A phylogeny-based sampling strategy and power calculator informs genome-wide associations study design for microbial pathogens

Genome Medicine

METHOD

A phylogeny-based sampling strategy and power calculator informs genome-wide associations study design for microbial pathogens

Maha R Farhat[1,5*], B Jesse Shapiro[2], Samuel K Sheppard[3], Caroline Colijn[4] and Megan Murray[5,6]

Abstract

Whole genome sequencing is increasingly used to study phenotypic variation among infectious pathogens and to evaluate their relative transmissibility, virulence, and immunogenicity. To date, relatively little has been published on how and how many pathogen strains should be selected for studies associating phenotype and genotype. There are specific challenges when identifying genetic associations in bacteria which often comprise highly structured populations. Here we consider general methodological questions related to sampling and analysis focusing on clonal to moderately recombining pathogens. We propose that a matched sampling scheme constitutes an efficient study design, and provide a power calculator based on phylogenetic convergence. We demonstrate this approach by applying it to genomic datasets for two microbial pathogens: *Mycobacterium tuberculosis* and *Campylobacter* species.

Background

In infectious disease, host and pathogen factors interact to result in the observed severity of illness. Genetic changes within pathogen populations can result in a spectrum of virulence, drug resistance, transmission rates, and immunogenicity - all highly relevant phenotypes in the study of infectious disease. Host variables that affect susceptibility to infection, such as age, immunodeficiency, and nutritional status are more easily measured and have been studied for some time, whereas the study of pathogen specific determinants of disease risk is more recent. One of the first to use the term molecular epidemiology and apply it to infectious disease agents was E. Kilbourne. In his 1973 paper 'Molecular epidemiology of influenza', he discussed antigenic variation as a cause of the influenza pandemics of the 20th century [1]. The ability to type molecular traits of pathogens, such as surface proteins or highly variable DNA segments, allowed the characterization of sufficient strain-to-strain variation to determine when transmission of disease occurred [2] as well as surveillance of the frequencies of different strain types over time [3]. As sequencing became sufficiently high throughput to allow for whole genome analysis, the typing resolution immediately reached the limit for heritable strain differences and has accordingly gained momentum in the study of infectious disease [4-7].

Molecular epidemiologic tools have not only enabled disease surveillance and the study of transmission chains, but also have facilitated the study of pathogen biology, by allowing researchers to compare the transmissibility, immunogenicity, or other phenotypes that vary among strain types or lineages and correlate these differences with specific changes in the genome [8,9]. Large numbers of pathogen samples are often gathered for clinical diagnostic purposes. For pathogens of high outbreak potential, samples may be collected for surveillance purposes. The short evolutionary times corresponding to outbreaks often mean that samples of transmitted pathogens are clonal. The availability of samples from diagnostic and outbreak setting, and the DNA sequences generated from them, means that investigators are faced with questions about which and how many pathogen isolates to sequence and which analytical techniques to use to maximize efficiency and power. These questions are especially relevant for studies of whole-genome sequences (WGS) that will

* Correspondence: mrfarhat@partners.org
[1]Department of Pulmonary and Critical Care, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
[5]Department of Global Health and Social Medicine, Harvard Medical School, 641 Huntington Avenue Suite 4A, Boston, MA 02115, USA
Full list of author information is available at the end of the article

generate thousands of potentially relevant mutations, the great majority of which will be noise, that is, neutral mutations not related to the phenotype of interest.

The methods underlying human genome-wide association studies (GWAS) and whole exome sequencing have advanced significantly in the past 10 years, and are now more rigorous and standardized across studies of different human traits and diseases [10,11]. These advancements have included recommendations on study design including subject selection strategies and sample size to uncover elements of varying frequency and effect sizes. These methods are most well developed for single nucleotide polymorphism (SNP) changes in typing data (as opposed to whole genome sequences) and make implicit assumptions about the human genomic structure, diploidy, and recombination rates [12-14]. The situation is different in bacteria where recombination and genetic mutation rates vary among species, from highly clonal organisms like *Mycobacterium tuberculosis* (MTB), to the rapidly recombining/sexual *Streptococcus pneumoniae*. In contrast to disease states in humans, pathogen phenotypes of interest are often those that provide a selective advantage for the organism. Several different methods are in current use for the study of genome wide variation of pathogens that, in contrast to human genetic association studies, can frequently leverage information about positive selection. Despite this, the field has not yet defined accepted methodologies and standards for statistical testing of variants on a whole genome scale. In this paper we review the literature on genotype-phenotype studies and analytical techniques focusing on MTB as an example. We propose a matched genome sampling and analysis strategy to optimize power for pathogens that are clonal to moderately sexual. We provide an associated power and sample size calculator and demonstrate and validate the method using two genomic datasets: one from MTB and one from *Campylobacter* species.

Methods

The methods outlined below were used for the application of the sampling strategy.

Strain isolation, culture, sequencing, and variant calling are detailed in the original publications [15,16].

Phylogeny construction

MTB: The phylogeny was constructed based on the whole genome multiple alignment. As MTB populations are considered to be predominantly clonal, most of the genome is thought to support a single consensus phylogeny that is not impacted significantly by recombination [17]. A superset of SNPs relative to reference strain H37Rv [18] was created across the clinical isolates from the variant caller SNP reports. SNPs occurring in repetitive elements including transposases, PE/PPE/PGRS genes, and phiRV1 members (273 genes, 10% of genome) (genes listed in reference [19]) were excluded to avoid any concern about inaccuracies in the read alignment in those portions of the genome. Furthermore, SNPs in an additional 39 genes previously associated with drug resistance [20] were also removed to exclude the possibility that homoplasy of drug resistance mutations would significantly alter the phylogeny. After applying these filters the remaining SNPs were concatenated and used to construct a parsimony phylogenetic tree using PHYLIP dnapars algorithm v3.68 [21] with KZN-DS [22] strain as an outgroup root. We constructed a phylogeny by two methods. First, using Bayesian Markov chain Monte Carlo (MCMC) methods as implemented in the package MrBayes v3.2 [23] using the GTR model and a maximum likelihood tree using PhyML v3.0 [24]. Second, using the GTR model with eight categories for the gamma model and the results were consistent with the PHYLIP Phylogeny.

Campylobacter

Using multi-local sequence typing data, a phylogeny was estimated using ClonalFrame [25], a model-based approach to determining microevolution in bacteria. This program differentiates mutation and recombination event on each branch of the tree based on the density of polymorphisms. ClonalFrame was run with 50,000 burn in iterations and 50,000 sampling iterations. The consensus tree represents combined data from three independent runs with 75% consensus required for inference of relatedness. Recombination events were defined as sequences with a length of >50 bp with a probability of recombination > =75% over the length, reaching 95% in at least one site.

Analysis

The number of mutations, insertions, or deletions (of any size) differing between each strain pair was summed across each locus for the eight strain pairs for each of the two datasets belonging to MTB or *Campylobacter*. The upper 95% confidence interval for the average number of mutations/locus across the eight pairs was used as a mean of the null Poisson distribution. All genes with larger counts than expected under this null distribution were considered to be significantly association with the resistance phenotype.

Results and Discussion

Literature search

We first defined five cornerstones of a systematically designed microbial genotype-phenotype association study: (1) a well-defined phenotype of interest, that can be measured/classified with negligible error; (2) some understanding of the effect size for that phenotype, for example is it influenced by many genetic variants each with small or incremental effect, or are there fewer variants with a large effect?; (3) estimates of the number of whole

genomes needed to achieve nominal power; (4) a sampling strategy that may include the sequencing of pathogens serially sampled over time from the same patient, the study of strains matched by some predefined characteristic, a 'random' subsample, or an 'exhaustive' complete sample; and (5) a defined statistical analysis strategy that maximizes power and minimizes the rate of false positives.

We performed a systematic search of the literature to determine which sampling and analytical strategies (the five components above) have been applied to the study of MTB biology using whole genome sequences. We sought articles studying one of the following aspects of MTB biology: immunogenicity, pathogenicity, virulence, transmissibility, drug resistance, or fitness using whole genome sequences. Search terms, inclusion and exclusion criteria are detailed in Table 1. We searched PubMed on 1 September 2013 and identified 216 abstracts, and included 16 studies (Figure 1, Table 2).

