Published online 16 July 2014

Nucleic Acids Research, 2014, Vol. 42, No. 15 ***e122***
doi: 10.1093/nar/gku585

Inferring population structure and relationship using minimal independent evolutionary markers in Y-chromosome: a hybrid approach of recursive feature selection for hierarchical clustering

Amit Kumar Srivastava[1], Rupali Chopra[1], Shafat Ali[1], Shweta Aggarwal[1], Lovekesh Vig[2] and Rameshwar Nath Koul Bamezai[3,*]

[1]National Centre of Applied Human Genetics, School of Life Sciences, Jawaharlal Nehru University, New Delhi 110067, India, [2]School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi 110067, India and [3]National Centre of Applied Human Genetics, School of Life Sciences, and School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi 110067, India

Received November 22, 2013; Revised June 10, 2014; Accepted June 16, 2014

ABSTRACT

Inundation of evolutionary markers expedited in Human Genome Project and 1000 Genome Consortium has necessitated pruning of redundant and dependent variables. Various computational tools based on machine-learning and data-mining methods like feature selection/extraction have been proposed to escape the curse of dimensionality in large datasets. Incidentally, evolutionary studies, primarily based on sequentially evolved variations have remained un-facilitated by such advances till date. Here, we present a novel approach of recursive feature selection for hierarchical clustering of Y-chromosomal SNPs/haplogroups to select a minimal set of independent markers, sufficient to infer population structure as precisely as deduced by a larger number of evolutionary markers. To validate the applicability of our approach, we optimally designed MALDI-TOF mass spectrometry-based multiplex to accommodate independent Y-chromosomal markers in a single multiplex and genotyped two geographically distinct Indian populations. An analysis of 105 world-wide populations reflected that 15 independent variations/markers were optimal in defining population structure parameters, such as F_{ST}, molecular variance and correlation-based relationship. A subsequent addition of randomly selected markers had a negligible effect (close to zero, i.e. 1 × 10^{-3}) on these parameters. The study proves efficient in tracing complex population structures and deriving relationships among world-wide populations in a cost-effective and expedient manner.

INTRODUCTION

Human population genetics has witnessed advances through inundation of thousands of evolutionary markers made known from Human Genome project (HGP) and the 1000 Genome Consortium (1000 GC) studies. Also, markers in haploid mitochondrial genome (1) and male-specific Y-chromosome (MSY) (2) are incidentally categorized under haplogroups on the basis of sequential events of ancestral and acquired mutations in a time frame of human evolution. However, these ever-increasing variations impose two major challenges to evolutionary studies in identifying population structure and their relationship. The abundant presence of redundant and inter-dependent variables gives rise to the problem of high dimensionality and high genotyping cost limiting the sample size for a study. An appropriate alternative to overcome these problems is to select and study highly informative independent variations, sufficient to infer populations' structure and relationship as precisely as inferred from a larger set of evolutionary markers. In the light of difficulties and proposed solution, pruning of redundant and dependent variations through adaptation and development of new approaches followed by low-cost genotyping technologies is essential.

In the past decade, various computational and statistical approaches based on Bayesian clustering (3–6), Wright–Fisher model (7) and machine learning and data mining methods (8,9) have revolutionized genetic studies to expedite processing of large datasets more precisely. However, most of the available models and algorithms inferring populations' structure and relationship consider vari-

[*]To whom correspondence should be addressed. Tel: +91 11 26704518; Fax: +91 11 26742211; Email: bamezai@hotmail.com

ables as independent events which remain partially true for sequentially evolved markers. Although few models exploiting machine learning and data mining-based feature selection/extraction methods have recently been proposed for minimizing redundancy and dependency in a variety of high dimensional biological data including genome-wide single nucleotide polymorphism (SNP) data (10–14), nevertheless evolutionary studies still suffer with the curse of dimensionality (15) due to absence of appropriate models/approaches dealing with sequentially evolved markers in haploid genome.

In view of a wide applicability of feature selection/extraction methods in high-dimensional biological data, current models dealing with genome-wide SNP data are based on either haplotype block-dependent pair-wise linkage disequilibrium (LD) (16,17) or haplotype block-independent *F*-test (18), *t*-test (18), χ^2-test and regression parameters (11,14). However, each of the proposed methods has its own strengths and limitations. Thus, there is a need for hybrid models exploiting both supervised and unsupervised machine learning methods.

In the current study, we used a correlation coefficient-based supervised feature selection method embedded with agglomerative hierarchical clustering based on prior knowledge of Y-chromosomal phylogeny. To validate our novel approach, we chose a model study based on real datasets of male-specific Y-chromosomal (MSY) variations generated in present and earlier studies. As per neutral theory of molecular evolution (7) and Kimura's step-wise mutation model (19), a major source of allelic diffusion in finite populations is fixation of neutral mutations by genetic drift, i.e. mutations occurring in steps are defined by state of variation occurred in the preceding generation. The same applies to Y-chromosome phylogeny as well, i.e. each haplogroup (combination of same or different haplotypes) is an outcome of one or more mutation event, which later on stabilizes under different evolutionary forces, such as migration, genetic drift, selection and admixture in a population or geographical region. Therefore, lower nodes in hierarchy appear in the background of already existing higher ones (Figure 1). In the background of the above fact, only few evolutionary markers which are most ancestral in their respective clades could be considered independent and rest are sequentially derived after the fixation and selection of ancestral ones (Figure 1).

At present, the hierarchical phylogeny of paternally inherited human Y chromosome with universal nomenclature by Y Chromosome Consortium (http://ycc.biosci.arizona.edu) consists of 20 major (A–T) and 311 divergent haplogroups, defined by 599 validated binary markers (20). This nomenclature denotes all major clades (haplogroups) by capital letters (e.g. A, B, C, etc.) and sub-clades either by numbers or small letters (e.g. H1a, H1b, R1a1, etc.) (21). However, an addition of 2870 variations in Y chromosome including two-third novel ones from the 1000 GC has differentiated further the already existing haplogroups/clades into more profound sub-haplogroups/sub-clades (21,22). In an ocean of a huge number of SNPs to be genotyped simultaneously and the limitations of the high-throughput technologies to provide desired outcome in a large dataset of diverse population groups, a scope of pruning of such variables is justified, even within Y chromosome alone. Additionally, the optimization of the procedure to genotype all independent markers in one go without compromising the quality of the results becomes critical.

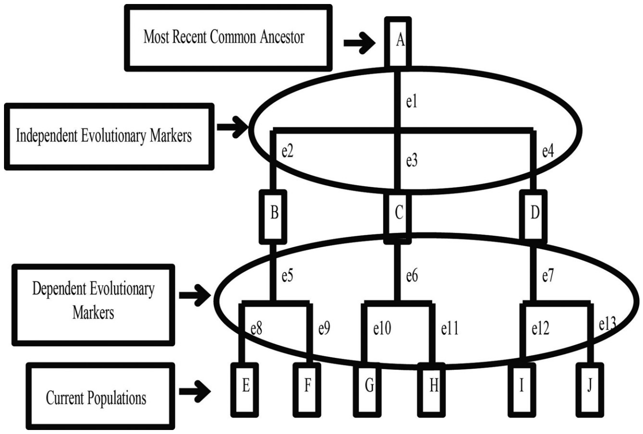
Most Recent Common Ancestor
A
Independent Evolutionary Markers
e1
e2
e3
e4
B
C
D
Dependent Evolutionary Markers
e5
e6
e7
e8
e9
e10
e11
e12
e13
Current Populations
E
F
G
H
I
J

Figure 1. Representative example of hierarchical events of mutations in evolution (as would happen say in the Y-chromosome) in human population. 'A' represents the most recent common ancestor with a genetic background with mutation e1. In the background of e1 three independent mutation events follow to give rise to three different clades 'B, C, D'. The variations originating in lower nodes later would represent the ancestors of their respective clades.

Generally, evolutionary studies prefer medium throughput techniques (suitable for hundreds of SNPs in large sample size) over high-throughput technologies (suitable for millions of SNPs in limited sample size), since evolutionarily conserved SNPs are limited in numbers and need to be genotyped in large sample size. Various medium-throughput technologies, e.g. matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) (23–33), TaqMan (34) and SNaPshot^{TM} (21,35–41) have been developed in the past few years and validated with respect to accuracy, sensitivity, flexibility in assay designing and cost per genotype (42–44). Based on the requirement and above-mentioned criterion, MALDI-TOF-MS-based iPLEX GOLD assay from SEQUENOM, Inc. (San Diego, CA, USA) was used for multiplex genotyping of Y-chromosome SNPs in the present study.

