GENETIC TESTING AND MOLECULAR BIOMARKERS
Volume 14, Number 6, 2010

Pp. 781–785
DOI: 10.1089/gtmb.2010.0095

Population Structure of Aggarwals of North India as Revealed by Molecular Markers

Vipin Gupta,[1] Rajesh Khadgawat,[2] Hon Keung Tony Ng,[3] Satish Kumar,[4] Vadlamudi Raghavendra Rao,[5] and Mohinder Pal Sachdeva[5]

Using molecular genetic data on Aggarwals (Vaish/Vysya), an endogamous population group of north India, we provide evidence of its homogeneous unstratified population structure. We found the mean average heterozygosity value of 0.33 for 14 single nucleotide polymorphisms belonging to four genes (*TCF7L2*-, *HHEX*-, *KCNJ11*-, and *ADIPOQ*-) in the Aggarwal population (sample of 184 individuals) and tried to evaluate the genomic efficiency of endogamy in this population with the help of clan-based stratified analysis. We concluded that the sociocultural identity of the endogamous population groups could act as a robust proxy maker for inferring their homogeneity and population structure in India, which is ideal also for population selection for future genome-wide association studies in the country.

Introduction

The entire Indian population structure can broadly be divided along the lines of caste and tribe status, religion, and geography. The appropriate selection of a homogenous population group in genetic-based association studies is always a pertinent problem. Its solution has been seen by the geneticists to select geography along with race as the two important variables for collecting their case and control samples; some also consider belongingness to one linguistic family as one of the important indicators of homogeneity of their samples. However, in India due to its high cultural diversity, the selection of population on the basis of race, linguistic affiliations, and even geography definitely proved to be loose criteria. In India, the ethnicity, which is a sociocultural similarity of the individuals, and the geography are the most robust measurements of the population homogeneity. The recent Indian Genome Variation Consortium (2008) revealed a high degree of genetic differentiation among Indian ethnic groups and suggests that pooling of endogamous populations without regard to ethnolinguistic factors will result in false inferences in association studies. They also criticized heavily the reference to people of India as "Indian" in many population genetic studies. The implication of such a blanket usage explicitly makes the Indian population genetically homogeneous, which is factually not correct, as indicated by their study. However, in the anthropological sense they also suggested that it is possible to identify large clusters of ethnic groups that have substantial genetic homogeneity. Gene-culture co-evolution views culture as a dynamic process that can shape the material world. These models have established that cultural processes can dramatically affect the rate of change of allele frequencies as a response to selection, sometimes speeding it up and sometimes slowing it down. Human genes exposed to culturally modified selection pressures reveal extraordinarily strong selection (Laland *et al.*, 2010).

Reich *et al.* (2009) had found that the Vysya population of south India is highly homogeneous due to maintenance of the impact of the founder effect for the last 100 generations. In this article we tried to evaluate the homogeneity of the same Vysya population in north India. This article is part of a case–control genetic association study for type 2 diabetes mellitus (T2DM), and we are presenting here the population genetic analysis of its 184 control samples (Table 1) and describing its homogeneous population structure by clan-based stratified analysis of both cases and control samples (Table 2). Reich *et al.* (2009) have conducted a genome-wide study on 4 south Indian samples and in this study we have genotyped 14 single nucleotide polymorphisms (SNPs) on 184 control north Indian samples. In the parent study, we genotyped 14 markers related to T2DM in 219 cases and 184 control subjects and tried to avoid false-positive association signals due to population stratification, by selecting an endogamous population, that is Aggarwals. This work shows that we need different research designs to face methodological challenges for conducting genome-wide association studies (GWAS) in India.

[1]South Asia Network for Chronic Disease (SANCD), Public Health Foundation of India, New Delhi, India.
[2]Department of Endocrinology and Metabolism, All India Institute of Medical Sciences, New Delhi, India.
[3]Department of Statistical Science, Southern Methodist University, Dallas, Texas.
[4]Department of Genetics, Southwest Foundation for Biomedical Research, San Antonio, Texas.
[5]Department of Anthropology, University of Delhi, Delhi, India.

