Exploiting Homoplasy in Genome-Wide Association Studies to Enhance Identification of Antibiotic-Resistance Mutations in Bacterial Genomes

Evolutionary Bioinformatics
Volume 16: 1–16

DOI: 10.1177/1176934320944932

Yi-Pin Lai and Thomas R Ioerger

Department of Computer Science and Engineering, Texas A&M University, College Station, TX, USA.

ABSTRACT: Many antibacterial drugs have multiple mechanisms of resistance, which are often represented simultaneously by a mixture of resistance mutations (some more frequent than others) in a clinical population. This presents a challenge for Genome-Wide Association Studies (GWAS) methods, making it difficult to detect less prevalent resistance mechanisms purely through (weak) statistical associations. Homoplasy, or the occurrence of multiple independent mutations at the same site, is often observed with drug resistance mutations and can be a strong indicator of positive selection. However, traditional GWAS methods, such as those based on allele counting or linear regression, are not designed to take homoplasy into account. In this article, we present a new method, called ECAT (for Evolutionary Cluster-based Association Test), that extends traditional regression-based GWAS methods with the ability to take advantage of homoplasy. This is achieved through a preprocessing step which identifies hypervariable regions in the genome exhibiting statistically significant clusters of distinct evolutionary changes, to which association testing by a linear mixed model (LMM) is applied using GEMMA (a well-established LMM-based GWAS tool). Thus, the approach can be viewed as extending GEMMA from the usual site- or gene-level analysis to focusing on clustered regions of mutations. This approach was evaluated on a large collection of more than 600 clinical isolates of multidrug-resistant (MDR) *Mycobacterium tuberculosis* from Lima, Peru. We show that ECAT does a better job of detecting known resistance mutations for several antitubercular drugs (including less prevalent mutations with weaker associations), compared with (site- or gene-based) GEMMA, as representative of existing GWAS methods. The power of the multiphase approach in ECAT comes from focusing association testing on the hypervariable regions of the genome, which reduces complexity in the model and increases statistical power.

KEYWORDS: homoplasy, bacterial evolution, GWAS, mycobacterium tuberculosis, antibiotic-resistance mutations

RECEIVED: April 3, 2020. **ACCEPTED:** June 30, 2020.

TYPE: Antimicrobial Resistance - Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by a U19 CETR grant (AI109755) from NIH/NIAID.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Thomas R Ioerger, Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843, USA.
Email: ioerger@cs.tamu.edu

Introduction

Several methods have recently been developed for identifying genetic variants (mutations) associated with phenotypic traits in bacteria, such as drug resistance, host specificity, and virulence in bacteria.[1,2] This has led to the discovery of new mechanisms of resistance from the analysis of single-nucleotide polymorphisms (SNPs) in collections of genome sequences from drug-resistant clinical isolates of bacterial pathogens.[3-5] One approach to inferring the relationship between genotypes and phenotypes is Genome-Wide Association Studies (GWAS). A commonly used GWAS approach tests whether one allele at a polymorphic site is overrepresented in cases than controls using contingency table tests, such as chi-square test or Fisher exact test.[6] Another approach to GWAS uses linear regression to fit a predictive model for phenotypes by regressing against genotypes. The coefficients in the linear model (LM) reflect statistical correlations between the presence of an allele and the occurrence of a particular phenotype. Coefficients in the LM are fit for each SNP and drug combination in parallel, which are then tested for significance, for example, using a Wald test.[6,7] The GWAS methods have been adapted for bacteria by incorporating genetic relatedness relationships in the form of a kinship matrix as random effects in a linear mixed model (LMM), to account for the more clonal nature of prokaryotes over eukaryotes (ie, population structure).[8] Similarly, *bugwas* incorporates lineage-specific effects into an LMM by decomposing the kinship to principal components.[9] The lineage-specific effects could also be estimated by *pyseer* using multidimensional scaling of a distance matrix as a covariate in a regression model.[10]

As an alternative approach to GWAS, phylogeny-based approaches have also been developed, including PhyC,[11] PhyOverlap,[4] and TreeWAS.[12] The advantage of phylogeny-based methods over LMMs is that they look at individual mutation events based on a phylogenetic tree, accounting for the evolutionary relationships among strains, rather than just calculating statistics based on raw proportions (overlap) of strains with a particular allele. PhyC (Phylogenetic Convergence test) uses a permutation test to determine the significance of association of observed nucleotide changes at a site (on branches inferred in a phylogenetic tree) and acquisition of resistance (inferred at internal nodes using maximum parsimony) by computing how likely this observation would occur by chance.[11] Still, another approach that has been proposed is k-mer-based methods,[13-15] which analyze the association of short nucleotide fragments (20-30 bp [base pairs] sequences,

possibly including SNPs) with phenotypes. Machine Learning algorithms such as random forests and gradient-boosting have been used to identify k-mers that are significantly associated with drug resistance phenotypes and predictive of minimum inhibitory concentrations (MICs), as was applied to identify antibacterial resistance mutations in *Salmonella*.[15] In addition, rule-based learning models employing decision trees and logical rules are developed for k-mer-based bacterial GWAS.[16]

Most of the methods above focus on identifying the strongest effects independently and work best when there is a single (or dominant) well-defined genetic explanation for a given phenotype. However, drug resistance can be multifactorial. There can be multiple mechanisms of resistance (loss of activators, upregulation of efflux pumps or detoxification enzymes, adaptations in the metabolic network, changes in cell wall permeability, etc.),[17-19] along with compensatory mutations[20] and epistatic effects,[21] leading to a mixture of resistance-associated sites with weak signals (because, individually, they each are only a partial explanation of the phenotype). Using a "burden test" can help by combining multiple allelic sites within a gene, which are pooled before the association test.[22-24] This can strengthen the association by pooling together mutations at several distinct nucleotides (eg, corresponding to multiple amino acids in an active site) that can confer resistance, but a burden test also risks diluting the signal by combining with other nonresistance-related mutations in the gene.

An approach to increasing these signals and improving the detection of genotype-phenotype associations over these prior methods is to take into account *homoplasy*. Homoplasy occurs when the same mutation arises independently in different lineages (branches of the evolutionary tree). Homoplasy can be a strong indicator of positive selection.[25] For example, it has been widely observed that mutations in catalase (*KatG* S315T) and RNA polymerase (*RpoB* S450L) have occurred multiple times independently in *Mycobacterium tuberculosis*, even within outbreak regions, and are not just inherited through transmission of a single clone.[11,26,27] Grandjean et al detected homoplasic polymorphisms across the entire *M. tuberculosis* genome using a convergence method based on the disruption of a phylogenetic tree. Many well-known drug-resistant loci are observed as being homoplasic among the data set of a high proportion of multidrug-resistant (MDR) *M. tuberculosis* strains,[28] suggesting they are under selection pressure. Although not all homoplasy is due to drug resistance, we show it can be exploited to enhance association testing.

The LMM-based GWAS methods are not generally designed to take homoplasy into account. The statistical association between alleles and phenotypes is generally assessed without regard to the number of (inferred) mutational events that produced the allele, and hence these methods are not directly sensitive to signals of selection pressure. To identify regions harboring mutations under selection pressure, we propose a method for identifying regions with clusters of mutations as a preprocessing step to association testing. Mutations are treated not just as allelic sites, but as distinct evolutionary changes inferred phylogenetically. The significance assessment of such clusters is based on the assumption that genomes evolve generally under neutral theory, with mutations occurring spontaneously in random locations distributed throughout the chromosome. The null hypothesis is that mutations/nucleotide substitutions occurring within a given span of sites in the DNA sequence follow a Poisson distribution (allowing for variations in the local mutation rate[29]). If nucleotide substitutions observed in a region (as distinct evolutionary events) are more abundant than expected, then the mutations are unlikely to have occurred by chance within the region, and the clustering of mutations within the region suggests the effects of positive selection. By evaluating the local rate of changes (as opposed to polymorphic sites), homoplasic sites have an advantage, and indeed, even a single site with multiple changes could appear significant compared with other clusters of SNPs spanning larger regions. The significant clusters can then be tested for association with the phenotype by LMM-based GWAS via a burden test. The advantages of this approach over conventional GWAS methods are as follows: (1) the cluster analysis allows association testing to focus on regions of evident variability, rather than testing either all individual sites (typically tens-of-thousands) or all genes (thousands); (2) regions can be of varying size (eg, focusing on just the most variable portions of a gene, rather than requiring a burden test on all SNPs in the whole open reading frame [ORF]); and (3) the identification of a small subset of variable regions throughout the genome for testing increases the statistical power by reducing the total number of genes/regions that have to be analyzed, which affects the sensitivity during post hoc multiple-tests correction (*P* value adjustment to control the overall false discovery rate [FDR][30]).

In this article, we describe a new approach called ECAT (Evolutionary Cluster-based Association Test) that is an adaptation of LMMs to incorporate homoplasy information for enhancing the identification of genes associated with drug resistance. ECAT involves 3 phases. First, the nucleotide changes at each polymorphic site are inferred by mapping them (using maximum parsimony) onto a phylogenetic tree constructed from the genome sequences of a collection of clinical isolates. Second, the changes are clustered to identify a subset of loci exhibiting a local excess of mutations. Here, homoplasy is exploited to increase the detection of clustered (ie, hypervariable) regions. Third, the clustered regions are analyzed for association with a drug resistance phenotype using an LMM (with mutations combined in each clustered region by simple collapsing). We show that this approach is effective in identifying known genes implicated in drug resistance for in a collection of drug-resistant clinical isolates of *M. tuberculosis* (Mtb). We show that ECAT outperforms traditional LMM-based GWAS approaches (represented by GEMMA) by detecting

associations of some drug resistance loci not identified by site-based or gene-based analyses.

Materials and Methods

Genomic data preprocessing for empirical data sets

We start by assembling each genome in the collection using a comparative assembly approach by aligning all reads (.fastq files) to a reference genome using BWA (Burrows-Wheeler Aligner).[31] We exclude mixed, contaminated or low-coverage (<×30) strains. Next, we call genetic variants of polymorphisms (nonsynonymous and synonymous) and short indels (<10 bp, extracted from .sam files) from the assembled genome for each strain using in-house scripts. Then, we obtain a multiple sequence alignment of SNPs and indels, ignoring those that are ambiguous (<70% base-call homogeneity), in repetitive regions, or in regions with large-scale deletions in some strains. Small indels (<10 bp) with the same sequence and coordinates are identified together. Finally, we build a phylogenetic tree based on SNPs using a maximum parsimony method.[32]

Three-phase ECAT

The ECAT involves 3 phases, including homoplasy signals inference, clustered region identification, and association testing. It first computes the homoplasy count for each polymorphic site, then identifies the clustered regions based on the homoplasy counts (using the Poisson distribution), and finally tests the associations between the clustered regions and resistance to a drug of interest (using an LMM). Given a data set of t strains of a genome of size n bp and a total of m polymorphic sites, ECAT takes 2 input files that include a multiple sequence alignment of nonsynonymous SNPs or small indels, and a phylogenetic tree built from both synonymous and nonsynonymous SNPs. The details of the 3 phases of ECAT are described as follows.

