

1 **The advent of genome-wide association studies for bacteria**

2

3 Peter E. Chen<sup>a</sup> & B. Jesse Shapiro<sup>a</sup>

4

5 <sup>a</sup>Département de sciences biologiques, Université de Montréal, Montréal, QC H3C 3J7, Canada

6

7

8 Corresponding author: B. Jesse Shapiro, e-mail: [jesse.shapiro@umontreal.ca](mailto:jesse.shapiro@umontreal.ca);

9

## 10 **Highlights**

- 11 • The advent of the genome-wide association study (GWAS) approach provides a promising framework  
12 for dissecting the genetic basis of bacterial or archaeal phenotypes.
- 13 • Bacterial genomes tend to be shaped by stronger positive selection, stronger linkage disequilibrium and  
14 stronger population stratification than humans, with implications for GWAS power and resolution.
- 15 • An example GWAS in *Mycobacterium tuberculosis* genomes highlights the potentially confounding  
16 effects of linkage disequilibrium and population stratification.
- 17 • A comparison of the traditional GWAS approach versus a somewhat orthogonal method based upon  
18 evolutionary convergence (phyC) shows strengths and weaknesses of both approaches.

19

## 20 **Abstract**

21 Significant advances in sequencing technologies and genome-wide association studies (GWAS) have  
22 revealed substantial insight into the genetic architecture of human phenotypes. In recent years, the  
23 application of this approach in bacteria has begun to reveal the genetic basis of bacterial host preference,  
24 antibiotic resistance, and virulence. Here, we consider relevant differences between bacterial and human  
25 genome dynamics, apply GWAS to a global sample of *Mycobacterium tuberculosis* genomes to highlight  
26 the impacts of linkage disequilibrium, population stratification, and natural selection, and finally compare  
27 the traditional GWAS against phyC, a contrasting method of mapping genotype to phenotype based upon  
28 evolutionary convergence. We discuss strengths and weaknesses of both methods, and make suggestions  
29 for factors to be considered in future bacterial GWAS.

## 30 **Introduction**

31 A central goal of biology is to understand how DNA, the primary sequence, gives rise to observable traits.  
32 Historically, much effort has gone into deciphering the primary sequence of eukaryotes, primarily *Homo*  
33 *sapiens*. As of August 8, 2014, the National Human Genome Research Institute (NHGRI) reported 1,961  
34 publications of genome-wide association studies (GWAS). Within these studies, a total of 14,014 single  
35 nucleotide polymorphisms (SNPs) are associated with over 600 phenotypes. The advent of GWAS in  
36 bacteria has mainly occurred in the last two years [1\*\*, 2\*\*, 3\*\*, 4\*\*, 5\*\*, 6\*\*], and provides an  
37 unbiased "top-down" framework [7] to dissect the genetic basis of bacterial phenotypes. In principle, any  
38 measurable bacterial phenotype (or archaeal phenotype, although here our focus is on bacteria) can be  
39 dissected with a GWAS approach. To date, bacterial GWAS have focused on clinically-relevant  
40 phenotypes such as virulence and antibiotic resistance, but there is also great potential to investigate  
41 environmentally or industrially relevant phenotypes as well.

42

## 43 **Bacterial genomes experience strong linkage, strong stratification and strong selection**

44 Are bacterial genetic mapping studies any different from eukaryotic studies? Although there are many  
45 fundamental differences, this review highlights three features that are most germane to GWAS. The  
46 impact of the first two differences, in linkage and population stratification, have been recognized before  
47 [6\*\*, 7], but we identify the strength of natural selection relative to drift as a third and under-appreciated  
48 factor to consider in bacterial GWAS.

49

50 First, unlike eukaryotic recombination which occurs predominantly via the crossing-over of two  
51 homologous chromosomes during meiosis, bacterial recombination occurs via gene conversion of  
52 relatively short stretches of DNA. In bacteria, recombination is not coupled with reproduction, and can  
53 occur multiple times within a cell's lifespan, or not at all. Without any recombination, purely clonal  
54 transmission of DNA leaves the entire bacterial chromosome in complete linkage (in strong linkage  
55 disequilibrium; LD). As with eukaryotic genomes, bacterial recombination events break this linkage, but

56 the landscape of LD is markedly different from that seen in eukaryotes; gene conversion events leave a  
57 “patchwork” of recombined tracts on top of a genomic background of linked regions called a clonal frame  
58 [8]. In contrast to eukaryotic LD patterns, all regions of the clonal frame are in complete linkage, and  
59 these regions may be quite distant from one another. The clonal frame phenomenon limits the utility of  
60 classic genetic mapping methods mainly by obscuring the true causal variant from the rest of the linked  
61 sites in the clonal frame. Here, we define a variant as causal if it plays a functional role in the phenotype  
62 of interest, as opposed to only being correlated with the phenotype.

63  
64 Second, as with eukaryotes, bacterial genomic diversity may be shaped by population stratification.  
65 Stratification refers to a “situation in which the population of interest includes subgroups of individuals  
66 that are on average more related to each other than to other members of the wider population” [9]. These  
67 subpopulations give rise to spurious associations when "cases" (with phenotype A) are on average more  
68 closely related with each other than with "controls" (without phenotype A); in other words, associations  
69 due to genetic relatedness rather than causality for the phenotype of interest. The problem of population  
70 stratification is particularly acute in highly clonal (rarely recombining) bacteria, and in those with  
71 separate geographic or host-associated subpopulations [6\*\*].

