MICROBIAL GENOMICS

METHOD

Saund and Snitkin, *Microbial Genomics*

DOI 10.1099/mgen.0.000469

Hogwash: three methods for genome-wide association studies in bacteria

Katie Saund[1] and Evan S. Snitkin[1,2,*]

Abstract

Bacterial genome-wide association studies (bGWAS) capture associations between genomic variation and phenotypic variation. Convergence-based bGWAS methods identify genomic mutations that occur independently multiple times on the phylogenetic tree in the presence of phenotypic variation more often than is expected by chance. This work introduces hogwash, an open source R package that implements three algorithms for convergence-based bGWAS. Hogwash additionally contains two burden testing approaches to perform gene or pathway analysis to improve power and increase convergence detection for related but weakly penetrant genotypes. To identify optimal use cases, we applied hogwash to data simulated with a variety of phylogenetic signals and convergence distributions. These simulated data are publicly available and contain the relevant metadata regarding convergence and phylogenetic signal for each phenotype and genotype. Hogwash is available for download from GitHub.

DATA SUMMARY

1. hogwash is available from GitHub under the MIT license (https://github.com/katiesaund/hogwash) and can be installed using the R commands

```
install.packages("devtools")
devtools::install_github("katiesaund/hogwash")
```

2. The simulated data used in this manuscript and the code to generate them are available from GitHub (https://github.com/katiesaund/simulate_data_for_convergence_based_bGWAS).

INTRODUCTION

Bacterial genome-wide association studies (bGWAS)

bGWAS infer statistical associations between genotypes and phenotypes. Seminal bGWAS papers identified novel variants associated with antibiotic resistance in *Mycobacterium tuberculosis* and host specificity in *Campylobacter* [1, 2]. Since then, there have been numerous applications of bGWAS that have further highlighted the potential of this approach to identify genetic pathways underlying phenotypic variation and provide insights into the evolution of phenotypes of interest. Association studies can use various genetic data types, including single-nucleotide polymorphisms (SNPs), k-mers, copy number variants, accessory genes, insertions and deletions. To improve the power and interpretability of bGWAS, inclusion criteria or weighting can be applied to these variants based on predicted functional impact, membership in pathways of interest, or other user preferences [3, 4]. Differences between human and bacterial genome-wide association studies (GWAS) have been reviewed extensively by Power *et al.* [5]. Of note, clonality and horizontal gene transfer complicate the application of human GWAS methodology to bacteria. However, bGWAS approaches can leverage unique features of bacterial evolution, including frequent phenotypic convergence and genotypic convergence, to identify phenotype–genotype correlations.

bGWAS software

Several different variations of bGWAS approaches have been applied, including methods for SNPs, accessory genes (Scory) [6], or k-mers (pyseer) [7], methods using

Received 19 April 2020; Accepted 16 October 2020; Published 18 November 2020

Author affiliations: [1]Department of Microbiology and Immunology, University of Michigan, Ann Arbor, Michigan, USA; [2]Department of Internal Medicine, Division of Infectious Diseases, University of Michigan, Ann Arbor, Michigan, USA.

***Correspondence:** Evan S. Snitkin, esnitkin@med.umich.edu

Keywords: GWAS; bacterial genomics; convergent evolution; software.

Abbreviation: bGWAS, bacterial genome-wide association study.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Six supplementary figures are available with the online version of this article.

regression (pyseer) [7, 8] or phylogenetic convergence (PhyC, treeWAS) [1, 9], and methods designed for humans (PLINK) [10] or specifically for bacteria [7, 9]. Differences between standard and convergence-based bGWAS were expertly reviewed by Chen and Shapiro [11]. Convergence-based methods identify events where a genomic mutation arises independently on different edges of a phylogeny more often in the presence of the phenotype of interest than expected by chance (Fig. 1c). Convergence-based methods can yield higher significance with a smaller sample size, but may fail to identify some statistical associations that traditional GWAS approaches would identify when the population is clonal [11]. Additionally, convergence-based methods are limited to smaller datasets because of their large memory requirements and computational time relative to traditional methods [12], but can surmount issues of clonality.

Impact Statement

We introduce hogwash, an R package with three methods for bacterial genome-wide association studies. There are two methods for handling binary phenotypes, including an implementation of PhyC [1], as well as one method for handling continuous phenotypes. We formulate novel indices quantifying the relationship between phenotype convergence and genotype convergence on a phylogenetic tree. These indices shape an intuitive understanding for the ability of hogwash to detect significant intersections of phenotype convergence and genotype convergence and how to interpret hogwash outputs.

Objective

As the popularity of bGWAS increases there is a need for more widely available software that addresses specific aspects of bacterial evolution and is appropriate for various kinds of datasets. This work introduces two novel methods for convergence-based bGWAS with these needs in mind: the Synchronous Test and the Continuous Test. Users can implement these methods using hogwash, a new R package available on GitHub. Hogwash also contains an implementation of PhyC, which is a bGWAS algorithm introduced by Farhat *et al.* [1]. The Synchronous Test is a stringent variation of PhyC, requiring a tighter relationship between the genotype and phenotype. We describe the algorithms and evaluate them on a set of simulated data. The hogwash wiki contains further explanation of bGWAS, a more conceptual introduction to these three algorithms and specific user instructions for hogwash on a set of data provided with the software package (https://github.com/katiesaund/hogwash/wiki).

