Annals of human genetics

doi: 10.1111/ahg.12028

Variation at Diabetes- and Obesity-Associated Loci May Mirror Neutral Patterns of Human Population Diversity and Diabetes Prevalence in India

Srilakshmi M. Raj[1,6*], Pradeep Halebeedu[2], Jayarama S Kadandale[3], Marta Mirazon Lahr[4], Irene Gallego Romero[5], Jamuna R. Yadhav[3], Mircea Iliescu[6], Niraj Rai[7], Federica Crivellaro[4], Gyaneshwer Chaubey[8], Richard Villems[8], Kumarasamy Thangaraj[7*], Kalappagowda Muniyappa[9], H. Sharat Chandra[3] and Toomas Kivisild[6,8*]

[1]*Department of Molecular Biology and Genetics, 101 Biotechnology Building, Cornell University, Ithaca, NY 14853, USA*
[2]*Department of Studies in Microbiology, University of Mysore, Manasagangotri, Mysore 570006, Karnataka, India*
[3]*Centre for Human Genetics, 1st Phase, Electronic City, Bangalore 560100, Karnataka, India*
[4]*Leverhulme Centre for Human Evolutionary Studies, Henry Wellcome Building, Fitzwilliam Street, Cambridge CB2 1QH, UK*
[5]*Department of Human Genetics, Cummings Life Science Center 920 E. 58th Street, University of Chicago, Chicago, IL 60637, USA*
[6]*Division of Biological Anthropology, Henry Wellcome Building, Fitzwilliam Street, Cambridge CB2 1QH, UK*
[7]*CSIR-Centre for Cellular and Molecular Biology, Uppal Road, Hyderabad 500007, Andhra Pradesh, India*
[8]*Department of Evolutionary Biology, University of Tartu, Estonian Biocentre, Riia Str. 23, 51010, Tartu, Estonia*
[9]*Department of Biochemistry, Indian Institute of Science, Bangalore 560012, Karnataka, India*

Summary

South Asian populations harbor a high degree of genetic diversity, due in part to demographic history. Two studies on genome-wide variation in Indian populations have shown that most Indian populations show varying degrees of admixture between ancestral north Indian and ancestral south Indian components. As a result of this structure, genetic variation in India appears to follow a geographic cline. Similarly, Indian populations seem to show detectable differences in diabetes and obesity prevalence between different geographic regions of the country. We tested the hypothesis that genetic variation at diabetes- and obesity-associated loci may be potentially related to different genetic ancestries. We genotyped 2977 individuals from 61 populations across India for 18 SNPs in genes implicated in T2D and obesity. We examined patterns of variation in allele frequency across different geographical gradients and considered state of origin and language affiliation. Our results show that most of the 18 SNPs show no significant correlation with latitude, the geographic cline reported in previous studies, or by language family. Exceptions include *KCNQ1* with latitude and *THADA* and *JAK1* with language, which suggests that genetic variation at previously ascertained diabetes-associated loci may only partly mirror geographic patterns of genome-wide diversity in Indian populations.

Keywords: Human genetic variation, India, type 2 diabetes, obesity, population genetics

Introduction

Disentangling the contribution of environment and genetics to complex disease risk requires large amounts of genetic data on large numbers of individuals, the usage of appropriate statistical models, and information on environment and phenotype. Aspects of these issues have proven to be a challenge especially for non-European populations (Need & Goldstein, 2009; Bustamante et al., 2011). Yet, often, these populations exhibit different etiologies and greater risk of certain complex diseases (Kumar et al., 2010; Gravel et al., 2011). Indians

*Corresponding authors: Srilakshmi M Raj, 101 Biotechnology Building, Cornell University, Ithaca, NY 14853. Tel: +1 607 255 2556; Fax: +1 607 255 6249; E-mail: smr46@cornell.edu. Toomas Kivisild, Leverhulme Centre for Human Evolutionary Studies, University of Cambridge, The Henry Wellcome Building, Fitzwilliam Street, Cambridge CB2 1QH, UK. Tel: +44 (0)1223 764703; Fax: +44 (0) 1223 764710; E-mail: tk331@cam.ac.uk. Kumarasamy Thangaraj, CSIR-Centre for Cellular and Molecular Biology, Hyderabad 500 007, India. Tel: +91 40 27192828; Fax: +91 40 27160591; E-mail: thangs@ccmb.res.in

in general do not tend to develop high BMI compared to other global populations, yet have a high risk of type 2 diabetes (T2D) and have among the highest number of cases in the world, totaling over 51 million (McKeigue, 1989; McKeigue et al., 1991; International Diabetes Federation, 2009; Diamond, 2011; Finucane et al., 2011). This trend may be due in part to higher visceral fat deposition in Indians, suggesting an underlying biological basis for high T2D risk in Indians (McKeigue et al., 1991).

Why Indian populations exhibit such high risk of T2D remains an open question, however. Some studies suggest that Indians have a "thrifty phenotype," which indicates that risk is predominantly due to environmental factors such as low birth weight and maternal nutrition status (Hales & Barker, 1992; Yajnik, 2000, 2004). Others have suggested a "thrifty genotype" in which evolutionary adaptations to harsh environmental conditions molded a genetic predisposition to energy thrift, which has become maladaptive in the presence of caloric abundance (Neel, 1962, 1999). Studies that have tested the thrifty genotype hypothesis have thus far not yielded candidate genes that appear to be thrifty in the context of T2D, with the possible exception of *PPARGC1A,* a gene that is associated with BMI in Tongans and may be under positive selection in that population (Paradies et al., 2007; Southam et al., 2009; Myles et al., 2011).

A critical step toward understanding the genetic basis of disease etiology is the understanding of local versus global patterns of genetic diversity. Yet, only a handful of attempts to study Indian genetic variation on a genome-wide scale have been published so far (Indian Genome Variation Consortium, 2008; Reich et al., 2009; Metspalu et al., 2011). One of the first genome-wide studies on Indian populations included 132 individuals from 25 populations across India (Reich et al., 2009). The study demonstrated that most Indian populations are derived from a mixture of two major groups, ancestral north Indians (ANI) and ancestral south Indians (ASI), with proportions of the ANI component varying from 39% to 71%. The pattern of two major ancestry components has been confirmed in a separate study including 142 samples from 30 Indian populations (Metspalu et al., 2011). Long-term genetic isolation among populations, possibly amplified by the social structuring of the caste system, may have heightened the effects of genetic drift, contributing to the high degree of population structure observed. This substructure implies that Indians may have an excess of certain recessive genetic disorders compared with other populations (Reich et al., 2009).

In addition to possible consequences for disease predisposition, genetic diversity across India may follow a geographic cline. Thus far, evidence of a latitudinal cline in India has been mixed. One study on candidate disease-associated SNPs showed that genetic variation does not appear to vary along a latitudinal cline within India (Pemberton et al., 2008). A genome-wide study of genetic variation in India, however, reported a geographic (northwest to southeast) gradient of relatedness extending from Europe to India, which they call "the India cline," perhaps reflecting a gradient in ANI-ASI admixture proportions (Reich et al., 2009). Supporting the hypothesis of a genetic basis for T2D susceptibility in India is the appearance of a north-south gradient in diabetes prevalence, mirroring the genetic variation-based India cline. Cities in the state of Kerala in south India have up to threefold higher T2D prevalence than the northern-most state, Kashmir (Ramachandran et al., 2001; Deepa et al., 2003; Mohan et al., 2006; Fig. S1). The distribution of BMI in India shows a different trend, with higher BMI values among both men and women in north and south Indian states, but lower BMI values in central regions of India (Fig. S1).

These two clines reflected in genetic and T2D prevalence data may indicate a relationship between diabetes susceptibility and genetic variation in India. We studied the distribution of genetic variants associated with susceptibility to T2D and obesity in Europeans, in Indian populations sampled at a fine geographical scale. This was conducted with the aim of assessing whether obesity and T2D risk alleles follow geographic patterns within India consistent with known distributions of disease prevalence and genetic ancestry. Compared to many previous studies, which have focused on specific populations, this study uses over 3200 individuals from 61 different populations sampled across India and its north-south cline. Our sample includes populations from diverse ethnic, linguistic, geographic, and cultural backgrounds.

Materials and Methods

Populations Selected for Genotyping within Karnataka and other States of India

At the national level, 1530 individuals belonging to 38 populations outside Karnataka state were genotyped for SNPs associated with T2D and obesity (Table 1; Table S1). Besides the cross-national level, we focused on genetic variation within the single state of Karnataka in India to minimize cultural, geographic, and linguistic differences among populations. We collected over 1500 saliva samples of reportedly unrelated individuals (separated by at least two generations) belonging to 14 populations across Karnataka; 1447 of these individuals were included in the final analysis. Populations were selected to represent a diverse cross-section of variation in Karnataka and included all five major caste groups and two major tribal groups. All samples were collected with the informed written consent of the donors and the study was approved by the Institutional Ethical Committee of the CCMB.

Table 1 Description of the data. Table 1a. The number of individuals and populations included in the paper. Further information is available in Tables S1, S3, and S4. Table 1b. List of SNPs genotyped in Indian populations.

Population category	Number of individuals	Number of populations
Populations genotyped for 18 SNPs		
Within Karnataka	1447	23
Outside Karnataka	1530	38
State	2977	16 (states)
Language family	2977	4 (language families)
Total	2977	61
Publicly available datasets (total 506,306 SNPs)		
World, interpolated maps	1898	94
Reich cline, outside India	829	5 (geog. regions)
Within India	311	9
Total	1898	94

SNP	Gene	Chr	Disease	Discovery	Risk allele	Derived allele	Reference
rs10146997 (A>G)	*NRXN3*	14	Obesity	GWAS—waist circumference	**G**	G	(Heard-Costa et al., 2009)
rs10229583 (G>A)	*PAX4*	7	T2D	1% iHS South Indians	A	A	(Gaulton et al., 2008)
rs10811661 (T>C)	*CDKN2A/B*	9	T2D	GWA	**T**	T	(Zeggini et al., 2008)
rs11208534 (A>G)	*JAK1*	1	T2D	5% iHS South Indians	G	G	(Gaulton et al., 2008)
rs12330015 (A>G)	*PPARA*	22	T2D	1% iHS South Indians	G	A	(Gaulton et al., 2008)
rs12970134 (G>A)	*MC4R*	18	Obesity	GWA—waist circumference in Indians	**A**	G	(Chambers et al., 2008)
rs13220810 (T>C)	*FOXO3A*	6	T2D	Highly conserved Role in ageing	C	C	(Willcox et al., 2008)
rs1349498 (G>A)	*RAPGEF4*	2	T2D	1% iHS South Indians	A	G	(Gaulton et al., 2008)
rs1713222 (C>T)	*APOB*	2	T2D	1% iHS South Indians	T	T	(Gaulton et al., 2008)
rs17647588 (C>T)	*NFE2L2*	2	T2D	1% iHS South Indians	T	T	(Gaulton et al., 2008)
rs17782313 (T>C)	*MC4R*	18	Obesity	GWA	**C**	C	(Loos et al., 2008)
rs2237892 (C>T)	*KCNQ1*	11	T2D	GWA in Asians	**C**	T	(Unoki et al., 2008; Yasuda et al., 2008)
rs6802898 (C>T)	*PPARG*	3	T2D	1% iHS South Indians; GWA	**C**	C	(Altshuler et al., 2000; Gaulton et al., 2008)
rs7578597 (T>C)	*THADA*	2	T2D	GWA	**C**	C	(Zeggini et al., 2008)
rs7903146 (C>T)	*TCF7L2*	10	T2D	GWA	**T**	C	(Saxena et al., 2006)
rs985694 (C>T)	*ESR1*	6	T2D	Gaulton (2008) candidate	T	C	(Gaulton et al., 2008)
rs9911630 (G>A)	*BRCA1*	17	Breast cancer	Potential candidate	A	A	(Miki et al., 1994; Larsson et al., 2007)
rs9939609 (T>A)	*FTO*	16	Obesity	GWA	**A**	T	(Frayling et al., 2007)

The "Discovery" column refers to the reasons that the SNP was chosen for genotyping. In the risk allele column, the actual risk alleles are indicated in bold while the rest are minor alleles, unless otherwise indicated.

