Genomics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Classification of pathogenic microbes using a minimal set of single nucleotide polymorphisms derived from whole genome sequences

Tanmoy Roychowdhury[a,1], Vinod Kumar Singh[a], Alok Bhattacharya[a,b,*]

[a] School of Computational & Integrative Sciences, Jawaharlal Nehru University, New Delhi, India
[b] School of Life Sciences, Jawaharlal Nehru University, New Delhi, India

ARTICLE INFO

Keywords:
Minimal set
Random forest
Jensen-Shannon divergence
Shewhart control chart
Phylogroup
Pathogroup

ABSTRACT

In a context specific manner, Intra-species genomic variation plays an important role in phenotypic diversity observed among pathogenic microbes. Efficient classification of these pathogens is important for diagnosis and treatment of several infectious diseases. NGS technologies have provided access to wealth of data that can be utilized to discover important markers for pathogen classification. In this paper, we described three different approaches (Jensen-Shannon divergence, random forest and Shewhart control chart) for identification of a minimal set of SNPs that can be used for classification of organisms. These methods are generic and can be implemented for analysis of any organism. We have shown usefulness of these approaches for analysis of *Mycobacterium tuberculosis* and *Escherichia coli* isolates. We were able to identify a minimal set of 18 SNPs that can be used as molecular markers for phylogroup based classification and 8 SNPs for pathogroup based classification of *E. coli*.

1. Introduction

Several pathogenic microorganisms display a range of disease phenotypes varying from asymptomatic to highly invasive. Different isolates/strains of the same species are thought to be associated with different pathophysiologies including virulence, tissue tropism and are referred to as pathotypes or pathogroups. On the other hand, lineages or phylogroups can be derived from genomic analysis of different isolates of the same species. Relationships among these phylogroups and pathogroups are often very complex. However, efficient classification based on genomic information is important for understanding principles behind molecular basis of pathogenesis and clinical management of these diseases.

A number of experimental methods are available for classification of pathogenic bacteria based on their genotype. Most of these methods use polymerase chain reaction (PCR) for detection of large sequence polymorphisms (LSP), such as, number and sites of insertion sequences [1] and variable number of tandem repeats [2]. Some of these methods tend to be useful for phylogenetic classification of a specific organism. For example, spoligotyping based on presence or absence of spacers in between repetitive sequences (DR), has been used for classification of *Mycobacterium tuberculosis* complex organisms (MTBC) [3]. On the other hand, multilocus sequence typing (MLST) [4] of a few housekeeping genes, has been found to be useful for categorizing a large number of different bacterial species. Developments in whole genome sequencing (WGS) have revolutionized the way genome sequences are obtained and it has become possible to classify organisms using WGS. WGS is slowly replacing MLST as it is better in handling deeper branches and can detect subtle genomic differences between organisms with phenotypic variability. However, the bottleneck is cost, time and expertise needed for data analysis. A set of genomic markers (for example, a minimal set of few SNPs) that truly reflect strain/isolate divergence can become a useful alternative to WGS/MLST. Few methods [5,6] were developed for identification of sequence type from MLST datasets. This has limited application due to choice of small number of specific genes while WGS allows identification of all possible variations. Once a minimal set of SNPs are identified from WGS data of small number of isolates, classification of an uncharacterized isolate can be performed using experimental identification of small number SNPs.

WGS based analyses are being increasingly applied for classification of MTBC as well as other pathogenic bacteria. Phylogroup based classification of MTBC has been found to be highly useful in understanding genotype-phenotype relationship [7]. This classification is thought to reflect observed variability in emergence of drug-resistance, immune response and disease severity [8]. Roychowdhury et al. showed that it is possible to classify MTBC isolates into their respective phylogroups

* Corresponding author.
E-mail address: alok.bhattacharya@gmail.com (A. Bhattacharya).
[1] Present address: Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA.

https://doi.org/10.1016/j.ygeno.2018.02.004
Received 23 September 2017; Received in revised form 4 February 2018; Accepted 8 February 2018
0888-7543/ © 2018 Published by Elsevier Inc.

