RESEARCH ARTICLE

Comparative analyses of genetic risk prediction methods reveal extreme diversity of genetic predisposition to nonalcoholic fatty liver disease (NAFLD) among ethnic populations of India

ANKITA CHATTERJEE[1], ANALABHA BASU[1], ABHIJIT CHOWDHURY[2,3], KAUSIK DAS[2], NEETA SARKAR-ROY[1], PARTHA P. MAJUMDER[1] and PRIYADARSHI BASU[1,3]*

[1]*National Institute of Biomedical Genomics, Netaji Subhas Sanatorium (T. B. Hospital), Kalyani 741 251, India*
[2]*Institute of Post Graduate Medical Education and Research, J. C. Bose Road, Kolkata 700 020, India*
[3]*Biomedical Genomics Center, PG Polyclinic Building, Kolkata 700 020, India*

Abstract

Nonalcoholic fatty liver disease (NAFLD) is a distinct pathologic condition characterized by a disease spectrum ranging from simple steatosis to steato-hepatitis, cirrhosis and hepatocellular carcinoma. Prevalence of NAFLD varies in different ethnic groups, ranging from 12% in Chinese to 45% in Hispanics. Among Indian populations, the diversity in prevalence is high, ranging from 9% in rural populations to 32% in urban populations, with geographic differences as well. Here, we wished to find out if this difference is reflected in their genetic makeup. To date, several candidate genes and a few genomewide association studies (GWAS) have been carried out, and many associations between single nucleotide polymorphisms (SNPs) and NAFLD have been observed. In this study, the risk allele frequencies (RAFs) of NAFLD-associated SNPs in 20 Indian ethnic populations (376 individuals) were analysed. We used two different measures for calculating genetic risk scores and compared their performance. The correlation of additive risk scores of NAFLD for three Hapmap populations with their weighted mean prevalence was found to be high ($R^2 = 0.93$). Later we used this method to compare NAFLD risk among ethnic Indian populations. Based on our observation, the Indian caste populations have high risk scores compared to Caucasians, who are often used as surrogate and similar to Indian caste population in disease gene association studies, and is significantly higher than the Indian tribal populations.

[Chatterjee A., Basu A., Chowdhury A., Das K., Sarkar-Roy N., Majumder P. P. and Basu P. 2015 Comparative analyses of genetic risk prediction methods reveal extreme diversity of genetic predisposition to nonalcoholic fatty liver disease (NAFLD) among ethnic populations of India. *J. Genet*. **94**, 105–113]

Introduction

Nonalcoholic fatty liver disease (NAFLD) is the hepatic manifestation of metabolic syndrome occurring when fat is deposited (steatosis) in the liver; but not due to excessive alcohol use, or viral (HBV and HCV) infection. It is characterized by a disease spectrum ranging from steatosis to steato-hepatitis (steatosis in addition to necro-inflammation), cirrhosis and hepatocellular carcinoma (HCC) (Wilfred de Alwis and Day 2008; Cohen *et al.* 2011). NAFLD prevalence is 20–30% in North America and Western countries, where it is related to modern lifestyle with additional complication due to rising incidence of type 2 diabetes mellitus (DM) and obesity (Angulo and Lindor 2002; Williams 2006; Torres and Harrison 2008). There is also substantial ethnic and temporal variability in prevalence. In Dallas Heart Study, hepatic steatosis was found in 45% of Hispanics, 33% of Caucasians and 24% of African Americans (Browning *et al.* 2004; Cohen *et al.* 2011). Epidemiological studies suggest prevalence of NAFLD to be around 17–32% in urban populations of India (table 1), with higher prevalence in those with obesity and diabetes. In rural India the prevalence is 9% (Das *et al.* 2010; Duseja 2010).

NAFLD is a common and genetically complex disorder with an estimated heritability of 39% (Schwimmer *et al.* 2009). Several candidate genes studies and three genome-wide association studies (GWAS) have been performed to identify the associated genes for NAFLD (Cohen *et al.* 2011; Hernaez 2011). There is evidence that while disease associated with single-nucleotide polymorphisms (SNPs) are not significantly more differentiated between populations than random SNPs in the genome on average, risk allele frequencies do show substantial variation across human populations and may account for differences in disease

*For correspondence. E-mail: pb1@nibmg.ac.in.

Keywords. nonalcoholic fatty liver disease; risk model; ethnicity; single-nucleotide polymorphism.

Table 1. Prevalence of NAFLD reported among different populations.

Population	Prevalance (%)	Sample size	References	Weighted mean prevalence (for multiple studies)
Northern India (urban)	32	121	Bajaj *et al.* (2009)	NA
Eastern Indian coastal (semiurban)	24.5	159	Singh *et al.* (2004)	NA
Western India (urban)	17	1168	Amarapurkar *et al.* (2007)	NA
Southern India (urban)	32	541	Mohan *et al.* (2009)	NA
Eastern India (rural)	9	1911	Das *et al.* (2010)	NA
Hispanics	45	2287	Browning *et al.* (2004)	NA
Caucasians	33	2287	Browning *et al.* (2004)	NA
African Americans	24	2287	Browning *et al.* (2004)	NA
Japanese	29	1950	Jimba *et al.* (2005)	NA
	9.3	3432	Omagari *et al.* (2002)	22.9%
	29.7	5075	Eguchi *et al.* (2012)	
Chinese	15	3543	Zhou *et al.* (2007)	14.8%
	17.3	3175	Fan *et al.* (2005)	
	12.9	4009	Shen *et al.* (2003)	

NA, not applicable.

prevalence between human populations (Guthery *et al.* 2007; Myles *et al.* 2008). This is even more important in the Indian context as ethnic diversity in India is immense (Indian Genome Variation Consortium 2008; Reich *et al.* 2009). Linguistically, Indians can be divided into Indo-European, Austro-Asiatic, Tibeto-Burman and Dravidian speakers. Historically, the Hindu caste populations are organized into four social rungs or strata, with Brahmins occupying the highest rung (upper caste) Although the ethnic groups within and across social rungs have been, by and large, genetically isolated due to centuries of endogamy, the nature and extent of admixture across social rungs have been variable (Majumder 1998; Bamshad *et al.* 2001; Roychoudhury *et al.* 2001). The 20 populations selected in this study are representative of the ethnic, linguistic and geographic diversity of India.

