



Journal of Experimental Biology and Agricultural Sciences

<http://www.jebas.org>

ISSN No. 2320 – 8694

Ancestry Specific variation in neuropsychological disorders among the South Asian population

Swathy Krishna M¹, Gaana Rukmini SM², Nishmitha K², Ranajit Das^{1*}

¹Yenepoya Research Centre, Yenepoya (Deemed to be) University, Mangalore, 575018, Karnataka.

²Department of Biosciences, Mangalore University, Mangalore, 575008, Karnataka.

Received – July 31, 2021; Revision – September 16, 2021; Accepted – November 28, 2021

Available Online – February 28, 2022

DOI: [http://dx.doi.org/10.18006/2022.10\(1\).248.253](http://dx.doi.org/10.18006/2022.10(1).248.253)

KEYWORDS

Neuropsychological disorders

Population-specific disorder

Indian ancestry

Indus valley civilization connection

Disease based clustering

ABSTRACT

The enormous genetic diversity in South Asia resulting from a long and complex admixture history resulted in the emergence of variation in various traits and variations in disease susceptibility. Neuropsychological disorders are one such example that shows variation at the population level. In this study, we aimed at understanding the variation in neuropsychological disorders at the population level among South Asian populations by curating, comparing and contrasting single nucleotide polymorphisms (SNPs), known to be associated with the same. Whole-genome data comprising of 1662 South Asians, belonging to 241 distinct populations were obtained from the database of Dr. David Reich, Harvard Medical School, USA. Principal Component Analysis (PCA) revealed that the Ancestral Tibeto Burman (ATB) genomes form a distinct and distinguishable cluster for the SNPs known to be associated with neuropsychological disorders. Identical By Descent (IBD) analysis showed that out of the top seven populations in terms of IBD sharing, six are from Southern India indicating that these populations may have undergone a recent selective sweep for these SNPs. Further, out of the top ten genomes, according to the number of genomes fixed for the minor alleles, seven were from Southern India. Furthermore, several indigenous populations from South India depicted high F values (>0.25) for SNPs associated with neuropsychological disorders, indicating higher susceptibility for neuropsychological disorders among these South Indian populations. Interestingly, we found that most of the SNPs, fixed for the alternative alleles, were also found to be fixed among the ancient genomes from Indus Valley Civilization (IVC), indicating that these SNPs likely got transmitted to various modern-day South Indian populations from IVC.

* Corresponding author

E-mail: das.ranjit@gmail.com (Ranjit Das)

Peer review under responsibility of Journal of Experimental Biology and Agricultural Sciences.

Production and Hosting by Horizon Publisher India [HPI]
(<http://www.horizonpublisherindia.in/>).
All rights reserved.

All the articles published by [Journal of Experimental Biology and Agricultural Sciences](#) are licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](#) Based on a work at www.jebas.org.



1 Introduction

South Asia, comprised of India, Sri Lanka, Pakistan, Bangladesh, Bhutan, and Nepal has a millennia-old history of mixing of gene-pools from various parts of the world including Southeast Asia, West Eurasia, and a hunter-gatherer lineage from South Asia who likely arrived here via 'southern exit route' from Africa (Shinde et al. 2019). These South Asian hunter-gatherers, referred to as Ancient Ancestral South Indians (AASI), have genetic similarities with present-day Andamanese people. Modern-day South Asian genomes largely belong to four ancestries: Ancestral South Indian (ASI), Ancestral North Indian (ANI), Ancestral Austro-Asiatic (AAA), and Ancestral Tibeto-Burman (ATB) comprising of various combinations of West Eurasian, Southeast Asian, and AASI ancestry fractions (Basu et al. 2016). The enormous genetic diversity has been maintained by the age-old practice of endogamy (Nakatsuka et al. 2017). Genetic admixture among various populations across the world, followed by the practice of endogamy has resulted in the emergence of variation in susceptibility towards various diseases and conditions among South Asian populations.

Neuropsychological disorders are one of the notable examples that show population-specific variation. More than 100 million people are suffering from various psychological disorders such as clinical depression and anxiety disorders in India (Sagar et al. 2020). Over the last decade, several studies have shown population-specific variation in neuropsychological disorders among South Asian populations. Dutta (2013) aimed at understanding the disproportionately higher number of cases of alcoholic cirrhosis among Indians compared to the Caucasian populations. He speculated the association of the 10109G allele of *PNPLA3* with the susceptibility towards alcoholic cirrhosis among North Indian populations.

