MICROBIAL GENOMICS

METHOD

Saund *et al.*, *Microbial Genomics* 2020;6
DOI 10.1099/mgen.0.000368

MICROBIOLOGY SOCIETY

OPEN DATA OPEN MICROBIOLOGY

prewas: data pre-processing for more informative bacterial GWAS

Katie Saund[1]†, Zena Lapp[2]†, Stephanie N. Thiede[1]†, Ali Pirani[1] and Evan S. Snitkin[1,3,*]

Abstract

While variant identification pipelines are becoming increasingly standardized, less attention has been paid to the pre-processing of variants prior to their use in bacterial genome-wide association studies (bGWAS). Three nuances of variant pre-processing that impact downstream identification of genetic associations include the separation of variants at multiallelic sites, separation of variants in overlapping genes, and referencing of variants relative to ancestral alleles. Here we demonstrate the importance of these variant pre-processing steps on diverse bacterial genomic datasets and present prewas, an R package, that standardizes the pre-processing of multiallelic sites, overlapping genes, and reference alleles before bGWAS. This package facilitates improved reproducibility and interpretability of bGWAS results. prewas enables users to extract maximal information from bGWAS by implementing multi-line representation for multiallelic sites and variants in overlapping genes. prewas outputs a binary SNP matrix 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 enable users to maximize the power and evolutionary interpretability of their bGWAS studies. prewas is available for download from GitHub.

DATA SUMMARY

1. prewas is available from CRAN and can be installed using the command install.packages('prewas').

2. Code to perform analyses is available from GitHub under the MIT License (URL: https://github.com/Snitkin-Lab-Umich/prewas_manuscript_analysis).

3. All genomes are publicly available on NCBI [see Table S1, (available in the online version of this article) for more details].

INTRODUCTION

Bacterial genome-wide association studies (bGWAS) are frequently used to identify genetic variants associated with variation in microbial phenotypes such as antibiotic resistance, host 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 use defined variant classes such as SNPs, gene presence/absence, or insertions/deletions (indels) as features (e.g. [4, 8–12]). Unlike k-mer-based bGWAS, variant-based bGWAS can be performed using individual variants or by grouping variants into genes or pathways (i.e. performing a burden test). Additionally, variant-based bGWAS may be preferred over k-mer-based methods for ease of biological interpretation. While there have been efforts to standardize variant identification 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 (Fig. 1a); however, the ideas and methods we discuss with respect to SNPs can be extended to other genetic variants.

One aspect of pre-processing for SNP-based bGWAS is handling multiallelic sites. A site in the genome is considered multiallelic when more than two alleles are present

Received 02 January 2020; Accepted 31 March 2020; Published 20 April 2020
Author affiliations: [1]Department of Microbiology and Immunology, University of Michigan, Ann Arbor, Michigan, USA; [2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA; [3]Department of Internal Medicine/Division of Infectious Diseases, University of Michigan, Ann Arbor, Michigan, USA.
***Correspondence:** Evan S. Snitkin, esnitkin@med.umich.edu
Keywords: software; GWAS; multiallelic loci; overlapping genes; reference allele; data pre-processing.
Abbreviations: bGWAS, bacterial genome-wide association studies; GFF, general feature format; indels, insertions/deletions; kmers, k-length nucleotide sequences; SNP, single nucleotide polymorphism; VCF, variant call format.
†These authors contributed equally to this work
Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. One supplementary table and six supplementary figures are available with the online version of this article.

at that locus (Fig. 1b). Multiallelic sites do not fit neatly into the framework of most bGWAS methods, which often require a binary input [3, 4]. Furthermore, the alternative minor alleles at a single site may impact the encoded protein to different extents, and therefore considering them separately may allow users to uncover otherwise masked relationships between genotype and phenotype.

Grouping SNPs by genes or metabolic pathways (Fig. 1d) prior to performing bGWAS can increase power and reduce collinearity [3, 15, 16]. However, power is only increased when grouping SNPs if the grouped SNPs affect the gene in the same direction, even if not to the same degree. When performing gene-based analyses, two pre-processing steps may include choosing a reference allele for each SNP (Fig. 1c) and assigning SNPs in overlapping gene pairs. The reference allele is the nucleotide relative to which variants are defined. Choice of reference allele is particularly important when grouping SNPs by gene to ensure that the direction of evolution for each SNP is preserved. Additionally, overlapping genes are common in bacteria [17, 18]. SNPs shared by overlapping gene pairs may be assigned to both genes in a gene-based analysis.

