The core genome and beyond: comparative bacterial pathogenomics and functional gene analysis in foodborne pathogens

PJ Biggs[1,2,3], T Blackmore[4], AD Reynolds[5], AC Midwinter[2], J Marshall[2] and NP French[1,2]

1. Allan Wilson Centre for Molecular Ecology and Evolution,
2. [m]EpiLab, Institute of Veterinary, Animal and Biomedical Sciences,
3. New Zealand Genomics Ltd (NZGL – as Massey Genome Service)
   all at Massey University, Palmerston North, New Zealand.
4. Capital and Coast District Health Board, Wellington, New Zealand.
5. AgResearch, Hopkirk Research Institute, Palmerston North, New Zealand.

OIE Collaborating Centre for Veterinary Epidemiology and Public Health

mEpiLab.massey.ac.nz

Acknowledgements

- members of mEpiLab, IVABS, Massey University
  - Lynn Rogers
- Massey Genome Service (MGS; a part of NZGL)
  - Lorraine Berry & Mauro Truglio
- Dept. of Zoology, University of Oxford, Oxford, UK
  - Keith Jolley & Martin Maiden
- Funding
  - Wellington Hospital Laboratory Education and Research Fund
  - IVABS, Massey University

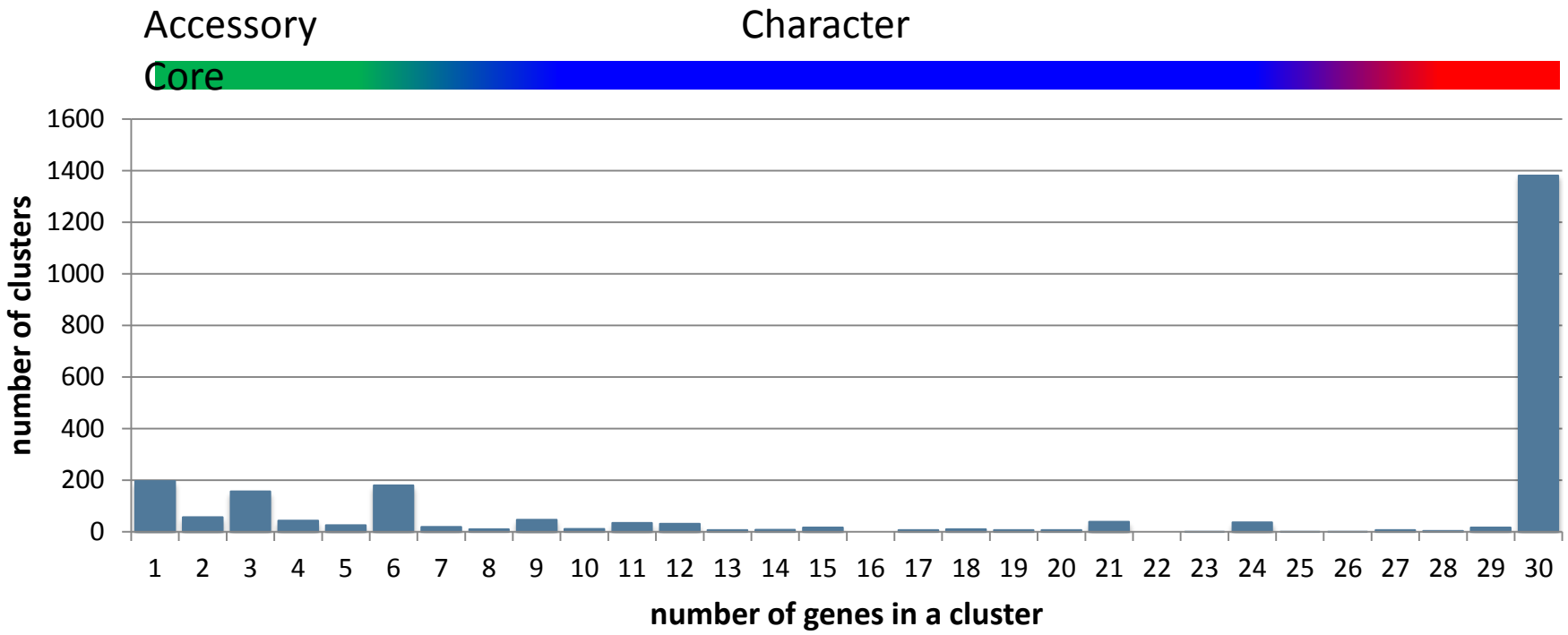
OIE Collaborating Centre for Veterinary Epidemiology and Public Health

