doi:10.1093/nar/gkm480

Aberrant allele frequencies of the SNPs located in microRNA target sites are potentially associated with human cancers

Zhenbao Yu[1,†], Zhen Li[1,2,†], Normand Jolicoeur[1,†], Linhua Zhang[1], Yves Fortin[1], Edwin Wang[1], Meiqun Wu[1] and Shi-Hsiang Shen[1,2,*]

[1]Health Sector, Biotechnology Research Institute, National Research Council of Canada, 6100 Royalmount Avenue, Montréal, Québec, Canada, H4P 2R2 and [2]Department of Medicine, McGill University, Montréal, Québec, Canada, H3G 1A4

Received October 3, 2006; Revised May 11, 2007; Accepted May 31, 2007

ABSTRACT

MicroRNAs (miRNAs) are a class of noncoding small RNAs that regulate gene expression by base pairing with target mRNAs at the 3′-terminal untranslated regions (3′-UTRs), leading to mRNA cleavage or translational repression. Single-nucleotide polymorphisms (SNPs) located at miRNA-binding sites (miRNA-binding SNPs) are likely to affect the expression of the miRNA target and may contribute to the susceptibility of humans to common diseases. We herein performed a genome-wide analysis of SNPs located in the miRNA-binding sites of the 3′-UTR of various human genes. We found that miRNA-binding SNPs are negatively selected in respect to SNP distribution between the miRNA-binding 'seed' sequence and the entire 3′-UTR sequence. Furthermore, we comprehensively defined the expression of each miRNA-binding SNP in cancers versus normal tissues through mining EST databases. Interestingly, we found that some miRNA-binding SNPs exhibit significant different allele frequencies between the human cancer EST libraries and the dbSNP database. More importantly, using human cancer specimens against the dbSNP database for case-control association studies, we found that twelve miRNA-binding SNPs indeed display an aberrant allele frequency in human cancers. Hence, SNPs located in miRNA-binding sites affect miRNA target expression and function, and are potentially associated with cancers.

INTRODUCTION

MicroRNAs (miRNAs) are noncoding small (~22 nt) RNAs that regulate the expression of target mRNAs (1,2). MiRNAs are encoded in the chromosomal DNA and transcribed as longer stem-loop precursors, termed pri-miRNAs (3–5). Upon transcription, pri-miRNA is converted to mature miRNA duplex through sequential processing by the RNaseIII family of endonucleases Drosha/DGCR8 and Dicer (3–5). One strand of the processed duplex is incorporated into a silencing complex and guided to target sequences located at the 3′-terminal untranslated regions (3′-UTRs) of mRNAs by base pairing (6), resulting in the cleavage of target mRNAs or repression of their productive translation (6). Over the past few years, several hundred miRNAs were identified in animals and plants. It is currently estimated that miRNAs account for ~1% of the predicted genes in higher eukaryotic genomes (7).

A growing body of evidence revealed that miRNAs are involved in a variety of biological processes (8), such as embryonic development, cell proliferation, cell differentiation, apoptosis and insulin secretion. Moreover, several reports indicate that miRNAs are involved in human tumorigenesis. A systematic search for correlation between the genomic position of miRNAs and the location of cancer-associated regions revealed that over half of the mapped miRNAs in chronic lymphocytic leukemias (CLL) are located at fragile chromosome regions involved in human cancers (9). MiRNA expression profiles also indicated that most miRNAs had lower expression levels in tumors compared with normal tissues (10). For example, Let-7, targeting the oncogene RAS, is downregulated in lung cancers (11) and miR-15 and 16, targeting the antiapoptotic factor

*To whom correspondence should be addressed. Tel: +514 496 6318; Fax: +514 496 6319; Email: shi.shen@cnrc-nrc.gc.ca
Correspondence may also be addressed to Zhenbao Yu. Tel: +514 496 6377; Fax: +514 496 6319; Email: zhenbao.yu@nrc.ca

[†]The authors wish to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

BCL2, are downregulated in CLL (12). Furthermore, a germ-line mutation in the *pri-miR-16-1/15a* precursor was found to cause its reduced transcription in a patient with familial CLL (13).

