Whole genome sequencing shows a low proportion of tuberculosis disease is attributable to close contacts in rural Malawi

Judith R Glynn[1]*, José Afonso Guerra-Assunção[1], Rein MGJ Houben[1], Lifted Sichali[2], Themba Mzembe[2], Lorrain K Mwaungulu[2], Nimrod J Mwaungulu[2], Ruth McNerney[3], Palwasha Kahn[1], Julian Parkhill[4], Amelia C Crampin[1], Taane G Clark[1,3]

* Corresponding author

[1] Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London UK

[2] Karonga Prevention Study, Malawi

[3] Faculty of Infectious Diseases, London School of Hygiene & Tropical Medicine, London UK

[4] Wellcome Trust Sanger Institute

Running head: Tuberculosis not from close contacts

Word count: Abstract 248; Text 2481

Funding: The Wellcome Trust [grant number 096249/Z/11/B]

Corresponding author:

Prof Judith Glynn, LSHTM, Keppel St, London WC1E 7HT

Judith.glynn@lshtm.ac.uk

Tel: 0207 927 2423

Alternate: Dr Taane Clark, LSHTM, Keppel St, London WC1E 7HT

Taane.clark@lshtm.ac.uk

Tel: 0207 927 2484

Short summary: [40 words]

In a large long-term study in rural Malawi, whole genome sequencing showed that most smear-positive prior contacts identified by tuberculosis patients were not the sources of their infection, and overall less than 10% of tuberculosis was attributable to known contacts.

Abstract

Background
The proportion of tuberculosis attributable to transmission from close contacts is not well known. Comparison of the genome of strains from index patients and prior contacts allows transmission to be confirmed or excluded.

Methods
In Karonga District, Malawi, all tuberculosis patients are asked about prior contacts with others with tuberculosis. All available strains from culture-positive patients were sequenced. Up to 10 single nucleotide polymorphisms (SNPs) between index patients and their prior contacts were allowed for confirmation, and ≥ 100 for exclusion. The population attributable fraction was estimated from the proportion of confirmed transmissions and the proportion of patients with contacts.

Results
From 1997-2010 there were 1907 new culture-confirmed tuberculosis patients, of whom 32% reported at least one family contact and an additional 11% had at least one other contact; 60% of contacts had smear-positive disease. Among case- contact pairs with whole genome sequences available, transmission was confirmed from 38% (62/163) smear-positive prior contacts and 0/17 smear-negative prior contacts. Confirmed transmission was more common in those related to the prior contact (42.4%, 56/132) than in non-relatives (19.4%, 6/31, p=0.02), and in those with more intense contact, to younger index cases, and in more recent years. The proportion of tuberculosis attributable to known contacts was 9.4% overall.

Conclusions
In this population known contacts only explained a small proportion of tuberculosis cases. Even those with a prior family contact with smear positive tuberculosis were more likely to have contracted their infection elsewhere. Contact tracing is likely to have minimal impact.

Background

Understanding where *M.tuberculosis* transmission is occurring is key to tuberculosis control. Close contact with someone with tuberculosis is a known risk factor for infection and hence disease, so contact tracing is often recommended.[1,2] However the proportion of disease due to such contacts is not certain, particularly in high prevalence areas. It can be investigated using traditional epidemiological techniques, comparing the contact histories of cases of tuberculosis and controls,[3] but this assumes that the increased risk in the cases is attributable to the contacts rather than shared risk factors, which can be difficult to adjust for. Older DNA fingerprinting techniques such as RFLP or MIRU-VNTR have improved on this by ensuring that the contacts share fingerprint patterns,[4-6] but the level of discrimination is limited if some DNA fingerprint strains are common, and it is impossible to exclude a common source.

Whole genome sequencing allows a more accurate approach:[7-9] the combination of a history of contact and a small number of single nucleotide polymorphisms (SNPs) between the strain in the index patient and their prior contact makes transmission highly likely, and can allow pinpointing of the most likely source if there is more than one. If the proportion of tuberculosis patients with contacts of different types is also known, an accurate estimate can be made of the proportion of tuberculosis attributable to these contacts.

In Karonga District, Malawi, we have been collecting data on all tuberculosis patients and their prior contacts since 1997. We have now used whole genome sequencing on all available case-contact pairs to improve understanding of transmission.

