Evidence for selection in the abundant accessory gene content of a prokaryote pangenome

Fiona J Whelan\textsuperscript{a}, Rebecca J Hall\textsuperscript{a,b}, James O McInerney\textsuperscript{a}\textsuperscript{*}

May 3, 2021

\textsuperscript{a}School of Life Sciences, University of Nottingham, East Dr., Nottingham, NG7 2TQ, United Kingdom
\textsuperscript{b}Institute of Microbiology and Infection, College of Medical and Dental Sciences, University of Birmingham, Birmingham, B15 2TT, United Kingdom

ORCIDs:
FJ Whelan 0000-0001-9165-4859
RJ Hall 0000-0001-6701-5408
JO McInerney 0000-0003-1885-2503

*Corresponding author:
James O. McInerney,
School of Life Sciences,
Faculty of Medicine and Health Sciences,
University of Nottingham,
B41 Life Sciences Building, East Dr
Nottingham,
NG7 2TQ,
United Kingdom
+44 (0)115 823 0141
james.mcinerney@nottingham.ac.uk

Keywords: pangenome, microbial genomics, selection

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Abstract

A pangenome is the complete set of genes (core and accessory) present in a phylogenetic clade. We hypothesize that a pangenome’s accessory gene content is structured and maintained by selection. To test this hypothesis, we interrogated the genomes of 40 Pseudomonas species for statistically significant coincident (i.e. co-occurring/avoiding) gene patterns. We found that 86.7% of common accessory genes are involved in ≥1 coincident relationship. Further, genes that co-occur and/or avoid each other - but are not vertically inherited - are more likely to share functional categories, are more likely to be simultaneously transcribed, and are more likely to produce interacting proteins, than would be expected by chance. These results are not due to coincident genes being adjacent to one another on the chromosome. Together, these findings suggest that the accessory genome is structured into sets of genes that function together within a given strain. Given the similarity of the Pseudomonas pangenome with open pangenomes of other prokaryotic species, we speculate that these results are generalizable.

Introduction

The mechanisms governing the existence of the pangenome - the totality of genes across a given set of genomes [Tettelin et al.2005] - have been debated, with evidence being presented for both neutral and selective processes [McInerney et al.2017a, Shapiro2017, McInerney et al.2017b]. Some evidence suggests that the accessory gene content within pangenomes has arisen as a consequence of extensive horizontal gene transfer (HGT) coupled with large effective population size and thus evolve neutrally [Andreani et al.2017]. In contrast, others argue that accessory genome evolution is dominated by selective pressures, and that diversity is maintained by selection acting on gene gain by horizontal acquisition, as well as gene loss [McInerney et al.2017a, Goyal2018, Bobay and Ochman2018]. Gene content changes therefore enable and are facilitated by population differentiation and niche adaptation [McInerney et al.2017a, Goyal2018].

Within, we argue that one way to determine the evolutionary pressures at play on the pangenome is to focus on gene-gene relationships within the accessory genome. If the pangenome is governed by neutrality, we would expect any observed structure in the accessory genome - including, for example, the co-occurrence of co-functional genes - to have arisen by chance and thus to be rare. In contrast, if the accessory gene content is primarily or substantially shaped by natural selection, we would expect the accessory genome to
be structured into groups of genes that work well together. Similarly, it would be reasonable to expect, at least in some cases, that genes whose interaction would be detrimental to the host to avoid being in the same genome. In this way, we expect that genes which consistently co-occur or avoid each other across a large pangenome to be under selective pressures to maintain these patterns. As such, we use gene pair information to ask whether a portion of the accessory gene content is governed by selective pressure.

To answer this question, we use a previously published software called Coinfinder [Whelan et al.2020] to focus on gene-gene association (i.e. co-occurrence) and dissociation (i.e. avoidance) patterns, collectively referred to as coincident relationships. We argue that, if evolving neutrally, we would not expect to see more coincident genes in the pangenome than would be expected by chance. In contrast, selective pressure would manifest as a significant proportion of the pangenome consisting of coincident gene relationships. In this case, we might further ask whether the assigned functionalities, gene expression patterns and known protein-protein interaction partners of these genes also suggest co-selection. To conduct these analyses rigorously, we exclude genes that are vertically acquired. Coincident genes that are clade-specific (i.e. lineage dependent genes) are likely to be coincident because they have remained within a single clade for the duration of their evolutionary history. Similarly, we exclude coincident gene pairs that share significant physical linkage (i.e. are co-localized on the genome) in order to ensure that the analysis isn’t focused on genes which form functional units such as operons. Removing both of these types of genes provides us with a stringent set of lineage-independent coincident gene pairs with which to answer our research question.

In this paper, we focus on the genus *Pseudomonas*. Although pangenome analyses are typically conducted at the species-level, the *Pseudomonas* genus shares properties with other well-studied open pangenomes, including the ability to persist in a variety of niches [Stanier et al.1966] and containing ample accessory gene content (approximately 81% in *P. aeruginosa*) [Kung et al.2010, Ding et al.2018]. For example, although estimates vary based on dataset size and analysis method, the accessory gene content (i.e. percentage of accessory vs. core genes across a genome set) of *Escherichia coli* is estimated to be between 86-91% [Ding et al.2018, Decano and Downing2019], *Streptococcus pneumoniae* 68-90% [Ding et al.2018], [Hiller and Sá-Leão2018], and *Bacillus subtilis* 86% [Ding et al.2018, Wu et al.2020]. In our analysis, we use coincident genes to ask whether the abundant accessory gene content of this microbial pangenome is maintained by selection. We identify coincident gene presence-absence patterns that deviate from random expectation, and build network representations of the data to identify sets of coincident genes. We find that 86.7% of abundant accessory genes form ≥1 significant gene association/dissociation relationship. Co-occurring gene pairs are more likely to share functionality, be transcribed together, and to encode proteins...
that interact with each other more often than randomly paired accessory genes. Together, these results provide consilient lines of evidence supporting the hypothesis that selection on genome content drives the evolution of the abundant accessory pangenome of this prokaryote.

Results

Species and gene distribution in the *Pseudomonas sp.* dataset

209 complete assemblies of *Pseudomonas* species were obtained from pseudomonas.com. The genomes were distributed across 40 *Pseudomonas* species, the most prevalent of which were *P. aeruginosa* (n=81), *P. putida* (n=18), *P. fluorescens* (n=15), *P. syringe* (n=13), and *P. stutzeri* (n=10) (Supplementary Figure 1a). 25 species were represented by a single genome within the dataset. Furthermore, a total of 22 genomes were included that do not have a species identification.

Across these 40 species, we identified a total of 96,694 orthologous gene clusters (Supplementary Figure 1a). Of these, only 1,365 (1.41%) were identified in ≥90% of strains (i.e. “core” genes). The mean number of genes per genome was 5,530, meaning that in a given strain, an average of 24.9% of its genes are core. PAO1 – a commonly studied *P. aeruginosa* lab strain [Klockgether et al.2010] – was found to contain 5,601 genes (compared to 5,688 as annotated on pseudomonas.com), of which 1,494 are core genes. A total of 88,792 (91.8%) genes were found in ≤15% of genomes (Supplementary Figure 1a). While the number of accessory genes varies across strains, the number of core genes is remarkably stable (Supplementary Figure 1b).

The *Pseudomonas* pangenome contains an abundance of coincident gene relationships

Using the gene annotations provided by pseudomonas.com and gene clusters identified with Roary [Page et al.2015], the 96,694 orthologous gene clusters (herein referred to as gene clusters) were used to identify coincident gene relationships within the pangenome. Any gene cluster that was considered core or present in ≤5% of strains were culled from coincident analyses, leaving 13,864 gene clusters across 209 genomes for testing. From these analyses (detailed in the Methods), we identified a significantly associating dataset comprised of 293,123 co-occurring gene pairs. We build a network representation of the gene pairs such that each gene is represented by a node which is connected to another gene iff those genes co-occur with each other. In this way, we identify 433 connected components or gene sets (Figure 1a). The 433 associating gene sets
are well dispersed across the *Pseudomonas* sp. core gene phylogeny and none are species-specific, indicating the effect of culling lineage-dependent genes from the analysis (Supplementary Figure 2). Similarly, we determined the significantly dissociative dataset which contains 421,080 dissociative gene pairs organized into 13 connected components (Figure 1b).

