OPEN ACCESS Freely available online

PLoS one

Towards Eliminating Bias in Cluster Analysis of TB Genotyped Data

Cari van Schalkwyk[1*☯], Madeleine Cule[2,3☯], Alex Welte[1], Paul van Helden[4], Gian van der Spuy[4], Pieter Uys[1]

1 The South African Department of Science and Technology/National Research Foundation (DST/NRF) Centre of Excellence in Epidemiological Modelling and Analysis, Faculty of Science, University of Stellenbosch, Stellenbosch, South Africa, **2** African Institute for Mathematical Sciences, Muizenberg, South Africa, **3** Department of Statistics, University of Oxford, Oxford, United Kingdom, **4** DST/NRF Centre of Excellence in Biomedical Tuberculosis Research/MRC Centre of Molecular and Cellular Biology, Division of Molecular Biology and Human Genetics, Faculty of Health Sciences, University of Stellenbosch, Stellenbosch, South Africa

Abstract

The relative contributions of transmission and reactivation of latent infection to TB cases observed clinically has been reported in many situations, but always with some uncertainty. Genotyped data from TB organisms obtained from patients have been used as the basis for heuristic distinctions between circulating (clustered strains) and reactivated infections (unclustered strains). Naïve methods previously applied to the analysis of such data are known to provide biased estimates of the proportion of unclustered cases. The hypergeometric distribution, which generates probabilities of observing clusters of a given size as realized clusters of all possible sizes, is analyzed in this paper to yield a formal estimator for genotype cluster sizes. Subtle aspects of numerical stability, bias, and variance are explored. This formal estimator is seen to be stable with respect to the epidemiologically interesting properties of the cluster size distribution (the number of clusters and the number of singletons) though it does not yield satisfactory estimates of the number of clusters of larger sizes. The problem that even complete coverage of genotyping, in a practical sampling frame, will only provide a partial view of the actual transmission network remains to be explored.

Citation: van Schalkwyk C, Cule M, Welte A, van Helden P, van der Spuy G, et al. (2012) Towards Eliminating Bias in Cluster Analysis of TB Genotyped Data. PLoS ONE 7(3): e34109. doi:10.1371/journal.pone.0034109

Editor: Madhukar Pai, McGill University, Canada

Received January 17, 2012; **Accepted** February 23, 2012; **Published** March 29, 2012

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: carivs@sun.ac.za

☯ These authors contributed equally to this work.

Introduction

In order to better understand the epidemiology of tuberculosis (TB), recent infection needs to be distinguished from the reactivation of latent disease, for instance to assess the success of intervention programs. To this end, molecular techniques of DNA 'fingerprinting' such as restriction fragment length polymorphism (RFLP) are commonly used. Typically, bacteria from a sample of infected individuals are typed, and classified as either 'clustered' or 'unique'. Unique cases each form what is termed a 'singleton cluster'. Two cases yielding the same type, and hence in the same cluster, are usually considered likely to be directly 'linked' in the following sense: either one case is the 'descendant' of the other, or they share a common 'ancestor' [1,2]. Hence, the proportion of clustered individuals is used as an indicator of the proportion of on-going or recent transmission.

There are two common rules of thumb for estimating the proportion of cases due to recent transmission: the 'n method' and the 'n-1 method'. The former uses the proportion of cases in clusters as a proxy for the proportion of cases due to recent transmission. In the latter, one case from each cluster is assumed to be an index case, and the proportion of non-index cases is used as a measure of recent transmission (thus the 'n-1 method' always leads to a lower estimate of the proportion of recent transmission).

It is unlikely that one will be able to identify every active case in a community. Moreover, among sputum confirmed TB subjects encountered (typically self-reporting to clinics), not all sputum samples will be successfully typed. However, the proportion successfully typed (sampling rate) is, of course, known.

It has previously been shown [3,4] that naïve estimates of clustering exhibit a systematic bias, leading to underestimation of the proportion of clustered individuals. There are three components to this problem of bias: 1) the imperfect view of the epidemic in a community provided by considering only the reported TB cases for a given finite period of time e.g. bias caused by under-diagnosis, partial contact tracing or the restriction of the time window. For these and other various logistical reasons, the reported cases do not represent a random sample from all TB cases in the community. 2) The sample of genotyped cases is not necessarily a random sample of the reported cases due to the diagnostic probability of culture-positivity being dependent on age and HIV status. To reduce this bias, children should be excluded from the study population. In settings where HIV prevalence is high, this bias will not be negligible. 3) Bias in the number of unique cases exists due to contributions resulting from sampling the larger clusters, e.g. 10 clusters of size 4 when sampled at a rate of 0.6 may present as 4 singleton clusters (uniques) together with 3 doublet clusters, 1 triplet cluster and 2 clusters of size 4. For the same reason, bias arises in the total number of clusters. These

contributions to total bias are not insignificant. This third source of bias will be called frequency distribution bias.