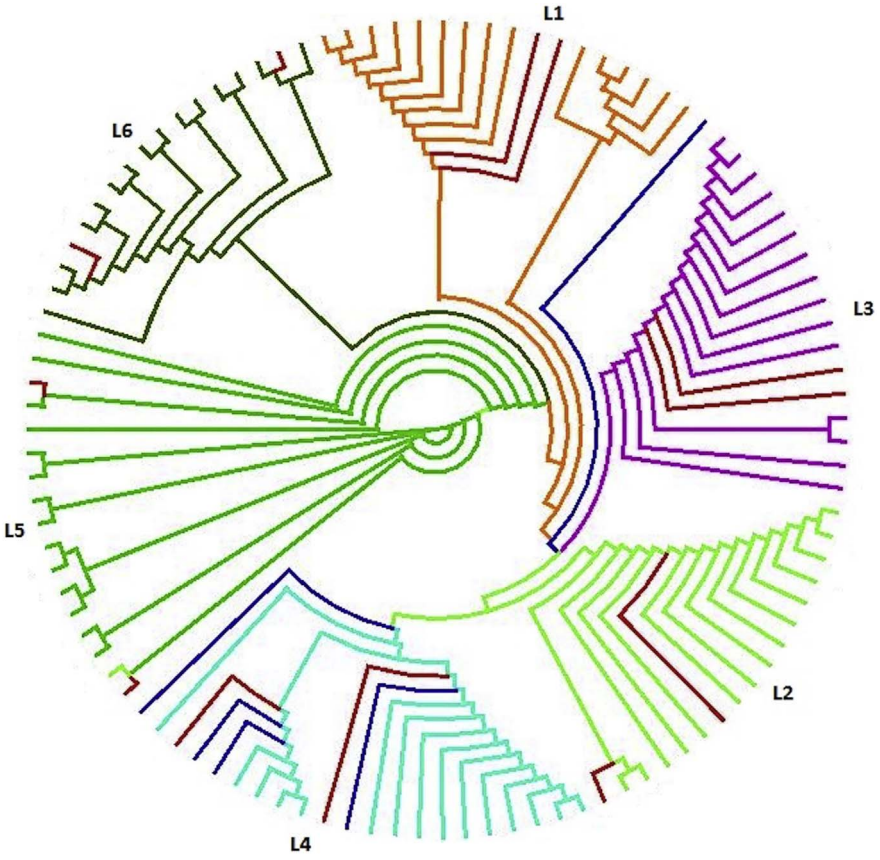

Fig. 1. Phylogenetic tree of 107 *M. tuberculosis* isolates based on 65 SNPs (selected using Random Forest model). Ninety of these isolates are part of 1362 isolates which were originally used to generate the SNP set (15 from each of 6 phylogroups). Different colors represent different phylogroups such as, L1: Orange; L2: Light Green; L3: Purple; L4: Cyan; L5: Green; L6: Dark Green. Seventeen other isolates were used to demonstrate the usability of minimal set. Twelve isolates from SRA (red) with known phylogroup information. Five isolates of unknown phylogroup were also included (blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

using random sampling of genomic regions from next generation sequencing (NGS) data [9]. In another study, WGS was used to classify 1601 MTBC isolates based on region of differences (RD) [10]. A total of 91,648 SNPs were identified in this study and these were used to infer a phylogenetic tree that showed proper separation of different RDs. Using fixation index (F_{ST}) and nature of SNPs (synonymous/non-synonymous), a total of 62 SNPs were identified that were able to correctly classify these organisms. Similarly, a set of SNPs was able to discriminate *P. falciparum* isolates based on geographical origin [11]. WGS data also allows detection of repetitive elements, such as insertion sequences. These elements can also be used for organism classification, for example, IS*6110* element insertion sites were detected in a set of 1377 MTBC isolates and the data was used for identification of phylogroups [12].

Escherichia coli can be classified in three broad pathogroups such as, commensal, diarrhoeagenic (ETEC, EAEC, EHEC, EPEC) and uropathogenic (UPEC, ABU) based on pathogenicity and site of infection [13]. Initially, *E. coli* was classified in six phylogroups (A, B1, B2, C, D, E; C is no longer used) using multilocus enzyme electrophoresis [14]. Clermont et al. developed a PCR multiplexing method for detecting three genomic regions for rapid identification of four phylogroups [15]. It was shown that correct phylogenetic grouping can be determined for 80–85% of type-able strains [16,17]. Later, single gene and MLST based methods were developed and these studies showed that B1 and B2 are not sister taxa as described earlier [14]. However, there are problems associated with MLST based phylogenetic analysis due to factors such as, influence of arbitrary choice of genes, overlap with recombination hotspots and inclusion of some of the antibiotic resistance related genes. Using WGS, *E. coli* isolates have been classified in six major phylogroups, namely A, B1, B2, D, E and F (D and F are alternatively known as D1 and D2) [18]. Each of these phylogroups behaves differently in terms of pathogenicity or ecological niche. A direct association between the site of infection and specific phylogroup has not been seen yet, though over-representation of some phylogroups has been observed at specific sites [19,20]. We believe that currently there is no suitable method available for phylogroup or pathogroup classification of an unknown isolate of *E. coli* based on a small set of SNPs.

In this report, we describe three different approaches for identification of a minimal set of SNPs from WGS data for classification of microbial pathogens. Our major goal is to find a minimum set of markers (SNP in this case) that can be used to correctly classify a microorganism with respect to phylogroup/pathogroup. The methods are based on Jensen-Shannon divergence (JSD) [21], random forest (RF) [22] and Shewhert Control Chart (SCC) [23]. Though we have applied these methods for classification of different isolates of *M. tuberculosis* and *E. coli,* these are generic in nature and can be used for any organism. These two organisms were chosen partly based on availability of well characterized information about their phylogroup/pathogroup.

2. Materials and methods

Publicly available WGS datasets of *M. tuberculosis* and *E. coli* isolates were used for this analysis. Supplementary Table 1 summarizes information regarding organisms that have been used for this study. Whole genome data of *M. tuberculosis* isolates were available as NGS reads. Short sequencing reads were aligned with *M. tuberculosis* H37Rv genome using Bowtie [24] and SNPs were identified using Samtools [25]. We obtained assembled genomes for *E. coli* isolates. MUMmer [26] was used to identify SNPs from all *E. coli* isolates used in this study (reference *E. coli* K12 MG1665).

