Comparative analysis of bacterial genomes: identification of divergent regions in mycobacterial strains using an anchor-based approach

Anchal Vishnoi[1,*], Rahul Roy[2] and Alok Bhattacharya[1,3]

[1]Center for Computational Biology and Bioinformatics, School of Information Technology, [2]Indian Statistical Institute, New Delhi 110016, India and [3]School of Life Sciences, Jawaharlal Nehru University, New Delhi 110067, India

ABSTRACT

Comparative genomic approaches are useful in identifying molecular differences between organisms. Currently available methods fail to identify small changes in genomes, such as expansion of short repetitive motifs and to analyse divergent sequences. In this report, we describe an anchor-based whole genome comparison (ABWGC) method. ABWGC is based on random sampling of anchor sequences from one genome, followed by analysis of sampled and homologous regions from the target genome. The method was applied to compare two strains of *Mycobacterium tuberculosis* CDC1551 and H37Rv. ABWGC was able to identify a total of 104 indels including 20 expansion of short repetitive sequences and five recombination events. It included 18 new unidentified genomic differences. ABWGC also identified 188 SNPs including eight new ones. The method was also used to compare *M. tuberculosis* H37Rv and *M. avium* genomes. ABWGC was able to correctly pick 1002 additional indels (size >100 nt) between the two organisms in contrast to MUMmer, a popular tool for comparative genomics. ABWGC was able to identify correctly repeat expansion and indels in a set of simulated sequences. The study also revealed important role of small repeat expansion in the evolution of *M. tuberculosis* strains.

INTRODUCTION

Comparative analysis of fully sequenced genomes is a powerful approach to detect and measure diversity among organisms. It has become apparent in the last few years that many biological properties including clinical features can be inferred successfully from the analysis of full genome sequences (1–4). In parallel a number of experimental studies have been initiated to document differences among closely related organisms and field isolates in the form of sequence differences, such as single nucleotide polymorphisms (SNPs), repetitive sequence-based polymorphisms, variable transposon insertions, recombination events, etc (5). Both these approaches complement each other. In the context of pathogenic organisms the results from these studies can help in developing newer methods for diagnosis and identification of drug and vaccine targets (6–8).

One of the major problems in understanding host–pathogen relationship in many infections is to explain the variety of clinical features ranging from asymptomatic infection to different forms of invasive disease. Many of the differences in the clinical features can be attributed to genetic differences among pathogenic strains. For example, comparative genome sequence analysis identified 1500 distinct genes present in pathogenic *Escherichia coli* O157, H7 but absent in non-pathogenic *E. coli* strain K-12 (9). These include genes involved in colonization and toxin production, responsible for disease pathology. Among mycobacteria, SNPs, insertion elements and genomic deletions have been associated with clinical features of different strains and species (10). In these organisms, unlike many others, there is no evidence for the presence of toxin genes which can be directly associated with virulence. Comparative analysis of a number of different strains and species of mycobacteria indicate that many of the sequence polymorphisms arise from specific deletion patterns. The genes affected by the deletions have important roles in the biology of these organisms (11). Since, deletions tend to be irreversible events (12,13), the pattern of deletions can be used to deduce the phylogeny of the mycobacteria. The distribution of deletions suggests that *Mycobacterium tuberculosis*

H37Rv has not originated from *M. bovis* (14,15) as thought previously. Also, deletions along with other mutational analysis can be used as markers to study the evolution of genomes. The SNPs have also been used to carry out phylogenetic analysis of *M. tuberculosis* strains (16). Some of the identified SNPs in *M. tuberculosis* alter activities of enzymes thought to be involved in pathogenesis. The results show that this species is highly clonal, without detectable lateral gene transfer. Attempts have also been made to associate virulence with insertion of IS elements and repetitive polymorphic sequences (17). Different markers have been deployed for typing clinical isolates of *M. tuberculosis*, for example IS6110-based restriction fragment length polymorphism (RFLP) (18), spoliogotyping, a PCR-based method using repetitive elements at a single locus (19) and variable number tandem repeat (VNTR) typing, a tool based on repetitive elements (20). While these experimental approaches are useful for detection of a few differences, they cannot display total genomic diversity. Microarray hybridization has also been used to map variations among strains and isolates of mycobacteria (21,22). Though this approach involves a large number of loci spanning the genome subtle differences are likely to be missed, as coding regions are normally used for making the arrays and detection is based on hybridization. Computational methods based on whole genome sequences may be more useful for identification of small changes among genomes.

The existing computational approaches used for identification of the diverged sites and regions between genomes are based on complete genome alignment. Several algorithms have been developed for alignment of large sequences. These can be categorized as local alignment and global alignment-based methods. Anchor identification is the first step in global alignment-based methods, such as DIALIGN (23), DBA (24), LAGAN (25), GLASS (26) and AVID (27). While GLASS and AVID use exact k-mers as anchors, substitutions or mismatches are allowed in DIALIGN and LAGAN. The consecutive anchor pairs are then aligned. These methods are not able to identify genomic rearrangements, such as inversions and translocations (28). On the other hand local alignment techniques, such as MUMmer (29), WABA (30), BLASTZ (31) and SSAHA (32) are able to locate translocations and rearrangements. These methods also use anchor-based strategies employing slightly different approaches for anchor identification. While MUMmer identifies exact matches, WABA allows mismatches at the wobble positions. The local alignment-based tools fail to identify short repeat expansions as these are either shown as overlapping regions (MUMmer) or the two overlapping regions are merged (WABA). Many of these methods are not suitable for analysis of a pair of diverged sequences as the anchors are not identified properly or very few anchors are identified. Therefore, there is a need for new genome comparison tools that can analyse both closely related and diverged genomes with capability to find indels, repeat expansions and rearrangements, and other mechanisms contributing to genome diversity.

In this article, we present a new method anchor-based whole genome comparison (ABWGC) for identification of divergent regions between genomes. We have tested our method on different strains of a pathogen *M. tuberculosis*. We have also analysed two different species of mycobacteria to show that this method can also identify divergent sequences, a feature not available in many other comparative genomic tools. Based on the comparison with other alignment tools we conclude that ABWGC is a preferred method for finding small changes in genomes such as small repeat expansion, indels, rearrangement between closely related genomes and analysis of divergent genomes.

METHODS

Let S (the query) and T (the target) be two genomes of length N and M, respectively. We first select some random positions on the query genome. Each of these positions would be starting points of the anchors. The anchors are of fixed length m and we require that these anchors be non-overlapping. As such we need to ensure that there was a minimum distance, L, between two successive random positions, where $L \geq m$. We obtain this as follows.

Let $x_1, x_2, \ldots, x_N$ be a random permutation of the numbers $1, 2, \ldots, N$, where each permutation is equally likely to occur. This random permutation is obtained by the Mersenne Twister programme (http,//www.math.sci.hiroshima-u.ac.jp/ m-mat/MT/emt.html). The random positions of the anchors are constructed according to the following iterative scheme,

let $y_1 = x_1$;
and $y_2 = x_{k_1}$, where $k_1 = \min\{j > 1, \mid x_j - y_1 \mid \geq L\}$;
having defined y_i and k_{i-1} let
$y_{i+1} = x_{k_i}$
where $k_i = \min\{j > k_{i-1}, \mid x_j - y_l \mid \geq L$ for all $l \leq i\}$.

We terminate this iterative scheme when it is not possible to define any further y. Let $\{y_1, y_2, \ldots, y_n\}$ be the set of all possible y's obtained by the above scheme. We note here that $y_1, y_2, \ldots, y_n$ need not be in either an increasing or a decreasing order. However, with a slight abuse of notation assume that $y_1, y_2, \ldots, y_n$ are in an increasing order.

Let λ_i^j denote the nucleotide at the position $y_i + j$ in the query genome S. Thus, for example $\lambda_i^j = A$ if the nucleotide at the $(y_i + j)$th position in the query genome S is A, etc. The string

$$A(i) = \lambda_i^0, \lambda_i^1, \ldots, \lambda_i^{m-1}$$

represents the string consisting of m consecutive nucleotides of the genome S starting at the y_ith position.

The strings $A(1), A(2), \ldots, A(n)$ represents our anchors at position $y_1, y_2, \ldots, y_n$ on the genome S. The choice of y_i's ensure that these anchors do not overlap.

Based on these anchors we obtain a set of strings $B(1), B(2), \ldots, B(n)$ from the target genome T. The string $B(i)$ is that segment of T which gives the highest BLAST score when compared with the string $A(i)$ of the query genome S.