Methods

The Karonga Prevention study in northern Malawi has been studying tuberculosis in the whole of Karonga district (current population approximately 300,000) since the 1980s. Cases are identified through enhanced passive case finding with project staff based at the district hospital and major health centres to identify those with chronic cough or other symptoms suggestive of tuberculosis. Each patient is interviewed and at least three sputum specimens are taken at diagnosis, with further specimens at follow-up and at the end of treatment. Cultures from sputum (and other specimens if indicated) are set up in the laboratories in Malawi. Those that resemble *M.tuberculosis* macroscopically are sent to the UK mycobacterium reference laboratory for species identification and drug resistance testing.

Since March 1997 we have asked all patients about prior contacts with tuberculosis, in their family or household (ever), or other contacts (within the past 5 years). If they report contacts we ask further detail including names, the duration and location of contact, and whether they had contact when the first case was ill. If these prior contacts were treated within the district they will already be known to us, allowing us to confirm the type of tuberculosis (smear positive, smear negative pulmonary, or extrapulmonary) and other details.

Approval for the study was given by the ethics committee of the London School of Hygiene & Tropical Medicine and the Malawian National Health Sciences Research Committee. Informed consent was obtained from all participants.

SNP differences

We have carried out whole genome sequencing of all available cultures from case-prior contact pairs to establish whether transmission can be confirmed or not. The sequencing was carried out at the Sanger Institute, using Illumina HiSeq 2000 technology with paired-end reads of length 100 base-pairs. We used trimmomatic software (http://www.usadellab.org/cms/?page=trimmomatic) to remove low-quality reads and reads <50 base-pairs long. We mapped reads to the H37Rv reference

genome (Genbank assession: AL123456.3), using the *BWA-mem* algorithm (http://bio-bwa.sourceforge.net/) and excluded samples with an average genomic coverage <10-fold.

We identified SNP positions using *SAMtools* (http://samtools.sourceforge.net/). If alleles at a position were not identical we took the majority allele if it had a frequency of ≥ 75% and the position was supported by ≥20-fold coverage; otherwise we classified the position as missing (thus ignoring heterozygous calls). We excluded samples with >15% missing genotype calls, to remove possible contaminated or mixed samples or technical errors. (The proportion of mixed strains is low in this setting.[10]) We excluded genome positions with >15% missing genotypes, and those in highly repetitive and variable regions (e.g. PE/PPE genes). This quality control left 94% of the *M.tuberculosis* genome to be analysed for variants. Median coverage was 88-fold, mean 127. Spoligotyping was performed *in silico* using SpolPred.[11] Lineages were defined from spoligotype families.[12] We calculated SNP distances between sequences using the ape library in the R statistical package (http://cran.r-project.org/).

Based on the number of SNPs between samples in patients with multiple isolates,[13] and between patients with likely transmission in other analyses[8,9,14] and in this dataset (see below), we made the following rules: 0-10 SNPs transmission is likely; 11-99 SNPs transmission uncertain; ≥ 100 SNPs no transmission.

Confirming transmission

We excluded index cases who had had previous tuberculosis, and prior contacts with extrapulmonary tuberculosis (since this is not infectious). For those with more than one contact we selected the one with fewest SNPs different (and then the closest contact if there was still more than one). Index case-prior contact pairs with ≤10 SNPs were taken as confirmed. The mutation rate in these pairs was estimated using linear regression.

Cases had a higher chance of being included in the analysis if they named more contacts since they were included if sequence was available for at least one contact. However, assuming that tuberculosis is acquired from one source, those with more contacts are less likely to have transmission confirmed from each one. This could underestimate the proportion of confirmed transmissions. We attempted to minimise this bias by selecting the closest contact, as above, and assessed the extent of any remaining bias by comparing the proportion of confirmed transmissions among those with one or more than one named contact.

Contact analysis

We analysed risk factors associated with confirmation of transmission from the closest contact identified using logistic regression, after excluding pairs with 11-99 SNPs, and taking those with 0-10 SNPs as confirmed. Risk factors included: characteristics of the index case and the contact (age, sex, HIV status); characteristics of the strain (isoniazid resistance, *M.tuberculosis* lineage); and characteristics of the contact (relationship, intensity of contact, and time interval between the case and the contact). Intensity of contact was defined as high if the contact was prolonged, indoors and on more than one day, and very high if the case had nursed the prior contact while they were ill.