An estimation method to eliminate the bias in 3) only is demonstrated. This method makes no attempt to address the bias in 1) above, nor the bias in 2). Should, however, the notified cases form a random sample of TB cases in the community, and the genotyped cases a random sample of notified cases, then the present analysis could be used to make inferences about transmission in the community. In the remainder of this paper, it is assumed that genotyped cases do form a random sample of the notified cases so that the method addresses the question of the proportion of transmission represented among the notified cases only.

Four existing datasets are used to illustrate this method. In addition to a less biased estimate for the amount of clustering, an estimate of the variance, and hence confidence intervals, is obtained. However it must be noted that these results effectively assume no bias of type 2.

In subsequent sections the term 'population' will refer to all individuals in a community for whom a sputum-based positive TB diagnosis was made. The group for whom sputum samples were successfully typed will be called the 'sample'. Bias refers to the frequency distribution bias.

Methods

Notation and preliminaries

Note the following definitions that hold for the population:

- Each case in the population is a member of a cluster.
- Let M be the number of clusters, with the typical cluster indexed by $i = 1,\ldots, M$, and let a_i be the size of the i^{th} cluster.
- $a = (a_1, \ldots , a_M)$ therefore represents the sizes of the clusters in the population.
- The total number of cases is given by $A = \sum_{i=1}^{M} a_i$.
- Let A_k be the total number of clusters of size k, ($k = 1, \ldots, N = \max(a_i)$).
- The population vector of cluster size frequencies is then given by $\boldsymbol{A} = (A_1, \ldots , A_N)$.

S of the A total cases are typed and clustered using genotyping. It is assumed that the sampling process is independent of the clusters i.e. each case is equally likely to be typed, irrespective of which cluster it belongs to. This gives rise to observed clusters of size $\boldsymbol{s} = (s_1, \ldots, s_M)$. The investigator is of course unaware of the existence of clusters which have an observed size of zero. Nevertheless, $\boldsymbol{s}$ can be thus defined and has a multivariate hypergeometric distribution, with mass function

$$p(\boldsymbol{s}) = \frac{\Pi_{i=1}^{M} \binom{a_i}{s_i}}{\binom{A}{S}}$$

The value

$$S_k = \sum_{i=1}^{M} 1_{\{s_i = k\}}$$

represents the total number of clusters of size k observed in the typed sample. 1 denotes the indicator function, that is,

$$1_{\{s_i = k\}} = \begin{cases} 1 & if \quad s_i = k \\ 0 & else \end{cases}$$

The sample vector of cluster frequencies (a histogram of observed cluster sizes) is given by: $S = (S_1,\ldots,S_N)$. Of course, it is not possible to know N, the true size of the largest cluster. Thus, a truncated vector $\tilde{\boldsymbol{S}} = (S_1,\ldots,S_{\tilde{N}})$ is observed, where $\tilde{N} = \max(s_i)$ is the largest observed cluster size.

Let $\mathbf{P}$ denote the matrix

$$\mathbf{P} = \mathbf{P}(N) = \begin{bmatrix} p(1,1) & p(1,2) & p(1,3) & \ldots & p(1,N) \\ 0 & p(2,2) & p(2,3) & & p(2,N) \\ 0 & 0 & p(3,3) & & p(3,N) \\ \vdots & & & \ddots & \vdots \\ 0 & & \cdots & & p(N,N) \end{bmatrix}$$

where

$$p(k,n) = \frac{\binom{n}{k}\binom{A-n}{S-k}}{\binom{A}{S}}$$

is the hypergeometric probability mass function, which represents the probability that a population cluster of size n presents as a cluster of size k in the sample.

It is additionally useful to define the probability that population clusters of size m and n present as sample clusters of size j and k respectively:

$$q(j,k,m,n) = \frac{\binom{m}{j}\binom{n}{k}\binom{A-n-m}{S-j-k}}{\binom{A}{S}}$$

Quantities of interest

The main quantities of interest are:

1. M , the total number of clusters
2. A_1, the number of unclustered cases (i.e. the number of singleton clusters)
3. $p_n = \frac{A - A_1}{A}$, the proportion of cases not in singleton clusters. According to the 'n method' heuristic, this is the proportion of recent transmissions.
4. $p_{n-1} = \frac{A - M}{A}$, the proportion of cases which are not the first case in a cluster. According to the 'n-1 method' heuristic, this is the proportion of recent transmissions.

Naïvely, one would estimate $p_n = \frac{S - S_1}{S}$ and $p_{n-1} = \frac{S - \sum S_k}{S}$. These estimators have been shown to be biased [3,4]. In the subsequent section, unbiased estimators are derived for M and A_1, whence estimators for p_n and p_{n-1} may be directly derived by substitution into their definitions (A is simply the reported number of positive TB diagnoses in the study). The uncertainty (standard error) inherent in the estimator is also analyzed in order to obtain confidence intervals.