Table 1. Distribution of Alleles of 14 Single Nucleotide Polymorphisms and Estimates of Heterozygosity (h) and Average Heterozygosity (H) Among Aggarwals

SNPs	*Allele A*	*Allele B*	*Chi-square probability for HWE*[a]	*Total F_{IS}*	*Heterozygosity (h)*
TCF7L2	A	T			
rs4506565 (A/T)	0.6339	0.3661	0.851792	0.0111	0.4641
TCF7L2	C	T			
rs7903146 (C/T)	0.6374	0.3626	0.559647	−0.0460	0.4622
TCF7L2	G	T			
rs12256372 (G/T)	0.6489	0.3511	0.088860	0.1246	0.4557
HHEX	T	C			
rs1111875 (T/C)	0.6723	0.3277	0.081331	0.1281	0.4406
HHEX	C	T			
rs5015480 (C/T)	0.3408	0.6592	0.268315	0.0799	0.4493
KCNJ11	C	T			
rs5219 (C/T)	0.4777	0.5223	0.113312	0.1156	0.4490
ADIPOQ					
rs1681194 (A/G)	0.9855	0.0145	0.029848	−0.0147	0.0286
ADIPOQ	G	A			
rs17300539 (G/A)	0.9913	0.0087	0.925053	−0.0088	0.0172
ADIPOQ	C	G			
rs266729 (C/G)	0.6753	0.3247	0.541690	0.0433	0.4385
ADIPOQ	C	T			
188042502 (C/T)	0.9500	0.0500	0.321037	0.0712	0.0950
ADIPOQ	T	G			
rs2241766 (T/G)	0.9083	0.0917	0.182990	−0.1009	0.1666
ADIPOQ	G	T			
rs1501299 (G/T)	0.8588	0.1412	0.340205	0.0684	0.2425
ADIPOQ	G	A			
rs182052 (G/A)	0.6308	0.3692	0.583592	0.0389	0.4658
ADIPOQ	C	T			
rs62292784 (C/T)	0.6599	0.3401	0.274290	0.0804	0.4489
Average Heterozygosity (*H*)					0.3303(SD = 0.1786)

[a]Level of significance: 0.05.
HWE, Hardy–Weinberg equilibrium; F_{IS}, Wright's fixation index; SD, standard deviation; SNP, single nucleotide polymorphism.

Materials and Methods

Study subjects and phenotypic definitions

Two hundred nineteen unrelated cases and 184 control subjects belonging to the Aggarwal population (belong to the Vysya/Vaish category in the *Varna* system of caste) of Delhi were recruited for this study from the All India Institute of Medical Sciences (AIIMS) and Maharaja Agresen Hospital, New Delhi. This is the first study in which population was selected on the basis of five criteria: (i) race (Caucasian), (ii) linguistic affinity (Indo-European family), (iii) geography (National Capital Region of Delhi), (iv) geographic ancestry (ancestors of both cases and control subjects belong to Haryana and Rajasthan states of India), and (v) ethnicity (the Aggarwals are an endogamous caste population marrying among 18 existing clans). Because of the uniqueness and stringency of criteria for sample selection, our field work took more than 18 months. We recruited subjects on the basis of our inclusion criteria on the basis of WHO guidelines for the diagnosis of T2DM (WHO, 1999). Subjects were recruited after ethics clearance from AIIMS New Delhi.

Genotyping

Genotypes were obtained after direct sequencing (using ABI-3730 Genetic Analyzer; Applied Biosystem) of purified polymerase chain reaction product (primers available on request) of nine genomic regions belonging to four genes (*TCF7L2*-3SNPs, *HHEX*-2SNPs, *KCNJ11*-1SNP, and *ADIPOQ*-8SNPs). Allele frequencies, Hardy–Weinberg equilibrium, heterozygosities (Nei, 1973), and Wright's fixation index at the respective locus were estimated using POPGENE software (Yeh and Young, 1999). HAPLOVIEW was used for haplotype analysis and to estimate linkage disequilibrium (LD) (Barrett *et al.*, 2005). For studying population stratification, a clan-based stratified analysis was done for testing the homogeneity of odds ratio across different clans using Breslow–Day Test with Tarone's correction.