Current study (Figure 2) has taken care of the problems of high-dimensionality and expensive genotyping methods simultaneously. The problem of high-dimensionality was attended to by the selection of highly informative independent Y-chromosomal markers (features) through a novel approach of 'recursive feature selection for hierarchical clustering (RFSHC)'. Our approach utilized recursive selection of features through variable ranking on the basis of Pearson's correlation coefficient (PCC) embedded with agglomerative (bottom up) hierarchical clustering based on judicious use of phylogeny of Y-chromosomal haplogroups. The approach was initially applied on a dataset of 50 populations. Later, observations from above dataset were confirmed on two datasets of 79 and 105 populations. Several computational analyses such as principal component analysis (PCA) plots, cluster validation, purity of clus-

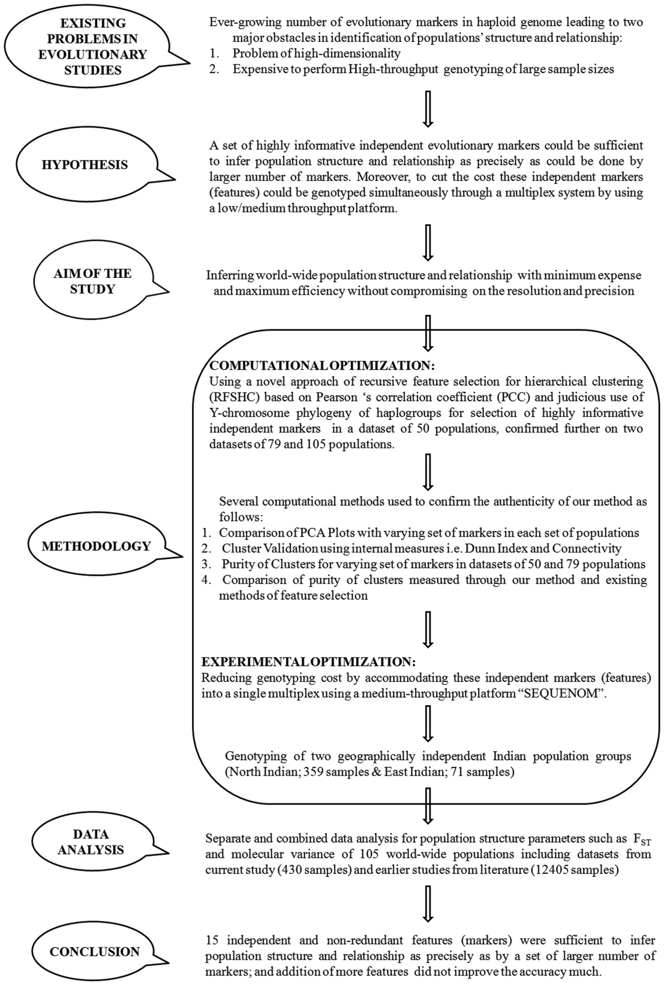
EXISTING PROBLEMS IN EVOLUTIONARY STUDIES
Ever-growing number of evolutionary markers in haploid genome leading to two major obstacles in identification of populations' structure and relationship:
1. Problem of high-dimensionality
2. Expensive to perform High-throughput genotyping of large sample sizes
HYPOTHESIS
A set of highly informative independent evolutionary markers could be sufficient to infer population structure and relationship as precisely as could be done by larger number of markers. Moreover, to cut the cost these independent markers (features) could be genotyped simultaneously through a multiplex system by using a low/medium throughput platform.
AIM OF THE STUDY
Inferring world-wide population structure and relationship with minimum expense and maximum efficiency without compromising on the resolution and precision
COMPUTATIONAL OPTIMIZATION:
Using a novel approach of recursive feature selection for hierarchical clustering (RFSHC) based on Pearson 's correlation coefficient (PCC) and judicious use of Y-chromosome phylogeny of haplogroups for selection of highly informative independent markers in a dataset of 50 populations, confirmed further on two datasets of 79 and 105 populations.
METHODOLOGY
Several computational methods used to confirm the authenticity of our method as follows:
1. Comparison of PCA Plots with varying set of markers in each set of populations
2. Cluster Validation using internal measures i.e. Dunn Index and Connectivity
3. Purity of Clusters for varying set of markers in datasets of 50 and 79 populations
4. Comparison of purity of clusters measured through our method and existing methods of feature selection
EXPERIMENTAL OPTIMIZATION:
Reducing genotyping cost by accommodating these independent markers (features) into a single multiplex using a medium-throughput platform "SEQUENOM".
Genotyping of two geographically independent Indian population groups (North Indian; 359 samples & East Indian; 71 samples)
DATA ANALYSIS
Separate and combined data analysis for population structure parameters such as F_{ST} and molecular variance of 105 world-wide populations including datasets from current study (430 samples) and earlier studies from literature (12405 samples)
CONCLUSION
15 independent and non-redundant features (markers) were sufficient to infer population structure and relationship as precisely as by a set of larger number of markers; and addition of more features did not improve the accuracy much.

Figure 2. A flow-chart representing problems in evolutionary studies, our hypothesis, aim of the study with step-wise methodology adopted and conclusion.

ters and their comparison with already existing methods of feature selection were performed to prove the authenticity of our novel approach. Further, to cut the cost as much as possible without compromising on the ability of estimating population structure, these independent markers were multiplexed together into a single multiplex by using a medium-throughput MALDI-TOF-MS platform 'SEQUENOM'. In addition, recently evolved haplogroups representing lower nodes in Y-chromosome hierarchy were accommodated in subsequent three multiplexes in a continent-specific manner to check even minor changes in the resolution of population structure and relationship, if any. Moreover, newly designed multiplexes consisting of highly informative-independent features were genotyped for two geographically independent Indian population groups (North India and East India) and data was analyzed along with 105 world-wide populations (datasets of 50, 79 and 105 populations) for population structure parameters such as population differentiation (F_{ST}) and molecular variance. The results illustrated that an optimal set of 15 independent Y-chromosomal markers was sufficient to infer populations' structure and relationship with equivalent resolution and precision as would be deduced after the use of a larger set of markers (Figure 2).

MATERIALS AND METHODS

Feature selection

As numerous evolutionary variations defining population structure and relationship include only few independent markers which are most ancestral in their respective clades, the inter-dependence of these markers is well established (Figure 1). In such circumstances, population structure and relationship could be defined as follows:

$$W = [W1 + a(W2)] - b \tag{1}$$

where W is final observation represented by population differentiation, diversity or variance, $W1$ is output from minimal-independent markers, $W2$ is output from randomly selected variables dependent on $W1$, i.e. $W2 = \mathrm{f}(W1)$, a and b are constants where a is no of randomly selected factors and b is factor of reduction.

In case of perfect dependence of representative markers on their ancestral ones, above equation would be shaped as

$$a(W2) = b \text{ or } a(W2) - b = 0 \tag{2}$$

i.e.

$$W = W1 \tag{3}$$

However, we cannot assume a practical situation where representative markers would be perfectly dependent on their progenitor, i.e. zero effect of these markers on population structure and relationship, utmost negligible effect could be considered, i.e.

$$a(W2) - b \approx 0 \tag{4}$$

With above background, we adopted an approach of feature selection through variable ranking based on PCC to minimize the inter-dependency derived redundancy of evolutionary (Y-chromosomal) markers and select highly informative-independent ones. Since, Y-chromosomal markers are usually genotyped in population-specific manner; availability of large dataset regarding these markers from world-wide populations was major limitation for validation of our novel approach. Therefore, markers were appropriately selected on the basis of prior knowledge of phylogeny of Y-chromosomal haplogroups. To avoid any bias during selection process, we extracted markers representing higher and lower nodes in population tree simultaneously. Therefore, let E be the collection of these evolutionary markers which is defined as

$$E = \{e1, e2, \ldots, eN\} \tag{5}$$

where $|E| = N$.

We first generated a square symmetrical Pearson's correlation matrix for appropriately selected variables ($N \times N$) using PCA. For example, matrix 'M' for random evolutionary markers/variables $e1$, $e2$, $e3$, ..., eN can be represented as follows:

$$M = \begin{pmatrix} \mathrm{corr}(e1e1)\mathrm{corr}(e1e2)\mathrm{corr}(e1e3)\ldots\mathrm{corr}(e1eN) \\ \mathrm{corr}(e2e1)\mathrm{corr}(e2e2)\mathrm{corr}(e2e3)\ldots\mathrm{corr}(e2eN) \\ \mathrm{corr}(e3e1)\mathrm{corr}(e3e2)\mathrm{corr}(e3e3)\ldots\mathrm{corr}(e3eN) \end{pmatrix}$$

Correlation among variables is denoted by PCC (45):

$$k_{e1e2} = \frac{\sum_{i=1}^{n} (fe1_i - \overline{fe1})(fe2_i - \overline{fe2})}{\sqrt{\sum_{i=1}^{n} (fe1_i - \overline{fe1})^2 (fe2_i - \overline{fe2})^2}} \qquad (6)$$

where k_{e1e2} is correlation coefficient for random variables $e1$ and $e2$ representing n number of populations, f_{e1i} and f_{e2i} are frequencies of $e1$ and $e2$ for population i, f_{e1} and f_{e2} are average frequencies of $e1$ and $e2$. From Equation (6), the correlation coefficient would be +1 in case of a perfect positive (increasing) linear relationship (correlation) and −1 in case of a perfect negative (decreasing) linear relationship (anti-correlation), whereas values between −1 and 1 in all other cases indicate the degree of linear dependence between the variables. As it approaches zero there would be minimum relationship (closer to uncorrelated), i.e. the closer the coefficient value to either −1 or 1, the stronger the correlation between the variables.

On the basis of PCC-based variable ranking, we observed that few markers, considered as independent signatures for diversification of male populations world-wide were highly correlated. However, we could not have merged two such markers providing independent signature for Y-chromosomal haplogroups, knowing the fact that these markers are located in non-recombining Y-chromosome which itself is haploid in nature representing a haplotype block and thereby, forms the basis for close correlation. This situation is unlike autosomal SNPs where both conditions, i.e. haplotype block-dependent and haplotype block-independent are considerable. Therefore, we embedded feature selection with agglomerative (bottom up) hierarchical clustering of haplogroups on the basis of the prior knowledge of phylogeny of Y-chromosomal haplogroups to minimize the redundancy generated by markers representing lower nodes in Y-chromosomal hierarchy and depending on the higher nodes of their respective clades (Figures 1 and 3). With this approach, sub-clades were clustered into their respective major clades and again pruned on the basis of PCC. The above step was repeated till we reached the most ancestral nodes (12 markers) of Y-chromosome phylogeny (Supplementary Table S1a–i) and the procedure named as RFSHC.

