

TECHNICAL NOTE

Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria

Harry A. Thorpe, Sion C. Bayliss, Samuel K. Sheppard and Edward J. Feil*

The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY

*Correspondence address. Edward J. Feil, The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY. Tel: +44 1225 383021; E-mail: e.feil@bath.ac.ukRRID: (Piggy, [RRID:SCR.015941](https://www.ebi.ac.uk/rrid/SCR.015941))

Abstract

Background: The concept of the “pan-genome,” which refers to the total complement of genes within a given sample or species, is well established in bacterial genomics. Rapid and scalable pipelines are available for managing and interpreting pan-genomes from large batches of annotated assemblies. However, despite overwhelming evidence that variation in intergenic regions in bacteria can directly influence phenotypes, most current approaches for analyzing pan-genomes focus exclusively on protein-coding sequences. **Findings:** To address this we present Piggy, a novel pipeline that emulates Roary except that it is based only on intergenic regions. A key utility provided by Piggy is the detection of highly divergent (“switched”) intergenic regions (IGRs) upstream of genes. We demonstrate the use of Piggy on large datasets of clinically important lineages of *Staphylococcus aureus* and *Escherichia coli*. **Conclusions:** For *S. aureus*, we show that highly divergent (switched) IGRs are associated with differences in gene expression and we establish a multilocus reference database of IGR alleles (igMLST; implemented in BIGSdb).

Keywords: pan-genome; accessory genome; genomics; whole-genome sequencing (WGS); bacteria; intergenic regions; igMLST; gene expression; *Staphylococcus aureus*, *Escherichia coli*

Findings

Introduction

Whole-genome sequencing has revealed that in many bacteria, individual strains frequently recruit new genes from a seemingly endless genetic reservoir [1, 2]. The total complement of genes observed across all strains, known as the pan-genome, often numbers tens of thousands, up to an order of magnitude more than the number of genes present in any single genome. In contrast, the “core-genome,” which refers to the complement of genes present in all (or the vast majority) of sampled isolates, can be significantly smaller than the total number of genes in any given genome [3, 4]. For example, a study of 328 *Klebsiella pneumoniae* isolates, each of which harbor 4000–5000 genes, revealed a pan-genome of 29 886 genes, only 1888 (6.8%) of which were universally present (core) [5]. Similarly, genome data for 228 *Escherichia coli* ST131 isolates revealed a pan-genome of 11 401 genes, of which 2722 (23.9%) were core [6]. The degree of gene content variation in the latter study is particularly striking as these isolates were all from the same sequence type (ST) and thus show limited nucleotide divergence in core genes and are descended from a recent common ancestor. More generally, the relationship between the size of the core and accessory genomes varies between species, with ecologically diverse species having large accessory genomes and ecologically restricted species (such as endosymbionts) having small accessory genomes [1, 2].

Received: 24 September 2017; Revised: 9 January 2018; Accepted: 16 February 2018

© The Author(s) 2018. Published by Oxford University Press. 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.

There is growing recognition that the acquisition of new genes through horizontal gene transfer (HGT) has a central role in ecological adaptation [7]. The emergence and spread of antibiotic resistance, underpinned by the transfer of plasmids and other mobile genetic elements (MGEs), is a pertinent example. The increasing availability of datasets that contain thousands of isolates thus offers an unprecedented opportunity for describing the genetic basis of bacterial adaptation, although the scale of these data presents serious logistic and conceptual challenges in terms of data management and analysis.

Pioneering pan-genome analysis tools such as PanOCT and PGAP relied on all-vs-all Basic Local Alignment Search Tool (BLAST) comparisons between protein sequences and scaled approximately quadratically with the number of isolates [8, 9]. The large-scale blast score ratio (LS-BSR) introduced a preclustering step that substantially reduced the number of BLAST comparisons, enabling it to be feasibly run on thousands of samples [10]. More recently, the Roary pipeline has rapidly gained popularity for scalable, user-friendly, pan-genome characterization [4].

The concept of the pan-genome, as described above, places an exclusive emphasis on genes or, more specifically, open reading frames with the potential to encode proteins. This gene-centric perspective has both shaped and been shaped by the bioinformatics tools developed to interrogate the pan-genome. For example, Roary works by taking individual protein-coding sequences, predefined using Prokka annotation [11], and assigning each to a single cluster of homologous sequences. This approach thus excludes non protein-coding intergenic regions (IGRs) that typically account for approximately 15% of the genome [12, 13]. This is clearly problematic for downstream attempts to identify genotype–phenotype links, as IGRs contain many important regulatory elements including, but not limited to, promoters, terminators, non-coding RNAs, and regulatory binding sites. Moreover, we have recently shown that IGRs are subject to purifying selection in the core-genomes of diverse bacterial species, even when known major regulatory elements are excluded [14, 15], and a recent study has shown that intergenic variation is positively selected during *Pseudomonas aeruginosa* infections [16].

Given that variation in IGRs can have profound phenotypic consequences, it is timely to consider how best to incorporate these sequences into pan-genome analyses. A key question is the degree to which protein-coding genes, and their cognate regulatory elements, should be considered a single “unit,” both selectively (in terms of co-adaptation) and in terms of physical linkage on the chromosome. If physical linkage is assumed to be highly robust, such that genes are mostly transferred along with their cognate IGRs, then in principle the definition of a “gene” could be expanded to include the upstream regulatory regions. On the other hand, if there is moderate or weak linkage between genes and IGRs, such that IGRs can occasionally transfer independently, then the purview of the pan-genome could be expanded to include the full complement of IGR alleles in addition to protein-coding sequences.

Consistent with the second model, which allows for independent transfer of IGRs, a landmark study demonstrated that *E. coli* genes can apparently be regulated by alternative IGRs that frequently share no sequence similarity to each other [17]. Moreover, the distribution of these IGRs was incongruent with gene trees, suggesting that recombination can act to replace one IGR with another, resulting in regulatory “switches”; a process referred to as horizontal regulatory transfer (HRT) [17]. It is important to note here that the term “switching” refers only to the replacement of an IGR by a nonhomologous or highly diver-

gent variant sequence. It does not specify that the replacement IGR has a particular origin and could therefore correspond to a transfer from elsewhere in the same genome or from another isolate. It was also noted that conserved flanking genes may facilitate this process by providing localized regions of homology. IGR switches can be accompanied by differential gene expression [17] and may provide a mechanism to offset the fitness costs of harboring plasmids and other MGEs [6], pointing to a central role for this process in adaptation.

Our current understanding of the evolutionary dynamics of IGRs in the context of bacterial pan-genomes leaves many open questions. Specifically, it is unclear how IGRs are distributed among isolates within bacterial populations, how commonly IGRs and their cognate genes are cotransferred, and how the frequency of HRT relates to different functional gene categories. A more complete understanding of bacterial adaptation clearly requires a careful consideration of gene presence and absence alongside gene regulation. Here, we address this by introducing a new pipeline called Piggy that closely emulates and complements the established pan-genome analysis pipeline Roary [4]. Input and output files for Piggy and Roary use the same format and run in a similar time on modest computing resources. Piggy provides a means to rapidly identify IGR switches and, more broadly, the means to examine the role of horizontal transfer in shaping the bacterial regulome. We demonstrate the utility of Piggy using large genome datasets for 2 bacterial species, both of which are of high public health importance—*Staphylococcus aureus* and *Escherichia coli*. Conventional pan-genome analyses are applied to analyze and compare core and accessory IGRs and genes in these lineages. In *S. aureus* we show an association between IGR switching and changes in gene expression and demonstrate proof-of-principle by establishing a multilocus IGR scheme (igMLST) in BIGSdb [18]. Piggy is available at [19] under the GPLv3 license.

