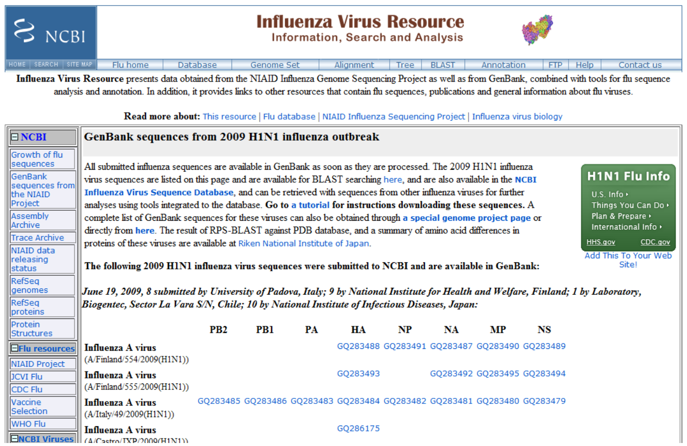
	PB2	PB1	PA	HA	NP	NA	MP	NS
Influenza A virus (A/Finland/554/2009(H1N1))				GQ283488	GQ283491	GQ283487	GQ283490	GQ283489
Influenza A virus (A/Finland/555/2009(H1N1))				GQ283493		GQ283492	GQ283495	GQ283494
Influenza A virus (A/Italy/49/2009(H1N1))	GQ283485	GQ283486	GQ283483	GQ283484	GQ283482	GQ283481	GQ283480	GQ283479
Influenza A virus (A/Castro/JXP/2009(H1N1))				GQ286175				

Genomics revealed a complex evolutionary history



Still includes some essentially human-adapted core genes

Gavin Smith et al, Nature 2009 & 2010



NCBI
HOME | SEARCH | SITE MAP
Flu home | Database | Genome Set | Alignment | Tree | BLAST | Annotation | FTP | Help | Contact us

Influenza Virus Resource
Information, Search and Analysis

Influenza Virus Resource presents data obtained from the NIAID Influenza Genome Sequencing Project as well as from GenBank, combined with tools for flu sequence analysis and annotation. In addition, it provides links to other resources that contain flu sequences, publications and general information about flu viruses.

Read more about: This resource | Flu database | NIAID Influenza Sequencing Project | Influenza virus biology

GenBank sequences from 2009 H1N1 influenza outbreak

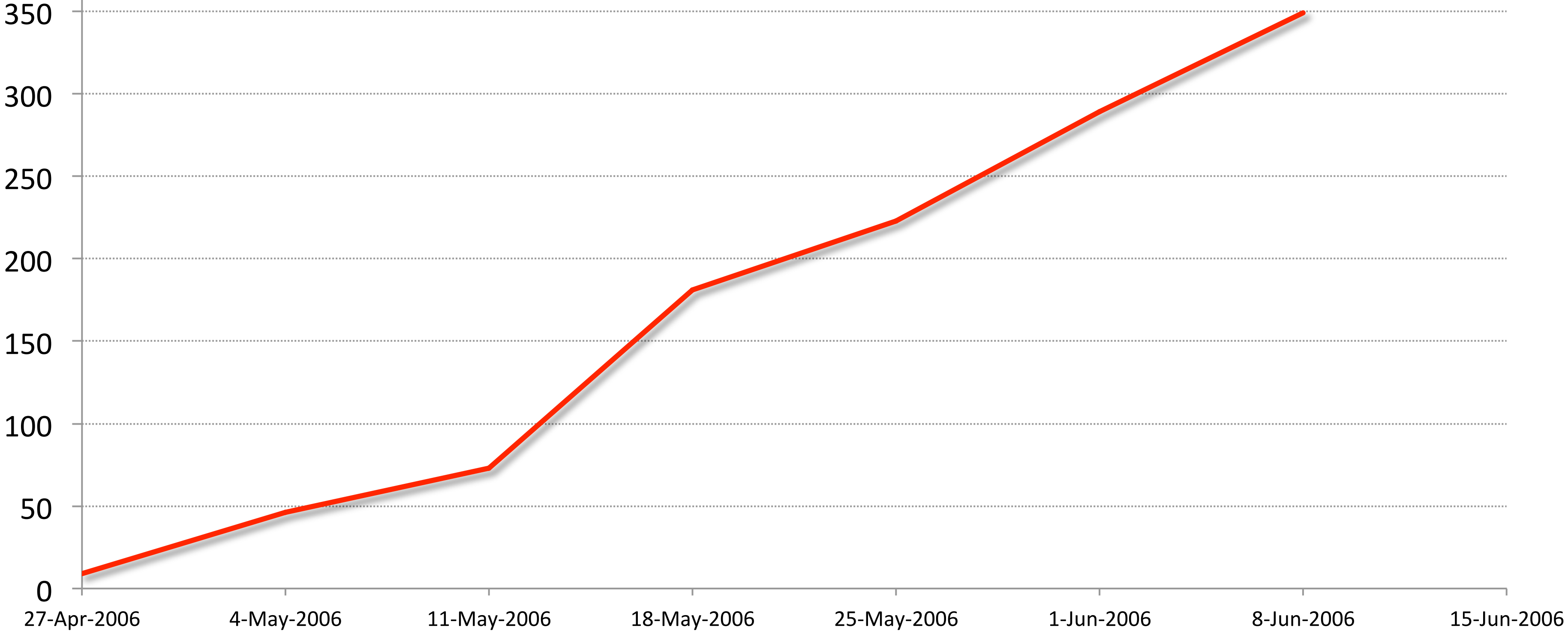
All submitted influenza sequences are available in GenBank as soon as they are processed. The 2009 H1N1 influenza virus sequences are listed on this page and are available for BLAST searching here, and are also available in the **NCBI Influenza Virus Sequence Database**, and can be retrieved with sequences from other influenza viruses for further analyses using tools integrated to the database. **Go to a tutorial for instructions downloading these sequences.** A complete list of GenBank sequences for these viruses can also be obtained through **a special genome project page** or directly from **here**. The result of RPS-BLAST against PDB database, and a summary of amino acid differences in proteins of these viruses are available at Riken National Institute of Japan.

The following 2009 H1N1 influenza virus sequences were submitted to NCBI and are available in GenBank:

June 19, 2009, 8 submitted by University of Padova, Italy; 9 by National Institute for Health and Welfare, Finland; 1 by Laboratory, Biogentec, Sector La Vara S/N, Chile; 10 by National Institute of Infectious Diseases, Japan:

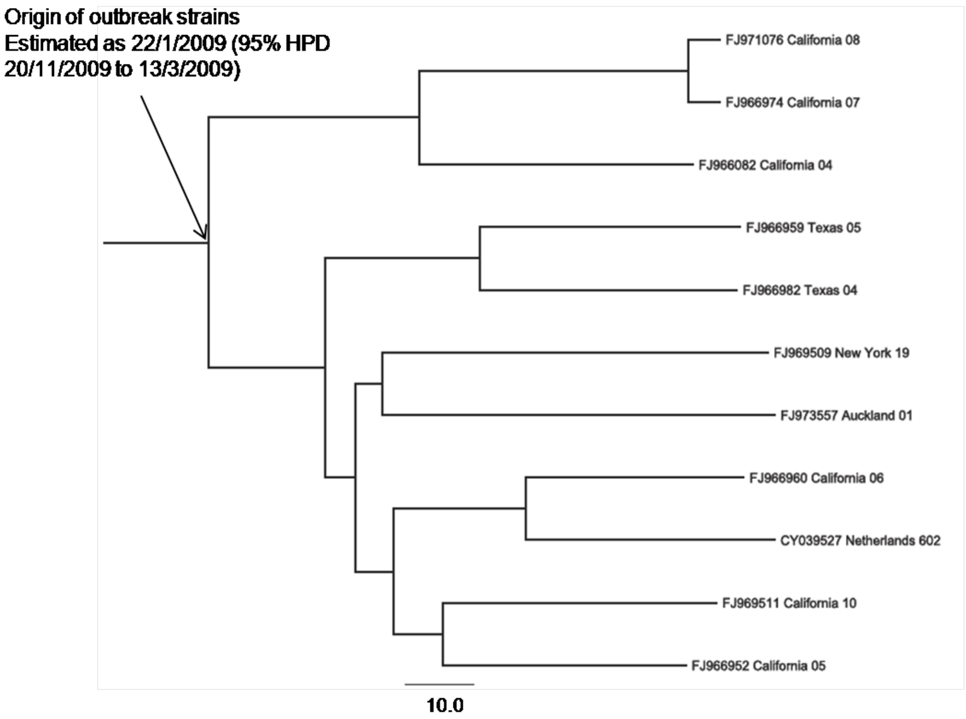
	PB2	PB1	PA	HA	NP	NA	MP	NS
Influenza A virus (A/Finland/554/2009(H1N1))				GQ283488	GQ283491	GQ283487	GQ283490	GQ283489
Influenza A virus (A/Finland/555/2009(H1N1))				GQ283493		GQ283492	GQ283495	GQ283494
Influenza A virus (A/Italy/49/2009(H1N1))	GQ283485	GQ283486	GQ283483	GQ283484	GQ283482	GQ283481	GQ283480	GQ283479
Influenza A virus (A/Castro/JXP/2009(H1N1))				GQ286175				