72  
73 Third, the phenotypes of most interest in bacterial GWAS are largely different from many human disease  
74 phenotypes. In particular, bacterial phenotypes tend to be shaped by strong natural selection (e.g. positive  
75 directional selection driving drug resistance), while many human disease phenotypes evolve largely by  
76 genetic drift owing to historically small effective population sizes (e.g. due to population bottlenecks); in  
77 this scenario, drift overpowers purifying selection and leaves slightly deleterious alleles in the population  
78 that underlie disease traits [10, 11]. This is not to say that bacteria do not experience genetic drift  
79 (particularly in frequently bottlenecked populations), but simply that many traits of interest (e.g.  
80 resistance, virulence, host-association) have evolved recently and under strong positive selection. These  
81 bacterial traits might also be controlled by mutations with large effect sizes on the phenotypes of interest.

82 If this is the case, relatively small samples of bacterial genomes should be sufficient to identify causal  
83 mutations [11, 12].

84

### 85 **Units of genetic and phenotypic variation**

86 The two basic requirements for GWAS are genotypic and phenotypic measurements from a sample of  
87 organisms. Phenotypes are usually broken into either discrete units (e.g. resistance/sensitive or high/low  
88 virulence) or continuous traits (e.g. human height). Phenotypes must be reproducible, and easy to  
89 measure, ideally in high-throughput if hundreds or thousands of samples are being studied. At the  
90 genotypic level, a set of bacterial genomes can be broken down into a “core” genome shared among  
91 nearly all members and an “accessory” genome composed of elements present in some strains but not  
92 others (typically including genes involved in environmental adaptation) [13, 14]. The genetic units of a  
93 GWAS may be variants in the core (e.g. single nucleotide polymorphisms (SNPs) or small indels) [2\*\*,  
94 3\*\*, 4\*\*, 5\*\*] or in the flexible genome (e.g. presence/absence of larger pieces of DNA including genes  
95 or operons [1\*\*, 15, 16, 17] (Table 1). While most bacterial GWAS to date have studied either SNPs or  
96 gene presence/absence, Sheppard et al. [1\*\*] described a method that uses n-mers ("words" of DNA) as  
97 the basic unit of association, allowing them to study both the core and flexible genome simultaneously.

98

### 99 **Allele counting and homoplasy counting approaches to GWAS**

100 GWAS approaches for bacteria can be broadly broken down into allele counting [1\*\*, 3\*\*, 4\*\*, 5\*\*] and  
101 homoplasy counting [2\*\*, 12] methods (Table 1 and Graphical Abstract). The primary association signal  
102 for allele counting methods is derived from an over-representation of an allele at the same site in cases  
103 relative to controls, which can later be corrected for population stratification. In contrast, homoplasy  
104 counting methods (in this case, phyC [2\*\*]) derives its evidence of association by counting repeated and  
105 independently emerged mutations occurring more often on branches of cases relative to controls.  
106 Homoplasy, as an indicator of convergent evolution, is a well-known signal of positive selection [28].

107 Combining this signal of selection with phenotypic associations (e.g. convergent mutations that occur  
108 only in cases and not in controls) provides the basis for homoplasmy-based association tests.

109

### 110 **Architecture of a strong association signal**

111 GWAS signals from allele counting and homoplasmy counting methods are not expected to perfectly  
112 overlap because each method represents different strengths and weaknesses. However, with a sufficiently  
113 large sample size, allele counting methods theoretically can detect all convergent sites (identified by  
114 homoplasmy counting methods) as well as non-convergent sites. Still, ever-increasing sample size does not  
115 directly address the confounding effects of both population stratification and LD on allele counting  
116 methods. Homoplasmy counting intrinsically accounts for these effects by virtue of its phylogenetic  
117 convergence criterion. In contrast, allele counting methods have no such phylogenetic requirement. Thus,  
118 a monophyletic group containing many cases with the same over-represented allele at the same site may  
119 provide a strong signal for allele counting while providing no signal for homoplasmy counting.

120 Conversely, homoplasmy counting requires a smaller count of homoplasmy events (versus allele counts) in  
121 order to reach statistical significance; thus, a relatively small sample size with a strong paraphyletic  
122 structure may provide homoplasmy counting with a much stronger signal than allele counting.

123

### 124 **A genome-wide association study of antibiotic drug resistance in *Mycobacterium tuberculosis***

125 To examine the impacts of clonal frames (strong LD) and population stratification, we performed a  
126 'traditional' GWAS using PLINK on a population of 123 *M. tuberculosis* (MTB) genomes that had been  
127 previously analyzed by phylogenetic convergence (phyC) [2\*\*]. Of the 123 strains, 47 (cases) are  
128 resistant to at least one antibiotic and 76 strains are sensitive to all antibiotics (controls). This dataset  
129 contains 11 'gold standard' experimentally-verified antibiotic resistance alleles, all of which were  
130 identified by phyC, along with 39 new phyC hits in nonsynonymous coding sites and intergenic regions,  
131 and 7 hits in synonymous sites. We chose this particular MTB dataset as it allows a comparison of the

132 results from traditional GWAS and phyC, and also because MTB genomes possess extensive LD and  
133 strong population structure, making them challenging subjects for traditional GWAS.