Computational approach

We initially generated a correlation matrix of 32 common Y-chromosomal markers from 50 populations using PCA. We observed that few markers such as 'H*', 'H1', 'J*' and 'O' were closely and significantly related to each other (correlation coefficient ≥ 0.78) (Supplementary Figure S1a). Similarly we observed two separate sets of close variables: 'C3', 'K*', 'R*'and 'NO*', 'Q' (correlation coefficient ≥ 0.68) (Supplementary Figure S1a). Since 'H', 'J', 'O', 'Q' and 'R' are major haplogroups of human Y chromosome phylogeny, random removal or merging of variables could disturb the harmony of Y-chromosomal haplogroups' phylogeny. Hence, we embedded feature selection with agglomerative hierarchical clustering of sub-haplogroups into major haplogroups on the basis of prior knowledge of phylogeny of Y-chromosomal haplogroups. This approach led to moving one level up in hierarchy and in the next step,

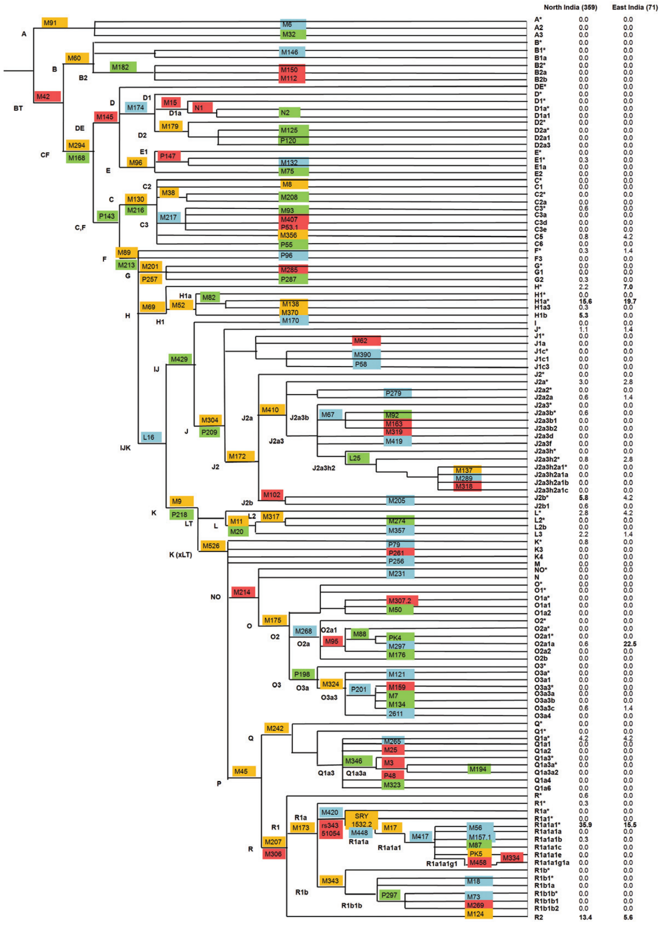

Figure 3. Hierarchical phylogeny based on 127 successfully worked Y-chromosome SNPs, genotyped through four systematically designed multiplexes, yellow highlighted SNPs represent PLEX 1, green highlighted SNPs represent the PLEX 2, blue highlighted SNPs represent the PLEX 3 and red highlighted SNPs represent the PLEX 4.

correlation matrix was generated on the basis of 25 variables obtained by merging of 'G1 and G2', 'H* and H1', 'J* and J1', 'J2*, J2a and J2b', 'L*, L2 and L3', 'R1a1* and R1a1a', 'R*, R1b* and R1b1' into their respective major clades 'G', 'H', 'J*', 'J2', 'L', 'R1a1' and 'R*' (Supplementary Table S1a and b). We observed that values of correlation coefficients decreased at this step (Supplementary Figure S1b). However, some of the above mentioned highly correlated variables at the step of 32 SNPs which could not be removed for being critical nodes in evolutionary tree were still closely related. Therefore, we again merged 'L, T, K*, NO*, N, O and Q' into a major clade 'K,T (xR)' and generated a correlation matrix on the basis of 15 markers (Supplementary Table S1c). The matrix showed minimum correlation among all the markers except 'F* and H' (correlation coefficient = 0.59, which itself is quite low) (Supplementary Figure S1c). To rule out any possibility of interdependence generated redundancy, the process was repeated till 12 most ancestral markers (Supplementary Table S1d, Supplementary Figure S1d) and evaluated in a set of 79 and 105 populations besides initial set of 50 populations (Supplementary Table S1e–i). Interestingly, we observed that a set of

15 markers was most optimal in all sets of populations for defining population structure and relationship as confirmed by PCA plots, cluster validation, purity of clusters and their comparison with two independent existing methods (information gain and χ^2) of feature selection.

SNP selection for multiplex designing

Initially, SNPs were chosen from non-recombining Y chromosome (NRY), based on their position in Y chromosome hierarchical phylogenetic tree and the distribution of paternal haplogroups in different geographic and ethnic groups. A total of 1551 polymorphisms including 599 SNPs depicting 311 haplogroups (20) along with new entries from International Society of Genetic Genealogy (ISOGG) and Family Tree (FT) DNA Database were used to precisely select 133 SNPs covering nearly all major world-wide haplogroups (A–R) and their sub-haplogroups. We aimed to accommodate nearly all independent haplogroups (A–R) in a single, i.e. first multiplex. Additionally, second, third and fourth multiplexes were designed for sub-clades/haplogroups, sub-subclades/haplogroups, respectively. Third and fourth multiplexes were specifically selected for Eurasian haplogroups and sub-haplogroups, e.g. H, J, O, R and their sub-clades, to examine the effect of recently evolved evolutionary markers on the resolution of populations' structure and relationship.

Multiplex designing

SEQUENOM, Inc. provides its own software 'MassARRAY® Assay Design 3.1' for multiplex primer designing which can accommodate upto 40 SNPs in one well till date. Multiplexing is a five step process: (i) rs sequence retriever: downloads flanking sequence of every known SNP from NCBI—dbSNP by using their rs ID, in case SNP does not have rs ID, the flanking sequence can be manually downloaded from NCBI (http://www.ncbi.nlm.nih.gov/mapview) database. (ii) ProxSNP: searches for any proximal SNP in the flanking region of desired SNP (usually 200 bp flank is provided for this step). (iii) PreXTEND: designs pre-extension PCR primers in the output of ProxSNP (usually 80–120 bp PCR product is optimum for further UEP designing). (iv) Assay design: designs extension primers for extension PCR within the amplicon of pre-extension PCR which binds to one nucleotide upstream to the polymorphic loci [locus]. Extension primers are highly specific to the polymorphic loci, as iPLEX reaction products have minimum 16 Da difference in mass (Supplementary Table S2) (46). (v) PleXTEND: validates multiplex assays.

Taking the advantage of these features, a total of 206 SNPs representing nearly all major clades and sub-clades of Y-chromosome phylogeny along with their 200 bp flanks were processed using online tools (ProxSNP and PreXTEND). However, 18 SNPs could not pass the criteria of software for multiplex assay designing and 188 SNPs were used for assay design software. Out of 188 SNPs, we first selected 15 highly informative independent SNPs to accommodate in a single multiplex. Since assay design software from SEQUENOM, Inc. allowed us to accommodate up to 40 SNPs in a single multiplex, we super-plexed the initial multiplex of the 15 independent variables with rest of the SNPs to accommodate 22 more SNPs representing major clades (haplogroups) or sub-clades (sub-haplogroups) for fill-in purpose only. However, in this process of fill-in, four independent SNPs were left out and accommodated into subsequent multiplexes. Once first multiplex was ready, subsequent multiplexes were designed by critical selection of important SNPs representing sub-clades and sub-subclades for affirmative purposes only. All four multiplexes together accommodated 133 SNPs whereas rest were included in many multiplexes consisting very low number of markers and therefore, left out. While assay designing the default settings of amplicon length in a range of 80–120, primer length (17–24) and T_m (45–60°C) were maintained to obtain maximum efficiency. Based on our multiplexing criteria (of systematic approach with cost-effectiveness and high-throughput precision) for high-resolution mapping of Y chromosome phylogeny, 133 critically important SNPs were chosen for generating four multiplexes, with 37 SNPs in PLEX 1, 36 SNPs in PLEX 2, 32 SNPs in PLEX 3 and 28 SNPs in PLEX 4 (Supplementary Table S3). Finally, all pre-extension and extension primers were checked for any cross-complementation throughout the genome and within primers to ensure perfect specificity.

DNA samples

Peripheral blood samples were drawn from 359 healthy controls from North India and 71 healthy controls from Orissa, India after seeking their consent and approval of JNU ethical Committee. DNA was isolated from 500 μl blood by phenol–chloroform method and dissolved in TE [Tris–HCl (pH 8.0) + ethylinediaminetetraacetatic acid (EDTA)] buffer for long storage. Genomic DNA was quantitated on Nanodrop (ND1000) and checked for quality through agarose gel (0.8%) electrophoresis.

Multiplex PCR amplification

PCR amplification was accomplished by using 20–25 ng DNA templates, 10× PCR buffer (1×/reaction), 25 mM $MgCl_2$ (2 mM/reaction), 25 mM deoxyribonucleotide (dNTPs) (500 μM/reaction), 1 μM extension primer mix (200 nM each/reaction) and 1 unit of Hot Start polymerase (all reagents were provided in PCR kit by SEQUENOM, Inc.) for 5 μl reaction in following amplification conditions: 94°C for 4 min followed by 40 cycles of 94°C for 20 s, 56°C for 30 s, 72°C for 1 min and final extension of 72°C for 3 min followed by 4°C-hold.