Grouped genotype analysis

Some phenotypes are not well correlated with commonly occurring genomic variants. In these cases, rare variants may provide some additional explanation for trait variability. There are multiple approaches to studying rare variants, including various burden testing methods that can group loci into meaningful groups, such as mapping SNPs to genes [13, 14]. Analysing aggregated loci can improve both the interpretability of GWAS results and the power to

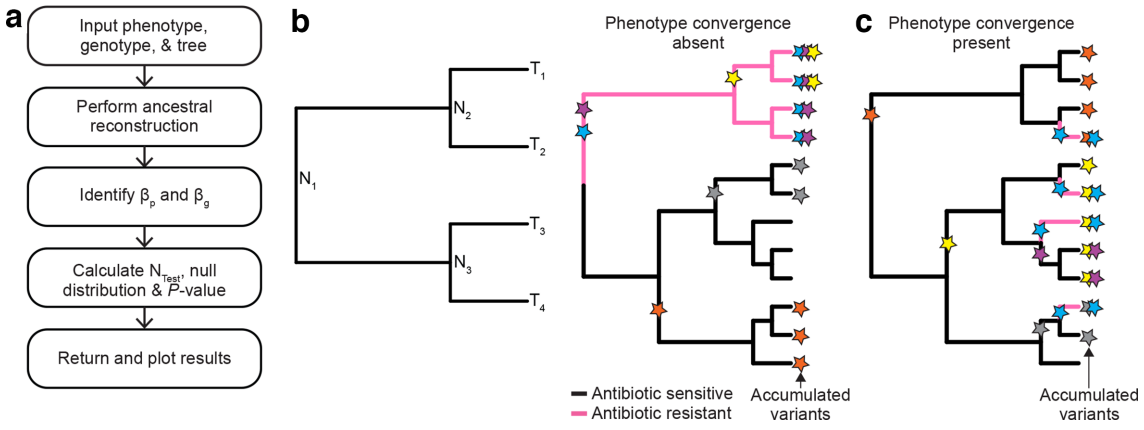

Fig. 1. Hogwash workflow, tree nomenclature and convergence example. (a) Software workflow. (b) In this example, phylogenetic tree N_1 is the root. Tree nodes are labelled N_1–N_3. Tree tips are labelled T_1–T_4. N_1 is a parent node to N_2 and N_3. N_2 is a child of N_1 and a parent to T_1 and T_2. Edges are lines connecting a parent node to a child node or a parent node to a tip. (c) A conceptual example of a phylogenetic tree with a phenotype that has arisen under two different scenarios. In the left tree antibiotic resistance, encoded by pink edges, arises once and therefore does not converge on the tree. In the right tree antibiotic resistance arises four times and therefore converges. Each coloured star represents a unique genomic variant, such as a single-nucleotide polymorphism (SNP), that arises. Stars on edges indicate the time at which the SNP is inferred to have arisen. The stars at each tip indicate the accumulated variants found in each sample. In both trees the blue variant occurs in 4/4 antibiotic-resistant isolates and 0/8 antibiotic-sensitive isolates. Convergence-based association methods could only ascertain the relationship between the blue variant and antibiotic resistance in the right tree.

detect associations [13–16]. Hogwash implements two such grouping approaches to improve convergence detection for related but weakly penetrant genotypes.

Data simulation

We evaluate hogwash results on simulated data generated to capture aspects of bacterial evolution pertinent to these bGWAS approaches. We simulated data with a range of phylogenetic signals and convergence distributions to highlight the critical impact of these features on bGWAS results. The simulated data are publicly available and could be used to compare the impact of convergence patterns within phenotypes and genotypes, and their intersection when benchmarking various convergence-based bGWAS methods.

Package description

We developed hogwash to allow users to implement three bGWAS methods, including an open source implementation of the previously described PhyC algorithm [1], and aggregate genotypes by user-defined groups of mutations. The hogwash function minimally requires a phenotype, a phylogenetic tree and a set of genotypes. An optional argument may be supplied to facilitate grouping genotypes. The genotypes and tree can be prepared from a multiVCF file by the variant preprocessing tool prewas [17]. Hogwash assumes that the genotype is encoded such that 0 refers to wild-type and 1 refers to a mutation and binary phenotypes are encoded such that 0 refers to absence and 1 refers to presence.

In brief, the hogwash workflow (Fig. 1a) begins with the user supplying a phenotype, a set of genotypes and a tree. Hogwash performs ancestral state reconstruction for the phenotype and genotypes to assign phenotype and genotype values to each tree edge (Fig. 1b). The interaction of the phenotype with the genotypes is uniquely defined for each of the three association tests. To establish the significance of the interaction, the genotypes are permuted and their intersection with the phenotype is recorded as a null distribution. Finally, we introduce an additional metric, ε, to capture the interaction between the convergence of the phenotype and genotypes.

Definitions

To describe the association algorithms, we introduce terms to characterize phenotypes and genotypes and their interactions. We evaluate node values in a phylogenetic tree through ancestral state reconstruction. β is vector where each element corresponds to an edge in this tree.

- $\dot{\beta}_p$ is a binary vector indicating phenotype presence, with a value of 1 for exactly the edges with a child node with value 1 and otherwise 0.
- $\overleftrightarrow{\beta_p}$ is a binary vector indicating phenotype transitions, with a value of 1 for exactly the edges where the parent differs from the child and otherwise 0.
- $\hat{\beta}_p$ is a continuous vector that has value $\Delta_{edge} = |phenotype_{parent\ node} - phenotype_{child\ node}|$ for each edge, where Δ_{edge} values are normalized from 0 to 1.
- $\overrightarrow{\beta_g^i}$ is a binary vector indicating a genotype arising on the tree. It has a value of 1 for exactly the edges where the parent node has value 0 and the child node has value 1, for each genotype i in the set of all genotypes.
- $\overleftrightarrow{\beta_g^i}$ is a binary vector indicating genotype transitions, with a value of 1 for exactly the edges where the parent differs from the child and otherwise 0, for each genotype i in the set of all genotypes.
- We define the elementwise sum of β as $\sum \beta$.