134

### 135 **Clonal frames and the resolution of GWAS signals**

136 MTB is considered to be a highly clonal pathogen, with very little detectable recombination [18].  
137 Consistent with this, we observe a clonal frame consisting of linked sites across the genome. This clonal  
138 frame is evident from the extensive genome-wide linkage (black or red in Figures 1a and b, respectively),  
139 interrupted by a few homoplastic sites (small white or black points, respectively) identified by the four-  
140 gamete test [19] or the  $D'$  measure [20] of linkage (Figure 1a and b). The  $r^2$  measure [21] does not directly  
141 measure recombination or homoplasmy, but rather how well the allelic state at one site in the genome can  
142 predict the allele present at another site. The  $r^2$  analysis confirms that MTB has extensive genome-wide  
143 LD, posing a challenge to pinpointing causal variants (Figure 1c and d). Other more highly recombining  
144 bacteria, such as *Streptococcus pneumoniae* (Figure 1e) have less long-range LD and more localized,  
145 shorter LD blocks (black triangles near the horizontal axis), facilitating GWAS [5\*\*]. Because the extent  
146 of genome-wide linkage is unlikely to be known *a priori*, an important first step before performing a  
147 bacterial GWAS is to characterize LD, as illustrated here (Figure 1).

148

### 149 **Correcting for population stratification**

150 The strong clonal nature of MTB also creates strong population substructures that in turn may lead to  
151 false positive associations. Without any population stratification correction we observe a substantial  
152 systematic inflation of the association test p-values (Figure 2a), likely due to both causal and non-causal  
153 resistance-associated mutations being linked on the same clonal frame. We assessed two classic methods  
154 of addressing population stratification. The first method, called genomic control [22], normalizes all p-  
155 values by a single inflation factor  $\lambda$ , which is the observed median chi-square divided by the expected  
156 median chi-square with 1 degree of freedom. Due to a relatively large observed inflation factor ( $\lambda =$   
157 12.20), genomic control seems to over-correct, leaving no statistically significant GWAS hits (Figure 2b).

158 A less conservative correction for population stratification is to infer ancestry by identifying genetic  
159 subpopulations within the overall population, and then testing for association conditional on these  
160 subpopulations. Subpopulations can be inferred using a variety of methods (e.g. multi-dimensional  
161 scaling in PLINK [23], principal component analysis in EIGENSTRAT[24], and BAPS [25]), then used  
162 as covariates in association testing (e.g. with the Cochran-Mantel-Haenszel test). Here, we defined  
163 subpopulations based on 14 previously defined MTB epidemiological clusters [2\*\*]. Using these epi-  
164 clusters as covariates reduced the inflation factor to 1, suggesting that it effectively controls for  
165 population stratification (Figure 2c). Although this procedure clearly changes the Manhattan plot (Figure  
166 2, right panels), producing more clean 'hits' that stand out from the average p-value, we note that none of  
167 these hits pass correction for multiple hypothesis testing. Therefore, correcting for population  
168 stratification can reduce GWAS power significantly – a problem that could potentially be overcome by  
169 using larger sample sizes (e.g. thousands rather than hundreds of genomes; [5\*\*]).

170

### 171 **Comparison of GWAS against convergence testing**

172 Despite the lack of significance after multiple testing correction, we identified 133 potential GWAS hits  
173 (SNPs) in 77 genes that stood out as outliers from the average genome-wide p-value (Figure 2c), which  
174 we will discuss for illustration purposes. These GWAS hits overlapped with 5 of 11 previously known  
175 'gold standard' resistance genes and 22 of 46 additional phyC candidate resistance genes. It is also  
176 evident that correcting for population stratification improves the overlap with known resistance genes and  
177 phyC hits (Figure 3).

178

### 179 **Potential new drivers of drug resistance**

180 Of these top 133 GWAS hits, 75 SNPs (in 50 genes) did not overlap with either known resistance genes  
181 nor with phyC candidates (Figure 3b). Due to long-range LD, it is not immediately clear without further  
182 analysis whether these 75 SNPs represent false positives due to their correlation with the true drivers of  
183 resistance, either 'gold standard' resistance genes or phyC candidates. However, 15 out of these 75 SNPs



184 were relatively uncorrelated ( $r^2 < 0.3$ ) with any of the other 133 top GWAS hits, suggesting they could  
185 play causal roles in resistance phenotypes. As an example to illustrate the importance of assessing LD  
186 patterns around GWAS hits, the top GWAS hit (a nonsense mutation in an oxidoreductase gene, Rv0197)  
187 can be viewed from two different perspectives:

188

- 189 1) The top GWAS hit may be a false positive because it is in moderate correlation ( $0.4 < r^2 < 0.5$ ) with  
190 two phyC candidates (PPE9 and PE\_PGRS4 genes) and two other GWAS hits (PE-PGRS30 and PE-  
191 PGRS46 genes), and does not represent a true causal variant.
- 192 2) The top GWAS hit may be driving the association. Although it is in moderate correlation with four  
193 other phyC or GWAS hits, all four hits reside within the PE/PGRS families of genes, which are  
194 highly polymorphic and might represent false positive associations [2\*\*].

