Detection of Deleted Genomic DNA Using a Semiautomated Computational Analysis of GeneChip Data

Hugh Salamon,[1–3] Midori Kato-Maeda,[1] Peter M. Small,[1] Jorg Drenkow,[2] and Thomas R. Gingeras[2]

[1]*Division of Infectious Diseases and Geographic Medicine, Department of Medicine, Stanford University, Stanford, California 94305, USA;* [2]*Affymetrix, Inc., Santa Clara, California 95051, USA*

Genomic diversity within and between populations is caused by single nucleotide mutations, changes in repetitive DNA systems, recombination mechanisms, and insertion and deletion events. The contribution of these sources to diversity, whether purely genetic or of phenotypic consequence, can only be investigated if we have the means to quantitate and characterize diversity in many samples. With the advent of complete sequence characterization of representative genomes of different species, the possibility of developing protocols to screen for genetic polymorphism across entire genomes is actively being pursued. The large numbers of measurements such approaches yield demand that we pay careful attention to the numerical analysis of data. In this paper we present a novel application of an Affymetrix GeneChip to perform genome-wide screens for deletion polymorphism. A high-density oligonucleotide array formatted for mRNA expression and targeted at a fully sequenced 4.4-million–base pair *Mycobacterium tuberculosis* standard strain genome was adapted to compare genomic DNA. Hybridization intensities to 111,000 probe pairs (perfect complement and mismatch complement) were measured for genomic DNA from a clinical strain and from a vaccine organism. Because individual probe-pair hybridization intensities exhibit limited sensitivity/specificity characteristics to detect deletions, data-analytical methodology to exploit measurements from multiple probes in tandem locations across the genome was developed. The TSTEP (Tandem Set Terminal Extreme Probability) algorithm designed specifically to analyze the tandem hybridization measurements data was applied and shown to discover genomic deletions with high sensitivity. The TSTEP algorithm provides a foundation for similar efforts to characterize deletions in many hybridization measures in similar-sized and larger genomes. Issues relating to the design of genome content screening experiments and the implications of these methods for studying population genomics and the evolution of genomes are discussed.

Genetic diversity among isolates of *Mycobacterium tuberculosis* may in part be caused by genetic deletions (Mahairas et al. 1996; Brosch et al. 1998; Behr et al. 1999). Genomic content polymorphism in related vaccine organisms has been investigated using parallel hybridization techniques using glass-slide fluorescent arrays (Behr et al. 1999).

In general, DNA hybridization arrays and chips permit rapid, parallel queries for the presence of thousands of sequence patterns in a sample. In most uses of this technology, levels of mRNA species are detected, providing a profile of gene transcripts expressed in populations of cells. The very same technology can be used to detect genetic polymorphism. The use of oligonucleotide arrays for identification of deleted ORFs (Open Reading Frames) in yeast has been investigated previously (Winzeler et al. 1999). Here, we use high-density oligonucleotide arrays to query sample genomes for >100,000 sequence patterns in the *M. tuberculosis* genome. The array used in this study was designed for gene transcript expression profiling. Thus, we both show the usability of such arrays for deletion finding and also discuss possibilities for improving the design of arrays for the purpose of deletion detection.

As individual oligonucleotide probe-pair queries are not reliable enough to call the presence or absence of DNA in the sample, multiple probe-pair results are useful when analyzed together. This is the basis for Affymetrix GeneChip design and use for gene expression profiling (Lipshutz et al. 1999). In this study we exploited multiple probe-pair results for detection of deleted genomic DNA. However, we wish to do more than simply identify that a deletion is very likely to exist in a given region of the genome; we wish to identify the boundaries of the deletion to the extent that the density of the probes permits. We seek a general solution to deletion-finding in probe-hybridization in-

[2]Present address: Department of Immunology, Berlex Biosciences, 15049 San Pablo Avenue, Richmond, CA 94804, USA
[3]Corresponding author.
E-MAIL Hugh_Salamon@Berlex.com; FAX (510) 669-4244.
Article published online before print: *Genome Res.,* 10.1101/gr.152900. Article and publication are at www.genome.org/cgi/doi/10.1101/gr.152900.

10:2044–2054 ©2000 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/00 $5.00; www.genome.org

tensity data, which is robust to variation in the hybridization level from experiment to experiment; we wish not to train on one experiment and assume that another experiment will show the same absolute intensity or intensity-ratio levels for present and absent DNA. The need for a method that defines the set of probes that exhibit hybridization patterns consistent with deleted DNA, with specific attention directed to their order in the genome, detection of the ends of deletions, and the desire to avoid requiring training sets of experiments, prompts us to develop a novel approach to data analysis of this probe-hybridization data.

RESULTS

For our negative control, the H37Rv genome from which the GeneChip was designed, a single region was assembled into a deletion call: 2237087–2237488, queried by 17 probe pairs, three with *P* values above our cutoff. This region was filtered out from subsequent screens of CDC1551 and BCG Pasteur; the region showed low intensity ratios in these screens as well, and thus low *P* values from TSTEP (Tandem Set Terminal Extreme Probability).

Figure 1a shows the ratio of the perfect match probe intensity to the corresponding mismatch probe intensity (I_{PM} / I_{MM}) in the screen of BCG Pasteur for probe pairs mapping to the region of the H37Rv genome from base pairs 1280000–1410000. By visual inspection alone, it is difficult to discriminate any regional distributions that are clear outliers from the overall variation in ratio values. In Figure 1b, the same probe pairs are shown, now plotting the *P* value calculated in the application of TSTEP. Immediately it is apparent that there are two regions of reasonable length that exhibit *P* values easily distinguished from the background values. These regions are even more easily distinguished visually when we plot the corrected *P* values (Fig. 1c). The black bars show the sequence-confirmed deletion regions, whereas the orange bars show the extent of the deletions as called by the assembly procedure. The short region around base pair 1,360,000 with low *P* values was not called deleted; the low *P* value probe pairs consist of a minority of measurements in a 100 bp region. Deletion assembly is discussed below.

Figure 2 shows that large deletions should be straightforward to discover; there is clearly enough information in the hybridization signals to discover longer deletions, such as this 12,734 bp deletion. The beginning edge of the deletion is missed by nearly 200 bp, reporting the start too early; four probe pairs before the true beginning of the deletion were included in the assembly. Only the first one was below the cutoff *P* value, the subsequent three being "skipped" in the assembly. At the far end of the deletion, there is an overestimate by only 26 bp; a single probe pair not overlapping the true deletion was included in the assembly.

