Design and evaluation of a panel of Single Nucleotide Polymorphisms (SNPs) in microRNA genomic regions for association studies in human disease

Margarita Muiños-Gimeno[1,2,5], Magda Montfort[3], Mònica Bayés[2,4], Xavier Estivill[1,2,3,5] Yolanda Espinosa-Parrilla[1,2]

[1] Genes and Disease Program, Center for Genomic Regulation (CRG), Barcelona, Catalonia, Spain

[2] The epidemiology and public health CIBER (CIBERESP), Instituto de Salud Carlos III, Madrid, Spain

[3] Barcelona Genotyping Node, CeGen-CRG, Barcelona, Catalonia, Spain

[4] Genomics Core Facility, CRG, Barcelona, Catalonia, Spain

[5] Experimental and Health Sciences department, Pompeu Fabra University (UPF), Barcelona, Catalonia, Spain

Corresponding authors:

Yolanda Espinosa Parrilla

Center for Genomic Regulation (CRG), Genes and Disease Program

Dr Aiguader, 88; PRBB building, 08003 Barcelona, Catalonia, Spain.

Tel. +34933160233, FAX +34933160099, E-mail: yolespinosa@gmail.com

Xavier Estivill

Center for Genomic Regulation (CRG), Genes and Disease Program

Dr. Aiguader 88; 08003 Barcelona, Catalonia, Spain

Tel: +34933160159, FAX +34933160099, E-mail: xavier.estivill@crg.cat

Running title: Genetic variability in miRNA genomic regions

Abstract

MicroRNAs are recognized posttranscriptional gene repressors involved in the control of almost every biological process. Allelic variants in these regions may be an important source of phenotypic diversity and contribute to disease susceptibility. We analyzed the genomic organization of 325 human microRNAs (release 7.1, miRBase) to construct a panel of 768 SNPs covering ~1 Mb of genomic DNA including 131 isolated microRNAs (40%) and 194 microRNAs arranged in 48 microRNA clusters, as well as their 5-kb flanking regions. Thirty-seven percent of the microRNAs were inside known protein-coding genes, which were significantly associated with biological functions regarding neurological, psychological or nutritional disorders. SNP coverage analysis revealed a lower SNP density in microRNAs compared with the average of the genome, with only 24 SNPs located in the 325 microRNAs studied. Further genotyping of 340 unrelated Spanish individuals showed that more than half of the SNPs in microRNAs were either rare or monomorphic, in agreement with the reported selective constraint on human microRNAs. Comparison of the minor allele frequencies between the Spanish and HapMap population samples confirmed the applicability of this SNP panel to the study of complex disorders among the Spanish population and revealed two microRNA regions, hsa-mir-26a-2 in the *CTDSP2* gene and hsa-mir-128-1 in the *R3HDM1* gene, showing geographical allelic frequency variation among the four HapMap populations, probably due to differences in natural selection. The designed microRNA SNP panel could help to the identification of still hidden links between microRNAs and human disease.

Keywords: miRNA, single nucleotide polymorphism, HapMap, population genetics

Introduction

The search for genetic factors predisposing to disease has traditionally focused on the study of protein-coding sequences. Nevertheless, increasing evidence indicates that genetic variation in regulatory regions could be a major contributor to phenotypic diversity in human populations.[1] In the case of psychiatric disorders, changes in regulatory elements leading to small variations in the dosage of proteins involved in neuronal pathways may play an important role in fine-tuning complex brain functions, and contribute to the development of these disorders. Recently, microRNAs (miRNAs) have emerged as important genomic regulators with a key role in the developing and in the adult nervous system, contributing to the correct calibration of neuronal gene expression.[2]