Our three methods use different combinations of β_p and $\underrightarrow{\beta_g}$. PhyC is concerned with presence and appearance ($\dot{\beta}_p$, $\overrightarrow{\beta_g^i}$). The Synchronous test is concerned with transitions ($\overleftrightarrow{\beta_p}$, $\overleftrightarrow{\beta_g^i}$). The Continuous Test is concerned with deltas and transitions ($\hat{\beta}_p$, $\overleftrightarrow{\beta_g^i}$).

The interaction of the phenotype and genotypes are summarized as N for each method.

- We define the number of edges where both a genotype arises and the phenotype is present as $N^i_{PhyC} = \sum \overrightarrow{\beta_g^i} \wedge \dot{\beta}_p$, for each genotype i in the set of all genotypes.
- We define the number of edges where both a genotype changes and the phenotype changes as $N^i_{Synchronous} = \sum \overleftrightarrow{\beta_g^i} \wedge \overleftrightarrow{\beta_p}$, for each genotype i in the set of all genotypes.
- We define the sum of the absolute value of phenotype change on only genotype transitions edges as $N^i_{Continuous} = \overleftrightarrow{\beta_g^i}^{\top} \hat{\beta}_p$, for each genotype i in the set of all genotypes.

PhyC

PhyC is a convergence-based bGWAS method introduced by Farhat *et al.* [1] that identified novel antibiotic resistance-conferring mutations in *M. tuberculosis*. To our knowledge, the original PhyC code is not publicly available, but the algorithm is well described in the original paper. The algorithm addresses the following question: does the genotype transition from wild-type, 0, to mutant, 1, occur more often than expected by chance on tree edges where the phenotype is present, 1, than where the phenotype is absent, 0? By requiring the overlap of the phenotype with the genotype transition, instead of genotype presence, associations are not inflated by clonal sampling and thus this approach controls for population structure. We implement the PhyC algorithm as described by Farhat *et al.* [1].

For permutation tests to determine the significance of associations genotype transitions are randomized on the tree with probability proportional to the branch length. The number of

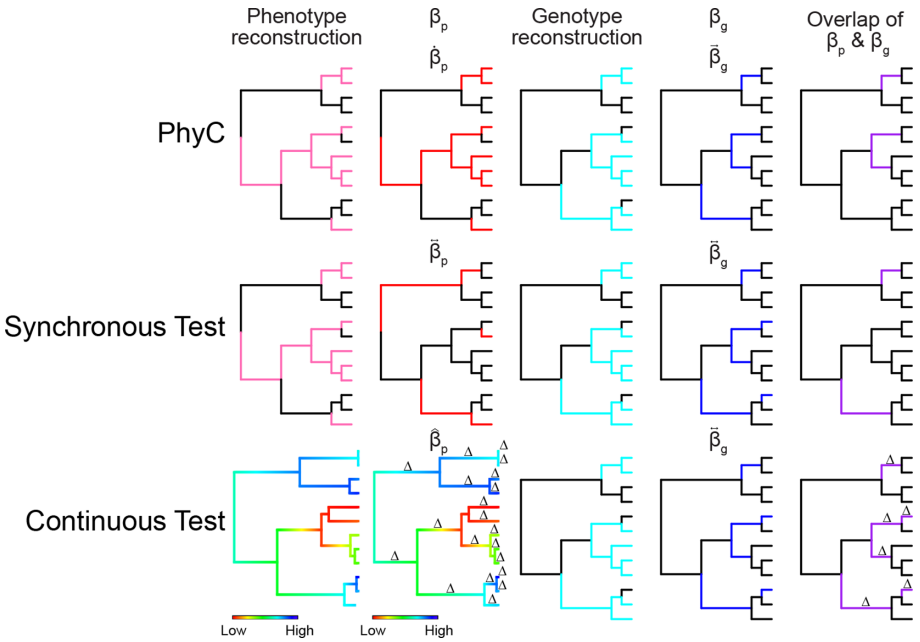

Fig. 2. Schematic of PhyC, Synchronous, and Continuous Tests. For all binary trees black indicates 0 and a solid colour indicates 1. The phenotype reconstruction indicates the ancestral state reconstruction for a simulated phenotype; either binary for PhyC and Synchronous test or a range of values for the Continuous Test. The β_p indicates the test-specific β_p value taken on each tree edge; 0 or 1 for PhyC and the Synchronous test or the normalized Δ_{edge} for the Continuous Test. The genotype reconstruction column indicates the ancestral state reconstruction for a simulated genotype; the values are 0 or 1 in all algorithms. The β_g indicates the test-specific β_g value taken on each tree edge; the values are 0 or 1 in all algorithms. The overlap of β_p and β_g represents the components of N_{test}. The variables β_p, β_g, and N_{test} are described in the definitions section.

edges where the permuted genotype mutation intersects with phenotype presence edges is recorded for each permutation; these permuted N^i_{PhyC} values create a null distribution. An empirical *P*-value is calculated based on the observed N^i_{PhyC} as compared to the null distribution.

Our PhyC implementation (Fig. 2) has several important differences from the original paper. First, multiple test correction in hogwash is performed with false discovery rate instead of the more stringent Bonferroni correction. Second, hogwash reduces the multiple testing burden by testing only those genotype–phenotype pairs for which convergence is detectable; genotypes with $\sum \vec{\beta^i_g} < 2$ are excluded and genotype–phenotype pairs with $N^i_{PhyC} < 2$ are assigned a *P*-value of 1. Third, ancestral state reconstruction for genotypes and phenotypes is performed using only maximum likelihood. Finally, users sacrifice some robustness in exchange for ease of use by supplying one phylogenetic tree instead of three.

Synchronous Test

This test (Fig. 2) is an extension of PhyC but requires more stringent association between the genotype and phenotype. The Synchronous Test addresses the question: do genotype transitions occur more often than expected by chance on phenotype transition edges than on phenotype non-transition edges? As in PhyC, the Synchronous Test is only appropriate for binary phenotypes.