Estimator of M and A_1

It can be shown that (see Supporting Information S1), for $k = 1, \ldots, N$,

$$\mathrm{E}(S_k) = \sum_{n \geq k} p(k,n) A_n$$

or, in matrix notation,

$$\mathrm{E}(\mathbf{S}) = \mathbf{P}\boldsymbol{A} \tag{1}$$

It can also be shown that the diagonal elements of the covariance matrix, $\mathrm{cov}(S)_{kk}$ are given by:

$$\sum_{m,n=k}^{N} (q(k,k,m,n) - p(k,m)p(k,n)) A_m A_n + \sum_{n=k}^{N} (p(k,n) - q(k,k,n,n)) A_n \tag{2}$$

And the off-diagonal elements $\mathrm{cov}(S)_{kl}$ by:

$$\sum_{m=k}^{N} \sum_{n=l}^{N} (q(k,l,m,n) - p(k,m)p(l,n)) A_m A_n - \sum_{n=\max(k,l)}^{N} q(k,l,n,n) A_n \tag{3}$$

From equation 1, a crude estimate $\hat{\boldsymbol{A}}$ of $\boldsymbol{A}$ may be obtained by simply matching the first moment, that is,

$$\hat{\boldsymbol{A}} = \mathbf{P}^{-1}\boldsymbol{S} \tag{4}$$

The covariance matrix of $\hat{\boldsymbol{A}}$ may be calculated using equations 2 and 3. Due to high variance and high correlation between components, this performs poorly as an estimate of $\boldsymbol{A}$. The origin and consequences of this inherent instability are discussed in the Supporting Information S2.

The crucial quantities of interest are the total number of clusters $M = \sum_{i=1}^{N} A_i$ and the number of singletons, A_1. An estimator of M

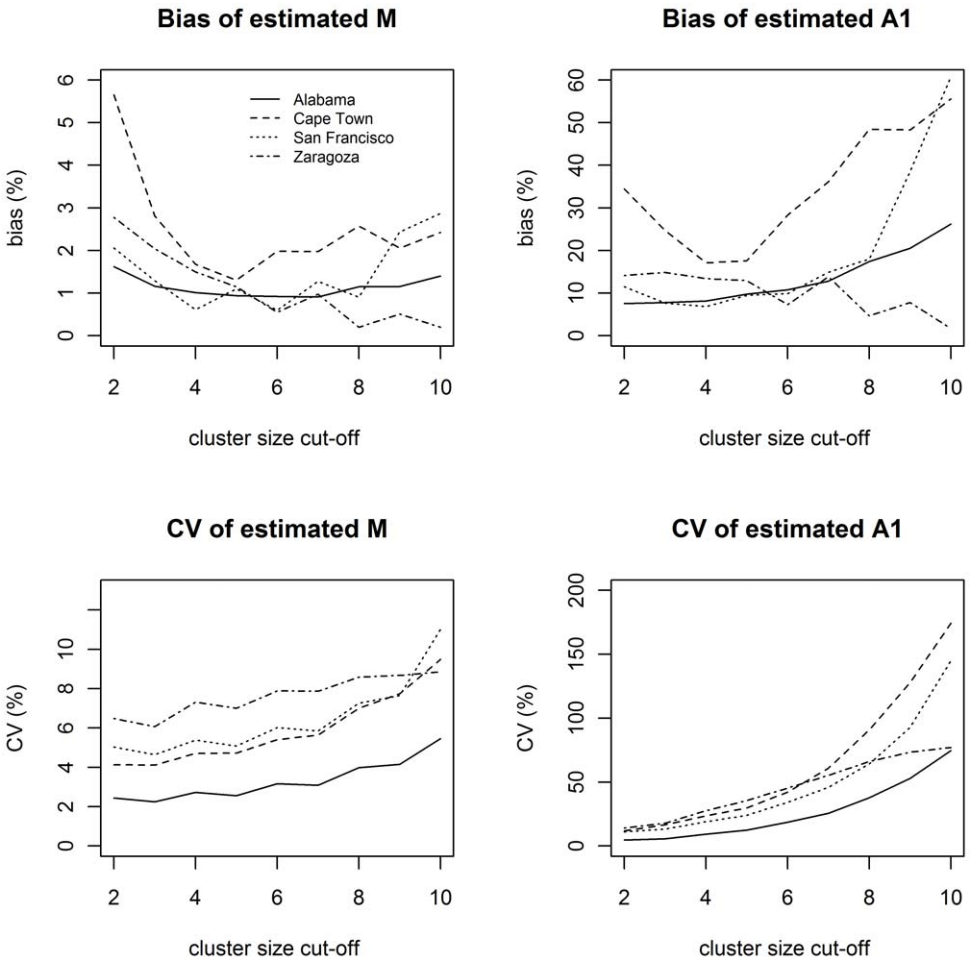

Figure 1. Normalized bias and coefficient of variation for estimated M and A_1 for sampling rate of 40%. The normalized bias alternates between positive and negative values at even and uneven truncations respectively. For ease of viewing, the absolute values of normalized bias are shown.
doi:10.1371/journal.pone.0034109.g001