Proportion of cases due to transmission from named contacts

The proportion of confirmed transmissions in the case-contact pairs is the attributable risk percent since it is the proportion of tuberculosis cases with named contacts who have acquired tuberculosis from that contact. To estimate the proportion of tuberculosis cases due to transmission from named contacts in the whole population (the population attributable fraction, PAF) the attributable risk percent was multiplied by the proportion of all new culture-confirmed cases naming at least one contact, and the proportion of named contacts who are smear positive.

Results

Between March 1997 and March 2010 there were 1907 patients with culture confirmed tuberculosis having their first episode of tuberculosis; 32.1% (555 of 1727 with recorded data) had had at least one previous tuberculosis case in their household or close family and 15.8% (270/1705, 68.8% of whom had no family/household contact) had other known contacts with tuberculosis (figure 1). Of the named contacts reported to have been treated in Karonga District since 1986, 82.0% were identified as TB cases, of whom 59.9% had confirmed smear positive pulmonary disease.

Whole genome sequences that passed the quality control were available for both members of the pair for 207 case-contact pairs, including 20 cases with more than one prior contact. After selecting the most likely source there were 187 pairs (170 with a smear positive prior contact and 17 with a smear negative prior contact). The number of SNPs between the 170 pairs is shown in figure 2: 62 had ≤10 SNPs, 9 had 10-99 SNPs and 116 had ≥100 SNPs. Since there was no confirmed transmission from smear negative prior contacts (the minimum SNP distance was 35) the remaining analysis was restricted to smear positive prior contacts. In the whole dataset over the period of the study there were 406 culture positive index cases with at least one prior culture-confirmed smear positive contact. The 170 included cases were similar to those without paired sequences available in terms of age, sex and intensity of contact, but those diagnosed before 2000 were less likely to have sequences available.

Of the 170 included pairs, 36% (62/170) had transmission confirmed based on SNPs, or 38% (62/163) after excluding those with uncertain transmission. Among the 163 cases, 83 had named one (identified) contact, and 80 more than one. The proportion with confirmed transmission was 35% in those with one contact and 41% in those with more than one contact.

Figure 3 shows the number of SNPs and the time difference between disease onset in the prior contact and the case for the 62 pairs with ≤10 SNPs. From the slope of the linear regression the mutation rate is estimated at 0.33 SNPs/ year (95% CI 0.18-0.49, r^2 24%, p<0.001).

The characteristics of the cases and prior smear positive contacts and the associations with transmission are shown in table 1. Much the strongest association with transmission was the intensity of the contact. Information on intensity was missing for 16 pairs; 5 because questions on intensity were not asked in the first year of the study. Confirmed transmission was more common in those related to the prior contact (42.4%, 56/132) than in non relatives (19.4%, 6/31, *p*=0.02), and especially from spouses and parents (61.1%). The proportion of confirmed transmissions decreased with increasing age of the case and was lower in earlier years of the study.

Since intensity of transmission was so strongly associated with the outcome the multivariable analysis was restricted to the 147 pairs with information on this exposure. Relationship was strongly correlated with intensity of transmission (for example, no non-relatives had high or very high intensity (nursing) contacts), and after adjusting for intensity, relation was no longer associated with transmission. After adjusting for intensity, the associations with age of the case and year of diagnosis of the case became stronger (table 2), and there were borderline associations with sex of the case (higher in males) and age of the contact (lower proportion confirmed from older contacts). None of the other factors were associated with transmission once adjusted for intensity and the other factors shown in table 2. The adjusted odds ratio was 0.56 (95%CI 0.23-1.4) for HIV positive contacts vs HIV negative contacts, and 0.51 (0.19-1.4) for HIV positive cases vs HIV negative cases.