Jensen-Shannon Divergence is an information theoretic measurement to quantify difference between two or more probability distributions [27]. One advantage of using information theoretic measurements for symbolic sequences (DNA) is that these don't require symbolic sequences to be mapped to numerical sequences. The dataset contained L number of SNPs of m isolates of specific bacteria and these isolates were

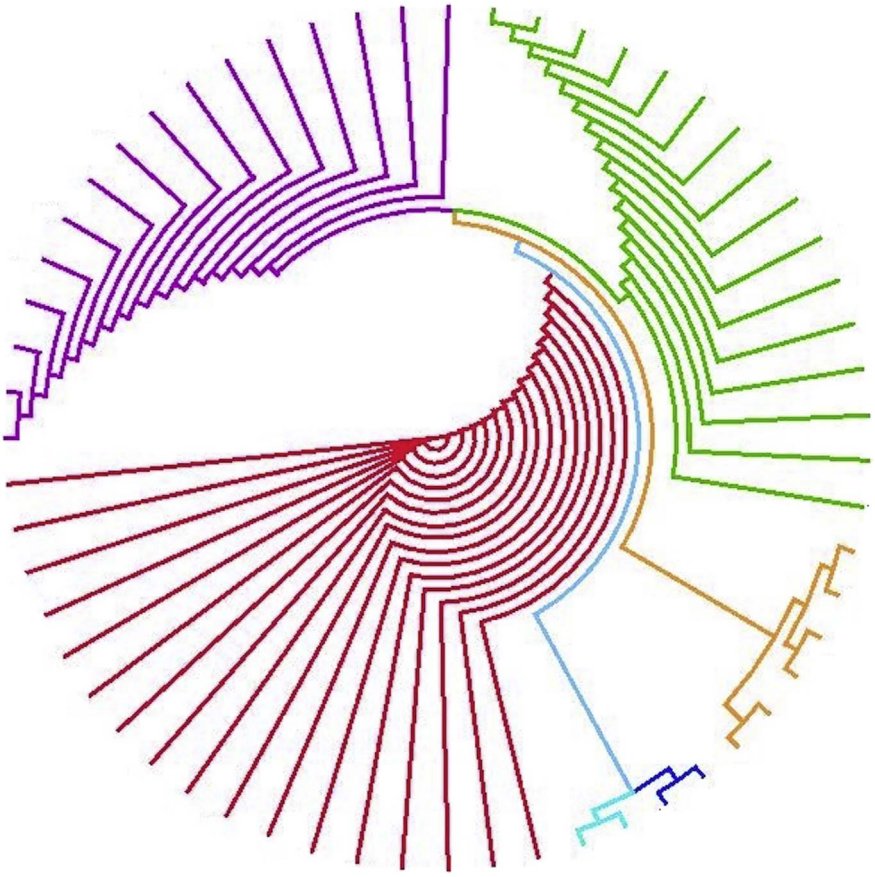

Fig. 2. Phylogenetic tree of 59 *E. coli* isolates based on 18 SNPs (selected using Random Forest model). Different colors represent different phylogroup such as, A: Green; B1: Purple; B2: Red; D: Cyan; E: Orange; F: Blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Genotype of 18 SNPs for phylogroup based classification of *E. coli*. Genomic locations are shown with respect to *E. coli* K-12 MG1665.

Location	A	B1	B2	D	E	F
17,752	A	A	G	A	A	A
680,931	C	A	C	C	G	C
695,049	A	G	G	G	G	G
920,637	G	A	G	G	G	G
978,008	A	G	G	G	G	G
992,919	C	T	T	T	C	C
995,131	G	G	G	G	G	A
1,103,680	G	A	G	G	G	G
1,119,235	T	T	G	T	T	T
1,225,750	T	C	C	C	C	C
1,378,692	C	T	C	C	C	C
1,835,547	A	C	C	C	C	C
1,957,803	G	G	G	A	G	A
2,494,090	G	G	C	C	C	C
3,247,973	C	T	C	C	C	C
3,252,205	T	T	C	T	T	T
3,432,363	A	T	T	T	T	T
3,680,276	T	C	C	C	C	C

further stratified into n phylogroups. For any specific SNP, P_1, P_2, ⋯, P_n are the probability distributions of four nucleotide bases of n distinct phylogroups at that location. The JSD for given probability distributions having prior weights π_1, π_2, ⋯, π_n.

$$JSD_{\pi_1,\pi_2,\cdots,\pi_n}(P_1,P_2,\cdots,P_n) = H\left(\sum_{i=1}^{n}\pi_i P_i\right) - \sum_{i=1}^{n}\pi_i H(P_i) \quad (1)$$

where, $H(P)$ is Shannon entropy of nucleotides is expressed by Eq. 2 and $p_{i,k}$ is the probability of base k in i^{th} phylogroup at any given SNP.

$$H(P_i) = -\sum_{k\in\{a,\ c,g,t\}} p_{i,k}\log_2 p_{i,k} \quad (2)$$

JSD is bound between 0 and $log_2(n)$ i.e., $0 \le JSD_{\pi_1}$, π_2, ⋯, $\pi_n(P_1,P_2,\cdots,P_n) \le log_2(n)$. With $JSD = 0$ if and only if $P_1 = P_1 = \cdots = P_n$. All phylogroups were given equal weightage for the study, hence, prior weights of distributions were considered equal i.e., $\pi_1 = \pi_1 = \cdots = \pi_n = \frac{1}{n}$. Informative SNPs were chosen based on higher JSD value. This was implemented in Matlab and the script used to run this analysis is available as supplementary information.