Phase 1: homoplasy inference. In the first phase of ECAT, we infer the degree of homoplasy for each polymorphic site by calculating a homoplasy index (*HI*) as the number of evolutionary changes inferred at each site.[33] Homoplasy index (*HI*) is defined as one plus the difference between the number of actual changes (c_{tree}) and the minimum number of changes ($c_{\min}$). The actual number of changes is inferred from the phylogenetic tree using Sankoff's algorithm.[34] For a polymorphic site (a character χ), the minimum number of changes ($c_{\min}$) is the number of observed character states (nucleotides) minus one. The homoplasy index of 1 for a site at position i represents that the site is homoplasy-free, whereas the higher value of HI_i represents multiple changes (independent evolutionary events) at the site. The *excess changes* at a site i (c_{excess_i}) is the difference between the number of actual changes (c^i_{tree}) and the minimum number of changes ($c_{\min}$). The homoplasy index *HI* at the site i is defined as the excess changes plus 1.

$$c_{excess_i} = c_{tree_i} - c_{\min_i}, \quad 1 \le i \le n \tag{1}$$

$$HI_i = c_{excess_i} + 1, \quad 1 \le i \le n \tag{2}$$

Phase 2: clustered regions identification. To identify the clustered regions based on homoplasy locally, we use a Poisson distribution for detecting hypervariable clusters proposed by Wagner.[29] Given m mutations occurring in a genome of n nucleotides, the mutation rate λ is estimated by the number of mutations evolved over the entire genome, which equals m/n. The probability of a region of size x bp containing k mutations can be modeled as a Pearson type III distribution.[29]

$$P(x) = \frac{\lambda}{\Gamma(k-1)}(\lambda x)^{k-2} e^{-\lambda x} \tag{3}$$

where $\lambda = m/n$ and $\Gamma(k) = (k-1)!$

The probability of k mutations occurring within regions smaller than the size of d_k can be estimated from the cumulative Pearson type III distribution, which is equivalent to the accumulated probabilities from a Poisson distribution where more abundant mutations ($>k$) occur within a given span of d_k.

$$\begin{aligned} P(d_k) &= 1 - \sum_{i=0}^{k-2} \frac{(\lambda d_k)^i}{i!} e^{-\lambda d_k} \\ \lambda &= \frac{m}{n} \end{aligned} \tag{4}$$

Here, in the second phase of our model, the probability of k evolutionary changes (sum of homoplasy indices, HI_i) occurring within a window of u consecutive sites, i to $i+u$, spanning dk nucleotides, under a local rate of changes (λ'), is estimated as

$$\begin{aligned} k &= \sum_{j=i}^{i+u} HI_j \\ P(d_k) &= 1 - \sum_{i=0}^{k-2} \frac{(\lambda' d_k)^i}{i!} e^{-\lambda' d_k} \\ \lambda' &= \frac{\sum_{i-w/2}^{i+w/2} c_{tree_i}}{w} \end{aligned} \tag{5}$$

where λ' is the local rate of changes within a smoothing window of size w bp.

For each polymorphic site across the genome, we group adjacent sites up to a given span of SNPs as a region and calculate its probability using equation (5). The maximum size of the sliding window of SNPs, u is determined by the mutation rate and the average size of genes in the data set. We compute the expected number of SNPs in the largest genes in the genome. The size of the largest genes is estimated as the mean size plus 2 standard deviations (to exclude outliers). For example, in *M. tuberculosis*, given the overall mutation rate and

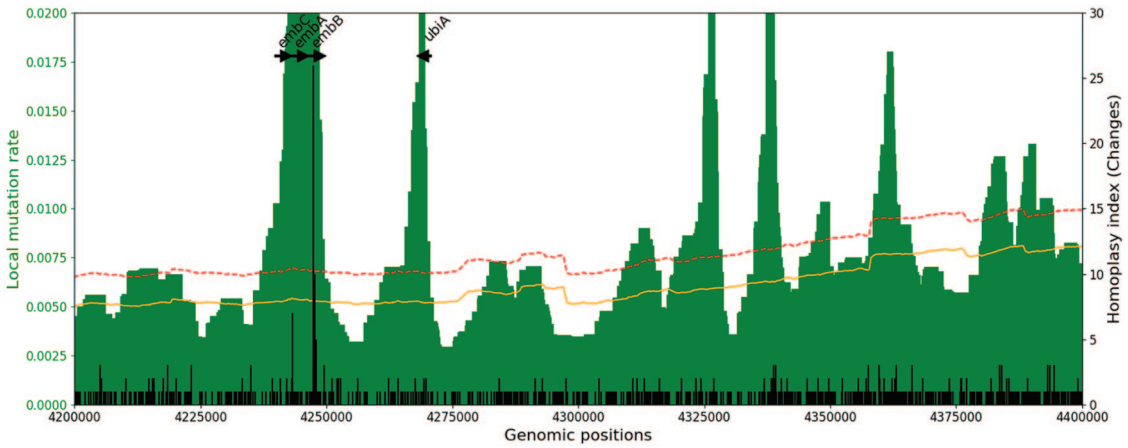

Figure 1. Analysis of single-nucleotide polymorphism (SNP) clusters in a 200 kb region. The green bars represent the local mutation rate and the black bars stand for the homoplasy indexes. Only regions where the local density of nucleotide changes exceeds the significance threshold (dashed red line) based on the Poisson distribution with the average mutation rate (yellow line) would be considered for association testing with drug resistance phenotypes.

gene-size distribution, we estimate that most genes will have at most 17.2 polymorphic sites on average. (In our experiments below, we test all windows of size 1-20 consecutive SNPs for each site to identify hypervariable regions.) For multiple test correction, the *P* values are adjusted for a 5% FDR using the Benjamini-Hochberg procedure.[30] We sort all regions by the adjusted *P* values $P_{adj.}$ and then apply a greedy algorithm to examine each ordered region to obtain nonoverlapping clustered regions, as follows. For each candidate region in sorted order, if the candidate region does not overlap with previously selected regions, then the region is marked as selected. Otherwise, the overlapping region is discarded from the list. After iterating through the entire list (up to $P_{adj.} = .05$), we obtain a set of significant, optimized, nonoverlapping clustered regions. Note, a gene or an intergenic region might have 0, 1, or more than 1 nonoverlapping clustered subregions.

The analysis of SNP clusters is illustrated in Figure 1 for a 200 kb region in the *M. tuberculosis* genome. The green bars represent the local mutation rate in overlapping windows of 10 consecutive SNPs. The black bars stand for the homoplasy indexes (number of evolutionary changes) for all polymorphic sites. The yellow line represents the average mutation rate (smoothed over 100 kb). It increases slightly from the left to the right of the region, which shows the advantage over using a smoothed average over a single, global mutation rate. The dashed red line represents the significance threshold using a *P* value cutoff of .003125 ($\alpha = .05/16\,000$ regions) based on the Poisson model (applying a simple Bonferroni correction for illustration). The peaks above the cutoff represent regions where changes are clustered more densely than expected by chance. The chromosomal region shown contains genes *embCAB* and *ubiA*, which are involved in ethambutol resistance. These genes show clear evidence of excess mutations (presumably selected as a result of exposure to ethambutol as a chemotherapeutic), and it is such clustered regions where association testing is focused.

Phase 3: association tests. In the third phase of ECAT, we test associations of clustered regions identified from the second phase against phenotypes of antibiotic resistance using an LMM.[7] As both homoplasic and nonhomoplasic polymorphic sites are used to identify hypervariable regions, clusters derived from phase 2 may include several polymorphic sites of $HI = 1$ within a small span of gene/intergenic region, or an individual site with a relatively high *HI* ($HI > 1$), or a region consisting of both sites with $HI = 1$ and $HI > 1$. The genotypes of clusters are determined by grouping allelic sites within each cluster using a binary collapsing method (burden test), which is categorized to the allele counting methods. Hence, correcting for the population structure is needed in this phase for association tests to discount mutations shared by genetically similar members of the population.

To correct for population stratification, regression-based methods could employ covariates in the regression models to account for structure effects. Among commonly used regression-based methods, LMMs are able to account for confounders using both fixed effects (covariates) and random effects (kinship/genetic relatedness). Thus, an LMM-based approach was chosen for performing association tests in this phase. The LMM implemented in GEMMA[7] was shown to perform better than the linkage agglomerative clustering approach in PLINK[6] and the dimensionality reduction method in pyseer[10] for controlling cofounders in terms of precision, recall, and F1 scores.[35] Hence, GEMMA is chosen as a foundation for evaluating statistical significance of clustered regions in ECAT.

For each clustered region, polymorphic sites are pooled within the region. If a strain that has at least one mutation among the sites within the boundaries of the region, the genotype of the strain within the region will be marked as having a mutation. The character state of strain *j* at site *i* ($\chi_{j,i}$) is set to 0 if it is the same as the reference state or converted to 1 if it is mutated. For a region of size *u* SNPs starting from the site at position *i*, the genotype (or "burden") of a strain *j* (x_j) within the region from SNP i to SNP $i+u$ is determined by

Input: Genome size of n nucleotides, an alignment of m polymorphisms, a phylogenetic tree, drug susceptibility phenotypes (DST), sliding window sizes w bp (for local mutation rate) and u (maximum region size in SNPs)

Phase 1 – Calculation of Homoplasy Indices and Local Mutation Rates

for all $i \leftarrow 1$ to m **do** ▷ for each polymorphic site
  $c_{excess_i} \leftarrow c_{tree_i}$ - c_{min_i} ▷ compute changes at site i using Sankoff's algorithm
  $HI_i \leftarrow c_{excess_i} + 1$ ▷ compute the homoplasy index (HI)
end for
for all $i \leftarrow 1$ to m **do** ▷ for each polymorphic site
  $\lambda'_i \leftarrow \frac{\sum_{i-w/2}^{i+w/2} c_{tree_i}}{w}$ ▷ compute the local rate of changes within a sliding window of size w bp
end for

Phase 2 – Identification of Clustered Regions

for all $p \leftarrow 1$ to m **do** ▷ for each polymorphic site in genome
  for all $q \leftarrow$ p to $p + u$ **do** ▷ all possible groupings up to u adjacent polymorphic sites
    $k \leftarrow \sum_{i=p}^{q} HI_i$ ▷ sum up homoplasy index of all sites within the region G_{pq}
    $d_k \leftarrow$ Coord(q)-Coord(p)+1 ▷ genomic distance in nucleotides between positions p and q
    $P(d_k) \leftarrow 1 - \sum_{i=0}^{k-2} \frac{(\lambda'_p d_k)^i}{i!} e^{-\lambda'_p d_k}$ ▷ Likelihood for the region G_{pq} by a Poisson model
  end for
end for
Apply FDR of 5% to adjust p values
Obtain non-overlapping clustered regions by a greedy algorithm
Identify significant clustered regions G_{pq} where $p_{adj.} \leq 0.05$

Phase 3 – Association Testing

for all selected regions G_{pq} **do**
  Group sites within the region G_{pq} as a genotype x_{pq} using a binary collapsing method (burden)
  Test associations between the genotype x_{pq} and DST using a linear mixed model (GEMMA)
end for
Adjust p values by the FDR correction (5%)
Identify regions associated with a particular drug by the positive effect size and the adjusted p values (cutoff = 0.05)

Figure 2. ECAT algorithm: 3-phase Evolutionary Cluster-based Association Test.

$$x_j = \begin{cases} 1 & \text{if } \sum_{k=i}^{i+u} \chi_{j,k} \geq 1 \\ 0 & otherwise \end{cases} \quad (6)$$

The significance of the association between a clustered region and a particular phenotype of n individuals such as drug resistance is determined by the Wald test using GEMMA.[7] GEMMA solves the following linear equation:

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X}\beta + \gamma + \varepsilon \\ \gamma &\sim \text{MVN}(0, \sigma_a^2 \boldsymbol{K}) \\ \varepsilon &\sim \text{MVN}(0, \sigma_e^2 \boldsymbol{I}_n) \end{aligned} \quad (7)$$

where MVN is a multivariate normal distribution, $\boldsymbol{Y}$ is a vector of phenotypes for each strain, $\boldsymbol{X}$ is a matrix encoding the genotypes (alleles) for each site/region/gene, coefficients β are the effect sizes of the genotypes, γ is a vector of random effects estimated from a genetic relatedness matrix $(\boldsymbol{K}_{n\times n})$, and ε is a vector of errors (assumed to be normally distributed, $\varepsilon \sim N(0, \sigma_e^2)$). Note that $\boldsymbol{Y}$ can be binary (eg, 0 = Sensitive or 1 = Resistant for each strain based on standard concentrations for drug susceptibility tests or cutoffs for MICs) or quantitative (eg, log of MICs). To account for population structure, a genetic relatedness (kinship) matrix $\boldsymbol{K}$ is used as a *random effect* in an LMM that captures genetic covariances (genotype correlations) between each pair of individuals. We apply both synonymous and nonsynonymous SNPs to calculate the kinship matrix but exclude well-known drug-resistant SNPs (see table below). The P values are subsequently adjusted for an FDR of 5% using a Benjamini-Hochberg procedure for multiple test correction.[30] The regions of negative effect sizes are ignored as we focus on the positive associations between the presence of mutations and drug resistance.

