MICROBIAL GENOMICS

RESEARCH ARTICLE

Whelan *et al.*, *Microbial Genomics* 2020 ;6

DOI 10.1099/mgen.0.000338

MICROBIOLOGY SOCIETY

OPEN DATA OPEN MICROBIOLOGY

Coinfinder: detecting significant associations and dissociations in pangenomes

Fiona Jane Whelan[1]†, Martin Rusilowicz[2]† and James Oscar McInerney[1,2,*]

Abstract

The accessory genes of prokaryote and eukaryote pangenomes accumulate by horizontal gene transfer, differential gene loss, and the effects of selection and drift. We have developed Coinfinder, a software program that assesses whether sets of homologous genes (gene families) in pangenomes associate or dissociate with each other (i.e. are 'coincident') more often than would be expected by chance. Coinfinder employs a user-supplied phylogenetic tree in order to assess the lineage-dependence (i.e. the phylogenetic distribution) of each accessory gene, allowing Coinfinder to focus on coincident gene pairs whose joint presence is not simply because they happened to appear in the same clade, but rather that they tend to appear together more often than expected across the phylogeny. Coinfinder is implemented in C++, Python3 and R and is freely available under the GNU license from https://github.com/fwhelan/coinfinder.

DATA SUMMARY

1. Coinfinder is freely available at https://github.com/fwhelan/coinfinder.

2. A list of the Identifiers of the genomes used within as well as all input/output files are available at https://github.com/fwhelan/coinfinder-manuscript.

INTRODUCTION

Pangenomes consist of core genes, common across all strains of a species, and accessory genes that are present in some but not all strains [1]. Accessory genes by definition are not essential to every strain of a species. Accessory genes are often pathogenicity islands, or associated with niche adaptation, or defence from predation, and so forth [2]. Why some members of a species might have some of these genes, while others do not is subject to debate [3, 4]. It is likely that some genes co-occur, or associate, because they positively influence each other's fitness in a particular set of host genomes. Similarly, we expect some genes to avoid, or dissociate with one another because their co-occurrence produces a negative fitness effect. We expect that genes whose products function together in a biochemical pathway, or that can combine to form a useful heteromeric protein complex, will appear together in the same genome more often than their observed frequency in the dataset would predict. For example, MYD88 consistently co-occurs with the genetic components of the MYD88-dependent TLR-signalling pathway in vertebrate species [5]. In contrast, genes that produce a toxic by-product when they are expressed in the same cell, or that perform the same function and therefore induce functional redundancy, are expected to appear together less often than their observed frequency in the dataset would predict. This is seen, for example, with siderophore biosynthetic gene clusters in *Salinispora* spp. where an isolate either has one iron-chelating siderophore or a different non-homologous system, but never both [6]. As a first step towards understanding these kinds of gene-to-gene interactions in the accessory pangenome, it is useful to identify genes that appear together or that avoid one another significantly more often than would be expected by chance.

Received 02 December 2019; Accepted 23 January 2020; Published 25 February 2020

Author affiliations: [1]School of Life Sciences, The University of Nottingham, Nottingham, UK; [2]Faculty of Biology, Medicine & Health, The University of Manchester, Manchester, UK.

***Correspondence:** James Oscar McInerney, mbzjom@exmail.nottingham.ac.uk

Keywords: pangenome; gene co-occurrence; gene association networks.

Abbreviations: ATP, adenosine triphosphate; COG, clustered orthologous groups; CPU, central processing unit; GEXF, graph exchange XML format; GPS, Global Pneumococcal Sequencing Project; MCL, the Markov cluster algorithm; ORF, open reading frame; pan-GWAS, pangenome genome-wide association studies; SNP, single nucleotide polymorphisms; TLR, toll-like receptor.

https://github.com/fwhelan/coinfinder

†These authors contributed equally to this work

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files.

000338 © 2020 The Authors

This is an open-access article distributed under the terms of the Creative Commons Attribution License. This article was made open access via a Publish and Read agreement between the Microbiology Society and the corresponding author's institution.