Table 3 shows estimates of the proportion of disease attributable to transmission from known contacts among patients with first episode culture-confirmed disease. It is assumed that the proportion of contacts with confirmed smear positive disease (59.9%) is the same in all groups, including those not identified in the database. Overall 9.4% of tuberculosis cases were attributable to transmission from known contacts. This was higher in younger individuals (11.8% aged <35) than

older individuals (7.7% aged 35+), in women (11.0%) than men (7.9%), and in 2004-2010 (10.3%) than in the earlier years (7.7%). Family contacts were much more important as a source than known outside contacts (87% overall) and this was particularly marked in women (96% vs 78% in men).

Discussion

This is the first study to calculate the proportion of tuberculosis attributable to transmission from known contacts established through whole genome sequencing. This technique provides the most accurate method available of verifying the source of transmission. We confirm the expected strong association of transmission with intensity of contact and smear positivity, but estimate that, overall, known smear positive prior contacts account for less than 10% of tuberculosis cases in this community, and that even for those with a prior contact with smear positive tuberculosis in their family, there was a >50% chance that they acquired their tuberculosis elsewhere.

The results are consistent with our earlier findings based on RFLP,[4] and a low proportion of tuberculosis attributable to household transmission has also been reported in other high prevalence settings using other techniques.[15] Using SNPs is a more accurate measure of similarity than RFLP, and the cut-off we used is in line with that used in other studies. The mutation rate in our study was also similar that found elsewhere measured between or within patients.[8,9] Using genomic similarity to confirm transmission assumes that neither the initial nor the subsequent episode is due to a mixed infection. We have previously found a low proportion of mixed infections in this setting, of around 3%.[10] We have assumed that the proportion of confirmed transmissions we found in the case-contact pairs for which we had samples and sequence available is applicable to all such pairs. Although cases had a higher chance of being included in the analysis if they named more contacts, there was no evidence that including those with multiple contacts lowered the proportion confirmed: those naming more contacts had a slightly higher proportion of confirmed transmissions.

The calculations of the population attributable fraction make the additional assumption that the proportion of contacts with smear positive disease overall is the same as that in those we identified in the database. This is probably an overestimate: many of those not identified may not have had tuberculosis at all. We have only included contacts known to and identified by the tuberculosis patients. This will underestimate the number of contacts but this is likely to be more of an issue for less close contacts who only contribute a small proportion of transmissions. Arbitrarily doubling the number of “other contacts” in those with no known family contact, for example, would only increase the proportion of tuberculosis attributable to known contacts from 9.4% to 10.7%.

As well as the variation by type of contact, we found variation by age and time period. The decreased proportion of confirmed transmissions to older cases is consistent with a higher proportion of reactivation disease with increasing age. And the higher proportion of confirmed transmissions in recent years is consistent with lower tuberculosis incidence and reduced transmission in the community.[16,17] There were also differences by sex, with non-family contacts being relatively more important for men, who spend more of their time away from home. We found only weak evidence of reduced transmission from (smear positive) HIV positive patients compared to HIV negative patients, suggesting that they are an important source of transmission.

Conclusions

In this setting, where tuberculosis is endemic, almost half of the individuals with culture-confirmed tuberculosis have had identified contact with previous patients with tuberculosis, often in their close family. Yet even those with a family contact with smear positive tuberculosis are likely to have acquired their tuberculosis elsewhere, and close contacts contribute less than 10% of sources of tuberculosis in the population. Contact tracing would be expected to have little impact on the burden of disease in this type of setting.

Raw sequence data are available from the European Nucleotide Archive (Accession numbers ERP000436 and ERP001072)

Funding

This work was supported by The Wellcome Trust [grant number 096249/Z/11/B]

Acknowledgements

We thank the Government of the Republic of Malawi for their interest in this Project and the National Health Sciences Research Committee of Malawi for permission to publish the paper. We thank the Wellcome Trust Sanger Institute core and pathogen sequencing and informatics teams.