can be derived as:

$$\hat{M} = \sum_{i=1}^{N} \hat{A}_i. \quad (5)$$

The first element of the vector $\hat{\boldsymbol{A}}$ is an estimate for A_1. These are unbiased as a consequence of equation 1. Expressions for the variances $\sigma^2_{\hat{M}}$ and $\sigma^2_{\hat{A}_1}$ may be derived from equations 2 and 3. These quantities still depend on the (unknown) $\boldsymbol{A}$. The estimate of $\hat{\boldsymbol{A}} = \mathbf{P}^{-1}\boldsymbol{S}$ can be used to obtain estimates $\hat{\sigma}^2_{\hat{M}}$ and $\hat{\sigma}^2_{\hat{A}_1}$ of $\sigma^2_{\hat{M}}$ and $\sigma^2_{\hat{A}_1}$.

Approximate confidence intervals

Since $\hat{\boldsymbol{M}}$ is a linear combination of random variables (albeit with some dependence), a normal approximation to the distribution of $\hat{\boldsymbol{M}}$ seems reasonable. Assuming this approximation is valid, the estimate $\hat{\sigma}^2_{\hat{M}}$ of $\sigma^2_{\hat{M}}$ can be used, which leads to the approximate (1 - α) - level confidence interval of

$$\left[\hat{M} - z_{\alpha/2}\hat{\sigma}_{\hat{M}}, \hat{M} + z_{\alpha/2}\hat{\sigma}_{\hat{M}}\right]$$

Similarly, an approximate (1 - α) - level confidence interval for A_1 is given by:

$$\left[\hat{A}_1 - z_{\alpha/2}\hat{\sigma}_{\hat{A}_1}, \hat{A}_1 + z_{\alpha/2}\hat{\sigma}_{\hat{A}_1}\right]$$

Computational considerations

Note that the estimates in equations 4 and 5 still involve an unknown quantity, namely N, the maximum population cluster size. However, from the upper triangular structure of $\mathbf{P}$ and the fact that $S_n = 0$ for $n > \tilde{N}$, the estimates are unchanged if $\mathbf{P}(N)$ is replaced by $\mathbf{P}(\tilde{N})$ and $\boldsymbol{S}$ by the observed vector $\tilde{\boldsymbol{S}}$.

The matrix $\mathbf{P}$ is close to singular when $\tilde{N}$ is large. A truncation approach to the use of the $\mathbf{P}$ matrix is now introduced. The observed vector $\tilde{\boldsymbol{S}}$ is divided into two parts $\boldsymbol{S}_0$ (the first C components, which lead to a numerically stable inversion of $\mathbf{P}$) and $\boldsymbol{S}_1$ (the remaining $\tilde{N} - C$components). The vector $\boldsymbol{S}_0$ is used as the input into the method outlined above, and the number of clusters in $\boldsymbol{S}_1$ is simply a known number of clusters to be added to any inferred total cluster count. The key question that arises is whether