To determine the importance of variant pre-processing methods for bGWAS, we investigated the prevalence of multiallelic sites, mismatches in reference allele choice, and SNPs in overlapping genes in nine bacterial datasets. Our analysis indicates that multiallelic sites are common in large, diverse bacterial datasets, there are frequently mismatches between different reference allele choices, and SNPs in overlapping genes often have discordant functional impacts. Therefore, pre-processing decisions have the potential to impact bGWAS results. We implemented a solution in the R package prewas to handle the nuances of variant pre-processing to enable more robust and reproducible bGWAS analyses (Fig. S1). The output of prewas can be directly input into bGWAS tools that require a binary matrix as an input (e.g. [3, 4]). prewas can be downloaded from GitHub.

Impact Statement

In between variant calling and performing bacterial genome-wide association studies (bGWAS) there are many decisions regarding processing of variants that have the potential to impact bGWAS results. We discuss the benefits and drawbacks of various variant pre-processing decisions and present the R package prewas to standardize SNP pre-processing, specifically to incorporate multiallelic sites and prepare the data for gene-based analyses. We demonstrate the importance of these considerations by highlighting the prevalence of multiallelic sites and SNPs in overlapping genes within diverse bacterial genomes and the impact of reference allele choice on gene-based analyses.

METHODS

Datasets

The collection of datasets we used for data analysis and the corresponding bioprojects are listed in Table S1 [19–30]. All of these datasets contain whole-genome sequences of the bacterial isolates.

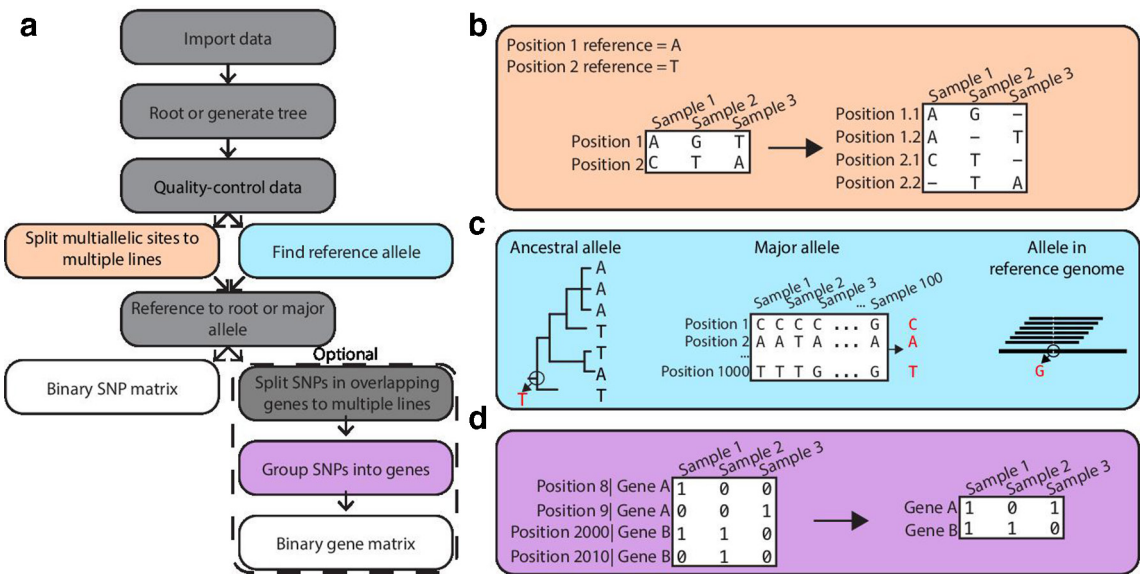

Fig. 1. prewas workflow. (a) Overview of the prewas workflow. Grey and colored boxes: processing steps. White boxes: output generated. (b) Multi-line representation of multiallelic sites. (c) Possible methods to find a reference allele. The ancestral allele method and the major allele method are implemented in prewas. (d) Grouping SNPs into genes.