Our 3-phase evolutionary cluster-based algorithm is summarized in Figure 2.

Results

Mycobacterium tuberculosis

Genetic variants, lineages distribution, and antitubercular drugs. To evaluate the ECAT method, we analyze a data set of 660 clinical isolates of *M. tuberculosis* with drug susceptibility data for 7 antibiotics.[5] *Mycobacterium tuberculosis* is the causative agent of tuberculosis (TB) that primarily infects the human lung. The *M. tuberculosis* genome is about 4.4 Mb in size and is believed to be highly clonal, with little evidence of recombination among isolates[36] and low genetic diversity worldwide.[37,38] To treat TB infection, current antitubercular drugs include 5 first-line drugs and several second-line drugs. The 5 first-line drugs are isoniazid (INH), rifampicin (RIF), streptomycin (STR), ethambutol (EMB), and pyrazinamide (PZA). Other second-line drugs include fluoroquinolones (ofloxacin, levofloxacin, moxifloxacin, and ciprofloxacin), ethionamide (ETH),

Table 1. Most frequent resistance mutations observed for several antitubercular drugs.

ANTIBIOTICS	RESISTANCE MUTATIONS
INH (isoniazid)	*katG*: S315T/R; *inhA* prom.: t-8c, c-15t, g-17t; *inhA*: S94A, I194T, I21T
RIF (rifampicin)	*rpoB*: RDRR (a.a. 435-450); *rpoC*
EMB (ethambutol)	*embB*: M306V, M306I, G406S, G406A; *embCA* intergenic region
STR (streptomycin)	*rpsL*: K43R, K88T; *gidB*: nonsynonymous mutations; *rrs*: A514C
PZA (pyrazinamide)	*pncA*: nonsynonymous mutations and indels
KAN (kanamycin)	*rrs*: A1401G; upstream of *eis*
CPX (ciprofloxacin)	*gryA*: A90, D94
ETH (ethionamide)	*ethA; inhA* promoter
PAS (*p*-aminosalic.)	*folC; thyA*

D-cycloserine (CS), amikacin (AMK), kanamycin (KAN), capreomycin (CAP), and *para*-aminosalicylic acid (PAS). Mechanism of resistance to several antibiotics in *M. tuberculosis* has been identified.[39] The well-known annotated loci associated with antitubercular drugs are listed in Table 1.

We ran ECAT on an empirical data set of *M. tuberculosis* clinical isolates from Lima, Peru (BioProject: PRJNA343736), which includes many MDR strains.[5] An MDR-TB strain is defined as being resistant to both INH and RIF, the 2 first-line antitubercular drugs, and many isolates are resistant to additional drugs. We collected a subset of 660 strains with drug susceptibility data for 7 antibiotics. The proportions of drug susceptibility for INH RIF, EMB, STR, PZA, KAN, and CPX (ciprofloxacin) are shown in Figure 3A. A heatmap of the correlations between pairs of antitubercular drugs is shown in Figure 3B. The correlation coefficients between many pairs of first-line drugs are larger than 0.5, suggesting that they have a high degree of overlap among resistant strains (ie, co-resistance), which is likely due to transmission of several MDR clones in the region. The highest correlation, between INH and RIF resistance, is 87%.

We aligned the genome sequences of the 660 isolates against the reference genome, H37Rv (GenBank: NC_000962.2), using MUMmer.[40] We obtained 22 441 polymorphic sites in the alignment where 15 485 are nonsynonymous, excluding gaps, ambiguous sites, and repetitive regions (including *PPE* and *PE_PGRS* genes). The overall mutation rate is 0.0051 per nucleotide (22 441 SNPs/4 411 532 bp). Only 1776 sites are homoplasic throughout the genome. All polymorphic sites, excepting well-known drug-resistant loci (Table 1), were used to estimate the phylogeny by maximum parsimony using PAUP.[32] We determined the family/lineage of each strain by examining lineage-specific biomarkers.[41] The phylogenetic

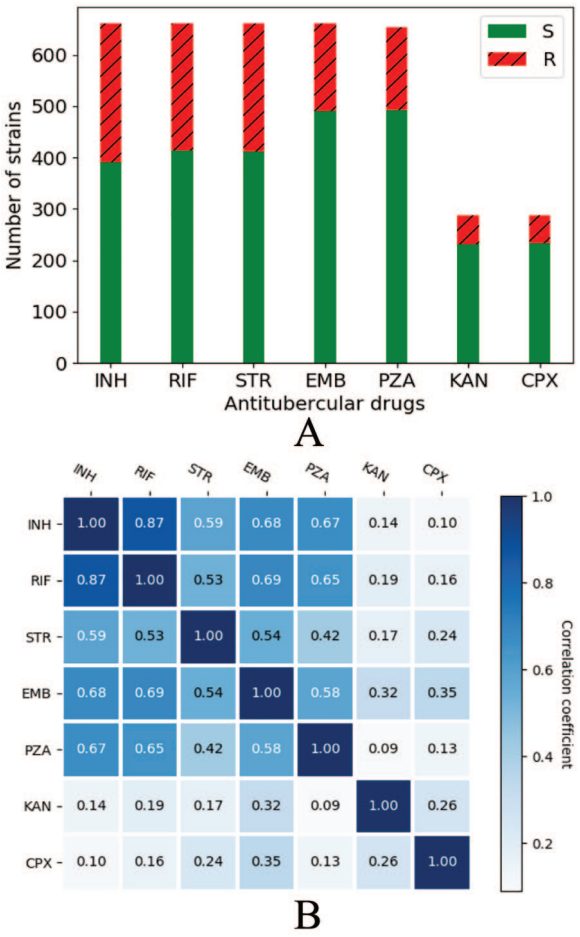

Figure 3. (A) Proportion of drug-resistant strains for 7 antitubercular drugs. The proportion ranges from 18.2% (CPX) to 40.8% (INH). KAN and CPX are available for only a subset of 286 strains. (B) Heatmap plot of pairwise correlations between drugs. Each cell represents the correlation between a pair of drug susceptibilities. Darker green presents stronger co-resistance between drugs for strains. The correlation between INH and RIF is 0.87, suggesting that many strains are resistant to both INH and RIF or sensitive to both of the drugs. CPX indicates ciprofloxacin; EMB, ethambutol; INH, isoniazid, KAN, kanamycin; PZA, pyrazinamide; RIF, rifampicin; STR, streptomycin.

tree labeled with lineages is shown in Figure 4, where most strains are categorized to lineage 2 (Beijing) or lineage 4 (LAM, Haarlem, X-clade, and T-clade).

Identification of optimized clusters of SNPs. *Mycobacterium tuberculosis* exhibits a global mutation rate of approximately 5 SNPs per kilobase (0.005 per nucleotide) in this data set. By applying the Poisson model to the homoplasy indices of nonsynonymous polymorphic sites in sliding windows of 20 SNPs, we obtained 596 clustered regions where the adjusted *P* values are less than .05 (Supplemental Table S1). The occurrence of clustered regions across the entire genome is shown in Figure 5. The median size of the clustered regions is 11 bp (range: 1-3397 bp, 75th percentile: 81 bp). Several well-known drug-resistant loci are represented by clusters and found to be highly homoplasic, including the genes *gyrA, embB, rpoB, rpoC, katG, rpsL, gidB*. As

an example, *rpoB* has a region with 111 changes spanning 1066 bp (encompassing the Region Determining Rifampicin Resistance, RDRR), implying a local mutation rate of 0.104 per nucleotide. We also identify some noncoding regions that are homoplasic and involved in antitubercular resistance such as *inhA* promoter region (*Rv1482c-fabG1*, coordinate: 1673423-1673432), the upstream of *eis* (coordinate: 2715340-2715346), and the intergenic region of *embC-embA* (coordinate: 4243217-4243228). Not all homoplasic sites are associated with drug resistance. For example, the gene *lldD2* is also identified as a homoplasic cluster, though it does not have any known relation to drug resistance.[42] These regions have a local excess of changes (phase 2 of ECAT), though they turn out not to be significant when assessed for association with drug-resistant phenotypes (phase 3).

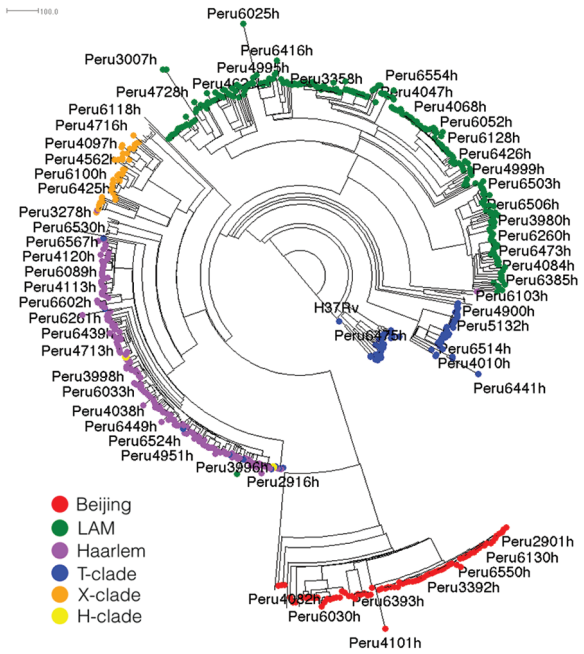

Figure 4. Phylogenetic tree and the distribution of lineages of 660 clinical isolates from Peru. The number of isolates and labeling color for each lineage is as follows: Red: Beijing (78); green: LAM (255); purple: Haarlem (167); blue: T-clade (82); orange: X-clade (42); yellow: H-clade (2); none: unrecognized (34).