Methods

Datasets

The *S. aureus* dataset was assembled from published genome sequences [20] available from the European Nucleotide Archive (ENA; study ERP001012). The *S. aureus* RNA-sequencing (seq) data were previously published [21] and are available from the ENA (study ERP009279). This was supplemented with the corresponding reference genomes, HO.5096.0412: HE681097, MRSA252: BX571856, Newman: AP009351, S0385: AM990992, available from the National Center for Biotechnology Information (NCBI). The *E. coli* ST131 dataset was from a previously published study [6] and is available at [22]. All complete genomes and assemblies were annotated with Prokka [11].

Roary and Piggy parameter settings

Roary [4] was run using default parameters except for the following: `-e -n` (to produce alignments with MAFFT [23]); `-i 90` (lower amino acid identity than the default); `-s` (to keep paralogs together); and `-z` (to keep intermediate files). Piggy was run using default parameters except for `-len_id`, which controls the percentage of IGR sequences that must share similarity in order to be clustered together. For the *S. aureus* and *E. coli* ST131 datasets, Piggy was run twice, once with `-len_id 10` and once with `-len_id 90`. The former was used for the pan-genome comparisons between genes and IGRs in order to be comparable with Roary. Use of a low length identity (`-len_id 10`) enabled

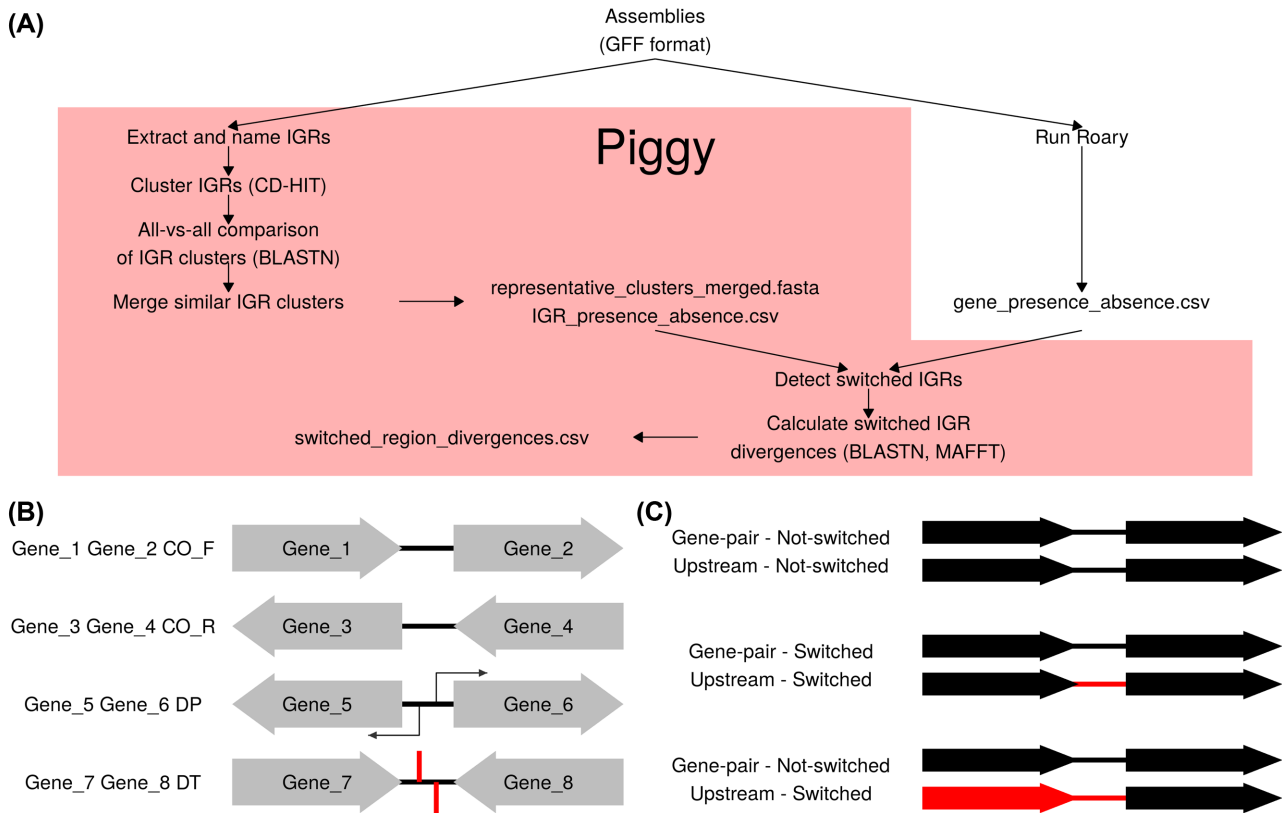


Figure 1: An overview of the Piggy pipeline. (a) A schematic to illustrate the Piggy pipeline and how it works alongside Roary [4]. (b) IGRs are named according to their flanking genes and their orientations (CO.F—Co-Oriented Forward, CO.R—Co-Oriented Reverse, DP—Double Promoter, DT—Double Terminator). This naming scheme enables Piggy to link genes with their associated IGRs and provides information on their orientations. (c) A schematic to illustrate the difference between the “gene-pair” and “upstream” methods used to identify candidate switched IGRs. For the gene-pair method, only the IGR between the 2 genes is nonhomologous (“switched”), and for the upstream method both the upstream IGR and gene may be nonhomologous to the downstream gene.

homologous sequences of varying lengths (e.g., a truncated sequence) to cluster together. Roary does not provide a similar setting and only requires that sequences have a minimum length of 120 base pairs (bp). Genes in the same clusters defined by Roary may vary considerably in length, either due to genuine truncations or assembly errors. A relaxed `-len.id` setting of 10 was therefore used in Piggy to provide consistency with Roary and to ensure that homologous IGRs are not erroneously placed in different clusters. A `-len.id` setting of 90 was subsequently used whenever “switched” IGRs were detected, as this enabled sequences to be subsequently filtered by either nucleotide or length identity.

RNA-seq analysis

Two biological replicates for each isolate were analyzed. Kallisto [24] was used to quantify transcripts (`-kmer-size 31` and `-bootstrap-samples 100`), and Sleuth [25] was used to normalize and filter the counts produced by Kallisto. These counts were then \log_{10} transformed, and major axis regression was performed. Rockhopper2 [26] was used to produce an operon map for each strain by grouping adjacent genes with similar expression profiles together into operons.

