Population Differences in the International Multi-Centre ADHD Gene Project

Benjamin M. Neale1, Pak C. Sham16,17, Shaun Purcell18,19, Tobias Banaschewski4, Jan Buitelaar5, Barbara Franke5,6, Edmund Sonuga-Barke8, Richard Ebstein10, Jacques Eisenberg11, Aisling Mulligan12, Michael Gill12, Iris Manoir10, Ana Miranda7, Fernando Mulas7, Robert D. Oades3, Herbert Roeyes13, Aribert Rothenberger4, Joseph A. Sergeant9, Hans-Christoph Steinhausen14, Eric A. Taylor3, Margaret Thompson8, Kaixin Zhou1, Philip Asherson1 and Stephen V. Faraone2,15


Abstract:
The International Multi-Centre ADHD Gene sample consists of 674 families from eight countries (Belgium, England, Germany, Holland, Ireland, Israel, Spain, and Switzerland) ascertained from clinics for combined-type attention deficit hyperactivity disorder in an offspring. 863 SNPs were successfully genotyped across 47 autosomal genes implicated in psychiatric disorders yielding a single nucleotide polymorphism (SNP) density of about one SNP per 2.5 kb. A global test of heterogeneity showed 269 SNPs were nominally significant (expected 43). Inclusion of the Israeli population accounted for about 70% of these nominally significant tests. Hardy-Weinberg equilibrium tests suggest that combining all these populations would induce stratification, but that the Northern European populations (Belgium, England, Germany, Holland, and Ireland) could be appropriate. Tag SNPs were generated using pair-wise and aggressive tagging from Carlson et al. [2004] and de Bakker et al. [2005], respectively, in each population and applied to the other populations. Cross-population performance across Northern Europe was consistent with within-population comparisons. Smaller sample size for each population tended to yield more problems for the generation of aggressive tags and the application of pair-wise tags. Any case-control sample employing an Israeli sample with Northern Europeans must consider stratification. A Northern European tag set, however, appears to be appropriate for capturing the variation across populations.

Key words: linkage disequilibrium; tagging; HapMap

Contract grant sponsor: NIH; Contract grant number: R01 MH62873.

Correspondence to Benjamin M. Neale, MRC Social Genetic Developmental and Psychiatry Centre, Institute of Psychiatry, London, United Kingdom. E-mail: b.neale@iop.kcl.ac.uk
INTRODUCTION

Population differences have generated a great deal of confusion in case-control association studies. In particular, population stratification can yield both positive and negative bias in association findings. As a result, family-based designs such as the Transmission Disequilibrium Test were developed [Spielman et al., 1993]. For case-control analyses, geneticists have created methods based on background markers such as genomic control, structure, and L-POP [Devlin and Roeder, 1999; Pritchard et al., 2000; Devlin et al., 2001; Purcell and Sham, 2004]. At the same time, association analysis has become more sophisticated with researchers now routinely employing haplotype-based methods, tagging single nucleotide polymorphisms (SNPs), and general patterns of linkage disequilibrium (LD) [Daly and Altshuler, 2005; Kruglyak, 1999; Palmer and Cardon, 2005; Risch, 2000]. These methods aim to improve efficiency by removing redundant genotyping and gather additional association information indirectly via LD, although the latter has also raised questions about the nature of population differences such as the consistency of LD structures across populations and the similarity of observed haplotypes [van den Oord and Neale, 2003; Evans and Cardon, 2005; Gonzalez-Neira et al., 2006].

