Genetic variations in the Dravidian population of South West coast of India: Implications in designing case-control studies

Anitha D'Cunha, Lekha Pandit & Chaithra Malli

Center for Advanced Neurological Research, KS Hegde Medical Academy, Nitte University, Mangaluru, India

Background & objectives: Indian data have been largely missing from genome-wide databases that provide information on genetic variations in different populations. This hinders association studies for complex disorders in India. This study was aimed to determine whether the complex genetic structure and endogamy among Indians could potentially influence the design of case-control studies for autoimmune disorders in the south Indian population.

Methods: A total of 12 single nucleotide variations (SNVs) related to genes associated with autoimmune disorders were genotyped in 370 healthy individuals belonging to six different caste groups in southern India. Allele frequencies were estimated; genetic divergence and phylogenetic relationship within the various caste groups and other HapMap populations were ascertained.

Results: Allele frequencies for all genotyped SNVs did not vary significantly among the different groups studied. Wright's F_{ST} was 0.001 per cent among study population and 0.38 per cent when compared with Gujarati in Houston (GIH) population on HapMap data. The analysis of molecular variance results showed a 97 per cent variation attributable to differences within the study population and <1 per cent variation due to differences between castes. Phylogenetic analysis showed a separation of Dravidian population from other HapMap populations and particularly from GIH population.

Interpretation & conclusions: Despite the complex genetic origins of the Indian population, our study indicated a low level of genetic differentiation among Dravidian language-speaking people of south India. Case-control studies of association among Dravidians of south India may not require stratification based on language and caste.

Key words Case-control studies - Dravidian population - single nucleotide variations

Single nucleotide variations (SNVs) are the most common DNA variations in the human genome. Genome-wide association studies (GWASs) have been performed to identify risk alleles attributable to complex disease traits using subsets of common SNVs seen among European populations. Most common SNVs are likely to be found in most populations, but allele frequencies will vary. The transferability of this information to non-Europeans with inherent differences in linkage disequilibrium patterns is not clearly established. Studies on non-European populations, particularly Indians and African-Americans, suggest that disease-related SNVs previously validated in European populations may be similar[1,2].

A case-control model is vital to the conduct of these studies. In White European population lacking significant genetic heterogeneity, it is possible to use common controls. The best example is the Wellcome Trust Case-Control Consortium which for the first time used shared controls effectively in GWASs[3]. This is not feasible in India which has rigid socio-cultural practices that allow division along caste and tribal and religious lines. There is also inbuilt endogamy which complicates the genetic origin[4]. The lack of genomic databases that have incorporated data from Indian populations is a serious drawback to designing association studies for complex disorders[5,6]. Indian population is represented in the Gujarati in Houston (GIH) subset (101 healthy Indians) in the HapMap 3[7]. Recent studies indicate that these data may not be representative for Indians on whole, particularly for the Dravidian population of southern India[8]. In view of the complex genetic structure and diversity, it is likely that while designing case-control studies in India, it is important to include factors related to caste, language and geography[9,10]. The present study was conducted to evaluate some SNVs related to immune response in a carefully selected group of healthy individuals from southern India to understand if any of these factors were important in the design of case-control studies in the Indian scenario.

Material & Methods

The present study was conducted at the Center for Advanced Neurological Research, KS Hegde Medical Academy, Nitte University, Mangalore, Karnataka, India, between August 2012 and December 2015. Healthy individuals selected for this study were from the control arm of an ongoing study on genetic susceptibility to multiple sclerosis (MS) in Indian population[1,11]. They were individuals matched for ethnicity, language, caste and area of living with individual MS patients. Healthy volunteers were from within a 500 km[2] radius of the investigation site based in the city of Mangalore in Karnataka. They belonged to the coastal towns of two adjoining southern States of Karnataka and Kerala in India.

In total 370 individuals were included in the study, and six prominent caste groups in the region were selected. From coastal Karnataka, there were Bunts (52) and Brahmins (50); from coastal Kerala, there were Thiyya (59) and Nairs (57). The Christians (52) and Malabar Muslims (100) from both Kerala and Karnataka were considered together. Through

a previously validated screening questionnaire[12], individuals with diabetes, bronchial asthma, arthritic disorders and thyroid disease were excluded. Individuals from known mixed marriages were also excluded from this study. Written informed consent was taken from all participants. The study protocol was approved by the Institutional Ethics Committee.

