1 prewas: Data pre-processing for more informative bacterial GWAS

- 2
- 3 **Authors:** Katie Saund¹* (0000-0002-6214-6713), Zena Lapp²* (0000-0003-4674-2176),
- 4 Stephanie N. Thiede^{1*} (0000-0003-0173-4324), Ali Pirani¹ (0000-0001-7810-0982), Evan S.
- 5 Snitkin^{1,3} (0000-0001-8409-278X)
- 6
- 7 *equal contribution

89 Affiliations

- 10 ¹Department of Microbiology and Immunology
- 11 ²Department of Computational Medicine and Bioinformatics
- 12 ³Department of Internal Medicine/Division of Infectious Diseases
- 13 University of Michigan, Ann Arbor, Michigan
- 14

17

15 Corresponding Author

16 Evan S. Snitkin, esnitkin@med.umich.edu

18 Keywords

- 19 software, gwas, multiallelic loci, overlapping genes, reference allele, data pre-processing
- 20

21 ABSTRACT

- 22 While variant identification pipelines are becoming increasingly standardized, less attention has
- 23 been paid to the pre-processing of variants prior to their use in bacterial genome-wide
- 24 association studies (bGWAS). Three nuances of variant pre-processing that impact downstream
- 25 identification of genetic associations include the separation of variants at multiallelic sites,
- 26 separation of variants in overlapping genes, and referencing of variants relative to ancestral
- 27 alleles. Here we demonstrate the importance of these variant pre-processing steps on diverse
- 28 bacterial genomic datasets and present prewas, an R package, that standardizes the pre-
- 29 processing of multiallelic sites, overlapping genes, and reference alleles before bGWAS. This
- 30 package facilitates improved reproducibility and interpretability of bGWAS results. Prewas
- 31 enables users to extract maximal information from bGWAS by implementing multi-line
- 32 representation for multiallelic sites and variants in overlapping genes. Prewas outputs a binary
- 33 SNP matrix that can be used for SNP-based bGWAS and will prevent the masking of minor
- 34 alleles during bGWAS analysis. The optional binary gene matrix output can be used for gene-
- 35 based bGWAS which will enable users to maximize the power and evolutionary interpretability
- 36 of their bGWAS studies. Prewas is available for download from GitHub.
- 37
- 38 39
- 40
- 41
- 42
- 43
- 44

45 DATA SUMMARY

46

53

- 47
 1. prewas is available from GitHub under the MIT License (URL: https://github.com/Snitkin 48
 Lab-Umich/prewas) and can be installed using the command
- 49 devtools::install_github("Snitkin-Lab-Umich/prewas")
- 50
 2. Code to perform analyses is available from GitHub under the MIT License (URL:
 51
 https://github.com/Snitkin-Lab-Umich/prewas_manuscript_analysis)
- 52 3. All genomes are publicly available on NCBI (see Table S1 for more details)

54 IMPACT STATEMENT

55 In between variant calling and performing bacterial genome-wide association studies (bGWAS)

- 56 there are many decisions regarding processing of variants that have the potential to impact
- 57 bGWAS results. We discuss the benefits and drawbacks of various variant pre-processing
- 58 decisions and present the R package prewas to standardize single nucleotide polymorphism
- 59 (SNP) pre-processing, specifically to incorporate multiallelic sites and prepare the data for gene-
- 60 based analyses. We demonstrate the importance of these considerations by highlighting the
- 61 prevalence of multiallelic sites and SNPs in overlapping genes within diverse bacterial genomes
- 62 and the impact of reference allele choice on gene-based analyses.
- 63

64 INTRODUCTION

- 65 Bacterial genome-wide association studies (bGWAS) are frequently used to identify genetic
- 66 variants associated with variation in microbial phenotypes such as antibiotic resistance, host
- 67 specificity, and virulence (1–4). bGWAS methods can be classified into two general categories:
- those that use k-length nucleotide sequences (kmers) as features (e.g. (3,5–7)), and those that
- 69 use defined variant classes such as single nucleotide polymorphisms (SNPs), gene
- 70 presence/absence, or insertions/deletions (indels) as features (e.g. 4,8–12). bGWAS can be
- 71 performed using individual variants or by grouping variants into genes or pathways (i.e.
- 72 performing a burden test). While there have been efforts to standardize variant identification
- 73 protocols (13,14), less attention has been paid to the downstream processing of variants prior to
- their use for applications like bGWAS. In this paper, we focus on pre-processing of SNPs
- 75 (Figure 1A); however, the ideas and methods we discuss with respect to SNPs can be extended
- 76 to other genetic variants.
- 77
- 78 One aspect of pre-processing for SNP-based bGWAS is handling multiallelic sites. A site in the
- 79 genome is considered multiallelic when more than two alleles are present at that locus (Figure
- 1B). Multiallelic sites do not fit neatly into the framework of most bGWAS methods, which often
- 81 require a binary input (e.g. 3,4). Furthermore, the alternative minor alleles at a single site may
- 82 impact the encoded protein to different extents, and therefore considering them separately may
- 83 allow users to uncover otherwise masked relationships between genotype and phenotype.
- 84
- 85 Grouping SNPs by genes or metabolic pathways (Figure 1D) prior to performing bGWAS
- 86 increases power and reduces collinearity (3,15,16). When performing gene-based analyses, two
- 87 pre-processing steps may include choosing a reference allele for each SNP (Figure 1C) and
- 88 assigning SNPs in overlapping gene pairs. The reference allele is the nucleotide relative to