To fix notation, let the string $B(i)$ start from the position t_i of the target genome T. Letting μ_i^j denote the nucleotide at the position $t_i + j$ in the target genome, we have

$$B(i) = \mu_i^0, \mu_i^1, \ldots, \mu_i^{m-1}.$$

We note that $B(i)$'s may be overlapping, and although $A(i)$'s are arranged in an increasing order according to their position in the genome S, $B(i)$'s need not preserve that order.

Let

$$p_i = y_{i+1} - y_i + 1$$

and

$$l_i = |t_{i+1} - t_i| + 1,$$

i.e. p_i is the inter-anchor distance in the S genome (including the end points) and l_i is the distance between the corresponding BLAST hits (the absolute value taken so as to ignore inversions in location of the hits). The positions of the anchors were recorded in order to find inversions. In case of duplications there will be multiple hits in the target genome T for an anchor $A(i+1)$. The expected position of $B(i+1)$ is estimated by adding p_i of the homologous anchor in the S genome with the position of $B(i)$. If an anchor is present in multiple copies, there would be a number of hits with nearly equal mismatch scores. The anchor, whose position is close to the expected computed position, is chosen for analysis.

Mismatch score calculation

The proportion of mismatches between the strings $A(i)$ and $B(i)$ was calculated between these strings. The mismatch score is based on a binary scheme, a nucleotide match between the string $A(i)$ and $B(i)$ was given a value of 1 and a mismatch was given as 0. The sum of these 1s and 0s normalized with the length of the anchor was a mismatch score of the anchor. More formally, the mismatch score is given by the quantity CNS (cumulative normalized score) given below. Let

$$\delta(A(i), B(i)) = \frac{1}{m}\sum_{j=0}^{m-1} d(\lambda_i^j, \mu_i^j)$$

where

$$d(\lambda_i^j, \mu_i^j) = \begin{cases} 0 & \text{if } \lambda_i^j = \mu_i^j \\ 1 & \text{otherwise.} \end{cases}$$

The mismatch score is

$$\text{CNS} = \frac{1}{n}\sum_{i=1}^{n} \delta(A(i), B(i)).$$

where n is the total number of anchors.

Estimation of anchor order

The gene-order approach used depends on the conservation of the genes (33), based on the same approach the anchor order was estimated as follows,

For $i = 1, 2, \ldots, n-2$ let

$$o(A(i), B(i)) = \begin{cases} 1 & \text{if } t_i < t_{i+1} < t_{i+2} \\ 0 & \text{otherwise,} \end{cases}$$

where $o(A(i), B(i))$ is 1 if the anchor order is preserved and it is 0 if there is a breakpoint in the synteny of the two genomes. This $o(A(i), B(i))$ is used instead $\delta(A(i), B(i))$ to obtain an equivalent CNS.

Identification of variable regions

All the p_i and l_i where the difference between them was >10 bp were extracted from genome S and genome T, respectively. The extracted sequences were then globally aligned to obtain the exact position of the insertion, deletion and duplication. The position were then mapped with the genome to get the coordinates of these events. For identification of divergent regions due to nucleotide alterations, a clustering approach was used. High scoring anchors above a threshold were clustered and a sampling procedure was carried out in the entire clustered region. If the samples also followed the high scoring criteria then the whole clustered region was a divergent region. The divergent region flanking an anchor its boundaries were extended in both the direction till a low scoring region was reached. To know the precise boundaries of the divergent region, the homologous regions from both the S and T genomes were aligned globally. The parameters such as gap open penalty and gap extension penalty were changed according to divergence between the genomes. The indels and duplications were then subsequently identified.

Data

The full genome sequences of *M. tuberculosis* strain CDC1551, H37Rv and *M. avium* subsubspecies paratuberculosis K-10 strain were obtained from NCBI (http,//www.ncbi.nlm.nih.gov/genomes/lproks.cgi). The three strains of *M. tuberculosis* F11, C and Haarlem have been isolated from South Africa, New York City and Netherland, respectively and the complete genome

sequences are not yet available. The partial genome sequence were obtained from Broad Institute (http,// www.broad.mit.edu/annotation/genome/mycobacterium-tuberculosis-spp/MultiDownloads.html). Three simulated data sets were also prepared in order to quantify the performance of ABWGC. The first simulated data set was a DNA sequence of 3 kb extracted from the genome of *M. tuberculosis* H37Rv in which 1 kb region was randomly mutated. Another data set for simulation consists of two DNA sequences which contains five and six copies of a tandem repeat, respectively. In the third simulated data set in a DNA sequence, two foreign sequences were inserted with a gap of 10 nt between the inserted regions.

Availability

The set of programmes which constitute ABWGC and detailed instruction of their use can be obtained from the corresponding author on request (anchalv@gmail.com, av1563@students.jnu.ac.in, alok0200@mail.jnu.ac.in).

RESULTS

Description of ABWGC

Random sampling and determination of optimal anchor length. The approach described here was based on random sampling of fully sequenced genome sequences. Short sequences of predefined fixed length were extracted from random positions of a given genome (*S*) with some restrictions as described in the Methods section. These samples were used as anchors. BLAST algorithm was used to get the homologous regions for each of the anchors in the target genome *T*. A score was calculated for each anchor, based on sequence mismatch with target anchor and a cumulative normalized score (CNS) was then derived encompassing all the anchors as described in the Methods section. The CNS is essentially the average score of all the anchors reflecting the level of diversity at the nucleotide sequence level. Detailed analysis has also shown that the final CNS score was independent of sampling (data not shown).

To determine the optimum length, anchors varying from 25 to 300 nt from *M. tuberculosis* strain H37Rv (S genome) were extracted and compared with *M. avium* (T genome). Figure 1 shows the change in CNS at every sampling point for an indicated anchor size. The value of CNS changed with increase in the number of samples eventually reaching a plateau after about 3000 samples. Sampling from a genome involved ~10% of the total size in terms of nucleotides. For all analysis, the value of CNS used was the one obtained when the curve reached a plateau. The value of CNS was similar if the anchor lengths varied from 100 to 300 nt. However, there was a marked difference in the score when the anchor size was less than 100 suggesting that 100 nt could be the optimum size of the anchor for determining divergence between genomes.

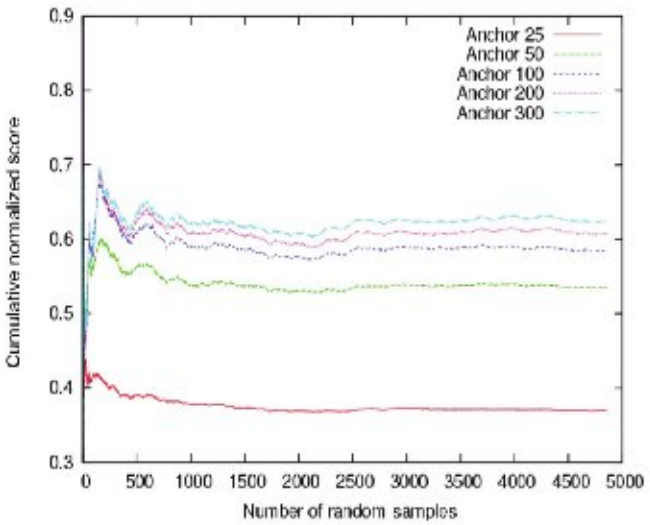

Figure 1. Cumulative normalized score of different sized anchors. Anchors, ranging from 25 to 300 nt, were extracted from random positions of *M. tuberculosis* H37Rv genome. The homologous anchors from *M. avium* were identified by BLAST. The mismatch score of each anchor pair was used to calculate CNS. The anchor numbers represent anchors that have been extracted sequentially in terms of the position in the genome.