Of the 13,864 accessory gene clusters identified in ≥5% of *Pseudomonas* strains (i.e. the abundant accessory genes tested by Coinfinder [Whelan et al.2020]), 8,007 (57.7%) were lineage-independent (see Methods, Supplementary Figure 3). Of these 8,007 clusters, 6,329 and 3,589 formed associating and dissociating relationships, respectively (Figure 1c, Supplementary Figure 4). Accounting for the genes involved in both types of relationships, a surprising 6,948 (86.7%) of abundant lineage-independent accessory genes were involved in ≥1 coincident relationship. While gene dissociations were identified across all three non-core gene categories, gene associations were only identified in the two more rare gene categories (Cloud and Shell genes; Figure 1c). Similar results were found when both lineage-independent and -dependent genes were considered (Supplementary Figure 5a).

Of the 6,329 genes forming coincident relationships identified, 2,970 (46.9%) are involved in both associating and dissociating relationships, meaning that they both co-occur with, and avoid other genes in the pangenome (Figure 1d; black nodes). These 2,970 dual-relationship genes account for 268,647 (91.6%) of all gene-gene associations and 418,698 (99.4%) of all gene-gene dissociations (Figure 1d). That is to say that almost half of the coincident genes account for the majority of coincident gene relationships. On average, associating genes form relationships with 94 and a median of 18 other genes (Figure 1e). However, the distribution is uneven, with 24.3% of genes forming fewer than five connections to other genes (1,542 genes < the 25th percentile; Figure 1e). The 624 association hubs (i.e. genes with >1.5x the upper interquartile range) each have ≥290 gene associations and account for 50.8% of the total observed gene association patterns. In contrast, dissociations in the *Pseudomonas* pangenome are driven by a small number of dissociating hub genes (n=3) that each form ≥1,110 gene dissociation relationships. Among the associating and dissociating hub genes are a diversity of functions including transcriptional regulators, transporter subunits, metabolic enzymes, and an abundance of hypothetical proteins. Interestingly, for those genes that were found to have both types of coincident relationships, no gene acts as both an associating and dissociating hub (Figure 1e). The number of hub genes increase when lineage-dependent genes are included in these analyses (Supplementary Figure 5b).
Co-localization of coincident genes

HGT and differential gene loss are the main contributing factors to pangenome formation [Azarian et al. 2020]. If functionally related gene pairs are found in close proximity on a genome, then they may have been acquired in a single HGT event, and their co-occurrence pattern might be a consequence of the HGT process, and not a consequence of natural selection. However, many known protein interactions occur between genes that are dispersed across the genome (for e.g. proteins produced by genes crr and ptsG form the the EII complex in enteric bacteria and are not in close proximity on the genome [Deutscher et al. 2006]). To explore whether co-localization and the simultaneous transfer of genes is responsible for gene association relationships in the pseudomonads, we compared the mean syntenic distance of associating genes, versus the mean syntenic distance of abundant accessory gene pairs chosen at random. The average chromosome length across the dataset is 6.2 Mbps; which, in addition to the chromosome being circular, means that the furthest away two genes could be from each other is $\sim 3.1$ Mbps. The mean syntenic distance between randomly paired abundant accessory genes is bell-shaped which fits our expectation of randomly dispersed genes. In contrast, associating gene pairs more often share significant localization (Figure 2a); however, only 8.6% of all co-occurring gene pairs have a mean syntenic distance of $< 150kbp$. This suggests that a proportion of, but not all, gene-gene co-occurrence is due to co-localized genes.

In order to ask whether the co-localization patterns of gene pairs generalize to that of gene sets, we next considered gene associations in terms of their connected component (i.e. associating gene set; Figure 1a). We observe 41 gene sets (26%) that are composed of pairs of genes with a mean pairwise syntenic distance of $\leq 150$ kbp (Figure 2b). We used PPanGGOLiN [Gautreau et al. 2020] to generate pangenome graphs of Pseudomonas sp. (Supplementary Figure 6) and the P. aeruginosa subset (Figure 2c) to visualize the genomic context of co-localized gene sets. For example, the P. aeruginosa pangenome graph includes a set of neighbouring co-occurring genes associated with flagellar assembly (Figure 2c, box 1). Interestingly, these genes are all encoded on the same strand, and this path in the pangenome graph bypasses a set of 16 genes which also show homology to flagellar assembly genes (Supplementary Table 1). A given genome may contain one but not both of these sets of genes, indicating possible redundancy of this function within the pangenome. We also observe gene sets that share very little physical linkage, such as a set of three unnamed genes involved in outer membrane permeability (Figure 2c, box 2; Supplementary Table 1). Still, other gene sets have mixed levels of co-localization amongst their membership. For example, a subset of P. aeruginosa strains contain three neighbouring genes (encoded on the same strand) that co-occur with a fourth gene sharing no physical linkage with the other three and encoded on the opposite strand (Figure 2c,
These four genes likely co-occur because they all function as components of the methionine salvage pathway (Supplementary Figure 7, Supplementary Table 1).

Coincident genes share functionality

The association (or dissociation) of genes alone does not infer a biological interaction between them (i.e. correlation does not infer causation; [Blanchet et al.2020]). If the accessory genome is influenced by selection, we could expect that coincident genes might be more likely to act together - for example, towards a shared functional goal - for the benefit of the host. Alternatively, genes might act together towards their own selection (e.g. in DNA secretion [Draghi and Turner2006] or integrative conjugative elements (ICEs; Wozniak and Waldor2010]). Using Gene ontology (GO) annotations as a proxy for gene functionality, we calculated the functional overlap of each coincident gene pair in comparison to randomly paired abundant accessory genes, indicating some structure in the accessory pangenome (Figure 3a). We identified a greater overlap in GO annotations between coincident gene pairs then randomly paired accessory genes. Specifically, 71.1% of associating and 69.4% of dissociating gene pairs shared GO annotations when compared to only 50.6 (±0.1)% of randomly paired accessory genes (Figure 3a). This indicates that coincident genes share function with each other more often than would be expected by chance. The percentage of shared GO annotations amongst associating genes increased to 74% when only non-syntenic genes were considered (Supplementary Figure 8). Given these results, we calculated whether particular GO terms were more likely to share annotation in a coincident gene pair compared to the expected term-sharing frequency (Figure 3b). 150 GO terms were found to be overrepresented in gene-gene associations, including pilus assembly (GO:0009297; p=1.41e-05), type II protein secretion system complex (GO:0015627; p=1.35e-08), and antibiotic biosynthetic process (GO:0017000; p=4.84e-10) (Figure 3b red points, Supplementary Table 2). In contrast, 60 GO terms were overrepresented in dissociation relationships, including ATP-binding cassette (ABC) transporter complex (GO:0043190; p=4.96e-52), and drug transmembrane transport (GO:0006855; p=2.16e-07) (Figure 3b blue points, Supplementary Table 2).

A subset of GO annotations was enriched in both associating and dissociating gene pairs (Figure 3b purple points; Supplementary Table 2). This appears counterintuitive, but may correspond to, for example, two multi-gene functional units that dissociate from one another but whose genes within the unit strongly associate with each other. For example, gene pairs annotated with transmembrane transporter activity (GO:0022857) were enriched in association (p=8.39e-06) and dissociation gene relationships (p=3.01e-28; Figure 3c). While some genes formed independent co-occurring cliques or solitary dissociation patterns
(not shown), the majority of genes clustered into groups of associating genes (Supplementary Figure 9a) that dissociated from each other (Figure 3c). Some of these cluster avoidance patterns appear to be largely due to species boundaries (e.g. clusters 7 and 15; Supplementary Figure 9b) but most are independent of phylogeny and syntenic relationships (Supplementary Figure 9bc). Although many of these genes are hypothetical or only loosely annotated, there are, for example, genes for an efflux pump (Resistance-nodulation-division (RND) family transporters) in cluster 2 that dissociate from genes for a different efflux pump (glutathione-regulated potassium-efflux system protein, KefB) in cluster 3 (Supplementary Table 3), indicating a possible example of functional redundancy or niche partitioning within this system. We also identify gene-gene association patterns between genes with known biological interactions such as bfmS and bfmR which form the BfmS/R two-component system (cluster 9; [Cao et al.2014]) and cynS and cynT which are involved in cyanate decomposition (cluster 11; [Guilloton et al.1993, Luque-Almagro et al.2008]).

