Big SARS-CoV-2 data: Processing of feature-extraction, dimensionality-reduction and clustering methods on over 5 million high coverage genetic sequences

Roberto Cahuantzi[1,5] Thomas House[1] Ian Hall[1] Katrina Lythgoe[2,3,4] Lorenzo Pellis[1]

[1]Department of Mathematics , U of Manchester [2]Department of Biology , U of Oxford [3]Big Data Institute, U of Oxford [4]Pandemic Sciences Institute, U of Oxford [5]United Kingdom Health Security Agency

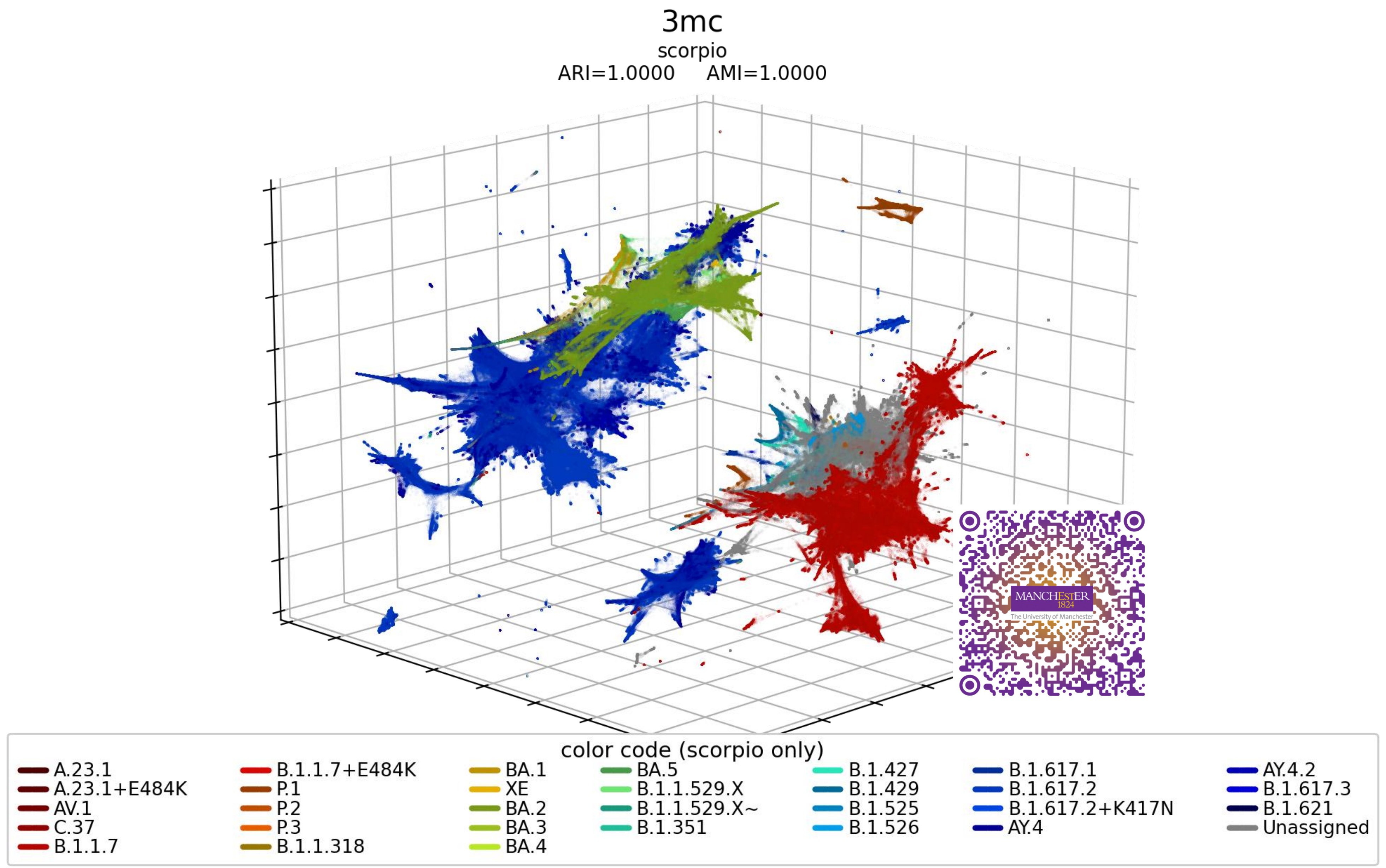
Introduction

Since its emergence in late 2019, SARS-CoV-2 has diversified into many lineages and globally caused multiple waves of infection. Novel lineages have the potential to spread rapidly and internationally if they have higher intrinsic transmissibility, can evade host immune responses. Phylogenetic methods provide the gold standard for representing the diversity and to identify newly emerging lineages of SARS-CoV-2. However, these methods are computationally expensive, struggle when datasets get too large, and require manual curation to designate new lineages. These challenges together with the increasing volumes of genomic data available provide a motivation to develop complementary methods that can incorporate all of the genetic data available, without down-sampling, to extract meaningful information rapidly. Here, we produce a proof of concept using word-statistics characterisation of sequences and dimensionality reduction, bringing speed, scalability, and interpretability to the construction of genetic topologies, while not serving as a substitute the proposed methods can be used as a complementary approach to identify and confirm new emerging variants.

Methods

- Downloading of all the genetic sequences from the GISAID database as up to January 19th, 2023, and filtering out low coverage sequences resulting on a usable dataset of approximately 5.7M sequences;
- Alignment and classification of sequences using Pangolin [8], to retrieve scorpio lineages reported to be considered as "ground truth";
- Charcterisation of the sequences through of *k-mer* counts spectrum [2], here labelled 3*mc*, since the *k* value is set to 3 to keep the computational costs low;
- The dimensionality reduction projection was produced using the PaCMAP [4] method, which proved to be robust for replication, while reliable on representing global and local structures;
- Two clustering algorithms were applied and evaluated, HDBSCAN [3] and CLASSIX [1], both yielded similar and consistent results;
- A hyperparametric grid exploration was made to a stratified sample of the GISAID dataset to find the optimal PaCMAP, HDBSCAN and CLASSIX values based on adjusted mutual information index (AMI) and adjusted Rand index (ARI) values of the projections, calculated based on the scorpio lineages. The most optimal hyperparameters settings for the whole dataset was labelled *GISAID2*;
- Another analysis with a subsample of the sequences from England was performed with accumulative slices and a temporal resolution of two weeks to discover the dynamics of the unsupervised emergence of clusters, to optimise for this subset a new hyperparameters exploration was perform optimising for correlation coefficient (r^2) to the scorpio labelling, this setting was labelled *England*;