miRNAs are a large class of single stranded small non-coding RNAs of 19-25 nucleotides in length in their mature form that act as post-transcriptional regulators of gene expression by either mRNA degradation or translational repression.[3] The recognition of target mRNAs is mediated by the complementarities between the miRNAs and the nucleotidic sequence of the target mRNAs. However, the most critical region for target recognition consists of nucleotides 2-7 of the miRNA sequence that is known as seed region.[4] miRNAs themselves are the final product of a multistep maturation process that starts with the generation of a transcript referred to as the primary miRNA (pri-miRNA) that hosts one or more miRNA precursors with a characteristic hairpin structure. Most pri-miRNAs are transcribed by RNA polymerase II and undergo capping, splicing and polyadenylation as regular mRNAs. miRNA genes can be either intergenic or located within protein-coding host genes, usually in the introns, and can be processed from the mRNAs of their host genes.[5,6] Since the discovery of the two first miRNAs, lin-4 and lin-7 in *c.elegans,*[7] hundreds of miRNAs in animals, plants and viruses have been identified and annotated in the miRBase sequence database (http://microrna.sanger.ac.uk/). Recent estimates indicate that miRNAs regulate at least 30% of all protein-coding genes, building complex regulatory

networks that control almost every cellular process.[3] In fact, deregulation of the miRNA regulatory pathways has already been involved in human disorders such as cancer or fragile X syndrome.[8,9]

Single nucleotide polymorphisms (SNPs) located within miRNA target sites have been demonstrated to affect the expression of the target gene and contribute to susceptibility to human diseases.[10] Even though many reports have corroborated the link between sequence variants in miRNA binding sites of target genes and complex diseases and phenotypes,[11-15] so far, only one common functional variant in a miRNA gene has been associated with disease; a C/G polymorphism (rs2910164) located in hsa-mir-146a, which has been recently found to contribute to genetic predisposition to papillary thyroid carcinoma.[16] Indeed, allelic changes as well as genomic variants involving either miRNAs or their regulatory machinery may be an important source of phenotypic variation and contribute to the susceptibility for complex disorders. Even though poorly considered until now, association studies using SNPs in miRNA genomic regions might help to evaluate the involvement of miRNAs in disease. With this aim in mind we have analyzed the genomic distribution and genetic variation of miRNAs-containing regions and constructed a panel of SNPs suitable for the study of complex disorders.

Material and Methods

Analysis of the genomic organization of miRNAs

The sequences and genomic coordinates of human miRNAs (miRBase release 7.1 and miRBase release 13.0) were obtained from the miRNA registry (http://microrna.sanger.ac.uk). Genomic locations and human genome annotations were obtained from the UCSC human genome browser assembly from Mar. 2006 build 36, hg 18.

Pathways analysis

Enrichment in biological functions, canonical pathways and molecular networks for miRNA host genes was analyzed with the Ingenuity Pathway Analysis Software (IPA)

version 6.3 (www.ingenuity.com) and the statistical significance of the associations were calculated with the right-tailed Fisher's Exact Test.

SNP selection

For the selection of tag-SNPs we used the HapMap project data set (HapMap Data Rel 19/phase II Oct 05, on NCBI B34 assembly, dbSNP 125) using the genotypes corresponding to the 60 individuals from the CEPH-30 – trios of European descent (http://www.hapmap.org). Only SNPs having a minor allele frequency (MAF) higher than 5% were considered for further analysis. Bins of common SNPs in strong linkage disequilibrium (LD), as defined by r^2 higher than 0.80, were identified within this data set by using the haploview v3.32 software[17] and the 'LD Select' method to process HapMap genotype dump format data corresponding to the selected regions. A total of 710 tag-SNPs were defined using the tagger implementation in haploview. In order to saturate the miRNA regions, 58 additional SNPs were selected from dbSNP (dbSNP 125) or Perlegen (http://genome.perlegen.com/browser/index.html) because of their location either within miRNA sequences or in the ~2kb nearby miRNAs regions with no restrictions on MAF or validation status (Supplementary Table 1).

DNA samples

DNA samples were obtained from 340 healthy blood donors recruited from the Blood and Tissue Bank of the Catalan Health Service, all were of Spanish origin (Catalonia, at the north-east of Spain) and gave informed consent. Genomic DNA was isolated from peripheral blood lymphocytes using automatic DNA extraction and Standard protocols.