Overview

- The core genome
- Invasive isolates of *Campylobacter spp*.
- Comparative analysis
  - rMLST, core SNPs
  - Core genome and beyond
- COG classification for functional analysis
  - “biological weighting”

The core genome

- Genes present in ALL strains are core genes
  - A proxy for analysing evolution
- Genes present in a subset of strains are accessory genes
  - Virulence?
  - Niche adaptation?



Clinical observations

- 10 invasive isolates from Wellington Hospital from 2010 to 2012:
  - Aged 19 to 89 years
  - 6 presented with diarrhoea
  - Others
    - Headache
    - Prosthetic hip infection
    - Exacerbation of chronic pulmonary disease
  - 9 samples from blood
  - 1 sample from joint aspirate
- Outcomes:
  - 6/10 treated with oral ciprofloxacin
  - 43 year old died
  - One had episode 4 years earlier

Campylobacter and bacteraemia

- Campy bacteraemia appears to originate from acute colitis
- Bacteraemia relatively subtle part of illness
  - Only one person with sepsis syndrome
- Complicated infection with discitis or prosthetic joint infection suggestive of acute seeding of previously abnormal tissue

- Bacteraemia population rate = 7.6 per million person years
  - Denmark = 2.9 per million person years[1]
- Bacteraemia:enteritis ratio = 0.4%
  - Finland = 0.3%[2]

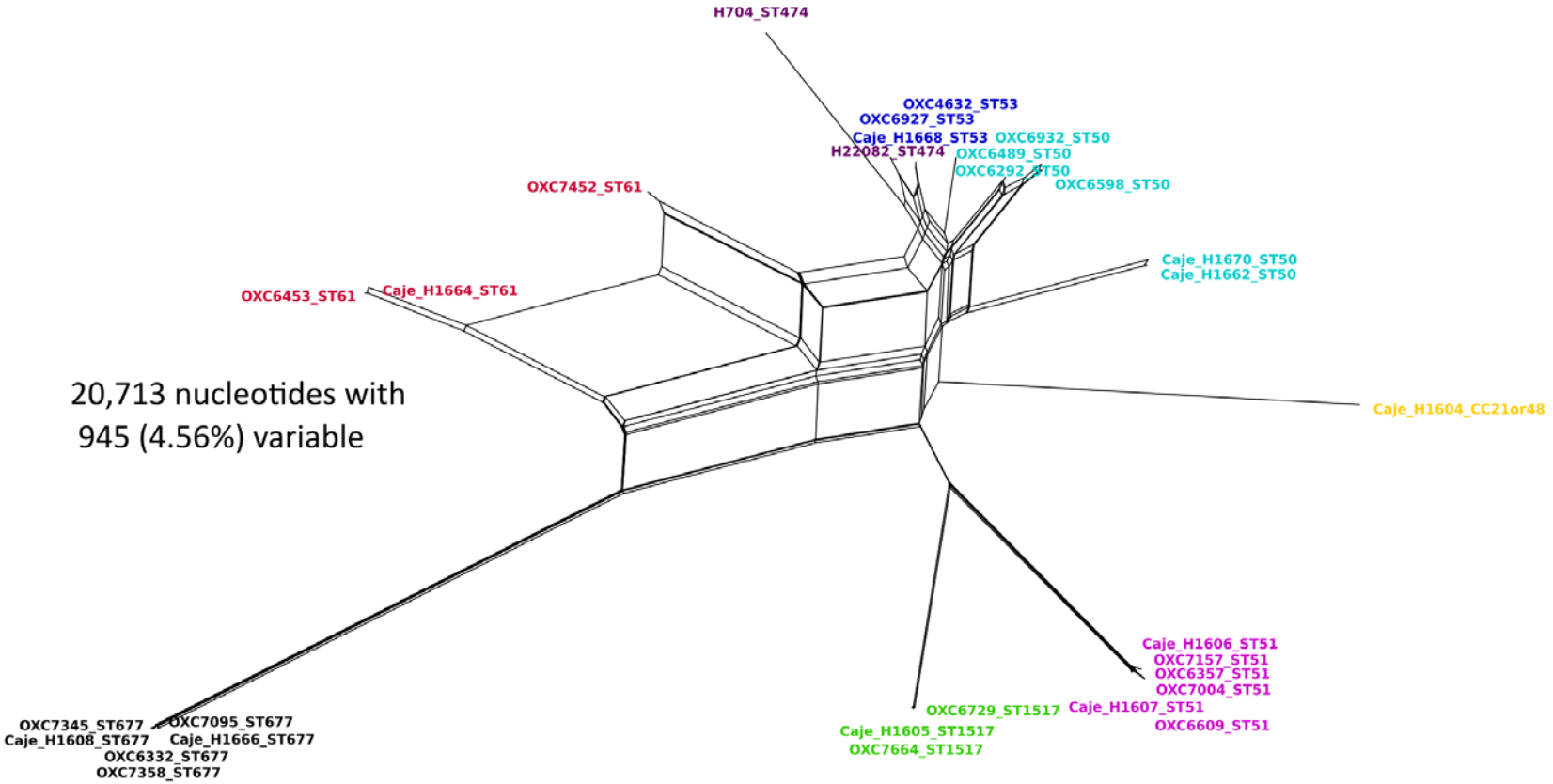
1: Nielsen et al. Clin Microbiol Infect 2010; 16: 57–61
2: Feodoroff et al. Clinical Infectious Diseases 2011;53(8):e99–e106