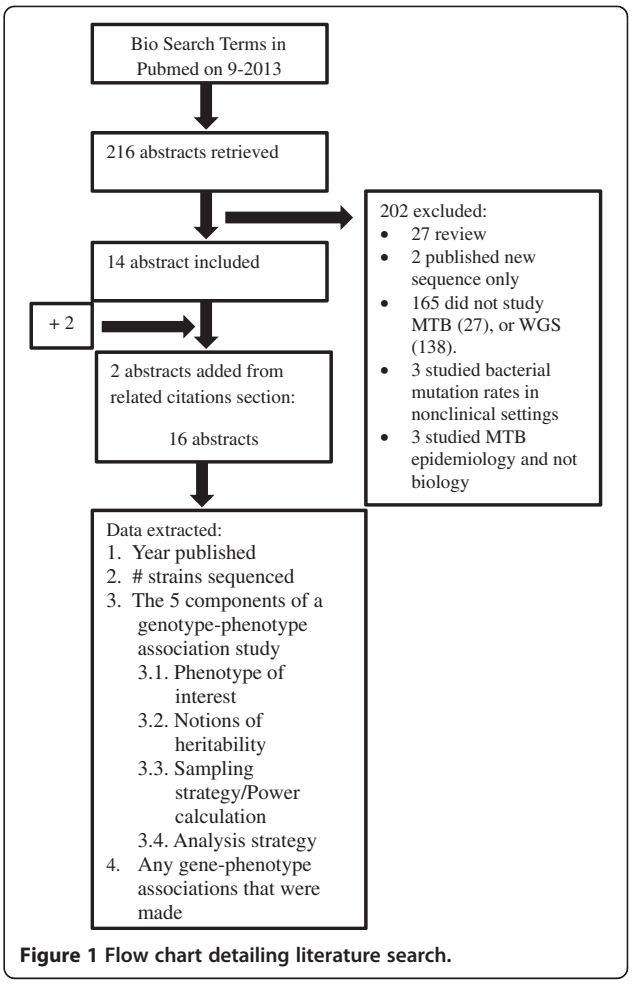

Figure 1 Flow chart detailing literature search.

Phenotype

Most of the studies (13/16) focused on the MTB resistance phenotype to a wide range of drugs. Three other studies examined other strains including: (1) strains causing extrapulmonary tuberculosis; (2) strains with a smooth phenotype; and (3) strains typed as Beijing using spoligotyping.

Effect sizes and *a priori* power calculations were not explicitly discussed in any of these studies.

Sampling

Half of the 16 studies sampled strains in time-course, either in laboratory-evolved strains (five studies), or in serial samples from the same patient (three studies). In all cases, strains were initially drug sensitive but later acquired a drug resistance phenotype. In the other eight studies, clinical MTB samples were obtained from different TB patients, and generally involved the study of more distantly-related strains than in the time-course studies. In general strains were sampled more or less randomly to include strains with and without the phenotype. Seven of the non-time-course studies were published within the last year.

Analysis

In the time-course studies, few mutations occurred and it was generally tractable to identify all novel mutations and infer their role in resistance. In the other studies, only two of eight were able to make specific genomic

Table 1 PubMed Search terms and inclusion and exclusion criteria

Search purpose	Search terms	Inclusion criteria	Exclusion criteria
Identify studies of Pathogen Biology using whole genome sequencing and analysis	'genome sequencing' AND 'tuberculosis' AND ('drug resistance' OR 'virulence' OR 'immunogenicity' OR 'transmissibility' OR 'fitness')	All abstracts describing the use of WGS data to identify genes related to pathogen immunogenicity, virulence, transmissibility, drug resistance, or fitness	(1) Review articles
			(2) Studies that published new sequence data only
	In all PubMed fields		(3) Studies that did not study MTB bacteria and its biology
			(4) Studies that only assess mutation rates in non-clinical settings

Table 2 Literature search results

Author	Reference	Year	Strains sequenced (n)	Stated study purpose[a]	Clinical strains?	Time series?	Method	Report specific genotypic association?
Zhang *et al.*	[26]	2013	161	Identify drug resistance genes	Yes	No	Phylogenetics and comparison of rates with poisson distribution	Yes; list of genes provided
Farhat *et al.*	[15]	2013	124	Identify drug resistance genes	Yes	No	Phylogenetics and convergence analysis	Yes; list of genes provided
Lin *et al.*	[27]	2013	2	Identify drug resistance genes	Yes	No	Comparison with reference mycobacterial strains	No
Wu *et al.*	[28]	2013	4	Identify Beijing associated pathways	Yes	No	COG enrichment of genes with snps	General pathways rather than individual genes
Das *et al.*	[29]	2013	5	Identify genes related to extrapulmonary TB	Yes	No	COG enrichment of genes with snps	General pathways rather than individual genes
Ilina *et al.*	[30]	2013	4	Identify drug resistance genes	Yes	No	Comparison with reference mycobacterial strains	No
Abrahams *et al.*	[31]	2013	-	Identify resistance targets(s) for novel imidazole	No	Yes: spontaneous mutants resistant to drug and their sensitive ancestor	Identification of all mutations	Yes *qcrb*
Supply *et al.*	[32]	2013	5	Identify genes associated with smooth TB phenotype	Yes	No	Comparison with reference mycobacterial strains	General pathways rather than individual genes
Hartkoorn *et al.*	[33]	2012	-	Identify resistance targets(s) for pyridomycin	No	Yes: spontaneous mutants resistant to drug and their sensitive ancestor	Identification of all mutations	Yes acyl-carrier-protein *inha*
G. Sun *et al.*	[34]	2012	7	Identify drug resistance genes	Yes	Yes: serial samples from same patient	Identification of all mutations	No; but list of potential candidates with new fixed mutations provided
Grzegorzewicz *et al.*	[35]	2012	-	Identify resistance targets(s) for novel compound Adamantyl Urea	Yes	Yes: serial samples from the same patient	Identification of all mutations	Yes *mmpl3*
Casali *et al.*	[36]	2012	59	Identify drug resistance genes	Yes	No	Phylogenetic tree and parallel evolution and convergence	Yes *rpoc*
Tahlan *et al.*	[37]	2012	-	Identify resistance targets(s) for novel compound SQ109	No	Yes: spontaneous mutants resistant to drug and their sensitive ancestor	Identification of all mutations	Yes *mmpl3*
La Rosa *et al.*	[38]	2012	-	Identify resistance target(s) for 1,5-diarylpyrrole derivative BM212	No	Yes: spontaneous mutants resistant to drug and their sensitive ancestor	Identification of all mutations	Yes *mmpl3*

Table 2 Literature search results *(Continued)*

Comas *et al.*	[39]	2011	10	Identify drug resistance genes	Yes	Yes: serial samples from the same patient	Identification of all mutations in rpoc. Assessment of convergence across different strain pairs	Yes confirmed rpoc
Manjunatha *et al.*	[40]	2006	-	Identify resistance targets(s) for PA-824	No	Yes: spontaneous mutants resistant to drug and their sensitive ancestor	Identification of all mutations	Yes Rv3547

[a]The term 'phenotype related genes' is used loosely here to describe genes that are associated with but not necessarily causative of the phenotype.

associations supported by formal assessments of statistical significance; both these studies sequenced a relatively large number of genomes (>100), and used phylogenetic ancestral reconstruction in their analysis of mutations relevant to the phenotype [15,26]. Two studies [15,36] used phylogenetic convergence (described below) to select candidates for association with the drug resistance phenotype. In the other six studies, the phenotype-genotype associations were of a more descriptive, less formal nature.

