Efficient Pangenome Construction through Alignment-Free Residue Pangenome Analysis (ARPA)

Arnav Lal[1], Ahmed Moustafa[2,3], and Paul J. Planet[3,4,5]

[1] School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA 19104, USA
[2] Division of Gastroenterology, Hepatology & Nutrition, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
[3] Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[4] Division of Pediatric Infectious Diseases, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
[5] Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY 10024, USA

Corresponding author: Planet, Paul J (planetp@chop.edu)

Abstract
Most pangenome analysis relies on large-scale alignments of genes or proteins, which makes them computationally expensive. We noted that protein sequences can be transformed into one dimensional twenty-value vectors (vector of residue counts; vRC), where each value is the count of each of the 20 possible amino acids in the protein. The vRC serves as unique protein identifier that makes mathematical comparisons much more tractable and retains information about homology. We use these vectors to develop an alignment-free method for pangenome clustering named Alignment-Free Residue Pangenome Analysis (ARPA; https://github.com/Arnavlal/ARPA). We test this method with large datasets of whole genome prokaryotic sequences and compare to an alignment-based pangenome approach (Roary). Our alignment-free approach yields much faster homolog clustering than BLAST-based alignment. While speed depends upon the dataset composition, ARPA is 70-90 times faster than Roary at clustering hundreds of intra-species genomes, and over 130 times faster when clustering hundreds of highly similar intra-species genomes. The specificity of ARPA compares with BLAST-based approaches to defining homologs, with approximately 88% sensitivity and 99.8% specificity. In addition to homolog clustering, vRCs can be used to separate paralogs (based on gene neighborhood) and to identify orthologs. The pangenome generated by ARPA offers an opportunity for quick calculation and visualization of allele-level differences, generating a foundation for allelic investigation within multi-genome analyses. We show that allelic diversity can be used for efficient and accurate phylogenetic inference. ARPA is an efficient and accurate platform for large-scale, pangenome alignment projects that can be used in comparative and phylogenetic analysis.

Introduction

With an increasing number of fully sequenced genomes available, computationally tractable tools are needed for efficient analysis of large genomic datasets. Global comparison of gene or species evolution often occurs through construction of a pangenome, namely the union of genes within all sampled genomes (Medini *et al.* 2005). Analyzing genetic change within many genomes is computationally challenging—the creation of a microbial pangenome is computationally NP-Hard (Nguyen *et al.* 2014). Attempts have been made to reduce the complexity of all-against-all comparison, most notably through pre-clustering of genomes first performed in LS-BSR (Sahl *et al.* 2014) and in Roary (Page *et al.* 2015). Nevertheless, current conventional pangenomic methods rely upon alignment software, mostly BLAST (Camacho *et al.* 2009), to compare and cluster genes. Current alignment-free genome clustering techniques utilize units smaller than whole genes, for example, through detection of k-mer frequency (Comin *et al.* 2015), which can be further integrated with markov chain modelling (Saw *et al.* 2019). Another alignment free technique that uses exact protein sequence matches, WhatsGNU (Moustafa *et al.* 2020), cannot identify close homologs without relying on an alignment-based platform.

We sought to develop an alignment free methodology that could identify homologs, orthologs, and paralogs and generate pangenomes for species without performing any sequence alignments. In this method, amino acid alignments are transformed into one dimensional, twenty value vectors (vRC) composed of 20 amino acid residue counts. Building an analogy from statistical mechanics (Schroeder 2005), this method identifies proteins through their macroscopic state of residue counts, as opposed to the microscopic state which is the ordered sequence of amino acids in the protein. We hypothesized that the macroscopic state could successfully identify similar sequences because actual amino acid sequence order is not consistently being reorganized. To our knowledge, the ability to uniquely identify homologs using counts of amino acids is novel. Beyond efficiency, vRC-based clustering offers the opportunity to quickly compare differences between clustered proteins at the allele-level and to efficiently retain these measurements in pangenome analysis.

Description

Residue Counts and Allele Compression

ARPA requires an input folder of ".faa" processed amino acid files corresponding to prokaryotic genomes. Encoded proteins are converted into a 1-by-20 row vector (vRC) with each term corresponding to the number of specific amino acid residues present within the sequence ordered arbitrarily (i.e., the first column corresponds to the glycine count within the protein; Fig. 1A). These 20 numbers, which surprisingly contain enough information to uniquely identify every predicted protein within any given *Staphyloccous aureus* genome even using partial vRCs (Fig. 1B), can be used identify encoded genes during pangenome clustering.

To improve computational efficiency, ARPA rapidly compresses the database by removing proteins with identical residue counts but retaining information about the genomes in which they are found. Proteins within a database of roughly 2,500 *S. aureus* sequences were compressed from 6,704,399 sequence to 161,445 unique sequences

(41.5-fold compression). A 4-fold compression is observed with 50 random *S. aureus* genomes, and 8-fold with 200 random genomes. When studying only genomes from the same clonal complex (CC1), over 20-fold compression (28,000 from 572,000 genes) was observed in 218 *S. aureus* CC1 genomes.

Clustering

Encoded genes are subsequently clustered into homologous groups, by comparing a vRC to other protein sequence vRCs. Because vRCs allow for rapid one-against-all comparison, we created and used an unsupervised bulk single linkage clustering (bSLC) method (described visually in Fig 1C). While the time complexity of single linkage clustering is $O(n^2)$ (Manning *et al.* 2008), the complexity of bulk clustering is faster than $O(n^2)$, but slower than $O(n)$. In comparison of vRCs, the metric of residue count similarity is the subtraction of the absolute value of each individual amino acid count (Fig. 1A). This intuitive comparison, which results in the specific number of residues differing between two proteins, allows for ARPA to utilize definite threshold cutoffs of similarity in pairwise comparisons.