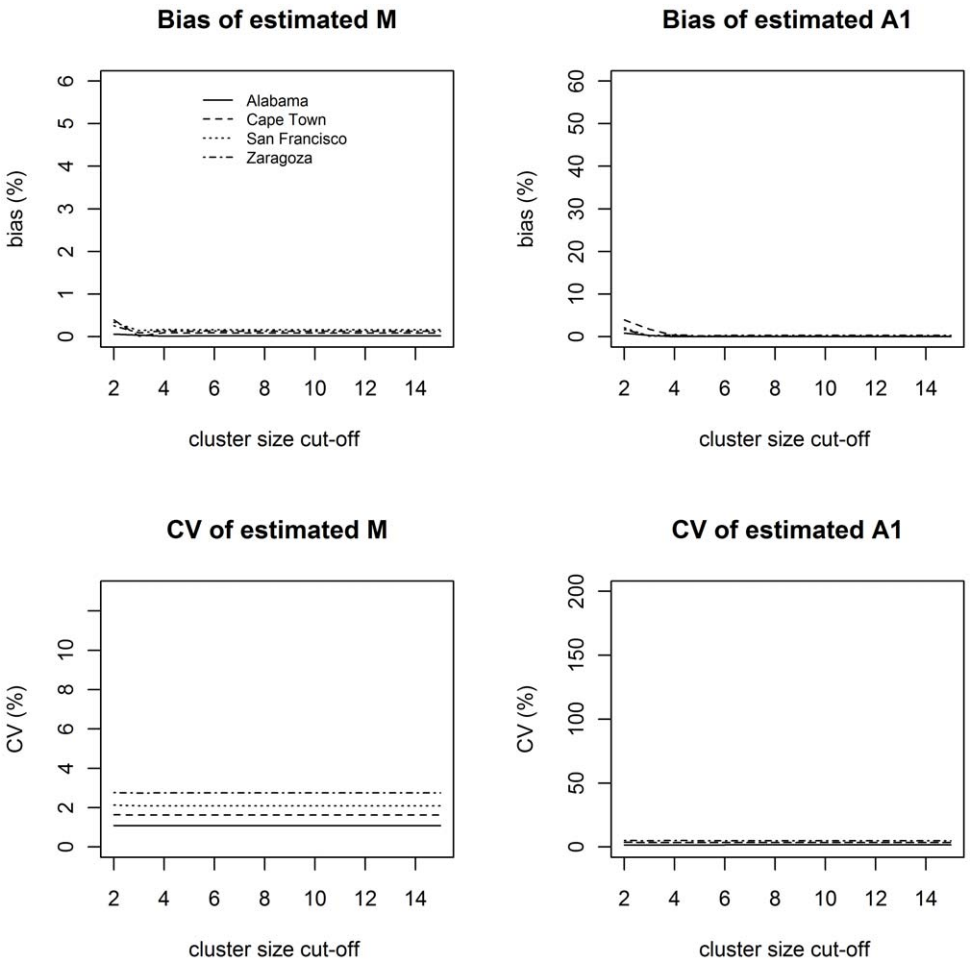

Figure 2. Normalized bias and coefficient of variation for estimated M and A_1 for sampling rate of 70%. The maximum normalized bias at truncation cluster size of 6 is 0.164% for M and 0.299% for A_1. The maximum CV at the same truncation is 2.75% for M and 4.92% for A_1. The normalized bias alternates between positive and negative values at even and uneven truncations respectively. For ease of viewing, the absolute values of normalized bias are shown.
doi:10.1371/journal.pone.0034109.g002

the truncation leads to stable estimates of the number of clusters, M, and the number of singletons, A_1, and this is investigated below.

The maximum cluster size for which the **P** matrix is still numerically non-singular will be the maximum cluster size at which the vector $\boldsymbol{S}_0$ can be truncated. Within this range, it is now possible to explore bias and variance of estimates of M and A_1. This is done for hypothetical populations at various sampling rates and truncations under Results.

Results

Exploring bias and variance of M and A_1

In order to explore this bias and variance, suitable hypothetical populations are required since actual populations are not known. Available data from 4 cities (Alabama [5], Cape Town [6], San Francisco [1], Zaragoza [7]) were used to generate hypothetical populations which produce samples clustered close to the observed data sets (see the method and data in Supporting Information S3). One thousand samples for each of these populations were obtained by sampling according to the multivariate hypergeometric distribution given in Equation 1. (The statistical software R version 2.14.1 was used for all analyses.) The **P** matrix inversion method described above was used to obtain estimates for the number of clusters, M , and the number of singletons, A_1. The coefficient of variation (CV) and the absolute of normalized bias for sampling rates of 40% and 70% are illustrated in Figure 1 and Figure 2, respectively.

At low sampling rates the estimates of the number of clusters, M, are more stable than the estimates of the number of singletons, A_1. The relative variability for both M and A_1 increases as the truncation increases. The bias for M is very small, and shows initial decrease with increasing truncation. The bias for A_1, which is considerably higher than the bias for M, increases with increasing truncation. This may be an indication that estimates for A_1 are not very reliable at low sampling rates.

In Figure 2 a higher sampling rate of 70% is considered. The graphs are plotted on the same scales as in Figure 1 to illustrate that bias and variability decreases dramatically for this higher sampling rate. Interestingly, the bias and variability also show no dependence on truncation.

Figure 3 illustrates the decrease in bias and variability at a fixed truncation of 6 for various sampling rates, showing that both are very small for sampling rates of 50% and greater.