Population admixture

To detect population admixture in our control sample, a structured association method was used to further test each sample set for stratification between cases and controls as previously described.[18] No allelic differences among the individuals from the Spanish population were observed and the highest log likelihood scores were obtained when the number of populations was set to 1.

Genotyping of the miRNA SNP panel

The selected 768 SNPs were genotyped using the GoldenGate assay on an Illumina BeadStation 500G in accordance with the manufacturer's standard recommendations. This technology is based on allele-specific primer extension and highly multiplex PCR with universal primers, as reviewed by Syvanen.[19] Allele calling were performed using the BeadStudio program (Illumina Inc, San Diego, CA). Nineteen HapMap individuals including 6 trios and one duplicated DNA sample were genotyped and used to help in the clustering and as a control of the genotyping process. The genotyped controls included 340 individual samples and 2 duplicated DNA samples. All SNPs were examined for standard quality control after genotyping, this evaluation resulted in elimination of a total of 54 SNPs from which 31 were excluded because low signal and 23 were eliminated due to poor clustering. These exclusions yielded a final cleaned data set of 714 SNPs typed (92.97%). Genotypes for the non-excluded SNPs were consistent with Hardy-Weinberg equilibrium (HWE) except for 2 SNPs that were eliminated (Supplementary Table 1). Both genotype concordance and correct Mendelian inheritance were verified, according to this, one sample was eliminated due to gender incoherencies in several SNPs in chromosome X.

Analytical methods

Minor allele frequencies (MAF) were estimated for the genotyped Spanish subjects and compared to those estimated by different HapMap populations (based on 60 European (CEU), 60 Chinese (CHT), 60 Japanese (JPT) and 60 Yoruba (YRI) individuals) via Pearson's χ^2 test. Pearson's correlation coefficient, R^2, was used to measure correlations in allele frequencies between samples taking into account sample's sizes. One sample t-test was also used to test whether the Spanish subjects sampled had allele frequencies equal to those published by HapMap. An adjusted p value threshold of 0.0000712 was used based on 702 independent loci according to Bonferroni correction for multiple testing.

Results

Genomic distribution of the whole collection of miRNAs

In order to select the miRNA genomic regions to be covered by the SNP panel we first studied the genomic distribution of 325 human miRNA genes (miRBase, Release 7.1) regarding their aggregation in clusters as well as their location in relation to other transcriptional units. The analysis of miRNAs distribution within chromosomes showed that miRNAs have a strong tendency to aggregate, with 111 miRNAs (34%) being located at distances of less than 1kb from other miRNAs, and more than half of the miRNAs (169 miRNAs) being less than 4kb apart from other miRNAs (Figure 1a). Taking these observations into account, we defined miRNA clusters as genomic regions containing at least two contiguous miRNAs with an inter-distance of less than 4kb. However, and overruling this criteria, we also considered that if a miRNA was located within the next 7kb of an already assigned miRNA cluster (no miRNAs were found at inter-distances from 7 to 10kb), this miRNA also belonged to this cluster. Finally we also considered that two miRNAs belonged to the same cluster if they were located in the same transcriptional unit, such as the same gene, independently of distance criteria. Following these criteria 60% (194 out of 325) of the miRNAs were organized in 48 clusters spanning 405kb of genomic DNA. Conversely, 40% (131 out of 325) of them were isolated (Figure 1b, Supplementary Table 2). Even though the median number of miRNAs per cluster is 2, some clusters contain large number of miRNAs being remarkable the case of two large clusters on chromosomes 14 and 19 containing 24 and 43 miRNAs, respectively (Table 1).