Published Genome-Wide Data Sources

Indian samples were grouped by geographic region, language family, or caste/tribe status (Table S4). Because Uttar Pradesh Brahmins and Gujaratis had larger sample sizes compared with other south Asian populations, they were not grouped into these broader categories but were analyzed separately. All populations had a minimum sample size of seven individuals. We estimated genome-wide average F_{ST} among populations from a combined dataset including 506,306 SNPs. PLINK software was used to assemble all genome-wide marker data (Purcell et al., 2007).

Data from 1898 individuals belonging to 94 distinct global populations drawn from six published sources were included in this study (Li et al., 2008; Behar et al., 2010; Rasmussen et al., 2010; The International HapMap Consortium, 2010; Gallego Romero et al., 2011; Metspalu et al., 2011; Table S3).

SNP Selection for Genotyping

Samples from Karnataka (1447 samples) and from the CCMB collection (1530 samples) were genotyped for 18 SNPs in gene regions with potential roles in T2D and obesity etiology in Indian populations. Seven of these variants have been confirmed to be associated with either T2D or obesity in GWA studies (Table 1). Because the association of these SNPs with T2D or obesity was determined in largely European GWA studies, and other SNPs may also serve as good candidates for these diseases in Indian populations, we used other methods to select SNPs that may be candidates for T2D and obesity risk in Indians. An additional 6 out of the 18 variants came from a list of 222 candidate genes involved in T2D (Gaulton et al., 2008), selected on the basis of evidence of scans of extended haplotype homozygosity applied on south Asian populations, using the linkage disequilibrium-based iHS statistic (Voight et al., 2006; Metspalu et al., 2011). We isolated genes belonging to the top 1% to top 5% of iHS scores for inclusion in the study. Other SNPs were chosen based on other biological indicators of their candidacy (Table 1). Of the 18 variants, the seven variants rs10811661, rs12970134, rs17782313, rs2237892, rs7578597, rs7903146, rs9939609 were also found to be associated with T2D, obesity, or related traits in Asian Indians (Bodhini et al., 2007; Chambers et al., 2008; Rees et al., 2008; Yajnik et al., 2009; Been et al., 2011; Rees et al., 2011; Taylor et al., 2011; Dwivedi et al., 2012; Gupta et al., 2012; Li et al., 2012; Vasan et al., 2012; Dwivedi et al., 2013).

The final criterion for SNP selection was compatibility in the multiplex design. Taken together, the full list of T2D- and obesity-associated SNPs, candidate SNPs in the Gaulton et al. (2008) list, as well as other candidate loci provide several hundred testable SNPs. The Sequenom genotyping platform (Sequenom GmbH, Hamburg, Germany) uses a multiplex-based system to type multiple SNPs at the same time. An algorithm provided by Sequenom was used to create optimal combinations of SNPs to minimize the chance of SNP genotyping failure.

Upon testing for Hardy-Weinberg equilibrium and applying a Bonferroni correction to all SNPs in all samples, only two populations showed significant deviation from HWE (Table S1). We decided to retain the populations and SNPs showing deviation from HWE in the analysis, because of: (1) confidence in the genotype scoring method, (2) large sample size, and (3) a potential role of the SNPs in T2D and obesity etiology in Indians. Genotyping these SNPs, and re-sequencing these regions in additional, independent cohorts of Reddy and Ao Naga populations is required to confirm the significance of the deviations from HWE.

DNA Isolation from Saliva

DNA isolation from saliva samples was carried out using two different protocols. The first involved DNA extraction kits (Oragene, DNA Genotek Inc., Kanata, Canada) that were used to collect saliva and extract DNA from approximately 400 of the 1500 participants from Karnataka state, India. Saliva collection and DNA extraction were carried out according to manufacturer's protocols. DNA pellets were dissolved in 100 μl of autoclaved double-distilled water, or autoclaved milliQ water.

The majority of saliva samples collected in Karnataka were processed using a noncommercial DNA extraction protocol (Quinque et al., 2006), which was further modified to accommodate variations in saliva-buffer solution amounts across subjects. For each milliliter of saliva-lysis buffer solution, 15 μl proteinase K, at a concentration of 30 mg/ml (Sigma-Aldrich, Bangalore, India), 75 μl 10% SDS, and 200 μl 5M NaCl were added into the conical tube containing the sample. The proteinase K was kept on ice, and the sample tubes and 10% SDS were kept at room temperature, prior to the above step. The tubes were then placed into a shaking water bath for 24 h at 53°C. Crude proteinase K was sometimes used. In this instance, proteinase K was incubated at room temperature for 10 min to allow it to remove its own proteases. The concentration was increased to 50 mg/ml. Samples were also incubated for 36 h instead of 24 h. DNA samples were subsequently stored in autoclaved, distilled water.

Genotyping Using Sequenom Platform

All samples were genotyped using the MassARRAY system (Sequenom GmbH) for 32 SNPs (32-plex system), although only 18 of these SNPs are included in the present study.

All reactions were carried out according to manufacturer's protocols. Approximately 5 ng of DNA was used for each reaction, corresponding to 1 well on a 384-well plate. A linear regression to calculate the appropriate primer dilutions and amounts of primer to be pooled for the 32-plex reaction was used for greater accuracy. Genotypes were inferred using the software provided with the instrument. We obtained >95% SNP calls for all 32 SNPs, with the exception of one of the 384-well plates, for which 14% of SNP calls were not available. Ambiguous genotypes were visually scored.

Geographic Analyses

ESRI ArcMap software (v. 9.2) was used to visualize spatial patterns of allele frequencies on geographic maps. Shapefiles of the world and India were obtained online (http://www.vdstech.com/map-data.aspx, http://www.diva-gis.org/gdata). Interpolated patterns of allelic variation on a global level were generated using the inverse-distance weighted method as implemented in the Spatial Analyst Tools function within ArcMap.

These interpolations were made based on the 12 nearest points to the region of interpolation, restricted to land-only boundaries. To extend the interpolation for full global coverage, four dummy points were created to represent the extreme points of the map, with dummy frequencies, which fell in the range of the frequency values. Any observed latitudinal or longitudinal patterns were confirmed by Spearman rank correlation, with Bonferroni-corrected p-values to calculate statistical significance.

A Mantel test was used to test for the relationship between genetic and geographic distance. For each pair of populations, we calculated geographic distance in kilometers based on great circle distances measured using the haversine formula (Sinnott, 1984).

Assuming that until very recently populations followed a land-only migration route from Africa and avoided crossing large bodies of water, obligatory waypoints were added to calculate pairwise population distances across continents. We incorporated the five waypoints used by Ramachandran et al. (2005). As we also included several Indian populations, we added two additional waypoints: Karachi, Pakistan (25.0, 69.0) and Kolkata, India (22.6, 88.4), through which all populations entering the Indian subcontinent from the west and the east, respectively, were forced to travel.

Descriptive Statistics

Estimates of descriptive statistics such as observed and expected levels of heterozygosity, and tests for Hardy-Weinberg equilibrium were calculated using GDA software (Lewis & Zaykin, 2001). Allele frequencies were calculated using Arlequin v. 3.5.1.2 software (Excoffier et al., 2005).

GDA implements the unbiased estimator of observed heterozygosity proposed by Nei (Nei, 1987). According to this formula, observed heterozygosity $(H_O) = 1 - \sum_1^k Xii$, in which X_{ii} is the relative frequency of each of the k possible homozygous genotypes at a given locus.

Expected heterozygosity is implemented according to Nei (1987), in which expected heterozygosity (H_S) in the sample (calculated as $1 - \sum_1^k pi^2$) is multiplied by the factor $\frac{2n}{2n-1}$ to account for variation in population sizes. Here, p_i is the frequency of the alleles observed at a given SNP.

Genetic distances, or degree of population differentiation, were measured using F_{ST}, for all pairs of populations for each SNP, using FSTAT software v. 2.9.3 (Weir & Cockerham, 1984; Goudet, 2001). The unbiased estimate of F_{ST} can sometimes have negative values, which do not have biological significance, or may result in error values if minor allele count is zero in a pair of populations. The negative and error F_{ST} values were thus set to zero.

F_{ST} values were calculated as $\frac{a}{a+b+c}$, where a, b, and c are determined by equations 2, 3, and 4 in Weir & Cockerham (1984). F_{ST} values estimated across multiple markers, or on the genome-wide set of markers were calculated using the mean of each of a, b, and c.

Other Statistical Analyses

The program R (v. 2.11.1) was used to calculate Fisher's Exact tests as well as Spearman rank correlations (R Development Core Team, 2006) between pairs of variables. P-values for significance of the Spearman rank correlations were corrected for multiple testing using Holm's method, equivalent to the Bonferroni method of correction.

Mantel tests for correlation between genetic and geographic distances, generated through FSTAT and the haversine, respectively, were calculated using GenAlEx v. 6.4 (Peakall & Smouse, 2006).

Results

Hardy-Weinberg equilibrium tests revealed that the 18 genotyped SNPs (Table 1) were in HWE in most of the 61 populations, with only two exceptions after Bonferroni correction (Tables S1 and S2).

To place frequency differences among the 61 Indian populations into a global geographic context, we examined allele frequency differences at several geographic scales: global, across a continental India cline, within India, among state and language groups within India and within the single state of Karnataka.

Table 2 Spearman rank correlation between allele frequencies, latitude, and longitude, with Bonferroni-corrected p-values.

