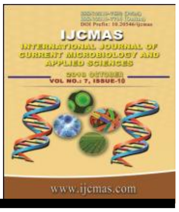
Int.J.Curr.Microbiol.App.Sci (2018) *7*(10): 103-110

EXCELLENT PUBLISHERS

International Journal of Current Microbiology and Applied Sciences
ISSN: 2319-7706 **Volume 7 Number 10 (2018)**
Journal homepage: http://www.ijcmas.com



Review Article https://doi.org/10.20546/ijcmas.2018.710.012

Ancestry Informative Markers: Getting a Lot Out of a Little

Supriya Chhotaray[1*], Aamir Bashir Wara[1], Dhan Pal[1], V. Bhanuprakash[1], Harshit Kumar[1], Snehasmita Panda[1], Mitek Tarang[1], Subhashree Parida[2], Bharat Bhushan[1] and Manjit Panigrahi[1]

[1]Division of Animal Genetics, [2]Division of Pharmacology, ICAR- Indian Veterinary Research Institute, Bareilly – 234122, India

**Corresponding author*

ABSTRACT

Keywords

Ancestry informative markers, Admixture, SNP

Article Info

Accepted:
04 September 2018
Available Online:
10 October 2018

The SNP markers have emerged as essential resources for performing genetic linkage, association, and admixture studies. Both academic and commercial groups are developing large numbers of genome-wide SNP datasets. These databases are utilised for several studies involving both structural and functional genomics. Some studies viz. admixture mapping, population structure identification and quantifying genetic diversity requires a genome-wide small informative panel of relatively evenly spaced markers. These informative markers can distinguish the ancestral origins of chromosomal segments in admixed individuals and can delineate the ancestral origin of a particular population. Thus present study reviews existing methods to select ancestry informative markers and their comparative efficiency in inferring individual ancestry level.

Introduction

Inference of individual ancestry from genetic markers is helpful in diverse situations, including admixture and association mapping, forensics, prediction of medical risks, wildlife management, food safety and studies of dispersal, gene flow, and evolutionary history (Shriver *et al.,* 1997; Davies *et al.,* 1999; Primmer *et al.,* 2000; Manel *et al.,* 2002; Bamshad *et al.,* 2003; Campbell *et al.,* 2003; Ziv and Burchard 2003). Most of the statistical methods developed for tracing ancestry use multi-locus genotypes and allele frequencies of the population, which is either specified in advance or estimated subsequently (Smouse *et al.,* 1982; Paetkau *et al.,* 1995; Rannala and Mountain 1997; Cornuet *et al.,* 1999; Pritchard *et al.,* 2000; Guinand *et al.,* 2002). The use of highly informative markers sometimes can reduce the genotyping rate, so, it is appropriate to measure the extent to which specific markers contribute to the ancestry inference.

The term ancestry refers to the proportion of genetic material that transcends each founder population. Ancestry-informative markers (AIMs) are those markers that can specify the likelihood of origin of a population from the

DNA sample where the source individual is not known or is unable to proclaim their ancestry (Phillips *et al.,* 2007). Autosomal AIMs are mainly used for studying admixture and inferring individual biogeographical ancestry (I-BGA) (Halder *et al.,* 2008). Availability of Different densities of SNP BeadChips has led us to generate a tremendous amount of genotypic datasets. Geographic ancestry can be easily extrapolated from this genotypic data produced through SNP genotyping (Rosenberg *et al.,* 2002; Tang *et al.,* 2005; Paschou *et al.,* 2007). While inferring genome-wide scans, most of the markers are found to be common and vary only to some extent among different populations. Markers which are present homogeneously within the population and possess highly differentiated frequencies are very informative for ancestry estimation (Shriver *et al.,* 2003). Basically, there are two approaches for studying genetic admixture and ancestry, i.e., model-based (Bayesian) (Rosenberg *et al.,* 2002; Pritchard *et al.,* 2000) and likelihood approach (based on the allele frequencies of the admixed population) (Liu *et al.,* 2013; Skotte *et al.,* 2013). However, the Bayesian approach and Principal component analysis (PCA) have been the main tools of choice for studying population structure, ancestry, and diversity (Paschou *et al.,* 2007).

