ARTICLE

The use of common mitochondrial variants to detect and characterise population structure in the Australian population: implications for genome-wide association studies

Enda M Byrne*[,1,2], Allan F McRae[1], Zhen-Zhen Zhao[3], Nicholas G Martin[3], Grant W Montgomery[3] and Peter M Visscher[1]

[1]Queensland Statistical Genetics Laboratory, Queensland Institute of Medical Research, Brisbane, Queensland, Australia; [2]School of Medicine, University of Queensland, Brisbane, Queensland, Australia; [3]Genetic and Molecular Epidemiology Laboratories, Queensland Institute of Medical Research, Brisbane, Queensland, Australia

There is an increasing recognition of the potential role of variants in mitochondrial DNA and nuclear-encoded mitochondrial proteins in modifying disease risk. This has led to a rise in the number of mitochondrial association studies being undertaken. The unique inheritance pattern of mitochondria makes mitochondrial DNA variation susceptible to having geographical structure. Such a structure may have a dramatic impact on mitochondrial association studies, particularly in heterogeneous populations. By combining self-reported ancestry data and mitochondrial genotype data for a sample of 3839 individuals from 1037 Australian families, population substructure is tested by looking for evidence of differences in mitochondrial haplogroup and single nucleotide polymorphism (SNP) frequencies between different ancestral groups in Australia. In addition, the substructure within ancestral groups is tested by comparing the similarity of mates to randomly drawn pairs of individuals from the same ancestral group. It is shown that there are significant differences in the frequency of variants both between European and non-European groups, and within Europe. This agrees with previous studies of European mitochondrial variation. No evidence was found for structure within ancestral groups. These results have implications for future association studies in the Australian population, and other populations of heterogeneous ancestry.
European Journal of Human Genetics (2008) 16, 1396–1403; doi:10.1038/ejhg.2008.117; published online 9 July 2008

Introduction

The central role of mitochondria in metabolism, allied to the continued rise in incidence of age-related metabolic and degenerative disease has led to common variants in mitochondrial DNA (mtDNA) becoming increasingly popular as targets in association studies.[1] Whole mitochondria association studies are attractive because of the low number of SNPs needed to be genotyped to get full coverage of all variants with >1% minor allele frequency in the mitochondrial genome.

Mitochondrial DNA has become an important resource for studying the history of populations. The lack of recombination in the mitochondria means that the mitochondrial genome is passed down from mother to offspring completely unchanged except for occasional spontaneous mutations. When mutations occur, they are

*Correspondence: EM Byrne, Queensland Statistical Genetics Laboratory, Queensland Institute of Medical Research, 300 Herston Road, Brisbane QLD 4029, Australia. Tel: +61 7 3362 0272; Fax +61 7 3362 0101; E-mail: enda.byrne@qimr.edu.au

in complete linkage with all other variants in the mtDNA. Over time, sets of linked variants called haplogroups form. Each haplogroup is related to all others through a mutational tree, which can be traced back to a common ancestor. By studying the patterns of haplogroup frequencies in different populations we can get an insight into the history of modern populations.

A number of diseases in which mitochondrial dysfunction has been demonstrated have differing prevalences depending upon the population examined. Genetic drift has led to differences in mtDNA variation between populations, whereas the contribution of selection to this variation is still under dispute.[2–4] For this reason, it is important to control for population substructure when performing mitochondrial association studies to avoid false-positive associations.

It has been known for some time that mitochondrial DNA shows less geographical structuring in Europe than classical markers and the Y chromosome.[5] Furthermore, variation between European populations is smaller compared with populations on other continents after controlling for geographic distance.[6] However, recent studies have shown that when large enough sample sizes are used, there is evidence for a distinct East–West gradient and a much less pronounced North–South gradient.[5,7] In addition, mitochondrial population substructure within European populations has also been demonstrated.[8,9]

A majority of the Australian population are descendents of immigrants from the British Isles, however there has been much migration from other areas of Europe and more recently from Asia. Given this population history, the Australian population may show evidence of population stratification. This would be manifest through there being differences in haplogroup frequencies between people of different geographical ancestry.