DNA isolation and quantification: Blood samples from 370 healthy individuals were collected in EDTA vials. Genomic DNA was extracted from the collected blood samples by conventional salting out technique[13]. After extraction, the DNA was quantified on a NANO DROP 2000 Spectrophotometer (Thermo Fisher Scientific Inc., USA) to determine the concentration, and its purity was examined using standard A260/A280 ratio. DNA samples were normalized to 5 ng/µl.

Single nucleotide variation (SNV) selection and genotyping: Twelve SNVs were selected[1] for genotyping of samples. SNV genotyping was performed using pre-designed TaqMan SNP genotyping assays (Applied Biosystems, Foster City, CA, USA). DNA amplification was carried out on an ABI7500 Real Time PCR genotyping platform. Four microlitres of normalized genomic DNA (5 ng/µl) was aliquoted into MicroAmp® Optical 96-Well Reaction Plates (Applied Biosystems). The DNA samples were dried down completely. Each PCR contained 20 ng DNA, 12.5 µl TaqMan Universal PCR Master Mix (2×), 1.25 µl single nucleotide polymorphism (SNP) genotyping assay (20×)[1] and 11.25 µl DNase-free water (Applied Biosystems). The PCR conditions were as follows: 60°C for one min, 95°C for 10 min, followed by 40 cycles of 95°C for 15 sec and 60°C for one min. After PCR amplification, endpoint plate read was performed on an Applied Biosystems 7500 Real-Time PCR System using Sequence Detection System software v 1.4.1.

Statistical analysis: Hardy-Weinberg equilibrium (HWE) and allele frequencies were determined. Wright's F_{ST} was performed to look at variation in allele frequency between populations. Analysis of molecular variance (AMOVA) was done to investigate the proportion of genetic variations within and among studied populations. The phylogenetic relationships were estimated with multidimensional scaling (MDS) analysis. HWE and Wright's F_{ST} were determined using Arlequin Program v 3.0[14]. Allele frequency, MDS analysis (after calculating Nei's[15] standard genetic distance using GDA software[16]) and comparison of

Table. Minor allele frequencies of 12 single nucleotide variations (SNV) related to autoimmune diseases

Gene name	Minor allele	SNV ID	Christian (n=52)	Muslim (n=100)	Thiyya (n=59)	Nair (n=57)	Bunt (n=52)	Brahmin (n=50)	*P*
IL7R	T	rs6897932	0.23	0.23	0.26	0.23	0.29	0.22	0.85
CLEC16A	G	rs12708716	0.37	0.47	0.44	0.42	0.35	0.37	0.28
CD226	C	rs763361	0.46	0.51	0.45	0.5	0.36	0.52	0.19
CLEC16A	G	rs6498169	0.23	0.21	0.21	0.25	0.33	0.28	0.25
IL2RA	C	rs2104286	0.17	0.17	0.22	0.23	0.15	0.16	0.57
CD6	G	rs17824933	0.02	0.05	0.04	0.03	0.08	0.07	0.54
RGS1	A	rs2760524	0.096	0.045	0.11	0.08	0.07	0.05	0.26
TMEM39A	T	rs1132200	0.19	0.16	0.15	0.096	0.13	0.18	0.47
CD58	G	rs2300747	0.33	0.35	0.34	0.34	0.41	0.25	0.28
MPHOSP9	G	rs1790100	0.22	0.27	0.22	0.25	0.15	0.26	0.31
KIF21B	A	rs12122721	0.1	0.18	0.14	0.14	0.21	0.11	0.15
IL12A	C	rs4680534	0.16	0.13	0.14	0.15	0.16	0.16	0.98

minor allele frequency (MAF) of Dravidian with HapMap population were done using linear-by-linear association Chi-square test in SPSS software v 20.0 (SPSS, Chicago, IL, USA).

Results

A total of 370 healthy individuals (male=171, female=199, mean age 37.4±11.90 yr) were genotyped. Among those genotyped, SNVs typed 10 were in HWE. Two SNVs (rs2760524 & rs1132200) showed a deviation which was <0.001 and were included after excluding typing error. The distribution of MAFs of all 12 gene-associated SNVs included in this study is summarized in the Table. There was no significant difference in MAF between the six different caste groups included in our study. The average Wright's F_{ST} over all loci among the six caste groups was 0.001 per cent. When all six caste groups were clumped together and compared with GIH population, the average F_{ST} was 0.38 per cent.