In the present study we aimed to find out whether the Indian populations differ in risk allele frequencies (RAFs) at NAFLD-associated candidate SNPs, and also to predict the genetic risk score of NAFLD in different Indian populations, as well as to compare such risk scores with other world populations. To achieve this comparison, we have defined and compared two methods of genetic risk score assessment. We have then computed the risk scores in both individual populations as well as population clusters, and compared them with other world populations.

Materials and methods

Whole genome genotyping was done on samples of 376 individuals on 988205 loci with informed consent from 20 Indian ethnic populations. Genotyping was performed using Illumina SNP array using Infinium Assay (Omni Quad Illumina, San Diego, USA) and scanned in iScan^{TM} System, Illumina (San Diego, USA). This protocol used 200 ng of DNA at concentration of 50 ng/μL as a starting material. Genotype calling was performed using Genome Studio 2011.1 Illumina, San Diego, USA with GenCall score threshold of 0.25. We have excluded DNA that failed to genotype for >90% of the SNPs on the chip. We identified several SNPs from literature survey that showed significant association with NAFLD in different populations. Genotyping data of 34 of these associated SNPs were extracted from the Whole Genome Genotyping data using PLINK software (ver. 1.07; http://pngu.mgh.harvard.edu/~purcell/plink/), in all 376 individuals across 20 Indian populations (table 2) distributed throughout India. Hardy–Weinberg equilibrium (HWE) was checked for each of these SNPs in every population using PLINK. All 34 loci tested were in HWE in all populations. Genotyping Data from four worldwide populations (Caucasian, CEU; Han Chinese, CHB; Japanese, JPT; Yoruba, YRI) were obtained from the Hapmap website (http://www.hapmap.org) and out of the 34 chosen SNPs, genotype information could be extracted for 22 SNPs from the worldwide dataset. The names, linguistic groupings, ethnicity, locations of habitats and occupation of the 20 ethnic groups included in the present study are provided in table 2 and the geographical locations are presented in figure 1.

Risk allele frequencies (RAFs) of the 34 SNPs were calculated using PLINK. Hierarchical cluster analysis approach was used to calculate phylogenetic relationship of 20 Indian populations on the basis of Nei's genetic distance method (Nei 1978), calculated from the risk allele frequencies of 34 loci, using the R software package. To show significant difference in RAF at a particular locus between any two populations, pairwise binomial proportion test was done among three clustered populations (obtained from phylogenetic analysis) across 34 loci. Out of the 22 SNPs, allelic odds ratio (OR) information of 18 SNPs and genotypic OR information of eight SNPs (common to both Indian clusters and other world populations) are available. Also, the RAFs of 18 loci (common for both Indian and world populations) were calculated for all Indian populations for comparison with the four world populations (table 1 in electronic supplementary material at http://www.ias.ac.in/jgenet/).

Table 2. Details of 20 populations across India arranged according to their linguistic grouping.

Code (no. of samples)	Name	State	Linguistic group	Caste/tribe	Occupation
BR2 (18)	West Bengal Brahmin	West Bengal (eastern India)	Indo-European	Upper caste	Traditionally priests, now various occupations
MT (7)	Maratha Brahmin	Maharashtra (western India)	Indo-European	Upper caste	Traditionally priests, now various occupations
GBR (20)	Gujarati Brahmin	Gujarat (western India)	Indo-European	Upper caste	Traditionally priests, now various occupations
K (19)	Khatri	Punjab (northern India)	Indo-European	Middle caste	Traditionally warriors, now various occupations
TH (20)	Tharu	Uttarakhand (northern India)	Indo-European, ancestrally Tibeto-Burman	Tribal	Agricultural labourers
MP (20)	Manipuri brahmin	Manipur (northeast India)	Tibeto-Burman	Upper caste	Traditionally priests, now various occupations
JAM (18)	Jamatia	Tripura (northeast India)	Tibeto-Burman	Tribal	Menial labourers
TRI (19)	Tripuri	Tripura (eastern India)	Tibeto-Burman	Tribal	Agricultural labourers
IR (20)	Iyer Brahmins	Tamil Nadu (southern India)	Dravidian	Upper caste	Traditionally priests, now various occupations
PL (20)	Pallan	Tamil Nadu (southern India)	Dravidian	Lower caste	Agricultural labourers
KA (20)	Kadar	Kerala (southern India)	Dravidian	Tribal	Hunter-gatherers
PY (18)	Paniya	Kerala (southern India)	Dravidian	Tribal	Hunter-gatherers
IL (20)	Irular	Karnataka, Tamil Nadu (southern India)	Dravidian	Tribal	Hunter-gatherers
GD (20)	Gond	Andhra Pradesh, Maharastra, (southern India)	Dravidian and Austro-Asiatic	Tribal	Agricultural labourers
SA (20)	Santal	West Bengal, Jharkhand, Bihar, Orissa (eastern India)	Austro-Asiatic	Tribal	Traditionally hunter-gatherers, now labourers
BIR (20)	Birhor	Bihar (eastern India)	Austro-Asiatic	Tribal	Hunter gatherers
HO (19)	Ho	Bihar (eastern India)	Austro-Asiatic	Tribal	Traditionally hunter-gatherers, now labourers
KO (20)	Korwa	Bihar (eastern India)	Austro-Asiatic	Tribal	Traditionally hunter-gatherers, now labourers
JW (20)	Jarawa	Andaman islands (Bay of Bengal)	Jarawa-Onge	Tribal	Hunter-gatherers
ONG (18)	Onge	Andaman islands (Bay of Bengal)	Jarawa-Onge	Tribal	Hunter-gatherers

The generalized formula of NAFLD risk score prediction for a particular population is described below. Risk score of NAFLD was predicted for each locus using two different formulae (where both allelic OR and genotypic OR information are available).

