

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

--Manuscript Draft--

Manuscript Number:	GIGA-D-17-00244R1	
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>Despite overwhelming evidence that variation in intergenic regions (IGRs) in bacteria can directly influence phenotypes, most current approaches for analysing pan-genomes focus exclusively on protein-coding sequences. To address this we present Piggy, a novel pipeline that emulates Roary except that it is based only on IGRs. We demonstrate the use of Piggy for pan-genome analyses of <i>Staphylococcus aureus</i> and <i>Escherichia coli</i> using large genome datasets. 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 and registered with SciCrunch (RRID: SCR_015941).</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 Feil	
Order of Authors Secondary Information:		
Response to Reviewers:	<p>Reviewer reports:</p> <p>Reviewer #1: Piggy represents a potentially valuable tool to the field of comparative genomics. In general, additional details on how the algorithm works would be helpful to understand the results.</p> <p>RESPONSE: We thank the reviewer for recognizing the value of our approach – we have added additional details concerning the algorithm throughout the manuscript as requested.</p> <p>P1,L16; bacteria "has" impacts</p> <p>RESPONSE: P2,L35-36: This line now reads "variation in intergenic regions (IGRs) in bacteria can directly influence phenotypes"</p>	

P2,L9: Add references to this first line

RESPONSE: P2,L46: Added references: McInerney et al. 2017; Andreani et al. 2017

P2,L14: Relationship between pan-genome and core will differ greatly on the organism chosen

RESPONSE: We agree with this point and have added the following text:

P2,L59-62: "More generally, the relationship between the size of the core and accessory genomes varies between species. Broadly, ecological generalists have large accessory genomes, whilst more ecologically restricted species, such as endosymbionts, have much smaller accessory genomes (McInerney et al. 2017; Andreani et al. 2017)."

P2,L30-34: This is a run-on sentence and could be broken up to improve clarity

RESPONSE: P3,L67-70: This sentence now reads:

The increasing availability of datasets containing 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.

P3,L11: I have several problems with this statement about LS-BSR. What do you mean that it is no longer specific. Specific to what? Also, you mention that this reduced specificity is a by-product of pre-clustering, but the next sentence indicates that Roary also uses pre-clustering. Why wouldn't that also affect the results?

RESPONSE: P3,L74-77: We apologise for the confusion, and on reflection agree with the referee that the text was not reflective of the relative performance of the two methods. We have changed the text accordingly.

P3,L16-17: You mention that Roary is "more accurate than LS-BSR" and this is likely based on one comparison in the Roary paper. This was the result of one simulated dataset, using an unknown version of USEARCH and unknown parameters for alignment. To be safe, if you want to still report these results, I would mention that Roary was more accurate than LS-BSR using one simulated dataset, although the details remain unclear. You could safely remove this statement and not detract from the rest of your manuscript.

RESPONSE: P3,L74-77: Again, we completely agree with the referee and have modified the text accordingly.

P3,L39: Reference for "15% of the genome" statement?

RESPONSE: P3,L86: Added references: Ochman and Caro-Quintero 2016; McCutcheon and Moran 2011

P13,L4-6: What lengths of IGRs do you consider? Is there a minimum length? What do you do at the beginning and ends of draft contigs? More detail here would be very helpful.

RESPONSE: We have provided more detail in the text as requested:

P7,L204-206: IGRs at the edge of contigs are excluded by default, but when they are included (using the --edges flag) the missing information is denoted by NA, for example 'Gene_1 NA NA'.

P7,L207-209: 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).

P13,L27: What BLASTN parameters do you use to merge similar clusters?

P7,L218-219: More detail provided: BLASTN defaults, except -word_size = 10

P13,L27: What thresholds do you decide on for presence/absence?

P7,L219-221: Thresholds are provided by --len_id and --nuc_id, and these are used to produce clusters. Once the clusters have been produced, the gene presence information is simply a matrix of these clusters vs strains.

Fig S1: These trees look to be unrooted, but am unsure of why

RESPONSE:The phandango tool provides a visual comparison between the relatedness based on core genome variation with differences in gene content. The use of an outgroup to root the tree is not required for this.

Reviewer #2: The manuscript entitled: "Piggy: A Rapid, Large-Scale Pan-Genome Analysis Tool for Intergenic Regions in Bacteria" introduces the pipeline Piggy for the analysis of intergenic regions (IGRs). The authors correctly point out that current approaches in pan-genome analysis focus purely on genes. They present a pipeline to address the remaining parts of the genome.

Based on published RNA-seq data the manuscript highlights that especially for the analysis of gene expression the state of the intergenic region can be relevant and should be considered carefully.

Since the presented pipeline equals to a great extent the approach of the software Roary, the main contribution of this work is the identification of switched IGRs. In particular, the handling of differently annotated gene borders is solved in a clever way. So far no standard file format for pangenomic data has established but the output format of Roary can be used by a bunch of analysis and visualization tools (panX, Phandango, FriPan).

It is thus reasonable to use this format for the output of Piggy.

Since for large parts of the intergenic regions in bacteria the function is unknown and most of these regions are very short, I am not sure how accurate the reconstruction of the "panIGRome" by Piggy currently is (see point 1. below).

However, before I can recommend accepting the manuscript there are some further points I would like to see addressed by the authors.:

Major points:

1. Intergenic regions in bacteria are usually much shorter than protein-coding sequences. Thus the clustering of these regions is potentially more vulnerable to wrongly aligned short sequences. Please add a part on the clustering performance to the manuscript.

RESPONSE:We thank the referee for this important point, and have spent considerable time addressing this issue in detail. Additional analyses on clustering performance are incorporated in the text (in both the Methods, P6,178-187, Results, P8-9,L252-271, and Discussion, P14,L445-458) as described below, and we feel this significantly improves the paper.

Our approach to examining clustering performance was based on truncating IGRs and re-clustering them with the original set of IGRs. This was based on the logic that if the truncation had no effect (i.e. if the same clusters were recovered), then this provides reassurance that the clustering is not confounded by the length of the sequences, at least within the relevant parameters we are using.

This approach confirmed that 20-30 bp represents a minimum length for reliable clustering of IGRs for *S. aureus*, but possibly slightly longer for *E. coli*. The incorrect clustering at these lengths was mostly driven by IGRs which are homologous to other IGRs over part, but not all of the sequence (as a result of rearrangements, HGT etc). In these cases when the IGR was truncated it could align equally well with multiple original IGR sequences, depending on which section of the sequence was retained during truncation. This may be a problem at the edge of contigs, but these IGRs are (now) removed by default (updated in the newest version of Piggy on GitHub) -

P7,L204-206. Due to the high number of incorrectly clustered IGRs when truncated to 10 bp, we recommend that these sequences are not included in the analysis at all.

2. page 16 line 27-39. Why did you use two different clusterings? One very loose clustering for Fig 2 and 3 and one more rigid for the rest of the manuscript? I do not see the point of using two different clusterings. Either two IGRs have the same origin or not. There should be an optimal value for --len_id where the clustering is close to the true relationship. And this one should be used for all subsequent analyses.

RESPONSE:With respect, we feel that there is no true --len_id which is appropriate for all situations, in the same way that there is no true --nuc_id. Of course it is true that either IGRs have the same origin or not, but when faced with real data the rules for assigning clusters are essentially pragmatic rather than grounded in biological certainties. Hence Piggy (and Roary, LS-BSR, PanOCT) use thresholds to define clusters. An IGR may acquire a deletion in one strain which means it is no longer the same length as the same IGR in other strains, despite sharing a common history.

The loose setting (--len_id 10) was used to enable a fair comparison with Roary results, where genes of different lengths are frequently clustered together. These can be the result of genuine truncations or assembly errors. Roary only requires that genes are >120 bp in length, and does not require genes to be similar in length in order to cluster together (fully explained on P5-6,L152-168). The stricter setting (--len_id 90) was used to detect switching, as this enables downstream filtering based on either length or nucleotide identity (P6,L166-168).

3. The text emphasizes that it is so far unknown whether genes and IGRs should be considered as independent or closely linked units. Likely this will depend on the context of the scientific question. Instead of separate genes g or IGRs i the set of both (i,g) can be considered. In this case one could get a first impression on the linkage of both. While the identification of switched IGRs in the manuscript uses the information of the flanking genes, I would have loved to read a bit more about this link in the two data examples. How many core genes are flanked by core IGRs? How many different genes can be found next to the same IGR and how many different IGR does a gene have? Even a first impression on these numbers would improve the quality of the manuscript.

RESPONSE:We agree that this is an important consideration, and so have done an analysis which is designed to be a first impression on these numbers. We analysed the number of core and accessory genes which are immediately upstream of core and accessory IGRs, and presented these data in a table (Table 2), and also in the text:

RESPONSE:P10,L302-312: We used the output of Piggy to investigate the degree of linkage between genes and IGRs. We identified all genomic loci consisting of an IGR flanked by two genes, and from these we identified all pairs of genes and IGRs where the IGR was upstream of the gene. 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 to strong background linkage between genes in IGRs in the genome.

P10,L324-327: 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).

In addition, please state how you proceeded with genes where a gene has an IGR > 30bp in one strain and an IGR < 30bp in another strain. Are those genes excluded from your analysis?

RESPONSE:When an IGR was > 30 bp in one strain and < 30 bp in another, then

those sequences > 30 bp would be included and the others would not. This is because the IGRs are selected before the clustering is done, and so the relationships between these sequences is not known.

4. The pan-genome can be studied at all levels of divergence from the level of single lineages within pathogenic strains up to the level of all bacteria. Piggy has been demonstrated in two closely related datasets based on a single lineage from *S. aureus* and *E. coli*, respectively. I am wondering if this is the envisaged distance of genomes to analyze and whether the pipeline can be used on more diverse datasets. In the former case, the manuscript should state more precisely that piggy is intended only for closely related bacterial strains. In the latter case, I would like to see the addition of some further more distantly related strains of *S. aureus* and/or *E. coli*.

RESPONSE: We have now included an additional analysis consisting of a diverse collection of 1500 *S. aureus* isolates (P9, L294, Fig 2b). This clearly shows that the size of the species-wide *S. aureus* pan-genome is much greater than that of ST22 (fourfold increase in the number of accessory genes, and fivefold increase in accessory IGRs) (Table 1). There was also a corresponding decrease in the number of core elements, although this was much more modest. That Piggy identified >2000 core genes and >1000 core IGRs suggests that Piggy can cope with diverse datasets (Table 1).

