Functional and evolutionary genomics of *Mycobacterium tuberculosis*: Insights from genomic deletions in 100 strains

Anthony G. Tsolaki*, Aaron E. Hirsh†, Kathryn DeRiemer*, Jose Antonio Enciso*‡, Melissa Z. Wong*, Margaret Hannan*, Yves-Olivier L. Goguet de la Salmoniere*, Kumiko Aman*, Midori Kato-Maeda*, and Peter M. Small*§

*Division of Infectious Diseases and Geographic Medicine, Stanford University Medical Center, Stanford, CA 94305; †Department of Biological Sciences, Stanford University, Stanford, CA 94305; and ‡Unidad de Investigación Médica en Enfermedades Infecciosas y Parasitarias–Pediatria, Instituto Mexicano del Seguro Social, Mexico City 06725, Mexico

Edited by Burton H. Singer, Princeton University, Princeton, NJ, and approved January 15, 2004 (received for review September 16, 2003)

To better understand genome function and evolution in *Mycobacterium tuberculosis*, the genomes of 100 epidemiologically well characterized clinical isolates were interrogated by DNA microarrays and sequencing. We identified 68 different large-sequence polymorphisms (comprising 186,137 bp, or 4.2% of the genome) that are present in H37Rv, but absent from one or more clinical isolates. A total of 224 genes (5.5%), including genes in all major functional categories, were found to be partially or completely deleted. Deletions are not distributed randomly throughout the genome but instead tend to be aggregated. The distinct deletions in some aggregations appear in closely related isolates, suggesting a genomically disruptive process specific to an individual mycobacterial lineage. Other genomic aggregations include distinct deletions that appear in phylogenetically unrelated isolates, suggesting that a genomic region is vulnerable throughout the species. Although the deletions identified here are evidently inessential to the causation of disease (they are found in active clinical cases), their frequency spectrum suggests that most are weakly deleterious to the pathogen. For some deletions, short-term evolutionary pressure due to the host immune system or antibiotics may favor the elimination of genes, whereas longer-term physiological requirements maintain the genes in the population.

Comparisons of complete microbial genome sequences have revealed significant differences in gene content and genome organization between closely related bacteria (1–5). In some species, comparisons have identified multigene clusters that appear to augment pathogenicity (2, 5). In other species, such as *Mycobacterium tuberculosis*, the comparison of complete genome sequences has identified large-sequence polymorphisms (LSPs) and single-nucleotide polymorphisms (SNPs), but the molecular basis of variability in virulence and transmissibility remains undefined (6, 7). A distinct but complementary approach to comparative genomics involves the interrogation of unsequenced genomes by DNA microarray to identify sequences present in a fully sequenced isolate, but absent from interrogated isolates (8–16). Although this approach is limited to the identification of relatively large sequence polymorphisms, it allows for the comparison of a large number of genomes, and thus provides information on the diversity and frequency of polymorphisms in the population.

Here we apply microarray-based comparative genomics to 100 epidemiologically well characterized isolates of *M. tuberculosis*. Because rates of SNPs are low in this species (6, 17, 18), large sequence differences detectable by microarray are likely to be an important source of genetic variation (6, 12). In addition, information on the frequency and diversity of polymorphisms allows for population and evolutionary genetic analyses, which may cast light on the relatively complex, and as yet poorly understood, basis of phenotypic variability in this important pathogen. To provide a useful data set for these analyses, we have substantially expanded the sample size of previous microarray-based comparisons (8–16). In addition, we have mapped LSPs to the base pair, allowing precise discrimination between deletions that are distinct but very close to each other in the genome. Such discrimination is critical for population genetic analyses (3, 19), but has not been pursued in most previous studies. We analyze the comprehensive set of precisely mapped LSPs in 100 clinical isolates to better understand the mechanisms of genomic deletions, their phenotypic effects, and their evolutionary consequences.

Materials and Methods

Selection of Bacterial Isolates. One hundred isolates were selected from a curated collection of epidemiologically well characterized isolates from San Francisco (20). A total of 2,498 cases of tuberculosis were reported in the city of San Francisco between January 1st, 1991, and December 31, 1999. Of these, 2,142 were culture-positive for *M. tuberculosis*. Isolates from 1,802 of the culture-positive cases were genotyped for IS6110 restriction fragment length polymorphism (RFLP) band pattern (21); in the event that an isolate had fewer than six bands, it was also genotyped for polymorphic GC-rich sequence (PGRS) RFLP pattern (22). We refer to isolates with identical IS6110 plus PGRS genotypes as members of the same "strain." Among the 1,802 genotyped isolates, 683 (38%) cases were due to isolates exhibiting a genotype found in at least one other isolate from San Francisco. We refer to these cases and the isolates that caused them as "clustered," because they were implicated in chains of transmission. The remaining 1,119 (62%) cases were caused by isolates that exhibited genotypes that were unique in the city of San Francisco. We refer to these isolates and the cases they were from as "unique" or "nonclustered." From the complete set of 1,802 genotyped isolates, we randomly selected 50 clustered and 50 unique strains. Our sample contains a slightly higher proportion of clustered isolates than the population at large. This bias reflects our interest in studying isolates that were implicated in chains of transmission, and for which we had clinical data from multiple cases.