195 Whether this GWAS hit is causal or not can only be firmly established with followup experiments.

196

### 197 **Future Directions**

198 We have shown the potential of GWAS for bacterial genomes while highlighting two key obstacles: long-  
199 range LD within the clonal frame and extensive bacterial population stratification both reduce our ability  
200 to pinpoint causal mutations with confidence. However, a third feature of bacterial genomes – the relative  
201 strength of positive selection – provides an opportunity to increase the resolution and confidence of  
202 GWAS hits. One could combine positive selection tests and GWAS, as has been done previously for traits  
203 shaped by positive selection [26, 27]. This approach may potentially address the problem of clonal frames  
204 obscuring true causal variants and making them indistinguishable from linked non-causal variants. This  
205 idea attempts to identify causal variants in two steps:

206

- 207 1. perform a genome-wide selection scan identifying any genomic regions that are putatively under  
208 positive selection
- 209 2. perform a “targeted” association study on each genomic region under positive selection

210

211 The rationale here is that each genomic region identified as being under positive selection effectively  
212 “unlinks” the putative causal variants from its background clonal frame, provided that the selection test  
213 itself can distinguish a positive selection region from the clonal frame upon which it occurred [28]. Since  
214 positive selection alone does not provide sufficient evidence that a region is associated with the  
215 phenotype of interest, step two targets each of the genomic regions identified in the selection scan and  
216 tests each one for association with a phenotype of interest. In phyC, the two steps are done  
217 simultaneously, using convergence as the signal of positive selection and the *specificity* of convergence to  
218 cases but not controls as the association signal. Future work might 'mix and match' different signals of  
219 selection and association.

220

221 As this new and growing field develops, we envision a future where multiple genetic mapping approaches  
222 – including GWAS, phyC and selection scans – are combined. Each method may harbor its own  
223 strengths and weaknesses so that when combined, each method provides distinct information, thus  
224 increasing the power to detect true and causal associations.

225

## 226 **Acknowledgements**

227 This work was funded by the Natural Sciences and Engineering Research Council, the Canadian Institutes  
228 for Health Research and the Canada Research Chairs program. We would like to thank Luis Barreiro and  
229 Jean-Baptiste Leducq for valuable feedback and discussions.

230

231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252

### Figure captions

#### Figure 1. Patterns of linkage disequilibrium (LD) in bacterial genomes assessed by different metrics.

The x-axis of heat maps (a-d) represents the physical position along the MTB genome; (e) shows the *S. pneumoniae* genome. Each square in the heat map represents a pairwise calculation of LD.

- a) Four-gamete test. White squares denote four observed haplotypes indicating recombination may have occurred between the two sites. Black squares denote three or fewer observed haplotypes (strong linkage).
- b) Pairwise  $|D'|$  measurements (range of  $|D'|$  values:  $0 \leq |D'| \leq 1$ ). Red squares denote  $|D'| = 1$  (strong linkage). Black squares denote  $|D'| < 1$ .
- c) Pairwise  $r^2$  measurements (range of  $r^2$  values:  $0 \leq r^2 \leq 1$ ). Black squares denote  $r^2 = 1$  (strong correlation). The lighter squares denote progressively smaller  $r^2$  values.
- d) Pairwise  $r^2$  measurements for the top 133 GWAS hits only. Black squares denote  $r^2 = 1$ . The lighter squares denote progressively smaller  $r^2$  values.
- e) Pairwise  $r^2$  measurements of beta-lactam resistance associated variants co-detected in two separate *S. pneumoniae* populations [5\*\*]. Black squares denote  $r^2 = 1$ . The lighter squares denote progressively smaller  $r^2$  values.

#### Figure 2. GWAS for antibiotic resistance in MTB.

GWAS was performed using Plink version 1.07 [23]. The x=y line (red in QQ plots; left) represents the null hypothesis of no association signal. In Manhattan plots (right), SNPs in 'Gold Standard' resistance genes are shown in red, and SNPs in phyC candidate genes in green (excluding synonymous sites). Different corrections for population stratification were applied:

- a) No population stratification correction.
- b) Population correction with genomic control.
- c) Population correction using “epi-clusters” and Cochran-Mantel-Haenszel 2x2xK test, where K = 14 epi-clusters.

#### Figure 3. Overlap among GWAS candidates, phyC candidates, and “Gold standard” resistance genes.

Numbers in Venn diagrams are in units of genes or intergenic regions.

- a) No population stratification correction.
- b) Population correction using “epi-clusters” and Cochran-Mantel-Haenszel 2x2xK test, where K = 14 epi-clusters.

268

269 **References**

270 1. Sheppard, S. K., Didelot, X., Méric, G., Torralbo, A., Jolley, K. A., Kelly, D. J., et al. (2013). Genome-  
271 wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*.  
272 Proceedings of the National Academy of Sciences, 110(29), 11923–11927.

273 doi:10.1073/pnas.1305559110

274 \*\* This method is unique in that it simultaneously addresses both the core and accessory genomes.

275

276 2. Farhat, M. R., Shapiro, B. J., Kieser, K. J., Sultana, R., Jacobson, K. R., Victor, T. C., et al. (2013).

277 Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium*  
278 *tuberculosis*. Nature Genetics, 45(10), 1183–1189. doi:10.1038/ng.2747

279 \*\* Currently, the only GWAS method to explicitly combine evolutionary convergence with association.

280

281 3. Laabei, M., Recker, M., Rudkin, J. K., Aldeljawi, M., Gulay, Z., Sloan, T. J., et al. (2014). Predicting  
282 the virulence of MRSA from its genome sequence. Genome Research, 24(5), 839–849.

283 doi:10.1101/gr.165415.113

284 \*\* Despite a relatively low sample size (90 strains) and a low-recombining population, a > 85%

285 predictive accuracy for a toxicity phenotype was achieved, thus highlighting the potential of GWAS.