Random forest is an ensemble learning method that boosts the performance of individual decision trees using a majority rule. For a discrete response variable, RF can be used for classification, where as regression is used for a continuous response variable. We used RF for classification as the response variables in this case are discrete, for e.g., 6 phylogroups for *M. tuberculosis* and *E. coli* each or 3 pathogroups for *E. coli*. Given a set of "m" training samples (*M. tuberculosis* or *E. coli* isolates) each with a vector of predictor variables (all SNPs) and response variables (phylogroup/pathogroup), RF generates "Ntree" number of decision trees. For all our analysis, we used "Ntree" as 500. For each tree, "k" number of samples are chosen using a bootstrap resampling to form a new set of "m" samples and the rest of the samples are known as out-of-bag (OOB) samples (m-k). While growing the decision tree, a small number of randomly selected "mtry" variables are used from the vector of predictor variables ie, subsets of SNPs are randomly selected, enabling random feature variable selection. We used "mtry" as square root of number of total SNPs. Unlike other machine learning methods, RF can estimate error rate based on classification of OOB samples of each tree (similarly as test sets in cross-validation). This variable selection feature of RF was used for analysis. While learning the model, a Gini index score is generated for each feature variable or SNP. Based on

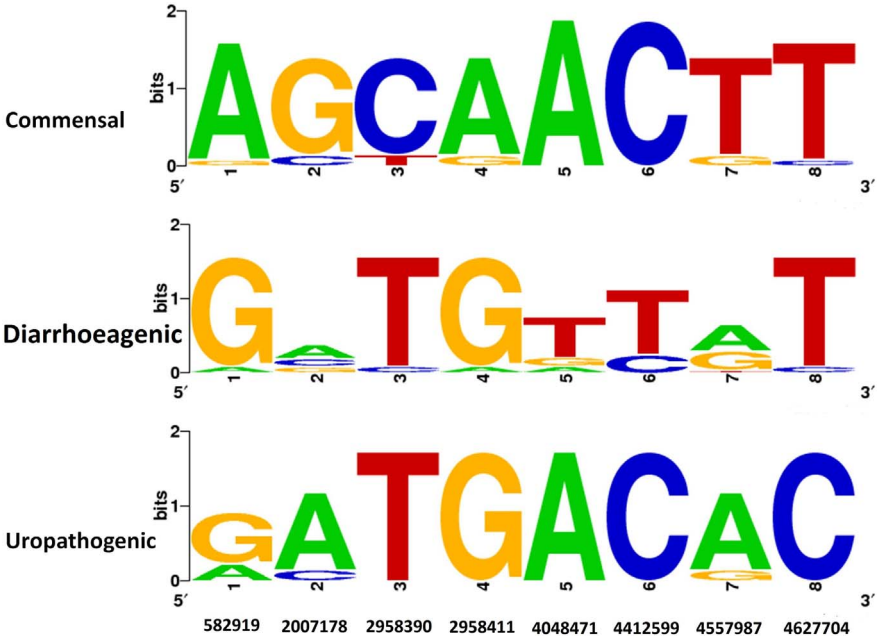

Fig. 3. Sequence logo representation of 8 SNPs associated with *E. coli* pathogenicity.

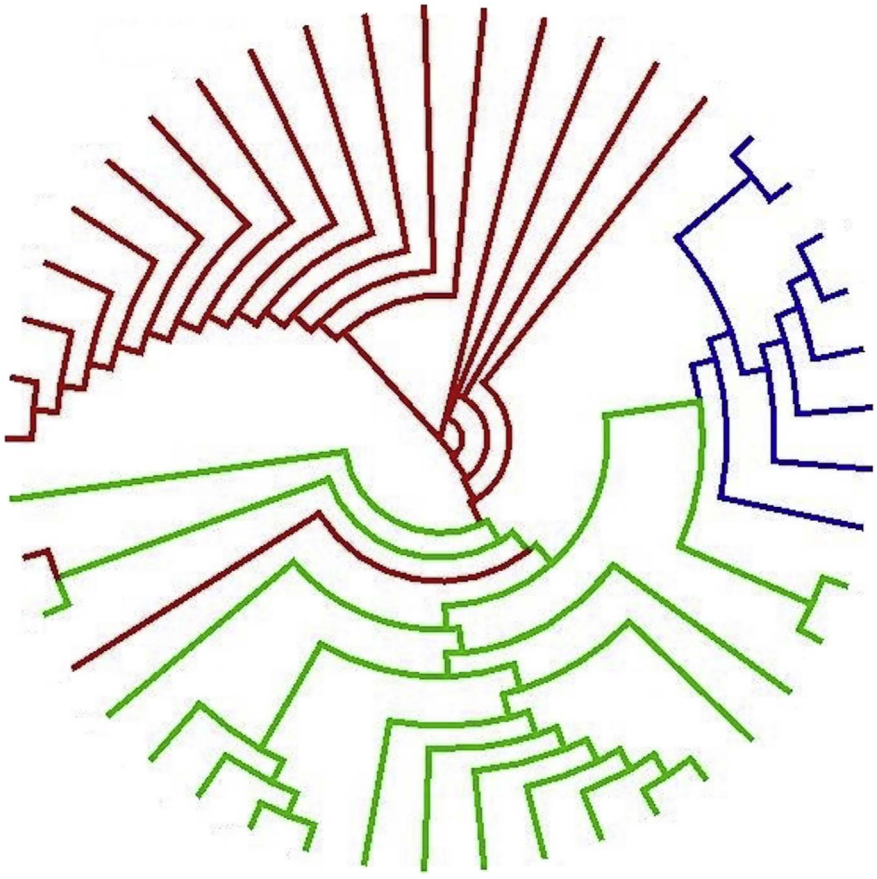

Fig. 4. Dendogram of 45 *E. coli* isolates based on 8 SNPs related to pathogroups. Different colors represent different pathogroups. Red: commensal, Green: Diarrhoeagenic, Blue: Uropathogenic. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

class-specific Gini scores, we identified most informative SNPs for each class. We used R package "randomForest" [22] for these calculations.

Genomic SNP hotspots were identified using Shewhart control chart described in Das et al. [28]. Shewhart control charts are routinely used for statistical quality control to identify outliers in production process. Basic properties of this chart are central line (CL), upper control limit (UCL) and lower control limit (LCL). In this analysis, genomic regions those are identified as outliers (above UCL) in terms of SNP frequency can be denoted as hotspots. Overall, whole genome was divided into continuous non-overlapping bins of 2000 bp. SNP frequency in each of

these genomic bins were calculated using SNPs those were found in at least one of the isolates being analyzed. SCC was plotted and properties were calculated using the qcc() function in R package “qcc”.