Identification of Genomic Deletions. Genomic hybridization to the Affymetrix (Santa Clara, CA) *M. tuberculosis* GeneChip was used to identify genomic sequences present in the fully sequenced genome, H37Rv, but putatively missing from our set of 100 selected clinical isolates. Mycobacterial DNA was extracted, and 8 μg of sheared genomic DNA was hybridized (12). Intensity data were analyzed with DELSCAN software (AbaSci, San Pablo, CA). Default settings were used in all analyses. To confirm putative deletions and

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: LSPs, large sequence polymorphisms; SNPs, single nucleotide polymorphisms; PE-/PPE-PGRS, proline glutamic acid/proline proline glutamic acid-polymorphic GC-rich sequence; FDR, false discovery rate.

See Commentary on page 4721.

§To whom correspondence should be addressed. E-mail: petersmall@stanford.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0305634101

determine their genomic locations to the base pair, PCR and sequencing were performed as described (12). If a putative deletion's boundary fell inside of a highly repetitive proline glutamic acid (PE-)/proline proline glutamic acid (PPE)-PGRS sequence, the precise address could not be confirmed, because PCR and sequencing proved unreliable in these sequences. For reasons outlined in the introduction, our objective was to obtain a data set of precisely defined deletions. Therefore, putative deletions with boundaries in PE-/PPE-PGRS sequences were set aside for future analysis. Deletions that spanned, but did not terminate within, PE-/PPE-PGRS sequences were not excluded.

Statistical and Sequence Analyses. Statistical tests were done by using scripts written in R (www.r-project.org) and MATHEMATICA-4 (Wolfram Research, Champaign, IL). In the analysis of genomic aggregation of deletions, the distance, w, between two deletions was the minimum number of base pairs between the 5′ boundary of one deletion and the 3′ boundary of the other, or 0 if the two deletions overlapped. The significance of deviation from the distribution of distances expected under the null hypothesis was assessed by a randomization test. Let $n[w]$ be the number of the 2,278 (68 choose two) possible pairs of deletions in which the two deletions are within a distance w of each other. Let $n[w]_{obs}$ be the value of this statistic observed in our data set. In each iteration of our randomization procedure, the observed 68 deletions in the data set were assigned random addresses. In the event that a randomly assigned address resulted in one or both boundaries within a region annotated as PE, PPE, or PGRS (23), that random address was discarded, and a new one was drawn. This procedure ensures that our null distribution of pairwise distances reflects potential biases caused by the exclusion from our data set of deletions that terminated in repetitive PE-/PPE-PGRS regions. From the list of random addresses, $n[w]$ was recalculated. The proportion of randomizations in which $n[w] > n[w]_{obs}$ provided an estimate of the P value for rejection of the null hypothesis. One thousand iterations of the randomization procedure were performed at each distance, w. The expected number of proximate pairs $n[w]_{exp}$ was estimated by the average across iterations of the randomization procedure.

In the analysis of association between genomic and phylogenetic locations of deletions, the phylogeny shown in figure 3c of ref. 19 was used. For a genomic aggregation of two deletions, the probability of these deletions occurring in isolates as closely related as those observed was estimated as follows. If we condition on one of the two deletions, the probability of the second deletion occurring in an isolate as closely related as observed can be found simply by counting the number of isolates in the minimal monophyletic group defined by the two deletions and dividing by the total number of isolates (100). For genomic aggregations containing $n > 2$ deletions, $P_{divergence} = (m/100)^{n-1}$, where m is the number of isolates in the minimal monophyletic group defined by all n deletions.

To determine whether deletions that occur only in clustered isolates have significantly higher mean frequency than deletions that occur only in nonclustered isolates, the list of isolate clustering status was permuted 1,000 times. For each permutation, we calculated the difference between the mean frequency of deletions occurring only in clustered isolates and the mean frequency of deletions occurring only in nonclustered isolates. Significance was estimated as the proportion of permutations generating a larger mean difference than that observed. This procedure accounts for the nonindependence among deletions caused by the complete linkage of the genome.

In tests of association between gene functional family and deletion probability, false discovery rate (FDR) was controlled by using the method of ref. 24.

Results and Discussion

Genomic Deletions. DELSCAN identified a total of 314 putatively different deleted sequences in our sample of 100 isolates. Fifty-one of these putative deletions had one or both boundaries within highly repetitive PE-/PPE-PGRS sequences and were therefore excluded due to difficulties in precisely determining deletion addresses (see *Materials and Methods*). It is unfortunate that some deletions involving PE and PPE genes had to be set aside, because these genes may represent a major source of antigenic variability (25). However, precise definition of deletion boundaries was critical to many of the analyses performed here and in ref. 19, so thorough coverage of this interesting gene family was killed to obtain a data set of unambiguously defined deletions. Not all PE and PPE genes were dropped from the set of deletions, as *wag22*, *PE-PGRS35*, *PPE9*, *PPE38*, and *PPE39* were spanned by deletions that did not terminate in repetitive regions. In PCR and sequence analysis of the remaining putative deletions, 87 were not found to be real deletions and 108 were found to be identical to other deletions. This left 68 distinct deleted sequences, and their locations were mapped to the bp. These sequences have been named regions of difference (RD) with respect to H37Rv, in keeping with the nomenclature of Brosch *et al.* (3). The relationship of these deletions to those previously described, their precise lengths, the genes they disrupt, and the isolates from which they are deleted are shown in Table 2, which is published as supporting information on the PNAS web site. The boundary sequences flanking the deletions and the PCR primer sequences used for confirmation and sequence analysis are presented in Table 3, which is published as supporting information on the PNAS web site.

