

MINI-REVIEW

Population stratification and genetic association studies in South Asia

Alan H Bittles

Centre for Human Genetics, Edith Cowan University, 100 Joondalup Drive, Perth WA 6027, Australia

*Correspondence to: Alan H Bittles, Email: a.bittles@ecu.edu.au, Tel: +61 8 6304 5623, Fax: +61 8 6304 5851

Journal of Molecular and Genetic Medicine (2005), 1(2), 43-48

© Copyright Alan H Bittles

(Received 25 November 2005; Accepted 13 December 2005; Available online 30 December 2005; Published 30 December 2005)

ABSTRACT

Population stratification and its influence on genetic association studies is a controversial topic. Although it has been suggested that stratification is unlikely to bias the results of association studies conducted in developed countries, convincing contrary empirical evidence has been published. However, it is in populations where historical ethnic, religious and language barriers exist that community subdivisions will predictably exert greatest genetic effect, and influence the organization of association studies. In many of the populations of the Indian sub-continent, these basic population divisions are compounded by a strict tradition of intra-community marriage and by marriage between close biological relatives. Data on the very significant levels of genetic diversity that characterize the populations of India and Pakistan, with some 50,000-60,000 caste and non-caste communities in India, and average first cousin marriage rates of 40%-50% in Pakistan, are presented and discussed. Under these circumstances, failure to explicitly control for caste/*biraderi* membership and the presence of consanguinity could seriously jeopardize, and may totally invalidate, the results of association/case control studies and clinical trials.

KEYWORDS: *Stratification, endogamy, consanguinity, association studies, India, Pakistan*

INTRODUCTION

The reliability and reproducibility of results obtained in genetic association studies have increasingly been questioned, especially when applied to complex diseases and in dealing with rare alleles with small effect sizes. In many cases poor study design has been identified as the main cause of misleading results (Cardon and Bell, 2001), with a lack of appropriate power calculations and insufficiently large datasets creating particular problems (Dahlman et al, 2002). Factors such as the effect size of susceptibility loci, the frequency of disease alleles, the frequency of marker alleles correlated with disease alleles, and the extent of linkage disequilibrium at or close to the region of the genome under investigation also have been cited as potential problem areas (Zondervan and Cardon, 2004), and the possible role(s) of epistasis remains to be investigated in appropriate detail (Carlborg and Haley, 2004).

Since the evolutionary history of haplotypes and patterns of linkage disequilibrium often vary widely between dif-

ferent ethnic populations, population stratification by ethnicity can cause problems in terms of the power of association studies (Cardon and Bell, 2001). Except in ethnically diverse samples, it has been suggested that population stratification may not be a major cause of spurious allelic associations (Cardon and Palmer, 2003). However, the influence of stratification on genetic association studies has been demonstrated even in well-designed protocols, with greatest effect in recently admixed populations and for diseases with variant prevalence rates in the ancestral source populations (Freedman et al, 2004).

Specific attention also has been drawn to the phenomenon of cryptic relatedness, *i.e.*, kinship among cases or controls that is unknown to or unrecognized by the investigator. Although not a problem in appropriately designed studies conducted on outbred populations, cryptic relatedness can result in false positives where there has been a sampling bias towards the recruitment of relatives and/or in founder populations that have undergone rapid, recent growth (Voight and Pritchard, 2005). Its undetected presence may

therefore be a major source of error in association studies when these conditions apply.

Estimates of the levels of genetic variation within human populations vary from lower bound values of 83%-88% (Excoffier and Hamilton, 2003) to 93%-95% of total genetic variance (Rosenberg et al, 2002). The full implications of these findings have largely been obscured by debate as to whether or not ethnicity should be considered in biomedical studies and clinical practice (Cooper et al, 2003; González Burchard et al, 2003), and because of this controversy the potential impact of intra-population allelic differences often seems to be overlooked or ignored in health-based studies. The aim of this mini-review is to highlight underlying causes of genomic variation in the populations of India and Pakistan that can be ascribed to their very diverse social, demographic and genetic sub-structures, and to consider the effects of this variation on the results of genetic association/case-control studies and clinical trials conducted in the sub-continent.