3. Results and discussion

An exhaustive analysis was carried out to identify all SNPs from the datasets. This yielded 103,066 unique SNPs from 1362 isolates of *M. tuberculosis* using H37Rv as reference genome and 343,929 from 59 isolates of *E. coli* using *E. coli* K12-MG1665 as reference genome. The number of SNPs per isolate was much higher in *E. coli* in comparison to *M. tuberculosis,* suggesting a higher level of genomic diversification in *E. coli.* That is also expected as *E. coli* displays extensive phenotypic diversity. Phylogenetic trees of *M. tuberculosis* [29] and *E. coli* [18] isolates using complete SNP datasets have been derived and the results clearly displayed distinct clusters consisting of isolates of different phylogroups. However, it is not feasible to either develop field based diagnostic or strain identification methods using a large dataset, like, list of all SNPs.

3.1. Identification of minimal set of SNPs for classification of M. tuberculosis isolates

We have used three different approaches for finding a minimal set that can be used for classification of MTBC from 103,066 SNPs identified from 1362 *M. tuberculosis* isolates. We believe that these methods are generic and not dependent on nature of genomes.

JSD was calculated for each SNP considering 6 distinct nucleotide distributions from 6 phylogroups. It is observed (Supplementary Fig. 1) that a very small proportion of SNPs are associated with phylogroup based classification of *M. tuberculosis* isolates. It is also evident that phylogroup-defining SNPs were distributed throughout the genome and not clustered in a few regions. Finally, 134 SNPs that displayed JSD $\geq$1.425, were empirically selected and are listed in Supplementary Table 2. The threshold was chosen to obtain optimum result after manual comparison of different trees. A phylogenetic tree based on these 134 SNPs classified the whole population in six phylogroups and is also consistent with classification as “ancient” and “modern” (Supplementary Fig. 2) [30].

For categorical response variables, Random Forest (RF) performs supervised classification based on learning from a training set. During the process of learning a variable importance score (Gini index) is calculated, for each predictor variable. We used this score to identify important SNPs associated with the different phylogroups of *M. tuberculosis* isolates. Variable importance was calculated for 103,066 SNPs using a RF model. Each isolate (from 1362 samples) is associated with one response variable (L1-L6) and all 1362 response variables were used in the model. The estimated out-of-bag (OOB) error rate for RF model was 0.005. Only 5213 SNPs had a non-zero importance score. Some of the variables with highest mean decrease in Gini index are depicted in the Supplementary Fig. 3. For each variable, RF model generates a Gini index score associated with each class. We have selected top 20 variables associated with each class that resulted in 65 unique SNPs (Supplementary Table 3). A phylogenetic tree (Supplementary Fig. 4) of this isolates was constructed from multiple sequence alignment of these SNPs. The results clearly showed that all phylogroups were clearly separated and formed distinct phylogroups. Random sampling based method NexABP showed that approximately 20% of the data is required for proper phylogenetic classification [9]. Interestingly, only 65 SNPs representing approximately 0.06% of the total SNPs were found to be sufficient for phylogroup prediction in this analysis. Further, SNPs associated with different spoligotypes were identified using RF and 34 different spoligotype classes were considered for the analysis. SNPs with Gini index greater than 1.5 (total 206) are shown in Supplementary Table 4. This analysis failed to identify SNPs associated with spoligotypes EAI2, EAI5, H2, H4 and LAM5.

In order to validate the usefulness of the derived minimal set of SNPs for *M. tuberculosis* classification, we analyzed WGS data from 17 isolates that were not part of the original set of 1362 isolates used for generating the SNP set. Phylogroup information was available for 12 of these isolates (two from each phylogroup). All these isolates (red) cluster within their respective phylogroups (Fig. 1). We used five additional isolates of an unknown type [31]. Three of these could be placed in specific clusters in an unambiguous manner.

An alternative approach for strain identification is sequencing specific regions of the genome. Generally house-keeping genes or genes with higher mutation rate are used in strain classification [32,33]. For this analysis, we used a set of genomic regions with high rate of mutation as identified using Shewhart control chart [28]. SCC reported 26 hotspots representing 3406 SNPs (Supplementary Table 5). A phylogenetic tree of MTBC based on these 26 genomic bins can classify isolates into distinct phylogroups (L1, L2, L3, L4, L5 and L6). The result is presented in Supplementary Fig. 5. This approach provides an alternative way to classify isolates utilizing a small set of genes/regions and is more useful in analyzing organisms with less variability. Identification of genomic hotspots is also important for understanding emergence of drug resistance as genes encoded in these regions show higher amount of variability in otherwise less diverse pathogens like *M. tuberculosis* [34–36].

These methods were chosen to have results which are mutually exclusive but serve a similar purpose. Both JSD and RF identify informative SNPs that are distributed throughout the genome. Although, a combination of these SNPs is able to classify each isolate, ability of a single SNP is different when identified by RF compared to JSD (Supplementary Fig. 6). RF based method looks for nucleotides that can differentiate one class from the rest. For instance, a specific nucleotide might differentiate L1 from the rest (i.e., SNP in only L1 isolates but conserved in L2, L3, L4, L5 and L6). On the other hand, nucleotides identified by JSD, classify each isolate into at least two subgroups, each subgroup consisting of one or more classes. For instance, a specific nucleotide might be able create two subgroups, one with L5 and L6 while the other one consists of L1, L2, L3 and L4. In the results presented here with respect to *M. tuberculosis* isolates, the number of SNPs common between these two approaches is very low (only 6). Hotspot based approach doesn't look for informative SNPs at all. This method looks for SNPs that are spatially clustered and represent genomic regions with higher mutation rate. Once hotspots are identified from few isolates, further analysis of other isolates involves only a fraction of total WGS data.