Number of AIMs required to study population structure

The informativeness of the markers decides the number of markers needed for studies on admixture levels, population structure or genetic diversity (Ding *et al.,* 2011). According to one study, more number of biallelic markers (SNPs) (4-10X) are required to achieve the same results in terms of efficiency and resolution as that of multi-allelic markers (microsatellites) in population structure and diversity studies (Morin *et al.,* 2004). This problem can be addressed by dimension reduction of variables using PCA. Lewis *et al.,* (2011) have elucidated that in most of the cases the number of genetic markers required for ancestry inference can be reduced to 1.5% of the original number of SNPs with 92% accuracy. In the study of PCA correlated SNPs for structure identification in nine human populations (Paschou *et al.,* 2007) distributed worldwide, 50 PCA-correlated SNPs can assign the individuals to their population of origin with 100% accuracy. Recently, Getachew *et al.,* (2017) have identified top ranked 74 AIMs on the basis of F_{ST} which could predict the admixture level in crossbred populations of Menz x Awassi and Wollo x Awassi Sheep. In an admixed population of Swiss Fleckvieh cattle, Frkonja *et al.,* (2012) selected 48 and 96 top ranked SNPs with $F_{ST}> 0.651$ and 0.623, respectively and found a correlation of 0.907 and 0.924 with admixture level estimated through pedigree method. Ding *et al.,* (2011) obtained 100 to 200 top informative SNPs using Fisher Information Content (FIC), Shannon Information Content (SIC), F statistics (F_{ST}), Informativeness for Assignment Measure (I_n), and Absolute Allele Frequency Differences (delta, δ). They observed that each method generated similar estimates of ancestry contribution in CEU and YRI admixed population. Kavakiotis *et al.,* (2017) recently developed a method i.e., Frequent Item Feature Selection (FIFS), which was tested on a dataset of pigs. This dataset consisted of 446 individuals belonging to 14 sub-populations, genotyped at 59,436 SNPs. The SNPs selected by different methods were FIFS: 28 SNPs, Delta: 70 SNPs, Pairwise FST: 70 SNPs, I_n: 100 SNPs. Their approach successfully dealt with the problem of informative marker selection in high dimensional genomic datasets.

Methods to find AIMs

There are basically two different approaches. One is to estimate the assignment power of each individual loci using available software (WHICHLOCI, Banks *et al.*, 2003) or combinations of the different software (i.e. GAFS, Topchy *et al.*, 2004 and BELS, Bromaghin, 2008). All these software possess few advantages and disadvantages (Helyar *et al.*, 2011), but their primary requirement is the computational efficiency. As Helyar *et al.*, (2011) mentioned, although, no limitations exists for the number of loci or individuals to be included in analysis, but it may be prohibitive on a desktop.

A second approach is to rank loci solely according to their informativeness, that is, the marker information content, which is the amount of information that a locus holds regarding the ancestry of an individual. The use of markers with high informativeness reduces the number of markers needed for correct assignment (Rosenberg *et al.*, 2003). Several measures/criteria of marker informativeness have been proposed (Rosenberg *et al.*, 2003; Ding *et al.*, 2011), such as Delta (Shriver *et al.*, 1997), pairwise Wright's F_{ST} given by Wright (1951), global Wright's F_{ST} by Wright (1951), pair wise Weir and Cockerham F_{ST} by Weir and Cockerham (1984), global pair wise Weir and Cockerham F_{ST} by Weir and Cockerham (1984), and informativeness for assignment (I_n) (Rosenberg *et al.*, 2003). The application of different methods for marker prioritization and decision making in the construction of SNP panels is becoming more important as large high-throughput assays become readily available. A novel data mining approach, called FIFS, based on the use of frequent items for selection of the most informative markers from population genomic data has also been developed by Kavakiotis *et al.*, (2017). It has two main components. For each sampled population the first component detects the most unique and frequently occurring genotypes while the second one picks the most appropriate among them and return the informative SNP subsets. TRES: Toolbox for Ranking and Evaluation of SNPs is another tool which is used for ranking markers and selecting set of SNPs on the basis of Wright's F_{ST}, informativeness for assignment (I_n), and Absolute allele frequency difference (δ) (Kavakiotis *et al.*, 2015). Existing approaches to select a subset of informative markers entails prior knowledge of the ancestry of individuals included in the study. PCA, a powerful dimensionality reduction technique is an emerging approach to identify a small panel of informative markers. Paschou *et al.*, (2007) developed a novel algorithm based on PCA that does not depend on any prior assumptions, which can be applied to datasets of hundreds of individuals and millions of markers.