Figure 4 shows that the bias in the estimate of the proportion of cases which are not singletons, p_n, using the **P** matrix inversion method, drops to almost zero for sampling rates of 50% and

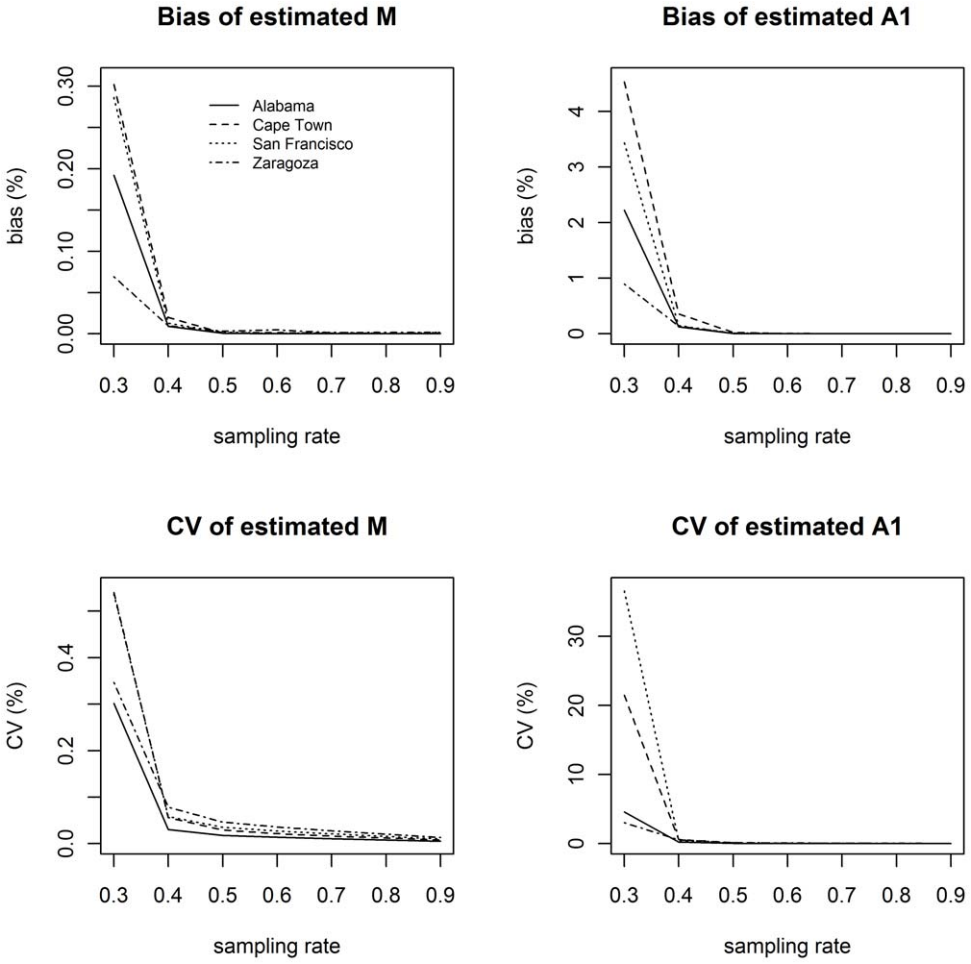

Figure 3. Normalized bias and coefficient of variation for estimated M and A_1 when the vector S is truncated at size 6. The maximum normalized bias at a sampling rate of 50% is 0.206% for M and 2.38% for A_1. The maximum CV at the same sampling rate is 4.72% for M and 13.93% for A_1.
doi:10.1371/journal.pone.0034109.g003