3.2. Identification of minimal set of SNPs for classification of E. coli isolates

In order to show that the methods described above are not specific for *M. tuberculosis*, we have used random forest to identify a minimal set of SNPs required for classification of *E. coli* isolates in separate phylogroups. Richards et al. reported a maximum likelihood phylogenetic tree of 59 isolates (these have also been used for this study) based on all SNPs [18]. In order to identify a minimal set of SNPs that can classify these isolates in distinct phylogroups, variable importance score was calculated for 343,929 SNPs using a RF model (OOB error rate 0). Top two SNPs from each phylogroup generated 26 unique SNPs. We further selected 18 SNPs that were completely conserved in each phylogroup. These SNPs were used for constructing phylogenetic tree of 59 isolates (Fig. 2). Comparison of this tree with that derived from all SNPs revealed similar phylogroup clustering [18]. The relationships among different phylogroups were also consistent with that observed with WGS data. Phylogroups A and B1 were identified as the most recently diverged sister taxa followed by D, F, E where as B2 was the most basal group. F taxon was also found to be closely related to D as observed earlier (D and F taxa are thought to be monophyletic [14]). Genotypes of these 18 SNPs are presented in Table 1. This set of 18 SNPs can be used as an alternative barcode for PCR based determination of *E. coli*

strain type. We generated the phylogenetic tree of these isolates based on MLST datasets (Supplementary Figs. 7 & 8). Although, larger number of SNPs (526 and 1008 respectively) were used to generate these trees, several incongruencies were observed in terms of relationship among phylogroups, for e.g., these trees predict phylogroup A as the ancestral/basal group. Moreover, the phylogenetic tree based on 526 SNPs was not consistent in classification of isolates in corresponding phylogroups in several occasions.

Due to inherent limitation of MLST data generation, these datasets can't be used to classify isolates based on pathogroups. We have extended this study further by identifying a minimal set of SNPs that can be used to classify *E. coli* isolates based on pathotypes. Broadly, *E. coli* strains can be classified in three categories: commensal, diarrhoeagenic and uropathogenic as mentioned before. Members of any specific pathotype are from different phylogroups but certain phylogroups are over-represented. For example, most of the commensal isolates are from phylogroups A or B1 whereas, most of the isolates causing extra-intestinal diseases are from phylogroups B2 or D and almost never from phylogroups A and B1. Diarrhoeagenic isolates are more diverse in terms of their presence in different phylogroups. We have used RF based method to identify SNPs associated with each pathogenic type and used 19, 18 and 8 isolates of commensal, diarrhoeagenic and uropathogenic *E. coli* respectively for this analysis. Eight SNPs were identified and conservation of nucleotides in each of them is depicted in Fig. 3. Seven of these SNPs are intragenic (Supplementary Table 6). These SNPs were found to be more conserved in commensal isolates than in pathogenic ones, though some degree of conservation is also observed in the later (Fig. 4). As a combination of these SNPs can be more efficient in isolate classification, a dendrogram based on these 8 SNPs was constructed. Commensal (except two commensal isolates from phylogroup B2 were clustered with diarrhoeagenic) and uropathogenic strains were found to be monophyletic whereas diarrhoeagenic isolates were not (Fig. 4). These 8 SNPs will be highly useful in strain typing and can be easily adapted for developing experimental methods. We believe that this is the first report for identification of useful SNP based markers without prior bias and can be immensely useful for future diagnostic purposes.

In this paper, we show usefulness of three different methods for identification of a minimal set of SNPs that can be used for bacterial classification into phylogroups or pathogroups. Though these three methods are different in design, the output appears to be similar, that is, finally yielding a small manageable set of SNP based markers that are able to identify organisms of different classes. A small set is amenable for rapid experimental manipulation reducing the cost and speed of diagnosis. Therefore the methods described here can be highly useful in mapping emergence of highly virulent strains in different geographical regions.

Acknowledgements

The authors thank Department of Biotechnology, Government of India for financial support, Department of Science and Technology, Government of India for J.C. Bose fellowship.

Author contribution

TR and AB conceptualized the study. TR performed the computational analysis. VS worked in the analysis regarding JSD. TR and AB wrote the manuscript. All authors reviewed the manuscript.

Data archiving

This article does not report any new empirical data or software.

Conflict of interest statement

None declared.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2018.02.004.

References

[1] R.C. Huard, L.C. Lazzarini, W.R. Butler, D. van Soolingen, J.L. Ho, PCR-based method to differentiate the subspecies of the Mycobacterium tuberculosis complex on the basis of genomic deletions, J. Clin. Microbiol. 41 (2003) 1637–1650.

[2] P. Supply, C. Allix, S. Lesjean, M. Cardoso-Oelemann, S. Rusch-Gerdes, E. Willery, E. Savine, P. de Haas, H. van Deutekom, S. Roring, P. Bifani, N. Kurepina, B. Kreiswirth, C. Sola, N. Rastogi, V. Vatin, M.C. Gutierrez, M. Fauville, S. Niemann, R. Skuce, K. Kremer, C. Locht, D. van Soolingen, Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of mycobacterium tuberculosis, J. Clin. Microbiol. 44 (2006) 4498–4510.