The above calculations of intersecting GO annotations rely on known gene information. While Pseudomonas sp. is a well-studied genus with well-annotated genomes, many of the identified coincident gene pairs involve interactions between hypothetical proteins or genes without a known GO association. 51,531 (17.6%) and 23,168 (7.9%) of the associating and dissociating gene pairs, respectively, involve at least one hypothetical gene (Figure 3d). Specifically, 95% of coincident gene pairs involving hypothetical genes are between hypothetical and annotated genes. Given our finding that many annotated coincident gene pairs share function, coincident relationships between hypothetical and annotated genes can help us generate hypotheses concerning the role these hypothetical proteins play in the Pseudomonas sp. pangenome. A subset of GO terms was found to be statistically more likely to be coincident with hypothetical genes when compared to the annotated coincident gene pairs (Supplementary Table 4). For example, the “beta-lactamase activity” (p=1.86e-06; GO:0008800) GO annotation was assigned to two genes that collectively associated with 120 annotated and 33 hypothetical genes. In particular, 42% of the genes that associate with an ampC homolog (most closely related to PDC-8 [Rodríguez-Martínez et al.2009]) were annotated as hypothetical proteins, and only seven had a gene name annotation in ≥1 genome (Figure 3e, Supplementary Table 5). This gene association cluster (including ampC) is present in ≥4 Pseudomonas species (4 named, 6 unnamed strains), and does not share considerable co-localization across the pangenome (Supplementary Figure 10). Similar investigations of the remaining hypothetical-annotated gene pairs may yield further hypotheses concerning the role of hypothetical proteins in this pangenome.
Gene co-occurrence is associated with co-transcription and protein-protein interactions

Using publicly available RNA-Seq transcription data, we examined how often associating gene pairs were transcribed together compared to randomly paired accessory genes. Due to limitations on the availability of good quality publicly available gene transcription data, we restricted our analysis to *P. aeruginosa* (81 of 209 genomes). Across the *P. aeruginosa* pangenome, we calculated the frequencies with which a given gene pair was transcribed together compared to that of only one of the two genes in a pair. We report this ratio of gene expression, such that a ratio of 1.0 indicates that - across the *P. aeruginosa* pangenome - it is as likely to see both genes transcribed together as it is for only one of the pair to be transcribed (Figure 4a). Across samples and experiments, associating gene pairs were more often co-transcribed than were randomly paired abundant accessory genes (Figure 4a), indicating a possible shared function or interaction between these genes. This result holds when only non-syntenic gene associations are considered (Supplementary Figure 11). Similar analyses of co-transcription could not be performed on the dissociating gene pairs as these pairs are not present within the same genomes.

Given the rate of co-transcription of associating genes, we asked how often coincident genes are involved in known protein-protein interactions. Using the STRING database [Szklarczyk et al.2019], we first identified the number of protein-protein interactions between randomly paired accessory genes as 1.4 (±0.03)%%. This percentage may seem low; however, we expect that documented protein-protein interactions are more likely to involve well-studied, abundant (likely core), house-keeping proteins, or those which share evolutionary histories with each other, which are precisely the genes which are excluded in our analyses of linkage-independent accessory genes. However, we identified protein-protein interactions between 9.4% of associating gene pairs (11.4% of all annotated associating pairs; Figure 4b). These data represent 2.5% of all known protein-protein interactions within *P. aeruginosa*; that is to say that - even when excluding core or vertically inherited genes - associating gene relationships recapitulates a percentage of all known protein interactions in this species. As expected, evidence of interactions between dissociating genes were identified at a rate less than randomly paired genes (Figure 4b).

Discussion

We recently developed a novel method for the identification of coincident gene presence-absence patterns within pangenomes [Whelan et al.2020]. Although pangenome analyses are usually focused on a particular
species, here we applied this approach to 209 publicly available *Pseudomonas sp.* genomes to ask whether a
portion of pangenome gene content is determined by selective pressure. Across the dataset, 86.7% of lineage-
independent, abundant accessory genes consistently associated with, or dissociated from, at least one other
gene in the pangenome. This represents a lot more genetic structure within the abundant accessory genome
than we would expect if neutral processes were driving pangenome formation. We found that these gene
pairs share functional annotations, are co-transcribed, and produce proteins that interact with each other
more often than expected when compared to randomly paired abundant accessory genes. These findings were
independent of whether the genes are lineage-dependent or are predominantly vertically transmitted and the
association remained even when co-localized genes were excluded. The fact that we found statistically sig-
nificant associations between non-syntenic genes is strong evidence for selection because it identifies genes
that share functionality despite being dispersed in the genome. Together, these data suggest that the assem-
blage of abundant accessory genes is structured in this pangenome is more likely explained by selection than
by neutral processes. This work has implications for our understanding of prokaryote pangenomes as a whole.

However, we should be mindful of the limitations of this approach. For example, because of the nature
of identifying coincident relationships, this analysis can only be conducted when a gene is present frequently
enough across a dataset. In this case, we used a threshold of >5% abundance, which equated to a focus
on the 13,864 most abundant accessory genes. Further, gene-gene co-occurrence does not necessarily infer
direct gene-gene interactions. Although such a high-throughput examination of gene-gene co-occurrence re-
lationships in pangenomes may be rare [Kim and Price2011, Press et al.2016, Cohen et al.2012], there is a
century of literature on species-species co-occurrence patterns [Blanchet et al.2020, Forbes1907, Michael1920,
Diamond1975, Connor and Simberloff1979]. In this research, it has been explicitly shown that in at least
some cases, species-species co-occurrence does not necessarily imply species-species ecological interactions.
In their recent Perspectives article, Blanchet *et al.* present seven arguments for why ecological interaction
should not be assumed from co-occurrence data [Blanchet et al.2020]. Although some of these arguments
are species-specific, many apply to gene-gene data as well. For example, the authors argue that in some
cases, species occurrences depend on the environment, and what appears as a species-species co-occurrence
pattern may actually be an indirect interaction of both species with their environment [Blanchet et al.2020];
similarly, *geneA* and *geneB* may co-occur due to a preference for a shared abiotic factor - environment,
nutrient, metabolite etc. - instead of indicating a direct gene-to-gene interaction. We suggest that further
*in vitro* investigations of gene pairs could help clarify these levels of interactions. Further, the methodology
used here - the identification of coincident gene relationships based on statistically similar or dissimilar gene
presence/absence patterns - will not identify all associations in the pangenome. For example, asymmetrical


dependencies will have been missed; in the case where geneA relies on geneB for an interaction but not vice versa, we would expect to see geneA present only in the presence of geneB, but that geneB could be present without geneA in a given genome. So called “event horizon genes” or those genes whose presence “leads the way” for the introduction of many other genes [McInerney et al.2020], will also not be identified by use of the Coinfinder software. Because these gene-gene patterns are hard to distinguish from random presence/absence patterns, their influence on the structure of the pangenome will be harder to determine.

With this caveat in mind, we sought to provide evidence for the possibility that a sizable subset of the gene-gene associations within the abundant accessory genes of the Pseudomonas pangenome may be due to direct interactions. The fact that many associating gene pairs tend to neighbour each other indicates this potential. Neighbouring genes often assemble into sets of co-transcribed genes which either physically interact to form protein complexes (e.g. manXYZ [Erni et al.1987]) or act as part of a shared metabolic pathway (e.g. the lac operon [Jacob and Monod1961]). However, many coincident genes which were not co-localized had overlapping functionality. This isn’t too surprising given previous evidence of co-selection of genes ([Cui et al.2018]) and single-nucleotide polymorphisms (SNPs; [Cui et al.2018], [Pensar et al.2019]) across disperse loci. These genes could still directly interact, although could also indicate a response to a shared abiotic factor (for e.g. genes present in response to a particular environmental niche). On the other hand, genes with shared functionalities which actively avoid each other would seem to suggest a more directed type of interaction. Further, strain-level variation in gene essentiality can also contribute to gene-gene association patterns. For example, it has been recently shown that the horizontal acquisition of specific genes in E. coli can make a gene that was already present in the genome as a non-essential accessory gene become newly essential ([Rousset et al.2021]). Either way, evidence for interactions at the protein level clearly indicate direct gene-gene interactions in the accessory pangenome.