SAP treatment

Amplification products were SAP (shrimp alkaline phosphatase) treated (37°C for 40 min, 85°C for 5 min, 4°C-hold) to remove unincorporated nucleotides. Attractiveness of this assay is that exonuclease treatment is not required for left-over primers as they have a 10-mer tag (5′-ACGTTGGATG-3′) at 5′ end which enables them to lie outside the window (4500 – 9000 Da).

UEP (un-extended primer) adjustment

For iPLEX Gold reaction, the concentration of extension primers was optimized as per instructions provided by SEQUENOM, Inc. to control signal-to-noise ratio. Prior to genotyping, a mix of un-extended primers (UEPs) was run on spectroCHIP and analyzed in Typer 4.0 providing an adjustment sheet. Based on the adjustment sheet, extension primers were divided into three sets: low mass UEP, medium mass UEP and high mass UEP. Concentration of high mass and medium mass UEPs was increased accordingly to get detectable peak intensity.

iPLEX extension reaction

iPLEX Gold reaction was set using 10× iPLEX Gold Buffer (1×/reaction), 10× iPLEX termination mix (1×/reaction), iPLEX enzyme (0.041 μl/reaction) and primer mix (all reagents were provided in PCR kit by SEQUENOM, Inc.) in 2 μl reaction volume. Reaction was cycled by using 200-short-step program including two loops of cycles. Initial denaturation was done at 94°C for 30 s followed by annealing at 52°C for 5 min and extension at 80°C for 5 min. The annealing and extension steps were repeated for 11 cycles and looped back to denaturing step (94°C for 5 min) for 40 cycles. A final extension was done at 72°C for 3 min followed by 4°C-hold.

Clean resin

iPLEX Gold reaction products were washed by resin beads (6 mg/well) to remove salts which interferes with MALDI-TOF MS.

MALDI-TOF mass spectrometry

Cleaned samples were spotted on SpectroCHIP by 'Nanodispenser' (spotting speed and volume is set according to the environment for example in moist environment comparatively less volume at slow speed should be dispensed). Finally, spotted SpectroCHIP was exposed to a LASER (light amplification by stimulated emission of radiation) beam and allelic discrimination was obtained by mass to charge (m/z) ratio. Output files were in .xml format and data was analyzed using Typer 4.0 software.

To check the accuracy of the technique, we genotyped randomly selected samples as positive controls using same platform and observed 100% concordance with the results obtained earlier. Additionally, 2 blank wells/96 wells were used to avoid any false positive result or cross-contamination. We also used two markers for 10 haplogroups, (CF, C, F, G, J, K, L, O, R, R1a and R1a1) and observed 100% genotyping concordance in results. Peak intensity of each SNP was sufficient for detection with average call rate of >95% (Supplementary Figure S2a, b, c and d).

Statistical analysis

Statistical analyses were carried out to show the applicability of our approach with designed multiplexes for refining populations' ancestry, homogeneity, differentiation and variance. Firstly, the data generated in the present study was analyzed separately. Later, this data was compared to a large dataset from 105 distinct world-wide populations including 12 835 samples on the basis of common haplogroup frequency (Supplementary Table S4). Pearson's correlation matrix of Y-chromosome SNPs/haplogroups was obtained by PCA through XLSTAT add-in (www.xlstat.com/). Homogeneity/stratification in present data was checked by using PCA through PLINK 1.07 (www.pngu.mgh.harvard.edu/~purcell/plink/) and three components representing highest cumulative variability were plotted by using SPSS 17.0. Also, populations' correlation in combined dataset was verified by PCA based on haplogroup frequency in MS-Excel by using XLSTAT add-in (www.xlstat.com/). Internal measures (Dunn index and connectivity) of population clusters ranging from three to seven were validated through different clustering methods, like Hierarchical clustering, K-medoid, K-means by using clValid R package (http://cran.r-project.org/web/packages/clValid/index.html/). Further, purity of most appropriate clusters obtained through above means was checked through hierarchical clustering using KKNN R package (http://cran.r-project.org/web/packages/kknn/index.html/). To prove the authenticity of our approach, the purity of appropriate clusters obtained through RFSHC was compared with two independent existing methods (information gain and χ^2) of feature selection for varying set of markers (32, 25, 15 and 12) in datasets of 50 and 79 populations. Population differentiation indices (F_{ST} values) were calculated by ARLEQUIN v3.0. Pair-wise F_{ST} values between populations were graphically represented on a color-coded heatmap by using Gplots R package (http://cran.r-project.org/web/packages/gplots/index.html/).

RESULTS

In order to reach an optimal number of independent variables (evolutionary markers/SNPs) for resolving the population structure and relationship world-wide, we applied a combined approach of feature selection and hierarchical clustering for pruning of variables in human Y-chromosome (Figure 3). At each step, optimization was validated by several computational simulations, such as comparison of PCA plots, evaluation of population clusters and their validation, scrutiny of the purity of the resulting clusters and their comparison with already existing methods of feature selection. Population clustering was performed through three different methods, namely hierarchical clustering, K-medoid and K-means. The most optimal cluster size for each population set was determined by considering the PCA plots of populations (Figure 4), followed by evaluation of the Dunn index (47) and connectivity (48) for all cluster sizes (3–7) with different sets of markers (Supplementary Figure S3a, b and c). Later, the purity of clusters was compared with different marker sets for the most appropriate cluster size in each population set (Figure 5). Purity of clusters (Y-axis) as a measure of varying number of markers (X-axis) is represented in Figure 6a and b for a set of 50 and 79 populations, respectively. Population clustering ability of our methodology was also compared with two existing feature selection methods of information gain and χ^2 (Table 1). These formed the basis for systematically

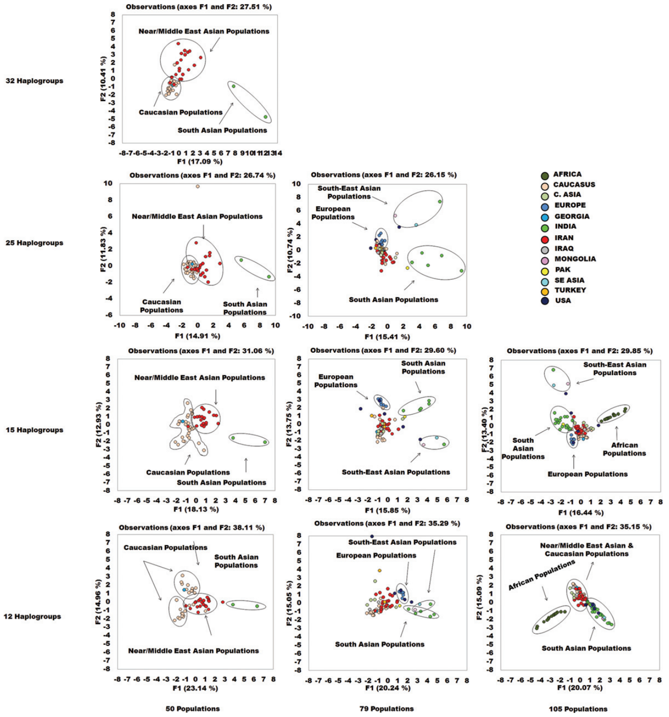

Figure 4. Structure of South Asian (different regions of India including our lab data; Sharma *et. al.,* (49) and Pakistan); Caucasus; Near/Middle East (Iran, Georgia and Turkey); Central Asian (Gulf Countries and Iraq); South East Asian including Mongolians and others; European; USA and African populations using principal component analysis (PCA), based on 15, 25 and 32 common haplogroups (variables) for a set of 50, 79 and 105 populations.

designing the multiplexes to accommodate independent Y-chromosome evolutionary markers in a single multiplex and generate three subsequent continent-specific multiplexes for recently evolved populations.

To validate the utility of our approach with the designed multiplexes, we genotyped two geographically distinct Indian populations (359 North Indian and 71 East Indian healthy controls) for all four multiplexes with the optimal number of 133 markers, of which 127 SNPs worked successfully, depicting 123 distinct Y-chromosome haplogroups including 2 super haplogroups, 17 major haplogroups, 29 subhaplogroups and 75 sub-subhaplogroups (Figure 3). We observed a total of 28 divergent haplogroups (excluding superhaplogroups and major haplogroups) with at least one sample in each group. The details of major contributors are provided in Figure 3. The data was also analyzed in 105 worldwide populations with a dataset of 12 835 samples (Supplementary Table S4).

Pruning of variables through correlation matrix among Y-chromosome markers

Initial analysis in a combined dataset of 50 populations (4682 samples from South Asia, Caucasus and Near/Middle East) indicated that correlation of variables decreased with present approach (Supplementary Figure S1). Matrix of precisely selected 32 Y-chromosome haplogroups including major and minor nodes from available data in literature represented many haplogroups in close correlation as discussed in computational approach. However, by embedding feature selection with agglomerative hierarchical clustering approach, we eventually reached an optimal set of 15 non-redundant and independent Y-chromosome haplogroups which could lead to a similar resolution of population structure as was obtained by higher number of variables say, 25, 32 or even 127 (present study). Later, analysis was repeated in a set of 79 populations (10 890 samples from diverse geographical regions, e.g. South Asia including major geographic regions of India (49) and Pakistan, Caucasus, Near/Middle East, Central Asia, South-East Asia, Russia, Europe and USA) and 105 populations (12 835 samples from diverse regions of world) (Supplementary Table S4) to confirm the results obtained in the initial analysis.