While a handful of small-scale studies revealed that substance abuse is highly prevalent among Northeast Indians (Medhi et al. 2006; Mahanta et al. 2016), the association between the intrinsic genetic makeup and the substance use among the people from Northeastern states has remained understudied so far. A recent public health survey revealed that substance consumption and abuse are discernibly higher among males from Northeast India compared to the other parts of the country (Saikia and Debbarma 2020). The higher instances of substance abuse among these people have been linked to socioeconomic factors such as poverty and illiteracy. Another survey across the states of Northeast India depicted that >50% of adults consume some forms of tobacco in Meghalaya, Manipur, Tripura, and Mizoram (Ladusingh et al. 2017). Further, an increase in the abuse of illegal substances such as opium and heroin in many parts of Northeast India has also been reported despite strict legal restrictions (Chaturvedi 2003). Further, a cross-sectional study on samples from Global Adult Tobacco Survey in India, took place

between 2009 and 10, comprising of more than 69,000 individuals, revealed that while the smoking prevalence is highest in Meghalaya, Mizoram has the highest prevalence of smoking among women in the country (Singh and Ladu Singh 2014). Finally, smokeless tobacco usage is highest in Nagaland (49.9%) compared to other Northeastern states (Sinha et al. 2003).

In this study, we aimed at understanding population-specific variation in neuropsychological disorders among South Asian populations by curating and comparing the single nucleotide polymorphisms (SNPs), associated with the same.

2 Materials and Methods

2.1 Data Collection

A total of 4617 SNPs associated with various neuropsychological disorders such as alcohol dependence, Alzheimer's disease, mental health and depression, smoking, nicotine dependence, schizophrenia, and substance abuse were curated from the University of California Santa Cruz (UCSC) genome browser (<http://genome.ucsc.edu/>). The potential link between the curated SNPs with neuropsychological disorders was confirmed by annotating them using SNPnexus web-based server (Dayem Ullah et al. 2018). The genome-wide data (*IndiaHO_dataforrelease*) of 1662 South Asian individuals, from 241 distinct populations were obtained from the database of Dr. David Reich, Harvard Medical School, Harvard University, USA (Accessed on 14.11.2017, <https://reich.hms.harvard.edu/datasets>). PLINK v1.9 (Purcell et al. 2007) was employed for all file conversions and manipulations. SNPs related to neuropsychological disorders were extracted from *IndiaHO_dataforrelease* using the --extract command in PLINK. Out of 4617 curated SNPs, 4233 SNPs were absent in this genotype file. Therefore, our final dataset assessed 1662 South Asian genomes for 384 SNPs related to neuropsychological disorders.

Most South Asian genomes are genetic mosaics of Indus periphery and AASI related ancestries with variable ancestral fractions from Mid-Late Bronze Age steppe genomes (Steppe MLBA) (Shinde et al. 2019). A total of 22 Steppe MLBA and four Indus periphery genomes, including the only genome obtained from the Indus valley, were analyzed to investigate whether any of the 384 SNPs, employed in this study, were fixed for the alternative alleles among these ancient populations. These ancient genomes were also obtained from the aforesaid database of Dr. David Reich (Accessed on 21.08.2019).

2.2 Clustering of South Asians based on Neuropsychological SNPs

Fine structure within South Asian genomes, based on 384 SNPs associated with neuropsychological disorders, was discerned

employing Principal Component Analysis (PCA) implemented in PLINK v1.9. PC1 and PC2, the two most informative PCs, were plotted in Microsoft® Excel v16.42.

2.3 Identical by Descent (IBD) analysis

Populations with high IBD sharing can be under natural selection since the latter is presumed to promote longer IBD segments. IBD sharing was calculated in PLINK v1.9 using the Pi-Hat statistic (Proportion IBD: $P(\text{IBD}=2) + 0.5 \cdot P(\text{IBD}=1)$). A pie chart depicting the highest Pi-Hat values was plotted in GraphPad Prism v9.

2.4 Genomic Data analysis

Cluster-stratified (population-wise) Minor Allele Frequencies (MAF) defined by the family IDs were estimated in PLINK v1.9. Observed and expected homozygosity estimation and subsequent method-of-moments F coefficient calculation for each individual in the dataset was performed in PLINK. Further, the fixation indices (FST) were estimated separately for all 384 markers employed in this study in PLINK v1.9. The geographical affinity of the individual genomes was indicated by the family ID (FID).