Identification of diverged regions at the level of nucleotide sequence. Anchors with high score in any pairwise comparison correspond to divergent DNA segments. For analysis of the diverged regions, the mismatch score of each anchor was plotted in order to obtain the overall distribution (Figure 2a). When two strains of *M. tuberculosis* were compared, most of the anchors had score less than 20 suggesting a high degree of identity (Figure 2a and c). A few anchors had high scores, suggesting that these may lie in diverged regions. The boundaries of the diverged regions were identified by clustering the neighbouring anchors with scores above a threshold. The threshold value for the identification of the diverged regions was obtained by the distribution analysis of the scores of anchors obtained when *M. tuberculosis* genome was compared with a randomly generated sequence of the same length and base composition (Figure 2b and d). None of the individual anchor score was less than 70. A threshold score of 20 was used for subsequent analysis involving closely related organisms, such as strains of *M. tuberculosis* H37Rv and CDC1551. More sampling was carried out from the entire clustered segment. If the anchors derived from this sampling were also found to have high scores above the threshold, then the entire clustered region was considered to be the divergent region. To find a divergent region around an anchor (a high score anchor flanked by anchors with low scores) sampling was carried out in its flanking regions. Sampling was stopped when a low scoring region was reached on both sides of the divergent anchor. Precise boundaries were then obtained by

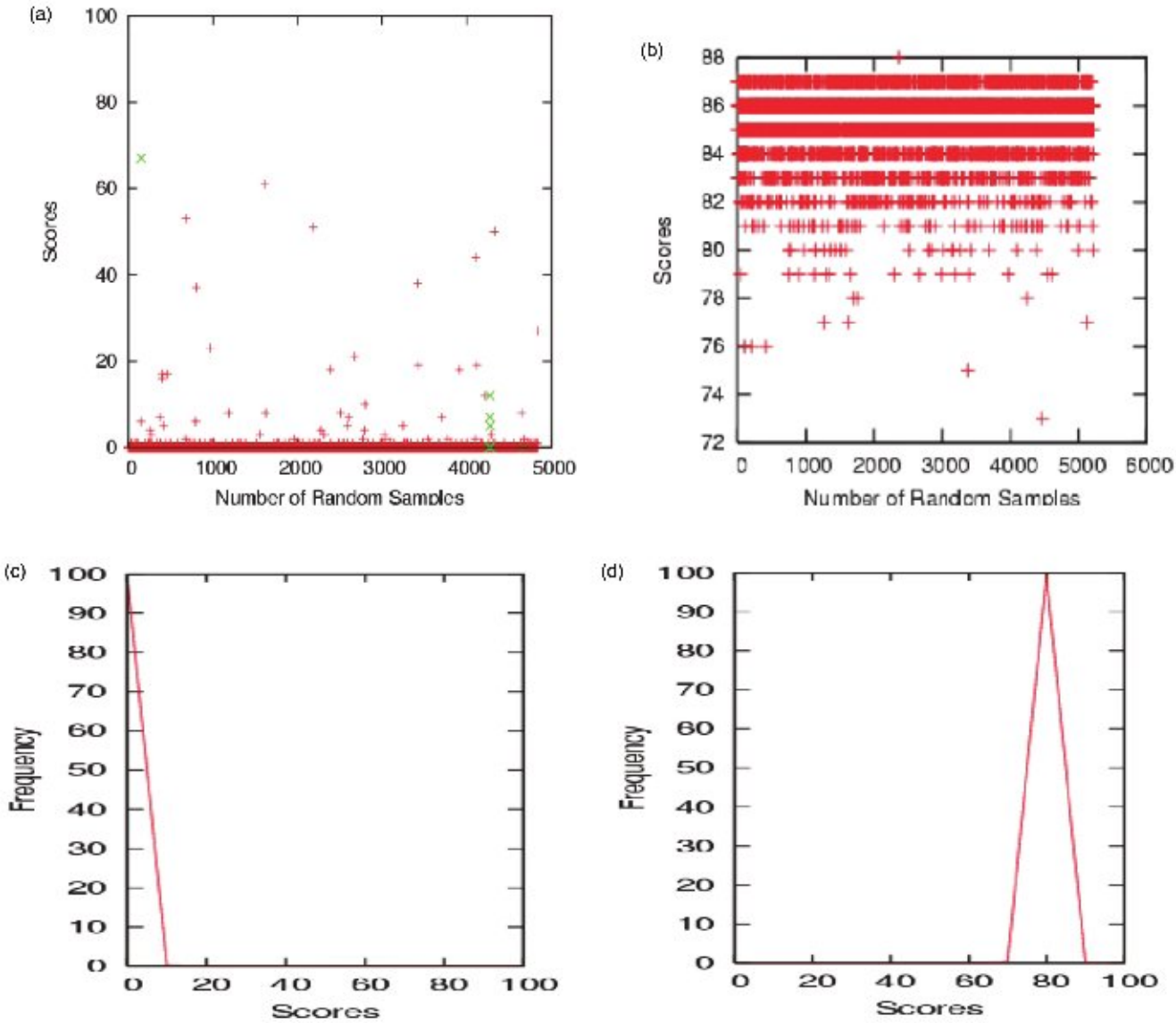

Figure 2. Distribution of individual mismatch score. For every randomly picked anchor and its homologous sequence in the target genome (T), a mismatch score was computed as described in the Methods section. The individual scores were plotted against anchors that have been extracted sequentially in terms of position in the genome (S). Homologous anchors on positive strand are shown by plus (+) and those in complementary strand by cross (×). (**a**) *Mycobacterium tuberculosis* CDC1551 (S) and *M. tuberculosis* H37Rv (T); (**b**) *M. tuberculosis* CDC1551 (S) and random genome (T). The frequency distribution plot of the scores of anchors (**c**) *M. tuberculosis* CDC1551 (S) and *M. tuberculosis* H37Rv; (**d**) *M. tuberculosis* CDC1551 (S) and random genome (T).

aligning the putative divergent regions using a combination of both local (Smith and Waterman alignment) and global alignment (Needle and Wunsch) methods (34,35). Apart from large variations it is also possible to find small changes (for example SNP) in nucleotide sequences. If an anchor mismatch score is above 0 it indicates sequence mismatches. Precise location of SNP was obtained by alignment of homologous anchors. As, the number of anchors used in the analysis covers 10% of the genome, the number of SNPs identified may be a fraction of the total numbers present. Increasing the number of samples would increase the detection rate.

Insertion, deletion, recombination and inversion. Inter-anchor regions were extracted and anchor order was determined from the nucleotide positions of the respective anchors in the two genomes. All these measures were then used to determine indels, duplications and disruption in synteny. The differences in the lengths of inter-anchor regions of S and the lengths between the corresponding

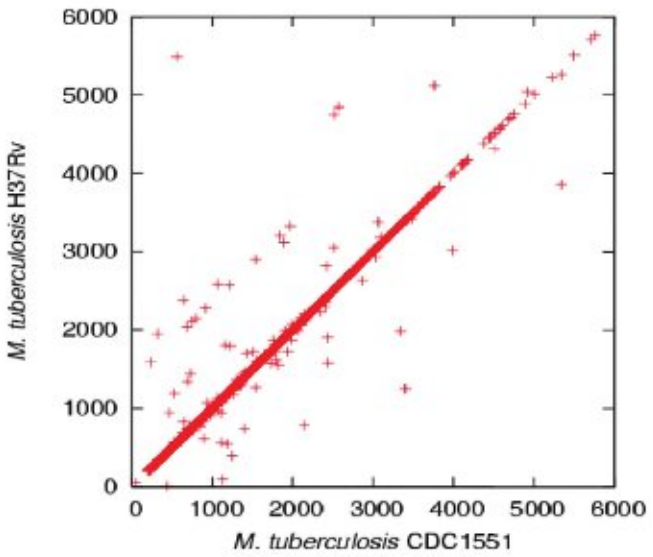

Figure 3. The inter-anchor length distribution. The anchors were extracted from *M. tuberculosis* CDC1551 randomly as described in the Methods section. These anchors were ordered in terms of their position in the genome. The homologous of these anchors were from *M. tuberculosis* H37Rv. The distance between two consecutive anchors were computed for both genomes. These inter-anchor lengths of two genomes were then plotted. The deviation from the diagonal represents the difference between the inter-anchor length of a pair of homologous anchors.

anchors of *T* genome were indicative of these events. The plot of the distribution of inter-anchor lengths of two closely related organisms is expected to contain most points on or around a diagonal, reflecting small or no change in the distance. The few points scattered around the diagonal line depict small changes. Large alterations can be identified as the points away from the diagonal (Figure 3). The inter-anchor length distribution analysis showed that >97% of the anchors were within 10 nt. Therefore, a threshold of 10 nt was used to identify those inter-anchor regions selected for further processing. The selected inter-anchor sequences were extracted from both *S* and *T* genomes and aligned using a global alignment method to identify the indels. Both the strands of the genome were taken into consideration while extracting homologous inter-anchor regions. All the indels obtained by ABWGC were checked again using a BLAST analysis against the target genome for their absence or presence in the target genome.

Identification of repetitive sequences. An insertion event would result from the addition of a stretch of nucleotides or an increase in the number of repeats of a repetitive sequence. In order to determine the nature of an insertion, the inserted region was extracted and BLAST analysis was carried out against the inter anchor region where the insertion was originally found. The presence of repetitive elements is indicated by multiple BLAST hits. Number of hits is equal to the number of copies of the repetitive sequence present in the region. The consecutive multiple hit depicts tandem repeats. The results were confirmed by tandem repeat finder (36).