Previously established methodology can identify various forms of co-occurrence patterns in prokaryotes. For example, many tools (e.g. [7–9]) and tool comparisons [10] are available for the identification of species–species co-occurrence patterns in microbial communities. For example, the program SparCC identifies correlations in compositional data, including species presence–absence patterns within microbial communities [11]. Other tools, such as NetShift [12], find differences in species association networks of microbial communities across datasets (e.g. healthy versus diseased states). Similarly, methods have been established to identify associations between genotypic and phenotypic traits in pangenomes (i.e. gene-trait co-occurrence). Usually called pangenome genome-wide association studies (pan-GWAS), tools such as bugwas [13] and Scoary [14] compare components of the pangenome to a user-provided list of phenotypic traits. New methods such as SpydrPick [15] identify single nucleotide polymorphisms (SNP)–SNP co-occurrence patterns by comparing SNPs in multiple sequence alignments of proteins in microbial population genomic datasets.

A few approaches have focussed on gene–gene co-occurrence. Pantagruel [16] uses gene and species trees to identify genes which have similar patterns of gain and loss in a pangenome to define co-evolved gene modules. Similarly, CoPAP [17] searches for correlated patterns of gene gain and loss across a species tree to find co-evolutionary interactions of clustered orthologous groups (COGs). While conceptually similar to Coinfinder, these methodologies are based on phyletic patterns; further, the dissociation of genes is not considered by either method. The most similar method to Coinfinder in concept is the identification of correlogs and anti-correlogs, genes which favour or disfavour co-occurrence within a genome, by Kim and Price [18]. However, this method was not packaged into publicly available software and was not coupled with the pangenome concept, instead focusing on global patterns of gene associations across the bacterial Domain.

Table 1. Description of Coinfinder output files

Suffix	File description
_pairs.tsv	Tab-delimited list of significant coincident gene pairs
_nodes.tsv	Node list of all unique coincident genes and their D value
_edges.tsv	Edge list of significant gene–gene pairs and the associated *P*-value
_network.gexf	GEXF (Graph Exchange XML Format) v1.2 formatted network file. Nodes are coloured by connected component (i.e. coincident gene set) and sized by D value; edge thickness is proportional to the *P*-value of the coincident relationship between any two connected genes
_components.tsv	Tab-delimited list of all connected components within the gene–gene coincident network
_heatmap[0-X].pdf	Heatmap images (R, ggplot2 [35], ggtree [36]) of the presence–absence patterns of coincident components across input genomes. The heatmap is split across multiple files when needed for ease of visibility

Impact Statement

Coinfinder identifies genes that co-occur (associate) with or avoid (dissociate) each other across the accessory genomes of a pangenome of interest. Genes that associate or dissociate more often than expected by chance, suggest that those genes have a connection (attraction or repulsion) that is interesting to explore. Identification of these groups of genes will further the field's understanding of the importance of accessory genes. Coinfinder is a freely available, open-source software, which can identify gene patterns locally on a personal computer in a matter of hours.

Currently, the identification of gene–gene coincident patterns is not part of pangenome analyses tools. Pangenome pipelines – such as Roary [19], PIRATE [20], or Pandora (https://github.com/rmcolq/pandora) – cluster open reading frames (ORFs) into homologous gene clusters and report a presence–absence matrix of these clusters in relation to each input genome. These pipelines also generate statistics as to the numbers of core and accessory genes, a core gene alignment (from which a phylogeny can be determined), and the distribution of genes across genomes; however, these pipelines fail to determine statistically significant gene–gene relationships.

Here, we present Coinfinder, a command line software program that identifies coincident (associating or dissociating) genes across a set of input genomes. Coinfinder can run in any Unix environment using a user-specified number of processing cores. Coinfinder can be used to investigate the structure of strain- or species-pangenomes and is not restricted to prokaryote or eukaryote genomic input.

THEORY AND IMPLEMENTATION

Input

Coinfinder accepts genome content data in one of two formats: (a) the gene_presence_absence.csv output from Roary [19]; or (b) as a tab-delimited list of the genes present in each strain. If option (b) is used, genes should be clustered into orthologous groups/gene clusters prior to using Coinfinder (for example, using BLAST [21] and a clustering algorithm, such as MCL [22, 23]). Additionally, Coinfinder requires a Newick-formatted phylogeny of the genomes in the dataset. We suggest that this phylogeny can be constructed using the core genes from the input genomes as produced using programs such as Roary, or using ribosomal RNA genes, or a similar approach [24].

