European Journal of Human Genetics (2010) 18, 218–226
© 2010 Macmillan Publishers Limited All rights reserved 1018-4813/10 $32.00
www.nature.com/ejhg

ARTICLE

Design and evaluation of a panel of single-nucleotide polymorphisms in microRNA genomic regions for association studies in human disease

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

MicroRNAs (miRNA) 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 miRNAs (release 7.1, miRBase) to construct a panel of 768 single-nucleotide polymorphisms (SNPs) covering ∼1 Mb of genomic DNA, including 131 isolated miRNAs (40%) and 194 miRNAs arranged in 48 miRNA clusters, as well as their 5-kb flanking regions. Of these miRNAs, 37% 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 miRNAs compared with the average of the genome, with only 24 SNPs located in the 325 miRNAs studied. Further genotyping of 340 unrelated Spanish individuals showed that more than half of the SNPs in miRNAs were either rare or monomorphic, in agreement with the reported selective constraint on human miRNAs. A comparison of the minor allele frequencies between 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 miRNA 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 because of differences in natural selection. The designed miRNA SNP panel could help to identify still hidden links between miRNAs and human disease.

European Journal of Human Genetics (2010) **18,** 218–226; doi:10.1038/ejhg.2009.165; published online 7 October 2009

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 have 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 development 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 noncoding RNAs of 19–25 nucleotides in length in their mature form that act as posttranscriptional regulators of gene expression by either mRNA degradation or translational repression.[3] The recognition of target mRNAs is mediated by the complementarities between miRNAs and the nucleotidic sequence of target mRNAs. However, the most critical region for target recognition consists of nucleotides 2–7 of the miRNA sequence that is known as the 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 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 *Caenorhabditis 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 miRNA regulatory pathways has already been involved in human disorders such as cancer or fragile X syndrome.[8,9]

[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] Experimental and Health Sciences Department, Pompeu Fabra University (UPF), Barcelona, Catalonia, Spain; [4] Barcelona Genotyping Node, CeGen-CRG, Barcelona, Catalonia, Spain; [5] Genomics Core Facility, CRG, Barcelona, Catalonia, Spain

*Correspondence: Dr Y Espinosa-Parrilla, Center for Genomic Regulation (CRG), Genes and Disease Program, Dr Aiguader, 88; PRBB building, 08003 Barcelona, Catalonia, Spain. Tel. +34 93 3160233; Fax +34 93 3160099; E-mail: yolespinosa@gmail.com or Dr X Estivill, Center for Genomic Regulation (CRG), Genes and Disease Program, Dr Aiguader 88; 08003 Barcelona, Catalonia, Spain. Tel: +34 93 3160159; Fax +34 93 3160099; E-mail: xavier.estivill@crg.cat

Received 5 May 2009; revised 11 August 2009; accepted 28 August 2009; published online 7 October 2009

Single-nucleotide polymorphisms (SNPs) located within miRNA target sites have been shown to affect the expression of the target gene and contribute to susceptibility to human diseases.[10] Although 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 recently been found to contribute toward a genetic predisposition to papillary thyroid carcinoma.[16] Indeed, allelic changes and genomic variants involving either miRNAs or their regulatory machinery may be important sources of phenotypic variation and contribute to the susceptibility for complex disorders. Although 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 analyzed the genomic distribution and genetic variation of miRNA-containing regions and constructed a panel of SNPs suitable for the study of complex disorders.

MATERIALS 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 March 2006 build 36, hg 18.

Pathways analysis

Enrichment in biological functions, canonical pathways and molecular networks for miRNA host genes was analyzed using the Ingenuity Pathway Analysis (IPA) Software version 6.3. (www.ingenuity.com) and the statistical significance of associations was calculated using the right-tailed Fisher's exact test.

SNP selection

For the selection of tagged-SNPs, we used the HapMap project data set (HapMap Data Rel 19/phase II October 2005, on NCBI B34 assembly, dbSNP 125) using genotypes that correspond 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 an r^2 value higher than 0.80, were identified within this data set by using 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 tagged-SNPs were defined using the tagger implementation in haploview. To saturate the miRNA regions, 58 additional SNPs were selected from dbSNP (dbSNP 125) or Perlegen (http://genome.perlegen.com/) because of their location either within miRNA sequences or in the ∼2 kb nearby miRNA 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 northeast of Spain) and gave an 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 Golden Gate assay on an Illumina BeadStation 500G (Illumina, San Diego, CA, USA) 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 was performed using the BeadStudio program (Illumina Inc). A total of 19 HapMap individuals including 6 trios and 1 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 the elimination of a total of 54 SNPs, from which 31 were excluded because of low signal and 23 were eliminated because of poor clustering. These exclusions yielded a final cleaned data set of 714 SNPs that were typed (92.97%). Genotypes for the nonexcluded SNPs were consistent with Hardy–Weinberg equilibrium (HWE), except for two SNPs that were eliminated (Supplementary Table 1). Both genotype concordance and correct Mendelian inheritance were verified; one sample was eliminated because of gender incoherencies in several SNPs in chromosome X.