A total of 224 genes, comprising 5.5% of the genes annotated in H37Rv, the reference strain (23), were partially or completely deleted. Only one deletion (RD206, in the direct repeat region) did not disrupt a coding sequence. The size of deleted sequences varied from 105 to 11,985 bp (median 2,278 bp), with eight deleted sequences >5,000 bp (Table 2). The lower bound of this size distribution is set by the limits of the GeneChip's sensitivity; there are undoubtedly many shorter polymorphisms that were not detected (6). In total, the 68 sequences that were deleted from one or more of the 100 isolates comprise 4.2% (186,137 bp) of the H37Rv genome.

Fig. 1 shows the distribution of deletions across the genome and among isolates: rows are ordered according to the phylogenetic relationships among isolates (19); columns are ordered according to the genomic address of deletions. As can be seen in Fig. 1, the number of LSPs detected per isolate ranged from zero (in seven isolates) to seven. The number of genes deleted per isolate ranged from 0 to 50, with a median of 19. A total of 40,861 bp were deleted from one isolate, which was multidrug-resistant. This magnitude of genomic variability caused by LSPs is comparable to that found between *Mycobacterium bovis* and *M. tuberculosis* H37Rv (13).

Comparison between the level of LSPs observed here and that found in other microbes (8–16) is approximate because of important differences in the numbers of isolates sampled, the rates of spurious deletions identified by distinct experimental procedures, and the frequency spectra of LSPs in different bacteria. Nonetheless, the striking contrast between the 5.5% of genes deleted in a sample of 100 *M. tuberculosis* isolates and the 22% of genes deleted in a sample of 15 *H. pylori* isolates (10), as well as in a sample of 36 *Staphylococcus aureus* isolates (11), suggests that levels of LSPs are relatively low in *M. tuberculosis*.

However, given the potentially large phenotypic effect of deletions and the low levels of nonsynonymous SNPs in *M. tuberculosis*, LSPs are likely to represent a relatively important cause of phenotypic variability. Comprehensive comparison of coding sequences in *M. tuberculosis* strains H37Rv and CDC1551 showed that 457 of the genes contained a nonsynonymous SNP (6). In the present interrogation of 100 isolates for large deletions, the average number of genes present in H37Rv but missing in an interrogated isolate was 20, and the maximum number was 50 (Table 2). Assuming that approximately the same number of genes is missing in H37Rv but present in an interrogated isolate, the difference in gene content