Clustering performance

We examined the clustering performance of Piggy by producing truncated variants of IGRs of lengths 10,15,20,30,50 bp and com-

paring how the lengths of the IGRs altered the resulting clustering. The IGRs were truncated from a random starting point in the sequence, and each length was analyzed separately. From the starting pool of IGRs from 10 randomly selected isolates, 1000 IGRs were chosen and truncated. These truncated variants were then added to the pool of IGRs, and Piggy was run on them. Clustering patterns based on the truncated and original IGRs were then compared, with truncated IGRs placed in the same cluster as their progenitor sequences being assigned as correctly clustered. This analysis was performed on both the *S. aureus* ST22 and *E. coli* ST131 datasets.

Statistical analyses

All statistical analyses were performed within R version 3.3.2 [27]. All plotting was performed with ggplot2 [28].

Results

Overview of the Piggy pipeline

Figure 1a shows an overview of the Piggy pipeline. The first step is to run Roary, as the gene presence–absence output file from Roary is used as an input for Piggy. Piggy is then run using the same annotated assemblies as Roary, specifically GFF3 format files such as those produced by Prokka [11]. Piggy extracts IGRs from these files and uses the flanking gene names and their orientations to name the IGRs (Fig. 1b). Each IGR name contains 3 pieces of information: the upstream gene, the downstream

gene, and their relative orientations (CO—Co-Oriented, DP—Double Promoter, DT—Double Terminator). For example, the IGR “Gene.1 Gene.2 DP” is flanked by Gene.1 and Gene.2, which are both downstream of the IGR (i.e., they are transcribed in opposite directions). IGRs at the edge of contigs are excluded by default; however, when they are included (using the `-edges` flag), the missing information is denoted by NA, e.g., “Gene.1 NA NA.” Including the gene neighborhood information gives context to the IGR and enables identification of switched IGRs. By default, only IGRs between 30–1000 bp in length are included by Piggy, though these lengths can be user defined using the `-size` flag (minimum length = 30 bp). The IGRs are then clustered with CD-HIT [29] at user-defined identity thresholds (`-nuc.id`—nucleotide identity, `-len.id`—length identity). The nucleotide identity is defined as single nucleotide polymorphisms (SNPs)/aligned sites, and the length identity is defined as shared sites/alignment length. These 2 flags allow the user to set the level of stringency for clustering. For example, a conservative approach is to set high values for both nucleotide and length identity such that IGRs must be similar in both nucleotide and length identity to cluster together. By relaxing the length identity while maintaining a high nucleotide identity threshold, highly related sequences still cluster even if one is truncated. The longest sequence from each cluster is then used to perform an all-vs-all BLASTN search [30]. This is used to merge similar clusters (BLASTN defaults, except `-word.size = 10`), which did not cluster with CD-HIT. These clusters are then used to produce an IGR presence-absence matrix (“IGR.presence.absence.csv”) in the same format as the gene presence-absence matrix (“gene.presence.absence.csv”) produced by Roary. Up until this point, the pipeline is very similar to Roary [4].

Switched IGR detection

Piggy identifies switched IGRs using 2 methods. For both methods, the term “switch” refers to 2 or more divergent IGR sequences that occupy the same locus as defined by flanking genes but does not specify an origin for the divergent IGR sequences [17]. The first method identifies adjacent genes on the same contig (gene-pairs) and searches for IGR clusters that lie between these gene-pairs (Fig. 1c). Instances where multiple IGR clusters correspond to the same gene-pair are identified as candidate switched IGRs. The second method identifies instances where multiple IGR clusters occupy a locus upstream of a single gene cluster. This is a less conservative approach as only 1 of the 2 genes flanking the IGR is taken into account (Fig. 1c). The gene-pair method is used by default as it controls against detecting switching (recombination) events that encompass more than a single IGR, e.g., cases where a mobile element has inserted between 2 genes. However, such cases remain relevant as the regulation of the downstream gene may still be affected.

To ensure that differences in gene annotation between isolates, specifically artifactual variation in the start and end points of each gene, are not erroneously assigned as switching events, the first and last 30 bp of each flanking gene are searched against the IGRs with BLASTN. Any matches from these searches indicate differences in annotation of gene borders (rather than genuine differences between the IGRs), and these sequences are disregarded. In order to confirm that they represent genuine switching events, candidate switched IGRs are searched against each other with BLASTN with low-complexity filtering turned off (`-dust no`). If there is no significant match, they are classed as “switched”; and if there is a significant match, they are aligned using MAFFT [23]. The resulting alignment is then used to cal-

culate nucleotide identity (SNPs/shared sites) and length identity (number of shared sites/alignment length). These values can then be used to define an appropriate threshold to identify switched IGRs. To aid this, Piggy calculates within-cluster divergences for both genes and IGRs, and these divergences can be used to calibrate Piggy with Roary.

Clustering performance

The shorter lengths of IGRs compared with genes poses potential problems for alignment accuracy. We tested the clustering performance of Piggy by producing truncated variants of IGRs, adding these to the total complement of IGRs in an analysis and then recording whether the truncated IGRs were clustered with their untruncated counterparts (see the Methods section). For *S. aureus* ST22, 82% of IGRs truncated to 10 bp clustered together with the corresponding full-length sequences, but this figure increased to >99% when the length of the truncated sequences was 20 bp (Supplementary Fig. S1a). A similar increase was observed for the *E. coli* ST131 data, although in this case, 50 bp was required for the percentage of correct assignments to be >99% (Supplementary Fig. S1b).

An inspection of the incorrectly clustered sequences from both datasets revealed that their progenitor sequences shared high sequence similarity in parts of their sequence to other IGR clusters but no sequence similarity in other parts of the sequence. This resulted in separate clusters that shared high sequence homology over parts of their sequences. When these sequences were truncated to assess the clustering, if the homologous part of the sequence was selected, then it could align to either of these progenitor IGR clusters. In many cases, these alignments were perfect matches, and so the IGR could not be unambiguously placed. This problem is likely to be a result of nonhomologous breaks at the edge of HGT events, and this is consistent with greater clustering accuracy in *S. aureus* ST22 compared with *E. coli* ST131, where the latter has a much larger pan-genome.