Starting with any arbitrary sequence vRC, we generate two thresholds: (1) a group of genes that immediately cluster with the initial sequence (we set a default threshold of 95% residue count similarity, but this is a modifiable setting) and (2) a group of genes that do not cluster with the initial sequence analyzed, but which are similar enough to potentially cluster with other sequences in the “immediate” cluster. The use of this exploratory secondary threshold throughout subsequent linkage clustering ensures that the single linkage clusters are consistent independent of the starting point. Fundamentally, this method utilizes the similarity in amino acid counts to generate clusters of similar proteins, and it allows for deviation from 95% identity (as may occur through gradual evolutionary change) in the case where there are protein sequences that link one to another within the 95% threshold. Therefore, the distance between any two genes may not meet the threshold value if they are connected by one or more intervening proteins. Running ARPA on 50 *S. aureus* genomes revealed that proteins outside of the initial cluster are incorporated during clustering analysis of biological datasets (Fig. 1D). Theoretically, the accuracy of this method is anticipated to improve with greater numbers of sequences that come from closely related genomes in that intermediate protein sequences may link divergent sequences.

Paralog Separation

Paralog separation exists as the final option in ARPA. Especially in a prokaryotic system, surrounding genes influence the regulation and phenotypic effects of a given gene, thus gene neighborhoods are one proxy metric for separation of paralogous genes (Esch and Merkl, 2020), which is simpler to implement than phylogenetic techniques for paralog separation (Chiu *et al.*, 2006). Having established gene homologs through clustering, the final (optional) procedure involves the separation of paralogous groups through analysis of conserved gene neighborhoods, an approach taken in Roary. ARPA uses pre-built clusters from homolog clustering as a starting point for neighborhood analysis. While Roary splits paralogous groups only when a single genome contains multiple homologs, ARPA identifies neighborhoods for all homologs and splits them into paralogous groups when their neighborhoods differ. While time

intensive, this separation strategy avoids missed paralogous copies that could result from differential loss of paralogs in each genome.

Output Files

ARPA presents three output files, two of which are visualizations of the pangenome. The most data-rich output is a ".csv" (or ".pickle" for large pangenomics) file, which lists proteins, labelled with observed protein identifiers, and a pangenome output table. The pangenomes table reports the presence or absence of proteins but also lists one of three "allele scores" for each present protein, that a user may specify. The default allele score is "most represented," which is calculated as the number of residues different from the protein sequence that is most represented within the homolog group. The alternatives output values are "all alleles," where the difference of the given allele is subtracted individually from each other allele within the homologous group, and this difference is subsequently averaged, and "all alleles normalized" where the difference from each unique allele is normalized by the number of times the allele is observed within the homolog group. Each of these metrics are designed to offer unique and/or informative identifiers for each allele.

Results

In order to assess the validity of clustering, we utilized a BLAST strategy that utilized one (arbitrary) clustered protein and compared this protein sequence against all sequences within and outside of the designated protein's cluster for 10 *S. aureus* genomes with a 90% similarity threshold in single linkage clustering. Pangenome construction for ARPA clustering demonstrates that vRCs separate inner cluster sequences with high BLAST identity from other sequences that have low BLAST identity to the query (Fig. 1E). However, percent identity may be artificially high if a subsection of a protein has high identity. Thus, we also performed this analysis using the metric of query coverage multiplied by percent identity (Fig. 1F). This analysis revealed that ARPA clustering has an approximately 88% sensitivity and 99.8% specificity. In separating sequence identity versus coverage, ARPA clustering strategy preserves sequences with lower identity but high coverage, while removing high identity low coverage groups (ones that would be kept by Roary; Fig 1G).

While a 10 genome *S. aureus* pangenome clustering with ARPA was conducted in 6.78 seconds, the BLAST confirmation analysis presented above required multiple hours. Regarding runtime, the clustering process of ARPA outcompetes Roary clustering in wall time when run on a personal computer, with the clustering time advantage increasing with the number of genomes being studied in the scale of hundreds of genomes (Fig 1H). Bulk single linkage clustering becomes increasingly complicated, and the time advantages of this version of ARPA clustering begin to decrease at large genomic sizes, for example ARPA is only 78-fold faster than Roary when analyzing 2,500 *S. aureus* genomes.

Retention of allele scores (see output files above) within the pangenome presence absence output is beneficial for further analysis of the pangenome. Maximum-likelihood phylogenetic inference using the presence/absence of genes within the pangenome results in similar phylogenetic relationships in a test case of 50 *S. aureus* genomes (Fig 2A). Creation of the pangenome by ARPA and visualization of allelic

differences from distinct subclades of *S. aureus* reveals that similar allelic changes occur within the core genome (a set of homologs within all, or nearly all, genomes) for each subclade (Fig. 2B).

The allele scores from ARPA also have potential utility in phylogenetic analysis. In phylogenetic analysis treating allele scores from ARPA for 50 *S. aureus* genomes (Fig. 3A) as morphological data, and labelling missing proteins as unknown, the topology of the resulting tree is highly similar to a tree generated from alignment of the core genome from a core set of genes (defined in this case as genes present in >96% of genomes: 1617 genes; Fig. 3B) and also when including a number of softer core genes (3000 total genes; Fig 3C). The normalized Jaccard-Robinson-Foulds metric comparing both alignment-free trees to the alignment-based tree was 0.24 and 0.20, respectively. In the future, of phylogenetic strategies designed specifically for vRC-based data may offer accurate phylogenetic inference.