Across all studies, a common theme was the use of tests for positive selection and phylogenetics to differentiate between genetic variation related to strain ancestry and those relevant to the phenotype [15,36]. There are also examples from non-TB pathogens [16,41]. In the phylogenetic convergence test mentioned above, a relatedness tree, constructed using the whole genome data is used to identify genes that accumulate frequent mutations synchronous with the acquisition of the phenotype of interest. Phylogenetic convergence has several advantages well-suited to the study of microorganisms. Most notably, by focusing only on the genetic changes that coincide with the independent appearances of the phenotype, it ignores false-positive associations due to clonal population structure, namely the genetic relatedness of the strains [15,16,36,41,42]. It can therefore be applied to both clonal and sexual/recombining pathogens as long as recombination is taken into account in the phylogenetic tree construction [43]. For highly recombining pathogens, the tools of human GWAS might be appropriate, with some modifications [44,45].

Sampling and analysis strategy

The literature review highlights the success of time-course WGS, either within patients or *in vitro*, to identify the genetic bases of clinically-important phenotypes. However time-course samples are often difficult to obtain, particularly in clinical settings, and may not always be generalizable to the larger population of pathogens [46]. In contrast to time-courses, 'cross-sectional' samples of strains routinely collected for patient diagnosis or public health surveillance are both easier to obtain and may provide a more comprehensive, global picture of a pathogen's adaptive landscape.

A major challenge posed by studying diverse clinical strains is that the sampled population of pathogens may contain population structure related to the shared ancestry of the strains. Populations are considered structured when they include subpopulations among which the frequency of genotypes differs systematically. Population structure, a form of non-independence of observations, can be seen when pathogen strains are isolated from disease outbreaks or direct transmission chains, or clusters, and compared with non-clustered strains; The study of pathogen subpopulations when they also preferentially share the phenotype of interest, can lead investigators to wrongly associate the subpopulation genotype, shared by virtue of ancestry alone, with the phenotype of interest. This type of confounding bias is a well-recognized problem in human GWAS [11,47-49].

Whereas different methods such as Principle Components analysis, mixed effects models and phylogenetic convergence can be used to correct for population structure [11,47-51], adopting a careful sampling strategy can minimize the impact of - or even capitalize on - population structure. Drawing parallels from case-control study design in epidemiology and human GWAS [47] we propose that sampling 'matched' pairs of closely-related strains with different phenotypes can not only control for population structure but can also deliver higher power relative to sampling randomly from strain collections. The matching procedure we propose addresses population structure and improves power by ignoring the shared variants within a subpopulation and focusing only on the recently evolved differences, thus reducing the number of variables tested and improving power. The sequence data generated using matched sampling can be analyzed using a simplified form of phylogenetic convergence by: (1) identifying the recently evolved mutations by pairwise alignment of a sequence from a strain with the phenotype of interest with a closely-related strain lacking the phenotype; (2) counting the number of mutations across several such pairs; and (3) comparing these counts either to a null distribution generated using a non-parametric permutation test [15], or simply to a Poisson or Binomial distribution, as we will discuss and demonstrate in the next sections.

Assuming a binary phenotype of interest that has been clearly defined, we propose to match strains using data from traditional strain typing such as pulsed-field gel electrophoresis and multi-locus sequence typing that is often already available for the banked strains, especially under surveillance for public health purposes. Using this lower resolution typing data, a phylogenetic tree can be constructed, accounting for recombination as needed using methods such as ClonalFrame [16,25]. Figure 2A displays a hypothetical tree topology obtained for a sample of 16 MTB clinical strains constructed using their MIRU-VNTR pattern [52]. Figure 2B demonstrates the matched sampling strategy. For each phenotype positive (ph+) strain, a neighboring phenotype negative (ph-) strain is selected such that the phylogenetic distance between the pair of strains is minimized. Only one ph- and one ph+ strain is sampled per clade. If more than one strain is equidistant, then one is selected at random. The larger phylogenetic tree is thus reduced to a set of matched ph+ and ph- pairs.

Power calculations to optimize genotype-phenotype association studies

To design a genotype-phenotype association study, knowledge about the optimal number of pathogen genomes to sample is necessary. Here we define the sample size n as

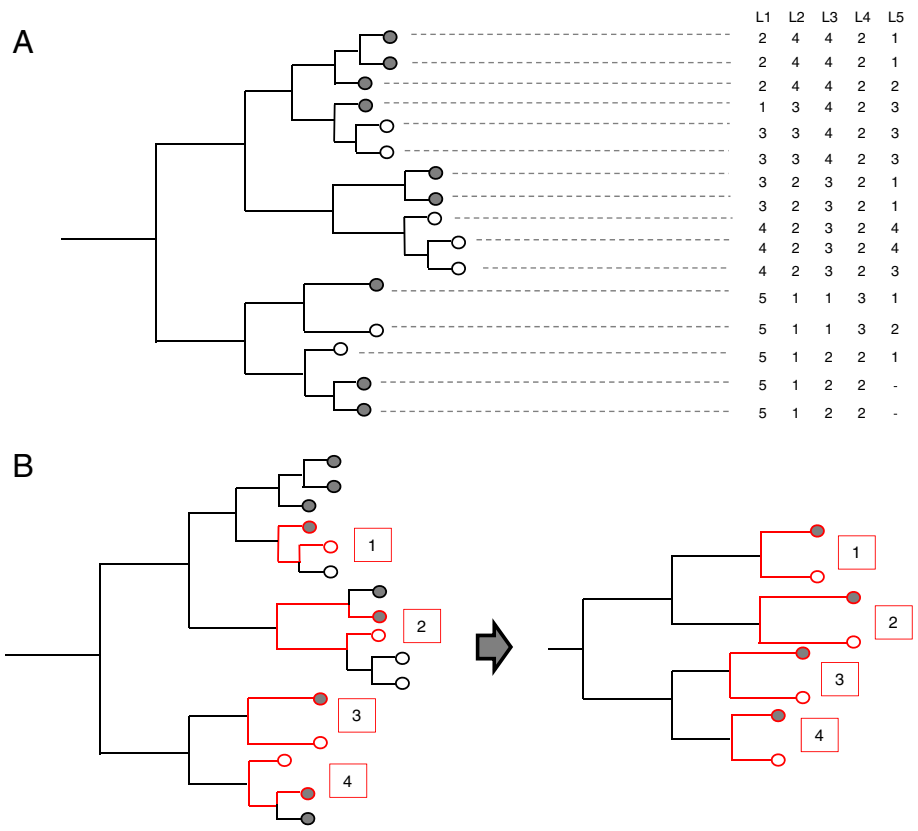

Figure 2 Demonstration of the selection strategy. (A) Example initial MIRU-VNTR phylogeny constructed for selection of strains for sequencing and analysis. Grey circles represent strains with the phenotype of interest (ph+ strains), the white circles represent strains without the phenotype of interest (ph- strains). The Table with columns L1-5 represent the variable number of tandem repeat at each locus L. **(B)** Example of selection methodology: For each ph+ strain (grey circle) a neighboring ph- strain is selected such that the distance between the two strains in the phylogeny is minimized. Each control or study strain is only sampled once. The resultant tree of selected strains will consist of matched study and control strains.

the number of matched genome pairs necessary to achieve a nominal power of >80% for detecting a true association, accepting a false positive association rate of no higher than 0.05. Our goal is to identify genomic variants, for example mutations or recombination events that confer a fitness advantage when the phenotype of interest such as antibiotic resistance, virulence, evolves under selective pressure. These positively selected variants are expected to be more prevalent in strains with the phenotype of interest (ph+). Below, we will describe two methods to identify genomic variants associated with this phenotype of interest. The first, 'site-level' method, uses individual nucleotide sites as the basic level of genetic variation. However, this method can also be applied to other levels of variation, including the presence of absence of genes, or clusters of mutations that are transferred together by recombination and can thus be considered as a unit. This method is therefore applicable to clonal pathogens that evolve almost entirely by point mutation, as well as to moderately recombining pathogens, in which recombinant parts of the genome can be identified computationally [53-55] and considered as a single 'site'. In the second, 'locus-level' method, we model a scenario in which different mutations within the same gene or locus can have a similar phenotypic effect, for example the loss of function by introducing stop codons at different points in the gene, providing additional evidence for the importance of that gene for a particular phenotype.

