

Piggy: A Rapid, Large-Scale Pan-Genome Analysis Tool for Intergenic Regions in Bacteria

--Manuscript Draft--

Manuscript Number:	GIGA-D-17-00244R2	
Full Title:	Piggy: A Rapid, Large-Scale Pan-Genome Analysis Tool for Intergenic Regions in Bacteria	
Article Type:	Technical Note	
Funding Information:	Medical Research Council (G1000803)	Not applicable
	United Kingdom Clinical Research Collaboration Translational Infection Research Initiative	Not applicable
Abstract:	<p>Abstract</p> <p>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 analysing pan-genomes focus exclusively on protein-coding sequences.</p> <p>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") IGRs upstream of genes. We demonstrate the use of Piggy on large datasets of clinically important lineages of <i>Staphylococcus aureus</i> and <i>Escherichia coli</i>.</p> <p>Conclusions For <i>S. aureus</i>, we show that highly divergent ("switched") IGRs are associated with differences in gene expression, and we establish a multi-locus reference database of IGR alleles (igMLST; implemented in BIGSdb). Piggy is available at https://github.com/harry-thorpe/piggy.</p>	
Corresponding Author:	Edward Feil UNITED KINGDOM	
Corresponding Author Secondary Information:		
Corresponding Author's Institution:		
Corresponding Author's Secondary Institution:		
First Author:	Harry A. Thorpe	
First Author Secondary Information:		
Order of Authors:	Harry A. Thorpe	
	Sion C. Bayliss	
	Samuel K. Sheppard	
	Edward J. Feil	
Order of Authors Secondary Information:		
Response to Reviewers:	<p>To the editor:</p> <p>Thank you for accepting our manuscript for publication in GigaScience.</p>	

	<p>We do not feel that it is necessary to host the supporting data in GigaDB as all the data is already published and available at public repositories as described in the manuscript.</p> <p>Although in principle Code Ocean looks to be valuable, we do not think it is appropriate for Piggy for a number of reasons. First, Piggy requires various dependencies (particularly Roary), which are not installed on Code Ocean. Second, the free plan only provides 5 GB of storage and 1 compute hour per month - these are not sufficient for any real analysis of whole-genome sequence data. We will however consider it in any future projects.</p> <p>We have reformatted the manuscript as per the guidelines.</p> <p>To the reviewer:</p> <p>We thank the reviewer for their detailed, constructive review, which helped us to improve the manuscript.</p>
Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics	Yes
<p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends.</p> <p>Have you included all the information requested in your manuscript?</p>	
Resources	Yes
<p>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?</p>	
Availability of data and materials	Yes
All datasets and code on which the conclusions of the paper rely must be	

either included in your submission or deposited in [publicly available repositories](#) (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.

Have you have met the above requirement as detailed in our [Minimum Standards Reporting Checklist](#)?

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Piggy: A Rapid, Large-Scale Pan-Genome Analysis Tool for Intergenic Regions in Bacteria.

Harry A. Thorpe¹, Sion C. Bayliss¹, Samuel K. Sheppard¹, Edward J. Feil^{1*}

¹The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY.

*Corresponding author

h.thorpe@bath.ac.uk

s.bayliss@bath.ac.uk

s.k.sheppard@bath.ac.uk

e.feil@bath.ac.uk

Keywords: Pan-genome, Accessory genome, Genomics, Whole-genome sequencing (WGS), Bacteria, Intergenic Regions, igMLST, Gene Expression, *Staphylococcus aureus*, *Escherichia coli*.

RRID: (Piggy, RRID:SCR_015941)

1
2
3
4 **Abstract**

5
6 **Background**

7 The concept of the “pan-genome”, which refers to the total complement of genes within a given
8 sample or species, is well established in bacterial genomics. Rapid and scalable pipelines are
9 available for managing and interpreting pan-genomes from large batches of annotated
10 assemblies. However, despite overwhelming evidence that variation in intergenic regions in
11 bacteria can directly influence phenotypes, most current approaches for analysing pan-
12 genomes focus exclusively on protein-coding sequences.
13
14
15
16

17
18 **Findings**

19 To address this we present Piggy, a novel pipeline that emulates Roary except that it is based
20 only on intergenic regions. A key utility provided by Piggy is the detection of highly divergent
21 (“switched”) IGRs upstream of genes. We demonstrate the use of Piggy on large datasets of
22 clinically important lineages of *Staphylococcus aureus* and *Escherichia coli*.
23
24
25

26 **Conclusions**

27 For *S. aureus*, we show that highly divergent (“switched”) IGRs are associated with differences
28 in gene expression, and we establish a multi-locus reference database of IGR alleles (igMLST;
29 implemented in BIGSdb). Piggy is available at <https://github.com/harry-thorpe/piggy>.
30
31
32
33

34 **Findings**

35
36 **Introduction**

37 Whole-genome sequencing has revealed that, in many bacteria, individual strains frequently
38 recruit new genes from a seemingly endless genetic reservoir [1,2]. The total complement of
39 genes observed across all strains, known as the pan-genome, often numbers tens of
40 thousands, up to an order of magnitude more than the number of genes present in any single
41 genome. In contrast, the “core-genome”, which refers to the complement of genes present in all
42 (or the vast majority) of sampled isolates, can be significantly smaller than the total number of
43 genes in any given genome [3,4]. For example, a study of 328 *Klebsiella pneumoniae* isolates,
44 each of which harbour 4-5,000 genes, revealed a pan-genome of 29,886 genes; only 1,888
45 (6.8%) of which were universally present (core) [5]. Similarly, genome data for 228 *Escherichia*
46 *coli* ST131 isolates revealed a pan-genome of 11,401 genes, of which 2,722 (23.9%) were core
47 [6]. The degree of gene content variation in the latter study is particularly striking as these
48 isolates were all from the same sequence type (ST), thus show limited nucleotide divergence in
49 core genes, and are descended from a recent common ancestor. More generally, the
50 relationship between the size of the core and accessory genomes varies between species, with
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 ecologically diverse species having large accessory genomes, and ecologically restricted
5 species (such as endosymbionts) having small accessory genomes [1,2].
6
7

8
9 There is growing recognition that the acquisition of new genes through horizontal gene transfer
10 (HGT) has a central role in ecological adaptation [7]. The emergence and spread of antibiotic
11 resistance, underpinned by the transfer of plasmids and other mobile genetic elements (MGEs),
12 is a pertinent example. The increasing availability of datasets containing thousands of isolates
13 thus offers an unprecedented opportunity for describing the genetic basis of bacterial
14 adaptation, although the scale of these data presents serious logistic and conceptual challenges
15 in terms of data management and analysis.
16
17
18
19
20
21

22 Pioneering pan-genome analysis tools, such as PanOCT and PGAP relied on all-vs-all BLAST
23 comparisons between protein sequences, and scaled approximately quadratically with the
24 number of isolates [8,9]. LS-BSR introduced a pre-clustering step which substantially reduced
25 the number of BLAST comparisons, enabling it to be feasibly run on thousands of samples [10].
26 More recently, the Roary pipeline has rapidly gained popularity for scalable, user-friendly, pan-
27 genome characterisation [4].
28
29
30
31
32
33

34 The concept of the pan-genome, as described above, places an exclusive emphasis on genes;
35 or, more specifically, open reading frames with the potential to encode proteins. This gene-
36 centric perspective has both shaped, and been shaped by, the bioinformatics tools developed to
37 interrogate the pan-genome. For example, Roary works by taking individual protein-coding
38 sequences, pre-defined using Prokka annotation [11], and assigning each to a single cluster of
39 homologous sequences. This approach thus excludes non protein-coding intergenic regions
40 (IGRs) which typically account for approximately 15% of the genome [12,13]. This is clearly
41 problematic for downstream attempts to identify genotype-phenotype links, as IGRs contain
42 many important regulatory elements including, but not limited to, promoters, terminators, non-
43 coding RNAs, and regulatory binding sites. Moreover, we have recently shown that IGRs are
44 subject to purifying selection in the core-genomes of diverse bacterial species, even when
45 known major regulatory elements are excluded [14,15], and a recent study has shown that
46 intergenic variation is positively selected during *Pseudomonas aeruginosa* infections [16].
47
48
49
50
51
52
53
54
55
56
57