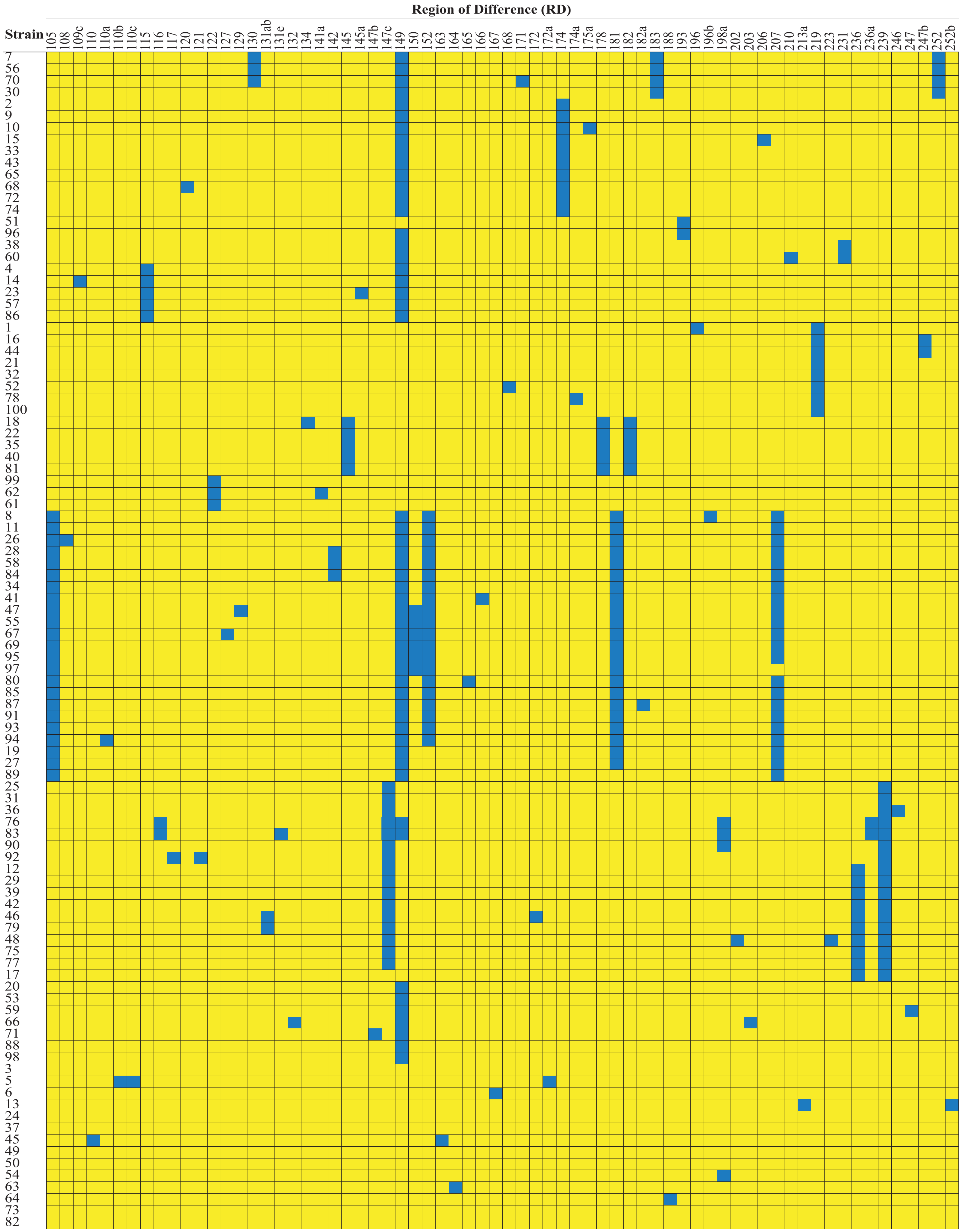

Fig. 1. Overall distribution of deleted sequences among 100 clinical isolates of *M. tuberculosis*. Sequences present in H37Rv and absent from the interrogated isolate are shown in blue. Each row represents an isolate, and each column is a region of difference. Columns are organized by genomic address; rows are organized according to phylogenetic relationships (19).

between two strains that are not closely related is likely to exceed 40 genes and may approach 100 genes. Given that the average phenotypic effect of a single amino acid substitution is likely to be substantially smaller than the average effect of deleting a gene altogether, the results presented here are consistent with the hypothesis that LSPs are a more important source of phenotypic variability than SNPs in *M. tuberculosis*.

Mechanisms of Deletion. Because sequence analysis was performed to discriminate between deletions that are very close to one another in the genome, the genomic and phylogenetic distributions of deletions could be analyzed to elucidate the mechanisms by which sequences are deleted. First, we investigated whether deletions are closer to each other in the genome than expected by chance. Statistical tests for nonrandomness in the distribution of deletions throughout the genome require definition of a null distribution that reflects potential deviations from randomness due merely to biases in the methods of detecting deletions. A potentially important source of bias in our method was the exclusion of putative deletions with a boundary inside a repetitive PE-/PPE-PGRS region. A null distribution was therefore defined to take this into account (see *Materials and Methods*). We used this biased null distribution to test whether more pairs of deletions are separated by $<w$ bp than would be expected under random placement of deletions on the genome. The probability, P, of observing a data set with more pairs of deletions separated by $<w$ bp was plotted against distance, w (Fig. 2).

For distances up to 16,000 bp, the probability that random placement of deletions would result in the observed number of proximate pairs is <0.001. Up to 46,000 bp, $P < 0.01$, and at 80,000 bp, the significant aggregation disappears: there are not significantly more pairs of deletions sharing a window of this size than expected by chance. Although Fig. 2 shows that deletion is a genomically local process, it does not allow us to define the distance of the effect precisely. Because the plot is cumulative, the significant aggregation at 60,000 bp is due in part to a highly significant effect at shorter distances. We therefore repeated the randomization procedure, but counted the number of pairs of deletions separated by a distance $>(w - 10{,}000)$ bp but $<w$ bp. This analysis revealed significant aggregation of deletions ($P < 0.05$) at distances of 0–5 kb and 10–15 kb, but not above ($P > 0.4$).

For several reasons, we believe that detection-related biases are very unlikely to be responsible for the pattern of tight aggregations we observe. First, the null distribution accounts for potential bias caused by our exclusion of deletions terminating in PE-/PPE-PGRS regions. Second, if the significant aggregation of deletions were due merely to increased sensitivity of the GeneChip in certain genomic regions, the frequency of LSPs in the *M. tuberculosis* genome would have to be phenomenally high. (In this scenario, aggregations of deletions would reveal the real rate of polymorphism, whereas stretches where deletions were not found would actually be genomic regions of low GeneChip sensitivity.) Such high rates of LSP would contradict results from whole genome comparison of CDC1551 with H37Rv (6). In fact, rates of LSP observed in the whole genome comparison are comparable with rates observed here. Third, the GeneChip's frequency of false negatives (failure to detect a deletion), conditional on a deletion's prior identification in another isolate, was low. Of 27 deletions with frequency >1 in the 100 isolates (Table 1), the GeneChip exhibited a false negative rate of 0 for all but one of the deletions. In the one remaining deletion (181) the GeneChip failed to detect almost half of the deletion's 22 occurrences. However, this deletion was found to be in an aggregation of deletions, the opposite of what one would expect if experimental bias were driving the observed patterns, and the other deletions in this genomic aggregation exhibited a false negative rate of 0.