Methods to measure marker information content

Absolute allele frequency difference (δ): For a biallelic locus, suppose allele one is the reference allele, then, δ=|p11−p21|. A marker with δ = 1 gives complete information about ancestry while a marker with δ = 0 carries no information. In an admixture model, this is related to the amount of linkage disequilibrium (Chakraborty and Weiss 1988); whereas in a multilocus no-admixture model this is related to the probability of correct assignment (Risch *et al.*, 2002); Fisher information curvature criterion and *ORCA* for K=2 (Rosenberg *et al.*, 2003). But it is optimum when only two populations are possible sources; and also does not take into account all available information about allele frequencies (Stephens *et al.*, 1999; Campbell *et al.*, 2003). Its statistical features also do not apply to the multi-allelic extension of δ (Shriver *et al.*, 1997).

F_{ST}: This is a measure of population substructure and genetic divergence among subpopulations. It is excess in the probability of identity of alleles from the same population compared with randomly chosen alleles.

$F_{ST} = (p_1j-p_2j)^2 / (p_1j+p_2j)(2-(p_1j+p_2j))$

Here, j = 1 or 2 is the reference allele. Values of F_{ST} can range from 0 to 1. A high F_{ST} value implies a considerable degree of differentiation between populations. This is related, for biallelic markers, to the quotient of expected posterior and prior variance of ancestry in a population equally admixed from two sources (McKeigue 1998; Molokhia *et al.,* 2003). But it performs only slightly better than random markers (Rosenberg *et al.,* 2001).

Expected heterozygosity: It performs better than random markers (Rosenberg *et al.,* (2001). It is good at measuring the amount of variation but not the differences across populations.

The number of alleles: It measures the amount of variation but not the differences across populations, hence is useful only for multi-allelic markers that have variation in the number of alleles such as microsatellites.

Fisher information curvature criterion: It enables prediction about approximate variances of ancestry estimates and the information matrix is additive across loci that are independent within populations.

It depends on unknown ancestry coefficients and requires computation for many possible parameter values; largest eigenvalue gives an upper bound that might not be generally applicable across the parameter space (Rosenberg *et al.,* 2003).

Shannon information criterion: The concept of entropy can be used to develop a measure of marker informativeness (Rosenberg *et al.,* 2003). Entropy is a measure of the uncertainty associated with a random variable and quantifies the expected information content contained in the data.

Pairwise Kullback-Leibler divergence: For K=2, it possesses an average potential and multi-locus extension, for assignment of an allele to one population. Contribution of specific alleles can also be measured. It requires that only two populations are possible sources and in small samples the estimate is upwardly biased (Rosenberg *et al.,* 2003).

Informativeness for assignment (I_n): This measure was produced by Rosenberg *et al.,* (2003). It potentially allocates an allele to one population and possess a natural multi-locus extension. It also facilitates measurement of specific allele contribution. Its ancestry assignment is better than random or highly heterozygous markers. But in small samples the estimate is upwardly biased.

Informativeness for ancestry coefficients (I_a): this was also formulated by Rosenberg *et al.,* (2003). It is similar to I_n in statistical features, but in samples with populations of equal sample size computational efficiency decreases.

Optimal rate of correct assignment (ORCA): Gives the probability of correct assignment of an allele using the decision rule with lowest risk; has a natural multi-locus extension; enables measurement of contributions of specific alleles. But in small samples estimates are upwardly biased (Rosenberg *et al.,* 2003).