Association test for clustered regions for individual drugs. The third phase of ECAT is to perform association testing on regions against resistance to individual drugs, as not all clustered polymorphic regions are associated with drug resistance (eg, *lldD2*). We test the association of antibiotic resistance with clustered regions by an LMM, controlling for confounders, using GEMMA.[7] For calculating the genetic relatedness matrix, we exclude SNPs at the canonical drug-resistant loci such as *katG, rpoB, embB* (loci listed in Table 1). We evaluate the performance of our method by comparing the results with GEMMA using genotypes defined at the individual site or gene levels. To assess the effect of LMM-based stratification corrections, we also apply an LM without controlling for any confounders (an accessory function implemented in GEMMA) on the same data set for association tests, and then compare their performance on correcting population stratifications.

Clustering enhances the detection of compensatory mutations in rpoC associated with RIF resistance. Rifampicin is a first-line antitubercular drug that inhibits DNA-dependent RNA synthesis (transcription). It binds to the β-subunit of the RNA polymerase, and the known mutations that confer RIF resistance are mostly located within the RDRR region of *rpoB* (amino acids 435-450).[43,44] In addition, Comas et al[20] observed that mutations in genes *rpoC* and *rpoA* have compensatory effects for *rpoB* mutations (ie, compensating for fitness cost),

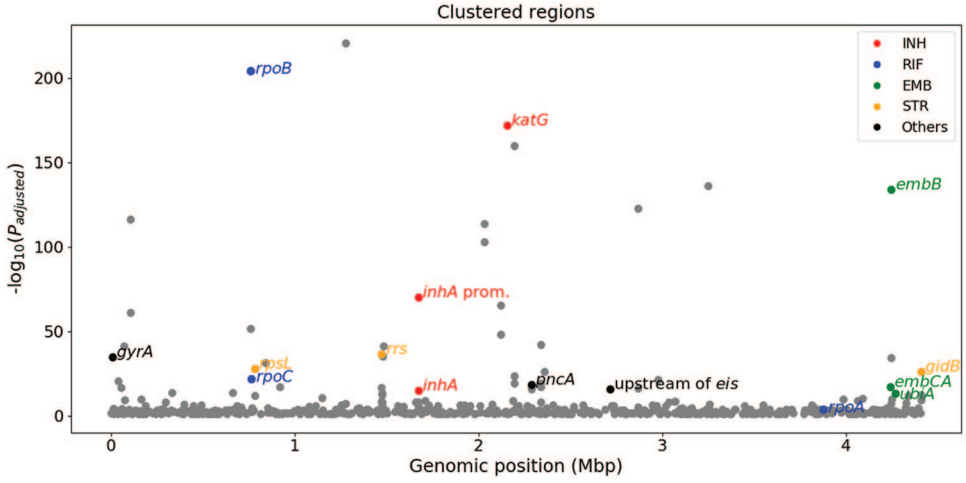

Figure 5. Manhattan plot showing nonoverlapping clustered regions across the genome in *Mycobacterium tuberculosis*. Clustered regions that involve known loci associated with INH, RIF, EMB, STR, and other drug resistance are labeled in red, blue, green, orange, and black, respectively. If multiple clusters in a gene are identified, only the most significant cluster is shown in the plot. EMB indicates ethambutol; INH, isoniazid, KAN, kanamycin; RIF, rifampicin; STR, streptomycin.

stemming from their physical interactions in the RNA polymerase complex. Compensatory mutations in *rpoA* tended to be clustered around amino acid 187, whereas they were distributed throughout the *rpoC*. While association of *rpoB* with RIF resistance is easy to detect with any method, only ECAT is able to detect the association of the secondary gene (*rpoC*) in our data set (see Table 2).

Figure 6A shows that there are 35 nonsynonymous polymorphic sites in the *rpoB* gene from codons P45S to K832E, and 9 SNPs are located in the RDRR region (codons 435-450). Our ECAT test identified 3 clustered regions in *rpoB* from phase 2, and 2 of them are found to be strongly associated with rifampicin resistance in phase 3, codon S450L by itself, and the group of 12 polymorphic sites from S428R to S450T that covers the RDRR region (see Table 3). S450L is highly homoplasic, with 66 changes. For the second region (S428R-S450T), there are 44 changes over 12 polymorphic sites, which is highly significant by the Poisson distribution. In the site-based analysis, codons D435V and S450L are identified to be associated with RIF resistance ($P_{adj.} < .05$). Clinically, these are the most frequently observed RIF-resistant mutations.[45] Taken together, they account for 79.3% of RIF-resistant strains (196/247) in our data set. There are also other sites in genes associated with resistance to other drugs (eg, *katG*:S315T), but this is probably due to co-resistance in the data set, where correlation is as high as 87% (for RIF and INH). Grouping all sites within *rpoB* together shows a significant association in the gene-based analysis with GEMMA ($P_{adj.} = 9.2 \times 10^{-95}$, rank = 1), even though multiple non-DR mutations occur outside of the RDRR region.

For *rpoC*, it exhibits 45 nonsynonymous SNPs from codons A17T to V1252L, where 9 SNPs are homoplasic and distributed throughout the gene at codons V483G, I491V, L507V, V517L, L527V, N698S, A734V, P1040A, and V1252L (Figure 6C). Codon V483G has the highest homoplasy, with 10 changes. We identified 5 clustered regions, including regions Q479R-A492P, L507V-L527V, V1039G-P1040R, and 2 sites by itself, N698S and V1252L. The region Q479R-A492P is identified to be strongly associated with RIF resistance ($P_{adj.} = 1.25 \times 10^{-3}$, rank = 9). Conversely, both site-based and gene-based analyses in GEMMA fail to show any significant association between *rpoC* and RIF resistance. The individual site in *rpoC* with the highest significance is codon V483G, which ranks as 24th with $P_{adj.} = .740$. Testing *rpoC* at the gene-wide level shows that its association with RIF resistance ranks it as 16th with $P_{adj.} = .106$. Grouping all sites within the gene enhances the significance of association, but it is still not significant enough by the default 5% FDR. This is because it may be affected by other nonresistance-related or lineage-specific mutations such as G594E. Codon G594E in *rpoC* is the marker of Haarlem family.[46] In this data set, 240 strains harbor mutations in this locus, but 170 are sensitive to rifampicin, which is close to the background frequency of 62.58%.

Table 2. Summary of analyses of loci associated with rifampicin resistance by 3 statistical methods.

LOCUS	ECAT	SITE-BASED GEMMA	GENE-BASED GEMMA
rpoB	**8.33×10^{-37}**	**2.16×10^{-35}**	**9.84×10^{-94}**
	(2)	**(2)**	**(1)**
rpoC	**0.00125**	0.74	0.106
	(9)	(24)	(16)
rpoA	0.435	0.983	0.997
	(36)	(977)	(396)

Adjusted *P* values for the highest-ranked region or site with each locus are given, along with rank (in parentheses). Significant associations are boldfaced in red. Abbreviation: ECAT, Evolutionary Cluster-based Association Test.

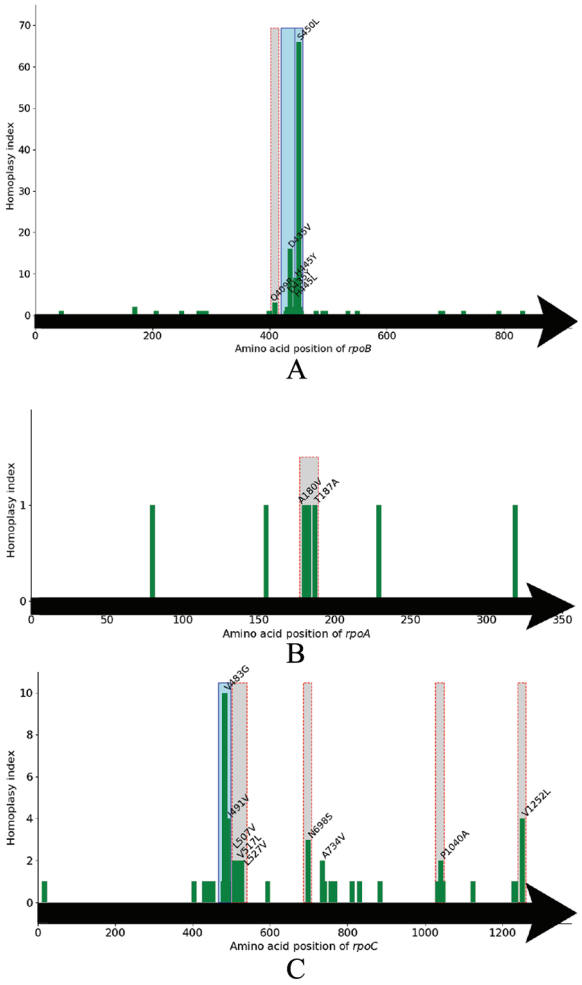

Figure 6. The distributions of homoplasy index for each polymorphic site in the genes (A) *rpoB*, (B) *rpoA*, and (C) *rpoC*. The *y*-axis presents number homoplasy index and the *x*-axis represents the amino acid position. A codon exhibiting more than 1 change (homoplasic site) is labeled in text. Clusters are boxed with solid blue borders for significant associations with rifampicin (RIF) resistance and dashed red borders for nonsignificant ones.

Table 3. Association of RIF resistance and genetic variants in regions identified by ECAT for *Mycobacterium tuberculosis*.

RANK	REGION	P_{adj}
1	*katG*:S315T-S315T	1.60×10^{-55}
2	***rpoB*:S450L-S450L**	8.33×10^{-37}
3	***rpoB*:S428R-S450T**	3.01×10^{-21}
4	*embB*:M306V-M306I	1.24×10^{-15}
5	*pncA*:S59F-upstream of *pncA*	1.16×10^{-14}
6	*rpsL*:K43R-K43R	1.12×10^{-6}
7	*gidB*:R96L-V65G	3.13×10^{-5}
8	*ethA*:R279*-C131Y	1.82×10^{-4}
9	***rpoC*:Q479R-A492P**	1.25×10^{-3}
10	Noncoding region between Rv3366-Rv3367	1.60×10^{-3}
. . .		
36	***rpoA*:T187A-A180V**	.435

The dashed line shows the significance threshold for FDR < 0.05.
Mutations known or suspected to be relevant to RIF resistance are bold-faced.
Abbreviations: ECAT, Evolutionary Cluster-based Association Test; FDR, false discovery rate; RIF, rifampicin.

Table 4. Summary of analyses of loci associated with ethambutol resistance by 3 statistical methods.

LOCUS	ECAT	SITE-BASED GEMMA	GENE-BASED GEMMA
embB	**1.72×10^{-38}**	**2.75×10^{-18}**	**1.98×10^{-82}**
	(2)	**(2)**	**(1)**
embCA	**0.0264**	0.792	0.257
	(26)	(104)	(33)
ubiA	**1.98e−08**	0.0823	**1.96×10^{-14}**
	(7)	(45)	**(5)**

Adjusted *P* values for the highest-ranked region or site with each locus are given, along with rank (in parentheses). Significant associations are boldfaced in red.
Abbreviation: ECAT, Evolutionary Cluster-based Association Test.