Haplogroup-based association tests assume that an associated variant is more likely to occur on a particular background or that the combination of variants that define the haplogroup may contribute to the phenotype. If however, the causative variant occurs on a number of different backgrounds, the haplogroup approach may give false-negative results,[8] and an SNP-based approach would have more power to detect true-positive associations, although they too may be confounded by population substructure.

Testing for differences in haplogroups and SNP frequencies between ancestral groups can inform about population structure, but there may also be population substructure within the ancestral groups. In a substructured population, we would expect that mates will be more likely to come from the same subpopulation, and therefore will be more similar to one another than random pairs from the population. Hatt and Parsons (1965)[10] have previously shown evidence for non-random mating among individuals of British Isles ancestry in Australia. Mates may also be more similar owing to assortative mating, which will also lead to population substructure.

Controlling for population structure is also important in genome-wide association studies (GWAS), and two major methods have been developed that use genotyping data to control structure in such analyses: 'genomic control'[11] and 'structured association'.[12] Population structure becomes more of a confounding factor as sample size increases, and even small levels of stratification can have a dramatic effect in the kind of large-scale studies that are needed to detect the variants of modest effect that have so far been found in GWAS.[13,14] Analysis of mitochondrial DNA variation can contribute to the detection of population stratification in GWAS cohorts.

In this study, we characterise the extent of maternal population structure in an Australian population sample as a precursor to carry out mitochondrial association studies, and show that common mitochondrial variants can be used to detect population structure in genome-wide association studies.

Materials and methods
Sample
In all 3839 individuals from 1037 Australian families were genotyped for a total of 69 mitochondrial SNPs. Families consisted of adolescent twins, their siblings and parents who had been recruited as part of two overlapping studies into the genetic and environmental influences on melanoma risk factors[15] and cognition.[16] Twins between the ages of 12 and 16 and their families were recruited by contacting principals of schools in the Brisbane area and through media appeals. The sample has been shown to be representative of the general Queensland population with respect to mole counts[16] and cognition.[17] Further information on the sample including recruitment strategy can be found in Zhu et al (2007)[18] and Duffy et al (2007).[19] The data set contained 343 pairs of monozygotic (genetically identical) twins, and 594 sets of dizygotic (non-identical) twins. Data for both parents were available in 622 of the families. There were 34 families for whom only the father's genotype was available, and 228 families in which only the mother's genotype was available. The remaining 153 families had no parental data available. There were a total of 345 non-twin siblings included.

Genotyping
Saxena et al (2006)[20] identified all coding region mtSNPs found at a frequency of >1% in European populations and found that all of these variants could be tagged at an $r^2 > 0.8$ by 64 tagging SNPs. We genotyped 61 of the 64 tag SNPs with an additional nine SNPs, which tag variation at $r^2 > 0.8$, that otherwise require the use of multi-SNP haplotypes using a method described in a previous study.[21] The last SNP is a common D-loop variant. Two SNPs were found to be monomorphic in our sample, leaving a total of 69 SNPs. SNPs were typed using iPLEX™ Gold chemistry

and analyzed using a Sequenom MassARRAY Compact Mass Spectrometer (Sequenom Inc., San Diego, CA, USA). A two-step 200 short cycles programme was used for the iPLEX Gold reaction. The products were spotted on a SpectroChip (Sequenom Inc., San Diego, CA, USA), and data were processed and analysed by MassARRAY TYPER 3.4 software (Sequenom Inc., San Diego, CA, USA).

Ancestry

Parents of the twins were asked to list the country of birth and ancestry of their parents and grandparents. In cases where the specific country of birth or ancestry was not known, they were asked to specify a region, for example, The British Isles. For our study, an individual's ancestry was taken as the named ancestry of the maternal great-grandmother, or in cases where this information was not available the maternal grandmother. Assignment to ancestral groups was done according to the criteria listed in Table 1.