In the site-level method, for an organism with genome of length k and an average distance (or number of variants) s between each pair of strains, we can define a null hypothesis for the distribution of the number of variants l_j at a particular neutral site (j) in the genome (in the ph+ relative to the ph- strains) across the n pairs. In particular, if the site j is not under selection, then s/k should be a reasonable estimate of the rate of neutral variation, and under the null hypothesis, l_j is a Binomial random variable corresponding to n trials with a success probability $p_{Null} = s/k$. Under the alternative hypothesis that site j is under positive selection, l_j is a binomial random variable with n trials and success probability f_{site} which is greater than s/k. f_{site} is related to the phenotypic effect size of the variant, as a higher frequency of a variant will result from stronger positive selection, that is, higher fitness of the variant in

ph+ relative to ph- strains [56]. An extreme example would be a selective sweep that results in all members of the ph+ population carrying the same variant in which case f_{site} would be 1. In a previous genotype-phenotype association study of drug resistance in MTB [15], the lowest frequency of a single nucleotide ('site level') variant with a known fitness advantage was estimated at 4% ($f = 0.04$) (*rpoB* codon 455 in rifampicin (RIF) resistant strains), whereas the highest was estimated at 52% ($f = 0.52$) (*rpoB* codon 450).

As observed for *rpoB*, more than one nucleotide site in a locus can carry a fitness conferring variant; we can thus formulate a locus-level test by defining a null distribution for the sum of the variant counts in a locus, l_{i_locus}. If locus i of length g_i is not under selection, with the same parameters s and k defined above, then the distribution of l_{i_locus} can be approximated by a Poisson distribution with a rate $= n\ s\ g_i\ /k$. Under the alternative hypothesis, this locus is under selection and the expected number of mutations is $n\ f_{locus}$, which is larger than $n\ s\ g_i\ /k$. Similar to f_{site}, f_{locus} is related to the collective fitness advantage conferred by its variants. For example, in the study cited above, f_{locus} was estimated to be 0.30 to 1.5/locus/ph+ strain for the *thyA* locus for MTB p-aminosalicylic resistance, and *rpoB* locus for RIF resistance, respectively [15]. The test will have a different power for different values of $f_{site/locus}$. Because this analysis involves testing all the sites and loci with observed variation, a correction for multiple testing is needed. We use the Bonferroni correction, assuming that the upper limit for the number of variable sites across the sample is $n\ s$, and the number of variable loci to be $1\text{-}\ e_i^{-n\ g\ s/k}$ (from the Poisson distribution). In Figures 3, 4, and 5, we provide power calculation results as a function of n, s and f using the 4.41 Mbp MTB genome as an example. Here we calculated the expected power by integrating across the distribution of locus lengths g_i for the MTB reference genome H37Rv. Based on previous data from fingerprint-matched MTB, our power calculations explored a range of between-strain genetic distances (s) from 50 to 300 mutations [4].

In the case of MTB, we found that high power (>80%) could be achieved by sequencing 50 to 100 strain pairs (matched at a distance of $s = 100$ variants) to detect a 'rare' drug resistance variant in >5% of the ph+ strains (f_{site} >0.05; Figure 3) or a locus with a low mutation rate of 0.25/locus/ph+ strain (f_{locus} >0.25; Figure 4). The advantage of performing a locus-level analysis is that we expect $f_{locus} > f_{site}$ because f_{locus} is proportional to the sum of f_{site} over all sites under selection in the locus. The number of tests performed in a locus-level analysis is several orders of magnitude lower than with a site-level analysis because a bacterial genome contains on the order of 10^6 sites, but only 10^3 genes (loci). We

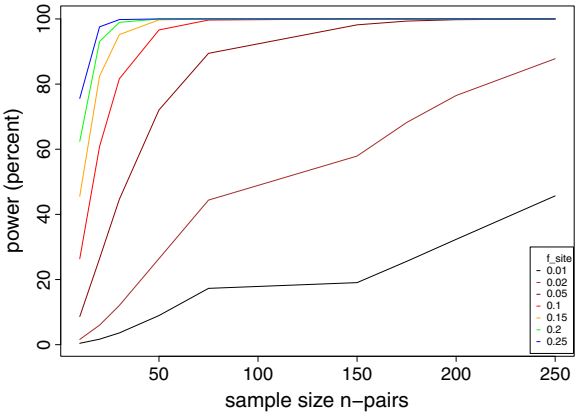

Figure 3 Power of the matched convergence test for identify nucleotide sites associated with a phenotype of interest. The average genetic distance between matched strains was set to an intermediate level of s = 100 mutations. Colors represent increasing values of site effect size f_{site}.

performed similar calculations for *Campylobacter* ($k =$ 1.64 Mbp), assuming a higher matching distance $s = 300$ that is expected for multi-locus sequence typing (MLST) of this pathogen [16]. With 50 to 100 strain pairs of Campylobacter the lowest f_{locus} that can be detected with >80% power is 0.60 (Additional file 1: Figure S1), higher than for MTB (Figure 4).

We next explored how power depends on the genetic distance between sampled genomes. Figure 5 demonstrates that considerable power gains can be achieved by sampling strain pairs that are close genetic relatives (low s). This is because, for a given value of f_{site} or f_{locus}, raising s decreases the ratio of selected to neutral variants, thereby decreasing the signal to noise ratio.

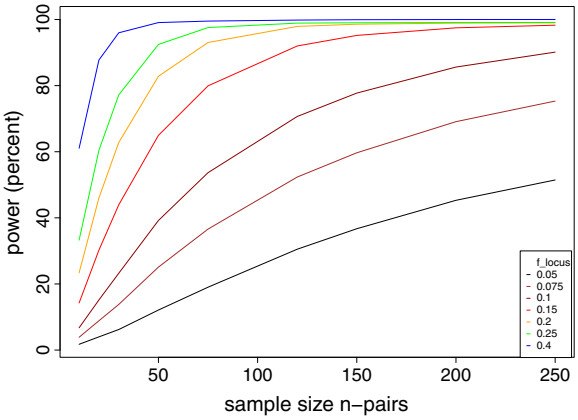

Figure 4 Power of the matched convergence test to identify loci associated with a phenotype of interest. The average distance between matched strains was set at s = 100 mutations. Colors represent increasing values of locus effect size f_{locus}.

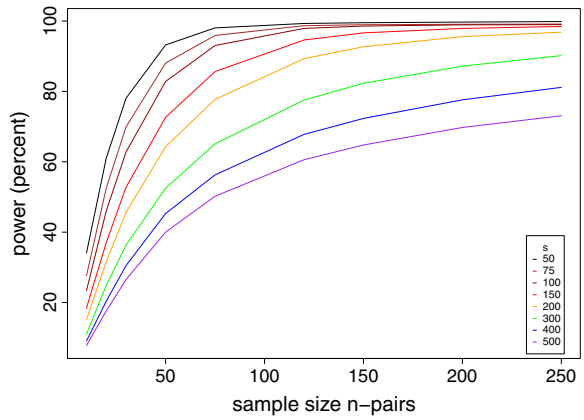

Figure 5 Power of the matched convergence test at the locus level as a function of genetic distance (*s*) between matched strains pairs. Smaller *s* indicates closer genetic relatedness between strain pairs.

The power calculator is provided with this manuscript as an R function (Additional file 2), and allows the user to tune all the parameters described to provide power estimates for different effect sizes, different pathogen genome sizes, and different levels of genetic relatedness.