			Spearman rank correlation					
			Latitude		Longitude		Mantel test	
Geographic level	SNP	Allele	ρ	p_{corr}	ρ	p_{corr}	r^2	p
World								
NRXN3	rs10146997 (A>G)	G	−0.14	1	**−0.73**	**$<10^{-4}$**	0.287	0.0001
RAPGEF4	rs1349498 (G>A)	A	0.22	1	**0.63**	**$<10^{-4}$**	0.152	0.0134
APOB	rs1713222 (C>T)	T	−0.08	1	**−0.54**	**$<10^{-4}$**	0.263	0.0003
NFE2L2	rs17647588 (C>T)	T	0.25	0.936	**−0.49**	**$<10^{-4}$**	0.399	0.0001
FOXO3A	rs13220810 (T>C)	C	**0.38**	**0.015**	−0.09	1	0.233	0.0007
ESR1	rs985694 (C>T)	T	−0.08	1	**0.74**	**$<10^{-4}$**	0.192	0.0012
BRCA1	rs9911630 (G>A)	A	**0.63**	**$<10^{-4}$**	−0.04	1	0.309	0.0024
THADA	rs7578597 (T>C)	C	−0.1	1	**−0.35**	**0.041**	0.227	0.0263
TCF7L2	rs7903146 (C>T)	T	−0.12	1	**−0.49**	**$<10^{-4}$**	0.386	0.0001
KCNQ1	rs2237892 (C>T)	C	−0.01	1	−0.33	0.092	0.425	$<10^{-4}$
Eurasia								
NRXN3	rs10146997 (A>G)	G	0.03	1	−0.75	$<10^{-4}$		
RAPGEF4	rs1349498 (G>A)	A	0.27	1	0.72	$<10^{-4}$		
APOB	rs1713222 (C>T)	T	−0.13	1	−0.66	$<10^{-4}$		
NFE2L2	rs17647588 (C>T)	T	0.15	1	−0.73	$<10^{-4}$		
ESR1	rs985694 (C>T)	T	−0.2	1	0.78	$<10^{-4}$		
BRCA1	rs9911630 (G>A)	A	**0.72**	**$<10^{-4}$**	**0.01**	**1**		
THADA	rs7578597 (T>C)	C	−0.11	1	−0.45	0.003		
TCF7L2	rs7903146 (C>T)	T	−0.23	1	−0.64	$<10^{-4}$		
KCNQ1	rs2237892 (C>T)	C	−0.26	1	−0.58	$<10^{-4}$		
MC4R	rs12970134 (G>A)	A	−0.26	1	−0.4	0.021		
India								
KCNQ1	rs2237892 (C>T)	C	**−0.51**	**0.005**	−0.26	1	−0.032	0.395
India language groups								
KCNQ1	rs2237892 (C>T)	C	−1	$<10^{-4}$	−0.5	$<10^{-4}$		
THADA	rs7578597 (T>C)	C	−0.5	$<10^{-4}$	−1	$<10^{-4}$		
NRXN3	rs10146997 (A>G)	G	−0.5	$<10^{-4}$	−1	$<10^{-4}$		
JAK1	rs11208534 (A>G)	G	0.5	$<10^{-4}$	−1	$<10^{-4}$		

Some of these values are also displayed in Figure 1. The numbers in bold reflect statistically significant correlations. Mantel correlations between F_{ST} and geographic distance are also given for all the world populations, as well as the India sequenom groups. For most populations, Mantel correlations between F_{ST} and geographic distance range between 0.2 and 0.4. These correlations are lower than the Mantel correlation of 0.8851 reported by Ramachandran et al. (2005) in an analysis of 783 microsatellites in 53 populations, 49 of which form a subset of the 94 global populations in this study (Ramachandran et al., 2005; Table S2).

Global Scale/Eurasia

On a global scale, using 94 populations across the world, several variants showed patterns which strongly correlated with longitude, as opposed to latitude (Table 2; Fig. 1). Mantel correlation between F_{ST} (Weir & Cockerham, 1984) and geographic distance with the inclusion of obligatory waypoints is between 0.2 and 0.4 ($p < 10^{-3}$), for the alleles listed in Table 2 and which show frequency differences along latitudinal or longitudinal gradients (Table 2). Many of the loci follow clear longitudinal gradients across Eurasia (Fig. 1). Spearman rank correlations between longitude, latitude, and allele frequencies in the Eurasian subset of the 94 global populations generally followed the same trend as the global populations, perhaps because most of the populations included in the global analysis came from Eurasia (Table 2). However, correlation between *FOXO3A* rs13220810 C and latitude disappeared, and instead correlation between *MC4R* rs12970134A and longitude became significant (Fig. 1; Table 2).

Patterns of Variation along the Indian Cline

The 62 Indian populations genotyped on the Sequenom platform show comprehensive geographic distribution across

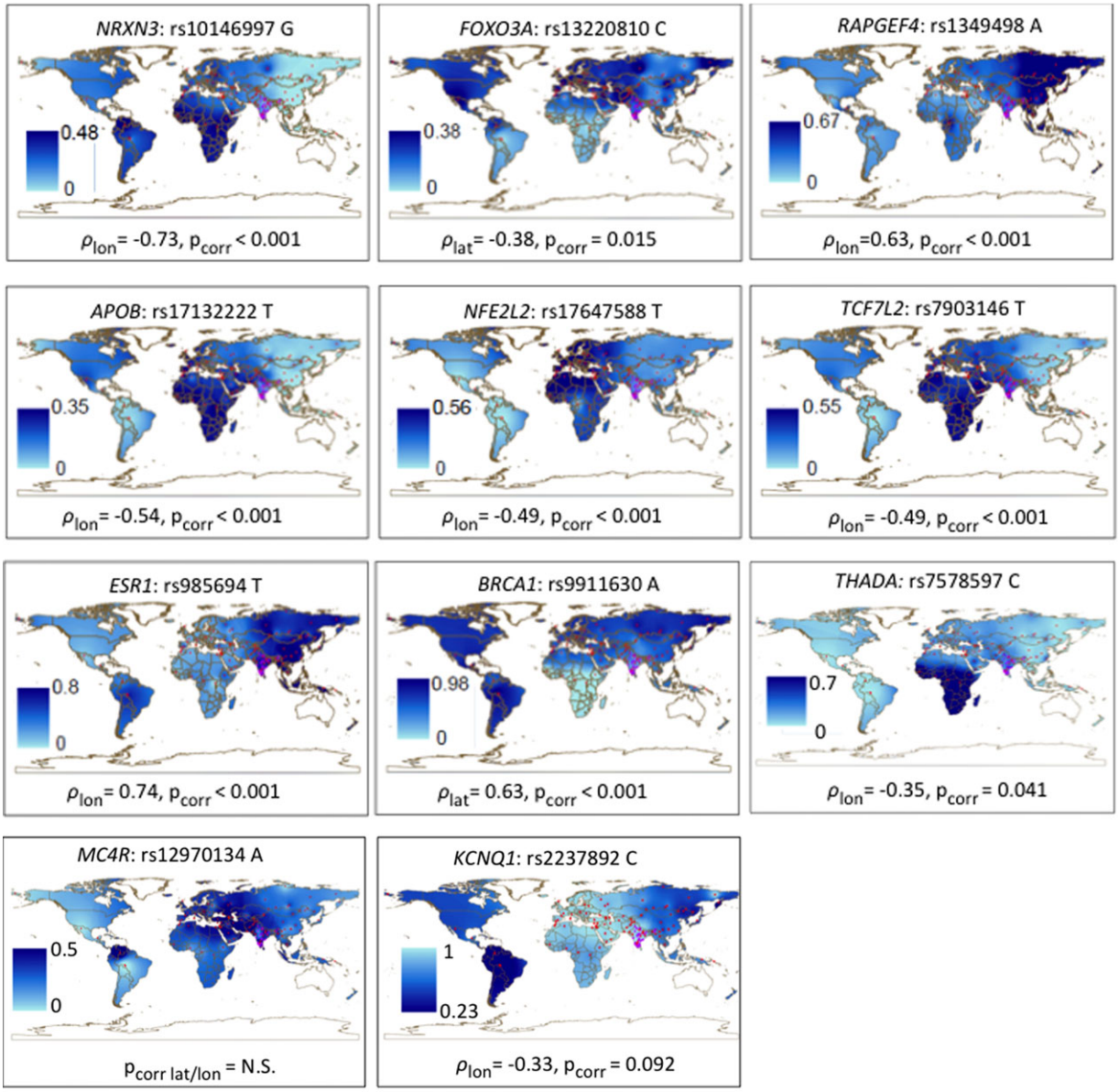

Figure 1 Distribution of allele frequencies across global populations. The line above each map specifies the allele and the line below the map gives the Spearman rank correlation coefficient and Bonferroni-corrected p-value for significance of correlation between latitude ("lat") and longitude ("lon") and allele frequency. The reference alleles for T2D-associated *TCF7L2, THADA, KCNQ1*, and obesity-associated *NRXN3* and *THADA* are risk alleles for the diseases. Global patterns and correlations for 11 out of the 18 SNPs are shown here because they were statistically significant. The two exceptions are *MC4R* rs12970134 and *KCNQ1* rs2237892, which may be significant within India (Fig. 2).

India, particularly on the north-south axis, allowing us to test for evidence of T2D- and obesity-associated allele frequency patterns following the northwest-southeast Indian cline revealed by genome-wide patterns (Reich et al. 2009) and the north-south cline in T2D prevalence (Fig. S1). Spearman rank correlations estimated on a set of seven Indian populations (excluding Austro-Asiatic and Sino-Tibetan Indians) and five non-Indian population groups ("Caucasus," "Central Asia," "Europe," "Near East," "Pakistan,"; Table S3) showed significant correlations between frequencies of

Table 3 Spearman rank and Mantel correlations among populations included in the "Indian cline" analysis.

			Spearman rank correlation					
			Latitude		Longitude		Mantel test	
Gene	SNP	Allele	ρ	p_{corr}	ρ	p_{corr}	r^2	p
NRXN3	rs10146997 (A>G)	G	0.36	1	**−0.83**	**0.014**	**0.35**	**0.026**
MC4R	rs12970134 (G>A)	A	−0.09	1	−0.24	1	−0.11	0.255
KCNQ1	rs2237892 (C>T)	C	−0.31	1	−0.23	1	**−0.22**	**0.028**
THADA	rs7578597 (T>C)	C	−0.03	1	−0.2	1	0.15	0.206
TCF7L2	rs7903146 (C>T)	T	−0.16	1	−0.51	1	−0.12	0.299
PPARG	rs6802898 (C>T)	C	−0.24	1	0.29	1	−0.04	0.476
PAX4	rs10229583 (G>A)	A	−0.53	1	−0.07	1	0.22	0.116
PPARA	rs12330015 (A>G)	G	0.68	0.605	−0.03	1	0.12	0.198
RAPGEF4	rs1349498 (G>A)	A	0.11	1	0.43	1	0.05	0.318
APOB	rs1713222 (C>T)	T	0.36	1	−0.43	1	−0.10	0.376
NFE2L2	rs17647588 (C>T)	T	0.54	1	**−0.82**	**0.018**	**0.86**	**$<10^{-4}$**
FOXO3A	rs13220810 (T>C)	C	0.35	1	−0.59	1	−0.16	0.161
ESR1	rs985694 (C>T)	T	−0.51	1	**0.85**	**0.006**	**0.81**	**$<10^{-4}$**
BRCA1	rs9911630 (G>A)	A	0.77	0.095	−0.29	1	0.40	0.012

The "World" populations show the groupings used for outside India populations, and "India" are the populations from India included in the analysis. India cline populations include the "World" populations grouped into "Caucasus," "Central Asia," "Pakistan," "Near East," and "Europe." The Indian populations are grouped into "UP Brahmins," "Central India tribe," "Gujaratis," "North India caste," "North India tribe," "South India caste," and "South India tribe." Details on the populations included in these groupings are provided in Table S3.