Genome summaries

- MLST results:
  - 5 clonal complexes found:
    - 2 from CC677, 2 from CC443, 4 from CC21, 1 from CC354 and 1 from CC61
- For comparison, 2 random isolates taken per ST from the Oxford surveillance project (OXC)
- Genomes:
  - When compared to the 20 random chosen OXC genomes:
  - No overall difference in:
    - Overall genome size: 1.654 Mb (± 22kb) vs. 1.659 Mb (± 14kb)
    - GC content: 30.38% (± 0.06%) vs. 30.38% (± 0.05%)
    - Number of predicted genes: 1705 (± 30) vs. 1714 (± 18)
  - Difference in:
    - tRNA number: 42.1 (± 1.16) vs. 37.3 (± 2.21)
    - but these are probably due to *de novo* assembly issues

Results – ribosomal MLST (rMLST[3])

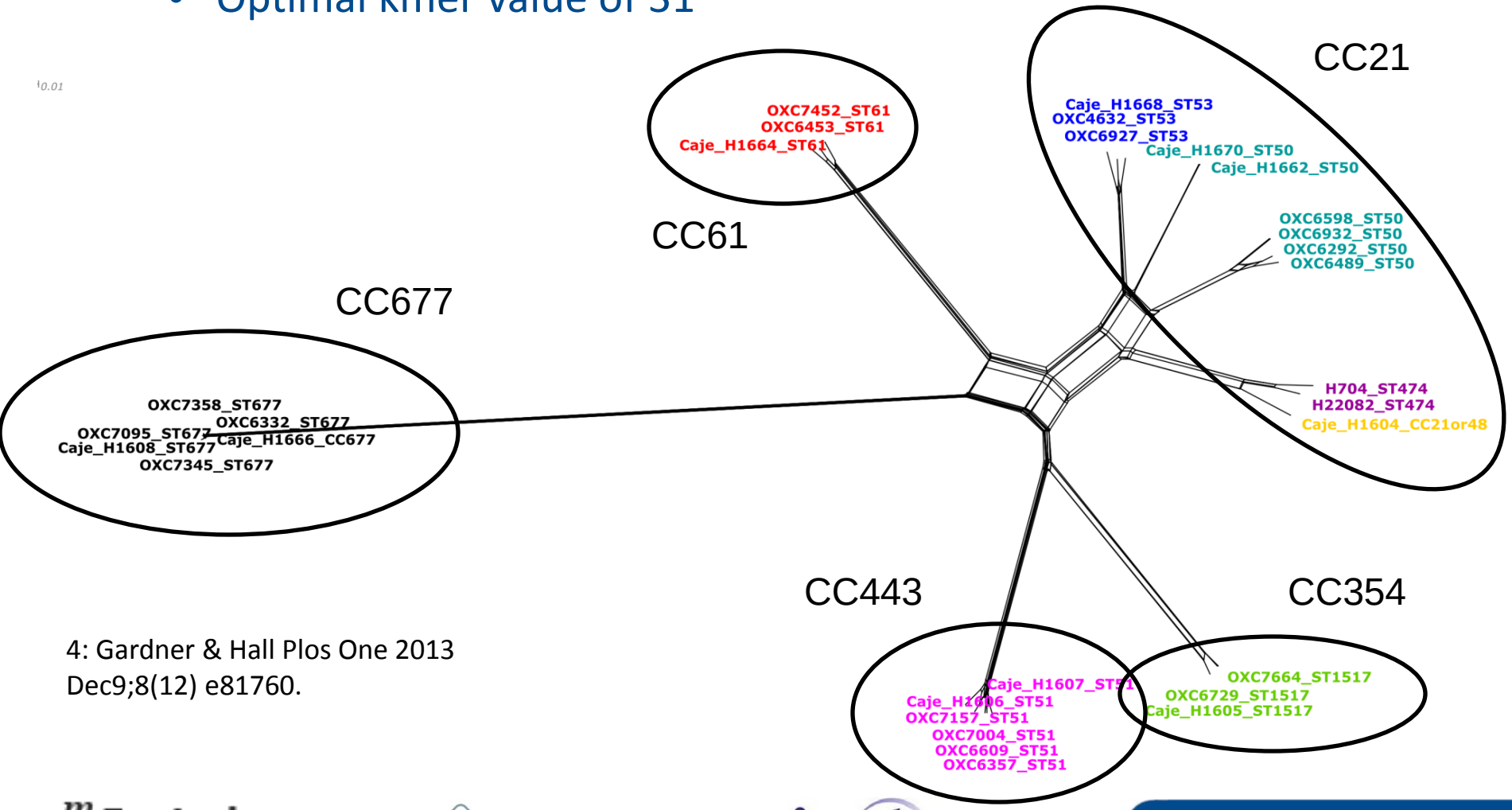
- 52 genes (no *rpm*D in order Campylobacterales)