- 89 which variants are defined. Choice of reference allele is particularly important when grouping
- 90 SNPs by gene to ensure that the direction of evolution for each SNP is preserved. Additionally,
- 91 overlapping genes are common in bacteria (17,18). SNPs shared by overlapping gene pairs
- 92 may be assigned to both genes in a gene-based analysis.
- 93
- 94 To determine the importance of variant pre-processing methods for bGWAS, we investigated 95 the prevalence of multiallelic sites, mismatches in reference allele choice, and SNPs in 96 overlapping genes in 9 bacterial datasets. Our analysis indicates that multiallelic sites are
- 97 common in large, diverse bacterial datasets, there are frequently mismatches between different 98 reference allele choices, and SNPs in overlapping genes often have discordant functional
- 99 impacts. Therefore, pre-processing decisions have the potential to impact to bGWAS results.
- 100 We implemented a solution in the R package prewas to handle the nuances of variant pre-
- 101 processing to enable more robust and reproducible bGWAS analyses (Figure S1). The output of
- 102 prewas can be directly input into bGWAS tools that require a binary matrix as an input (e.g.
- 103 (3,4)). Prewas can be downloaded from GitHub.
- 104 105 **METHODS**

106 Datasets

- 107 The collection of datasets we used for data analysis and the corresponding bioprojects are
- 108 listed in Table S1 (19–30). All of these datasets contain whole-genome sequences of the 109 bacterial isolates.
- 110

111 Variant calling & tree building

- 112 SNP calling and phylogenetic tree reconstruction were performed on each dataset as described
- 113 in (23). The variant calling pipeline can be found on GitHub (https://github.com/Snitkin-Lab-
- 114 Umich/variant calling pipeline). In short, variant calling was performed with samtools v0.1.18
- 115 (31) using the reference genomes listed in Table S1, and trees were built using IQ-TREE v1.5.5 (32).
- 116

117

118 **Functional impact prediction**

- 119 The functional impact of each SNP was predicted using SnpEff (33). Variants are categorized
- 120 by SnpEff as low impact (e.g. synonymous mutations), moderate impact (e.g. nonsynonymous
- 121 mutations), or high impact (e.g. nonsense mutations). Only variants in coding regions were
- 122 included in analyses.
- 123

124 Data analysis

- 125 Statistical analyses and modeling were conducted in R v3.6.1. The analysis code and data are
- 126 available at: github.com/Snitkin-Lab-Umich/prewas manuscript analysis. The R packages we
- 127 used can be found in the prewas.yaml file on GitHub (github.com/Snitkin-Lab-Umich/prewas;
- 128 34-43), and can be installed using miniconda (44).
- 129 Multiallelic sites Linear regressions were modeled with percentage of variants that are
- 130 multiallelic as the response variable and either number of samples or mean pairwise SNP
- distance as the predictor. R² values are reported. 131

- 132 Reference alleles For each dataset, the reference genome allele, major allele, and ancestral
- allele were identified and the number of mismatches between them was quantified. Ancestral
- reconstruction was performed in R using the ape::ace function with ape v5.3 (34).
- 135 Allele convergence We recorded the number of times each allele arises on the tree, as
- 136 inferred from ancestral reconstruction, and then subtracted 1 to calculate the number of
- 137 convergence events for each allele.
- 138

139 **RESULTS & DISCUSSION**

- 140 To maximize the potential for identifying genetic variation associated with a given phenotype
- 141 using bGWAS, care must be taken in the pre-processing stage. Here we focus on three aspects
- of variant pre-processing and evaluate their potential downstream importance for bGWAS
- analysis. In particular, we report on the prevalence of multiallelic sites, mismatches between
- reference allele choice, and variants in overlapping genes across 9 bacterial datasets from
- 145 various species and of varying genetic diversity (Table 1).
- 146