Variant calling and tree building

SNP calling and phylogenetic tree reconstruction were performed on each dataset as described in [23]. The variant-calling pipeline can be found on GitHub (https://github.com/Snitkin-Lab-Umich/variant_calling_pipeline). In short, variant calling was performed with samtools v1.9 [31] using the reference genomes listed in Table S1, and trees were built using IQ-TREE v1.6.12 [32].

Functional-impact prediction

The functional impact of each SNP was predicted using SnpEff v4.3T [33]. Variants are categorized by SnpEff as low impact (e.g. synonymous mutations), moderate impact (e.g. nonsynonymous mutations) or high impact (e.g. nonsense mutations). Only variants in coding regions were included in analyses.

Data analysis

Statistical analyses and modelling were conducted in R v3.6.1. The analysis code and data are available at: github.com/Snitkin-Lab-Umich/prewas_manuscript_analysis. The R packages we used can be found in the prewas.yaml file on GitHub (github.com/Snitkin-Lab-Umich/prewas; [34–43]), and can be installed using miniconda [44].

Multiallelic sites

Linear regressions were modelled with the percentage of variants that are multiallelic as the response variable and either the number of samples or mean pairwise SNP distance as the predictor. R^2 values are reported.

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].

Heritability analysis

Continuous Brownian motion and white-noise phenotypes following a normal distribution were generated for each dataset using both maximum-likelihood and neighbour-joining trees. The R function fastBM in phytools v0.6–99 [37] was used to generate Brownian motion phenotypes. White-noise phenotypes were generated by randomly shuffling the Brownian motion phenotype among the samples. Shared variants between pairs of samples were counted using the reference-genome allele, the major allele and the ancestral allele, with and without multiallelic sites. We consider each of these shared variant matrices to be a different possible kinship matrix. Heritability of the simulated phenotypes was calculated for kinship matrices with and without multiallelic sites using limix v2.0.0 [45].

Allele convergence

We recorded the number of times each allele arises on the tree, as inferred from ancestral reconstruction, and then subtracted 1 to calculate the number of convergence events for each allele.

Resource utilization

We ran prewas with one or ten cores, with or without ancestral reconstruction, and without or without providing a phylogenetic tree. We recorded total memory usage and run time.

RESULTS AND DISCUSSION

To maximize the potential for identifying genetic variation associated with a given phenotype 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 nine bacterial datasets from various species and of varying genetic diversity (Table 1). We chose to include a diverse array of important and emerging hospital pathogens

Table 1. Bacterial datasets

Name	Samples (count)	Multiallelic sites (count)	Mean SNP distance (BP)	SNPs in overlapping genes (count)	Reference
C. difficile no. 1	107	3527	18010.4	11511	[19]
C. difficile no. 2	247	2460	6840.8	7862	[20]
E. faecium	152	118	2976.5	8	[21, 22]
E. faecalis	157	201	5960.1	20	[21, 22]
K. pneumoniae	453	920	3825.4	76	[23]
L. crispatus	28	536	9501.5	34	[24, 25]
S. aureus no. 1	150	296	5195	74	[26]
S. aureus no. 2	267	391	5561.4	38	[21, 22]
S. maltophilia	149	3080	11243.4	32594	[27–30]