SNPs *NRXN3* rs10146997 ($\rho = -0.83$, $p_{corr} = 0.0142$), *NFE2L2* rs17647588 ($\rho = -0.82$, $p_{corr} = 0.0181$) and *ESR1* rs985694 ($\rho = 0.85$, $p_{corr} = 0.0062$) and longitude. None of these SNPs showed a significant correlation with latitude. In accordance with known geographic patterns in skin pigmentation, however, both *ESR1* and *BRCA1* showed geographic patterning along the Indian cline (Table 3), and *BRCA1* additionally correlated with latitude in global populations (Table 2; Jablonski & Chaplin, 2000). Mantel correlations, however, revealed a strong correlation between F_{ST} and geographic distance for *NFE2L2*, and *ESR1*, corresponding also to strong correlation between minor allele frequency and longitude (Table 3). Interestingly, *KCNQ1* shows negative Mantel correlation across the Indian cline, suggesting less genetic diversity with increased geographic distance. This result stands in contrast to the Mantel correlation estimated in the global analysis in which populations were not grouped by geographic region and involved a larger geographic range of populations (Table 3). The difference could be due to the nature of the population groupings in the Indian-cline analysis: populations in larger geographic regions (World) were grouped together, while populations at smaller geographic scales were left intact (India; Table 3). The grouping scheme used here suggests that correlations with geography may be significant at a macrolevel scale of population sampling, but may not be strong enough to reach significance at a microlevel scale.

Patterns of Variation within India

While many of the SNPs presently studied showed geographic patterns that mirrored latitudinal or longitudinal gradients on a global scale, these were absent or less pronounced in the Indian populations. Furthermore, Mantel test results show higher correlation of geographic and genetic distance at a global level than within India (Table 2). The pattern found here is consistent with frequencies of other disease-associated variants, which appear to vary along a latitudinal cline in world populations but not within India (Pemberton et al., 2008). Only a weak Mantel correlation between F_{ST} at *KCNQ1* rs2237892 C and geographic distance was observed. This correlation may be due to the inclusion of the Sino-Tibetan language-speaking Nyshi and Ao Naga populations of Northeast India, which show dramatically lower risk allele frequencies compared to populations in the rest of India, wherein the allele is close to or at fixation (Fig. 2).

Thus far, few studies have examined genome-wide variation among Indian populations by their geography (Reich et al., 2009; Metspalu et al., 2011). The Reich et al. (2009) study estimated genome-wide F_{ST} to be 0.01 among Indian populations, excluding Sino-Tibetans and other outlying populations, about three times higher than among European populations. When they adjusted their estimate to account for the effects of inbreeding, which could

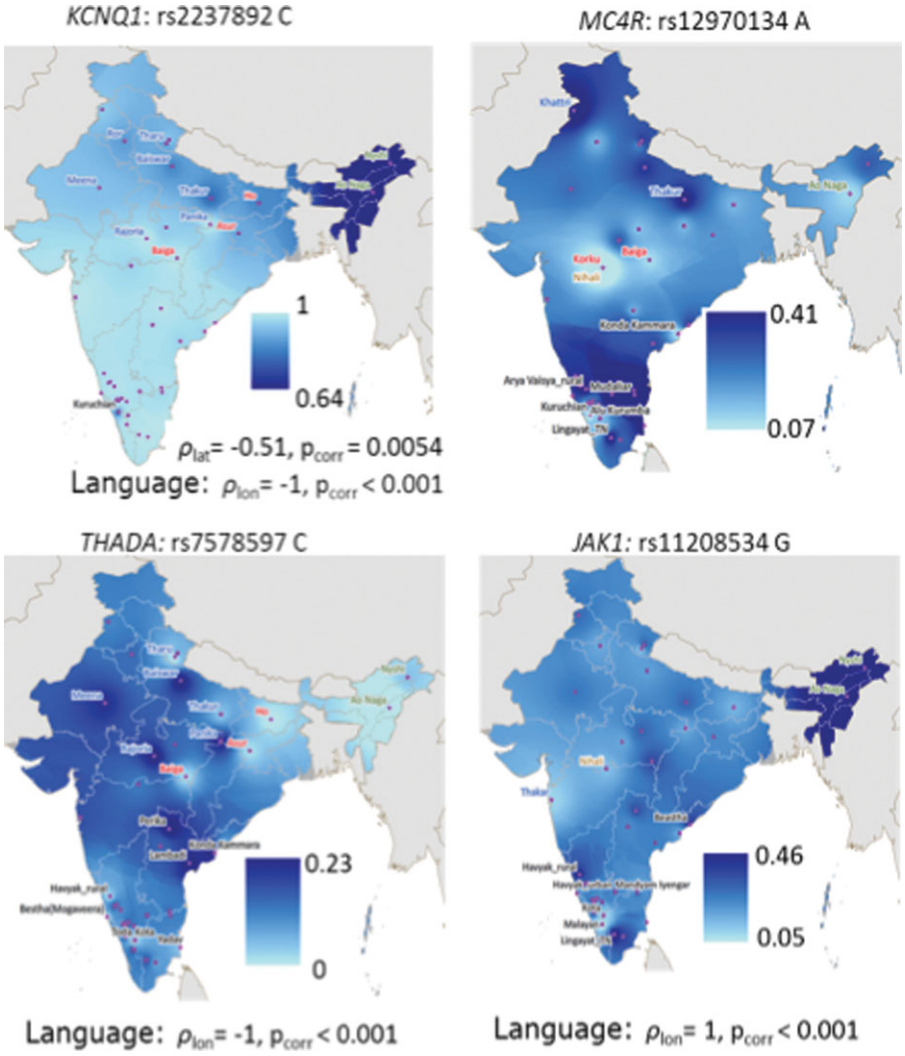

Figure 2 Distribution of rs12970134 A, rs2237892 C, rs7578597 C, and rs11208534 G alleles within India. Spearman rank correlation coefficients between rs12970134 allele frequency, latitude, and longitude were not significant on a global scale or within India. Red dots on the world map represent populations. Populations within India, showing unusual allele frequency differences, labeled in red speak Austro-Asiatic languages, those in blue speak Indo-European languages, green speak Sino-Tibetan languages, black speak Dravidian languages, and brown are a linguistic isolate (Nihali). For *KCNQ1*, the colors are reversed in the within-India group for clarity. Spearman rank correlations are provided both across all Indian populations, as well as populations grouped by language family. For the *JAK1* locus, only Spearman rank correlation within India is provided, because this locus was unavailable in the global dataset.

inflate differences between populations, the F_{ST} value decreased to 0.0069. To provide a comparative estimate based on the Illumina samples used in this study, we also calculated F_{ST} between north and south Indian population groups at 9942 SNPs sampled randomly from the genome (Table S5).

Pairwise F_{ST} differences between north and south Indian population groups in our genome-wide dataset showed values resembling the Reich et al. (2009) inbreeding-adjusted estimate from the Affymetrix data, although the F_{ST} values calculated between north and south Indian population groups from the Illumina data were not adjusted for inbreeding.

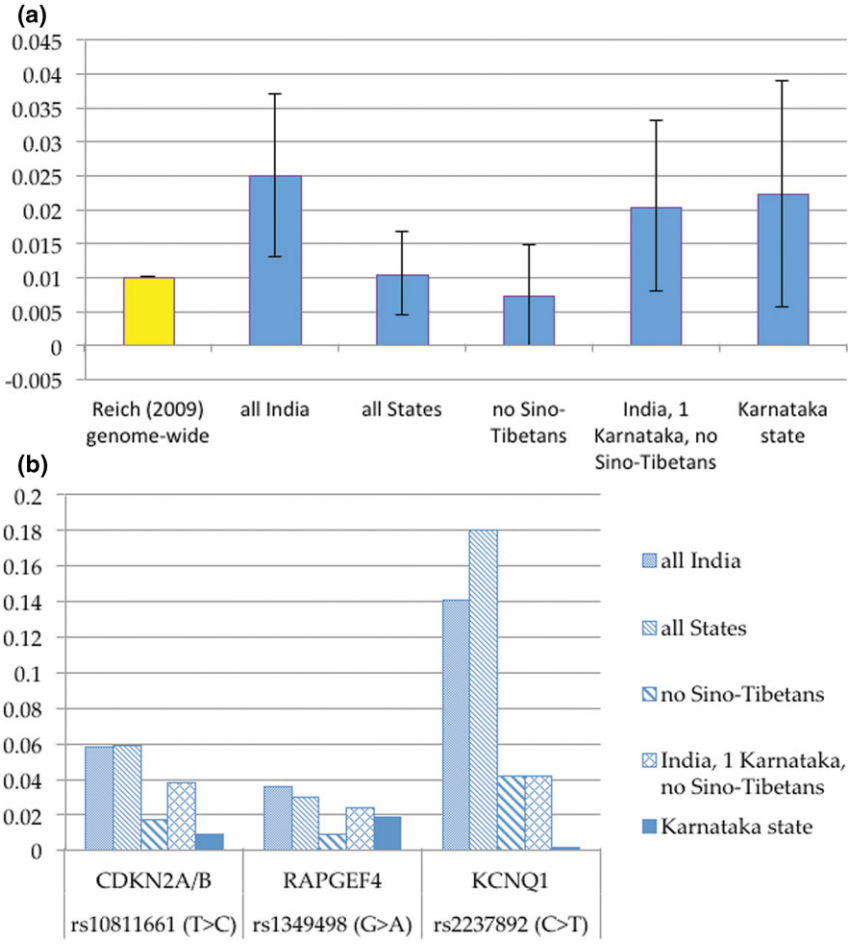

Figure 3 To compare the influence of both population groupings and population sampling at the level of India and also of Karnataka, we calculated F_{ST} values for: (A) all 62 populations, ungrouped; (B) all Indian populations minus the Sino-Tibetan populations; (C) all non-Karnataka Indian populations, only one Karnataka population (Gangadikaara Vokkaliga) and no Sino-Tibetan populations; and (D) Karnataka populations only. We also included published values of F_{ST} among Indians (Reich et al., 2009). Figure 3(A) shows F_{ST} differences among 15 of the 18 SNPs studied, and Figure 3(B) shows F_{ST} differences among the remaining three SNPs; these were separated because they show extreme F_{ST} differences between the different population grouping schemes. F_{ST} values are provided in Table S9.