147 Handling multiallelic sites

- 148 A multiallelic locus is a site in the genome with more than two alleles present and encompases
- both triallelic and quadallelic sites. bGWAS typically requires a binary input for each genotype
- (e.g. 3,4), and multiallelic sites are, by definition, not binary. Thus, special considerations must
 be taken to use multiallelic sites in bGWAS (see *Multi-line representation for multiallelic sites*).
- 152 We assessed the potential relevance of multiallelic SNPs to bGWAS on the basis of 1)
- 153 frequency, 2) differences in functional impact of alternative alleles at a single site, and 3)
- 154 environmente and a single site
- 154 convergence of multiallelic sites on phylogenetic tree.
- 155

156 Multiallelic site frequency

- 157 We expected that as the sample size increases the number of multiallelic sites would also
- 158 increase, as seen across human datasets of different sizes (45); however, this was not the case
- 159 when looking across different bacterial datasets (Figure S2A). We hypothesized that the lack of
- 160 correlation between the prevalence of multiallelic sites and dataset size was due to differences
- 161 in genetic diversity among the datasets (Table 1). Indeed, when we subsample from any single
- 162 dataset, the fraction of multiallelic sites increases as sample size increases until the diversity of
- 163 the dataset is exhausted (Figure 2A). Furthermore, datasets with higher sample diversity tend to
- have a larger fraction of multiallelic sites (Figure 2A,2B).
- 165

166 Differences in functional impact

- For multiallelic sites, considering each alternative allele at a single site allows for analyses to be performed on alleles based on their predicted functional impact on the encoded protein.
- Alternative alleles at a single site often have different predicted functional impacts (range across
- 170 datasets 0-18%, Figure 2C,S1C), and multiallelic sites include alleles with predicted high impact
- 171 mutations (Figure S2B). In light of these predicted allele-based functional differences, a bGWAS
- 172 user may want to only run bGWAS on alleles at multiallelic loci that are predicted to have a high
- 173 impact on the encoded protein.
- 174
- 175 Convergence on phylogenetic tree

- 176 For convergence-based bGWAS methods, a significant association between an allele and a
- phenotype requires that the allele converges on the phylogenetic tree (4,8). If alleles at
- multiallelic sites are convergent on the phylogeny, then they could potentially contribute to
- 179 genotype-phenotype associations. We found that single alleles from multiallelic sites are
- 180 convergent on the phylogeny as often as biallelic sites (Figure S1D), indicating that they could
- 181 potentially associate with phenotypes when using convergence-based bGWAS.
- 182
- 183 Multi-line representation for multiallelic sites
- 184 To use multiallelic sites in bGWAS, these sites typically must be represented as a binary input
- for each genotype (e.g. 3,4). Three ways multiallelic sites can be handled to fit with the binary
- 186 framework of bGWAS are: 1) remove them from the dataset prior to analysis, 2) group all minor
- alleles together, or 3) encode each minor allele separately. Excluding multiallelic sites is
 problematic if any of these sites determine the phenotype; in these cases, excluding multiallelic
- 189 sites will result in missed bGWAS hits. Furthermore, coding all minor alleles as one could
- 190 obscure true associations, particularly if the different minor alleles have dissimilar functional
- 191 impacts. Multi-line formatting of multiallelic SNPs provides more interpretability, more precise
- allele classification, and less information loss. For these reasons, multi-line representation is
- 193 increasingly important in certain human genetics analyses [12] and we propose this same
- 194 representation for bGWAS studies, particularly for large diverse datasets (Figure 1B).
- 195

196 Choosing a reference allele

197 Another aspect to consider when pre-processing SNPs for bGWAS is the allele referencing 198 method, which is critical for a uniform interpretation of variation at a gene locus when grouping 199 SNPs into genes. Three possible allele referencing methods are: the reference genome allele 200 from variant calling, the major allele, or the ancestral allele (Figure 1C). The reference genome 201 allele is the allele found in the reference genome when using a reference genome-based variant 202 calling approach. The major allele is the most common allele at a given locus in the dataset. 203 Neither of these methods encode the alleles with a consistent evolutionary direction. The 204 ancestral allele is the allele inferred to have existed at the most recent common ancestor of the 205 dataset. Given confident ancestral reconstruction, using the ancestral allele as the reference 206 allele allows for a uniform evolutionary interpretation of variants: there is a consistent direction 207 of evolution in that all mutations have arisen over time. We found that the three different 208 methods for identifying the reference allele frequently identify different alleles (range across 209 datasets 0-58%; Figure 3A). Thus, using the reference genome allele or the major allele as the 210 reference allele will not always maintain a consistent direction of evolution for each allele in a 211 gene, obscuring interpretation when grouping variants into genes.