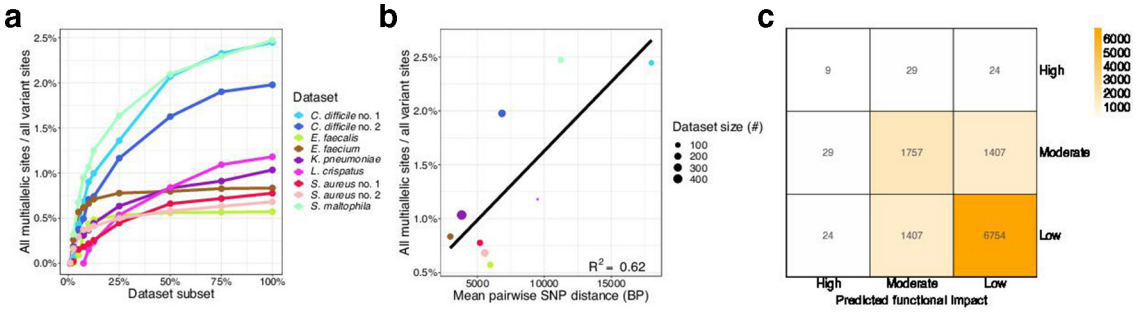

Fig. 2. Prevalence and predicted functional impact of multiallelic sites. (a) The number of multiallelic sites increases as sample size increases until the total diversity of the dataset is sampled. (b) More diverse samples have relatively more multiallelic sites. (c) Counts of predicted functional impact (mis)matches for pairs of alleles at triallelic sites (aggregated across all datasets). Alternative alleles often differ in impact.

and commensals to highlight genomic differences between various bacterial species.

Handling multiallelic sites

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). We assessed the potential relevance of multiallelic SNPs to bGWAS on the basis of (1) frequency, (2) differences in functional impact of alternative alleles at a single site and (3) convergence of multiallelic sites on the phylogenetic tree.

Multiallelic site frequency

We expected that as the sample size increases the number of multiallelic sites would also increase, as seen across human datasets of different sizes [46]; however, this was not the case when looking across different bacterial datasets (Fig. S2a). We hypothesized that the lack of correlation between the prevalence of multiallelic sites and dataset size was due to differences in genetic diversity among the datasets (Table 1). Indeed, when we subsample from any single dataset, the fraction of multiallelic sites increases as sample size increases until the diversity of the dataset is exhausted (Fig. 2a). Furthermore, datasets with higher sample diversity tend to have a larger fraction of multiallelic sites(Fig. 2a, b).

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 datasets 0–18 %, Figs 2c and S2c), and multiallelic sites include alleles with predicted high-impact mutations (Fig. S2b). In light of these predicted allele-based functional differences, a bGWAS user may want to only run bGWAS on alleles at multiallelic loci that are predicted to have a high impact on the encoded protein.

Convergence on phylogenetic tree

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 genotype-phenotype associations. We found that single alleles from multiallelic sites also converge on the phylogeny (Fig. S2d), indicating that they could potentially associate with phenotypes when using convergence-based bGWAS.

Multi-line representation for multiallelic sites

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 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 sites will result in missed bGWAS hits. Furthermore, coding all minor alleles as one could obscure true associations, particularly if the different minor alleles have dissimilar functional 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 increasingly important in certain human genetics analyses [12] and we propose this same representation for bGWAS studies, particularly for large diverse datasets (Fig. 1b).

Choosing a reference allele

Another aspect to consider when pre-processing SNPs for bGWAS is the allele referencing method. An allele referencing step has previously been implemented in bGWAS [5] and is critical for a uniform interpretation of variation at a gene locus when grouping SNPs into genes. Three possible allele