Results and Discussion

Population stratification refers to differences in allele frequencies between cases and control subjects due to systematic differences in ancestry rather than association of genes with a disease. In general, stratification exists when the total population has been formed by admixture between subpopulations and when admixture proportions (defined as the proportions of the genome that have ancestry from each subpopulation) vary between individuals. In the light of huge sociocultural diversity influencing mating pattern, the selection of a homogenous population with defined genetic structure, to overcome population stratification, for genetic association studies is a challenge in India. Here we have tried to show that population selection on the basis of sociocultural boundaries

Table 2. Clan-Based Stratified Analysis of Five Single Nucleotide Polymorphisms for Inferring Population Stratification

							Test for homogeneity of OR (B-D statistic with Tarone's correction)		*CMH test (p-Value)*	*Common OR [95% CI]*
Clan →	*Garg*	*Goyal*	*Singhal*	*Bansal*	*Mittal*	*Total*	*Statistic*	p-*Value*		
rs4506565										
Cs	54	55	23	16	18	166	3.896	0.4202	4.013 (0.04516)	1.616 [1.011–2.584]
Cn	50	37	18	16	13	134				
Total	104	92	41	32	32	300				
rs7903146										
Cs	54	56	24	16	19	169	7.21	0.1252	3.708 (0.05415)	1.848 [0.9858–3.464]
Cn	49	37	19	19	13	137				
Total	103	93	43	35	32	302				
rs1111875										
Cs	51	55	23	16	15	160	2.2	0.699	1.313 (0.2518)	1.316 [0.8242–2.102]
Cn	48	36	18	17	13	132				
Total	99	91	41	33	28	292				
rs5015480										
Cs	54	57	24	15	17	167	7.512	0.1112	1.52 (0.2177)	0.7498 [0.4738–1.187]
Cn	49	35	19	19	13	135				
Total	103	92	43	34	30	302				
rs5219										
Cs	53	51	24	16	16	160	4.306	0.3622	0.184 (0.6702)	0.8992 [0.5519–1.465]
Cn	48	36	18	19	13	134				
Total	101	87	42	35	29	294				

Selection of clans: Clans that have case–control sample >30 were selected for stratified analysis; therefore, only five clans were able to pass this criterion. None of the SNP showed significant results ($p < 0.05$).

Cs, case; Cn, control; OR, odds ratio; B-D test, Breslow–Day test (with Tarone's correction); CMH, Cochran–Mantel–Haenszel test; CI, confidence interval.

can act as a robust proxy for the genetic homogeneity of a population in India. Thus, we selected the Aggarwal population of north India, which is divided into 18 clans (Gotras), following the marital rule of clan exogamy (marrying between clans) and caste endogamy (marrying within caste). It would be true to say that 5.5% violations of the rule would mean that endogamy was not existent. Although violations are now definitely much more than this value, still the endogamy nature is preserved by and large (Channa, 1979).

Recently, the Indian population genetic history was comprehensively evaluated by the landmark study of Reich *et al.* (2009). They analyzed the Indian population structure at genome-wide scale and found that strong endogamy must have shaped marriage pattern in India for thousands of years. For instance, they detected the founder effect in their studied four individuals of Vysya group of south India (another name of Aggarwal population of north India) at more than 100 generations ago. Here in the same population (but in north India), we found the mean average heterozygosity of 0.33 (standard deviation of ± 0.1786) for all the 14 markers, and this moderately low heterozygosity among Aggarwal population clearly supports Reich *et al.* (2009) for the founder events of the Vysya population of south India. Saraswathy *et al.* (2010) studied the same population for the *DRD2* locus and found a similar average heterozygosity of 0.381. Finally, our clan-based stratified analysis on the basis of Breslow–Day test for homogeneity of odds ratio clearly provides empirical evidence that the Aggarwal is a homogenous population group and is not genetically stratified on the basis of its existing clans (Table 2).

