SCIENTIFIC REPORTS

SUBJECT AREAS:
BIOINFORMATICS
COMPUTATIONAL BIOLOGY
BIOTECHNOLOGY
MICROBIOLOGY

Received
16 January 2012

Accepted
8 February 2012

Published
2 March 2012

Correspondence and requests for materials should be addressed to H.K.P. (hkp1000@gmail.com) or A.B. (alok.bhattacharya@gmail.com)

Identification of Hot and Cold spots in genome of *Mycobacterium tuberculosis* using Shewhart Control Charts

Sarbashis Das[1], Priyanka Duggal[2], Rahul Roy[3], Vithal P. Myneedu[4,5], Digamber Behera[5], Hanumanthappa K. Prasad[2] & Alok Bhattacharya[1,6]

[1]School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India, [2]Department of Biotechnology, All India Institute of Medical Sciences (AIIMS), New Delhi, India, [3]Indian Statistical Institute, New Delhi, India, [4]Department of Microbiology, [5]LRS Institute of Tuberculosis and Respiratory Diseases, New Delhi, India, [6]School of Life Sciences, Jawaharlal Nehru University, New Delhi, India.

The organization of genomic sequences is dynamic and undergoes change during the process of evolution. Many of the variations arise spontaneously and the observed genomic changes can either be distributed uniformly throughout the genome or be preferentially localized to some regions (hot spots) compared to others. Conversely cold spots may tend to accumulate very few variations or none at all. In order to identify such regions statistically, we have developed a method based on Shewhart Control Chart. The method was used for identification of hot and cold spots of single-nucleotide variations (SNVs) in *Mycobacterium tuberculosis* genomes. The predictions have been validated by sequencing some of these regions derived from clinical isolates. This method can be used for analysis of other genome sequences particularly infectious microbes.

Genomic variations, such as single nucleotide variations (SNVs), insertion/deletion, copy number changes and changes in synteny are some of the major causes of genetic divergence and phenotypic differences among different strains and species[1]. Though the identification of these variants has become much easier and the underlying molecular mechanisms are getting revealed, it is still not clear if there is a pattern by which genomic changes occur[2]. Hot spots and cold spots are regions that display either higher or lower SNVs respectively, compared to the predicted normal frequencies[3]. Traditionally hot spots have been studied with respect to recombination frequencies and specific octamer DNA sequences (*e.g.* Chi sites 5′-GCTGGTGG-3′) were thought to be associated with these spots[4]. Most of the mutational analyses have been done with either individual genes such as *p53*[5–7] and some kinases[8], small genomes such as viruses[9–11], or extra chromosomal DNA elements like mitochordria[12] and chloroplast[13]. In general hot spots have been defined in terms of frequencies of variants arising at a single nucleotide position or a single amino acid. The frequencies are known to vary across genomes. Selection pressure, such as drug or immune pressure[14] plays an important role in determining which genomic regions that are likely to harbor the hot spots. Moreover, sites/genes that are hyper variable may be governed by evolution and are a result of the intricate relationships among genes, networks and environment[15]. Since identification of hot and cold spots can be highly useful in defining drug and vaccine targets it is important to develop tools that can identify these regions systematically. Of the few computational approaches available for identification of hot and cold spots, mutation spectrum analysis is one approach[16]. A mutation spectrum is a distribution of frequencies of every type of mutation along nucleotide sequences of a target gene. This is then transformed into distribution of observed mutational frequencies and compared with expected frequencies. However, these methods are designed for finding hotspot sites in a gene but not for scanning entire genomes in a short period of time. Moreover, there are a number of methods that can be used for mutational frequency analysis and it has been suggested that a combination of methods are needed to accurately identify hot spot sites[16]. Here we describe an approach based on "Shewhart Control Chart" for analysis of whole genome sequences of different strains of *Mycobacterium tuberculosis*, the causative agent of tuberculosis. Shewhart control chart is widely used in statistical quality control[17]. It has also been used in quality estimation in healthcare industries[18,19].The predictions we have made by using this method were validated by sequencing the putative regions amplified from clinical isolates.

Tuberculosis continues to be a major public health problem of the world[20]. It is an air borne infection and manifests predominantly as a pulmonary disease. Besides pulmonary tuberculosis, it can also occur, though less

frequently at extra-pulmonary sites. The strains isolated from both these clinical conditions have been investigated in the present study. In addition to variation in gene expression patterns during infection, intra-genomic variation among pathogenic strains has been recognized as a critical feature in pathogenesis of microorganisms.