Genotypes with $\sum \overleftrightarrow{\beta^i_g} < 2$ are removed, genotype–phenotype pairs with $N^i_{Synchronous} < 2$ are assigned a *P*-value of 1, and the remaining genotypes are permuted and a null distribution of $N^i_{Synchronous}$ is calculated to determine the significance of each genotype.

This test is similar to the Simultaneous Score in treeWAS [9]. The Simultaneous Score is derived from the number of edges on the tree where the genotype and phenotype transition in the same direction (both have a parent node of 0 and a child node of 1 or both have a parent node of 1 and child node of 0). In contrast, our newly developed Synchronous Test allows for the phenotype and genotype transition directions to mismatch, thus allowing for a genotype to have opposing effects on a phenotype. Such opposing effects of a genotype on a phenotype could arise when grouping mutations in the same gene that differentially impact gene function, or even for an individual mutation whose phenotypic impact may be dependent on genetic background.

Continuous test

The Continuous Test (Fig. 2) is a novel application of a convergence-based GWAS method to continuous

phenotypes. The Continuous Test addresses the question: does the phenotype change more than expected by chance on genotype transition edges than on genotype non-transition edges?

As above, the genotypes with $\sum \overleftrightarrow{\beta_g^i} < 2$ are removed; the remaining genotypes are permuted and a null distribution of the $N^i_{Continuous}$ is calculated to determine the significance of each genotype.

User inputs

The user must provide a phylogenetic tree, a set of genotypes and a phenotype. The user may optionally provide a key that maps individual genomic loci into groups in order to use hogwash's grouping feature. For a detailed description of the user inputs please see the Document S1.

Hogwash outputs

The package produces two files per test: data (.rda) and plots (.pdf). The data file contains many pieces of information, including *P*-values for each tested genotype. The plots are described in the Results section.

Grouping feature

To identify an association between a genomic variant and a phenotype, hogwash requires that a variant occur in multiple different lineages. Hogwash may classify some causal variants as independent of a phenotype if they are weakly penetrant. To surmount this issue, related genomic variants may be aggregated to capture larger trends at the grouped level. For example, a user may apply this method to group only nonsynonymous SNPs by gene to use hogwash to detect associations between the mutated gene and the phenotype. Grouping related variants can improve power through a reduction in the multiple testing correction penalty. However, the power benefits are dependent on grouping variants with similar effect directions.

By default, hogwash implements the grouping features by first performing ancestral state reconstruction for each individual locus (Fig. 3). Then those loci are joined as indicated in the user supplied key. Grouped loci with $\sum \beta_g^i < 2$ are excluded from analysis. After this point hogwash runs as previously described for non-grouped genotypes. Alternatively, users may group together related genomic variants prior to ancestral reconstruction (Supplementary Methods). The two grouping approaches are compared in Fig. S1 (available in the online version of this article).

METHODS

Data simulation

Trees

We simulated 8 random coalescent phylogenetic trees with 100 tips each; 4 trees were used for the Continuous Test and 4 trees were used for the binary tests.

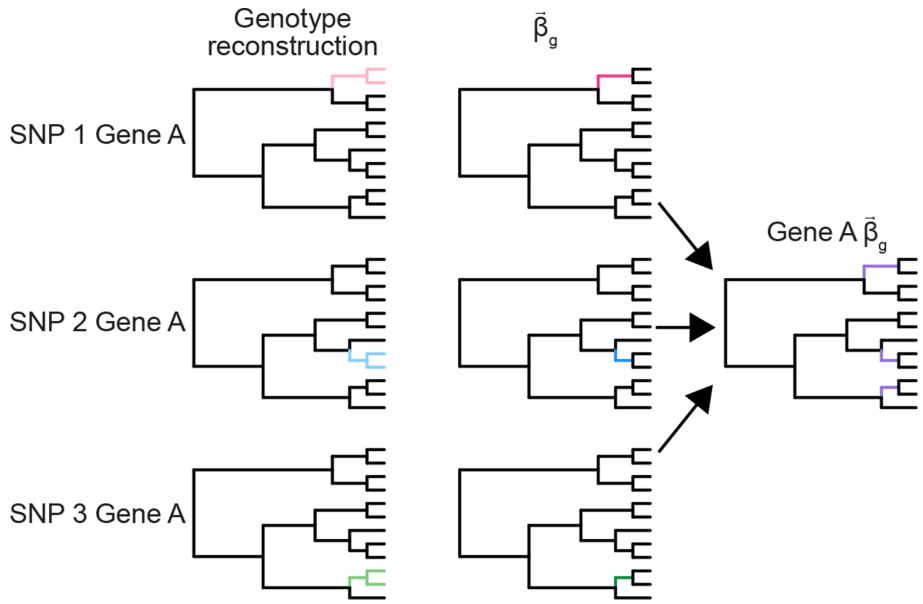

Fig. 3. Example of hogwash grouping feature on simulated data. In this case, three SNPs are found in the same gene (gene A). No individual SNP is convergent on the tree. Hogwash performs ancestral state reconstruction on each SNP. The edges where SNP presence is inferred are coloured. Next, hogwash identifies the transitions for each SNP (coloured edges). Finally, hogwash combines the three SNPs transitions together to create the gene A transitions (purple edges). When the SNPs are grouped into gene A the genotype converges on the tree. In this example, pre-ancestral reconstruction and post-ancestral reconstruction grouping results are identical. See Fig. S1 for scenarios illustrating differences in the two grouping approaches.

Tree edge filtering

Low-confidence edges are defined as those edges with low bootstrap support (default <70%), those that are more than 10% of the total tree length, or those with low genotype or phenotype ancestral reconstruction support (maximum likelihood <0.875). Low-confidence edges are ignored during permutation testing.

