Published online 6 December 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D51–D54***
doi:10.1093/nar/gkl797

PolymiRTS Database: linking polymorphisms in microRNA target sites with complex traits

Lei Bao[1,2], Mi Zhou[1,2], Ligang Wu[5], Lu Lu[2,3,6], Dan Goldowitz[2,3], Robert W. Williams[2,3,4] and Yan Cui[1,2,*]

[1]Department of Molecular Sciences, [2]Center of Genomics and Bioinformatics and [3]Department of Anatomy and Neurobiology, [4]Department of Pediatrics, University of Tennessee Health Science Center, Memphis, TN 38163, USA, [5]Skirball Institute of Biomolecular Medicine, New York University School of Medicine, NY, USA and [6]Key Laboratory of Nerve Regeneration, Nantong University, Nantong, Jiangsu Province, China

Received August 15, 2006; Revised September 20, 2006; Accepted October 2, 2006

ABSTRACT

Polymorphism in microRNA Target Site (PolymiRTS) database is a collection of naturally occurring DNA variations in putative microRNA target sites. PolymiRTSs may affect gene expression and cause variations in complex phenotypes. The database integrates sequence polymorphism, phenotype and expression microarray data, and characterizes PolymiRTSs as potential candidates responsible for the quantitative trait locus (QTL) effects. It is a resource for studying PolymiRTSs and their implications in phenotypic variations. PolymiRTS database can be accessed at http://compbio.utmem.edu/miRSNP/.

INTRODUCTION

Identification of causal genetic variants underlying complex traits is a major goal in genetic studies. Linkage analysis has long been used to discover chromosomal intervals harboring sequence variants that cause variations in quantitative traits. A typical quantitative trait locus (QTL) interval usually contains many genes ranging from several dozens to several hundreds, hence it is critical to be able to focus on genetic variants in the interval that are most likely to have functional impacts. Among them, nonsynonymous single nucleotide polymorphisms (SNPs) that alter protein sequences and regulatory polymorphisms that affect gene expression are natural high-priority candidates. Although regulatory polymorphisms are much more challenging to be identified and characterized, experimental and analytical tools are being actively developed for this purpose (1–4). Polymorphisms in miRNA target sites (PolymiRTS) represent a specific class of regulatory polymorphisms that may regulate posttranscriptional gene expression. A recent work reports that PolymiRTS can underlie the effects of physiological QTLs (pQTLs) that control classic higher order traits (5). MicroRNAs (miRNAs) are a family of small RNAs that pair to the transcripts of protein-coding genes and cause translational repression or mRNA destabilization (6,7). Hundreds of miRNAs have been identified in humans and mice and many of them have been shown to regulate their target genes that control diverse biological processes such as differentiation, proliferation, apoptosis and morphogenesis (7). PolymiRTS may affect the base-pairing process, hence affect the miRNA-mediated gene repression which in turn cause phenotypic variations. It has been found that miRNA-mediated target mRNA destabilization is widespread in mammals (8–10). Thus, for miRNAs acting by this mechanism, the PolymiRTS may lead to heritable variations in gene expression. Variations in gene expression across a population can be assessed by a newly developed genetical genomics approach (11–13). The genetical genomics approach treats gene expression level as quantitative trait. Linkage mapping is then used to discover the genetic loci regulating gene expression traits (eQTLs). PolymiRTS may induce a *cis*-acting eQTL that coincides with the gene's physical location. We proposed a simple conceptual model (Figure 1) that represents information flow from PolymiRTS to complex trait via *cis*-acting eQTL. Based on this model scheme, we create a database integrating SNP, phenotype and expression microarray data of human and mouse.

DATA SOURCES AND PROCESSING

Identifying and annotating PolymiRTS

The method of identifying and annotating PolymiRTS is shown in Figure 2. SNPs that are located in the 3′-untranslated regions (3′-UTRs) of all known genes by UCSC genome annotation (mouse: mm7 and human: hg18) (14) were extracted from dbSNP build 126 (15). Genomic locations of these SNPs were mapped onto mRNAs. For each SNP, we assessed whether its two alleles lead to different miRNA target sites. To be conservative, we only consider

[*]To whom correspondence should be addressed at Department of Molecular Sciences, University of Tennessee Health Science Center, 858 Madison Avenue, Memphis, TN 38163, USA. Tel: +1 9014483240; Fax: +1 9014487360; Email: ycui2@utmem.edu