One of the inspirations for this work was the recent suggestion that one way of better elucidating whether the pangenome is evolving neutrally or adaptively was to focus on the gene as the evolutionary unit [Shapiro2017]. Examining gene-gene relationships, as we have done here, is not the only gene-focused approach to understanding the evolutionary pressures present on prokaryote pangenomes. For example, analyses could be conducted to determine whether accessory genes are under selective pressures. Further, gene knockout and long-term evolutionary experiments could be combined to determine the effect of individual genes on the pangenome. Similarly, we could adapt the unit of observation to think about groups of genes; for example, we could ask whether operons, or genes which comprise multi-subunit proteins form coincident relationships. We propose these results concerning gene-gene coincident relationships as one line of evidence
for testing hypotheses of selective pressures on the accessory genome. We encourage further work in these areas to be contributed to this debate.

We focused our analysis on Pseudomonas sp. due to its diverse, well-studied pangenome [Kung et al.2010, Freschi et al.2019, Mosquera-Rendón et al.2016, Udaondo et al.2016, Dillon et al.2019], well-annotated genomes [Winsor et al.2016], and generalizability to other prokaryotic open pangenomes in terms of core-to-accessory gene ratios, and multiple environmental niches. Our results suggest genetic structure within this pangenome, and we hope that additional research, using different methodologies and pangenomes, will help further these findings.

Materials & Methods

Sequence acquisition & pangenome analysis

Genome annotations were retrieved from pseudomonas.com in GFF3 format [Winsor et al.2016] on 1 March 2019 and include 209 complete genome assemblies. Despite the availability of thousands of draft genomes, we restricted our study to completely assembled and curated strains, due to recent work suggesting that the quality of genome assembly can greatly impact predicted pangenome quality and size [Denton et al.2014].

Genes were clustered into gene families using Roary 3.12.0 [Page et al.2015] with a 70% BLASTP percentage identity cutoff. Definitions of core (90%≤x≤100%), soft core (89%≤x<90%), shell (15%≤x<89%), and cloud (x<15%) genes are as in Roary. All core genes (present in ≥90% of Pseudomonas genomes) were individually aligned using MAFFT v7.310 [Katoh and Standley2013], the alignments concatenated, and curated using Gblocks ([Castresana2000]; parameters as in [Creevey et al.2011], specifically allow gap positions = half, minimum length of block = 2). A core gene phylogeny was constructed from this curated and concatenated alignment using IQ-TREE [Nguyen et al.2015] using the GTR+I+G substitution model (as justified in [Abadi et al.2019]) and midpoint rooted. A total of 19 genome annotations contained plasmids which were not considered in these analyses.

Evaluation of gene coincident relationships

Coincident relationships between gene pairs were determined using Coinfinder [Whelan et al.2020]. Briefly, for each pair of genes in the input accessory genome, Coinfinder examines their presence/absence patterns to determine if they represent a coincident relationship (i.e. if they co-occur or avoid each other across the pangenome more often than expected by chance). Statistically significant coincident gene pairs were
determined by Coinfinder via a Bonferroni-corrected binomial exact test statistic, and the lineage dependence of each gene was calculated using a previously established phylogenetic measure of binary traits (D; [Fritz and Purvis2010]). Coinfinder was run with upper- and lower-filtering gene abundance thresholds of 90% and 5%, respectfully. A threshold of $D \geq -0.4$ was used based on the frequency of genes, their distribution across species in the core gene phylogeny, and the distribution of counts of coincident gene pairs (Supplementary Figure 3). The resulting associating and dissociating networks were visualized using Gephi [Bastian et al.2009]. Hub genes were defined as those with more gene-gene relationships than 1.5 times the upper interquartile range.

In order to determine whether coincident gene pairs were more likely to share functional annotations, gene expression patterns, or protein-protein interactions (see below), we compared these results against the null model by generating random abundant accessory gene pairs. To do so, accessory genes that were included in the Coinfinder analysis (i.e. were between 5-90% abundance with $D \geq -0.4$) were paired at random to match the mean number of associating/dissociating gene pairs ($n=357,102$) in 100 replicates (herein referred to as random abundant accessory gene pairs). This was accomplished by creating a list of all possible paired combinations of abundant accessory gene pairs and creating $n=100$ random permutations of the list to a length of 357,102. The specific use of these random abundant accessory gene pairs is outlined in the following Methods sections.

**Gene co-localization and pangenome structure analysis**

The physical linkage between genes in a gene pair was determined both for associating, and for random abundant accessory gene pairs. For a given gene pair, the physical distance between $geneA$ and $geneB$ was calculated for each genome for which both $geneA$ and $geneB$ reside. (For this reason, syntenic distance information could not be calculated for dissociating gene pairs.) From these $geneA$-$geneB$ distances for each genome, a mean syntenic distance was computed and plotted. In analyses of non-syntenic genes, only those gene pairs separated by a mean syntenic distance of $\geq 150$ kbp were considered.

A pangenome graph was created with PPanGGOLiN [Gautreau et al.2020]. In order to maintain consistency with the gene cluster information used throughout this study, PPanGGOLiN was provided with the gene clusters as determined by Roary. A Python script was used to redefine nodes in the pangenome graph to remain consistent with the definitions of core, soft core, shell, and cloud that are used by Roary. The nodes of the resulting graph were recoloured to represent the associating gene sets as determined by Coinfinder. The
network was visualized in Gephi [Bastian et al.2009]. KEGG was used to investigate metabolic pathways
[Kanehisa and Goto2000].

Functional annotations of coincident genes

Gene ontology (GO) term annotations for each of the 209 genomes were collected from pseudomonas.com on
22 March 2019. A minimum of one matching GO term annotation was necessary to consider a gene pair as
having overlapping function. Overlapping annotations were determined by examining only those gene pairs
for which both genes had a GO term annotation. After removing gene pairs for which GO term annotations
were missing for one or both genes, a total of 246,637 (84.1%) associating, and 379,439 (90.11%) dissociating
gene pairs remained. These were compared to 100 replicates of randomly paired abundant accessory genes
as described above. Bonferroni-corrected binomial tests (computed in R [Team2017]) were used to determine
which GO terms were under- or over-represented in the coincident gene pairs when compared to the random
abundant accessory gene pairs.

Separately, GO terms which were significantly associated with genes of hypothetical function was deter-
mined. Genes were defined as hypothetical if every instance of the gene across all genomes in which it was
found were annotated as “hypothetical protein”. Bonferroni-corrected binomial tests were used to determine
GO terms over-represented in gene pairs involving an annotated and hypothetical gene. Sub-networks of
specific gene-gene interaction pairs were displayed using Gephi [Bastian et al.2009].

Gene expression analysis

Short read archive (SRA) transcript data from the following P. aeruginosa RNA-Seq experiments (paired-end
reads with a range of 4,450,537 - 41,817,822 reads per sample) were used to test co-transcription levels of
gene-gene pairs: SRP163899 (n=2 samples), SRP215630 (n=9), and SRP191772 (n=8; [Zhang et al.2019]).
The reads from each RNA-Seq sample were mapped using Bowtie2 [Langmead and Salzberg2012] to the gene
content of the P. aeruginosa genomes in the dataset (n=81). In a given genome, a gene was considered
transcribed if ≥85% of the gene’s length was covered by ≥2 reads. Across the dataset, a gene cluster was
considered transcribed if it was transcribed in ≥75% of the genomes in which it was present. The ratio of
gene expression is the ratio of gene cluster pairs which are co-transcribed versus those in which only one of
the two genes were transcribed. Therefore, a ratio of 1.0 would mean that, across all P. aeruginosa genomes,
paired genes are just as likely to be co-transcribed as for exclusively one of the two genes to be transcribed; a
ratio of 2.0 would mean that paired genes are twice as likely to be transcribed together across the pangenome.

14
Protein interaction analysis

The STRING database [Szklarczyk et al.2019] was used to identify whether the protein products of associating, dissociating, and random abundant accessory gene pairs interact with each other. The protein network data and associated FASTA sequences for *P. aeruginosa* were obtained from https://string-db.org. The FASTA sequences for the proteins in this network were assembled into a BLAST database to map homologous gene clusters to the IDs in the STRING protein network, with the criteria of ≥85% coverage and ≥90% sequence identity. Calculations of the coincident gene pairs were compared to 100 replicates of randomly paired abundant accessory gene pairs as described above.

Acknowledgements

This research was supported by a Marie Skłodowska-Curie Individual Fellowship (GA no. 793818) awarded to FJW and BBSRC Responsive Mode Grant BB/N018044/1 awarded to JMcI. We would like to thank the tireless efforts of The *Pseudomonas* Genome Database and their funders. We acknowledge critical intellectual conversations with P. Mulhair and M.R. Domingo-Sananes.

