NeighbourGroups: a machine learning classification tool that assigns microbial multi-locus genotypes to clusters

Dessislava Veltcheva (dessislava.veltcheva@spc.ox.ac.uk)
University of Oxford
Stephen Richer
University of Bath
Samuel Sheppard
University of Oxford
Margaret Varga
University of Oxford
Frances Colles
University of Oxford
Michael Bonsall
University of Oxford
Martin Maiden
University of Oxford

Article

Keywords:

Posted Date: March 22nd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2666125/v1

Abstract

Robust microbial classification systems are essential, but their definition is complicated by the large size and high diversity of microbial populations combined with a widespread horizontal genetic exchange. Multi-locus approaches that index gene variation without explicit phylogenetic classification mitigates these problems, but reproducibly defining high-level groups remains problematic. We describe a generalisable machine learning approach, 'NeighbourGroups', that reproducibly, robustly, and rapidly classifies multi-locus sequence types with defined precision.

Full Text

Most bacteria exist in very large populations, and the combination of high growth rates, short generation times, extensive horizontal gene transfer (HGT), and strong selection can lead to very high diversity along with variable levels of clonality[1]. For many applications, notably infectious disease epidemiology, robust classification systems that pragmatically and reproducibly differentiate variants at high resolution are essential. Multi-locus sequence typing (MLST) was developed to solve this problem, indexing sequence variation using a limited number, often as few as seven, housekeeping gene fragments without explicitly classifying them phylogenetically[2]. The sequence variation of these fragments is recorded as alleles and combinations of alleles as sequence types (STs) that can be organised into groups or clonal complexes (ccs), sometimes referred to as 'eBurst groups' (sBGs)[3]. As sequence capacity has increased, additional schemes with more loci have been introduced, including ribosomal MLST (rMLST, indexing the 53 ribosomal protein genes) and core genome (cgMLST, indexing all shared genes in a particular population); however, seven-locus classifications remain widely understood and used as a cornerstone for bacterial typing[4,5].

Whilst defining alleles and sequence types are straightforward, as they are effectively summaries of sequence variation, representing higher-level groups, such as clonal complexes, is more problematic. In addition to HGT confusing purely phylogenetic approaches, the existence of intermediate variants can result in all variants merging into a single group. These problems are less intense for schemes with very large numbers of loci, but for seven-locus MLST schemes, pragmatic solutions have been adopted, such as defining clonal complexes with a central genotype[5]. However, while establishing a stable classification system, these approaches can misclassify STs into incorrect clonal complexes, as they rely on assumptions about the representativeness of the data set being analysed, which may or may not be correct. They can also be unstable to the addition of new data.

We have addressed this problem by leveraging the availability of large numbers of whole genome sequences and machine learning techniques. First, cgMLST data are analysed using the Neighbour Joining tree reconstruction method to establish clusters or 'Neighbour Groups' based on the similarity of their cgMLST profiles. Then, a supervised machine learning algorithm is used to optimally predict the membership of these clusters from fewer loci, such as the MLST loci. The trained algorithm enables a robust probabilistic assignment of a seven-locus genotype to a cluster defined with cgMLST data (Figure

1), which is especially helpful when whole genome sequence data are unavailable, for example, from clinical specimens, as WGS technology is not available or for legacy data. The algorithm is available as a command line tool accessible from https://github.com/bgrdessislava/NeighbourGroups.

An essential parameter for the NeighbourGroups model is the number of classification groups, which is user-defined and can be established empirically. For example, with a *Campylobacter* dataset of >10,000 isolates for which cgMLST data were available[6], we performed a grid search to assess model performance for two to 100 classification groups (Figure 2). Model performance was evaluated with an adjusted Rand score, which determined whether two clusters were similar between the 'testing tree' and 'true tree'. An adjusted Rand score of >0.90 was defined as an excellent prediction, 0.80-0.90 good recovery of groups, 0.65-0.80 moderate recovery, and <0.65 poor recoveries, with low confidence in the reproducibility of the classification. This analysis indicated that, for this dataset, 20 NGroups gave an optimum performance, with an adjusted Rand score of 0.895, showing high agreement between the 20 groups assigned from cgMLST with those assigned from the seven-locus MLST data.