Grouping by State of Origin

We evaluated allele frequency differences among populations grouped based on Indian state of origin to test for fine-scale patterns of local differentiation at alleles associated with T2D and obesity, although Spearman rank correlations with latitude and longitude and Mantel correlations with geographic distance did not reveal strong geographic patterns of allele frequencies across India among these groups. Examining patterns of allele frequency across Indian states may indicate a latitudinal cline in obesity-associated *MC4R* SNP rs12970134, although Spearman rank correlation did not reveal significant correlation with latitude (Table S7; Fig. 2). Comparisons of F_{ST} between Indian state groups and global populations show lower allele frequency differences among Indian states than among global regions across the 14 SNPs. However, F_{ST} differences corresponding to state groups are similar to the genome-wide F_{ST} value of 0.01 (Reich et al. 2009) suggesting that the studied obesity and T2D risk alleles as a group do not show reduced diversity as expected from their disease association (Fig. 3).

Language Family Grouping

Language families also show geographic clustering in India (Reich et al., 2009; Gallego Romero et al., 2011). Therefore, we carried out Spearman rank correlations tests between allele frequency, latitude, and longitude among populations grouped by language family within India. Strong correlations were observed for four alleles (Table 2; Fig. 2). Of these four alleles, *KCNQ1* rs2237892 and *THADA* rs7578597 were associated with T2D, *NRXN3* rs10146997 with waist circumference, and *JAK1* rs11208534 was found to be undergoing natural selection within India, based on the iHS statistic. When populations were regrouped according to state of residence, none of the allele frequencies correlated with latitude or longitude. This finding is consistent with previous studies based on regional languages, which approximately follow state boundaries, in India (Pemberton et al., 2008). Alternatively, all states were not comprehensively sampled across caste and tribe boundaries; while south Indian states were heavily represented by several castes and tribes, many north and northeast Indian states were only represented by two populations, the Ao Naga and Nyshi, which may have reduced our power to detect geographic patterns among Indian states.

F_{ST} was higher among language family groups than state groups, except for SNPs *NFE2L2* rs17647588, *MC4R* rs17782313, *THADA* rs7578597, and *ESR1* rs985694. Higher F_{ST} values may be attributable to strong differentiation between Sino-Tibetan populations and other Indian populations. Grouping populations by language family confirms strong differences between Sino-Tibetan populations and other Indian populations at almost all loci (Tables S7 and S8). Austro-Asiatic speakers sometimes show intermediate allele frequencies between Sino-Tibetan and Indo-European speakers, in accordance with their geographic distribution and demographic history involving some gene flow from the southeast (Chaubey et al., 2011; Table 2; Fig. 2). Average F_{ST} values for each SNP vary widely, from a minimum of 0.002 at *ESR1* rs985694 to a maximum of 0.277 for *KCNQ1* rs2237892 (Table S8). Variance at F_{ST} values among linguistic groups was slightly higher than variance among global groups, for the same set of 14 SNPs (global variance: 5.08×10^{-3}, language group variance: 5.1×10^{-3}). The high *KCNQ1* F_{ST} value is attributable to the inclusion of the Sino-Tibetan populations, in which the risk allele frequency of 0.66 is identical to the risk allele frequency in the East Asian population group (Tables S6 and S8).

Population Exclusions and Patterns of Variation within Karnataka State

Variation among Indian states was highest when Sino-Tibetan populations were included, especially for *CDKN2A/B, RAPGEF4*, and *KCNQ1* (Fig. 3). Grouping the 62 Indian populations by geographic region (e.g. State) did not always reveal large variation among populations (Fig. 3). On the other hand, substantial variation in allele frequencies among groups sampled at a fine-scale geographic level (e.g. within Karnataka only) suggests that these methods of grouping populations may be inadequate for accurately representing population variation (Raj et al., 2006, 2007).

In all except two instances, F_{ST} differences among Indian populations increased upon removal of all Karnataka populations except the Gangadikaara Vokkaliga population, which was chosen to represent Karnataka in the State-level analyses because it is one of the largest populations in Karnataka. However, removing nearly all of the Karnataka populations only had a minor impact on state-wide F_{ST} values compared with just removing Sino-Tibetan populations but keeping all Karnataka populations. Therefore, the Gangadikaara Vokkaliga population as one of the most common populations in Karnataka serves as a good representative of Karnataka population genetic variation and grouping populations based on state of origin may buffer against large vacillations in allele frequency across populations.

F_{ST} estimates within Karnataka populations were highly variable; at SNPs rs10229583, rs12330015, rs17647588, rs6802898, and rs985694 F_{ST} of Karnataka populations were higher than all other population groups, including F_{ST} of all Indian populations including Sino-Tibetan speakers. At T2D-associated locus *PPARG* rs6802898, the large F_{ST} value for Karnataka populations may be attributed to the Havyak and Arya Vaisya rural and urban populations having substantially lower risk allele frequency, at an average of 22 percentage points lower than other Karnataka populations. At loci such as rs10811661 and rs2237892, however, F_{ST} values of Karnataka populations were lower than among all other population groups, suggesting greater uniformity in allele frequency among populations within Karnataka at these loci (Fig. 3; Table S9). The observed high degrees of variability at disease-associated loci at a fine-scale geographic level (e.g. within Karnataka populations only) suggests that studies designed to investigate T2D and obesity risk, and also perhaps other complex diseases, in Indians must match cases and controls at a fine geographic scale.

Discussion

We studied the distribution of allelic variation at T2D- and obesity-associated loci in India to: (1) test if genetic variation at these loci mirrored the nation-wide distribution of obesity and T2D prevalence, including the variation at loci which have been identified as candidates of positive selection and (2) to test whether measures of population differentiation varied among groups.

We found that T2D- and obesity-associated alleles that show geographic variation on a global scale show less pronounced or no geographic patterning in India, inconsistent with known geographic variation in T2D and obesity prevalence in India. The appearance of predominantly longitudinal as opposed to latitudinal correlations of allele frequencies in the global dataset, and in the restricted Eurasian dataset, as well as statistically significant Mantel correlations between F_{ST} and geographic distance confirms established reports of a correlation between genetic and geographic distance at a broad geographic scale (Prugnolle et al., 2005; Ramachandran et al., 2005; Betti et al., 2009). Pairwise F_{ST} differences between north and south Indian population groups in our genome-wide dataset also showed values resembling the Reich et al. (2009) inbreeding-adjusted estimate, although our F_{ST} values calculated between north and south Indian population groups were not adjusted for inbreeding. There are several possible reasons for this discrepancy, including: (1) the SNPs on which F_{ST} differences are based represent nonrandom variation, (2) the Illumina population groups do not reflect all the geographic regions within India covered by the Reich et al. (2009) study, and (3) grouping several populations into north and south Indian population groups significantly impacts measures of F_{ST}.

Comparisons of allele frequencies across India and within a single state in India suggest that for some variants, differences within and among populations may be the same or greater within a single state than across India, and the degree of variation may depend on population sampling and grouping schemes. Across almost all alleles, inclusion of the Sino-Tibetan speaking populations created inflated estimates of variation (as measured by F_{ST} and AMOVA). Sino-Tibetan speaking populations are known to share closer ancestry with East Asian populations than with South Asians, which may explain this result. Excluding Sino-Tibetan populations from the analyses, however, did not drastically reduce variation at the loci. We employed the same strategy for Illumina samples, in which only one or two individuals were sampled from a single, endogamous population. The grouping scheme may not have provided accurate results, however, as genome-wide estimates of F_{ST} fell two orders of magnitude below published estimates of Indian F_{ST} values (Table S5). Alternatively, the randomly chosen alleles used to estimate genome-wide F_{ST} in the Illumina samples may not truly represent neutral variation.

Whether the SNPs investigated in this dataset represent neutral variation, disease-associated variation, or variants under selection in Indian populations may also influence patterns of genetic variation. Lack of correlation between allelic variation and T2D and obesity prevalence trends suggests that either these trends are influenced more by environmental than genetic factors, or by other SNPs that are yet to be determined. Association studies in Indian populations may suggest other variants that better explain T2D and obesity in Indian populations. Furthermore, most of the disease-associated alleles examined here also do not follow previously published patterns of neutral variation in India, referred to here as the "Indian cline," following a gradient in allele frequency variation from Europe to India (Reich et al. 2009). These results may not be entirely surprising, as not all neutral or disease-associated SNPs will be expected to follow the same geographic pattern. SNPs in the obesity-associated *NRXN3* and T2D-associated *KCNQ1* genes, however, somewhat follow the "Indian cline" (Table 3), although overall, less geographic variation was observed within India than across global populations. As already mentioned, these patterns could be due to either strong effects of selection, or drift at individual loci; note, however, that neither *NRXN3* nor *KCNQ1* that followed the geographic trend expected from genome-wide average data appeared to be under selection based on the local partial sweep iHS statistic. However, founder effects and genetic drift may be more pronounced in Indian populations than in other populations because many of them have a characteristically small size and high levels of endogamy (Reich et al., 2009).

We did not find any significant clinal patterns with *PPARG* variant rs6802898, chosen for genotyping because of its high iHS score ranking in the Indian populations. Unlike other variants that were selected as T2D candidates from Gaulton et al. (2008), rs6802898 is an intronic SNP in the *PPARG* gene, in which the Pro12Ala variant has been previously reported to be associated with T2D and obesity in Indian populations (Sanghera et al., 2010; Vimaleswaran et al., 2010; Prakash et al., 2012). The variant genotyped here was not previously reported to be associated with T2D. While some sharing of variants associated with T2D and obesity exists between European and Indian populations, there are a number of variants, which are associated only in Indian populations. Recent studies have identified new loci associated with T2D in South Asians, not previously found to be associated with T2D in other populations (Vimaleswaran et al., 2010; Kooner et al., 2011; Tabassum et al., 2012). It remains to be tested whether these newly discovered variants correlate with the north to south geographic patterns in India.

The sampling strategy used here comprehensively represented Indian populations on the north-south axis, but the east-west axis was less well-covered. Future studies may benefit from increased genetic information on Indian populations, additional studies to identify markers that specifically influence diabetes and obesity in Indians, and wider geographic sampling to gain a more complete understanding of the relationship among genetic and geographic variation.