(i) Allelic risk score at a particular locus for a particular population:
$R1 = [(p^2 \times 2OR) + (2pq \times OR) + q^2]$,
where $R1$ is risk score calculated using allelic OR, p is risk allele frequency at the specific locus, $q = (1 - p)$, OR is allelic OR of the particular SNP in that particular population. Allelic OR is calculated using allele counting model which essentially assumes an additive model, i.e. homozygous risk allele genotype has twice the risk of heterozygous genotype.

(ii) Genotypic risk score at a particular locus:
$R2 = [(p^2 \times OR_1) + (2pq \times OR_2) + q^2]$,
where $R2$ is risk score calculated using genotypic ORs, p is risk allele frequency at the specific locus, $q = (1 - p)$, OR_1 is ratio of odds of disease in an individual homozygous for the risk allele genotype to odds of disease homozygous for the nonrisk allele genotype in that particular population, OR_2 is ratio of odds of disease in an individual heterozygous for the risk allele genotype to odds of disease homozygous for the nonrisk allele genotype in that particular population.

Additive risk score was calculated for each population on the basis of 'n' disease-associated loci (where only allelic OR information is available) using the formula:

$$R_j = \sum_n [(p_i^2 \times 2OR_{ij}) + (2p_i q_i \times OR_{ij}) + q_i^2],$$

where R_j is additive risk score of jth population over 'n' loci, p_i is risk allele frequency at ith locus, $q_i = (1 - p_i)$, OR_{ij} is allelic OR of SNP at ith locus for the jth population, and $i = i$th locus ($1 \leq i \leq n$).

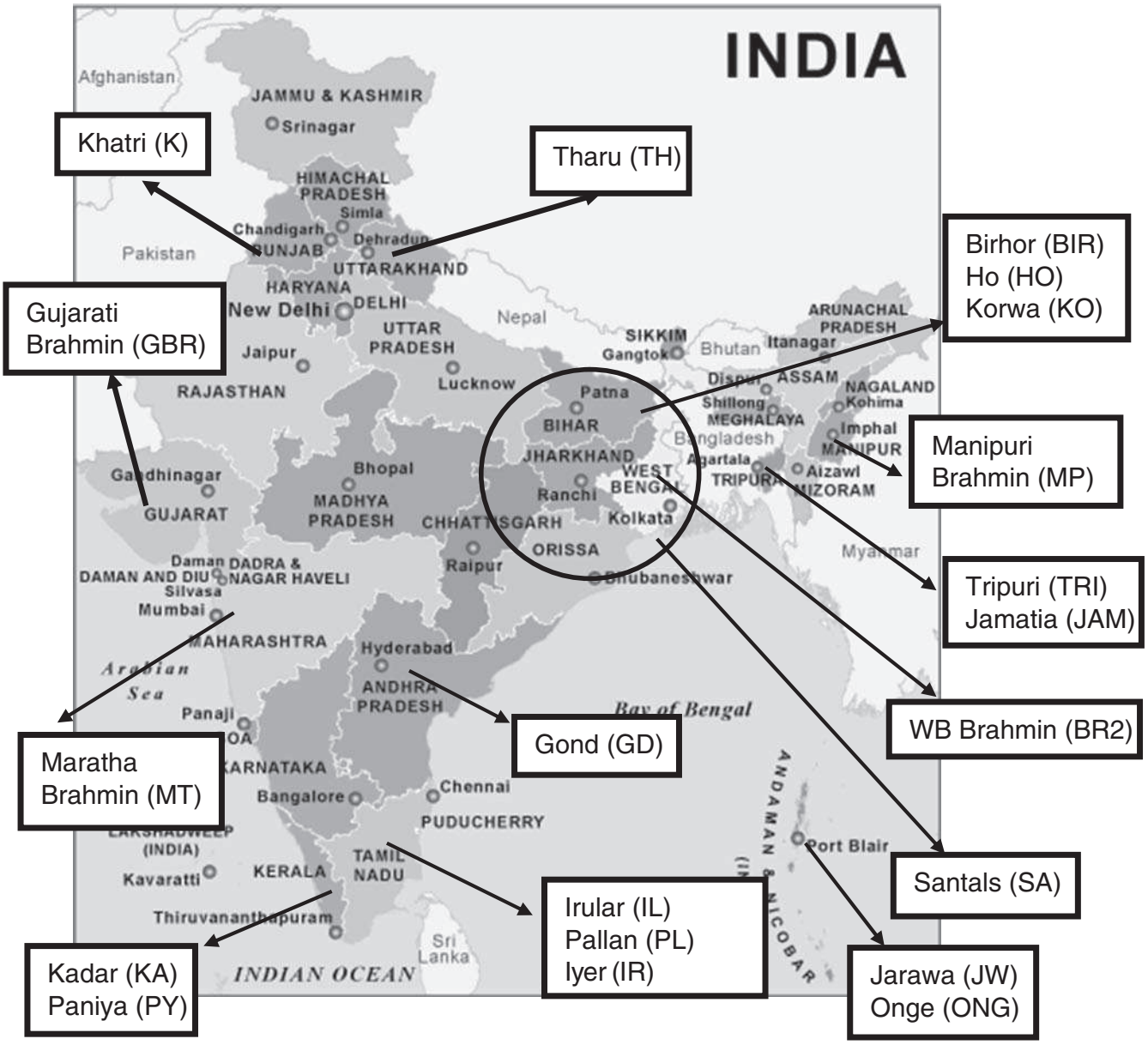

Figure 1. Geographical locations of 20 Indian populations studied.