Application to genomic data from MTB and *Campylobacter* species

We applied the sampling strategy described in Figure 2 to a set of 123 clinically isolated unmatched *MTB* genomes previously analyzed using phylogenetic convergence [15] (Additional files 3 and 4). Repetitive, transposon, and phage-related regions were removed as putatively recombinant or as error-prone regions of the alignment. Of the 123 strains, 47 were resistant to one or more drugs (ph+) and the rest were sensitive (ph-). As different fingerprinting methods were used for the different strains in this study and for demonstration purposes we used the phylogeny constructed using whole genome single nucleotide polymorphisms to match strains. We chose eight pairs of strains using this selection strategy (Figure 6). We then counted the recent mutational changes (single nucleotide polymorphisms; SNPs) between each pair of strains. The average distance (*s*) between pairs was 109 SNPs and was in the range of 12 to 254 SNPs. We calculated the number of changes per gene across the eight pairs and compared this number to a Poisson distribution of mutations randomly distributed across branches as the null distribution. We then identified the tail of the distribution, containing genes with a high number of changes highly associated with drug resistance (Figure 7). Overall, 12 genes and non-coding regions were found to be associated with drug resistance using only 16 out of 123 strains (13%) used in the original analysis. The analysis identified *katG*, *embB*, *rpoB* (well known drug resistance determinants) as well as top new candidates from the previous full analysis of all 123 genomes: *ponA1*, *ppsA*, *murD*, and *rbsk*. This selection strategy and analysis recovered 67% of the candidates identified with the full analysis, but used only 13% of the data, demonstrating the superior power of the matched convergence analysis to the general unmatched test.

Second, we applied the same method to a set of 192 *Campylobacter coli* and *jejuni* isolates used by Sheppard *et al.* in an association study to identify the factors responsible for adaptation to cattle and chickens [16] (Additional files 5 and 6). Sheppard *et al.* associated the presence or absence of unique 30 bp 'words' with the host specificity phenotype and controlled for population structure by comparing the real word counts with word counts generated along the tree through Monte Carlo simulations. We applied our method to a subset of 29 strains enriched in the phenotype of host switching that Sheppard *et al.* had used in their initial analysis. After correcting for recombination and constructing the phylogeny using ClonalFrame, we phylogenetically matched 8 pairs of strains that had undergone host switching (Figure 8). Five switches were estimated from cattle to bird or human, and three were from bird to human hosts. We counted the pairwise differences across the eight pairs, grouping insertions/deletions and mutations by gene and compared the distribution to the expected Poisson distribution (Figure 9). We associated two consecutive genes: *surE* and *Cj0294*, both of which were present in cattle-associated strains but absent in chicken-associated strains. These genes mapped to a vitamin B5 biosynthesis region, which Sheppard *et al.* had previously found to affect *Campylobacter* growth in the presence or absence of vitamin B5 [16]. In addition, our approach associated 105 additional genes (Additional file 7: Table S1). Thus, using the convergence method and focusing on genes rather than 30 bp words, we were able to detect the experimentally-validated vitamin B5 region of the *Campylobacter* genome, among other potential genes involved in host switching that had been observed by Sheppard *et al.* using a much smaller dataset.

Our power calculations rely on a well-defined phenotype that can be measured without error. The phenotype is also assumed to be binary, or at least divisible into two binary states; therefore, the calculations cannot be easily extended to quantitative traits. Knowledge about the expected effect size for different phenotypes is also important for these calculations and prospective study design. Among the studies reviewed, we found the effect size to be infrequently reported for MTB. Here we provide empirical effect sizes estimated from a previous MTB drug resistance study [15] as a reference point for future studies.

Our approach also assumes that a certain amount of previously collected antigen or genotyping data is available to allow for building a phylogeny and selecting pairs

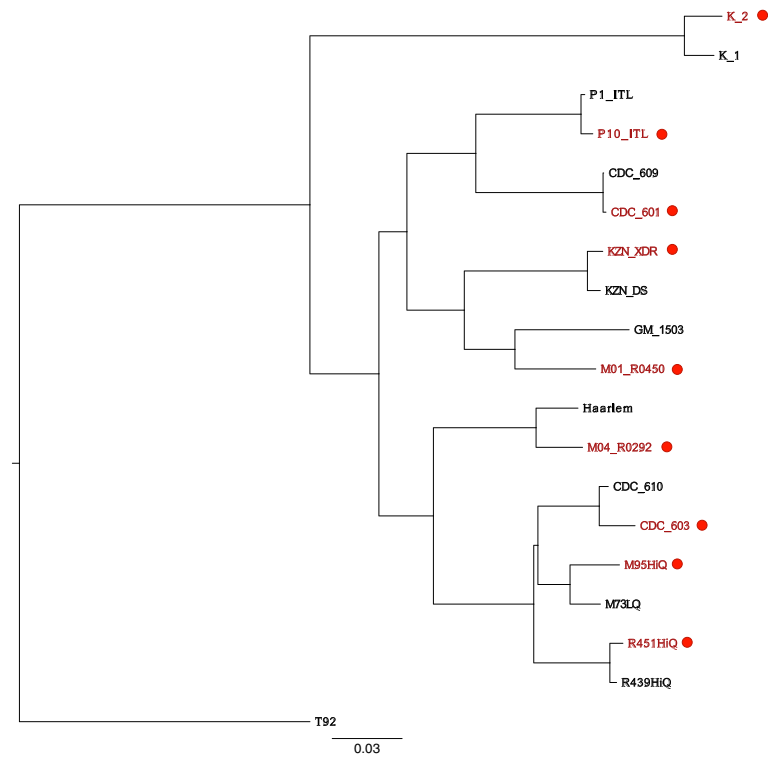

Figure 6 Phylogeny of MTB strains chosen for genotype-phenotype analysis. Dots indicate the presence of the drug resistant phenotype. The tree demonstrates the matching of strains with and without the drug resistance phenotype.

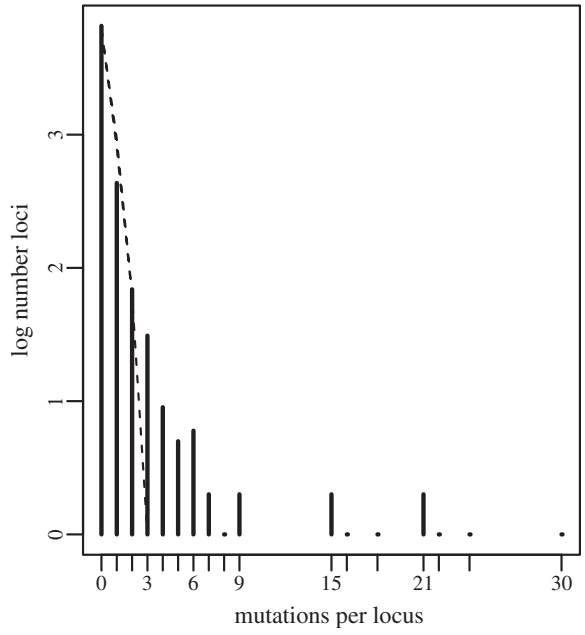

Figure 7 Distribution of SNPs/locus across the eight pairs of MTB genomes. Observed counts are represented by black bars. The dashed line represents the upper 95% confidence bounds on a Poisson distribution with the observed number of mutations.

of strains to sequence. When sequence data are already available, this method can still be used to select strains for paired analysis, providing a simple control for population structure and a more simplified analysis strategy. If no typing data are available, alternatives may still exist - for example, using epidemiological data that link strains within a particular outbreak. In each of these scenarios, perfect matching to form pairs of monophyletic strains may not always be possible, but given the relationship of the matching distance to power demonstrated above, we argue for matching as many strains as possible and as closely as possible. The analysis of the total dataset of all monophyletic and paraphyletic pairs can be performed via ancestral reconstruction and a more general phylogenetic convergence method ('phyC' [15]) rather than the simplified pairwise analysis described here.