Limitations

The development of heuristics like vRCs inevitably leads to loss of information. Without alignment, some types of sequence change may be missed, for instance when two sequences have residues that are swapped in location. Notably, protein sequences with large truncations (Fig. 1G) will not be clustered in ARPA. Further, it is possible that two distinct, unrelated proteins may have the same combination of residues in different orders. This is unlikely in real datasets, but it remains a possibility, and may result in two separate clusters being merged. Gene neighborhood analysis may be especially useful for separating these genes into distinct ortholog groups.

Methods

Sequences for this manuscript are available from GenBank, and accession numbers for each analysis is provided in supplementary data. Comparative pangenome construction and alignment-based core genome alignment were performed with Roary (Page *et al.* 2015) and MAAFT (Katoh *et al.* 2002), respectively. Sequence comparison utilized command line BLAST (Camacho *et al.* 2009), and phylogenetic tree comparisons were visualized with phylo.io (Robinson *et al.*, 2016), followed by adaptation and addition of tangle grams for visualization. Tree building was performed using IQTree (Nguyen *et al.* 2015).

Conclusions

ARPA is an entirely alignment-free pangenome construction technique that can be used as a pipeline for pangenomic and phylogenetic analysis with large numbers of sequences. The method provides enough computational simplicity to be performed on a regular modern computer, and can be used on hundreds of thousands of sequences.

References

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics*, *10*, 421. https://doi.org/10.1186/1471-2105-10-421

Chiu, J. C., Lee, E. K., Egan, M. G., Sarkar, I. N., Coruzzi, G. M., & DeSalle, R. (2006). OrthologID: automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics (Oxford, England)*, *22*(6), 699–707. https://doi.org/10.1093/bioinformatics/btk040

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

Comin, M., Leoni, A., & Schimd, M. (2015). Clustering of reads with alignment-free measures and quality values. *Algorithms for molecular biology : AMB*, *10*, 4. https://doi.org/10.1186/s13015-014-0029-x

Esch, R., & Merkl, R. (2020). Conserved genomic neighborhood is a strong but no perfect indicator for a direct interaction of microbial gene products. *BMC bioinformatics*, *21*(1), 5. https://doi.org/10.1186/s12859-019-3200-z

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Medini, D., Donati, C., Tettelin, H., Masignani, V., & Rappuoli, R. (2005). The microbial pan-genome. Current opinion in genetics & development, 15(6), 589–594. https://doi.org/10.1016/j.gde.2005.09.006

Moustafa, A. M., & Planet, P. J. (2020). WhatsGNU: a tool for identifying proteomic novelty. *Genome biology*, *21*(1), 58. https://doi.org/10.1186/s13059-020-01965-w

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Nguyen, N . *et al.* (2014) Building a pangenome reference for a population. In: Sharan,R. (ed.) Research in Computational Molecular Biology, Springer International Publishing, pp. 207–22

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics (Oxford, England)*, *31*(22), 3691–3693. https://doi.org/10.1093/bioinformatics/btv421

Robinson, O., Dylus, D., & Dessimoz, C. (2016). Phylo.io: Interactive Viewing and Comparison of Large Phylogenetic Trees on the Web. *Molecular biology and evolution*, *33*(8), 2163–2166. https://doi.org/10.1093/molbev/msw080

Sahl, J. et al. (2014) The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. PeerJ, 2, e332.

Saw, A. K., Raj, G., Das, M., Talukdar, N. C., Tripathy, B. C., & Nandi, S. (2019). Alignment-free method for DNA sequence clustering using Fuzzy integral similarity. *Scientific reports*, *9*(1), 3753. https://doi.org/10.1038/s41598-019-40452-6

Schroeder DV. chapter 2. In: *An introduction to thermal physics*. San Francisco, CA: Addison Wesley Longman, 2005

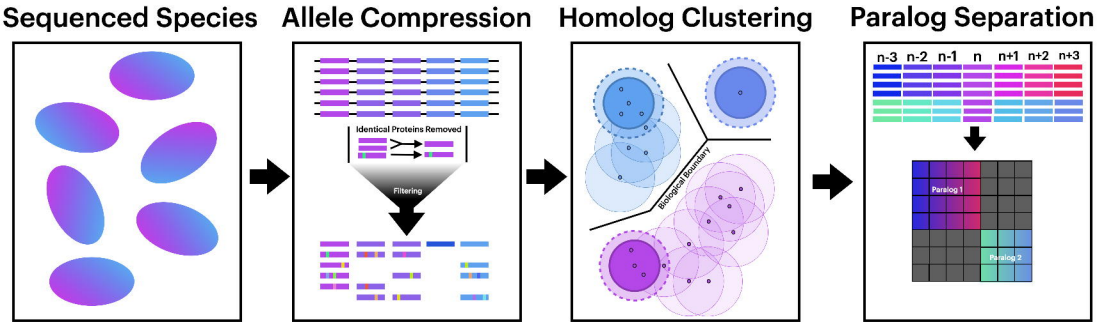
Figures

Graphical Summary. Workflow of Pangenome Construction. Amino acid sequences from imported genomes (.faa files) are compressed to yield unique alleles. Subsequent encoded genes are iteratively clustered through bulk single linkage clustering by comparing residue counts (vRC), with cluster boundaries relying upon biological differences in vRCs of different proteins. Resulting homologous proteins are separated by comparing gene neighborhoods, through a process that utilizes results from the prior clustering step, to yield orthologous groups.