Number of full length HA sequences:



Currently over 2,600 full length HA sequences (for H1N1pdm virus)

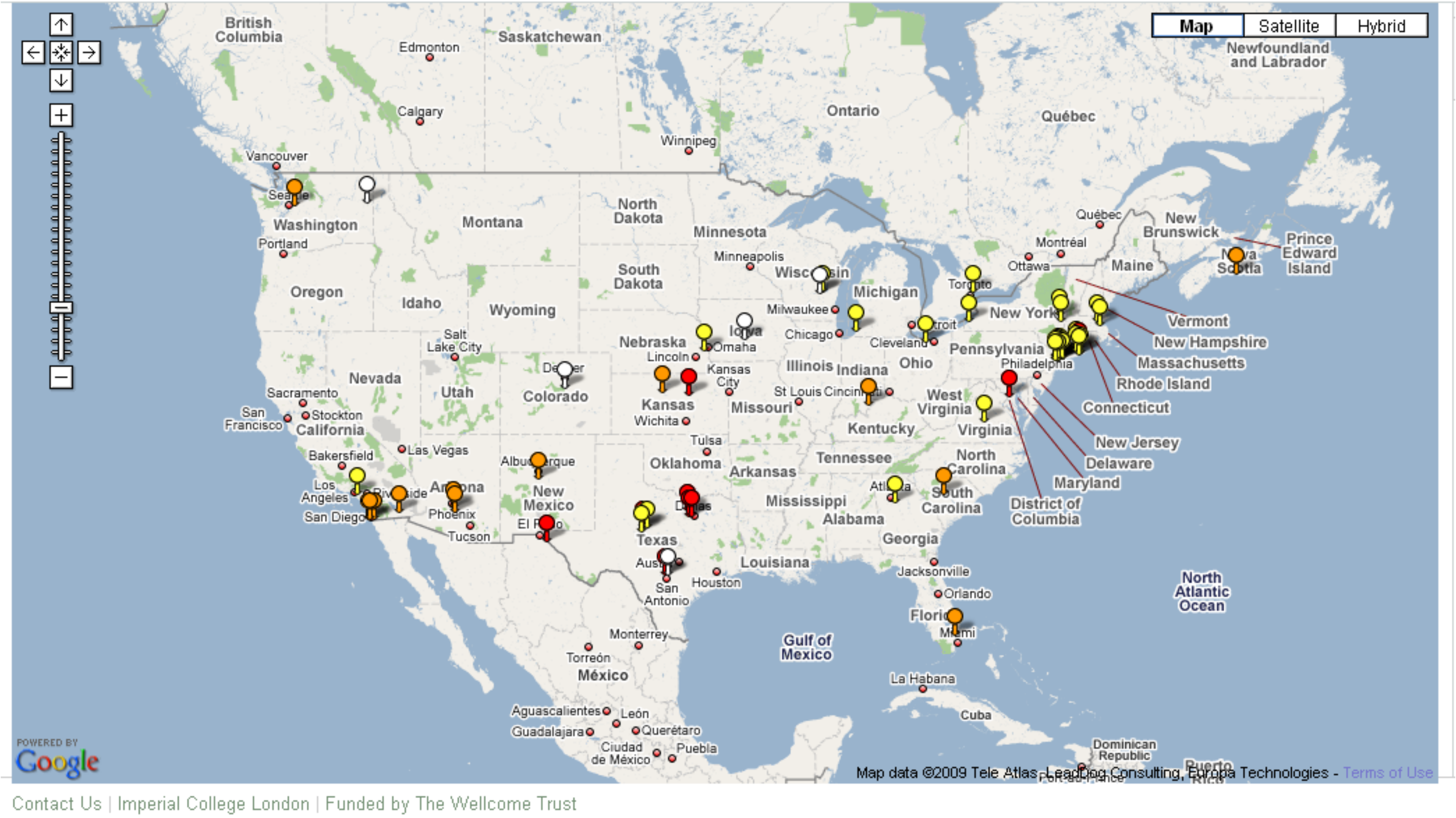
Earliest H1N1pdm trees (4 May)



Origin of outbreak strains
Estimated as 22/1/2009 (95% HPD
20/11/2009 to 13/3/2009)

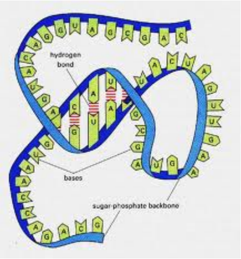
We can already tell from tree structure that the epidemic wasn't growing very fast…

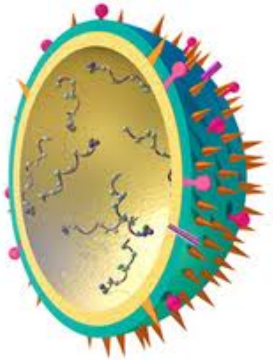
Molecular epidemiology to detect local transmission (EpiInfo)



Contact Us | Imperial College London | Funded by The Wellcome Trust

PHYLODYNAMICS:

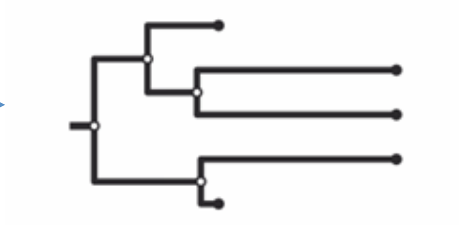


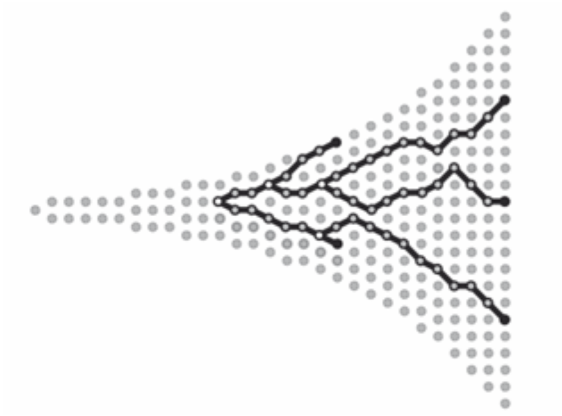


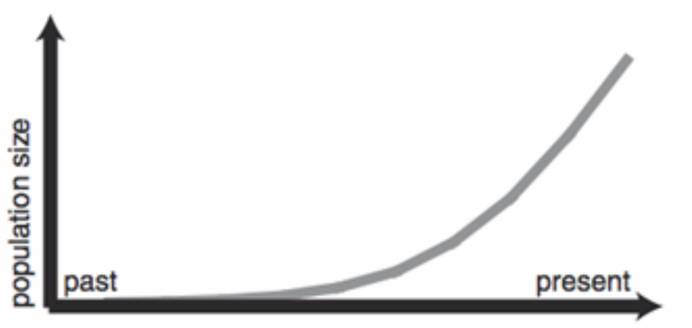
BEAST (Bayesian Evolutionary Analysis Sampling Trees)

aaaagcaaca aaaatgaagg caatactagt agttctgcta
tgcagacaca ttatgtatag gttatcatgc gaacaattcat
actagaaaag aatgtaacag taacacactc tgttaacctt
gaaactatgc aaactaagag gggtagcccc attgcatttg
ctggatcctg ggaaatccag agtgtgaatc actctccaca
tgtggaaaca tctagttcag acaatggaac gtgttaccca

