NIHR | Health Protection Research Unit in Gastrointestinal Infections at University of Liverpool

LineageCapture: Phylogenetic identification of bacterial clade members excluded by SNP clustering

Matthew P. Moore[1,2], Paolo Ribeca, Xavier Didelot[1,2]

1. School of Life Sciences, University of Warwick, United Kingdom
2. Department of Statistics, University of Warwick, United Kingdom
3. United Kingdom Health Security Agency

Background

As microbial genomes accumulate, particularly due to routine surveillance of bacterial pathogens, necessary comparisons scale quadratically. SNP cluster-based methods have expanded including core genome MLST allele distances and reference-based single linkage clustering (SLC) of SNPs. The advantage of these approaches is in their efficiency. SLC does incur systematic errors however, such as chaining, which can result in poor internal cluster validity. Conversely, nomenclature constraints can prevent cluster-merging.

It's also unclear to what extent SNP distances correlate with phylogenetic relatedness. The power of these approaches depends upon the accurate approximation of recent ancestry, of which phylogenetic methods are the most accurate. However, phylogenetic models can be prohibitively time and compute intensive.

How then, do we reliably capture all members of an outbreak lineage at scale?

Results

Polyphyly with cgMLST

In figures 1 and 2 we mapped cgMLST allelic difference SLC's (HierCC) at thresholds 0, 2, 5, 10, 20, 50 and 100. For the *Streptococcus pyogenes* set we observe polyphyly up to HC20 and for *Clostridioides difficile* HC50.

If our lineage of interest is at the HC5 level, we cannot know *a priori* which higher level clustering will capture all of those monophyletic with our HC5 genomes. Further, inclusion of all sequences of a higher-level cluster, within which ours is nested, is liable to include too many sequences of the major lineage. At this stage we have lost the benefit of SLC and retain the problem of requiring the full phylogeny.

With an example HC0 cluster of *C. difficile (*HC2.426), we need to gather all HC10.17 genomes to recover the lineage (figure 2). As the majority of genomes are compared at this level, we lose the benefit of SLC. Additionally, we only know that this is the smallest threshold sufficient because we have the full tree.

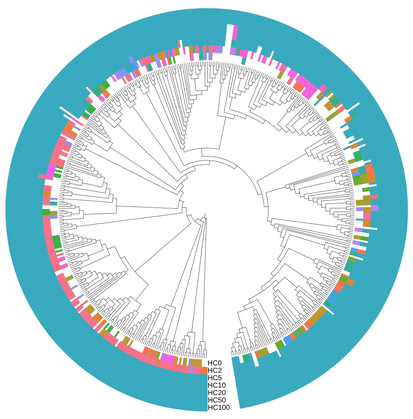
***LineageCapture* accuracy**

By simulating bacterial populations (phylogenies and alignments) we demonstrate the incongruency between SNPs and relatedness even when sequencing or bioinformatic errors are absent. With *LineageCapture* we attempt to gather all children of the cluster member MRCA and judge sensitivity and specificity accordingly (figure 3).

The algorithm

LineageCapture defines a lineage of interest by a SLC. The natural population to which the SLC belongs may be defined by MLST or MLST-level SNP clustering. All genomes are aligned to a reference (linear scaling). Subsets of this alignment are then iteratively combined with the SLC member sequences. Monophyly of non-members may then be tested. Once all genomes have been analysed the process repeats with the newly recovered sequences included in the SLC members.

Polyphyletic allele clusters



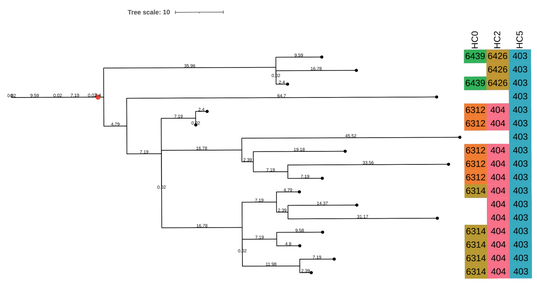
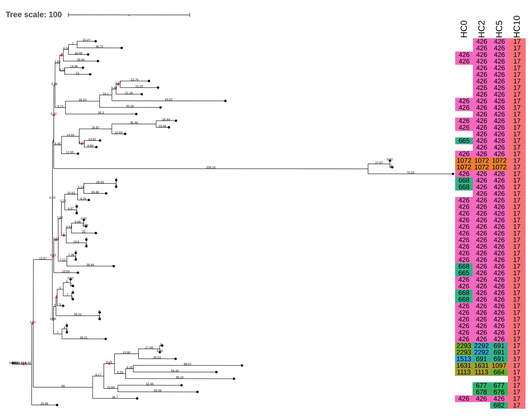