Staphylococcus aureus

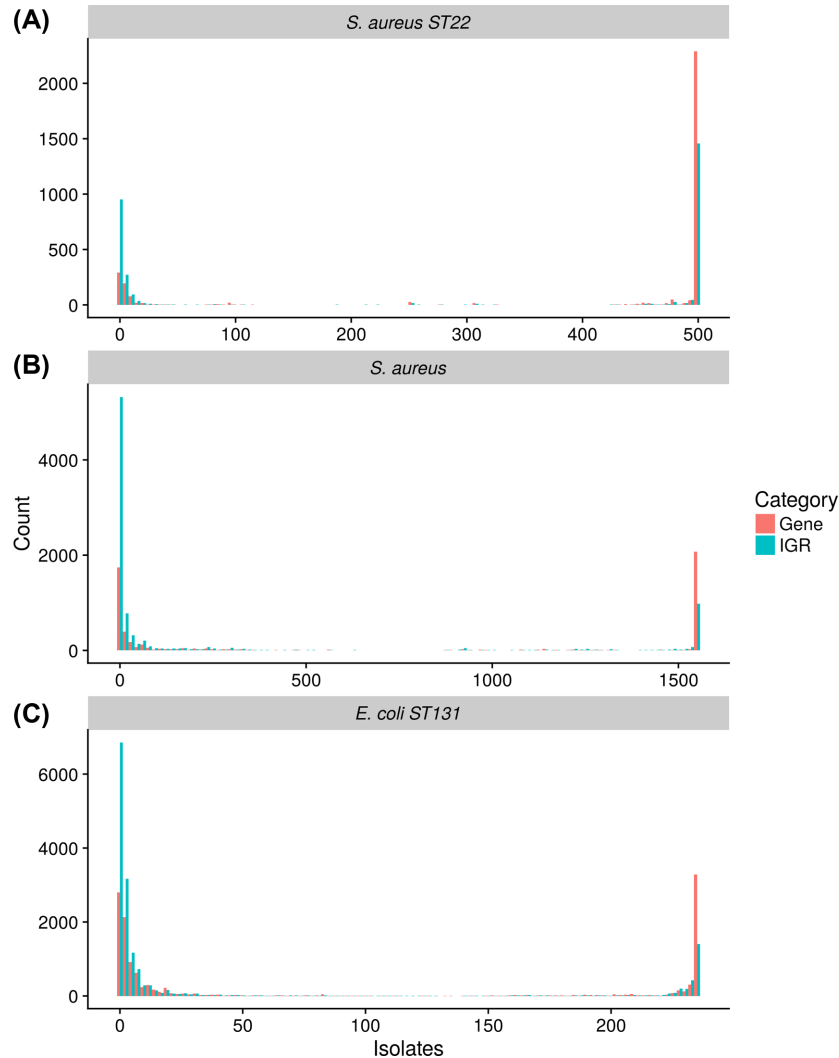
Staphylococcus aureus is an important skin-associated bacterium that is commonly carried asymptotically but can also cause a wide range of infections from minor skin infections to fatal bacteremias. It has a clonal population structure that consists of discrete lineages [31]. Although the core genome is relatively stable, phenotypic variation (e.g., resistance profiles, virulence traits, and host preference) is associated with a more dynamic accessory genome and the horizontal transfer of MGEs such as the *SCCmec* element, which confers resistance to β -lactam antibiotics [32].

Staphylococcus aureus ST22 (EMRSA-15) is a clinically important hospital-acquired methicillin-resistant strain that is common in the United Kingdom and is rapidly expanding elsewhere in Europe and globally [33]. Previous work has shown that *S. aureus* ST22 is clonal and shows relatively little variation in gene content [20, 33]. In order to compare the pan-genomes of *S. aureus* at different scales, we analyzed a diverse dataset of 1552 isolates from many lineages and a smaller dataset of 500 ST22 isolates subsampled from the larger dataset [20]. The size of the gene and IGR pan and core-genomes were compared by analyzing both datasets with Roary and Piggy. Frequency histograms were plotted for both genes and IGRs (Fig. 2a and b).

The gene-IGR frequency histogram for ST22 (Fig. 2a) shows that there are 2409 core genes and 1556 core IGRs, where core

Table 1: The numbers and percentages of core and accessory genes and IGRs in the three datasets examined.

Species	Core genes	Core IGRs	Accessory genes	Accessory IGRs	Percentage core genes per genome	Percentage core IGRs per genome
<i>S. aureus</i> ST22	2409	1556	816	1543	95	95
<i>S. aureus</i>	2129	1134	3446	8033	85	69
<i>E. coli</i> ST131	3930	2296	8876	14 133	84	77

**Figure 2:** Properties of the pan-genomes. Genes (red) and IGRs (blue) were analyzed with frequency histograms (the number of genes/IGRs present in any given number of isolates). The vast majority of genes/IGRs are either very rare or very common. (a) *S. aureus* ST22, (b) *S. aureus*, and (c) *E. coli* ST131.

is defined as gene presence in >95% of isolates (Table 1). When the whole species is considered, these numbers drop to 2129 and 1134, respectively. The fact that there are fewer core IGRs than core genes is in part due to the exclusion of IGRs <30 bp (many of which are intra-operonic) but also likely reflects faster evolution of IGRs. Both distributions conform to the U shape typically found in such analyses where the majority of genes and IGRs are either very common or very rare. However, the distribution of genes and IGRs is shifted toward the rare sequences when the whole species is considered rather than only ST22.

We used the output of Piggy to investigate the degree of linkage between genes and IGRs. We identified all genomic loci that

consists of an IGR flanked by 2 genes; from these we identified all pairs of genes and IGRs where the IGR was upstream of the gene start. We then grouped these according to whether the gene or IGR was core or accessory (Table 2). For the *S. aureus* ST22 data, 99.5% of core genes were immediately downstream of a core IGR, and 92.9% of the accessory genes were similarly downstream of an accessory IGR. When considering the wider *S. aureus* dataset, the figures were similar: 92.6% of core genes were downstream of a core IGR and 96.8% of accessory genes were downstream of an accessory IGR. Thus, the assignment of an IGR as core or accessory is strongly predictive of the corresponding assignment of the cognate downstream gene, which in turn points

Table 2: Linkage between genes and IGRs. The figures are percentages of core/accessory genes that are immediately downstream of core/accessory IGRs.

Species	Core gene: core IGR	Core gene: accessory IGR	Accessory gene: core IGR	Accessory gene: accessory IGR
<i>S. aureus</i> ST22	99.5	0.5	7.4	92.6
<i>S. aureus</i>	92.9	7.1	3.2	96.8
<i>E. coli</i> ST131	97.9	2.1	2.7	97.3

to strong background linkage between genes in IGRs in the genome.

Escherichia coli ST131

The utility of Piggy was further validated by re-analyzing data from a recent study on the widespread and clinically important *E. coli* lineage ST131 [6]. This dataset contains 228 clinical *E. coli* ST131 isolates from human, domesticated animal, and avian hosts. *Escherichia coli* is a more genetically diverse species than *S. aureus*, and unsurprisingly, *E. coli* ST131 has a larger pan-genome than *S. aureus* ST22, with 12 806 genes and 16 429 IGRs (Fig. 2c, Table 1). More surprisingly, *E. coli* ST131 has a larger pan-genome than the whole *S. aureus* species. Within *E. coli* ST131, 3930 genes and 2296 IGRs were core out of an average of 4689 genes and 2984 IGRs per isolate. Thus despite the differences between the 2 species in their level of diversity, there was a consistent signal of a lower number of core IGRs than core genes and a high number of accessory IGRs compared to accessory genes. There was tight linkage between genes and IGRs, with 97.9% of core genes being immediately downstream of core IGRs and 97.3% of accessory genes being similarly downstream of accessory IGRs; these results are consistent with those from *S. aureus* (Table 2).

The data from *S. aureus* and *E. coli* show a background of strong linkage between core genes and IGRs. However, this linkage is not perfect; some core genes are associated with accessory IGRs (and vice versa), and the linkage is weaker over long timescales (across the whole *S. aureus* species compared to within ST22). Previous work has examined this linkage and found evidence of widespread IGR regulatory switching, where genes are regulated by alternative IGRs in different isolates [17]. Piggy provides a list of candidate switching events together for both “gene-pair” and “upstream” approaches (see the Methods section) at different thresholds of nucleotide identity. For the *E. coli* ST131 data, the pipeline detected 61 cases of putative IGR switching using the most conservative settings (i.e., the conservative gene-pair method and the alternative IGRs, showing no sequence similarity by BLASTN). Relaxation of the threshold of sequence identity to <90% resulted in the identification of an additional 317 candidate switching events, though these possibly reflect either relaxed or positive selection.