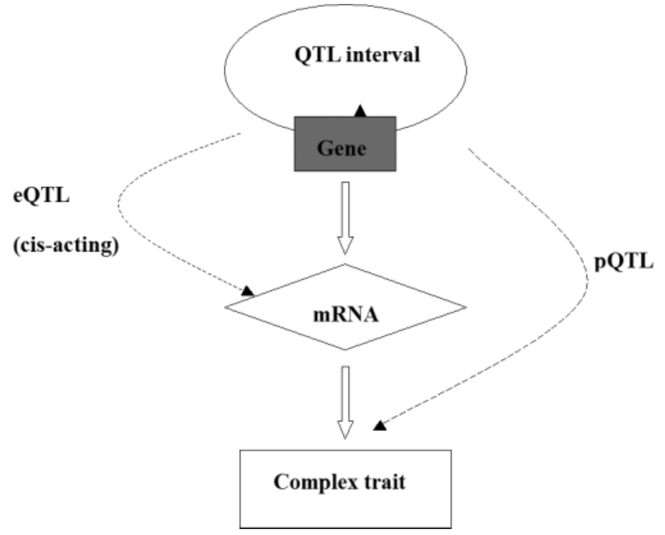

Figure 1. Conceptual QTL model. A PolymiRTS (triangle) may cause the gene expression variation (diamond) in segregating population and a *cis*-acting eQTL is observed. The variation in gene expression in turn may cause phenotype variation (rectangle) and a pQTL is observed.

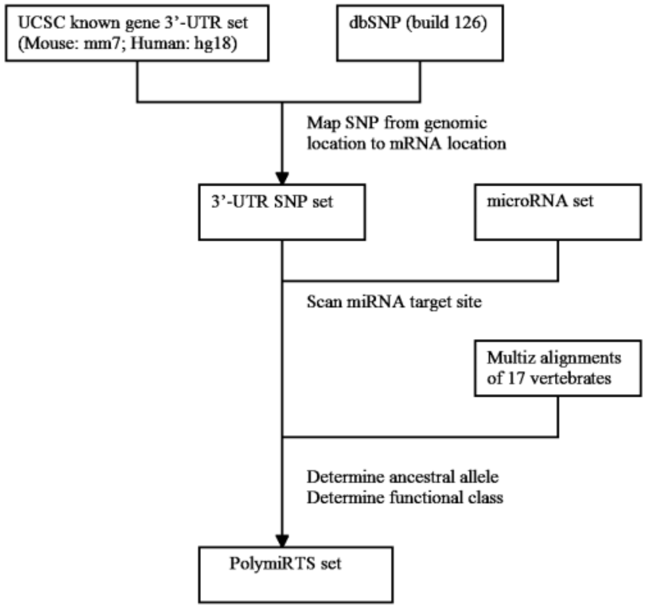
UCSC known gene 3'-UTR set (Mouse: mm7; Human: hg18)
dbSNP (build 126)
Map SNP from genomic location to mRNA location
3'-UTR SNP set
microRNA set
Scan miRNA target site
Multiz alignments of 17 vertebrates
Determine ancestral allele
Determine functional class
PolymiRTS set

Figure 2. Methods of identifying and annotating PolymiRTS.

the 3′-UTR SNPs that affect the match to the seed region of the miRNA. Mature miRNA sequences were downloaded from the miRBase (16). We used the criteria of TargetScanS (17) in the prediction of miRNA sites. Basically, besides requiring a perfect Watson–Crick match to the seed 2–7 nt of miRNA, we further require that there is either a perfect match to the 8th nt of miRNA, or an anchor adenosine immediately downstream the 2–7 seed in the target. We assigned the PolymiRTS to one of the four classes: 'D' (an allele disrupts a conserved miRNA site), 'N' (a derived allele disrupts a nonconserved miRNA site), 'C' (a derived allele creates a new miRNA site) and 'O' (other cases when the ancestral allele can not be determined unambiguously). PolymiRTS of class 'C' may cause abnormal gene repression and PolymiRTS of class 'D' may cause loss of normal repression control. These two classes of PolymiRTS are most likely to have functional impacts. We used the pre-calculated 17-way Multiz alignments of vertebrate genomes to derive the annotations. For a miRNA site to be conserved, we require that it is present in at least two other vertebrate genomes in addition to the query genome. For mouse SNPs, their ancestral alleles were determined by mouse versus rat (rn3) genome alignment. For human SNPs, their ancestral alleles were determined by human versus chimpanzee (panTro1) genome alignment. Additionally, we categorized PolymiRTS with A/G alleles because they are supposed to be less deleterious with their ability to form G:U wobble base-pairs with miRNAs.