Figure 1. (Left) circularised approximate maximum likelihood cladogram of 500 *Streptococcus pyogenes* genomes. The genomes were mapped to the reference genome and a relaxed core genome alignment extracted whereby sites were to have no more than 5% of taxa with a gap or an ambiguous base. Colour rings represent HierCC cgMLST allele clusterings mapped onto the tree for allele single linkage threshold levels 0, 2, 5, 10, 20, 50 and 100 (inner to outer). A unique colour was designated for each cluster identifier with at least 2 members. (Right) an example subtree from the larger sample tree with threshold levels 0, 2 and 5. For this subtree levels 10, 20, 50 and 100 were monophyletic. Branch lengths are shown in SNPs and nodes with bootstrap confidences <75% marked with a red circle



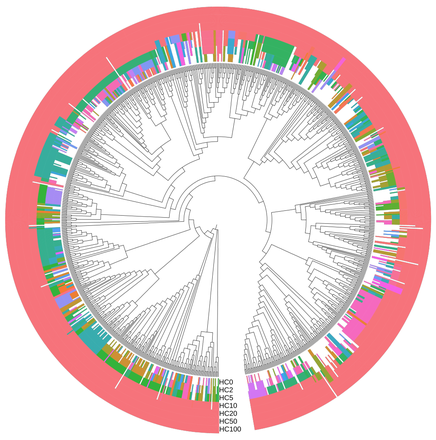


Figure 2. (Right) circularised approximate maximum likelihood cladogram of 999 RT017 *Clostridioides difficile* genomes. The genomes were mapped to the M68 reference genome and a relaxed core genome alignment extracted whereby sites were to have no more than 5% of taxa with a gap or an ambiguous base. Colour rings represent HierCC cgMLST allele clusterings mapped onto the tree for allele single linkage threshold levels 0, 2, 5, 10, 20, 50 and 100 (inner to outer). A unique colour was designated for each cluster identifier with at least 2 members. (Lower) an example subtree from the larger sample tree with threshold levels 0, 2, 5 and 10. For this subtree levels 20, 50 and 100 were also monophyletic. Branch lengths are shown in SNPs and nodes with bootstrap values <75% have a red circle. Horizontal scaling of the subtree is 1.5x

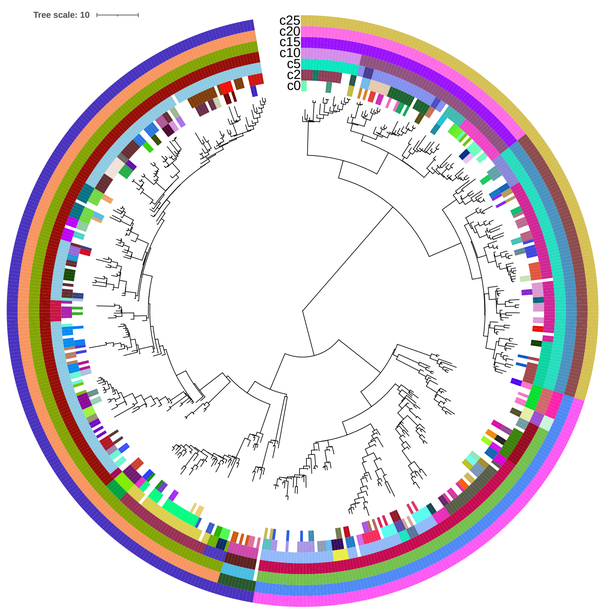
Discussion

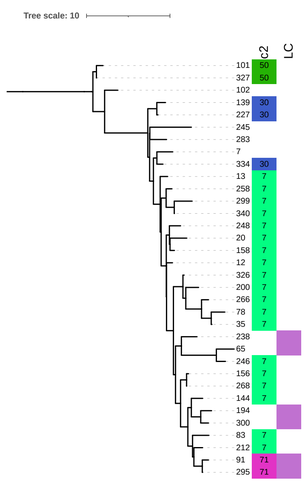
With an outbreak defined by a SNP cluster it's possible to check against global databases at scale. In doing so we can gain valuable information from outbreak members. However, SNP clustering can be incongruent with the phylogeny by incorrectly excluding outbreak members. In order to gather these additional genomes, we have two options: 1) increase the SNP threshold, 2) reconstruct the entire phylogeny. The first option may be precluded by unreliable clustering at greater discrimination than major lineages and the latter by prohibitive computational demands.

With *LineageCapture* we establish the first method to reliably gather these excluded genomes. Scalability is linear with the number of genomes (divided by batch size) in the major lineage and entirely parallelisable.