Figure 1. Methodology, Validation and Benchamarking for ARPA clustering. A. Mechanism for protein vRC vs. protein vRC analysis. Amino acid counts are compared in a one-versus-all mechanism, Numerical cutoffs are utilized in determination of clusters. **B.** Plot for the number of amino acids necessary in order to fully identify all unique proteins within a *S. aureus* genome, as a proof-of-concept metric for the use of vRCs. 200 lines corresponding to 200 *S. aureus* genomes are juxtaposed on the plot. Almost all proteins are uniquely identifiable using vRCs with lengths greater than 5 residues. **C.** Methodology for single linkage clustering employed in ARPA. **D.** Use of bulk single linkage clustering method within 50 *S. aureus* genomes reveals that clusters beyond the initial 95% threshold are actively employed when studying biological datasets. **E.** visualization of percentage identity, **F.** query coverage times percentage identity, and **G** query coverage plotted against percentage identity resulting in confirmation analysis during ARPA validation (blue: sequences within cluster, gold: sequences outside cluster). **H.** Plot of runtime for clustering process with ARPA versus Roary on a standard personal laptop (Apple Mac 2.6 GHz 6-Core Intel Core i7 with 16 GB RAM).

Figure 2. A. Maximum likelihood phylogeny and tangle gram of homolog presence absence derived from Roary versus ARPA programs. **B.** Maximum likelihood phylogeny of 218 CC1 genomes, and visualization of corresponding pangenome output from ARPA.

Figure 3. A. Visualization of pangenome output for 50 *S. aureus* genomes. Phylogeny derived from the core homolog allele deviations (**B**) and core and softer core homolog allele deviations (**C**) are plotted versus tree derived from alignment of core genomes for the same species.


Sequenced Species
Allele Compression
Identical Proteins Removed
Filtering
Homolog Clustering
Biological Boundary
Paralog Separation
n-3
n-2
n-1
n
n+1
n+2
n+3
Paralog 1
Paralog 2

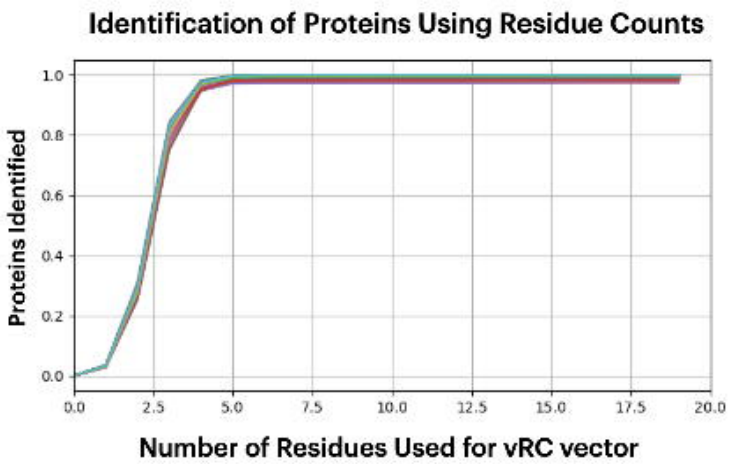
A

	G	P	A	V	L	I	M	C	F	Y	W	H	K	R	Q	N	E	D	S	T	ΔGene 1	ΔGene 1/len(Gene 1)
Gene 1	21	13	10	3	15	25	7	14	14	23	20	18	22	10	8	21	17	21	25	13	—	—
Gene 2	13	2	29	20	6	18	32	39	29	52	12	19	6	3	20	16	19	29	33	25	244	76%
Gene 3	23	13	10	5	15	25	7	18	14	23	20	20	23	11	8	21	18	21	25	14	14	4%
Gene 4	4	40	17	16	30	30	14	6	19	8	9	20	18	20	9	4	23	34	4	7	210	66%
Gene 5	15	2	31	20	8	18	32	41	29	48	12	21	6	3	22	16	19	30	33	25	245	77%

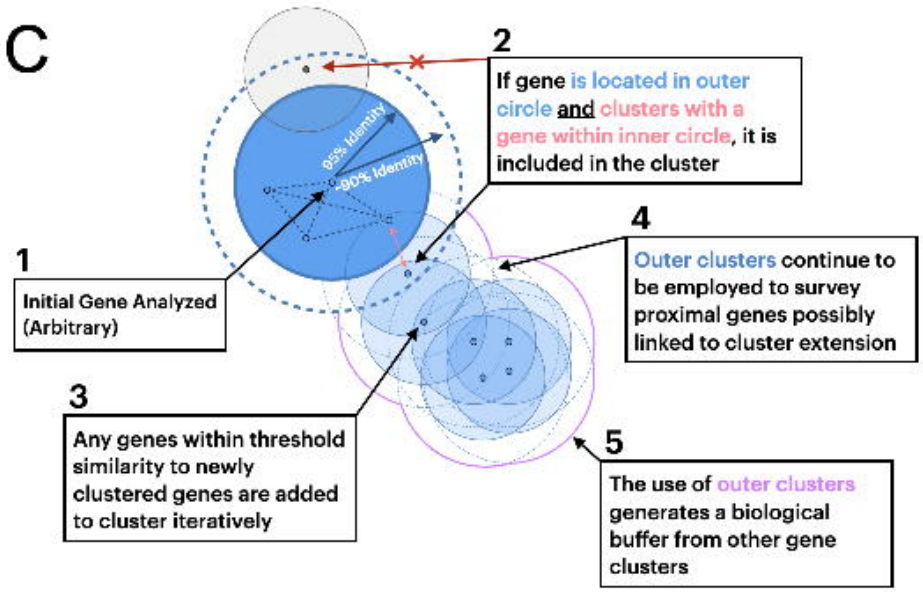
$$\sum_{res=0}^{19} |Gene\ 1_{res} - Gene\ 2_{res}|$$