Results

Hot & cold spots prediction using Shewhart Control Chart. Shewhart Control Chart (SCC) is one of the most popular techniques for maintaining process control in the field of statistical process control[17]. This chart is routinely used to monitor one or more variables that are directly or indirectly associated with the production process. This chart may instantly detect a large shift in the process level. Regardless of how carefully a process is maintained, a certain amount of natural variability does always exist. A process is said to be statistically "in control" when the amount of natural variability is within a certain limit. On the other hand, if the variability exceeds a certain limit, then the process is statistically "out of control". This chart graphically displays the quality of product or process based on characteristics of a sample in relation to sample number or time. Basic characteristics of these charts are Center Line (CL), the Upper Control Limit (UCL) and Lower Control Limit (LCL). In effect the use of Shewhart Control Chart in statistical process control mainly ensures that the statistical attributes of the process lie within the UCL & LCL. In our case SNV frequencies in the genome falling outside the control limit will satisfy hot spots. (see "Methods" for details)

We have defined hot and cold spots as regions of genomes (windows of 2000 nucleotides) that either display higher or lower than expected number of SNVs respectively in a population of isolates/strains. We have used ABWGAT (Anchor Based Whole Genome Analysis Tool) to carry out pair wise comparison of fully sequenced *M. tuberculosis* genomes in order to identify SNVs[21]. The distribution of SNVs identified by comparing *M. tuberculosis* CDC1551 and *M. tuberculosis* H37Rv strains across the genome is shown in Fig. 1. *M. tuberculosis* H37Rv strain was used as a reference strain. SNV counts were plotted using non-overlapping segment of 2000 nucleotides. We have also generated random SNVs and the positions of these are also depicted in Fig. 1. It is clear from the figure that the distribution of natural SNVs was non-uniform in comparison to randomly generated ones. The number of SNVs in a segment of 2000 is estimated to have a Poisson distribution with mean 0.4954. This was verified statistically by a Kolmogorov-Smirnov[22] test which yielded a D-statistic of 0.0337. (see Fig. 2).

SNVs generated by comparing two *M. tuberculosis* strains (see Fig. 1) were used to derive SCC(Fig. 3). The chart shows UCL, CL and LCL as dotted lines. The red color indicates out-of-control processes, that is, the genomic regions with high SNV frequencies. We have identified cold spots as those that show very few or negligible SNVs. In order to extend the studies to clinical isolates of *M. tuberculosis*, we have used two different strategies. In the first one we identified putative hot and cold spots from pair wise comparison of different strains and isolates using SCC and then mapped these with respect to each other in order to identify the common regions based on sequence. Only those regions that showed hot and cold spots in all the strains were considered for further analysis. In the second strategy, we considered all SNVs in all strains and isolates and mapped these to H37Rv sequence (reference sequence). This facilitated the generation of an average number of SNVs in each bin in the context of H37Rv. SCC of the binned average SNVs permitted the identification of hot and cold spots, (Supplementary Table 1, 2). Our results showed a total of 44 hot spots and 32 cold spots in *M. tuberculosis* genome. Some of the genes, in the hot spot regions, such as *Rv0064* and *Rv0095c* have been functionally characterized; however a large number of genes with unknown function are also located in these regions (hypothetical proteins). *Rv0064* has been annotated as a probable transmembrane protein based on sequence similarity with integral membrane proteins. A homolog of *Rv0064*, *(ML0644)* has been described as a conserved hypothetical transmembrane protein in *M. leprae* (http://www.ncbi.nlm.nih.gov/gene/909429).

In our analysis we did not consider those nucleotide variations of the reference strain H37Rv that are absent in all other strains and isolates. We also did not consider SNVs that mapped to multigene families, such as PPE/PGRS, and repetitive regions as these can skew SNV count.