5. paragraph starting at page 9 at line 44:

In this paragraph a resampling method is used to show that between certain strains of *S. aureus* genes linked to a switched IGR are on average more differentially expressed than other genes.

While the resampling approach is appropriate to produce p-values in this setting, I do not understand how these p-values have been adjusted. The Benjamini-Hochberg method is usually not used to change p-values, and one has to choose an acceptable false discovery rate. Which FDR did you choose? In addition, the observations need to be independent, which is clearly not the case in the 12 pairwise comparisons.

I would recommend to either just show the simulated p-values and choose a level of significance below 0.05 or explain much more detailed what has been adjusted and why.

In addition, please stick to lowercase "p" for the p-value. Also in Figure 4.

RESPONSE: P12, L384-393: The p-values have been left unadjusted, and those < 0.05 were deemed significant. Lowercase p was used throughout. "Independently" has been removed from the text.

6. I understand that the data provided by Piggy can be directly used to create an allele scheme. But I do not see the benefit of creating an allele scheme for IGRs compared to the wgMLST schemes. Could you please clarify how this scheme could be used and what would be the advantage compared to MLST, rMLST and wgMLST?

RESPONSE: The IGR scheme is not expected to be used in isolation, but rather can be combined with a scheme based on genes which may offer increased resolution in very closely related sets of strains. We have added some explanation of this:

P13, L421-424: "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."

Minor issues:

Please explain more clearly why IGRs < 30 bp are excluded. Is this due to problems with the clustering and how did you determine the border at 30 bp?

RESPONSE: The exclusion of IGRs < 30 bp is a conservative threshold as evidenced by the clustering assessment as described above.

Figure 1: The text in the flow diagram should be much larger.

RESPONSE: We have increased the size of the text in this figure.

	<p>Figure 2: In my opinion accumulation curves in pan-genome studies are not very informative and could easily be replaced by a simple table with the average number per genome and the total number in the pan-genome. I suggest to replace Fig 2b and Fig 3b by such a table and use the opportunity to replace vague statements about the gradient and the plateau of the accumulation curve in the text. The accumulation curves could still appear in the supplemental material.</p> <p>RESPONSE: Figures 2 and 3 have been merged into one (Figure 2), and the accumulation curves and vague statements have been removed. A new table (Table 1) has been created and the text adjusted.</p> <p>Figure 4: You could highlight the points in Figure 4 corresponding to the genes from Figure 5</p> <p>RESPONSE: Figure 5 only serves as an illustration of the data using some example genes. Highlighting these genes on Figure 4 may draw unnecessary attention to them, and this is not the message we are trying to convey, which is that there is a moderate and widespread effect of IGR divergence on gene expression which is not limited to a few hand-picked genes.</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

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

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

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

34 **Abstract**

35 Despite overwhelming evidence that variation in intergenic regions (IGRs) in bacteria can
36 directly influence phenotypes, most current approaches for analysing pan-genomes focus
37 exclusively on protein-coding sequences. To address this we present Piggy, a novel pipeline
38 that emulates Roary except that it is based only on IGRs. We demonstrate the use of Piggy for
39 pan-genome analyses of *Staphylococcus aureus* and *Escherichia coli* using large genome
40 datasets. For *S. aureus*, we show that highly divergent (“switched”) IGRs are associated with
41 differences in gene expression, and we establish a multi-locus reference database of IGR
42 alleles (igMLST; implemented in BIGSdb). Piggy is available at [https://github.com/harry-](https://github.com/harry-thorpe/piggy)
43 [thorpe/piggy](https://github.com/harry-thorpe/piggy) and registered with SciCrunch (RRID: SCR_015941).

45 **Background**

46 Whole-genome sequencing has revealed that, in many bacteria, individual strains frequently
47 recruit new genes from a seemingly endless genetic reservoir (McInerney, McNally, and
48 O’Connell 2017; Andreani, Hesse, and Vos 2017). The total complement of genes observed
49 across all strains, known as the pan-genome, often numbers tens of thousands, up to an order
50 of magnitude more than the number of genes present in any single genome. In contrast, the
51 “core-genome”, which refers to the complement of genes present in all (or the vast majority) of
52 sampled isolates, can be significantly smaller than the total number of genes in any given
53 genome (Medini et al. 2005; Page et al. 2015). For example, a study of 328 *Klebsiella*
54 *pneumoniae* isolates, each of which harbour 4-5,000 genes, revealed a pan-genome of 29,886
55 genes; only 1,888 (6.8%) of which were universally present (core) (Holt et al. 2015). Similarly,
56 genome data for 228 *Escherichia coli* ST131 isolates revealed a pan-genome of 11,401 genes,
57 of which 2,722 (23.9%) were core (McNally et al. 2016). The degree of gene content variation in
58 the latter study is particularly striking as these isolates were all from the same sequence type
59 (ST), thus show limited nucleotide divergence in core genes, and are descended from a recent
60 common ancestor. More generally, the relationship between the size of the core and accessory
61 genomes varies between species, with ecologically diverse species having large accessory
62 genomes, and ecologically restricted species (such as endosymbionts) having small accessory
63 genomes (McInerney, McNally, and O’Connell 2017; Andreani, Hesse, and Vos 2017).

64
65 There is growing recognition that the acquisition of new genes through horizontal gene transfer
(HGT) has a central role in ecological adaptation (Vos et al. 2015). The emergence and spread

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

67 of antibiotic resistance, underpinned by the transfer of plasmids and other mobile genetic
68 elements (MGEs), is a pertinent example. The increasing availability of datasets containing
69 thousands of isolates thus offers an unprecedented opportunity for describing the genetic basis
70 of bacterial adaptation, although the scale of these data presents serious logistic and
71 conceptual challenges in terms of data management and analysis.

72
73 Pioneering pan-genome analysis tools, such as PanOCT and PGAP relied on all-vs-all BLAST
74 comparisons between protein sequences, and scaled approximately quadratically with the
75 number of isolates (Fouts et al. 2012; Zhao et al. 2012). LS-BSR introduced a pre-clustering
76 step which substantially reduced the number of BLAST comparisons, enabling it to be feasibly
77 run on thousands of samples (Sahl et al. 2014). More recently, the Roary pipeline has rapidly
78 gained popularity for scalable, user-friendly, pan-genome characterisation (Page et al. 2015).

79
80 The concept of the pan-genome, as described above, places an exclusive emphasis on genes;
81 or, more specifically, open reading frames with the potential to encode proteins. This gene-
82 centric perspective has both shaped, and been shaped by, the bioinformatics tools developed to
83 interrogate the pan-genome. For example, Roary works by taking individual protein-coding
84 sequences, pre-defined using Prokka annotation (Seemann 2014), and assigning each to a
85 single cluster of homologous sequences. This approach thus excludes non protein-coding
86 intergenic regions (IGRs) which typically account for approximately 15% of the genome
87 (Ochman and Caro-Quintero 2016; McCutcheon and Moran 2011). This is clearly problematic
88 for downstream attempts to identify genotype-phenotype links, as IGRs contain many important
89 regulatory elements including, but not limited to, promoters, terminators, non-coding RNAs, and
90 regulatory binding sites. Moreover, we have recently shown that IGRs are subject to purifying
91 selection in the core-genomes of diverse bacterial species, even when known major regulatory
92 elements are excluded (Thorpe et al. 2017; Molina and Van Nimwegen 2008), and a recent
93 study has shown that intergenic variation is positively selected during *Pseudomonas aeruginosa*
94 infections (Khademi and Jelsbak 2017).

95
96 Given that variation in IGRs can have profound phenotypic consequences, it is timely to
97 consider how best to incorporate these sequences into pan-genome analyses. A key question is
98 the degree to which protein-coding genes, and their cognate regulatory elements, should be
99 considered a single “unit”, both selectively (in terms of co-adaptation) and in terms of physical

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

100 linkage on the chromosome. If physical linkage is assumed to be highly robust, such that genes
101 are mostly transferred along with their cognate IGRs, then in principle the definition of a “gene”
102 could be expanded to include the upstream regulatory regions. On the other hand, if there is
103 moderate or weak linkage between genes and IGRs, such that IGRs can occasionally transfer
104 independently, then the purview of the pan-genome could be expanded to include the full
105 complement of IGR alleles in addition to protein-coding sequences.

106
107 Consistent with the second model, which allows for independent transfer of IGRs, a landmark
108 study demonstrated that *E. coli* genes can apparently be regulated by alternative IGRs that
109 frequently share no sequence similarity to each other (Oren et al. 2014). Moreover, the
110 distribution of these IGRs was incongruent with gene trees, suggesting that recombination can
111 act to replace one IGR with another resulting in regulatory “switches”; a process they call
112 horizontal regulatory transfer (HRT) (Oren et al. 2014). It is important to note here that the term
113 “switching” refers only to the replacement of an IGR by a non-homologous or highly divergent
114 variant sequence. It does not specify that the replacement IGR has a particular origin, and could
115 therefore correspond to a transfer from elsewhere in the same genome, or from another isolate.
116 It was also noted that conserved flanking genes may facilitate this process by providing
117 localised regions of homology. IGR switches can be accompanied by differential gene
118 expression (Oren et al. 2014), and may provide a mechanism to offset the fitness costs of
119 harbouring plasmids and other MGEs (McNally et al. 2016), pointing to a central role for this
120 process in adaptation.

121
122 Our current understanding of the evolutionary dynamics of IGRs in the context of bacterial pan-
123 genome leave many open questions. Specifically, it is unclear how IGRs are distributed among
124 isolates within bacterial populations, how commonly IGRs and their cognate genes are co-
125 transferred, or how the frequency of HRT relates to different functional gene categories. A more
126 complete understanding of bacterial adaptation clearly requires a careful consideration of gene
127 presence/absence alongside gene regulation. Here we address this by introducing a new
128 pipeline called Piggy which closely emulates and complements the established pan-genome
129 analysis pipeline Roary (Page et al. 2015). Input and output files for Piggy and Roary use the
130 same format, and run in a similar time on modest computing resources. Piggy provides a means
131 to rapidly identify IGR switches, and more broadly the means to examine the role of horizontal
132 transfer in shaping the bacterial regulome. We demonstrate the utility of Piggy using large

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

133 genome datasets for single lineages within two bacterial species, both of which are of high
134 public health importance; *Staphylococcus aureus* and *Escherichia coli*. Conventional pan-
135 genome analyses are applied to analyse and compare core and accessory IGRs/genes in these
136 lineages. In *S. aureus* we show an association between IGR switching and changes in gene
137 expression, and demonstrate proof-of-principle by establishing a multilocus IGR scheme,
138 (igMLST) in BIGSdb (Jolley and Maiden 2010). Piggy is available at ([https://github.com/harry-
139 thorpe/piggy](https://github.com/harry-thorpe/piggy)) under the GPLv3 licence.