Single-nucleotide polymorphisms (SNPs) are the most frequent variation in the human genome, occurring once every several hundred base pairs throughout the genome. They have been studied extensively for defining the regions of disease candidate genes (14). Previously, in order to search for SNP-affected disease susceptibility and outcome, most researchers focused on specific genes, as resources and analytical tools were limited. Moreover, until recently, there was a profound interest in nonsynonymous SNPs because they shift the codons and often change the protein structure and function. However, the majority of SNPs in the genome are not nonsynonymous SNPs that occur in untranslated, intronic or intergenic regions. These SNPs could affect complex diseases through their effect on gene expression quantitatively. Unlike nonsynonymous SNPs, SNPs capable of affecting gene expression may not be easily identified because gene regulatory elements could not be accurately defined in a complex gene regulation process. However, since the thermodynamics of RNA–RNA binding plays an essential role in miRNA interaction with target mRNA, it is expected that sequence variations such as SNPs at miRNA-binding sites may affect the expression of miRNA targets. The rationale for this assumption is based on the principle of miRNA–target interaction. Accumulating evidence revealed that 7 nt at the 5′-terminus of miRNAs from position 2 to position 8, called 'seed' region, are essential for their function (15). Based on these discoveries, several computational methods have been developed to predict miRNA targets (16–23). Most of these methods have been biologically validated and proved to be very efficient and accurate. For example, up to 90% of the randomly selected miRNA targets predicted by Krek *et al.* (16) have been proved to be true targets (24,25). The accuracy of these methods has also been proved by gene expression profiling studies (26,27). These methods have yielded a large number of candidate targets in both plants and animals. The estimated human miRNA targets can account for up to one third of the human genes (16,20,24).

Consequently, a SNP located in the miRNA-binding site of a miRNA target (called miRNA-binding SNP in this study) is likely to disrupt miRNA–target interaction, resulting in the deregulation of target gene expression. In this regard, the effect of this type of SNPs on gene expression is predictable. Such SNP-associated deregulation of the expression of an oncogene or tumor suppressor might contribute to tumorigenesis. In this study, we conducted a genome-wide search for SNPs located in miRNA-binding sites of miRNA targets using the dbSNP database and comprehensively defined the display of each SNP in cancers versus normal tissues through mining the dbEST database. We thus identified a number of miRNA-binding SNPs with apparent cancer-associated aberrant allele frequencies, and confirmed through genotyping that some of these SNPs are indeed aberrantly present in tumors.

MATERIALS AND METHODS

Searching for SNPs located in the whole 3′-UTRs and SNPs complementary to the 'seed' sequences of miRNAs in the human genome

A 3′-UTR dataset and a miRNA target dataset of human genes were obtained from UCSC Genome browser (http://genome.ucsc.edu/cgi-bin/hgTables). The miRNA target dataset, developed by Krek *et al.* (16), contains the human genes, which, at 3′-UTR, have a 7-nt segment (called miRNA-binding 'seed' region here) complementary to the 'seed' region (the 7 nt from position 2 to position 8 at the 5′-terminus) of human miRNAs. A human SNP dataset (NCBI dbSNP Build 126) was obtained from NCBI databases (ftp://ftp.ncbi.nih.gov/snp). The genomic locations (chromosome number and nucleotide position) on human chromosome of the SNPs, the 3′-UTRs and the miRNA-binding 'seed' regions are all indicated in the individual datasets. The SNPs located in the 3′-UTRs and the SNPs located in the miRNA-binding 'seed' regions of human genes were identified using the chromosomal location information.

Allele distribution analysis of the SNPs located in miRNA-targeting sites in cancer EST libraries versus dbSNP database

Human expressed sequence tag (EST) libraries and EST sequences were obtained from NCBI databases (http://www.ncbi.nlm.nih.gov/projects/dbEST/). The EST libraries were manually curated and cataloged into cancer EST libraries and normal tissue EST libraries. Totally, 2.2 millions of EST sequences were obtained from 3721 cancer EST libraries and 1.9 millions of EST sequences from 2010 normal tissue EST libraries. MiRNA targets predicted by Krek *et al.* (16) were obtained. The SNPs located in the miRNA-binding regions (~30-nt long), called miRNA-binding SNPs, were identified by blast searching for the dbSNP database using the miRNA target sequences. The EST fragments representing each allele of miRNA-binding SNPs were then identified by blast-searching for the cancer EST libraries and normal tissue EST libraries respectively, using a 30-nt sequence surrounding the SNP site. The sequences for both alleles were used separately for searching. Only the sequences with 100% identity were picked up. For each SNP, the total number of ESTs corresponding to each allele identified from the cancer EST libraries and from normal tissue EST libraries were counted respectively and compared with the number found in the dbSNP database. Fisher's exact test was used to determine the significant difference of the allele frequency of each SNP by comparing the cancer EST libraries with that of the dbSNP database. The allelic distributions of some miRNA-binding SNPs were found in SNP500Cancer database (http://snp500cancer.nci.nih.gov/home_1.cfm) that contains the allele frequency of the SNPs of cancer-related genes in the four populations (28). We also performed Fisher's exact test in order to compare the allele frequencies found in cancer EST libraries and each population in the SNP500Cancer database.