Table 1 describes deletions called for the CDC1551 screen and shows that our application of TSTEP with the assembly procedure was sensitive to finding deletions >350 bp; even deletions in regions with large portions consisting of repetitive sequence families were discovered. In this screen, the procedure resulted in no false deletion calls; the hybridization and computational deletion finding is quite specific.

As the CDC1551 screen was used to some extent to tune the assembly parameters, it is not an entirely fair test. Table 2 shows the results for the BCG Pasteur strain screen. Here, all 14 previously identified deletions were discovered, along with three new deletions confirmed by PCR and sequencing. Three deletion calls could not be confirmed. On follow-up investigation, these false deletions appear to be a result of violating one of our assumptions implicit in the specific function F_p that we used: Our function assumes that tandem $z(j)$ values are independent. In the regions of the false calls, stacking of the probe-pair genomic regions (PPGRs) is evident; the probe pairs query strongly overlapping regions (Figure 3a).

DISCUSSION

Deletion polymorphism, that is, polymorphism in the genomic content among individuals, is probably of consequence to phenotypic diversity in many organisms. To measure genomic material missing from individual sampled organisms, a laboratory protocol and computational approach are developed in this article, permitting the use of GeneChip data to detect deletions larger than a few hundred base pairs in length. This groundwork opens the door to characterizing deletion polymorphism for multiple samples from populations.

Note that only a subset of deletion polymorphisms is being investigated: Those that appear as deletions relative to the laboratory strain H37Rv. Any genomic regions present in the queried strain, but absent from H37Rv, will not be discovered by the procedure described here. (However, experiments implementing the GeneChip may be designed to detect insertions interrupting a PPGR.) This issue has implications in the design of genomic content screening technology. The design of deletion-detection probes implicitly defines our understanding of what is important to measure and leads us to consider evolutionary issues. The phenomenon that the genome of a single organism is smaller than the genome of an entire population is likely to be especially pronounced in microbial populations; no individual is likely to have homologous genetic material with every portion of every other individual organism in a population. For simplicity of discussion, we define the ratio G_{IP}, the ratio of an

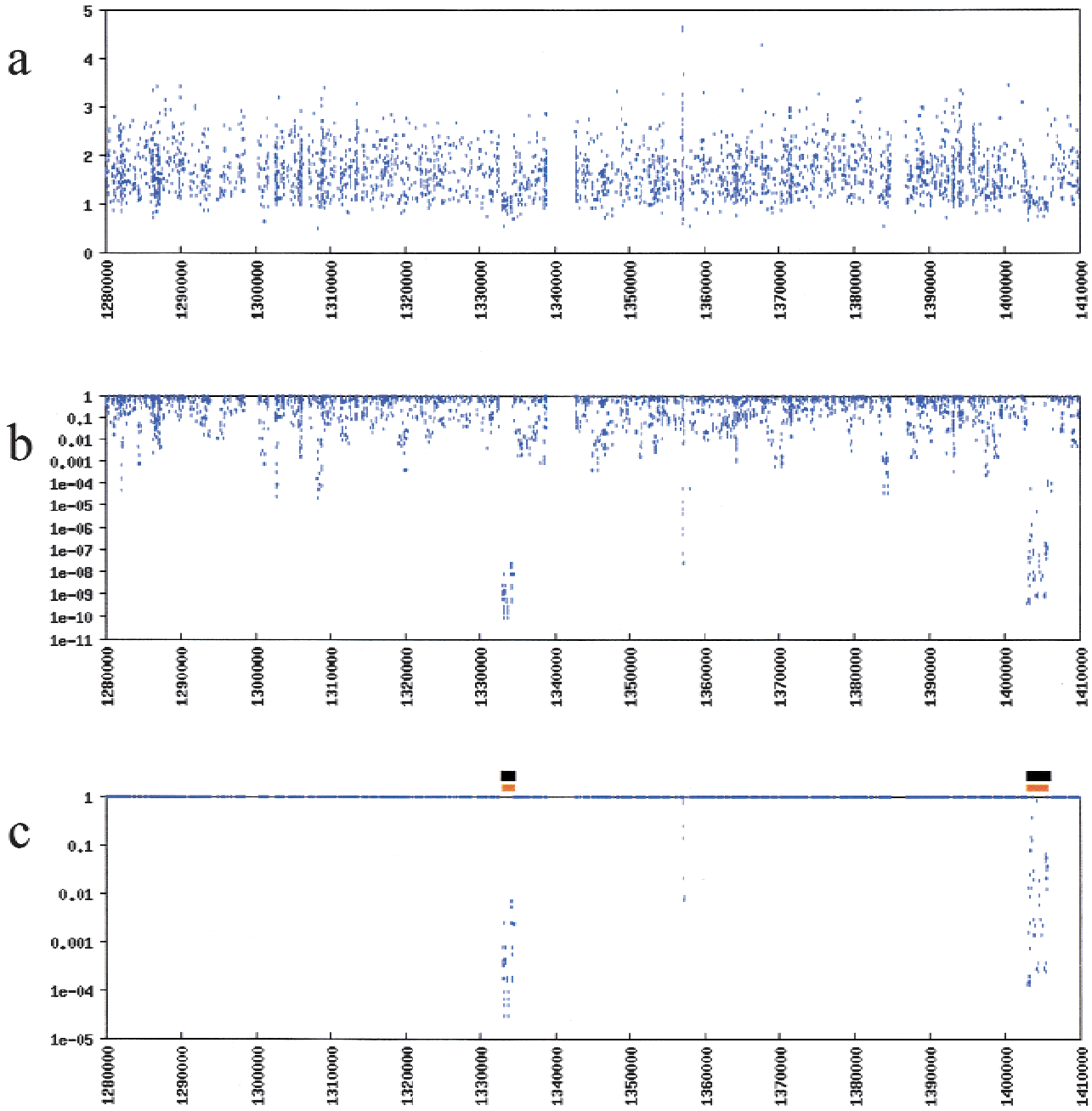

Figure 1 (*a*) Individual ratios of I_{PM} to I_{MM} (y-axis) plotted against genomic address (x-axis) reveal noise that masks deletions, hindering discovery. (*b*) The problem of identifying deleted regions is dramatically facilitated by investigating sets of tandem values using the TSTEP algorithm, which yields *P* values (y-axis) for each ratio. The x-axis is the same portion of the genome as in *a*. (*c*) Corrected *P* values, which account for testing more than 111,000 probe-pair hybridization ratios, are plotted against genomic address. Black bars indicate the regions of sequence-confirmed deletion. Orange bars indicate the regions predicted by a heuristic to assemble putative deletion intervals from *P* values calculated by TSTEP.