Haplogroup assignment

Owing to familial relationships in the sample, there were 1693 independent mitochondrial haplotypes. Haplogroups were assigned according to a previously published method.[17] To ensure greater accuracy in the assignment of haplogroups, haplotypes that contained missing data were discarded for the haplogroup analyses. A total of 1314 haplotypes had no missing data. Thirteen haplotypes could not be unambiguously assigned to a haplogroup, a majority of which were haplotypes of individuals with non-European ancestry (76.9%). The frequency of the haplogroups was calculated for each ancestry group (Table 2).

Between groups structure test

As each individual has been typed for a panel of mitochondrial tagging SNPs designed to tag all common European variation,[20] we would expect to see significant differences in mitochondrial DNA variation between European and non-European groups. However, there may also be differences in haplogroup and SNP frequencies between groups from different regions of Europe. For the most part, we would expect that the SNPs that are found to have differential frequencies between groups will be correlated with one another as they are found in the same haplotype. However, the fact that there is sufficiently low

Table 1 Ancestral groups in the sample countries, which are included in those groups, and the number of individuals assigned to each ancestral group

Ancestral group	Countries[a]	No. of individuals
British Isles	Ireland, Scotland, Wales, England	1227
Northern Europe	Germany, the Netherlands, Denmark, Norway, Sweden, Austria, Iceland	181
Southern Europe	Italy, Spain, Greece	33
Central-Eastern Europe	Poland, Turkey, Serbia, Croatia, Hungary, Romania, Bulgaria, Russia	57
Middle East	Syria, Lebanon, Libya, Jordan, Israel	9
Asia	China, Japan, Thailand, Vietnam, Indonesia, India, Pakistan	25
South Sea Islands	Fiji, Tonga, Papua New Guinea	16
Aboriginal	Australia	4
Unknown		141

[a]Not all countries listed here are represented in the sample.

Table 2 Distribution (%) of inferred haplogroups by ancestry groups

Haplogroup	Total population	UK/ Ireland	Northern Europe	Eastern Europe	Southern Europe	Middle East	Asia	South Sea Islands+Aboriginal	Unknown
H	42.0	43.0	42.9	30.8	65.8	14.3	5.9	6.2	39.4
I	4.1	4.5	3.6	0	0	14.3	0	6.2	3.5
J	11.0	12.3	9.3	0	5.3	0	0	0	8.8
K	8.4	8.9	6.4	0	0	14.3	5.9	0	12.3
M	1.5	0.4	1.4	0	2.6	0	47.0	43.8	3.5
T	11.0	10.9	12.9	27.0	15.8	28.5	0	0	7.0
U	10.4	10.6	12.1	0	5.3	0	11.8	0	11.4
V	5.9	4.9	5.7	23.3	0	0	0	0	7.9
W	2.8	2.1	4.3	11.5	0	14.3	11.8	0	0
X	1.8	1.7	0.1	3.8	2.6	14.3	0	0	3.5
Unknown	1.0	0.7	0.1	3.6	2.6	0	17.6	43.8	0
Total individuals	1314	956	140	26	38	7	17	16	114

linkage disequilibrium between many of these SNPs, all of which needed to be typed to tag most of the variation mitochondrial DNA as shown in McRae *et al*,[17] suggests that an SNP by SNP analysis may give a finer picture of stratification in the sample.

Significant differences in haplogroup frequencies between groups were tested for using a contingency table test for count data in the statistical package R.[22] As the contingency table being analysed contained cells with small values, *P*-values were estimated by Monte Carlo simulation using 100 000 replicates.

Significant differences in individual SNP frequencies between ancestral groups were also tested for using a contingency table test, with a simulated *P*-value. Although the haplogroup-based test excluded individuals with missing data at any SNP, this test did not.