At the time of writing, there were more than 150 MLST schemes available for a wide range of microbial species, primarily bacteria, with hundreds of thousands of isolates typed to the level of seven loci MLST and, in many cases, also with cgMLST. Most of these can be found on the PubMLST website (https://www.pubmlst.org)[7]. From some MLST databases, notably those for *Neisseria* species[8], *Campylobacter jejuni. Campylobacter coli* [6], and *Salmonella enterica*[3], clonal complexes (or eBurst groups), have been defined using a variety of approaches, but for most data collections, such groups have not been rigorously defined or maintained. Given the variability of bacterial population structures, the number of different schemes and the number of isolates available, there is a need for a rational and automated approach to defining groups which can be applied to whole genome and MLST data. This is especially the case for pathogens for which it may not be possible to generate reliable whole genome sequence data from clinical specimens but where this information is beneficial. In addition, the Neighbour Group approach is easily implemented, and its assumptions easily understood, providing a pragmatic complement to other analysis approaches, many of which require whole genome sequences and high-capacity computing [9,10]. A final advantage is that the approach indicates the confidence with which seven locus data can be assigned to whole genome sequence groups.

Methods

Isolates acquisition and tree creation

10,359 high-quality *Campylobacter jejuni* U.K isolates were downloaded from PubMLST [https://publmlsts.org](Jolley, Bray and Maiden, 2018) using the following search query: *'Species=Campylobacter jejuni' AND 'Country=UK' AND '1990>=Year<=2020' AND 'N50>=20,000' AND '1.4Mb<=Genome Size <= 1.8 Mb' AND 'Contigs<=50' AND 'source=human_stool'.*

The output of this query is available at: https://pubmlst.org/bigsdb?db=pubmlst_campylobacter_isolates&page=query&project_list=110&submit=1

Tree Construction

The NeighbourGroups methodology requires the construction of two phylogenetic trees. The first is a "True Tree" containing all isolates, and the second is a "Training Tree" containing a subset of isolates used for model training. The model's classification accuracy is then assessed using testing isolates from the "True Tree". This step ensures the testing isolates are unseen by the model during training and helps prevent data leakage. Neighbour-joining trees were constructed using the coreMLST values (1,343 loci in *C. jejuni*) using GrapeTree (Zhou *et al.*, 2018). The "True Tree" contained 10,359 isolates, and the "Training Tree" had a randomly selected 80% subset. Trees were output in Newick format.

NeighbourGroup Assignment

The two Newick format trees were initially processed to generate linkage matrices using Python. Briefly, Newick files were read using ete3 (Huerta-Cepas, Serra and Bork, 2016). Following this, trees were converted to cophenetic matrices. Next, cophenetic matrices were converted to condensed distance matrices and, subsequently, linkage matrices using SciPy (Virtanen *et al.*, 2020). Finally, linkage matrices were passed to fcluster to perform hierarchical clustering and extract a predetermined number of groups.

CatBoost Classifier Training

Following group assignment, the training isolates were passed to the CatBoost supervised learning multi-class classification algorithm. CatBoost is a gradient-boosting algorithm for decision trees and works well with categorical values. The training was performed using the seven housekeeping MLST genes for *Campylobacter jejuni* (*aspA, glnA, gltA, glyA, pgm, tkt* and *uncA*) as features and the NeighbourGroup assignment as the target.

Testing Classifier Performance

Following training, model performance was assessed using the 20% unseen isolates excluded from the training step. Predicted groups were compared against the "true" groups defined by the "True Tree" using an adjusted Rand Index. The Rand Index is a measure of similarity between two data clustering's.

Determining Optimal Group Number

The maximum number of classification groups (NGroups) is a user-defined hyper-parameter of the NeighbourGroups methodology. A range of neighbour groups was tested between 2 and 100 to identify an optimal number of NeighbourGroups. For each NGroup number, the classifier was retrained, and the adjusted Rand score was computed.

Retraining and Model Deployment

Following the validation of model performance, a final NeighbourGroups model was built using the complete set of isolates. The final trained model can be deployed to classify novel isolates. Each prediction outputs a predicted group and an associated probability. Appropriate probability thresholds may be determined to categorise unknown isolates. The Neighbour Groups methodology described here is highly generalisable and can be applied across a wide range of bacterial species.