58 Given that variation in IGRs can have profound phenotypic consequences, it is timely to
59 consider how best to incorporate these sequences into pan-genome analyses. A key question is
60
61
62
63
64
65

1
2
3
4 the degree to which protein-coding genes, and their cognate regulatory elements, should be
5 considered a single “unit”, both selectively (in terms of co-adaptation) and in terms of physical
6 linkage on the chromosome. If physical linkage is assumed to be highly robust, such that genes
7 are mostly transferred along with their cognate IGRs, then in principle the definition of a “gene”
8 could be expanded to include the upstream regulatory regions. On the other hand, if there is
9 moderate or weak linkage between genes and IGRs, such that IGRs can occasionally transfer
10 independently, then the purview of the pan-genome could be expanded to include the full
11 complement of IGR alleles in addition to protein-coding sequences.
12
13
14
15
16
17
18

19 Consistent with the second model, which allows for independent transfer of IGRs, a landmark
20 study demonstrated that *E. coli* genes can apparently be regulated by alternative IGRs that
21 frequently share no sequence similarity to each other [17]. Moreover, the distribution of these
22 IGRs was incongruent with gene trees, suggesting that recombination can act to replace one
23 IGR with another resulting in regulatory “switches”; a process they call horizontal regulatory
24 transfer (HRT) [17]. It is important to note here that the term “switching” refers only to the
25 replacement of an IGR by a non-homologous or highly divergent variant sequence. It does not
26 specify that the replacement IGR has a particular origin, and could therefore correspond to a
27 transfer from elsewhere in the same genome, or from another isolate. It was also noted that
28 conserved flanking genes may facilitate this process by providing localised regions of homology.
29 IGR switches can be accompanied by differential gene expression [17], and may provide a
30 mechanism to offset the fitness costs of harbouring plasmids and other MGEs [6], pointing to a
31 central role for this process in adaptation.
32
33
34
35
36
37
38
39
40
41

42 Our current understanding of the evolutionary dynamics of IGRs in the context of bacterial pan-
43 genomes leaves many open questions. Specifically, it is unclear how IGRs are distributed
44 among isolates within bacterial populations, how commonly IGRs and their cognate genes are
45 co-transferred, or how the frequency of HRT relates to different functional gene categories. A
46 more complete understanding of bacterial adaptation clearly requires a careful consideration of
47 gene presence/absence alongside gene regulation. Here we address this by introducing a new
48 pipeline called Piggy which closely emulates and complements the established pan-genome
49 analysis pipeline Roary [4]. Input and output files for Piggy and Roary use the same format, and
50 run in a similar time on modest computing resources. Piggy provides a means to rapidly identify
51 IGR switches, and more broadly the means to examine the role of horizontal transfer in shaping
52 the bacterial regulome. We demonstrate the utility of Piggy using large genome datasets for
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 two bacterial species, both of which are of high public health importance; *Staphylococcus*
5 *aureus* and *Escherichia coli*. Conventional pan-genome analyses are applied to analyse and
6 compare core and accessory IGRs/genes in these lineages. In *S. aureus* we show an
7 association between IGR switching and changes in gene expression, and demonstrate proof-of-
8 principle by establishing a multilocus IGR scheme, (igMLST) in BIGSdb [18]. Piggy is available
9 at (<https://github.com/harry-thorpe/piggy>) under the GPLv3 licence.
10
11
12
13
14

15 **Methods**

16 **Datasets**

17
18 The *S. aureus* dataset was assembled from published genome sequences [19] available from
19 the European Nucleotide Archive (ENA), study number ERP001012. The *S. aureus* RNA-seq
20 data was previously published [20], and is available from the ENA, study number ERP009279.
21 This was supplemented with the corresponding reference genomes, HO_5096_0412:
22 HE681097, MRSA252: BX571856, Newman: AP009351, S0385: AM990992, available from the
23 National Center for Biotechnology Information (NCBI). The *E. coli* ST131 dataset was from a
24 previously published study [6], and is available at [21]. All complete genomes and assemblies
25 were annotated with Prokka [11].
26
27
28
29
30
31
32
33

34 **Roary and Piggy parameter settings**

35
36 Roary [4] was run using default parameters except for the following: -e -n (to produce
37 alignments with MAFFT [22]); -i 90 (lower amino acid identity than the default); -s (to keep
38 paralogs together); -z (to keep intermediate files). Piggy was run using default parameters
39 except for --len_id, which controls the percentage of IGR sequences which must share similarity
40 in order to be clustered together. For the *S. aureus* and *E. coli* ST131 datasets, Piggy was run
41 twice, once with --len_id 10 and once with --len_id 90. The former was used for the pan-genome
42 comparisons between genes and IGRs (Fig 2) in order to be comparable with Roary. Using a
43 low length identity (--len_id 10) enabled homologous sequences of varying lengths (for example
44 a truncated sequence) to cluster together. Roary does not provide a similar setting, and only
45 requires that sequences have a minimum length of 120 bp. Genes in the same clusters defined
46 by Roary may vary considerably in length, either due to genuine truncations or assembly errors.
47 A relaxed --len_id setting of 10 was therefore used in Piggy to provide consistency with Roary
48 and to ensure that homologous IGRs are not erroneously placed in different clusters. A --len_id
49 setting of 90 was subsequently used whenever “switched” IGRs were detected, as this enabled
50 sequences to be subsequently filtered by either nucleotide or length identity.
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

RNA-seq analysis

Two biological replicates for each isolate were analysed. Kallisto [23] was used to quantify transcripts (--kmer-size 31 and --bootstrap-samples 100), and Sleuth [24] was used to normalise and filter the counts produced by Kallisto. These counts were then \log_{10} transformed, and major axis (MA) regression was performed. Rockhopper2 [25] 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 comparing 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 analysed 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 analysis

All statistical analysis was performed within R version 3.3.2 [26]. All plotting was performed with ggplot2 [27].

Results

Overview of the Piggy pipeline

Fig 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 intergenic sequences (IGRs) from these files, and uses the flanking gene names and their orientations to name the IGRs (Fig 1b). Each IGR name contains three 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

1
2
3
4 contigs are excluded by default, but when they are included (using the --edges flag) the missing
5 information is denoted by NA, for example “Gene_1 NA NA”. Including the gene neighbourhood
6 information gives context to the IGR and enables identification of “switched” IGRs. By default,
7 only IGRs between 30-1000 bp in length are included by Piggy, though these lengths can be
8 user-defined using the --size flag (minimum length = 30 bp). The IGRs are then clustered with
9 CD-HIT [28] at user-defined identity thresholds (--nuc_id - nucleotide identity, --len_id - length
10 identity). The nucleotide identity is defined as SNPs / aligned sites, and the length identity is
11 defined as shared sites / alignment length. These two flags allow the user to set the level of
12 stringency for clustering. For example, a conservative approach is to set high values for both
13 nucleotide and length identity such that IGRs must be similar in both nucleotide and length
14 identity to cluster together. By relaxing the length identify whilst maintaining a high nucleotide
15 identity threshold, highly related sequences still cluster even if one is truncated. The longest
16 sequence from each cluster is then used to perform an all-vs-all BLASTN search [29]. This is
17 used to merge similar clusters (BLASTN defaults, except -word_size = 10), which did not cluster
18 with CD-HIT. These clusters are then used to produce an IGR presence absence matrix
19 (“IGR_presence_absence.csv”), in the same format as the gene presence absence matrix
20 (“gene_presence_absence.csv”) produced by Roary. Up until this point, the pipeline is very
21 similar to Roary [4].
22
23
24
25
26
27
28
29
30
31
32
33
34
35