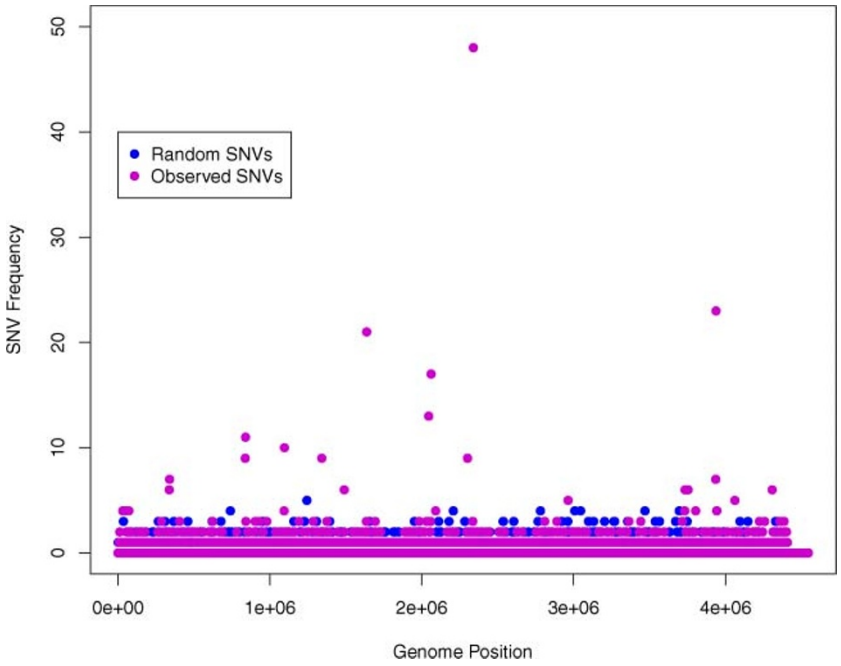

Figure 1 | Distribution of SNVs across whole genome. Pink dots indicate frequency of SNVs identified by comparison between *M. tuberculosis* CDC1551 and *M. tuberculosis* H37Rv were mapped on H37Rv genome using a bin size of 2000 nucleotides. Blue dots indicate distribution of randomly generated SNVs on H37Rv genome. X-axis represents whole genome position. Y-axis represents SNV frequency.

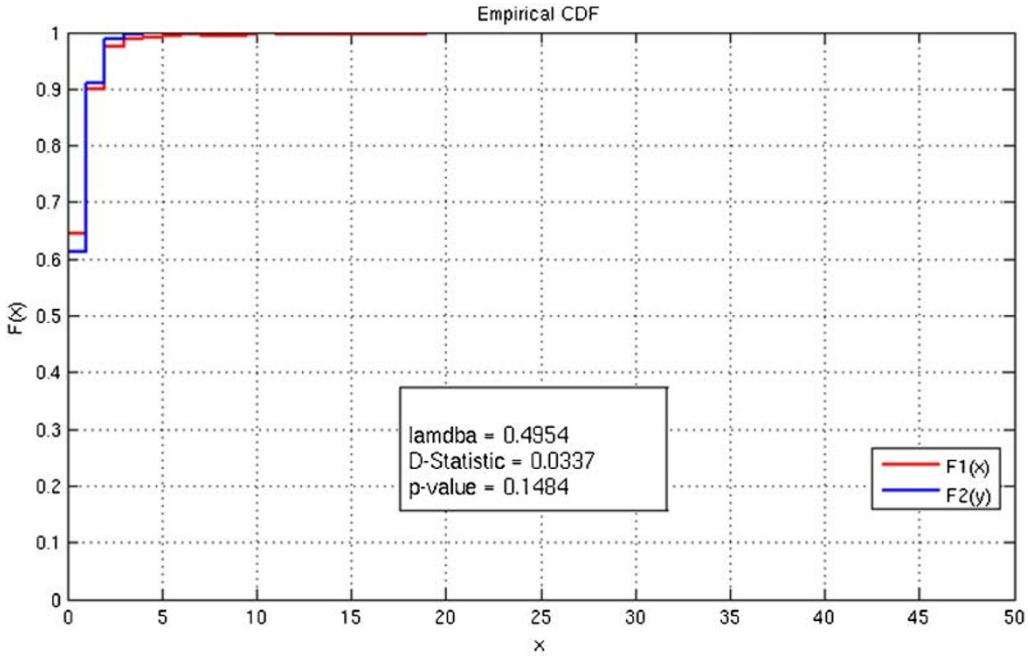

Figure 2 | Kolmogorov-Smirnov test to check if SNV distribution follows Poisson distribution. The function F1 is the empirically observed cumulative distribution of the SNVs and the function F2 is the cumulative distribution of a Poisson random variable with parameter 0.4954.

Sequencing of hot and cold spots of clinical isolates. Our identification of hot and cold spots is based on completely sequenced genomes. Though we have also taken into consideration sequences from *M. tuberculosis* isolates that have not been assembled, it is still likely that the changes observed by us may be present only in these selected isolates and not relevant in a global context. We tested the methodology for its reliability and robustness to predict hyper and hypovariable regions by sequencing two representative predicted hot and cold spot regions from a large number of clinical isolates. While the hot spot regions displayed 38 and 4 SNVs in *Rv0095c* and *Rv0064* respectively per 500 nucleotides in 40 isolates, no SNV was detected in the cold spots of any isolate, validating the strategy used in the present study for prediction of hot and cold spots. Multiple sequence alignments of a part of the sequenced regions of one of the hot spots and cold spots of clinical isolates are shown in Fig. 4. We have also analyzed published data on SNP distribution in 89 individual genes from 99 human adapted *M. tuberculosis* strains[23]. GyrB, that falls in a hot spot was among the genes sequenced and displayed 15 SNPs. On the other hand there were 3 SNPs in one of the cold spot genes PstS1. These results validate predictions made using SCC.