Analysis of two *M. tuberculosis* strains

ABWGC was applied to catalogue the genomic differences between the two strains of *M. tuberculosis*. The strains H37Rv and CDC1551 were used as *S* and *T* genomes, respectively. The majority of the anchors displayed scores in the range of 0–10 confirming a close evolutionary relationship between the two genomes. However, high scores, that is more mismatches, were observed for some of the anchors, suggesting that these anchors correspond to regions of divergence (Figure 2a). The anchors (17) that map to the complementary strand of the *T* genome reflect flipping of the sequences. The number of such events was found to be five between these two strains (Table 1). The two strains of *M. tuberculosis* H37Rv and CDC1551 are 99% identical in nucleotide sequence. Only two divergent regions were obtained by our analysis which mapped to MT1499 gene in *M. tuberculosis* CDC1551 and Rv3343c in H37Rv. These genes code for PE-PGRS and PPE, respectively. We also identified SNPs that map to the anchors. Eight of these SNPs have not been reported before though the SNPs observed by our analysis at the default level of sampling, that is 10% of the genome, is ~6.5% of that identified by MUMmer (Table 2). Inability of the MUMmer to identify these eight SNPs may be due to their locations in regions not aligned by MUMmer.

Analysis of the inter-anchor regions helped to identify 104 indels between the two strains. The majority of the variations (~85%) were restricted to the 5′ or the 3′ end of the coding regions. Only a small fraction (15%) of the insertions in *M. tuberculosis* H37Rv was mapped to the intergenic regions. Among the protein-coding genes (37), PE-PGRS and PPE genes accounted for the largest family (17) of proteins affected. Previous studies have shown that a number of large indels (> 500 nt) are due to insertion elements, such as IS6110 (37). IS elements are not included in the results as these are easier to identify. Eight non-IS elements-derived indels were also detected. For example, an insertion of 653 bp in *M. tuberculosis* H37Rv was located in Rv1091 gene, a member of PE-PGRS family. Similarly, an insertion of 2273 bp spanning the genes Rv2123 and Rv2124 was also found in *M. tuberculosis* H37Rv. These changes are likely to alter the functions of the genes.

The nature of the indels was identified by a detailed analysis as described before. The results revealed that many of the observed indels were actually due to an increase in copy number of small tandem repeats (Table 3). In *M. tuberculosis* H37Rv, 22% (10/45) of the insertions were due to the increase in copy number of small repeats and ten such expansion of repeats were identified in *M. tuberculosis* CDC1551. Some of these alterations do not change the reading frame of the

Table 1. *M. tuberculosis* CDC1551 anchors that mapped to the complementary strand of *M. tuberculosis* H37Rv

Anchor start position	Anchor end position	Score	Strand	CDS	Function
674519	674487	67	Minus	Rv0578c	PE-PGRS
1788224	1788125	0	Minus	Rv1587c	Partial REP13E12 repeat protein
1787355	1787256	0	Minus	Rv1586c	Probable phiRv1 integrase
1787238	1787139	0	Minus	Rv1586c	Probable phiRv1 integrase
1786632	1786533	0	Minus	Rv1585c	Possible phage phiRv1
1786225	1786126	0	Minus	Rv1583c	Possible phage phiRv1
1785944	1785845	0	Minus	Rv1582c	Possible phage phiRv1
1785225	1785126	0	Minus	Rv1582c	Possible phage phiRv1
1785007	1784908	0	Minus	Rv1582c	Possible phage phiRv1
1783666	1783567	0	Minus	Rv1579c	Possible phage phiRv1
1783346	1783247	0	Minus	Rv1579c	Possible phage phiRv1
1782637	1782538	0	Minus		lies in intergenic region
1781837	1781737	0	Minus	Rv1576c	Possible phage phiRv1
1781292	1781193	0	Minus	Rv1576c	Possible phage phiRv1
1780682	1780587	5	Minus	Rv1576c	Possible phage phiRv1
1532953	1532854	0	Minus	Rv1361c	PPE FAMILY PROTEIN
1469637	1469538	0	Minus	Rv1313c	POSSIBLE TRANSPOSASE

Table 2 Unique SNPs of *M. tuberculosis* H37Rv and *M. tuberculosis* CDC1551

SNP position in H37Rv	SNP	SNP position in CDC1551
1789448	T-C	1780353
1789513	G-A	1780418
2266507	.-A	2267858
2266511	A-T	2267862
2266515	T-.	2267866
2266520	G-T	2267871
2266522	C-G	2267874
2338775	G-A	2341102

gene, resulting in a small insertion in the protein. For example, an 18-mer repetitive element was present in *M. tuberculosis* CDC1551 as two tandem copies. Three tandem copies of the same element was found in *M. tuberculosis* H37Rv (Figure 4a). Since the repetitive motif (18) is a multiple of 3 the reading frame is maintained. Deletions in *M. tuberculosis* CDC1551 involving PPE and PE-PGRS family (Rv0747, Rv01818c, Rv3343, Rv2741) followed an interesting pattern. The deletions involved a stretch of nucleotides containing repetitive sequences (Figure 4b), suggesting that recombination between the repeats may have been instrumental in the process of deletion. Some of the differences between the two strains of *M. tuberculosis*, reported here, were not identified by any previous study (Tables S1 and S2 in Supplementary Data). All the small repeats identified by ABWGC were essentially confirmed by tandem repeat finder (Table 4). The differences were minor due to doubling of the size of the repetitive sequences predicted by ABWGC and mismatches allowed in tandem repeats by tandem repeat finder.

An analysis was carried out with *M. tuberculosis* strains F11, C and Haarlem to check if the variations observed between the strains H37Rv and CDC1551 are also seen in other strains. Since strains F11, C and Haarlem are not fully sequenced, detailed analysis using ABWGC was not possible. In this analysis anchors defining all the known insertions of H37Rv and CDC1551 (*S*) were used as query sequences against sequences of strains F11, C and Haarlem (*T*). The inserts from CDC1551 genome were extracted and the presence/absence of the 46 inserted regions were observed in all the 3 *T* genomes. Twenty seven deletion sites were identical in size between H37Rv and F11 and variation was observed at two sites (Table 5). Strain C and Haarlem showed 18 and 15 identical deletion sites, respectively. This suggests that F11 is closest to H37Rv compared to other strains. This was further confirmed by taking inserted regions from H37Rv and comparing with the three strains.

Alteration in anchor order. Recombination events lead to reorganization in genomes which can be observed as conservation of genic synteny (38,39). Anchors can also be used to determine reorganization in genomes. The disruption in anchor order can easily be computed in a pairwise comparison. The anchors which map directly within transposable elements have been excluded from the analysis. The results presented clearly showed that such events occurred in mycobacterial genomes (Table 6). In CDC1551, the breakpoints in anchor order were due to transposable elements. Four breakpoints in anchor order in H37Rv were found in comparison to CDC1551. These regions map to genes Rv1316c, Rv2020c and Rv3020c. All of these genes contained repetitive sequences. Homologous recombination between these repetitive sequences may be responsible for the changes in anchor order (Table 6).

Analysis of the anchors in H37Rv and CDC1551 revealed a rearranged region in H37Rv. This region contained a phiRv1 prophage within the biotin operon (10). There is a repetitive element REP13E12 flanking the rearranged region. Biotin is required as the growth

Table 8. Partial list of insertions sites (size more than 100) in *M. tuberculosis* H37Rv and *M. avium* identified by ABWGC and MUMmer

H37Rv (ABWGC)	*M. avium* (ABWGC)	H37Rv (MUMmer)	*M. avium* (MUMmer)
274317–274819 274878–275039	– –	274206–275017	4075151–4075297
400011–400242 401167–401445	– –	399261–401994	4279373–4281116
490659–490760	–	479482–490852	4341052–4344911
561809–562411 562459–562728	– –	561712–562821	4425498–4425734
921821–921921	683450–683552 683619–683770	919347–923827	682088–686944

due to changes in the number of repetitive sequences compared to other currently available methods.

The results of analysis of *M. tuberculosis* strain H37Rv and *M. avium* are shown in Table 8. Only indels >100 nt are reported for comparison. The alignment of the two genomes by MUMmer is represented as clusters of MUMs. The regions between consecutive clusters are those which MUMmer failed to align. ABWGC was able to align these regions and identified probable positions of differences. For example, an indel of 990 nt was identified by MUMmer when H37Rv (starting from 399261) was aligned with *M. avium* (starting from 4279373). Our results showed two separate insertions in H37Rv of sizes 231 (PE family) and 278 (13E12 family). In addition to these genes, this region encoded aspartate aminotransferase gene in both genomes (percent identity of 87%). A partial list of some of the differences between the predictions from the two methods is given in Table 8. Overall it appears that ABWGC is a method of choice when divergent genomes have to be compared.