A number of studies have examined the consistency of LD information in European samples. A key component of such work is tag selection from the LD information of the Centre d’Etude du Polymorphisme Humain (CEPH) sample. Mueller et al. densely genotyped four gene regions (PLAU, SNCA, LMNA, and FKBP5) at a frequency of one SNP per ~2–4 kb in strictly European populations to gauge populations differences [Mueller et al., 2005]. In general, mild allele frequency differences were detected. Aside from basic allele frequency differences, Mueller also looked at haplotype frequency and block boundary position as outcome measures, broadly noting that geographical distance correlated quite strongly with observed differences between the two populations. From a tagging perspective, CEPH samples captured variation accurately in two genes, and performed poorly in the other two, indicating variability in the patterns of tagging. de Bakker and colleagues recently examined the extent to which population differences exist in a comparison of various populations. For cross-European tagging, the tags from the CEPH sample were assessed in a Finish sample from the multiethnic cohort study. Generally, the power to detect association from tag SNPs is maintained in the Finnish population [de Bakker et al., 2006]. Nejentsev et al. examined the vitamin D receptor gene in four European populations and an African population. All of the European populations showed dramatically similar LD properties in the gene, and that tags generated from the CEPH sample performed reasonably well. The authors conclude that the similarity of LD is consistent with the model predicting no major founder effect for the Europeans [Nejentsev et al., 2004]. Recently, Ribases et al. published an examination of population differences between a Spanish sample and the CEPH sample. The correlations for allele frequency and LD were 0.91 and 0.95, respectively. Baucet and colleagues explore the population structure of Europeans as compared to the CEPH and Coriell samples. Broadly, Europe splits into two main sets, a northern (Polish, Irish, English, Germans, and some Italians) and a south-eastern (Greeks, Armenians, Jews, and some Italians) [Bauchet et al., 2007]. Finally, Evans and Cardon compared the recombination rates across a region in chromosome 20 in the CEPH and a UK
sample. The results were quite similar for the two populations, so the close ancestry appears to predict similarity in LD structure [Evans and Cardon, 2005].

In this paper, we present analysis of the similarities and differences of populations across Europe and Israel from the International Multi-Centre ADHD Gene (IMAGE) sample, exploring 47 autosomal genes, many of which have been implicated in psychiatric disorders.

**METHODS**

The IMAGE sample used in this study has been detailed elsewhere [Brookes et al., 2006]. In brief, the sample is composed of 674 families collected in eight countries from 11 clinical centers: Belgium, Germany, Holland, Ireland, Israel, Spain, Switzerland, and United Kingdom. We considered our sample by looking at each ascertainment centre separately. As Germany, Holland, and Israel each had two centers, for some analyses we combined these sets of families, yielding eight sets of data (with number of families in parentheses): Swiss (24), German (291805109), Spanish (52), Israeli (261 975123), Dutch (841105194), Belgian (22), England (98), and Ireland (52).

Ethical approval for the study was obtained from Ethical Review Boards within each country and informed consent obtained for the use of the samples for analyses relating to the genetic investigation of attention deficit hyperactivity disorder (ADHD). All ADHD probands and their siblings were aged 5–17 years at the time of entry into the study and access was required to one or both biological parents for DNA collection. Diagnosis of ADHD was standardized at the outset of the sample collection, so all probands are based on the same phenotypic definition. Exclusion criteria applying to both probands and siblings included autism, epilepsy, IQo70, brain disorders, and any genetic or medical disorder associated with externalizing behaviors that might mimic ADHD. DNA was available for both parents in 598 families (88.7%) and from one parent in 73 families (10.8%). 664 (85.6%) of the ADHD cases were males. The age range for both probands and siblings was 5–17 years with a mean age of 11.2 years (SD 5 2.7) for probands and 11.2 years (SD 53.1) for siblings.

Gene selection focused on “biological systems” by nominating 45 genes that were likely to exert an effect through regulation of dopamine, serotonin, and norepinephrine neurotransmission. Six additional genes were selected for their role in circadian rhythm. Of these, four genes were sex-linked and excluded from this analysis.