Assortative mating analyses

Each set of parents was tested to see if there is an association between the haplogroups of mates, that is, a pair of individuals who have produced offsprings in the sample. In families where only the father's genotype was available, the mother's genotype was assumed to be identical to the children's. A 10×10 contingency table of haplogroups was constructed (Table 4) and the association was tested for using a contingency table test with a simulated *P*-value. In a substructured population in which there is assortative mating, we might expect that individuals tend to mate with partners who have the same haplogroup as themselves, or that mates will tend to have a particular combination of haplogroups.

Similarity of mates was further tested by comparing all sets of parents at each of the 69 loci. There were 656 sets of parents in the sample. Three tests were performed to test for the genotypic similarity of mates. Firstly, a similarity score was calculated by comparing the number of sites at which the parents had the same allele and dividing by the total number of sites compared. If one of the parents had a missing genotype for an SNP, no comparison was made for that SNP and so the number of SNPs compared decreased by one.

$$S = \frac{\sum_{i=1}^{n} \delta_i}{n}$$

where S is the summary statistic, n is the total number of non-missing SNPs compared between the pair and δ_i equals one if the pair are concordant or δ_i equals 0 if the pair are discordant for the ith SNP.

An overall mean score was calculated from the similarity scores of each pair of parents. A total of 656 pairs of individuals were then drawn randomly from the population of all parents and a similarity score calculated for each. This was done 1000 times and each time an overall mean was calculated. The mean of the parents was then compared with the distribution of means for the randomly drawn pairs to calculate the significance.

The above test fails to take into account the sampling variance of each allele that is the probability of a pair sharing a given allele is dependent upon the allele frequency. To take into account the sampling variance and the differing allele frequencies, the test was performed again, but each time the parents share the same allele, δ_i is multiplied by the inverse of the variance, $\frac{1}{H_i(1-H_i)}$ where $H = 2p\,(1-p)$, the probability of the pair being discordant for an allele. This is then divided by the inverse sum of the variance of all SNPs compared for each pair.

$$T = \frac{\sum_{i=1}^{n} \frac{\delta i}{H_i(1-H_i)}}{\sum_{i=1}^{n} \frac{1}{H_i(1-H_i)}}$$

where T is the weighted sum statistic, n is the total number of non-missing SNPs compared between the pair, and δ_i equals one if the pair are concordant or δ_i equals 0 if the pair are discordant for the ith SNP.

A third test was carried out which takes into account the significance of sharing a rare variant. Each SNP is weighted by frequency by multiplying δ_i by $\frac{1}{\sqrt{p}}$ where p is the allele frequency of the shared allele.

The analyses were then repeated using only pairs drawn from the British Isles. There were 380 such pairs. There were insufficient number of pairs from other ancestral groups to perform the analyses.

Results

Table 1 shows the distribution of self-reported maternal ancestry in our sample. Given the history of immigration to Australia, it is not unexpected to find that a vast majority of the sample have ancestry from the British Isles.

Figure 1 shows the distribution of minor allele frequencies in the sample. SNPs were selected for genotyping if they have minor allele frequencies $>1\%$ frequency in
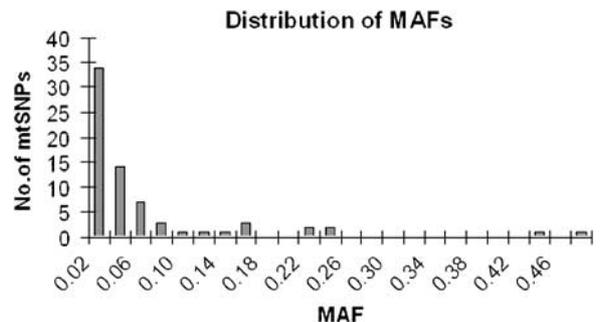

Figure 1 The distribution of minor allele frequencies of mtSNPs in the sample.

Europeans. There are 16 SNPs with a minor allele frequency of 5% or more.