Determination of SNP allele frequency in cancer patients

About 200 tumor tissue specimens from Caucasian patients with various cancers were obtained from the Cooperative Human Tissue Network (CHTN) in the USA. Genomic DNAs were extracted from freshly frozen specimens using DNeasy^R Tissue kit following the manufacturer's protocol (Qiagen). As controls, 1000 genomic DNAs of normal subjects were obtained from the British 1958 birth cohort that is based on all persons born in Britain during one week in 1958, and additional 200 genomic DNAs (Caucasian) were from Coriell Institute for Medical Research. Collection and use of the tissue and genomic DNA samples were approved by the National Research Council Canada. For SNP allele identification, ~300 bp DNA fragments flanking the SNP of interest were amplified by PCR using the genomic DNAs. The PCR products were purified using MinElute 96 UF plates (Qiagen) and subjected to genotyping using one of the following methods. If one of the two alleles of a SNP can be digested by a specific restriction enzyme, a restriction enzyme digestion method was used. Otherwise, DNA sequencing method was used. For restriction enzyme digestion method, the digested PCR products were analyzed by agarose gel electrophoresis, which can distinguish the digested and undigested DNA fragments, for calculation of allele frequencies.

RESULTS

SNPs located in the miRNA-binding motifs of miRNA targets are negatively selected at the 3′-UTRs of human genes

Although hundreds of miRNAs have been identified, only a few of them have been functionally characterized, thus the biological functions of miRNAs are largely unknown. If the functions of miRNAs are, as assumed currently, critical for basic biological processes, we expect that the sequence variations, such as SNPs, in miRNA-binding sites of miRNA targets should undergo a purifying selection during evolution. To this end, we downloaded the miRNA targets predicted by Krek *et al.* (http://pictar.bio.nyu.edu/). Each of the predicted miRNA targets contains at least one 7-nt segment (called 'seed' region here) at the 3′-UTRs, which is crucial for the recognition of miRNAs. The SNPs located in the 'seed' regions are most likely to affect miRNA–target interaction and thus target expression. We therefore searched for the SNPs located in the 'seed' regions of human miRNA targets using the dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/). For comparison, we also searched for the SNPs that are located in the whole 3′-UTRs of human genes in the same database. We counted the density of the SNPs found in the miRNA-binding 'seed' regions (7 nt/seed) versus the whole 3′-UTRs. As shown in Table 1, 265 SNPs from 1 400 000 nucleotides, namely 0.182 SNP/kb, were located in the miRNA-binding 'seed' regions of 200 000 miRNA targets, whereas 20 588 SNPs from 96 484 523 nucleotides (the total number of nucleotides at the 3′-UTRs of human genes obtained for this study), namely 0.213 SNP/kb, were identified in the whole 3′-UTRs of human genes, indicating that SNPs arise less frequently in the miRNA-binding sites of miRNA targets than in the entire 3′-UTR ($P < 0.022$). This result suggests that SNPs at the miRNA-binding 'seed' regions are negatively selected under evolutional pressure. Since purifying selection (negative selection) eliminates the mutations that have deleterious effects on function, the relative low density of miRNA-binding SNPs at the 3′-UTRs of human genes supports the important role of miRNA–target interaction.

Table 1. Comparison of the abundance of SNPs located in the miRNA-binding sites versus all the 3′-UTRs in the human genome

	SNP number	Base number	SNP/kilobase	*P*-value
3′UTR	20 588	96 484 523	0.213	0.022
miRNA target	265	1 400 000	0.182	

Note: *P*-value was calculated using Fisher's exact test.

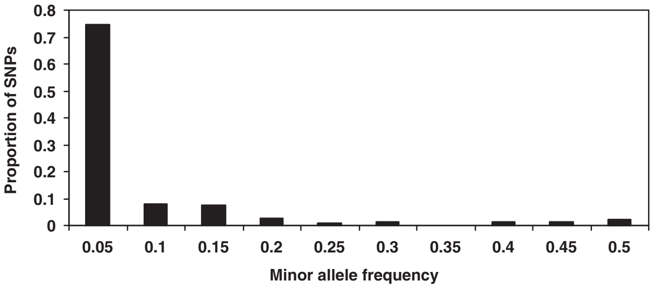

Figure 1. Distribution of the minor allele frequencies of miRNA-binding SNPs. The number of each allele of a miRNA-binding SNP found in the dbSNP database was counted and the frequency of the minor allele (ratio of the number of minor allele to total number of the two alleles) was calculated. The SNPs were then grouped according to their minor allele frequencies.