Author Contributions

FJW is the primary author of this prepared manuscript. FJW collected, processed, and analysed all data. RJH provided key intellectual insights and Figures for all metabolic pathways considered within. FJW, and JOM conceptualized the experimental outline. FJW conducted all data analyses and wrote this manuscript. All authors edited and approved the manuscript.

Competing Interest

The authors declare no competing interests.

Data Availability

All raw data, including genome and gene identifiers, used in this work is available as a SQL Schema from github.com/fwhelan/pseudomonas-manuscript including maps between genomes, genes, gene clusters, and GO term annotations. An R markdown file, pseudomonas_manuscript.Rmd, available at github.com/fwhelan/pseudomonas-manuscript details how each Figure was generated from the available raw data. The
set of Python scripts and SQL queries used to generate data matrices, and an R Markdown file of the R code
used to generate all Figures are available from github.com/fwhelan/pseudomonas-manuscript.

References

[Abadi et al.2019] Abadi, S., Azouri, D., Pupko, T., and Mayrose, I. 2019. Model selection may not
be a mandatory step for phylogeny reconstruction. Nature Communications 10:934.

[Andreani et al.2017] Andreani, N. A., Hesse, E., and Vos, M. 2017. Prokaryote genome fluidity is
dependent on effective population size. ISME Journal 11:1719–1721.

[Azarian et al.2020] Azarian, T., Huang, I.-T., and Hanage, W. P. 2020. Structure and Dynamics
of Bacterial Populations: Pangeneome Ecology, pp. 115–128. In H. Tettelin and D. Medini (eds.), The
Pangenome: Diversity, Dynamics and Evolution of Genomes. Springer International Publishing, Cham.

[Bastian et al.2009] Bastian, M., Heymann, S., and Jacomy, M. 2009. Gephi: An open source software
for exploring and manipulating networks. BT - International AAAI Conference on Weblogs and Social.
International AAAI Conference on Weblogs and Social Media pp. 361–362.

[Blanchet et al.2020] Blanchet, F. G., Cazelles, K., and Gravel, D. 2020. Co-occurrence is not
evidence of ecological interactions. Ecology Letters.

[Bobay and Ochman2018] Bobay, L.-M. and Ochman, H. 2018. Factors driving effective population size
and pan-genome evolution in bacteria. BMC Evolutionary Biology 18:153.

[Cao et al.2014] Cao, Q., Wang, Y., Chen, F., Xia, Y., Lou, J., Zhang, X., Yang, N., Sun, X.,
Zhang, Q., Zhuo, C., Huang, X., Deng, X., Yang, C. G., Ye, Y., Zhao, J., Wu, M., and
Lan, L. 2014. A Novel Signal Transduction Pathway that Modulates rhl Quorum Sensing and Bacterial
Virulence in Pseudomonas aeruginosa. PLoS Pathogens.

[Castresana2000] Castresana, J. 2000. Selection of Conserved Blocks from Multiple Alignments for Their
Use in Phylogenetic Analysis. Molecular Biology and Evolution 17:540–552.

[Cohen et al.2012] Cohen, O., Ashkenazy, H., Burstein, D., and Pupko, T. 2012. Uncovering the
coevolutionary network among prokaryotic genes. Bioinformatics 28:389–394.

[Connor and Simberloff1979] Connor, E. F. and Simberloff, D. 1979. The Assembly of Species Com-
munities: Chance or Competition? Ecology.
Universally Distributed Single-Copy Genes Indicate a Constant Rate of Horizontal Transfer. *PLoS ONE* 6:e22099.

Cui, Y., Yang, C., Qiu, H., Wang, H., Yang, R., and Falush, D. 2018. The landscape of coadaptation in Vibrio parahaemolyticus.

Decano, A. G. and Downing, T. 2019. An *Escherichia coli* ST131 pangenome atlas reveals population structure and evolution across 4,071 isolates. *Scientific Reports*.

Denton, J. F., Lugo-Martinez, J., Tucker, A. E., Schrider, D. R., Warren, W. C., and Hahn, M. W. 2014. Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. *PLoS Computational Biology*.

Deutscher, J., Francke, C., and Postma, P. W. 2006. How Phosphotransferase System-Related Protein Phosphorylation Regulates Carbohydrate Metabolism in Bacteria. *Microbiology and Molecular Biology Reviews*.

Diamond, J. 1975. Assembly of Species Communities, pp. 342–344. In J. Diamond and M. Cody (eds.), Ecology and Evolution of Communities. Harvard University Press, Boston.

Dillon, M. M., Thakur, S., Almeida, R. N., Wang, P. W., Weir, B. S., and Guttman, D. S. 2019. Recombination of ecologically and evolutionarily significant loci maintains genetic cohesion in the *Pseudomonas syringae* species complex. *Genome Biology* 20.

Ding, W., Baumdicker, F., and Neher, R. A. 2018. panX: pan-genome analysis and exploration. *Nucleic Acids Research* 46.

Draghi and Turner, J. A. and Turner, P. E. 2006. DNA secretion and gene-level selection in bacteria. *Microbiology*.

Erni, B., Zanolari, B., and Kochers, H. P. 1987. The Mannose Permease of *Escherichia coli* Consists of Three Different Proteins. *The Journal of biological chemistry*.

Forbes, S. A. 1907. On the Local Distribution of Certain Illinois Fishes: An Essay in Statistical Ecology. *Illinois Natural History Survey Bulletin* 7:273–297.

Freschi, L., Vincent, A. T., Jeukens, J., Emond-Rheault, J. G., Kukavica-Ibrulj, I., Dupont, M. J., Charette, S. J., Boyle, B., and Levesque, R. C. 2019. The Pseu-
domonas aeruginosa Pan-Genome Provides New Insights on Its Population Structure, Horizontal Gene Transfer, and Pathogenicity. *Genome Biology and Evolution*.

[Fritz and Purvis2010] Fritz, S. A. and Purvis, A. 2010. Selectivity in mammalian extinction risk and threat types: A new measure of phylogenetic signal strength in binary traits. *Conservation Biology* 24:1042–1051.

[Gautreau et al.2020] Gautreau, G., Bazin, A., Gachet, M., Planel, R., Burlot, L., Dubois, M., Perrin, A., Médigue, C., Calteau, A., Cruveiller, S., Matias, C., Ambroise, C., Rocha, E. P., and Vallenet, D. 2020. PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLoS Computational Biology*.

[Goyal2018] Goyal, A. 2018. Metabolic adaptations underlying genome flexibility in prokaryotes prokaryotes. *PLoS Genetics* pp. 1–27.

[Guilloton et al.1993] Guilloton, M. B., Lamblin, A. F., Kozliak, E. I., Gerami-Nejad, M., Tu, C., Silverman, D., Anderson, P. M., and Fuchs, J. A. 1993. A physiological role for cyanate-induced carbonic anhydrase in Escherichia coli. *Journal of Bacteriology*.

[Hiller and Sá-Leão2018] Hiller, N. L. and Sá-Leão, R. 2018. Puzzling Over the Pneumococcal Pangenome. *Frontiers in Microbiology* 9:2580.

[Jacob and Monod1961] Jacob, F. AND MONOD, J. 1961. Genetic regulatory mechanisms in the synthesis of proteins.

[Kanehisa and Goto2000] Kanehisa, M. AND GOTO, S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes.

[Katoh and Standley2013] Katoh, K. AND STANDLEY, D. M. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* 30:772–780.

[Kim and Price2011] Kim, P.-J. AND PRICE, N. D. 2011. Genetic Co-Occurrence Network across Sequenced Microbes. *PLoS Comput Biol* 7:1002340.

[Klockgether et al.2010] Klockgether, J., Munder, A., Neugebauer, J., Davenport, C. F., Stanke, F., LARBIG, K. D., Heeb, S., Schöck, U., Pohl, T. M., WIEHLMANN, L., AND TÜMMLER, B. 2010. Genome diversity of Pseudomonas aeruginosa PAO1 laboratory strains. *Journal of Bacteriology*.
[Kung et al.2010] KUNG, V. L., OZER, E. A., AND HAUSER, A. R. 2010. The Accessory Genome of Pseudomonas aeruginosa. *Microbiology and Molecular Biology Reviews*.

[Langmead and Salzberg2012] LANGMEAD, B. AND SALZBERG, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9:357–359.