On simulated sequences. The results presented so far suggest that ABWGC is a better method for comparative analysis of genomes for finding indels, particularly in diverged genomes. This was further confirmed using a set of simulated sequences. The first simulation data contained DNA sequence of 3 kb in which 1 kb region was mutated randomly to make a divergent region. MUMmer identified the divergent region as an indel when the simulated data set was compared with the original sequence. In another simulated data set, six tandem copies of a repetitive motif was inserted at a given position. This sequence was compared with the same sequence but containing five copies of the repeat. MUMmer failed to identify the tandem copy correctly. Both the sequences were correctly analysed by ABWGC. In the third simulated example, two foreign sequences were inserted with a gap of 10 nt. MUMmer reported it as one large insertion whereas ABWGC reported as two insertions. The results confirmed that ABWGC is likely to be a better method for comparing genomes in order to find and characterize indels and expansion of repeats.

DISCUSSION

ABWGC fixes anchors randomly, followed by processing of the anchor coordinates to get comparative identification of diverged regions in terms of indels, repeats, inversion, etc. Many genome alignment tools also use anchors as the starting step. However, ABWGC differs from other methods in terms of the way anchors are generated and further processing of anchor information. Though the method is not capable of identification of all SNPs, it helps in identification of some of the SNPs that map within an anchor region. In this study, anchors totaling 10% of the total genome were sampled. Increasing the amount of sampling will increase coverage of the genome and it will be possible to find more SNPs. Our preliminary results have shown that some of the measures in ABWGC can be used to get a distance estimate between genomes (manuscript in preparation). Pairwise distance estimates can then be used to draw genome trees, a feature not present in any other genome alignment tools.

The results presented here using both real and simulated data clearly show that ABWGC is a method of choice in finding specific differences among genomes compared to other genome alignment methods. The nature of the indels is deciphered by further processing of the data using global alignment tools. Since random anchors usually flank indels, the alignment tools are able to find correctly the indels and their positions unlike in other methods. ABWGC is also useful in finding changes in genomes due to increase or decrease in the number of short repetitive sequences. MUMmer, AVID and GLASS use exact matches for finding anchors. It is difficult to find exact matches in diverged sequences. The problem is more serious for non-coding regions which are more divergent than coding regions. Since ABWGC allows mismatches it is possible to identify homologous anchors in a pair of diverged genomes.

Previous studies have indicated that SNPs and indels have played a significant role in the evolution of *M. tuberculosis* strains (40–46). Indels were found to be mainly due to insertion elements (10,47). The major finding of this study is the identification of all the indels varying from one to hundreds of nucleotides.

In this study, detailed analysis of indels of < 10 nt was not included. But our limited analysis has shown that, like SNPs these are important for overall functional diversity of an organism. For example, one nucleotide insertion in coding region can change the reading frame unless compensated otherwise. A large fraction of the indels was due to expansion of short repetitive motifs. The amplification/deletion of the repeat sequences is thought to be due to replication slippage or unequal recombination events or by single-stranded annealing pathway (47). These processes may be responsible for the observed repeat expansion. Presence of additional copies of these motifs would in principle change the proteome of mycobacterial cells as many of this map to the coding regions and have a subtle effect on clinical features. The surprising finding of this study is the large number of variations found in the two strains of *M. tuberculosis*. Since the strain *M. tuberculosis* H37Rv is being cultivated for a long time and the strain *M. tuberculosis* CDC1551 is relatively a new isolate, the variations may be due to long-term *in vitro* cultivation of *M. tuberculosis* H37Rv. In order to rule out this possibility, the strain *M. tuberculosis* H37Rv was compared with the recently cultivated isolates *M. tuberculosis* F11, C and Haarlem and results showed that *M. tuberculosis* H37Rv is much closer to *M. tuberculosis* F11 compared to *M. tuberculosis* CDC1551. Therefore, the observed indels in the strain *M. tuberculosis* H37Rv were not due to long-term cultivation.

A number of experimental and a few computational studies have been carried out to find genetic variations in different strains and isolates of *M. tuberculosis* (10). ABWGC was not only able to find all the reported differences but also a number of other variations that have not been reported so far. It will be interesting to see the distribution of these variations among a number of clinical isolates from patients with different clinical manifestations. Our studies agree with published results which clearly show that *M. tuberculosis* undergoes genomic changes characterized by SNPs, indels, repeat expansion, etc. These can explain the level of phenotypic diversity seen among clinical isolates (46,48). A number of studies in prokaryotes have shown that multiple copies of tandem repeats play critical role in the evolution of bacterial genome. Occasionally, these changes can be correlated with the phenotypic properties of the organism. For example, the analysis of multiple locus variable number of tandem repetitive sequences showed that the Dutch *Bordetella pertussis* strain rapidly changed in late 1990s (49). The change in repeat number has given a clue of evolution. These changes influence antigenic variation leading to altered virulence (49). Such tandem repetitive elements have also been reported in *Francisella tularensis* (50) and *Neisseria meningitidis* (51). Therefore, it appears that changes in the number of tandem repeats may be a general mechanism by which organism evolve. Our findings about *M. tuberculosis* may be part of an overall strategy for bacterial evolution. Identification of such differences can help in identifying new diagnostic markers and targets for vaccine and drug discovery. In conclusion, our results show that ABWGC can be a useful tool in comparative genomics, providing features that are not available in other genome analysis tools.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

This work was supported by Department of BioTechnology and R. R. was supported by grant from Department of Science and Technology. We thank Ram Ramaswamy, S. Bhattacharya and Vivek Thakur for discussion and valuable comments. The authors thank the anonymous referees for valuable suggestions. Funding to pay the Open Access publication charges for this article was waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Tomb,J.F., White,O., Kerlavage,A.R., Clayton,R.A., Sutton,G.G., Fleischmann,R.D., Ketchum,K.A., Klenk,H.P., Gill,S. *et al.* (1997) The complete genome sequence of the gastric pathogen Helicobacter pylori. *Nature*, **388**, 539–547.
2. Ferretti,J.J., McShan,W.M., Ajdic,D., Savic,D.J., Savic,G., Lyon,K., Primeaux,C., Sezate,S., Suvorov,A.N. *et al.* (2001) Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proc. Natl. Acad. Sci. USA*, **98**, 4658–4663.
3. Behr,M.A., Wilson,M.A., Gill,M.A., Salamon,H., Schoolnik,G.K., Rane,G.K. and Small,P.M. (1999) Comparative Genomics of BCG Vaccines by Whole Genome DNA Micoarray, *Science*, **284**, 1520–1523.
4. Shimizu,T., Ohtani,K., Hirakawa,H., Ohshima,K., Yamashita,A., Shiba,T., Ogasawara,N., Hattori,M., Kuhara,S. *et al.* (2002) Complete genome sequence of Clostridium perfringens an anaerobic flesh eater. *Proc. Natl. Acad. Sci. USA*, **99**, 996–1001.
5. Whittam,T.S. and Bumbaugh,A.C. (2002) Inferences from whole genome sequences of bacterial pathogens, *Curr. Opin. Genet. Dev.*, **12**, 719–725.
6. Wizemann,T.M., Heinrichs,J.H., Adamou,J.E., Erwin,A.L., Kunsch,C., Choi,G.H., Barash,S.C., Rosen,C.A., Masure,H.R. *et al.* (2001) Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* Infection. *Infect. Immun.*, **69**, 1593–1598.
7. Jin,Q., Yuan,Z., Xu,J., Wang,Y., Shen,Y., Lu,W., Wang,J., Liu,H., Yang,J. *et al.* (2002) Genome sequence of *Shigella flexneri* 2a, insights into pathogenecity through comparison with genomes of Escherichia coli K12 and O157. *Nucleic Acids Res.*, **30**, 4432–4441.
8. Beres,S.B., Sylva,G.L., Barbian,K.D., Lei,B., Hoff,J.S., Mammarella,N.D., Liu,M.Y., Smoot,J.C., Porcella,S.F. *et al.* (2002) Genome sequence of serotype M3 strain of group A Streptococcus, phage encoded toxins, the high virulence phenotype and clone emergence, *Proc. Natl. Acad. Sci. USA*, **99**, 10078–10083.
9. Hayashi,T., Makino,K., Ohnishi,M., Kurokawa,K., Ishii,K., Yokoyama,K., Han,C.G., Ohtsubo,E., Nakayama,K. *et al.* (2001) Complete genome sequence of enterohemorrahagic *Escherichia coli* O157,H7 and genomic comparison with a laboratory strain K-12. *DNA Res.*, **8**, 11–22.
10. Fleischman,R.D., Alland,D., Eisen,J.A., Carpenter,L., White,O., Peterson,J., DeBoy,R., Dodson,R., Gwinn,M. *et al.* (2002) Whole genome comparison of Mycobacterium tuberculosis clinical and laboratory strains. *J. Bacteriol.*, **184**, 5479–5490.
11. Mostowy,S., Inwald,J., Gordon,S., Martin,C., Warren, R., Kremer, K., Cousins,D. and Behr,M.A. (2005) Revisiting the evolution of *Mycobacteria bovis*. *J. Bacteriol.*, **187**, 6386–6395.