- 212
- 213 Although ancestral reconstruction is the most interpretable option for reference allele choice,
- 214 this method is not feasible for some datasets. For example, sometimes we cannot confidently
- 215 predict the most likely ancestral root allele for many loci, as in the *Lactobacillus crispatus*
- 216 dataset (Figure 3B); in this case, it is not a reliable method to use to define the reference allele.
- 217 Other limitations of using the ancestral allele as the reference allele are that ancestral
- 218 reconstruction requires an accurate phylogenetic tree and may be computationally intensive for
- 219 large datasets. An alternative approach is to use the major allele as the reference allele as this

220 method does not require a tree and thus avoids ancestral reconstruction. When the ancestral 221 allele is not feasible, using the major allele is better than using the reference genome allele 222 when grouping variants into genes because using the major allele leads to less masking of 223 variation at the gene level (Figure S3).

224

225 Grouping variants into genes

226 Grouping variants into genes prior to performing bGWAS has two advantages for users: 1) 227 improved power to detect genotype-phenotype relationships due to reduced multiple testing 228 burden, and 2) enhanced interpretability as gene function may be clearer than the function of a 229 SNP. Grouping variants into genes may be a particularly helpful approach to bGWAS for 230 datasets with low penetrance of single variants but with convergence at the gene level (Figure 231 1D). To perform analysis of genomic variants grouped into genes, it is important to consider the 232 choice of reference allele (addressed above), assignment of variants in overlapping genes, and 233 functional impact of the variants.

234

235 It is important to ensure that variants in overlapping genes are assigned to each gene that the 236 variant is in to prevent information loss and because the functional impact of a SNP in one gene 237 may be different than its impact on the other gene(s). There are many overlapping genes that 238 share SNPs in each genome (Figure S4A,S4B). Furthermore, there are many sites where the 239 SNP has a different functional impact in the two overlapping genes (cumulative range across 240 datasets 50-70%; Figure 4). The functional impact of variants can be used to select what 241 variants to include in a gene-based analysis. For instance, researchers could subset to only 242 those SNPs most likely to affect gene function (e.g. start loss and stop gain mutations).

243

244 PACKAGE DESCRIPTION

245 We developed prewas to standardize the inclusion and representation of multiallelic sites, 246 choice of reference allele, and SNPs in overlapping genes (Figure 1A) for downstream use in 247 bGWAS analyses. Installation may be performed from GitHub (https://github.com/Snitkin-Lab-248 Umich/prewas). This R package is an easy-to-use tool with a function that minimally takes a 249 multiVCF input file. The multiVCF encodes the variant nucleotide alleles for all samples. The 250 outputs of the prewas function are matrices of variant presence and absence with multi-line 251 representation of multiallelic sites. Multiple optional files may be used as additional inputs to the 252 prewas function: a phylogenetic tree, an outgroup, and a GFF file. The phylogenetic tree may be 253 added when the user wants to identify ancestral alleles for the allele referencing step. The GFF 254 file contains information on gene location in the reference genome used to call variants and is 255 necessary to generate a binary matrix of presence and absence of variants in each gene.

- 256 Variants in overlapping genes are assigned to both genes. The matrix outputs from prewas can
- 257 be directly input into bGWAS tools such as treeWAS (4).
- 258

259 Generating a binary variant matrix including multiallelic sites (Figure 1B)

260 The multiVCF file is read into prewas and converted into an allele matrix with single-line

- 261 representation of each genomic position. Next, a reference allele is chosen for each variant
- 262 position (see section below). Then, the reference alleles are used to convert the allele matrix
- 263 into a binary matrix with multi-line representation of each multiallelic site. For each line in the

matrix, a 1 represents a single alternate allele, and a 0 represents either the reference allele or
 any other alternate alleles if the position is a multiallelic site. This binary matrix is output by
 prewas.

267

268 Identifying reference alleles (Figure 1C)

We have implemented two methods to identify appropriate reference alleles (see Results &Discussion for more details).

271

Ancestral allele approach. The reference allele may be defined as the ancestral allele at each
 genomic position. In this approach, we identify the most likely allele of the most recent common
 ancestor of all samples in the dataset by performing ancestral reconstruction. This allele is then
 always set to 0 in the binary variant matrix. Here, any 1 in the binary variant matrix represents a
 mutation that has arisen over time, assuming confident ancestral reconstruction results.

277

278 *Major allele approach.* The reference allele may also be defined as the major allele at each

279 genomic position. In this case, the most common allele in the dataset is the reference allele.

280 This choice improves the performance speed of prewas as compared to using the ancestral

- allele at the cost of evolutionary interpretability.
- 282

283 Grouping variants by gene (Figure 1D)

If a GFF file is provided as input to prewas, variants will be grouped by gene. First, variants found in overlapping genes will be split into multiple lines where each line corresponds to one of the overlapping genes. This ensures that the variant is assigned to each of the genes in which it occurs. Next, variants are collapsed into genes such that the output is a binary matrix with each line corresponding to a single gene and each entry within the matrix is the presence or absence of any variant within that gene.