Acknowledgements

We would like to thank all the participants for providing saliva samples for the DNA analysis, and over 80 individuals and organizations that helped in the process. In particular, the authors would like to acknowledge Mr. and Mrs. H. B. Rajagopal, Mahadeva, Mrs. Poornima Rangappa and Mr. Girijashankar for their help in coordinating sample collection. Maggie Bellatti, Krishnendu Khan, Jasbeer Singh, Charles Spurgeon, and Kranthi Kumar provided support in the laboratory. Drs Gabriel Amable and Paco Bertolani assisted in the generation of the interpolated maps. Finally, funding for this work came from the UK-India Education and Research Initiative, Gates Cambridge Trust, Centre for Human Genetics and Indian Institute of Science (Bangalore, India), the Bridget's Trust, Gonville and Caius College, the Cambridge-India Partnership Fund, as well as CardioMed-BSC0122 of Council of Scientific and Industrial Research (CSIR), Government of India.

References

Altshuler, D., Hirschhorn, J. N., Klannemark, M., Lindgren, C. M., Vohl, M. C., Nemesh, J., Lane, C. R., Schaffner, S. F., Bolk, S., Brewer, C., Tuomi, T., Gaudet, D., Hudson, T. J., Daly, M., Groop, L. & Lander, E. S. (2000) The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet* **26**, 76–80.

Been, L. F., Ralhan, S., Wander, G. S., Mehra, N. K., Singh, J., Mulvihill, J. J., Aston, C. E. & Sanghera, D. K. (2011) Variants in KCNQ1 increase type II diabetes susceptibility in South Asians: A study of 3,310 subjects from India and the US. *BMC Med Genet* **12**, 18.

Behar, D. M., Yunusbayev, B., Metspalu, M., Metspalu, E., Rosset, S., Parik, J., Rootsi, S., Chaubey, G., Kutuev, I., Yudkovsky, G., Khusnutdinova, E. K., Balanovsky, O., Semino, O., Pereira, L., Comas, D., Gurwitz, D., Bonne-Tamir, B., Parfitt, T., Hammer, M. F., Skorecki, K. & Villems, R. (2010) The genome-wide structure of the Jewish people. *Nature* **466**, 238–242.

Betti, L., Balloux, F., Amos, W., Hanihara, T. & Manica, A. (2009) Distance from Africa, not climate, explains within-population phenotypic diversity in humans. *Proc Biol Sci* **276**, 809–814.

Bodhini, D., Radha, V., Dhar, M., Narayani, N. & Mohan, V. (2007) The rs12255372(G/T) and rs7903146(C/T) polymorphisms of the TCF7L2 gene are associated with type 2 diabetes mellitus in Asian Indians. *Metabolism* **56**, 1174–1178.

Bustamante, C. D., Burchard, E. G. & De La Vega, F. M. (2011) Genomics for the world. *Nature* **475**, 163–165.

Chambers, J. C., Elliott, P., Zabaneh, D., Zhang, W., Li, Y., Froguel, P., Balding, D., Scott, J. & Kooner, J. S. (2008) Common genetic variation near MC4R is associated with waist circumference and insulin resistance. *Nat Genet* **40**, 716–718.

Chaubey, G., Metspalu, M., Choi, Y., Magi, R., Romero, I. G., Soares, P., Van Oven, M., Behar, D. M., Rootsi, S., Hudjashov, G., Mallick, C. B., Karmin, M., Nelis, M., Parik, J., Reddy, A. G., Metspalu, E., Van Driem, G., Xue, Y., Tyler-Smith, C., Thangaraj, K., Singh, L., Remm, M., Richards, M. B., Lahr, M. M., Kayser, M., Villems, R. & Kivisild, T. (2011) Population genetic structure in Indian Austroasiatic speakers: The role of landscape barriers and sex-specific admixture. *Mol Biol Evol* **28**, 1013–1024.

Deepa, M., Pradeepa, R., Rema, M., Mohan, A., Deepa, R., Shanthirani, S. & Mohan, V. (2003) The Chennai Urban Rural Epidemiology Study (CURES) – study design and methodology (urban component) (CURES-I). *J Assoc Physicians India* **51**, 863–870.

Diamond, J. (2011) Medicine: Diabetes in India. *Nature* **469**, 478–479.

Dwivedi, O. P., Tabassum, R., Chauhan, G., Ghosh, S., Marwaha, R. K., Tandon, N. & Bharadwaj, D. (2012) Common variants of FTO are associated with childhood obesity in a cross-sectional study of 3,126 urban Indian children. *PLoS One* **7**, e47772.

Dwivedi, O. P., Tabassum, R., Chauhan, G., Kaur, I., Ghosh, S., Marwaha, R. K., Tandon, N. & Bharadwaj, D. (2013) Strong influence of variants near MC4R on adiposity in children and adults: A cross-sectional study in Indian population. *J Hum Genet* **58**, 27–32.

Excoffier, L., Laval, G. & Schneider, S. (2005) Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evol Bioinform Online* **1**, 47–50.

Finucane, M. M., Stevens, G. A., Cowan, M. J., Danaei, G., Lin, J. K., Paciorek, C. J., Singh, G. M., Gutierrez, H. R., Lu, Y., Bahalim, A. N., Farzadfar, F., Riley, L. M. & Ezzati, M. (2011) National, regional, and global trends in body-mass index since 1980: Systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9.1 million participants. *Lancet* **377**, 557–567.

Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., Perry, J. R., Elliott, K. S., Lango, H., Rayner, N. W., Shields, B., Harries, L. W., Barrett, J. C., Ellard, S., Groves, C. J., Knight, B., Patch, A. M., Ness, A. R., Ebrahim, S., Lawlor, D. A., Ring, S. M., Ben-Shlomo, Y., Jarvelin, M. R., Sovio, U., Bennett, A. J., Melzer, D., Ferrucci, L., Loos, R. J., Barroso, I., Wareham, N. J., Karpe, F., Owen, K. R., Cardon, L. R., Walker, M., Hitman, G. A., Palmer, C. N., Doney, A. S., Morris, A. D., Smith, G. D., Hattersley, A. T. & Mccarthy, M. I. (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889–894.

Gallego Romero, I., Basu Mallick, C., Liebert, A., Crivellaro, F., Chaubey, G., Itan, Y., Metspalu, M., Eaaswarkhanth, M., Pitchappan, R., Villems, R., Reich, D., Singh, L., Thangaraj, K., Thomas, M. G., Swallow, D. M., Lahr, M. M. & Kivisild, T. (2011) Herders of Indian and European cattle share their predominant allele for lactase persistence. *Mol Biol Evol* **29**, 249–260.

Gaulton, K. J., Willer, C. J., Li, Y., Scott, L. J., Conneely, K. N., Jackson, A. U., Duren, W. L., Chines, P. S., Narisu, N., Bonnycastle, L. L., Luo, J., Tong, M., Sprau, A. G., Pugh, E. W., Doheny, K. F., Valle, T. T., Abecasis, G. R., Tuomilehto, J., Bergman, R. N., Collins, F. S., Boehnke, M. & Mohlke, K. L. (2008) Comprehensive association study of type 2 diabetes and related quantitative traits with 222 candidate genes. *Diabetes*, **57**, 3136–3144.

Goudet, J. (2001) FSTAT, a program to estimate and test genetic diversities and fixation indices (version 2.9.3), Available at: http://www.unil.ch/izea/software/fstat.html. Accessed 9 May 2013.

Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., Yu, F., Gibbs, R. A. & Bustamante, C. D. (2011) Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci USA* **108**, 11983–11988.

Gupta, V., Vinay, D. G., Rafiq, S., Kranthikumar, M. V., Janipalli, C. S., Giambartolomei, C., Evans, D. M., Mani, K. R., Sandeep,

M. N., Taylor, A. E., Kinra, S., Sullivan, R. M., Bowen, L., Timpson, N. J., Smith, G. D., Dudbridge, F., Prabhakaran, D., Ben-Shlomo, Y., Reddy, K. S., Ebrahim, S., Chandak, G. R. & Indian Migration Study, G. (2012) Association analysis of 31 common polymorphisms with type 2 diabetes and its related traits in Indian sib pairs. *Diabetologia* **55**, 349–357.

Hales, C. N. & Barker, D. J. (1992) Type 2 (non-insulin-dependent) diabetes mellitus: The thrifty phenotype hypothesis. *Diabetologia* **35**, 595–601.

Heard-Costa, N. L., Zillikens, M. C., Monda, K. L., Johansson, A., Harris, T. B., Fu, M., Haritunians, T., Feitosa, M. F., Aspelund, T., Eiriksdottir, G., Garcia, M., Launer, L. J., Smith, A. V., Mitchell, B. D., Mcardle, P. F., Shuldiner, A. R., Bielinski, S. J., Boerwinkle, E., Brancati, F., Demerath, E. W., Pankow, J. S., Arnold, A. M., Chen, Y. D., Glazer, N. L., Mcknight, B., Psaty, B. M., Rotter, J. I., Amin, N., Campbell, H., Gyllensten, U., Pattaro, C., Pramstaller, P.P., Rudan, I., Struchalin, M., Vitart, V., Gao, X., Kraja, A., Province, M. A., Zhang, Q., Atwood, L. D., Dupuis, J., Hirschhorn, J. N., Jaquish, C. E., O'donnell, C. J., Vasan, R. S., White, C. C., Aulchenko, Y. S., Estrada, K., Hofman, A., Rivadeneira, F., Uitterlinden, A. G., Witteman, J. C., Oostra, B. A., Kaplan, R. C., Gudnason, V., O'connell, J. R., Borecki, I. B., Van Duijn, C. M., Cupples, L. A., Fox, C. S. & North, K. E. (2009) NRXN3 is a novel locus for waist circumference: A genome-wide association study from the CHARGE consortium. *PLoS Genet* **5**, e1000539.

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

International Diabetes Federation (2009) Diabetes atlas. In: *Diabetes Atlas*, 4th ed. (eds. N. Unwin, D. Whiting, D. Gan, O. Jacqmain & G. Ghyoot). Brussels, Belgium: International Diabetes Federation.

Jablonski, N. G. & Chaplin, G. (2000) The evolution of human skin coloration. *J Hum Evol* **39**, 57–106.

Kooner, J. S., Saleheen, D., Sim, X., Sehmi, J., Zhang, W., Frossard, P., Been, L. F., Chia, K. S., Dimas, A. S., Hassanali, N., Jafar, T., Jowett, J. B., Li, X., Radha, V., Rees, S. D., Takeuchi, F., Young, R., Aung, T., Basit, A., Chidambaram, M., Das, D., Grunberg, E., Hedman, A. K., Hydrie, Z. I., Islam, M., Khor, C. C., Kowlessur, S., Kristensen, M. M., Liju, S., Lim, W. Y., Matthews, D. R., Liu, J., Morris, A. P., Nica, A. C., Pinidiyapathirage, J. M., Prokopenko, I., Rasheed, A., Samuel, M., Shah, N., Shera, A. S., Small, K. S., Suo, C., Wickremasinghe, A. R., Wong, T. Y., Yang, M., Zhang, F., Abecasis, G. R., Barnett, A. H., Caulfield, M., Deloukas, P., Frayling, T. M., Froguel, P., Kato, N., Katulanda, P., Kelly, M. A., Liang, J., Mohan, V., Sanghera, D. K., Scott, J., Seielstad, M., Zimmet, P. Z., Elliott, P., Teo, Y. Y., Mccarthy, M. I., Danesh, J., Tai, E. S. & Chambers, J. C. (2011) Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat Genet* **43**, 984–989.