Analytical methods

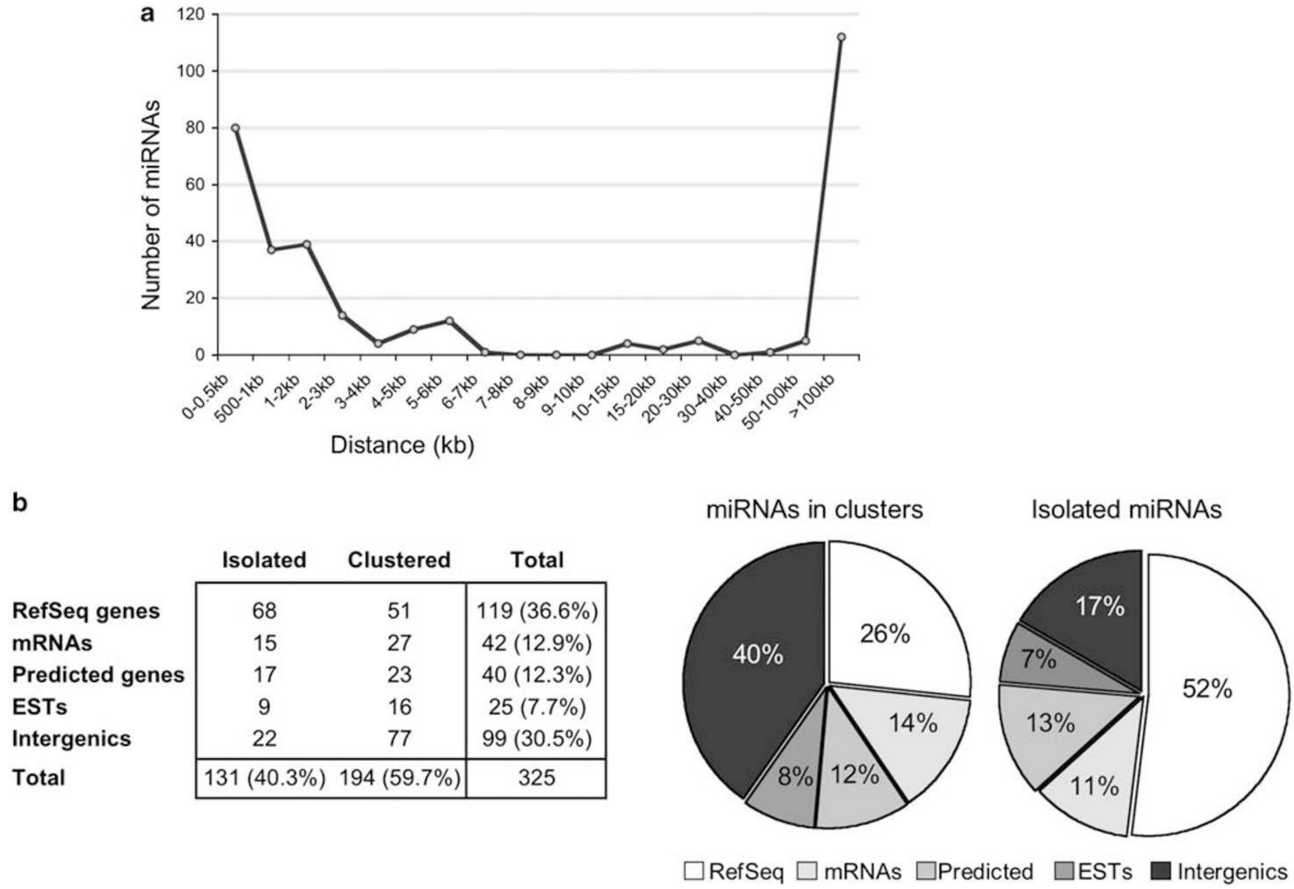
Minor allele frequencies were estimated for the genotyped Spanish subjects and were compared with those estimated by different HapMap populations (on the basis of 60 European (CEU), 60 Chinese (CHT), 60 Japanese (JPT) and 60 Yoruba (YRI) individuals) using Pearson's χ^2-test. Pearson's correlation coefficient, R^2, was used to measure correlations in allele frequencies between samples by taking into account the sizes of samples. 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 on the basis of 702 independent loci according to Bonferroni's correction for multiple testing.