individual's genome size to the sum of nonredundant genetic material in a population. First, the distribution of G_{IP} ratios in a population will determine how much data collection is required to design a deletion-detection array of a predefined sensitivity. Second, estimates of this ratio itself may be obtained with the technology we describe. We speculate that genomic content screens may provide researchers with indicators of the evolutionary stability of a genome. For example, an organism such as *M. tuberculosis*, which may have recently moved into the environment of infecting human beings, may exhibit a low G_{IP} if individual

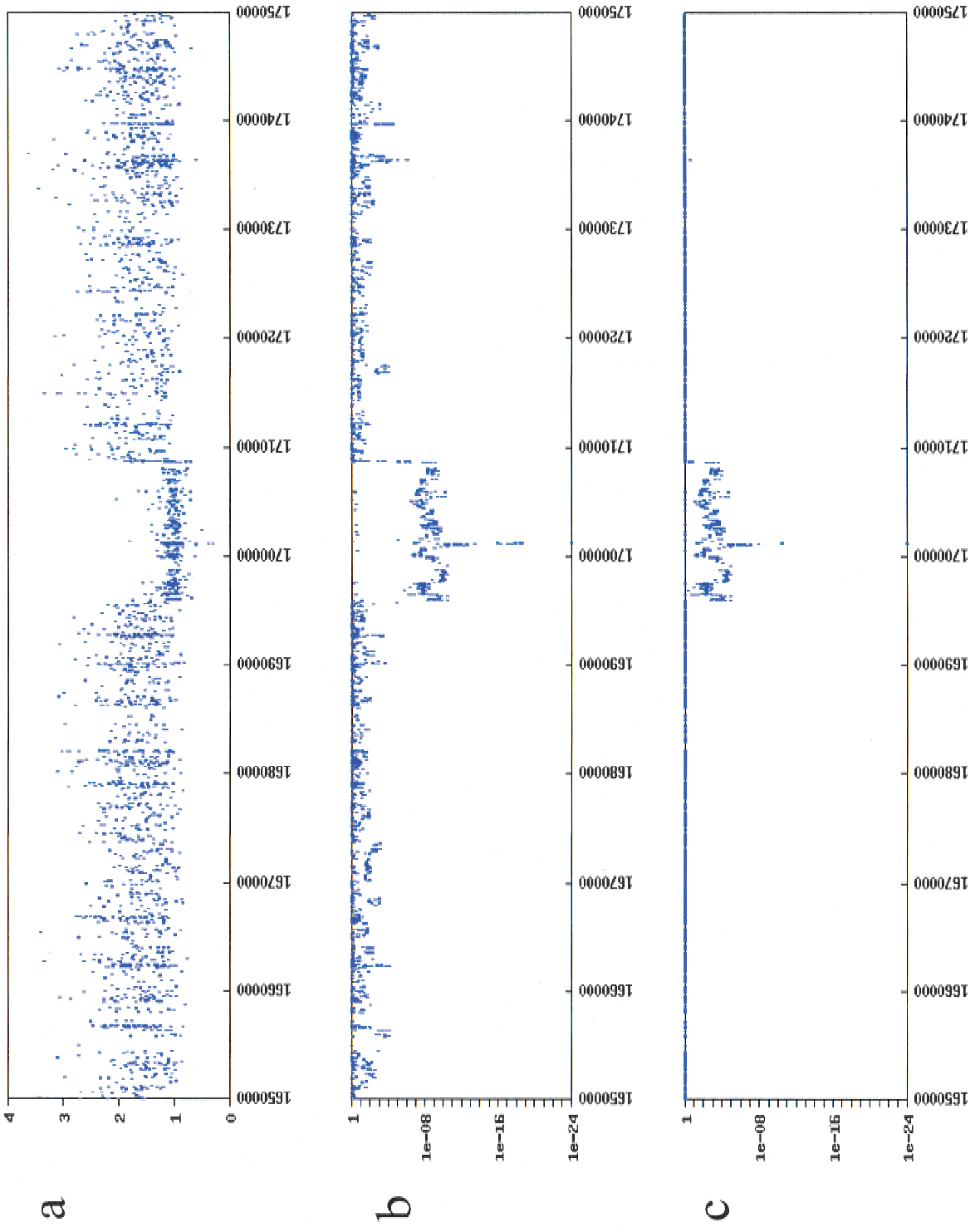

Figure 2 Longer deletions are not a great challenge to identify: Plotted against the genomic address (x-axis) the y-axis values are (*a*) I_{PM}/I_{MM}, (*b*) TSTEP-calculated *P* values, (*c*) *P* values corrected for multiple tests. The interval of the deletion is closely predicated by our assembly to be 1695829–1708776 (length 12948). The sequence confirmed interval is 1696017–1708750 (length 12734).

Table 1. Deletion Screen of CDC1551 Shows that TSTEP/Deletion Assembly Finds All Deletions >350 bp in Nonrepetitive Regions, with No False Deletion Cells