Discussion

Genome sequence divergence facilitates organisms to adapt to varying environmental conditions[24]. Periodic alternation and fluctuations in the host milieu are acknowledged features which pathogenic microorganisms encounter following infection. For example, the transition in microenvironment of the infectious tubercle bacilli, from the droplet state (free living) to the intracellular environment of the host macrophage is demanding, requiring efficient adaptation. Elucidating patterns in genome variations can help in establishing comprehensive strategies in formulating appropriate vaccines and therapeutics against pathogens[25]. Development of new sequencing technologies has made available genome sequences from a large number of organisms, particularly different species/strains and isolates. These provide major resources for deriving patterns of genome variations. Genome sequencing also provides a simple way to map mutations and this has tremendous potential in mapping drug resistance[26]. Identification of the regions that either have SNV clusters (hot spots) or lack any SNVs (cold spots) can lead to knowledge about rapidly evolving or conserved regions of genomes. In this article we have described a simple computational method which can be used to identify hot and cold spots in sequenced genomes. For this, we have analyzed fully assembled genomic sequences of laboratory strains and non assembled next generation sequence data of twenty isolates from a recent study to derive a composite prediction[27]. We provide experimental evidence to support our predictions.

SCCs are used routinely for quality control in manufacturing processes and to our knowledge this is the first example of its use in computational and comparative genomics. It is highly useful to find outliers from large sequential data and we have exploited that to find hot and cold spots which are also essentially outliers in genomes. In this analysis we have identified two genes *Rv005* and *Rv006* that map to hot spot regions and encode the gyrase gene. Gyrase genes are known to be associated with drug resistance, and mutations are often found in these genes in drug resistant strains[28]. Our results suggest that this gene is in hyper variable region and is likely to undergo variations leading to drug resistant phenotype. We have also found *Rv3919c*, a gene involved in resistance to streptomycin in the hot spot region[29]. We did not find any gene associated with drug resistance in the cold spots. On the other hand *Rv2986c*, a housekeeping gene encoding a histone like-DNA binding protein was found in the cold spot region. This protein is a conserved protein which is required for survival of the organism (http://www.tbdb.org)[30]. Our prediction and validation strategy involved sequencing only the protein coding genes that fall within selected and predicted hyper and hypovariable regions from a number of clinical isolates. This was done to see if the selected regions computed on the basis of sequenced genomes were also outliers in terms of presence of SNVs in Indian clinical isolates. The experimental results supported our predictions.

There are many reasons why hot and cold spots exist in genomes[15]. While mutations occur more or less randomly, SNVs appear as clusters because of positive selection in some regions due to adaptive advantage. It is also possible that these regions have unusual structural features that promote errors of different types[31]. It is also likely that SNVs may occur more commonly around pre-existing mutations as a result of DNA repair system[31]. Whatever the reason,