We also analyzed the localization of miRNAs regarding other transcriptional units annotated at the UCSC genome browser. We found that 37% of the miRNAs (119 out of 325) were located in known protein-coding genes from RefSeq, although only 96 were located in the same orientation of the host gene. According to the criteria of inclusion inside known genes, the other 206 miRNAs (63%) could be considered intergenic, however, when taking into account other transcriptional units annotated at

the UCSC genome browser such as mRNAs from GeneBank, Aceview or Ensembl predicted genes and ESTs, only 99 of the 325 miRNAs were in fact purely intergenic (30% Figure 1b). From the 96 miRNAs located in the same orientation of 77 host genes, most were located in introns and only 3 miRNAs (hsa-mir-22, hsa-mir-155 and hsa-mir-198) were in exons, intron boundaries or un-translated gene regions (Supplementary Table 2). We analyzed the association of the set of host genes containing miRNAs with a given biological process or pathway using the Ingenuity Pathway Analysis software. The program was interrogated for enrichment in biological functions, canonical pathways and molecular networks and the statistical significance of the associations was calculated with the right-tailed Fisher's Exact Test. As shown in Figure 2, some of the most significant associations for miRNA host genes with biological functions were found with several disorders regarding neurological, psychological or nutritional disease, significant associations (-log(p-value)>3) were also found with carbohydrate metabolism, molecular transport and small molecule biochemistry. When the analysis was repeated using data from the last miRbase release (13.0, March 2009) results obtained were in general very similar and, noticeably, the power of the associations increased for neurological and psychological disorders (Supplementary Table 3). As far as canonical pathways are concerned, the most significant associations were found with pathways involved in pantothenate and CoA biosynthesis and GABA receptor signalling. Finally, we also analyzed the molecular networks in which these host genes containing the miRNAs of the SNP panel interact, the highest score was for a gene network involving 33 miRNA host genes related to gene expression, neurological disease, skeletal and muscular system development and function (Figure 2).

Selection of SNPs in miRNA regions

For the selection of the panel of SNPs we considered ~1Mb of genomic DNA corresponding to 131 isolated miRNAs and 48 miRNA clusters, the selected region also includes a flanking region of 5kb upstream and downstream of the specific miRNA

or miRNA cluster. Prior to the selection of SNPs we studied the SNP coverage on miRNAs sequences according to dbSNP database (dbSNP 125). We could only map 24 SNPs within miRNAs sequences (Table 2). This represents a density of 0.86 SNPs per kb (24 SNPs per 27,7kb) at miRNA regions compared with the observed SNP density of 3.99 SNPs/kb for the rest of the genome ($11.96x10^6$ SNPs per $3x10^6$kb). Overall, 93.3% of human pre-miRNAs had no reported SNPs and only 2 of the observed SNPs were located in the mature miRNA region, rs34059726 in hsa-mir-124a-3 and rs12975333 in the seed region of hsa-mir-125a. Due to this low SNP density at miRNA regions and for an optimal selection of informative SNPs, we combined a classical tag-SNP approach (r^2=0.8, MAF>0.05) using information on the European population panel of Hapmap (release 20, Phase II) with the selection of other SNPs according to its putative functional relevance. Finally, the panel included a total of 768 SNPs (Supplementary Table 1) from which 576 were SNPs tagging miRNA gene regions, 19 were SNPs located in miRNA sequences (five out of the 24 SNPs within miRNAs were not included due to technical incompatibilities), 39 at a nearby miRNA location (independently of their MAF or validation status) and 134 were SNPs tagging the promoter regions of the miRNAs host genes. The latter were included to more precisely map the genomic regions involved in future putative associations, to take into account regions that may putatively be involved in miRNA biogenesis (genic miRNAs) and for the interest of these genes *per se*, according to the association found with them and gene networks related to neurological disease.

Genotyping and applicability of the miRNA SNP panel in the Spanish population

A Spanish control sample formed by 340 Spanish unrelated individuals was genotyped using a custom Golden Gate assay from Illumina. Three out of the 19 genotyped SNPs located in miRNA sequences failed in the genotyping. Analysis of the allele frequencies of the other 16 miRNA SNPs showed that 9 of them (56.25%) are monomorphic (Table 2). Next, we studied the applicability of our miRNA SNP panel, constructed based on the information about genetic variability for the European population (CEU) of the