Switched IGRs influence gene expression in *S. aureus*

To determine whether switches in IGRs affect the expression of cognate (downstream) genes, we used a previously published RNA-seq dataset based on 4 reference *S. aureus* isolates HO_5096.0412 (ST22), Newman (CC8), MRSA252 (CC36), and S0385 (CC398) [21]. Each of these *S. aureus* reference isolates rep-

resents a distinct major clonal complex, and all were grown under identical conditions with each experiment being replicated. Thus these data provide evidence of the natural variation in gene expression within the *S. aureus* population. By analyzing these data alongside the output from Piggy, it is possible to test the extent to which IGR switches between these 4 genomes can account for the observed variation in gene expression between clonal complexes. First, Roary was used to identify a set of 2094 single copy core genes present in all 4 isolates; expression of these core genes was then quantified using Kallisto [24]. To do this, we used RNA-seq data for 2 replicates for each of the 4 reference genomes. The transcripts per kilobase million values for each gene are given in Supplementary Table S1. We then used Sleuth [25] to normalize and filter these counts.

To check the consistency of the data between biological replicates, we first plotted 2 replicates for each isolate against each other (e.g., Newman replicate 1 vs Newman replicate 2) (Fig. 3). These plots were tightly correlated (mean $R^2 = 0.98$), confirming that the expression values for individual genes were consistent between replicates. We then plotted between-isolate comparisons, again using both replicates for each genome (e.g., Newman replicate 1 vs MRSA252 replicate 1, and Newman replicate 2 vs MRSA252 replicate 2) (Fig. 3). These comparisons revealed considerably more scatter, with R^2 values ranging from 0.76 to 0.9. Given the extremely high R^2 values for within-isolate comparisons, the decrease in R^2 for between-isolate comparisons reflects genuine differences in expression between the isolates. We note that a small number of genes show very striking differences in expression between the clonal complexes. For example, the expression of *mepA*, which encodes a multidrug efflux pump, was approximately 250 fold higher in Newman compared with the other isolates.

The genomes of each pair of isolates were analyzed using Roary and Piggy to identify switched IGRs with a nucleotide identity threshold of <90% for IGR clusters. For each pair of isolates, we then identified all genes immediately downstream of a switched IGR. As a single switched IGR might impact on the expression of more than 1 co-transcribed downstream gene, we also considered all genes linked in a single operon that could be impacted by a single switching event upstream affecting a shared promoter. For each pair of isolates, we thus identified all core genes putatively affected by upstream IGR switches. We then tested whether these genes showed a higher degree of differential expression by conducting Monte Carlo permutation tests on the residuals from the regressions (Fig. 3). For each pairwise comparison of isolates, we summed the residuals of the genes with switched IGRs (shown as red points in Fig. 3) and compared this to a distribution obtained by resampling (without replacement) 100 000 random sets of the same number of genes and summing their residuals. We computed a one-tailed *P* value by dividing the number of permutations with summed residuals greater than the observed value by 100 000 (Fig. 3). Because we used both replicates separately (e.g., Newman replicate 1 vs S0385 replicate 1, and Newman replicate 2 vs S0385 replicate 2), each comparison between pairs of isolates was tested twice. In 9/12 pairwise comparisons, the observed residuals of the genes downstream of switched IGRs were significantly ($P < 0.05$) greater than expected from the resampled data, indicating that genes with switched IGRs were more differentially expressed than those without. Of the 3 remaining comparisons, 2 corresponded to comparisons between HO_5096.0412 and S0385 ($P = 0.17$ and $P = 0.055$) and 1 between HO_5096.0412 and Newman ($P = 0.054$). The second comparison between HO_5096.0412 and Newman was the most weakly significant result ($P = 0.025$).