286

287 4. Alam, M. T., Petit, R. A., Crispell, E. K., Thornton, T. A., Conneely, K. N., Jiang, Y., et al. (2014).

288 Dissecting Vancomycin-Intermediate Resistance in *Staphylococcus aureus* Using Genome-Wide  
289 Association. Genome Biology and Evolution, 6(5), 1174–1185. doi:10.1093/gbe/evu092

290 \*\* An example that pinpoints one resistance gene, *rpoB*, using both allele and homoplasy counting  
291 methods.

292

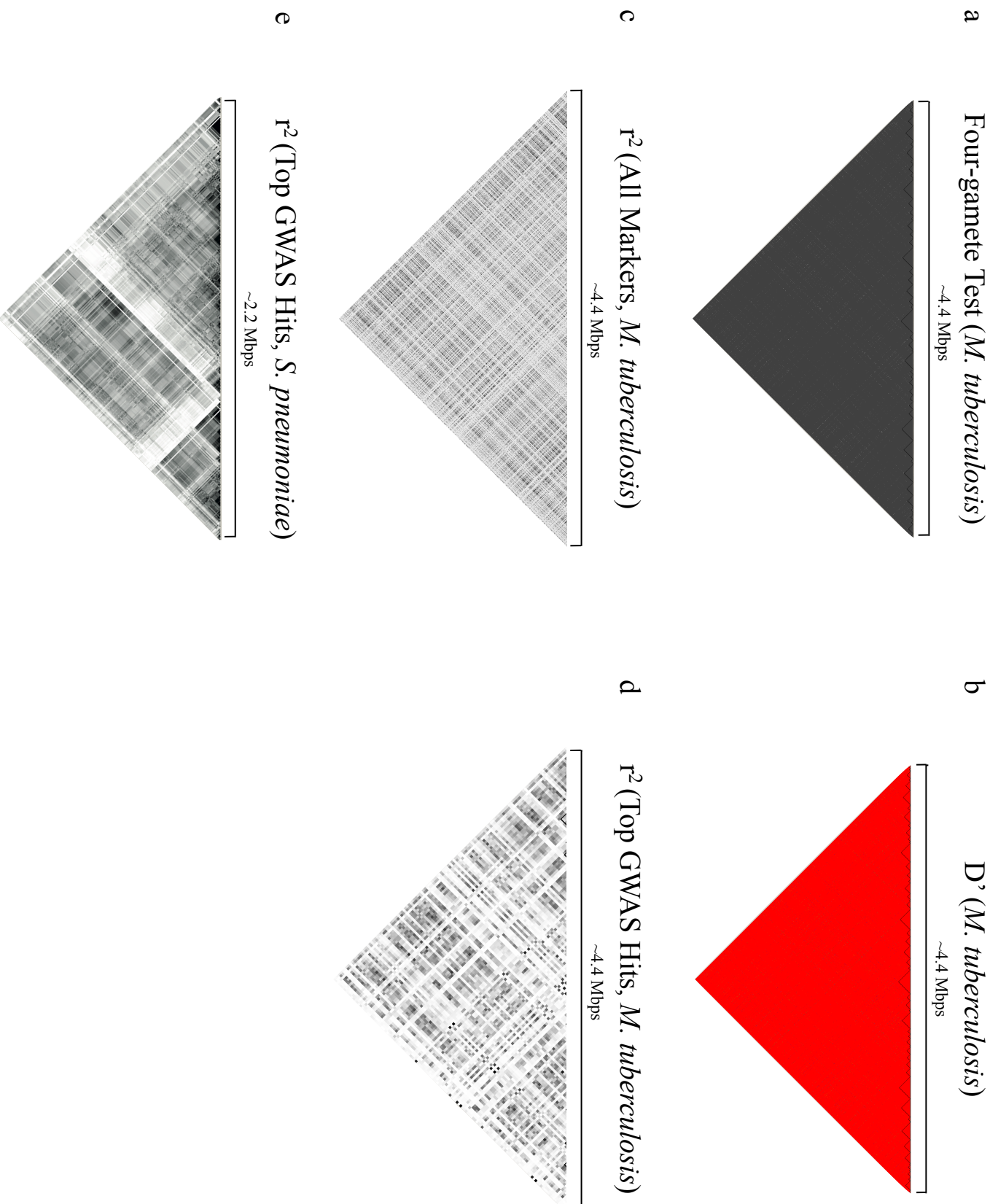
293 5. Chewapreecha, C., Marttinen, P., Croucher, N. J., Salter, S. J., Harris, S. R., Mather, A. E., et al.