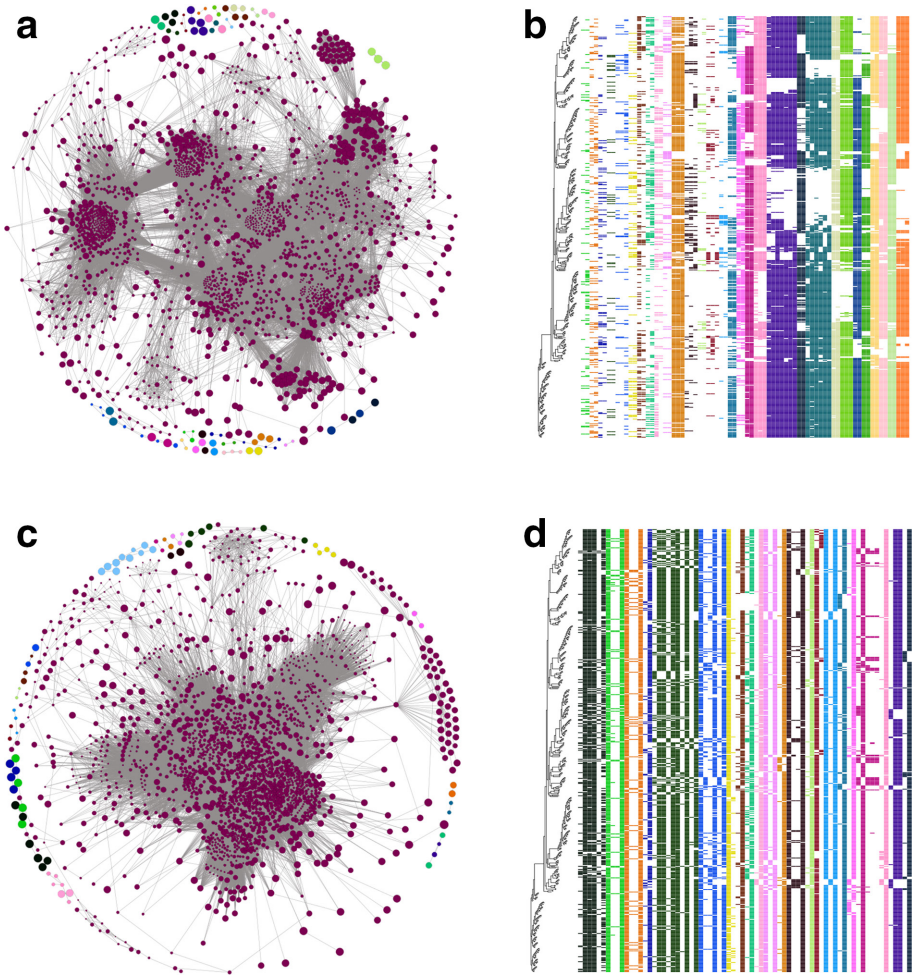

Fig. 1. Example of Coinfinder output. The network (a,c) and heatmap (b,d) outputs from Coinfinder executed on 534 *Streptococcus pneumoniae* genomes. (a, c) The resultant gene association (a) and dissociation (c) networks. Each gene (node) is connected to (edge) another gene if they statistically associate/dissociate with each other in the pangenome. Nodes are coloured by connected component (i.e. coincident gene sets) and the colours correspond to those used in the heatmap outputs. The network file Coinfinder generates includes all node and edge colouring; Gephi [37] was used to apply the Fruchterman Reingold layout. (b,d) A portion of the heatmaps of the presence/absence patterns of the associating (b) and dissociating (d) gene sets. Similar to the network, each set of coincident genes are co-coloured. Genes are displayed in relation to the input core gene phylogeny. Here the phylogeny tip and gene cluster labels have been removed from the output for clarity. Additionally, the largest connected component in the network (wine colour) has been omitted from the heatmap for ease of display.