MiRNA-binding SNPs display a rare minor allele frequency

Since SNPs are mutations that occur throughout evolution, natural selection should limit the alleles that have deleterious effects on function with time and thus limit the frequency of these harmful alleles. As shown in Figure 1, we found that more than 70% of the miRNA-binding SNPs have a minor allele frequency less than 0.5% in the dbSNP database. This level is higher than the average level of all SNPs found in the same dbSNP database (~10 and 25% observed in the HapMap and ENCODE datasets, respectively) and the expected distribution under the standard neutral model (~50%). This result further confirms the important biological function of miRNAs.

SNPs in the miRNA-binding motifs affects miRNA target expression

To further confirm the essential biological function of miRNAs as elucidated above by the SNP density analysis, we determined the expression variation of different alleles of miRNA-binding SNPs via mining the dbEST database. Because the EST libraries were constructed from cDNAs, the relative frequency of the two alleles for each SNP

Table 2. Comparison of allele frequency of miRNA-binding SNPs in the dbEST database and in the dbSNP database

Number of SNPs with a higher frequency of the non-target alleles in the dbEST database than in the dbSNP database	59
Number of SNPs with a higher frequency of the non-target alleles in the dbSNP database than in the dbEST database	35
Average ratio of non-target allele frequency to target allele frequency in the dbEST database	0.127
Average ratio of non-target allele frequency to target allele frequency in the dbSNP database	0.086

found in the dbEST database should be similar to that found in the dbSNP database, if the SNP does not affect gene expression. However, if a SNP affects gene expression, the two alleles of this SNP are likely differentially expressed and thus the relative frequency of the two alleles present in the dbEST database might be different from that found in the dbSNP database. Since the 7-nt ‘seed’ region of a miRNA is the most important sequence for miRNA–target interaction and miRNA function, further analysis was conducted by focusing on these regions. For convenience, we termed the allele of a miRNA-binding SNP that has the sequence complementary to the ‘seed’ region of the miRNA as ‘target allele’ and the other with one mismatch as ‘non-target allele’. Among the 930 miRNA-binding SNPs we identified, 297 are located in the miRNA-binding ‘seed’ regions. Half of these 297 SNPs (129) have allele frequency data available in the dbSNP database. We then calculated the relative frequency of the target allele and the non-target allele for each of these 129 SNPs in the dbEST database and in the dbSNP database, respectively. We found that 35 out of the 129 SNPs have a zero frequency of non-target allele in both databases, indicating that the non-target alleles (the minor alleles) of these SNPs are too rare to be detected. Therefore, these 35 were not used for comparison. However, 59 (63%) of the remaining 94 have a higher frequency of non-target alleles in the dbEST database than in the dbSNP database, whereas only 35 (37%) of them have a higher frequency of non-target alleles in the dbSNP database than in the dbEST database (Table 2). More importantly, the average frequency ratio of the non-target alleles over target alleles for these SNPs is significantly higher in the dbEST database than in the dbSNP database (0.127 versus 0.086) (Table 2), suggesting that the average expression level of the non-target alleles of miRNA-binding SNPs is higher than that of the target alleles. These results indicate that the SNPs/mutations located in the miRNA-binding ‘seed’ regions of miRNA targets can disrupt miRNA–target interaction and lead to up-regulation of miRNA target expression.

Identification of miRNA targets with aberrant SNP allele frequency in human tumors

To identify the miRNA-binding SNPs that may contribute to cancer susceptibility, we carried out a genome-wide scale *in silico* analysis of the dbSNP database in conjunction with the human dbEST database. To do so, we downloaded the human EST sequences derived from both the cancer EST libraries and the normal tissue EST libraries. In total, 2.2 million EST sequences from 3721 human cancer EST libraries and 1.9 million EST sequences from 2010 human normal tissue EST libraries were obtained. We then searched for the ESTs that correspond to each allele of those SNPs located in the miRNA-binding regions of miRNA targets in the cancer EST libraries and in the normal tissue EST libraries (Table S1). The miRNA targets were obtained from Krek *et al.*’s report (16). A 30-nt segment for each miRNA-binding site was entered in the original database and used for SNP searching in this study. The allele frequency of each SNP was also obtained from the dbSNP database (Table S1). To determine whether the allele frequency of these SNPs in the cancer EST libraries is similar to or different from that found in the general population, we counted the number of ESTs for each SNP allele found in cancer EST libraries and compared it with that of the general population as found in the dbSNP database. Fisher’s exact test was used to determine the statistical significance of the variations observed (Figure 2). The SNPs above the horizontal line of the figure have a *P*-value less than 0.01 (−log (*P*-value) > 2, Figure 2).