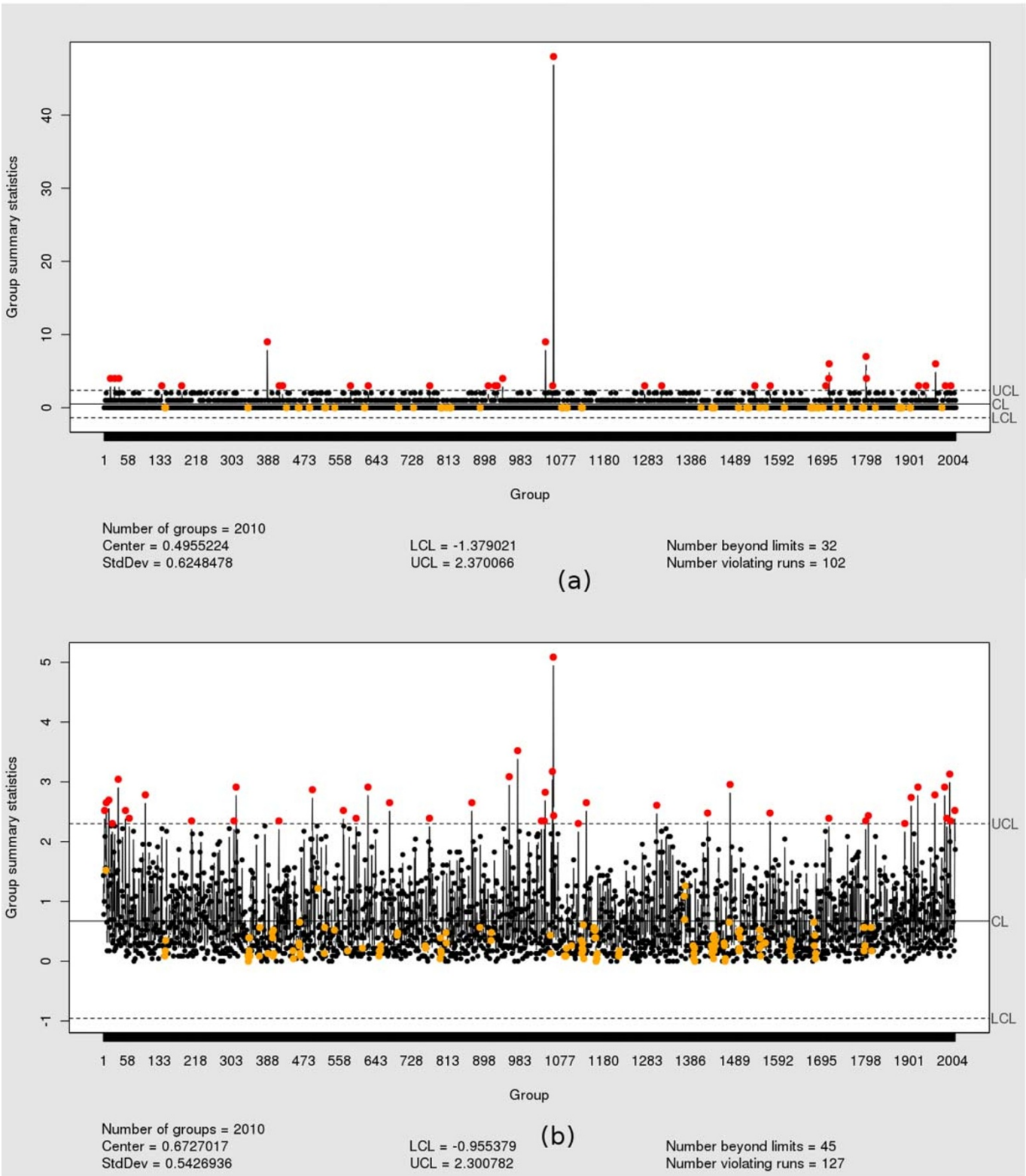

Figure 3 | Shewhart Control Chart: (a) Chart was derived using SNV frequencies from Fig. 1. (b) Average SNV frequencies in all the strains and isolates. Red and black dots indicate out-of-control (''hot spots'') and in-control respectively. Yellow dots indicate violating runs.

occurrence of hyper and hypo variable regions suggests that different regions of *M. tuberculosis* genomes are changing at different rates. Identification of these regions may be helpful in deciphering future therapeutic targets. In conclusion, we have shown that Shewhart Control Chart can be useful to identify hot and cold spots in microbial genomes.

Methods

Datasets. The sequences used were complete genome sequences of *M. tuberculosis* [H37Rv (NC_000962.2), CDC1551 (NC_002755.2)] and whole genome re-sequencing short reads of 20 clinical isolates of *M. tuberculosis* downloaded from SRA (http://www.ncbi.nlm.nih.gov/sra). The data were generated on a high throughput sequencing platform (Illumina) with an average depth of 50x and read length of 52 nucleotides[27].