Table 2. Real computational time for Coinfinder executed on a 534 genome dataset consisting of 2,813 accessory genes using different numbers of CPUs (GenuineIntel; Intel Xeon Gold 6142 CPU @ 2.60 GHz)

No. of CPUs	Real computer clock time
2	31m16.265s
4	17m56.973s
8	11m15.469s
16	7m44.942s
32	6m16.218s

Identifying coincident genes

For each set of genes in the input genomes, Coinfinder examines the presence/absence pattern of the gene pair to determine if they represent a coincident relationship; i.e. if

Table 3. Number of gene–gene associations identified with different sized subsets of the original 534 genome dataset

Iteration	*n*=400	*n*=300	*n*=200	*n*=100	*n*=50
1	75 586	52 038	24 196	1137	0
2	71 977	50 420	21 167	1389	0
3	75 190	51 459	25 545	1382	0

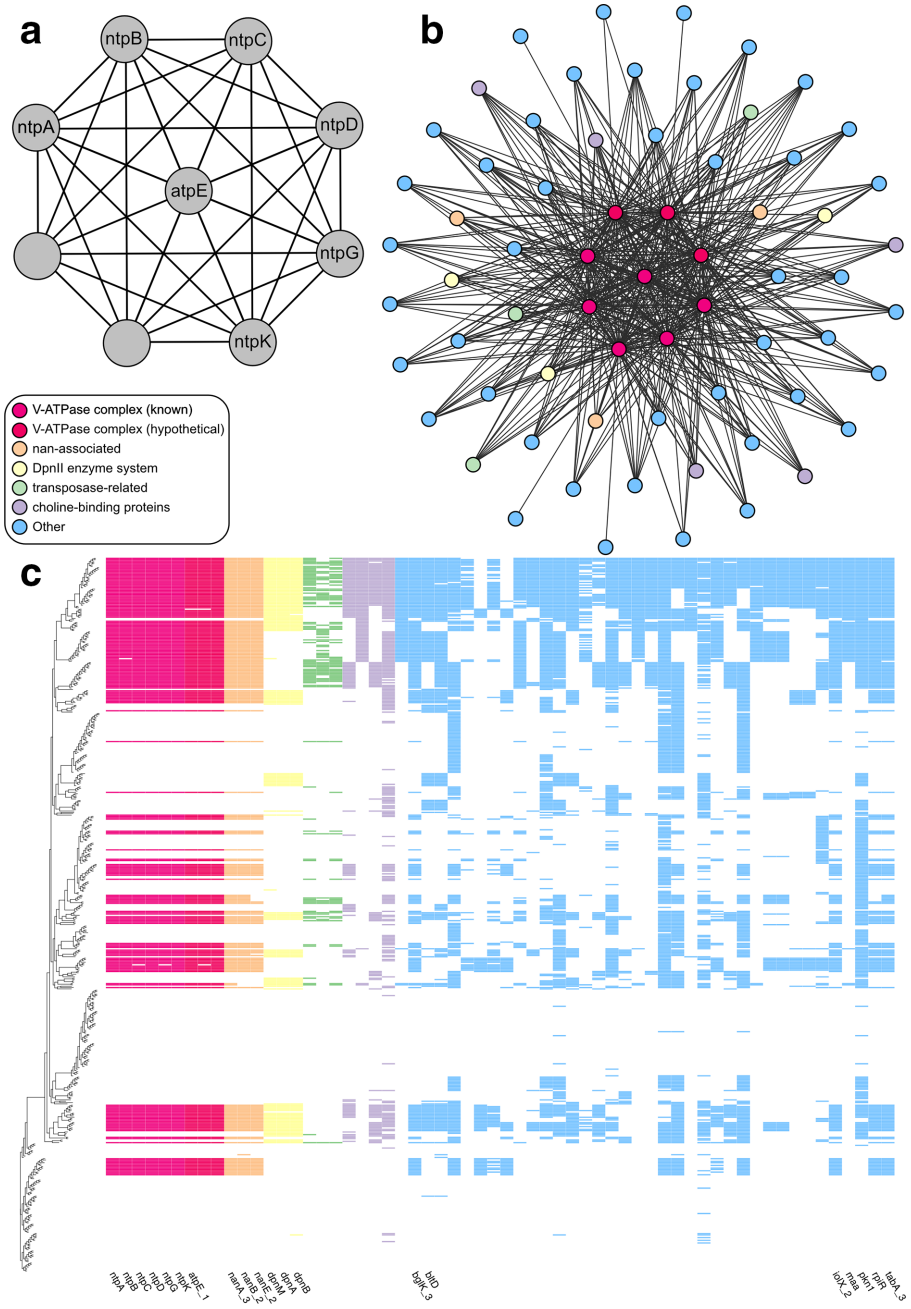

Fig. 2. Example of the association relationships Coinfinder can identify. (a) A clique of genes in the ntp operon which was identified within the association network (Fig. 1a). Six of these genes were correctly labelled with their gene names via the Prokka/Roary pipeline; one gene was given an alternative gene name often used as a synonym in the literature; a further two genes were listed as 'hypothetical proteins'. Collectively, the nine genes that compose the V-ATPase/ntp operon form cliques with an additional 51 genes. These cliques are shown as a network (b) and as a presence–absence heatmap (c). In the heatmap, unlabelled gene columns represent unnamed hypotheticals.

gene i and *gene j* are observed together or apart in the input genomes more often than would be expected by chance.

As a pre-processing step, the input gene set is culled for high- and low-abundance genes. Genes present in every genome (i.e. core genes) are removed as they cannot statistically associate or dissociate (i.e. be coincident with) another gene more or less often than expected. Similarly, genes whose presence is constrained to a small number of genomes will not produce significant associations, therefore low-abundance genes can be removed from the input at a user-determined cutoff. Coinfinder's default is to remove any gene present in less than 5 % of the input genomes.

Coinfinder has two modes for identifying coincident relationships: association and dissociation. When testing for gene associations, Coinfinder evaluates whether *gene i* and *gene j* of a given gene pair are observed together in the input genomes more often than would be expected by chance. More formally, for a set of genomes *N*, we define the probability of observing *gene i* as

$$P_i = N_i / N$$

where N_i is the number of occurrences of *gene i* in the dataset. The expected rate of association, E_A, of *gene i* with *gene j*, is then defined as:

$$E_A(ij) = P_i * P_j * N$$

and the observed rate of association, O_A, as:

$$O_A(ij) = N_{ij}$$

where N_{ij} is the number of times *gene i* and *gene j* are present within the same genome.