+



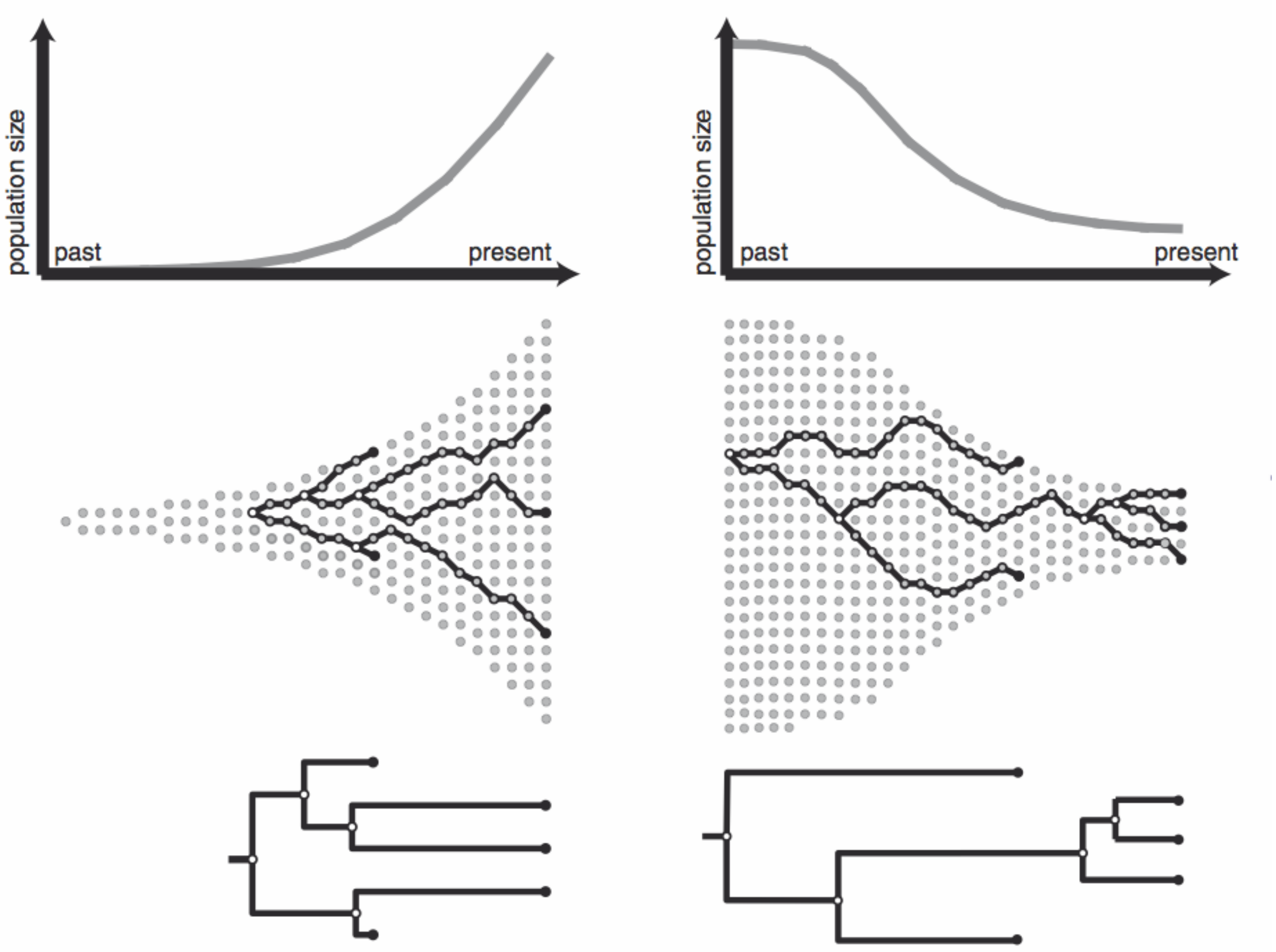




Bayesian
Skyline
Plot

BSP

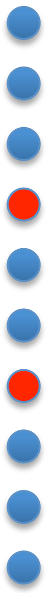
The coalescent with variable population size



Lemey et al, AIDS Rev 2006

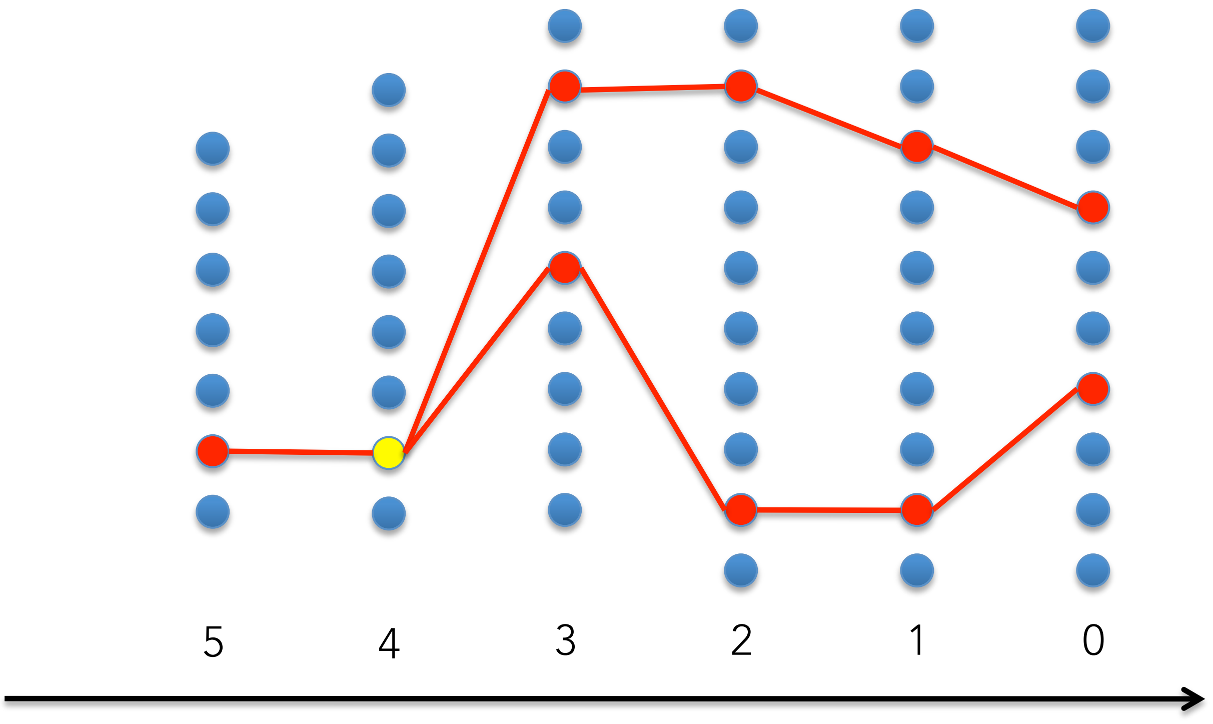
The coalescent

Consider 2 infected people randomly chosen from N$_t$ total infected people



The coalescent

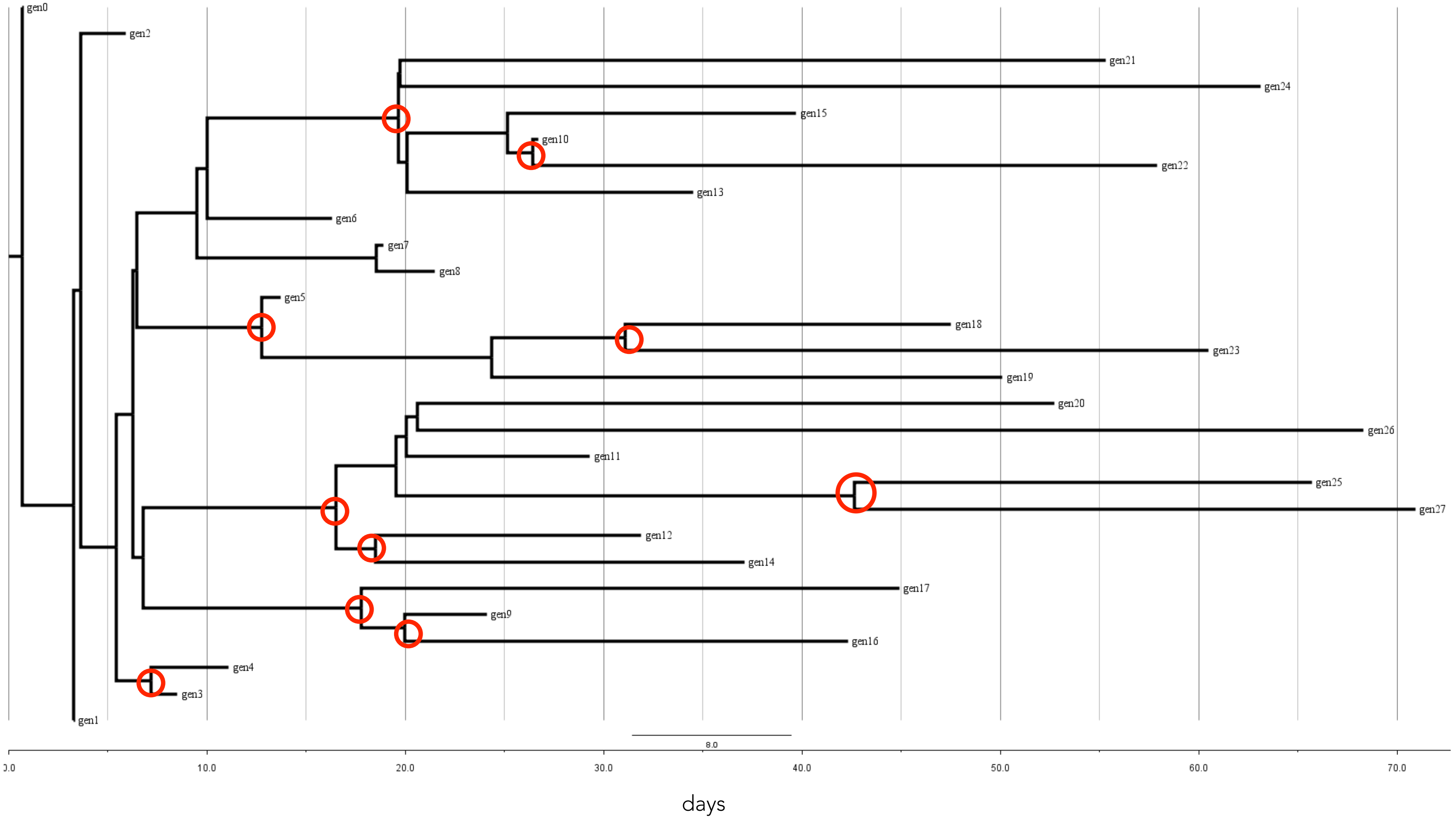
Probability they shared a common ancestor time T ago is $$\frac{1}{N_T}\prod_{x=1}^{(T-1)}\left(1-\frac{1}{N_x}\right)$$