RESULTS

Genomic distribution of the whole collection of miRNAs

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) with regard to their aggregation in clusters, as well as their location in relation to other transcriptional units. The analysis of miRNA distribution within chromosomes showed that miRNAs have a strong tendency to aggregate, with 111 miRNAs (34%) being located at distances of <1 kb from other miRNAs, and more than half of the miRNAs (169 miRNAs) being <4 kb 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 interdistance of <4 kb. However, and overruling these criteria, we also considered that if a miRNA was located within the next 7 kb of an already assigned miRNA cluster (no miRNAs were found at interdistances from 7 to 10 kb), 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 miRNAs were organized into 48 clusters spanning 405 kb of genomic DNA. Conversely, 40% (131 out of 325) of them were isolated (Figure 1b, Supplementary Table 2). Although the median number of miRNAs per cluster is two, some clusters contain a large number of miRNAs; a remarkable case is that of two large clusters on chromosomes 14 and 19 containing 24 and 43 miRNAs, respectively (Table 1).

We also analyzed the localization of miRNAs with regard to other transcriptional units annotated at the UCSC genome browser. We found that 37% of miRNAs (119 out of 325) were located in known protein-coding genes from RefSeq (NCBI), although only 96 were located in the same orientation of the host gene. According to the



	Isolated	Clustered	Total
RefSeq genes	68	51	119 (36.6%)
mRNAs	15	27	42 (12.9%)
Predicted genes	17	23	40 (12.3%)
ESTs	9	16	25 (7.7%)
Intergenics	22	77	99 (30.5%)
Total	131 (40.3%)	194 (59.7%)	325

Figure 1 Genomic localization of the whole collection of miRNAs (miRBase, release 7.1). (**a**) The number of miRNAs located closer than a given distance from other miRNAs, in kb, is plotted. (**b**) Distribution of miRNAs according to their location in relation to 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.

criterion 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 exon–intron boundaries or in untranslated 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 IPA software. The program was interrogated for enrichment in biological functions, canonical pathways and molecular networks, and the statistical significance of associations was calculated using 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 in relation to neurological, psychological or nutritional disease; significant associations ($-\log(P\text{-value})>3$) 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), the results obtained were in general very similar and, noticeably, the power of 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 signaling. 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 ~1 Mb of genomic DNA corresponding to 131 isolated miRNAs and 48 miRNA clusters; the selected region also includes a flanking region of 5 kb upstream and downstream of the specific miRNA or miRNA cluster. Before the selection of SNPs, we studied the SNP coverage on miRNA sequences according to the dbSNP database (dbSNP 125). We could only map 24 SNPs within miRNA sequences (Table 2). This represents a density of 0.86 SNPs per kilobase (24 SNPs per 27.7 kb) at miRNA regions compared with the observed SNP density of 3.99 SNPs per kilobase for the rest of the genome (11.96×10^6 SNPs per 3×10^6 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