The AMOVA was used to study the genetic differentiation among studied population. The AMOVA results showed that irrespective of any grouping; about 97 per cent variation was attributable to differences within the studied populations. There was only a minimal variation of less than one per cent among different caste groups.

Comparison of MAF of Dravidian with HapMap population showed that majority of SNVs tested were significantly (*P*<0.001) different between Dravidian and HapMap population except for *CLEC16A* (rs12708716) and *RGS1* (rs2760524).
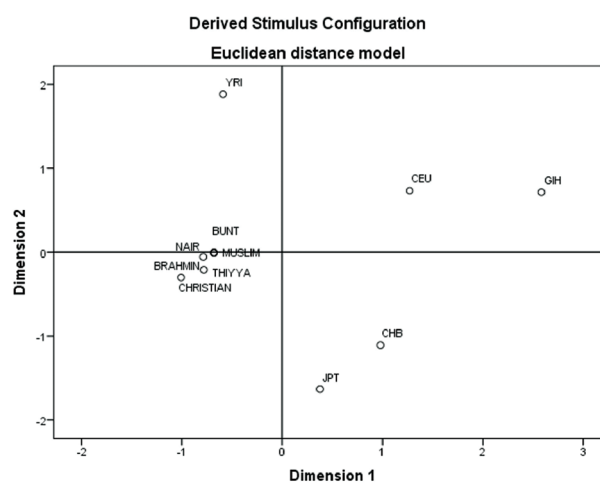

Figure. Multidimensional scaling plot depicting genetic distances among the studied population with HapMap populations.

To understand the phylogenetic relationship between our data and world populations, the present genotyped data were compared with HapMap Phase 3 populations [Yoruba people of Ibadan, Nigeria; Utah residents with Northern and Western European ancestry from the CEPH collection (CEU); Han Chinese from Beijing; Japanese from Tokyo and GIH, USA]. A MDS analysis was performed. Separation between the two Indian populations (our study populations and GIH populations) was evident in MDS plot (Figure). We observed that the south Indian Dravidian language-speaking population was clustered separately from rest of the HapMap population in both first and second dimensions. The GIH population clearly aligned closer to CEU population.

Discussion

Within India, two broad sub-groups exist, namely the Indo-European language-speaking (represented in the GIH population in the HapMap) and the south Indian Dravidian language-speaking populations[8]. Genetically, the ancestral North Indian (ANI) who had affiliations with the Middle Eastern and Europeans populations is distinct from the ancestral South Indian (ASI). The present-day populations in India are a mixture of both and have ANI ancestry of 39-71 per cent[10]. Allele frequencies vary among populations within India[17] and are much higher than reported in Europeans[18]. The difference in allele frequencies may be due to many common founder effects and limited gene flow.

In our previous study[1], genetic susceptibility for MS was evaluated among Indians using a case-control model. Twelve SNVs (analysis of risk allele frequency, which is most often the major allele frequency) were genotyped in this study, which were previously validated through GWASs and shown to be associated with MS among White Europeans[19-22]. Our study[1] and others[2] have shown that irrespective of ethnic diversity, there is remarkable overlap in risk allele frequencies among the populations studied and implicated genes associated with immune response.

The MAFs of all 12 SNVs included in this study showed no significant difference between the six different caste groups studied. Wright's F_{ST} showed that the genetic differentiation between the various subgroups was very low and similar to the previously published studies on Dravidian populations[17]. However, we observed a large variation when we compared our study population with GIH population represented in the HapMap. A similar result was obtained in another study that compared Dravidian language-speaking Tamil Indians (INS) from the Singapore Genome Variation Project with GIH population[8]. The AMOVA results showed that irrespective of any grouping, about 97 per cent variation was attributable to differences within the study population. There was a genetic divergence of the Dravidian population from other HapMap populations including GIH population. The Dravidian language-speaking individuals included in this study segregated distinctly into a tight cluster away from Indo-Aryan language speakers including GIH and CEU populations. In the MDS plot, the GIH population was aligned close to CEU population and supported the existing view of European influences in the genetic composition of north Indian populations.