THE BASES OF GENETIC SUB-DIVISION IN INDIA

Potential problems with respect to stratification in association studies especially apply to populations in the Indian sub-continent because of the multiple ethnic, language, religious and socio-demographic boundaries that historically have restricted inter-community gene flow. In terms of language differentials it is estimated that there are just four major language families in India, *i.e.*, Austric, Dravidian, Indo-European and Sino-Tibetan (Gadgil et al, 1998), but 15 major languages are spoken which can be further subdivided into 4,647 'mother tongues' used by specific communities (Bhasin et al, 1992; Gadgil et al, 1998).

More importantly from a genetic perspective, the stratification effects of language divisions are exacerbated by caste differentials which apply to the 82% of Indians who are Hindu. The caste system is believed to have been introduced to India by Indo-European invaders some 3,500 bp. Since that time caste has traditionally defined the type of work undertaken by individuals and hence has governed their position and status within society (Thapar, 1966; Basham, 1967; Gadgil and Guha, 1995). As caste membership is hereditary, and to the present day virtually all Hindu marriages are contracted within caste boundaries, caste acts as a potent and long-established mechanism of genetic subdivision within Hindu society. At the same time, since the patterns of employment and living conditions of most individuals within Indian society continue to be largely dependent on their caste status, caste also can be regarded as an 'environmental' variable, which has important implications for the design of complex trait association studies.

Current estimates suggest that there are approximately 3,000 castes, together with more than 1,600 'scheduled tribes and castes' whose members effectively live outside the caste system (Bhasin et al, 1992). In the early days of the caste system entire non-Indian communities could be accorded caste status within Hindu society (Thapar, 1966; Basham, 1967), and during the 18th and 19th centuries the splitting of castes into endogamous sub-castes was quite

commonly observed (Census of Mysore, 1871; Basham, 1967). Further, some tribal communities are reported to have attained lower caste status on adopting settled agriculture practices (Bhattacharya et al, 1999). During the last century caste boundaries became more inflexible, although with continuing opportunities for hypergamy, *i.e.*, the change of caste by a woman of lower caste on marriage to a husband of higher caste. This may explain reports of limited female but not male gene flow between closely ranked castes (Bamshad et al, 1998; Bhattacharya et al, 1999). However, male gene flow from caste groups to tribal communities has been demonstrated (Ramana et al, 2001; Cordaux et al, 2004).

Some 130 million of the non-Hindu population of India are Muslim, and there also are major Christian, Sikh, Buddhist and Jain communities which individually number in the millions. Although caste restrictions generally do not apply to these non-Hindu communities, they often exhibit stringent religious and social stratification. As a result of these various population subdivisions it has been estimated that there are 50,000-60,000 separate endogamous communities in India as a whole (Gadgil et al, 1998), ranging in total numbers from fewer than a hundred individuals to hundreds of thousands.

Given the potential relationship between founder effect and cryptic relatedness (Voight and Pritchard, 2005), it is important to note the very rapid growth of the Indian population during the course of the 20th century, from an estimated 238 million in 1901 (Dyson, 2001) to the current population of 1,104 million (PRB, 2005). In addition, in past generations the total number of communities across the Indian sub-continent may have been closer to 75,000 (Cavalli-Sforza et al, 1994), many of which would have been numerically very small. A combination of restricted initial effective population sizes, rapid population growth and strict marital endogamy would be expected to significantly increase the opportunities for founder effect and random drift at community level, with population bottlenecks also potentially affecting specific communities, *e.g.*, following major epidemics. Because of these past events, high levels of genomic homozygosity can be observed even in communities which proscribe intra-familial marriage, as evidenced by studies on U.K. migrants of north Indian origin (Overall and Nicholls, 2001).