Declarations

Data Availability

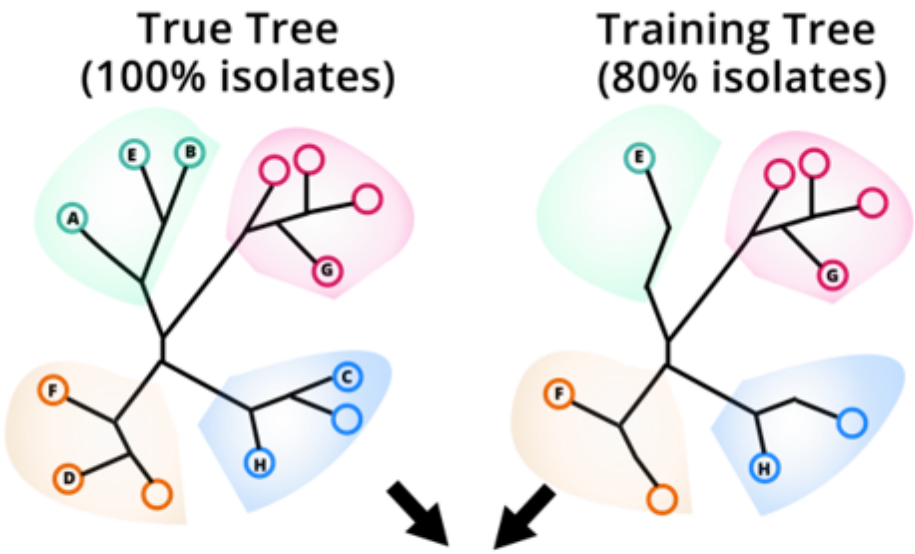
All the data are freely available and can be found on the PubMLST database (https://pubmlst.org/organisms/campylobacter-jejunicoli/) (Jolley, Bray and Maiden, 2018). The datasets used during the current study are available by accessing the publicly available project No.110: 1998-2018 UK human UK isolates (n=10,359). https://pubmlst.org/bigsdb?db=pubmlst_campylobacter_isolates&page=query&project_list=110&submit=1.

References

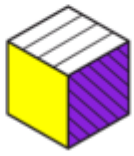
1. VanInsberghe, D., Arevalo, P., Chien, D. & Polz, M. F. How can microbial population genomics inform community ecology? *Philos Trans R Soc Lond B Biol Sci* **375**, 20190253, doi:10.1098/rstb.2019.0253 (2020).
2. Maiden, M. C. J. *et al.* Multi-locus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci USA* **95**, 3140-3145, doi:10.1073/pnas.95.6.3140 (1998).
3. Achtman, M. *et al.* Multilocus Sequence Typing as a Replacement for Serotyping in *Salmonella enterica*. *Plos Pathogens* **8**, e1002776, doi:10.1371/journal.ppat.1002776 (2012).
4. Maiden, M. C. *et al.* MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat. Rev. Microbiol.* **11**, 728-736, doi:10.1038/nrmicro3093 (2013).
5. Maiden, M. C. Multi-locus Sequence Typing of Bacteria. *Annu. Rev. Microbiol.* **60**, 561-588 (2006).
6. Cody, A. J., Bray, J. E., Jolley, K. A., McCarthy, N. D. & Maiden, M. C. J. Core Genome Multi-locus Sequence Typing Scheme for Stable, Comparative Analyses of *Campylobacter jejuni* and *C. coli* Human Disease Isolates. *J Clin Microbiol* **55**, 2086-2097, doi:10.1128/JCM.00080-17 (2017).
7. Jolley, K. A., Bray, J. E. & Maiden, M. C. J. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res* **3**, 124, doi:10.12688/wellcomeopenres.14826.1 (2018).
8. Bratcher, H. B., Corton, C., Jolley, K. A., Parkhill, J. & Maiden, M. C. A gene-by-gene population genomics platform: *de novo* assembly, annotation and genealogical analysis of 108 representative *Neisseria meningitidis* genomes. *BMC Genomics* **15**, 1138, doi:10.1186/1471-2164-15-1138 (2014).

9. Didelot, X. & Wilson, D. J. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLoS Computational Biology* **11**, e1004041, doi:10.1371/journal.pcbi.1004041 (2015).
10. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* **43**, e15, doi:10.1093/nar/gku1196 (2015).

Figures



	MLST							
Isolate	1	2	3	4	5	6	7	NGroup
A	4	20	22	1	53	23	1	?
B	4	20	42	1	30	2	3	?
C	2	3	51	1	2	4	3	?
D	15	2	23	1	42	10	52	?
E	4	20	5	1	20	23	1	Green
F	20	10	34	1	42	4	11	Orange
G	3	1	29	1	30	2	3	Pink
H	2	3	25	1	32	4	23	Blue



	MLST									
Isolate	1	2	3	4	5	6	7	NGroup True	NGroup Predicted	NGroup Probability
A	4	20	22	1	53	23	1	Green	Green	0.85
B	4	20	42	1	30	2	3	Green	Pink	0.4
C	2	3	51	1	2	4	3	Blue	Blue	0.9
D	15	2	23	1	42	10	11	Orange	Orange	0.87
↓	... all other testing isolates (20%)									
	Adjusted Rand Score 0.89									

Figure 1

Schematic illustration of the NeighbourGroup (NGroup) classification pipeline. Two phylogenetic trees are constructed: (1) the "True Tree", which is reconstructed from all the isolates and (2) the training tree, which uses 80% of the isolates. Isolates A, B, C, and D are test isolates excluded from the training tree, for which the model predicts Neighbour Group (NGroup) membership based on a limited number of loci. Isolates E, F, G, and H are training isolates assigned to Neighbour Groups. The table shows the results obtained from the model, with predicted NGroup membership as established by the "true tree" and predicted from the training tree. Neighbour Group membership probability is estimated and reflects a level of uncertainty in the classification. Adjusted Rand score shows how the model has performed: a score of 0 indicates the performance is no better than chance, and a score of 1 indicates perfect classification.

