Knowledge-based analysis of functional impacts of mutations in microRNA seed regions

ANINDYA BHATTACHARYA and YAN CUI

Department of Microbiology, Immunology and Biochemistry
and
Center for Integrative and Translational Genomics,
University of Tennessee Health Science Center, Memphis, TN 38163, USA

Fax, +1 901 448 7360; Email, AB – abhatta3@uthsc.edu; YC – ycui2@uthsc.edu

MicroRNAs are a class of important post-transcriptional regulators. Genetic and somatic mutations in miRNAs, especially those in the seed regions, have profound and broad impacts on gene expression and physiological and pathological processes. Over 500 SNPs were mapped to the miRNA seeds, which are located at position 2–8 of the mature miRNA sequences. We found that the central positions of the miRNA seeds contain fewer genetic variants and therefore are more evolutionary conserved than the peripheral positions in the seeds. We developed a knowledge-based method to analyse the functional impacts of mutations in miRNA seed regions. We computed the gene ontology-based similarity score GOSS and the GOSS percentile score for all 517 SNPs in miRNA seeds. In addition to the annotation of SNPs for their functional effects, in the present article we also present a detailed analysis pipeline for finding the key functional changes for seed SNPs. We performed a detailed gene ontology graph-based analysis of enriched functional categories for miRNA target gene sets. In the analysis of a SNP in the seed region of hsa-miR-96 we found that two key biological processes for progressive hearing loss 'Neurotrophin TRK receptor signaling pathway' and 'Epidermal growth factor receptor signaling pathway' were significantly and differentially enriched by the two sets of allele-specific target genes of miRNA hsa-miR-96.

[Bhattacharya A and Cui Y 2015 Knowledge-based analysis of functional impacts of mutations in microRNA seed regions. *J. Biosci.* **40** 791–798]
DOI 10.1007/s12038-015-9560-2

1. Introduction

MicroRNAs (miRNAs) are small, single-stranded RNA molecules that function as post-transcriptional regulators for many genes. It is estimated that over 60% of protein-coding genes are potential targets of miRNAs (Siomi and Siomi 2010). The miRNA seed region, positions 2–8 of the mature miRNAs, is particularly important for miRNA target recognition (Lewis *et al.* 2005). Many microRNAs are highly conserved across species (Bartel 2004). The seed regions are more conserved than the flanking regions, indicating their functional importance (Wheeler *et al.* 2009). Consistently, the seed regions contain fewer genetic variants compared to their flanking regions (Wheeler *et al.* 2009). But still, with the rapid advance of genome sequencing technologies in recent years, many genetic variants, including SNPs and INDELs, have been identified in miRNA seed regions (Bhattacharya *et al.* 2014). A recent study reports >60–70% changes in miRNA targets one nucleotide change within miRNA seeds (Hill *et al.* 2014). Another recent study on livestock species pointed out the potential regulatory polymorphisms from analyses of polymorphic microRNA genes (Jevsinek Skok

Keywords. Gene ontology; microRNA; mutation; semantic similarity

Supplementary materials pertaining to this article are available on the *Journal of Biosciences* Website at *http://www.ias.ac.in/jbiosci/oct2015/supp/Bhattacharya.pdf*

Published online: 28 September 2015

et al. 2013). Somatic mutations have also been identified in miRNA seed regions in cancer genomes (Bhattacharya *et al.* 2013). These genetic variants and somatic mutations may disrupt the interactions between miRNAs and their targets or create new targets, and therefore, rewire the miRNA regulatory networks. Genetic and somatic mutations in miRNA seed regions may affect gene expression and cause diseases. Recent studies have linked some of these mutations to various diseases. For example, a mutation in the seed region of miR-96 is associated with hearing loss (Mencia *et al.* 2009), and a mutation in the seed region of miR-184 is associated with the EDICT (endothelial dystrophy, iris hypoplasia, congenital cataract, and stromal thinning) Syndrome (Iliff *et al.* 2012).