time

Time is measured in generations of infection

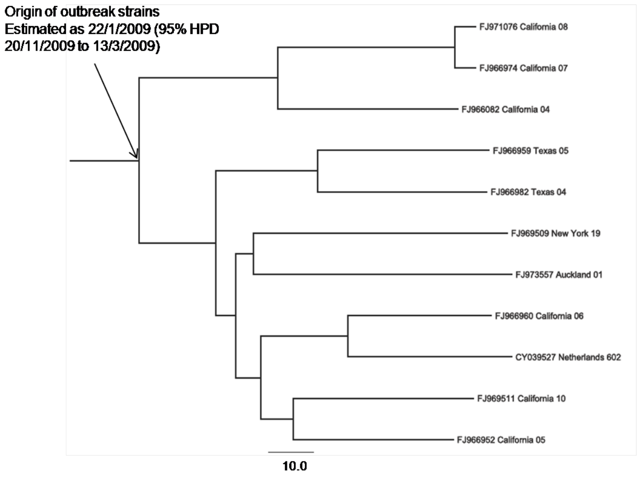
The coalescent can thus be used to reconstruct the number infected from a phylogeny…



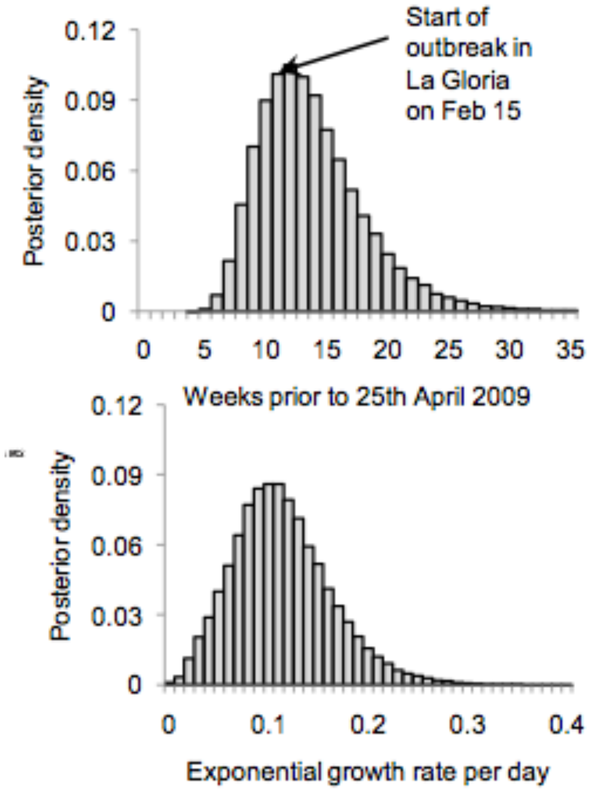
○ coalescent events

…provided a suitable model is used to relate mutations to time (*molecular clock model*)

Earliest H1N1pdm trees (4 May)



Origin of outbreak strains
Estimated as 22/1/2009 (95% HPD
20/11/2009 to 13/3/2009)

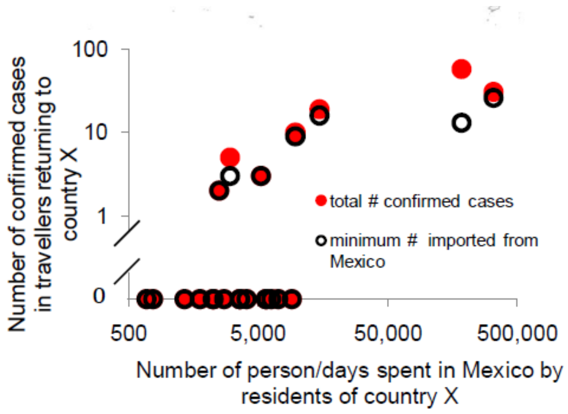


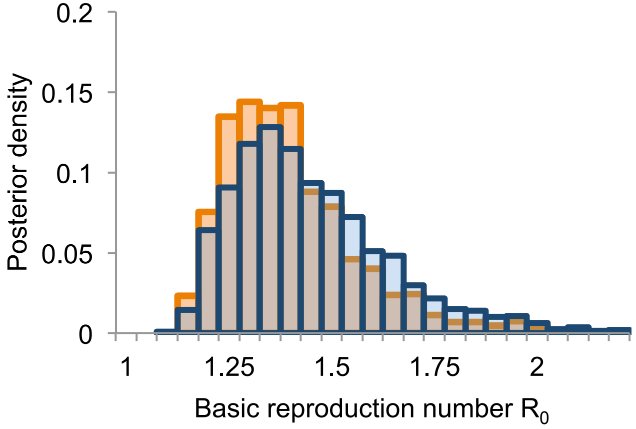
Start of
outbreak in
La Gloria
on Feb 15

Can already tell that the epidemic wasn’t growing very fast
Based on assuming exponential growth: strong assumptions → robust estimates

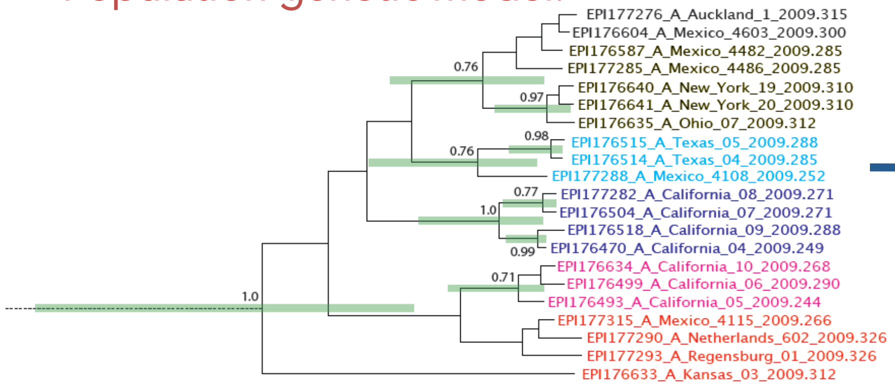
9 May: Updated trees

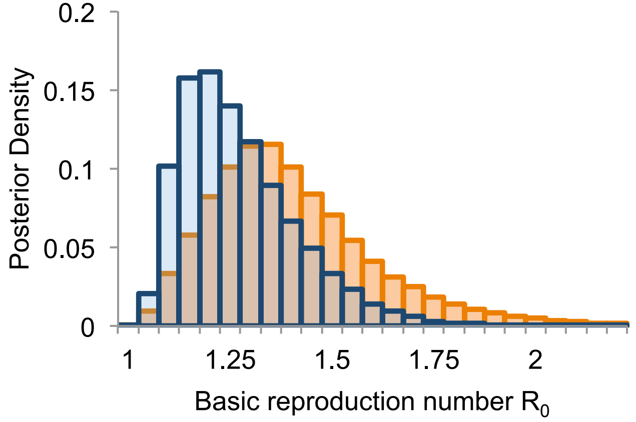
Epidemiological model:





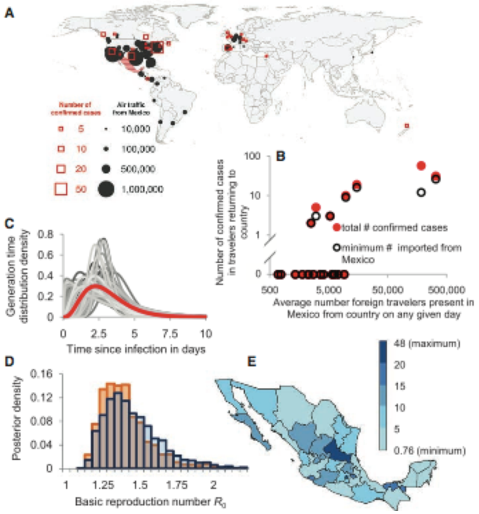
Population genetic model:


EPI177276_A_Auckland_1_2009.315
EPI176604_A_Mexico_4603_2009.300
EPI176587_A_Mexico_4482_2009.285
EPI177285_A_Mexico_4486_2009.285
EPI176640_A_New_York_19_2009.310
EPI176641_A_New_York_20_2009.310
EPI176635_A_Ohio_07_2009.312
EPI176515_A_Texas_05_2009.288
EPI176514_A_Texas_04_2009.285
EPI177288_A_Mexico_4108_2009.252
EPI177282_A_California_08_2009.271
EPI176504_A_California_07_2009.271
EPI176518_A_California_09_2009.288
EPI176470_A_California_04_2009.249
EPI176634_A_California_10_2009.268
EPI176499_A_California_06_2009.290
EPI176493_A_California_05_2009.244
EPI177315_A_Mexico_4115_2009.266
EPI177290_A_Netherlands_602_2009.326
EPI177293_A_Regensburg_01_2009.326
EPI176633_A_Kansas_03_2009.312
0.76
0.97
0.98
0.76
0.77
1.0
0.99
0.71
1.0

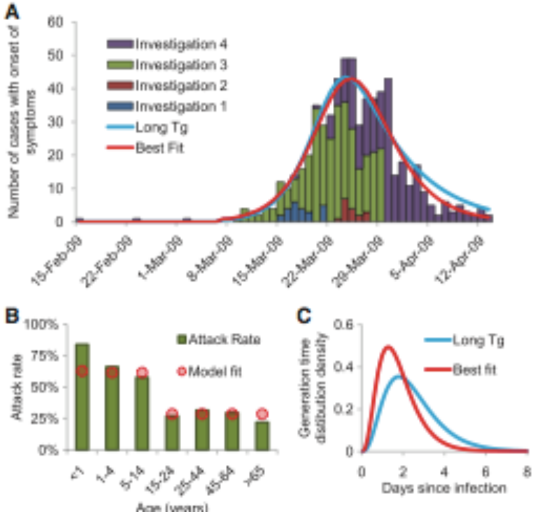


Orange – first iteration
Blue – updated – more effort to obtain 'random' unlinked sample

Fraser et al, Science 2009

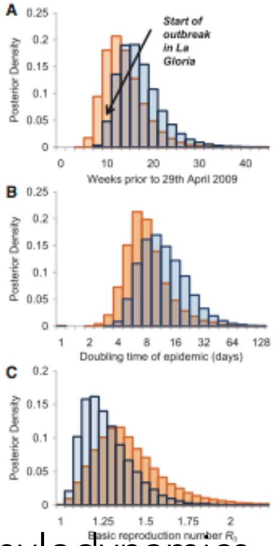


Back-calculation

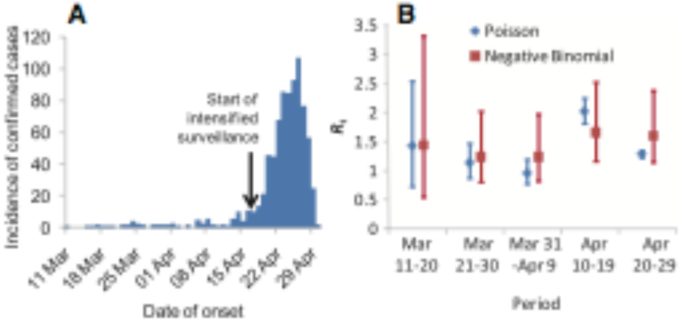


Outbreak investigation

Effective reproduction number



Phylodynamics

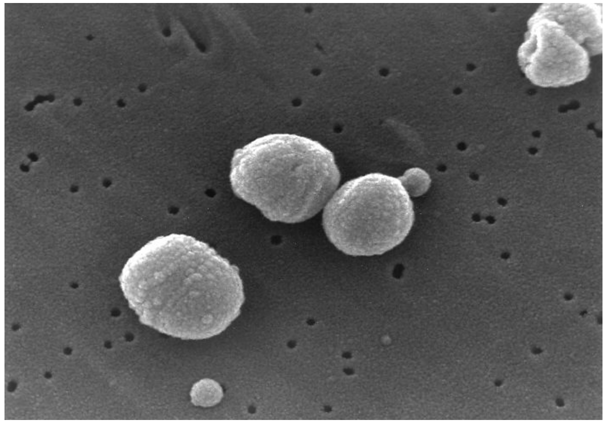


Time-series analysis

- The 2009 H1N1 influenza pandemic

- Tracking bacterial spread and evolution

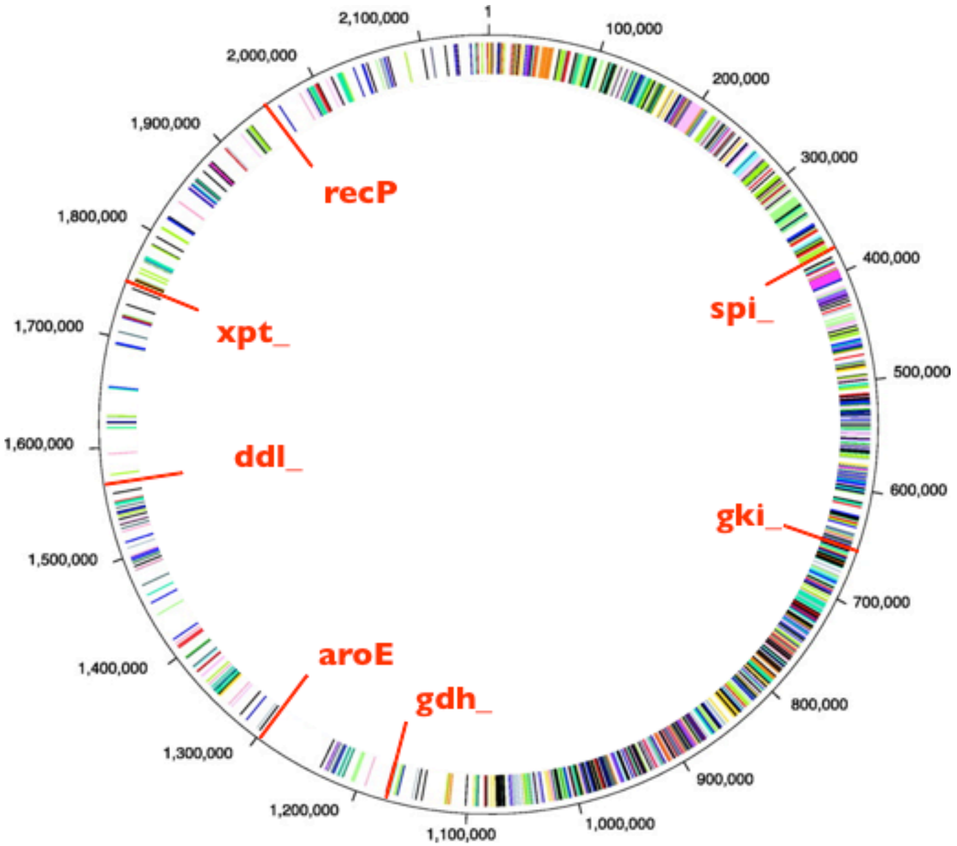
Streptococcus pneumoniae



- Gram +ve, commonly carried.
- Near ubiquitous in children.
- Causes otitis, pneumonia, invasive disease, meningitis, ….
- Causes 10% of all paediatric mortality.
- Diverse patterns of virulence and resistance.
- Antigenically diverse (92 serotypes).
- 7- and 13-valent vaccines now available.
- Naturally competent – recombinogenic.

Vaccine-caused Serotype Replacement

- Serotype replacement was complete in US in <10 years (Hanage et al, Epidemics 2009).
- Disease levels decreased approx 2/3 in US, but very little decline in the UK (HPA UK).
- Huge ecological perturbation, with unknown effect on antibiotic resistance and virulence factors
- What are implications for global roll-out?