H37Rv regions of at least 300 bp with no identical match in CDC1551 (to the nearest 60 bp)			Repetitive sequence family in probe set annotation	Similarity with any region by blast to CDC1551 (%)	Edges detectable by blast to CDC1551 (corresponding address in H37Rv is given)			Deletion in nonrepetitive region, with detectable edges >350 bp	Complete TSTEP/ deletion assembly results for screen of CDC1551		
Start	Stop	Length			Start	Stop	Length		Start	Stop	Length
333540	336300	2760	PE/PGRS	99.3				No			
336360	339000	2640	PE/PGRS	97.1				No			
835920	841080	5160	PE/PGRS	99.6				No			
886500	887460	960			886541	887414	873	Yes	886427	887553	1127
1217460	1218180	720	PE/PGRS	72.9[a]				No			
2045040	2045400	360	PE/PGRS	95.0				No			
2060280	2060700	420		87.4				No			
2339220	2339640	420			2339258	2339503	245	No	[b]		
2381400	2383740	2340	Overlaps PPE[c]		2381413	2383683	2270	Yes	2380893	2383514	2622
2704260	2704920	660			2704308	2704805	497	Yes	2704242	2704695	454
3054660	3054960	300	PE/PGRS		3054720	3054906	186	No			
3122580	3123300	720			3122828	3123145	517	Yes	3122606	3123248	643
3501300	3501720	420	PPE		3501331	3501662	331	No			
3730860	3732180	1320	PPE		3730999	3732075	1076	No			
3732600	3735720	3120	PPE		3732653	3735632	2979	No			
3842280	3846600	4320	Overlaps PPE[d]		3842305	3846512	4207	Yes	3842882	3846825	3944
3894600	3894900	300	PPE	91.3				No			
3895020	3895320	300	PPE	86.4				No			
3933480	3934020	540	PE/PGRS	96.7				No			
3934080	3936660	2580	PE/PGRS	89.9				No			
3939720	3945120	5400	PE/PGRS	99.0				No			
3946860	3950340	3480	PE/PGRS	69.2				No			
3955440	3956100	660			3955464	3956100	636	Yes	3955252	3956433	1182

(TSTEP) Tandem Set Terminal Extreme Probability.
[a]More than one region with 60% or more identity is found in this interval. A weighted average is reported.
[b]There are only three probe pairs querying the 245-bp interval. A weak signal of P value <0.01 is measured for these probe pairs.
[c]PPE locus at 2381069–2382490.
[d]PPE loci at 3842235–3842765 and 3843032–3843730.

Table 2. Deletion Screen of BCG Pasteur Shows that TSTEP/Deletion Assembly Finds Known Deletions plus Three New Deletions

Complete TSTEP/deletion assembly results for screen of BCG Pasteur						Sequencing confirmation results			
Start	End	Length	No. of probe pairs	Proportion of probe pairs above the cutoff *p* value	RD number[a]	Confirmation result	Start	End	Length
264681	266633	1953	35	0.03	RD4	Deletion	264752	266658	1907
1058107	1058601	495	16	0.19		Present			
1333051	1334446	1396	29	0.03	N-RD18	Deletion	1332920	1334466	1547
1402897	1405631	2735	51	0.08	RD10	Deletion	1402932	1405939	3008
1483164	1483836	673	13	0.15		Present			
1695829	1708776	12948	430	0.04	RD6	Deletion	1696017	1708750	12734
1779216	1789408	10193	380	0.03	RD3	Deletion	1779276	1788525	9250
1998716	2007316	8601	210	0.02	RD14	Deletion	1998225	2007297	9073
2196766	2197597	832	23	0.09		Present			
2208107	2231883	23777	705	0.06	RD2, RD15	Deletion	2208003	2231848	23846
2329435	2332234	2800	58	0.05	RD12	Deletion	2330073	2332104	2032
2626662	2637681	11020	271	0.04	RD7	Deletion	2626070	2635032	8963
2969694	2980802	11109	310	0.07	RD13	Deletion	2969988	2981195	11208
3485154	3487575	2422	98	0.05	RD5	Deletion	3484737	3487512	2776
3842882	3846520	3639	119	0.05	RD11	Deletion	3842653	3847540	4888
3896832	3897792	961	27	0.11	N-RD17	Deletion	3897069	3897783	715
4056690	4062741	6052	108	0.03	RD9	Deletion	4056837	4062732	5896
4188975	4190798	1824	29	0.17	N-RD25	Deletion	4189605	4190757	1153
4349983	4359538	9556	127	0.03	RD1	Deletion	4350262	4359720	9459

(TSTEP) Tandom Set Terminal Extreme Probability.
[a]After the deletion naming convention in Behr et al. (1999). New deletions are indicated with N-RD.

bacteria losing (or gaining) a variety of genes have a selective advantage. The ratio may also vary as consequence of the changes to genomic content and structure brought about by the action of transposable elements, such as IS*6110* in *M. tuberculosis*. Thus, addressing the theoretical issues relating to sampling genomes, measuring genomic material, and the evolutionary genetic models that would lead to different distributions of G_{IP} will be very important for interpreting data for many genomic screens and future work in population genomics.

The use of genomic deletion screens should be defined by a clear understanding of which population is to be studied, and the desired sensitivity to discover deletions. For example, were one to implement genomic deletion screening to simply define some genetic diversity, say in an epidemiologic study, it might not be critical to find all genetic deletions, only those relative to a reference genome. On the other hand, in order to study the phenotypic consequence of total genome content, it may be far more important to include in the genomic screens a larger proportion of the genetic material likely to be observed in individual samples. In light of concerns regarding sensitivity limitations caused by genomic content screening experimental design, any deletions identified in H37Rv relative to other *M. tuberculosis* strains, most notably those relative to the well-studied CDC1551 strain, should be considered in future work on the biology of the *M. tuberculosis* genome and its evolutionary history.

By applying TSTEP, probe-pair hybridization measurements were assigned small probability scores only if they provided supporting evidence that the neighboring values to the right, or to the left, were improbably low. This approach was used to improve the detection of the deletion boundaries. However, in our particular implementation of TSTEP, there was an implicit assumption that most of the probe pairs provided signal for present DNA; the set of *LRI* values served as a null distribution (more accurately, the set of *LRI* values permitted estimation of a mean and standard error, which defined a normal distribution serving as our null). In a case in which deletions cover more than a small proportion of PPGRs, this approach to defining the null distribution would not be appropriate.