Preferential consanguineous marriage is, however, a potent additional factor that would be expected to facilitate genetic stratification in many Indian communities and thus influence the results of association studies. The National Family and Health Survey conducted in India during 1992-1993 indicated that 11.9% of marriages were consanguineous, equivalent to a mean coefficient of inbreeding, $\alpha = 0.0075$ (IIPS, 1995). But this composite figure masks the very marked ethnic, regional and religious differences in the prevalence of consanguineous marriages across the country. In the majority Hindu population, consanguineous unions are proscribed for North Indian communities (Kapadia, 1958; Bittles et al, 1991), but uncle-niece ($F = 0.125$), with progeny homozygous at 12.5% of loci, and first cousin marriages ($F = 0.0625$), with progeny homozygous at 6.25% of loci, are especially popular in the three

southern states of Andhra Pradesh, Karnataka and Tamil Nadu, which have a combined population approaching 200 million. In these three states, an estimated 29.7%-38.2% of marriages are consanguineous, with $\alpha = 0.0180-0.0266$ (IIPS, 1995; Bittles, 1998, 2002), and 7.5%, 10.6% and 21.0% consanguinity additionally was reported in neighbouring Kerala, Goa and Maharashtra (IIPS, 1995). It should be stressed that these estimates apply to a single generation only. Where there has been a longstanding tradition of consanguineous marriage across generations, the resultant level of cumulative homozygosity would predictably be very much higher (Bittles et al, 1991; Bittles, 2001).

Among non-Hindu minorities, first cousin unions nationally account for 20.8% of Muslim marriages (Bittles and Hussain, 2000), and uncle-niece and first cousin marriages are favoured by many south Indian Christian groups, e.g., with 18.6% consanguineous marriage ($\alpha = 0.0173$) among Christians in Karnataka (Bittles et al, 1991). The specific patterns of consanguineous marriage contracted vary by religion, with uncle-niece marriages proscribed for Muslims but double first cousin unions (also $F = 0.125$) permitted within Islam. Similarly, almost all Hindu and Christian first cousin marriages are cross-cousin, most commonly between a man and his mother's brother's daughter, whereas in Muslim communities all four types of first cousin marriage are permissible. While this difference would have no major effect at autosomal loci, it could significantly influence the expression of genes located on the X-chromosome (Bittles, 2001).

PATTERNS OF POPULATION SUB-DIVISION IN PAKISTAN

A comparable picture of community endogamy exists in Pakistan but with different emphases on the underlying causes of population stratification. Some 98% of the current population of 162 million are Muslim, however major language and ethnic differences exist with some 18 ethnic groups and more than 60 languages identified (Qamar et al, 2002; Hussain, 2005). Rather than the straightforward Sunni/Shia divide generally perceived outside Pakistan, major subdivisions exist within the different branches of Islam. The Sunni majority is divided into four major endogamous religious groups, the Hanafi, Shafei, Maliki and Hanbali, based on different schools of Islamic jurisprudence, and the Shia minority is similarly subdivided into the Ishnahary, Ismailis, and Dawoodi Bohras (Hussain, 2005). Sufism is quite widespread, and there are also other minor, non-Muslim, doctrinal subgroups, such as the Ahmadis and Qadianis.

Superimposed on these basic population divisions, marriage in Pakistan is usually contracted within traditional social and occupational groups, variously termed *biraderis*, *quoms* or *zats* (Shami et al, 1994; Hussain, 2005), which to an extent parallel the Hindu caste system. Besides inter-ethnic genetic differences (Mohyuddin et al, 2001; Qamar et al, 2002), although marriages between spouses from different *biraderis* can occur, genomic studies have shown major differences between the members of co-resident *biraderis* that appear indicative of very limited levels of past *biraderi* intermarriage (Wang et al, 2000).