Phenotypes

Motivation for simulating phenotypes under two evolutionary models

For each tree we simulated phenotypes under different evolutionary models: either Brownian motion or white noise. A phenotype modelled well by Brownian motion follows a random walk along the tree. A phenotype modelled well with white noise appears to be independent of tree structure and may suggest a role for horizontal gene transfer, gene loss, or convergent evolution [18]. A white noise phenotype may be better suited to the hogwash algorithms than a phenotype modelled by Brownian motion, given the requirement for phylogenetic convergence.

Calculation of phylogenetic signal

Phylogenetic signal is a metric that captures the tendency for closely related samples on a tree to be more similar than random samples. Phylogenetic signal is calculated by different metrics for continuous and binary traits; continuous traits are measured by λ, while binary traits are measured by D (Fig. S2). A continuous phenotype that is modelled well by Brownian motion has a λ near 1 while a white noise phenotype has a λ near 0 [19]. In contrast, a binary phenotype that is modelled well by Brownian motion has a D near 0 while a white noise phenotype has a D near 1 [20].

Simulation of phenotypes on trees

For each tree we simulated four phenotypes fitting a Brownian motion model and four phenotypes fitting a white noise model. For phenotypes modelling Brownian motion, binary phenotypes were restricted to $-0.05<D<0.05$ and continuous phenotypes to $0.95<\lambda<1.05$. For phenotypes modelling white noise, binary phenotypes were restricted to $0.95<D<1.05$ and continuous phenotypes to $-0.05<\lambda<0.05$.

Genotypes

For each simulated tree a set of unique binary genotypes were generated. We generated genotypes that span a range of phylogenetic signals, degree of similarity to the phenotype and prevalence.

Genotypes used in PhyC and the Synchronous Test

First, 25000 binary genotypes were generated using ape::rTraitDisc; these genotypes have a range of phylogenetic signals [21]. Second, these genotypes were duplicated and randomized with the following approach to reduce their phylogenetic signal: one quarter had 10% of tips changed, one quarter had 25% of tips changed, one quarter had 40% of tips changed and one quarter were entirely redistributed. Third, we removed any simulated genotypes present in 0, 1, $N-1$, or N samples. Fourth, we subset the genotypes to keep only unique presence/absence patterns. Fifth, we subset genotypes to only those within a range of $-1.5<D<1.5$. These filtering steps reduced the dataset size (range 2214–2334).

Genotypes used in the Continuous Test

In addition to the five steps above we added two more data generation steps. First, we made all possible genotypes based on the rank of the continuous phenotype. Second, we made genotypes based on which edges of the tree had high Δ_{edge}. The filtering steps reduced the dataset size (range 1234–1310).

Hogwash on simulated data

We ran hogwash for each of the tree–phenotype–genotype sets. In addition to generating P-values for each tested genotype, hogwash also reports convergence information. We ran hogwash with the following settings for single-locus analysis: permutations=50000; false discovery rate=0.0005 (binary), 0.05 (continuous); bootstrap value=0.70; no genotype grouping key was provided. For grouped analyses the settings were identical except that a grouping key was generated, and hogwash was run with both grouping methods (pre- and post-ancestral reconstruction). For the grouped analyses only PhyC was run on simulated Brownian motion phenotype 1, simulated genotype 1 and simulated tree 1. The grouping key assigned approximately three unique simulated variants to each created 'gene', resulting in approximately one-third as many input genotypes when compared to the single-locus analysis.

Calculation of ε

We introduce ε to quantify the degree of shared phenotype convergence and genotype convergence. Low values of ε indicate a lack of overlap in the edges where the phenotype and genotype converge. High values of ε indicate many instances of overlap in the edges where the phenotype and genotype converge. By reducing these patterns of convergence into a simple number, ε, we can more easily contextualize convergence-based bGWAS results. We define an ε for each algorithm.

- $\varepsilon^{i}_{PhyC} = \frac{2 \times N^{i}_{PhyC}}{\sum \vec{\beta}^{i}_{g} + \sum \dot{\beta}_{p}}$, for each genotype i in the set of all genotypes.
- $\varepsilon^{i}_{Synchronous} = \frac{2 \times N^{i}_{Synchronous}}{\sum \overleftrightarrow{\beta}^{i}_{g} + \sum \overleftrightarrow{\beta}_{p}}$, for each genotype i in the set of all genotypes.
- $\varepsilon^{i}_{Continuous} = \frac{N^{i}_{Continuous}}{\sum \overleftrightarrow{\beta}^{i}_{g} + \sum \dot{\beta}_{p} - N^{i}_{Continuous}}$, for each genotype i in the set of all genotypes.
- For each ε, $0 \leq \varepsilon \leq 1$.

Data analysis

Statistical analyses were conducted in R v3.6.2 [22]. The R packages used can be found in the simulate_data.yaml file on GitHub [21, 23–27] and can be installed using miniconda [28].