Also implicit in our definition of F_p is that the values $z(i)$ are independent. This need not be so, as we could build into our definition of F_p appropriate probabilities incorporating well-characterized dependencies in the data. For example, we know that in some regions the PPGRs overlap considerably. Combined with our assembly, which was designed to be very sensitive to detecting deletions, the overlapping PPGRs lead to false deletion calls and, sometimes, overestimation of the length of deletions. Pursuit of F_p functions that do not assume independence of overlapping

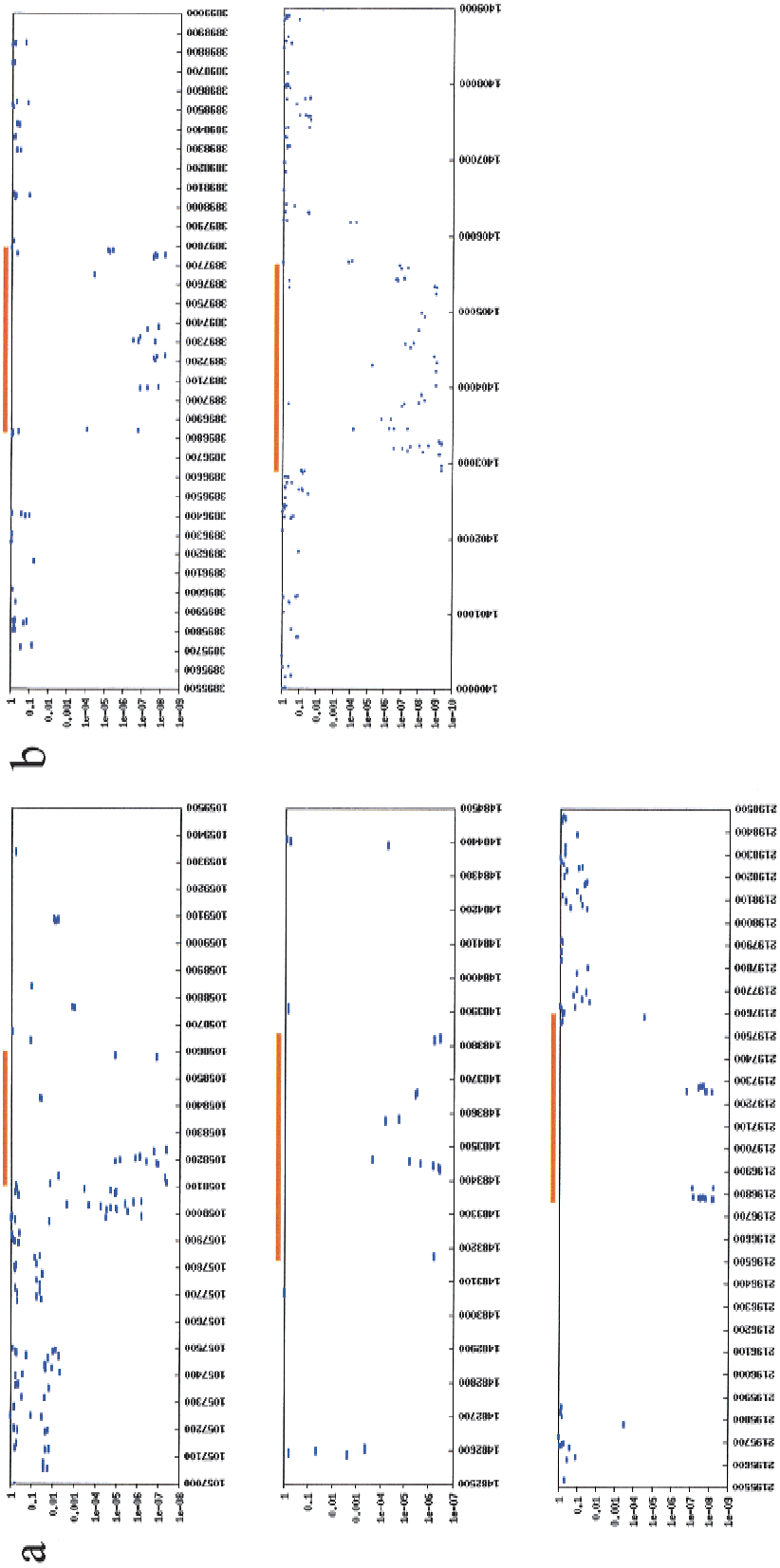

Figure 3 Three falsely called deletions (*a*) show characteristic stacking and overlapping of PPGRs. These are called as a result of the incorrect assumption of independent ratio values in the function used to evaluate the probability of a set of values (equation 3). For comparison, two other shorter true deletions are shown (*b*).

PPGRs should greatly improve the application of TSTEP.

The implementation of TSTEP here incorporated a parametric approach to assigning probabilities to hybridizaton measurements. Nonparametric approaches could be used with TSTEP in the case that nonnormality in the data leads to erroneous results. A rank-based probability was investigated for use with TSTEP but was abandoned given the much higher sensitivity of the distributional approach presented here. Nevertheless, nonparametric approaches, approaches to using the empirical distribution of intensity ratios, and the incorporation of information regarding levels of intensities (in addition to intensity ratios), could all be explored with TSTEP by designing appropriate F_p functions.

Despite the limitations of the implementation of a new computational approach, we were able to facilitate identification of deletions from high-density oligonucleotide array results. An important point addressed by the TSTEP algorithm, when it is applied with an F_p function not exploiting parameters estimated in any other experiment, is that we identify only regions that show unusually poor hybridization for that particular experiment. Thus, the implementation of TSTEP is robust against experimental error that leads to lower intensity values for an entire experiment; we will lose sensitivity, but not specificity for calling deletions if hybridization is poor for an individual experiment. A discriminant-based analysis, using a training set of experiments, could lead to many false deletion calls in the latter scenario.

Three possibilities for avoiding false deletion calls with TSTEP are (1) design or use only probes targeted at nonoverlapping portions of the genome, (2) visually examine plots of hybridization values and P values in the regions of deletion calls, rejecting calls obviously caused by overlap, or (3) use F_p functions that do not assume measurements are independent.

In the design of probe hybridization arrays for deletion finding, we provide the following suggestions. First, regions with strong likelihood of cross-hybridizing, especially polymorphic repetitive sequence systems, should be excluded from the screen, or at least investigated separately. Given the challenge to simply measure whether a (short) sequence is present or absent, the cross-hybridizing regions may be extremely difficult to characterize. A high density of evenly spaced probes will provide the most even ability to characterize deletions of the shortest detectable length. If overlapping probe regions are used, specific attention to the hybridization intensity dependency of these probes must be taken into account in the computation of deletion calls. Finally, if sequence contains self-hybridizing portions, as is the case for genes encoding tRNAs, be aware that hybridization to the array will again be intermediate between present and absent DNA, similar to polymorphic repeat systems, as these nucleotide sequences are likely to be sequestered from hybridizing the to the array.

In addition to showing successful identification of deletions by using GeneChip measurements, the TSTEP computation has been used in a study of 16 well-characterized clones of *M. tuberculosis*, in which the relationship of deletion genotype to clinical and biological phenotype is studied (P. Small, pers. comm., the authors). The work presented on TSTEP also outlines a general approach to characterizing deletions, using any DNA hybridization intensity profiling technology.