141 **Methods**

142 **Datasets**

143 The *S. aureus* dataset was assembled from published genome sequences (Reuter et al. 2015)
144 available at <http://www.ebi.ac.uk/ena> (study number ERP001012). The *S. aureus* RNA-seq data
145 was previously published (Warne et al. 2016), and is available at (<http://www.ebi.ac.uk/ena>,
146 study number ERP009279). This was supplemented with the corresponding reference
147 genomes, HO_5096_0412: HE681097, MRSA252: BX571856, Newman: AP009351, S0385:
148 AM990992, available at (www.ncbi.nlm.nih.gov). The *E. coli* ST131 dataset was also from a
149 previously published study (McNally et al. 2016), and is available at
150 (<http://datadryad.org/resource/doi:10.5061/dryad.d7d71>). All complete genomes and assemblies
151 were annotated with Prokka (Seemann 2014).

153 **Roary and Piggy parameter settings**

154 Roary (Page et al. 2015) was run using default parameters except for the following: -e -n (to
155 produce alignments with MAFFT (Katoh and Standley 2013)); -i 90 (lower amino acid identity
156 than the default); -s (to keep paralogs together); -z (to keep intermediate files). Piggy was run
157 using default parameters except for --len_id, which controls the percentage of IGR sequences
158 which must share similarity in order to be clustered together. For the *S. aureus* and *E. coli*
159 ST131 datasets, Piggy was run twice, once with --len_id 10 and once with --len_id 90. The
160 former was used for the pan-genome comparisons between genes and IGRs (Fig 2) in order to
161 be comparable with Roary. Using a low length identity (--len_id 10) enabled homologous
162 sequences of varying lengths (for example a truncated sequence) to cluster together. Roary
163 does not provide a similar setting, and only requires that sequences have a minimum length of
164 120 bp. Genes in the same clusters defined by Roary may vary considerably in length, either
165 due to genuine truncations or assembly errors. A relaxed --len_id setting of 10 was therefore

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

166 used in Piggy to provide consistency with Roary and to ensure that homologous IGRs are not
167 erroneously placed in different clusters. A --len_id setting of 90 was subsequently used
168 whenever “switched” IGRs were detected, as this enabled sequences to be subsequently
169 filtered by either nucleotide or length identity.

171 **RNA-seq analysis**

172 Two biological replicates for each isolate were analysed. Kallisto (Bray et al. 2016) was used to
173 quantify transcripts (--kmer-size 31 and --bootstrap-samples 100), and Sleuth (Pimentel et al.
174 2017) was used to normalise and filter the counts produced by Kallisto. These counts were then
175 log₁₀ transformed, and major axis (MA) regression was performed. Rockhopper2 (Tjaden 2015)
176 was used to produce an operon map for each strain by grouping adjacent genes with similar
177 expression profiles together into operons.

179 **Clustering performance**

180 We examined the clustering performance of Piggy by producing truncated variants of IGRs of
181 lengths 10, 15, 20, 30, 50 bp, and comparing how the lengths of the IGRs altered the resulting
182 clustering. The IGRs were truncated from a random starting point in the sequence, and each
183 length was analysed separately. From the starting pool of IGRs from 10 randomly selected
184 isolates, 1000 IGRs were chosen and truncated. These truncated variants were then added to
185 the pool of IGRs and Piggy was run on them. Clustering patterns based on the truncated and
186 original IGRs were then compared, with truncated IGRs placed in the same cluster as their
187 progenitor sequences being assigned as correctly clustered. This analysis was performed on
188 both the *S. aureus* ST22 and *E. coli* ST131 datasets.

190 **Statistical analysis**

191 All statistical analysis was performed within R version 3.3.2 (<https://www.r-project.org>). All
192 plotting was performed with ggplot2 (Wickham 2009).

194 **Results**

195 **Overview of the Piggy pipeline**

196 Fig 1a shows an overview of the Piggy pipeline. The first step is to run Roary, as the gene
197 presence absence output file from Roary is used as an input for Piggy. Piggy is then run using
198 the same annotated assemblies as Roary, specifically GFF3 format files such as those

1
2
3
4 199 produced by Prokka (Seemann 2014). Piggy extracts intergenic sequences (IGRs) from these
5
6 200 files, and uses the flanking gene names and their orientations to name the IGRs (Fig 1b).
7
8 201
9 202 Each IGR name contains three pieces of information: the upstream gene, the downstream gene,
10
11 203 and their relative orientations (CO - co-oriented, DP - double promoter, DT - double terminator).
12
13 204 For example, the IGR “Gene_1 Gene_2 DP” is flanked by Gene_1 and Gene_2, which are both
14
15 205 downstream of the IGR (i.e. they are transcribed in opposite directions). IGRs at the edge of
16
17 206 contigs are excluded by default, but when they are included (using the --edges flag) the missing
18
19 207 information is denoted by NA, for example “Gene_1 NA NA”. Including the gene neighbourhood
20
21 208 information gives context to the IGR and enables identification of “switched” IGRs. By default,
22
23 209 only IGRs between 30-1000 bp in length are included by Piggy, though these lengths can be
24
25 210 user-defined using the --size flag (minimum length = 30 bp). The IGRs are then clustered with
26
27 211 CD-HIT (Fu et al. 2012) at user-defined identity thresholds (--nuc_id - nucleotide identity, --
28
29 212 len_id - length identity). The nucleotide identity is defined as SNPs / aligned sites, and the
30
31 213 length identity is defined as shared sites / alignment length. These two flags allow the user to
32
33 214 set the level of stringency for clustering. For example, a conservative approach is to set high
34
35 215 values for both nucleotide and length identity such that IGRs must be similar in both nucleotide
36
37 216 and length identity to cluster together. By relaxing the length identify whilst maintaining a high
38
39 217 nucleotide identity threshold, highly related sequences still cluster even if one is truncated. The
40
41 218 longest sequence from each cluster is then used to perform an all-vs-all BLASTN search
42
43 219 (Camacho et al. 2009). This is used to merge similar clusters (BLASTN defaults, except -
44
45 220 word_size = 10), which did not cluster with CD-HIT. These clusters are then used to produce an
46
47 221 IGR presence absence matrix (“IGR_presence_absence.csv”), in the same format as the gene
48
49 222 presence absence matrix (“gene_presence_absence.csv”) produced by Roary. Up until this
50
51 223 point, the pipeline is very similar to Roary (Page et al. 2015).
52
53 224

54 225 **Switched IGR detection**

55 226 Piggy identifies “switched” IGRs using two methods. For both methods, the term “switch” refers
56
57 227 to two or more divergent IGR sequences occupying the same locus as defined by flanking
58
59 228 genes, but does not specify an origin for the divergent IGR sequences (Oren et al. 2014). The
60
61 229 first method identifies adjacent genes on the same contig (gene-pairs), and searches for IGR
62
63 230 clusters which lie between these gene-pairs (Fig 1c). Instances where multiple IGR clusters
64
65 231 correspond to the same gene-pair are identified as candidate switched IGRs. The second

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

232 method identifies instances where multiple IGR clusters occupy a locus upstream of a single
233 gene cluster. This is a less conservative approach as only one of the two genes flanking the
234 IGR is taken into account, (Fig 1c). The gene-pair method is used by default as it controls
235 against detecting “switching” (recombination) events that encompass more than a single IGR,
236 for example, cases where a mobile element has inserted between two genes. However such
237 cases remain relevant as the regulation of the downstream gene may still be affected.

238
239 To ensure that differences in gene annotation between isolates, specifically artifactual variation
240 in the start and end points of each gene, are not erroneously assigned as switching events, the
241 first and last 30 bp of each flanking gene are searched against the IGRs with BLASTN. Any
242 matches from these searches indicate differences in annotation of gene borders (rather than
243 genuine differences between the IGRs), and these sequences are disregarded. In order to
244 confirm that they represent genuine switching events, candidate switched IGRs are searched
245 against each other with BLASTN with low complexity filtering turned off (-dust no). If there is no
246 significant match they are classed as “switched”, and if there is a significant match they are
247 aligned using MAFFT (Kato and Standley 2013). The resulting alignment is then used to
248 calculate nucleotide identity (SNPs / shared sites), and length identity (number of shared sites /
249 alignment length). These values can then be used to define an appropriate threshold to identify
250 “switched” IGRs. To aid this, Piggy calculates within-cluster divergences for both genes and
251 IGRs, and these divergences can be used to calibrate Piggy with Roary.

252

253 **Clustering performance**

254 The shorter lengths of IGRs compared with genes poses potential problems for alignment
255 accuracy. We tested the clustering performance of Piggy by producing truncated variants of
256 IGRs, adding these to the total complement of IGRs in an analysis, and then recording whether
257 the truncated IGRs were clustered with their untruncated counterparts (Methods). For *S. aureus*
258 ST22, 82% of IGRs truncated to 10 bp clustered together with the corresponding full length
259 sequences, but this figure increased to > 99% when the length of the truncated sequences was
260 20 -bp. (Fig S1a). A similar increase was observed for the *E. coli* ST131 data, although in this
261 case 50 bp was required for the percentage of correct assignments to be > 99%. (Fig S1b).

262
263 An inspection of the incorrectly clustered sequences from both datasets revealed that their
264 progenitor sequences shared high sequence similarity in parts of their sequence to other IGR

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

265 clusters, but no sequence similarity in other parts of the sequence. This resulted in separate
266 clusters which shared high sequence homology over parts of their sequences. When these
267 sequences were truncated to assess the clustering, if the truncated part of the sequence was
268 selected, then it could align to either of these IGR clusters. In many cases these alignments
269 were perfect matches, and so the IGR could not be unambiguously placed. This problem is
270 likely to be a result of non-homologous breaks at the edge of HGT events, and this is consistent
271 with greater clustering accuracy in *S. aureus* ST22 compared with *E. coli* ST131, where the
272 latter has a much larger pan-genome.