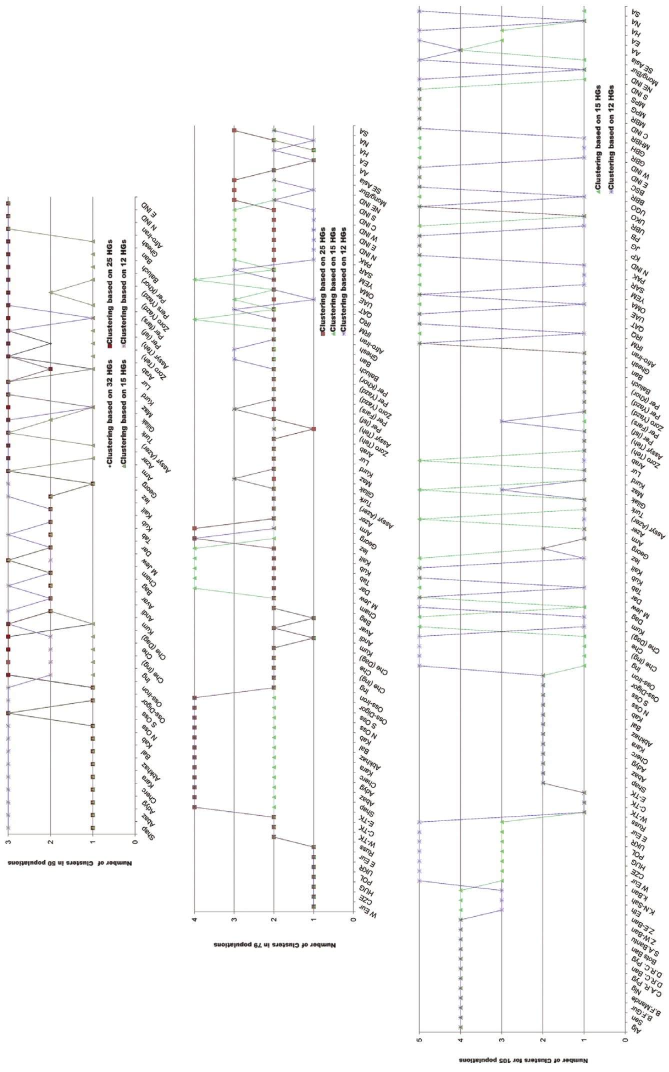

Figure 5. Agglomerative hierarchical clustering of different set of populations (50, 79 and 105) with varying set of markers (32, 25, 15 and 12) using average distance method. X-axis and Y-axis denote populations and number of clusters respectively. Based on the result of cluster validation and PCA plots, 3, 4 and 5 clusters were defined for 50, 79 and 105 populations, respectively.

Table 1. Comparison of the proposed method of ‘RFSHC’ and two already existing independent methods of feature selection

(a) Comparative tables of purity of clusters (weighted average) derived from independent haplogroups

Method of feature selection	Number of independent features		Purity of clusters (%)	
	50 Populations	79 Populations	50 Populations	79 Populations
Information gain	14	13	67.48	68.33
χ^2	14	12	67.48	68.33
RFSHC	15	15	86.03	80.68

(b) Details of independent haplogroups derived from proposed and existing methodologies of feature selection

RFSHC	Information gain	χ^2	RFSHC	Information gain	χ^2
Independent HGs in a dataset of 50 populations			Independent HGs in a dataset of 79 populations		
A	E	E	A	E	E
B	G2	G2	B	C3	G
D	I	I	D	G	H
E	J1	J1	E	H	I
C*	J2a	J2a	C*	I	J*
C3	J2b	J2b	C3	J*	J2
F*	L*	L*	F*	J2	L
G	L2	L2	G	L	T
H	L3	L3	H	T	Q
IJ*	T	T	IJ*	Q	R1*
I	Q	Q	I	R1*	R1a1
J*	R1a1a	R1a1a	J*	R1a1	R2
J2	R1b1	R1b1	J2	R2	×
K,T	R2	R2	K,T	×	×
R	×	×	R	×	×

Comparison of PCA Plots in combined datasets

A combined data analysis of world-wide populations was performed on the basis of 32, 25, 15 and 12 common haplogroups in 50 populations (Supplementary Table S5a–d); 25, 15 and 12 common haplogroups in 79 populations (Supplementary Table S5e, f and g), and 15, 12 common haplogroups for 105 populations (Supplementary Table S5h and i). Comparison of PCA plots was made in two ways: (i) with different set of markers for same number of population and (ii) with different set of populations for same number of common markers. All four sets of markers, i.e. 32, 25, 15 and 12 common haplogroups could only be used for the first dataset of 50 populations. Due to limitation of data available from literature, we could not include higher number of markers in subsequent steps of analysis. Comparison of the PCA plots based on 32, 25, 15 and 12 common haplogroups for 50 populations [4682 samples from South Asia (India (49) and Pakistan), Caucasus and Near/Middle East (Iran and Georgia)] depicted the retention of three clusters of populations up to 15 markers, which was completely distorted with 12 markers. Although cluster of Caucasian populations was quite sparse in the PCA plot using 15 markers, these formed a single cluster, as observed in PCA plots with 25 or 32 markers; whereas PCA plot with 12 markers depicted two distinct clusters of Caucasian populations (Figure 4). This was more evident in further PCA plots based on 25, 15 and 12 common markers in the set of 79 populations (four clusters), and 15, 12 common markers in a set of 105 populations (5 clusters), representing similar resolution of population structure with a set of 25 or 15 markers but substantially deteriorated with a set of 12 markers in the same dataset (Figure 4). On the other hand, a comparison of PCA plots with increasing number of populations for the same number of common haplogroups showed an increase in the resolution of population structure with increasing number of populations (Figure 4).

Cluster validation and purity of clusters

Of the three essential measures: (i) internal, (ii) stability, (iii) biological (50) for cluster validation in any kind of clustering method, internal measures were used in this study for validation of clustering of population groups at different steps. The Dunn index (47) and connectivity (48) are popular internal measures of cluster quality indicating the maximization of inter-cluster distance, minimization of intra-cluster distance and consistency of nearest neighbor assignments, respectively. For an ideal clustering, Dunn index should be high and connectivity low.

For each population set the three clustering algorithms were run for different cluster sizes (3–7). Taking the PCA plots into consideration, cluster sizes of 3, 4 and 5 were observed to yield adequately high values of Dunn index and low values of connectivity for the set 50, 79 and 105 populations, respectively. Supplementary Figure S3a, b and c shows the relative values of these scores obtained for different cluster sizes. After determining the appropriate cluster sizes, i.e. 3, 4 and 5 for the set of 50, 79 and 105 populations, respectively, we examined the values of Dunn index and connectivity for these clusters using different marker sets (12, 15, 25 and 32 markers). It was observed that a set of 15 markers yielded the best results for the above parameters (Supplementary Figure S3a, b and c) and also for cluster

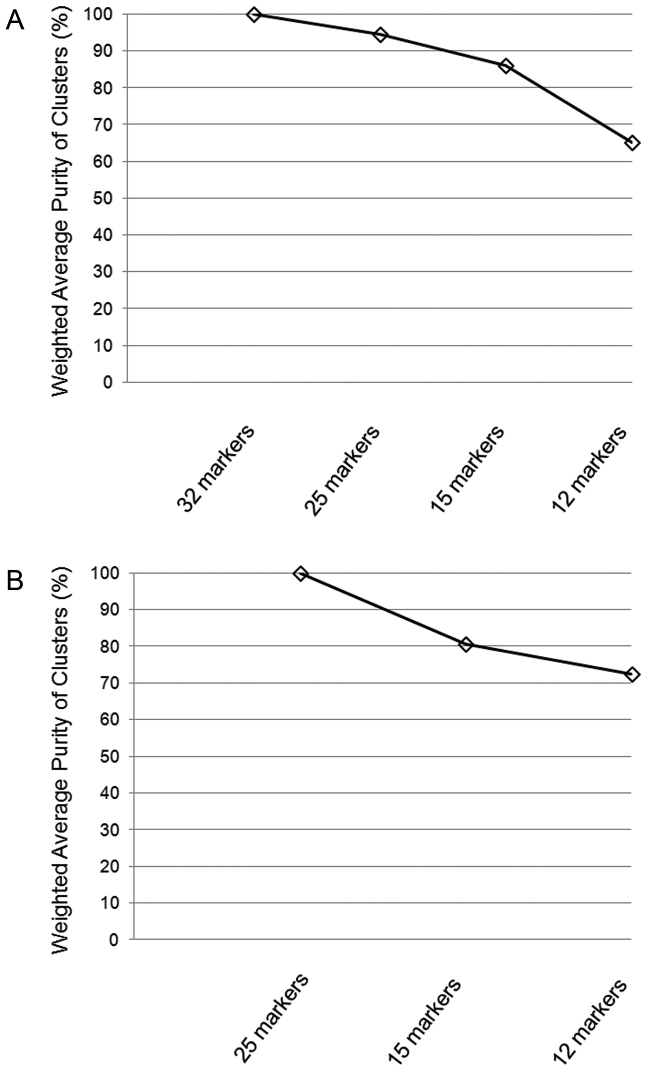

Figure 6. (**a** and **b**) A scatter plot of purity of clusters, as a measure of varying number of markers (32, 25, 15 and 12 for a set 50 populations) and (25, 15 and 12 for a set of 79 populations), respectively.

purity (Figures 5 and 6a and b) in each population set. The purity measure of most appropriate clusters was evaluated for population sizes 50 (with respect to 32 and 25 markers) and 79 (with respect to 25 markers) for different number of markers by using following formulae:

$$\text{purity measure} = \frac{\text{number of populations in a cluster by lower set of markers}}{\text{number of populations in the same cluster by higher set of markers}}$$

$$\text{weighted average purity of cluster in a population set} = \frac{\sum \text{purity measure} \times \text{number of populations in a given cluster}}{\text{total number of populations}}$$

Again, with the above calculation a set of 15 markers provided superior results in comparison to 12 markers and almost equivalent to 25 and 32 marker sets in all set of populations (Figures 5 and 6a and b). Though purity of clusters in a dataset of 50 populations was marginally higher for 25 markers than 15 markers, nevertheless the set of 15 markers justified its use for the identification of populations' structure and their relationship in various other ways. These were, better internal measures, i.e. higher Dunn index and lower connectivity, in comparison to 25 markers (Supplementary Figure S3a, b and c); retention of population structure in PCA plots as compared with 25 and 32 markers; and cost-effective genotyping as compared with the larger set of markers (25 or 32). The above features were suggestive of the use of 15 markers as an optimal set of Y-chromosomal markers.