For risk score prediction methods, our assumptions were as follows: (i) all loci were in HWE; (ii) the contribution of a risk allele in disease prediction is positively correlated with the magnitude of OR value of that SNP at the particular locus, i.e., a SNP with high OR, will contribute more in risk prediction; (iii) that the contribution of risk alleles in predicting disease phenotype follows an additive model, i.e., the homozygous risk allele has twice the risk than heterozygous individuals.

In this study, we calculated both allelic and genotypic risk scores of eight disease-associated loci, where both the allelic and genotypic OR information are available (see table 2 in electronic supplementary material) and found a strong correlation between the results obtained from the two formulae. Since allelic OR information were available for more number of loci ($n = 18$), additive risk score of each of 20 Indian populations and four world populations were furthercalculated using allelic OR information by the additive risk score formula mentioned previously.

The limitation of our analysis is that we did not have OR information for every disease associated SNP in each of our study populations. Thus, apart from the above mentioned assumptions, we have taken an additional assumption for this study; the degree of association of the disease with a particular SNP is same across populations, i.e., we assumed that the OR of a SNP at a particular locus is invariant across populations. However, since additive risk score formula mentioned above is a more generalized one, it can effectively be used for risk

score prediction for any disease in any population, using the information of SNPs associated with that particular disease.

To calculate the significance of the risk score provided by 18 disease-associated loci (where only allelic OR information is available) for each population subclusters, we created an empirical null distribution of risk by randomly choosing 18 nondisease associated SNPs from the whole genome data (∼1 million SNPs). We calculated the score statistic (S_i) of the additive risk score 10,000 times, and tested whether the cluster specific score values fall on the right 5% of the distribution. For constructing the null distribution we assumed that the randomly chosen locus provides no risk to the population, i.e., OR = 1.

To check the significance of differences in risk scores between any two populations, we again created an empirical null distribution of differences in risk scores by randomly choosing 18 random SNPs not associated with the disease, from the whole genome SNP-Chip data (∼1 million SNPs), and repeating the procedure 10,000 times. This provides us with the empirical cutoffs for 1% significance. We assumed that the randomly chosen locus provides no risk to the population, i.e., OR = 1. Each time we recalculated the additive risk score for the two chosen populations and taken the difference in risk scores between them $Di = (S_{1i} - S_{2i})$, where S_{1i} = additive risk under null model for population 1, for the ith draw, S_{2i} = additive risk under null model for population 2, for the ith draw ($1 \leq I \leq 10{,}000$). We created the empirical distribution with these 10,000 risk difference values (D1, D2…D10,000). Risk difference values of any two chosen clusters/populations calculated from the additive risk values of the 18 disease-associated SNPs were considered significant if they fell outside the critical region, i.e., smaller than 0.5 percentile value or larger than 99.5 percentile value.

Principal component analysis (PCA) was done using the R software package to analyse the extent of genetic relatedness among populations using Nei's distance matrix over 18 loci. Next, we calculated the correlation between the PC1 values and the risk scores.

Results

Indian population clustering and heterogeneity in RAF of NAFLD-associated loci across populations

To observe the genetic difference between 20 different Indian populations, pairwise genetic distance were calculated on the basis of 34 disease-associated SNPs and a dendrogram was drawn (figure 2). Based on phylogenetic analyses, the Indian ethnic groups form three major clusters. The clusters correlate well with their geographical proximity and linguistic grouping (table 2), but not with social hierarchy (tribal and caste). To show heterogeneity in RAF, pairwise binomial proportion test was done among the three population clusters across 34 loci (table 3). Twenty-nine out of 34 loci showed significant differences in RAF across three clusters (figure 1 in electronic supplementary material).

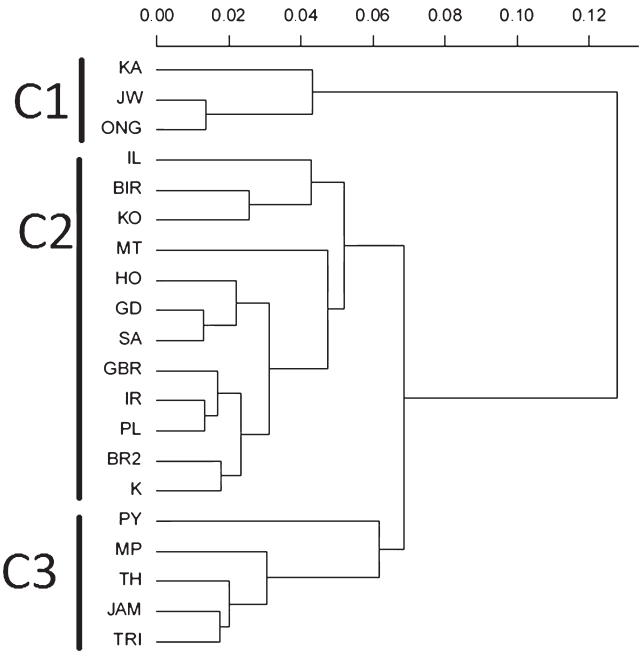

Figure 2. Phylogenetic tree constructed on the basis of pairwise Nei's genetic distance on 34 NAFLD-associated SNPs between 20 Indian populations. Cluster 1 (C1) comprises of three populations: Kadar, KA; Jarawa, JW; Onge, ONG. Cluster 2 (C2) comprises of 12 populations: Irular, IL; Birhor, BIR; Korwa, KO; Maratha Brahmin, MT; Ho, HO; Gond, GD; Santal, SA; Gujarati Brahmin, GBR; Iyer, IR; Pallan, PL; West Bengal Brahmin, BR2; Khatri, K. Cluster 3 (C3) comprises of five populations: Paniya, PY; Manipuri Brahmin, MP; Tharu, TH; Jamatia, JAM; Tripuri, TRI.