273

274 ***Staphylococcus aureus***

275 *S. aureus* is an important skin-associated bacterium which is commonly carried
276 asymptotically, but can also cause a wide range of infections from minor skin infections to
277 fatal bacteraemias. It has a clonal population structure consisting of discrete lineages (Feil et al.
278 2003). Although the core genome is relatively stable, phenotypic variation (e.g. resistance
279 profiles, virulence traits, and host preference) is associated with a more dynamic accessory
280 genome and the horizontal transfer of MGEs, such as the *SCCmec* element which confers
281 resistance to β -lactam antibiotics (Lindsay and Holden 2004).

282

283 *S. aureus* ST22 (EMRSA-15) is a clinically important hospital-acquired methicillin resistant strain
284 which is common in the UK and is rapidly expanding elsewhere in Europe and globally (Holden
285 et al. 2013). Previous work has shown that *S. aureus* ST22 is clonal and shows relatively little
286 variation in gene content (Holden et al. 2013; Reuter et al. 2015). In order to compare the pan-
287 genomes of *S. aureus* at different scales, we analysed a diverse dataset of 1552 isolates from
288 many lineages, and a smaller dataset of 500 ST22 isolates subsampled from the larger dataset
289 (Reuter et al. 2015). The size of the gene and IGR pan and core-genomes were compared by
290 analysing both datasets with Roary and Piggy. Frequency histograms were plotted for both
291 genes and IGRs (Fig 2a-b).

292

293 The gene-IGR frequency histogram for ST22 (Fig 2a) shows that there are 2,409 core genes
294 and 1,556 core IGRs, where core is defined as gene presence in > 95% of isolates (Table 1).
295 When the whole species is considered, these numbers drop to 2,129 and 1,134, respectively.
296 The fact that there are fewer core IGRs than core genes is in part due to the exclusion of IGRs
297 < 30 bp (many of which are intra-operonic), but also likely reflects faster evolution of IGRs. Both

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

298 distributions conform to the U-shape typically found in such analyses where the majority of
299 genes/IGRs are either very common or very rare, however the distribution of genes and IGRs is
300 shifted towards the rare sequences when the whole species is considered rather than only
301 ST22.

302
303 We used the output of Piggy to investigate the degree of linkage between genes and IGRs. We
304 identified all genomic loci consisting of an IGR flanked by two genes, and from these we
305 identified all pairs of genes and IGRs where the IGR was upstream of the gene start. We then
306 grouped these according to whether the gene or IGR was core or accessory (Table 2). For the
307 *S. aureus* ST22 data, 99.5% of core genes were immediately downstream of a core IGR, and
308 92.9% of the accessory genes were similarly downstream of an accessory IGR. When
309 considering the wider *S. aureus* dataset the figures were similar; 92.6% of core genes were
310 downstream of a core IGR, and 96.8% of accessory genes were downstream of an accessory
311 IGR. Thus, the assignment of an IGR as core or accessory is highly predictive of the
312 corresponding assignment of the cognate downstream gene, which in turn points to strong
313 background linkage between genes in IGRs in the genome.

314
315 ***Escherichia coli* ST131**

316 The utility of Piggy was further validated by re-analysing data from a recent study on the
317 widespread and clinically important *E. coli* lineage ST131 (McNally et al. 2016). This dataset
318 contains 236 clinical *E. coli* ST131 isolates from human, domesticated animal, and avian hosts.
319 *E. coli* is a more genetically diverse species than *S. aureus*, and unsurprisingly *E. coli* ST131
320 has a larger pan-genome than *S. aureus* ST22, with 12,806 genes and 16,429 IGRs (Fig 2c,
321 Table 1). More surprisingly, *E. coli* ST131 has a larger pan-genome than the whole *S. aureus*
322 species. Within *E. coli* ST131, 3,930 genes and 2,296 IGRs were core out of an average of
323 4,689 genes and 2,984 IGRs per isolate. Thus despite the differences between the two species
324 in their level of diversity there was a consistent signal of a lower number of core IGRs than core
325 genes, and a high number of accessory IGRs compared to accessory genes. There was tight
326 linkage between genes and IGRs, with 97.9% of core genes being immediately downstream of
327 core IGRs and 97.3% of accessory genes being similarly downstream of accessory IGRs; these
328 results are consistent with those from *S. aureus* (Table 2).

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

330 The data from *S. aureus* and *E. coli* shows a background of strong linkage between genes and
331 IGRs. However, this linkage is not perfect; some core genes are associated with accessory
332 IGRs (and vice-versa), and the linkage is weaker over long timescales (across the whole *S.*
333 *aureus* species compared to within ST22). Previous work has examined this linkage and found
334 evidence of widespread IGR regulatory switching, where genes are regulated by alternative
335 IGRs in different isolates (Oren et al. 2014). Piggy provides a list of candidate switching events
336 together for both “gene-pair” and “upstream” approaches (see Methods) at different thresholds
337 of nucleotide identity. For the *E. coli* ST131 data, the pipeline detected 61 cases of putative IGR
338 switching using the most conservative settings (i.e. the conservative gene-pair method, and the
339 alternative IGRs showing no sequence similarity by BLASTN). Relaxing the threshold of
340 sequence identity to < 90% resulted in the identification of an additional 317 candidate switching
341 events, though these possibly reflect either relaxed or positive selection.

342

343 **Switched IGRs influence gene expression in *S. aureus***

344 To examine whether switches in IGRs affect the expression of cognate (downstream) genes, we
345 used a previously published RNA-seq dataset based on four reference *S. aureus* isolates
346 HO_5096_0412 (ST22), Newman (CC8), MRSA252 (CC36), and S0385 (CC398) (Warne et al.
347 2016). Each of these *S. aureus* references isolate represents a distinct major clonal complex,
348 and all were grown under identical conditions with each experiment being replicated. Thus these
349 data provide evidence of the natural variation in gene expression within the *S. aureus*
350 population. By analysing these data alongside the output from Piggy, it is possible to test the
351 extent to which IGR switches between these four genomes can account for the observed
352 variation in gene expression between clonal complexes. First Roary was used to identify a set of
353 2094 single copy core genes present in all four isolates, and then expression of these core
354 genes was quantified using Kallisto (Bray et al. 2016). To do this we used RNA-seq data for two
355 replicates for each of the four reference genomes. The tpm (Transcripts per Kilobase Million)
356 values for each gene are given in Table S1. We then used Sleuth (Pimentel et al. 2017) to
357 normalise and filter these counts.

358

359 To check the consistency of the data between biological replicates, we first plotted two
360 replicates for each isolate against each other (e.g. Newman replicate 1 vs Newman replicate 2)
361 (Fig 3). These plots were tightly correlated (mean $R^2 = 0.98$), confirming that the expression
362 values for individual genes were consistent between replicates. We then plotted between-isolate

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

363 comparisons, again using both replicates for each genome (e.g. Newman replicate 1 vs
364 MRSA252 replicate 1, and Newman replicate 2 vs MRSA252 replicate 2) (Fig 3). These
365 comparisons revealed considerably more scatter, with R^2 values ranging from 0.76 to 0.9. Given
366 the extremely high R^2 values for within-isolate comparisons, the decrease in R^2 for between-
367 isolate comparisons reflects genuine differences in expression between the isolates. We note
368 that a small number of genes show very striking differences in expression between the clonal
369 complexes. For example, the expression of *mepA*, which encodes a multidrug efflux pump, was
370 ~250 fold higher in Newman compared with the other isolates.

371
372 The genomes of each pair of isolates were analysed using Roary and Piggy to identify switched
373 IGRs with a nucleotide identity threshold of < 90% for IGR clusters. For each pair of isolates, we
374 then identified all genes immediately downstream of a switched IGR. As a single switched IGR
375 might impact on the expression of more than one co-transcribed downstream genes we also
376 considered all genes linked in a single operon that could be impacted by a single switching
377 event upstream affecting a shared promoter. For each pair of isolates, we thus identified all core
378 genes putatively affected by upstream IGR switches. We then tested whether these genes
379 showed a higher degree of differential expression by conducting Monte Carlo permutation tests
380 on the residuals from the regressions (Fig 3). For each pairwise comparison of isolates, we
381 summed the residuals of the genes with switched IGRs (shown as red points in Fig 3), and
382 compared this to a distribution obtained by resampling (without replacement) 100,000 random
383 sets of the same number of genes and summing their residuals. We computed a one-tailed p-
384 value by dividing the number of permutations with summed residuals greater than the observed
385 value by 100,000 (Fig 3). Because we used both replicates separately (e.g. Newman replicate 1
386 vs S0385 replicate 1, and Newman replicate 2 vs S0385 replicate 2), each comparison between
387 pairs of isolates was tested twice. In 9/12 pairwise comparisons, the observed residuals of the
388 genes downstream of switched IGRs were significantly ($p < 0.05$) greater than expected from
389 the resampled data, indicating that genes with switched IGRs were more differentially
390 expressed than those without. Of the three remaining comparisons, two corresponded to
391 comparisons between HO_5096_0412 and S0385 ($p = 0.17$, and $p = 0.055$), and one between
392 HO_5096_0412 and Newman ($p = 0.054$). The second comparison between HO_5096_0412
393 and Newman was the most weakly significant result ($p = 0.025$). Thus, the two replicates for
394 each individual pairwise comparison were largely concordant with each other.

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

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

402

403 **Compatibility and scalability**

404 We have so far demonstrated that Piggy can be used to analyse the intergenic component of
405 the pan-genome and identify IGR switches, and shown that these switches have biological
406 relevance with respect to gene expression. Importantly, Piggy is designed such that the output
407 files are compatible with existing software and databases. The “IGR_presence_absence.csv”
408 file has an identical format to the “gene_presence_absence.csv” file produced by Roary, and
409 can be loaded directly into the interactive browser-based viewer phandango (Hadfield et al.
410 2017) (Fig S2). It can also be used as input, along with a traits file, to Scoary (Brynildsrud et al.
411 2016) to test for associations between IGRs and phenotypic traits. Moreover, the
412 “representative_clusters_merged.fasta” file can be loaded directly into BIGSdb (Jolley and
413 Maiden 2010) to create an allele scheme for IGRs. In order to provide proof-of-principle, we
414 created a multilocus IGR (igMLST) scheme in BIGSdb. Briefly, 2631 unique IGR sequences
415 with length ≥ 30 bp, from 7 *S. aureus* reference genomes, were entered into the database locus
416 list. Using functionality within the database, these sequences were grouped as a searchable
417 scheme (S_aureus_Intergenic_PIGGY), comparable to MLST, rMLST and wgMLST schemes
418 (Maiden et al. 2013; Jolley et al. 2012; Sheppard, Jolley, and Maiden 2012). The distribution of
419 IGRs was analysed for all isolates in the database, identifying IGRs as present in the respective
420 genome if a hit was recorded with nucleotide identity $\geq 70\%$ over $\geq 50\%$ of the sequence using a
421 BLAST word size of 7 bp. The scheme can be found at <https://sheppardlab.com/resources/>.
422 Although we do not expect a typing scheme based solely on IGRs to be widely used,
423 supplementing protein-coding regions with IGR alleles may provide additional information
424 regarding links between genotype and phenotype, as well as increased epidemiological and
425 phylogenetic resolution.