Most importantly within the Pakistan context, major national and regional studies have determined that 46.2%-61.2% of marriages in the current generation are consanguineous ($\alpha = 0.0242-0.0332$), with first cousin unions accounting for 40%-50% of all marriages (NIPS, 1992; Bittles et al, 1993; Yaqoob et al, 1993; Hussain and Bittles, 1998). Varying levels of consanguineous marriage have, however, been reported in studies conducted in different communities, ranging from 31.1% ($\alpha = 0.0163$) in urban Swat (Wahab and Ahmad, 1996) to 77.1% ($\alpha = 0.0414$) among Army recruits (Hashmi, 1997), and there also is significant variation in the overall levels of consanguinity observed in different *biraderis* (Shami et al, 1994). In general, consanguineous marriages have been found to be much less common in the small Christian and Hindu communities (Hussain and Bittles, 1998).

Equivalently high levels of consanguineous marriage have been reported in the U.K. Pakistani community that mainly originates from the Mirpur district of Kashmir (Darr and Modell, 1988; Bunday et al., 1990; Corry, 2002). The degree to which consanguinity adversely affects the health of this community continues to arouse controversy in the medical literature, with divergent opinions often expressed (Bunday et al. 1992; Ahmad, 1994; Modell and Darr, 2002; Devereux et al., 2004), and recently the subject also has attracted wider public attention (Dyer, 2005). Unfortunately, discussion of the role of population sub-division is almost always missing from these disputations, despite the fact that within Pakistani society *biraderi* membership can circumscribe, and thus effectively define, the spectrum of mutations within the different constituent gene pools.

POPULATION SUBDIVISION AND GENETIC ASSOCIATION STUDIES IN INDIA AND PAKISTAN

Association/case-control studies implicitly assume the absence of significant confounding variables. With such strong evidence of population stratification, and random and preferential inbreeding in both India and Pakistan, it is very surprising that the potential influence of either caste/*biraderi* membership or consanguinity has been rarely mentioned or apparently considered in studies conducted in either country. This is illustrated in Table 1, using information compiled from a search of PubMed that employed association study, case-control study and clinical trial as keywords, and matching of these terms with caste or *biraderi* and consanguinity. According to the data accessed, the term caste was cited in a maximum of 6.6% Pakistani and 4.2% of Indian studies, with consanguinity listed in a maximum of 4.6% and 0.2% of studies in Pakistan and India, respectively.

In the absence of explicit authors' statements to the contrary, it is unlikely that caste/*biraderi* or consanguinity were included as explanatory variables other than in the small minority of the studies identified in Table 1. It also seems highly improbable that the apparent lack of control for these core variables could have been based on access to pre-existing information demonstrating non-significant effects of caste/*biraderi* membership and consanguinity on allele profiles and frequencies at the loci investigated.

It has been proposed that genetic studies which recruit participants from a small number of geographically proximate sites may be partially protected from the effects of population stratification, and thus benefit from a level of *de facto* control over environmental variation between cases and controls (Foster and Sharp, 2004). However, the scale and multiple levels of internal subdivision that characterize both the Indian and Pakistani populations suggest that this tactic alone would be insufficient to nullify the effects of population subdivisions or potential 'environmental' variables in genetic association studies conducted in either country.

DISCUSSION

From a population perspective the South Asian region is dominated by the populations of India and Pakistan which comprise 20% of the global total, and with large Indian and Pakistani migrant communities now permanently resident abroad. An additional 144 million people live in

Bangladesh and 20 million in Sri Lanka (PRB, 2005). As in the rest of the sub-continent, major ethnic and religious boundaries exist in both of these countries and consanguineous marriage is a widely permissible option in most of their constituent communities.

Under these circumstances, urgent attention needs to be paid to the possible effects of population stratification in genetic association studies, especially since India has a rapidly expanding and increasingly international pharmaceutical industry, with growing numbers of clinical trials conducted and reported. An initiative of this nature will first require a thorough understanding and appreciation of the highly complex demographic and social structures of the numerous constituent sub-populations of the sub-continent and their population genetics, a caveat that equally applies to the overseas migrant communities. Unless and until it can be shown that caste/*biraderi* differentials do not significantly influence patterns of genomic variation in local and regional populations, *e.g.*, by using

Table 1. Frequency of citation of caste, *biraderi* and consanguinity in association studies, case-control studies and clinical trials in Indian and Pakistan*