Relationship between two algorithms for calculating population risk score of NAFLD, correlation of predicted risk score with known prevalence of NAFLD

There is high correlation ($R^2 = 0.78$) between the risk values obtained from two different formulae (figure 2a in electronic supplementary material). The correlation coefficients within individual clusters C1, C2 and C3 also show high correlation of 0.83, 0.8 and 0.72, respectively (figure 2b in electronic supplementary material). We have thus ascertained that both formulae are equally good estimates for risk score prediction. For further analyses of additive risk scores, we used the allelic OR information, since this information was available for 18 SNPs, for both Indian as well as world populations.

To check whether our risk score prediction model correctly predicts the chance of disease occurrence in a population, we took three world populations: CEU, JPT and CHB (table 1) and calculated the correlation between their weighted mean prevalence (33%, 24.3% and 15.1%, respectively) and their respective additive risk scores (37.35, 36.01 and 35.85, respectively). The correlation was found to be very high ($R^2 = 0.93$). Hence, we concluded that our model correctly predicts the chance of occurrence of the disease in a particular population.

Table 3. Details of 34 NAFLD-associated SNPs. Prediction of RISK SCORE was done with 18 SNPs whose allelic OR information was available.

SNP	Gene	Full name of gene	RA (dbSNP alleles)	Odds ratio
rs1044498	*ENPP1*	Lysophospholipase-like 1	C(C/A)	1.55
rs12137855	*LYPLAL1*	Ectonucleotide pyrophosphatase/phosphodiesterase 1	C(C/T)	1.37
rs1227756	*COL13A1*	Collagen, type XIII, alpha 1	G(A/G)	UA
rs1501299	*ADIPOQ*	Adiponectin, C1Q and collagen domain containing	G(G/T)	UA
rs16944	*IL1B*	Interleukin 1 beta	T(A/C)	UA
rs17222723	*CFTR*	Cystic fibrosis transmembrane conductance regulator (atp-binding cassette subfamily C, member 7)	T(T/A)	UA
rs1799724	*TNF-α*	Tumour necrosis factor alfa	A(A/T)	UA
rs1799883	*FABP2*	Fatty acid-binding protein 2	A(C/T)	UA
rs1799945	*HFE*	Hemochromatosis	G(A/C/G/T)	1.03
rs1799964	*TNF-α*	Tumour necrosis factor alfa	C(C/G)	UA
rs1800588	*LIPC*	Lipase, hepatic	C(C/T)	1.28
rs1800630	*TNF-α*	Tumour necrosis factor alfa	A(C/T)	UA
rs1801131	*MTHFR*	Methylenetetrahydrofolate reductase (NAD(P)H)	C(A/C)	2.7
rs1801278	*IRS1*	Insulin receptor substrate 1	A(A/C)	1.57
rs1801282	*PPARG*	Peroxisome proliferator-activated receptor gamma	G(A/C/G/T)	1.43
rs2031920	*CYP2E1*	Cytochrome P450, family 2, subfamily E, polypeptide 1	T(A/G)	1.75
rs2287622	*ABCB11*	Atp-binding cassette, subfamily B (MDR/TAP), member 11	C(C/T)	1.32
rs2569190	*CD14*	CD14 molecule	T(A/C/G/T)	UA
rs2645424	*FDFT1*	Farnesyl-diphosphate farnesyltransferase 1	A(A/G)	UA
rs343064	*PDGFA*	Platelet-derived growth factor alpha polypeptide	A(C/T)	UA
rs3750861	*KLF6*	Kruppel-like factor 6	G(C/T)	2.88
rs3816873	*MTTP*	Microsomal triglyceride transfer protein	T(C/T)	1.458
rs3856806	*PPARG*	Peroxisome proliferator-activated receptor gamma	T(C/T)	2.28
rs4880	*SOD2*	Superoxide dismutase 2, mitochondrial	T(C/T)	UA
rs4994	*ADRB3*	Adrenergic, beta-3-, receptor	A(C/T)	UA
rs6503695	*STAT3*	Signal transducer and activator of transcription 3 (acute-phase response factor)	T(C/T)	2.32
rs6666089	*ADIPOR1*	Adiponectin receptor 1	A(C/T)	UA
rs738409	*PNPLA3*	Patatin-like phospholipase domain containing 3	G(A/G)	3.26
rs780094	*GCKR*	Glucokinase (hexokinase 4) regulator	T(A/G)	1.16
rs7903146	*TCF7L2*	Transcription factor 7-like 2 (t-cell specific, hmg-box)	T(A/G)	2.2
rs7946	*PEMT*	Phosphatidylethanolamine n-methyltransferase	T(C/T)	2.28
rs8187710	*CFTR*	Cystic fibrosis transmembrane conductance regulator (atp-binding cassette subfamily C, member 7)	G(C/T)	1.99
rs887304	*EFCAB4B*	Ef-hand calcium-binding domain 4B	T(A/G)	UA
rs909253	*TNF-α*	Tumour necrosis factor alfa	A(A/G)	UA

UA, unavailable.

Comparative risk of NAFLD in Indian populations, comparison with world populations

We calculated the additive risk scores of each of 20 Indian populations and four world populations over 18 loci (figure 3a). RAF data for individual populations for each locus is presented in table 1 in electronic supplementary material. The risk scores vary across populations, with Khatri, an Indo-European speaking caste population, having the highest risk score of 37.91 and Ho, an Austro-Asiatic speaking tribal population having the lowest risk score of 34.46. Among the world populations, CEU have the highest risk score of 37.35 and the YRI have the lowest risk score of 34.57. In general, Indo-European speaking caste populations like Maratha Brahmin, West Bengal Brahmin, Gujrati Brahmin, Khatri as well as the Manipuri Brahmin (Tibeto-Burman speaking upper caste population) have a high genetic risk score of NAFLD, compared to the tribal populations. The exception is the Indo-European speaking Tharu tribe of Mongoloid ancestry, which also has a high risk.