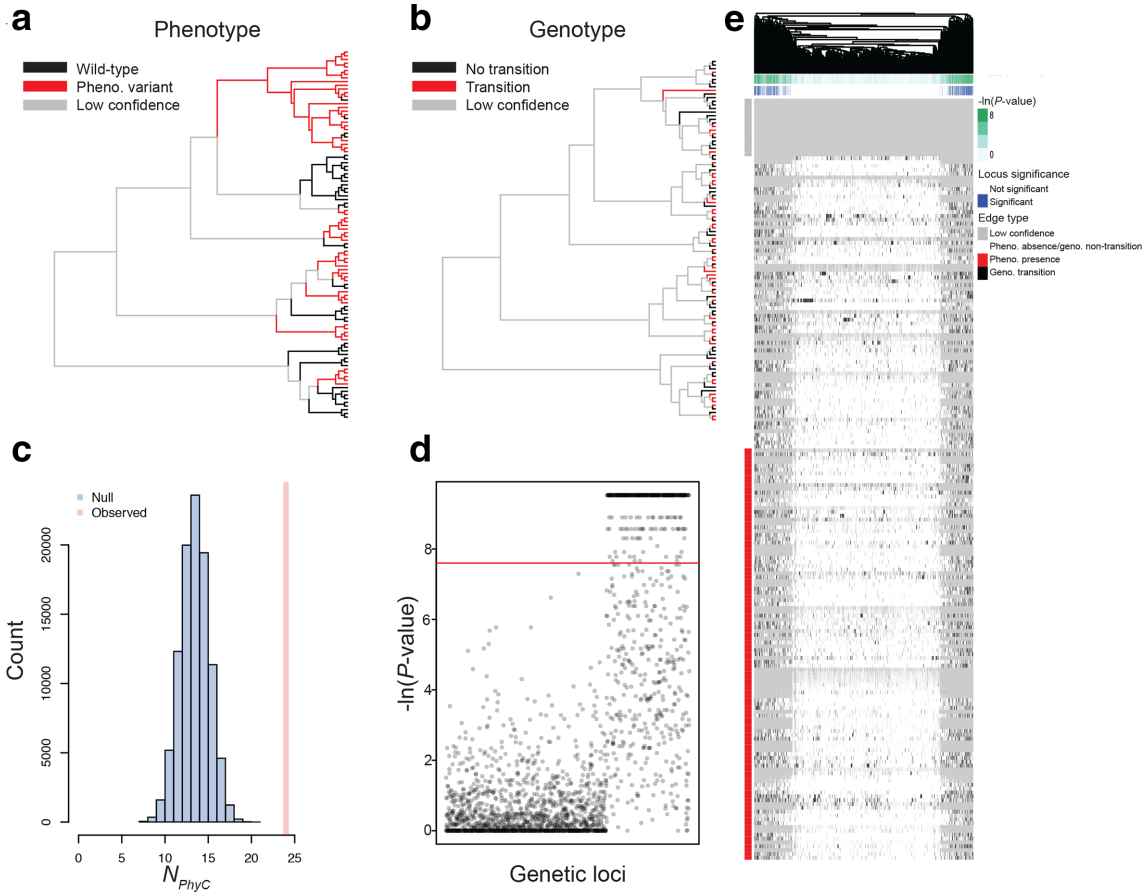

Fig. 4. Example output from hogwash PhyC results from simulated data. (a) Phenotype reconstruction. Edges with: phenotype presence in red; phenotype absent in black; low confidence in tree or low-confidence phenotype ancestral state reconstruction in grey. (b) Genotype transitions. Edges with: genotype mutations that arose in red; genotype mutation did not arise in black; low confidence in tree or low-confidence genotype ancestral state reconstruction in grey. (c) Null distribution of N_{PhyC} . (d) Manhattan plot. The genetic loci were simulated to achieve a range of phylogenetic signals. The leftmost two-thirds of genetic loci were simulated under Brownian motion models (mean D=0.16), while the remaining third were modelled by white noise (mean D=0.99). (e) Heatmap with tree edges in the rows and genotypes in the columns. The genotypes are hierarchically clustered . The genotypes are classified as being a transition edge in black or non-transition edge in white. The column annotations pertain to loci significance; green indicates the *P*-value, while blue indicates that the *P*-value is more significant than the user-defined threshold. The row annotation classifies the phenotype at each edge; red indicates phenotype presence and white indicates phenotype absence. Grey indicates a low-confidence tree edge; low confidence can be due to low phenotype ancestral state reconstruction likelihood, low genotype ancestral state reconstruction likelihood, low tree bootstrap value, or long edge length.

RESULTS

Motivation for evaluating Hogwash on simulated data

Given our lack of comprehensive knowledge of the genetic variation contributing to any phenotype, it is not feasible to quantify sensitivity/specificity on real data. We therefore generated data that simulate genotype and phenotype distributions covering a spectrum of realistic evolutionary scenarios (spanning Brownian motion to white noise). Our goal is not to validate the premise of using convergence-based approaches, as these have been previously shown to provide useful biological insights, but rather to understand how our approach detects convergence for phenotypes with different evolutionary regimes. The following analyses can guide users in the appropriate use cases and the applicability of this method to their data. In particular, these results provide context for interpreting the strength of the observed associations.

Hogwash output for simulated data

Hogwash outputs two sets of results: a data file and a PDF file with plots. Each run of PhyC produces at least three plots: the phenotype reconstruction (Fig. 4a), a Manhattan plot (Fig. 4d) and a heatmap of genotypes (Fig. 4e). The phenotype presence is highlighted on the tree (Fig. 4a). The Manhattan plot shows the distribution of *P*-values from the hogwash run (Fig. 4d). The heatmap shows the genotype reconstruction and phenotype reconstruction for each tree edge (rows) and