None of the 3 methods detects the association of mutations in *rpoA* with RIF resistance. However, ECAT comes closest, identifying a cluster in *rpoA* that ranks much higher than by either method with GEMMA. Nine nonsynonymous SNPs in total occur in *rpoA* from codons L80V to E319K (Figure 6B). None of them are homoplasic, as the SNPs are mostly represented by a single strain at each site. Our cluster-based approach focuses on the region of 5 SNPs between amino acids 180 and 187 spanning 21 bp clustered from the second-phase analysis ($P_{adj.} = 9.60 \times 10^{-5}$). Although the region would not be identified to be associated with RIF resistance given the FDR adjustment ($P_{adj.} = .435$; see Table 3), it ranks highly as 36th out of 596 regions. Seven strains exhibit at least 1 SNP within the region and they are all resistant to RIF. In the site-based analysis with GEMMA, no association is identified for any individual SNPs in *rpoA* (highest rank: 977th out of 22 441 sites). Similarly, grouping all SNPs within the gene *rpoA* suggests no association from the gene-based analysis (rank: 396th out of 4657 genes or intergenic regions with SNPs).

Thus, for cases such as *rpoB*, where changes in an individual site are abundant enough to be strongly linked with resistance, our cluster-based method performs as well as other methods in terms of identifying the best grouping of SNPs that maximizes the association. For other cases such as *rpoC* and *rpoA*, where changes at an individual site are not enough to be identified from the association test, optimal grouping of SNPs within a clustered region helps identify resistant-related mutations (Table 4).

Clustering enhances the detection of secondary resistance mutations associated with ethambutol. Ethambutol is a first-line drug involved in the inhibition of cell wall synthesis by targeting arabinogalactan biosynthesis (a cell wall glycolipid in *M. tuberculosis*). Mutations in the *embB* gene (especially codons 306, 406, and 497, which are the most frequent) are primarily responsible for mediating EMB resistance.[47,48] Recently, it has been reported that mutations in the *embC-embA* intergenic region are also associated with EMB resistance with low frequency, and these have been shown to affect expression of genes in the *embCAB* operon.[49] *embA, embB*, and *embC* form a cell wall complex that is involved in transferring lipoarabinomannan (LAM) precursors to the outer membrane. Brossier et al[49] found that about 70% of ethambutol-resistant clinical isolates harbored mutations in *embB* and 15% of them had mutations in the *embC-embA* intergenic region (especially positions −8 to −21), affecting expression. Growing evidence suggests that *ubiA* is also associated with EMB resistance, especially high-level EMB resistance.[50] *ubiA* encodes a decaprenyl-phosphate 5-phosphoribosyltransferase in the pathway for synthesizing LAM.[51] All 3 GWAS methods detect the association between EMB resistance and *embB* and *ubiA*, but only ECAT detects the association with the intergenic region of *embC-embA*.

There are 31 nonsynonymous SNPs in the *embB* gene, and SNPs at codons M306 and G406 are highly homoplasic, with 51 changes and 16 changes, respectively (inferred from the tree using Sankoff's algorithm) (Figure 7A). These 2 residues are each identified by ECAT as separate clustered regions in *embB* and are found to be strongly associated with EMB resistance (Table 5). They each consist of a pair of adjacent nucleotides with many inferred changes, producing small clusters which are statistically significant due to homoplasy. On testing for association with EMB resistance, the ranks of the regions at codons 306 and 406 are second and sixth with the adjusted *P* values 1.72×10^{-38} and 1.79×10^{-9}, respectively. The other 2 methods (GEMMA applied to individual sites or collapsing them over

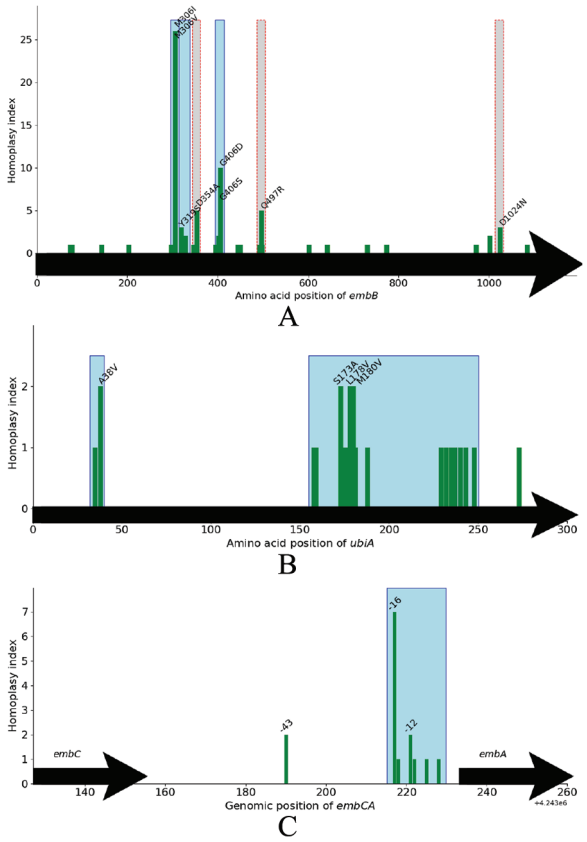

Figure 7. The distributions of homoplasy index for each polymorphic site in the genes (A) *embB*, (B) *ubiA*, and the intergenic region (C) *embCA*. The *y*-axis presents number homoplasy index and the *x*-axis represents the amino acid position. A codon exhibiting more than 1 change (homoplasic site) is labeled in text. Clusters are boxed with solid blue borders for significant associations with ethambutol (EMB) resistance and dashed red borders for nonsignificant ones.

the whole gene) capture the association between *embB* and EMB resistance as well (Table 6).

Seven polymorphisms exist within the *embC-embA* intergenic region spanning 39 bp from coordinates 4 243 190 to 4 243 228 (−43 to −5 bp upstream of *embA*), and 3 of the SNPs are homoplasic (see Figure 7C). In the second phase of ECAT, a clustered region is obtained consisting of 6 SNPs spanning 12 bp from coordinates 4 243 217 (−16) to 4 243 228 (−5). The region has 13 changes in total and is identified to be significantly associated with EMB resistance ($P_{adj.} = .026$, rank = 25; see Table 5). However, the other 2 methods (site-based and gene-based analysis with GEMMA) do not detect any association in *embC-embA*. The highest association at the site level occurs at nucleotide −11 within the intergenic region, where its rank is 104 and $P_{adj.} = .792$. We observed that the allele frequency is not high enough for any single site within *embC-embA*, as only 1.97% of strains have mutations (13/660) at nucleotide −16, which is the most frequent site in this locus. Grouping all sites within *embC-embA* in GEMMA shows no association with EMB resistance; its rank is 33 and $P_{adj.} = .257$. We found that 23 strains have at least one of the 7 SNPs in the region, but 7 strains are sensitive to EMB, so the gene-based approach with GEMMA does not detect this association either.

For the *ubiA* (Rv3806c) gene, there are 20 nonsynonymous SNPs from A35S to E273D spanning 717 bp, where 4 are homoplasic, including A38V, S173A, L178V, and M180V (Figure 7B). With ECAT, 2 clustered regions within *ubiA* are found to be associated with EMB resistance. The first region, F248L-L158S, consists of 18 SNPs spanning 272 bp and the second region clusters 2 codons A35S and A38V together. None of the mutations in *ubiA* is identified to be involved in EMB resistance by the site-based test in GEMMA with 5%

Table 5. Association of EMB resistance and genetic variants in regions identified by ECAT for *Mycobacterium tuberculosis*.

RANK	REGION	P_{adj}
1	*katG*:S315T-S315T	1.63×10^{-50}
2	***embB*:M306V-M306I**	1.72×10^{-38}
3	*rpoB*:S428R-S450T	5.93×10^{-28}
4	***embB*:E405D-G406D**	1.79×10^{-9}
5	*rpoB*:S450L-S450L	1.79×10^{-9}
6	*pncA*:S59F-upstream of *pncA*	2.94×10^{-9}
7	***ubiA*:F248L-L158S**	1.98×10^{-8}
8	***embB*:Y319S-D328G**	1.77×10^{-6}
9	*pncA*:V139A-V93L	2.52×10^{-6}
10	*rrs*:A1401G-A1401G	4.92×10^{-6}
19	*rrs*:C513T-C517T	1.32×10^{-3}
26	***embCA*:−5 to −16 bp upstream of *embA***	.0264

Mutations known or suspected to be relevant to RIF resistance are bold-faced.
Abbreviations: ECAT, Evolutionary Cluster-based Association Test; EMB, ethambutol.

Table 6. Summary of analyses of loci associated with streptomycin resistance by 3 statistical methods.

LOCUS	ECAT	SITE-BASED GEMMA	GENE-BASED GEMMA
rrs	**0.021**	0.854	**4.04e-06**
	(12)	(35)	**(6)**
rpsL	**1.70×10^{-9}**	**1.11×10^{-7}**	**4.53×10^{-10}**
	(5)	**(2)**	**(5)**
gidB	**4.76×10^{-13}**	0.773	0.966
	(2)	(10)	(1719)

Adjusted *P* values for the highest-ranked region or site with each locus are given, along with rank (in parentheses). Significant associations are boldfaced in red.

FDR. The highest ranked site occurs at A35S with rank = 45 and $P_{adj.} = .082$. However, at the gene level with GEMMA, combining all changes in *ubiA* together (gene-level test) shows a strong association of *ubiA* with ethambutol resistance, where the rank = 5 and $P_{adj.} = 1.96 \times 10^{-14}$.

Overall, our ECAT method detected associations between EMB resistance and mutations in *embB*, the intergenic region of *embC-embA* and *ubiA*, whereas other methods (site-based and gene-based analyses with GEMMA) failed to identify the association of the *embC-embA* intergenic region with EMB resistance. The failure of site-based and gene-level GWAS analysis is likely due to low allele frequency, whereas ECAT takes advantage of clustering sites together optimally to enhance significance.

Clustering helps identify the role of gidB in streptomycin resistance. Streptomycin is a first-line anti-TB drug that binds to ribosomal protein S12 and the 16S ribosomal RNA (*rrs*) to inhibit translation (protein synthesis). Resistance to streptomycin is mediated by nucleotide substitutions of A514C in *rrs* (16S rRNA) and codons K43R and K88T in *rpsL* (gene encoding the S12 ribosomal protein).[52,53] Also, nonsynonymous mutations at *gidB* (a 16S rRNA methyltransferase) have been reported to confer streptomycin resistance.[54] The methylation of the ribosome is needed for optimal binding of streptomycin, so loss-of-function mutations in the methyltransferase mediate resistance to STR. While all 3 GWAS methods detect associations between streptomycin resistance and genes *rpsL* and *rrs, gidB* is harder to detect because mutations are spread throughout the gene, and are interspersed with nonassociated SNPs, for example, universal mutation S100F (relative to H37Rv) or lineage-specific mutations such as L16R and E92D.[55,56] L16R is a marker of LAM strains, whereas E92D is a marker of Beijing family.[55]

For *rpsL*, only 2 nonsynonymous SNPs are observed in the alignment, K43R and K88R/T. Both are homoplasic, with 11 and 6 changes at each site. As these 2 mutations occurring in *rpsL* are separated from each other by 135 bp, each is locally clustered by itself in the second phase of ECAT. In the third phase, the codon K43R is identified to be associated with STR resistance (rank = 5 and $P_{adj.} = 1.70 \times 10^{-9}$), where 27 strains harbor the mutation and 26 are resistant (see Table 7). The association between the codon K88T and STR resistance is not significant, probably because the allele frequency is relatively low (2.3%). The site-based analysis with GEMMA identifies codon K43R in *rpsL* (rank = 2 and $P_{adj.} = 1.11 \times 10^{-7}$). The gene-based analysis reports the gene *rpsL* to be associated with STR resistance (rank = 5 and $P_{adj.} = 4.53 \times 10^{-10}$).