Country	Keywords	No of papers	Percentage
India	association study	948	
	association study caste	39	4.2
	association study consanguinity	10	1
	case-control study	2030	
	case-control study caste	37	1.9
	case-control study consanguinity	4	0.2
	clinical trial	1381	
	clinical trial caste	8	0.5
	clinical trial consanguinity	1	<0.1
	Pakistan	association study	152
association study caste		10	6.6
association study <i>biraderi</i>		0	0
association study consanguinity		7	4.6
case-control study		381	
case-control study caste		7	1.9
case-control study <i>biraderi</i>		0	0
case-control study consanguinity		15	4.0
clinical trial		166	
clinical trial caste		1	0.6
clinical trial <i>biraderi</i>	0	0	
clinical trial consanguinity	0	0	

*Data collated from PubMed on 8 November 2005

unlinked genetic markers (Pritchard and Rosenberg, 1999), association studies that effectively ignore caste differentials in their recruitment schedules must be open to serious criticism. Conversely, the results of association studies conducted on members of a single caste may have little relevance to other caste and non-caste communities in the wider population.

A similar situation exists with respect to consanguinity, especially in Pakistan where the majority of marriages are preferential intra-familial unions. Consanguinity is thus an important factor in the expression of genes at the family level and, as previously noted, within specific communities it can serve to reinforce the actions of other important population genetic influences, such as founder effect and drift. Failure to control for these genetic variables could result in cryptic relatedness and thus invalidate many of the data that are derived from association studies, with no guarantee either that the results reported are biologically or clinically meaningful or are capable of replication.

These reservations apply equally to the populations of many other populous regions, including West and Central Asia, the Middle East, and North and sub-Saharan Africa, which also retain strong traditions of clan and tribal endogamy and consanguineous marriage (Bittles, 2005; www.consang.net). If pharmacogenetic profiling is to fulfil the promise of highly specific, customized treatment protocols, ongoing advances are clearly needed in our knowledge of the patterns and structure of genetic variation across the genome, assisted by collaborative initiatives such as the HapMap project (McVean et al, 2005). At the same time, there seems to be a major gap in comprehending the importance of the basic parameters that govern the transmission of genes within and between human populations, which makes complementary detailed empirical studies into this topic an equally important and urgent issue.

STATEMENT OF COMPETING INTERESTS

The author declared no competing interests.

REFERENCES

Ahmad WIU. 1994. Reflections on the consanguinity and health outcome debate. *J Pub Hlth Med*, 16, 423-428.

Bamshad MJ, Watkins WS, Dixon ME, Jorde LB, Rao BB, Naidu JM. et al. 1998. Female gene flow stratifies Hindu castes. *Nature*, 395, 651-652.

Basham AL. 1967. Society: class, family and individual. *The Wonder that was India*. New Delhi, Rupa, vol 1, pp. 137-188.

Bhasin MK, Walter H and Danker-Hopfe H. 1992. *The Distribution of Genetical, Morphological and Behavioural Traits among the Peoples of the Indian Region*. Kamla-Raj, Delhi, pp. 14-35.

Bhattacharya NP, Basu P, Das M, Pramanik S, Banerjee R, Roy B et al. 1999. Negligible male gene-flow across ethnic boundaries in India, revealed by analysis of Y-chromosomal DNA polymorphisms. *Genome Res*, 9, 711-719.

Bittles AH. 1998. *Empirical Estimates of the Prevalence of Consanguineous Marriage in Contemporary Societies*. Working Report no. 74, Morrison Institute for Population and Resource Studies, Stanford, Stanford University.

Bittles AH. 2001. Consanguinity and its relevance to clinical genetics. *Clin Genet*, 60, 89-98.

Bittles AH. 2002. Endogamy, consanguinity and community genetics. *J Genet*, 81, 91-98.

Bittles AH. 2005. Endogamy, consanguinity and community disease profiles. *Commun Genet*, 8, 17-20.

Bittles AH, Grant JC and Shami SA. 1993. An evaluation of consanguinity as a determinant of reproductive behaviour and mortality in Pakistan. *Intl J Epidemiol*, 22, 463-467.