Table 2 shows the distribution of inferred haplogroups by ancestry in our sample, and these agree with those previously reported in the Australian population.[23] Haplogroup H, the most common haplogroup in Europe, is also the most common haplogroup in the Australian population, with a frequency of 42.0%. The next three most common haplogroups J, T and U, together comprise a further 32.4% of the population.

As would be expected, a significant difference in the distribution of haplogroup frequencies between individuals of European and non-European ancestry was found ($P = 0.00001$). When the four European groups are compared, there is evidence for a significant difference in haplogroup frequencies between them ($P = 0.00012$). Most of this difference appears to be caused by differences between Western and Eastern populations. The UK and Ireland, Northern European and Southern European populations are all significantly different from the Eastern European group ($P = 0.002, 0.0016, 0.001$, respectively). This observation is in agreement with a previous study which showed that the first principal component of variation in haplogroup frequencies in Europe is along an east–west axis.[5] Even with the small sample size of Eastern Europeans in our sample, the lower frequency of haplogroup H in that region can be detected. There is also evidence for a difference in frequencies between Northern and Southern populations, although these differences only approach significance ($P = 0.06$). There is no evidence for differences between the British Isles and Northern European groups ($P = 0.48$). It is clear that there is significant evidence for population substructure among European ancestral groups in Australia.

Our SNP panel is composed of SNPs that tag all coding region variants found within Europe at a frequency $>1\%$,

so we would expect to see significant differences in the allele frequencies between the European and Asian groups. Using Fisher's exact test, we find that several SNPs (Table 3) are significant even after correction for multiple testing (5% significance with 69 tests $= 0.0007$). Bonferroni correction is an appropriate strategy here because of the low LD between markers.

Several of these SNPs define major European haplogroups. mt4216 tags haplogroups J and T, whereas mt7028 defines haplogroup H, the most common European haplogroup. A number of the other SNPs tag super haplogroup M, which is found primarily in Asia. Some of these SNPs have low frequencies in the European population (marked in the haplogroups column in Table 3) but are much more variant in the South Sea Island and Aboriginal group.

The same analysis was performed on only the European ancestral populations. When all four populations are included, five SNPs are found to be significant (Table 3).

None of these SNPs are significant after Bonferroni correction. However, mt10398, mt14798, and mt12372 approach significance. The three SNPs that approach significance are strong predictors of haplogroups in our sample. The mt10398A allele is found at a much lower frequency in the Eastern European group relative to the other European populations. It is a predictor of haplogroup T, which is found to be enriched in Eastern Europeans in our sample. The mt10398G allele is found in haplogroups U and I, both of which are not found in the Eastern European sample. The mt12372A allele is found only in haplogroups U and K, both of which are almost absent in Southern and Eastern Europeans. The mt14798C allele is found in low frequency in individuals of Southern or Eastern European ancestry. This allele is found in individuals with haplogroups J and K, both of which are at much lower frequency in Eastern and Southern Europe. It is clear

Table 3 mtSNPs that have significant frequency differences between European and non-European groups and that are significantly different between European groups

SNP	Europe–non-Europe P-value	Between European groups P-value	Haplogroup(s) tagged
mt4216	5.7×10^{-03}	—	T
mt5465[a]	8.1×10^{-13}	—	—
mt6719[a]	2.8×10^{-11}	—	—
mt7028	1.7×10^{-06}	0.024	H
mt9123[a]	1.1×10^{-08}	—	—
mt10238	1.3×10^{-06}	—	M+W
mt11719	9.0×10^{-08}	—	V
mt12705	1.8×10^{-07}	—	M
mt15043	5.5×10^{-06}	—	M
mt16189[a]	1.9×10^{-05}	—	—
mt8705	—	0.010	—
mt10398	—	0.003	T
mt12372	—	0.004	U+K
mt14798	—	0.003	J+K

[a]SNPs that are found at low frequency in Europeans but high frequency in Aboriginal and South Sea Island populations.

that the SNP by SNP analysis complements the haplogroup analysis. The SNPs that have significant differences in frequency between groups are also those that contribute to the significant differences in haplogroup frequencies.