2.5 Negative Control

To assess whether the population structure revealed by the SNPs associated with neuropsychological disorders is an artifact of the small number of SNPs under evaluation, we qualitatively compared the PCA plot produced by these SNPs with the ones generated by 100 SNP panels of randomly sampled 384 SNPs from the whole genome dataset.

3 Results

3.1 Clustering of populations from South Asia based on Neuropsychological SNPs

South Asians depicted an AAA –ASI –ANI cline along X-axis (PC1) with Juang, Koya, Gadaba, Ho, Santhal, Bondo, Oraon, Batudi, Kharia, Kondakamari, Asur, Porja, Kondh, Palliyar, Ulladan, and Irula genomes clustering at one end of the cline, and Ashkenazi Jews, Cochin Jews, Kamboj, Makrani and Kalash genomes clustering at the other end (Figure 1). The Ancestral Tibeto Burman (ATB) genomes including Magar, Sherpa, Chakshanega, Kusunda, and Tharu populations formed a distinct cluster, differentiated from AAA and ASI along the Y-axis (PC2).

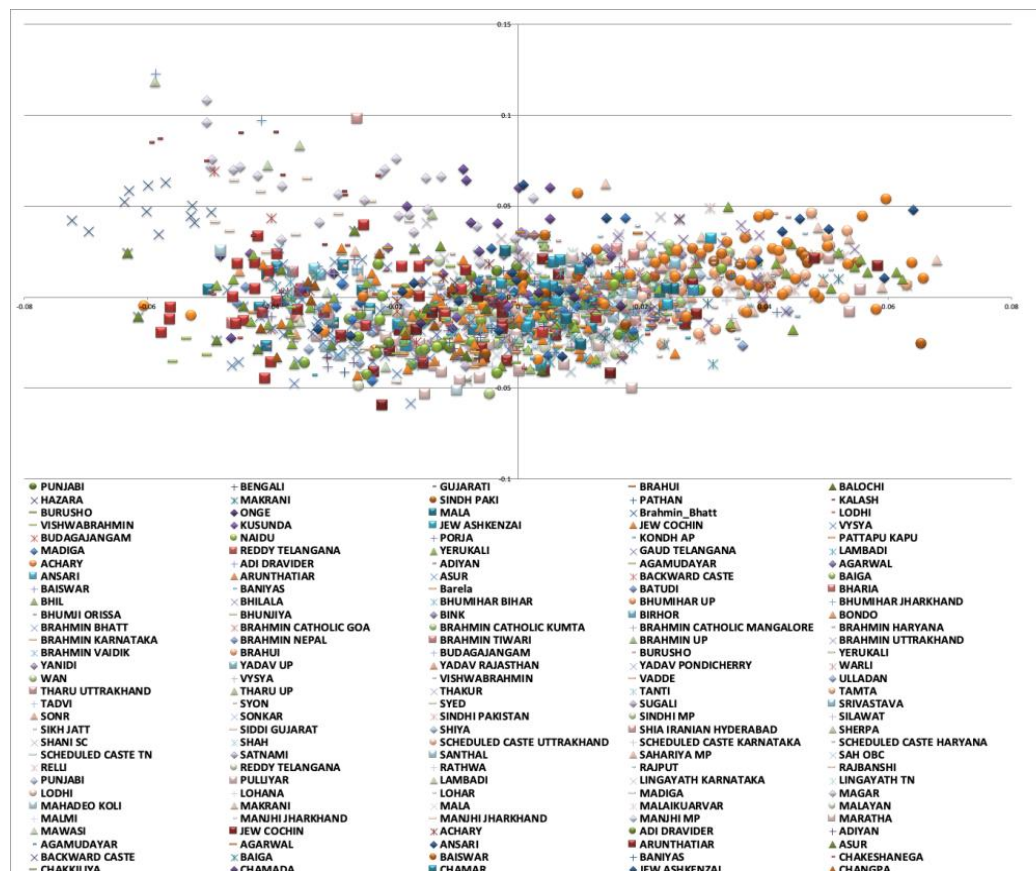


Figure 1 Principal Component Analysis (PCA) of South Asian populations based on SNPs, associated with neuropsychological disorders. PCA plot revealing the genetic differentiation among South Asian genomes based on 384 SNPs associated with neuropsychological disorders

Interestingly, Rajbanshi genomes from West Bengal formed a bridge between the ATB and AAA clusters for the study of SNPs associated with neuropsychological disorders.