Table 1 Summary of the characteristics of miRNA clusters

CLUSTERS	*Size (bp)*	*Genomic position*	*N*	*miRNAs*	*Location*	*Strand*
cl1.1	1983	chr1:1092347–1094330	3	miR-200b, miR-200a, miR-429	Intergenic	+
cl1.2	3017	chr1:40992614–40995631	2	miR-30e, miR-30c-1	*NFYC*	+
cl1.3	5846	chr1:170374561–170380407	2	miR-214, miR-199a-2	*DNM3(–)*	–
cl1.4	280	chr1:197094625–197094905	2	miR-181b-1, miR-213	Predicted gene	–
cl1.5	671	chr1:206041820–206042491	2	miR-29c, miR-29b-2	mRNA	–
cl1.6	388	chr1:218357818–218358206	2	miR-215, miR-194-1	*IARS2(–)*	–
cl2.1	6092	chr2:56063606–56069698	2	miR-217, miR-216	EST	–
cl3.1	561	chr3:49032585–49033146	2	miR-425, miR-191	*DALRD3*	–
cl3.2	237	chr3:161605070–161605307	2	miR-15b, miR-16-2	*SMC4L1*	+
cl4.1	683	chr4:113788479–113789162	5	miR-367, miR-302b	*LARP7(–)*	–
cl5.1	1815	chr5:148788674–148790489	2	miR-143, miR-145	EST/mRNA	+
cl7.1	514	chr7:99529119–99529633	3	miR-25, miR-93, miR-106b	*MCM7*	–
cl7.2	4631	chr7:129197459–129202090	3	miR-182, miR-96, miR-183	Intergenic	–
cl7.3	792	chr7:130212046–130212838	2	miR-29a, miR-29b-1	Predicted gene	–
cl9.1	2963	chr9:95978060–95978536	3	let-7a-1, let-7f-1, let-7d	Predicted gene	+
cl9.2	880	chr9:96887311–96888191	3	miR-23b, miR-27b, miR-24-1	*C9orf3*	+
cl9.3	1356	chr9:126494542–126495898	2	miR-181a, miR-181b-2	*NR6A1(–)*	+
cl11.1	302	chr11:64415185–64415487	2	miR-192, miR-194-2	Predicted gene	
cl11.2	577	chr11:110888873–110889450	2	miR-34b, miR-34c	mRNA	+
cl11.3	5786	chr11:121475675–121528226	3	let-7a-2, miR-100, miR-125b-1	*LOC399959*	–
cl12.1	492	chr12:6943123–6943615	2	miR-200c, miR-141	EST	+
cl13.1	228	chr13:49521110–49521338	2	miR-16-1, miR-15a	*DLEU2*	–
cl13.2	786	chr13:90800860–90801646	6	miR-17, miR-92-1	mRNA	+
cl14.1	15723	chr14:100405150–100420873	7	miR-493, miR-136	*RTL1(–)*	+
cl14.2	43925	chr14:100558156–100602081	24	miR-379, miR-410	Intergenic	+
cl16.1	5404	chr16:14305325–14310729	2	miR-193b, miR-365-1	Predicted gene	+
cl17.1	472	chr17:1899952–1900424	2	miR-132, miR-212	Intergenic	–
cl17.2	407	chr17:6861658–6862065	2	miR-195, miR-497	mRNA	–
cl17.3	249	chr17:24212513–24212762	2	miR-451, miR-144	EST	–
cl18.1	3390	chr18:17659657–17663047	2	miR-133a-1, miR-1-2	*MIB1(–)*	–
cl19.1	372	chr19:13808101–13808473	3	miR-24-2, miR-27a, miR-23a	EST	–
cl19.2	312	chr19:13846513–13846825	2	miR-181c, miR-181d	Predicted gene	+
cl19.3	727	chr19:56887677–56888404	3	miR-99b, let-7e, miR-125a	Intergenic	+
cl19.4	95751	chr19:58861745–58957496	43	miR-512-1, miR-519a-2	Intergenic	+
cl19.5	1098	chr19:58982741–58983839	3	miR-371, miR-372, miR-373	Intergenic	+
cl20.1	10707	chr20:60561958–60572665	2	miR-1-1, miR-133a-2	*C20orf166*	+
cl21.1	51236	chr21:16833280–16884516	3	miR-99a, let-7c, miR-125b-2	*C21orf34*	+
cl22.1	1019	chr22:44887293–44888312	2	let-7a-3, let-7b	mRNA	+
clX.1	945	chrX:45490529–45491474	2	miR-221, miR-222	Predicted gene	–
clX.2	11166	chrX:49654849–49666015	5	miR-188, miR-502	*CLCN5*	+
clX.3	1051	chrX:53599909–53600960	2	miR-98, let-7f-2	*HUWE1*	–
clX.4	86225	chrX:76056092–76142317	2	miR-384, miR-325	mRNA	–
clX.5	900	chrX:133131074–133131974	6	miR-363, miR-106a	Intergenic	–
clX.6	6370	chrX:133502037–133508407	4	miR-450-1, miR-450-2, miR-503, miR-424	Intergenic	–
clX.7	11201	chrX:146115036–146126237	4	miR-513-2, miR-508	Intergenic	–
clX.8	12378	chrX:146161545–146173923	4	miR-510, miR-514-3	Intergenic	–
clX.9	1134	chrX:150877706–150878840	2	miR-224, miR-452	*GABRE*	–
clX.10	2273	chrX:151311347–151313620	2	miR-105-1, miR-105-2	*GABRA3*	–

N, number of miRNAs included in each cluster, when this number is higher than four, only the two miRNAs located at the 3′ and 5′ ends of the cluster are shown.
(–) Indicates that the miRNA is located in the opposite orientation of the host gene.