Conflicts of interest

None of the authors have any conflicts of interest

Table 1: Number of confirmed transmissions by characteristics of the prior contact, the case, and the relationship between them

Characteristic		n/N	%	*P**
Prior Contact				
Age	<25	12/22	54.6	0.3
	25-34	21/61	34.4	
	35-44	15/38	39.5	
	45+	14/42	33.3	
Sex	Female	40/94	42.6	0.2
	Male	22/69	31.9	
Isoniazid	Resistant	5/11	45.5	0.6
	Sensitive	57/152	37.5	
Lineage	1	4/18	22.2	0.4
	2	3/9	33.3	
	3	11/24	45.8	
	4	44/112	39.3	
HIV status	HIV-	29/63	46.0	0.4
	HIV+ no ART	24/66	36.4	
	HIV+ ART	3/5	60.0	
Index Case				
Age	<25	17/28	60.7	0.04
	25-34	20/56	35.7	
	35-44	17/49	34.7	
	45+	8/30	26.7	
Sex	Female	38/97	39.2	0.7
	Male	24/66	36.4	
HIV status	HIV-	24/49	49.0	0.1
	HIV+ no ART	23/75	30.7	
	HIV+ ART	2/7	28.6	
Year	1997-2001	15/55	27.3	0.1
	2002-2006	32/74	43.2	
	2007-2010	15/34	44.1	
Relationship				
Relation	Spouse	11/18	61.1	0.01
	Parent	11/18	61.1	
	Child	1/6	16.7	
	Sibling	15/35	42.9	
	Other relation	18/55	32.7	
	Not related	6/31	19.4	
Intensity	Low	16/81	19.8	<0.001
	High	17/30	56.7	
	Nursing	24/36	66.7	
Interval	<1 year	20/44	45.5	0.1
	1-1.99 yrs	11/38	29.0	
	2-4.99 yrs	17/54	31.5	
	5+ yrs	14/27	51.9	

* from X^2 tests or Fisher's exact tests if numbers are small

Table 2: Factors associated with confirmation of transmission in the multivariable analysis

	Univariable			Multivariable, adjusted for the other factors in the table		
	OR	95% CI	P (lrtest)	OR	95% CI	P (lrtest)
High intensity (vs low)	5.3	2.1-13.1		3.1	1.1-8.8	
Nursing (vs low)	8.1	3.4-19.6	<0.001	11.6	4.2-32.1	<0.001
Age case (per year)	0.96	0.93-0.99	0.003	0.94	0.90-0.98	0.002
Age contact (per year)	0.98	0.96-1.0	0.1	0.97	0.4-1.0	0.06
Male case (vs female)	0.89	0.46-1.7	0.7	2.2	0.92-5.3	0.07
Year case (per year)	1.1	0.99-1.2	0.07	1.1	1.0-1.3	0.04

Table 3: Estimate of the proportion of first episode culture-confirmed cases attributable to known smear positive contacts

		A Proportion with contact (from data)	B Proportion with smear+ve contact (A x 0.599)	C Proportion smear+ve transmitting (from data)	D PAF (BxC)	PAF for any contact
Overall						
	Family	32.1%	19.2%	42.4%	8.2%	
	Other*	10.9%	6.5%	19.4%	1.3%	9.4%
Age < 35						
	Family	35.1%	21.0%	47.1%	9.9%	
	Other	11.1%	6.6%	28.6%	1.9%	11.8%
Age 35+						
	Family	31.4%	18.8%	37.1%	7.0%	
	Other	10.6%	6.3%	11.8%	0.7%	7.7%
Female						
	Family	39.6%	23.7%	44.6%	10.6%	
	Other	9.4%	5.6%	7.1%	0.4%	11.0%
Male						
	Family	26.5%	15.9%	38.8%	6.2%	
	Other	10.0%	6.0%	29.4%	1.8%	7.9%
1997-2003						
	Family	33.6%	20.1%	28.6%	5.8%	
	Other	14.0%	8.4%	23.1%	1.9%	7.7%
2004-2010						
	Family	32.7%	19.6%	50.0%	9.8%	
	Other	6.2%	3.7%	13.3%	0.5%	10.3%