In conclusion, our study suggested that Dravidian language-speaking populations of south India did not show significant patterns of genetic differentiation. Among non-tribal populations of south India, evaluation of risk alleles associated with disease may not require stratification for language or caste, while the same may not hold good for studies that involve mixed populations from the north and south of the country.

Our study being preliminary in nature had a limitation of small sample size. Only a few subsets of Dravidians from coastal Karnataka and Kerala were included. A larger sample size is required representing all Dravidian subsets in the region in the future to confirm these finding.

Conflicts of Interest: None.

References

1. Pandit L, Ban M, Sawcer S, Singhal B, Nair S, Radhakrishnan K, *et al.* Evaluation of the established non-MHC multiple sclerosis loci in an Indian population. *Mult Scler* 2011; *17* : 139-43.

2. Isobe N, Madireddy L, Khankhanian P, Matsushita T, Caillier SJ, Moré JM, *et al.* An ImmunoChip study of multiple sclerosis risk in African Americans. *Brain* 2015; *138* (Pt 6) : 1518-30.

3. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007; *447* : 661-78.

4. Tamang R, Singh L, Thangaraj K. Complex genetic origin of Indian populations and its implications. *J Biosci* 2012; *37* : 911-9.

5. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* 2005; *307* : 1072-9.

6. Packer BR, Yeager M, Burdett L, Welch R, Beerman M, Qi L, *et al.* SNP500Cancer: A public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. *Nucleic Acids Res* 2006; *34* : D617-21.

7. International HapMap3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* 2010; *467* : 52-8.

8. Ali M, Liu X, Pillai EN, Chen P, Khor CC, Ong RT, *et al.* Characterizing the genetic differences between two distinct migrant groups from Indo-European and Dravidian speaking populations in India. *BMC Genet* 2014; *15* : 86.

9. Indian Genome Variation Consortium. Genetic landscape of the people of India: A canvas for disease gene exploration. *J Genet* 2008; *87* : 3-20.

10. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. *Nature* 2009; *461* : 489-94.

11. Pandit L, Malli C, Singhal B, Wason J, Malik O, Sawcer S, *et al.* HLA associations in South Asian multiple sclerosis. *Mult Scler* 2016; *22* : 19-24.

12. Malli C, Pandit L, D'Cunha A, Mustafa S. Environmental factors related to multiple sclerosis in Indian population. *PLoS One* 2015; *10* : e0124064.

13. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* 1988; *16* : 1215.

14. Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evol Bioinform Online* 2007; *1* : 47-50.

15. Nei M. *Molecular evolutionary genetics*. New York: Columbia University Press; 1987.

16. Lewis PO, Zaykin D. Genetic Data Analysis: Computer Program for the Analysis of Allelic Data. Release 1.1. Connecticut, United States of America: Department of Ecology and Evolution, University of Connecticut; 2002.

17. Rosenberg NA, Mahajan S, Gonzalez-Quevedo C, Blum MG, Nino-Rosales L, Ninis V, *et al.* Low levels of genetic divergence across geographically and linguistically diverse populations from India. *PLoS Genet* 2006; *2* : e215.

18. Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, *et al.* Correlation between genetic and geographic structure in Europe. *Curr Biol* 2008; *18* : 1241-8.

19. International Multiple Sclerosis Genetics Consortium (IMSGC). Risk alleles for multiple sclerosis identified by a genomewide study. *N Engl J Med* 2007; *357* : 851-62.

20. De Jager PL, Jia X, Wang J, de Bakker PIW, Ottoboni L, Aggarwal NT, *et al*. Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSFIA as new multiple sclerosis susceptibility loci. *Nat Genet* 2009; *41* : 776-82.

21. De Jager PL, Baecher-Allan C, Maier LM, Arthur AT, Ottoboni L, Barcellos L, *et al*. The role of the CD58 locus in multiple sclerosis. *Proc Natl Acad Sci U S A* 2009; *106* : 5264-9.

22. International Multiple Sclerosis Genetics Consortium (IMSGC). Comprehensive follow-up of the first genome-wide association study of multiple sclerosis identifies KIF21B and TMEM39A as susceptibility loci. *Hum Mol Genet* 2010; *19* : 953-62.

Reprint requests: Dr Lekha Pandit, KS Hegde Medical Academy, Nitte University, Mangaluru 575 018, Karnataka, India
e-mail: panditmng@gmail.com