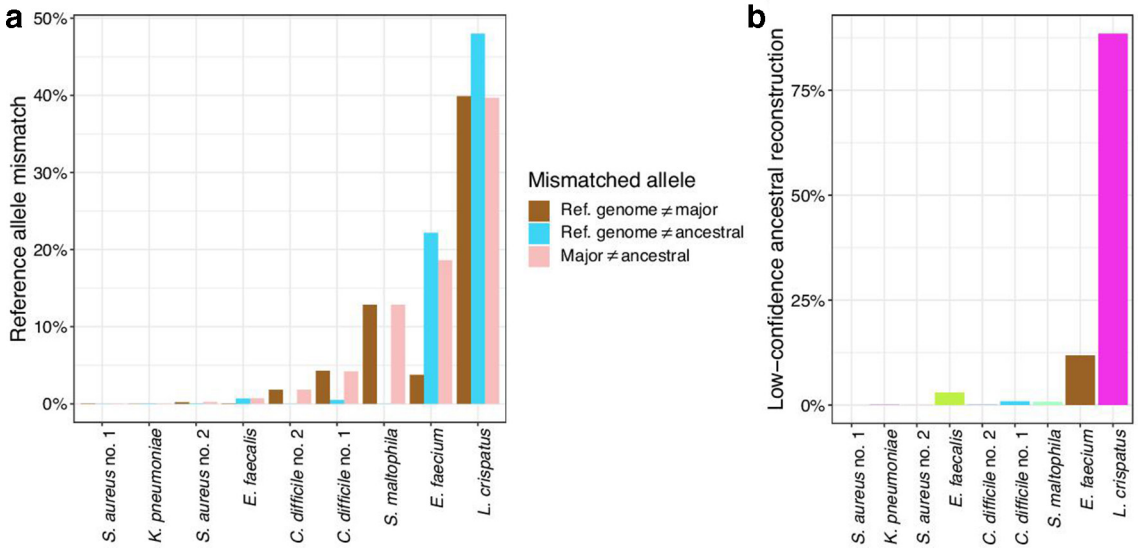

Fig. 3. Methods to determine the reference allele identify different alleles. (a) The fraction of variant positions where the identified reference allele varies between two methods. Only high-confidence ancestral reconstruction sites (>=87.5 % confidence in the ancestral root allele by maximum likelihood) are included. (b) Fraction of low-confidence ancestral reconstruction sites for each dataset (<87.5 % confidence in the ancestral root allele by maximum likelihood).

referencing methods are: the reference-genome allele from variant calling, the major allele or the ancestral allele (Fig. 1c). The reference-genome allele is the allele found in the reference genome when using a reference-genome-based variant-calling approach. The major allele is the most common allele at a given locus in the dataset. Neither of these methods encode the alleles with a consistent evolutionary direction. The ancestral allele is the allele inferred to have existed at the most recent common ancestor of the dataset. Given confident ancestral reconstruction, using the ancestral allele as the reference allele allows for a uniform evolutionary interpretation of variants: there is a consistent direction of evolution in that all mutations have arisen over time. We found that the three different methods for identifying the reference allele frequently identify different alleles (range across datasets 0–58 %; Fig. 3a). Thus, using the reference-genome allele or the major allele as the reference allele will not always maintain a consistent direction of evolution for each allele in a gene, obscuring interpretation when grouping variants into genes.

Although ancestral reconstruction is the most interpretable option for reference allele choice, this method is not feasible for some datasets. For example, sometimes we cannot confidently predict the most likely ancestral root allele for many loci, as in the *Lactobacillus crispatus* dataset (Fig. 3b); in this case, it is not a reliable method to define the reference allele. Other limitations of using the ancestral allele as the reference allele are that ancestral reconstruction requires an accurate phylogenetic tree and may be computationally intensive for large datasets. An alternative approach is to use the major allele as the reference allele as this method does not require a tree and thus avoids ancestral reconstruction. When the ancestral allele is not feasible, using the major allele is better than using the reference-genome allele when grouping variants into genes because using the major allele leads to less masking of variation at the gene level (Fig. S3).

Researchers are often interested in determining the heritability of a certain trait using shared variants (e.g. [47]). prewas allows users to re-reference variants and include multiallelic sites, which may influence the predicted heritability of a certain trait. We found no difference between heritability estimates using the different referencing methods, with and without multiallelic sites (Fig. S4). Note that one limitation of our analysis is the use of simulated phenotypes; results could differ for observed phenotype data.

Grouping variants into genes

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

It is important to ensure that variants in overlapping genes are assigned to each gene that the variant is in to prevent information loss and because the functional impact of a SNP

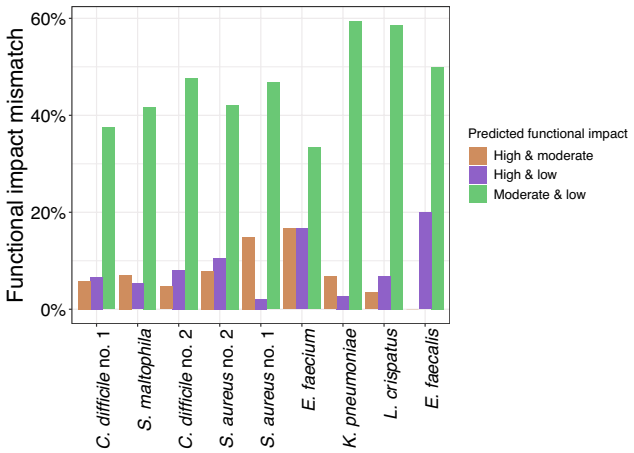

Fig. 4. SNPs in overlapping sites can have distinct functional impacts in each gene of the gene pair. The fraction of overlapping variant positions where the SNP has a different predicted functional impact in each of the two overlapping genes.