Results

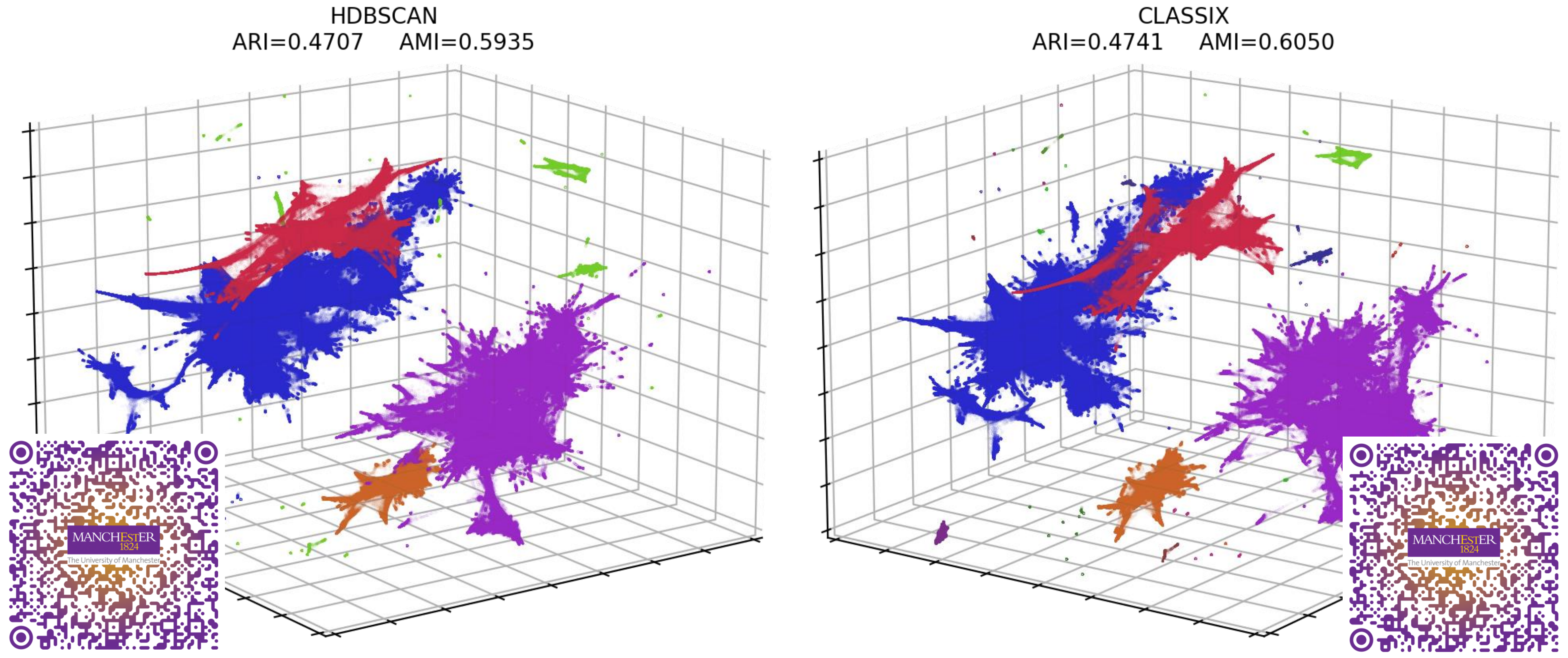

3mc

scorpio

ARI=1.0000 AMI=1.0000

color code (scorpio only)

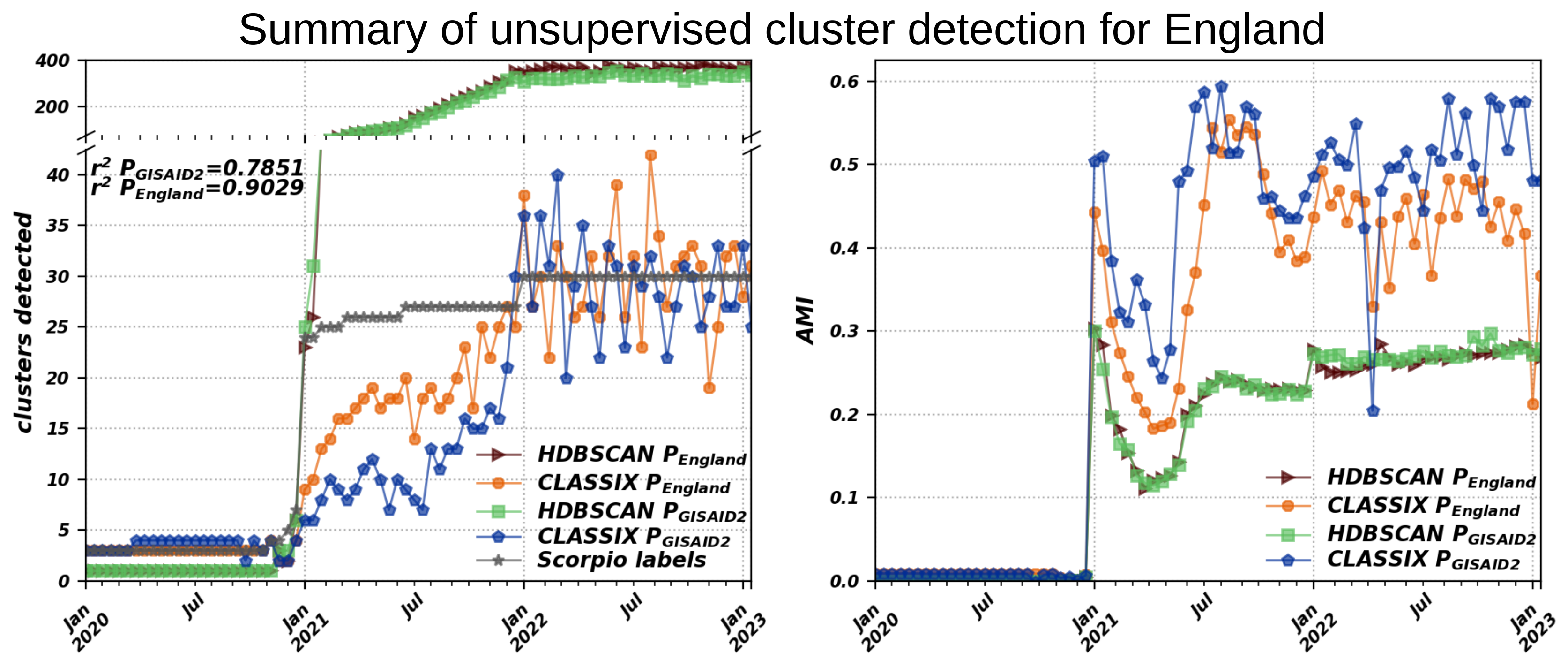
A.23.1	B.1.1.7+E484K	BA.1	BA.5	B.1.427	B.1.617.1	AY.4.2
A.23.1+E484K	P.1	XE	B.1.1.529.X	B.1.429	B.1.617.2	B.1.617.3
AV.1	P.2	BA.2	B.1.1.529.X~	B.1.525	B.1.617.2+K417N	B.1.621
C.37	P.3	BA.3	B.1.351	B.1.526	AY.4	Unassigned
B.1.1.7	B.1.1.318	BA.4				