Kumar, R., Seibold, M. A., Aldrich, M. C., Williams, L. K., Reiner, A. P., Colangelo, L., Galanter, J., Gignoux, C., Hu, D., Sen, S., Choudhry, S., Peterson, E. L., Rodriguez-Santana, J., Rodriguez-Cintron, W., Nalls, M. A., Leak, T. S., O'meara, E., Meibohm, B., Kritchevsky, S. B., Li, R., Harris, T. B., Nickerson, D. A., Fornage, M., Enright, P., Ziv, E., Smith, L. J., Liu, K. & Burchard, E. G. (2010) Genetic ancestry in lung-function predictions. *N Engl J Med* **363**, 321–330.

Larsson, S.C., Mantzoros, C. S. & Wolk, A. (2007) Diabetes mellitus and risk of breast cancer: A meta-analysis. *Int J Cancer* **121**, 856–862.

Lewis, P.O. & Zaykin, D. (2001) GDA (genetic data analysis): Computer program for the analysis of allelic data (University of Connecticut, Storrs, CT.). Version 1.0 d16c. Available at: http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php. Accessed 9 May 2013.

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L. & Myers, R. M. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104.

Li, H., Kilpelainen, T. O., Liu, C., Zhu, J., Liu, Y., Hu, C., Yang, Z., Zhang, W., Bao, W., Cha, S., Wu, Y., Yang, T., Sekine, A., Choi, B. Y., Yajnik, C. S., Zhou, D., Takeuchi, F., Yamamoto, K., Chan, J. C., Mani, K. R., Been, L. F., Imamura, M., Nakashima, E., Lee, N., Fujisawa, T., Karasawa, S., Wen, W., Joglekar, C. V., Lu, W., Chang, Y., Xiang, Y., Gao, Y., Liu, S., Song, Y., Kwak, S. H., Shin, H. D., Park, K. S., Fall, C. H., Kim, J. Y., Sham, P. C., Lam, K. S., Zheng, W., Shu, X., Deng, H., Ikegami, H., Krishnaveni, G. V., Sanghera, D. K., Chuang, L., Liu, L., Hu, R., Kim, Y., Daimon, M., Hotta, K., Jia, W., Kooner, J. S., Chambers, J. C., Chandak, G. R., Ma, R. C., Maeda, S., Dorajoo, R., Yokota, M., Takayanagi, R., Kato, N., Lin, X. & Loos, R. J. (2012) Association of genetic variation in FTO with risk of obesity and type 2 diabetes with data from 96,551 East and South Asians. *Diabetologia* **55**, 981–995.

Loos, R. J., Lindgren, C. M., Li, S., Wheeler, E., Zhao, J. H., Prokopenko, I., Inouye, M., Freathy, R. M., Attwood, A. P., Beckmann, J. S., Berndt, S. I., Jacobs, K. B., Chanock, S. J., Hayes, R. B., Bergmann, S., Bennett, A. J., Bingham, S. A., Bochud, M., Brown, M., Cauchi, S., Connell, J. M., Cooper, C., Smith, G. D., Day, I., Dina, C., De, S., Dermitzakis, E. T., Doney, A. S., Elliott, K. S., Elliott, P., Evans, D. M., Sadaf Farooqi, I., Froguel, P., Ghori, J., Groves, C. J., Gwilliam, R., Hadley, D., Hall, A. S., Hattersley, A. T., Hebebrand, J., Heid, I. M., Lamina, C., Gieger, C., Illig, T., Meitinger, T., Wichmann, H. E., Herrera, B., Hinney, A., Hunt, S. E., Jarvelin, M. R., Johnson, T., Jolley, J. D., Karpe, F., Keniry, A., Khaw, K. T., Luben, R. N., Mangino, M., Marchini, J., Mcardle, W. L., Mcginnis, R., Meyre, D., Munroe, P. B., Morris, A. D., Ness, A. R., Neville, M. J., Nica, A. C., Ong, K. K., O'rahilly, S., Owen, K. R., Palmer, C. N., Papadakis, K., Potter, S., Pouta, A., Qi, L., Randall, J. C., Rayner, N. W., Ring, S. M., Sandhu, M. S., Scherag, A., Sims, M. A., Song, K., Soranzo, N., Speliotes, E. K., Syddall, H. E., Teichmann, S. A., Timpson, N. J., Tobias, J. H., Uda, M., Vogel, C. I., Wallace, C., Waterworth, D. M., Weedon, M. N., Willer, C. J., Wraight Yuan, X., Zeggini, E., Hirschhorn, J. N., Strachan, D. P., Ouwehand, W. H., Caulfield, M. J., Samani, N. J., Frayling T. M., Vollenweider, P., Waeber, G., Mooser V., Deloukas P., McCarthy, M. I., Wareham, N. J., Barroso, I., Jacobs K. B., Chanock, S. J., Hayes, R. B., Lamina, C., Gieger, C., Illig, T., Meitinger, T., Wichmann, H. E., Kraft, P., Hankinson, S. E., Hunter, D. J., Hu, F. B., Lyon, H. N., Voight, B. F., Ridderstrale, M., Groop, L., Scheet, P., Sanna, S., Abecasis, G. R., Albai, G., Nagaraja, R., Schlessinger, D., Jackson, A. U., Tuomilehto, J., Collins, F. S., Boehnke, M., Mohlke, K. L. (2008) Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat Genet* **40**, 768–775.

McKeigue, P. M. (1989) Disturbances of insulin in British Asian and white men. *BMJ* **299**, 1161–1162.

McKeigue, P. M., Shah, B. & Marmot, M. G. (1991) Relation of central obesity and insulin resistance with high diabetes prevalence and cardiovascular risk in South Asians. *Lancet* **337**, 382–386.

Metspalu, M., Gallego Romero, I., Yunusbayev, B., Chaubey, G., Mallick, C. B., Hudjashov, G., Nelis, M., Magi, R., Metspalu, E., Remm, M., Pitchappan, R., Singh, L., Thangaraj, K., Villems, R. & Kivisild, T. (2011) Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. *Am J Hum Genet* **89**, 731–744.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P.A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L. M., Ding, W. & Et, A. I. (1994) A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266**, 66–71.

Mohan, V., Deepa, M., Deepa, R., Shanthirani, C. S., Farooq, S., Ganesan, A. & Datta, M. (2006) Secular trends in the prevalence of diabetes and impaired glucose tolerance in urban South India–the Chennai Urban Rural Epidemiology Study (CURES-17). *Diabetologia* **49**, 1175–1178.

Myles, S., Lea, R. A., Ohashi, J., Chambers, G. K., Weiss, J. G., Hardouin, E., Engelken, J., Macartney-Coxson, D. P., Eccles, D. A., Naka, I., Kimura, R., Inaoka, T., Matsumura, Y. & Stoneking, M. (2011) Testing the thrifty gene hypothesis: The Gly482Ser variant in PPARGC1A is associated with BMI in Tongans. *BMC Med Genet* **12**, 10.

Need, A. C. & Goldstein, D. B. (2009) Next generation disparities in human genomics: Concerns and remedies. *Trends Genet* **25**, 489–494.

Neel, J.V. (1962) Diabetes mellitus: A "thrifty" genotype rendered detrimental by "progress"? *Am J Hum Genet* **14**, 353–362.

Neel, J. V. (1999) Diabetes mellitus: A "thrifty" genotype rendered detrimental by "progress"? 1962. *Bull World Health Organ* **77**, 694–703; discussion 692–693.

Nei, M. (1987) *Molecular evolutionary genetics*. New York: Columbia University Press.

Paradies, Y. C., Montoya, M. J. & Fullerton, S. M. (2007) Racialized genetics and the study of complex diseases: The thrifty genotype revisited. *Perspect Biol Med*, **50**, 203–227.

Peakall, R. & Smouse, P. E. (2006) Genalex 6: Genetic analysis in Excel. Population genetic software for teaching and research. *Mol Ecol Notes*, **6**, 288–295.

Pemberton, T. J., Mehta, N. U., Witonsky, D., Di Rienzo, A., Allayee, H., Conti, D. V. & Patel, P. I. (2008) Prevalence of common disease-associated variants in Asian Indians. *BMC Genet*, **9**, 13.

Prakash, J., Srivastava, N., Awasthi, S., Agarwal, C., Natu, S., Rajpal, N. & Mittal, B. (2012) Association of PPAR-gamma gene polymorphisms with obesity and obesity-associated phenotypes in North Indian population. *Am J Hum Biol* **24**, 454–459.

Prugnolle, F., Manica, A. & Balloux, F. (2005) Geography predicts neutral genetic diversity of human populations. *Curr Biol* **15**, R159–R160.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J. & Sham, P.C. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–575.

Quinque, D., Kittler, R., Kayser, M., Stoneking, M. & Nasidze, I. (2006) Evaluation of saliva as a source of human DNA for population and association studies. *Anal Biochem* **353**, 272–277.

R Development Core Team (2006) *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: http://www.R-project.org. Accessed 20 July 2013.

Raj, S. M., Chakraborty, R., Wang, N. & Govindaraju, D. R. (2006) Linkage disequilibria and haplotype structure of four SNPs of the interleukin 1 gene cluster in seven Asian Indian populations. *Hum Biol* **78**, 109–119.

Raj, S. M., Govindaraju, D. R. & Chakraborty, R. (2007) Genetic variation and population structure of interleukin genes among seven ethnic populations from Karnataka, India. *J Genet* **86**, 189–194.

Ramachandran, A., Snehalatha, C., Kapur, A., Vijay, V., Mohan, V., Das, A. K., Rao, P. V., Yajnik, C. S., Prasanna Kumar, K. M. & Nair, J. D. (2001) High prevalence of diabetes and impaired glucose tolerance in India: National Urban Diabetes Survey. *Diabetologia* **44**, 1094–1101.

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W. & Cavalli-Sforza, L. L. (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci USA* **102**, 15942–15947.

Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J. S., Albrechtsen, A., Moltke, I., Metspalu, M., Metspalu, E., Kivisild, T., Gupta, R., Bertalan, M., Nielsen, K., Gilbert, M. T., Wang, Y., Raghavan, M., Campos, P. F., Kamp, H. M., Wilson, A. S., Gledhill, A., Tridico, S., Bunce, M., Lorenzen, E. D., Binladen, J., Guo, X., Zhao, J., Zhang, X., Zhang, H., Li, Z., Chen, M., Orlando, L., Kristiansen, K., Bak, M., Tommerup, N., Bendixen, C., Pierre, T. L., Gronnow, B., Meldgaard, M., Andreasen, C., Fedorova, S. A., Osipova, L .P., Higham, T. F., Ramsey, C. B., Hansen, T. V., Nielsen, F. C., Crawford, M. H., Brunak, S., Sicheritz-Ponten, T., Villems, R., Nielsen, R., Krogh, A., Wang, J. & Willerslev, E. (2010) Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757–762.