When testing gene dissociation, Coinfinder evaluates whether *gene i* and *gene j* of a given gene pair are observed separately in the input genomes more often than would be expected by chance. Formally, the expected rate of dissociation, E_D, is defined as

$$E_D(ij) = [P_i(1 - P_j) + P_j(1 - P_i)] * N$$

and the observed rate of dissociation, O_D, as

$$O_D(ij) = N_i + N_j - 2N_{ij}$$

In each mode, Coinfinder's default behaviour is to use a Bonferroni-corrected binomial exact test statistic (adapted from https://github.com/chrchang/stats) of the expected and observed rates to evaluate whether each gene pair are significantly coincident with each other.

Coincident genes that share an evolutionary history are more likely to have indirect correlations with each other. For example, if two genes are found to associate and each is observed only within a particular clade, the most parsimonious explanation for the observation is that the last common ancestor of the clade obtained both genes at the same evolutionary step. These two genes may, or may not, have a functional relationship with one another, and are of potential interest. However, non-monophyletic – or lineage-independent – genes that are dispersed throughout a phylogeny and are found to be significantly coincident are more likely to have a direct relationship with each other – their patchy phylogenetic distribution, combined with their statistically significant rate of association is *prima facie* evidence that they interact in some way, prefer a particular ecological niche, or have some other direct association with each other. Thus, Coinfinder focuses on identifying coincident relationships between lineage-independent accessory genes. To do this, Coinfinder uses a previously established phylogenetic measure of binary traits (D, as coded into the R function phylo.d [25]) to determine the lineage-dependence of each coincident gene. D is a measure of phylogenetic signal strength of a binary trait, which quantifies the amount of dispersion of the trait – here, the presence of a gene – over a phylogenetic tree [25]. Coinfinder does not implement a particular threshold for lineage independence but instead reports the D value of each gene in the output for the user to consider in conjunction with their input phylogeny. The calculations for gene association/dissociation and lineage independence are conducted independently of one another.

Output

Coinfinder visualizes the results of its analysis in two ways. First, Coinfinder produces a network in which each node is a gene family and each edge is a statement of significant gene association (corrected for lineage effects) or significant gene dissociation. The size of a node is proportional to the gene's D value. Second, Coinfinder generates a presence–absence heatmap, indicating the presence of coincident genes in the context of the input phylogeny. The genes in the heatmap are ordered by D value (from most lineage-independent to least) and are coloured according to coincident patterns.