Our power calculations, like all models, make necessary simplifications and assumptions. For example, we assume that neutral variants are distributed randomly across the whole genome. This may not necessarily be the case as some pathogen genes may contain mutation or recombination hot spots. Some adjustment for such a scenario could be made by using a higher average rate of variation than the one expected, that is, testing power under a pairwise distance $s^{\dagger}$ amplified by a factor $m > 1$

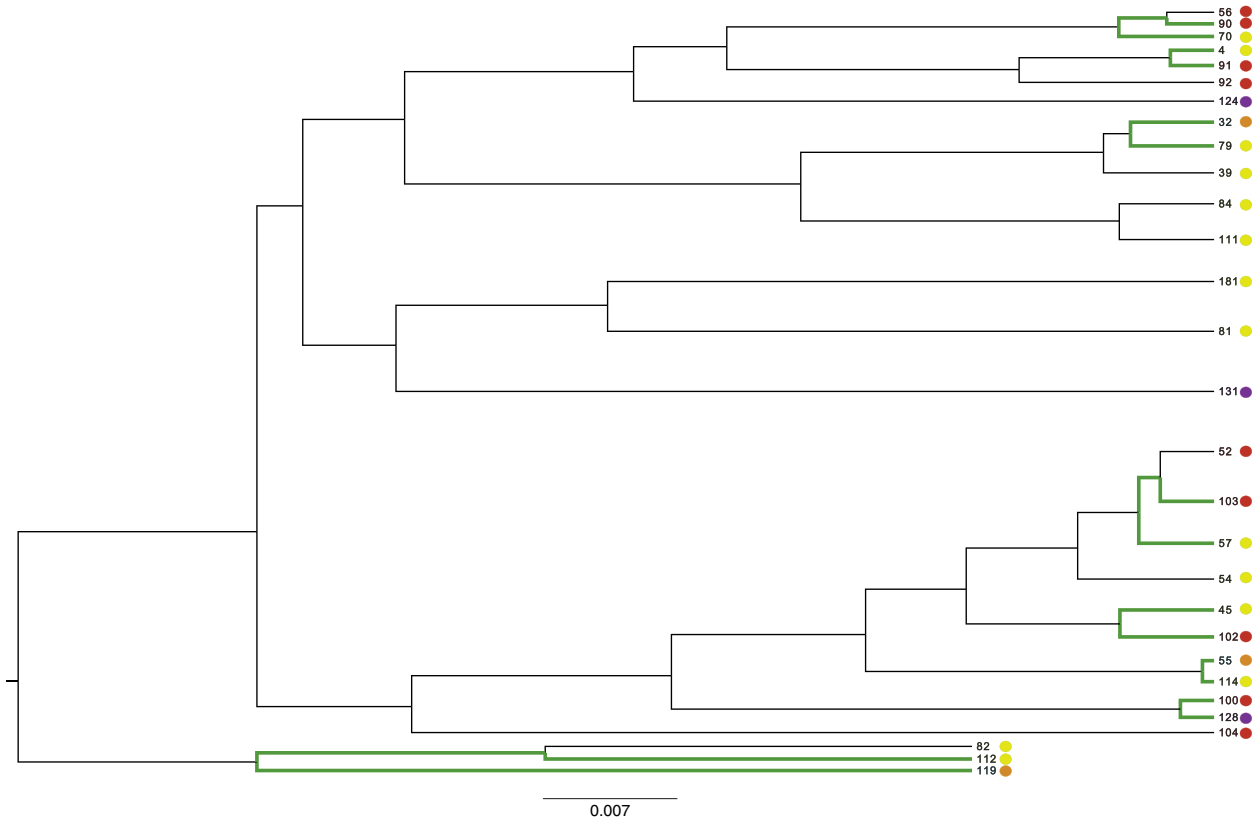

Figure 8 Phylogeny of Campylobacter strains. Branches highlighted in green lead up to the strain pairs chosen for genotype-phenotype association. Colored circles denote host specificity: red = cattle, green = chicken, purple = wild bird/non-host, orange = human.

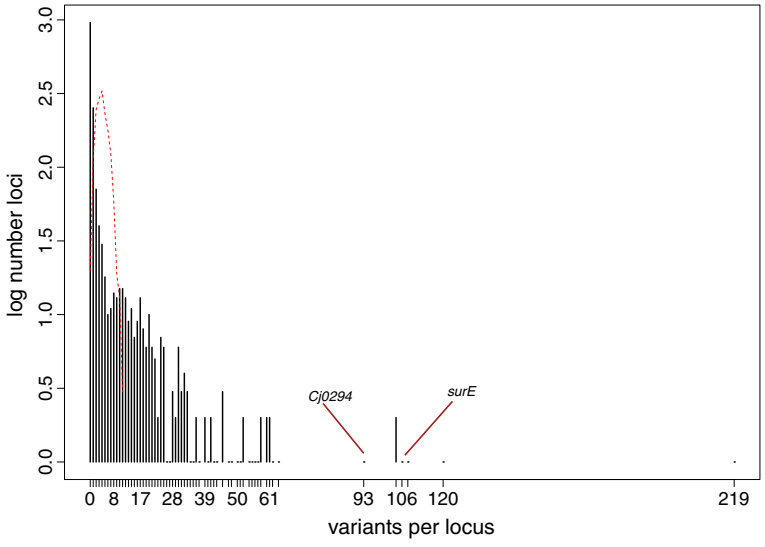

Figure 9 Distribution of variants/locus across the eight pairs of *Campylobacter* genomes. Observed counts are represented by black bars. The dashed red line represents the upper 95% confidence bounds on a Poisson distribution with the observed number of variants. Variant counts per locus for *surE* and *Cj0294* are highlighted.

where $s^{\dagger} = m\ s_{expected}$ for a range of m. The framework and power calculations presented here represent a step toward more systematic and prospective genotype-phenotype study design for microbial pathogens, and can provide the basis for more refined power calculations (for example, accounting for continuous rather than binary phenotypes, or for analysis of un-matched strains).

Conclusions

The improved ability to study the evolution of clinical strains will be an important advance for the study of pathogens as they spread. Thus far, most of our understanding of infectious disease has focused on the epidemiological study of host risk factors, or on the *in vitro* study of the pathogen. The rich information contained in whole genomes of clinical pathogens - isolated as they adapt to their host and cause disease - provides a new and complementary perspective on pathogen biology. Here we have shown how clonal to moderately sexual strain collections, originally assembled for epidemiological purposes, using appropriate sub-sampling schemes, can empower genome-level association studies and reveal genotype-phenotype associations, increasing our understanding of pathogen biology and adaptation.

Additional files

Additional file 1: Figure S1. Power of matched convergence test to identify phenotype associated loci. The average distance between matched strains was set at s = 300 variants. Colors represent increasing values of locus effect size f_{locus}.

Additional file 2: Power calculator as R function.

Additional file 3: *Mycobacterium tuberculosis* strains multiple sequence alignment file.

Additional file 4: *Mycobacterium tuberculosis* strains drug resistance profile.

Additional file 5: *Campylobacter* strain multiple sequence alignment file.

Additional file 6: *Campylobacter* strain host specificity phenotype.

Additional file 7: Table S1. *Campylobacter* genes associated with host switching using the matched selection strategy and the proposed analysis.

Abbreviations

GTR: Generalized Time Reversible substitution model; GWAS: Genome Wide Association Study; MIRU-VNTR: Mycobacterial interspersed repetitive units-variable number tandem repeats; MLST: Multi-locus sequence typing; MTB: Mycobacterium tuberculosis; SNPs: Single nucleotide changes; TB: Tuberculosis; WGS: Whole-genome sequencing or sequences.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MF and MM were responsible for the conception and design of this study. MF conducted the analysis and drafted the original manuscript. BJS and CC contributed to the design and made key manuscript edits. SS contributed to the *Campylobacter* analysis and provided key manuscript edits. All authors read and approved the final manuscript.

Acknowledgments

This work was funded by the Parker B. Francis Foundation (MF) and the NIH U19-AI109755 (MF and MM), NIH U19 A1-076217 (MM), the Canadian Institutes for Health Research, the Natural Sciences and Engineering Research Council of Canada, and the Canada Research Chairs program (BJS), the Engineering and Physical Sciences Research Council EPSRC EP/K026003/1 (CC). SS is funded by the UK Biotechnology and Biological Sciences Research Council (BBSRC), the Wellcome Trust, and the UK Medical Research Council (MRC) - under the Cloud Infrastructure for Microbial Bioinformatics (CLIMB) project. We thank Dr. Alkes Price, Harvard School of Public Health Department of Biostatistics, for his helpful feedback on the manuscript and methods.