hsa-mir-125a. Owing to this low SNP density at miRNA regions and for an optimal selection of informative SNPs, we combined a classical tagged-SNP approach (r^2=0.8, MAF$>$0.05) with the selection of other SNPs according to its putative functional relevance, using information from the European population panel of Hapmap (release 20, Phase II). 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 (5 out of the 24 SNPs within miRNAs were not included because of technical incompatibilities), 39 at a nearby miRNA location (independently of their MAF or validation status) and 134 were SNPs tagging the promoter regions of miRNA host genes. The latter

a TOP BIOLOGICAL FUNCTIONS

Diseases and disorders	p-value	Molecules
Neurological Disease	3.46E-05 - 4.62E-02	14
Psychological Disorders	3.46E-05 - 4.08E-02	7
Nutritional Disease	3.80E-05 - 3.94E-02	6
Developmental Disorder	1.71E-05 - 1.77E-02	3
Skeletal and Muscular Disorders	2.25E-05 - 3.94E-02	5
Molecular and Cellular Functions		
Carbohydrate Metabolism	4.56E-05 - 4.37E-02	6
Molecular Transport	4.56E-05 - 4.37E-02	10
Small Molecule Bichemistry	4.56E-05 - 4.37E-02	10
Gene Expression	5.83E-05 - 4.37E-02	4
Cellular Assembly and Organization	4.03E-04 - 4.80E-02	10
Physiological System Development and Function		
Nervous System	4.03E-04 - 4.37E-02	7
Skeletal ad Muscular System	6.87E-04 - 4.37E-02	6
Tissue Developmet	6.87E-04 - 4.37E-02	4
Endocrine System	1.04E-03 - 4.37E-02	3
Cardiovascular System	4.25E-03 - 4.80E-02	6

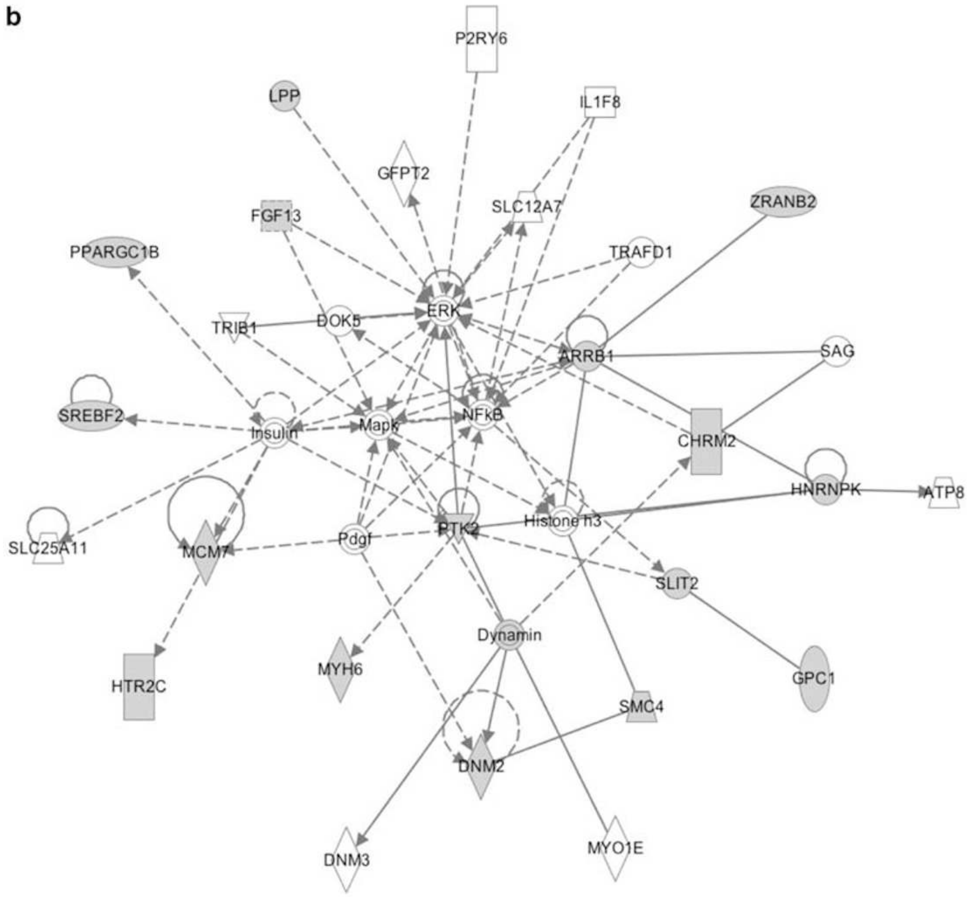


Figure 2 Association of miRNA host genes with biological processes or molecular networks according the Ingenuity Pathway Analysis software. (**a**) The five most significant associations of host genes with different categories of biological functions are shown. (**b**) Diagram of the molecular network 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 gray-filled shapes).

were included to map the genomic regions involved in future possible associations more precisely, 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 with gene networks related to neurological disease.