Generally, since the caste populations were found to have higher risk scores than the tribes, we subdivided each cluster into respective caste and tribal subclusters (C1, C2/C, C2/T, C3/C, C3/C), and looked at the average additive risk among subclusters. C2 and C3 caste populations (C2/C and C3/C) show higher risk scores compared to any of the tribal clusters (C1, C2/T or C3/T) (figure 3b). The Indo-European speaking (C2/C comprising of West Bengal Brahmin, Maratha Brahmin, Gujrati Brahmin and Khatri) and Tibeto-Burman speaking (C3/C comprising Manipuri Brahmin) caste populations show an additive risk scores similar to CEU and higher than that in Chinese, Japanese and African populations. The tribal populations in C1 and C2 show an additive risk scores between African and Mongoloid (Chinese

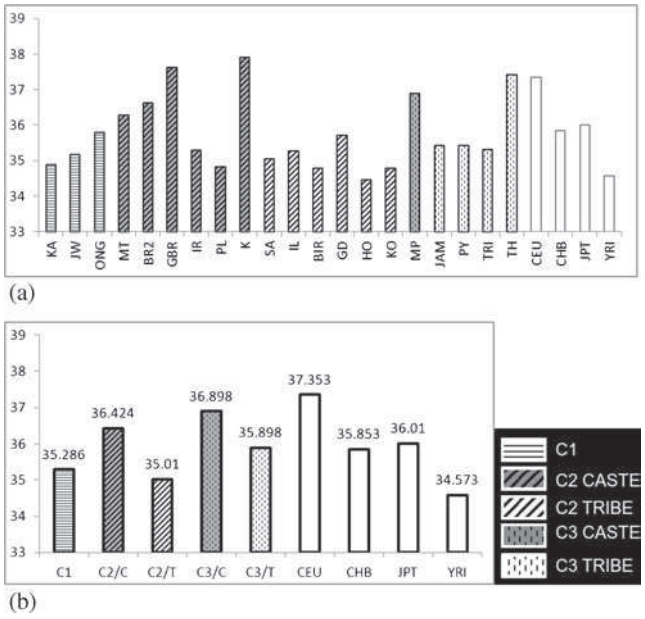

Figure 3. (a) Additive risk scores of 20 Indian populations and four world populations over 18 loci. The caste and tribal populations are marked. (b) Additive risk scores of the caste and tribal populations of India and four world populations. The caste populations generally show higher risk scores than the tribal populations. The Indo-European speaking (West Bengal Brahmin, Maratha Brahmin, Gujrati Brahmins and Kahtri) and Tibeto-Burman speaking (Manipuri Brahmin) caste populations show additive risk scores similar to Caucasians (CEU), and higher than that in Chinese, Japanese and African populations. The tribal populations in C1 and C2 show additive risk scores between African and Mongoloid (Chinese and Japanese) populations.

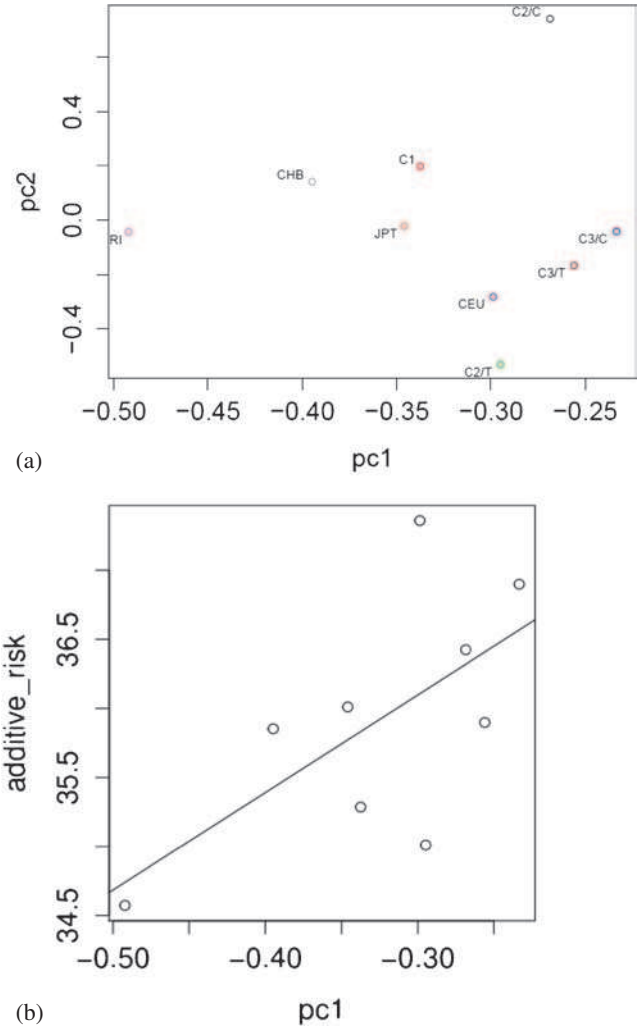

Figure 4. (a) PCA over 18 loci shows genetic relatedness among populations (C1, C2/C, C2/T, C3/C, C3/C, CEU, CHB, JPT and YRI). The graph shows that the Indian population clusters are genetically closer to the Caucasians than the African populations. (b) Correlation between the additive risk scores for three Indian clusters and four world populations and their corresponding values on PC1.

and Japanese) populations. It is interesting to note that the tribal populations in C3, which are of Mongoloid ethnicity, show similar risk scores with Japanese and Chinese populations, which are also of Mongoloid ethnicity.

We calculated the significance of the risk score between population clusters (table 4). The caste populations showed significantly higher risk scores than the tribal populations. There is no significant difference in risk scores between Indo-European and Tibeto-Burman caste populations but in case of the tribal populations of C2 and C3, a significant difference was observed. C1 comprising of only tribal

Table 4. Result of significance test for difference in risk scores between any two studied Indian population subclusters.