Probing the core genome with Multi-Locus Sequence Typing (MLST)

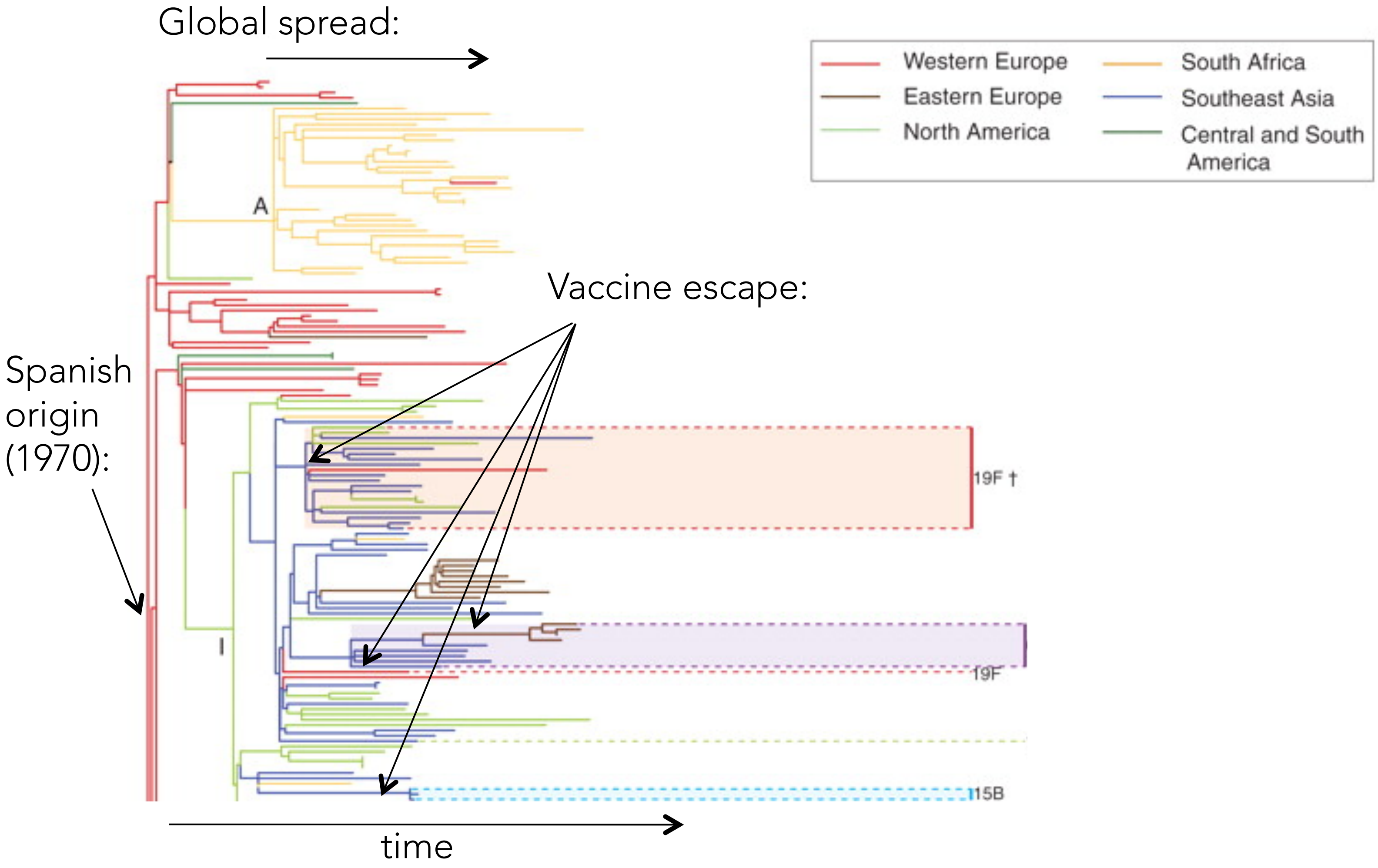
Now cheaper to sequence all 2,150,000 base pairs (whole genome with next generation sequencing) than 3,500 base pairs (MLST with capillary sequences).

A population genomic analysis

- Focus on PMEN-1 lineage:
  - Earliest recognised multi-drug resistant lineage of *Streptococcus pneumoniae* (penicillin, chloramphenicol, tetracycline, occasionally: fluoroquinolones & rifampicin, …)
  - Predominantly serotype 23F/ST81
  - Caused 40% of invasive disease in USA in 1990s
  - Member of the highly mosaic cluster (*based on BAPS/MLST analysis*)

- Full genomes from 241 isolates

Rapid Pneumococcal Evolution in Response to Clinical Interventions



Global spread:

Vaccine escape:

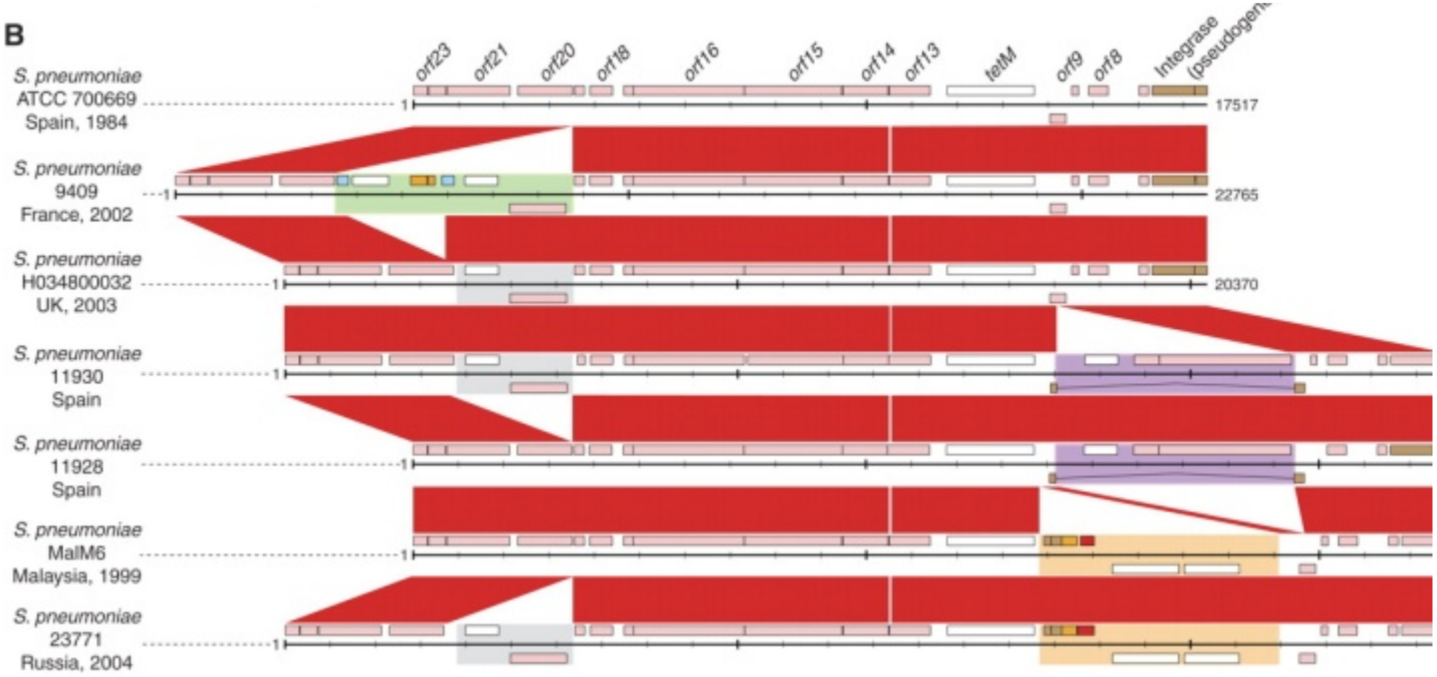
Spanish origin (1970):