[Luque-Almagro et al.2008] LUQUE-ALMAGR, V. M., HUERTAS, M. J., SÁEZ, L. P., LUQUE-ROMERO, M. M., MORENO-VIVIÁN, C., CASTILLO, F., ROLDÁN, M. D., AND BLASCO, R. 2008. Characterization of the Pseudomonas pseudoalcaligenes CECT5344 cyanase, an enzyme that is not essential for cyanide assimilation. *Applied and Environmental Microbiology*.

[McInerney et al.2017a] McINERNEY, J. O., McNALLY, A., AND O’CONNELL, M. J. 2017a. Why prokaryotes have pangenomes. *Nature Microbiology* 2:17040.

[McInerney et al.2017b] McINERNEY, J. O., McNALLY, A., AND O’CONNELL, M. J. 2017b. Reply to ‘The population genetics of pangenomes’. *Nature Microbiology* 2:1575–1575.

[McInerney et al.2020] McINERNEY, J. O., WHELAN, F. J., DOMINGO-SANANES, M. R., McNALLY, A., AND O’CONNELL, M. J. 2020. Pangenomes and selection: The public goods hypothesis. In The Pangene: Diversity, Dynamics and Evolution of Genomes. Springer, Cham.

[Michael1920] MICHAEL, E. L. 1920. Marine Ecology and the Coefficient of Association: A Plea in Behalf of Quantitative Biology. *The Journal of Ecology*.

[Mosquera-Rendón et al.2016] MOSQUERA-RENDÓN, J., RADA-BRAVO, A. M., CÁRDENAS-BRITO, S., CORREDOR, M., RESTREPO-PINEDA, E., AND BENÍTEZ-PÁEZ, A. 2016. Pangenome-wide and molecular evolution analyses of the Pseudomonas aeruginosa species. *BMC Genomics*.

[Nguyen et al.2015] NGUYEN, L.-T., SCHMIDT, H. A., VON HAESLER, A., AND MINH, B. Q. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* 32:268–274.

[Page et al.2015] PAGE, A. J., CUMMINS, C. A., HUNT, M., WONG, V. K., REUTER, S., HOLDEN, M. T., FOOKES, M., FALUSH, D., KEANE, J. A., AND PARKHILL, J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31:3691–3693.

[Pensar et al.2019] PENSAR, J., PURANEN, S., ARNOLD, B., MACALASDAIR, N., KURONEN, J., TONKIN-HILL, G., PESONEN, M., XU, Y., SIPOLA, A., SÁNCHEZ-BUSÓ, L., LEE, J. A., CHEWAPREECHA, C., BENTLEY, S. D., HARRIS, S. R., PARKHILL, J., CROUCHER, N. J., AND CORANDER, J. 2019. Genome-wide epistasis and co-selection study using mutual information. *Nucleic acids research*.
[Press et al.2016] PRESS, M. O., QUEITSCH, C., AND BORENSTEIN, E. 2016. Evolutionary assembly patterns of prokaryotic genomes. *Genome Research* p. gr.200097.115.

[Rodríguez-Martínez et al.2009] RODRÍGUEZ-MARTÍNEZ, J. M., POIREL, L., AND NORDMANN, P. 2009. Extended-spectrum cephalosporinases in Pseudomonas aeruginosa. *Antimicrobial Agents and Chemotherapy*.

[Rouset et al.2021] ROUSET, F., CABELAS-CABALLERO, J., PIASTRA-FACON, F., FERNÁNDEZ-RODRÍGUEZ, J., CLERMONT, O., DENAMUR, E., ROCHA, E. P., AND BIKARD, D. 2021. The impact of genetic diversity on gene essentiality within the Escherichia coli species. *Nature Microbiology*.

[Shapiro2017] SHAPIRO, B. J. 2017. The population genetics of pangenomes. *Nature Microbiology* 2:1574–1574.

[Stanier et al.1966] STANIER, R. Y., PALLERONI, N. J., AND DOUDOROFF, M. 1966. The aerobic pseudomonads: a taxonomic study. *Journal of general microbiology*.

[Szklarczyk et al.2019] SZKLARCZYK, D., GABLE, A. L., LYON, D., JUNGE, A., WYDER, S., HUERTA-CEPAS, J., SIMONOVIC, M., DONCHEVA, N. T., MORRIS, J. H., BORK, P., JENSEN, L. J., AND MERING, C. 2019. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* 47:D607–D613.

[Team2017] TEAM, R. C. 2017. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*.

[Tettelin et al.2005] TETTELIN, H., MASIGNANI, V., CIESLEWICZ, M. J., DONATI, C., MEDINI, D., WARD, N. L., ANGIUOLI, S. V., CRABTREE, J., JONES, A. L., DURKIN, A. S., DEBOY, R. T., DAVIDSEN, T. M., MORA, M., SCARSELLI, M., MARGARIT Y ROS, I., PETERSON, J. D., HAUSER, C. R., SUNDARAM, J. P., NELSON, W. C., MADUPU, R., BRINKAC, L. M., DODSON, R. J., ROSOVITZ, M. J., SULLIVAN, S. A., DAUGHERTY, S. C., HAFT, D. H., SELENGUT, J., GWINN, M. L., ZHOU, L., ZAFFAR, N., KHOURI, H., RADUNE, D., DIMITROV, G., WATKINS, K., O’CONNOR, K. J. B., SMITH, S., UTTERBACK, T. R., WHITE, O., RUBENS, C. E., GRANDI, G., MADOFF, L. C., KASPER, D. L., TELFORD, J. L., WESSELS, M. R., RAPPUOLI, R., AND FRASER, C. M. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences* 102:13950–13955.

[Udaondo et al.2016] UDAONDO, Z., MOLINA, L., SEGURA, A., DUQUE, E., AND RAMOS, J. L. 2016. Analysis of the core genome and pangenome of Pseudomonas putida. *Environmental Microbiology*.
[Whelan et al.2020] Whelan, F. J., Rusilowicz, M., and McInerney, J. O. 2020. Coinfinder: Detecting significant associations and dissociations in pangenomes. *Microbial Genomics* 6.

[Winsor et al.2016] Winsor, G. L., Griffiths, E. J., Lo, R., Dhillon, B. K., Shay, J. A., and Brinkman, F. 2016. Enhanced annotations and features for comparing thousands of *Pseudomonas* genomes in the Pseudomonas genome database. *Nucleic Acids Research* 44:D646–D653.

[Wozniak and Waldor2010] Wozniak, R. A. and Waldor, M. K. 2010. Integrative and conjugative elements: Mosaic mobile genetic elements enabling dynamic lateral gene flow.

[Wu et al.2020] Wu, H., Wang, D., and Gao, F. 2020. Toward a high-quality pan-genome landscape of Bacillus subtilis by removal of confounding strains. *Briefings in Bioinformatics*.