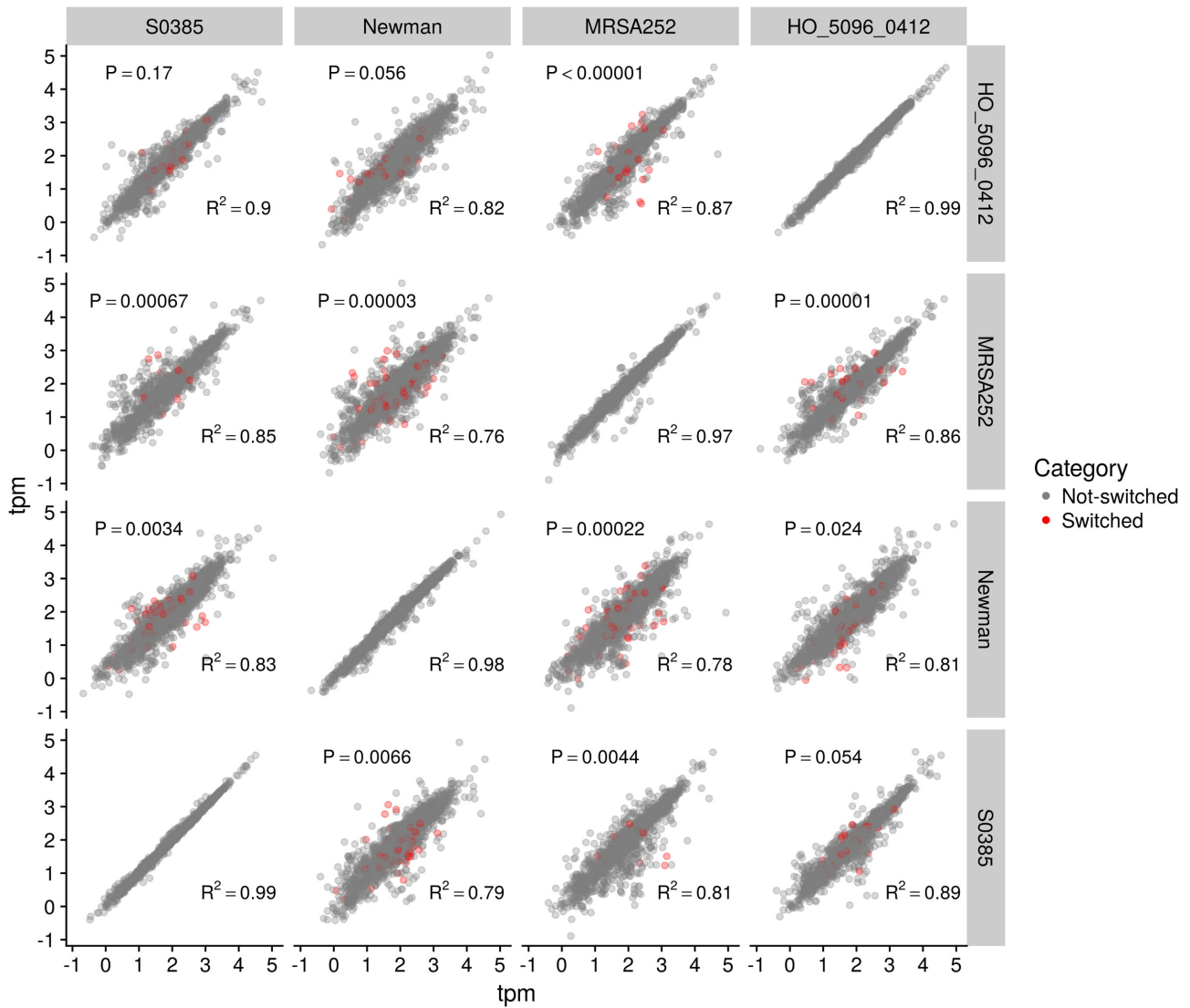


Figure 3: *Staphylococcus aureus* gene expression data. Pairwise RNA-seq comparisons between 4 *S. aureus* isolates, where 2 biological replicates were used for each isolate. The top-left of the diagonal corresponds to comparisons between replicate 1 from different isolates (e.g., SO385 replicate 1 vs HO.5096.0412 replicate 1). The bottom-right of the diagonal corresponds to comparisons between replicate 2 from different isolates (e.g., SO385 replicate 2 vs HO.5096.0412 replicate 2). The diagonal corresponds to comparisons between the 2 biological replicates from the same isolate. The 2094 core genes were analyzed in each comparison, and transcripts per kilobase million (tpm) was used to quantify expression. The genes were separated into 2 categories: switched (red) and not switched (gray) based on their upstream IGRs. The R^2 value corresponds to all the genes. The P value corresponds to a Monte Carlo permutation test comparing the residuals of the 2 groups of genes, where a significant score indicates that the genes downstream of switch IGRs are associated with a higher degree of differential expression (i.e., higher residuals).

Thus, the 2 replicates for each individual pairwise comparison were largely concordant with each other.

Our analysis confirms that genes downstream of switched IGRs are on average more likely to be differentially expressed than genes not associated with IGR switches, as identified using Piggy. To illustrate the genomic context and expression differences of genes with switched IGRs, we selected 3 of the most differentially expressed genes with IGR switches for the Newman vs MRSA252 comparison and plotted nucleotide identity across the IGR (calculated as a 20-bp sliding window) alongside gene expression (Fig. 4).

Compatibility and scalability

We have so far demonstrated that Piggy can be used to analyze the intergenic component of the pan-genome and identify IGR

switches, and we have shown that these switches have biological relevance with respect to gene expression. Importantly, Piggy is designed such that the output files are compatible with existing software and databases. The “IGR_presence_absence.csv” file has an identical format to the “gene_presence_absence.csv” file produced by Roary and can be loaded directly into the interactive browser-based viewer Phandango [34] (Supplementary Fig. S2). It can also be used as input, along with a traits file, to Scoary [35] to test for associations between IGRs and phenotypic traits. Moreover, the “representative_clusters_merged.fasta” file can be loaded directly into BIGSdb [18] to create an allele scheme for IGRs. In order to provide proof of principle, we created a multilocus IGR (igMLST) scheme in BIGSdb. Briefly, 2631 unique IGR sequences with lengths ≥ 30 bp from 7 *S. aureus* reference genomes were entered into the database locus list. Using functionality within the database, these sequences were grouped as a

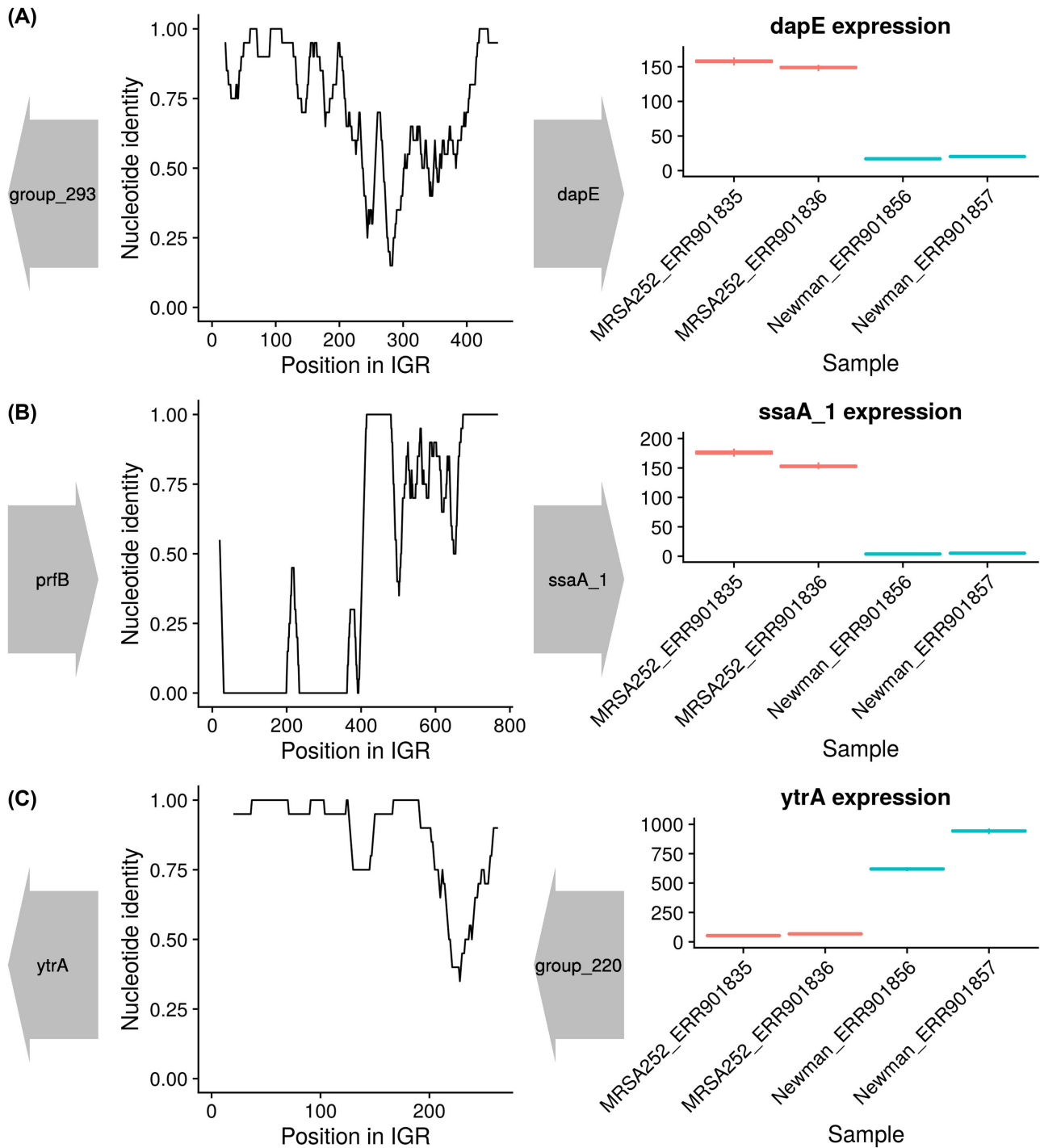


Figure 4: A detailed view of the genomic neighborhood and expression data for selected genes in Newman vs MRSA252. Nucleotide identity was calculated using a 20-bp sliding window across the IGR, and this is shown alongside the flanking genes in their correct orientation (left). The corresponding expression data for the gene of interest was also shown (right), with the 2 box plots per isolate corresponding to the 2 biological replicates. (a) *dapE*, (b) *ssaA.1*, and (c) *ytrA*.

searchable scheme (S.aureus Intergenic.PIGGY), comparable to MLST, rMLST, and wgMLST schemes [36–38]. The distribution of IGRs was analyzed for all isolates in the database, identifying IGRs as present in the respective genome if a hit was recorded with nucleotide identity $\geq 70\%$ over $\geq 50\%$ of the sequence using a BLAST word size of 7 bp. The scheme can be found at [39]. Although we do not expect a typing scheme based solely on IGRs to be widely used, supplementing protein-coding regions with

IGR alleles may provide additional information regarding links between genotype and phenotype, as well as increased epidemiological and phylogenetic resolution.