Author details

[1]Department of Pulmonary and Critical Care, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [2]Département de sciences biologiques, Université de Montréal, Montréal, QC, Canada. [3]Institute of Life Science, College of Medicine, Swansea University, Swansea SA2 8PP, UK. [4]Department of Mathematics, Imperial College London, London, UK. [5]Department of Global Health and Social Medicine, Harvard Medical School, 641 Huntington Avenue Suite 4A, Boston, MA 02115, USA. [6]Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA.

Published online: 15 November 2014

References

1. Kilbourne ED: **The molecular epidemiology of influenza.** *J Infect Dis* 1973, **127:**478–487.
2. Alland D, Kalkut GE, Moss AR, McAdam RA, Hahn JA, Bosworth W, Drucker E, Bloom BR: **Transmission of tuberculosis in New York City – an analysis by DNA fingerprinting and conventional epidemiologic methods.** *N Engl J Med* 1994, **330:**1710–1716.
3. Streicher EM, Müller B, Chihota V, Mlambo C, Tait M, Pillay M, Trollip A, Hoek KGP, Sirgel FA, van Pittius NCG, van Helden PD, Victor TC, Warren RM: **Emergence and treatment of multidrug resistant (MDR) and extensively drug-resistant (XDR) tuberculosis in South Africa.** *Infect Genet Evol* 2012, **12:**686–694.
4. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW, Parkhill J, Harris D, Walker AS, Bowden R, Monk P, Smith EG, Peto TEA: **Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study.** *Lancet Infect Dis* 2013, **13:**137–146.
5. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJM, Brinkman FSL, Brunham RC, Tang P: **Whole-genome sequencing and social-network analysis of a tuberculosis outbreak.** *N Engl J Med* 2011, **364:**730–739.
6. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS, Pichon B, Baker S, Parry CM, Lambertsen LM, Shahinas D, Pillai DR, Mitchell TJ, Dougan G, Tomasz A, Klugman KP, Parkhill J, Hanage WP, Bentley SD: **Rapid pneumococcal evolution in response to clinical interventions.** *Science* 2011, **331:**430–434.
7. Kumar V, Sun P, Vamathevan J, Li Y, Ingraham K, Palmer L, Huang J, Brown JR: **Comparative genomics of Klebsiella pneumoniae strains with different antibiotic resistance profiles.** *Antimicrob Agents Chemother* 2011, **55:**4267–4276.
8. Kato-Maeda M, Shanley CA, Ackart D, Jarlsberg LG, Shang S, Obregon-Henao A, Harton M, Basaraba RJ, Henao-Tamayo M, Barrozo JC, Rose J, Kawamura LM, Coscolla M, Fofanov VY, Koshinsky H, Gagneux S, Hopewell PC, Ordway DJ, Orme IM: **Beijing sublineages of Mycobacterium tuberculosis differ in pathogenicity in the guinea pig.** *Clin Vaccine Immunol CVI* 2012, **19:**1227–1237.
9. Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, Johnston JC, Gardy J, Lipsitch M, Fortune SM: **Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis.** *Nat Genet* 2013, **45:**784–790.