3: Jolley et al. Microbiology 2012 Apr;158(Pt 4):1005-15

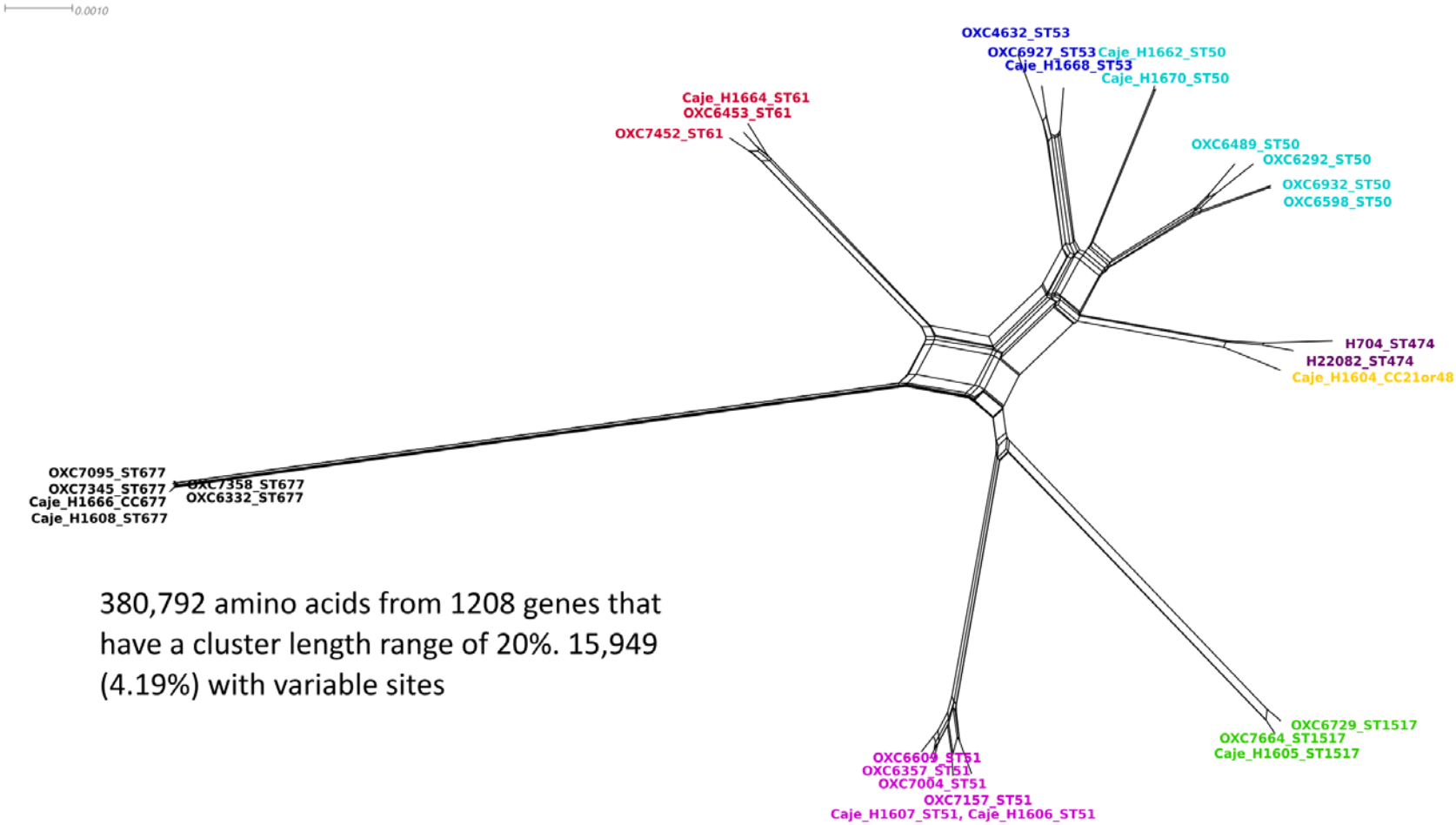
Results – core SNPs

- e.g. kSNP2[4] analysis based on 13,260 core SNPs
  - Optimal kmer value of 31



4: Gardner & Hall Plos One 2013 Dec9;8(12) e81760.