METHODS

Array Design

A high-density oligonucleotide array designed to monitor the expression of *M. tuberculosis* genes was used for these studies. The sequence for *M. tuberculosis* H37Rv with ORF annotations (Cole et al. 1998) served as a source for probe selection. Each annotated ORF and IG (Intergenic Region) was interrogated with oligonucleotide probe pairs. Twenty 25-base sequences complementary to the target were selected within each ORF or IG; these we call the PM (Perfect-Match) probe. The sequence of the PM probe with a single substitution at the middle base was also designed and called the MM (Mismatch) probe. The MM probe serves as a negative control for hybridization to the PM probe. The PM probes and their respective MM probes constitute a probe pair, and the genomic sequence they interrogate we call the PPGR. The arrays were originally designed to measure quantitative changes in mRNA expression. Thus, each ORF or IG was analyzed at 20 different loci, called a probe set. However, PPGRs were not uniformly spaced throughout each region being interrogated. For some ORFs and IGs longer than 2000 base pairs, multiple probe sets were designed. A total of 236,360 probes for *M. tuberculosis* sequence were synthesized on the array.

The design reflects the typical use of the chips to provide gene transcript expression profiles. For this study, however, all that matters to us is that we have probe pairs for PPGRs along the length of the entire genome. We use in this study 111,488 PPGRs for which we can query the presence in a sample, after removing PPGRs in repetitive sequence families (PGRS, PPE), tRNA, and rRNA genes.

Selected Bacteria

To test the data-analytical methodology, we used a clinical isolate of *M. tuberculosis* named CDC1551 and the BCG vaccine strain Pasteur. CDC1551 was isolated from an outbreak in a small community in Kentucky–Tennessee region (Valway et al. 1998). Because an unusually high infectivity was observed, this strain was chosen to be sequenced by TIGR (http://www.tigr.org). BCG Pasteur is the strain that gave rise to the group of BCG vaccine organisms that are currently used around the world. Behr et al. (1999) studied the genome content of the group of BCG organisms by using a spotted cDNA hybridization microarray.

Sample Preparation and Hybridization Conditions

The clinical sample (CDC1551) of *M. tuberculosis* was grown in Dubos with albumin medium. After 14–21 d, the DNA was

extracted using a standard procedure based on lysozyme and proteinase K[(1)]. Six micrograms of genomic DNA was partially digested with 0.1 units of DNase I (Gibco BRL Life Technologies, Rockville, MD) for 5 min at 37 °C in 1 × One-Phor-All buffer PLUS (Amersham-Pharmacia). The reaction was then heated to 99°C for 10 min to deactivate the DNase I and placed on ice, and an aliquot was loaded on a 1% agarose gel to ensure that fragments between 50–200 bp long were generated. The DNA fragments were end labeled in a 70-µL reaction with 25 nmoles biotin-N6-dideoxyadenosine triphosphate (NEN) using 100 units of terminal transferase (Roche Pharmaceuticals) in 1 × reaction buffer containing 2.5 mM $CoCl_2$ for 2 h at 37 °C. The reaction was directly used in the hybridization to high-density arrays. Hybridization solutions contained 3.0 M tetramethylammonium chloride, 0.1 M MES, pH 6.6, and 0.01% Triton X-100, 0.1 mg/mL herring sperm DNA (Life Technologies), and 0.5 mg/mL acetylated BSA (Life Technologies). A control oligonucleotide (control oligo B2, Affymetrix, Inc.) was added to a final concentration of 50 pM. Hybridization samples were heated to 99°C for 5 min followed by a 5-min incubation at 45°C and placed in the GeneChip cartridge. Hybridization was performed at 45°C for 18–20 h in a heating oven (Affymetrix, Inc.) with rotation set at 60 rpm. After hybridization, the solutions were removed, the arrays were rinsed with 6 × SSPET (0.9 M NaCl, 60 mM NaH2PO4, 6 mM EDTA, 0.01% Triton X-100, pH 7.6), washed with 0.1 × MES (100 mM MES, 0.1 M NaCl, 0.01% Triton X-100, pH 6.6) at 50°C for 30 min, and finally rinsed with 1 × MES (1.0 M NaCl, 0.1M MES, pH 6.6, and 0.01% Triton X-100). After the washes, the hybridized biotinylated DNA was fluorescently labeled by incubation with SAPE staining solution (1 ×MES, 2 mg/mL acetylated BSA, 10 µ/mL streptavidin-phycoerythrin) at 40 ° C for 15 min. Unbound streptavidin-phycoerythrin was removed by rinsing with 6 × SSPET at room temperature before scanning.

Scanning

The arrays were read at a resolution of 3 µm by using a confocal scanner with argon laser instrument (Hewlett-Packard). Photoemission was detected by a photomultiplier tube through a 570-nm longpass filter. An Affymetrix G2500A GeneArray Scanner was used for the scanning.

Intensity Data Collection

After a scan of the array surface, the computer-generated image of the array was overlayed with a virtual grid. This allowed for each feature to be defined, and the interrogating features aligned to: (1) known dimensions of the array, and (2) the features at the corner and edge regions that served as control markers. The pixels (~68 pixels per synthesis feature) within each feature were averaged, after discarding outliers and pixels near feature boundaries. The intensity information for each interrogating oligonucleotide (PM or MM) was exported as a text file. Determination of whether a hybridization result at each probe pair signaled the presence or absence of the interrogated nucleotide was determined using the TSTEP algorithm described below.

Data Analysis

We applied the calculation developed below to the genomic screen data. Our null hypothesis is that probe pairs measure present DNA; when we can reject this at a low *P* value, we have evidence that supports there being a deletion. The calculation analyzes windows of tandem probe-pair measurements. We calculated a probability of observing the set of values (here, we used the Z-scores for the log ratios of probe pairs) in a specific window, assuming that the values measure present DNA. Then, we calculated the probability of observing the set of values in the window, excluding the probe-pair measurement at the end of the window. Using the criteria as outlined in Figure 4, we assigned probe pairs at the end of deletions (as well as those in the middle of deletion regions) low probabilities (of being in a region of present DNA), because the hybridization intensities of the neighboring probe pairs in the deletion contributed to the low score assigned to the end probe pair. How many neighboring values were considered depended on the window size used in the TSTEP program.