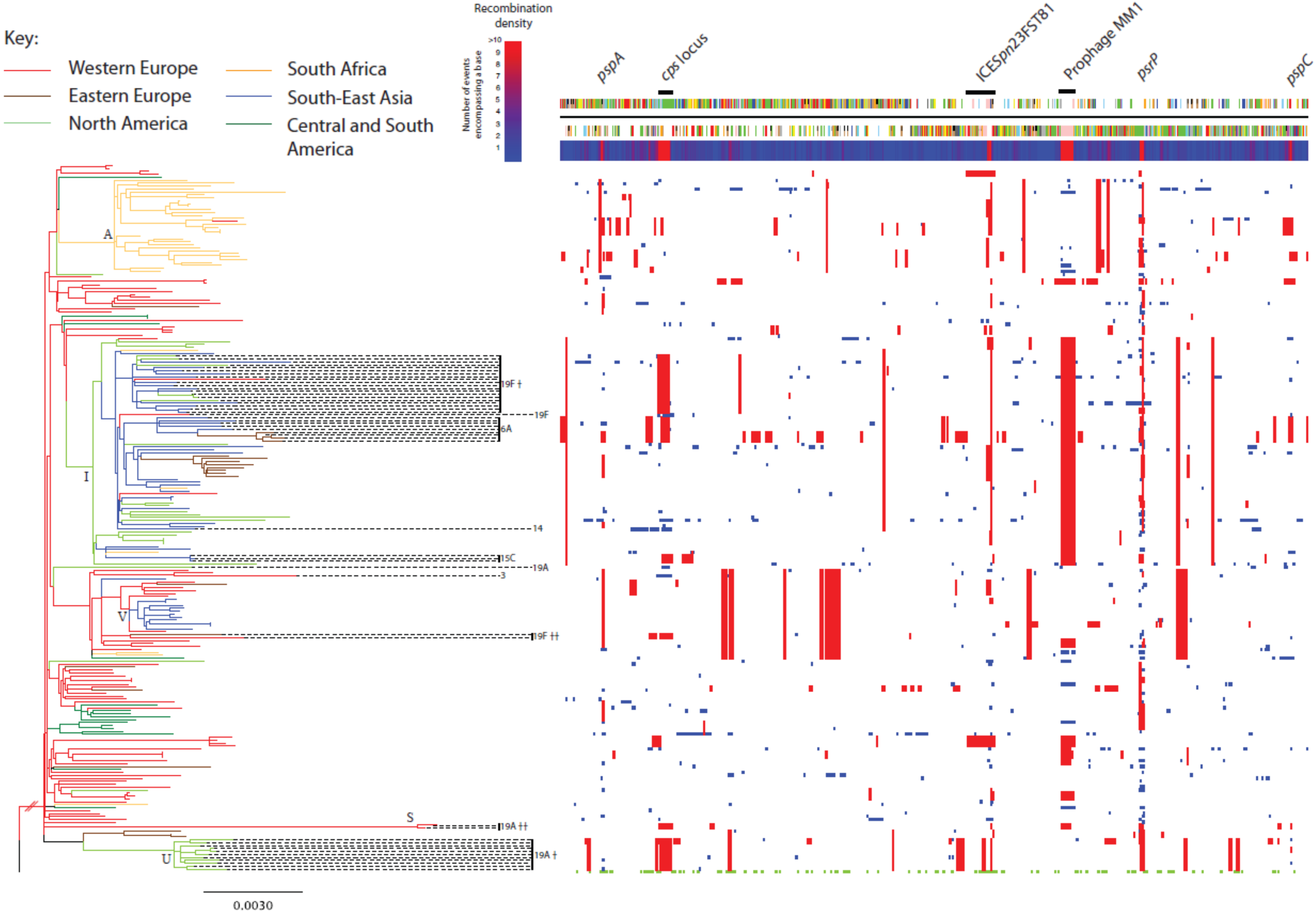
time

Croucher et al, Science 2011

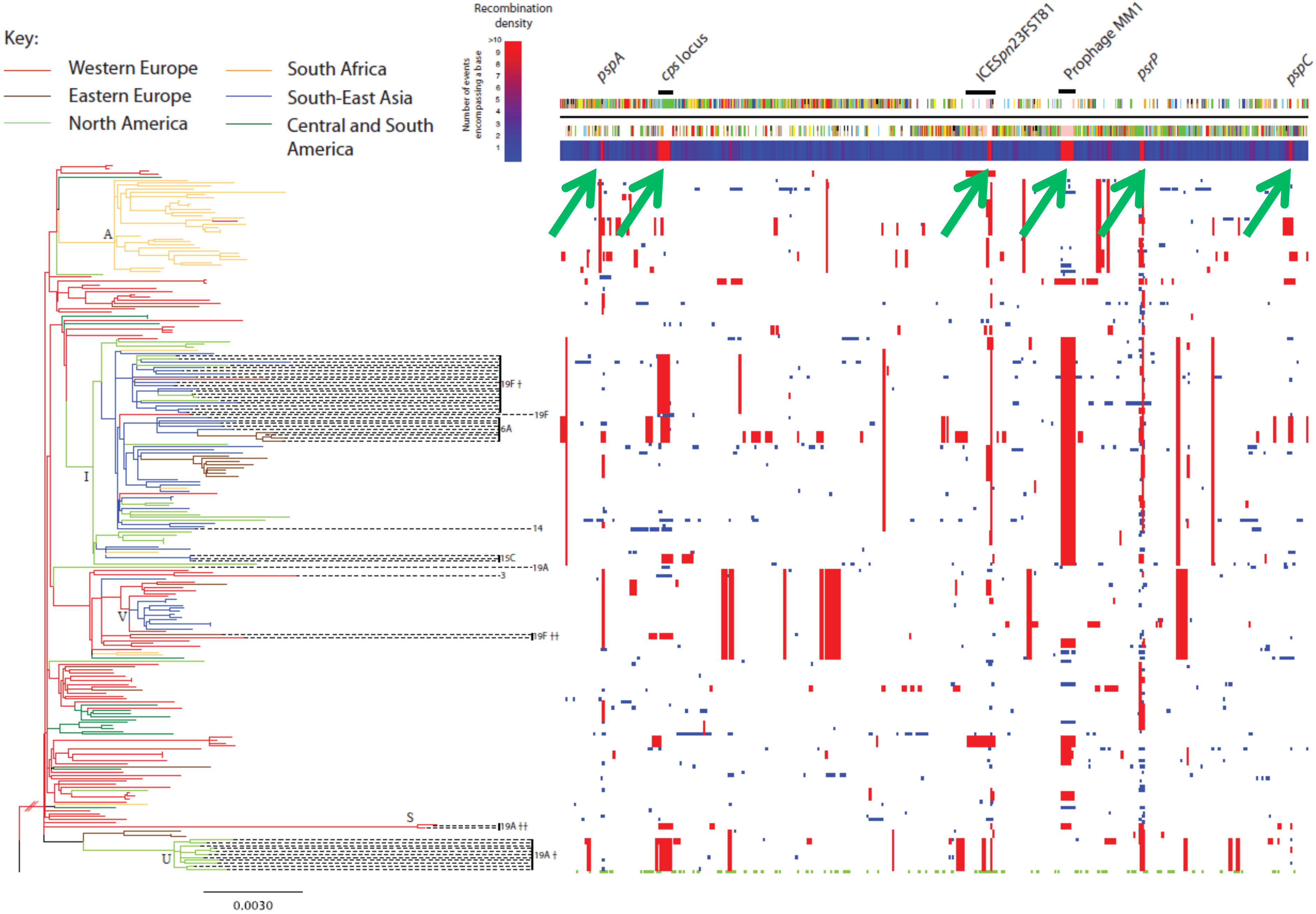
Multiple acquisition & loss of antibiotic resistance

- Whole lineage is resistant to penicillin, chloramphenicol & tetracycline.
- Fluoroquinolone resistance mutations acquired & lost, seemingly random.
- Macrolide resistance cassettes acquired repeatedly through horizontal gene transfer