36 **Switched IGR detection**

37 Piggy identifies “switched” IGRs using two methods. For both methods, the term “switch” refers
38 to two or more divergent IGR sequences occupy the same locus as defined by flanking genes,
39 but does not specify an origin for the divergent IGR sequences [17]. The first method identifies
40 adjacent genes on the same contig (gene-pairs), and searches for IGR clusters which lie
41 between these gene-pairs (Fig 1c). Instances where multiple IGR clusters correspond to the
42 same gene-pair are identified as candidate switched IGRs. The second method identifies
43 instances where multiple IGR clusters occupy a locus upstream of a single gene cluster. This is
44 a less conservative approach as only one of the two genes flanking the IGR is taken into
45 account, (Fig 1c). The gene-pair method is used by default as it controls against detecting
46 “switching” (recombination) events that encompass more than a single IGR, for example, cases
47 where a mobile element has inserted between two genes. However such cases remain relevant
48 as the regulation of the downstream gene may still be affected.
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 To ensure that differences in gene annotation between isolates, specifically artefactual variation
5 in the start and end points of each gene, are not erroneously assigned as switching events, the
6 first and last 30 bp of each flanking gene are searched against the IGRs with BLASTN. Any
7 matches from these searches indicate differences in annotation of gene borders (rather than
8 genuine differences between the IGRs), and these sequences are disregarded. In order to
9 confirm that they represent genuine switching events, candidate switched IGRs are searched
10 against each other with BLASTN with low complexity filtering turned off (-dust no). If there is no
11 significant match they are classed as “switched”, and if there is a significant match they are
12 aligned using MAFFT [22]. The resulting alignment is then used to calculate nucleotide identity
13 (SNPs / shared sites), and length identity (number of shared sites / alignment length). These
14 values can then be used to define an appropriate threshold to identify “switched” IGRs. To aid
15 this, Piggy calculates within-cluster divergences for both genes and IGRs, and these
16 divergences can be used to calibrate Piggy with Roary.
17
18
19
20
21
22
23
24
25
26

27 **Clustering performance**

28
29 The shorter lengths of IGRs compared with genes poses potential problems for alignment
30 accuracy. We tested the clustering performance of Piggy by producing truncated variants of
31 IGRs, adding these to the total complement of IGRs in an analysis, and then recording whether
32 the truncated IGRs were clustered with their untruncated counterparts (Methods). For *S. aureus*
33 ST22, 82% of IGRs truncated to 10 bp clustered together with the corresponding full length
34 sequences, but this figure increased to > 99% when the length of the truncated sequences was
35 20 bp. (Fig S1a). A similar increase was observed for the *E. coli* ST131 data, although in this
36 case 50 bp was required for the percentage of correct assignments to be > 99%. (Fig S1b).
37
38
39
40
41
42
43

44 An inspection of the incorrectly clustered sequences from both datasets revealed that their
45 progenitor sequences shared high sequence similarity in parts of their sequence to other IGR
46 clusters, but no sequence similarity in other parts of the sequence. This resulted in separate
47 clusters which shared high sequence homology over parts of their sequences. When these
48 sequences were truncated to assess the clustering, if the homologous part of the sequence was
49 selected, then it could align to either of these progenitor IGR clusters. In many cases these
50 alignments were perfect matches, and so the IGR could not be unambiguously placed. This
51 problem is likely to be a result of non-homologous breaks at the edge of HGT events, and this is
52 consistent with greater clustering accuracy in *S. aureus* ST22 compared with *E. coli* ST131,
53 where the latter has a much larger pan-genome.
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4
5
6 ***Staphylococcus aureus***

7 *S. aureus* is an important skin-associated bacterium which is commonly carried
8 asymptotically, but can also cause a wide range of infections from minor skin infections to
9 fatal bacteraemias. It has a clonal population structure consisting of discrete lineages [30].
10 Although the core genome is relatively stable, phenotypic variation (e.g. resistance profiles,
11 virulence traits, and host preference) is associated with a more dynamic accessory genome and
12 the horizontal transfer of MGEs, such as the SCC*mec* element which confers resistance to β -
13 lactam antibiotics [31].
14
15
16
17
18
19
20

21 *S. aureus* ST22 (EMRSA-15) is a clinically important hospital-acquired methicillin resistant strain
22 which is common in the UK and is rapidly expanding elsewhere in Europe and globally [32].
23 Previous work has shown that *S. aureus* ST22 is clonal and shows relatively little variation in
24 gene content [19,32]. In order to compare the pan-genomes of *S. aureus* at different scales, we
25 analysed a diverse dataset of 1552 isolates from many lineages, and a smaller dataset of 500
26 ST22 isolates subsampled from the larger dataset [19]. The size of the gene and IGR pan and
27 core-genomes were compared by analysing both datasets with Roary and Piggy. Frequency
28 histograms were plotted for both genes and IGRs (Fig 2a-b).
29
30
31
32
33
34
35

36 The gene-IGR frequency histogram for ST22 (Fig 2a) shows that there are 2,409 core genes
37 and 1,556 core IGRs, where core is defined as gene presence in > 95% of isolates (Table 1).
38 When the whole species is considered, these numbers drop to 2,129 and 1,134, respectively.
39 The fact that there are fewer core IGRs than core genes is in part due to the exclusion of IGRs
40 < 30 bp (many of which are intra-operonic), but also likely reflects faster evolution of IGRs. Both
41 distributions conform to the U-shape typically found in such analyses where the majority of
42 genes/IGRs are either very common or very rare, however the distribution of genes and IGRs is
43 shifted towards the rare sequences when the whole species is considered rather than only
44 ST22.
45
46
47
48
49
50
51

52 We used the output of Piggy to investigate the degree of linkage between genes and IGRs. We
53 identified all genomic loci consisting of an IGR flanked by two genes, and from these we
54 identified all pairs of genes and IGRs where the IGR was upstream of the gene start. We then
55 grouped these according to whether the gene or IGR was core or accessory (Table 2). For the
56 *S. aureus* ST22 data, 99.5% of core genes were immediately downstream of a core IGR, and
57
58
59
60
61
62
63
64
65

1
2
3
4 92.9% of the accessory genes were similarly downstream of an accessory IGR. When
5 considering the wider *S. aureus* dataset the figures were similar; 92.6% of core genes were
6 downstream of a core IGR, and 96.8% of accessory genes were downstream of an accessory
7 IGR. Thus, the assignment of an IGR as core or accessory is strongly predictive of the
8 corresponding assignment of the cognate downstream gene, which in turn points to strong
9 background linkage between genes in IGRs in the genome.
10
11
12
13
14

15 ***Escherichia coli* ST131**

16 The utility of Piggy was further validated by re-analysing data from a recent study on the
17 widespread and clinically important *E. coli* lineage ST131 [6]. This dataset contains 228 clinical
18 *E. coli* ST131 isolates from human, domesticated animal, and avian hosts. *E. coli* is a more
19 genetically diverse species than *S. aureus*, and unsurprisingly *E. coli* ST131 has a larger pan-
20 genome than *S. aureus* ST22, with 12,806 genes and 16,429 IGRs (Fig 2c, Table 1). More
21 surprisingly, *E. coli* ST131 has a larger pan-genome than the whole *S. aureus* species. Within
22 *E. coli* ST131, 3,930 genes and 2,296 IGRs were core out of an average of 4,689 genes and
23 2,984 IGRs per isolate. Thus despite the differences between the two species in their level of
24 diversity there was a consistent signal of a lower number of core IGRs than core genes, and a
25 high number of accessory IGRs compared to accessory genes. There was tight linkage between
26 genes and IGRs, with 97.9% of core genes being immediately downstream of core IGRs and
27 97.3% of accessory genes being similarly downstream of accessory IGRs; these results are
28 consistent with those from *S. aureus* (Table 2).
29
30
31
32
33
34
35
36
37
38
39
40