12. Kato-Maeda,M., Rhee,J.T., Gingeras,T.R., Salamon,H., Drenkow,J., Smittipat,N. and Small,P.M. (2001) Comparing genomes within the species *Mycobacterium tuberculosis*. *Genome Res.*, **11**, 547–554.
13. 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**, 1520–523.
14. Mostowy,S., Cousins,D., Brinkman,J., Aranaz,A. and Behr,M.A. (2002). Genomic deletions suggest a phylogeny for the Mycobacterium tuberculosis complex. *J. Infect. Dis.*, **186**, 74–80.
15. Brosch,R., Gordon,S.V., Marmiesse,M., Brodin,P., Buchrieser,C., Eiglmeier,K., Garnier,T., Gutierrez,C., Hewinson,G. *et al.* (2002). A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc. Natl. Acad. Sci. USA*, **99**, 3684–3689.
16. Alland,D., Whittam,T.S., Murray,T.S., Cave,M.D., Hazbon,M.H., Dix,K., Kokoris,M., Duesterhoeft,A., Eisen,J.A. *et al.* (2003) Modelling bacterial evolution with comparative genome based marker systems, Application to *Mycobacterium tuberculosis* evolution and pathogenesis. *J. Bacteriol.*, **185**, 3392–3399.
17. Gordon,S.V., Brosch,R., Billault,A., Barrell,B. and Cole,S.T. (1999) New insertion sequences and a novel repeated sequences in the genome of *Mycobacterium tuberculosis* H37Rv. *Microbiology*, **145**, 881–892.
18. Kremer,K., Soolingen,D.V., Frothingham,R., Haas,W.H., Hermans,P.W.M., Martn,C., Palittapongarnpim,P., Plikaytis,B.B., Riley,L.W. *et al.* (1999) Comparison of methods based on different molecular epidemiological markers for typing of *Mycobacterium tuberculosis* complex strains: interlaboratory study of discriminatory power and reproducibility. *J. Clin. Microbiol.*, **37**, 2607–2618.
19. Kamerbeek,J., Schouls,L., Kolk,A., van Agterveld,M., van Soolingen,D., Kujjper,S., Bunschoten,A., Molhuizen,H., Shaw,R. *et al.* (1997). Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J. Clin. Microbiol.*, **35**, 907–914.
20. Frothingham,R. and Meeker-OConnell,W.A. (1998). Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers tandem repeats. *Microbiology*, **144**, 1180–1196.
21. Tsolaki,A.G., Hirsh,A.E., DeRiemer,K., Enciso,J.A., Wong,M.Z., Hannan,M., Yves-Olivier,L., de la Salmoniere,G., Aman,K. *et al.* (2004). Functional and evolutionary genomics of Mycobacterium tuberculosis: insights from genomic deletions in 100 strains. *Proc. Natl. Acad. Sci. USA*, **101**, 4865–4870.
22. Gagneuxa,S., DeRiemerb,K., Vanb,T., Kato-Maedab,M., de Jongb,B.C., Narayanang,S., Nicolh,M., Niemann,S., Kremerj,K. *et al.* (2006). Variable hostpathogen compatibility in Mycobacterium tuberculosis, *Proc. Natl. Acad. Sci. USA*, **103**, 2869–2873.
23. Morgenstern,B., Frech,K., Dress,A. and Werner,T. (1998). DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics*, **14**, 290–294.
24. Jareborg,N., Birney,E. and Durbin,R. (1999) Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pair. *Genome Res.*, **9**, 815–824.
25. Brudno,M., Do,C.B., Cooper,G.M., Kim,M.F., Davydov,E., NISC Comparative Sequencing Program, Green,E.D., Sidow,A. and Batzoglou,S. (2003) LAGAN and multi-LAGAN, efficient tools for large-scale multiple alignment of genomic DNA, *Genome Res.*, **13**, 721–731.
26. Batzoglou,S., Pachter,L., Mesirov,J.P., Berger,B. and Lander,E.S. (2000) Human and mouse gene structure, comparative analysis and application to exon prediction. *Genome Res.*, **10**, 950–958.
27. Bray,N., Dubchak,I. and Pachter,L. (2003). AVID: A global alignment program. *Genome Res.*, **13**, 97–102.
28. Chain,P., Kurtz,S., Ohlebusch,E. and Slezak,T. (2003) An applications-focused review of comparative genomics tools: capabilities, limitations and future challenges. *Brief. Bioinformatics*, **4**.
29. Delcher,A.L., Kasif,S., Fleischmann,R.D., Peterson,J., White,O. and Salzberg,S.L. (1999) Alignment of whole genomes. *Nucleic Acids Res.*, **27**, 2369–2376.
30. Kent,W.J. and Zahler,A.M. (2000) Conservation, regulation, synteny, and introns in a large scale *C.briggsae-C-elegans* genomic alignment. *Genome Res.*, **10**, 1115–1125.
31. Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D. and Miller,W. (2003) Human-mouse alignments with BLASTZ. *Genome Res.*, **13**, 103–107.
32. Ning,Z., Cox,A.J. and Mullikin,J.C. (2001) SSAHA: a fast search method for large DNA databases. *Genome Res.*, **11**, 1725–1729.
33. Suyama,M. and Bork,P. (2001). Evolution of prokaryotic gene order: genome rearrangements in closely related species. *Trends Genet.*, **17**, 10–13.
34. Needleman,S.B. and Wunsch,C.D. (1970) A general method application to search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.*, **48**, 442–453.
35. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
36. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
37. McHugh,T.D. and Gillespie,S.H. (1998) Nonrandom association of IS6110 and *Mycobacterium tuberculosis*, implications for molecular epidemiological studies. *J. Clin. Microbiol.*, **36**, 1410–1413.
38. Pal,C. and Hurst,L.D. (2003) Evidence of co-evolution of gene order and recombination events. *Nature Genet.*, **33**, 392–395.
39. Akhunov,E.D., Akhunov,A.R., Linkiewicz,A.M., Dubcovsky,J., Hummel,D., Lazo,G.R., Chao,S., Anderson,O.D., David,J. *et al.* (2003) Synteny perturbations between wheat homeologous chromosomes caused by locus duplications and deletions correlate with recombination rates, *Proc. Natl. Acad. Sci. USA*, **19**, 10836–10841.
40. Cole,S.T., Brosch,R., Parkhill,J., Garnier,T., Churcher,C., Harris,D., Gordon,S.V., Eigimeier,K., Gas,S. *et al.* (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, **393**, 537–544.
41. Warren,R.M., Sampson,S.L., Richardson,M., van der Spuy,G.D., Lombard,C.J., Victor,T.C. and van Helden,P.D. (2000) Mapping of IS6110 flanking regions in clinical isolates of *Mycobacterium tuberculosis* demonstrates genome plasticity. *Mol. Microbiol.*, **37**, 1405–1416.
42. Gordon,S.V., Brosch,R., Billault,A., Garnier,T., Eiglmeier,K. and Cole,S.T. (1999) Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Mol. Microbiol.*, **32**, 643–655.
43. Sreevatsan,S., Pan,X., Stockbauer,K.E., Connelldagger,N.D., Kreiswirth,B.N., Whittam,T.S. and Musser,J.M. (1997) Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionary recent global dissemination, *Proc. Natl. Acad. Sci. USA*, **19**, 9869–9874.
44. Brosch,R., Pym,A.S., Gordon,S.V. and Cole,S.T. (2001) The evolution of mycobacterial pathogenicity, clues from comparative genomics. *Trends Microbiol.*, **9**, 452–458.
45. Maeda,M.K., Rhee,J.T., Gingeras,T.R., Salamon,H., Drenkow,J., Smittipat,N. and Small,P.M. (2001) Comparing genomes within the species *Mycobacterium tuberculosis*. *Genome Res.*, **11**, 547–554.
46. Manca,C., Tsenova,L., Barry III,C.E., Bergtold,A., Freeman,S., Haslett,P.A.J., Musser,J.M., Freedman,V.H. and Kaplan G. (1999) *Mycobacterium tuberculosis* CDC1551 induces a more vigorous host response in vivo and in vitro, but in not more virulent than other clinical isolates. *J. Immunol.*, **162**, 6470–6746.
47. Bzymek,M. and Lovett,S.T. (2001) Instability of repetitive DNA sequences, the role of replication in multiple mechanisms. *Proc. Natl. Acad. Sci. USA*, **98**, 8319–8325.
48. Valway,S.E., Sanchez,M.P.C., 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 *Mycobacteria tuberculosis*. *N. Engl. J. Med.*, **338**, 633–639.
49. Schouls,L.M., Van der Heide,H.G.J., Vauterin,L., Vauterin,P. and Mooi,F.R. (2004). Multiple locus variable-number tandem repeat analysis of Dutch *Bordetella pertussis* strains reveals

rapid genetic changes with clonal expansion during the late 1990s. *J. Bacteriol.*, **186**, 5496–5505.
50. Johansson,A., Farlow,J., Larsson,P., Dukerich,M., Chambers,E., Bystrom,M., Fox,J., Chu,M., Forsman,M. *et al.* (2004) Worldwide genetic relationships among *Francisella tularensis* isolates determined by multiple locus variable-number tandem repeat analysis. *J. Bacteriol.*, **186**, 5808–5818.
51. Schouls,L.M., van der Ende,A., Damen,M. and van de Pol,I. (2006) Multiple-locus variable-number tandem repeat analysis of *Neisseria meningitidis* yields groupings similar to those obtained by multilocus sequence typing. *J. Clin. Microbiol.*, **44**, 1509–1518.