The study of Helgason *et al.* (2007) accounted the frequency of around 2% for rs7903146 (risk allele) of the *TCF7L2* gene in East Asian populations (East Asian HapMap group) and suggested a kind of selection sweep by their haplotype A (HapA, containing wild allele on rs7903146), almost to fixation in East Asian populations. In contrast to this, due to the endogamous nature of the population we found relatively high frequencies of risk alleles for all of the three genotyped SNPs of *TCF7L2* (allele frequencies of rs4506565, rs7903146, and rs12256372 were 0.37, 0.37, and 0.37, respectively, in the homogeneous Aggarwal community of the heterogeneous Indo-European linguistic group). Moreover, the frequencies reported here are consistently higher than the earlier reported frequencies of 0.29, 0.29, and 0.22, respectively, in the whole Indo-European linguistic group (Chandak *et al.*, 2007). Further, Chandak *et al.* (2007) also suggested that SNP rs4506565 was the best proxy for other two SNPs of *TCF7L2* (rs7903146 and rs12256372) in the HapMap II data, but the LD pattern of these SNPs in the present Aggarwal population found rs7903146 as the best proxy for the remaining two SNPs (the desirable D' values of 0.82 and 0.82 between SNPs rs7903146 and rs12256372 and rs4506565 and rs7903146, respectively, suggest LD, and undesirable D' value between rs4506565 and rs12256372 was 0.67, suggesting a low LD if any). We also found the fully mutated haplotype of the *TCF7L2* gene with a frequency of 25.6% (Table 3). We found a high frequency (48%) of minor allele (T) of SNP rs5210 (KCNJ11) among Aggarwals in comparison to 38% among the Khatris of Punjab (Sanghera *et al.*, 2008). Similarly, for SNPs rs1111875 (C) and rs5015480 (T) on the *HHEX* gene we found a high allele frequency of 33% and 66%, respectively; in contrast, Wu *et al.* (2008) found the frequency of 18% and 30%, respectively, in Beijing and 27% and 13.8%, respectively, in Shanghai. Further, the frequencies of mutated haplotypes of the *ADIPOQ* and

Table 3. Haplotype Frequencies of *TCF7L2*, *ADIPOQ*, and *HHEX* Genes Among 184 Control Aggarwal Samples

Haplotype	*Frequency*
TCF7L2 gene	
ACG	0.553
TTT	0.256
TTG	0.050
ATT	0.047
TCG	0.044
ACT	0.035
ADIPOQ gene	
AGCCGCTG	0.419
AGGCATTG	0.274
AGCCGCTT	0.086
AGGCATGG	0.047
AGCTGCTG	0.046
AGCCGCGG	0.047
AGGCATTT	0.016
AGCCATTG	0.014
HHEX gene	
TT	0.583
CC	0.254
TC	0.090
CT	0.073

Order of SNPs in haplotypes (TCF7L2): rs4506565; rs7903146; rs12256372; SNPs (ADIPOQ): rs16861194; rs17300539; rs266729; 188042502; rs182052; rs62292784; rs22241766; rs1501299; SNPs (HHEX): rs1111875; rs5015480.

HHEX genes are 1.4%–27% and 7%–25%, respectively (Table 3). These comparisons showed the power of endogamy-based population structure in India.

The methodological challenges faced by Teo *et al.* (2010) in conducting GWAS in Africa suggested that it has a genetically more diverse population than European and Asian. This is because groups migrating out of Africa experienced severe population bottlenecks, resulting in a reduction of genetic diversity in descendant populations (Campbell and Tishkoff, 2008). In Africa low levels of LD reduce the likelihood that a causal variant will have a sufficient level of correlation with nearby SNPs to show significant genotype–phenotype associations unless it is directly typed (Teo *et al.*, 2010). In India, however, endogamy has managed strong LD due to a high degree of allele frequency differentiation in Indian populations that is at least as large as that between northern and southern Europeans (Reich *et al.*, 2009). This will not reduce the problem because as in Africa different study sites may have different allele frequencies and/or patterns of LD (Teo *et al.*, 2010). In fact, diversity in Indian endogamous groups also showed a significantly different LD pattern (Indian Genome Variation Consortium, 2008; Reich *et al.*, 2009), which reduced the likelihood of reproducing associations in multicentric studies. For an African population it might be relatively easy to localize the causal variant because it is in weak LD in neighbouring SNPs and will therefore stand out at the peak of the association signal, but in Indian endogamous population groups it would be as difficult as in an European population due to strong LD. In India, nonendogamous populations like Scheduled Castes, due to their socially unrestricted mating pattern, might pose similar methodological problems as faced by Teo *et al.* (2010) in Africa.