To assess whether this AAA-ASI-ANI cline is unique for the SNPs associated with neuropsychological disorders or an artifact of a small sample size, we compared the PCA plot generated by these SNPs with those generated by 100 SNP panels of randomly sampled 384 SNPs from the whole genome dataset. No PCA plot generated by the random SNPs could recapitulate this cline and depicted random clustering of populations, irrespective of their ancestral associations. Therefore, we surmise that the AAA-ASI-ANI cline depicted by the SNPs associated with neuropsychological disorders is likely reflective of the clustering of South Asian genomes according to their propensity towards these diseases.

3.2 Identical by Descent (IBD) analysis

The highest Pi-Hat value obtained was 0.5276. For stringency, we made a cut-off of the Pi-Hat value at 0.3 and calculated the number of pairs sharing high Pi-Hat values in various populations under study. We found 19 populations sharing 10 or more high Pi-Hat values, indicating high IBD sharing. The highest IBD sharing was observed among Yerukali genomes from Andhra Pradesh and

Telangana, closely followed by the native Burusho people from Pakistan, Yadavs from Pondicherry, and Vishwabrahmins from Southern India (Figure 2), a likely indication of a recent selective sweep in these SNPs, among the aforementioned populations. Notably, out of the top seven populations in terms of IBD sharing, six are from Southern India.

3.3 Genomic Data analysis

Out of 384 SNPs assessed, ≥ 35 was fixed for the alternate allele among Ho (N=40), Kondakamari (N=38), Malmi (N=38), Agamudayar (N=37), Kamsali (N=37), Gadaba (N=36), Gowli (N=36), Sonr (N=35) genomes respectively. Overall, we found 25 genomes with at least 10 alleles of MAF=1. To note, seven out of the top ten genomes according to the number of fixed alleles were from Southern India, and six out of ten were tribal groups.

Onge genomes due to their genetic distinctness, expectedly had the highest discrepancies between the observed and expected heterozygosities and thus had the highest F values (>0.25). Among others, Kusundas from Nepal, several indigenous populations, and tribal groups from South India including Palliyar, Ulladan, Narikuruvar, and Kallar depicted high F values (>0.25) for the study SNPs. Among these, Palliyars from Kerala had the maximum number of individuals with significant differences

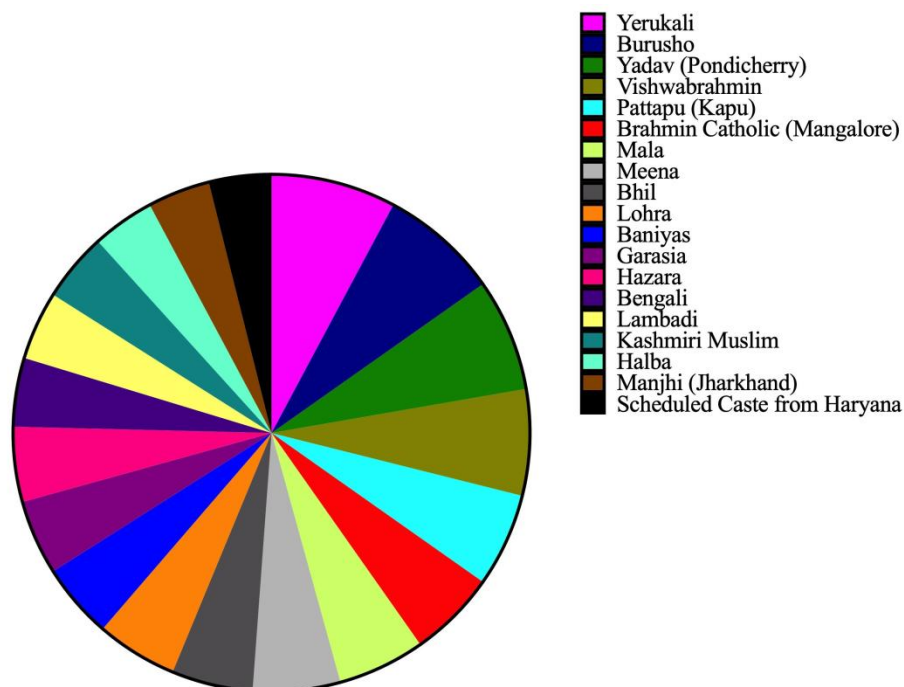


Figure 2 Identical by Descent (IBD) analysis of genomes from South Asia genomes. 19 populations with strong IBD sharing (>10 high Pi-Hat values) are plotted. The pie chart indicates that maximum number of highest IBD scores was found among the Yerukali population from Andhra Pradesh (Pink), followed by the Burusho people from Pakistan (Navi blue) and Yadavs from Pondicherry (Dark green).