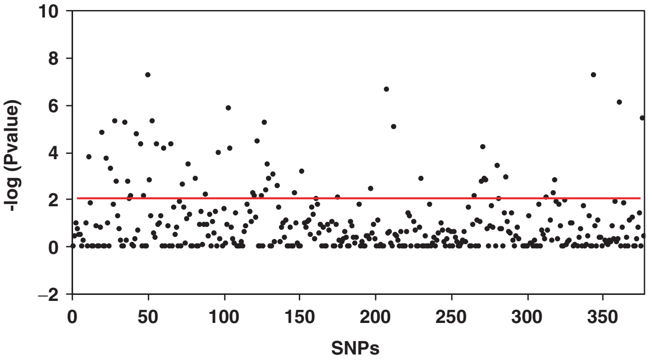

Figure 2. Comparison of the allele frequency of miRNA-binding SNPs in the cancer EST libraries and in the dbSNP database. EST sequences containing 30 nt with 100% homology to one of the two SNP alleles were identified from the cancer EST libraries. For each individual SNP, the total number of ESTs with each allele was counted and compared with that found in the dbSNP database. A Fisher’s exact test was performed to determine the significance of the different distribution of the two alleles in the cancer EST libraries versus the control population found in the dbSNP database. The samples above the red line have a *P*-value less than 0.01.

Experimental validation of miRNA target SNPs with an aberrant SNP allele frequency in human tumor tissues

The *in silico* analysis of the dbEST database and dbSNP database provided us with some potential candidates for miRNA-binding SNPs with an aberrant allele frequency present in the human cancer EST database and filtered out many SNPs (the majority) that have similar allele distribution in both databases. We next genotyped genomic DNAs derived from human cancer tissues in order to experimentally validate these potential miRNA-binding SNPs. We compared the allele frequency of each SNP found in cancer tissues with that found in the

Table 3. MiRNA-binding SNPs with an aberrant SNP allele frequency in human tumor tissues

SNP	miRNA target		Count in dbSNP		Count in cancers		MAF (ctrl)	MAF (case)	Odds ratio [95% CI]	*P*-value
	Gene name	Gene ID	Major allele	Minor allele	Major allele	Minor allele				
rs1044129[a]	RYR3	NM_001036	250	118	208	172	0.3207	0.4526	1.75 [1.30, 2.36]	4.9×10^{-10}
rs4901706	C14orf101	NM_017799	644	4	346	18	0.0062	0.0495	8.37 [3.34, 21.00]	1.1×10^{-5}
rs12583	DAG1	NM_00493	375	225	162	46	0.3750	0.2212	0.47 [0.33, 0.68]	1.5×10^{-5}
rs1128665	STK40	NM_032017	353	15	356	0	0.0408	0		5.4×10^{-5}
rs16917496	SETD8	NM_020382	287	81	248	136	0.2201	0.3542	1.94 [1.41, 2.68]	5.5×10^{-5}
rs17703261	AFF1	NM_005935	62	26	336	48	0.2955	0.1250	0.34 [0.20, 0.58]	2.5×10^{-4}
rs1053667	KIAA0423	NM_015091	681	19	174	16	0.0271	0.0842	3.29 [1.72, 6.32]	1.0×10^{-3}
rs11337	GOLGA7	NM_00102296	690	14	310	18	0.0199	0.0549	2.86 [1.45, 5.66]	3.6×10^{-3}
rs14109	MATR3	NM_108834	596	12	362	20	0.0197	0.0524	2.74 [1.36, 5.53]	8.6×10^{-3}
rs3660	KRT81	NM_002281	385	281	175	177	0.4219	0.5028	1.39 [1.07, 1.80]	0.014
rs17107469	SH2D4B	NM_207372	326	0	188	4	0	0.0208		0.018
rs10463	USP9X	NM_001039590	478	26	171	19	0.0516	0.1000	2.04 [1.11, 3.74]	0.025

*Fisher's exact test was used to calculate the *P*-values based on the minor allele frequencies found in the human tumor specimens against the total Caucasian subjects present in the dbSNP database. Minor allele frequencies (MAF) and odds ratios with 95% confidence intervals (95% CI) are shown for each SNP except rs1128665 and rs17197469, the counts of the minor alleles of which are zero.