Here, we used the example of a simulated population. We generated ideal data for SNP clustering with no recombination, no difference in genome length and a known (true) phylogeny. Nonetheless, clustering mappings were polyphyletic. With *LineageCapture* we were able to gather the vast majority (98.65%) of excluded lineage members whilst only falsely including 6% false positives. Crucially, these false positives were always derived from sister taxa. In ongoing work we're evaluating the impact of recombination

LineageCapture examples





- Sensitivity and specificity across all clusters was 98.65% and 93.97% respectively

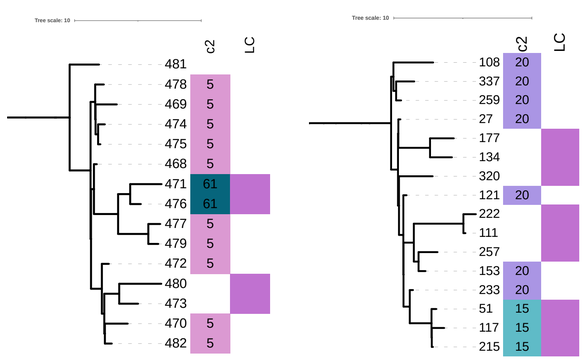
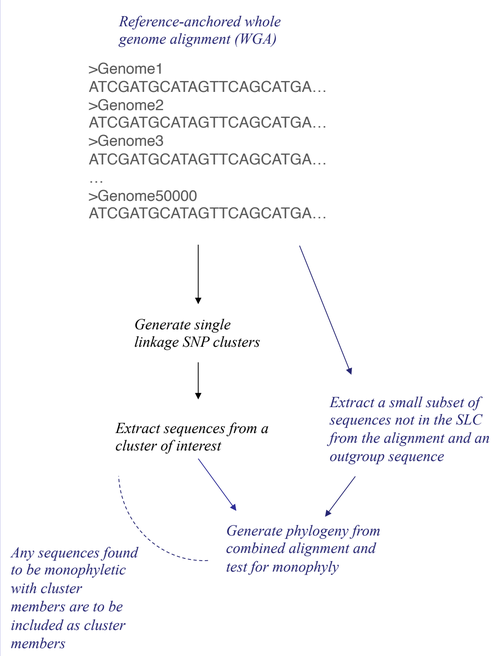
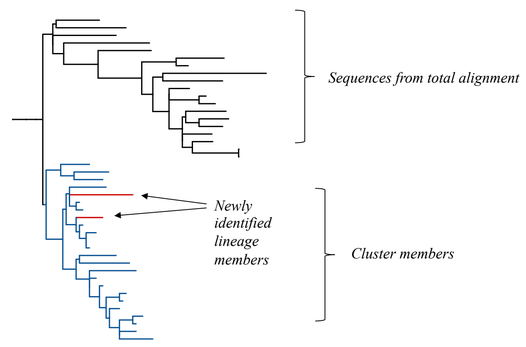


Figure 3. (Left) simulated population phylogeny with SNPs and no homologous recombination (n=500 taxa). Inner to outer colour strips show SNP SLC thresholds 0, 2, 5, 10, 15, 20 and 25. Those with >2 cluster members are shown. (Right) selected subtrees where the 2 SNP cluster threshold resulted in polyphyletic mappings. The second column 'LC' shows the sequences identified by *LineageCapture*

LineageCapture algorithm



- Starting with the sequences of an SLC of interest (outbreak linked perhaps), maybe a SNP threshold of 5 (t5), we then gather all the genomes of the major lineage (defined by MLST for instance)
- If this full alignment is too large to construct a tree (>20,000 O157:H7 genomes in Enterobase, for example), then use *LineageCapture*
- The algorithm (left) is run until all sequences of the alignment have been examined.
- Then, the entire process is repeated over multiple iterations, shuffling the order of sequences each iteration



Methods

Lineages of *Streptococcus pyogenes* (n=500) and *Clostridiodes difficile* (n=999) were downloaded from Enterobase[1] on Tuesday 17th October 2023. All genomes were reference-aligned with MUMmer[2] and a SNP-only whole genome alignment (WGA) extracted. The whole genome tree was approximated with FastTree[3]. *LineageCapture* is written in python 3 and uses the package ETE3[4]. It wraps IQ-TREE[5] for subtree approximations and *BactCore* (https://github.com/moorembioinfo/BactCore).

LineageCapture was run to a maximum of 8 iterations with batches of 20 sequences not in the cluster of interest.

1. Zhou, Z., Alikhan, N. F., Mohamed, K., Fan, Y. & Achtman, M. The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. *Genome Res* (2020) doi:10.1101/gr.251678.119.
2. Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. MUMmer: comparative applications\nFast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* **30**, 2478–83 (2002).
3. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
4. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol* **33**, 1635–1638 (2016).
5. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268–274 (2015).

The research was funded by the National Institute for Health Research Protection Research Unit (NIHR HPRU) in Gastrointestinal Infections at University of Liverpool in partnership with the UK Health Security Agency (UKHSA), in collaboration with University of Warwick. The views expressed are those of the author(s) and not necessarily the NIHR, the Department of Health and Social Care or the UK Health Security Agency.