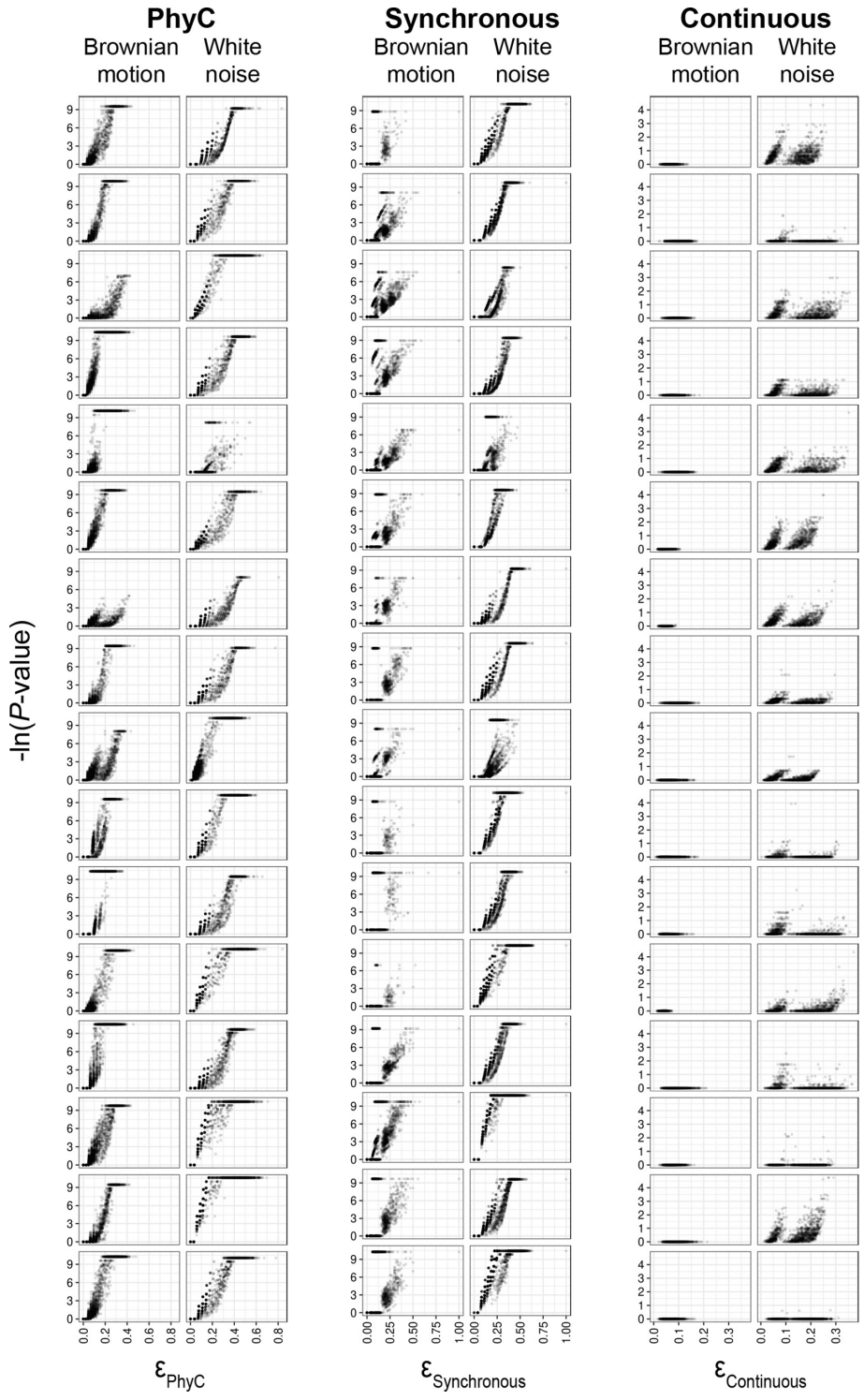

Fig. 5. High ε values correlate with increased significance. Each plot is a tree–phenotype pair. Each point represents one genotype–phenotype pair. Brownian motion and white noise refer to the evolutionary regime modelled by the phenotype. The genotypes span a range of phylogenetic signals.

Table 1. Mean Spearman's rank correlation coefficient for -ln(*P*-value) versus ε from hogwash run on simulated data. The ρ could not be calculated for the results from the Continuous Test on the Brownian motion phenotypes because, after multiple testing correction, all *P*-values are identical

	Phenotype	
	Brownian motion	White noise
PhyC	0.91	0.93
Synchronous Test	0.60	0.94
Continuous Test	NA	0.08

genotype (columns) (Fig. 4e). The genotypes are clustered by the presence/absence pattern. Two additional plots are produced for each genotype that is significantly associated with the phenotype: a phylogenetic tree showing the genotype transition edges (Fig. 4b) and the null distribution of N_{PhyC} (Fig. 4c). As expected, the two grouping approaches identified different associations as compared to non-grouped PhyC results (Fig. S3).

The Synchronous Test and Continuous Test output plots reflect their test-specific β and N definitions (Figs S4 and S5). Running hogwash on 100 samples required <3 h and <2 GB of memory for binary data and <5 h and <2 GB of memory for continuous data (Fig. S6).

Hogwash evaluation on simulated data

To help users identify optimal use cases and also interpret hogwash results we describe the behaviour of hogwash on simulated data. We note that this assessment is not meant to convey performance in the sense of calculating sensitivity and specificity, but rather evaluate whether hogwash can robustly detect the association between phenotypic and genotypic convergence. To guide our assessment, we compared the relationship between the *P*-value and ε values produced by hogwash on sets of simulated data constructed using different evolutionary models (Fig. 5). ε is a quantification of the relationship between phenotype convergence and genotype convergence. Low ε values indicate little to no intersection of phenotype convergence and genotype convergence, while higher ε values indicate their increased intersection. The ε value is always a fraction between 0 and 1 and therefore obscures information about the sample size; to account for the number of samples in the tree we recommend always interpreting ε value for any locus with its *P*-value.

For binary phenotypes, we observe an overall strong positive association between -ln(*P*-value) and ε, demonstrating that as the intersection of phenotype convergence and genotype convergence increase, hogwash predicts that it is less likely that they intersect due to chance (Table 1). In other words, below a certain ε_{binary} threshold (ε_{binary} is ε_{PhyC} or $\varepsilon_{Synchronous}$), hogwash attributes the association between the genotype convergence and phenotype convergence to chance; from Fig. 5 the user can get a sense for the range of this ε_{binary} threshold under different evolutionary regimes.