For SNP selection, we aimed at a comprehensive analysis of each gene from two perspectives: functionality and tagging. For the functional markers, we targeted SNPs located within coding regions (synonymous and non-synonymous), 50 and 30 untranslated regions, intron sequences within 300 bp from intron/exon boundaries, and one SNP per kb covering 5 kb upstream from the start of transcription or known 5 regulatory regions including the premotor. For tagging we selected a non-redundant set of tagging SNPs that showed r2 > 0.80 with SNPs with minor allele frequency (MAF) ≥ 0.05, using the CEPH panel from the HapMap [2003] database. Tagging SNPs were selected using two methods. We used the “CompleteLink routine” within CLUSTAG that implements a hierarchical clustering algorithm (hkumath.hku.hk/web/link/CLUSTAG/CLUSTAG.html) [Ao et al., 2005]. The second method used the default algorithm in Haploview (www.broad.mit.edu/mpg/haploview) from Gabriel et al. [2002]. 39.6% (173 out of 437 tSNPs) of the
CLUSTAG SNPs and 71.8% (173 out of 241 tSNPs) of the Gabriel method SNPs were shared. 7.0% (77 out of 1,105, functional SNPs) were included in the tag criteria. To avoid redundancy of the marker information, where the two methods recommended selection of two different SNPs that fell within the same cluster defined by CLUSTAG, we preferentially selected the tagging SNP recommended by the CLUSTAG algorithm. If the constraints of the Illumina technology prevented the use of an identified tagging SNPs nominated by CLUSTAG, then an alternative SNP located in the same cluster was selected. At the time of SNP selection, it was unclear what the most effective approach for tagging a region is. As a result, we aimed to err on the side of SNP inclusion to ensure that the variation in each gene was well characterized.

Blood samples were sent to Rutgers University Cell and DNA repository, New Jersey (RUCDR). These were either used to generate lymphocyte cell lines from which DNA was extracted, or DNA was extracted directly from a portion of the blood sample and lymphocytes cryopreserved for future recovery. In a few cases where individuals were not able to supply a blood sample, we used a mouth swab sampling technique and extracted the DNA at the SGDP laboratory in London [Freeman et al., 2003]. DNA stocks for the entire dataset were collated in London where they were stored, organized and plated out for further analysis. Geneservice Ltd Cambridge (UK) performed whole-genome amplification on all samples with o100 mg stock DNA, using the REPLI-g kit (Qiagen Ltd, Crawley, Sussex). DNA samples were arrayed into 96-well plates at a concentration of 50 ng/mL and delivered to Illumina Inc. (San Diego, CA).

SNP genotyping was completed with Illumina high-throughput BeadArrayTM technology (http://www.illumina.com). We set a lower limit for MAF of 0.05, but also included a high proportion of non-validated SNPs and SNPs with unknown heterozygosity from “functional” regions. Additionally, we excluded variation on the sex chromosomes. For this analysis we finished with a set of 863 SNPs across 47 genes. The average SNP density was approximately one SNP every 2.5 kb.

Ten samples of the total of 2,937 samples (<0.04%) could not be genotyped even after multiple attempts and were removed from the dataset. For the remaining samples the average genotype drop out rate was 0.02% (not including the 34 markers that failed on the whole-genome amplification DNA). The overall genotyping error was estimated to be lower than 0.065% (0.06%+0.005%). We identified 47 families with potential pedigree errors using PEDCHECK [O’Connell and Weeks, 1998]. Within these families the numbers of Mendelian errors ranged from 16 to 312 and could not be explained by genotyping errors. Forty errors could be corrected by considering different familial relationships but seven of these families exhibited unknown pedigree errors and were removed from the analysis. The final dataset for association analysis included 674 families that contained 156 sporadic Mendelian errors from 987 autosomal markers. The overall detection rate of Mendelian errors is therefore 0.02%, which is consistent with the estimated efficiency of SNP markers [13–75%; Douglas et al., 2002] to detect such errors by identification of Mendelian errors and an estimated overall genotype error rate o0.065%. Pedigree errors may be explained by the complicated procedures inherent within a large international cooperative process. We also imposed a threshold of the Illumina quality control score of 0.5 for all SNPs.
We employed three approaches to determine the level of similarity across the sub-populations; heterogeneity tests, Hardy-Weinberg equilibrium (HWE) tests, and tag SNPs comparisons (the number of tags and the cross-population efficiency). Selecting only markers that were polymorphic in all populations, a w2 test was conducted on the allele counts by centre site and population group. The w2 was conducted IMAGE Project 3 Genet. Epidemiol. DOI 10.1002/gepi using the program R 2.1.0, and the significance was simulated using 2,000 replicates for each test. Each center site or population group was subsequently dropped from the heterogeneity analysis, to determine the effect on the distribution of results. In addition to global heterogeneity analysis, pair-wise analysis was conducted for every possible pair of populations.