10. Wang Z, Liu X, Yang B-Z, Gelernter J: **The role and challenges of exome sequencing in studies of human diseases.** *Stat Genet Methodol* 2013, **4:**160.
11. Price AL, Zaitlen NA, Reich D, Patterson N: **New approaches to population stratification in genome-wide association studies.** *Nat Rev Genet* 2010, **11:**459–463.
12. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies.** *Nat Genet* 2006, **38:**904–909.
13. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES: **A unified mixed-model method for association mapping that accounts for multiple levels of relatedness.** *Nat Genet* 2006, **38:**203–208.
14. Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D: **Improved linear mixed models for genome-wide association studies.** *Nat Methods* 2012, **9:**525–526.
15. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, Warren RM, Streicher EM, Calver A, Sloutsky A, Kaur D, Posey JE, Plikaytis B, Oggioni MR, Gardy JL, Johnston JC, Rodrigues M, Tang PKC, Kato-Maeda M, Borowsky ML, Muddukrishna B, Kreiswirth BN, Kurepina N, Galagan J, Gagneux S, Birren B, Rubin EJ, Lander ES, Sabeti PC, Murray M: **Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis.** *Nat Genet* 2013, **45:**1183–1189.
16. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, Bentley SD, Maiden MCJ, Parkhill J, Falush D: **Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter.** *Proc Natl Acad Sci U S A* 2013, **110:**11923–11927.
17. Namouchi A, Didelot X, Schöck U, Gicquel B, Rocha EPC: **After the bottleneck: Genome-wide diversification of the Mycobacterium tuberculosis complex by mutation, recombination, and natural selection.** *Genome Res* 2012, **22:**721–734.
18. Lew JM, Kapopoulou A, Jones LM, Cole ST: **TubercuList–10 years after.** *Tuberc Edinb Scotl* 2011, **91:**1–7.
19. Comas I, Chakravartti J, Small PM, Galagan J, Niemann S, Kremer K, Ernst JD, Gagneux S: **Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved.** *Nat Genet* 2010, **42:**498–503.
20. Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM, Murray MB: **Tuberculosis drug resistance mutation database.** *PLoS Med* 2009, **6:**e2.
21. Felsenstein J: **PHYLIP - Phylogeny Inference Package (Version 3.2).** *Cladistics* 1989, **5:**164–166.
22. Ioerger TR, Koo S, No E-G, Chen X, Larsen MH, Jacobs WR, Pillay M, Sturm AW, Sacchettini JC: **Genome Analysis of Multi- and Extensively-Drug-Resistant Tuberculosis from KwaZulu-Natal.** *South Africa. PLoS ONE* 2009, **4:**e7778.
23. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP: **MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space.** *Syst Biol* 2012, **61:**539–542.
24. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O: **New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.** *Syst Biol* 2010, **59:**307–321.
25. Didelot X, Falush D: **Inference of bacterial microevolution using multilocus sequence data.** *Genetics* 2007, **175:**1251–1266.
26. Zhang H, Li D, Zhao L, Fleming J, Lin N, Wang T, Liu Z, Li C, Galwey N, Deng J, Zhou Y, Zhu Y, Gao Y, Wang T, Wang S, Huang Y, Wang M, Zhong Q, Zhou L, Chen T, Zhou J, Yang R, Zhu G, Hang H, Zhang J, Li F, Wan K, Wang J, Zhang X-E, Bi L: **Genome sequencing of 161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance.** *Nat Genet* 2013, **45:**1255–1260.
27. Lin N, Liu Z, Zhou J, Wang S, Fleming J: **Draft genome sequences of two super-XDR isolates of M. tuberculosis from China.** *FEMS Microbiol Lett* 2013, **347:**93–96.
28. Wu W, Zheng H, Zhang L, Wen Z, Zhang S, Pei H, Yu G, Zhu Y, Cui Z, Hu Z, Wang H, Li Y: **A genome-wide analysis of multidrug-resistant and extensively drug-resistant strains of Mycobacterium tuberculosis Beijing genotype.** *Mol Genet Genomics MGG* 2013, **288:**425–436.
29. Das S, Roychowdhury T, Kumar P, Kumar A, Kalra P, Singh J, Singh S, Prasad HK, Bhattacharya A: **Genetic heterogeneity revealed by sequence analysis of Mycobacterium tuberculosis isolates from extra-pulmonary tuberculosis patients.** *BMC Genomics* 2013, **14:**404.
30. Ilina EN, Shitikov EA, Ikryannikova LN, Alekseev DG, Kamashev DE, Malakhova MV, Parfenova TV, Afanas'ev MV, Ischenko DS, Bazaleev NA, Smirnova TG, Larionova EE, Chernousova LN, Beletsky AV, Mardanov AV, Ravin NV, Skryabin KG, Govorun VM: **Comparative genomic analysis of Mycobacterium tuberculosis drug resistant strains from Russia.** *PLoS One* 2013, **8:**e56577.
31. Abrahams KA, Cox JAG, Spivey VL, Loman NJ, Pallen MJ, Constantinidou C, Fernández R, Alemparte C, Remuiñán MJ, Barros D, Ballell L, Besra GS: **Identification of novel imidazo[1,2-a]pyridine inhibitors targeting M. tuberculosis QcrB.** *PLoS One* 2012, **7:**e52951.
32. Supply P, Marceau M, Mangenot S, Roche D, Rouanet C, Khanna V, Majlessi L, Criscuolo A, Tap J, Pawlik A, Fiette L, Orgeur M, Fabre M, Parmentier C, Frigui W, Simeone R, Boritsch EC, Debrie A-S, Willery E, Walker D, Quail MA, Ma L, Bouchier C, Salvignol G, Sayes F, Cascioferro A, Seemann T, Barbe V, Locht C, Gutierrez M-C, *et al*: **Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of Mycobacterium tuberculosis.** *Nat Genet* 2013, **45:**172–179.
33. Hartkoorn RC, Sala C, Neres J, Pojer F, Magnet S, Mukherjee R, Uplekar S, Boy-Röttger S, Altmann K-H, Cole ST: **Towards a new tuberculosis drug: pyridomycin - nature's isoniazid.** *EMBO Mol Med* 2012, **4:**1032–1042.
34. Sun G, Luo T, Yang C, Dong X, Li J, Zhu Y, Zheng H, Tian W, Wang S, Barry CE 3rd, Mei J, Gao Q: **Dynamic population changes in Mycobacterium tuberculosis during acquisition and fixation of drug resistance in patients.** *J Infect Dis* 2012, **206:**1724–1733.
35. Grzegorzewicz AE, Pham H, Gundi VAKB, Scherman MS, North EJ, Hess T, Jones V, Gruppo V, Born SEM, Korduláková J, Chavadi SS, Morisseau C, Lenaerts AJ, Lee RE, McNeil MR, Jackson M: **Inhibition of mycolic acid transport across the Mycobacterium tuberculosis plasma membrane.** *Nat Chem Biol* 2012, **8:**334–341.
36. Casali N, Nikolayevskyy V, Balabanova Y, Ignatyeva O, Kontsevaya I, Harris SR, Bentley SD, Parkhill J, Nejentsev S, Hoffner SE, Horstmann RD, Brown T, Drobniewski F: **Microevolution of extensively drug-resistant tuberculosis in Russia.** *Genome Res* 2012, **22:**735–745.
37. Tahlan K, Wilson R, Kastrinsky DB, Arora K, Nair V, Fischer E, Barnes SW, Walker JR, Alland D, Barry CE 3rd, Boshoff HI: **SQ109 targets MmpL3, a membrane transporter of trehalose monomycolate involved in mycolic acid donation to the cell wall core of Mycobacterium tuberculosis.** *Antimicrob Agents Chemother* 2012, **56:**1797–1809.
38. La Rosa V, Poce G, Canseco JO, Buroni S, Pasca MR, Biava M, Raju RM, Porretta GC, Alfonso S, Battilocchio C, Javid B, Sorrentino F, Ioerger TR, Sacchettini JC, Manetti F, Botta M, De Logu A, Rubin EJ, De Rossi E: **MmpL3 is the cellular target of the antitubercular pyrrole derivative BM212.** *Antimicrob Agents Chemother* 2012, **56:**324–331.
39. Comas I, Borrell S, Roetzer A, Rose G, Malla B, Kato-Maeda M, Galagan J, Niemann S, Gagneux S: **Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes.** *Nat Genet* 2012, **44:**106–110.
40. Manjunatha UH, Boshoff H, Dowd CS, Zhang L, Albert TJ, Norton JE, Daniels L, Dick T, Pang SS, Barry CE 3rd: **Identification of a nitroimidazo-oxazine-specific protein involved in PA-824 resistance in Mycobacterium tuberculosis.** *Proc Natl Acad Sci U S A* 2006, **103:**431–436.
41. Sokurenko EV, Feldgarden M, Trintchina E, Weissman SJ, Avagyan S, Chattopadhyay S, Johnson JR, Dykhuizen DE: **Selection footprint in the FimH adhesin shows pathoadaptive niche differentiation in Escherichia coli.** *Mol Biol Evol* 2004, **21:**1373–1383.
42. Chattopadhyay S, Paul S, Dykhuizen DE, Sokurenko EV: **Tracking recent adaptive evolution in microbial species using TimeZone.** *Nat Protoc* 2013, **8:**652–665.
43. Shapiro BJ, David LA, Friedman J, Alm EJ: **Looking for Darwin's footprints in the microbial world.** *Trends Microbiol* 2009, **17:**196–204.
44. Alam MT, Petit RA, Crispelll EK, Thornton TA, Conneely KN, Jiang Y, Satola SW, Read TD: **Dissecting vancomycin intermediate resistance in Staphylococcus aureus using genome-wide association.** *Genome Biol Evol* 2014, **6:**1175–1185.
45. Park DJ, Lukens AK, Neafsey DE, Schaffner SF, Chang H-H, Valim C, Ribacke U, Van Tyne D, Galinsky K, Galligan M, Becker JS, Ndiaye D, Mboup S, Wiegand RC, Hartl DL, Sabeti PC, Wirth DF, Volkman SK: **Sequence-based association and selection scans identify drug resistance loci in the Plasmodium falciparum malaria parasite.** *Proc Natl Acad Sci U S A* 2012, **109:**13052–13057.
46. Wichman HA, Badgett MR, Scott LA, Boulianne CM, Bull JJ: **Different trajectories of parallel evolution during viral adaptation.** *Science* 1999, **285:**422–424.

47. Luca D, Ringquist S, Klei L, Lee AB, Gieger C, Wichmann H-E, Schreiber S, Krawczak M, Lu Y, Styche A, Devlin B, Roeder K, Trucco M: **On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants.** *Am J Hum Genet* 2008, **82:**453–463.
48. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-Y, Freimer NB, Sabatti C, Eskin E: **Variance component model to account for sample structure in genome-wide association studies.** *Nat Genet* 2010, **42:**348–354.
49. Zhou X, Stephens M: **Genome-wide efficient mixed-model analysis for association studies.** *Nat Genet* 2012, **44:**821–824.
50. Jombart T, Devillard S, Balloux F: **Discriminant analysis of principal components: a new method for the analysis of genetically structured populations.** *BMC Genet* 2010, **11:**94.
51. Limpiti T, Intarapanich A, Assawamakin A, Shaw PJ, Wangkumhang P, Piriyapongsa J, Ngamphiw C, Tongsima S: **Study of large and highly stratified population datasets by combining iterative pruning principal component analysis and structure.** *BMC Bioinformatics* 2011, **12:**255.
52. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rüsch-Gerdes S, Willery E, Savine E, De Haas P, Van Deutekom H, Roring S, Bifani P, Kurepina N, Kreiswirth B, Sola C, Rastogi N, Vatin V, Gutierrez MC, Fauville M, Niemann S, Skuce R, Kremer K, Locht C, Van Soolingen D: **Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of Mycobacterium tuberculosis.** *J Clin Microbiol* 2006, **44:**4498–4510.
53. Didelot X, Lawson D, Darling A, Falush D: **Inference of homologous recombination in bacteria using whole-genome sequences.** *Genetics* 2010, **186:**1435–1449.
54. Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, Corander J: **Detection of recombination events in bacterial genomes from large population samples.** *Nucleic Acids Res* 2012, **40:**e6.
55. Yahara K, Didelot X, Ansari MA, Sheppard SK, Falush D: **Efficient inference of recombination hot regions in bacterial genomes.** *Mol Biol Evol* 2014, **31:**1593–1605.
56. Eyre-Walker A, Keightley PD: **The distribution of fitness effects of new mutations.** *Nat Rev Genet* 2007, **8:**610–618.

doi:10.1186/s13073-014-0101-7
Cite this article as: Farhat *et al.*: **A phylogeny-based sampling strategy and power calculator informs genome-wide associations study design for microbial pathogens.** *Genome Medicine* 2014 **6**:101.