290

291 Future directions

In a future version of prewas, we plan to implement an option to allow users to select which
SNPs they want to include in the binary output matrices based on SnpEff functional impact (e.g.
only output predicted high functional impact mutations). When considering the predicted
functional impact of each SNP, it is important to use multi-line representation of multiallelic sites
even when grouping SNPs by genes because sometimes different alleles at the same site have
different predicted functional impacts. Furthermore, prewas could also be extended to process
other genomic variants such as indels and structural variants.

- 299
- 300

301 CONCLUSION

We have developed prewas, an easy-to-use R package, that handles multiallelic sites and grouping variants into genes. The prewas package provides a binary SNP matrix output that can be used for SNP-based bGWAS and will prevent the masking of minor alleles during bGWAS analysis. The optional binary gene matrix output can be used for gene-based bGWAS which will

analysis. The optional binary gene matrix output can be used for gene-based bowks which we are a set of gene-based bowks when we are a set of gene-based bowks we are a set of gene-based bowks when we are a set of gene-based bowks we are a set

307 bGWAS.

308

309 AUTHOR CONTRIBUTIONS

- 310 The study was conceptualized by KS, ZL, SNT, and ESS. Software design and implementation,
- 311 formal analysis, original draft preparation, and visualization were performed by KS, ZL, and
- 312 SNT. Data was curated by AP, KS, ZL, and SNT. All authors performed editing and review, and
- 313 ESS supervised the project.
- 314

315 CONFLICTS OF INTEREST

- 316 The authors declare that there are no conflicts of interest.
- 317

318 FUNDING

- 319 KS was supported by the National Institutes of Health (T32GM007544). ESS and KS were
- 320 supported by the National Institutes of Health (1U01AI124255). SNT was supported by the
- 321 Molecular Mechanisms of Microbial Pathogenesis training grant (NIH T32 Al007528). ZL
- 322 received support from the National Science Foundation Graduate Research Fellowship Program
- 323 under Grant No. DGE 1256260. Any opinions, findings, and conclusions or recommendations
- expressed in this material are those of the authors and do not necessarily reflect the views of
- 325 the National Science Foundation.
- 326

327 ACKNOWLEDGEMENTS

328 We thank Shawn Hawken for coining the name prewas.

330 **REFERENCES**

331

329

- Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: lessons
 from human GWAS. Nature Reviews Genetics. 2017 Jan;18(1):41–50.
- Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pangenome-wide association studies with Scoary. Genome Biology. 2016 Nov 25;17(1):238.
- Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. Stegle O, editor. Bioinformatics. 2018 Dec 15;34(24):4310–2.
- Collins C, Didelot X. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. PLOS Computational Biology. 2018 Feb 5;14(2):e1005958.
- 5. Earle SG, Wu C-H, Charlesworth J, Stoesser N, Gordon NC, Walker TM, et al. Identifying
 lineage effects when controlling for population structure improves power in bacterial
 association studies. Nat Microbiol. 2016 Apr 4;1:16041.
- Lees JA, Vehkala M, Välimäki N, Harris SR, Chewapreecha C, Croucher NJ, et al.
 Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. Nat Commun. 2016 Sep 16;7(1):1–8.
- Jaillard M, Lima L, Tournoud M, Mahé P, Belkum A van, Lacroix V, et al. A fast and
 agnostic method for bacterial genome-wide association studies: Bridging the gap between
 k-mers and genetic events. PLOS Genetics. 2018 Nov 12;14(11):e1007758.
- Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. Genomic
 analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. Nature Genetics. 2013 Oct;45(10):1183–9.
- 354 9. Alam MT, Petit RA, Crispell EK, Thornton TA, Conneely KN, Jiang Y, et al. Dissecting