For the 16S rRNA, our ECAT method identifies the association between streptomycin resistance and A514C at *rrs* within a clustered region of 3 consecutive SNPs (a.a. C513T-C517T; rank = 12 and $P_{adj.} = .021$). It is not detected by the site-based method (rank = 35 and $P_{adj.} = .854$) probably due to low allele frequency (1.1%). Seven strains harbor mutation at A514C, and 5 are resistant to STR. Gene-level analysis with GEMMA shows *rrs* is significantly associated with STR resistance by collapsing all SNPs together, but this is probably due to kanamycin co-resistance (the primary kanamycin resistance mutation is *rrs:a1401g*). Among 56 KAN-resistant strains, 47 are also resistant to STR, which constitutes 83.9% co-resistance. Thus, site-based analysis shows that A1401G at *rrs* is associated with STR resistance (rank = 5 and $P_{adj.} = .0085$), even though this particular mutation only confers KAN resistance. The A1401G mutation itself occurs in 36 strains, of which 32 are (coincidentally) STR-resistant (out of 249 STR-resistant strains in total).

Table 7. Association of STR resistance and genetic variants in regions identified by ECAT for *Mycobacterium tuberculosis*.

RANK	REGION	P_{adj}
1	*katG*:S315T-S315T	8.91×10^{-32}
2	***gidB*:R96L-V65G**	4.76×10^{-13}
3	*embB*:M306V-M306I	4.17×10^{-12}
4	*rpoB*:S428R-S450T	6.45×10^{-12}
5	***rpsL*:K43R-K43R**	1.70×10^{-9}
6	*rrs*:A1401G-A1401G	2.74×10^{-4}
7	*pncA*:S59F-upstream of *pncA*	3.63×10^{-3}
12	***rrs*:C513T-C517T**	.021
15	***gidB*:I179V-L108R**	.0785
21	***rpsL*:K88T-K88T**	.245

The dashed line shows the significance threshold for FDR < 0.05.
Mutations known or suspected to be relevant to RIF resistance are bold-faced.
Abbreviations: ECAT, Evolutionary Cluster-based Association Test; FDR, false discovery rate; STR, streptomycin.

Figure 8 shows the distribution of polymorphic sites in *gidB*, where 55 nonsynonymous mutations are spread out from codons L16R to L196F, spanning 542 bp. In total, 11 sites are homoplasic, where codons L145F and P75S are the top 2, with 5 and 4 changes, respectively. The ECAT detects a strong association of one clustered region within *gidB*, R96L-V65G (rank = 2 and $P_{adj.} = 4.76 \times 10^{-13}$). The other region from L108R to I179V ranks 15th with $P_{adj.} = .0785$, which is just beyond the FDR threshold (Table 7). However, no association between *gidB* and STR resistance is detected by either site-based or gene-based analysis with GEMMA. The highest association at the individual site level for *gidB* occurs in D67G, $P_{adj.} = .773$. The association by the gene-based analysis with GEMMA shows that *gidB* is not significantly associated with STR resistance, as its rank is 1719 and $P_{adj.} = .966$. This is because *gidB* is a ribosome methyltransferase, and resistance is conferred by loss-of-function mutations, which can occur anywhere throughout the ORF. Therefore, individual sites might not have high enough allele frequency. Also, not all SNPs are

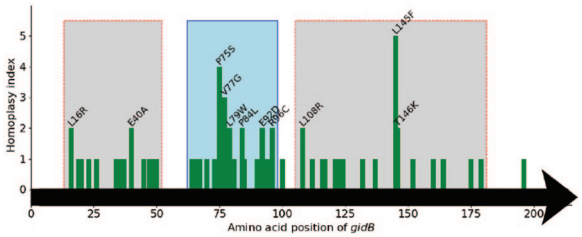

Figure 8. The distribution of homoplasy index for each polymorphic site in the gene *gidB*. The *y*-axis presents number homoplasy index and the *x*-axis represents the amino acid position. A codon exhibiting more than 1 change (homoplasic site) is labeled in text. Clusters are boxed with solid blue borders for significant associations with streptomycin (STR) resistance and dashed red borders for nonsignificant ones.

involved in resistance and some may be lineage-specific. Thus, grouping all SNPs without excluding these SNPs causes GEMMA to fail to report a significant association. In fact, when we remove common and lineage-specific SNPs (S100F, L16R, E92D), the association between *gidB* and STR resistance becomes significant by the burden test (top rank). Hence, our cluster-based method does a much better job identifying the significance of association of mutations in *gidB* with STR resistance than other GWAS methods.

Effect of population stratification corrections. To characterize the effect of population stratification corrections on ECAT (which accounts for kinship relationships in an LMM), we compared it with the effect of an LM-based approach (without random effects) by applying an LM to regress genotypes against phenotype of interests directly without correcting for confounders based on genetic relatedness. This is achieved by running GEMMA in a mode in which it ignores the kinship matrix and does not attempt to estimate random effects. The significant hits estimated from both models using a cutoff of adjusted *P* values <.05 for associations between 3 types of genotypes and resistance to one of the drugs, RIF, EMF, or STR, are listed in Tables 8 to 10, respectively. When using an LM, the number of significant hits is much higher than the number of hits estimated from an LMM in general, regardless of genotype. The Quantile-Quantile plot (Q-Q plot) of associations between cluster-based genotypes and RIF resistance estimated from an LM is more inflated than the ones from an LMM (shown in Figure 9), suggesting a higher type I error rate. The Q-Q plots for site- and gene-based tests look similar.

Novel resistance mutations. In our analysis of the collection of MDR-TB clinical isolates from Peru, ECAT identified only 4 other loci potentially associated with resistance to one or more of the antitubercular drugs: *spoU, idsA2, ppsA*, and Rv2571c (see Table 11). In contrast, many of these associations were not identified by either site-based or gene-based approaches using GEMMA. The homoplasic mutations in *spoU* occur 20 bp downstream from the stop codon (potentially affecting transcriptional termination). The association is significant for

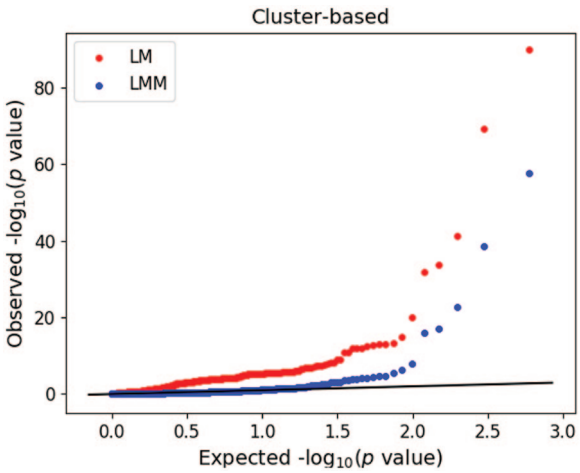

Figure 9. The Quantile-Quantile plots (QQ plots) of unadjusted *P* values estimated from a linear model (LM) (labeled in red) and from a linear mixed model (LMM) (labeled in blue) for association tests between rifampicin (RIF) resistance and the cluster-based genotype.

Table 8. Summary of total number of significant hits (adjusted *P* values < .05) from association tests between RIF resistance and 3 genotypes using LMM and LM models.

MODELS	CLUSTER-BASED (ECAT)	SITE-BASED (GEMMA)	GENE-BASED (GEMMA)
LMM	21	7	14
LM	138	1153	436

Abbreviations: ECAT, Evolutionary Cluster-based Association Test; LM, linear model; LMM, linear mixed model.

Table 9. Summary of total number of significant hits (adjusted *P* values < .05) from association tests between EMB resistance and 3 genotypes using LMM and LM models.

MODELS	CLUSTER-BASED (ECAT)	SITE-BASED (GEMMA)	GENE-BASED (GEMMA)
LMM	29	27	14
LM	112	703	332

Abbreviations: ECAT, Evolutionary Cluster-based Association Test; LM, linear model; LMM, linear mixed model.

Table 10. Summary of total number of significant hits (adjusted *P* values < .05) from association tests between STR resistance and 3 genotypes using LMM and LM models.

MODELS	CLUSTER-BASED (ECAT)	SITE-BASED (GEMMA)	GENE-BASED (GEMMA)
LMM	14	6	8
LM	72	319	168

Abbreviations: ECAT, Evolutionary Cluster-based Association Test; LM, linear model; LMM, linear mixed model.

resistance to RIF and EMB, and 20 out of 660 strains have such mutations. This association was also observed in another analysis of the same data set.[57] *spoU* is a putative tRNA/

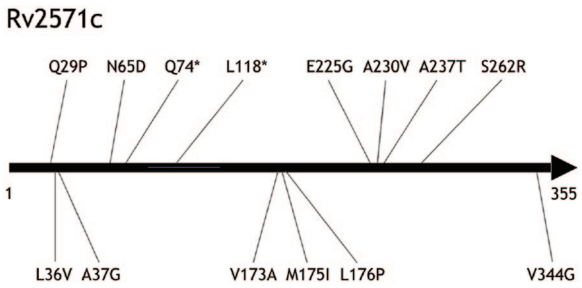

Figure 10. Relative locations of 14 observed changes within the Rv2571c in the data set of 660 strains from Peru. Rv2571c has 355 amino acids. Stop codons are marked with asterisks.

Table 11. Summary of analyses of novel loci associated with antitubercular resistance by 3 statistical methods.

DRUG: LOCUS	ECAT	SITE-BASED GEMMA	GENE-BASED GEMMA
RIF: *spoU*	**0.002**	0.052	**0.015**
	(10)	(8)	**(11)**
EMB: *spoU*	**4.97×10^{-4}**	**0.015**	0.195
	(17)	**(14)**	(22)
INH: *idsA2*	**0.011**	0.990	0.066
	(11)	(50)	(11)
EMB: *idsA2*	**0.006**	0.379	0.058
	(24)	(85)	(15)
INH: *ppsA*	**0.003**	0.084	0.167
	(10)	(8)	(17)
RIF: *ppsA*	**0.003**	0.083	0.312
	(12)	(11)	(20)
STR: *ppsA*	**0.012**	0.351	**0.028**
	(8)	(7)	**(8)**
EMB: Rv2571c	**0.005**	0.792	0.063
	(22)	(145)	(20)

Adjusted *P* values for the highest-ranked region or site with each locus are given, along with rank (in parentheses). Significant associations are boldfaced in red. Abbreviations: ECAT, Evolutionary Cluster-based Association Test; EMB, ethambutol; INH, isoniazid; RIF, rifampicin; STR, streptomycin.