426

427 **Discussion**

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

428 Whole-genome sequence datasets consisting of hundreds or even thousands of bacterial
429 isolates have revealed pan-genomes of many thousands of genes and large differences in gene
430 content between isolates of the same species. Currently, pan-genome diversity is considered
431 almost exclusively in terms of protein-coding genes, despite overwhelming evidence that
432 variation within IGRs impacts on phenotypes. Here we address this by introducing Piggy, a
433 pipeline specifically designed to incorporate IGRs into routine pan-genome analyses by working
434 in close conjunction with Roary (Page et al. 2015).

435
436 The utility of this approach is demonstrated using large datasets of *S. aureus* and *E. coli* ST131.
437 Consistent with previous analyses of protein-coding regions (Holden et al. 2013; McNally et al.
438 2016), the IGR component of the ST131 pan-genome (the “panIGRome”) is considerably larger
439 than that for *S. aureus* ST22, and surprisingly is also larger than the pan-genome of the whole
440 *S. aureus* species. There was more diversity within IGRs than genes in both species. While
441 some IGRs may be essential for expression of multiple genes, IGRs are broadly subject to
442 weaker purifying selection than protein coding genes (Thorpe et al. 2017). The maintenance of
443 core IGRs in both bacterial genome datasets is consistent with selection acting to conserve
444 them and allows alignment and analysis in much the same way as protein-coding regions.

445
446 The current exclusion of IGRs from routine pan-genome or cgMLST analyses may in part reflect
447 perceived difficulties in the alignment and subsequent cluster definition, particularly if the
448 sequences are very short. We therefore validated the pipeline by investigating clustering
449 accuracy as a function of sequence length by truncating the IGR sequences and recording
450 whether they remained in the same cluster as their full-length counterparts. For *S. aureus*, the
451 data showed that truncated IGRs > 20 bp almost always remained in the original cluster,
452 confirming that the minimum length permitted in the pipeline of 30-bp is conservative. For *E.*
453 *coli*, truncating the sequences had greater impact on cluster assignments, and a minimum
454 length of 50 bp would be a safer setting in this case. The problems with clustering shorter
455 sequences in *E. coli*, compared to *S. aureus*, are not due to the length of the sequence *per se*
456 but reflect the higher rate of recombination in this species. This means that the IGRs are more
457 likely to be chimeric in structure, with localised regions within the IGRs showing a high level of
458 homology to different clusters. This leads to cluster assignment being dependant not so much
459 on length, but on which part of the truncated sequence happened to be retained.

460

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

461 Variation within regulatory elements located within IGRs can impact on the expression of the
462 downstream gene (Oren et al. 2014). Piggy (alongside Roary) provides the means to combine
463 information on genes and their cognate IGRs thus facilitating the detection of “switched” IGRs
464 and downstream genes that are potentially affected. We have shown that in *S. aureus*, genes
465 with switched upstream IGRs show a higher degree of differential expression than those
466 without. This is consistent with previous work on *E. coli* (Oren et al. 2014), and suggests that the
467 identification of IGR switches using Piggy can provide a useful indication of differential gene
468 expression, even in the absence of RNA-seq data. However, we note that high divergence
469 within IGRs does not necessarily imply selection for differential gene expression, and may
470 instead simply reflect weaker selective constraints. A clear direction for future work is to make
471 constructs consisting of genes with alternative IGRs, in order to directly measure the effect of
472 natural IGR variants on gene expression. Similar experiments have previously been performed
473 in *E. coli* based on variation within promoters (Shimada et al. 2014), and IGRs more broadly
474 (Oren et al. 2014). The importance of changes in gene expression mediated by intergenic
475 variation as a route of adaptation is currently unknown, but one recent study suggested that
476 intergenic changes are strongly positively selected in *Pseudomonas aeruginosa* during infection
477 in patients with cystic fibrosis, and more work is required to test the generality of these findings
478 (Khademi and Jelsbak 2017).

479
480 **Conclusions**

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

492
493 **Declarations**

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

494 Ethics approval and consent to participate: Not applicable

495 Consent for publication: Not applicable

496 Availability of data and material: The *S. aureus* ST22 dataset was assembled from published
497 genome sequences of the clinically important lineage ST22 (EMRSA-15) (Reuter et al. 2015)
498 available at <http://www.ebi.ac.uk/ena> (study number ERP001012). The *S. aureus* RNA-seq data
499 was previously published (Warne et al. 2016), and is available at (<http://www.ebi.ac.uk/ena>,
500 study number ERP009279). This was supplemented with the corresponding reference
501 genomes, all available at (www.ncbi.nlm.nih.gov), HO_5096_0412: HE681097, MRSA252:
502 BX571856, Newman: AP009351, S0385: AM990992. The *E. coli* ST131 dataset (McNally et al.
503 2016) is available at (<http://datadryad.org/resource/doi:10.5061/dryad.d7d71>).

504 Piggy is available at (<https://github.com/harry-thorpe/piggy>) under the GPLv3 licence.

505 Competing interests: Not applicable

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

516

517 Authors' contributions: HAT designed and implemented the pipeline, and carried out the majority
518 of the analyses, with input from EJF, SCB and SKS. HAT and EJF wrote the manuscript with
519 input from SKS and SCB.

520

521 Acknowledgements: We are very grateful to Torsten Seemann, Andrew Page and João Carriço
522 for encouragement and helpful feedback. We are also grateful to Matt Holden for provision of
523 the *S. aureus* RNA-seq data, to Sandra Reuter for help with the *S. aureus* ST22 data, and to
524 Alan McNally for the *E. coli* ST131 data. This work also greatly benefitted from access to the
525 Medical Research Council funded Cloud Infrastructure for Microbial Bioinformatics (MRC-
526 CLIMB).

1
2
3
4 527

5
6 528 **References**

- 7
8
9 529 Andreani, Nadia Andrea, Elze Hesse, and Michiel Vos. 2017. "Prokaryote Genome Fluidity Is
10 530 Dependent on Effective Population Size." *The ISME Journal* 11 (7):1719–21.
11 531 Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. "Near-Optimal
12 532 Probabilistic RNA-Seq Quantification." *Nature Biotechnology* 34 (5):525–27.
13 533 Brynildsrud, Ola, Jon Bohlin, Lonneke Scheffer, and Vegard Eldholm. 2016. "Rapid Scoring of
14 534 Genes in Microbial Pan-Genome-Wide Association Studies with Scoary." *Genome Biology*
15 535 17 (1):238.
16 536 Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos,
17 537 Kevin Bealer, and Thomas L. Madden. 2009. "BLAST+: Architecture and Applications."
18 538 *BMC Bioinformatics* 10 (December):421.
19 539 Feil, Edward J., Jessica E. Cooper, Hajo Grundmann, D. Ashley Robinson, Mark C. Enright,
20 540 Tony Berendt, Sharon J. Peacock, et al. 2003. "How Clonal Is *Staphylococcus Aureus*?"
21 541 *Journal of Bacteriology* 185 (11):3307–16.
22 542 Fouts, Derrick E., Lauren Brinkac, Erin Beck, Jason Inman, and Granger Sutton. 2012.
23 543 "PanOCT: Automated Clustering of Orthologs Using Conserved Gene Neighborhood for
24 544 Pan-Genomic Analysis of Bacterial Strains and Closely Related Species." *Nucleic Acids*
25 545 *Research* 40 (22):e172.
26 546 Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. "CD-HIT: Accelerated
27 547 for Clustering the next-Generation Sequencing Data." *Bioinformatics* 28 (23):3150–52.
28 548 Hadfield, James, Nicholas J. Croucher, Richard J. Goater, Khalil Abudahab, David M.
29 549 Aanensen, and Simon R. Harris. 2017. "Phandango: An Interactive Viewer for Bacterial
30 550 Population Genomics." *bioRxiv*. <https://doi.org/10.1101/119545>.
31 551 Holden, Matthew T. G., Li-Yang Hsu, Kevin Kurt, Lucy A. Weinert, Alison E. Mather, Simon R.
32 552 Harris, Birgit Strommenger, et al. 2013. "A Genomic Portrait of the Emergence, Evolution,
33 553 and Global Spread of a Methicillin-Resistant *Staphylococcus Aureus* Pandemic." *Genome*
34 554 *Research* 23 (4):653–64.
35 555 Holt, Kathryn E., Heiman Wertheim, Ruth N. Zadoks, Stephen Baker, Chris A. Whitehouse,
36 556 David Dance, Adam Jenney, et al. 2015. "Genomic Analysis of Diversity, Population
37 557 Structure, Virulence, and Antimicrobial Resistance in *Klebsiella Pneumoniae*, an Urgent
38 558 Threat to Public Health." *Proceedings of the National Academy of Sciences of the United*
39 559 *States of America* 112 (27):E3574–81.
40 560 Jolley, Keith A., Carly M. Bliss, Julia S. Bennett, Holly B. Bratcher, Carina Brehony, Frances M.
41 561 Colles, Helen Wimalarathna, et al. 2012. "Ribosomal Multilocus Sequence Typing:
42 562 Universal Characterization of Bacteria from Domain to Strain." *Microbiology* 158 (Pt
43 563 4):1005–15.
44 564 Jolley, Keith A., and Martin C. J. Maiden. 2010. "BIGSdb: Scalable Analysis of Bacterial
45 565 Genome Variation at the Population Level." *BMC Bioinformatics* 11 (1):595.
46 566 Katoh, Kazutaka, and Daron M. Standley. 2013. "MAFFT Multiple Sequence Alignment
47 567 Software Version 7: Improvements in Performance and Usability." *Molecular Biology and*
48 568 *Evolution* 30 (4):772–80.
49 569 Khademi, Hossein, and Lars Jelsbak. 2017. "Host Adaptation Mediated by Intergenic Evolution
50 570 in a Bacterial Pathogen." *bioRxiv*. <https://doi.org/10.1101/236000>.
51 571 Lindsay, Jodi A., and Matthew T. G. Holden. 2004. "Staphylococcus Aureus: Superbug, Super
52 572 Genome?" *Trends in Microbiology* 12 (8):378–85.
53 573 Maiden, Martin C. J., Melissa J. Jansen van Rensburg, James E. Bray, Sarah G. Earle,