The "hotspots" of deletion that produce the effect illustrated in Fig. 2 could be generated in two distinct ways. An event specific to a particular lineage of isolates, such as the insertion of a mobile element (26, 27), might render a genomic region vulnerable to deletion in all isolates belonging to that lineage. Alternatively, a genomic region might have properties that render it vulnerable in all members of the species. For example, selective neutrality of sequences in the region, or selective advantages associated with disruption of the region, would have such an effect. We can use the genomic and phylogenetic distributions of deletions to determine whether a given aggregation of deletions is likely to be due to a genetic event specific to a lineage, or is instead likely to reflect a species-wide property of the genomic region. If a lineage-specific event leads to an aggregation of deletions, those distinct but neighboring deletions are expected to occur in isolates that are closely related phylogenetically. By contrast, if a hotspot is caused by a species-wide property of the genomic region, local genomic deletions are not expected to occur in closely related isolates, but should instead be distributed randomly throughout the phylogeny of 100 isolates. It is important to emphasize here that we are referring to the phylogenetic distribution of deletions that have distinct but neighboring genomic addresses. We are not suggesting that hotspots should lead to homoplasy in the phylogeny; deletions generally behave as unique event polymorphisms, and therefore do not exhibit homoplasy (19).

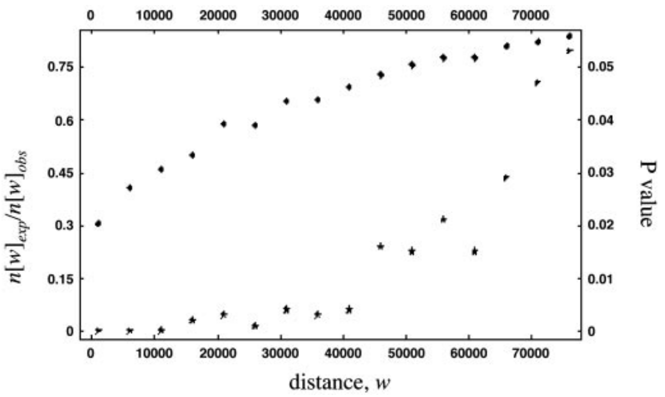

Fig. 2. Significance of deletion proximity. Stars indicate the probability that random placement on the genome of 68 deletions of the observed sizes results in at least the observed number of pairs of deletions separated by $<w$ bp. Diamonds indicate the ratio of expected to observed proximate pairs, $n[w]_{exp}/n[w]_{obs}$.

We used the unambiguous, maximum parsimony phylogeny for all 100 isolates (19) to investigate whether deletions that are aggregated in the genome occur in closely related isolates. As an approximate index of divergence among the isolates bearing the deletions belonging to a given aggregation, we estimated $P_{divergent}$, the probability that the deletions would occur by chance in a set of isolates equally close, or closer, on the phylogeny (Table 1; see *Materials and Methods*). $P_{divergent} = 1$ indicates that the deletions in a genomic aggregation occur in isolates that are not more closely related than expected by chance. Smaller $P_{divergent}$ values reflect a higher level of relatedness among the isolates bearing the deletions in a given aggregation. Thus, aggregations 110b–110c and 236–239 are very likely to be due to genetic events resulting in local genomic instability in specific lineages. Aggregation 236–239 is confined to the so-called "Manila" clade (28), and is likely to result from a genetic event that occurred in the ancestor of these isolates.

By contrast, in most other genomic aggregations of deletions, distinct but neighboring deletions occur in phylogenetically unrelated isolates. These aggregations of deletions are very likely to mark genomic regions that are vulnerable to deletion throughout the species. Functional annotation and epidemiological data that will be discussed below suggest that two of these aggregations contain genes whose deletion may be positively selected. Aggregation 163–168 contains *katG* as well as *furA*, which regulates *katG*. Deletion of these genes confers isoniazid resistance (29). Aggregation 171–175a contains genes that are part of a regulon induced

Table 1. Phylogenetic distribution of deletions that are aggregated in the genome

Aggregation	*P* (divergence)
105,108	0.23
110b,110c	0.01
115,116	1
121,122	1
131ab,131e	0.17
145,145a	1
163,164,165,166,167,168	1
171,172,172a,174,174a,175a	1
181,182,182a	1
196,196b	1
202, 203, 206, 207, 210	1
236, 236a, 239	0.0289
246, 247, 247b	1
252, 252b	1

Each row contains a genomic aggregation of deletions. *P* is an estimate of the probability of the deletions in the aggregation belonging, by chance, to a group as closely related as that observed. Thus, $P = 1$ indicates that the isolates bearing the deletions in a given genomic aggregation are not more related than expected by chance, whereas smaller *P* values indicate a greater degree of relatedness among isolates bearing the deletions in an aggregation.

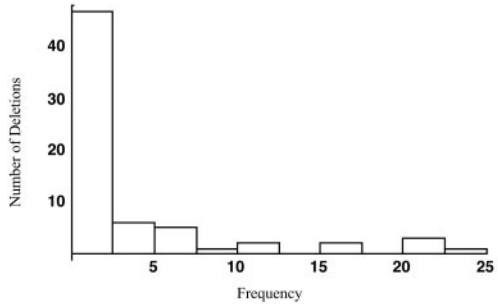

Fig. 3. The frequency spectrum of genomic deletions. Deletion frequency (occurrences out of 100 isolates) is shown on the *x* axis. The number of deletions exhibiting each frequency is shown on the *y* axis. Phage-associated deletion 149 is not shown; it was present in 54 of 100 isolates.

by hypoxia, a physiological response that is important in the latent state of the pathogen (30). Deletion of these "latency genes" may promote rapid progression to disease and transmission. Indeed, as we discuss below, several of the deletions in this aggregation may be associated with chains of transmission. Interestingly, the genes *ctpG* and *Rv1993c* are not known to be part of the hypoxia regulon, but they are part of aggregation 171–175a, and they are involved in three independent deletion events (172, 172a, and 174). It would seem that either there is a very high rate of deletion in the vicinity of these genes, or their loss is positively selected. Finally, aggregation 202–210 contains the direct repeat region and many hypothetical proteins. This genomic region may be unstable and the deletions in aggregation 202–210 may be selectively neutral.