Coinfinder produces a number of output files, with the default prefix of *coincident_*, as described in Table 1. Examples of the network and heatmap outputs of Coinfinder are shown in Fig. 1.

RESULTS

As an example, Coinfinder was executed using 534 *Streptococcus pneumoniae* genomes as input, a subset of the Global Pneumococcal Sequencing Project (GPS; https://www.pneumogen.net/gps/) whose ORFs were identified using Prokka [26] and clustered into orthologous gene families using Roary [19]. Coinfinder took 7.2 min (using 20 cores; see Table 2 for more runtime details) to examine the relationships between 2 813 gene families across 534 genomes (3 957 891 pairwise tests in total). Coinfinder identified 104 944 associating gene pairs, which clustered into 32 connected components or sets of genes that associate with each other. Similarly, Coinfinder took 7.5 min using 20 cores to identify 98 461 dissociate gene relationships within this dataset. The network and heatmap outputs of Coinfinder from this example set are shown in Fig. 1.

Although the availability of sequenced genomes is increasing rapidly, it is still rare to have access to such a large species-level pangenomic dataset. As such, the user could consider analyses at the genus- or family-level to increase dataset size. In order to identify the effect of input dataset size on Coinfinder's ability to identify gene–gene associations, we randomly subsetted the 534 genome *S. pneumoniae* dataset into datasets sized between 400 and 50 genomes (Table 3). Analyses of these data with Coinfinder returned less gene–gene associations than observed in the full dataset, and the number of associations observed decreased substantially with smaller numbers of genome inputs, culminating with no associations identified with an input of 50 genomes. Although this provides an estimate of the necessary number of input

genomes to Coinfinder, it should be noted that the power of Coinfinder will vary based on the average number of genes per genome as well as the diversity of genes within the dataset (i.e. the 'openness' of the pangenome).

Of the gene associations and dissociations that Coinfinder identified, many are in line with previous investigations of *S. pneumoniae* pangenomes. For example, we identify a large number of associations between widely dispersed genes, which agrees with evidence that *S. pneumoniae* has an extensive set of 'soft core' genes in its pangenome [27]. Further, many genes involved in coincident relationships are lineage independent, which is expected given the high natural competency of the species and, therefore, affinity for horizontal gene transfer events [28]. As an example, we focus on a V-ATPase present in *S. pneumoniae*. V-ATPases are enzymes which transport protons across the cell membrane in a process which hydrolyses ATP [29]. V-ATPases are intricate protein complexes, providing an excellent use case for Coinfinder's potential to identify the genes expected to co-occur as part of a multi-protein enzyme. While the V-ATPase in *S. pneumoniae* has been understudied, it has been well-documented in *S. pyogenes* and sister taxon *Enterococcus hirae* [29, 30]. In *E. hirae* the V-ATPase consists of 10–11 proteins organized into the ntp operon: ntpFIKECGABD(H)J [29]. In *S. pneumoniae,* the V-ATPase complex is predicted to contain nine proteins (KEGG pathway *spx_M00159* [29]). In the annotation of *S. pneumoniae* that we performed here, only six genes of the ntp operon were annotated successfully: *ntpA, ntpB, ntpC, ntpD, ntpG* and *ntpK*. Coinfinder identified consistent co-occurrence relationships between these six genes, forming a clique (i.e. a complete subgraph of gene associations; Fig. 2a). However, these six genes also co-occurred with other genes in the dataset; we extended our analyses to determine whether any other genes consistently co-occurred with all six genes of this operon. In doing so, we identified three genes – atpE, and two unnamed genes – with homology to *ntpE, ntpI* and *ntpG/H*, respectively, that consistently co-occur with the rest of the ntp operon (Fig. 2a). An additional 51 genes formed cliques with the genes of the ntp operon. Of the 51 genes, three encode neuraminadase genes from *nan* gene clusters (Fig. 2b–c). Another three genes co-occurring with the V-ATPase complex belong to the *dpnMAB* operon which encode the DpnII system implicated in DNA transformation (among other functions) [31] and an additional three are homologous to transposase IS66-related domains, perhaps suggesting how this operon has been horizontally transferred in this species (Fig. 2b–c). Additionally, four of these proteins contained a putative cell wall binding repeat ('*CW_binding_1*') which has been implicated in choline binding [32]. Choline-binding proteins (CBPs) contain a choline-binding module/domain which allows them to bind to the cell wall of *S. pneumoniae*, functioning as essential elements of cell division, as well as strong determinants of virulence [32, 33]. It is unknown why four CBPs co-occur with the V-ATPase complex; in eukaryotes, it has been shown that acetylcholine can be transmitted via the V-ATPase complex of vacuoles [34] but the result has not been generalized to prokaryotic cell membranes. A further 11 genes are of uncharacterized function. This example shows the power of Coinfinder in (a) identifying gene associations between proteins in a known protein complex; (b) being able to overcome poor gene annotations by looking for patterns in gene co-occurrence and gene association networks; and (c) being able to extrapolate those results to other genes with known protein interactions.