Comparison of different methods

The simulation study conducted by Ding *et al.,* (2011) revealed that I_n was the best as compared to δ, F_{ST}, FIC, and SIC in selecting the set of AIMs giving the least bias and mean square error in ancestry estimation. FIC had the least overlap with other measures in selecting markers. F_{ST}, FIC, and I_n gave relatively smaller and similar marker panels, whereas SIC gave a very small panel of AIMs. FIC and SIC were more alike in picking a set of SNPs, while δ, F_{ST}, and I_n were more likely to pick the same set of SNPs, and FIC was more likely to choose SNPs that were not chosen by the other measures. FIC and SIC measures necessitate a pre-defined ancestral proportions in an admixed population, whereas F_{ST}, δ, and I_n do not. Combined information from more than one method may provide a reliable means, in selecting markers for ancestry inference (Ding *et al.*, 2011).

Use of AIMs in detecting meat adulteration and mislabelling

Wilkinson *et al.*, (2012) developed a panel of 96 SNPs that has the ability to authenticate pork products labelled with traditional breed names and thus expose mislabelled products. Assignment of individuals to particular breeds was extremely accurate for traditional breeds. When they studied 40 market samples with the panel of 96 SNPs, observed that 2 samples not assigned to claimed breed origin but assigned to another breed, indicating possibly mislabelled meat. The assignment to traditional breeds was also accurate for processed pork product though genotyping rate fell to 88%. Similarly, Orrù *et al.*, (2009) found 18 SNPs, those are useful in tracing beef to the most consumed breeds in the markets of Italy. Suekawa *et al.*, (2010) developed five breed specific DNA markers to correctly discriminate between Japanese and imported cattle for food safety based on bovine 50K SNP array.

Although millions of SNPs have been identified through which population structure, admixture, traceability and diversity is studied, the number of markers can be reduced to minimal for the same purpose based on their information content. The small panel of ancestry informative markers may greatly reduce the genotyping cost while retaining the power for assigning individuals to accurate ancestral populations. A choice of method to select AIMs, with least biasness and root mean square error plays an important role in admixture studies which in turn avoids false positive genotype-phenotype associations. The set of informative markers are also useful in meat traceability for food safety, and checking mislabelling of the meat products. The choice of method mostly depends upon the number of populations in the dataset, sample size, type of marker and number of loci involved, but we suggest a combination of all these methods to measure the marker information content and to select the desired small AIM panel accordingly.

References

Bamshad, M.J., Wooding, S., Watkins, W.S., Ostler, C.T., Batzer, M.A., and Jorde, L.B. (2003) Human population genetic structure and inference of group membership. Am. J. Human Genet., 72: 578–589.

Banks, M. A., Eichert, W. Olsen, J. B. (2003) Which genetic loci have greater population assignment power? Bioinformatics. 19: 1436–1438.

Bromaghin, J.F. (2008) BELS: backward elimination locus selection for studies of mixture composition or individual assignment. MolEcolResour., 8: 568–571.

Campbell, D., Duchesne, P., and Bernatchez, L. (2003) AFLP utility for population assignment studies: Analytical investigation and empirical comparison with microsatellites. Mol. Ecol., 12: 1979–1991.

Chakraborty, R. and Weiss, K.M., (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci.

Proceedings of the National Academy of Sciences. 85(23): 9119-9123.

Cornuet, J.M., Piry, S., Luikart, G., Estoup, A. and Solignac, M., (1999) New methods employing multilocus genotypes to select or exclude populations as origins of individuals. Genetics. 153(4): 1989-2000.

Davies, N., Villablanca, F.X., and Roderick, G.K. (1999) Determining the source of individuals: Multilocus genotyping in non-equilibrium population genetics. Trends Ecol. Evol., 14: 17–21.

Ding, L., Wiener, H., Abebe, T., Altaye, M., Go, R.C., Kercsmar, C., Grabowski, G., Martin, L.J., Hershey, G.K.K., Chakorborty, R. and Baye, T.M. (2011) Comparison of measures of marker informativeness for ancestry and admixture mapping. BMC genomics. 12(1): 622.