Table 2 SNPs located in miRNAs and allele frequencies in a Spanish population

SNP	*Genomic position*	*miRNA*	*miRNA location*	*Alleles*	*Validation*	*Genotyping*
rs35301225	chr1:9134389	miR-34a	Predicted gene	A/C/T	Hapmap	Not considered
rs2292832	chr2:241044176	miR-149	*GPC1*	T/C	Hapmap	Fail
rs2910164	chr5:159844996	miR-146a	Predicted gene	C/G	Hapmap	C: 0.2734
rs10061133	chr5:54502301	miR-449	*CDC20B*	A/G	Cluster,2hit-2allele	Not considered
rs13232101	chr7:1029100	miR-339	*C7orf50*	G/T	Unknown	Monomorphic-G: 1.00
rs12355840	chr10:134911103	miR-202	mRNA	C/T	Unknown	C: 0.198
rs11231898	chr11:64415412	miR-194-2	Predicted gene	A/G	Hapmap	A: 0.0015
rs11614913	chr12:52671866	miR-196a-2	Predicted gene	C/T	Hapmap	T: 0.3687
rs2289030	chr12:93752417	miR-492	mRNA	C/G	Hapmap	G: 0.0649
rs9589207	chr13:90801590	miR-92-1	mRNA	A/G	Hapmap	Monomorphic-G: 1.00
rs12884005	chr14:100417161	miR-431	*RTL1(–)*	A/C	Unknown	Monomorphic-G: 1.00
rs7205289	chr16:68524506	miR-140	*WWP2*	A/C	Unknown	Monomorphic-C: 1.00
rs6505162	chr17:25468309	miR-423	*CCDC55*	A/C	Hapmap	Fail
rs895819	chr19:13808292	miR-27a	EST	C/T	Hapmap	T: 0.2224
rs11671784	chr19:13808296	miR-27a	EST	A/G	Unknown	Not considered
rs12975333[a]	chr19:56888340	miR-125a	EST	G/T	Unknown	Monomorphic-G: 1.00
rs7255628	chr19:58902546	miR-520c	Intergenic	G/T	Unknown	Monomorphic-G: 1.00
rs13382089	chr19:58911666	miR-521-2	Intergenic	C/G	Hapmap	Monomorphic-G: 1.00
rs2569389	chr19:58951814	miR-516-1	EST	A/G	Unknown	Fail
rs3746444	chr20:33041912	miR-499	*MYH7B*	C/T	Cluster, frequency	C: 0.1973
rs7267163	chr20:33041937	miR-499	*MYH7B*	C/T	Unknown	Not considered
rs6122014	chr20:60561960	miR-1-1	*C20orf166*	C/T	Hapmap	Monomorphic-C: 1.00
rs34059726[a]	chr20:61280352	miR-124a-3	Intergenic	G/T	Unknown	Not considered
rs2504172	chrX:146147969	miR-509	Intergenic	A/G	Unknown	Monomorphic-G: 1.00

[a]SNPs located in miRNA seed regions.
(–) Indicates that the miRNA is located in the opposite orientation of the host gene.
The last column shows results after genotyping and the minor allele frequencies (MAF) when possible.
Five of the SNPs were not considered for the genotyping panel due to technical incompatibilities.

Genotyping and applicability of the miRNA SNP panel in a 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 on the basis of information regarding genetic variability of the European population (CEU) of the HapMap database, to the study of complex diseases among the Spanish population. Allele frequencies were calculated after confirmation of HWE, and MAFs were subjected to pair-wise comparisons between the Spanish and the HapMap CEU, as well as Asiatic and Yoruba populations. Comparisons were carried out 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, the results were very consistent, and a high positive correlation (R^2) between the Spanish and CEU samples was observed (R^2=0.864, $P \ll 1\times10^{-6}$). Conversely, we observed low correlations between the allele frequencies of the Spanish and the Asiatic (R^2=0.247), and the Spanish and the Yoruba (R^2=0.155) samples. 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 comparisons were corrected for multiple testing (702 independent loci, $P<7.12\times10^{-5}$), only allele frequencies for 4 out of the 702 analyzed SNPs remained significantly different between the Spanish and CEU HapMap samples (Table 3). Furthermore, these four SNPs, three 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 a strong geographical genetic variation among the Yoruba, the Asiatic and the CEU populations from HapMap (Table 3).