Coinfinder uses parallel processing to compute pairwise tests of coincident relationships. The most time-consuming step is the determination of the lineage-dependence of each gene; consequently, we have programmed this part of the software to run in parallel for only those genes that are found in statistically significant coincident relationships. For the *S. pneumoniae* example, using the input set of 2 813 accessory gene families, the lineage-dependence calculation was only necessary on the 1 961 genes deemed to be in coincident relationships. Using these data, the computation time varied from 6 to 31 min when using 32 to 2 CPUs, respectively (Table 2).

Conclusions

Coinfinder is an accurate and efficient tool for the identification of coincident gene relationships within pangenomes. Coinfinder is open-source software available from https://github.com/fwhelan/coinfinder.

Funding information
J.O.M. was awarded funding from the BBSRC no. BB/N018044/1 to support the work of M.J.R. and F.J.W. F.J.W. has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 793818.

Acknowledgements
The authors would like to thank the members of the McInerney research group for valuable input, as well as the Global Pneumococcal Sequencing Project for their dedication to open-source sequencing data.

Author contributions
F.J.W., M.R., and J.O.M., conceptualized this work. F.J.W., and M.R., built the software. F.J.W., validated and visualized the output data. F.J.W., wrote the original draft; F.J.W, M.R., and J.O.M., reviewed and edited the manuscript. J.O.M., acquired the funding and conducted project administration.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D *et al*. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A* 2005;102:13950–13955.
2. Brockhurst MA, Harrison E, Hall JPJ, Richards T, McNally A *et al*. The ecology and evolution of Pangenomes. *Curr Biol* 2019;29:R1094–R1103.
3. McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. *Nat Microbiol* 2017;2:17040.
4. Shapiro BJ. The population genetics of pangenomes. *Nat Microbiol* 2017;2:1574.