Bittles AH and Hussain R. 2000. An analysis of consanguineous marriage in the Muslim population of India at regional and state levels. *Ann Hum Biol*, 27, 163-171.

Bittles AH, Mason WM, Greene J and Appaji Rao N. 1991. Reproductive behavior and health in consanguineous marriages. *Science*, 252, 789-794.

Bundey S, Alam H, Kaur A, Mir S and Lancashire RJ. 1990. Race, consanguinity and social features in Birmingham babies: a basis for a prospective study. *J Epidemiol Commun Hlth*, 44, 13-135.

Bundey S, Alam H, Kaur A, Mir S and Lancashire RJ. 1992. Why do UK-born Pakistani babies have high perinatal and neonatal mortality rates? *Paediat Perinat Epid*, 5, 101-114.

Cardon LR and Bell JI. 2001. Association study designs for complex diseases. *Nature Rev Genet*, 2, 91-99.

Cardon LR and Palmer LJ. 2003. Population stratification and spurious allelic association. *The Lancet*, 361, 598-604.

Carlborg Ö and Haley CS. 2004. Epistasis: too often neglected in complex trait studies? *Nature Rev Genet*, 5, 618-625.

Cavalli-Sforza LL, Menozzi P and Piazza A. 1994. *The History and Geography of Human Genes*. Princeton, Princeton University Press, pp. 210-213.

Census of Mysore. 1871. *Part 1, Report*. Bangalore, Government Central Printing Office.

Cooper RS, Kaufman JS and Ward R. 2003. Race and genomics. *Science*, 348, 1166-1170.

Cordaux R., Auger R., Bentley G., Nasidze I, Sirajuddin SM and Stoneking M. 2004. Independent origins of Indian caste and tribal paternal lineages. *Current Biol*, 4, 231-235.

Corry P. 2002. Intellectual disability and cerebral palsy in a UK community. *Commun Genet*, 5, 210-204.

Dahlman I, Eaves IA, Kosoy R, Morrison VA, Heward J, Gough SCL et al. 2002. Parameters for reliable results in genetic association studies in common disease. *Nature Genet*, 30, 149-150.

Darr A and Modell B. 1988. The frequency of consanguineous marriage among British Pakistanis. *J Med Genet*, 25, 186-190.

Devereux G, Stellitano L, Verity CM, Nicoll A, Will RG and Rogers P. 2004. Variations in neurodegenerative disease across the UK: findings from the national study of Progressive Intellectual and Neurological Deterioration (PIND). *Arch Dis Childh*, 89, 8-12.

Dyer O. 2005. MP is criticized for saying that marriage of first cousins is a health problem. *Brit Med J*, 331, 1292.

Dyson T. 2001. The preliminary demography of the 2001 Census of India. *Pop Dev Rev*, 27, 341-356.

Excoffier L and Hamilton G. 2003. Comment on 'Genetic structure of human populations'. *Science*, 300, 1877.

Foster MW and Sharp RR. 2004. Beyond race: towards a whole-genome perspective on human populations and genetic variation. *Nature Rev Genet*, 5, 790-796.

Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N. et al. 2004. Assessing the impact of population stratification on genetic association studies. *Nature Genet*, 36, 388-393.

Gadgil M and Guha R. 1995. *This Fissured Land: an Ecological History of India*. Delhi, Oxford University Press.

Gadgil M, Joshi NV, Manoharan S, Patil S and Prasad UVS. 1998. Peopling of India. Balasubramaniam D and Appaji Rao N. (Eds) *The Indian Human Heritage*. Hyderabad, Universities Press, pp. 101-129.

González Burchard E, Ziv E, Coyle N, Gomez SL, Tan H, Kartner AJ. et al. 2003. The importance of race and ethnic background in biomedical research and clinical practice. *Science*, 348, 1170-1175.