DISCUSSION

Genome-wide association studies using SNP genotyping constitute the standard approach for identifying 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 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

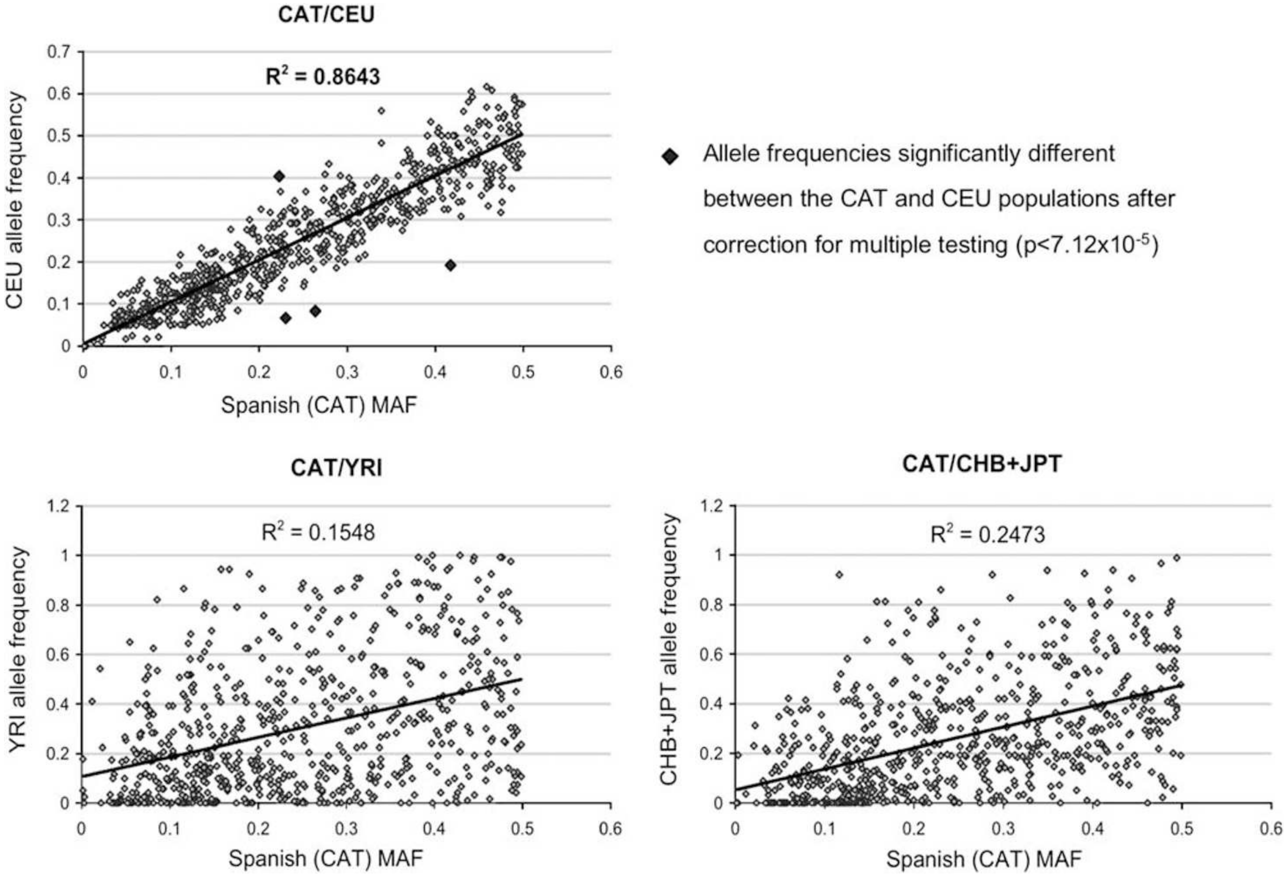


Figure 3 Pair-wise comparisons of the allele frequencies for 702 SNPs between the Spanish population (CAT) and the CEU, YRI and CHB+JPT populations from Hapmap. The MAFs of the CAT sample were tested against the allele frequencies of the other three population samples.