Comparison of purity of clusters obtained through RFSHC with existing methods of feature selection

Purity of clusters obtained through the proposed 'RFSHC' approach was compared with two already existing methods of; information gain and chi square. The results represented that purity obtained through our approach was better in comparison to that obtained through above mentioned existing methods (Table 1a). Further, observation of independent haplogroups derived from different methods illustrated that a few major haplogroups skipped during the selection of independent features, which would disturb the Y-chromosomal phylogeny and therefore compromise the inference of population structure and their relationship.

Validation of approach through population structure parameters

To show the applicability of our approach and systematically designed multiplexes, we also compared population structure and stratification in present dataset independently with different number of variables (evolutionary markers) in uniform sample size (430 samples). In the background of initial ancestry information obtained through genotyping of 133 Y-chromosomal markers, the structure of populations under present study was further dissected through PCA using 127 (six markers did not work in multiplex), 32, 25, 15 and 12 Y chromosome binary markers. In the study, despite substantial frequency difference in certain haplogroups among North Indian and East Indian populations, we did not observe stratification in different plots generated through PCA on the basis of 127, 32, 25, 15 and 12 markers (Figure 7). Except few outliers, samples were distributed in different clusters, each representing samples from the two studied populations and indicating a possible admixture event after initial settlement of both population groups in India. Utility of the set of 15 markers in dissecting population structure and arriving at similar conclusions, as with larger number of markers, was evident.

Besides population structure and stratification, haplogroup frequency data based on genotyping through these multiplexes could also define population complexity parameters, such as Genetic differentiation (F_{ST}) and molecular variance. Through a combined data analysis, we attempted to define these parameters for each set of above-mentioned populations which were further categorized as 'No Group (all populations under analysis were grouped together)', 'Demographic Group (populations under analysis were grouped according to their geographic niche)' and 'Ethno-linguistic Group (populations under analysis were

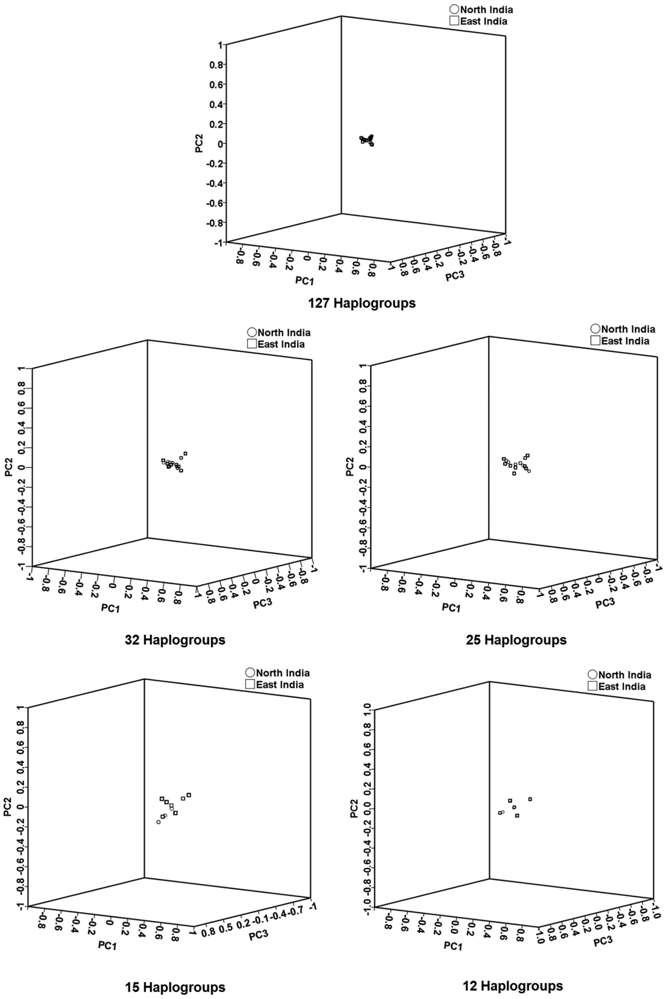

Figure 7. PCA plots reflecting population structure of samples included in the present study from North India (359 samples) and East India (71 samples), using 127, 32, 25, 15 and 12 Y-chromosome SNPs.

grouped according to their ethnicity and languages)'. Comparison of results among different categories of population groups for each set of markers indicated negligible effect (10^{-2}) of such groupings (Table 1a–i). Pair-wise F_{ST} among populations in each set of markers, therefore, is represented only for 'No Group' category using color coded heatmaps (Supplementary Figure S4a–i). Analysis of F_{ST} and molecular variance indicated that additional number of variables over independent ones has negligible effect (1×10^{-3}) on final observation in a constant population size (Table 2).

In conclusion, results based on various computational simulations, PCA, calculation of population differentiation (F_{ST}) and analysis of molecular variance (AMOVA), reflected that resolution of populations' structure and their relationship with other global populations based on 15 highly informative independent variables (markers), represented in the phylogenetic map of Y-chromosomal haplogroups (Figure 8), were optimal, and the set of markers <15 affected the results substantially.

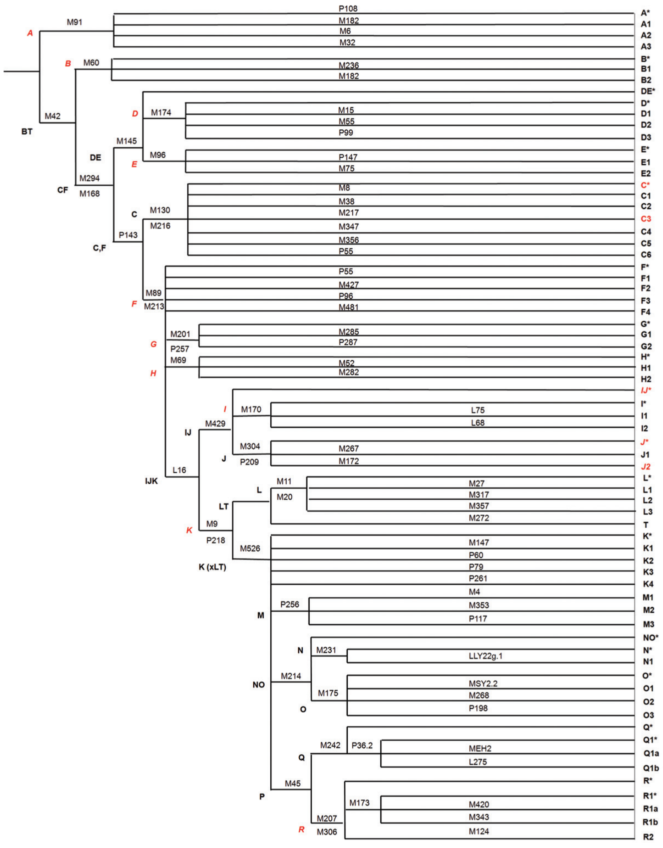

Figure 8. Hierarchical phylogeny of Y-chromosomal haplogroups (up to sub-haplogroup level), depicting positioning of 15 highly informative independent markers (marked in red), sufficient to infer population structure and relationship in a precise and efficient manner. All the 15 markers could be used in a single plex under similar conditions as depicted in 'Materials and Methods' section, without super-plexing with more markers.

DISCUSSION

In evolutionary context, phylogeny of human Y-chromosomal haplogroups plays key roles in deciphering population structure and history. In the background of Wright–Fisher neutral theory of evolution (7) and step wise mutation model (19), the ancestral mutations lay the structural basis for additional variations to generate further sub-haplogroups within major haplogroups. With the basic idea of haplogroup generation, only a few evolutionary markers which are most ancestral in their respective clades could be considered as independent; whereas rest are dependent on their background(s), leaving a scope for hierarchical grading and pruning of redundant variables. Therefore, an appropriate approach for pruning of the ever increasing variables is essential to optimize the number of SNPs which is sufficient to draw the same conclusion about population structure and ancestry as is possible by the inclusion of a larger number of variables.