41 The data from *S. aureus* and *E. coli* shows a background of strong linkage between core genes
42 and IGRs. However, this linkage is not perfect; some core genes are associated with accessory
43 IGRs (and vice-versa), and the linkage is weaker over long timescales (across the whole *S.*
44 *aureus* species compared to within ST22). Previous work has examined this linkage and found
45 evidence of widespread IGR regulatory switching, where genes are regulated by alternative
46 IGRs in different isolates [17]. Piggy provides a list of candidate switching events together for
47 both “gene-pair” and “upstream” approaches (see Methods) at different thresholds of nucleotide
48 identity. For the *E. coli* ST131 data, the pipeline detected 61 cases of putative IGR switching
49 using the most conservative settings (i.e. the conservative gene-pair method, and the alternative
50 IGRs showing no sequence similarity by BLASTN). Relaxing the threshold of sequence identity
51 to < 90% resulted in the identification of an additional 317 candidate switching events, though
52 these possibly reflect either relaxed or positive selection.
53
54
55
56
57
58
59
60
61
62
63
64
65

Switched IGRs influence gene expression in *S. aureus*

To examine whether switches in IGRs affect the expression of cognate (downstream) genes, we used a previously published RNA-seq dataset based on four reference *S. aureus* isolates HO_5096_0412 (ST22), Newman (CC8), MRSA252 (CC36), and S0385 (CC398) [20]. Each of these *S. aureus* references isolate represents 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 analysing these data alongside the output from Piggy, it is possible to test the extent to which IGR switches between these four 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 four isolates, and then expression of these core genes was quantified using Kallisto [23]. To do this we used RNA-seq data for two replicates for each of the four reference genomes. The tpm (Transcripts per Kilobase Million) values for each gene are given in Table S1. We then used Sleuth [24] to normalise and filter these counts.

To check the consistency of the data between biological replicates, we first plotted two 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 ~250 fold higher in Newman compared with the other isolates.

The genomes of each pair of isolates were analysed 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 one co-transcribed downstream genes we also considered all genes linked in a single operon that could be impacted by a single switching

1
2
3
4 event upstream affecting a shared promoter. For each pair of isolates, we thus identified all core
5 genes putatively affected by upstream IGR switches. We then tested whether these genes
6 showed a higher degree of differential expression by conducting Monte Carlo permutation tests
7 on the residuals from the regressions (Fig 3). For each pairwise comparison of isolates, we
8 summed the residuals of the genes with switched IGRs (shown as red points in Fig 3), and
9 compared this to a distribution obtained by resampling (without replacement) 100,000 random
10 sets of the same number of genes and summing their residuals. We computed a one-tailed p-
11 value by dividing the number of permutations with summed residuals greater than the observed
12 value by 100,000 (Fig 3). Because we used both replicates separately (e.g. Newman replicate 1
13 vs S0385 replicate 1, and Newman replicate 2 vs S0385 replicate 2), each comparison between
14 pairs of isolates was tested twice. In 9/12 pairwise comparisons, the observed residuals of the
15 genes downstream of switched IGRs were significantly ($p < 0.05$) greater than expected from
16 the resampled data, indicating that genes with switched IGRs were more differentially
17 expressed than those without. Of the three remaining comparisons, two corresponded to
18 comparisons between HO_5096_0412 and S0385 ($p = 0.17$, and $p = 0.055$), and one between
19 HO_5096_0412 and Newman ($p = 0.054$). The second comparison between HO_5096_0412
20 and Newman was the most weakly significant result ($p = 0.025$). Thus, the two replicates for
21 each individual pairwise comparison were largely concordant with each other.
22
23
24
25
26
27
28
29
30
31
32
33
34
35

36 Our analysis confirms that genes downstream of switched IGRs are on average more likely to
37 be differentially expressed than genes not associated with IGR switches as identified using
38 Piggy. To illustrate the genomic context and expression differences of genes with switched
39 IGRs, we selected three of the most differentially expressed genes with IGR switches for the
40 Newman vs MRSA252 comparison, and plotted nucleotide identity across the IGR (calculated
41 as a 20-bp sliding window) alongside gene expression (Fig 4).
42
43
44
45
46
47

48 **Compatibility and scalability**

49 We have so far demonstrated that Piggy can be used to analyse the intergenic component of
50 the pan-genome and identify IGR switches, and shown that these switches have biological
51 relevance with respect to gene expression. Importantly, Piggy is designed such that the output
52 files are compatible with existing software and databases. The “IGR_presence_absence.csv”
53 file has an identical format to the “gene_presence_absence.csv” file produced by Roary, and
54 can be loaded directly into the interactive browser-based viewer Phandango [33] (Fig S2). It can
55 also be used as input, along with a traits file, to Scoary [34] to test for associations between
56
57
58
59
60
61
62
63
64
65

1
2
3
4 IGRs and phenotypic traits. Moreover, the “representative_clusters_merged.fasta” file can be
5 loaded directly into BIGSdb [18] to create an allele scheme for IGRs. In order to provide proof-
6 of-principle, we created a multilocus IGR (igMLST) scheme in BIGSdb. Briefly, 2,631 unique
7 IGR sequences with length ≥ 30 bp, from 7 *S. aureus* reference genomes, were entered into the
8 database locus list. Using functionality within the database, these sequences were grouped as a
9 searchable scheme (S_aureus_Intergenic_PIGGY), comparable to MLST, rMLST and wgMLST
10 schemes [35–37]. The distribution of IGRs was analysed for all isolates in the database,
11 identifying IGRs as present in the respective genome if a hit was recorded with nucleotide
12 identity $\geq 70\%$ over $\geq 50\%$ of the sequence using a BLAST word size of 7 bp. The scheme can
13 be found at [38]. Although we do not expect a typing scheme based solely on IGRs to be widely
14 used, supplementing protein-coding regions with IGR alleles may provide additional information
15 regarding links between genotype and phenotype, as well as increased epidemiological and
16 phylogenetic resolution.
17
18
19
20
21
22
23
24
25
26

27 **Discussion**

28
29 Whole-genome sequence datasets consisting of hundreds or even thousands of bacterial
30 isolates have revealed pan-genomes of many thousands of genes and large differences in gene
31 content between isolates of the same species. Currently, pan-genome diversity is considered
32 almost exclusively in terms of protein-coding genes, despite overwhelming evidence that
33 variation within IGRs impacts on phenotypes. Here we address this by introducing Piggy, a
34 pipeline specifically designed to incorporate IGRs into routine pan-genome analyses by working
35 in close conjunction with Roary [4].
36
37
38
39
40
41
42

43 The utility of this approach is demonstrated using large datasets of *S. aureus* and *E. coli* ST131.
44 Consistent with previous analyses of protein-coding regions [6,32], the IGR component of the
45 ST131 pan-genome is considerably larger than that for ST22, and surprisingly is also larger
46 than the pan-genome of the whole *S. aureus* species. There was more diversity within IGRs
47 than genes in both species. While some IGRs may be essential for expression of multiple
48 genes, IGRs are broadly subject to weaker purifying selection than protein coding genes [14].
49 The maintenance of core IGRs in both bacterial genome datasets is consistent with selection
50 acting to conserve them and allows alignment and analysis in much the same way as protein-
51 coding regions.
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 The current exclusion of IGRs from routine pan-genome or cgMLST analyses may in part reflect
5 perceived difficulties in the alignment and subsequent cluster definition, particularly if the
6 sequences are very short. We therefore validated the pipeline by investigating clustering
7 accuracy as a function of sequence length by truncating the IGR sequences and recording
8 whether they remained in the same cluster as their full-length counterparts. For *S. aureus*, the
9 data showed that truncated IGRs > 20 bp almost always remained in the original cluster,
10 confirming that the minimum length permitted in the pipeline of 30-bp is conservative. For *E.*
11 *coli*, truncating the sequences had greater impact on cluster assignments, and a minimum
12 length of 50 bp would be a safer setting in this case. The problems with clustering shorter
13 sequences in *E. coli*, compared to *S. aureus*, are not due to the length of the sequence per se
14 but reflect the higher rate of horizontal gene transfer in this species. This means that the IGRs
15 are more likely to be chimeric in structure, with localised regions within the IGRs showing a high
16 level of homology to different clusters. This lead to cluster assignment being dependant not so
17 much on length, but on which part of the truncated sequence happened to be retained.
18
19
20
21
22
23
24
25
26
27
28