Data Preparation

The relative values of the PM and MM hybridization intensities (I_{PM} and I_{MM}, respectively) provide information relevant to detecting the presence of a 25 bp PPGR. In general, if the interrogated DNA was present in the sample, the value of the I_{PM} should have been higher than the I_{MM}. Alternatively, if the interrogated DNA was absent, the value of the I_{PM} should have be similar to the I_{MM} value. Thus, the ratio of I_{PM} to I_{MM} was calculated for each probe pair. The logarithms of these ratios were calculated, to provide better symmetry in the distribution about a ratio of one. Specifically, we defined the log ratio intensities

$$LRI(i) = \log_{10}[I_{PM}(i)/I_{MM}(i)] \quad (1)$$

for each probe pair, *i,* in order of the corresponding PPGRs along the genome, beginning at the origin of replication.

The log ratio tends to be higher for DNA present at a PPGR, and lower when DNA queried is absent. The log ratio data was then normal-transformed. The mean (μ_{LRI}) and standard error (σ_{LRI}) of the distribution of *LRI* values was com-

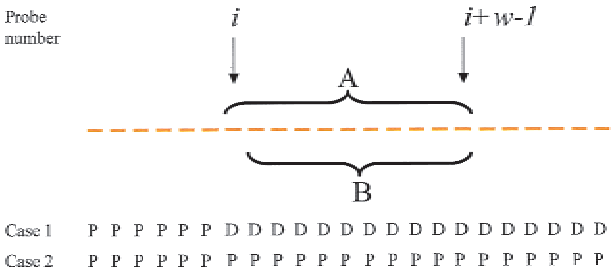

Figure 4 Enhancement of scores for hybridization by TSTEP is used to help define deletion regions. One iteration of the TSTEP calculation proceeds as follows: If the joint probability of (tandem) measurements A is less than the joint probability of (tandem) measurements B, and is also less than the probability of measurement *i* itself, then assign the probability of measurements A to position *i*. Note that A is the set of *w* measurements beginning at the *i*th measurement, where *w* is the window size. The score reassignment will tend to be different depending on whether probes query present (P) or deleted (D) regions. For example, in case 1, probe *i* is the first probe to query in a deleted region (from left to right); it is quite likely that the probe will be reassigned a very low probability score. In case 2, where all probes query present DNA, probe *i* is likely either to not be reassigned a score, or to be re-assigned a score which does not correspond to a very low probability.

puted. The normal deviate for each probe pair was calculated as

$$z(i) = (LRI(i) - \mu_{LRI})/\sigma_{LRI}. \quad (2)$$

This normal deviate measures the departure from the mean PM to MM intensity ratio. Each $z(i)$ value is associated with a probability of observing a value at or below the observed $LRI(i)$ value, using the standard function for the area to the left of a value under the normal curve, ϕ. The calculation of probabilities is valid only if one assumes that all probe pairs measure present DNA (this is our null hypothesis) and that the noise in log ratios for present DNA is normal. Departure from normality will mean that the P values calculated are not valid. However, as long as low $z(i)$ values are rare and correlated with deleted DNA, the algorithm developed below will still aid us in discovering deletion regions.

Individually, the presence or absence of a single PPGR in a sample is difficult to determine; the sensitivity and specificity characteristics of single probe-pair *LRI* values is poor. Furthermore, low *LRI* values may indicate either absence of hybridizing DNA in the sample, or a probe pair that is not performing well. We thus introduce the following method, a computational algorithm we call the Tandem Set Terminal Extreme Probability algorithm (TSTEP), to analyze the normal-transformed *LRI* data. In addition, we develop heuristics to assemble deletion intervals from the output of TSTEP.

TSTEP

TSTEP requires a function F_p, which evaluates a set of w tandem values $T(i,w)$, beginning at the ith value in the data. The function F_p must report a score that decreases monotonically with decreasing probability of observing a set of values; we must have a way to score the probability of observing the values found in any window of values. First, we set the individual values in a complete list of N tandem (ordered) values, $s(i) = F_p(T(i,1))$, $i = 1, \ldots, N$ (in the application to genomic probe data, we initialized the values to the Z-scores defined in equation 2, see Application of TSTEP to GeneChip data, below).

TSTEP reassigns scores in the following way. Consider some window size $w > 1$. For each window, evaluate the subset of values without the leftmost value; i.e., calculate $F_p(T[i + 1, w - 1])$. If $F_p(T[i,w]) < F_p(T[i + 1, w - 1])$ and $F_p(T[i,w]) < s(i)$ then reset $s(i)$ to $F_p(T[i,w])$. That is, if the set of w values including the ith value is less probable than the set of $w - 1$ values to the right, and if the set is less probable than the ith value itself, assign the group score to the ith value. If the value at the left of a window contributes to an improbable, extreme set of values, then it is assigned the score of that extreme set. In this way, values that could be the left-hand end of a run of extreme values are enhanced by assigning them the low scores of the entire window. The right-hand side is treated in the analogous way: If $F_p(T[i,w]) < F_p(T[i, w - 1])$ and $F_p(T[i,w])$ $s(i + w - 1)$ then reset $s(i + w - 1)$ to $F_p(T[i,w])$.

For all windows of size w in the data, that is, for $T(i,w)$ where $i = 1, \ldots, N - w + 1$, this enhancement is performed for values in regions of extreme sets of values.

Application of TSTEP to GeneChip Data

We applied TSTEP to the normal deviate values that were calculated in preparing the data for analysis. Thus, we define $T(i,w) = \{z(i), \ldots, z(i + w - 1)\}$, the array of tandem normal deviates in a window size w in the data, beginning at PPGR i. We define the function required by TSTEP as follows:

$$F_p(T(i,w)) = (1/\sqrt{w}) \sum_{j=1}^{j=i+w-1} - z(j). \quad (3)$$

We may combine normal deviates (Z-scores) in this manner under the assumption of independence of the $z(i)$. The expectation of each variable $z(i)$ is zero and its variance is one. As the expectation of a sum is the sum of expectations, and the variance of a sum of independent random variables is the sum of the variances of the random variables, the left-hand side of equation 3 is the normal deviate of the sum of the w normal deviates in the window. This function has the required property of decreasing strictly with decreasing probability. Furthermore, it is computationally inexpensive to work with the z-scores without conversion to the P values (which would require calculating approximations to the ϕ function for each window and subwindow investigated).