Same Cluster

B



C



D

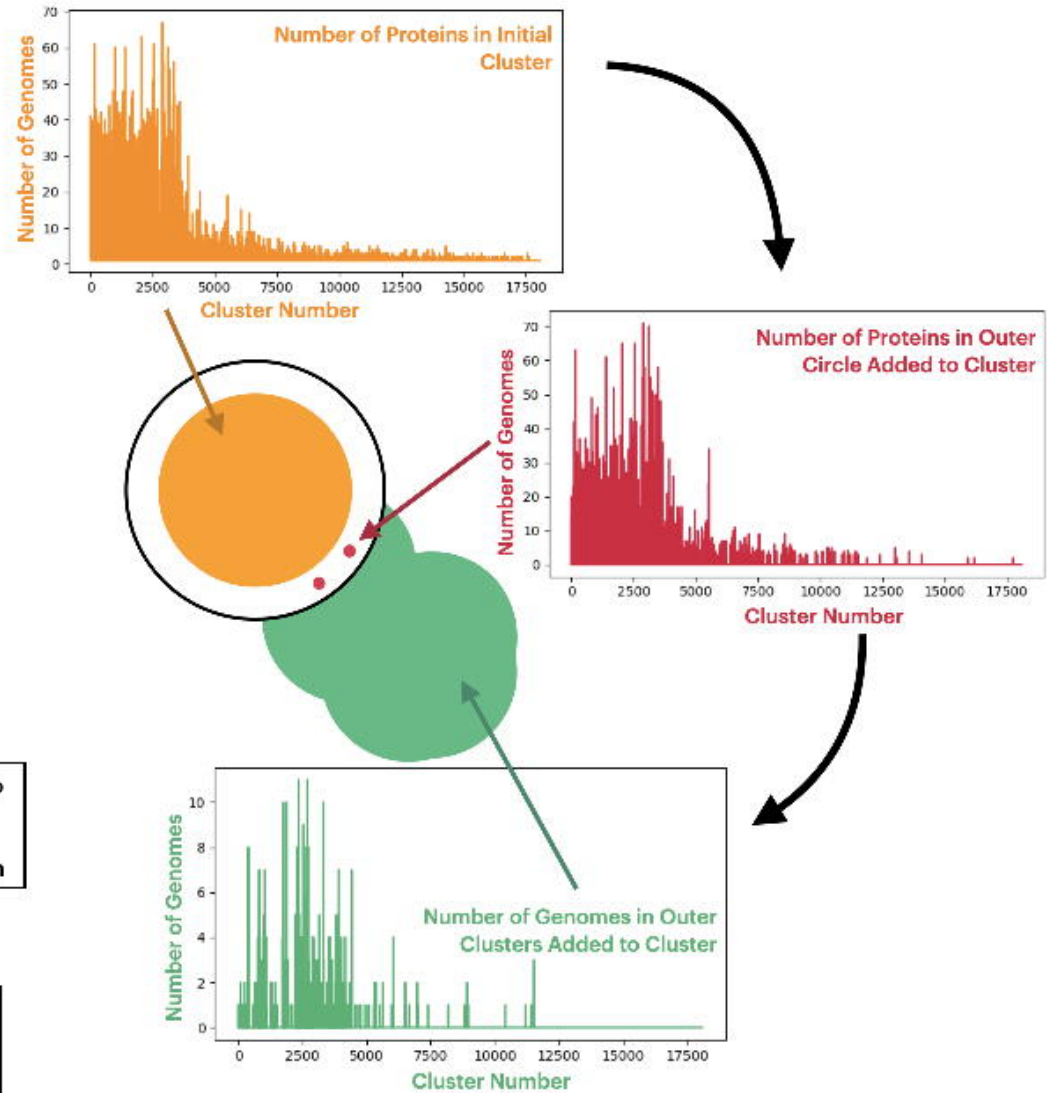


Figure 1

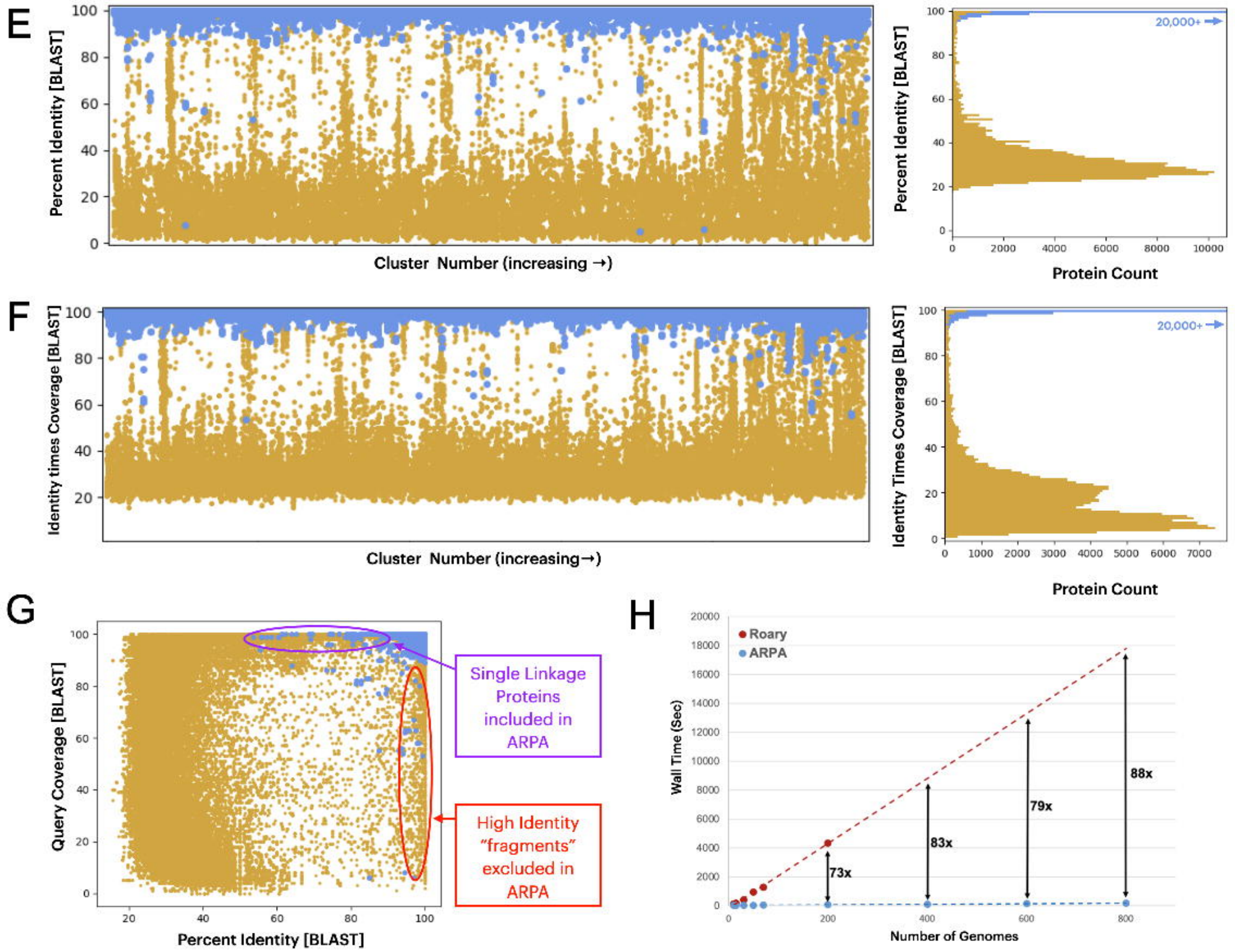

E
Percent Identity [BLAST]
Cluster Number (increasing →)
20,000+
Protein Count
F
Identity times Coverage [BLAST]
Cluster Number (increasing→)
Identity Times Coverage [BLAST]
G
Query Coverage [BLAST]
Percent Identity [BLAST]
Single Linkage Proteins included in ARPA
High Identity "fragments" excluded in ARPA
H
Roary
ARPA
Wall Time (Sec)
Number of Genomes
73x
83x
79x
88x