The pie-chart was plotted in GraphPad Prism v9.

between observed and expected heterozygosities, indicating their genetic uniqueness for aforesaid SNPs.

The highest F_{st} (0.26) among the South Asian genomes assessed, was observed for rs11825659 at chromosome 11, which is associated with unipolar depression and alcohol dependence, followed by rs11038167 at chromosome 11 (F_{st} =0.13) and rs7683704 at chromosome 4 (F_{st} =0.08), which is associated with schizophrenia and alcoholism respectively (Shokraeian et al. 2019, Gizer et al. 2011). While none of the 384 SNPs were fixed for the alternative alleles among the Steppe MLBA genomes, 45 SNPs were fixed among the Indus Periphery genomes.

4 Discussions

In the current study, we aimed at understanding the variation in neuropsychological disorders at the population level among populations from South Asia. PCA on SNPs associated with neuropsychological disorders revealed a distinct cluster of Ancestral Tibeto Burman (ATB) genomes including Magar, Sherpa, Chakshanega, Kusunda, and Tharu populations, significantly differentiated from the ANI, ASI, and AAA clusters (Figure 1). The genomic distinctness of ATB genomes for the study SNPs can be presumed to be indicative of their higher susceptibility to neuropsychological disorders. This echoes our previous study, where we showed that the SNPs related to neuropsychological disorders are under strong selection pressure among individuals from East Asia (Hande et al. 2021).

Besides ATB genomes, genomes of several tribal and indigenous populations from South India revealed a higher propensity towards neuropsychological disorders. For instance, out of the top seven populations in terms of IBD sharing, six were from Southern India hinting at a recent selective sweep in terms of the SNPs associated with neuropsychological disorders among these populations. Further, out of the top ten genomes fixed for the alternate allele seven were from Southern India and six were from tribal groups. Several indigenous populations from Southern India also revealed high F values, indicating their genetic higher propensity for neuropsychological disorders. Congruent with our findings, a recent comprehensive study carried out by the Indian state-level disease burden initiative revealed that Southern Indian states such as Telangana, Tamil Nadu, Kerala, Karnataka, and Andhra Pradesh revealed a discernibly high prevalence of adulthood mental disorders including depression and anxiety (Sagar et al. 2020).

Out of 384 SNPs assessed in this study, 45 were fixed for the alternative alleles among ancient Indus Periphery genomes. Interestingly, among the modern-day populations, these SNPs were found to be fixed among the tribal groups and indigenous populations from Southern India with high AASI ancestry. This indicates that these SNPs were likely fixed for the alternative

alleles among people of AASI or Indus periphery ancestry and subsequently got transmitted to modern-day Indian populations, especially to those who had little or no Steppe MLBA admixture. To assess whether Indus Periphery genomes had signatures for susceptibility towards neuropsychological disorders, we annotated all SNPs of these genomes that were fixed for the alternative allele ($MAF=1$). A total of 90,781 SNPs were found to be fixed for the alternative allele, uniquely among the Indus Periphery genomes. Among them, 2273 are associated with various molecular pathways governing the neuronal system. It can be speculated that the residents of ancient Indus Valley likely had unique genetic signatures that made them more susceptible to neuropsychological disorders and might have propagated substance abuse and addiction among them. While not much is known about the substance abuse among individuals from the Indus valley civilization, Saraswat and Pokharia (2002) speculated that these people might have grown and used plants such as Ephedra and Datura for recreational purposes (Saraswat and Pokharia 2002).

Our study shines a light on the putative genetic signatures that may be associated with neuropsychological disorders among South Asian populations. The major limitation of our study is the genomic data employed for the analyses. This study was carried out using publicly available genomic data and therefore the disease status of these individuals is unknown. We did not recruit any patient genome in this study or perform any genome-wide association study (GWAS) to identify SNPs linked to neuropsychological disorders, both of which can be extremely crucial in understanding individual and/or population-specific variation in disease susceptibility. We surmise here that a more detailed understanding of this topic will require robust sequencing/microarray genotyping endeavors of patients suffering from neuropsychological disorders from varied ancestries and geographical locations.