The Aggarwal population was 2,718,390, according to the Census of India 1911 (Channa, 1979), and now after almost 100 years (i.e., four generations) we expect its population size in India might raise to10–15 million or even more. Therefore, the selection of case–control samples with desirable power must not be a problem. Anthropologically, the majority of the Indian population is composed of >4000 endogamous caste groups constituting about three-fourths of its population. Most importantly, the caste endogamy and geography are the only substantive basis of estimating marital distances between populations in rural an urban India, and any other criteria will play putative role. Both large studies, the Indian Genome Variation Consortium (2008) and the study by Reich *et al.* (2009), explicitly confirm the conventional assertions of anthropologists, that is, genomic efficiency of endogamy, and this also strengthened the utility of anthropological models (defined ethnicity by culture rather than race) of studying populations for the Indian subcontinent. Further, the social properties of caste-like endogamy (which effectively maintains caste boundaries), commensal restrictions, hierarchal arrangements, pollution by various forms of contact, and association with a traditional occupation coupled with exchange of goods and services in a relatively stable patron–client relationship (Channa, 1979) provide an excellent research model for studying gene-culture co-evolution. For instance, the human amylase gene is a clear example of cultural variation in human diet explaining some of the adaptive genetic differences between human populations (Laland *et al.*, 2010). We conclude that for conducting GWAS, geneticists must respect the endogamous nature of the Indian population structure while designing their epidemiological studies.

Acknowledgments

This study was supported by grants from Indian Council of Medical Research and Anthropological Survey of India, South Asia Network for Chronic Disease (New Delhi), Maharaja Agresen Hospital (New Delhi), and AIIMS (New Delhi).

Disclosure Statement

No competing financial interests exist.

References

Barrett JC, Fry B, Maller J, *et al.* (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265.

Campbell MC, Tishkoff SA (2008) African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genomics Hum Genet 9:403–433.

Chandak GR, Janipalli CS, Bhaskar S (2007) Common variants in the TCF7L2 gene are strongly associated with type 2 diabetes mellitus in the Indian population. Diabetologia 50:63–67.

Channa VC (1979) Caste: Identity and Continuity. B. R. Publishing Corporation, New Delhi.

Helgason A, Pa'lsson S, Thorleifsson G, *et al.* (2007) Refining the impact of TCF7L2 gene variant on type 2 diabetes and adaptive evolution. Nat Genet 39:218–225.

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

Laland KN, Odling-Smee J, Myles S (2010) How culture shaped the human genome: bringing genetics and the human sciences together. Nat Rev Genet 11:137–148.

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 70:3321–3323.

Reich D, Thangaraj K, Patterson N, *et al.* (2009) Reconstructing Indian population history. Nature 461:489–494.

Sanghera DK, Ortega L, Han S, *et al.* (2008) Impact of nine common type 2 diabetes risk polymorphisms in Asian Indian Sikhs: PPARG2 (Pro12Ala), IGF2BP2, TCF7L2 and FTO variants confer a significant risk. BMC Med Genet 9:59.

Saraswathy KN, Meitei SY, Gupta V, *et al.* (2010) Brief communication: Allelic and haplotypic structure at the DRD2 locus among five North Indian caste populations. Am J Phys Anthropol 141:651–657.

Teo Y, Small KS, Kwiatkowski DP (2010) Methodological challenges of genome-wide association analysis in Africa. Nat Rev Genet 11:149–160.

WHO (1999) Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications. Report of a WHO Consultation. Part I. Diagnosis and Classification of Diabetes Mellitus. World Health Organization, Geneva.

Wu Y, Li H, Loos RJF, *et al.* (2008) Common variants in CDKAL1, CDKN2A/B, IGF2BP2, SLC30A8, and HHEX/IDE genes are associated with type 2 diabetes and impaired fasting glucose in a Chinese Han population. Diabetes 57:2834–2842.

Yeh F, Yang R (1999) POPGENE. Microsoft Window-based freeware for population genetic analysis Version 1.31. University of Alberta, Edmonton, Alberta. Available at www.ualberta.ca/~fyeh/download.htm.

Address correspondence to:
Vadlamudi Raghavendra Rao, Ph.D.
Department of Anthropology
University of Delhi
Delhi 110 007
India

E-mail: vrraoasi@gmail.com