HDBSCAN

ARI=0.4707 AMI=0.5935

CLASSIX

ARI=0.4741 AMI=0.6050


Summary of unsupervised cluster detection for England

Conclusions and Future work

- The PaCMAP resulted robust for replication showing a consistent global structure, however a hyperparametric exploration was still required to optimise the most meaningful AMI metric;
- The low dimension projection of 3*mc* characterisation seem to rescue important information to identify new variants through unsupervised clustering methods;
- The novel CLASSIX clustering method outperformed HDBSCAN in processing times roughly four times fold, while also marginally outperforming it on AMI;
- The proposed method for monitoring of emergence of clusters has the potential of making possible to add supportive evidence to police the emergence of variants of concern, as shown in the in the fourth figure.
- Within this framework it seems to be a trade-off between r^2 and AMI with respect to the scorpio labelling;
- A temporal analysis of the dynamics of consistency of clusters should be conducted, this is to say, how datapoints of a given cluster produce the emergence or merging into new clusters. To build a stronger case on the usefulness of these tools on the detection of emergent variants of concern;
- Other characterisations with more biological bases such as: protein *NV* [5], or even 3mc characterisation by coding regions could give it more interpretability;
- The applicability of other features could be explored, like graphic representation [6], or Fourier power spectrum [7];
- It would be also of interest to assess different metrics of distance like: Kull-Leiber, Yaus-Hausdorff [6] or Wasserstein, among others; to explore possibilities for the better clustering of relevant lineages.

References

[1] X. Chen and S. Guettel, 'Fast and explainable clustering based on sorting', 2022. DOI: 10.48550/ARXIV.2202.01456.

[2] S. Vinga and J. Almeida, 'Alignment-free sequence comparison—a review', 2002. DOI: 10.1093/bioinformatics/btg005

[3] L. McInnes and J. Healy, 'Accelerated hierarchical density based clustering', Nov. 2017. DOI: 10.1109/ICDMW.2017.12

[4] Huang, H., Wang, Y., Rudin, C. et al. 'Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization'. Commun Biol 5, 719 (2022). https://doi-org.manchester.idm.oclc.org/10.1038/s42003-022-03628-x.

[5] C. Yu, et al., 'Protein space: A natural method for realizing the nature of protein universe', Journal of Theoretical Biology, vol. 318, 2013

[6] K Tian, et al., 'Two Dimensional Yau-Hausdorff Distance with Applications on Comparison of 417 DNA and Protein Sequences', PLoS ONE, vol. 10, 2015.

[7] C. Yin, et al., 'A measure of DNA sequence similarity by Fourier Transform with applications on hierarchical clustering', Journal of Theoretical Biology, vol. 359, 2014.

[8] A. O'Toole et al. 'Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool', Virus evolution, 7(2), veab064. https://doi.org/10.1093/ve/veab064

Contact email:
roberto.cahuantzi@manchester.ac.uk

GitHub repo:
github.com/robcah/dimredcovid19