Conclusion

Numerous admixture events among various populations across the globe have shaped the South Asian genome, which has been maintained by the century-old practice of intra-community marriage (endogamy). The socio-cultural and genetic diversity in South Asia, which is often reflected through the variation in different life-history traits and differential susceptibility towards various diseases and conditions. In the current study, we aimed at unraveling variation in susceptibility towards neuropsychological disorders among South Asian populations. We found that individuals from Northeast India, indigenous populations, and tribal groups from Southern India are genetically more susceptible to neuropsychological disorders compared to other populations in South Asia.

Overall, our study strongly advocates the adoption of pharmacogenomic approaches for developing population and individual specific therapeutics utilizing whole-genome sequencing data from people of diverse ancestries, performing GWAS on the same, and identifying population/individual specific disease loci. Adoption of a population genetics-driven strategy in neuropsychological disorders will potentially facilitate our understanding of patient genetic attributes that impact the variabilities in the disease presentations and help in the repurposing of existing drugs for the mitigation of neuropsychological disorders worldwide.

References

- Basu, A., Sarkar-Roy, N., & Majumder, P.P. (2016). Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. *Proceedings of the National Academy of Sciences*, 6, 1594–1599.
- Chaturvedi, H. K. (2003). The association of selected sociodemographic factors and differences in patterns of substance use: a pilot study in selected areas of Northeast India. *Substance Use & Misuse*, 9, 1305-22.
- DayemUllah, A.Z., Oscanoa, J., Wang, J., et al. (2018). SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. *Nucleic Acids Research*, 46, W109-W113.
- Dutta, A. K. (2013). Genetic factors affecting susceptibility to alcoholic liver disease in an Indian population. *Annals of Hepatology*, 6, 901-7.
- Gizer, I.R., Edenberg, H.J., Gilder, D.A., et al. (2011). Association of alcohol dehydrogenase genes with alcohol-related phenotypes in a Native American community sample. *Alcoholism: Clinical and Experimental Research*, 11, 2008–18.
- Hande, S.H., Krishna S.M., Sahote, K.K., et al. (2021). Population genetic variation of SLC6A4 gene, associated with neurophysiological development. *Journal of Genetics*, 100, 16.
- Ladusingh, L., Dhillon P., & Narzary P.K. (2017). Why Do the Youths in Northeast India Use Tobacco? *Journal of Environmental and Public Health*, 2017, 1391253.
- Mahanta, B., Mohapatra, P., Phukan, N., et al. (2016). Alcohol use among school-going adolescent boys and girls in an industrial town of Assam, India. *Indian Journal of Psychiatry*, 2, 157–63.
- Medhi, G.K., Hazarika, N.C., & Mahanta, J. (2006). Correlates of alcohol consumption and tobacco use among tea industry workers of Assam. *Substance Use and Misuse*, 5, 691–706.
- Nakatsuka, N., Moorjani, P., Rai, N., et al. (2017). The promise of discovering population-specific disease-associated genes in South Asia. *Nature Genetics*, 9, 1403–7.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3), 559-575.
- Sagar, R., Dandona, R., Gururaj, G., et al. (2020). The burden of mental disorders across the states of India: the Global Burden of Disease Study 1990–2017. *The Lancet Psychiatry*, 2, 148–61.
- Saikia, N., & Debbarma, B. (2020). The socioeconomic correlates of substance use among male adults in Northeast India. *Clinical epidemiology and global health*, 1, 149–57.
- Saraswat, K.S., & Pokharia, A.K. (2002). Palaeoethnobotanical Investigations. *Early Harappan Kunal*, 719, 105–140.
- Shinde, V., Narasimhan, V.M., Rohland N., et al. (2019). An Ancient Harappan Genome Lacks Ancestry from Steppe Pastoralists or Iranian Farmers. *Cell*, 3, 729-735.
- Shokraeian, P., Daneshmandpour, Y., Jamshidi, J., et al. (2019). Genetic analysis of rs11038167, rs11038172 and rs835784 polymorphisms of the TSPAN18 gene in Iranian schizophrenia patients. *Meta Gene*, 22, 100609
- Singh, A., & Ladusingh, L. (2014). Prevalence and Determinants of Tobacco Use in India: Evidence from Recent Global Adult Tobacco Survey Data. *PloS one*, 9(12), e114073.
- Sinha, D.N., Gupta, P.C., & Pednekar, M.S. (2003). Tobacco use among students in the eight North-eastern states of India. *Indian Journal of Cancer*, 2, 43-59.