Considering the right-hand thumb rule of haplogroup generation in human evolution, we attempted to decipher population structure by optimizing evolutionary markers using a novel approach 'RFSHC' which is a combination of variable ranking-based feature selection and agglomerative

Table 2. Genetic differentiation (F_{ST}, F_{IS} and F_{IT} values) and molecular variance in each set (50, 79 and 105) of population with different categories (no group, demographic groups and ethno-linguistic groups)

	32 Variables						25 Variables						15 Variables						12 Variables					
	No Group		Demographic Group		Ethno-linguistic Group		No Group		Demographic Group		Ethno-linguistic Group		No Group		Demographic Group		Ethno-linguistic Group		No Group		Demographic Group		Ethno-linguistic Group	
Source of var	SS	Var comp	SS	Var comp	SS	Var comp	SS	Var comp	SS	Var comp	SS	Var comp	SS	Var comp	SS	Var comp	SS	Var comp	SS	Var comp	SS	Var comp	SS	Var comp
(a) 50 populations based on 32, 25, 15 and 12 haplogroup frequencies data																								
Among groups			75.4	0.02	349.3	0.08			71.6	0.02	340.7	0.08			79.0	0.02	338.1	0.08			107.3	0.04	325.5	0.08
Among populations	453.9	0.10	378.5	0.09	104.6	0.03	442.9	0.09	371.3	0.09	102.2	0.03	444.5	0.09	365.5	0.08	106.4	0.03	400.2	0.09	292.9	0.07	74.7	0.02
Within populations	1546.7	0.33	1546.7	0.33	1546.7	0.33	1523.0	0.33	1523.0	0.33	1523.0	0.33	1453.1	0.31	1453.1	0.31	1453.1	0.31	1249.1	0.27	1249.1	0.27	1249.1	0.27
Total	2000.6	0.43	2000.6	0.44	2000.6	0.44	1965.9	0.42	1965.9	0.43	1965.9	0.43	1897.6	0.41	1897.6	0.42	1897.6	0.42	1649.3	0.36	1649.3	0.37	1649.3	0.37
Fix Ind																								
F_{IS}			0.21		0.07				0.21		0.07				0.21		0.08				0.20		0.06	
F_{ST}	*0.22*		*0.24*		*0.25*		*0.22*		*0.24*		*0.24*		*0.23*		*0.25*		*0.25*		*0.24*		*0.28*		*0.26*	
F_{IT}			0.04		0.19				0.04		0.19				0.05		0.19				0.10		0.21	
(b) 79 populations based on 25, 15 and 12 haplogroup frequencies data																								
Among groups									608.7	0.06	774.4	0.07			568.6	0.05	848.0	0.08			617.2	0.06	847.2	0.09
Among populations							1352.6	0.13	698.6	0.08	578.2	0.07	1269.9	0.12	701.3	0.07	421.9	0.05	1165.8	0.11	548.7	0.06	318.6	0.04
Within populations							3566.3	0.33	3480.6	0.33	3566.3	0.33	3307.7	0.31	3307.7	0.31	3307.7	0.31	2790.8	0.26	2790.8	0.26	2790.8	0.26
Total							4918.9	0.46	4787.8	0.47	4918.9	0.47	4577.6	0.42	4577.6	0.43	4577.6	0.44	3956.6	0.37	3956.6	0.38	3956.6	0.38
Fix Ind																								
F_{IS}									0.19		0.17				0.20		0.13				0.18		0.12	
F_{ST}							0.28		0.29		0.29		0.28		0.29		0.30		0.30		0.32		0.32	
F_{IT}									0.13		0.15				0.12		0.19				0.16		0.22	
(c) 105 populations based on 15 and 12 haplogroup frequencies data																								
Among groups															1060.0	0.09	1355.2	0.11			1177.3	0.10	1410.1	0.12
Among populations													1799.5	0.14	739.5	0.07	444.3	0.04	1753.2	0.14	575.9	0.05	343.0	0.03
Within populations													3635.2	0.29	3635.2	0.29	3635.2	0.29	3077.8	0.24	3077.8	0.24	3077.8	0.24
Total													5434.7	0.43	5434.7	0.44	5434.7	0.44	4831.0	0.38	4831.0	0.39	4831.0	0.40
Fix Ind																								
F_{IS}															0.19		0.13				0.17		0.12	
F_{ST}													0.33		0.35		0.35		0.36		0.39		0.39	
F_{IT}															0.20		0.26				0.26		0.31	

SS, sum of squares; Var comp, variation component.

hierarchical clustering. Current approach relies on the fact that although, evolutionary markers are generated through random mutation events, these events are sequential and not independent of each other. Our results based on Pearson's correlation matrix indicated that correlation among variables decreases by employing the above approach and eventually lead to independent and non-redundant evolutionary markers which could infer world-wide populations' structure and relationship as efficiently and precisely as a set of larger number of evolutionary markers. Though slight changes in the resolution of population structure may occur in recently evolved populations like Europe, South Asia where recently generated markers have refined the structure to some extent, our approach proves suitable for them too in broader perspective. A comparison of population clustering and purity of most optimal clusters based on four sets of markers in three different sets of populations illustrated that proposed set of 15 independent markers provide similar results as is obtained by a set of larger number of variables, such as 25, 32 in combined datasets and as large as 127 variables in the present dataset, which proves far better than the results obtained from a set of markers below 15 (12 markers). Further analysis of present and combined datasets regarding population structure parameters based on PCA, F_{ST} and AMOVA have clearly indicated negligible effect of additional (>15) evolutionary markers used for analysis, whereas a substantial change in these parameters was observed with a set of 12 markers. Some differences observed in population structure resolution at the level of 50 populations were population-specific and justified by further analysis with increased number of populations and other more specific parameters. Interestingly, we observed that results based on a set of 12 markers have little difference with that based on a set of 15 markers in a small number of populations, however, the discrepancy becomes evident with increasing number of populations, approving proposed 15 markers as an optimum set for tracing populations' structure and relationship in world-wide populations.

Further, to deal with the large sample size as required in evolutionary studies, we need highly efficient, precise and cost-effective methods. During last decade, various methods have emerged to provide moderate to high efficiency. However, most of the available methods seem to be limiting in one or other above-mentioned aspects. Though high-throughput genotyping methods carry potential to genotype millions of SNPs at a time, evolutionary studies involve hundreds to thousands of SNPs which need to be genotyped in large sample sizes. This specific requirement provides an extra advantage to moderately efficient techniques over high-throughput techniques for evolutionary and forensic purposes. Here, we adopted the advantage of a moderately efficient MALDI-TOF mass spectrometry-based iPLEX Gold Assay (SEQUENOM, Inc.). In addition, our systematic multiplexing provided a step-wise gradation of major male-related haplogroups and their sub-haplogroups in a continent-specific manner. These multiplexes based on RF-SHC approach add new dimensions to moderate genotyping techniques by offering cost-effectiveness for deep resolution of populations' structure, ancestry and relationship in large scale evolutionary studies.

Although MALDI-TOF-based SEQUENOM platform facilitated this study in a speedy manner. Nevertheless, these highly informative independent evolutionary markers selected through our method could also be genotyped by any other low or medium throughput platform, such as allele specific hybridization, PCR and single base extension methods, RFLP, sequencing based methods and TaqMan assay, etc. As the focus of study was to present an optimized methodology that could infer the population structure and relationship with minimum expense and maximum efficiency, an objective successfully accomplished by the selection of 15 highly informative independent Y-chromosomal markers, the selection of a platform or an approach for genotyping of these markers remains solely the choice of a researcher.

In conclusion, the combined approach of feature selection and hierarchical clustering for pruning of variables in human Y-chromosome provides a highly efficient and cost-effective method to expedite the process of understanding different global populations' structure and their relationship by using the proposed and validated sets of multiplexes. The strength of current approach lies in elegantly designed multiplex of a small set of independent markers leading to low redundancy and high efficiency. At the same time, this method is limited to sequentially evolved markers, and can only be applied to those markers for which hierarchical phylogeny is available. However, this feature of our approach makes it more specific for evolutionary studies.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

We would like to thank Mr Vibhor Kant and Mr Randhir Kumar for extended help in mathematical modeling.

FUNDING

No funding.
Conflict of interest. None declared.

REFERENCES

1. Behar,D.M., Villems,R., Soodyall,H., Blue-Smith,J., Pereira,L., Metspalu,E., Scozzari,R., Makkan,H., Tzur,S., Comas,D. *et al.* (2008) The dawn of human matrilineal diversity. *Am. J. Hum. Genet.*, **82**, 1130–1140.
2. Jobling,M.A. and Tyler-Smith,C. (2003) The human Y chromosome: an evolutionary marker comes of age. *Nat. Rev. Genet.*, **4**, 598–612.
3. Francois,O. and Durand,E. (2010) Spatially explicit Bayesian clustering models in population genetics. *Mol. Ecol. Resour.*, **10**, 773–784.
4. Santafe,G., Lozano,J.A. and Larranaga,P. (2008) Inference of population structure using genetic markers and a Bayesian model averaging approach for clustering. *J. Comput. Biol.*, **15**, 207–220.
5. Corander,J., Sirén,J. and Arjas,E. (2007) Bayesian spatial modeling of genetic population structure. *Comput. Stat.*, **23**, 111–129.
6. Corander,J., Waldmann,P. and Sillanpaa,M.J. (2003) Bayesian analysis of genetic differentiation between populations. *Genetics*, **163**, 367–374.
7. Tran,T.D., Hofrichter,J. and Jost,J. (2013) An introduction to the mathematical structure of the Wright-Fisher model of population genetics. *Theory Biosci.*, **132**, 73–82.