HapMap database, to the study of complex diseases among the Spanish population. The allele frequencies were calculated after confirmation of HWE and MAFs were subjected to pair-wise comparisons between the Spanish and the HapMap CEU, Asiatic and Yoruba populations. Comparisons were performed for the 702 SNPs out of the 768 SNPs of the panel for which genotyping information on HapMap populations and on our population was available (Figure 3). When the MAFs of the Spanish sample were tested against the allele frequencies of the other three population samples, results were very consistent and a high positive correlation (R^2) between the Spanish and the CEU samples was observed (R^2=0.864, $P<<1x10^{-6}$). Conversely, we observed low correlations between the allele frequencies of the Spanish and Asiatic (R^2=0.247) and the Spanish and Yoruba (R^2=0.155) samples. Furthermore, in the case of 36 SNPs (5.12%), the less frequent allele in the CEU HapMap population was found to be the more frequent allele in the Spanish sample (points above the 0.5 horizontal dotted line in Figure 3). Further, we compared the allele frequencies between the CEU Hapmap and the Spanish populations using a Pearson χ^2 test. According to this, allele frequencies for 129 SNPs showed to be significantly different between both populations ($p<0.05$), although when the results of the comparisons were corrected for multiple testing (702 independent loci, $p<7.12x10^{-5}$) only allele frequencies for 4 out of the 702 analyzed SNPs remained significantly different between the Spanish and the CEU HapMap samples (Table 3). Furthermore, these 4 SNPs, 3 located in the same genomic region corresponding to hsa-mir-128-1 (within an intron of *R3HDM1*, R3H domain containing 1), and one located in the region corresponding to hsa-mir-26a2 (within an intron of *CTDSP2*, nuclear LIM interactor-interacting factor 2), also showed strong geographical genetic variation among the Yoruba, Asiatic and CEU populations from HapMap (Table 3).

Discussion

Genome wide association studies using SNP genotyping constitute the standard

approach to identify the genetic component underlying complex traits. The HapMap Project has generated a bulk of genetic information that has become essential for genotyping purposes,[20] providing the required LD information for custom design of SNP panels that have maximal power to capture the genetic variation in a specific genomic region of interest. In this study we have designed a panel of SNPs for the evaluation of miRNA regions as candidate loci for disease susceptibility in association studies. In particular, the panel is addressed to the study of psychiatric disorders for which the identification of susceptibility genes has been less successful than in other complex disorders.[21] The approach proposed here is based on the study of genetic variation in these regulatory elements as possible contributors to psychiatric disease susceptibility. Investigation of how complex gene regulatory networks evolve, and how this results in phenotypic alterations, may represent a useful approach towards the understanding of human evolution and disease.

SNPs are the best characterized source of genetic variation in the human genome and SNP density can be used to measure conservation of DNA sequences. The miRNA regions studied here revealed low SNP density, which could indicate that, as previously suggested,[22-24] miRNA conservation is important and that changes in these regions may contribute to human disease susceptibility. This is further supported by the fact that only 6 of the SNPs located in miRNAs were found to be common SNPs with MAF>0.05 in the studied population. It would be interesting to analyze if the lack of SNPs in miRNAs is indeed due to natural selection or if other factors, such as mutation rate bias on these genomic regions or the fact that many miRNAs are located in still poorly studied regions, are on the back of the low SNP density observed. However, the low number of SNPs and the lack of population frequency information for many of them make this analysis technically difficult to afford nowadays. The vertiginous acquisition of sequence data from different individuals on many ongoing ultra-sequencing projects, together with the increase in the number of new discovered miRNAs, could make it possible in the near future. In fact, very recently, 117 miRNAs have been extensively

re-sequenced in 4 different human populations in an effort to assess natural selection of small RNAs during recent human evolution. This analysis reported a lower SNP density in miRNAs than in other non-coding regions, which were shown to be twice as dense.[23] This study has also shown that strong purifying selection affects the sequence corresponding to the mature miRNA as well as the complementary miRNA sequence (miRNA*), stem region and loop, indicating that mutations in miRNA hairpins are likely to be deleterious and may have severe phenotypic consequences on human health. Unfortunately, only 117 out of the actual 718 miRNAs could be resequenced in their study. Nevertheless, as it happened in our case, the fast increase in the number of new discovered miRNAs is being one of the main handicaps that researchers are facing. Remarkably, since the time that we started the project until now, the number of identified miRNAs has been doubled. However, when analyzing the localization of the actual number of 718 miRNAs (miRBase, Release 13.0) we observed that our SNP panel (considering 325 of the miRBase, Release 7.1), accounts for variability around 100 of the new discovered miRNAs. This is likely due to the fact that many of the new miRNAs are in close vicinity to already known miRNAs.