* The proportion with other contacts excludes those with family contacts as well

Figure 1: Flowchart of patients included in the study

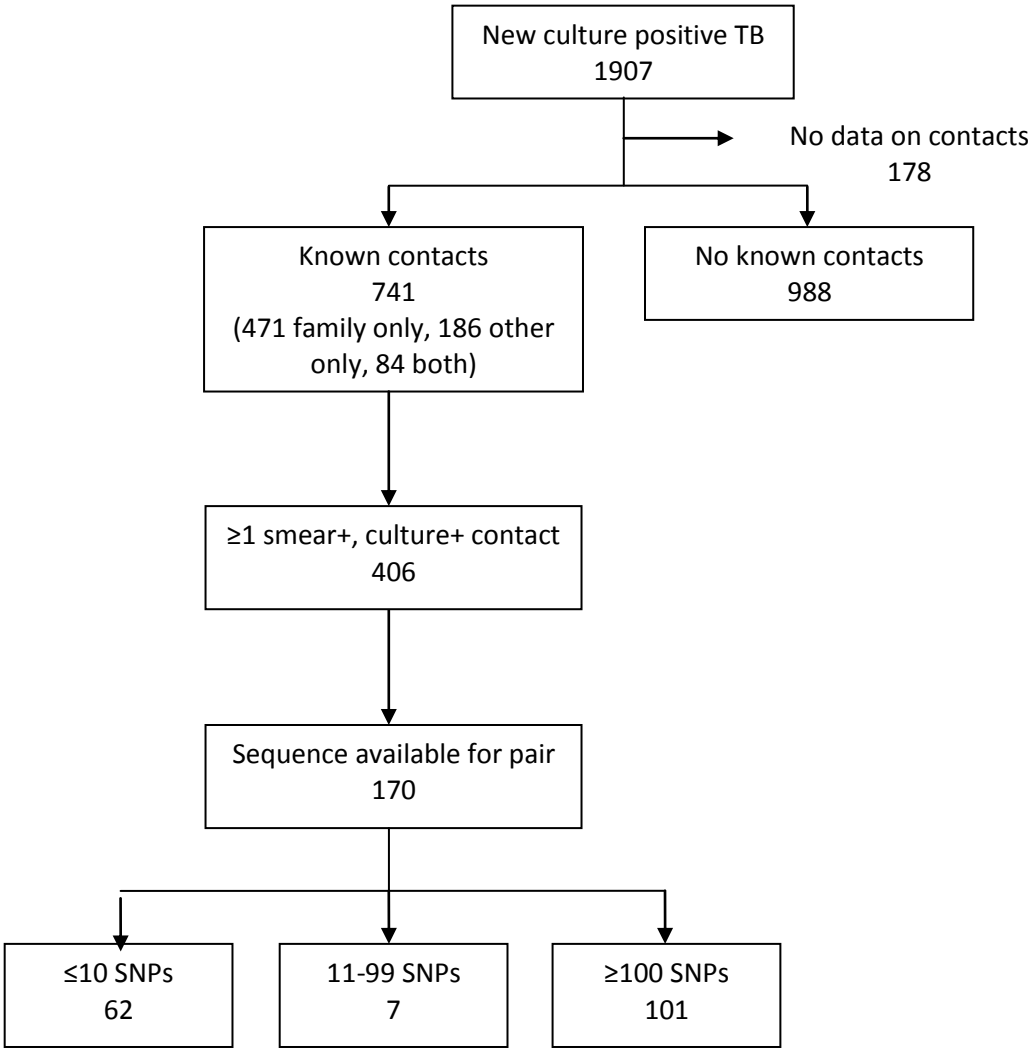


Figure 2: Number of SNPs between index patients and identified prior contacts, by sputum smear status of the prior contact.