Figure 1

A

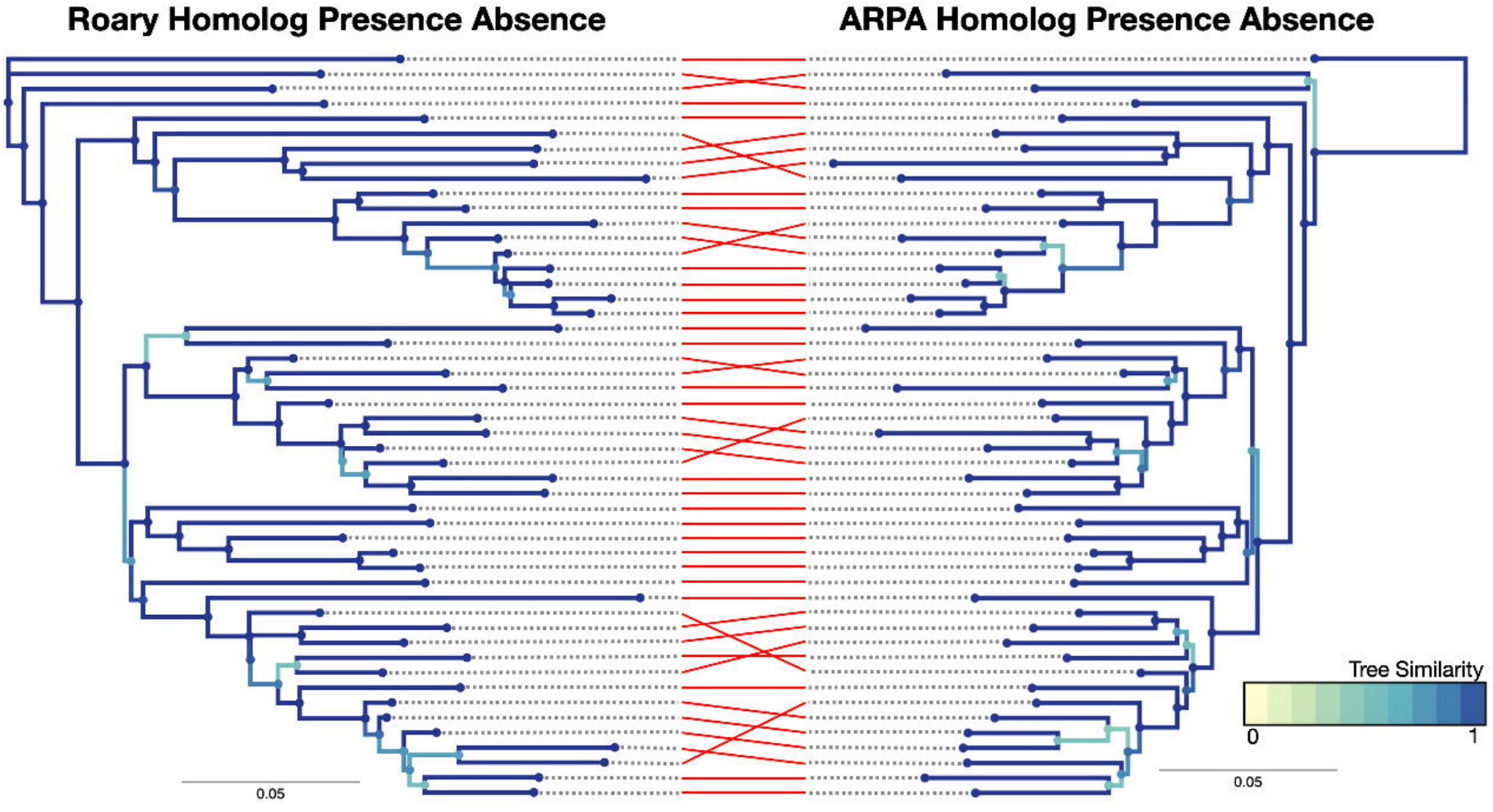

Roary Homolog Presence Absence
ARPA Homolog Presence Absence
Tree Similarity
0
1
0.05
0.05