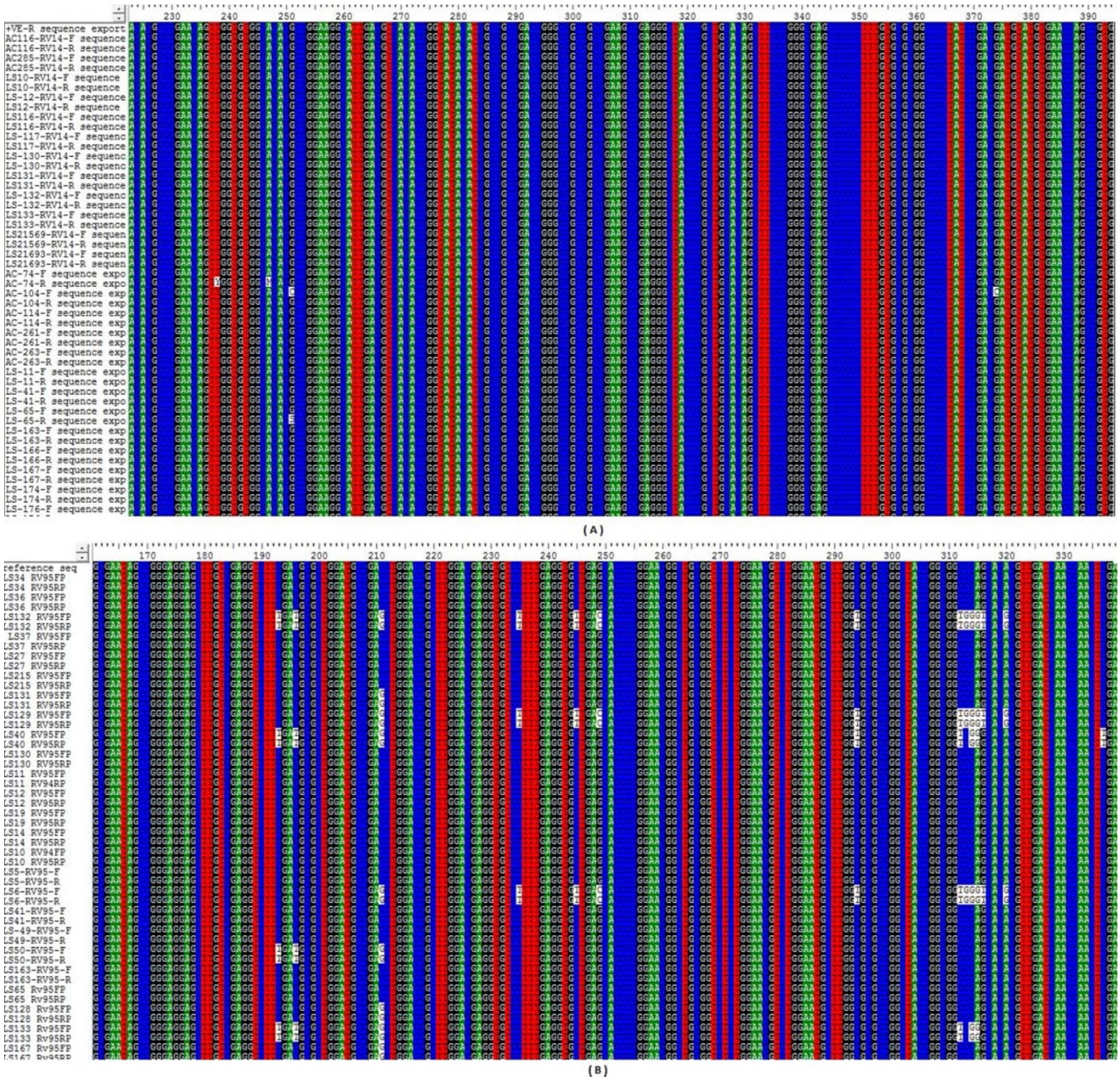

Figure 4 | Multiple sequence alignment of representative amplified re-sequenced cold spot and hot spot regions from clinical isolates. Left side of the alignments is the isolates names with F/R indicating forward/reverse stands. (A) Cold spot; (B) Hot spot.

Identification of single nucleotide variations (SNVs). Identification of single nucleotide variations (SNVs) was done separately for complete genome and next generation sequencing data. We used published genome sequence of *M. tuberculosis* H37Rv as reference genome in both data sets. SNVs were identified from genome data using Anchor Based Whole Genome Analysis Tool (ABWGAT)[21], an online sever for identification of genetic variations from whole genome sequences. Output from the server is a list of SNVs in tabular format including reference genome position, nucleotide change, COG, functions *etc*. For re-sequencing short reads data, we mapped these with respect to reference genome individually using MAQ[32] allowing at most two mismatches. SNV calling parameters used were minimum read depth 3, maximum read depth 256 and consensus quality score 20. We filtered the low score SNVs using SNPFilter, a module of MAQ to get high confidence SNVs.

Prediction of Hot and Cold spots using Shewhart Control Chart. We divided the reference genome into bins of size of 500 to 5000 nucleotides and calculated SNV frequency in each bin. The frequency in each bin was then plotted to find the distribution of SNVs over the genome. We have found the optimal bin size of 2000 in order to get on an average a single SNV in a bin as the total number of SNVs in different datasets was found to be around 1500–2500.

To identify hot and cold spots we used a statistical quality control method called Shewhart Control Chart[17]. Quality control is a technique to monitor a process with the goal of making it more efficient. Shewhart control chart can easily identify outliers in a production process. For our study the presence of outliers indicates hot spots. The chart contains three lines, named UCL -upper control limit, CL -control limit and LCL -lower control limit. (See Fig. 2).

$$CL = \mu \tag{1}$$

$$UCL = \mu + 3\sigma \tag{2}$$

$$LCL = \mu - 3\sigma \tag{3}$$

Where μ = mean,
σ = standard deviation

Analysis of *M. tuberculosis* isolates from patients. DNA extracted from mycobacterial cultures maintained/stored at −20°C on Lowenstein-Jensen (LJ)

media in the TB immunology laboratory, Biotechnology department, AIIMS, New Delhi and in the Microbiology Department, Lala Ram Sarup Institute of Tuberculosis and Respiratory Diseases, Meharuli, New Delhi, was used in the study. The 40 mycobacterial isolates included in the study have been derived from a variety of clinical samples, which include sputum and extra-pulmonary samples such as cerebral spinal fluid, pleural fluid, fine needle lymph node aspirates, endometrial biopsies etc.