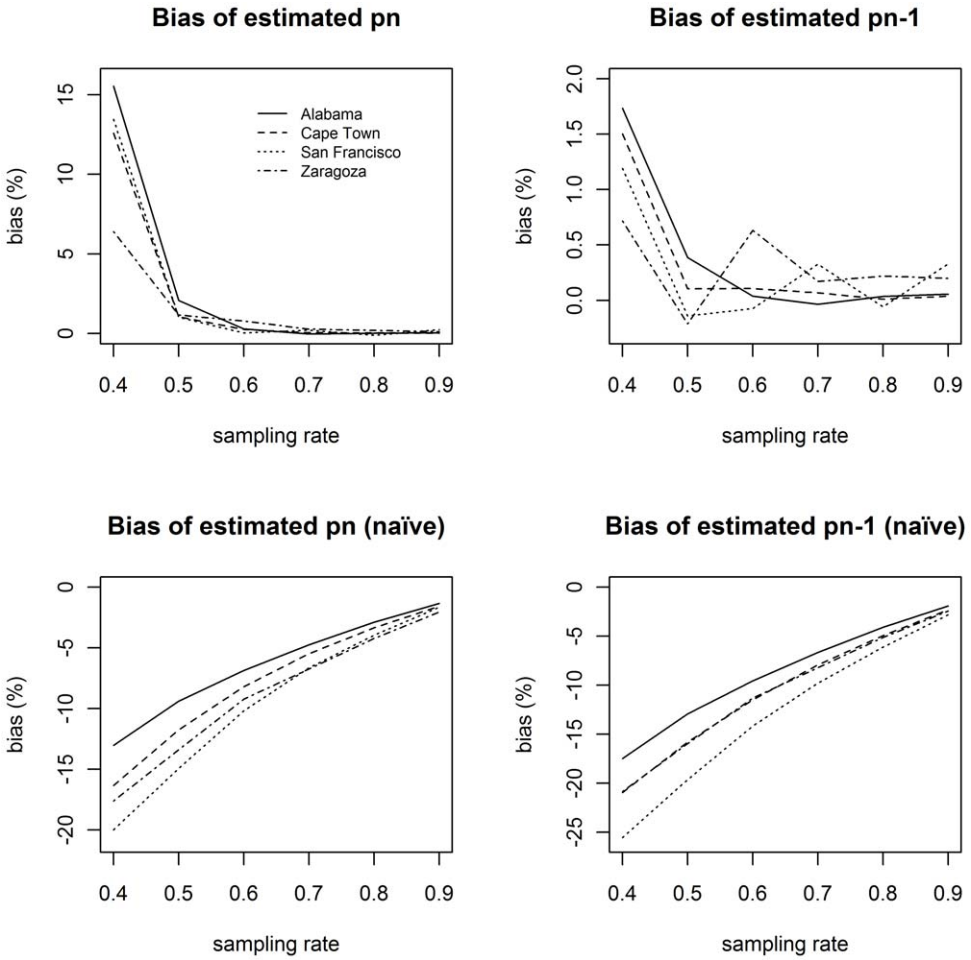

Figure 4. Bias in proportions using the naïve method, or the P matrix inversion method when the vector S is truncated at size 6.
doi:10.1371/journal.pone.0034109.g004

higher, while the decrease in bias for the naïve method is much slower. Similarly, the bias in the estimate of p_{n-1}, by the **P** matrix inversion method, is less than 2% at the low sampling rate of 40% and drops to almost zero for sampling rates of 50% and higher, while the decrease in bias for the naïve method is again much slower.

The instabilities introduced by this approach could in principle be addressed by adopting a maximum likelihood or fully Bayesian approach to the inference. Given that 1) the method is quite stable in the needed regime, 2) the elements of **P** are 'cluster level' likelihoods and the full 'cluster size histogram' level likelihood analysis is therefore considerably more complicated, and 3) this additional complexity would still not address the inherent limitations of the sampling frame, there is probably no real benefit in pursuing these more general approaches.

Applying method to existing data

The **P** matrix inversion method is applied to the same datasets considered above. In these datasets, 70% or more of TB diagnoses were typed. Based on the conclusion drawn in the previous section, any truncation can be chosen with negligible risk of introducing significant bias or variance. The results obtained effectively assume no bias of type 2). Table 1 shows the fraction of patients sampled, followed by estimates (and standard errors) for the total number of

Table 1. Estimated quantities, with standard errors for truncation = 6.

Datasets	A	r	$\hat{M}$	$\hat{A}$	$\hat{p}_n$	$\hat{p}_{n-1}$	$\hat{p}_{n(naive)}$	$\hat{p}_{n-1(naive)}$
Alabama	2204	0.8	1408 (11.7)	1271 (14.8)	0.422 (0.0067)	0.359 (0.0053)	0.41	0.345
Cape Town	2093	0.7	895 (14.7)	636 (23.04)	0.696 (0.011)	0.572 (0.007)	0.653	0.528
San Francisco	585	0.81	391 (5.86)	339 (7.79)	0.419 (0.0133)	0.33 (0.01)	0.404	0.311
Zaragoza	486	0.93	276 (3.04)	227 (4.09)	0.534 (0.0083)	0.434 (0.0062)	0.526	0.427

Note: These estimates are for the notified cases only, under the assumption of no type 2 bias.
doi:10.1371/journal.pone.0034109.t001

clusters, the number of clusters of size one, and the fraction, p, of transmitted cases, according to the 'n method' and 'n-1 method'. In addition to the **P** matrix inversion method, these fractions are also calculated with the naïve method. The observed vector $\boldsymbol{S}$ is truncated at cluster size 6. The R code used to produce Table 1 is provided in the Supporting Information S3.

Discussion

Cluster analysis is used to estimate the proportion of transmission of tuberculosis in a community. However, it is subject to limitations, many of which have been discussed elsewhere [3,4,8]. Previous attempts at assessing the magnitude of bias in the proportion of transmission failed to adequately distinguish between three distinct sources of bias: 1) the cases reporting to clinics are unlikely to represent a random sample of all active TB cases in the community, 2) the genotyped cases are not necessarily a random sample of the reported cases, 3) frequency distribution bias. The extent of bias of types 1) and 2) relative to bias of type 3) will vary according to local conditions and no general statement concerning this extent is possible here.

The present work identifies these separate problems, but solves the third problem only, under the assumption that the subset is random.