rRNA-methyltransferase. Currently, its role in drug resistance is not known, though the fact that *spoU* orthologs in other organisms can methylate the ribosome suggests a possible link to resistance to translation inhibitors.[58] *idsA2* is a putative geranylgeranyl pyrophosphate synthetase, which is on the pathway for synthesis of cell wall glycolipids. It might be that these mutations affect cell wall permeability, which could explain how mutations in *idsA2* associates with resistance to INH and EMB (though it is important to remember that there is a great deal of co-resistance in this data set, so it is difficult to say with certainty which specific drug *idsA2* is associated with). Mutations in *ppsA* could have a similar effect on permeability, as this gene is involved in synthesis of PDIM (phthiocerol dimycocerosate), another cell wall glycolipid. For example, mutations in PDIM synthesis genes have been shown to confer resistance to pyrazinamide.[59] Perhaps the most intriguing case is Rv2571c, which is most strongly associated with EMB resistance ($P_{adj.} = .0051$). Mutations are scattered throughout the ORF (see Figure 10). In the Peru collection, 25 strains have nonsynonymous SNPs in Rv2571c distributed over 14 polymorphic sites spanning 701 bp (though none are homoplasic), and 2 have indels (frameshifts). Of these, 16 of these strains (64%) are resistant to EMB, which is more than twice as high as the background resistance rate for the whole population (26%, Figure 3A). Rv2571c is annotated only as a transmembrane protein that is Ala/Val/Leu-rich, and its function is unknown. Further tests will be needed to evaluate the effect of mutations (including loss-of-function) in Rv2571c on resistance to various drugs. Mutations in Rv2571c and *idsA* have not previously been reported as associated with drug resistance in other collections of drug-resistant TB clinical isolates,[4,5,26,60] and they do not appear in TBDReaMDB, a summarized database of known TB drug resistance mutations.[61]

Discussion

We have presented a new method, called ECAT, to extend GWAS with the ability to take advantage of homoplasy. The motivation behind ECAT is that drug resistance mutations are frequently observed to occur multiple times independently in a clinical population, which is a strong indicator of positive selection. As traditional GWAS methods are not sensitive to this information, this was implemented via a preprocessing phase to identify significantly clustered regions of mutations (evolutionary changes), which are then subjected to association testing using an LMM. Importantly, the clustering step assesses the clustering of distinct evolutionary changes (inferred from a phylogenetic tree using maximum parsimony) rather than just polymorphic sites, so homoplasic sites (with multiple changes) are counted with higher weight. The boundaries of clusters are optimized to focus on just the most variable parts of genes, without limiting analysis to either individual SNPs or entire ORFs. In fact, some genes might not have any significant clusters, and hence do not need to included in association testing.

This approach has several advantages. First, the preprocessing step focuses the association testing on variable regions of the genome, which is where drug resistance mutations are expected to occur (at least in a collection of drug-resistant clinical isolates). It is more computationally efficient to evaluate only statistically significant clusters, rather than the naive approach of doing association testing on all sliding windows of SNPs throughout the genome. In our case, there were >20 000 allelic sites (and hence >400 000 possible overlapping windows of size 1-20), but only 596 regions were selected as

significantly clustered (using the Poisson model). Filtering out regions without significant clusters of mutations also helps maintain sensitivity during the multiple-tests (FDR) correction at the end, as the association testing is only done on a small subset of hypervariable regions.

Second, the clustered regions can be of varying size. Traditionally, GWAS has been applied to individual sites (SNPs) or entire genes (ie, collapsing all SNPs together in an ORF, as often done with burden tests[22]). However, resistance mutations for many drugs are often clustered within the target ORFs, such as in active sites. Often, there are only a handful of prevalent mutations in drug targets (such as the RDRR in RpoB [residues 435-450], and the Ala90-Gly94 region of GyrA), and the rest of the mutations in the ORF are not relevant to drug resistance (eg, lineage-specific mutations, like GyrA:S95A). The clustering approach we describe (second phase of ECAT) allows optimized local variable regions within ORFs to be singled out and tested for association, which can strengthen the signal.

Finally, because the clustering is based on distinct evolutionary changes, homoplasic sites have an advantage. A site with multiple changes (spanning just 1 bp) can often be significant on its own or can seed a cluster combined with other nearby changes (based on the Poisson distribution). This contrasts with traditional GWAS methods, such as those based on allele counting or LMs, which do not take into account the number of changes that led to an allelic site, and hence are insensitive to homoplasy.

The results obtained by ECAT on the *M. tuberculosis* data set from Peru show that most of the major known drug resistance mechanisms can be detected, including less prevalent ones not detected by GEMMA. In some cases, individual sites with high homoplasy are identified, such as *KatG*:S315T and *EmbB*:M306V/I. Similarly, *RpsL*:K43R and K88R/T are individually detected as significant for streptomycin resistance. In other cases, larger clusters of mutations are identified, such as the RDRR region in RpoB, as well as in RpoC. Mutations in the latter have been shown to be able to compensate for fitness costs of mutations in RpoB, though they are less frequent, making the statistical association harder to detect. The ECAT also detects an association of the intergenic region between *embA* and *embC* with resistance to ethambutol, which involves 6 allelic sites spanning 12 nucleotides (2 of which are homoplasic, representing a total of 13 distinct changes in this locus). Although substitutions in the *embC-embA* intergenic region are less prevalent than missense mutations in EmbB, they have also been shown to increase resistance to ethambutol through upregulating expression.[49] In still other cases, clusters of mutations can be spread throughout a gene, such as in PncA (activator of pyrazinamide)[62] and GidB (ribosome methyltransferase, which influences the binding of streptomycin). In such genes, resistance is typically conferred through disruption, that is, loss-of-function mutations, which can occur at many different positions in the ORF, making them difficult to detect through site-based analysis because; individually, they often occur at low allele frequencies and might be interspersed with other polymorphisms not relevant to resistance. However, gene-level analysis by tools like GEMMA force all mutations to be pooled in a burden test, even though not all mutations in such genes necessarily cause loss-of-function, diluting the signal.

The fact that, aside from these known resistance-associated mutations for TB drugs, only 4 other loci (*spoU, idsA2, ppsA*, and Rv2571c) were identified as significantly associated with one of the 7 drugs in our data set suggests that the false-positive rate for ECAT is very low, which is a consequence of combining rigorous statistical filtering at both stages, clustering and association testing.

The ECAT method has several limitations. First, high levels of co-resistance among multiple drugs can cause ambiguity over which polymorphisms are associated with which drugs. In our MDR-TB data set from Peru, the overlap in resistance to drugs such as rifampicin and isoniazid was as high as 87%. This resulted in cases where SNPs like RpoB:S450L (which confers resistance to rifampicin) showed up on the list of significant loci for isoniazid resistance, and KatG:S315T (confers resistance to isoniazid) showed up on the list of hits for rifampicin resistance. An expanded data set with more independent clinical isolates would be needed to correctly deconvolve these genotype-phenotype associations. Co-resistance is a well-known problem for traditional GWAS as well.[5]

Second, our method does not specifically address epistatic interactions. Similar to traditional GWAS, our method assesses the statistical association of each locus independently from all the others. However, in some cases, resistance (or the level of resistance) can be determined by alleles at multiple positions in the genome.[63] For example, mutations related to toxin production were found to influence the growth rate of mupirocin-resistant strains of methicillin-resistant *Staphylococcus aureus* (MRSA) with mutations in *ileS*.[64] Vogwill et al[65] found that the fitness cost for nearly 50% of rifampicin resistance mutations in *rpoB* differed among strains of *Pseudomonas*, implying the influence of other (lineage-specific) mutations. And, Knopp and Andersson[66] demonstrated an interaction among resistance mutations to different antibiotics in *Salmonella*; in some cases, the fitness effects of different combinations of mutations were strain-dependent, including dependence on mutations at other loci, such as *lon* (protease), *ompR* (porin), and *marR* (regulator). As typically applied, LMMs do not account for interactions where the association at one site is dependent on an allele at another site, and instead, other analytical methods targeted at identifying linkage disequilibrium[6,67] are required. A common form of epistasis is compensatory mutations. The ECAT was fortunately able to detect the weak association of compensatory mutations in RpoC with rifampicin resistance in our data set without explicitly taking RpoB mutations into account. However, there may well be other cases where interactions among genomic loci (exemplified by *rpoB* and *rpoC*) might not be detected.

Third, an accurate phylogenetic tree is required for estimating the homoplasy counts for SNPs. However, it remains challenging to reconstruct a global phylogeny for recombinant species such as *Streptococcus pneumoniae*[68] and *Klebsiella pneumoniae*.[69] Without a reliable phylogeny, sites located within the recombined regions may tend to look more homoplasic with respect to a global phylogeny based on SNPs across the whole genome, resulting in more false positives. Recent methods for locating chromosomal boundaries where recombination events occur during evolution may allow us to obtain more reliable trees accounting for the recombined regions[70] or to build up separate local trees for the recombined regions based on inferred recombination breakpoints.[71]

Finally, as ECAT is designed for analyzing the association between chromosomal mutations and drug resistance, it is not expected to work for cases where resistance is acquired through lateral transfer of plasmid-borne resistance genes or other mobile genes. For example, the presence or absence of a plasmid carrying the *mecA* gene is the primary determinant of methicillin resistance in *S. aureus* (MRSA). Because it is not chromosomally encoded, it would escape detection by ECAT. However, k-mer-based methods are particularly suitable for detecting associations of drug resistance with mobile genes[14,15] as they analyze associations with short DNA fragments derived from the entire genetic content of an organism and are not just restricted to analysis of genetic variants in the core genome (on the chromosome).

Conclusions

We presented a new method, called ECAT, that extends traditional regression-based GWAS methods with the ability to take advantage of homoplasy, which is an important signal of positive selection. This is achieved through a preprocessing step which identifies hypervariable regions in the genome consisting of statistically significant clusters of evolutionary changes, which are then subjected to association testing by an LMM. The efficacy of this approach was demonstrated by showing that it outperforms simple site-based and gene-based GWAS analysis (using GEMMA) in identifying known resistance mutations associated with drug resistance in a collection of MDR clinical isolates of *M. tuberculosis*. The improved sensitivity is attributed to focusing the analysis on optimized clusters of SNPs.

Author Contributions

Y-PL developed the method, implemented the scripts, and performed the experiments. Y-PL and TRI wrote the paper.

Availability of Data and Materials

The ECAT method, which is implemented as a Python script, is freely available at https://github.com/yplai/ECAT