- vancomycin-intermediate resistance in staphylococcus aureus using genome-wide
 association. Genome Biol Evol. 2014 Apr 30;6(5):1174–85.
- 10. Chewapreecha C, Marttinen P, Croucher NJ, Salter SJ, Harris SR, Mather AE, et al.
 Comprehensive identification of single nucleotide polymorphisms associated with betalactam resistance within pneumococcal mosaic genes. PLoS Genet. 2014
 Aug;10(8):e1004547.
- 11. Desjardins CA, Cohen KA, Munsamy V, Abeel T, Maharaj K, Walker BJ, et al. Genomic
 and functional analyses of Mycobacterium tuberculosis strains implicate ald in D cycloserine resistance. Nat Genet. 2016;48(5):544–51.
- 12. Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, Sloan TJ, et al. Predicting the virulence of MRSA from its genome sequence. Genome Res. 2014 May;24(5):839–49.
- 366 13. Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, et al. Best practices for
 367 evaluating single nucleotide variant calling methods for microbial genomics. Front Genet
 368 [Internet]. 2015 Jul 7 [cited 2019 Dec 10];6. Available from:
 369 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4493402/
- Yoshimura D, Kajitani R, Gotoh Y, Katahira K, Okuno M, Ogura Y, et al. Evaluation of SNP
 calling methods for closely related bacterial isolates and a novel high-accuracy pipeline:
 BactSNP. Microbial Genomics, 2019;5(5):e000261.
- 373 15. Zhan X, Chen S, Jiang Y, Liu M, Iacono WG, Hewitt JK, et al. Association Analysis and
 374 Meta-Analysis of Multi-allelic Variants for Large Scale Sequence Data. bioRxiv [Internet].
 375 2017 Oct 3 [cited 2019 Nov 26]; Available from:
 376 http://biorxiv.org/lookup/doi/10.1101/197913
- 376 Inttp://bioixiv.org/lookup/doi/10.1101/197913
 377 16. Farhat MR, Freschi L, Calderon R, Ioerger T, Snyder M, Meehan CJ, et al. GWAS for
- 378quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance379genes and regulatory regions. Nature Communications. 2019 May 13;10(1):1–11.
- Johnson ZI, Chisholm SW. Properties of overlapping genes are conserved across
 microbial genomes. Genome Res. 2004 Nov 1;14(11):2268–72.
- 18. Huvet M, Stumpf MP. Overlapping genes: a window on gene evolvability. BMC Genomics.
 2014 Aug 27;15(1):721.
- Carlson PE, Walk ST, Bourgis AET, Liu MW, Kopliku F, Lo E, et al. The relationship
 between phenotype, ribotype, and clinical disease in human Clostridium difficile isolates.
 Anaerobe. 2013 Dec;24:109–16.
- Saund K, Rao K, Young VB, Snitkin ES. Genetic determinants of trehalose utilization are
 not associated with severe *Clostridium difficile* infection [Internet]. Infectious Diseases
 (except HIV/AIDS); 2019 Oct [cited 2019 Nov 6]. Available from:
 http://medrxiv.org/lookup/doi/10.1101/19008342
- Mody L, Krein SL, Saint S, Min LC, Montoya A, Lansing B, et al. A Targeted Infection
 Prevention Intervention in Nursing Home Residents With Indwelling Devices: A
 Randomized Clinical Trial. JAMA Intern Med. 2015 May 1:175(5):714–23.
- Mody L, Foxman B, Bradley S, McNamara S, Lansing B, Gibson K, et al. Longitudinal
 Assessment of Multidrug-Resistant Organisms in Newly Admitted Nursing Facility Patients:
 Implications for an Evolving Population. Clin Infect Dis. 2018 Aug 31;67(6):837–44.
- Han JH, Lapp Z, Bushman F, Lautenbach E, Goldstein EJC, Mattei L, et al. WholeGenome Sequencing To Identify Drivers of Carbapenem-Resistant *Klebsiella pneumoniae*Transmission within and between Regional Long-Term Acute-Care Hospitals. Antimicrob
 Agents Chemother. 2019 Aug 26;63(11):e01622-19, /aac/63/11/AAC.01622-19.atom.
- 401 24. Bassis CM, Bullock KA, Sack DE, Saund K, Pirani A, Snitkin ES, et al. Evidence that
 402 vertical transmission of the vaginal microbiota can persist into adolescence [Internet].
 403 Microbiology; 2019 Sep [cited 2019 Nov 6]. Available from:
- 404 http://biorxiv.org/lookup/doi/10.1101/768598
- 405 25. Sun Z, Harris HMB, McCann A, Guo C, Argimón S, Zhang W, et al. Expanding the