- Hashmi MA. 1997. Frequency of consanguinity and its effect on congenital malformations – a hospital based study. *J Pak Med Assoc*, 47, 75-78.
- Hussain R. 2005. The effect of religious, cultural and social identity on population genetic structure among Muslims in Pakistan. *Ann Hum Biol*, 32, 145-154.
- Hussain R and Bittles AH. 1998. The prevalence and demographic characteristics of consanguineous marriages in Pakistan. *J Biosoc Sci*. 30, 261-275.
- IIPS. 1995. *National Family and Health Survey, India, 1992-93*. Bombay, International Institute for Population Sciences.
- Kapadia KM. 1958. *Marriage and Family in India*, 2nd ed. Calcutta, Oxford University Press, pp. 117-137.
- McVean G., Spencer CCA and Cahix R. 2005. Perspectives on human genetic variation from the HapMap Project. *PLoS Genet*, 1(4): DOI: 10.1371/journal.pgen.0010054.
- Modell B and Darr A. 2002. Genetic counselling and customary consanguineous marriage. *Nature Rev Genet*, 3, 225-229.
- Mohyuddin A, Ayub Q, Qamar R, Zerjal T, Helgason A, Mehdi SQ, et al. 2001. Y-chromosomal STR haplotypes in Pakistani populations. *Forensic Sci Intl*, 118, 141-146.
- NIPS. 1992. *Pakistan Demographic and Health Survey 1990-1991*. Calverton MD, Macro Systems.
- Overall ADJ and Nichols RA. 2001. A method for distinguishing consanguinity and population substructure using multilocus genotype data. *Mol Biol Evol*, 18, 2048-2056.
- PRB. 2005. *World Population Data Sheet*. Washington DC., Population Reference Bureau.
- Pritchard JK and Rosenberg NA. 1999. Use of unlinked genetic markers to detect population stratification in association studies. *Amer J Hum Genet*, 65, 220-228.
- Qamar R, Ayub Q, Mohyuddin A, Helgason A, Mazhar K, Mansoor A. et al. 2002. Y-chromosomal DNA variation in Pakistan. *Amer J Hum Genet*, 70, 1107-1124.
- Ramana GV, Su B, Jin L, Wang N, Underhill P and Chakraborty R. 2001. Y-chromosome gene SNP haplotypes suggest evidence of gene flow among caste, tribe, and the migrant Siddi populations of Andhra Pradesh, South India. *Eur J Hum Genet*, 9, 695-700.
- Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA. et al., 2002. Genetic structure of human populations. *Science*, 298, 2381-2385.
- Shami SA, Grant JC and Bittles AH. 1994. Consanguineous marriage within social/occupational class boundaries in Pakistan. *J Biosoc Sci*, 26, 91-96.
- Thapar R. 1966. *A History of India*, vol. 1. London, Penguin Books.
- Wahab A and Ahmad M. 1996. Biosocial perspective of consanguineous marriages in rural and urban Swat, Pakistan. *J Biosoc Sci*, 28, 305-313.
- Wang W, Sullivan SG, Ahmed S, Chandler D, Zhivotovsky LA and Bittles AH. 2000. A genome-based study of consanguinity in three co-resident endogamous Pakistan communities. *Ann Hum Genet*, 64, 41-49.
- Yaqoob M, Gustavson K.-H, Jalili F, Karlberg J and Iselius L. 1993. Early child death in Lahore, Pakistan: II. Inbreeding. *Acta Paediat Scand*, 390 Suppl., 17-26.
- Voight BF and Pritchard JK. 2005. Confounding from cryptic relatedness in case-control association studies. *PLoS Genet*, 1(3), DOI: 10.1371/journal.pgen.0010032.
- Zondervan KT and Cardon LR. 2004. The complex interplay among factors that influence allelic association. *Nature Rev Genet*, 5, 89-101.

commercial use, distribution and reproduction of the article, provided the original work is appropriately acknowledged with correct citation details.

SHORT COPYRIGHT STATEMENT

This is an open access article, published under the terms of the Licence for Users available at <http://www.libpubmedia.co.uk/MedJ/LicenceForUsers.pdf>. This licence permits non-