Assortative mating analyses

The mean similarity score for all pairs is 0.9105 (SD = 0.046, SE of mean = 0.0018). The mean score for the 1000 means of randomly drawn pairs is 0.9092 (SE = 0.002). Comparing the mean of the true mates to the empirical distribution of random pair means gives a *P*-value of 0.22, providing no evidence of increased allele sharing among mates. This analysis included all pairs of mates regardless of their population of ancestry. We would assume that in a population in which there is stratification that mates will more likely be from the same subpopulation. As such, there will be a higher probability that a randomly drawn pair will be from different populations and hence, will share less alleles in common. For this reason, we might expect that mates will have more similar genotypes than a set of randomly drawn pairs. However, this is not what is found.

When the sampling variance is taken into account, the mean similarity score for mates is 0.9741 (SD = 0.018, SE of mean = 0.0007), whereas the mean score of the replicates for random pairs is 0.9729 (SE = 0.001). The empirical *P*-value is 0.11. For the third test that takes allele frequencies into account, the mean similarity score for mates was 0.9459 (SD = 0.07, SE = 0.0027) and the mean similarity score of random pairs is 0.9435 (SD = 0.002), and the *P*-value is 0.07.

We then restricted the analysis to only mates and randomly drawn pairs from the British Isles. As there was no evidence of greater similarity among mates when all populations were included, we expect to find no evidence when we restrict it to one ancestral population. This is because randomly drawn pairs are more likely to be similar as they come from the same population. The mean similarity score for mates was 0.9093, and the mean of the 1000 samples of randomly drawn pairs is 0.9092 (SD = 0.002). The empirical *P*-value was 0.48. There is no evidence for increased allele sharing among mates of British Isles ancestry using this test. There is also no evidence for increased sharing when the tests take sampling variance (*P* = 0.52) and allele frequency into account (*P* = 0.17).

Mates were also compared to see if they tend to share the same haplogroup or a pair of haplogroups more often than would be expected by chance. A 10 × 10 table of male and female haplogroups was constructed (Table 4). The association between haplogroups was tested for using Fisher's exact test.[24] No significant association was found (*P* = 0.784).

It is clear that there is little evidence to suggest that, in our sample, individuals tend to mate with individuals whose mitochondrial genotype is more similar to their own.

Table 4 Assigned haplogroups of all pairs of parents in the sample

Female Male	H	I	J	K	M	T	U	V	W	X
H	64	11	22	21	3	19	19	12	3	3
I	5	1	2	0	1	5	1	0	1	0
J	15	2	3	6	0	5	3	3	1	1
K	15	2	3	3	2	4	8	3	1	1
M	2	1	0	0	1	0	1	0	0	2
T	18	2	5	3	1	6	7	2	4	2
U	22	1	7	2	1	5	5	4	0	0
V	12	1	2	1	3	4	3	2	2	0
W	2	0	2	0	0	1	1	1	0	0
X	4	1	0	0	0	1	1	1	0	0

Discussion

The epidemic of metabolic disease in Western countries and increasing emphasis on aging research has lead to increasing focus on mitochondrial variants as potential modifiers of disease risk. The data gleaned from mitochondrial association studies will also be very useful for population genetics analysis. As a precursor to carrying out association studies, we have investigated whether any population substructure can be detected in the maternal line of the Australian population.

Our study is limited by the small sample size of some of the ancestral groups, but we believe our sample to be representative of the ancestral make-up of the Australian population. Although most of the population trace their ancestors back to the British Isles, a much smaller number are of native Australian ancestry, and there are large numbers of smaller groups whose ancestors came to Australia from various regions of the world. Even with small sample sizes, significant differences can be found in the distribution of mitochondrial variants in different populations. In populations of European ancestry, these differences follow the same trend as seen previously in Europe.