294 (2014). Comprehensive Identification of Single Nucleotide Polymorphisms Associated with Beta-

- 295 lactam Resistance within Pneumococcal Mosaic Genes. *PLoS Genetics*, 10(8), e1004547.  
296 doi:10.1371/journal.pgen.1004547.s008
- 297 \*\* An example of a high-powered, high-resolution study (3,085 strains) in populations with high levels of  
298 recombination.
- 299
- 300 6. Falush, D., & Bowden, R. (2006). Genome-wide association mapping in bacteria? *Trends in*  
301 *Microbiology*, 14(8), 353–355. doi:10.1016/j.tim.2006.06.003
- 302 \*\* The first paper to broach the potential and challenges of GWAS in bacterial genomes.
- 303
- 304 7. Read, T. D., & Massey, R. C. (2014). Characterizing the genetic basis of bacterial phenotypes using  
305 genome-wide association studies: a new direction for bacteriology. *Genome Medicine* 6:109  
306 doi:10.1186/s13073-014-0109-z
- 307 8. Milkman, R., & Bridges, M. M. K. (1990). Molecular evolution of the *Escherichia coli* chromosome.  
308 III. Clonal frames. *Genetics*, 126(3), 505–517.
- 309 9. Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature*  
310 *Reviews Genetics*, 7(10), 781–791. doi:10.1038/nrg1916
- 311 10. Gherman, A., Chen, P. E., Teslovich, T. M., Stankiewicz, P., Withers, M., Kashuk, C. S., et al.  
312 (2007). Population bottlenecks as a potential major shaping force of human genome architecture.  
313 *PLoS Genetics*, 3(7), e119. doi:10.1371/journal.pgen.0030119
- 314 11. Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al.  
315 (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265), 747–753.  
316 doi:10.1038/nature08494
- 317 12. Farhat, M. R., Shapiro, B. J., Sheppard, S. K., Colijn, C., & Murray, M. (2014). A phylogeny-based  
318 sampling strategy and power calculator informs genome-wide associations study design for microbial  
319 pathogens. *Genome Medicine* 6:101. doi:10.1186/s13073-014-0101-7

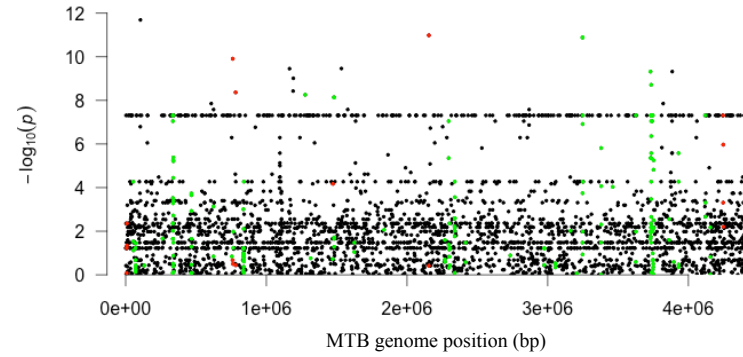
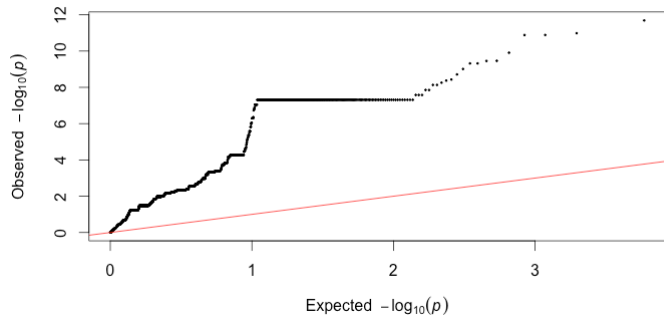
- 320 13. Lapiere, P., & Gogarten, J. P. (2009). Estimating the size of the bacterial pan-genome. Trends in  
321 Genetics : TIG, 25(3), 107–110. doi:10.1016/j.tig.2008.12.004
- 322 14. Vernikos, G., Medini, D., Riley, D. R., & Tettelin, H. (2015). Ten years of pan-genome analyses.  
323 Current Opinion in Microbiology, 23, 148–154. doi:10.1016/j.mib.2014.11.016
- 324 15. van Hemert, S., Meijerink, M., Molenaar, D., Bron, P. A., de Vos, P., Kleerebezem, M., et al. (2010).  
325 Identification of *Lactobacillus plantarum* genes modulating the cytokine response of human peripheral  
326 blood mononuclear cells. BMC Microbiology, 10(1), 293. doi:10.1186/1471-2180-10-293
- 327 16. Salipante, S. J., Roach, D. J., Kitzman, J. O., Snyder, M. W., Stackhouse, B., Butler-Wu, S. M., et al.  
328 (2014). Large-scale genomic sequencing of extraintestinal pathogenic *Escherichia coli* strains.  
329 Genome Research. doi:10.1101/gr.180190.114
- 330 17. Chaston, J. M., Newell, P. D., & Douglas, A. E. (2014). Metagenome-Wide Association of Microbial  
331 Determinants of Host Phenotype in *Drosophila melanogaster*. mBio, 5(5), e01631–14–e01631–14.  
332 doi:10.1128/mBio.01631-14
- 333 18. Smith NH, Gordon SV, de la Rúa-Domenech R, Clifton-Hadley RS, Hewinson RG (2006).  
334 Bottlenecks and broomsticks: the molecular evolution of *Mycobacterium bovis*. Nature Reviews  
335 Microbiology, 4(9), 670–681. doi:10.1038/nrmicro1472
- 336 19. Hudson, R. R., & Kaplan, N. L. (1985). Statistical properties of the number of recombination events  
337 in the history of a sample of DNA sequences. Genetics, 111(1), 147–164.
- 338 20. Lewontin, R. C. (1964). The Interaction of Selection and Linkage. I. General Considerations;  
339 Heterotic Models. Genetics, 49(1), 49–67.
- 340 21. Hill, W. G. & Robertson, A. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38,  
341 226–231 (1968).
- 342 22. Devlin, B., & Roeder, K. (1999). Genomic control for association studies. Biometrics, 55(4), 997–  
343 1004.