As part of the comprehensive study of the genomic localization of miRNAs, we observed that around half of the miRNAs are located inside coding genes. Since it has been suggested that miRNAs and their host genes are co-expressed and that their action must be coordinated,[25] we also wanted to study if there was enrichment for a particular kind of genes among those that host miRNAs. Intriguingly, we found an enrichment for genes involved in psychiatric disorders, such as the serotonin receptor gene, *HTR2C*;[26] the acetylcholine receptor gene, *CHRM2*;[27] the glutamate receptor ionotropic delta 1, *GRID1*; or two of the inhibitory neurotransmitter GABA receptor genes (*GABRA3* and *GABRE*).[28] This is of particular interest for the goal of our study since the design of the panel of SNPs is mainly addressed to the study of psychiatric disorders. In fact, promoter regions of miRNA host genes were included among the studied regions because, besides their own interest, inclusion of these regions will also

allow to dissect the contribution of these particularly conspicuous genes in association studies and therefore, to evaluate the potential involvement of miRNAs in the putative associations. Moreover, although the biogenesis of genic miRNAs remains still unclear, intragenic miRNAs seem to be transcribed as part of their hosting transcription units, with the exception of those miRNAs located in antisense orientation to the “host” gene. Thus, transcription of the host gene itself, controlled partially by promoter regions, may be important for miRNA production, and variation in these regions could affect the expression of the hosted miRNA. The comprehensive study of the genomic localization of miRNAs that we have performed shows that 93% of the miRNAs are located either inside previously described or predicted transcriptional units or in clusters (only 22 miRNAs could be considered purely isolated and intergenic). It is known that one miRNA can target as many as several hundred genes, but it is also known that one gene can be targeted synergistically by more than one miRNA. Considering such winding regulatory networks it is tempting to speculate that it might be favourable for the cell to cluster into the same transcriptional unit those miRNAs and/or genes that should act on the same developmental or metabolic pathways.

Finally, another aim of this study was to investigate how well the HapMap European data represents our specific North-East Spanish (Catalan) population. Comparison of allele frequencies not only confirmed the applicability of our SNP panel but also pointed to two genomic regions that show geographic genetic variation among populations. The most likely cause for these marked geographical differences is natural selection. This is clearly the case for the 3 SNPs located in the genomic region corresponding to hsa-mir-128-1 within the *R3HDM1* gene, which is in fact in the vicinity (within 1 Mb) of a LD region containing the lactase gene (*LCT*), for which selection-based evolutionary change in humans has already been established. The other SNP is located in a genomic region for which positive selection has not been demonstrated. In fact, analyzing the extension of the regions showing geographical differences as well as the ancestral alleles would be of importance in order to discern if the real cause for these

differences is natural selection. Apart from evolutionary significance, the study of the possible phenotypic consequences of genetic variation within these regions, such as it could be differential expression of a particular miRNA, may be a matter of concern for disease, in the case associations with those regions were found. Nevertheless, caution must be kept when facing the analysis of these particular regions, in order to avoid spurious associations as it was the case for the reported association between the lactase gene (*LCT*) and tall/short status in a European American sample.[31]

In conclusion, we performed a comprehensive analysis of the genomic organization of miRNAs and their SNP coverage in order to build a panel of SNPs for the analysis of complex disease. Aside from limitations imposed from the fast discovery of miRNAs, which makes difficult to cover the actual number of these regulators, the employment of the designed miRNA SNP panel for association studies should help to elucidate the molecular basis of several disorders by means of the identification of still hidden links between miRNAs and human disease.