Frkonja, A., Gredler, B., Schnyder, U., Curik, I. and Soelkner, J. (2012) Prediction of breed composition in an admixed cattle population. Anim. Genet., 43(6): 696-703.

Getachew, T., Huson, H.J., Wurzinger, M., Burgstaller, J., Gizaw, S., Haile, A., Rischkowsky, B., Brem, G., Boison, S.A., Mészáros, G. and Mwai, A.O. (2017) Identifying highly informative genetic markers for quantification of ancestry proportions in crossbred sheep populations: implications for choosing optimum levels of admixture. BMC genetics. 18(1): 80.

Guinand, B., Topchy, A., Page, K.S., Burnham-Curtis, M.K., Punch, W.F. and Scribner, K.T., (2002) Comparisons of likelihood and machine learning methods of individual classification. J. Hered., 93(4): 260-269.

Halder, I., Shriver, M., Thomas, M., Fernandez, J.R. and Frudakis, T. (2008) A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. Hum. Mutat., 29(5): 648-658.

Helyar, S. J., Hemmer-Hansen, J., Bekkevold, D. (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. MolEcolResour., 11: 123–136.

Kavakiotis, I., Samaras, P., Triantafyllidis, A. and Vlahavas, I. (2017) FIFS: A data mining method for informative marker selection in high dimensional population genomic data. Comput. Biol. Med., 90: 146-154.

Kavakiotis, I., Triantafyllidis, A., Ntelidou, D., Alexandri, P., Megens, H.J., Crooijmans, R.P., Groenen, M.A., Tsoumakas, G. and Vlahavas, I. (2015) TRES: identification of discriminatory and informative SNPs from population genomic data. J. Hered., 106(5): 672-676.

Lewis, J., Abas, Z., Dadousis, C., Lykidis, D., Paschou, P. and Drineas, P. (2011) Tracing cattle breeds with principal components analysis ancestry informative SNPs. PLoS ONE. 6(4):18007.

Liu, Y., Nyunoya, T., Leng, S., Belinsky, S.A., Tesfaigzi, Y. and Bruse, S. (2013) Softwares and methods for estimating genetic ancestry in human populations. Human genomics. 7(1): 1.

Manel, S., Berthier, P. and Luikart, G., (2002) Detecting wildlife poaching: identifying the origin of individuals with Bayesian assignment tests and multilocus genotypes. Conserv Biol., 16(3): 650-659.

McKeigue, P.M., (1998) Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. Am. J. Human Genet., 63(1): 241-251.

Molokhia, M.A.R.I.A.M., Hoggart, C.L.I.V.E., Patrick, A.L., Shriver, M.A.R.K., Parra, E.S.T.E.B.A.N., Ye, J., Silman, A.J. and McKeigue, P.M., (2003) Relation of risk of systemic lupus erythematosus to west African admixture in a Caribbean population. Hum. Genet., 112(3): 310-318.

Morin, P.A., Luikart, G. and Wayne, R.K. (2004) SNPs in ecology, evolution and conservation. Trends Ecol. Evol., 19(4): 208-216.

Orru, L., Catillo, G., Napolitano, F., De Matteis, G., Scata, M.C., Signorelli, F. and Moioli, B., (2009) Characterization of a SNPs panel for meat traceability in six cattle breeds. Food Control, 20(9): 856-860.

Paetkau, D., Calvert, W., Stirling, I. and Strobeck, C., (1995) Microsatellite analysis of population structure in Canadian polar bears. Mol. Ecol., 4(3): 347-354.

Paschou, P., Ziv, E., Burchard, E.G., Choudhry, S., Rodriguez-Cintron, W., Mahoney, M.W. and Drineas, P. (2007) PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet., 3(9): 160.

Phillips, C., Salas, A., Sanchez, J.J., Fondevila, M., Gomez-Tato, A., Alvarez-Dios, J., Calaza,M., de Cal, M.C., Ballard, D., Lareu, M.V. and Carracedo, A. (2007) Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Sci. Int. Genet., 1(3): 273-280

Primmer, C.R., Koskinen, M.T., and Piironen, J. (2000) The one that did not get away: Individual assignment using microsatellite data detects a case of fishing competition fraud. Proc. R. Soc. Lond. B 267, 1699–1704.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. Genetics. 155(2): 945-959.