29 Variation within regulatory elements located within IGRs can impact on the expression of the
30 downstream gene [17]. Piggy (alongside Roary) provides the means to combine information on
31 genes and their cognate IGRs thus facilitating the detection of “switched” IGRs and downstream
32 genes that are potentially affected. We have shown that in *S. aureus*, genes with switched
33 upstream IGRs show a higher degree of differential expression than those without. This is
34 consistent with previous work on *E. coli* [17], and suggests that the identification of IGR
35 switches using Piggy can provide a useful indication of differential gene expression, even in the
36 absence of RNA-seq data. However, we note that high divergence within IGRs does not
37 necessarily imply selection for differential gene expression, and may instead simply reflect
38 weaker selective constraints. A clear direction for future work is to make constructs consisting of
39 genes with alternative IGRs, in order to directly measure the effect of natural IGR variants on
40 gene expression. Similar experiments have previously been performed in *E. coli* based on
41 variation within promoters [39], and IGRs more broadly [17]. The importance of changes in gene
42 expression mediated by intergenic variation as a route of adaptation is currently unknown, but
43 one recent study suggested that intergenic changes are strongly positively selected in
44 *Pseudomonas aeruginosa* during infection in patients with cystic fibrosis, and more work is
45 required to test the generality of these findings [16].
46
47
48
49
50
51
52
53
54
55
56
57
58

59 **Conclusions**

60
61
62
63
64
65

1
2
3
4 Driven by recent technical advances in high-throughput sequencing, large whole-genome
5 datasets have provided powerful evidence concerning the genetic determinants that underlie
6 complex multifactorial phenotypes such as virulence. Moreover, associating variation in core
7 and accessory genes with phenotype data is providing new fundamental insight into the ecology
8 and evolution of bacteria. However, in much the same way that non-protein coding DNA in the
9 human genome was initially dismissed as “junk”, omitting IGRs from bacterial genome analysis
10 severely limits our ability to draw inferences on the regulation of gene expression and
11 associated phenotypic consequences. By developing Piggy as an easy-to-use bioinformatics
12 tool with output files that are compatible with existing software and databases (eg Roary,
13 Phandango; Figure S1, Scoary, BIGSdb) we envisage that combined information from genes
14 and their cognate IGRs will vastly improve our understanding of genome evolution in bacteria.
15
16
17
18
19
20
21
22
23

24 **Availability of supporting source code**

25 Project name: Piggy

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

27 Operating system(s): Linux

28 Programming language: Perl, R

29 Other requirements: Roary

30 License: GPLv3

31 RRID: (Piggy, RRID:SCR_015941)
32
33
34
35
36
37
38
39
40

41 **Availability of supporting data**

42 The *S. aureus* dataset was assembled from published genome sequences [19] available from
43 the European Nucleotide Archive (ENA), study number ERP001012. The *S. aureus* RNA-seq
44 data was previously published [20], and is available from the ENA, study number ERP009279.
45 This was supplemented with the corresponding reference genomes, HO_5096_0412:
46 HE681097, MRSA252: BX571856, Newman: AP009351, S0385: AM990992, available from the
47 National Center for Biotechnology Information (NCBI). The *E. coli* ST131 dataset was from a
48 previously published study [6], and is available at [21]. An archival copy of the Piggy source
49 code is available via the *GigaScience* repository GigaDB[40].
50
51
52
53
54
55
56
57

58 **Declarations**

59 Ethics approval and consent to participate: Not applicable
60
61
62
63
64
65

1
2
3
4 Consent for publication: Not applicable

5
6 Competing interests: Not applicable
7
8

9 Funding: The *Staphylococcus aureus* genome sequences were generated as part of a study
10 supported by a grant from the United Kingdom Clinical Research Collaboration Translational
11 Infection Research Initiative and the Medical Research Council (grant number G1000803, held
12 by Sharon Peacock) with contributions from the Biotechnology and Biological Sciences
13 Research Council; the National Institute for Health Research on behalf of the Department of
14 Health; and the Chief Scientist Office of the Scottish Government Health Directorate, on which
15 E.J.F. was a principal investigator and S.C.B. was a postdoctoral researcher. H.A.T. is funded
16 by a University of Bath research studentship. The funders had no role in study design, data
17 collection and analysis, decision to publish, or preparation of the manuscript.
18
19
20
21
22
23
24
25

26 Authors' contributions: HAT designed and implemented the pipeline, and carried out the majority
27 of the analyses, with input from E.J.F., S.C.B. and S.K.S. HAT and E.J.F. wrote the manuscript with
28 input from S.K.S. and S.C.B.
29
30
31

32 Acknowledgements: We are very grateful to Torsten Seemann, Andrew Page and João Carriço
33 for encouragement and helpful feedback. We are also grateful to Matt Holden for provision of
34 the *S. aureus* RNA-seq data, to Sandra Reuter for help with the *S. aureus* data, and to Alan
35 McNally for the *E. coli* ST131 data. This work also greatly benefitted from access to the Medical
36 Research Council funded Cloud Infrastructure for Microbial Bioinformatics (MRC-CLIMB) [41].
37
38
39
40
41

42 **References**

- 43
44
45 1. McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. *Nat Microbiol.*
46 2017;2:17040.
47
48 2. Andreani NA, Hesse E, Vos M. Prokaryote genome fluidity is dependent on effective
49 population size. *ISME J.* 2017;11:1719–21.
50
51 3. Medini D, Donati C, Tettelin H, Massignani V, Rappuoli R. The microbial pan-genome. *Curr.*
52 *Opin. Genet. Dev.* 2005;15:589–94.
53
54 4. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-
55 scale prokaryote pan genome analysis. *Bioinformatics.* 2015;31:3691–3.
56
57 5. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, et al. Genomic
58 analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella*
59
60
61
62
63
64
65

- 1
2
3
4 pneumoniae, an urgent threat to public health. *Proc. Natl. Acad. Sci. U. S. A.* 2015;112:E3574–
5 81.
6
7
8 6. McNally A, Oren Y, Kelly D, Pascoe B, Dunn S, Sreecharan T, et al. Combined Analysis of
9 Variation in Core, Accessory and Regulatory Genome Regions Provides a Super-Resolution
10 View into the Evolution of Bacterial Populations. *PLoS Genet.* 2016;12:e1006280.
11
12 7. Vos M, Hesselman MC, te Beek TA, van Passel MWJ, Eyre-Walker A. Rates of Lateral Gene
13 Transfer in Prokaryotes: High but Why? *Trends Microbiol.* 2015;23:598–605.
14
15 8. Fouts DE, Brinkac L, Beck E, Inman J, Sutton G. PanOCT: automated clustering of orthologs
16 using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely
17 related species. *Nucleic Acids Res.* 2012;40:e172.
18
19 9. Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J. PGAP: pan-genomes analysis pipeline.
20 *Bioinformatics.* 2012;28:416–8.
21
22 10. Sahl JW, Caporaso JG, Rasko DA, Keim P. The large-scale blast score ratio (LS-BSR)
23 pipeline: a method to rapidly compare genetic content between bacterial genomes. *PeerJ.*
24 2014;2:e332.
25
26 11. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics.* 2014;30:2068–9.
27
28 12. Ochman H, Caro-Quintero A. Genome Size and Structure, Bacterial. In: Kliman RM, editor.
29 *Encyclopedia of Evolutionary Biology.* Oxford: Academic Press; 2016. p. 179–85.
30
31 13. McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. *Nat. Rev.*
32 *Microbiol.* 2011;10:13–26.
33
34 14. Thorpe HA, Bayliss SC, Hurst LD, Feil EJ. Comparative Analyses of Selection Operating on
35 Non-translated Intergenic Regions of Diverse Bacterial Species. *Genetics [Internet].* 2017;
36 Available from: <http://dx.doi.org/10.1534/genetics.116.195784>
37
38
39 15. Molina N, Van Nimwegen E. Universal patterns of purifying selection at noncoding positions
40 in bacteria. *Genome Res.* 2008;18:148–60.
41
42 16. Khademi H, Jelsbak L. Host adaptation mediated by intergenic evolution in a bacterial
43 pathogen [Internet]. *bioRxiv.* 2017 [cited 2017 Dec 20]. p. 236000. Available from:
44 <https://www.biorxiv.org/content/early/2017/12/19/236000.article-info>
45
46
47 17. Oren Y, Smith MB, Johns NI, Kaplan Zeevi M, Biran D, Ron EZ, et al. Transfer of noncoding
48 DNA drives regulatory rewiring in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* 2014;111:16112–7.
49
50 18. Jolley KA, Maiden MCJ. BIGSdb: Scalable analysis of bacterial genome variation at the
51 population level. *BMC Bioinformatics.* 2010;11:595.
52
53 19. Reuter S, Török EM, Holden MTG, Reynolds R, Raven KE, Blane B, et al. Building a
54 genomic framework for prospective MRSA surveillance in the United Kingdom and the Republic
55 of Ireland. *Genome Res.* [Internet]. 2015; Available from:
56 <http://genome.cshlp.org/content/early/2015/12/15/gr.196709.115.abstract>
57
58
59 20. Warne B, Harkins CP, Harris SR, Vatsiou A, Stanley-Wall N, Parkhill J, et al. The Ess/Type
60 VII secretion system of *Staphylococcus aureus* shows unexpected genetic diversity. *BMC*
61
62
63
64
65