ACKNOWLEDGEMENTS

We would like to thank A. Carreras, G. Escaramís and M. Gratacòs for their help in technical and scientific support. We would also like to thank all blood donors participating in this study. This work was supported by the “Instituto Carlos III and Fondo de Investigaciones Sanitarias” [CIBER-CB06/02/0058, CIBER-SAM, FIS/ISCIII:P1052565, ISCIII:GO3/184], the “Fundació la Marató-TV3” [014331], the “Departament d’Universitats Innovació i Empresa, Generalitat de Catalunya” [2005SGR00008] and the European Union Sixth Framework Programme Integrated Project SIROCCO [Grant LSHG-CT-2006-037900]. The Spanish National Genotyping Center (CeGen) is supported by “Genoma España”. M.M.G is a recipient of a FIS fellowship [FI05/0006]. Y.E.P was supported by the “Ramón y Cajal” Program [Spanish Ministry of Science and Innovation]. M.M. was supported by the "Juan de la Cierva" Program [Spanish Ministry of Science and Innovation].

Titles and legends to figures

Figure 1. Genomic localization of the whole collection of miRNAs (miRBase, release 7.1). **a**, the number of miRNAs located closer from other miRNAs than a given distance, in kb, is plotted. **b**, Distribution of miRNAs according to their location regarding other transcriptional units and their aggregation in clusters. Clustered and isolated miRNAs are classified depending on their genomic localization in relation to RefSeq genes, mRNAs, Predicted genes and ESTs.

Figure 2. Association of miRNA host genes with biological processes or molecular networks according the Ingenuity Pathway Analysis software. **a**, The five most significant associations of host genes with different categories of biological functions are shown. **b**, Diagram of the molecular network showing the highest score, it is associated with functions on gene expression, neurological disease, skeletal and muscular system development and function (miRNA host genes included in the network are represented as grey filled shapes).

Figure 3. Pair-wise comparisons of the allele frequencies for 702 SNPs between the Spanish population (CAT) and the CEU, YRY and CHB+JPT populations from Hapmap. The MAFs of the CAT sample were tested against the allele frequencies of the other three population samples.

Supplementary information is available at the European Journal of Human Genetics' website

References

1 Feng Y, Vetro A, Kiss E *et al*: Association of the Neurotrophic Tyrosine Kinase Receptor 3 (NTRK3) Gene and Childhood-Onset Mood Disorders. *Am J Psychiatry* 2008; **165**: 610-616.

2 Kosik KS: The neuronal microRNA system. *Nat Rev Neurosci* 2006; **7**: 911-920.

3 Filipowicz W, Bhattacharyya SN, Sonenberg N: Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet* 2008; **9**: 102-114.

4 Brennecke J, Stark A, Russell RB, Cohen SM: Principles of microRNA-target recognition. *PLoS Biol* 2005; **3**: e85.

5 Kim VN, Nam JW: Genomics of microRNA. *Trends Genet* 2006; **22**: 165-173.

6 Kim YK, Kim VN: Processing of intronic microRNAs. *Embo J* 2007; **26**: 775-783.

7 Ruvkun G: Molecular biology. Glimpses of a tiny RNA world. *Science* 2001; **294**: 797-799.

8 Gong H, Liu CM, Liu DP, Liang CC: The role of small RNAs in human diseases: potential troublemaker and therapeutic tools. *Med Res Rev* 2005; **25**: 361-381.

9 Melo SA, Ropero S, Moutinho C *et al*: A TARBP2 mutation in human cancer impairs microRNA processing and DICER1 function. *Nat Genet* 2009; **41**: 365-370.

10 Borel C, Antonarakis SE: Functional genetic variation of human miRNAs and phenotypic consequences. *Mamm Genome* 2008; **19**: 503-509.

11 Jensen KP, Covault J, Conner TS, Tennen H, Kranzler HR, Furneaux HM: A common polymorphism in serotonin receptor 1B mRNA moderates regulation by miR-96 and associates with aggressive human behaviors. *Mol Psychiatry* 2008.