[a] The minor allele of rs1044129 is 'A' and the major allele is 'G' in the dbSNP database, but 'G' is the minor allele and 'A' is the major allele in the cancer samples. The minor allele frequency (MAF) shown in the table for this SNP is the frequency of 'A' in dbSNP and the frequency of 'G' in the cancer population. The frequency of 'A', which is 0.32 in the dbSNP database and 0.55 in the cancer population, was used to calculate odds ratio.

dbSNP database and in the normal subjects. From 65 SNPs tested, we found that 12 have a significantly aberrant SNP allele frequency ($P < 0.05$) in human cancer tissues as compared with the control present in the dbSNP database (Table 3 and Table S2). The aberrant allele frequencies of these SNPs were also found in unaffected tissues, located near these cancer tissues examined, suggesting their germ-line origin. Figure 3 shows the representative sequences of SNP rs16917496 derived from two colon tumor samples. To determine if the number of SNPs (12 out of 65) that have an aberrant allele frequency in cancer samples compared to the dbSNP database is statistically significant, we did genotyping analysis of 16 of the 65 SNPs using normal subjects. We found that none of the 16 SNPs have a significant difference of allele frequency in the normal subjects from that in the dbSNP database ($P > 0.1$ for all of them, Table S2), whereas 2 of the 16 SNPs have an aberrant allele frequency in cancer population (Table S2, rs1044129 and rs17107469). These results suggest that the *in silico* analysis method provides us with useful preliminary data for the identification of miRNA-binding SNPs that may affect miRNA target expression, possibly leading to cancer susceptibility. These SNPs with aberrant allele frequencies in cancer can be used as potential tumor markers and drug targets for early cancer detection and prevention.

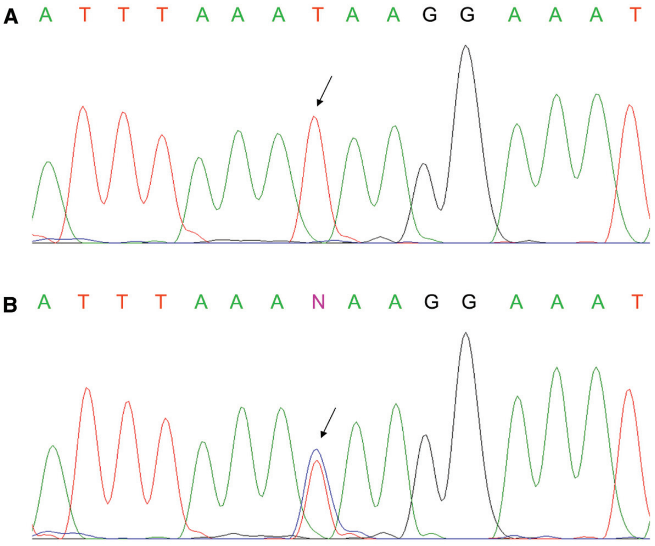

Figure 3. Representative sequencing results of two colon tumor samples validating the aberrant allele frequency of SNP rs16917496. (**A**) A colon tumor sample displays a normal homozygous T/T allele. (**B**) A colon tumor sample displays a heterozygous T/G allele. Arrows indicate the SNP position.

DISCUSSION

Under the neutral theory of molecular evolution, the majority of DNA variations observed in a population are due to random drift of neutral or nearly neutral mutations. Natural selection, such as purifying selection, may eliminate those mutations that have deleterious effects on function. This will reduce the ratio of those mutations over neutral mutations observed in the present population. In this study, we first found that SNP density is lower in the miRNA-binding motifs ('seed' regions) than in the 3′-UTRs, potentially caused by purifying selection. In addition, since SNPs are the result of mutations that occurred one time in human history, natural selection may also lead to rare allele frequencies of the mutations that could not be completely eliminated under rapid population expansion, despite their deleterious effects on function. Indeed, we found that the frequencies of the minor alleles (non-target alleles) of the miRNA-binding SNPs are extremely low (0.079) based on the dbSNP database analysis, which may reflect the deleterious effect of nucleotide substitutions at miRNA-binding sites. Taken together, both the negative selection (lower density) and the rare allele frequencies of miRNA-binding SNPs at the miRNA-binding sites, reflect the importance of miRNAs.