The goal of the heterogeneity analysis is to assess allele frequency differences across the populations. The global test of heterogeneity determines whether the entire set of populations can be grouped together, and by extension whether combined association analysis is appropriate. The pair-wise tests aim to illuminate the relative similarity between populations. In particular, the question of which populations can be grouped together with little cost to association analysis. These pair-wise tests are not attempting to find significant differences, but rather explore how similar these various populations are. Furthermore, r2 is affected by allele frequency, so allele frequency differences are an important factor for consideration in consistency of LD information. As a convenient way to interpret the results broadly, we report the number of SNPs with global heterogeneity test P value of 0.05. The expectation for the number of tests significant at this level is 5%, so for 863 SNPs, we expect to observe about 43 differences significant at the 0.05 level.

HWE tests were conducted in each subpopulation and combined within countries. PLINK was used to calculate the HWE test, using the w2 test [Purcell et al., 2007]. Consistent deviations from Hardy-Weinberg are a hallmark of population stratification.

Using the Tagger program implemented in Haploview, we determined the number of tag SNPs to capture all information across the region at an r2 of 0.8 or better [Carlson et al., 2004; Patil et al., 2001]. We employed both “pair-wise tagging” from Carlson et al. and the “aggressive tagging” from de Bakker et al. considering all variation [Carlson et al., 2004; de Bakker et al., 2005]. We examined the average number of tags necessary across all populations. We elected to use the number of tag SNPs rather than specific tag SNPs as mild variation in the observed r2 can cause different SNP selection. Additionally, the number of SNPs rather than specific SNPs ought to serve as an indication of the amount of LD across each region. The efficacy of the tags selected is measured by the cross-population tagging performance. Each population was used as a training set to generate tags and then those tags were tested in every other population. Both tagging approaches outlined above, pair-wise and aggressive, were employed.

RESULTS

Comparing all population centers jointly yielded 269 out of 863 SNPs with a P-value of 0.05. A kernel density estimator plot of the distribution of P-values can be found in Fig.1. Clearly, the overall distribution of tests is biased heavily toward smaller P-values. Such a result is an indication of major population differences in allele frequencies. Allele frequency differences are a major component of population stratification, and important in the consideration of association evidence.
Table I shows the number and percentage of tests significant at a nominal 0.05 P-value level for all pair-wise population comparisons and for the global test conditional on dropping each centre. In instances of multiple centers in a country, these centers were combined and dropped. Fig. 2 shows the distribution of HWE tests for each population, and country. Additionally, global and Northern European (Belgium, England, Holland, Germany, and Ireland) sets were tested. Extensive data cleaning has already taken place (see above), so some evidence for differences may have already been removed.

Fig. 3 shows the HWE tests for the Northern European set. These 2 figures show the radically different amount of deviation from HWE observed in the combined sample, versus the Northern European sample. Deviation from HWE is a hallmark of population stratification, and so these results further confirm the relative difference of the Israeli populations from the Northern Europeans.

Table II shows the number of SNPs necessary to characterize the entire set of markers using both the pair-wise and aggressive tagging approaches. For both of these approaches, a minimum R2 threshold of 0.8 for all markers in the region was set. The correlations between the number of tag SNPs and sample size are 0.28 and 0.81 for pair-wise and aggressive tagging, respectively.

Tables III contains the cross-tagging performance in terms of percentage of

![Fig.1. Distribution of p-values for 863 markers in population heterogeneity analysis.](image)
Fig. 2. Centre and country Hardy-Weinberg test distributions. Each centre’s distribution of Hardy-Weinberg tests. For Germany, Holland, and Israel the combined plot refers to pooling the samples and recalculating all the Hardy-Weinberg tests.

Fig. 3. Distribution of Hardy-Weinberg tests in the Northern European and all populations. The distribution of Hardy-Weinberg tests across the 863 markers for Northern Europe and all populations combined. Northern Europe is composed of Belgium, Germany 1 and 2, Holland 1 and 2, England, and Ireland.