Effects of Genomic Deletions on Microbial Fitness. The extent to which genetic variation among strains of *M. tuberculosis* contributes to heterogeneity in their epidemiological and clinical behavior has been the subject of extensive discussion (31–34). We first address this question at the most general level, investigating whether deletions have consistent and detectable effects on epidemiology. We then examine the question in terms of functional categories of genes, and finally at the level of individual genomic deletions and their phenotypic consequences.

To address the question generally, we statistically investigate whether deletions are consistently associated with certain phenotypes. If this is the case, then a deletion that is found in an isolate that is clustered is more likely to be common in the population, whereas a deletion that is found in an isolate that is unique is more likely to be rare. Note that no two isolates in our sample of size 100 are associated with the same cluster, so a single isolate's status as clustered does not directly affect the frequency of that isolate's deletions in our sample. We therefore use the frequencies of deletions in our sample as estimates of their population frequencies without introducing an obvious association between isolate clustering and deletion frequency.

Comparison of the frequencies of deletions found in clustered isolates to the frequencies of deletions that are not found in clustered isolates is statistically biased: frequent deletions tend to be in clustered and unique isolates alike, so deletions not found in clustered isolates are more likely to be rare. To avoid this bias, we test whether deletions found only in clustered isolates have a higher frequency than deletions found only in unique isolates. As expected under the hypothesis that some deletions are consistently associated with certain phenotypes, deletions found only in clustered isolates are on average more frequent than deletions found only in unique isolates. However, the difference is only marginally significant. (Mean frequency of 26 deletions found only in nonclustered isolates = 0.0108; mean frequency of 19 deletions found only in clustered isolates = 0.0142; $P = 0.03$.) This finding suggests that the deletions we observe may have effects on isolate phenotype, and thus on disease epidemiology, but the effects are not strong. In short, many deletions may be nearly neutral.

Are most observed deletions advantageous or deleterious? The shape of the frequency spectrum of deletions (Fig. 3), and particularly the abundance of singletons, suggests that the average effect of a deletion is slightly deleterious. Under mutation–selection balance and additive or multiplicative effects of multiple mutations, the proportion of individuals, n, who carry m mutations is given by $n(m, u, s) = (u/s)^m(1/m!)e^{-u/m}$, where u is the mutation rate per genome per generation and s is the effect of each mutation on individual fitness. The best fit of this distribution to our observed frequency spectrum of deletions is given by $s/u = 0.55$, a per-deletion deleterious selective effect that is about half the rate of occurrence of new deletions per genome per generation. However, without a frequency spectrum for other, putatively neutral polymorphisms in the *M. tuberculosis* genome, we cannot control for other processes that would tend to skew the frequency spectrum toward rare deletions. Rapid population expansion and sweeps by strongly favored mutants are also likely to contribute to the relative abundance of rare polymorphisms.

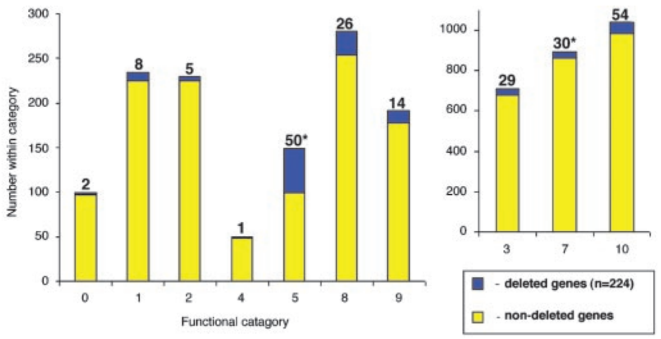

Fig. 4. Distribution of deleted genes by functional category. Number of deleted genes (blue) and nondeleted genes (yellow) are shown for each category. An asterisk indicates that a functional category was statistically overrepresented among deletions after controlling for FDR. Functional categories: 0, virulence, detoxification, adaptation; 1, lipid metabolism; 2, information pathways; 3, cell wall and cell processes; 4, stable RNAs; 5, insertion sequences and phages; 6, PE/PPE (not studied); 7, intermediary metabolism and respiration; 8, unknown; 9, regulatory proteins; 10, conserved hypotheticals.

Functional Effects of Specific Deletions. Although a large proportion of deletions appear to be deleterious, deletions have a wide range of effects, and some may be associated with an increased probability of transmission. Every major functional category (23) is represented among the 224 deleted genes, but certain functional categories are represented disproportionately (Fig. 4). Genes of mobile genetic elements are deleted more frequently than expected by chance (50 deleted out of a total of 149 annotated in the H37Rv genome; Fisher's exact test, $P = 0.00018$; controlling for FDR, $P < 0.005$). This is likely to reflect a combination of distinctive mutational processes and the neutral or slightly advantageous selective effects of deleting these genes.