Discussion

Whole-genome sequence datasets that consist of hundreds or even thousands of bacterial isolates have revealed pan-genomes

of many thousands of genes and large differences in gene content between isolates of the same species. Currently, pan-genome diversity is considered almost exclusively in terms of protein-coding genes, despite overwhelming evidence that variation within IGRs impacts on phenotypes. Here, we address this by introducing Piggy, a pipeline specifically designed to incorporate IGRs into routine pan-genome analyses by working in close conjunction with Roary [4].

The utility of this approach is demonstrated using large datasets of *S. aureus* and *E. coli* ST131. Consistent with previous analyses of protein-coding regions [6, 33], the IGR component of the ST131 pan-genome is considerably larger than that for ST22 and, surprisingly, is also larger than the pan-genome of the whole *S. aureus* species. There was more diversity within IGRs than genes in both species. While some IGRs may be essential for expression of multiple genes, IGRs are broadly subject to weaker purifying selection than protein-coding genes [14]. The maintenance of core IGRs in both bacterial genome datasets is consistent with selection acting to conserve them and allows alignment and analysis in much the same way as protein-coding regions.

The current exclusion of IGRs from routine pan-genome or wgMLST analyses may in part reflect perceived difficulties in the alignment and subsequent cluster definition, particularly if the sequences are very short. We therefore validated the pipeline by investigating clustering accuracy as a function of sequence length by truncating the IGR sequences and recording whether they remained in the same cluster as their full-length counterparts. For *S. aureus*, the data showed that truncated IGRs >20 bp almost always remained in the original cluster, confirming that the minimum length permitted in the pipeline of 30 bp is conservative. For *E. coli*, truncating the sequences had greater impact on cluster assignments, and a minimum length of 50 bp would be a safer setting in this case. The problems with clustering shorter sequences in *E. coli*, compared to *S. aureus*, are not due to the length of the sequence per se but reflect the higher rate of horizontal gene transfer in this species. This means that the IGRs are more likely to be chimeric in structure, with localized regions within the IGRs showing a high level of homology to different clusters. This led to cluster assignment being dependent not so much on length but on which part of the truncated sequence happened to be retained.

Variation within regulatory elements located within IGRs can impact on the expression of the downstream gene [17]. Piggy (alongside Roary) provides the means to combine information on genes and their cognate IGRs, thus facilitating the detection of “switched” IGRs and downstream genes that are potentially affected. We have shown that in *S. aureus* genes with switched upstream IGRs show a higher degree of differential expression than those without. This is consistent with previous work on *E. coli* [17] and suggests that the identification of IGR switches using Piggy can provide a useful indication of differential gene expression, even in the absence of RNA-seq data. However, we note that high divergence within IGRs does not necessarily imply selection for differential gene expression and may instead simply reflect weaker selective constraints.

A clear direction for future work is to make constructs that consist of genes with alternative IGRs in order to directly measure the effect of natural IGR variants on gene expression. Similar experiments have previously been performed in *E. coli* based on variation within promoters [40] and IGRs more broadly [17]. The importance of changes in gene expression mediated by intergenic variation as a route of adaptation is currently unknown. However, a recent study suggested that intergenic changes are

strongly positively selected in *P. aeruginosa* during infection in patients with cystic fibrosis, and more work is required to test the generality of these findings [16].

Conclusions

Driven by recent technical advances in high-throughput sequencing, large whole-genome datasets have provided powerful evidence concerning the genetic determinants that underlie complex multifactorial phenotypes such as virulence. Moreover, associating variation in core and accessory genes with phenotype data is providing new fundamental insight into the ecology and evolution of bacteria. However, in much the same way that non protein-coding DNA in the human genome was initially dismissed as “junk,” omitting IGRs from bacterial genome analysis severely limits our ability to draw inferences on the regulation of gene expression and associated phenotypic consequences. By developing Piggy as an easy-to-use bioinformatics tool with output files that are compatible with existing software and databases (e.g., Roary, Phandango; Supplementary Fig. S1, Scoary, BIGSdb), we envisage that combined information from genes and their cognate IGRs will vastly improve our understanding of genome evolution in bacteria.

Availability of supporting source code

Project name: Piggy

Project home page: <https://github.com/harry-thorpe/piggy>

Operating system(s): Linux

Programming language: Perl, R

Other requirements: Roary

License: GPLv3

RRID: (Piggy, [RRID:SCR.015941](https://eutils.ncbi.nlm.nih.gov/entrez/eutils/rrid.cgi?db=SCRR))

Availability of supporting data

The *S. aureus* dataset was assembled from published genome sequences [20] available from the ENA (study ERP001012). The *S. aureus* RNA-seq data was previously published [21] and is available from the ENA (study ERP009279). This was supplemented with the corresponding reference genomes, HO_5096.0412: HE681097, MRSA252: BX571856, Newman: AP009351, and S0385: AM990992, available from the NCBI. The *E. coli* ST131 dataset was from a previously published study [6] and is available at [22]. An archival copy of the Piggy source code is available via the GigaScience repository GigaDB [41].

Additional file

Figure S1. Clustering performance. The clustering performance was assessed by truncating IGR sequences and reclustering them with the pool of original sequences. Truncated IGRs which were placed into the same cluster as their progenitor sequences were deemed to be correctly clustered. a) *S. aureus* b) *E. coli*.

Figure S2. The IGR pan-genome as visualized using Phandango. A neighbor-joining phylogenetic tree was imported into Phandango [34] alongside the IGR_presence_absence.csv file. Each row corresponds to an isolate, and each column corresponds to an IGR, with the IGRs ordered from the left in order of decreasing frequency within the sample. The line graph at the bottom shows the frequency of the IGRs within the sample. a) *S. aureus* ST22 b) *E. coli* ST131.

Table_S1.xlsx

Abbreviations

BLAST: Basic Local Alignment Search Tool; bp: base pair; ENA: European Nucleotide Archive; HGT: horizontal gene transfer; HRT: horizontal regulatory transfer; igMLST: multilocus IGR scheme; IGR: intergenic region; MGE: mobile genetic element; NCBI: National Center for Biotechnology Information; RNA-seq: RNA-sequencing; SNP: single nucleotide polymorphism; ST: sequence type

Competing interests

The authors declare that they have no competing interests.