First, we initialized $s(i) = F_p(T(i,1))$, for $i = 1, \ldots, N$, which assigned $z(i)$ to $s(i)$ for each PPGR numbered in genomic order from 1 to N (here N is 111,488). We applied TSTEP to these values five times, using window sizes 8, 9, 10, 11, and 12, keeping the lowest score for each PPGR. This application of TSTEP with multiple window sizes was performed on the BCG and on CDC1551 data.

The enhanced z-scores obtained from TSTEP were converted to pseudoprobabilities by using the standard ϕ function. The resulting list of probabilities, $p(i)$, for each PPGR i, were used in the assembly of genetic deletion regions. In Figure 1, the enhancement of the probabilities using TSTEP is clearly evident.

A simple implementation of TSTEP in a script is available at http://molepi.stanford.edu/TSTEP. Please send inquiries regarding software to tstep@molepi.stanford.edu.

A very conservative correction for multiple tests was applied to the probabilities to test if this strict approach is useful: $p_c(i) = 1 - (1 - p[i])^N$. A p_c value attaining a significance level of say, 0.05, indicates that a value so low should be found once or more in only 1 of 20 complete genomic screens. Although there is loss of sensitivity with this extreme correction (data not shown), deletions typically contain some PPGRs that are significant at the 0.05 level. However, if we expect to measure multiple deletions in a genomic screen, we may well want to risk more false deletions to gain sensitivity. In fact, our heuristic approach outlined below was designed to be more sensitive by using a more liberal cutoff than the corrected value of 0.05.

Assembly of Putative Deletions

We do not have a distribution of deletion lengths that we expect to find in these samples. A probabilistic calculation to determine deletion boundaries is desirable; however, given the limitations on our knowledge of deletions, determination of putative deleted regions simply involved a computational approach incorporating the following information:

- A threshold probability for the $p(i)$, below which PPGRs were considered putatively deleted
- The total sequence length covered by assuming tandem probe pairs and intervening sequences were deleted
- The sequence distance between putatively deleted PPGRs (unqueried sequence length)
- The number of tandem PPGRs not called putatively deleted

(i.e., above the threshold) in a region of other PPGRs that were called putatively deleted
- The proportion of probe pairs individually called putatively deleted within a region of multiple PPGRs.

The final deletions that were called by our computer scripts consisted of genomic intervals in which:

1. A minimum of 80% of probe pairs exhibited $p(i) < 0.00005$.
2. A total sequence length (in H37Rv) of at least 350 bp was defined.
3. A maximum of 2000 bp existed between putatively deleted PPGRs; i.e., final deletion calls contained a maximum length of 2000 bp of contiguous unqueried sequence.
4. A maximum of three PPGRs in tandem with $p(i)$ not below the threshold were permitted, as long as 1–3 above were also satisfied.

These reported intervals are expected to be somewhat shorter than the true deletions, as the beginning of the called deletion begins at one end of a PPGR and ends at the far end of the last PPGR in the interval; unless a PPGR contains the deletion break, we can expect to miss the end of the true deletions.

Target Genomes and Confirmation of Deletion Calls

For the CDC1551 strain, 42 sequenced and assembled contigs acquired at the Institute for Genomic Research (TIGR) web site in October 1998 were used as a proxy for the complete genome (data released as of March 1999 was identical). The H37Rv complete genomic sequence, acquired from the Sanger Center web site in October 1998, was split into 60 bp segments to quickly obtain a list of the deleted DNA sequences in CDC1551 relative to H37Rv; each segment was subsequently searched for presence in the CDC1551 sequences. In each case that five or more 60–-bp segments of H37Rv in a row failed to find a perfect match in the CDC1551 sequence, the region was investigated further. The absence or mismatch of the regions, consisting of at least 300 bp, was ascertained by BLAST search; the segments were BLAST-searched against the CDC1551 sequence at TIGR. In addition, mismatched segments containing annotated repetitive sequences, such as those in PPE or PE-PGRS families were noted. Those segments that were found to exhibit <50% similarity with any region and not consisting of a majority of repeat family sequence were considered true deletions, and thus fair targets for the GeneChip deletion screen of CDC1551.

For the BCG Pasteur strain, 14 previously reported deletions were considered fair targets for the GeneChip deletion screen of BCG Pasteur. Additional called deletions were investigated using a previously described PCR strategy (Behr et al. 1999).

As H37Rv DNA should all be present in the GeneChip screen, the H37Rv screen served as a negative control for deletions. Those regions with low probability scores for the H37Rv screen using TSTEP were eliminated from further investigation in the screens of both CDC1551 and BCG Pasteur.

ACKNOWLEDGMENTS

We thank Daryl Thomas for assistance with data acquisition and Hong-Tao Lu for comments on this manuscript.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Behr, M.A., Wilson, M.A., Gill, W.P., Salamon, H., Schoolnik, G.K., Rane, S., and Small, P.M. 1999. Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* **284:** 1520–1523.

Brosch, R., Gordon, S.V., Billault, A., Garnier, T., Eiglmeier, K., Soravito, C., Barrell, B.G., and Cole, S.T. 1998. Use of a Mycobacterium tuberculosis H37Rv bacterial artificial chromosome library for genome mapping, sequencing, and comparative genomics. *Infect. Immun.* **66:** 2221–2229.

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry, C.E., et al. 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* **393:** 537–544.

Lipshutz, R.J., Fodor, S.P., Gingeras, T.R., and Lockhart, D.J. 1999. High density synthetic oligonucleotide arrays. *Nat. Genet.* **21:** 20–24.

Mahairas, G.G., Sabo, P.J., Hickey, M.J., Singh, D.C., and Stover, C.K. 1996. Molecular analysis of genetic differences between Mycobacterium bovis BCG and virulent M. bovis. *J. Bacteriol.* **178:** 1274–1282.

Valway, S.E., Sanchez, M.P., Shinnick, T.F., Orme, I., Agerton, T., Hoy, D., Jones, J.S., Westmoreland, H., and Onorato, I.M. 1998. An outbreak involving extensive transmission of a virulent strain of Mycobacterium tuberculosis. *N. Engl. J. Med.* **338:** 633–639.

Winzeler, E.A., Lee, B., McCusker, J.H., and Davis, R.W. 1999. Whole genome genetic-typing in yeast using high-density oligonucleotide arrays. *Parasitology* **118:** S73–S80.

Received June 19, 2000; accepted in revised form September 18, 2000.