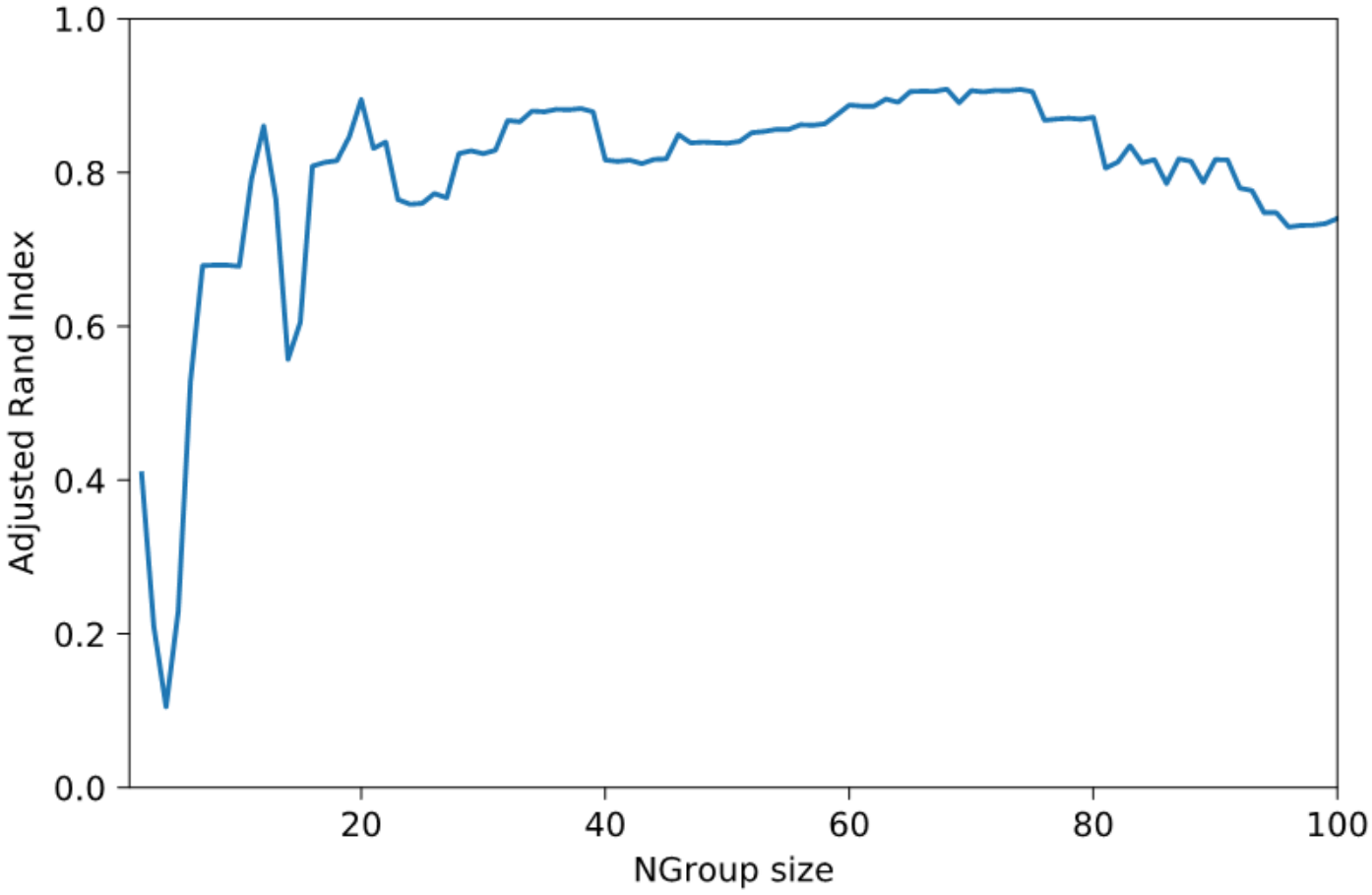


Figure 2

Variation in adjusted Rand Index with number of NGroups for a dataset of 10,359 *Campylobacter jejuni*WGSs. The adjusted Rand score is shown for 2 to 100 NGroups. Model performance was poor (low adjusted Rand Index) with fewer NGroups and reached a maximum (0.895) for 20 NGroups, with little added value for higher numbers of NGroups.