Table 3 SNPs with significant allele frequency differences between the Spanish and the European HapMap populations

						Reference allele frequencies			
SNP	*Genomic position*	*Genomic region*	*Alleles*	*Ref allele*	*P-value*	*CAT*	*CEU*	*YRI*	*CHB+JPT*
rs1446584	chr2: 136136431	hsa-mir-128-1-*R3HDM1*	C/T	C	4.03E−8	0.417	0.192	0.95	0.861
rs1374330	chr2: 136137378	hsa-mir-128-1-*R3HDM1*	G/T	G	1.66E−8	0.23	0.067	0.017	0.461
rs2289959	chr2: 136140374	hsa-mir-128-1-*R3HDM1*	A/G	A	5.77E−8	0.264	0.083	0.358	0.663
rs1599750	chr12: 56528812	hsa-mir-26a-2-*CTDSP2*	G/T	G	9.96E−8	0.223	0.404	—	0.778

R3HDM1, R3H domain containing 1 gene; *CTDSP2*, nuclear LIM interactor-interacting factor 2 gene.
Spanish (Catalan, CAT), European (CEU), Yoruba (YRI), Chinese (CHB) and Japanese (JPT) populations.
Significant differences in the allele frequencies between the CAT and CEU populations after Bonferroni's correction ($P<7$, 12×10^6, 702 independent tests) are shown.

phenotypic alterations, may represent a useful approach toward understanding 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 the conservation of DNA sequences. The miRNA regions studied here revealed a 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 six 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 whether the lack of SNPs in miRNAs is indeed because of natural selection or whether 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 the cause for 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 ultrasequencing projects, together with the increase in the number of newly discovered miRNAs, could make it more affordable in the near future. In fact, very recently, 117 miRNAs have been extensively resequenced in four different human populations in an effort to assess the natural selection of small RNAs during recent human evolution. This analysis reported a lower SNP density in miRNAs than in other noncoding 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 that study. Nevertheless, as it happened in our case, the fast increase in the number of newly discovered miRNAs is one of the main handicaps that researchers face. Remarkably, since the time that we started the project until now, the number of identified miRNAs has 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 a variability of around 100 of the newly discovered miRNAs. This is likely because of the fact that many of the new miRNAs are in close vicinity to already known miRNAs.

As a part of the comprehensive study of the genomic localization of miRNAs, we observed that approximately half of the miRNAs are located inside coding genes. As it has been suggested that miRNAs and their host genes are coexpressed and that their action must be coordinated,[25] we also wanted to study whether there was enrichment for a particular kind of gene 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, as 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 hence evaluate the potential involvement of miRNAs in putative associations. Moreover, although the biogenesis of genic miRNAs is 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 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 favorable for the cell to cluster into the same transcriptional unit as those miRNAs and/or genes that act on the same developmental or metabolic pathways.

Finally, another aim of this study was to investigate how well the HapMap European data represent our specific Northeast 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 a geographical genetic variation among populations. The most likely cause for these marked geographical differences is natural selection. This is clearly the case for the three 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 an LD region containing the lactase gene (*LCT*), for which selection-based evolutionary change in humans has already been established.[29] The other SNP is located in a genomic region for which positive selection has not been shown. In fact, analyzing the extension of the regions showing geographical differences, as well as the ancestral alleles, is important to discern whether 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 a differential expression of a particular miRNA, may be a matter of concern for disease, in case associations with those regions were found. Nevertheless, caution must be exercised when facing the analysis of these particular regions to avoid spurious associations, as it was the case for the reported association between the lactase gene (*LCT*) and the tall/short status in a European American sample.[30]

In conclusion, we performed a comprehensive analysis of the genomic organization of miRNAs and their SNP coverage to build a panel of SNPs for the analysis of complex disease. Aside from limitations imposed from the fast discovery of miRNAs, which makes it difficult to cover the actual number of these regulators, the use 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 thank A Carreras, G Escaramís and M Gratacòs for their help in technical and scientific support. We also 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’. MMG is a recipient of an FIS fellowship [FI05/0006]. YEP was supported by the ‘Ramón y Cajal’ Program [Spanish Ministry of Science and Innovation]. MM was supported by the ‘Juan de la Cierva’ Program [Spanish Ministry of Science and Innovation].

1 Knight JC: Regulatory polymorphisms underlying complex disease traits. *J Mol Med* 2005; **83**: 97–109.
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; **14**: 381–389.
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 asthma. *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 USA* 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 USA* 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; July 1 [E-pub ahead of print].

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

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

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)