The high rate of deletion of genes involved in intermediary metabolism and respiration is more surprising (30 deleted out of 895; Fisher's exact test, $P = 0.00089$; controlling for FDR, $P < 0.01$). One possible explanation for the deletion of some of these genes, as well as a number of the cell wall and cell process genes, is that selective pressures of the host immune system favor elimination of potential antigens. For example, the intermediary metabolism genes *plcA* (on deletion 147c) and *plcD* (on deletion 152) are known to encode important antigens (35). Similarly, a number of deleted cell wall genes, such as the lipoprotein genes *lpqS* (deletion 132), *lprP* (deletion 134), *lppC* (deletion 166), *lppA* (deletion 196), *lppB* (deletion 196 b), and *lpqH* (deletion 246), are likely to encode antigenic proteins (see for example refs. 36 and 37). Deletion of such genes might confer a selective advantage during certain stages of infection or transmission. However, if deletion of a gene is not a viable strategy in the long term, if, for instance, the gene is intermittently essential, then the gene could be maintained in the population despite sporadic, positively selected deletion.

Several other deletions have noteworthy, specific effects on isolate phenotype. Deletion 166 completely deletes the *katG* gene, which confers isoniazid resistance (29). A number of genes involved in the hypoxia-induced regulon (30) are deleted (Rv1996 and *ctpF* on deletion 174; *nrdZ*, Rv0571c, and Rv0572 on deletion 121). Interestingly, 9 of 11 isolates exhibiting these deletions were clustered. It is conceivable that disruption of the hypoxia-induced regulon hinders latency, making active disease and transmission more likely, but a focused study with larger sample size will be required to further investigate this hypothesis.

Concluding Remarks. A previously conducted comprehensive evaluation of 20 variable genomic regions among members of the *M. tuberculosis* complex provided fundamental insight into the evolutionary history of this group of organisms (3). Here we have applied comparative genomics to investigate the mechanisms and consequences of genomic deletions in the most clinically significant member of the complex, *M. tuberculosis*.

Deletions were found to be tightly aggregated in the genome, and analysis of the genomic distribution of deletions in the context of phylogenetic relationships among isolates suggested that there are two distinct causes behind this pattern. Some aggregations are caused by a genetic event specific to a single mycobacterial lineage, whereas others reveal regions of genomic vulnerability throughout the species. Taken together, our analyses suggest that although the majority of polymorphic deletions are slightly deleterious to the pathogen, there may be intriguing exceptions. Some deletions are likely to offer short-term advantages of escape from the host immune system. Others reduce the microbe's load of mobile genetic elements. Still others confer strong advantages, such as antibiotic resistance, or curtail latency and thus promote transmission.

We acknowledge Dmitri Petrov and Marc Feldman for thoughtful comments and discussions about this work, Hugh Salamon for the use of DELSCAN software, Melvin Javonillo and Neil Abernethy for assistance with database management and programming, and Jorg Drenkow and Tom Gingeras (Affymetrix, Santa Clara, CA) for material and informatic support. A.E.H. was supported by National Institutes of Health Grants GM 28016 and 28424 (to M. W. Feldman). K.D. was supported by National Institutes of Health Grant K01 TW000001. M.H. was supported by Wellcome Trust Grant 051034. J.A.E. was supported by National Institutes of Health Grant D43TW01135 and by Sistema Nacional de Investigadores–Consejo Nacional de Ciencia y Tecnologia. This study was funded by National Institute of Allergy and Infectious Diseases Grant AI 34238 (to P.M.S.).