12 Martin MM, Buckenberger JA, Jiang J *et al*: The human angiotensin II type 1 receptor +1166 A/C polymorphism attenuates microrna-155 binding. *J Biol Chem* 2007; **282**: 24262-24269.

13 Sethupathy P, Borel C, Gagnebin M *et al*: Human microRNA-155 on chromosome 21 differentially interacts with its polymorphic target in the AGTR1 3' untranslated region: a mechanism for functional single-nucleotide polymorphisms related to phenotypes. *Am J Hum Genet* 2007; **81**: 405-413.

14 Tan Z, Randall G, Fan J *et al*: Allele-specific targeting of microRNAs to HLA-G and risk of a sthma. *Am J Hum Genet* 2007; **81**: 829-834.

15 Wang G, van der Walt JM, Mayhew G *et al*: Variation in the miRNA-433 binding site of FGF20 confers risk for Parkinson disease by overexpression of alpha-synuclein. *Am J Hum Genet* 2008; **82**: 283-289.

16 Jazdzewski K, Murray EL, Franssila K, Jarzab B, Schoenberg DR, de la Chapelle A: Common SNP in pre-miR-146a decreases mature miR expression and predisposes to papillary thyroid carcinoma. *Proc Natl Acad Sci U S A* 2008; **105**: 7269-7274.

17 Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005; **21**: 263-265.

18 Gratacos M, Soria V, Urretavizcaya M *et al*: A brain-derived neurotrophic factor (BDNF) haplotype is associated with antidepressant treatment outcome in mood disorders. *Pharmacogenomics J* 2008; **8**: 101-112.

19 Syvanen AC: Toward genome-wide SNP genotyping. *Nat Genet* 2005; **37 Suppl**: S5-10.

20 Frazer KA, Ballinger DG, Cox DR *et al*: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; **449**: 851-861.

21 Burmeister M, McInnis MG, Zollner S: Psychiatric genetics: progress amid controversy. *Nat Rev Genet* 2008; **9**: 527-540.

22 Lu M, Zhang Q, Deng M *et al*: An analysis of human microRNA and disease associations. *PLoS ONE* 2008; **3**: e3420.

23 Quach H, Barreiro LB, Laval G *et al*: Signatures of purifying and local positive selection in human miRNAs. *Am J Hum Genet* 2009; **84**: 316-327.

24 Saunders MA, Liang H, Li WH: Human polymorphism at microRNAs and microRNA target sites. *Proc Natl Acad Sci U S A* 2007; **104**: 3300-3305.

25 Baskerville S, Bartel DP: Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *Rna* 2005; **11**: 241-247.

26 Gurevich I, Tamir H, Arango V, Dwork AJ, Mann JJ, Schmauss C: Altered editing of serotonin 2C receptor pre-mRNA in the prefrontal cortex of depressed suicide victims. *Neuron* 2002; **34**: 349-356.

27 Wang JC, Hinrichs AL, Stock H *et al*: Evidence of common and specific genetic effects: association of the muscarinic acetylcholine receptor M2 (CHRM2) gene with alcohol dependence and major depressive syndrome. *Hum Mol Genet* 2004; **13**: 1903-1911.

28 Craddock N, Jones L, Jones IR *et al*: Strong genetic evidence for a selective influence of GABA(A) receptors on a component of the bipolar disorder phenotype. *Mol Psychiatry* 2008.

29 Beja-Pereira A, Luikart G, England PR *et al*: Gene-culture coevolution between cattle milk protein genes and human lactase genes. *Nat Genet* 2003; **35**: 311-313.

30 Hollox E: Evolutionary genetics: genetics of lactase persistence--fresh lessons in the history of milk drinking. *Eur J Hum Genet* 2005; **13**: 267-269.

31 Campbell CD, Ogburn EL, Lunetta KL *et al*: Demonstrating stratification in a European American population. *Nat Genet* 2005; **37**: 868-872.