- 406 biotechnology potential of lactobacilli through comparative genomics of 213 strains and 407 associated genera. Nature Communications. 2015 Sep 29;6(1):1–13.
- Popovich KJ, Snitkin ES, Zawitz C, Aroutcheva A, Payne D, Thiede SN, et al. Frequent
 Methicillin-Resistant Staphylococcus aureus Introductions Into an Inner-city Jail:
- Indications of Community Transmission Networks. Clin Infect Dis [Internet]. [cited 2019
 Dec 20]; Available from: https://academic.oup.com/cid/advance-
- 412 article/doi/10.1093/cid/ciz818/5551540
- Roach DJ, Burton JN, Lee C, Stackhouse B, Butler-Wu SM, Cookson BT, et al. A Year of
 Infection in the Intensive Care Unit: Prospective Whole Genome Sequencing of Bacterial
 Clinical Isolates Reveals Cryptic Transmissions and Novel Microbiota. PLoS Genet. 2015
 Jul;11(7):e1005413.
- Sichtig H, Minogue T, Yan Y, Stefan C, Hall A, Tallon L, et al. FDA-ARGOS is a database
 with public quality-controlled reference genomes for diagnostic use and regulatory science.
 Nature Communications. 2019 Jul 25;10(1):1–13.
- 420 29. Lira F, Berg G, Martínez JL. Double-Face Meets the Bacterial World: The Opportunistic
 421 Pathogen Stenotrophomonas maltophilia. Front Microbiol. 2017;8:2190.
- 422 30. Esposito A, Pompilio A, Bettua C, Crocetta V, Giacobazzi E, Fiscarelli E, et al. Evolution of
 423 Stenotrophomonas maltophilia in Cystic Fibrosis Lung over Chronic Infection: A Genomic
 424 and Phenotypic Population Study. Front Microbiol. 2017;8:1590.
- 425 31. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
 426 Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.
- 32. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective
 stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015
 Jan;32(1):268–74.
- 33. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for
 annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in
 the genome of Drosophila melanogaster strain w ¹¹¹⁸□; iso-2; iso-3. Fly. 2012 Apr;6(2):80–
 92.
- 434 34. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary
 435 analyses in R. Bioinformatics. 2019 Feb 1;35(3):526–8.
- 436 35. Bengtsson H, R Core Team. future.apply: Apply Function to Elements in Parallel using
 437 Futures [Internet]. 2019 [cited 2019 Dec 10]. Available from: https://CRAN.R438 project.org/package=future.apply
- 439 36. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011 Feb 15;27(4):592–
 440 3.
- 441 37. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things).
 442 Methods in Ecology and Evolution. 2012;3(2):217–23.
- 443 38. Knaus BJ, Grünwald NJ. vcfr: a package to manipulate and visualize variant call format
 444 data in R. Molecular Ecology Resources. 2017;17(1):44–53.
- Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the
 Tidyverse. Journal of Open Source Software. 2019 Nov 21;4(43):1686.
- 447 40. Wickham H. Reshaping Data with the reshape Package. Journal of Statistical Software.
 448 2007 Nov 13;21(1):1–20.
- 449 41. Kolde R. pheatmap: Pretty Heatmaps [Internet]. 2019 [cited 2019 Dec 10]. Available from:
 450 https://CRAN.R-project.org/package=pheatmap
- 451 42. Xie Y. animation: An R Package for Creating Animations and Demonstrating Statistical
 452 Methods. Journal of Statistical Software. 2013 Apr 21;53(1):1–27.
- 43. Pagès H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: Efficient manipulation of
 biological strings [Internet]. Bioconductor version: Release (3.10); 2019 [cited 2019 Dec
 455 10]. Available from: https://bioconductor.org/packages/Biostrings/
- 456 44. Anaconda | The World's Most Popular Data Science Platform [Internet]. Anaconda. [cited

- 457 2019 Dec 10]. Available from: https://www.anaconda.com/
- 458 45. Campbell IM, Gambin T, Jhangiani S, Grove ML, Veeraraghavan N, Muzny DM, et al.
- 459 Multiallelic Positions in the Human Genome: Challenges for Genetic Analyses. Hum Mutat.
 460 2016 Mar;37(3):231–4.
- 462 **Data Bibliography**
- 463 See Table S1.
- 464

461

- 465 TABLES
- 466 Table 1: Bacterial datasets
- 467

Name	Samples (Count)	Multiallelic Sites (Count)	Mean SNP Distance (BP)	SNPs in overlapping genes (Count)	Reference
C. difficile #3	107	3527	18010.4	11511	19
C. difficile #4	247	2460	6840.8	7862	20
E. faecium #1	152	118	2976.5	8	21, 22
E. faecalis #1	157	201	5960.1	20	21, 22
K. pneumoniae #1	453	920	3825.4	76	23
L. crispatus #1	28	536	9501.5	34	24, 25
S. aureus #1	150	296	5195.0	74	26
S. aureus #2	267	391	5561.4	38	21, 22
S. maltophilia #1	149	3080	11243.4	32594	27-30

468 469



Figure 1: prewas workflow. (A) Overview of the prewas workflow. Grey and colored boxes:
processing steps. White boxes: output generated. (B) Multi-line representation of multiallelic
sites. (C) Possible methods to find a reference allele. The ancestral allele method and the major

allele method are implemented in prewas. (D) Grouping SNPs into genes.

- 475
- 476
- 477











485

486 Figure 3. Methods to determine the reference allele identify different alleles. (A) The

487 fraction of variant positions where the identified reference allele varies between two methods.