Given genotyped data on at least a majority of positive TB diagnoses (within a defined sampling frame) this work presents a method of inferring, robustly, the number of singletons and clusters that would have been observed if all positive sputa had been genotyped. This leads to an unbiased estimate of the proportion of transmission among the notified TB cases in the community.

It should be noted however, that the genotype methods currently used do not have perfect sensitivity and specificity. The choice of method necessarily results in a compromise between an evolutionary rate that is fast enough so as to provide sufficient discriminatory power between unrelated disease cases and yet still link related cases. Therefore measures of recent transmission that account for genetic heterogeneity and fingerprint pattern change rate need to be developed to ensure that the sample cluster distribution accurately represents the reality in the population [9]. Scott et al [10] investigated and compared three measures – IS*6110* RFLP, both dichotomous and continuous (nearest genetic distance) and PCR-based. They concluded that the poor sensitivity of the standard IS6110 RFLP test leads to estimates of clustering that are likely too low yet IS6110 typing remains the best method, at least in a low-incidence setting where the population of M. tuberculosis isolates shows a high degree of genetic diversity. This is in large part because IS6110 typing has the slowest evolution rate. The **P** matrix inversion method assumes that typing is accurate.

It should, moreover, also be noted that not all cases in a cluster are necessarily related by infection events. It is possible for a case to be the result of re-activation, i.e. to be endogenous, and to be misinterpreted. Thus bias lowering the number of singletons may be present. These considerations are investigated by Pretorius et al [11] and do not form part of the present work.

Knowledge of the relative impact of transmission, versus reactivation disease, can be used to design and evaluate transmission reduction programs, and target vulnerable locations or regions with appropriate interventions. This may be particularly appropriate for investigating antibiotic resistance worldwide, to explore the extent to which resistant strains are actively circulating in the community, or emerging de novo in sub-optimally treated patients. Cluster analysis is an invaluable tool to assist such investigations.

Supporting Information

Supporting Information S1 This file describes the derivation of Equations (1)–(3).
(DOC)

Supporting Information S2 This file describes the origin and consequences of the inherent instability in estimating the vector A with equation (4).
(DOC)

Supporting Information S3 This file provides the algorithm for creating hypothetical populations for a given sample and code to reproduce the results in Table 1.
(DOC)

Author Contributions

Conceived and designed the experiments: PU PvH GvdS. Analyzed the data: CvS MC PU. Wrote the paper: CvS MC AW PU.

References

1. Small PM, Hopewell PC, Singh SP, Paz A, Parsonnet J, et al. (1994) The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. N Engl J Med 330: 1703–1709.
2. Haddad MB, Diem LA, Cowan LS, Cave MD, Bettridge J, et al. (2007) Tuberculosis genotyping in six low-incidence States, 2000–2003. Am J Prev Med 32: 239–243.
3. Glynn JR, Vynnycky E, Fine PE (1999) Influence of sampling on estimates of clustering and recent transmission of Mycobacterium tuberculosis derived from DNA fingerprinting techniques. Am J Epidemiol 149: 366–371.
4. Murray M (2002) Sampling bias in the molecular epidemiology of tuberculosis. Emerg Infect Dis 8: 363–369.
5. Kempf MC, Dunlap NE, Lok KH, Benjamin WH, Jr., Keenan NB, et al. (2005) Long-term molecular analysis of tuberculosis strains in alabama, a state characterized by a largely indigenous, low-risk population. J Clin Microbiol 43: 870–878.
6. van der Spuy GD, van Helden PD, Warren RM (2009) Effect of study duration on the interpretation of tuberculosis molecular epidemiology investigations. Tuberculosis (Edinb) 89: 238–242.
7. Lopez-Calleja AI, Lezcano MA, Vitoria MA, Iglesias MJ, Cebollada A, et al. (2007) Genotyping of Mycobacterium tuberculosis over two periods: a changing scenario for tuberculosis transmission. Int J Tuberc Lung Dis 11: 1080–1086.
8. Borgdorff MW, van den Hof S, Kalisvaart N, Kremer K, van Soolingen D (2011) Influence of sampling on clustering and associations with risk factors in the molecular epidemiology of tuberculosis. Am J Epidemiol 174: 243–251.
9. Benedetti A, Menzies D, Behr MA, Schwartzman K, Jin Y (2010) How close is close enough? Exploring matching criteria in the estimation of recent transmission of tuberculosis. Am J Epidemiol 172: 318–326.
10. Scott AN, Joseph L, Belisle P, Behr MA, Schwartzman K (2008) Bayesian modelling of tuberculosis clustering from DNA fingerprint data. Stat Med 27: 140–156.
11. Pretorius C, Dodd P, Wood R (2011) An investigation into the statistical properties of TB episodes in a South African community with high HIV prevalence. J Theor Biol 270: 154–163.