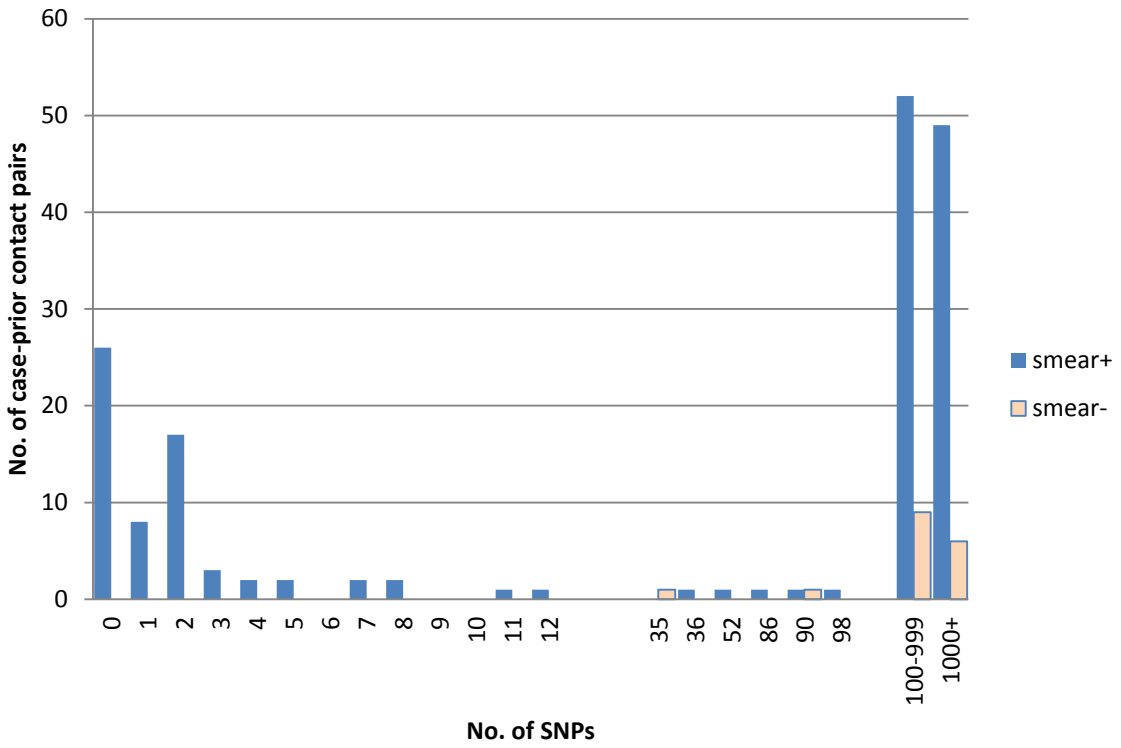
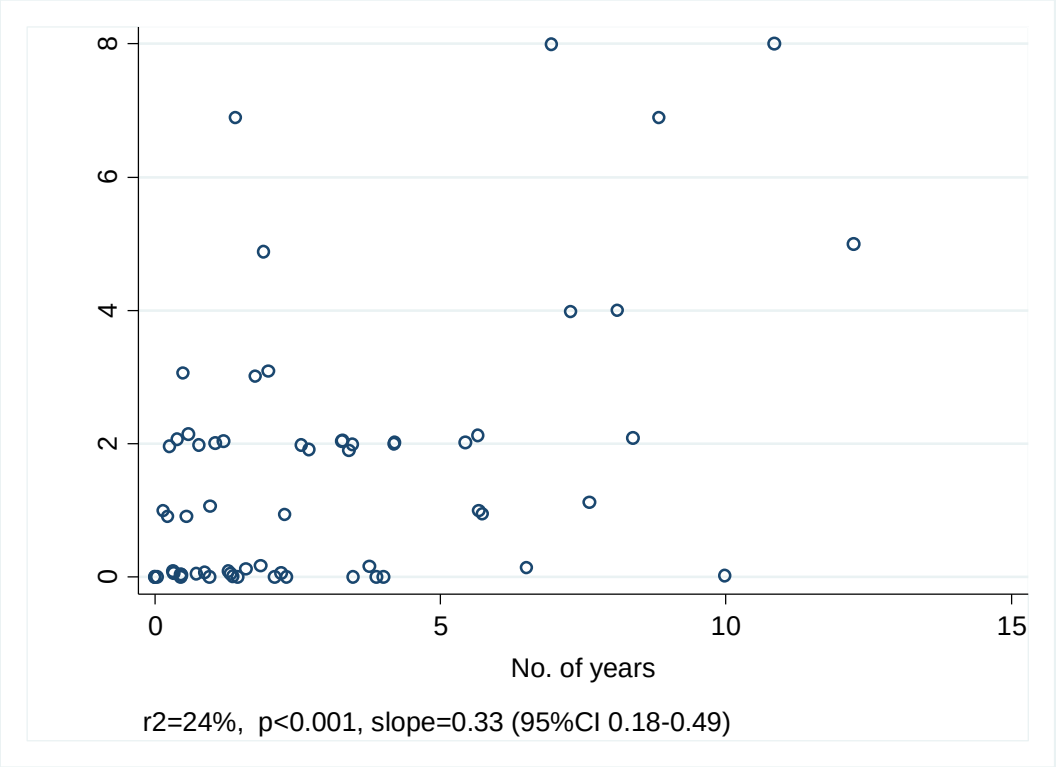