SNPs captured by the pre-defined tagging set. Across the datasets, we have a range of sample sizes. In general, larger populations show more pair-wise tags and fewer aggressive tags. The correlation between tag number and sample size is considerably stronger in the aggressive tagging approach.
DISCUSSION AND CONCLUSION

As shown in Table I, the allelic heterogeneity analysis shows clear allele frequency differences between the Israeli populations and the rest of the European populations. Further, allele frequency comparisons between Northern European populations show considerably less heterogeneity. The Northern European populations may still hold population differences at a finer level which are not picked up, given the sample size and subset of all genome variation genotyped in our analysis. As indicated by Bauchet et al. South-eastern and Jewish populations ought not to be grouped together with Northern European populations as there are significant differences [Bauchet et al., 2007].

The HWE tests presented in Figure 2 show that each country is largely behaving in an appropriate fashion from a genotype distribution perspective. Additionally, the HWE tests suggest that the Northern Europeans are largely similar given the relative uniform nature of the P-value distribution.

For the cross-tagging performance, the three countries with two centers each can be considered a convenient benchmark for comparison. Accordingly, the Belgian, English, German, Dutch, Irish, and to a lesser extent Swiss appear similar for pairwise tagging, with the Spanish, and Israelis more distinct. Other work in this vein indicates that the CEPH sample, which is of Northern European ancestry, provides a good tagging set [de Bakker et al., 2006; Gonzalez-Neira et al., 2006; Nejentsev et al., 2004]. However, the sample size appears to be the distinguishing factor in terms of performance when using aggressive tagging. Considering the two centres in each population, the smaller populations perform as better training sets for the larger populations for pairwise tagging, than vice versa. One possible reason is that the error for the estimate of the allele frequency is greater in smaller populations. As a result, rare SNPs are more likely to be included and more common SNPs are more likely to be excluded from the smaller populations. Aggressive tagging suffers more in cross-population comparisons because of the increased importance of LD information.

Work on population differences using measures of similarity such as heterogeneity tests and the number of tagging SNPs may be colored by phenotypic considerations. Such measures could potentially differ because of ascertainment biases, allelic or locus heterogeneity. Even with that qualification, results in this paper may be confounded with association signals or imposed selection. For example, Zaykin et al. [2006] have recently published a method to examine differences in LD structure as evidence for association.

<table>
<thead>
<tr>
<th>Population</th>
<th>No. of families</th>
<th>No. of pair-wise tags</th>
<th>No. of aggressive tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>BE</td>
<td>22</td>
<td>534</td>
<td>448</td>
</tr>
<tr>
<td>CH</td>
<td>24</td>
<td>530</td>
<td>448</td>
</tr>
<tr>
<td>DE1</td>
<td>29</td>
<td>534</td>
<td>448</td>
</tr>
<tr>
<td>DE2</td>
<td>80</td>
<td>540</td>
<td>441</td>
</tr>
<tr>
<td>EN</td>
<td>98</td>
<td>537</td>
<td>439</td>
</tr>
<tr>
<td>ES</td>
<td>52</td>
<td>544</td>
<td>439</td>
</tr>
<tr>
<td>IL1</td>
<td>26</td>
<td>543</td>
<td>449</td>
</tr>
<tr>
<td>IL2</td>
<td>97</td>
<td>555</td>
<td>467</td>
</tr>
<tr>
<td>IR</td>
<td>52</td>
<td>527</td>
<td>441</td>
</tr>
<tr>
<td>NL1</td>
<td>84</td>
<td>526</td>
<td>449</td>
</tr>
<tr>
<td>NL2</td>
<td>110</td>
<td>538</td>
<td>435</td>
</tr>
<tr>
<td>Total</td>
<td>674</td>
<td>542</td>
<td>*</td>
</tr>
</tbody>
</table>

BE: Ghent, Belgium; CH: Zurich, Switzerland; DE1: Essen, Germany; DE2: Gottingen, Germany; EN: London, England; ES: Valencia, Spain; IL1: Jerusalem, Israel; IL2: Tel Aviv, Israel; IR: Dublin, Ireland; NL1: Amsterdam, the Netherlands; NL2: Nijmegen, the Netherlands.