[Zhang et al.2019] Zhang, Y., Zhou, C. M., Pu, Q., Wu, Q., Tan, S., Shao, X., Zhang, W., Xie, Y., Li, R., Yu, X. J., Wang, R., Zhang, L., Wu, M., and Deng, X. 2019. *Pseudomonas aeruginosa* regulatory protein AnvM controls pathogenicity in anaerobic environments and impacts host defense. *mBio*.
Figure 1: **Network of coincident relationships in the Pseudomonas sp. accessory pangenome.** Relationships between significantly associating (a) and dissociating (b) gene pairs are shown as gene-gene networks. Only nodes with a D ≥ -0.4 (i.e. sufficiently lineage-independent) are displayed. Nodes (i.e. gene clusters) are connected to other nodes if-and-only-if there is a significant coincident relationship between them. Nodes are coloured by the connected component which they belong to; in other words, nodes are coloured by significantly coincident gene sets. The size of the node is proportional to the D-value of the gene cluster (the larger the node, the more lineage-independent the gene is); the thickness of the edge is reversely proportional to the p-value associated with the coincident relationship. **c.** The counts of the genes present in the pangenome, the abundant accessory subset, as well as those which associate, dissociate, or form either type of relationship (i.e. are coincident). **d.** A gene-gene network of all lineage-independent coincident gene relationships. Edges are coloured by association (red) and dissociation (blue) relationships. Genes which form both association and dissociation relationships are represented by black nodes, genes which only associate by white, and genes which only dissociate by yellow. **e.** The distribution of gene-gene relationships across genes. Boxplots display the first and third quartiles, with a horizontal line to indicate the median, and whiskers extend to 1.5 times the interquartile range. Associating and dissociating “hub” genes are coloured.
Figure 2: Co-localization amongst associating gene pairs. a. Associating genes are more likely to be co-localized than are randomly assigned abundant accessory gene pairs on Pseudomonas sp. chromosomes. b. 26% of all sets of associating genes (i.e. connected components of genes which share co-occurrence patterns) do not share significant physical linkage as defined by the mean syntenic distance between all genes within a gene set. Coloured gene sets correspond to labelled boxes in part C. c. The pangenome graph of the P. aeruginosa subset of the Pseudomonas dataset. The pangenome graph of the full dataset is available in Supplementary Figure 5. Labelled boxes show examples of gene association clusters that are co-localized (box 1, turquoise genes), are not co-localized (boxes 2, red genes), and have variable levels of syntenic distance (boxes 3, green genes). For visibility, cloud genes are not shown.
Figure 3: Coincident (associating and dissociating) gene pairs have more overlapping GO term annotations when compared with random gene pairs. a. 71.1% of associating gene pairs share the same GO annotations compared with 50.6 (± 0.1)% of randomly paired genes. b. Triangular plots of GO term annotation within coincident gene space. Each GO term is represented by a point whose location is determined by how frequently genes with that term are found in the associating, dissociating, and random gene pair categories. GO terms which are significantly overrepresented in any category are coloured c. Coincident gene relationships for genes annotated with transmembrane transporter activity (GO:002857). Edges are coloured by the type of interaction (associating, red; dissociating, blue). A Figure showing only the associating edges is provided in Supplementary Figure 8a. d. The proportion of coincident gene pair relationships which exist between annotated and hypothetical genes. e. A network of gene (node) association relationships (edges) depicting the associations of ampC (orange) with hypothetical (gray) and annotated (yellow) genes.
Figure 4: **Associating genes are more likely to be co-transcribed.** a. The ratio of gene expression between associating gene pairs and random abundant accessory gene pairs. The ratio is calculated as the proportion of times that both genes in a gene pair are consistently co-transcribed across *P. aeruginosa* genomes versus the proportion of times that only one of the two genes is transcribed. Symbols represent different publicly-available RNA-Seq experimental projects. b. Protein-protein interaction pairs as compared to the STRING database indicate more interactions in the associating gene pairs compared to the dissociating and random gene-gene data. 100 replicates of randomly paired genes were used to obtain a mean of 1.4 (±0.03)%.
Supplementary Figure 1: Distribution of *Pseudomonas* species and genes in the dataset. a. The number of genomes per species present in the dataset (n=209). a, inset. The total gene cluster counts (n=96,694) across all strains broken down into the following categories: Cloud genes (present in 0% ≤ strains < 15%), Shell genes (15% ≤ strains < 89%), Soft core genes (89% ≤ strains < 90%), and Core genes (90% ≤ strains ≤ 100%). Collectively, cloud, shell, and soft core genes are referred to as accessory genes. b. The distribution of core and accessory genes across all strains.
Supplementary Figure 2: **Mean presence of the genes comprising each connected component in the association dataset.** Each connected component (i.e. set of genes which associate with each other) is well dispersed across the core gene phylogeny indicating that vertically inherited, or lineage-dependent genes, have been successfully culled from the dataset. Each connected component is indicated by a single, randomly coloured column in the heatmap.
Supplementary Figure 3: D-value distribution across the nodes found to have significant co-occurring (a) and avoidance (b) relationships in the complete genome dataset (n=209). The D-value indicates the amount of lineage-independence of a particular gene compared to a phylogeny. Based on the counts and distribution of genes across the core gene phylogeny, a D-value cutoff of -0.4 was chosen, resulting in the inclusion of 6329 and 3589 gene clusters (i.e. nodes), respectively, in the co-occurrence and avoidance datasets. c. An example of the distribution of 2 genes across the core gene phylogeny with D values on either side of the chosen threshold. d. The distribution of counts of coincident gene pairs identified by Coinfinder at various cutoffs of lineage-independence. The vertical dashed line indicates the cutoff used in this analysis.
Supplementary Figure 4: A non-log transformed version of the gene counts per category of gene cluster type. A continuous representation of the counts shown in Figure 1c.

Supplementary Figure 5: Coincident gene relationships between lineage-dependent and -independent genes. a. Similar patterns of coincident gene relationships are seen when lineage-dependent genes are included when compared to counts of only lineage-independent genes (Figure 1c). b. However, the distribution of relationships per gene differs when lineage-dependent genes are included, including an increase in the number of hub genes when compared to those identified between lineage-independent genes (Figure 1e). Boxplots display the first and third quartiles, with a horizontal line to indicate the median, and whiskers extend to 1.5 times the interquartile range.
Supplementary Figure 6: **The graph of the *Pseudomonas sp.* pangenome.** The pangenome graph of the 209 *Pseudomonas sp.* genomes as determined by PPanGGOLiN. Each gene cluster (node) is coloured by its connected component with the same colour scheme used in **Figure 1a.** For visibility, cloud genes are not shown.

Supplementary Figure 7: **The location of the co-occurring gene set within the simplified pathway of methionine salvage.** Homologs of the gene clusters in question are bolded and the steps they are involved with are coloured. These 4 genes were found to significantly co-occur but not all genes within the gene set are co-localized on the genome. This visual of select reactions and metabolites of the methionine salvage pathway is based off of information obtained from KEGG.
Supplementary Figure 8: **Non-syntenic coincident gene pairs share functionality.** Similar to the analysis of all coincident gene pairs (Figure 3a), the results hold when only those with a mean syntenic distance of $\geq 150,000$bp between genes are considered. Because dissociating genes by definition are not found in the same genome, this analysis could not be performed on dissociating gene pairs.
Supplementary Figure 9: **Genes annotated as transmembrane transporters are over represented in both gene association and dissociation relationships.**

**a.** The gene association relationships amongst transmembrane transporters; this network is the same as that displayed in Figure 3c but with the dissociation edges hidden for visibility. Labels represent gene cluster numbers.

**b.** The mean genomic distance between genes with the GO annotation of transmembrane transporters is evenly distributed across the pangenome.

**c.** The presence absence patterns of the gene clusters across the *Pseudomonas sp.* phylogeny. Clusters are coloured and numbered as in part A and Figure 3c.
Supplementary Figure 10: More hypothetical genes co-occur with an *ampC* homolog than would be expected by chance. **a.** The network of genes associating with *ampC*. Genes with named annotations are labelled. **b.** These genes do not share significant co-localization with each other when compared with randomly chosen abundant accessory genes. **c.** The presence/absence patterns of these genes across the *Pseudomonas sp.* phylogeny indicates a lack of evolutionary history of these genes. Genes are coloured as to whether they are annotated (yellow) or hypothetical (gray). The *ampC* homolog is coloured orange.
Supplementary Figure 11: **Gene co-expression ratios of non-syntenic gene pairs.** The gene co-expression ratio when only genes with $\geq 150,000$bp between them were considered.
Supplementary Table 1: Gene names and descriptions of the gene sets used as examples of high, low, and variable physical distance in Figure 2. Gene cluster descriptions are based on the descriptions of each gene within the cluster; parentheses indicate how often a given description was observed.