These analyses demonstrate that population substructure may be a confounding factor in mitochondrial association studies in the Australian population. It would be advisable to get an estimate of population structure in the sample and to adjust the test accordingly. In this study, we have only analysed the frequencies of the major European haplogroups, many of which can be subdivided into subhaplogroups. Analysis of these subhaplogroups may lead to the detection of a finer scale substructuring in the population.

There are clear implications for mitochondrial association studies in the Australian population, but there are also implications for mitochondrial association studies in any

population that has had a significant immigration or is admixed. Although the non-recombining nature of mitochondria means that differences between populations are more pronounced and easily detected when looking at mitochondrial variants, there may also be significant implications for nuclear association studies. An equivalent study using autosomal SNPs is much more difficult; however, as it requires accurate ancestry information for a larger number of individuals in each participant's family, and only leads to an estimate of what proportion of the genome is admixed, in cases where an individual's parents or grandparents are of differing ancestry. For association studies in the Australian population, it may be necessary to subdivide people of European ancestry when it comes to matching cases and controls.

In addition to the stratification analysis, the correlation between mtSNPs and nuclear SNPs (nSNPs) was examined (results not shown). Although they have differing modes of inheritance, it is likely that there are significant interactions between variants in nuclear and mitochondrial DNA. Mitochondrial DNA codes for 13 proteins that are subunits of four of the five large complexes, are essential for oxidative phosphorylation. As such, proteins that are encoded in the mitochondria do not act independently and efficient oxidative phosphorylation requires interaction with nuclear encoded proteins.[25] These interactions have led to adaptive coevolution between the two genomes.[26] This coevolution may have led to certain combinations of nuclearDNA (nDNA) and mtDNA variants being highly correlated in the population. Even in the absence of selective pressures, we would expect that mitochondrial variants that define region-specific haplogroups would be correlated with ancestry informative nSNPs. Therefore, the correlation between mitochondrial and nuclear variants may have implications for the design of association studies, and can give an insight into the relationship between nuclear and mitochondrial variants.

A recent large-scale case–control study in the United Kingdom[14] showed that it is possible to separate individuals from the North and South of the country when a large enough panel of SNPs is used. In addition, nonrandom mating was previously demonstrated among individuals of British Isles ancestry in Australia. As the data set does not contain information on which region of each country individuals can trace their ancestry to, it is difficult for us to test within the population substructure using mitochondrial data. We have performed a crude test for population structure by testing whether mates tend to be more similar to one another at the genetic level than random pairs drawn from the population. In a structured population, we expect that individuals from the same subpopulation will be more similar to one another at the genetic level. Although when examining mitochondria, the genotype of the father has no bearing on that of his children, his sons would still be more likely to mate with someone from the same population, and will therefore have a similar mitochondrial genotype to himself. We find no evidence for increased sharing among mates, even after correction for allele frequency. The only ancestral group for which there was a large enough sample size to perform this test was the British Isles group. It is possible that other groups will show evidence of increased sharing among mates if a large enough sample size is examined. Despite being unable to detect any within-group stratification in our sample, it may still be a confounding factor in mitochondrial association studies. The knowledge gleaned from these analyses will greatly aid the interpretation of association studies that will be carried out on the same sample.