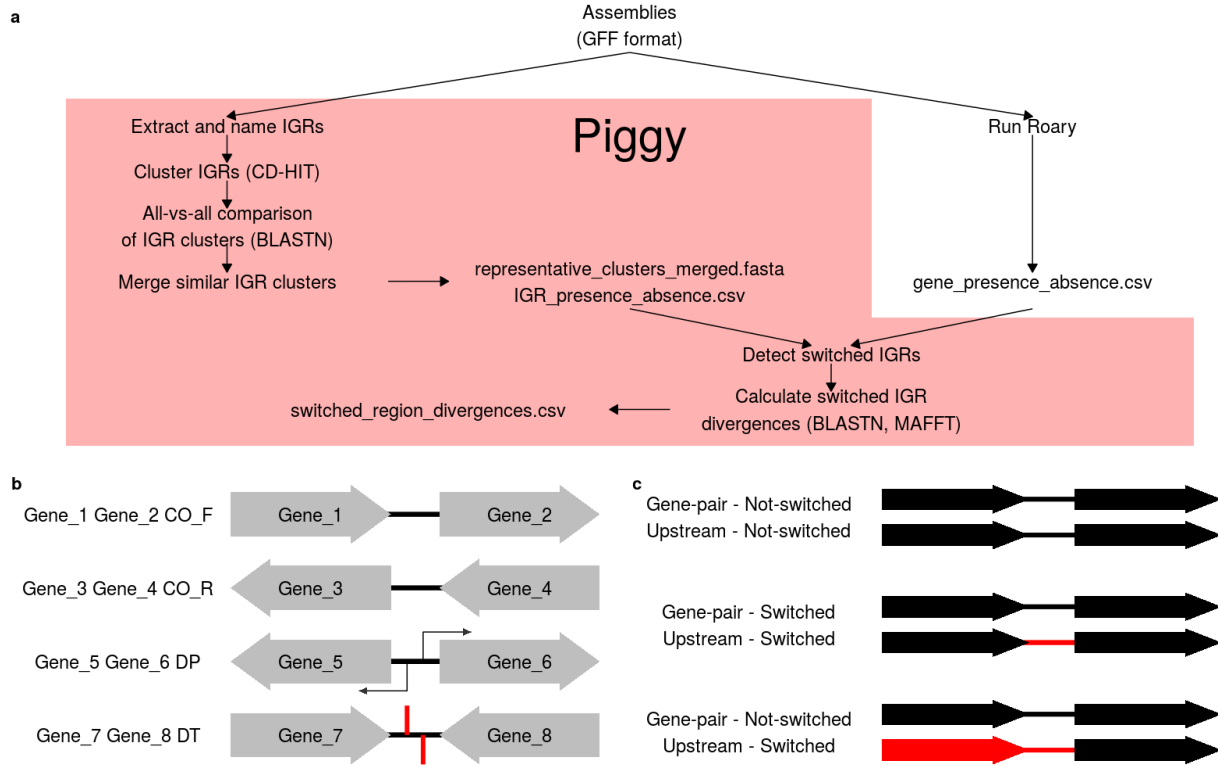
1
2
3
4 574 Suzanne A. Ford, Keith A. Jolley, and Noel D. McCarthy. 2013. "MLST Revisited: The
5 575 Gene-by-Gene Approach to Bacterial Genomics." *Nature Reviews. Microbiology* 11
6 576 (10):728–36.
7 577
8 578 McCutcheon, John P., and Nancy A. Moran. 2011. "Extreme Genome Reduction in Symbiotic
9 579 Bacteria." *Nature Reviews. Microbiology* 10 (1):13–26.
10 580
11 581 McInerney, James O., Alan McNally, and Mary J. O'Connell. 2017. "Why Prokaryotes Have
12 582 Pangenomes." *Nature Microbiology* 2 (March):17040.
13 583
14 584 McNally, Alan, Yaara Oren, Darren Kelly, Ben Pascoe, Steven Dunn, Tristan Sreecharan, Minna
15 585 Vehkala, et al. 2016. "Combined Analysis of Variation in Core, Accessory and Regulatory
16 586 Genome Regions Provides a Super-Resolution View into the Evolution of Bacterial
17 587 Populations." *PLoS Genetics* 12 (9):e1006280.
18 588
19 589 Medini, Duccio, Claudio Donati, Hervé Tettelin, Vega Massignani, and Rino Rappuoli. 2005. "The
20 590 Microbial Pan-Genome." *Current Opinion in Genetics & Development* 15 (6):589–94.
21 591
22 592 Molina, Nacho, and Erik Van Nimwegen. 2008. "Universal Patterns of Purifying Selection at
23 593 Noncoding Positions in Bacteria." *Genome Research* 18 (1):148–60.
24 594
25 595 Ochman, H., and A. Caro-Quintero. 2016. "Genome Size and Structure, Bacterial." In
26 596 *Encyclopedia of Evolutionary Biology*, edited by Richard M. Kliman, 179–85. Oxford:
27 597 Academic Press.
28 598
29 599 Oren, Yaara, Mark B. Smith, Nathan I. Johns, Millie Kaplan Zeevi, Dvora Biran, Eliora Z. Ron,
30 600 Jukka Corander, Harris H. Wang, Eric J. Alm, and Tal Pupko. 2014. "Transfer of Noncoding
31 601 DNA Drives Regulatory Rewiring in Bacteria." *Proceedings of the National Academy of
32 602 Sciences of the United States of America* 111 (45):16112–17.
33 603
34 604 Page, Andrew J., Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew
35 605 T. G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, and Julian Parkhill. 2015.
36 606 "Roary: Rapid Large-Scale Prokaryote Pan Genome Analysis." *Bioinformatics* 31
37 607 (22):3691–93.
38 608
39 609 Pimentel, Harold, Nicolas L. Bray, Suzette Puente, Páll Melsted, and Lior Pachter. 2017.
40 610 "Differential Analysis of RNA-Seq Incorporating Quantification Uncertainty." *Nature
41 611 Methods* 14 (7):687–90.
42 612
43 613 Reuter, Sandra, Estee M. Török, Matthew T. G. Holden, Rosy Reynolds, Kathy E. Raven, Beth
44 614 Blane, Tjibbe Donker, et al. 2015. "Building a Genomic Framework for Prospective MRSA
45 615 Surveillance in the United Kingdom and the Republic of Ireland." *Genome Research*,
46 616 December. <https://doi.org/10.1101/gr.196709.115>.
47 617
48 618 Sahl, Jason W., J. Gregory Caporaso, David A. Rasko, and Paul Keim. 2014. "The Large-Scale
49 619 Blast Score Ratio (LS-BSR) Pipeline: A Method to Rapidly Compare Genetic Content
50 620 between Bacterial Genomes." *PeerJ* 2 (April):e332.
51 621
52 622 Seemann, Torsten. 2014. "Prokka: Rapid Prokaryotic Genome Annotation." *Bioinformatics* 30
53 623 (14):2068–69.
54 624
55 625 Sheppard, Samuel K., Keith A. Jolley, and Martin C. J. Maiden. 2012. "A Gene-by-Gene
56 626 Approach to Bacterial Population Genomics: Whole Genome MLST of *Campylobacter*."
57 627 *Genes* 3 (2):261–77.
58 628
59 629 Shimada, Tomohiro, Yukiko Yamazaki, Kan Tanaka, and Akira Ishihama. 2014. "The Whole Set
60 630 of Constitutive Promoters Recognized by RNA Polymerase RpoD Holoenzyme of
61 631 *Escherichia Coli*." *PloS One* 9 (3):e90447.
62 632
63 633 Thorpe, Harry A., Sion C. Bayliss, Laurence D. Hurst, and Edward J. Feil. 2017. "Comparative
64 634 Analyses of Selection Operating on Non-Translated Intergenic Regions of Diverse Bacterial
65 635 Species." *Genetics*, March. <https://doi.org/10.1534/genetics.116.195784>.
66 636
67 637 Tjaden, Brian. 2015. "De Novo Assembly of Bacterial Transcriptomes from RNA-Seq Data."
68 638 *Genome Biology* 16 (January):1.