8. Oquendo,M.A., Baca-Garcia,E., Artes-Rodriguez,A., Perez-Cruz,F., Galfalvy,H.C., Blasco-Fontecilla,H., Madigan,D. and Duan,N. (2012) Machine learning and data mining: strategies for hypothesis generation. *Mol. Psychiatry*, **17**, 956–959.
9. Amigo,J., Phillips,C., Salas,A. and Carracedo,A. (2009) Viability of in-house datamarting approaches for population genetics analysis of SNP genotypes. *BMC Bioinformatics*, **10**(Suppl. 3), S5.
10. Wu,Q., Ye,Y., Liu,Y. and Ng,M.K. (2012) SNP selection and classification of genome-wide SNP data using stratified sampling random forests. *IEEE Trans. Nanobiosci.*, **11**, 216–227.
11. Wang,W.B. and Jiang,T. (2008) A new model of multi-marker correlation for genome-wide tag SNP selection. *Genome Inform. Int. Conf. Genome Inform.*, **21**, 27–41.
12. Saeys,Y., Inza,I. and Larranaga,P. (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**, 2507–2517.
13. Hao,K. (2007) Genome-wide selection of tag SNPs using multiple-marker correlation. *Bioinformatics*, **23**, 3178–3184.
14. Grover,D., Woodfield,A.S., Verma,R., Zandi,P.P., Levinson,D.F. and Potash,J.B. (2007) QuickSNP: an automated web server for selection of tagSNPs. *Nucleic Acids Res.*, **35**, W115–W120.
15. Bellman,R. and Kalaba,R. (1959) A mathematical theory of adaptive control processes. *Proc. Natl. Acad. Sci. U.S.A.*, **45**, 1288–1290.
16. Pe'er,I., de Bakker,P.I., Maller,J., Yelensky,R., Altshuler,D. and Daly,M.J. (2006) Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat. Genet.*, **38**, 663–667.
17. Barrett,J.C. and Cardon,L.R. (2006) Evaluating coverage of genome-wide association studies. *Nat. Genet.*, **38**, 659–662.
18. Zhou,N. and Wang,L. (2007) Effective selection of informative SNPs and classification on the HapMap genotype data. *BMC Bioinformatics*, **8**, 484.
19. Kimura,M. and Ohta,T. (1978) Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proc. Natl. Acad. Sci. U.S.A.*, **75**, 2868–2872.
20. Karafet,T.M., Mendez,F.L., Meilerman,M.B., Underhill,P.A., Zegura,S.L. and Hammer,M.F. (2008) New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res.*, **18**, 830–838.
21. Geppert,M. and Roewer,L. (2012) SNaPshot(R) minisequencing analysis of multiple ancestry-informative Y-SNPs using capillary electrophoresis. *Methods Mol. Biol.*, **830**, 127–140.
22. 1000 Genomes Project Consortium, Abecasis,G.R., Altshuler,D., Auton,A., Brooks,L.D., Durbin,R.M., Gibbs,R.A., Hurles,M.E. and McVean,G.A. (2010) A map of human genome variation from population-scale sequencing. Nature, **467**, 1061–1073.
23. Millis,M.P. (2011) Medium-throughput SNP genotyping using mass spectrometry: multiplex SNP genotyping using the iPLEX(R) Gold assay. *Methods Mol. Biol.*, **700**, 61–76.
24. Meyer,K. and Ueland,P.M. (2011) Use of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry for multiplex genotyping. *Adv. Clin. Chem.*, **53**, 1–29.
25. Looi,M.L., Sivalingam,M., Husin,N.D., Radin,F.Z., Isa,R.M., Zakaria,S.Z., Hussin,N.H., Alias,H., Latiff,Z.A., Ibrahim,H. *et al.* (2011) Multiplexed genotyping of beta globin mutations with MALDI-TOF mass spectrometry. *Clin. Chim. Acta*, **412**, 999–1002.
26. Thongnoppakhun,W., Jiemsup,S., Yongkiettrakul,S., Kanjanakorn,C., Limwongse,C., Wilairat,P., Vanasant,A., Rungroj,N. and Yenchitsomanus,P.T. (2009) Simple, efficient, and cost-effective multiplex genotyping with matrix assisted laser desorption/ionization time-of-flight mass spectrometry of hemoglobin beta gene mutations. *J. Mol. Diagn.*, **11**, 334–346.
27. Ragoussis,J., Elvidge,G.P., Kaur,K. and Colella,S. (2006) Matrix-assisted laser desorption/ionisation, time-of-flight mass spectrometry in genomics research. *PLoS Genet.*, **2**, e100.
28. Paracchini,S., Arredi,B., Chalk,R. and Tyler-Smith,C. (2002) Hierarchical high-throughput SNP genotyping of the human Y chromosome using MALDI-TOF mass spectrometry. *Nucleic Acids Res.*, **30**, e27.
29. Bray,M.S., Boerwinkle,E. and Doris,P.A. (2001) High-throughput multiplex SNP genotyping with MALDI-TOF mass spectrometry: practice, problems and promise. *Hum. Mutat.*, **17**, 296–304.
30. Griffin,T.J. and Smith,L.M. (2000) Single-nucleotide polymorphism analysis by MALDI-TOF mass spectrometry. *Trends Biotechnol.*, **18**, 77–84.
31. Li,J., Butler,J.M., Tan,Y., Lin,H., Royer,S., Ohler,L., Shaler,T.A., Hunter,J.M., Pollart,D.J., Monforte,J.A. *et al.* (1999) Single nucleotide polymorphism determination using primer extension and time-of-flight mass spectrometry. *Electrophoresis*, **20**, 1258–1265.
32. Ross,P., Hall,L., Smirnov,I. and Haff,L. (1998) High level multiplex genotyping by MALDI-TOF mass spectrometry. *Nat. Biotechnol.*, **16**, 1347–1351.
33. Haff,L.A. and Smirnov,I.P. (1997) Single-nucleotide polymorphism identification assays using a thermostable DNA polymerase and delayed extraction MALDI-TOF mass spectrometry. *Genome Res.*, **7**, 378–388.
34. Martinez-Cruz,B., Ziegle,J., Sanz,P., Sotelo,G., Anglada,R., Plaza,S. and Comas,D. (2011) Multiplex single-nucleotide polymorphism typing of the human Y chromosome using TaqMan probes. *Invest. Genet.*, **2**, 13.
35. van Oven,M., van den Tempel,N. and Kayser,M. (2012) A multiplex SNP assay for the dissection of human Y-chromosome haplogroup O representing the major paternal lineage in East and Southeast Asia. *J. Hum. Genet.*, **57**, 65–69.
36. van Oven,M., Ralf,A. and Kayser,M. (2011) An efficient multiplex genotyping approach for detecting the major worldwide human Y-chromosome haplogroups. *Int. J. Legal Med.*, **125**, 879–885.
37. Sanchez,J.J., Phillips,C., Borsting,C., Balogh,K., Bogus,M., Fondevila,M., Harrison,C.D., Musgrave-Brown,E., Salas,A., Syndercombe-Court,D. *et al.* (2006) A multiplex assay with 52 single nucleotide polymorphisms for human identification. *Electrophoresis*, **27**, 1713–1724.
38. Onofri,V., Alessandrini,F., Turchi,C., Pesaresi,M., Buscemi,L. and Tagliabracci,A. (2006) Development of multiplex PCRs for evolutionary and forensic applications of 37 human Y chromosome SNPs. *Forensic Sci. Int.*, **157**, 23–35.
39. Brion,M., Sobrino,B., Blanco-Verea,A., Lareu,M.V. and Carracedo,A. (2005) Hierarchical analysis of 30 Y-chromosome SNPs in European populations. *Int. J. Legal Med.*, **119**, 10–15.
40. Brion,M., Sanchez,J.J., Balogh,K., Thacker,C., Blanco-Verea,A., Borsting,C., Stradmann-Bellinghausen,B., Bogus,M., Syndercombe-Court,D., Schneider,P.M. *et al.* (2005) Introduction of an single nucleodite polymorphism-based "Major Y-chromosome haplogroup typing kit" suitable for predicting the geographical origin of male lineages. *Electrophoresis*, **26**, 4411–4420.
41. Brion,M. (2005) Y chromosome SNP analysis using the single-base extension: a hierarchical multiplex design. *Methods Mol. Biol.*, **297**, 229–242.
42. Sobrino,B., Brion,M. and Carracedo,A. (2005) SNPs in forensic genetics: a review on SNP typing methodologies. *Forensic Sci. Int.*, **154**, 181–194.
43. Muro,T., Iida,R., Fujihara,J., Yasuda,T., Watanabe,Y., Imamura,S., Nakamura,H., Kimura-Kataoka,K., Yuasa,I., Toga,T. *et al.* (2011) Simultaneous determination of seven informative Y chromosome SNPs to differentiate East Asian, European, and African populations. *Leg. Med. (Tokyo)*, **13**, 134–141.
44. Berniell-Lee,G., Sandoval,K., Mendizabal,I., Bosch,E. and Comas,D. (2007) SNPlexing the human Y-chromosome: a single-assay system for major haplogroup screening. *Electrophoresis*, **28**, 3201–3206.
45. Joseph Lee Rodgers,W.A.N. (1988) Thirteen ways to look at the correlation coefficient. *Am. Stat.*, **42**, 59–66.
46. Paul Oeth,M.B., Park,Christopher, Kosman,Dominik, del Mistro,Guy, van den Boom,Dirk and Jurinke+,Christian. (2006) iPLEX^{TM} assay: increased plexing efficiency and flexibility for MassARRAY® system through single base primer extension with mass-modified terminators. *SEQUENOM Applic. Note*, **4**, 8876–006
47. Dunn,J.C. (1974) Well separated clusters and optimal fuzzy partitions. *J. Cybernet.*, **4**, 95–104.
48. Rousseeuw,P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.
49. Sharma,Swarkar, Rai,Ekta, Sharma,Prithviraj, Jena,Mamata, Singh,Shweta, Darvishi,Katayoon, Bhat,Audesh K., Bhanwer,A.J.S., Tiwari,Pramod Kumar and Bamezai,Rameshwar N.K. (2009) The Indian origin of paternal haplogroup R1a1* substantiates the autochthonous origin of Brahmins and the caste system. *Journal of human genetics*, **54**, 47–55.

50. Guy Brock,V.P., Dutta,Susmita and Dutta,Somnath (2008) clValid: an R package for cluster validation. *J. Stat. Softw.*, **25**, 1–22.