Figure 2

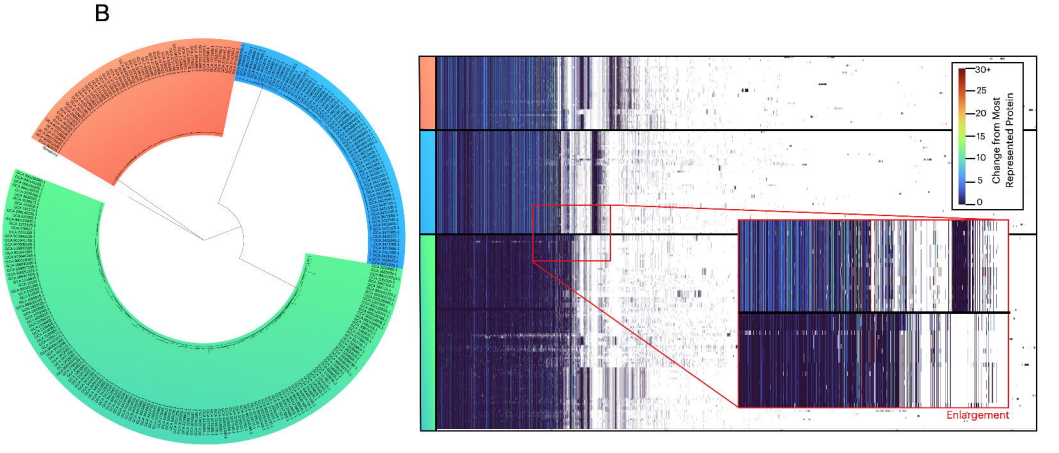

B
30+
25
20
15
10
5
0
Change from Most Represented Protein
Enlargement

Figure 2

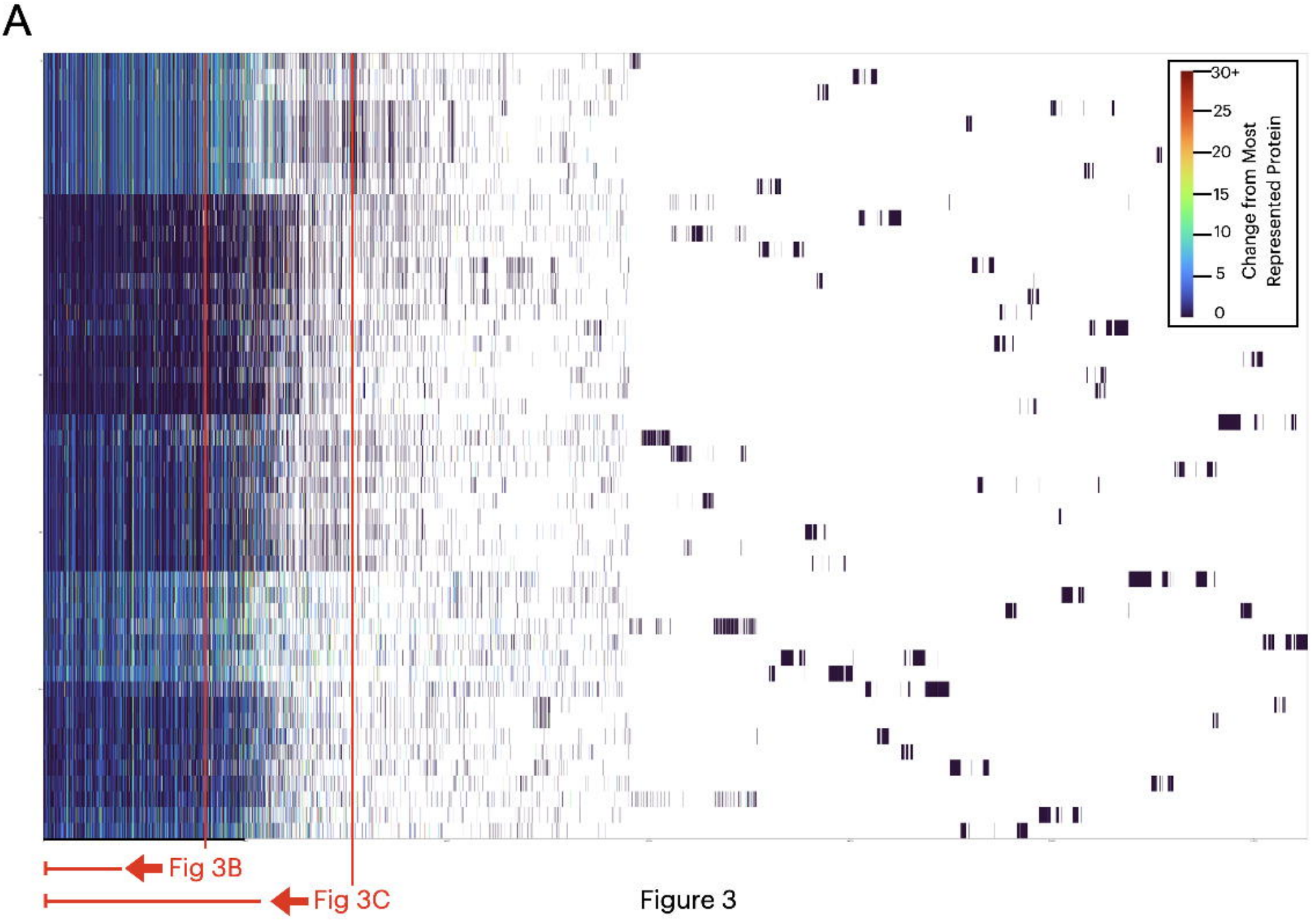

A
30+
25
20
15
10
5
0
Change from Most Represented Protein
Fig 3B
Fig 3C

Figure 3

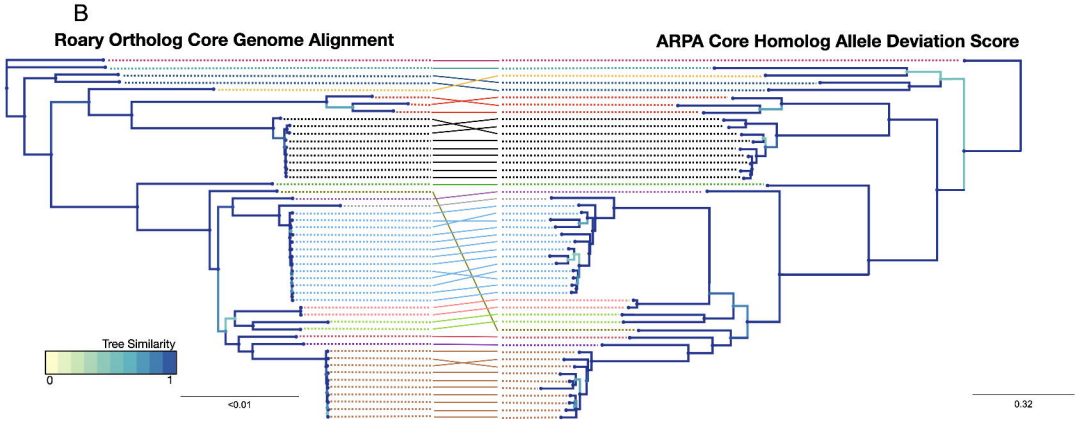