Population 1 vs population 2	0.5% cut-off	99.5% cut-off	Observed risk difference	Significance
C1 vs C2/C**	−0.8263	0.5922	−1.138	Significant
C1 vs C2/T	−0.7525	0.5796	0.276	NS
C1 vs C3/C**	−0.8387	0.6469	−1.611	Significant
C1 vs C3/T	−0.7758	0.6189	−0.612	NS
C2/C vs C2/T**	−0.4648	0.5200	1.414	Significant
C2/C vs C3/C	−0.5420	0.6094	−0.473	NS
C2/C vs C3/T	−0.5386	0.608	0.526	NS
C2/T vs C3/C**	−0.5601	0.5712	−1.887	Significant
C2/T vs C3/T**	−0.4784	0.4692	−0.887	Significant
C3/C vs C3/T**	−0.5069	0.4686	1.00	Significant

**Comparisons where significant difference was observed; NS, not significant.

populations has significantly lower risk scores from the caste populations.

Subsequently, we used PCA approach to analyse the extent of genetic relatedness among populations (C1, C2/C, C2/T, C3/C, C3/C, CEU, CHB, JPT and YRI), using Nei's genetic distance over 18 loci. The first principal component PC1 explains 60% of the total variation (figure 4a). On plotting the additive risk values against the PC1 values, the correlation was found to be 0.637 (figure 4b). Hence we concluded that the risk scores correlated well with the genetic relatedness among populations, i.e., genetically close populations have similar chance of occurrence of the disease.

Discussion

Differences in genetic risk for complex diseases across human populations may have arisen at least partially due to population migration and subsequent adaptation to the new habitat (Young *et al*. 2005; Myles *et al*. 2008; Pemberton *et al*. 2009; Hancock *et al*. 2011). In a recent study, Corona *et al*. (2013) analysed SNPs known to modify disease susceptibility in the context of population migration, and reported that distribution of the variants associated with complex diseases cannot be explained by random variation. However, these studies do not have information regarding risk distribution of disease in ethnic groups of India. In addition, there is no data on population-based risk of NAFLD in the literature, which is rapidly becoming a health burden in western and developing countries. In this study we defined a model of disease risk score prediction for different populations that can predict the chances of occurrence of NAFLD in the studied populations. The high correlation between the weighted mean prevalence of the three world populations and their respective risk scores supports the fact that our model predicts correctly the chance of occurrence of the disease. We have also observed that populations genetically close to the disease loci might have similar chance of disease predisposition

Most anthropologists agree that the tribal populations are the indigenous populations of India, among whom the Austro-Asiatic speaking groups, who are exclusively tribal, may be the oldest and may have entered India about 60,000–70,000 year before present (ybp); the Tibeto-Burman speaking tribals of Mongoloid origin were later migrants to India (around 4500 to 11000 years before present (ybp). This is supported by genetic evidence as well. The formation of the caste system took place within the last 3500–4000 years; most caste groups have evolved from the existing population groups and admixture with later migrants. The tribal and caste groups in India are primarily intramarrying; the extent of admixture across ethnic barriers is negligible, but variable (Majumder 1998; Bamshad *et al*. 2001; Roychoudhury *et al*. 2001; Basu *et al*. 2003; Reich *et al*. 2009).

Although the people of India are referred to as 'Indians' in many studies on population genetics, implying genetic homogeneity, in reality, Indian populations form a continuum of genetic spectrum between CEU and JPT/CHB HapMap populations (Indian Genome Variation Consortium 2008; Reich *et al.* 2009). India has a huge genetic diversity shaped by ethnicity, geography, linguistics and social hierarchy further complicated by migration and admixture. In this study we also found significant differences in RAF across Indian population clusters. The Indian tribal populations show significantly lower predicted risk score of NAFLD than caste populations. The caste populations show an additive risk similar to CEU. Since CEU have the highest prevalence of NAFLD among the four world populations, we may say that Indian caste populations will have a high genetic predisposition to developing NAFLD.

Acknowledgements

The authors would like to thank all individuals throughout India who have participated in this study. A Chatterjee is supported by a fellowship from the Council of Scientific and Industrial Research fellowship. We would also like to thank Dr. Ankur Mukherjee for guidance and suggestions. The work was sponsored by an NIBMG internal grant.

References

Amarapurkar D., Kamani P., Patel N., Gupte P., Kumar P., Agal S. *et al.* 2007 Prevalence of nonalcoholic fatty liver disease: population based study. *Ann. Hepatol.* **6**, 161–163.

Angulo P. and Lindor K. D. 2002 Non-alcoholic fatty liver disease. *J. Gastroenterol. Hepatol.* **17**, 186–190.

Bajaj S., Nigam P., Luthra A., Pandey R. M., Kondal D., Bhatt S. P. *et al.* 2009 A case-control study on insulin resistance, metabolic co-variates and prediction score in non-alcoholic fatty liver disease. *Indian J. Med. Res.* **129**, 285–292.

Bamshad M., Kivisild T., Watkins W. S., Dixon M. E., Ricker C. E., Rao B. B. *et al.* 2001 Genetic evidence on the origins of Indian caste populations. *Genome Res.* **11**, 994–1004.

Basu A., Mukherjee N., Roy S., Sengupta S., Banerjee S., Chakraborty M. *et al.* 2003 Ethnic India: a genomic view, with special reference to peopling and structure. *Genome Res.* **13**, 2277–2290.

Browning J. D., Szczepaniak L. S., Dobbins R., Nuremberg P., Horton J. D., Cohen J. C. *et al.* 2004 Prevalence of hepatic steatosis in an urban population in the United States: impact of ethnicity. *Hepatology* **40**, 1387–1395.

Cohen J. C., Horton J. D. and Hobbs H. H. 2011 Human fatty liver disease: old questions and new insights. *Science* **332**, 1519–1523.