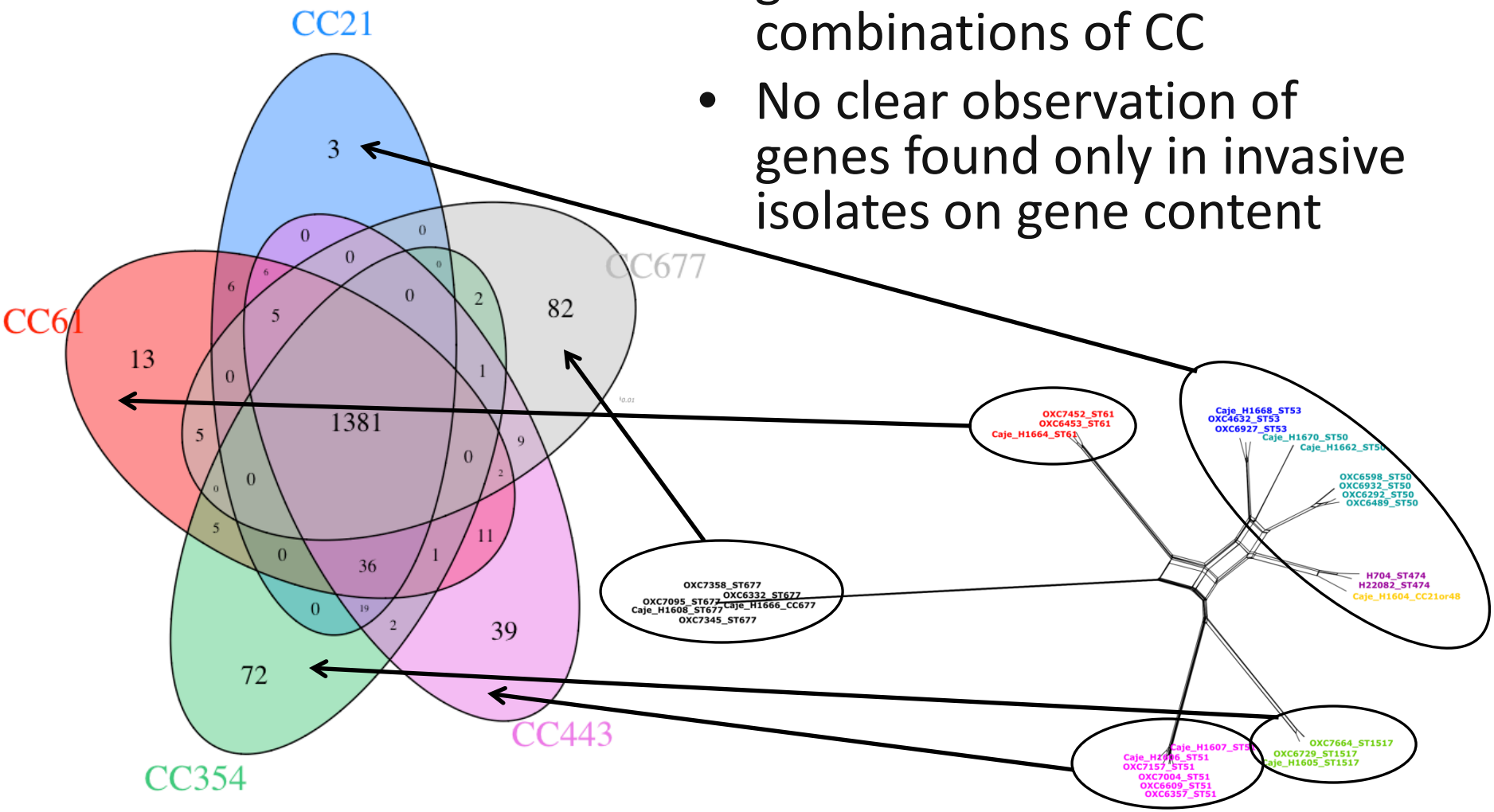
Results – core genome



380,792 amino acids from 1208 genes that have a cluster length range of 20%. 15,949 (4.19%) with variable sites

Beyond the core

- From OrthoMCL, different genes are found in different combinations of CC
- No clear observation of genes found only in invasive isolates on gene content



Functional analysis by COGs

- Clusters of Orthologous Groups – COGs
- 1997 - concept by Koonin et al. as a framework for functional and evolutionary genome analysis[5]
- Hierarchical process wherein:
  - Gene function defined into 26 codes (A to Z)
    - e.g. P = 'Inorganic ion transport and metabolism'
  - Each of the 4632 curated orthologous group of genes gets a COG classification, and with it a functional code
    - e.g. COG2847 = 'Copper(I)-binding protein'
- Idea: use COGnitor software to analyse the function of predicted genes