1. da Silva, A. C., Ferro, J. A., Reinach, F. C., Farah, C. S., Furlan, L. R., Quaggio, R. B., Monteiro-Vitorello, C. B., Van Sluys, M. A., Almeida, N. F., Alves, L. M., *et al.* (2002) *Nature* **417,** 459–463.
2. Alm, R. A., Ling, L. S., Moir, D. T., King, B. L., Brown, E. D., Doig, P. C., Smith, D. R., Noonan, B., Guild, B. C., deJonge, B. L., *et al.* (1999) *Nature* **397,** 176–180.
3. Brosch, R., Gordon, S. V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., Garnier, T., Gutierrez, C., Hewinson, G., Kremer, K., *et al.* (2002) *Proc. Natl. Acad. Sci. USA* **99,** 3684–3689.
4. Perrin, A., Bonacorsi, S., Carbonnelle, E., Talibi, D., Dessen, P., Nassif, X. & Tinsley, C. (2002) *Infect. Immun.* **70,** 7063–7072.
5. Makino, K., Oshima, K., Kurokawa, K., Yokoyama, K., Uda, T., Tagomori, K., Iijima, Y., Najima, M., Nakano, M., Yamashita, A., *et al.* (2003) *Lancet* **361,** 743–749.
6. Fleischmann, R. D., Alland, D., Eisen, J. A., Carpenter, L., White, O., Peterson, J., DeBoy, R., Dodson, R., Gwinn, M., Haft, D., *et al.* (2002) *J. Bacteriol.* **184,** 5479–5490.
7. Gutacker, M. M., Smoot, J. C., Migliaccio, C. A., Ricklefs, S. M., Hua, S., Cousins, D. V., Graviss, E. A., Shashkina, E., Kreiswirth, B. N. & Musser, J. M. (2002) *Genetics* **162,** 1533–1543.
8. Chan, K., Baker, S., Kim, C. C., Detweiler, C. S., Dougan, G. & Falkow, S. (2003) *J. Bacteriol.* **185,** 553–563.
9. Anjum, M. F., Lucchini, S., Thompson, A., Hinton, J. C. & Woodward, M. J. (2003) *Infect. Immun.* **71,** 4674–4683.
10. Salama, N., Guillemin, K., McDaniel, T. K., Sherlock, G., Tompkins, L. & Falkow, S. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 14668–14673.
11. Fitzgerald, J. R., Sturdevant, D. E., Mackie, S. M., Gill, S. R. & Musser, J. M. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 8821–8826.
12. Kato-Maeda, M., Rhee, J. T., Gingeras, T. R., Salamon, H., Drenkow, J., Smittipat, N. & Small, P. M. (2001) *Genome Res.* **11,** 547–554.
13. Behr, M. A., Wilson, M. A., Gill, W. P., Salamon, H., Schoolnik, G. K., Rane, S. & Small, P. M. (1999) *Science* **284,** 1520–1523.
14. Dziejman, M., Balon, E., Boyd, D., Fraser, C. M., Heidelberg, J. F. & Mekalanos, J. J. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 1556–1561.
15. Porwollik, S., Wong, R. M. & McClelland, M. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 8956–8961.
16. Wolfgang, M. C., Kulasekara, B. R., Liang, X., Boyd, D., Wu, K., Yang, Q., Miyada, C. G. & Lory, S. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 8484–8489.
17. Musser, J. M., Amin, A. & Ramaswamy, S. (2000) *Genetics* **155,** 7–16.
18. Hughes, A. L., Friedman, R. & Murray, M. (2002) *Emerg. Infect. Dis.* **8,** 1342–1346.
19. Hirsh, A. E., Tsolaki, A. G., DeRiemer, K., Feldman, M. W. & Small, P. M. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 4871–4876.
20. Jasmer, R. M., Hahn, J. A., Small, P. M., Daley, C. L., Behr, M. A., Moss, A. R., Creasman, J. M., Schecter, G. F., Paz, E. A. & Hopewell, P. C. (1999) *Ann. Intern. Med.* **130,** 971–978.
21. van Embden, J. D., Cave, M. D., Crawford, J. T., Dale, J. W., Eisenach, K. D., Gicquel, B., Hermans, P., Martin, C., McAdam, R., Shinnick, T. M., *et al.* (1993) *J. Clin. Microbiol.* **31,** 406–409.
22. Chaves, F., Yang, Z., el Hajj, H., Alonso, M., Burman, W. J., Eisenach, K. D., Dronda, F., Bates, J. H. & Cave, M. D. (1996) *J. Clin. Microbiol.* **34,** 1118–1123.
23. Camus, J. C., Pryor, M. J., Medigue, C. & Cole, S. T. (2002) *Microbiology* **148,** 2967–2973.
24. Benjamini, Y. & Liu, W. (2003) *A Distribution-Free Multiple Test Procedure That Controls the False Discovery Rate* (Department of Statistics, Tel Aviv Univ., Tel Aviv), RP-SOR-99-3.
25. Banu, S., Honore, N., Saint-Joanis, B., Philpott, D., Prevost, M. C. & Cole, S. T. (2002) *Mol. Microbiol.* **44,** 9–19.
26. Fang, Z., Doig, C., Kenna, D. T., Smittipat, N., Palittapongarnpim, P., Watt, B. & Forbes, K. J. (1999) *J. Bacteriol.* **181,** 1014–1020.
27. Ho, T. B., Robertson, B. D., Taylor, G. M., Shaw, R. J. & Young, D. B. (2000) *Yeast* **17,** 272–282.
28. Douglas, J. T., Qian, L., Montoya, J. C., Musser, J. M., Van Embden, J. D., Van Soolingen, D. & Kremer, K. (2003) *J. Clin. Microbiol.* **41,** 2723–2726.
29. Heym, B., Alzari, P. M., Honore, N. & Cole, S. T. (1995) *Mol. Microbiol.* **15,** 235–245.
30. Sherman, D. R., Voskuil, M., Schnappinger, D., Liao, R., Harrell, M. I. & Schoolnik, G. K. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 7534–7539.
31. Kato-Maeda, M., Bifani, P. J., Kreiswirth, B. N. & Small, P. M. (2001) *J. Clin. Invest.* **107,** 533–537.
32. Rhee, J. T., Piatek, A. S., Small, P. M., Harris, L. M., Chaparro, S. V., Kramer, F. R. & Alland, D. (1999) *J. Clin. Microbiol.* **37,** 1764–1770.
33. Bifani, P. J., Mathema, B., Liu, Z., Moghazeh, S. L., Shopsin, B., Tempalski, B., Driscol, J., Frothingham, R., Musser, J. M., Alcabes, P. & Kreiswirth, B. N. (1999) *J. Am. Med. Assoc.* **282,** 2321–2327.
34. Murray, M. & Nardell, E. (2002) *Bull. W. H. O.* **80,** 477–482.
35. Falla, J. C., Parra, C. A., Mendoza, M., Franco, L. C., Guzman, F., Forero, J., Orozco, O. & Patarroyo, M. E. (1991) *Infect. Immun.* **59,** 2265–2273.
36. Hovav, A. H., Mullerad, J., Davidovitch, L., Fishman, Y., Bigi, F., Cataldi, A. & Bercovier, H. (2003) *Infect. Immun.* **71,** 3146–3154.
37. Neufert, C., Pai, R. K., Noss, E. H., Berger, M., Boom, W. H. & Harding, C. V. (2001) *J. Immunol.* **167,** 1542–1549.