For the simulated continuous data an $\varepsilon_{Continuous}$ threshold that separates meaningful genotype–phenotype associations from associations by chance is less apparent. Higher ε, low significance values demonstrate that some overlap of β_g and β_p is likely by chance given the data. Low ε, high significance genotype–phenotype pairs demonstrate that sometimes small amounts of β_g and β_p overlap are unlikely, but that does not necessarily suggest that these hits are the best candidates for *in vitro* follow-up. We suspect that these associations are largely driven by poor exploration of the sampling space, despite running many permutations, because of the edge length-based sampling probability of the permutation method. Therefore, it is essential that *P*-values be interpreted within the context of ε. Notably, the Continuous Test was only able to detect significant genotype–phenotype associations for phenotypes modelled by white noise, suggesting that this method is particularly sensitive to the phenotype's evolutionary model. We observe for both the binary and Continuous Tests that ε is more tightly correlated with -ln(*P*-value) for phenotypes characterized by white noise than by Brownian motion (Table 1), indicating that hogwash performs better under a white noise model. Therefore, we suggest using the report_phylogenetic_signal function on the phenotype prior to running hogwash to ascertain the appropriateness of these algorithms for the dataset.

DISCUSSION

We have developed two algorithms for convergence-based bGWAS that are particularly well suited for phenotypes modelled by white noise. Hogwash is straightforward to instal in R, accepts easy-to-format data inputs (described in detail on the wiki), and provides publication ready plots of the GWAS results. Hogwash also implements grouping features to aggregate related genomic variants to increase detection of convergence for weakly penetrant genotypes. Hogwash is best used for datasets comprising binary or continuous phenotypes, phenotypes fitting white noise models, situations where convergence may occur at the level of genes or pathways and with datasets whose size can be accommodated given the time and memory constraints of convergence methods.

The results of running hogwash on simulated data suggest that after a certain ε threshold, it unlikely that the intersection between phenotype convergence and genotype convergence occurs by chance, particularly for white noise phenotypes. Given the variability in results within each method, as shown in Fig. 5, users may want to contextualize the statistical significance of the tested genetic loci with the amount of convergence possible for any one particular dataset; to facilitate this the hogwash output includes both *P*-values and ε.

The simulated dataset presented here is published to serve as a resource or template for future work focused on benchmarking convergence-based bGWAS software, as such a dataset has not yet, as far as we are aware, been made available [29]. The simulated dataset is available on GitHub and includes convergence information for each phenotype and genotype and their intersection.

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).

Acknowledgements
We thank Brad Saund for his help in formalizing the continuous algorithm ε definition.

Author contributions
K. S. and E. S. S., conceptualized the project and edited the manuscript. K. S., designed and implemented the software, performed the analysis, prepared the original draft, and visualized the data. E. S. S. supervised the project.

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

References

1. 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.
2. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA *et al*. Genome-Wide association study identifies vitamin b5 biosynthesis as a host specificity factor in Campylobacter. *Proc Natl Acad Sci U S A* 2013;110:11923–11927.
3. Sveinbjornsson G, Albrechtsen A, Zink F, Gudjonsson SA, Oddson A *et al*. Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat Genet* 2016;48:314–317.
4. Hendricks AE, Bochukova EG, Marenne G, Keogh JM, Atanassova N *et al*. Rare variant analysis of human and rodent obesity genes in individuals with severe childhood obesity. *Sci Rep* 2017;7:1–14.
5. Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: lessons from human GWAS. *Nat Rev Genet* 2017;18:41-50.
6. 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.
7. 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.
8. 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.
9. 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.
10. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;81:559–575.
11. Chen PE, Shapiro BJ. The advent of genome-wide association studies for bacteria. *Current Opinion in Microbiology*, 25; 2015. pp. 17–24.
12. Corander J, Croucher NJ, Harris SR, Lees JA, Tonkin-Hill G. Bacterial Population Genomics. *Handbook of Statistical Genomics*. Wiley; 2019. pp. 997–1020.
13. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* 2014;95:5–23.
14. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 2008;83:311–321.
15. Mooney MA, Wilmot B. Gene set analysis: a step-by-step guide. *Am J Med Genet B Neuropsychiatr Genet* 2015;168:517–527.
16. White MJ, Yaspan BL, Veatch OJ, Goddard P, Risse-Adams OS *et al*. Strategies for pathway analysis using GWAS and WGS data. *Curr Protoc Hum Genet* 2019;100:e79.
17. Saund K, Lapp Z, Thiede SN, Pirani A, Snitkin ES. prewas: data preprocessing for more informative bacterial GWAS. *Microb Genom* 2020;6 .
18. Van Assche A, Álvarez-Pérez S, de Breij A, De Brabanter J, Willems KA *et al*. Phylogenetic signal in phenotypic traits related to carbon source assimilation and chemical sensitivity in Acinetobacter species. *Appl Microbiol Biotechnol* 2017;101:367–379.
19. Pagel M. Inferring the historical patterns of biological evolution. *Nature* 1999;401:877–884.
20. Fritz SA, Purvis A. Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. *Conserv Biol* 2010;24:1042–1051.
21. Paradis E, Schliep K. *Phylogenetics ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R*, 35; 2019. pp. 526–528.
22. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2018.
23. Orme D. The caper package : comparative analysis of phylogenetics and evolution in R. R Packag version 05, 2 2013:1–36.
24. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol* 2012;3:217–223.
25. Wickham H. tidyverse: Easily Install and Load the "Tidyverse." 2017.
26. Wickham H, Seidel D. Scales: scale functions for visualization 2019.
27. Auguie B. gridExtra: Miscellaneous Functions for "Grid". *Graphics* 2017.
28. Anaconda. Data science technology forgroundbreaking research.a competitive edge.a better world.human sensemaking. [cited 2020 Feb 21]. Available from: https://www.anaconda.com/.
29. Saber MM, Shapiro BJ. Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes. *Microb Genom* 2020;6.