in one gene may be different than its impact on the other gene(s). There are many overlapping genes that share SNPs in each genome (Fig. S5). Furthermore, there are many sites where the SNP has a different functional impact in the two overlapping genes (cumulative range across datasets 50–70 %; Fig. 4). The functional impact of variants can be used to select what variants to include in a gene-based analysis. For instance, researchers could subset to include only those SNPs most likely to affect gene function (e.g. start loss and stop gain mutations).

Package description

We developed prewas to standardize the inclusion and representation of multiallelic sites, choice of reference allele and SNPs in overlapping genes (Fig. 1a) for downstream use in bGWAS analyses. Installation may be performed from GitHub (https://github.com/Snitkin-Lab-Umich/prewas). This R package is an easy-to-use tool with a function that minimally takes a multiVCF input file. The multiVCF encodes the variant nucleotide alleles for all samples and can be generated using bcftools merge [31]. The outputs of the prewas function are matrices of variant presence and absence with multi-line representation of multiallelic sites. Multiple optional files may be used as additional inputs to the prewas function: a phylogenetic tree, an outgroup and a GFF file. The phylogenetic tree may be added when the user wants to identify ancestral alleles for the allele referencing step. The GFF file contains information on gene location in the reference genome used to call variants and is necessary to generate a binary matrix of presence and absence of variants in each gene. Variants in overlapping genes are assigned to both genes. The matrix outputs from prewas can be directly input into bGWAS tools such as treeWAS [4].

Generating a binary variant matrix including multiallelic sites (Fig. 1b)

The multiVCF file is read into prewas and converted into an allele matrix with single-line representation of each genomic position. prewas handles multiVCF files with SNPs and/or indels. Next, a reference allele is chosen for each variant position (see section below). Then, the reference alleles are used to convert the allele matrix 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. Missing alleles are also coded as 0. This binary matrix is output by prewas. Note that bcftools (bcftools norm -m<multiVCF>) [31] also has the functionality to split multiallelic sites into biallelic sites given a multiVCF file. We also included this functionality in the prewas package to provide a standardized, all-in-one R package to pre-process data before bGWAS.

Identifying reference alleles (Fig. 1c)

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

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.

Major allele approach

The reference allele may also be defined as the major allele at each genomic position. In this case, the most common allele in the dataset is the reference allele. This choice improves the performance speed of prewas as compared to using the ancestral allele at the cost of evolutionary interpretability.

For both approaches missing data is treated as a fifth allele at each site. If, when using the ancestral reconstruction approach, the root allele is identified as missing data, then that site is removed. If, when using the major allele approach, the major allele for a site is missing data, then that site is removed. Furthermore, while prewas references alleles, users should be aware that some downstream bioinformatic software may automatically recode alleles.

Grouping variants by gene (Fig. 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.

Users may want to only aggregate variants into groups if the variants have similar impacts on the encoding protein. When prewas is supplied with a multiVCF that contains SnpEff predicted functional-impact [33] annotations then variants can be filtered by predicted functional impact (HIGH, MODERATE, LOW and MODIFIER) prior to a gene-based analysis. prewas will produce multiple binary gene matrices: one matrix with all variants, one matrix subset to each functional impact, and an optional matrix made from a user provided subset of predicted functional impacts (e.g. HIGH and MODERATE together).

Resource utilization

prewas runs most quickly and with the lowest memory requirements when using the major allele as the reference allele as ancestral reconstruction is a computationally intensive process (Fig. S6). Tree building within prewas also increases resource utilization. Ancestral reconstruction supports multithreading, so the use of multiple cores shortens run times.

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 enable microbial genomics researchers to maximize the power and interpretability of their bGWAS.