PolymiRTSs within *cis*-acting expression QTL (eQTL) intervals

The genes with both *cis*-acting eQTL and PolymiRTS are featured in the database. First, gene expression levels in cerebellum, hippocampus, striatum, eye, whole brain and hematopoietic cell (18) were assessed in recombinant inbred mouse strains (BXD) derived from two parental strains C57BL/6J and DBA/2J. Gene expression levels were treated as quantitative traits and were mapped onto genomic regions (eQTL) using standard marker regression. A gene is said to have a suggestive (significant) *cis*-acting eQTL, if the LOD (log of odds) peak location is within 10 MB from the gene's physical location and the LOD >2.8 (>4.3) (19). Second, gene expression levels in lymphoblastoid cells of 194 human individuals from 14 CEPH families (20) were downloaded from the GEO database (21) and the raw data were processed by using the RMA protocol (22). Genotypes for 1628 autosomal SNP markers were downloaded from The SNP Consortium database (23). We used Merlin (24) to remove genotype errors and perform family-based linkage analysis. A gene is said to have a *cis*-acting eQTL, if the LOD peak location is within 10 MB from the gene's physical location and the *P*-value is <0.05.

PolymiRTSs in physiological QTL (pQTL) intervals

We first mapped the QTLs (with a LOD >2.8) for more than 800 published BXD phenotypes (physiological/behavioral traits) (18). For each QTL, we linked it with genes that are physically located in the QTL interval and have at least one PolymiRTS. These genes are candidate causal genes underlying the pQTL.

DATABASE CONTENT AND ACCESS

Table 1 shows the major data fields for a typical PolymiRTS record. The users can access the database by several options. First, a web interface is implemented for browsing the entries. Second, a text search interface is designed for query by SNP ID, miRNA ID, GenBank accession, HUGO gene identifier and gene description. Third, a chromosome location search is offered so that the users can specify a genomic interval of interest and search all the PolymiRTSs within the interval. For mouse, we also provide the inbred strain comparison option. By combining strain comparison with the range

Table 1. Field description

Field	Description
Location	SNP location in the mRNA transcript
SNPID	Link to dbSNP
Wobble base-pair	Whether the SNP can form a G:U wobble basepair with the miRNA. Y: Yes; N: No
Ancestral allele	If applicable, ancestral allele is denoted
Allele	Two alleles of the SNP in the mRNA transcript
Strain	Genotypes of two mouse inbred strains to be compared
miRID	Link to miRBase
Support	Occurrence of miRNA site in other vertebrate genomes
Function Class	C: derived allele creates a new miRNA sites N: derived allele disrupts a nonconserved miRNA site D: allele disrupts conserved a miRNA site O: other cases when the ancestral allele cannot be determined unambiguously
miRSite	Sequence context of the miRNA site. Seed region are in capital letters and SNPs are highlighted in red

search, the user can retrieve all the PolymiRTSs that are candidate variants underlying the pQTL identified by them. Finally, we provide flat file downloads for all the data.

DISCUSSIONS AND FUTURE WORK

miRNAs regulate posttranscriptional gene expression through mRNA destabilization or translational repression. Destabilization of mRNA may cause variations in the transcript levels while translational repression does not (25–27). The effects of miRNAs that act through translational repression cannot be detected by expression microarrays. Therefore, in the PolymiRTS database, the sequence variations in the target sites of miRNAs that act through translational repression may be linked to pQTLs but not to eQTLs.

A recent study shows that there are two broad categories of miRNA target sites: (i) 5′-dominant sites with sufficient 5′-pairing, and (ii) 3′-compensatory sites with insufficient 5′-pairing which require strong 3′-pairing (28). By using the TargetScanS algorithm alone, we are likely to miss many target sites from the second category. Because some mismatches in the 5′-pairing region of 3′-compensatory target sites are tolerated which is difficult to estimate the influence of SNPs on the function of 3′-compensatory target sites. Hence, at this stage we want to be conservative and focus on the major and well-studied category of 5′-dominant sites.

Currently, we did not take miRNA gene expression pattern into consideration due to lack of large-scale miRNA expression data. However, in many cases, such information would be very valuable. For example, it is known that a set of genes evolutionarily avoid having miRNA target site in their 3′-UTRs (8). They are called miRNA antitargets. Obviously, creation of a new miRNA site in an antitarget is most likely to have severe consequences. Since antitargets tend to be highly and specifically expressed in the tissue where the miRNA is expressed (8), we can use this information to pick PolymiRTS in the putative antitargets and further prioritize strong functional candidates. With the anticipated accumulation of such miRNA profiling data, we would expect a much better annotation of PolymiRTS in the near future.

ACKNOWLEDGEMENTS

This work was supported by a PhRMA Foundation grant (Y.C.), and NIH grants HD052472 (D.G.), AA014425 (L.L.) and DA021131 (R.W.W.). We thank the two anonymous reviewers for their helpful suggestions. Funding to pay the Open Access publication charges for this article was provided by a PhRMA Foundation grant (Y.C.), and NIH grants HD052472 (D.G.), AA014425 (L.L.) and DA021131 (R.W.W.).