- 344 23. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007).  
345 PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The*  
346 *American Journal of Human Genetics*, 81(3), 559–575. doi:10.1086/519795
- 347 24. Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006).  
348 Principal components analysis corrects for stratification in genome-wide association studies. *Nature*  
349 *Genetics*, 38(8), 904–909. doi:10.1038/ng1847
- 350 25. Corander, J., Marttinen, P., Sirén, J., & Tang, J. (2008). Enhanced Bayesian modelling in BAPS  
351 software for learning genetic structures of populations. *BMC Bioinformatics*, 9(1), 539.  
352 doi:10.1186/1471-2105-9-539
- 353 26. Cheeseman, I. H., Miller, B. A., Nair, S., Nkhoma, S., Tan, A., Tan, J. C., et al. (2012). A Major  
354 Genome Region Underlying Artemisinin Resistance in Malaria. *Science*, 336(6077), 79–82.  
355 doi:10.1126/science.1215966
- 356 27. Karlsson, E. K., Harris, J. B., Tabrizi, S., Rahman, A., Shlyakhter, I., Patterson, N., et al. (2013).  
357 Natural selection in a bangladeshi population from the cholera-endemic ganges river delta. *Science*  
358 *Translational Medicine*, 5(192), 192ra86. doi:10.1126/scitranslmed.3006338
- 359 28. Shapiro, B. J., David, L. A., Friedman, J., & Alm, E. J. (2009). Looking for Darwin's footprints in the  
360 microbial world. *Trends in Microbiology*, 17(5), 196–204. doi:10.1016/j.tim.2009.02.002

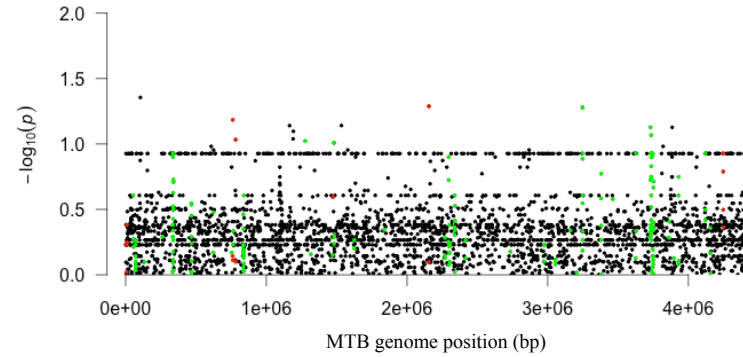
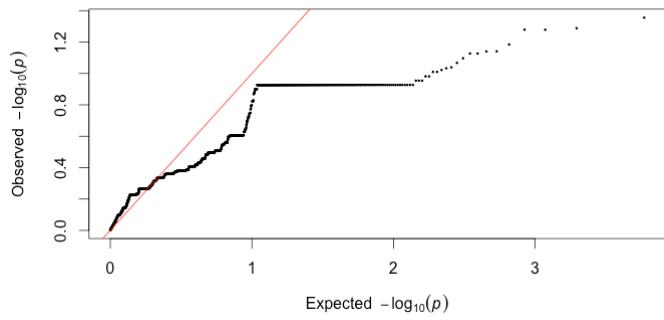




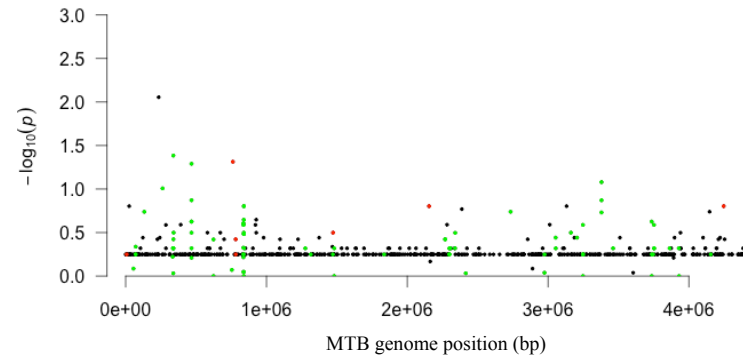
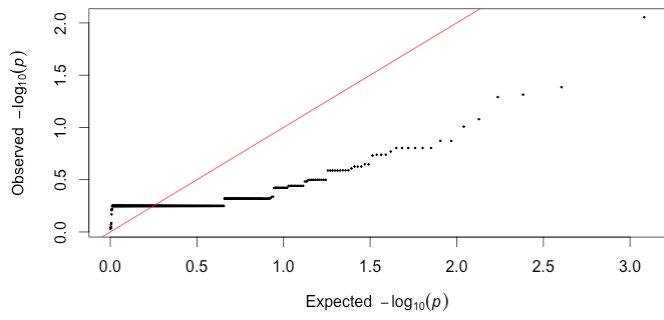
### a No Population Stratification Correction