References

1 Wallace DC: A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. *Annu Rev Genet* 2005; **39**: 359–407.
2 Mishmar D, Ruiz-Pesini E, Golic P *et al*: Natural selection shaped regional mtDNA variation in humans. *Proc Natl Acad Sci USA* 2003; **100**: 171–176.
3 Ruiz-Pesini E, Mishmar D, Brandon M, Procaccio V, Wallace DC: Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science* 2004; **303**: 223–226.
4 Ingman M, Gyllensten U: Rate variation between mitochondrial domains and adaptive evolution in humans. *Hum Mol Genet* 2007; **16**: 2281–2287.
5 Richards M, Macaulay V, Torroni A, Bandelt HJ: In search of geographical patterns in European mitochondrial DNA. *Am J Hum Genet* 2002; **71**: 1168–1174.
6 Comas D, Calafell F, Mateu E, Perez-Lezaun A, Bosch E, Bertranpetit J: Mitochondrial DNA variation and the origin of Europeans. *Hum Genet* 1997; **99**: 443–449.
7 McEvoy B, Richards M, Forster P, Bradley DG: The longue duree of genetic ancestry: multiple genetic marker systems and Celtic origins on the Atlantic façade of Europe. *Am J Hum Genet* 2004; **75**: 693–702.
8 Pereira L, Goncalves J, Goios A, Rocha T, Amorim A: Human mtDNA haplogroups and reduced male fertility: real association or hidden population substructuring. *Int J Androl* 2005; **28**: 241–247.
9 Feder J, Ovadia O, Glaser B, Mishmar D: Ashkenazi Jewish mtDNA haplogroup distribution varies among distinct subpopulations: lessons of population substructure in a closed group. *Eur J Hum Genet* 2007; **15**: 498–500.
10 Hatt D, Parsons PA: Association between surnames and blood groups in the Australian population. *Acta Genet* 1965; **15**: 309–318.

11 Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 1999; **55**: 997–1004.
12 Pritchard JK, Donnelly P: Case–control studies of association in structured or admixed populations. *Theor Pop Biol* 2001; **60**: 227–237.
13 Marchini J, Cardon LR, Phillips MS, Donnelly P: The effects of human population structure on large genetic association studies. *Nat Genet* 2004; **36**: 512–517.
14 The Wellcome Trust Case–Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. *Nature* 2007; **447**: 661–678.
15 Zhu G, Duffy DL, Eldridge A *et al*: A major quantitative-trait locus for mole density is linked to the familial melanoma gene CDKN2A:a maximum-likelihood combined linkage and association analysis in twins and their sibs. *Am J Hum Gen* 1999; **65**: 483–492.
16 Wright MJ, Martin NG: Brisbane Adolescent Twin Study: outline of study methods and research projects. *Aust J Psychol* 2004; **56**: 65–78.
17 McRae AF, Byrne EM, Zhao ZZ, Montgomery GW, Visscher PM: Power and SNP typing in whole-mitochondrial genome association studies. *Genome Res* 2008; **18**: 911–917.
18 Zhu G, Montgomery GW, James MR *et al*: A genome-wide scan for naevus count: linkage to CDKN2A and to other chromosome regions. *Eur J Hum Genet* 2007; **15**: 94–102.

19 Duffy DL, Montgomery GW, Chen W *et al*: A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. *Am J Hum Genet* 2007; **80**: 241–252.
20 Saxena R, de Bakker PI, Singer K *et al*: Comprehensive association testing of common mitochondrial DNA variation in metabolic disease. *Am J Hum Genet* 2006; **79**: 54–61.
21 Zhao ZZ, Nyholt DR, Le L *et al*: KRAS variation and risk of endometriosis. *Mol Hum Reprod* 2006; **12**: 671–676.
22 R Development Core Team: *R: A language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2005; available at http://www.R-project.org.
23 Manwaring N, Jones MM, Wang JJ, Rochtchina E, Mitchell P, Sue CM: Prevalence of mitochondrial DNA haplogroups in an Australian population. *Aust Int Med J* 2006; **36**: 530–533.
24 Fisher RA: On the interpretation of χ^2 from contingency tables, and the calculation of *P*. *J R Stat Soc* 1922; **85**: 87–94.
25 Schmidt TR, Wu W, Goodman M, Grossman LI: Evolution of nuclear- and mitochondrial-encoded subunit interaction in cytochrome c oxidase. *Mol Biol Evol* 2001; **18**: 563–569.
26 Bayona-Bafaluy MP, Mueller S, Moraes CT: Fast adaptive coevolution of nuclear and mitochondrial subunits of ATP synthetase in orang-utan. *Miol Biol Evol* 2005; **22**: 716–724.