5. Tassia MG, Whelan NV, Halanych KM. Toll-Like receptor pathway evolution in deuterostomes. *Proc Natl Acad Sci U S A* 2017;114:7055–7060.
6. Bruns H, Crüsemann M, Letzel A-C, Alanjary M, McInerney JO *et al*. Function-Related replacement of bacterial siderophore pathways. *Isme J* 2018;12:320–329.
7. Ruan Q, Dutta D, Schwalbach MS, Steele JA, Fuhrman JA *et al*. Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics* 2006;22:2532–2538.
8. Faust K, Raes J. CoNet app: inference of biological association networks using Cytoscape [version 2; referees: 2 approved]. *F1000 Res [Internet]* 2016.
9. Ling Y, Watanabe Y, Okuda S. The human gut microbiome is structured to optimize molecular interaction networks. *Comput Struct Biotechnol J* 2019;17:1040–1046.
10. Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J *et al*. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *Isme J* 2016;10:1669–.
11. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. *PLoS Comput Biol* 2012;8:e1002687.
12. Kuntal BK, Chandrakar P, Sadhu S, Mande SS. 'NetShift': a methodology for understanding 'driver microbes' from healthy and disease microbiome datasets. *ISME J* 2019;13:442–454.
13. Wu C-H, Wu C-H, Charlesworth J, Stoesser N, Gordon NC *et al*. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat Microbiol [Internet]* 2016;1.
14. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol* 2016;17.
15. Pensar J, Puranen S, Arnold B, MacAlasdair N, Kuronen J *et al*. Genome-Wide epistasis and co-selection study using mutual information. *Nucleic Acids Res* 2019;47:e112.
16. Lassalle F, Veber P, Jauneikaite E, Didelot X. Automated reconstruction of all gene histories in large bacterial pangenome datasets and search for co-evolved gene modules with Pantagruel. *bioRxiv [Internet]* 2019;19:495–586.
17. Cohen O, Ashkenazy H, Levy Karin E, Burstein D, Pupko T. CoPAP: coevolution of presence-absence patterns. *Nucleic Acids Res* 2013;41:W232–.
18. Kim P-J, Price ND. Genetic co-occurrence network across sequenced microbes. *PLoS Comput Biol* 2011;7:1002340.
19. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S *et al*. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015;31:3691–3693.
20. Bayliss SC, Thorpe HA, Coyle NM, Sheppard SK, Feil EJ. PIRATE: a fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. *Gigascience* 2019;8.
21. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.
22. Dongen S. *Performance Criteria for Graph Clustering and Markov Cluster Experiments*. Amsterdam, The Netherlands: CWI (Centre for Mathematics and Computer Science); 2000.
23. Dongen S. *A Cluster Algorithm for Graphs*. Amsterdam, The Netherlands, The Netherlands: CWI (Centre for Mathematics and Computer Science); 2000.
24. Yarza P, Richter M, Peplies J, Euzeby J, Amann R *et al*. The All-Species living tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. *Syst Appl Microbiol* 2008;31:241–250.
25. Fritz SA, Purvis A. Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. *Conserv Biol* 2010;24:1042–1051.
26. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.
27. Obolski U, Gori A, Lourenço J, Thompson C, Thompson R *et al*. Identifying genes associated with invasive disease in *S. pneumoniae* by applying a machine learning approach to whole genome sequence typing data. *Sci Rep* 2019;9.
28. Hiller NL, Sá-Leão R. Puzzling over the pneumococcal Pangenome. *Front Microbiol* 2018;9:2580.
29. Lolkema JS, Chaban Y, Boekema EJ, Composition S. Structure, and distribution of bacterial V-type ATPases. vol. 35. *J Bioenerg Biomembr* 2003;35:323–335.
30. Wang B, Qin W, Ren Y, Zhou X, Jung M-Y *et al*. Expansion of Thaumarchaeota habitat range is correlated with horizontal transfer of ATPase operons. *Isme J* 2019.
31. Johnston C, Polard P, Claverys J-P. The DpnI/DpnII pneumococcal system, defense against foreign attack without compromising genetic exchange. *Mob Genet Elements* 2013;3:e25582.
32. Maestro B, Sanz JM. *Choline binding proteins from Streptococcus pneumoniae: A dual role as enzybiotics and targets for the design of new antimicrobials*, 5. Antibiotics: MDPI AG; 2016.
33. Gosink KK, Mann ER, Guglielmo C, Tuomanen EI, Masure HR. Role of novel choline binding proteins in virulence of Streptococcus pneumoniae. *Infect Immun* 2000;68:5690–5695.
34. Peters C, Bayer MJ, Bühler S, Andersen JS, Mann M *et al*. Trans-complex formation by proteolipid channels in the terminal phase of membrane fusion. *Nature* 2001;409:581–588.
35. Wickham H. *ggplot2: Elegant Graphics for Data Analysis [Internet]*. New York: Springer-Verlag New York; 2009.
36. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree : an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 2017;8:28–36.
37. Bastian M, Heymann S, Jacomy M. Gephi: An open source software for exploring and manipulating networks. BT - International AAAI Conference on Weblogs and Social. *Int AAAI Conf Weblogs Soc Media* 2009:361–362.

Five reasons to publish your next article with a Microbiology Society journal

1. The Microbiology Society is a not-for-profit organization.
2. We offer fast and rigorous peer review – average time to first decision is 4–6 weeks.
3. Our journals have a global readership with subscriptions held in research institutions around the world.
4. 80% of our authors rate our submission process as 'excellent' or 'very good'.
5. Your article will be published on an interactive journal platform with advanced metrics.

Find out more and submit your article at microbiologyresearch.org.