5: Tatusov et al. Science 1997 Oct 24;278(%338):631-7

Whole genome analysis by function

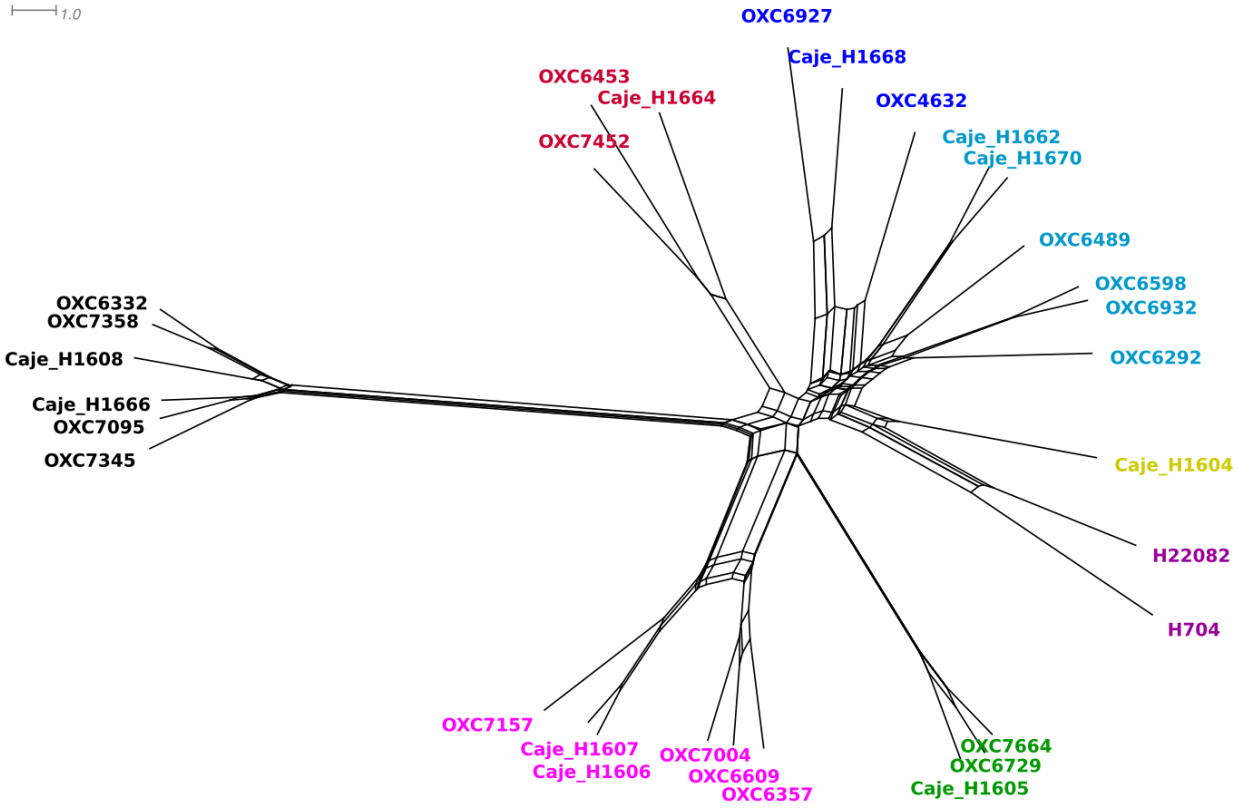
- Allows analysis of complete dataset – i.e. not just the core genome
- Analysis as a two step process:
  - For each isolate under investigation:
    - BLAST the PROKKA gene predictions against COG database of ~1.8 million genes using PSIBLAST
    - Parse output with COGnitor and generate counts for each COG in the genome
  - For all isolates:
    - Detect COGs where there is a difference in number across isolates
    - Use these values to generate a distance matrix
    - Convert to Nexus format and view resulting tree as a NeighborNet in SplitsTree

Insights from function?

- Is the difference between absence/presence more important than that between the number of paralogues?
  - i.e.: should 1,0,0,1,2,5 be recoded as:
    - binary (1,0,0,1,1,1),
    - something else e.g. 1,0,0,1,1.3,1.7?
- Can look at this by changing weighting of the data in the distance matrix
- Values of weight *w* can vary between 0 and 1:
  - $w = 0$
    - Data is binary, i.e. absence or presence of a COG member
  - $w = 1$
    - Data has its values, i.e. 4 paralogues of a COG member provide 4x more weight than 1 member
- What is the value of *w* that is biologically relevant:
  - 0, 1, or somewhere in-between?

COG function – Euclidean (p = 2)

- What is the effect on genome datasets?
- Values of *w* varied from 0 to 1
- Correlation with ST rather than disease state