Funding information
K.S. was supported by the National Institutes of Health (T32GM007544). E.S.S. and K.S. were supported by the National Institutes of Health (1U01Al124255). S.N.T. was supported by the Molecular Mechanisms of Microbial Pathogenesis training grant (NIH T32 AI007528). Z.L. received support from the National Science Foundation Graduate Research Fellowship Program 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 the National Science Foundation.

Acknowledgements
We thank Shawn Hawken for coining the name prewas.

Author contributions
The study was conceptualized by K.S., Z.L., S.N.T. and E.S.S. Software design and implementation, formal analysis, original draft preparation and visualization were performed by K.S., Z.L. and S.N.T. Data was curated by A.P., K.S., Z.L. and S.N.T. All authors performed editing and review, and E.S.S. supervised the project.

Conflicts of interest
The authors declare that there are no conflicts of interest.

Data Bibliography
See Table S1.

References

1. Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: lessons from human GWAS. *Nat Rev Genet* 2017;18:41–50.
2. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol* 2016;17:238.
3. Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics* 2018;34:4310–4312.
4. Collins C, Didelot X. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLoS Comput Biol* 2018;14:e1005958.
5. Earle SG, Wu C-H, Charlesworth J, Stoesser N, Gordon NC *et al*. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat Microbiol* 2016;1:16041.
6. Lees JA, Vehkala M, Välimäki N, Harris SR, Chewapreecha C *et al*. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat Commun* 2016;7:1–8.
7. Jaillard M, Lima L, Tournoud M, Mahé P, van Belkum A, van BA *et al*. A fast and agnostic method for bacterial genome-wide association studies: bridging the gap between k-mers and genetic events. *PLoS Genet* 2018;14:e1007758.
8. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR *et al*. Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nat Genet* 2013;45:1183–1189.
9. Alam MT, Petit RA, Crispell EK, Thornton TA, Conneely KN *et al*. Dissecting vancomycin-intermediate resistance in Staphylococcus aureus using genome-wide association. *Genome Biol Evol* 2014;6:1174–1185.
10. Chewapreecha C, Marttinen P, Croucher NJ, Salter SJ, Harris SR *et al*. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS Genet* 2014;10:e1004547.
11. Desjardins CA, Cohen KA, Munsamy V, Abeel T, Maharaj K *et al*. Genomic and functional analyses of Mycobacterium tuberculosis strains implicate ALD in D-cycloserine resistance. *Nat Genet* 2016;48:544–551.
12. Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z *et al*. Predicting the virulence of MRSA from its genome sequence. *Genome Res* 2014;24:839–849.
13. Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW *et al*. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Front Genet* 2015;6:235.
14. Yoshimura D, Kajitani R, Gotoh Y, Katahira K, Okuno M *et al*. Evaluation of SNP calling methods for closely related bacterial isolates and a novel high-accuracy pipeline: BactSNP. *Microb Genom* 2019;5:e000261.
15. Zhan X, Chen S, Jiang Y, Liu M, Iacono WG *et al*. Association analysis and meta-analysis of multi-allelic variants for large scale sequence data. *bioRxiv [Internet]* 2017.
16. Farhat MR, Freschi L, Calderon R, Ioerger T, Snyder M *et al*. Gwas for quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions. *Nat Commun* 2019;10:1–11.
17. Johnson ZI, Chisholm SW. Properties of overlapping genes are conserved across microbial genomes. *Genome Res* 2004;14:2268–2272.
18. Huvet M, Stumpf MPH. Overlapping genes: a window on gene evolvability. *BMC Genomics* 2014;15:721.
19. Carlson PE, Walk ST, Bourgis AET, Liu MW, Kopliku F *et al*. The relationship between phenotype, ribotype, and clinical disease in human Clostridium difficile isolates. *Anaerobe* 2013;24:109–116.