[3] K. Brudey, J.R. Driscoll, L. Rigouts, W.M. Prodinger, A. Gori, S.A. Al-Hajoj, C. Allix, L. Aristimuno, J. Arora, V. Baumanis, L. Binder, P. Cafrune, A. Cataldi, S. Cheong, R. Diel, C. Ellermeier, J.T. Evans, M. Fauville-Dufaux, S. Ferdinand, D. Garcia de Viedma, C. Garzelli, L. Gazzola, H.M. Gomes, M.C. Guttierez, P.M. Hawkey, P.D. van Helden, G.V. Kadival, B.N. Kreiswirth, K. Kremer, M. Kubin, S.P. Kulkarni, B. Liens, T. Lillebaek, M.L. Ho, C. Martin, C. Martin, I. Mokrousov, O. Narvskaia, Y.F. Ngeow, L. Naumann, S. Niemann, I. Parwati, Z. Rahim, V. Rasolofo-Razanamparany, T. Rasolonavalona, M.L. Rossetti, S. Rusch-Gerdes, A. Sajduda, S. Samper, I.G. Shemyakin, U.B. Singh, A. Somoskovi, R.A. Skuce, D. van Soolingen, E.M. Streicher, P.N. Suffys, E. Tortoli, T. Tracevska, V. Vincent, T.C. Victor, R.M. Warren, S.F. Yap, K. Zaman, F. Portaels, N. Rastogi, C. Sola, Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology, BMC Microbiol. 6 (2006) 23.

[4] M.C. Maiden, Multilocus sequence typing of bacteria, Annu. Rev. Microbiol. 60 (2006) 561–588.

[5] I. Filliol, A.S. Motiwala, M. Cavatore, W. Qi, M.H. Hazbon, M. Bobadilla del Valle, J. Fyfe, L. Garcia-Garcia, N. Rastogi, C. Sola, T. Zozio, M.I. Guerrero, C.I. Leon, J. Crabtree, S. Angiuoli, K.D. Eisenach, R. Durmaz, M.L. Joloba, A. Rendon, J. Sifuentes-Osornio, A. Ponce de Leon, M.D. Cave, R. Fleischmann, T.S. Whittam, D. Alland, Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set, J. Bacteriol. 188 (2006) 759–772.

[6] G.A. Robertson, V. Thiruvenkataswamy, H. Shilling, E.P. Price, F. Huygens, F.A. Henskens, P.M. Giffard, Identification and interrogation of highly informative single nucleotide polymorphism sets defined by bacterial multilocus sequence typing databases, J. Med. Microbiol. 53 (2004) 35–45.

[7] S. Gagneux, K. DeRiemer, T. Van, M. Kato-Maeda, B.C. de Jong, S. Narayanan, M. Nicol, S. Niemann, K. Kremer, M.C. Gutierrez, M. Hilty, P.C. Hopewell, P.M. Small, Variable host-pathogen compatibility in *Mycobacterium tuberculosis*, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 2869–2873.

[8] C.B. Ford, R.R. Shah, M.K. Maeda, S. Gagneux, M.B. Murray, T. Cohen, J.C. Johnston, J. Gardy, M. Lipsitch, S.M. Fortune, Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis, Nat. Genet. 45 (2013) 784–790.

[9] T. Roychowdhury, A. Vishnoi, A. Bhattacharya, Next-generation anchor based phylogeny (NexABP): constructing phylogeny from next-generation sequencing data, Sci. Rep. 3 (2013) 2634.

[10] F. Coll, R. McNerney, J.A. Guerra-Assuncao, J.R. Glynn, J. Perdigao, M. Viveiros, I. Portugal, A. Pain, N. Martin, T.G. Clark, A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains, Nat. Commun. 5 (2014) 4812.

[11] M.D. Preston, S. Campino, S.A. Assefa, D.F. Echeverry, H. Ocholla, A. Amambua-Ngwa, L.B. Stewart, D.J. Conway, S. Borrmann, P. Michon, I. Zongo, J.B. Ouedraogo, A.A. Djimde, O.K. Doumbo, F. Nosten, A. Pain, T. Bousema, C.J. Drakeley, R.M. Fairhurst, C.J. Sutherland, C. Roper, T.G. Clark, A barcode of organellar genome polymorphisms identifies the geographic origin of *Plasmodium falciparum* strains, Nat. Commun. 5 (2014) 4052.

[12] T. Roychowdhury, S. Mandal, A. Bhattacharya, Analysis of IS6110 insertion sites provide a glimpse into genome evolution of *Mycobacterium tuberculosis*, Sci. Rep. 5 (2015) 12567.

[13] M.A. Croxen, B.B. Finlay, Molecular mechanisms of *Escherichia coli* pathogenicity, Nat. Rev. Microbiol. 8 (2010) 26–38.

[14] R.R. Chaudhuri, I.R. Henderson, The evolution of the *Escherichia coli* phylogeny, Infect. Genet. Evol. 12 (2012) 214–226.

[15] O. Clermont, S. Bonacorsi, E. Bingen, Rapid and simple determination of the *Escherichia coli* phylogenetic group, Appl. Environ. Microbiol. 66 (2000) 4555–4558.

[16] D.M. Gordon, O. Clermont, H. Tolley, E. Denamur, Assigning *Escherichia coli* strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method, Environ. Microbiol. 10 (2008) 2484–2496.

[17] S.M. Turner, R.R. Chaudhuri, Z.D. Jiang, H. DuPont, C. Gyles, C.W. Penn, M.J. Pallen, I.R. Henderson, Phylogenetic comparisons reveal multiple acquisitions of the toxin genes by enterotoxigenic *Escherichia coli* strains of different evolutionary lineages, J. Clin. Microbiol. 44 (2006) 4528–4536.