DNA extraction, PCR amplification and sequencing. DNA extraction from the isolates was carried out as described before[33]. Briefly, a single colony of a *M. tuberculosis* was picked and suspended in 100 µl of 0.1% Triton –X 100.The suspension was boiled in a dry bath at 90°C for 45 min and centrifuged at 10,000 rpm for 10 minutes. The supernatant was used as template DNA in PCRs.

Amplifications were carried using reagents obtained from Fermentas AB, Vilnius, Lithuania, using a thermocycler (Applied Biosystems , USA). The amplicons were analyzed in a 1.5% agarose gel. Specific DNA bands corresponding to the estimated amplicon size were cut and DNA extracted as per the manufacturer's recommendation, (Real Biotech Corporation,Tawian). Sequencing of the extracted amplicons was done commercially, (GCC Biotech, (India) Pvt. Ltd., Kolkata) for both forward and reverse strands.

SNV calling. Sequences were aligned with *M. tuberculosis* H37Rv genome sequence as reference using CLUSTALW multiple alignment tool[34]. Any nucleotide change was marked as an SNV if the change was observed in both the forward and reverse strands. Otherwise it was considered as a sequencing error.

1. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nature reviews. Genetics* **7**, 85–97 (2006).
2. Dowell, R. D., Ryan, O., Jansen, A. *et al.* Genotype to phenotype: a complex problem. *Science* **328**, 469 (2010).
3. Rogozin, I. B., Pavlov, Y. I. Theoretical analysis of mutation hotspots and their DNA sequence context specificity. *Mutation Research/Reviews in Mutation Research* **544**, 65–85 (2003).
4. Amundsen, S. K. & Smith, G. R. Chi hotspot activity in Escherichia coli without RecBCD exonuclease activity: implications for the mechanism of recombination. *Genetics* **175**, 41–54 (2007).
5. Walker, D. R., Bond, J. P., Tarone, R. E. *et al.* Evolutionary conservation and somatic mutation hotspot maps of p53: correlation with p53 protein structural and functional features. *Oncogene* **18**, 211–218 (1999).
6. Chen, P., Lin, S., Wang, C. *et al.* "Hot spots" mutation analysis of p53 gene in gastrointestinal cancers by amplification of naturally occurring and artificially created restriction sites. *Clin. Chem* **39**, 2186–2191 (1993).
7. Glazko, G. V. Babenko, V. N. Koonin, E. V. Rogozin, I. B. Mutational hotspots in the TP53 gene and, possibly, other tumor suppressors evolve by positive selection. *Biology direct* **1**, 4 (2006).
8. Dixit, A. Yi, L., Gowthaman, R. *et al.* Sequence and structure signatures of cancer mutation hotspots in protein kinases. Selvarajoo K, ed. *PloS one* **4**, e7485 (2009).
9. Lin, X., Xu, X., Huang, Q.-L. *et al.* Biological impacts of "hot-spot" mutations of hepatitis B virus X proteins are genotype B and C differentiated. *World journal of gastroenterology: WJG* **11**, 4703–4708 (2005).
10. Liu, Q., Hoi, S. C. H., Chinh, S. T. T. *et al.* Structural analysis of the hot spots in the binding between H1N1 HA and the 2D1 antibody: do mutations of H1N1 from 1918 to 2009 affect much on this binding? *Bioinformatics (Oxford, England).*, btr437- (2011).
11. Wilson, J. B., Hayday, A., Courtneidge, S. & Fried, M. A frameshift at a mutational hotspot in the polyoma virus early region generates two new proteins that define T-antigen functional domains. *Cell* **44**, 477–487 (1986).
12. Jandova, J., Eshaghian, A., Shi, M. *et al.* Identification of an mtDNA Mutation Hot Spot in UV-Induced Mouse Skin Tumors Producing Altered Cellular Biochemistry. *The Journal of investigative dermatology* (2011).
13. Ogihara, Y., Terachi, T. & Sasakuma, T. Molecular analysis of the hot spot region related to length mutations in wheat chloroplast DNAs. I. Nucleotide divergence of genes and intergenic spacer regions located in the hot spot region. *Genetics* **129**, 873–884 (1991).
14. Chattopadhyay, S., Weissman, S. J., Minin, V. N. *et al.* High frequency of hotspot mutations in core genes of Escherichia coli due to short-term positive selection. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 12412–12417 (2009).
15. Stern, D. L. & Orgogozo, V. Is genetic evolution predictable? *Science (New York, N.Y.)* **323**, 746–751 (2009).