20. Saund K, Rao K, Young VB, Snitkin ES. Genetic determinants of trehalose utilization are not associated with severe *Clostridium difficile* infection [Internet]. *Infectious Diseases* 2019.
21. Mody L, Krein SL, Saint S, Min LC, Montoya A *et al*. A targeted infection prevention intervention in nursing home residents with indwelling devices: a randomized clinical trial. *JAMA Intern Med* 2015;175:714–723.
22. Mody L, Foxman B, Bradley S, McNamara S, Lansing B *et al*. Longitudinal assessment of multidrug-resistant organisms in newly admitted nursing facility patients: implications for an evolving population. *Clin Infect Dis* 2018;67:837–844.
23. Han JH, Lapp Z, Bushman F, Lautenbach E, Goldstein EJC *et al*. Whole-genome sequencing to identify drivers of carbapenem-resistant *Klebsiella pneumoniae* transmission within and between regional long-term acute-care hospitals. *Antimicrob Agents Chemother* 2019;63:e01622-19.
24. Bassis CM, Bullock KA, Sack DE, Saund K, Pirani A *et al*. Evidence that vertical transmission of the vaginal microbiota can persist into adolescence [Internet]. *Microbiology* 2019.
25. Sun Z, Harris HMB, McCann A, Guo C, Argimón S *et al*. Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. *Nat Commun* 2015;6:1–13.
26. Popovich KJ, Snitkin ES, Zawitz C, Aroutcheva A, Payne D *et al*. Frequent methicillin-resistant Staphylococcus aureus introductions into an inner-city jail: indications of community transmission networks. *Clin Infect Dis*;352.
27. Roach DJ, Burton JN, Lee C, Stackhouse B, Butler-Wu SM *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;11:e1005413.
28. Sichtig H, Minogue T, Yan Y, Stefan C, Hall A *et al*. FDA-ARGOS is a database with public quality-controlled reference genomes for diagnostic use and regulatory science. *Nat Commun* 2019;10:1–13.
29. Lira F, Berg G, Martínez JL. Double-face meets the bacterial world: the opportunistic pathogen *Stenotrophomonas maltophilia*. *Front Microbiol* 2017;8:2190.
30. Esposito A, Pompilio A, Bettua C, Crocetta V, Giacobazzi E *et al*. Evolution of *Stenotrophomonas maltophilia* in cystic fibrosis lung over chronic infection: a genomic and phenotypic population study. *Front Microbiol* 2017;8:1590.
31. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J *et al*. The sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
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;32:268–274.
33. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T *et al*. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 2012;6:80–92.
34. Paradis E, Schliep K. Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 2019;35:526–528.
35. Bengtsson H, R Core Team. future.apply: apply function to elements in parallel using futures [Internet]. 2019 [cited 2019 Dec 10]. Available from: https://CRAN.R-project.org/package=future.apply.
36. Schliep KP. phangorn: phylogenetic analysis in R. *Bioinformatics* 2011;27:592–593.
37. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 2012;3:217–223.
38. Knaus BJ, Grünwald NJ. vcfr: a package to manipulate and visualize variant call format data in R. *Mol Ecol Resour* 2017;17:44–53.
39. Wickham H, Averick M, Bryan J, Chang W, McGowan L *et al*. Welcome to the Tidyverse. *Journal of Open Source Software* 2019;4:1686.
40. Wickham H. Reshaping data with the reshape package. *J Stat Softw* 2007;21:1–20.
41. Kolde R. pheatmap: Pretty Heatmaps [Internet]. 2019 [cited 2019 Dec 10]. Available from: https://CRAN.R-project.org/package=pheatmap.
42. Xie Y. animation: an R Package for creating animations and demonstrating statistical methods. *J Stat Softw* 2013;53:1–27.
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 10]. Available from: https://bioconductor.org/packages/Biostrings/.
44. Anaconda. Anaconda | The World's Most Popular Data Science Platform [Internet]. Anaconda. [cited 2019 Dec 10]. Available from: https://www.anaconda.com/.
45. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI *et al*. Fast linear mixed models for genome-wide association studies. *Nat Methods* 2011;8:833–835.
46. Campbell IM, Gambin T, Jhangiani S, Grove ML, Veeraraghavan N *et al*. Multiallelic positions in the human genome: challenges for genetic analyses. *Hum Mutat* 2016;37:231–234.
47. Lees JA, Ferwerda B, Kremer PHC, Wheeler NE, Serón MV *et al*. Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. *Nat Commun* 2019;10:1–14.