Corona E., Chen R., Sikora M., Morgan A. A., Patel C. J., Ramesh A. *et al.* 2013 Analysis of the genetic basis of disease in the context of worldwide human relationships and migration. *PLoS Genet.* **9**, e1003447.

Das K., Das K., Mukherjee P. S., Ghosh A., Ghosh S., Mridha A. R. *et al.* 2010 Non-obese population in a developing country has a high prevalence of nonalcoholic fatty liver and significant liver disease. *Hepatology* **51**, 1593–1602.

Duseja A. 2010 Nonalcoholic fatty liver disease in India – a lot done, yet more required! *Indian J. Gastroenterol.* **29**, 217–225.

Eguchi Y., Hyogo H., Ono M., Mizuta T, Ono N., Fujimoto K. *et al.* 2012 Prevalence and associated metabolic factors of non-alcoholic fatty liver disease in the general population from 2009

to 2010 in Japan: a multicenter large retrospective study. *J. Gastroenterol.* **47**, 586–595.

Fan J. G., Zhu J., Li X. J., Chen L., Lu Y. S., Li L. *et al.* 2005 Fatty liver and the metabolic syndrome among Shanghai adults. *J. Gastroenterol. Hepatol.* **20**, 1825–1832.

Guthery S. L., Salisbury B. A., Pungliya M. S., Stephens J. C. and Bamshad M. 2007 The structure of common genetic variation in United States populations. *Am. J. Hum. Genet.* **81**, 1221–1231.

Hancock A. M., Witonsky D. B., Alkorta-Aranburu G., Beall C. M., Gebremedhin A., Sukernik R. *et al.* 2011 Adaptations to climate-mediated selective pressures in humans. *PLoS Genet.* **7**, e1001375.

Hernaez R. 2011 Genetic factors associated with the presence and progression of nonalcoholic fatty liver disease: A narrative review. *Gastroenterol. Hepatol.* **35**, 32–41.

Indian Genome Variation Consortium 2008 Genetic landscape of the people of India: a canvas for disease gene exploration. *J. Genet.* **87**, 3–20.

Jimba S., Nakagami T., Takahashi M., Wakamatsu T., Hirota Y., Iwamoto Y. and Wasada T. 2005 Prevalence of non-alcoholic fatty liver disease and its association with impaired glucose metabolism in Japanese adults. *Diabet. Med.* **22**, 1141–1145.

Majumder P. P. 1998 People of India: Biological diversity and affinities. *Evol. Anthropol.* **6**, 100–110.

Mohan V., Farooq S., Deepa M., Ravikumar R. and Pitchumoni C.S. 2009 Prevalence of non-alcoholic fatty liver disease in urban south Indians in relation to different grades of glucose intolerance and metabolic syndrome. *Diabetes Res. Clin. Pract.* **84**, 84–91.

Myles S., Davison D., Barrett J., Stoneking M. and Timpson N. 2008 Worldwide population differentiation at disease-associated SNPs. *BMC. Med. Genomics* **1**, 22.

Nei M. 1978 The theory of genetic distance and evolution of human races. *Jap. J. Human Genet.* **23**, 341–369.

Omagari K., Kadokawa Y. and Masuda J. 2002 Fatty liver in non-alcoholic non-overweight Japanese adults: incidence and clinical characteristics. *J. Gastroenterol. Hepatol.* **17**, 1098–1105.

Pemberton T. J., Sandefur C. I, Jakobsson M. and Rosenberg N. A. 2009 Sequence determinants of human microsatellite variability. *BMC Genomics* **10**, 612.

Reich D., Thangaraj K., Patterson N., Price A. L. and Singh L. 2009 Reconstructing Indian population history. *Nature* **461**, 489–494.

Roychoudhury S., Roy S., Basu A., Banerjee R., Vishwanathan H., Usha Rani M. V. *et al.* 2001 Genomic structures and population histories of linguistically distinct tribal groups of India. *Hum. Genet.* **109**, 339–350.

Schwimmer J. B., Celedon M. A., Lavine J. E., Salem R., Campbell N., Schork N. J. *et al.* 2009 Heritability of nonalcoholic fatty liver disease. *Gastroenterology* **136**, 1585–1592.

Shen L., Fan J. G., Shao Y., Zeng M. D., Wang J. R., Luo G. H. *et al.* 2003 Prevalence of nonalcoholic fatty liver among administrative officers in Shanghai: an epidemiological survey. *World J. Gastroenterol.* **9**, 1106–1110.

Singh S. P., Nayak S., Swain M., Rout N., Mallik R. N., Agrawal O. *et al.* 2004 Prevalence of nonalcoholic fatty liver disease in coastal eastern India: a preliminary ultrasonographic survey. *Indian J. Gastroenterol.* **25**, 76–79.

Torres D. M. and Harrison S. A 2008 Diagnosis and therapy of nonalcoholic steatohepatitis. *Gastroenterology* **134**, 1682–1698.

Wilfred de Alwis N. M. and Day C. P. 2008 Genes and nonalcoholic fatty liver disease. *Curr. Diab. Rep.* **8**, 156–163.

Williams R. 2006 Global challenges in liver disease. *Hepatology* **44**, 521–526.

Young J. H., Chang Y. P., Kim J. D., Chretien J. P., Klag M. J., Levine M. A. *et al.* 2005 Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion. *PLoS Genet.* **1**, e82.

Zhou Y.-J., Li Y.-Y., Nie Y.-Q., Ma J. X., Lu L. G., Shi S. L. *et al.* 2007 Prevalence of fatty liver disease and its risk factors in the population of South China. *World J. Gastroenterol.* **13**, 6419–6424.

Received 12 March 2014, in revised form 17 October 2014; accepted 3 November 2014
Unedited version published online: 11 November 2014
Final version published online: 12 March 2015