Aggressive tagging refers to the Carkon et al., 2004 method. Aggressive tagging is more aggressive than the de Bakker et al., 2005 method. Both of these tags algorithms were implemented using Haploviev and Tagger.

Mean number of tags across the centres is 537.0 for pair-wise tagging and 464.6 for aggressive tagging. The correlation between the number of families and number of tags across centres is 0.28 for pair-wise tagging and –0.82 for aggressive tagging.

*Computer memory restrictions prevented a global analysis of aggressive tagging.
The marker selection approach utilized for tag selection creates an obvious bias toward markers with a more common than average MAF. First, current SNP databases are enriched for more common markers, as they are easier to detect. Additionally, we used a tagging approach with a condition on the MAF of the SNPs necessary to be tagged. More common markers might show a different pattern of variation between populations than rare alleles. Also, private alleles for some subpopulations, such as the Spanish and the Israeli, might be underrepresented in HapMap samples, yielding to a bias away from observable population differences. Finally, the marker selection protocol for the tags at the outset of the project causes bias in the relative tagging efficiency in these regions.

We selected our families based on the presence of a clinical phenotype and looked at genes with a strong functional hypothesis in relation to the selected phenotype. While this may limit the generalizability of some of the conclusions to the general population, most of the samples in genetics are of clinical populations. So, while this is a limitation from a general population perspective, it also is advantageous from the perspective of seeing how much clinical populations vary on heterozygosity and tagging. The pooling of samples from different sites within the same country could potentially deflate the differences that we could observe.

In general, variation between populations tends to diminish as the populations become more homogenous. The German, Dutch, Belgian, Irish, and English cohorts show relatively little difference between each other, while the Spanish, Swiss, and Israeli samples tend to differ more greatly from the Northern Europeans. However, this difference is far smaller than that observed in surveys of more diverse populations (e.g. Africans vs. Caucasian Europeans) [Crawford et al., 2004]. Finally, while variation changes between populations it appears that the amount of LD structure in each region is relatively similar.

Any case-control association study using these populations together ought to consider using genomic control or a population classification system in conjunction with a statistical technique for combining evidence across groups, such as the Mantel-Haenszel technique. Additionally, the Israeli sample requires more tags than the other populations, after correcting for sample size, indicating that the CEPH sample defined tag set used to define the marker selection did not capture all the variation in the Israeli population. In general, most investigations into LD differences find that Northern Europeans can be grouped together for the purposes of generating and applying tag SNPs [Mueller et al., 2005; Gonzalez-Neira et al., 2006; de Bakker et al., 2006]. Southern Europeans and Israelis tend to form a different population cluster [Bauchet et al., 2007]. Our results are largely consistent with these observations in the context of genes implicated in psychiatric disorders.
ACKNOWLEDGMENTS

The IMAGE project is a multi-site, international effort supported by NIH Grant R01MH62873 to SV Faraone. Site Principal Investigators are Philip Asherson, Tobias Banaschewski, Edmund Sonuga-Barke, Jan Buitelaar, Richard P Ebstein, Stephen V Faraone, Michael Gill, Ana Miranda, Robert Oades, Herbert Roeyers, Aribert Rothenberger, Joseph Sergeant, and Hans-Christoph Steinhausen. Senior co-investigators are Margaret Thompson, Pak Sham, Peter McGuffin, Robert Plomin, Ian Craig and Eric Taylor. Chief Investigators at each site are Rafaela Marco, Nanda Rommelse, Fernando Mulas, Wai Chen, Henrik Uebel, Hanna Christiansen, U Mueller, Cathelijne Buschgens, Barbara Franke, and Lamprini Psychiou. Other investigators are Marieke Altink, Ellen Fliers, Ruud Minderaa, and Alysa Doyle. We thank all the families who kindly participated in this research.

REFERENCES


IMAGE Project 9 Genet. Epidemiol. DOI 10.1002/gepi


Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, de Bakker PI, Daly MJ, Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575.