1
2
3
4 Genomics. 2016;17:222.
5

6 21. McNally et al. 2016 data repository [Internet]. Dryad. [cited 2016 Oct 25]. Available from:
7 <http://datadryad.org/resource/doi:10.5061/dryad.d7d71>
8

9 22. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7:
10 improvements in performance and usability. *Mol. Biol. Evol.* 2013;30:772–80.
11

12 23. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq
13 quantification. *Nat. Biotechnol.* 2016;34:525–7.
14

15 24. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq
16 incorporating quantification uncertainty. *Nat. Methods.* 2017;14:687–90.
17

18 25. Tjaden B. De novo assembly of bacterial transcriptomes from RNA-seq data. *Genome Biol.*
19 2015;16:1.
20

21 26. RDevelopment CORE TEAM R, Others. R: A language and environment for statistical
22 computing [Internet]. R foundation for statistical computing Vienna, Austria; 2008. Available
23 from: <http://www.R-project.org>
24

25 27. Wickham H. *Ggplot2: Elegant Graphics for Data Analysis*. 2nd ed. Springer Publishing
26 Company, Incorporated; 2009.
27

28 28. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation
29 sequencing data. *Bioinformatics.* 2012;28:3150–2.
30

31 29. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+:
32 architecture and applications. *BMC Bioinformatics.* 2009;10:421.
33

34 30. Feil EJ, Cooper JE, Grundmann H, Robinson DA, Enright MC, Berendt T, et al. How clonal
35 is *Staphylococcus aureus*? *J. Bacteriol.* 2003;185:3307–16.
36

37 31. Lindsay JA, Holden MTG. *Staphylococcus aureus*: superbug, super genome? *Trends*
38 *Microbiol.* 2004;12:378–85.
39

40 32. Holden MTG, Hsu L-Y, Kurt K, Weinert LA, Mather AE, Harris SR, et al. A genomic portrait
41 of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus*
42 pandemic. *Genome Res.* 2013;23:653–64.
43

44 33. Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, Harris SR. Phandango:
45 an interactive viewer for bacterial population genomics. *Bioinformatics* [Internet]. 2017; Available
46 from: <http://dx.doi.org/10.1093/bioinformatics/btx610>
47

48 34. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-
49 genome-wide association studies with Scoary. *Genome Biol.* 2016;17:238.
50

51 35. Maiden MCJ, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, et al. MLST
52 revisited: the gene-by-gene approach to bacterial genomics. *Nat. Rev. Microbiol.* 2013;11:728–
53 36.
54

55 36. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, et al. Ribosomal
56 multilocus sequence typing: universal characterization of bacteria from domain to strain.
57
58
59
60
61
62
63
64
65

1
2
3
4 Microbiology. 2012;158:1005–15.
5

6 37. Sheppard SK, Jolley KA, Maiden MCJ. A gene-by-gene approach to bacterial population
7 genomics: Whole genome MLST of *Campylobacter*. *Genes* . 2012;3:261–77.
8

9 38. Sheppard lab resources [Internet]. [cited 2018 Jan 16]. Available from:
10 <https://sheppardlab.com/resources>
11

12 39. Shimada T, Yamazaki Y, Tanaka K, Ishihama A. The whole set of constitutive promoters
13 recognized by RNA polymerase RpoD holoenzyme of *Escherichia coli*. *PLoS One*.
14 2014;9:e90447.
15

16 40. Thorpe HA, Bayliss SC, Sheppard SK, Feil EJ. Supporting data for ‘Piggy: A Rapid, Large-
17 Scale Pan-Genome Analysis Tool for Intergenic Regions in Bacteria’. GigaScience database
18 2018. <http://dx.doi.org/10.5524/100410>
19
20

21 41. Connor TR, Loman NJ, Thompson S, Smith A, Southgate J, Poplawski R, et al. CLIMB (the
22 Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical
23 microbiology community. *Microbial Genomics* [Internet]. Microbiology Society; 2016 [cited 2016
24 Nov 1];2. Available from:
25 <http://mgen.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000086>
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Figure legends

Fig 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 two genes is non-homologous (“switched”), and for the “upstream” method both the upstream IGR and gene may be non-homologous to the downstream gene.

Fig 2: Properties of the pan-genomes. Genes (red) and IGRs (blue) were analysed 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* c) *E. coli* ST131.

Fig 3: *S. aureus* gene expression data. Pairwise RNA-seq comparisons between four *S. aureus* isolates, where two 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 two biological replicates from the same isolate. 2094 core genes were analysed in each comparison, and tpm (Transcripts per Kilobase Million) was used to quantify expression. The genes were separated into two categories: Switched (red), and Not-switched (grey), 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 two groups of genes, where a significant score indicates that the genes downstream of switch IGRs are associated with a higher degree of differential expression (ie higher residuals).

Fig 4: A detailed view of the genomic neighbourhood 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

1
2
3
4 orientation (left). The corresponding expression data for the gene of interest was also shown
5 (right), with the two boxplots per isolate corresponding to the two biological replicates. a) *dapE*
6 b) *ssaA_1* c) *ytrA*.
7
8
9

10
11 **Fig. S1: Clustering performance.** The clustering performance was assessed by truncating IGR
12 sequences and reclustering them with the pool of original sequences. Truncated IGRs which
13 were placed into the same cluster as their progenitor sequences were deemed to be correctly
14 clustered. a) *S. aureus* b) *E. coli*.
15
16
17
18

19
20 **Fig S2: The IGR pan-genome as visualised using Phandango.** A neighbour-joining
21 phylogenetic tree was imported into Phandango [33] alongside the *IGR_presence_absence.csv*
22 file. Each row corresponds to an isolate, and each column corresponds to an IGR, with the IGRs
23 ordered from the left in order of decreasing frequency within the sample. The line graph at the
24 bottom shows the frequency of the IGRs within the sample. a) *S. aureus* ST22 b) *E. coli* ST131.
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Sheet1

Species	Core genes	Core IGRs	Accessory genes	Accessory IGRs	Percentage core genes	Percentage core IGRs
<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	14133	84	77

Table 1

Sheet1

Species	Core gene, Core IGR	Core gene, Accessory	Accessory gene, Core IGR
<i>S. aureus</i> ST22	99.5	0.5	7.4
<i>S. aureus</i>	92.9	7.1	3.2
<i>E. coli</i> ST131	97.9	2.1	2.7