Rare SNP alleles may have severe consequences and thus cause various human diseases. By mining the dbEST database and dbSNP database, we identified a number of miRNA-binding SNPs that have aberrant allele frequency in the cancer EST libraries. However, using EST libraries for this analysis has some limitations caused by the quality of the EST sequences, biased sampling and biased ethnic origin. To reduce the effect of these limitations on the accuracy of the results, we performed further analysis. First, we manually searched for these interesting miRNA target genes that have an aberrant allele frequency in the cancer EST libraries compared to that in the dbSNP database, as identified from the *in silico* screen. We manually detected the quality of each EST sequence to ensure that only the ESTs with high quality are used for our final statistical analysis of these interesting SNPs. In addition, to determine the effect of EST quality on the results, we searched for the EST sequences that contain any two other nucleotides ('non-SNP' nucleotides) to replace the defined SNP alleles at the SNP position. We assumed that any EST sequence with the substitution of two other 'non-SNP' nucleotides was caused by sequencing errors and that the number of sequencing errors causing change of one SNP allele to the other SNP allele is similar to the number of errors leading to the change of a SNP allele to a 'non-SNP' allele (non-existing allele). Using these data, we can calculate the contribution of EST sequence quality to the errors of SNP allele distribution. We found that among 18 845 ESTs located in the two SNP alleles, 24 ESTs were substituted by the 'non-SNP' nucleotides, revealing a 0.13% error in our analysis (0.65% for the minor alleles and 0.07% for the major alleles, Table 4). To reduce sampling bias, we counted only one EST for a specific allele of a SNP if more than one EST for the allele were found in the same library. However, many different EST libraries might be constructed with the cDNAs derived from the same donor. EST sequences representing a specific SNP allele found in different libraries derived from the same donor may be counted several times although only one count should be used for statistical analysis. We could not avoid this kind of sampling bias since the identity of donors for the EST library was not available. For the same reason, it is hard to exclude the effect of bias of ethnic origin. Nevertheless, to reduce the limitation caused by the lack of information of the ethnic origin of donors in our statistical analysis, we first compared the allele frequency found in the cancer EST database with that of the dbSNP database from all four populations and also with that from each individual population, if data from different populations was available. In addition, we used the SNP500Cancer database where the SNP allele frequency data from all four populations are available (Table S3). In spite of these limitations, we assessed the efficacy of *in silico* analysis through experimental studies using human cancer samples. We successfully confirmed by experimental validation 12 out of 65 miRNA-binding SNPs with an aberrant allele frequency in the cancer EST libraries. These results indicate that the aberrant allele frequencies of the miRNA-binding SNPs present in tumors, might be an important factor contributing to tumorigenesis. In addition, these results demonstrated the usefulness of the primary *in silico* analysis, as it was able to filter out the majority of the miRNA-binding SNPs that showed no difference of allele frequencies between the cancer EST database and the dbSNP database.

Table 4. Contribution of EST sequence quality to the errors of the analysis

	Major alleles	Minor alleles	Non-SNP alleles
Number of ESTs	17016	1829	24
Percentage of errors*	0.07%	0.65%	
	0.13%		

*Errors were calculated as:
$24/(17016 + 1829) \times 100\% = 0.13\%$;
$(24/2)/17016 \times 100\% = 0.07\%$;
$(24/2)/1829 \times 100\% = 0.65\%$.

SUPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Denis L'Abbe for technical assistance and Dr Krikor Bijian for comments on the manuscript. We acknowledge use of DNA from the British 1958 Birth Cohort collection, funded by the Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02, and thank D. Strachan, R. Jones and all other members of the 1958 Cohor Biomedical Study Team. This work was supported in part by the Canadian Institute of Health Research Grant MOP82807. Funding to pay the Open Access publication charges for this article was provided by the National Research Council Canada.

Conflict of interest statement. None declared.

REFERENCES

1. Ambros,V. (2001) microRNAs: tiny regulators with great potential. *Cell*, **107**, 823–826.
2. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
3. Cullen,B.R. (2004) Transcription and processing of human microRNA precursors. *Mol. Cell*, **16**, 861–865.
4. Kim,V.N. (2005) MicroRNA biogenesis: coordinated cropping and dicing. *Nat. Rev. Mol. Cell Biol.*, **6**, 376–385.
5. Carmell,M.A. and Hannon,G.J. (2004) RNase III enzymes and the initiation of gene silencing. *Nat. Struct. Mol. Biol.*, **11**, 214–218.
6. Meister,G. and Tuschl,T. (2004) Mechanisms of gene silencing by double-stranded RNA. *Nature*, **431**, 343–349.
7. Griffiths-Jones,S., Grocock,R.J., van,D.S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
8. Ambros,V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350–355.