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

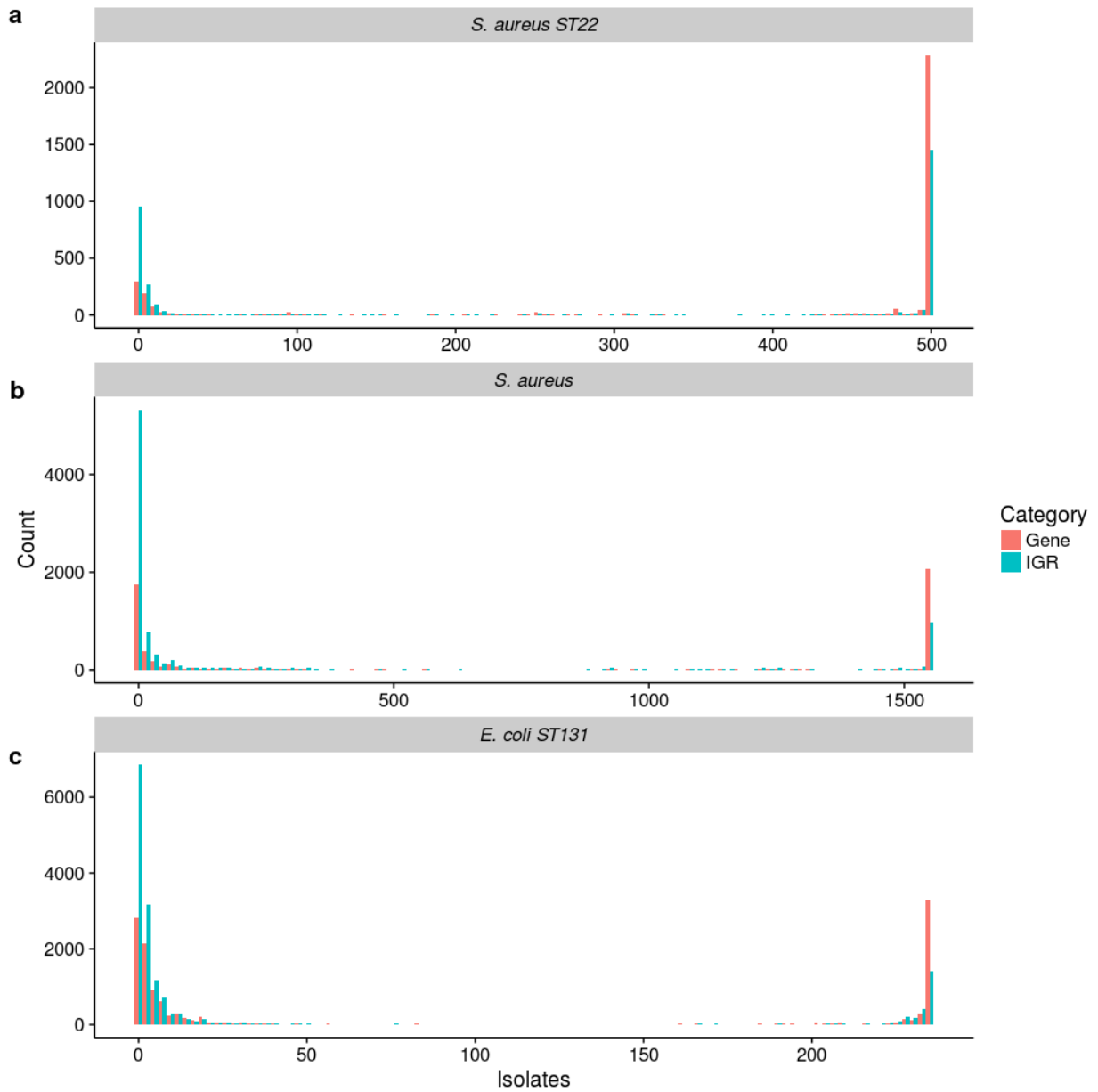
623 Vos, Michiel, Matthijn C. Hesselman, Tim A. te Beek, Mark W. J. van Passel, and Adam Eyre-
624 Walker. 2015. "Rates of Lateral Gene Transfer in Prokaryotes: High but Why?" *Trends in*
625 *Microbiology* 23 (10):598–605.
626 Warne, Ben, Catriona P. Harkins, Simon R. Harris, Alexandra Vatsiou, Nicola Stanley-Wall,
627 Julian Parkhill, Sharon J. Peacock, Tracy Palmer, and Matthew T. G. Holden. 2016. "The
628 Ess/Type VII Secretion System of Staphylococcus Aureus Shows Unexpected Genetic
629 Diversity." *BMC Genomics* 17 (March):222.
630 Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. 2nd ed. Springer
631 Publishing Company, Incorporated.
632 Zhao, Yongbing, Jiayan Wu, Junhui Yang, Shixiang Sun, Jingfa Xiao, and Jun Yu. 2012.
633 "PGAP: Pan-Genomes Analysis Pipeline." *Bioinformatics* 28 (3):416–18.

634
635

636 **Figures**



637
638
639 **Fig 1: An overview of the Piggy pipeline.** a) A schematic to illustrate the Piggy pipeline and
640 how it works alongside Roary. b) IGRs are named according to their flanking genes and their
641 orientations. This naming scheme enables Piggy to link genes with their associated IGRs, and
642 provides information on their orientations. c) A schematic to illustrate the difference between the
643 “gene-pair” and “upstream” methods used to identify candidate switched IGRs.

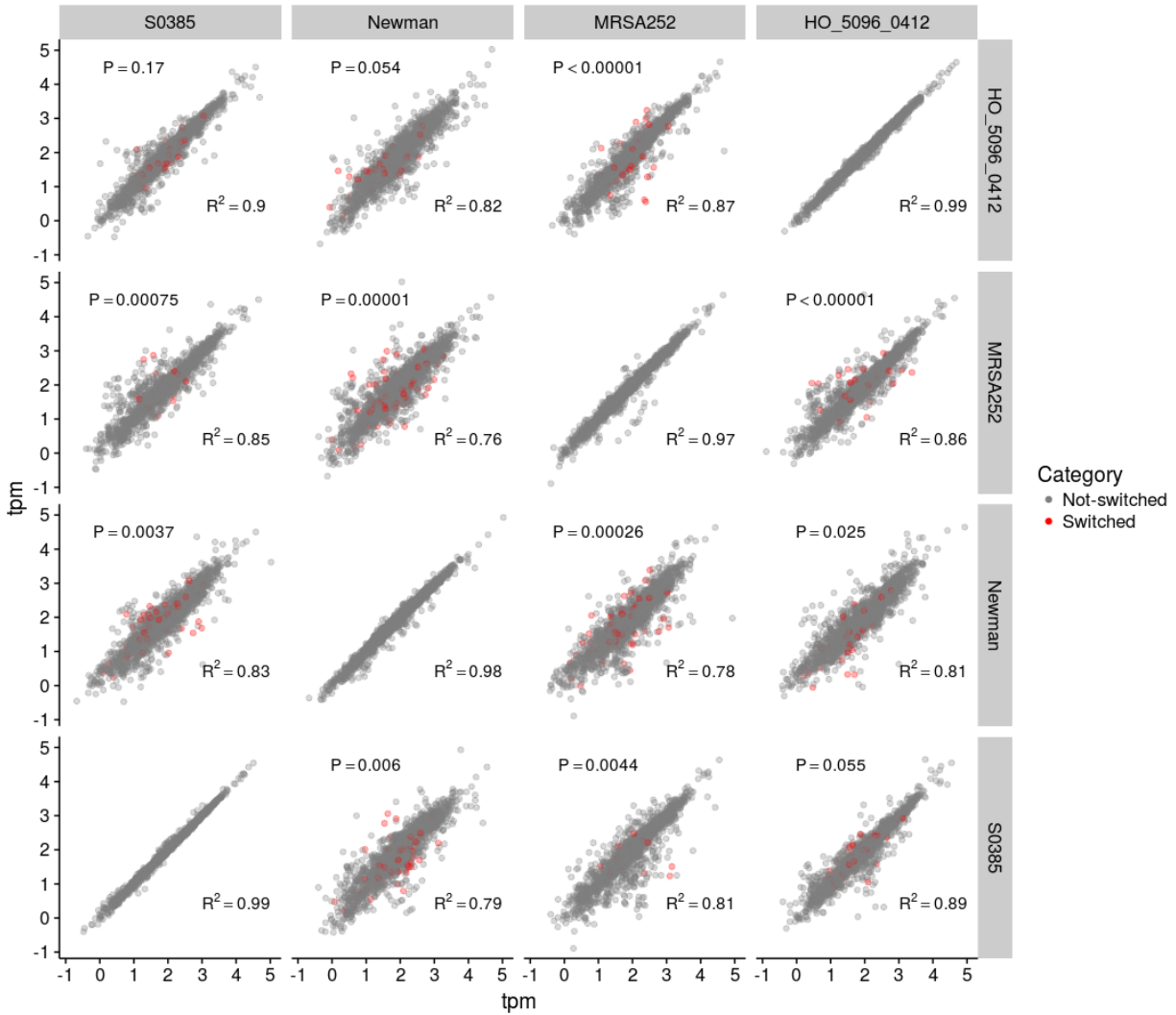


646

647

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

652



653

654

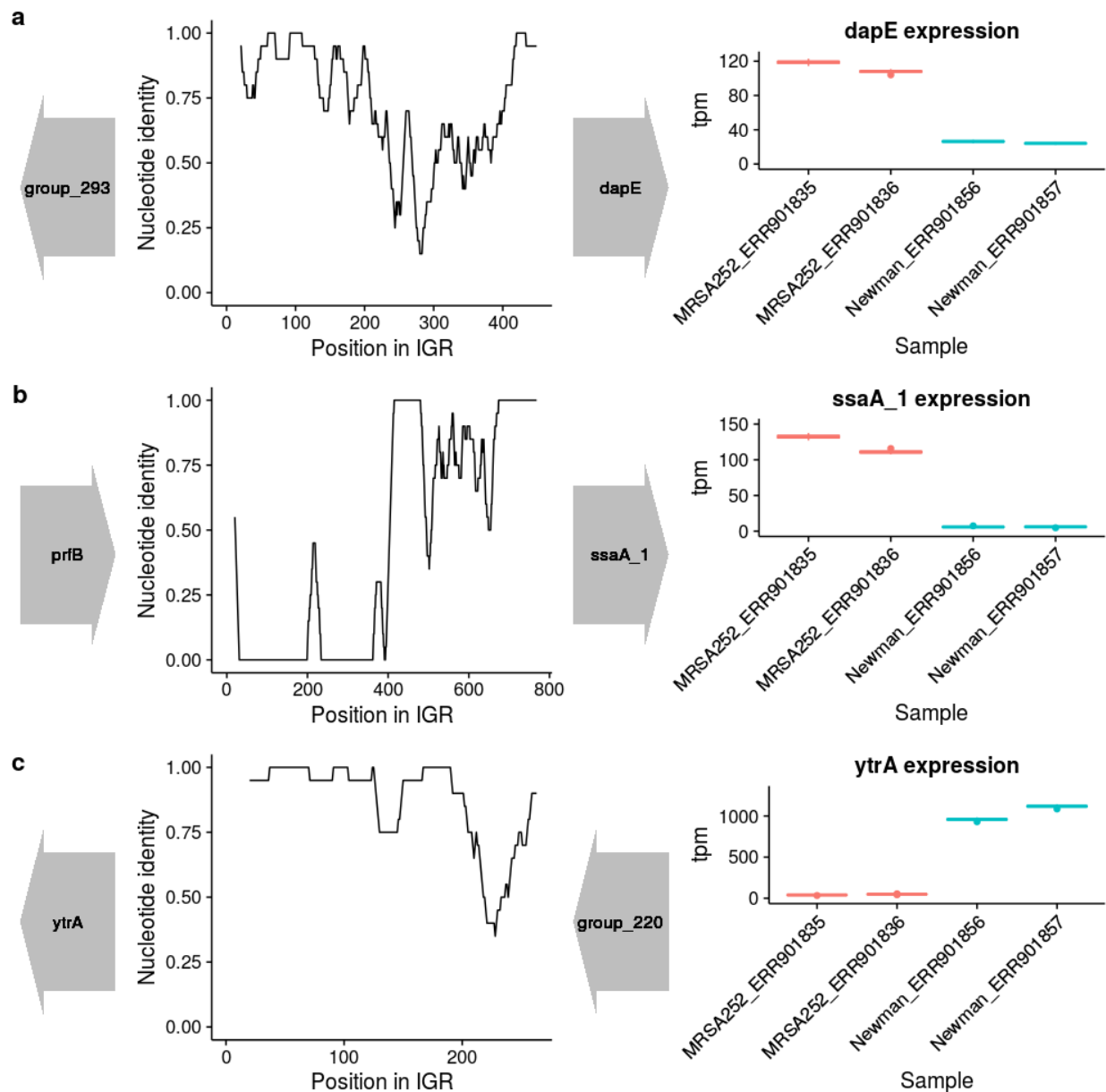
655 **Fig 3: S. aureus gene expression data.** Pairwise RNA-seq comparisons between four *S.*
 656 *aureus* isolates, where two biological replicates were used for each isolate. The top-left of the
 657 diagonal corresponds to comparisons between replicate 1 from different isolates (e.g. SO385
 658 replicate 1 vs HO_5096_0412 replicate 1). The bottom-right of the diagonal corresponds to
 659 comparisons between replicate 2 from different isolates (e.g. SO385 replicate 2 vs
 660 HO_5096_0412 replicate 2). The diagonal corresponds to comparisons between the two
 661 biological replicates from the same isolate. 2094 core genes were analysed in each comparison,