Figure 3: Number of SNPs by time interval between successive cases in 62 case-contact pairs with confirmed transmission. (Random noise has been introduced to allow those with identical numbers of SNPs to be visualised)



References

1. Lonnroth K, Corbett E, Golub J, et al. Systematic screening for active tuberculosis: rationale, definitions and key considerations. *Int J Tuberc Lung Dis* 2013;**17**(3)**:**289-98.
2. Fox GJ, Barry SE, Britton WJ, Marks GB. Contact investigation for tuberculosis: a systematic review and meta-analysis. *Eur Respir J* 2013;**41**(1)**:**140-56.
3. Crampin AC, Glynn JR, Floyd S, et al. Tuberculosis and gender: exploring the patterns in a case control study in Malawi. *Int J Tuberc Lung Dis* 2004;**8**(2)**:**194-203.
4. Crampin AC, Glynn JR, Traore H, et al. Tuberculosis transmission attributable to close contacts and HIV status, Malawi. *Emerg Infect Dis* 2006;**12**(5)**:**729-35.
5. Anderson LF, Tamne S, Brown T, et al. Transmission of multidrug-resistant tuberculosis in the UK: a cross-sectional molecular and epidemiological study of clustering and contact tracing. *Lancet Infect Dis* 2014;**14**(5)**:**406-15.
6. Verver S, Warren RM, Munch Z, et al. Proportion of tuberculosis transmission that takes place in households in a high-incidence area. *Lancet* 2004;**363:**212-4.
7. Walker TM, Lalor MK, Broda A, et al. Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study. *Lancet Respir Med* 2014;**2**(4)**:**285-92.
8. Walker TM, Ip CL, Harrell RH, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *Lancet Infect Dis* 2013;**13**(2)**:**137-46.
9. Bryant JM, Schurch AC, van Deutekom H, et al. Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data. *BMC Infect Dis* 2013;**13**(1)**:**110.
10. Mallard K, McNerney R, Crampin AC, et al. Molecular detection of mixed infections of Mycobacterium tuberculosis strains in sputum samples from patients in Karonga District, Malawi. *J Clin Microbiol* 2010;**48**(12)**:**4512-8.
11. Coll F, Mallard K, Preston MD, et al. SpolPred: rapid and accurate prediction of Mycobacterium tuberculosis spoligotypes from short genomic sequences. *Bioinformatics* 2012;**28**(22)**:**2991-3.
12. Demay C, Liens B, Burguiere T, et al. SITVITWEB--a publicly available international multimarker database for studying Mycobacterium tuberculosis genetic diversity and molecular epidemiology. *Infect Genet Evol* 2012;**12**(4)**:**755-66.
13. Guerra Assunção JA, Houben RMGJ, Crampin AC, et al. Relapse or reinfection with tuberculosis: a whole genome sequencing approach in a large population-based cohort with high HIV prevalence and active follow-up. *J Infect Dis* 2014 (in press).
14. Perez-Lago L, Comas I, Navarro Y, et al. Whole Genome Sequencing Analysis of Intrapatient Microevolution in Mycobacterium tuberculosis: Potential Impact on the Inference of Tuberculosis Transmission. *J Infect Dis* 2014;**209**(1)**:**98-108.
15. Narain R, Nair SS, Ramanath Rao G, Chandrasekhar P. Distribution of tuberculous infection and disease among households in a rural community. *Bull WHO* 1966;**34:**639-654.
16. Mboma SM, Houben RM, Glynn JR, et al. Control of (multi)drug resistance and tuberculosis incidence over 23 years in the context of a well-supported tuberculosis programme in rural Malawi. *PLoS One* 2013;**8**(3)**:**e58192.
17. Zachariah R, Bemelmans M, Akesson A, et al. Reduced tuberculosis case notification associated with scaling up antiretroviral treatment in rural Malawi. *Int J Tuberc Lung Dis* 2011;**15**(7)**:**933-7.