9. Calin,G.A., Sevignani,C., Dumitru,C.D., Hyslop,T., Noch,E., Yendamuri,S., Shimizu,M., Rattan,S., Bullrich,F. *et al.* (2004) Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc. Natl. Acad. Sci. USA*, **101**, 2999–3004.
10. Lu,J., Getz,G., Miska,E.A., Alvarez-Saavedra,E., Lamb,J., Peck,D., Sweet-Cordero,A., Ebert,B.L., Mak,R.H. *et al.* (2005) MicroRNA expression profiles classify human cancers. *Nature*, **435**, 834–838.
11. Takamizawa,J., Konishi,H., Yanagisawa,K., Tomida,S., Osada,H., Endoh,H., Harano,T., Yatabe,Y., Nagino,M. *et al.* (2004) Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. *Cancer Res.*, **64**, 3753–3756.
12. Calin,G.A., Dumitru,C.D., Shimizu,M., Bichi,R., Zupo,S., Noch,E., Aldler,H., Rattan,S., Keating,M. *et al.* (2002) Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc. Natl. Acad. Sci. USA*, **99**, 15524–15529.
13. Calin,G.A., Ferracin,M., Cimmino,A., Di,L.G., Shimizu,M., Wojcik,S.E., Iorio,M.V., Visone,R., Sever,N.I. *et al.* (2005) A microRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *N. Engl. J. Med.*, **353**, 1793–1801.
14. Bernig,T. and Chanock,S.J. (2006) Challenges of SNP genotyping and genetic variation: its future role in diagnosis and treatment of cancer. *Expert. Rev. Mol. Diagn.*, **6**, 319–331.
15. Brennecke,J., Stark,A., Russell,R.B. and Cohen,S.M. (2005) Principles of microRNA-target recognition. *PLoS Biol.*, **3**, e85.
16. Krek,A., Grun,D., Poy,M.N., Wolf,R., Rosenberg,L., Epstein,E.J., MacMenamin,P., da Piedade,I., Gunsalus,K.C. *et al.* (2005) Combinatorial microRNA target predictions. *Nat. Genet.*, **37**, 495–500.
17. Enright,A.J., John,B., Gaul,U., Tuschl,T., Sander,C. and Marks,D.S. (2003) MicroRNA targets in Drosophila. *Genome Biol.*, **5**, R1.
18. John,B., Enright,A.J., Aravin,A., Tuschl,T., Sander,C. and Marks,D.S. (2004) Human MicroRNA targets. *PLoS Biol.*, **2**, e363.
19. Kiriakidou,M., Nelson,P.T., Kouranov,A., Fitziev,P., Bouyioukos,C., Mourelatos,Z. and Hatzigeorgiou,A. (2004) A combined computational-experimental approach predicts human microRNA targets. *Genes Dev.*, **18**, 1165–1178.
20. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
21. Rhoades,M.W., Reinhart,B.J., Lim,L.P., Burge,C.B., Bartel,B. and Bartel,D.P. (2002) Prediction of plant microRNA targets. *Cell*, **110**, 513–520.
22. Robins,H., Krasnitz,M., Barak,H. and Levine,A.J. (2005) A relative-entropy algorithm for genomic fingerprinting captures host-phage similarities. *J. Bacteriol.*, **187**, 8370–8374.
23. Stark,A., Brennecke,J., Russell,R.B. and Cohen,S.M. (2003) Identification of Drosophila MicroRNA targets. *PLoS Biol.*, **1**, E60.
24. Rajewsky,N. (2006) MicroRNA target predictions in animals. *Nat. Genet.*, **38**(Suppl.), S8–S13.
25. Stark,A., Brennecke,J., Bushati,N., Russell,R.B. and Cohen,S.M. (2005) Animal microRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. *Cell*, **123**, 1133–1146.
26. Lim,L.P., Lau,N.C., Garrett-Engele,P., Grimson,A., Schelter,J.M., Castle,J., Bartel,D.P., Linsley,P.S. and Johnson,J.M. (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, **433**, 769–773.
27. Krutzfeldt,J., Rajewsky,N., Braich,R., Rajeev,K.G., Tuschl,T., Manoharan,M. and Stoffel,M. (2005) Silencing of microRNAs in vivo with 'antagomirs'. *Nature*, **438**, 685–689.
28. Packer,B.R., Yeager,M., Burdett,L., Welch,R., Beerman,M., Qi,L., Sicotte,H., Staats,B., Acharya,M. *et al.* (2006) SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. *Nucleic Acids Res.*, **34**, D617–D621.