WEBSITE REFERENCE

1. The NCBI Genome Page [http,//www.ncbi.nlm.nih.gov/genomes/lproks.cgi]
2. The Broad Institute [http,//www.broad.mit.edu/annotation/genome/mycobacteriumtuberculosis-spp/MultiDownloads.html]
3. The Mersenne Twister Home Page [http,//www.math.sci.hiroshima-u.ac.jp/ mmat/MT/emt.html]

Table 3. Partial list of genomic alterations between *M. tuberculosis* CDC1551 and *M. tuberculosis* H37Rv

Nature of event	Start position	Size[a]	CDS	Function
(A) *M. tuberculosis* CDC1551 compared to *M. tuberculosis* H37Rv				
Insertion	71529	36		Intergenic region (2)
Insertion	150882	179	MT0132	PE PGRS family protein (1)
Insertion	483384	1357	MT0413	IS6110 (2)
Insertion	483384	1357	MT0414	IS6110 (2)
Insertion	483384	1357	MT0415	Hypothetical protein (2)
Insertion	624648	83	MT0556	PE-PGRS family protein (1)
Insertion	744075	532	MT0676	Glycosyl hydrolase, family 5 (1)
Insertion	804401	215	MT0730	50S ribosomal protein L23 (2)
Insertion	804401	215	MT0731	50S ribosomal protein L2 (2)
Insertion	960065	105		Intergenic region (2)
Insertion	1094076	11	MT1006.1	PE PGRS family protein (2)
Insertion	1096183	11	MT1008	PE PGRS family protein (2)
Insertion	1121702	14	MT1033	Hypothetical protein (1)
Insertion	1191499	192	MT1097	PE-PGRS family protein (1)
Insertion	1213836	44		Intergenic region(1)
Insertion	1442915	55		Insertion lies in the intergenic region(2)
Insertion	1480513	1674	MT1360	Adenylate cyclase (1)
Insertion	1612509	21	MT1479	Hypothetical protein (1)
Insertion	1632400	26	MT1497.1	PE-PGRS family protein (1)
(B) *M. tuberculosis* H37Rv compared to *M. tuberculosis* CDC1551				
Expansion of repeat	24704	18 (2)	Rv0020c	Hypothetical protein (2)
Insertion	32351	36	Rv0029	Hypothetical protein (1)
Insertion	206812	56	Rv0175	Probable conserved mce associated membrane protein (1)
Expansion of repeat	335812	18 (3)	Rv0278c	PE-PGRS family protein (2)
Insertion	337806	32	Rv0279c	PE-PGRS family protein (2)
Expansion of repeat	427312	15 (4)	Rv0355c	PPE family protein (2)
Expansion of repeat	428188	59 (3)	Rv0355c	PPE family protein (2)
Insertion	577286	57	Rv0487	Hypothetical protein (2)
Insertion	577286	57	Rv0488	Probable conserved integral membrane protein (2)
Insertion	840167	47	Rv0747	PE-PGRS family protein (1)
Insertion	1212109	77	Rv1087	PE-PGRS family protein (2)
Insertion	1217495	653	Rv1091	PE-PGRS family protein (1)
Insertion	1267172	57		Insertion lies in the intergenic region
Insertion	1895353	113		Insertion lies in the intergenic region
Insertion	2062036	89	Rv1818c	PE-PGRS family protein (1)
Insertion	2074436	111	Rv1829	Hypothetical protein (2)
Insertion	2074436	111	Rv1830	Hypothetical protein (2)
Expansion of repeat	2163731	69 (3)	Rv1917c	PPE family protein (2)

Reported by
(1) Fleischman *et al.*, 2002.
(2) ABWGC.
[a]Number of copies is shown in brackets.

supplement in some strains of *M. tuberculosis*. Horizontal gene transfer may have played a role in the integration of the prophage and this event may have occurred before *M. tuberculosis* became intracellular (40). The prophage intergration may be the cause of rearrangement of the genomic region in H37Rv. This region was also identified as a flipped region where the homologous anchor mapped to the complementary strand.

Analysis of two divergent species *M. tuberculosis* and *M. avium*

In order to find out if homologous anchors can be unambiguously identified in divergent genomes the method was applied to two different species of Mycobacterium, namely *M. tuberculosis* H37Rv and *M. avium*. The results of the analysis are shown in Figure 5. The majority of anchor scores were above 10, showing that the two genomes were not closely related. The analysis revealed presence of 1811 indels in H37Rv of size >100 nt. Moreover, four inverted regions were identified. Some of the results are shown in the Supplementary Table S3. It is clear from the analysis that it is possible to apply anchor-based method to the analysis of genomes of different species.

Evaluation of Performance

Mycobacterial genomes. A number of tools for comparing genomes have been reported. These include DIALIGN (23), DBA (24), LAGAN (25), GLASS (26), AVID (27), MUMmer (29), WABA (30), BLASTZ (31) and SSAHA (32). Out of all these MUMmer has been extensively used to align closely related bacterial genomes, such as strains

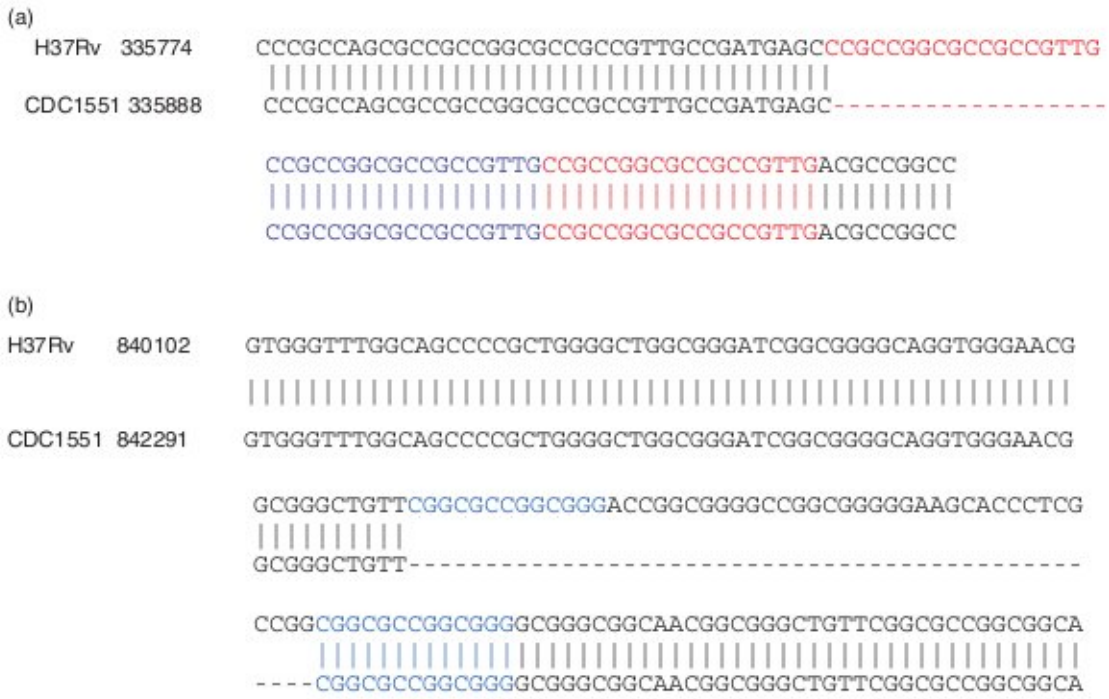

Figure 4. Variation in repetitive motifs in *M. tuberculosis* H37Rv in comparison with *M. tuberculosis* CDC1551. Homologous inter-anchor regions of *M. tuberculosis* CDC1551 and *M. tuberculosis* H37Rv were aligned. Repetitive motifs were identified automatically. (**a**) The 18-mer repetitive motif is highlighted in red and blue colour; (**b**) Deletion of segment of nucleotides containing repetitive elements. Repetitive elements are highlighted in blue colour.