16. Rogozin, I. B., Babenko, V. N., Milanesi, L., Pavlov, Y. I. Computational analysis of mutation spectra. *Briefings in bioinformatics* **4**, 210–227 (2003).
17. Koutras, M. V., Bersimis, S., Maravelakis, P. E. Statistical Process Control using Shewhart Control Charts with Supplementary Runs Rules. *Methodology and Computing in Applied Probability* **9**, 207–224 (2007).
18. Benneyan, J. C., Lloyd, R. C. & Plsek, P. E. Statistical process control as a tool for research and healthcare improvement. *Quality & safety in health care* **12**, 458–464 (2003).
19. Harrison, W. N., Mohammed, M. A., Wall, M. K. & Marshall, T. P. Analysis of inadequate cervical smears using Shewhart control charts. *BMC public health* **4**, 25 (2004).
20. WHO. *Global tuberculosis control 2011*. Geneva, Switzerland: World Health Organization; 2011:246.
21. Das, S., Vishnoi, A. & Bhattacharya, A. ABWGAT: anchor-based whole genome analysis tool. *Bioinformatics (Oxford, England)* **25**, 3319–3320 (2009).
22. Stephens, M. a. EDF Statistics for Goodness of Fit and Some Comparisons. *Journal of the American Statistical Association* **69**, 730 (1974).
23. Hershberg, R., Lipatov, M., Small, P. M. *et al.* High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography. *PLoS biology* **6**, e311 (2008).
24. Weissman, S. J., Beskhlebnaya, V., Chesnokova, V. *et al.* Differential stability and trade-off effects of pathoadaptive mutations in the Escherichia coli FimH adhesin. *Infection and immunity* **75**, 3548–3555 (2007).
25. Fleischmann, R. D., Alland, D., Eisen, J. A. *et al.* Whole-genome comparison of Mycobacterium tuberculosis clinical and laboratory strains. *Journal of bacteriology* **184**, 5479–5490 (2002).
26. Ford, C. B., Lin, P. L., Chase, M. R. *et al.* Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. *Nature genetics* **43**, 482–486 (2011).
27. Comas, I., Chakravartti, J., Small, P. M. *et al.* Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. *Nature genetics* **42**, 498–503 (2010).
28. Takiff, H. E., Salazar, L., Guerrero, C. *et al.* Cloning and nucleotide sequence of Mycobacterium tuberculosis gyrA and gyrB genes and detection of quinolone resistance mutations. *Antimicrobial agents and chemotherapy* **38**, 773–780 (1994).
29. Sandgren, A., Strong, M., Muthukrishnan, P. *et al.* Tuberculosis drug resistance mutation database. *PLoS medicine* **6**, e2 (2009).
30. Sassetti, C. M., Boyd, D. H., Rubin, E. J. Genes required for mycobacterial growth defined by high density mutagenesis. *Molecular microbiology* **48**, 77–84 (2003).
31. Amos, W. Even small SNP clusters are non-randomly distributed: is this evidence of mutational non-independence? *Proceedings. Biological sciences / The Royal Society* **277**, 1443–1449 (2010).
32. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research* **18**, 1851–1858 (2008).
33. Kumar, P., Sen, M. K., Chauhan, D. S. *et al.* Assessment of the N-PCR assay in diagnosis of pleural tuberculosis: detection of M. tuberculosis in pleural fluid and sputum collected in tandem. Mokrousov I, ed. *PloS one* **5**, e10220 (2010).
34. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research* **22**, 4673–4680 (1994).

Acknowledgments

The Department of Biotechnology, Government of India for financial support, the Council of Scientific & Industrial Research, India for research fellowship to S. Das, the technical help of Mr. K.P. Singh, Shailendra Kumar, Inderesh Kumar and Surender Singh is acknowledged. The authors thank Prof. Sudha Bhattacharya for critically reading the manuscript.

Author contribution

AB and SD conceptualized the study. AB, SD and HKP wrote the manuscript. SD performed the computational works. RR helped in statistical analysis. HKP, PD performed experiments and analysis of clinical isolates. VPM and DB have characterized and isolated the clinical isolates. All authors reviewed the manuscript.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Das, S. *et al.* Identification of Hot and Cold spots in genome of *Mycobacterium tuberculosis* using Shewhart Control Charts. *Sci. Rep.* **2**, 297; DOI:10.1038/srep00297 (2012).