Conflict of interest statement. None declared.

REFERENCES

1. Cowles,C.R., Hirschhorn,J.N., Altshuler,D. and Lander,E.S. (2002) Detection of regulatory variation in mouse genes. *Nature Genet*., **32**, 432–437.
2. Knight,J.C. (2005) Regulatory polymorphisms underlying complex disease traits. *J. Mol. Med*., **83**, 97–109.
3. Knight,J.C., Keating,B.J., Rockett,K.A. and Kwiatkowski,D.P. (2003) *In vivo* characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. *Nature Genet*., **33**, 469.
4. Ronald,J., Akey,J.M., Whittle,J., Smith,E.N., Yvert,G. and Kruglyak,L. (2005) Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. *Genome Res*., **15**, 284–291.
5. Clop,A., Marcq,F., Takeda,H., Pirottin,D., Tordoir,X., Bibe,B., Bouix,J., Caiment,F., Elsen,J.M., Eychenne,F. *et al*. (2006) A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nature Genet*., **38**, 813–818.
6. Ambros,V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350–355.
7. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
8. Farh,K.K.-H., Grimson,A., Jan,C., Lewis,B.P., Johnston,W.K., Lim,L.P., Burge,C.B. and Bartel,D.P. (2005) The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science*, **310**, 1817–1821.
9. 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.
10. Wu,L., Fan,J. and Belasco,J.G. (2006) MicroRNAs direct rapid deadenylation of mRNA. *Proc. Natl Acad. Sci. USA*, **103**, 4034–4039.
11. Bao,L., Wei,L., Peirce,J.L., Homayouni,R., Li,H., Zhou,M., Chen,H., Lu,L., Williams,R.W., Pfeffer,L.M. *et al*. (2006) Combining gene expression QTL mapping and phenotypic spectrum analysis to uncover gene regulatory relationships. *Mamm. Genome*, **17**, 575–583.
12. Jansen,R.C. and Nap,J.P. (2001) Genetical genomics: the added value from segregation. *Trends Genet*., **17**, 388–391.
13. Schadt,E.E. (2005) Exploiting naturally occurring DNA variation and molecular profiling data to dissect disease and drug response traits. *Curr. Opin. Biotechnol*., **16**, 647–654.
14. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res*, **12**, 996–1006.
15. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*., **29**, 308–311.
16. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*., **34**, D140–D144.
17. 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.
18. Chesler,E.J., Lu,L., Wang,J., Williams,R.W. and Manly,K.F. (2004) WebQTL: rapid exploratory analysis of gene expression and genetic networks for brain and behavior. *Nature Neurosci*., **7**, 485–486.
19. Lander,E. and Kruglyak,L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genet*., **11**, 241–247.

20. Morley,M., Molony,C.M., Weber,T.M., Devlin,J.L., Ewens,K.G., Spielman,R.S. and Cheung,V.G. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743–747.
21. Barrett,T., Suzek,T.O., Troup,D.B., Wilhite,S.E., Ngau,W.-C., Ledoux,P., Rudnev,D., Lash,A.E., Fujibuchi,W. and Edgar,R. (2005) NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic. Acids Res.*, **33**, D562–D566.
22. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
23. Thorisson,G.A. and Stein,L.D. (2003) The SNP Consortium website: past, present and future. *Nucleic Acids Res.*, **31**, 124–127.
24. Abecasis,G.R., Cherny,S.S., Cookson,W.O. and Cardon,L.R. (2002) Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genet.*, **30**, 97–101.
25. Bhattacharyya,S.N., Habermacher,R., Martine,U., Closs,E.I. and Filipowicz,W. (2006) Relief of microRNA-mediated translational repression in human cells subjected to stress. *Cell*, **125**, 1111–1124.
26. 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.
27. Schratt,G.M., Tuebing,F., Nigh,E.A., Kane,C.G., Sabatini,M.E., Kiebler,M. and Greenberg,M.E. (2006) A brain-specific microRNA regulates dendritic spine development. *Nature*, **439**, 283–289.
28. Brennecke,J., Stark,A., Russell,R.B. and Cohen,S.M. (2005) Principles of microRNA-target recognition. *PLoS Biol*, **3**, e85.
29. Krek,A., Grun,D., Poy,M.N., Wolf,R., Rosenberg,L., Epstein,E.J., MacMenamin,P., da Piedade,I., Gunsalus,K.C., Stoffel,M. *et al.* (2005) Combinatorial microRNA target predictions. *Nature Genet.*, **37**, 495–500.