[18] V.P. Richards, T. Lefebure, P.D. Pavinski Bitar, B. Dogan, K.W. Simpson, Y.H. Schukken, M.J. Stanhope, Genome based phylogeny and comparative genomic analysis of intra-mammary pathogenic *Escherichia coli*, PLoS One 10 (2015) e0119799.

[19] A.S. Bukh, H.C. Schonheyder, J.M. Emmersen, M. Sogaard, S. Bastholm, P. Roslev, *Escherichia coli* phylogenetic groups are associated with site of infection and level of antibiotic resistance in community-acquired bacteraemia: a 10 year population-based study in Denmark, J. Antimicrob. Chemother. 64 (2009) 163–168.

[20] C. Carlos, M.M. Pires, N.C. Stoppe, E.M. Hachich, M.I. Sato, T.A. Gomes, L.A. Amaral, L.M. Ottoboni, *Escherichia coli* phylogenetic group determination and its application in the identification of the major animal source of fecal contamination, BMC Microbiol. 10 (2010) 161.

[21] I. Grosse, P. Bernaola-Galvan, P. Carpena, R. Roman-Roldan, J. Oliver, H.E. Stanley, Analysis of symbolic sequences using the Jensen-Shannon divergence, Phys. Rev. E Stat. Nonlinear Soft Matter Phys. 65 (041905) (2002).

[22] A. Liaw, M. Wiener, Classification and regression by randomForest, R News 2 (2002) 18–22.

[23] M.V. Koutras, S. Bersimis, P.E. Maravelakis, Statistical process control using Shewhart control charts with supplementary runs rules, Methodol. Comput. Appl. Probab. 9 (2007) 207–224.

[24] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol. 10 (2009) R25.

[25] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, S. Genome Project Data Processing, The sequence alignment/map format and SAMtools, Bioinformatics 25 (2009) 2078–2079.

[26] A.L. Delcher, S. Kasif, R.D. Fleischmann, J. Peterson, O. White, S.L. Salzberg, Alignment of whole genomes, Nucleic Acids Res. 27 (1999) 2369–2376.

[27] L.S. Hibbard, Region segmentation using information divergence measures, Med. Image Anal. 8 (2004) 233–244.

[28] S. Das, P. Duggal, R. Roy, V.P. Myneedu, D. Behera, H.K. Prasad, A. Bhattacharya, Identification of hot and cold spots in genome of *Mycobacterium tuberculosis* using Shewhart control charts, Sci. Rep. 2 (2012) 297.

[29] I. Comas, J. Chakravartti, P.M. Small, J. Galagan, S. Niemann, K. Kremer, J.D. Ernst, S. Gagneux, Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved, Nat. Genet. 42 (2010) 498–503.

[30] R. Brosch, S.V. Gordon, M. Marmiesse, P. Brodin, C. Buchrieser, K. Eiglmeier, T. Garnier, C. Gutierrez, G. Hewinson, K. Kremer, L.M. Parsons, A.S. Pym, S. Samper, D. van Soolingen, S.T. Cole, A new evolutionary scenario for the *Mycobacterium tuberculosis* complex, Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 3684–3689.

[31] S. Das, T. Roychowdhury, P. Kumar, A. Kumar, P. Kalra, J. Singh, S. Singh, H.K. Prasad, A. Bhattacharya, Genetic heterogeneity revealed by sequence analysis of *Mycobacterium tuberculosis* isolates from extra-pulmonary tuberculosis patients, BMC Genomics 14 (2013) 404.

[32] I. Comas, S. Homolka, S. Niemann, S. Gagneux, Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies, PLoS One 4 (2009) e7815.

[33] S. Homolka, M. Projahn, S. Feuerriegel, T. Ubben, R. Diel, U. Nubel, S. Niemann, High resolution discrimination of clinical *Mycobacterium tuberculosis* complex strains based on single nucleotide polymorphisms, PLoS One 7 (2012) e39855.

[34] N. Casali, V. Nikolayevskyy, Y. Balabanova, S.R. Harris, O. Ignatyeva, I. Kontsevaya, J. Corander, J. Bryant, J. Parkhill, S. Nejentsev, R.D. Horstmann, T. Brown, F. Drobniewski, Evolution and transmission of drug-resistant tuberculosis in a Russian population, Nat. Genet. 46 (2014) 279–286.

[35] H. Zhang, D. Li, L. Zhao, J. Fleming, N. Lin, T. Wang, Z. Liu, C. Li, N. Galwey, J. Deng, Y. Zhou, Y. Zhu, Y. Gao, T. Wang, S. Wang, Y. Huang, M. Wang, Q. Zhong, L. Zhou, T. Chen, J. Zhou, R. Yang, G. Zhu, H. Hang, J. Zhang, F. Li, K. Wan, J. Wang, X.E. Zhang, L. Bi, Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance, Nat. Genet. 45 (2013) 1255–1260.

[36] M.R. Farhat, B.J. Shapiro, K.J. Kieser, R. Sultana, K.R. Jacobson, T.C. Victor, R.M. Warren, E.M. Streicher, A. Calver, A. Sloutsky, D. Kaur, J.E. Posey, B. Plikaytis, M.R. Oggioni, J.L. Gardy, J.C. Johnston, M. Rodrigues, P.K. Tang, M. Kato-Maeda, M.L. Borowsky, B. Muddukrishna, B.N. Kreiswirth, N. Kurepina, J. Galagan, S. Gagneux, B. Birren, E.J. Rubin, E.S. Lander, P.C. Sabeti, M. Murray, Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*, Nat. Genet. 45 (2013) 1183–1189.