488 Only high confidence ancestral reconstruction sites (>=87.5% confidence in the ancestral root

allele by maximum likelihood) are included. (B) Fraction of low confidence ancestral

490 reconstruction sites for each dataset (<87.5% confidence in the ancestral root allele by



maximum likelihood).



495 Figure 4: SNPs in overlapping sites can have distinct functional impacts in each gene of

496 the gene pair. The fraction of overlapping variant positions where the SNP has a different497 predicted functional impact in each of the two overlapping genes.

511512 SUPPLEMENT





Supplementary Figure 1: Detailed prewas workflow. 515

519 Supplementary Figure 2: Multiallelic Sites (A) Independence observed between sample size 520 and prevalence of multiallelic sites. (B) Prevalence of multiallelic sites compared to variant sites 521 with each subset to the various predicted functional impacts. Any multiallelic site with specific 522 impact is compared to any variant site with the same predicted impact. (C) Multiallelic sites with 523 discordant predicted functional impact among alternative alleles. (D) The relative frequency of 524 the number of times an allele arises on the tree. At multiallelic sites, all minor alleles are treated 525 separately.

- 526
- 527



528

529

530 Supplementary Figure 3: Masking variation at the gene level when grouping into genes.

531 When not confident in the ancestral reconstruction or ancestral reconstruction is not

532 computationally feasible, we suggest referencing to the major allele. In this example,

533 referencing to the reference genome allele masks variation at the gene level. When referencing

to the reference genome allele, the variation in Position 2 gets masked by the variation in
 Position 1 when grouped by gene, leading to a likely lack of association. However, if instead we

reference to the major allele, the variation in Gene A is maintained, allowing for potential associations to be detected.





539

540

541 Supplementary Figure 4. Overlapping genes with SNPs. (A) SNP loci found in positions
542 shared by overlapping genes. (B) Overlapping genes with SNPs found in the overlapping
543 positions.

- 544
- 545
- 546

547 Table S1: Sources for bacterial datasets

Name	Dataset Description	Bioproject	Bioproject_link	Reference Genome Biosample	Ref.
C. difficile #3	Clinical infection isolates	PRJNA594943	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA594943	SAMEA1705932	19
C. difficile #4	Clinical infection isolates	PRJNA561087	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA561087	SAMEA1705932	20
E. faecium #1	Healthcare-associated colonization isolates	PRJNA435617	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA435617	SAMN10039001	21, 22
E. faecalis #1	Healthcare-associated colonization isolates	PRJNA435617	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA435617	SAMN10039299	21, 22
K. pneumoniae #1	Healthcare-associated clinical isolates	PRJNA415194	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA415194	SAMN01057611	23
L. crispatus #1	Publicly available genomes	PRJNA547620, PRJNA50051, PRJNA50057, PRJNA50067, PRJNA50067, PRJNA50165, PRJNA50167, PRJNA52107, PRJNA52105, PRJNA272101, PRJNA272101, PRJEB8104, PRJNA316969, PRJNA379934, PRJEB22112	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA547620, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA50051, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA50067, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA50165, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA50165, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA50167, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA50167, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA50165, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA52107, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA52105, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA52105, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA272101, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA316969, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA316969, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA316969, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA316969, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA316969, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA316969,	SAMEA2272191	24, 25
S. aureus #1	MRSA jail colonization isolates	PRJNA530184	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA530184	SAMN00253845	26
S. aureus #2	Healthcare-associated colonization isolates	PRJNA435617	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA435617	SAMN10038895	21, 22
S. maltophilia #1	Publicly available genomes	PRJDB3841, PRJNA267549, PRJNA231221, PRJNA380601, PRJNA380601, PRJNA380620, PRJNA38096, PRJNA483996, PRJNA489399, PRJNA268101, PRJNA344912	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJDB3841, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA267549, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA231221, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA360601, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA350620, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA350620, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA390523, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA483996, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA483999, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA489399, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA484912	SAMEA1705934	27-30









Predicted Functional Impact

High & Moderate High & Low Moderate & Low



Reference SNPs

Group by gene

Name	Samples (Count)	Multiallelic Sites (Count)	Mean SNP Distance (BP)	SNPs in overlapping genes (Count)	Reference
C. difficile #3	107	3527	18010.4	11511	19
C. difficile #4	247	2460	6840.8	7862	20
E. faecium #1	152	118	2976.5	8	21, 22
E. faecalis #1	157	201	5960.1	20	21, 22
K. pneumoniae #1	453	920	3825.4	76	23
L. crispatus #1	28	536	9501.5	34	24, 25
S. aureus #1	150	296	5195.0	74	26
S. aureus #2	267	391	5561.4	38	21, 22
S. maltophilia #1	149	3080	11243.4	32594	27-30