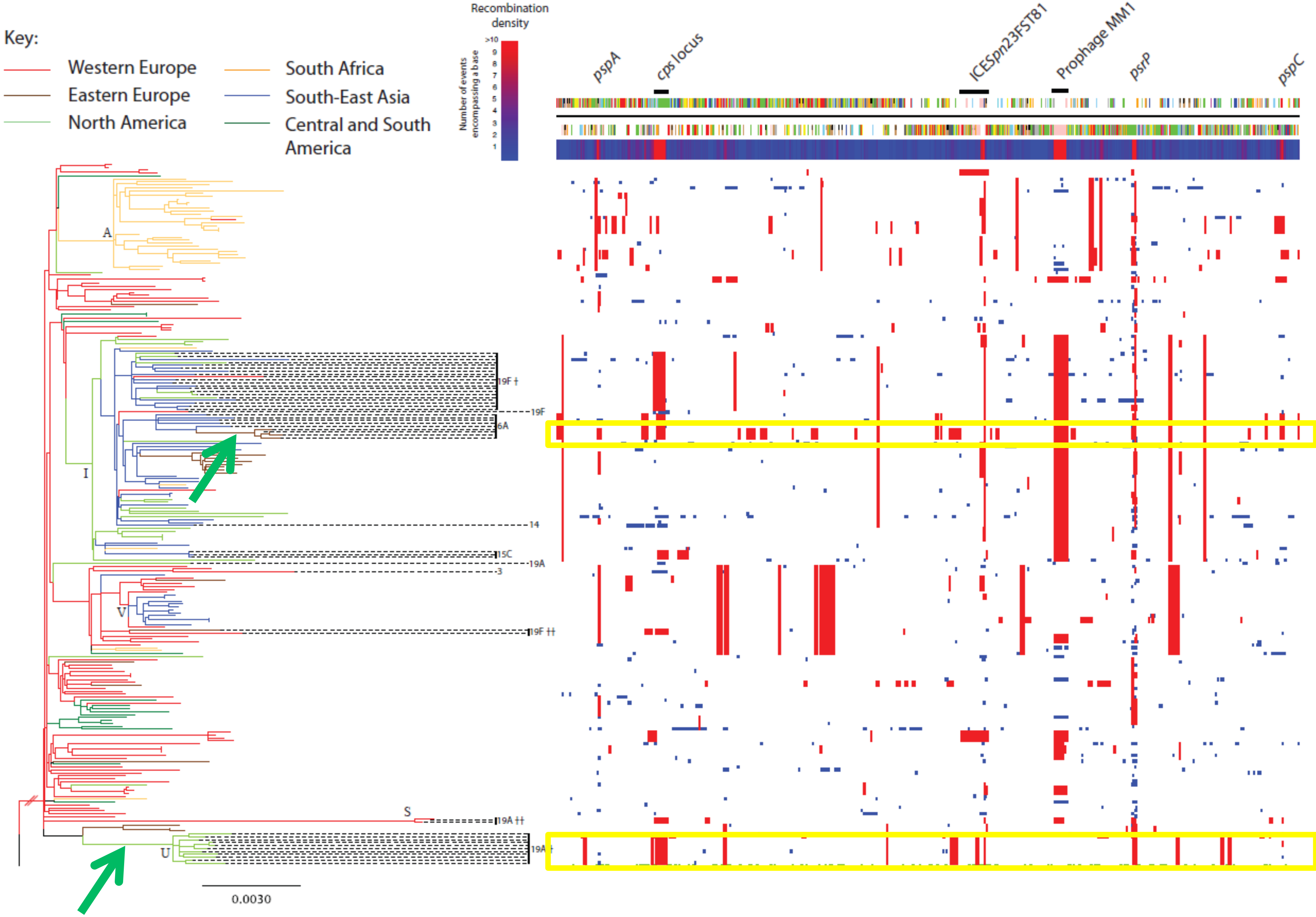




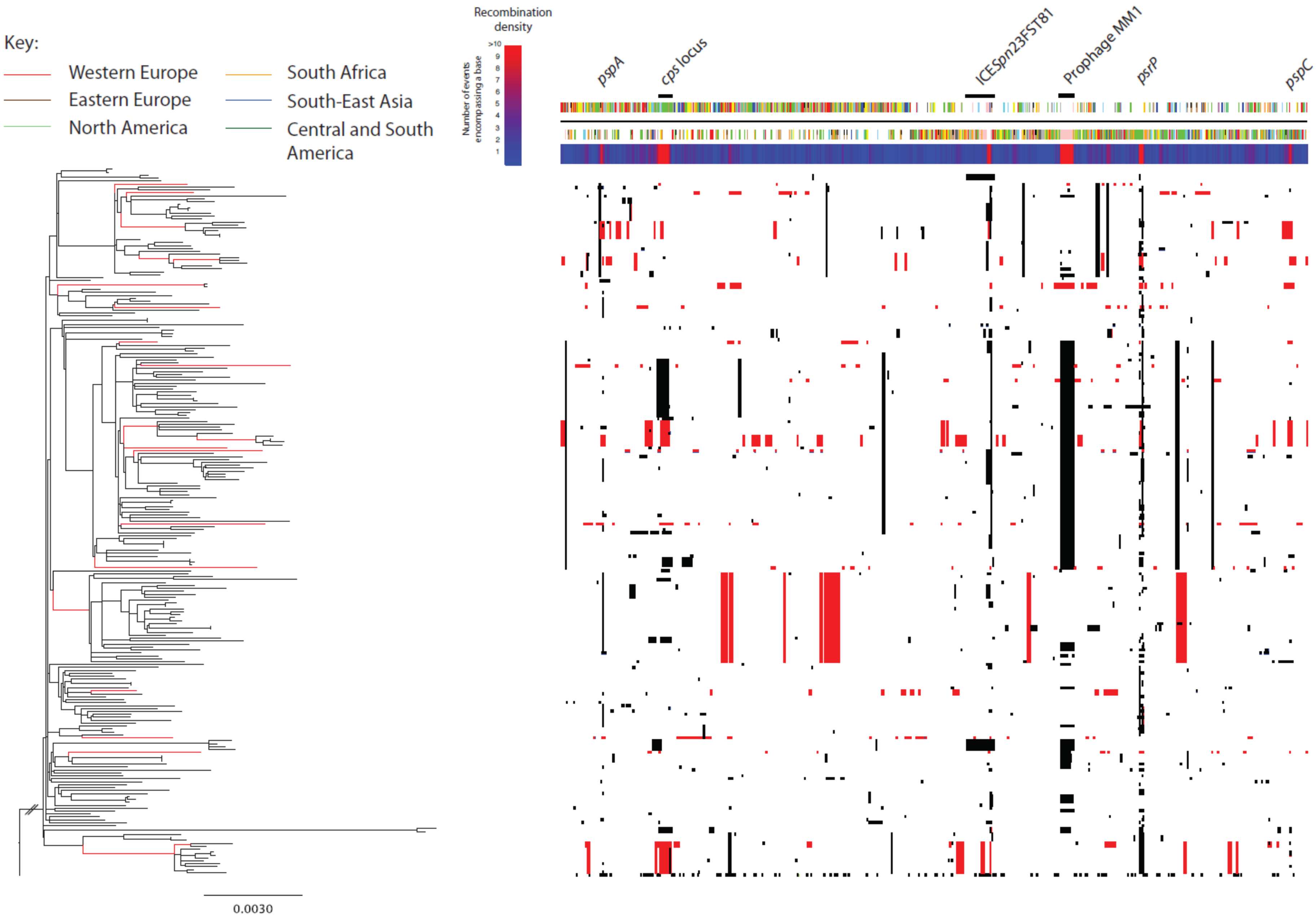
90% of polymorphisms acquired by recombination, covering 75% of genome (identified by SNP density)



Recombination not uniform along genome, but marked by hotspots

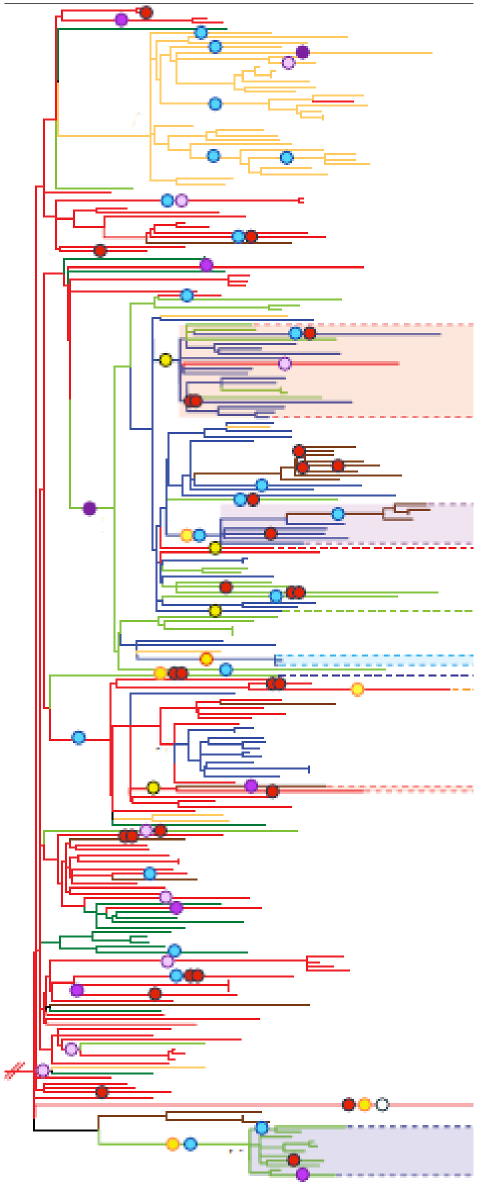


Some suggestion of lineages having experienced hyper-recombination?



Hyper-recombination accounts for over 50% of polymorphisms – role still unclear

Phylogeny reveals a rich adaptive history



Blue: hyper-recombination events
Yellow: serotype switches
Purple: Acquisition of macrolide resistance cassettes
Red: Fluoroquinolone resistance mutations
White: Abrogation of competence.

With whole genome data, we can appreciate that genetic, genomic and selective events all occur concurrently, not in isolation. This will require new theory.

Summary

- PMEN-1 lineage defined by acquisition of an accessory multidrug-resistant gene (ICE), and rapid spread since 1970
- Further changes in lineage driven by rapid switching in accessory genome (inc. antibiotic resistance and vaccine escape)
- New understanding of mechanisms of recombination

Conclusions

- Sequencing is cheap, and will soon be standard.
- Most infections will be 'typed'.
- If nothing else, cheapest & quickest way of determining antibiotic resistance profile.
- Interpretation & analysis will remain challenging.
- Data storage & sharing, and linkage to meta-data will be extremely challenging, but necessary.