COG summary:
Absent: 3282
Present: 866
Variable: 484

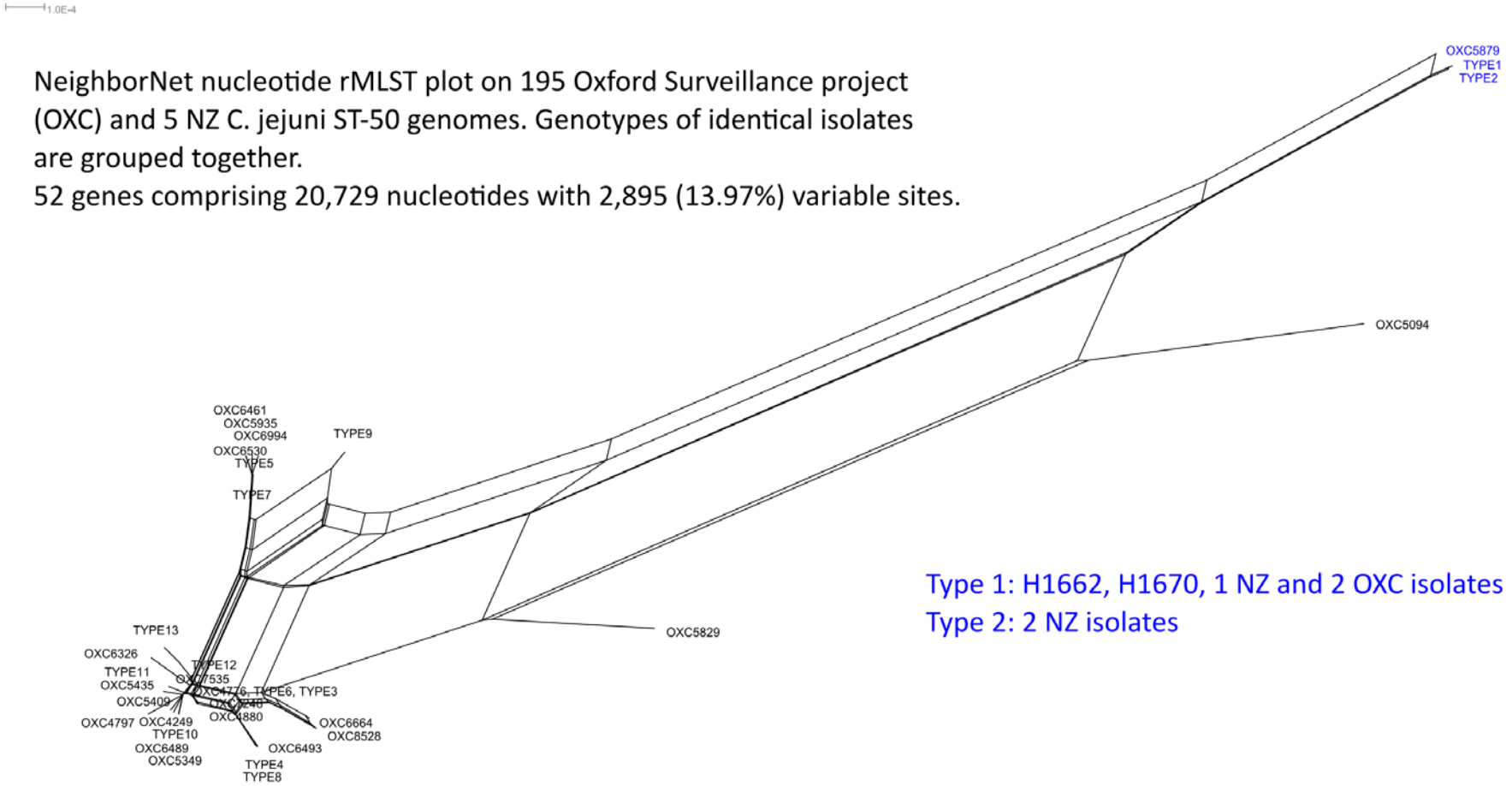
***w* = 1.00**

Summary

- 10 NZ invasive strains whole genome sequenced and 5 clonal complexes found
- Comparison with 20 OXC gastroenteritis isolates:
  - Isolates grouped by sequence type rather than phenotype at levels of:
    - rMLST, core SNP, core genome and COG function
- Differences in phenotype of invasive isolates are not at overall level of genome
  - Further work required to investigate the role of sequence differences

ST50 – slightly different

NeighborNet nucleotide rMLST plot on 195 Oxford Surveillance project (OXC) and 5 NZ C. jejuni ST-50 genomes. Genotypes of identical isolates are grouped together.
52 genes comprising 20,729 nucleotides with 2,895 (13.97%) variable sites.



Type 1: H1662, H1670, 1 NZ and 2 OXC isolates
Type 2: 2 NZ isolates

Summary

- 10 NZ invasive strains whole genome sequenced and 5 clonal complexes found
- Comparison with 20 OXC gastroenteritis isolates:
  - Isolates grouped by sequence type rather than phenotype at levels of:
    - rMLST, core SNP, core genome and COG function
- ST50 isolates from NZ show slight difference to initial OXC isolates
  - Further rMLST analysis showed most OXC ST50 isolates (193/200) are unlike NZ isolates
- Differences in phenotype of invasive isolates are not at overall level of genome
  - Further work required to investigate the role of sequence differences