B
Roary Ortholog Core Genome Alignment
ARPA Core Homolog Allele Deviation Score
Tree Similarity
0
1
<0.01
0.32

Figure 3

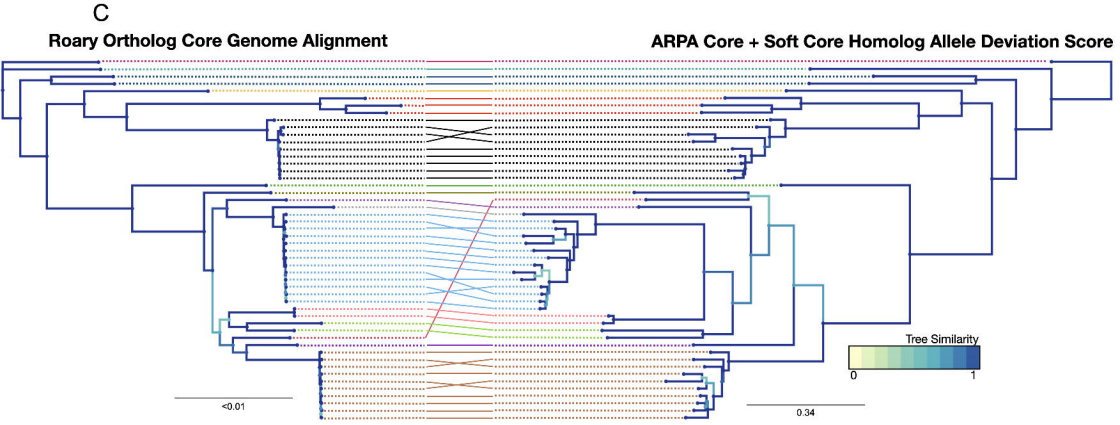

C
Roary Ortholog Core Genome Alignment
ARPA Core + Soft Core Homolog Allele Deviation Score
Tree Similarity
0
1
<0.01
0.34

Figure 3