| Box # | Gene cluster ID | Gene name/description |
|-------|----------------|-----------------------|
| 1     | group_31493    | hypothetical protein (25); HAD family hydrolase (5); conserved hypothetical protein (1) |
| 1     | group_37342    | nucleotidyltransferase (24); hypothetical protein (2); sugar nucleotidyltransferase (2); putative sugar nucleotidyltransferase (1) |
| 1     | group_49093    | hypothetical protein (26); flagellar assembly protein FlIT (3) |
| 1     | group_49094    | flagellar protein FliS (23); flagellar biosynthesis protein FliS (3); B-type flagellar protein FliS (2); hypothetical protein (1) |
| 1     | group_49095    | B-type flagellar hook-associated protein 2 (18); branched-chain alpha-keto acid dehydrogenase subunit E2 (7); flagellar capping protein FliD (4) |
| 1     | group_49096    | hypothetical protein (29) |
| 1     | group_49097    | glycosyl transferase family 1 (14); hypothetical protein (9); flagellar glycosyl transferase, FgtA (3); flagellar glycosyl transferase FgtA (2); glycosyl transferase (1) |
| 1     | group_49098    | hypothetical protein (18); methyltransferase type 12 (10) |
| 1     | group_49099    | flagellar hook-associated protein FlgL (28); flagellar hook-associated protein type 3 FlgL (1) |
| 2     | group_1102     | hypothetical protein (40); dehydrogenase (17); aldose sugar dehydrogenase YliI (9); oxidoreductase (4); putative dehydrogenase (2); aldose sugar dehydrogenase (1); Soluble aldose sugar dehydrogenase, PQQ-dependent (1) |
| 2     | group_1963     | porin (70); Glucose/carbohydrate outer membrane porin OprB precursor (6); Glucose-selective porin OprB (2); carbohydrate porin (1); outer membrane porin OprB (1) |
| 2     | group_7677     | outer membrane protein OprG (66); outer membrane protein W (7); Outer membrane protein OprG precursor (5); Outer membrane protein W precursor (2); OmpW family protein (1) |
|   | group   | Description                                                                                                                                 |
|---|---------|---------------------------------------------------------------------------------------------------------------------------------------------|
| 3 | group_11613 | S-adenosylmethionine decarboxylase proenzyme (66); S-adenosylmethionine decarboxylase (13); S-adenosylmethionine decarboxylase proenzyme, prokaryotic class 1A (1) |
| 3 | group_6267 | enolase-phosphatase E1 (30); haloacid dehalogenase (23); 2,3-diketo-5-methylthio-1-phosphopentane phosphatase (13); enolase-phosphatase E-1 (4); 2,3-diketo-5-methylthiopentyl-1-phosphate enolase-phosphatase (2); enolase-phosphatase (2) |
| 3 | group_7302 | methylthioribulose-1-phosphate dehydratase (42); methylthioribulose 1-phosphate dehydratase (38); probable sugar aldolase (1)                        |
| 3 | group_7553 | spermidine synthase (75); polyamine aminopropyltransferase 1 (6)                                                                          |

Supplementary Table 2: **Summary of statistically enriched GO terms in the association and dissociation datasets.** See SupTable2.xlsx.

Supplementary Table 3: **Summary of genes with transmembrane transporter activity over represented in both association and dissociation gene relationships.** See SupTable3.xlsx.

Supplementary Table 4: **Summary of statistically enriched GO terms with hypothetical pairs.** See SupTable4.xlsx.
Supplementary Table 5: **Summary of genes co-occurring with an ampC homolog.** Gene cluster descriptions are based on the descriptions of each gene within the cluster; parentheses indicate how often a given description was observed.

| Gene cluster ID | Gene name/description                                                                 |
|-----------------|--------------------------------------------------------------------------------------|
| group_10590     | hypothetical protein (13)                                                            |
| group_1110      | spermidine/putrescine ABC transporter substrate-binding protein PotF (4); putrescine/spermidine ABC transporter substrate-binding protein (3); spermidine/putrescine ABC transporter substrate-binding protein (2); extracellular solute-binding protein (1); polyamine ABC transporter substrate-binding protein (1); putrescine-binding periplasmic protein (1) |
| group_13151     | 2′-5′ RNA ligase (17)                                                                |
| group_13686     | flagellar hook-length control protein (12); flagellar hook-length control protein FliK (2) |
| group_13726     | hypothetical protein (16)                                                            |
| group_13727     | twin-arginine translocation pathway signal protein (16); isoquinoline 1-oxidoreductase subunit beta (1) |
| group_13748     | GNAT family acetyltransferase (10); GNAT family N-acetyltransferase (2); GCN5-like N-acetyltransferase (1); N-acetyltransferase GCN5 (1) |
| group_13811     | hypothetical protein (14)                                                            |
| group_13818     | phosphodiesterase (12); phosphodiesterase/alkaline phosphatase D-like protein (2)   |
| group_13819     | hypothetical protein (16)                                                            |
| group_13820     | hybrid sensor histidine kinase/response regulator (8); histidine kinase (3); multisensor hybrid histidine kinase (2); hypothetical protein (1) |
| group_14115     | formate dehydrogenase accessory protein FdhE (13); formate dehydrogenase accessory protein (1) |
| group_14332     | glycosyltransferase (9); glycosyl transferase (3); hypothetical protein (1)           |
| group_14389     | aldo/keto reductase (14)                                                              |
| group_14455     | peptide transporter (12); amino acid adenylation:thioester reductase (1)              |
| group_14473     | chorismate mutase (13)                                                                |
| group_15113     | hypothetical protein (13)                                                             |
| group_15142     | hypothetical protein (12)                                                             |
| group | Group Size | Description |
|-------|------------|-------------|
| group_15152 | 15 | hypothetical protein (10); thiolase (2) |
| group_15235 | 15 | DUF3077 domain-containing protein (9); hypothetical protein (5) |
| group_15241 | 15 | serine protease (11); hypothetical protein (2) |
| group_15274 | 15 | hypothetical protein (12) |
| group_15276 | 15 | transcriptional regulator CynR (8); transcriptional regulator (4); DNA-binding transcriptional regulator CynR (1) |
| group_15277 | 15 | hypothetical protein (14) |
| group_15280 | 15 | ATPase (6); hypothetical protein (6) |
| group_15281 | 15 | GCN5 family acetyltransferase (9); GNAT family N-acetyltransferase (2); GCN5-like N-acetyltransferase (1); N-acetyltransferase GCN5 (1) |
| group_15282 | 15 | GDP-mannose pyrophosphatase (8); GDP-mannose pyrophosphatase nudK (2); nucleoside diphosphate pyrophosphatase (2); ADP-ribose pyrophosphatase (1) |
| group_15283 | 15 | transcriptional regulator (11); Cro/CI family transcriptional regulator (1) |
| group_15284 | 15 | hypothetical protein (10); delta-60 repeat domain-containing protein (2) |
| group_15285 | 15 | hypothetical protein (11); lipoprotein (1) |
| group_15286 | 15 | isochorismatase (10); cysteine hydrolase (1); isochorismatase hydrolase (1) |
| group_15287 | 15 | calpastatin (9); hypothetical protein (3) |
| group_15288 | 15 | polyketide cyclase (10); MxaD family protein (2); hypothetical protein (1) |
| group_15938 | 15 | hypothetical protein (12) |
| group_16075 | 15 | methylase (9); hypothetical protein (2); SAM-dependent methyltransferase (1) |
| group_16076 | 15 | general secretion pathway protein GspH (10); general secretion pathway protein H (1) |
| group_16077 | 15 | hypothetical protein (11) |
| group_16078 | 15 | hypothetical protein (9); membrane protein (3) |
| group_16082 | 15 | hypothetical protein (9); lipoprotein (2) |
| group_16083 | 15 | hypothetical protein (12) |
| group_16084 | 15 | hypothetical protein (11) |
| group_16085 | 15 | hypothetical protein (12) |
| group_16086 | 15 | hypothetical protein (12) |
| group_16087 | 15 | NUDIX hydrolase (11) |
| group_16088 | 15 | glycerol-3-phosphatase (10); HAD family hydrolase (1); phosphatase (1) |
| group     | Members                                                                 |
|-----------|-------------------------------------------------------------------------|
| group_16091 | acetyltransferase (9); GCN5-like N-acetyltransferase (1); GNAT family N-acetyltransferase (1) |
| group_16092 | class C beta-lactamase (10); beta-lactamase (2)                        |
| group_16908 | hypothetical protein (11)                                               |
| group_17033 | hypothetical protein (12)                                               |
| group_17040 | hypothetical protein (10); DUF4160 domain-containing protein (1)        |
| group_17041 | hypothetical protein (11); lipoprotein (1)                             |
| group_17048 | glutamate-5-semialdehyde dehydrogenase (12)                            |
| group_17049 | cupin (6); hypothetical protein (5)                                     |
| group_17050 | glutathione S-transferase (10); glutathione S-transferase-like protein (2) |
| group_18423 | hypothetical protein (6); endoribonuclease L-PSP (4); endoribonuclease L-PSP family protein (1) |
| group_28806 | tRNA-Arg (3); tRNA-Arg;Dbxref=GeneID:14051617 (1); tRNA-Arg;Dbxref=GeneID:3712881 (1) |
| group_4768  | beta-ketoacyl-[acyl-carrier-protein] synthase II (5); beta-ketoacyl-ACP synthase II (4); 3-oxoacyl-ACP synthase (2); 3-oxoacyl-(acyl carrier protein) synthase II (1) |
| group_6199  | hypothetical protein (11)                                               |