Rannala, B. and Mountain, J.L., (1997) Detecting immigration by using multilocus genotypes. Proceedings of the National Academy of Sciences, 94(17): 9197-9201.

Risch, N., Burchard, E., Ziv, E. and Tang, H. (2002) Categorization of humans in biomedical research: genes, race and disease. Genome Biol., 3(7): comment 2007-1.

Rosenberg, N.A., Burke, T., Elo, K., Feldman, M.W., Freidlin, P.J., Groenen, M.A., Hillel, J., Mäki-Tanila, A., Tixier-Boichard, M., Vignal, A. and Wimmers, K., (2001) Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. Genetics. 159(2): 699-713.

Rosenberg, N.A., Li, L.M., Ward, R. and Pritchard, J.K., (2003) Informativeness of genetic markers for inference of ancestry. Am. J. Human Genet., 73(6): 1402-1422.

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A. and Feldman, M.W. (2002) Genetic structure of human populations. Science. 298(5602): 2381-2385.

Shriver, M.D., Parra, E.J., Dios, S., Bonilla, C., Norton, H., Jovel, C., Pfaff, C., Jones, C., Massac, A., Cameron, N. and Baron, A. (2003) Skin pigmentation, biogeographical ancestry and admixture mapping. Hum.Genet., 112(4): 387-399.

Shriver, M.D., Smith, M.W., Jin, L., Marcini, A., Akey, J.M., Deka, R. and Ferrell, R.E. (1997) Ethnic-affiliation estimation by use of population-specific DNA markers. Am J Hum Genet., 60(4): 957.

Skotte, L., Korneliussen, T.S. and Albrechtsen, A. (2013) Estimating individual admixture proportions from next generation sequencing data. Genetics. 195(3): 693-702.

Smouse, P.E., Spielman, R.S. and Park, M.H., (1982) Multiple-locus allocation of individuals to groups as a function of the genetic variation within and differences among human populations. Am. Nat., 119(4): 445-463.

Stephens, P.A. and Sutherland, W.J., (1999) Consequences of the Allee effect for behaviour, ecology and conservation. Trends Ecol. Evol., 14(10): 401-405.

Suekawa, Y., Aihara, H., Araki, M., Hosokawa, D., Mannen, H. and Sasazaki, S., (2010) Development of breed identification markers based on a bovine 50K SNP array. Meat Sci., 85(2): 285-288.

Tang, H., Quertermous, T., Rodriguez, B., Kardia, S.L., Zhu, X., Brown, A., Pankow, J.S., Province, M.A., Hunt, S.C., Boerwinkle, E. and Schork, N.J. (2005) Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. Am J Hum Genet., 76(2): 268-275.

Topchy, A., Scribner, K., Punch, W. (2004) Accuracy-driven loci selection and assignment of individuals. Mol. Ecol. Notes., 4: 798–800.

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

Wilkinson, S., Archibald, A.L., Haley, C.S., Megens, H.J., Crooijmans, R.P., Groenen, M.A., Wiener, P. and Ogden, R., (2012) Development of a genetic tool for product regulation in the diverse British pig breed market. BMC Genom., 13(1): 580.

Wright, S. (1951) The genetical structure of populations. Ann Eugen., 15: 323–354.

Ziv, E. and Burchard, E.G., (2003) Human population structure and genetic association studies. Pharmacogenomics. 4(4): 431-441.

How to cite this article:

Supriya Chhotaray, Aamir Bashir Wara, Dhan Pal, V. Bhanuprakash, Harshit Kumar, Snehasmita Panda, Mitek Tarang, Subhashree Parida, Bharat Bhushan and Manjit Panigrahi. 2018. Ancestry Informative Markers: Getting a Lot Out of a Little. *Int.J.Curr.Microbiol.App.Sci.* 7(10): 103-110. doi: https://doi.org/10.20546/ijcmas.2018.710.012