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

662 and tpm (Transcripts per Kilobase Million) was used to quantify expression. The genes were
663 separated into two categories: Switched (red), and Not-switched (grey), based on their
664 upstream IGRs. The R^2 value corresponds to all the genes. The P-value corresponds to a
665 Monte Carlo permutation test comparing the residuals of the two groups of genes, where a
666 significant score indicates that the genes downstream of switch IGRs are associated with a
667 higher degree of differential expression (ie higher residuals).

668

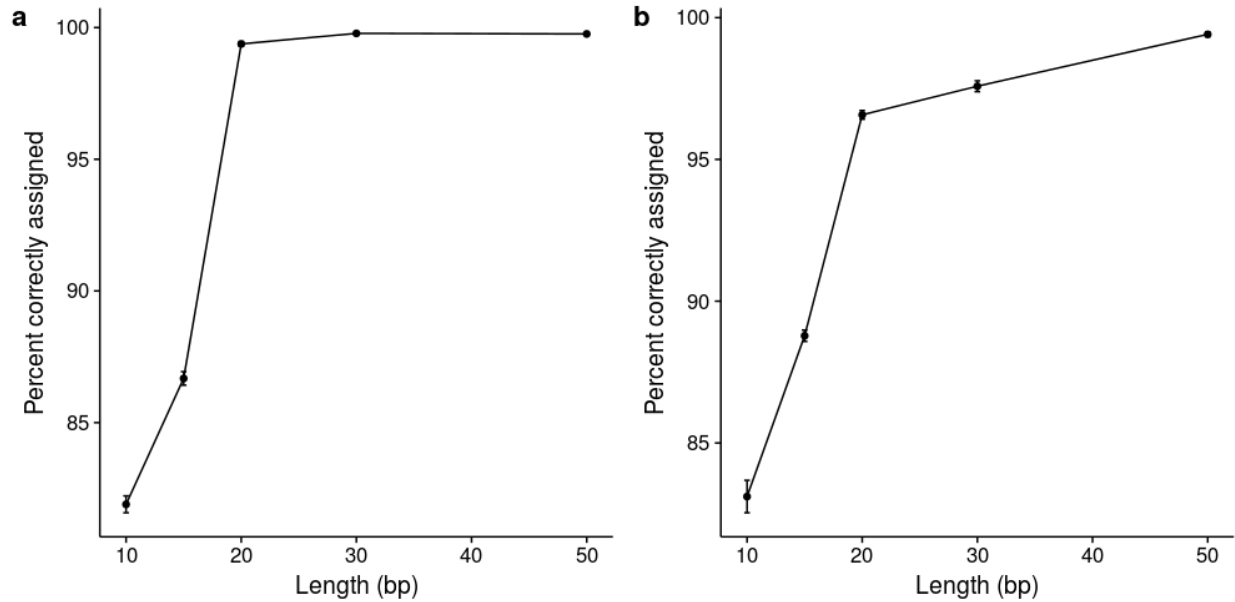
669



670

671 **Fig 4: A detailed view of the genomic neighbourhood and expression data for selected**
 672 **genes in Newman vs MRSA252.** Nucleotide identity was calculated using a 20 bp sliding
 673 window across the IGR, and this is shown alongside the flanking genes in their correct
 674 orientation (left). The corresponding expression data for the gene of interest was also shown
 675 (right), with the two boxplots per isolate corresponding to the two biological replicates. a) dapE
 676 b) ssaA_1 c) ytrA.

677



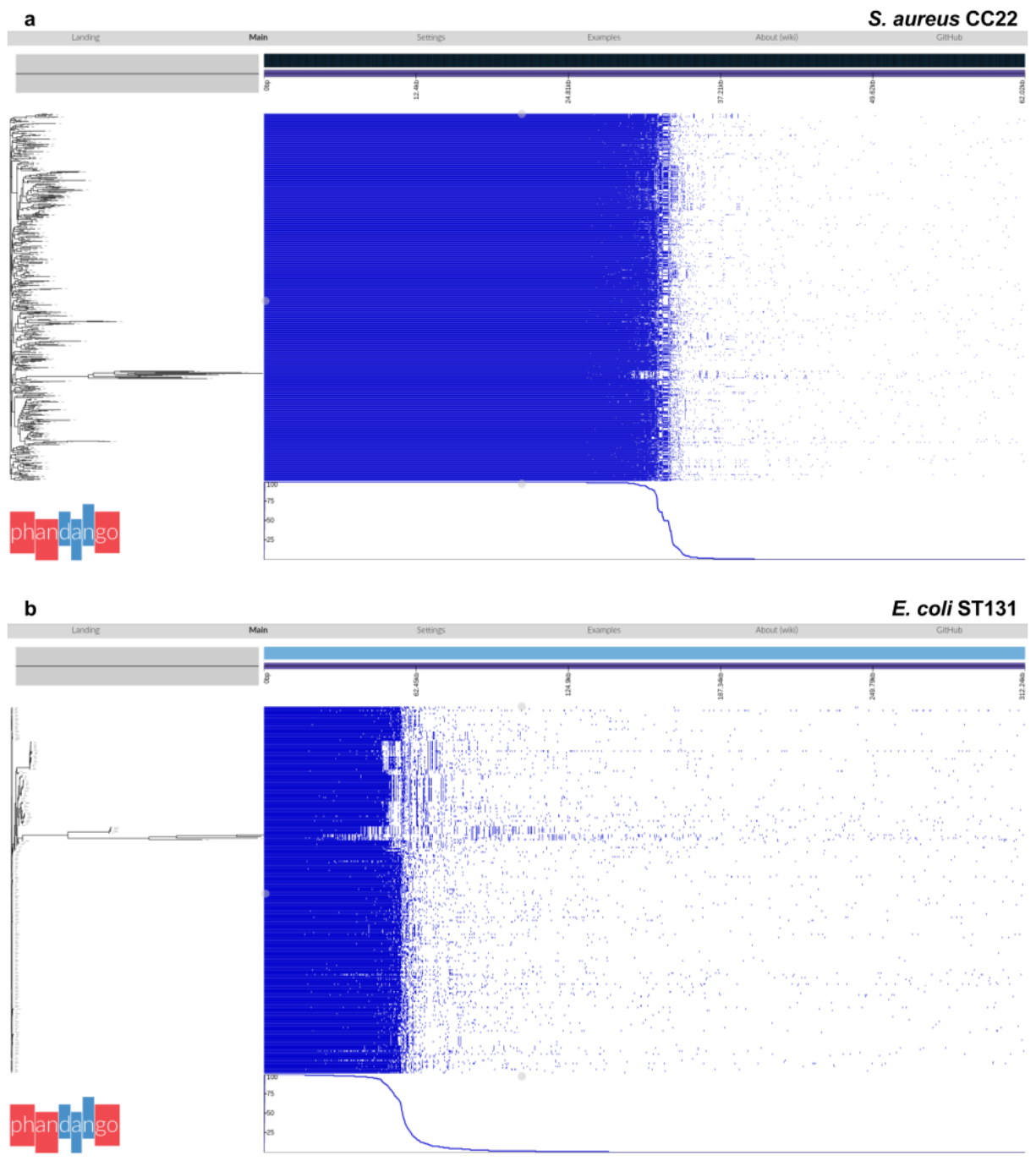
678

679

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

684

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



685

686

687 **Fig S2: The IGR pan-genome (“panIGRome”) as visualised using Phandango.** A
688 neighbour-joining phylogenetic tree was imported into Phandango alongside the
689 IGR_presence_absence.csv file. Each row corresponds to an isolate, and each column
690 corresponds to an IGR, with the IGRs ordered from the left in order of decreasing frequency

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

691 within the sample. The line graph at the bottom shows the frequency of the IGRs within the
692 sample. a) *S. aureus* ST22 b) *E. coli* ST131.

Sheet1

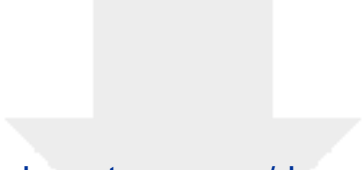
Species	Core genes	Core IGRs	Accessory genes	Accessory IGRs	Percentage core genes	Percentage core IGRs
<i>S. aureus</i> ST22	2409	1556	816	1543	95	95
<i>S. aureus</i>	2129	1134	3446	8033	85	69
<i>E. coli</i> ST131	3930	2296	8876	14133	84	77

Sheet1

Species	Core gene, Core IGR	Core gene, Accessory IGR	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

Accessory gene, Accessory IGR

92.6
96.8
97.3



Click here to access/download
Supplementary Material
Table_S1 (1).xlsx