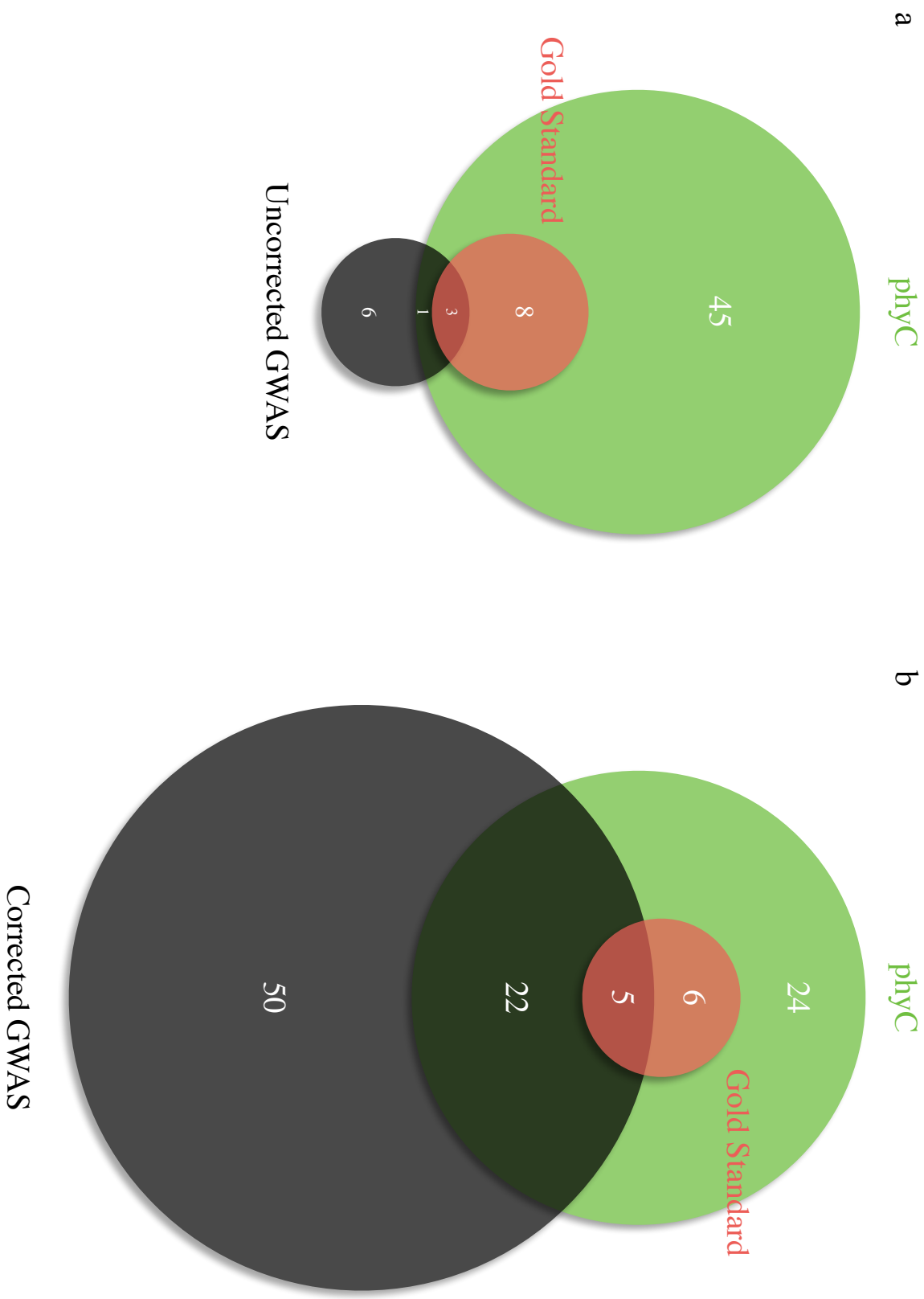


### b Genomic Control



### c Epi-clusters





Study	Year	Taxa	Relative recombination rate	# genomes	Phenotype	Association method	Addresses accessory genome?	Unit of genetic variation studied	# of variants	Correction for population stratification
Sheppard et al. [1]	2013	<i>C. jejuni</i>	moderate	29 (+ validation in 161)	host specificity	allele counting	yes	30-bp DNA sequences (words)	>10,000 words (?)	simulation of word gain/loss along the phylogenetic tree
Farhat et al. [2]	2013	<i>M. tuberculosis</i>	low	123	antibiotic resistance	homoplasy counting	no	SNPs	~25,000	implicit in phylogenetic convergence criterion
Chen & Shapiro (This review)	2015	<i>M. tuberculosis</i>	low	123	antibiotic resistance	allele counting	no	SNPs	~3,000 MAF > 0.05)	inferred ancestry clusters
Laabei et al. [3]	2014	<i>S. aureus</i>	low	90	virulence	allele counting	no	SNPs & small indels	~3000	genomic control
Alam et al. [4]	2014	<i>S. aureus</i>	low	75	antibiotic resistance	allele counting and homoplasy counting	no	SNPs	~55,000	inferred ancestry clusters
Chewapreecha et al. [5]	2014	<i>S. pneumoniae</i>	high	3085 (+ validation in 616)	antibiotic resistance	allele counting	no	SNPs	~400,000 (MAF > 0.01)	inferred ancestry clusters
Salipante et al. [16]	2014	<i>E. coli</i>	low-moderate	312	antibiotic resistance	allele counting	yes	gene presence/absence	~15,000 genes	inferred ancestry clusters
Chaston et al. [17]	2014	41 strains	N/A	41	host development time and triglyceride content	allele counting	yes	gene presence/absence	~12,000 genes	consideration of genes with unique phylogenetic distributions
van Hemert et al. [15]	2010	<i>L. plantarum</i>	low	42	host immune response	allele counting	yes	gene presence/absence	? (CGH)	none

SNP = single nucleotide polymorphism, MAF = minor allele frequency, CGH = comparative genomic hybridization