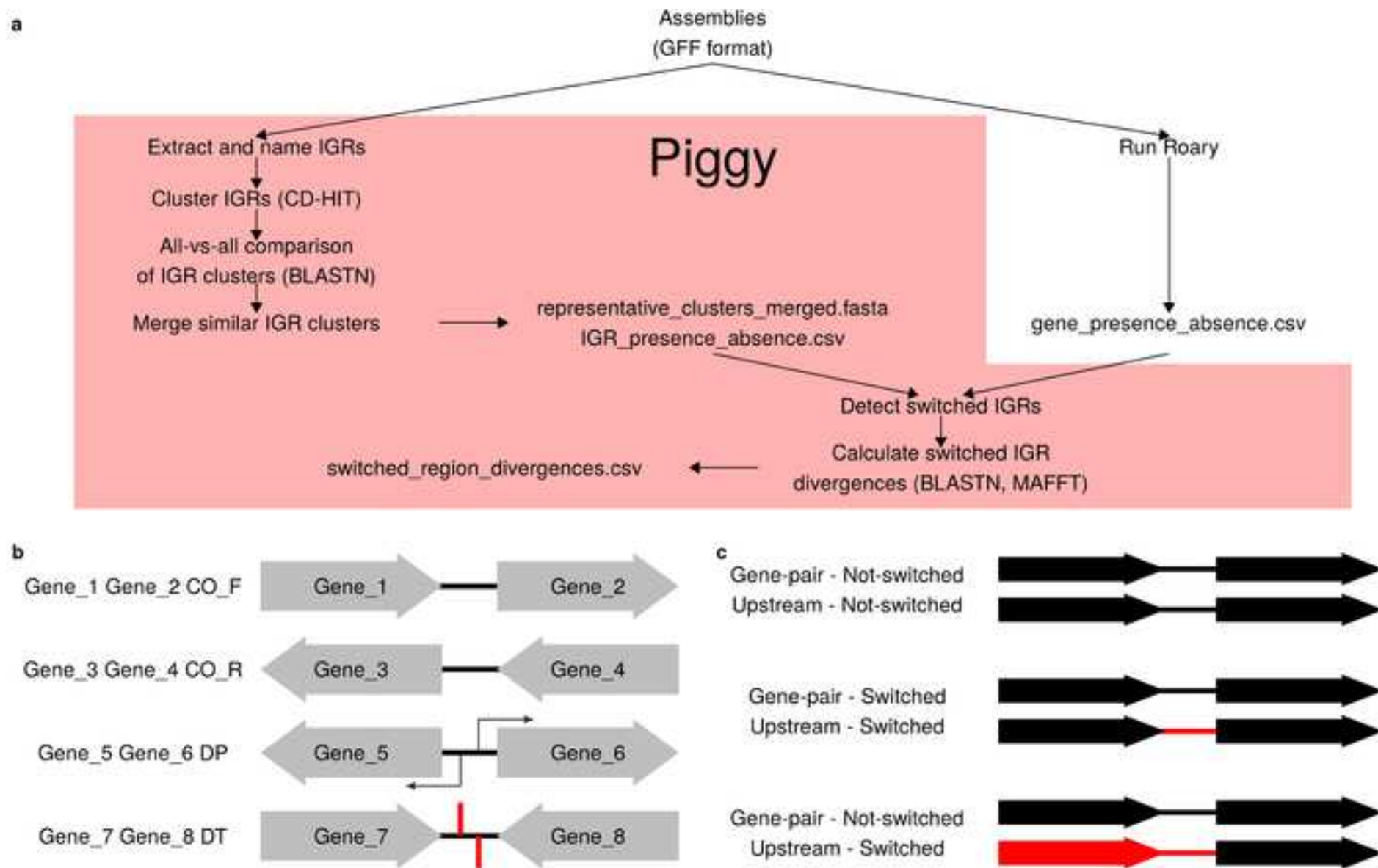
Table 2

Accessory gene, Accessory IGR

92.6

96.8

97.3



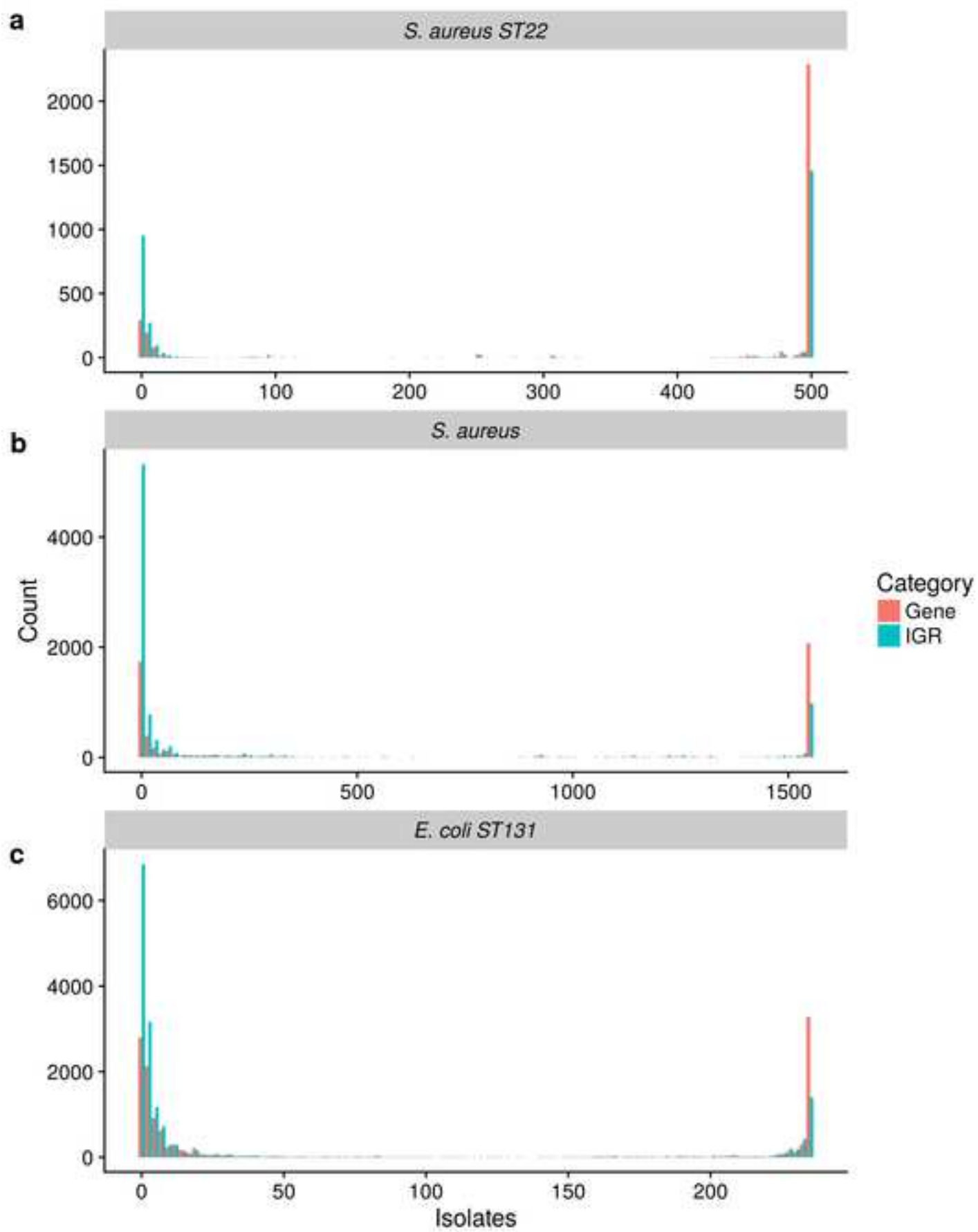
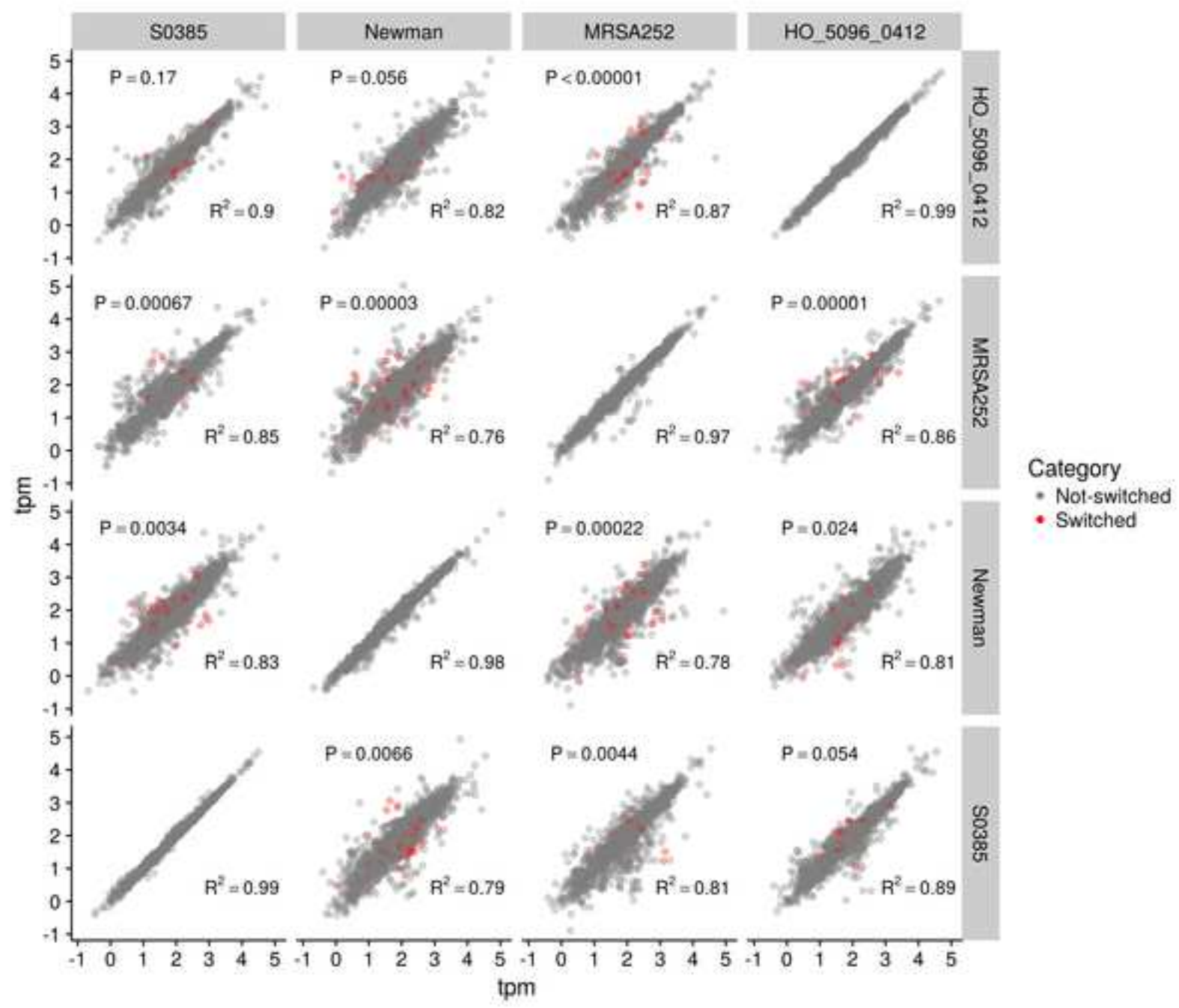
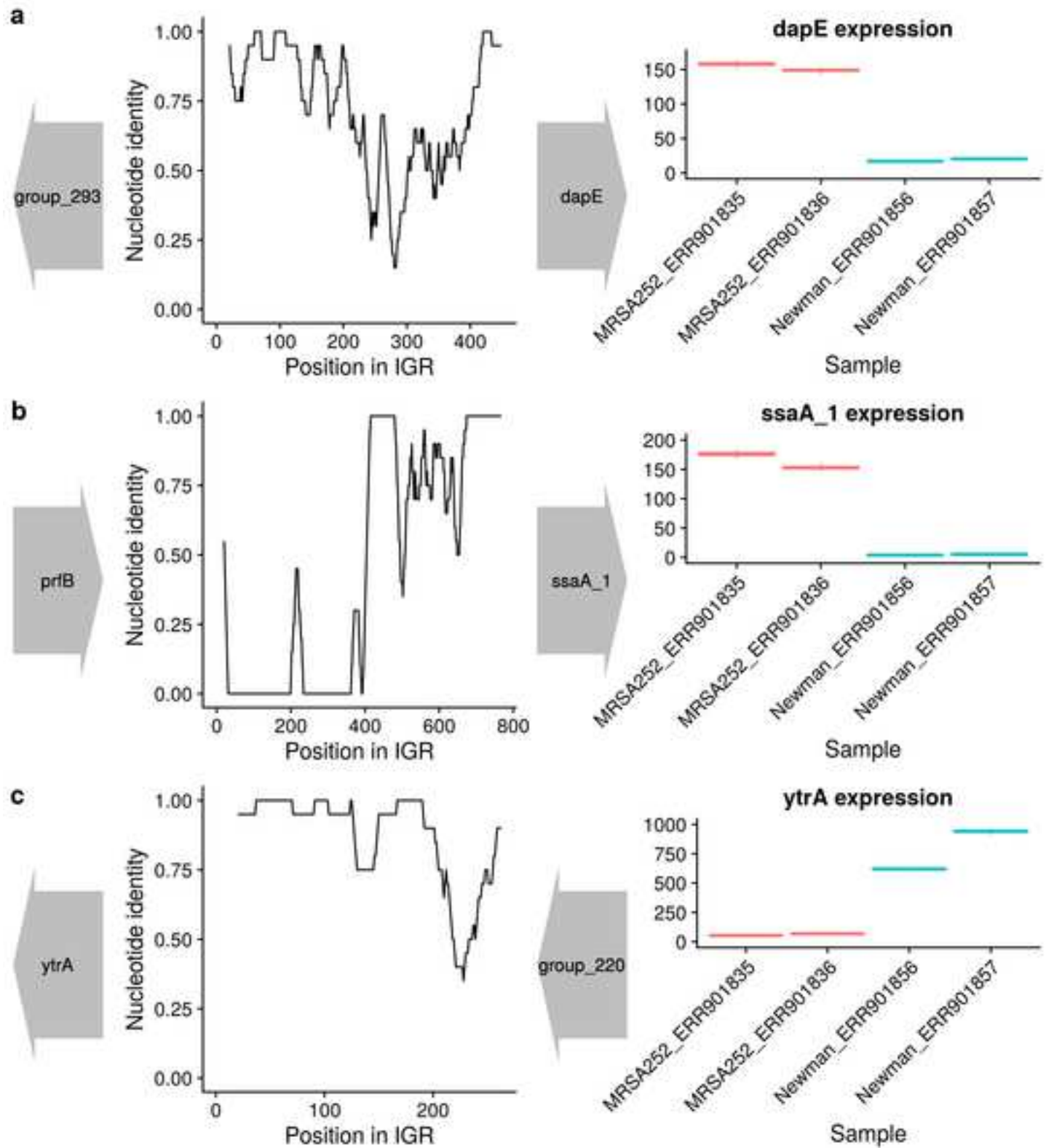
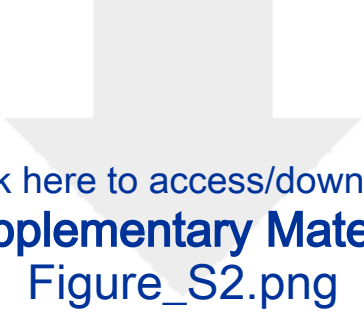


Figure 3


[Click here to download Figure Figure_3.png](#)









Click here to access/download
Supplementary Material
Figure_S2.png





Click here to access/download
Supplementary Material
Table_S1.xlsx



Click here to access/download
Supplementary Material
Figure_S1.png