ORCID iD

Thomas R Ioerger https://orcid.org/0000-0001-8702-9102

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

1. Read TD, Massey RC. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. *Genome Med*. 2014;6:109.
2. Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: lessons from human GWAS. *Nat Rev Genet*. 2017;18:41-50.
3. Chen PE, Shapiro BJ. The advent of genome-wide association studies for bacteria. *Curr Opin Microbiol*. 2015;25:17-24.
4. Hicks ND, Yang J, Zhang X, et al. Clinically prevalent mutations in *Mycobacterium tuberculosis* alter propionate metabolism and mediate multidrug tolerance. *Nat Microbiol*. 2018;3:1032-1042.
5. Farhat MR, Freschi L, Calderon R, et al. GWAS for quantitative resistance phenotypes in *Mycobacterium tuberculosis* reveals resistance genes and regulatory regions. *Nat Commun*. 2019;10:2128. doi:10.1038/s41467-019-10110-6.
6. Purcell S, Neale B, Todd-Brown K, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81:559-575.
7. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet*. 2012;44:821.
8. Sul JH, Martin LS, Eskin E. Population structure in genetic studies: confounding factors and mixed models. *PLoS Genet*. 2018;14:e1007309.
9. Earle SG, Wu CH, Charlesworth J, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat Microbiol*. 2016;1:16041.
10. Lees JA, Galardini M, Bentley SD, et al. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics*. 2018;34:4310-1312. doi:10.1093/bioinformatics/bty539.
11. Farhat MR, Shapiro BJ, Kieser KJ, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nat Genet*. 2013;45:1183-1189.
12. 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.
13. Lees JA, Vehkala M, Välimäki N, et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat Commun*. 2016;7:12797. doi:10.1038/ncomms12797.
14. Jaillard M, Lima L, Tournoud M, 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. doi:10.1371/journal.pgen.1007758.
15. Nguyen M, Long SW, McDermott PF, et al. Using machine learning to predict antimicrobial mics and associated genomic features for nontyphoidal salmonella. *J Clin Microbiol*. 2019;57:e01260-18. doi:10.1128/JCM.01260-18.
16. Drouin A, Letarte G, Raymond F, et al. Interpretable genotype-to-phenotype classifiers with performance guarantees. *Sci Rep*. 2019;9:4071. doi:10.1038/s41598-019.
17. Dever LA, Dermody TS. Mechanisms of bacterial-resistance to antibiotics. *Arch Intern Med*. 1991;151:886-895.
18. Munita JM, Arias CA. Mechanisms of antibiotic resistance. *Microbiol Spectr*. 2016;4:1-24. doi:10.1128/microbiolspec.VMBF-0016-2015.
19. Gygli SM, Borrell S, Trauner A, et al. Antimicrobial resistance in *Mycobacterium tuberculosis*: mechanistic and evolutionary perspectives. *FEMS Microbiol Rev*. 2017;41:354-373. doi:10.1093/femsre/fux011.
20. Comas I, Borrell S, Roetzer A, et al. Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. *Nat Genet*. 2011;44:106-110. doi:10.1038/ng.1038.
21. Trindade S, Sousa A, Xavier KB, Dionisio F, Ferreira MG, Gordo I. Positive epistasis drives the acquisition of multidrug resistance. *PLoS Genet*. 2009;5:e1000578. doi:10.1371/journal.pgen.1000578.
22. Dering C, Hemmelmann C, Pugh E, Ziegler A. Statistical analysis of rare sequence variants: an overview of collapsing methods. *Genet Epidemiol*. 2011;35:S12-S17.
23. Lee S, Abecasis GR, Boehnke M, et al. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet*. 2014;95:5-23.
24. Vishcheva GR, Belonogova NM, Zorkoltseva IV, et al. Gene-based association tests using GWAS summary statistics. *Bioinformatics*. 2019;35:3701-3708.
25. Brandley MC, Warren DL, Leache AD, McGuire JA. Homoplasy and clade support. *Syst Biol*. 2009;58:184-198.
26. Walker TM, Kohl TA, Omar SV, et al. Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect Dis*. 2015;15:1193-1202.
27. Mortimer TD, Weber AM, Pepperell CS. Signatures of selection at drug resistance loci in *Mycobacterium tuberculosis*. *Msystems*. 2018;3:e00108-17. doi:10.1128/mSystems.00108-17.

28. Grandjean L, Gilman RH, Iwamoto T, et al. Convergent evolution and topologically disruptive polymorphisms among multidrug-resistant tuberculosis in Peru. *PLoS ONE*. 2017;12:e0189838.
29. Wagner A. Rapid detection of positive selection in genes and genomes through variation clusters. *Genetics*. 2007;176:2451-2463.
30. Benjamini Y, Hochberg Y. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J R Stat Soc Ser B: Stat Methodol*. 1995;57:289-300.
31. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25:1754-1760.
32. Swofford DL. *Paup*. Phylogenetic Analysis Using Parsimony. Version 4*. Sunderland, MA: Sinauer Associates; 2003.
33. Kluge AG, Farris JS. Quantitative phyletics and the evolution of anurans. *Syst Zool*. 1969;18:1-32. http://www.jstor.org/stable/2412407.
34. Sankoff D. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J Appl Math*. 1985;45:810-825.
35. Saber MM, Shapiro BJ. Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes. *Microb Genom*. 2020;6:1-15. doi:10.1099/mgen.0.000337.
36. Didelot X, Maiden MC. Impact of recombination on bacterial evolution. *Trends Microbiol*. 2010;18:315-322.
37. Supply P, Warren RM, Banuls AL, et al. Linkage disequilibrium between minisatellite loci supports clonal evolution of *Mycobacterium tuberculosis* in a high tuberculosis incidence area. *Mol Microbiol*. 2003;47:529-538.
38. Dos Vultos T, Mestre O, Rauzier J, et al. Evolution and diversity of clonal bacteria: the paradigm of *Mycobacterium tuberculosis*. *PLoS ONE*. 2008;3:e1538.
39. Palomino JC, Martin A. Drug resistance mechanisms in *Mycobacterium tuberculosis*. *Antibiotics*. 2014;3:317-340. doi:10.3390/antibiotics3030317.
40. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. Mummer4: a fast and versatile genome alignment system. *PLoS Comput Biol*. 2018;14:e1005944. doi:10.1371/journal.pcbi.1005944.
41. Coll F, Preston M, Guerra-Assunção JA, et al. Polytb: a genomic variation map for *Mycobacterium tuberculosis*. *Tuberculosis*. 2014;94:346-354. doi:10.1016/j.tube.2014.02.005.
42. Osorio NS, Rodrigues F, Gagneux S, et al. Evidence for diversifying selection in a set of *Mycobacterium tuberculosis* genes in response to antibiotic- and nonantibiotic-related pressure. *Mol Biol Evol*. 2013;30:1326-1336. doi:10.1093/molbev/mst038.
43. Telenti A, Imboden P, Marchesi F, et al. Detection of rifampicin-resistance mutations in *Mycobacterium tuberculosis*. *Lancet*. 1993;341:647-651.
44. Kapur V, Li LL, Iordanescu S, et al. Characterization by automated DNA sequencing of mutations in the gene (rpob) encoding the RNA polymerase beta subunit in rifampin-resistant *Mycobacterium tuberculosis* strains from New York city and Texas. *J Clin Microbiol*. 1994;32:1095-1098.
45. Ramaswamy S, Musser JM. Molecular genetic basis of antimicrobial agent resistance in *Mycobacterium tuberculosis*: 1998 update. *Tuber Lung Dis*. 1998;79:3-29.
46. Li Q J, Jiao WW, Yin QQ , et al. Compensatory mutations of rifampin resistance are associated with transmission of multidrug-resistant *Mycobacterium tuberculosis* Beijing genotype strains in china. *Antimicrob Agents Chemother*. 2016;60:2807-2812. doi:10.1128/AAC.02358-15.
47. Telenti A, Philipp WJ, Sreevatsan S, et al. The EMB operon, a gene cluster of *Mycobacterium tuberculosis* involved in resistance to ethambutol. *Nat Med*. 1997;3:567-570.
48. Sreevatsan S, Stockbauer KE, Pan X, et al. Ethambutol resistance in *Mycobacterium tuberculosis*: critical role of EMBB mutations. *Antimicrob Agents Chemother*. 1997;41:1677-1681.
49. Brossier F, Sougakoff W, Bernard C, et al. Molecular analysis of the embcab locus and embr gene involved in ethambutol resistance in clinical isolates of *Mycobacterium tuberculosis* in France. *Antimicrob Agents Chemother*. 2015;59:4800-4808. doi:10.1128/AAC.00150-15.
50. Safi H, Lingaraju S, Amin A, et al. Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-β-d-arabinose biosynthetic and utilization pathway genes. *Nat Genet*. 2013;45:1190-1197. doi:10.1038/ng.2743.
51. He L, Wang X, Cui P, et al. ubia (rv3806c) encoding DPPR synthase involved in cell wall synthesis is associated with ethambutol resistance in *Mycobacterium tuberculosis*. *Tuberculosis*. 2015;95:149-154.
52. Finken M, Kirschner P, Meier A, Wrede A, Böttger EC. Molecular basis of streptomycin resistance in *Mycobacterium tuberculosis*: alterations of the ribosomal protein s12 gene and point mutations within a functional 16s ribosomal RNA pseudoknot. *Mol Microbiol*. 1993;9:1239-1246. doi:10.1111/j.1365-2958.1993.tb01253.x.
53. Okamoto S, Tamaru A, Nakajima C, et al. Loss of a conserved 7-methylguanosine modification in 16S rRNA confers low-level streptomycin resistance in bacteria. *Mol Microbiol*. 2007;63:1096-1106.
54. Wong SY, Lee JS, Kwak HK, Via LE, Boshoff HI, Barry CE 3rd. Mutations in gidB confer low-level streptomycin resistance in *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother*. 2011;55:2515-2522.
55. Spies FS, Ribeiro AW, Ramos DF, et al. Streptomycin resistance and lineage-specific polymorphisms in *Mycobacterium tuberculosis* gidB gene. *J Clin Microbiol*. 2011;49:2625-2630. doi:10.1128/JCM.00168-11.
56. Wang Q, Lau SKP, Liu F, et al. Molecular epidemiology and clinical characteristics of drug-resistant *Mycobacterium tuberculosis* in a tuberculosis referral hospital in China. *PLoS ONE*. 2014;9:e110209. doi:10.1371/journal.pone.0110209.
57. Dixit A, Freschi L, Vargas R, et al. Whole genome sequencing identifies bacterial factors affecting transmission of multidrug-resistant tuberculosis in a high-prevalence setting. *Sci Rep*. 2019;9:5602.
58. Mosbacher TG, Bechthold A, Schulz GE. Structure and function of the antibiotic resistance-mediating methyltrans-ferase AviRb from *Streptomyces viridochromogenes*. *J Mol Biol*. 2005;345:535-545.
59. Gopal P, Yee M, Sarathy J, et al. Pyrazinamide resistance is caused by two distinct mechanisms: prevention of coenzyme A depletion and loss of virulence factor synthesis. *ACS Infect Dis*. 2016;2:616-626.
60. Zhang H, Li D, Zhao L, et al. Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance. *Nat Genet*. 2013;45:1255-1260.
61. Sandgren A, Strong M, Muthukrishnan P, et al. Tuberculosis drug resistance mutation database. *PLoS Med*. 2009;6:e2.
62. Yadon AN, Maharaj K, Adamson JH, et al. Comprehensive characterization of pncA polymorphisms conferring resistance to pyrazinamide. *Nat Commun*. 2017;8:588.
63. Moura de Sousa J, Balbontfn R, Durao P, Gordo I. Multidrug-resistant bacteria compensate for the epistasis between resistances. *PLoS Biol*. 2017;15:e2001741.
64. Yokoyama M, Stevens E, Laabei M, et al. Epistasis analysis uncovers hidden antibiotic resistance-associated fitness costs hampering the evolution of MRSA. *Genome Biol*. 2018;19:94.
65. Vogwill T, Kojadinovic M, MacLean RC. Epistasis between antibiotic resistance mutations and genetic background shape the fitness effect of resistance across species of Pseudomonas. *Proc Roy Soc B*. 2016;283:20160151.
66. Knopp M, Andersson DI. Predictable phenotypes of antibiotic resistance mutations. *mBio*. 2018;9:e00770-18.
67. Niel C, Sinoquet C, Dina C, Rocheleau G. A survey about methods dedicated to epistasis detection. *Front Genet*. 2015;6:285.
68. Chaguza C, Cornick JE, Everett DB. Mechanisms and impact of genetic recombination in the evolution of *Streptococcus pneumoniae*. *Comput Struct Biotechnol J*. 2015;13:241-247. doi:10.1016/j.csbj.2015.03.007.
69. Wyres KL, Wick RR, Judd LM, et al. Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of *Klebsiella pneumoniae*. *PLoS Genet*. 2019;15:e1008114.
70. Didelot X, Wilson DJ. Clonalframeml: efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol*. 2015;11:e1004041.
71. Lai YP, Ioerger TR. A statistical method to identify recombination in bacterial genomes based on SNP incompatibility. *BMC Bioinformatics*. 2018;19:450.