Rees, S. D., Bellary, S., Britten, A. C., O'hare, J. P., Kumar, S., Barnett, A. H. & Kelly, M. A. (2008) Common variants of the TCF7L2 gene are associated with increased risk of type 2 diabetes mellitus in a UK-resident South Asian population. *BMC Med Genet* **9**, 8.

Rees, S. D., Islam, M., Hydrie, M. Z., Chaudhary, B., Bellary, S., Hashmi, S., O'hare, J. P., Kumar, S., Sanghera, D. K., Chaturvedi, N., Barnett, A. H., Shera, A. S., Weedon, M. N., Basit, A., Frayling, T. M., Kelly, M. A. & Jafar, T. H. (2011) An FTO variant is associated with Type 2 diabetes in South Asian populations after accounting for body mass index and waist circumference. *Diabet Med* **28**, 673–680.

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

Sanghera, D. K., Demirci, F. Y., Been, L., Ortega, L., Ralhan, S., Wander, G. S., Mehra, N. K., Singh, J., Aston, C. E., Mulvihill, J. J. & Kamboh, I. M. (2010) PPARG and ADIPOQ gene polymorphisms increase type 2 diabetes mellitus risk in Asian Indian Sikhs: Pro12Ala still remains as the strongest predictor. *Metabolism* **59**, 492–501.

Saxena, R., Gianniny, L., Burtt, N. P., Lyssenko, V., Giuducci, C., Sjogren, M., Florez, J. C., Almgren, P., Isomaa, B., Orho-Melander, M., Lindblad, U., Daly, M. J., Tuomi, T., Hirschhorn, J. N., Ardlie, K. G., Groop, L. C. & Altshuler, D. (2006) Common single nucleotide polymorphisms in TCF7L2 are reproducibly associated with type 2 diabetes and reduce the insulin response to glucose in nondiabetic individuals. *Diabetes* **55**, 2890–2895.

Sinnott, R.W. (1984) Virtues of the Haversine. *Sky Telesc* **68**, 159.

Southam, L., Soranzo, N., Montgomery, S. B., Frayling, T. M., Mccarthy, M. I., Barroso, I. & Zeggini, E. (2009) Is the thrifty genotype hypothesis supported by evidence based on confirmed

type 2 diabetes- and obesity-susceptibility variants? *Diabetologia* **52**, 1846–1851.

Tabassum, R., Chauhan, G., Dwivedi, O. P., Mahajan, A., Jaiswal, A., Kaur, I., Bandesh, K., Singh, T., Mathai, B. J., Pandey, Y., Chidambaram, M., Sharma, A., Chavali, S., Sengupta, S., Ramakrishnan, L., Venkatesh, P., Aggarwal, S. K., Ghosh, S., Prabhakaran, D., Srinath, R. K., Saxena, M., Banerjee, M., Mathur, S., Bhansali, A., Shah, V. N., Madhu, S. V., Marwaha, R. K., Basu, A., Scaria, V., Mccarthy, M. I., Diagram, I., Venkatesan, R., Mohan, V., Tandon, N. & Bharadwaj, D. (2012) Genome-wide association study for type 2 diabetes in Indians identifies a new susceptibility locus at 2q21. *Diabetes* **62**, 977–986.

Taylor, A. E., Sandeep, M. N., Janipalli, C. S., Giambartolomei, C., Evans, D. M., Kranthi Kumar, M. V., Vinay, D. G., Smitha, P., Gupta, V., Aruna, M., Kinra, S., Sullivan, R. M., Bowen, L., Timpson, N. J., Davey Smith, G., Dudbridge, F., Prabhakaran, D., Ben-Shlomo, Y., Reddy, K. S., Ebrahim, S. & Chandak, G. R. (2011) Associations of FTO and MC4R variants with obesity traits in Indians and the role of rural/urban environment as a possible effect modifier. *J Obes*, 2011, 307542.

The International Hapmap Consortium (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58.

Unoki, H., Takahashi, A., Kawaguchi, T., Hara, K., Horikoshi, M., Andersen, G., Ng, D. P., Holmkvist, J., Borch-Johnsen, K., Jorgensen, T., Sandbaek, A., Lauritzen, T., Hansen, T., Nurbaya, S., Tsunoda, T., Kubo, M., Babazono, T., Hirose, H., Hayashi, M., Iwamoto, Y., Kashiwagi, A., Kaku, K., Kawamori, R., Tai, E. S., Pedersen, O., Kamatani, N., Kadowaki, T., Kikkawa, R., Nakamura, Y. & Maeda, S. (2008) SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nat Genet* **40**, 1098–1102.

Vasan, S. K., Fall, T., Neville, M. J., Antonisamy, B., Fall, C. H., Geethanjali, F. S., Gu, H. F., Raghupathy, P., Samuel, P., Thomas, N., Brismar, K., Ingelsson, E. & Karpe, F. (2012) Associations of variants in FTO and near MC4R with obesity traits in South Asian Indians. *Obesity (Silver Spring)*, **20**, 2268–2277.

Vimaleswaran, K. S., Radha, V., Jayapriya, M. G., Ghosh, S., Majumder, P. P., Rao, M. R. & Mohan, V. (2010) Evidence for an association with type 2 diabetes mellitus at the PPARG locus in a South Indian population. *Metabolism* **59**, 457–462.

Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. (2006) A map of recent positive selection in the human genome. *PLoS Biol* **4**, e72.

Weir, B. S. & Cockerham, C. (1984) Estimating F-statistics for the analsyis of population structure. *Evolution* **38**, 1358–1370.

Willcox, B. J., Donlon, T. A., He, Q., Chen, R., Grove, J. S., Yano, K., Masaki, K. H., Willcox, D. C., Rodriguez, B. & Curb, J. D. (2008) FOXO3A genotype is strongly associated with human longevity. *Proc Natl Acad Sci USA* **105**, 13987–13992.

Yajnik, C. (2000) Interactions of perturbations in intrauterine growth and growth during childhood on the risk of adult-onset disease. *Proc Nutr Soc* **59**, 257–265.

Yajnik, C. S. (2004) Early life origins of insulin resistance and type 2 diabetes in India and other Asian countries. *J Nutr* **134**, 205–210.

Yajnik, C. S., Janipalli, C. S., Bhaskar, S., Kulkarni, S. R., Freathy, R. M., Prakash, S., Mani, K. R., Weedon, M. N., Kale, S. D., Deshpande, J., Krishnaveni, G. V., Veena, S. R., Fall, C. H., Mccarthy, M. I., Frayling, T. M., Hattersley, A. T. & Chandak, G. R. (2009) FTO gene variants are strongly associated with type 2 diabetes in South Asian Indians. *Diabetologia* **52**, 247–252.

Yasuda, K., Miyake, K., Horikawa, Y., Hara, K., Osawa, H., Furuta, H., Hirota, Y., Mori, H., Jonsson, A., Sato, Y., Yamagata, K., Hinokio, Y., Wang, H. Y., Tanahashi, T., Nakamura, N., Oka, Y., Iwasaki, N., Iwamoto, Y., Yamada, Y., Seino, Y., Maegawa, H., Kashiwagi, A., Takeda, J., Maeda, E., Shin, H. D., Cho, Y. M., Park, K. S., Lee, H. K., Ng, M. C., Ma, R. C., So, W. Y., Chan, J. C., Lyssenko, V., Tuomi, T., Nilsson, P., Groop, L., Kamatani, N., Sekine, A., Nakamura, Y., Yamamoto, K., Yoshida, T., Tokunaga, K., Itakura, M., Makino, H., Nanjo, K., Kadowaki, T. & Kasuga, M. (2008) Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. *Nat Genet* **40**, 1092–1097.

Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T., De Bakker, P. I., Abecasis, G. R., Almgren, P., Andersen, G., Ardlie, K., Bostrom, K. B., Bergman, R. N., Bonnycastle, L. L., Borch-Johnsen, K., Burtt, N. P., Chen, H., Chines, P. S., Daly, M. J., Deodhar, P., Ding, C. J., Doney, A. S., Duren, W. L., Elliott, K. S., Erdos, M. R., Frayling, T. M., Freathy, R. M., Gianniny, L., Grallert, H., Grarup, N., Groves, C. J., Guiducci, C., Hansen, T., Herder, C., Hitman, G. A., Hughes, T. E., Isomaa, B., Jackson, A. U., Jorgensen, T., Kong, A., Kubalanza, K., Kuruvilla, F. G., Kuusisto, J., Langenberg, C., Lango, H., Lauritzen, T., Li, Y., Lindgren, C. M., Lyssenko, V., Marvelle, A. F., Meisinger, C., Midthjell, K., Mohlke, K. L., Morken, M. A., Morris, A. D., Narisu, N., Nilsson, P., Owen, K. R., Palmer, C. N., Payne, F., Perry, J. R., Pettersen, E., Platou, C., Prokopenko, I., Qi, L., Qin, L., Rayner, N. W., Rees, M., Roix, J. J., Sandbaek, A., Shields, B., Sjogren, M., Steinthorsdottir, V., Stringham, H. M., Swift, A. J., Thorleifsson, G., Thorsteinsdottir, U., Timpson, N. J., Tuomi, T., Tuomilehto, J., Walker, M., Watanabe, R. M., Weedon, M. N., Willer, C. J., Illig, T., Hveem, K., Hu, F. B., Laakso, M., Stefansson, K., Pedersen, O., Wareham, N. J., Barroso, I., Hattersley, A. T., Collins, F. S., Groop, L., Mccarthy, M. I., Boehnke, M. & Altshuler, D. (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* **40**, 638–645.

Supporting Information

Additional supporting information may be found in the online version of this article:

Figure S1. Obesity and T2D prevalence in India.

Table S1. Deviation from Hardy-Weinberg equilibrium in India samples.

Table S2. Observed and expected heterozygosities for SNPs that showed significant deviation from HWE at a threshold of $p < 0.005$.

Table S3. Description of global populations and numbers of individuals.

Table S4. Groupings of Indian populations genotyped on the Illumina platform.

Table S5. F_{ST} differences between North and South Indian population groups, at a genome-wide level.

Table S6. Allele frequencies in global populations, grouped by geographic region.

Table S7. Allele frequencies in Indian populations we genotyped, grouped based on state of origin.

Table S8. Allele frequencies of Indian populations we genotyped, grouped based on language family.

Table S9. FST values between all populations, given the population subdivisions.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be reorganized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

Received: 22 November 2012
Accepted: 9 April 2013