Funding

The *S. aureus* genome sequences were generated as part of a study supported by a grant from the United Kingdom Clinical Research Collaboration Translational Infection Research Initiative and the Medical Research Council (grant G1000803, held by Sharon Peacock) with contributions from the Biotechnology and Biological Sciences Research Council, the National Institute for Health Research on behalf of the Department of Health, and the Chief Scientist Office of the Scottish Government Health Directorate, on which E.J.F. was a principal investigator and S.C.B. was a postdoctoral researcher. H.A.T. is funded by a University of Bath research studentship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors contributions

H.A.T. designed and implemented the pipeline and carried out the majority of the analyses, with input from E.J.F., S.C.B., and S.K.S. H.A.T. and E.J.F. wrote the manuscript with input from S.K.S. and S.C.B.

Acknowledgements

We are very grateful to Torsten Seemann, Andrew Page, and João Carriço for encouragement and helpful feedback. We are also grateful to Matt Holden for provision of the *S. aureus* RNA-seq data, Sandra Reuter for help with the *S. aureus* data, and Alan McNally for the *E. coli* ST131 data. This work also greatly benefited from access to the Medical Research Council-funded Cloud Infrastructure for Microbial Bioinformatics [42].

References

- McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. *Nat Microbiol* 2017;2:17040.
- Andreani NA, Hesse E, Vos M. Prokaryote genome fluidity is dependent on effective population size. *ISME J* 2017;11:1719–21.
- Medini D, Donati C, Tettelin H et al. The microbial pangenome. *Current Opinion in Genetics & Development* 2005;15:589–94.
- Page AJ, Cummins CA, Hunt M et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015;31:3691–3.
- Holt KE, Wertheim H, Zadoks RN et al. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. *Proc Natl Acad Sci U S A* 2015;112:E3574–81.
- McNally A, Oren Y, Kelly D et al. Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. *PLoS Genet* 2016;12:e1006280.
- Vos M, Hesselman MC, te Beek TA et al. Rates of lateral gene transfer in prokaryotes: high but why? *Trends Microbiol* 2015;23:598–605.
- Fouts DE, Brinkac L, Beck E et al. PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic Acids Res* 2012;40:e172–.
- Zhao Y, Wu J, Yang J et al. PGAP: pan-genomes analysis pipeline. *Bioinformatics* 2012;28:416–8.
- Sahl JW, Caporaso JG, Rasko DA et al. The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. *Peer J* 2014;2:e332.
- Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–9.
- Ochman H, Caro-Quintero A. Genome Size and Structure, Bacterial. In: Kliman RM, (ed.). *Encyclopedia of Evolutionary Biology*. Oxford: Academic Press; 2016, p. 179–85.
- McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. *Nat Rev Micro* 2012;10:13–26.
- Thorpe HA, Bayliss SC, Hurst LD et al. Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species. *Genetics* 2017; Available from: <http://dx.doi.org/10.1534/genetics.116.195784>.
- Molina N, Van Nimwegen E. Universal patterns of purifying selection at noncoding positions in bacteria. *Genome Res* 2008;18:148–60.
- Khademi H, Jelsbak L. Host adaptation mediated by intergenic evolution in a bacterial pathogen. *bioRxiv* 2017 [cited 2017 Dec 20]. p. 236000. Available from: <https://www.biorxiv.org/content/early/2017/12/19/236000.article-info>.
- Oren Y, Smith MB, Johns NI et al. Transfer of noncoding DNA drives regulatory rewiring in bacteria. *Proc Natl Acad Sci U S A* 2014;111:16112–7.
- Jolley KA, Maiden MCJ. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* 2010;11:595.
- <https://github.com/harry-thorpe/piggy>.
- Reuter S, Török EM, Holden MTG et al. Building a genomic framework for prospective MRSA surveillance in the United Kingdom and the Republic of Ireland. *Genome Res* 2015; Available from: <http://genome.cshlp.org/content/early/2015/12/15/gr.196709.115.abstract>.
- Warne B, Harkins CP, Harris SR et al. The Ess/Type VII secretion system of *Staphylococcus aureus* shows unexpected genetic diversity. *BMC Genomics* 2016;17:222.
- McNally et al. 2016 data repository [Internet]. Dryad. [cited 2016 Oct 25]. Available from: <http://datadryad.org/resource/doi:10.5061/dryad.d7d71>.
- Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013;30:772–80.
- Bray NL, Pimentel H, Melsted P et al. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 2016;34:525–7.
- Pimentel H, Bray NL, Puente S et al. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat Meth* 2017;14:687–90.
- Tjaden B. De novo assembly of bacterial transcriptomes from RNA-seq data. *Genome Biol* 2015;16:1.

27. RDevelopment CORE TEAM R, Others. R: a language and environment for statistical computing [Internet]. R Foundation for Statistical Computing, Vienna, Austria; 2008. Available from: <http://www.R-project.org>.
28. Wickham H. Ggplot2: Elegant Graphics for Data Analysis. 2nd ed. Springer Publishing Company, Incorporated; 2009.
29. Fu L, Niu B, Zhu Z et al. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**:3150–2.
30. Camacho C, Coulouris G, Avagyan V et al. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;**10**:421.
31. Feil EJ, Cooper JE, Grundmann H et al. How clonal is *Staphylococcus aureus*? *J Bacteriol* 2003;**185**:3307–16.
32. Lindsay JA, Holden MTG. *Staphylococcus aureus*: superbug, super genome? *Trends Microbiol* 2004;**12**:378–85.
33. Holden MTG, Hsu L-Y, Kurt K et al. A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic. *Genome Res* 2013;**23**:653–64.
34. Hadfield J, Croucher NJ, Goater RJ et al. Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics* 2017; Available from: <http://dx.doi.org/10.1093/bioinformatics/btx610>.
35. Brynildsrud O, Bohlin J, Scheffer L et al. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol* 2016;**17**:238.
36. Maiden MCJ, Jansen van Rensburg MJ, Bray JE et al. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat Rev Micro* 2013;**11**:728–36.
37. Jolley KA, Bliss CM, Bennett JS et al. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology* 2012;**158**:1005–15.
38. Sheppard SK, Jolley KA, Maiden MCJ. A gene-by-gene approach to bacterial population genomics: whole genome MLST of *Campylobacter*. *Genes* 2012;**3**:261–77.
39. Sheppard lab resources [Internet]. [cited 2018 Jan 16]. Available from: <https://sheppardlab.com/resources>.
40. Shimada T, Yamazaki Y, Tanaka K et al. The whole set of constitutive promoters recognized by RNA polymerase RpoD holoenzyme of *Escherichia coli*. *PLoS One* 2014;**9**:e90447.
41. Thorpe HA, Bayliss SC, Sheppard SK et al. Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. *GigaScience database* 2018. <http://dx.doi.org/10.5524/100410>.
42. Connor TR, Loman NJ, Thompson S et al. CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community. *Microbial Genomics* [Internet]. Microbiology Society; 2016 [cited 2016 Nov 1];**2**. Available from: <http://mgen.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000086>.