Table 4. The tandem repeats identified in *M. tuberculosis* CDC1551 by ABWGC and tandem repeat finder

ABWGC			Tandem repeat finder		
Indices	Size	Copy number	Indices	Size	Copy number
1121703–1121718	15	4	1121702–1121771	15	4.7
1612487–1612508	21	2	1612426–1612545	21	5.7
1946383–1946440	57	2	1946383–1946530	57	2.6
1974051–1974207	78	8	1973745–1974339	78	7.6
2143312–2143357	45	2	2143312–2143406	45	2.1
2160645–2160921	69	8	2160645–2161098	69	6.6
3730841–3730859	18	2	3730827–3730875	9	5.4
3940714–3940750	36	2	3940650–3940756	18	5.9
3941039–3941114	75	2	3941039–3941193	75	2.1
4149054–4149113	59	3	4149054–4149281	59	3.9

of *M. tuberculosis* (10). The alignments can then be used to identify and document the differences, such as SNPs, indels, duplications and inversions. The results obtained using ABWGC were compared with those obtained by LAGAN and MUMmer. For this, different *M. tuberculosis* strains and species were used. The results are shown in Table 7. MUMmer was used with default settings (run-mummer 3 was used for the analysis). While it identified 86 indels in the strain H37Rv (10), ABWGC and LAGAN (25) located 104 such events, which included the indels identified by MUMmer. MUMmer identifies indels as the regions between two consecutive MUMs that cannot be aligned. When two strains of *M. tuberculosis* were aligned by MUMmer, some overlapping MUMs were observed.

Table 5. Presence of indels in different strains of *M. tuberculosis*

Size[S]	H37Rv	F11	C	Haarlem
(A) Insertion present in *M. tuberculosis* CDC1551[a]				
37(71529)	P	N	N	N
180(150882)	P	NA	N	N
103(424203)	P	N	P	P
83(624648)	P	N	N	N
533(744075)	P	N	N	N
216(804401)	P	P	P	P
106(960065)	P	P	P	N
12(1094076)	P	P	N	N
12(1096183)	P	P	P	P
15(1121703)	P	P	N	P
119(1191499)	P	P	N	N
45(1213836)	P	P	N	P
56(1442915)	P	P	N	P
21(1612487)	P	N	N	N
26(1632401)	P	N	NA	NA
207(1633340)	P	P	P	P
53(1644353)	P	N	N	P
10(1885207)	P	N	N	N
57(1946384)	P	N	N	N
156(1974210)*	P	P	P	P
6808(1978716)	P	P	P	P
15(2130692)	P	N	N	N
45(2143313)	P	N	N	N
276(2160446)*	P	P	P	N
5000(2266058)	P	N	N	N
115(2400403)	P	P	N	N
940(2629977)	P	P	P	P
767(2633468)	P	P	P	P
21(2701714)	P	N	N	N
676(2862694)	P	P	P	N
55(2985372)	P	P	P	N

(continued)

Table 5. Continued

Size[S]	H37Rv	F11	C	Haarlem
71(3114712)	P	P	N	N
59(3331026)	P	P	P	P
68(3418973)	P	P	P	N
2148(3524160)	P	P	P	N
4059(3705273)	P	N	N	P
18(3730860)	P	P	P	P
78(3733427)	P	N	P	N
75(3926618)	P	P	NA	NA
15(3928764)	P	P	NA	NA
35(3940688)	P	P	NA	NA
75(3940715)	P	P	NA	NA
117(3942726)	P	P	P	P
18(4086509)	P	P	P	NA
Size[S]	**CDC1551**	**F11**	**C**	**Haarlem**
(B) Insertion present in *M. tuberculosis* H37Rv[a]				
18(24699)	P	P	P	P
37(32351)	P	N	P	N
57(206812)	P	N	N	N
18(335811)	P	P	NA	NA
33(337805)	P	N	NA	NA
15(427311)	P	P	P	P
60(428187)	P	N	P	N
58(577286)	P	N	P	P
48(840166)	P	N	NA	NA
874(886541)	P	N	N	N
79(1212108)	P	N	NA	NA
654(1217494)	P	N	P	P
90(2062023)	P	N	P	NA
75(2165327)	P	N	N	N
23(2180797)	P	N	P	P
57(2372437)	P	P	N	N
2273(2381411)	P	N	N	N
499(2704307)	P	N	N	N
213(3054706)	P	P	NA	N
54(3171467)	P	N	P	N
312(3501334)	P	P	N	N
63(3663826)	P	P	P	P
3197(3732759)	P	P	N	N
46(3935411)	P	P	NA	NA
189(3739700)	P	N	N	NA
30(3936241)	P	P	NA	NA
32(3934871)	P	P	NA	NA
640(3955464)	P	N	N	N
18(4359134)	P	P	P	P

[a]Insertion sites used in this analysis were defined by pairwise analysis of the two strains H37Rv and CDC1551.
[S]Insertion site.
P: The presence of deletion.
N: Absence of deletion.
NA: Data not available.
*: *M. tuberculosis* H37Rv and *M. tuberculosis* F11 differ in deletion pattern.

These overlapping MUMs correspond to repetitive elements, and exact position could not be located. The output from LAGAN and MUMmer could not be processed to reveal that the insertions, were due to change in copy number of small tandem-repetitive motif. The local alignment tool BLASTZ was also used to test its performance. The indels identified by BLASTZ were fragmented as many short indels rather than a single large indel. For example, an insertion of 180 nt at 150882 in *M. tuberculosis* CDC1551 was identified by ABWGC and MUMmer but not by BLASTZ.

Table 6. Alteration in anchor order in *M. tuberculosis* H37Rv as compared with *M. tuberculosis* CDC1551

Preceding anchor in CDC1551/H37Rv	Anchor in CDC1551/H37Rv	Succeeding anchor in CDC1551/H37Rv
1481173/1481632	1482185/1480970	1483260/1482045
2265308/2268165	2267841/2266488	2271735/2269402
2629407/2633562	2631401/2633513	2634602/2637271
3887611/3893778	3888939/1532953	3890381/3895380
4244934/4252621	4245184/1469637	4245791/4253468

CDC1551 — *M. tuberculosis* CDC1551.
H37Rv — *M. tuberculosis* H37Rv. The numbers represent position in the respective genome.

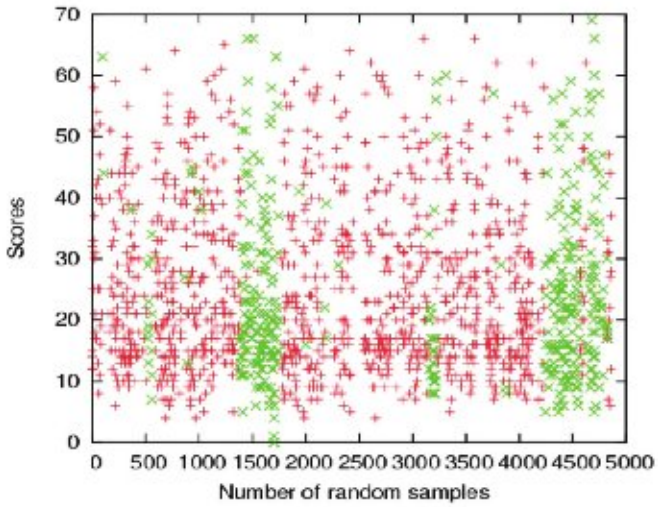

Figure 5. Distribution of individual mismatch score of anchors of *M. tuberculosis* H37Rv compared to *M. avium*. For every randomly picked anchor from *M. tuberculosis* H37Rv (*S*) and its homologous sequence in the *M. avium* (*T*), a mismatch score was computed as described in the Methods section. The individual scores were plotted against anchors that have been extracted sequentially in terms of position in the genome (*S*). Homologous anchors on positive strand are shown by plus and those in complementary strand by cross.

Table 7. Repeats detected by ABWGC but not by MUMmer and LAGAN. Data represent comparison of *M. tuberculosis* H37Rv with *M. tuberculosis* CDC1551

Site of repeat	Size of repeat	Number of copies present
24699	18	2
335812	18	3
427312	15	4
428188	60	2
2163731	69	5
2165328	75	3
2347415	58	3
3171467	54	2
33663826	63	2
3948753	603	2

For closely related genomes such as H37Rv and CDC1551, complete analysis by ABWGC took <3 min on Pentium IV 3 GHz CPU. The results clearly showed that ABWGC is better in finding indels particularly those