We have created databases, PolymiRTS (Bao *et al.* 2007; Ziebarth *et al.* 2012; Bhattacharya *et al.* 2014) and SomamiR (Bhattacharya *et al.* 2013), for the analysis of genetic variants and somatic mutations in miRNAs and their binding sites. In this work, we developed a knowledge-based computational method to assess the functional impacts of genetic and somatic mutations in miRNA seed regions. First, we identified 517 SNPs in the seed regions of microRNAs. The second step was allele-specific prediction of miRNA

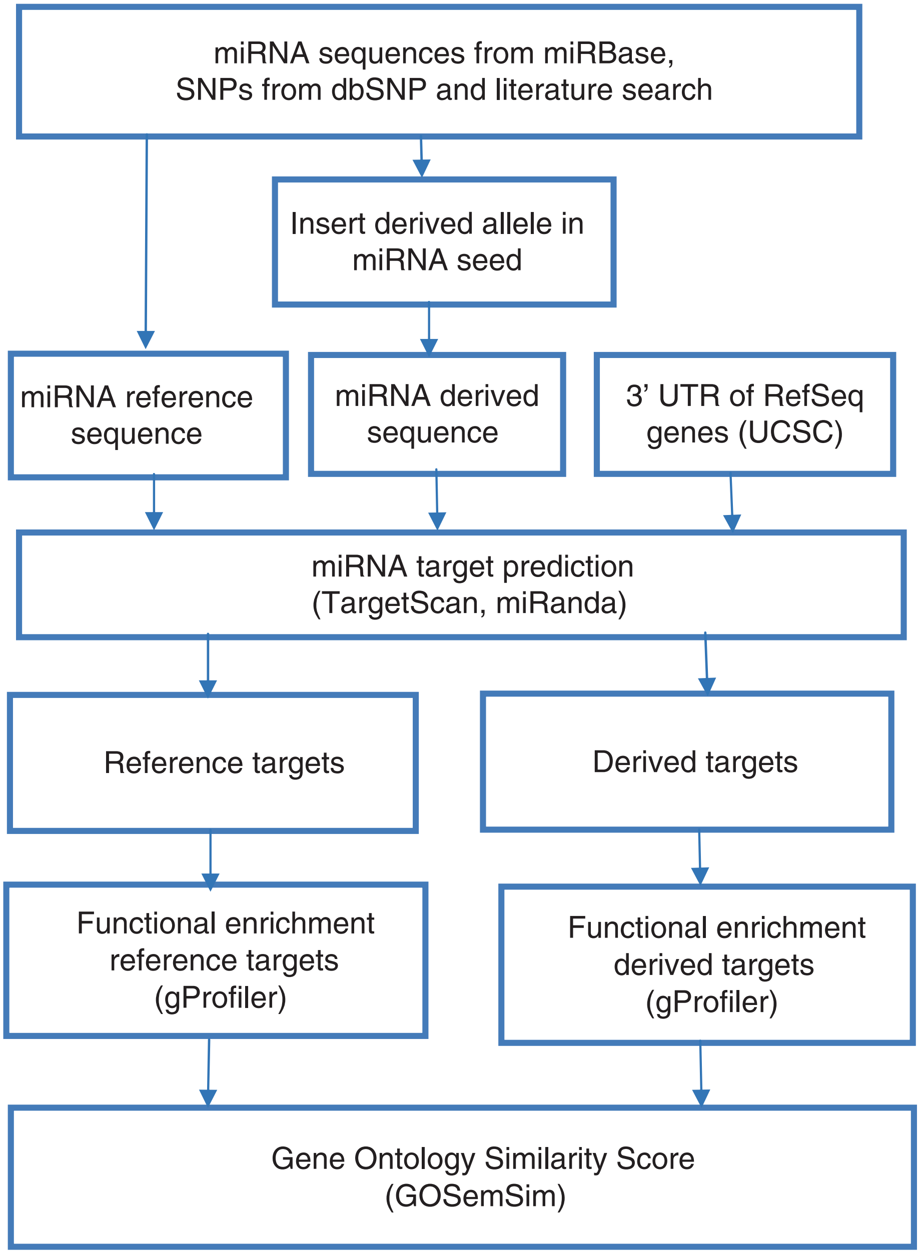
miRNA sequences from miRBase,
SNPs from dbSNP and literature search
Insert derived allele in miRNA seed
miRNA reference sequence
miRNA derived sequence
3' UTR of RefSeq genes (UCSC)
miRNA target prediction (TargetScan, miRanda)
Reference targets
Derived targets
Functional enrichment reference targets (gProfiler)
Functional enrichment derived targets (gProfiler)
Gene Ontology Similarity Score (GOSemSim)

Figure 1. Computational work flow for GOSS.

target sites for each SNP. Then we compared different sets of enriched functional annotations for target genes of reference (wild-type) and derived (minor) alleles. A score based on the semantic similarity of the annotation terms was calculated to quantify the overall impacts of each mutation. We also visualized the effects of the mutations by mapping the enriched annotations of allele-specific target genes and common target genes onto the Gene Ontology graphs.

2. Materials and methods

2.1 *Data sources and prediction tools*

Genomic locations of mature miRNAs were downloaded from miRBase (Kozomara and Griffiths-Jones, 2014). Genomic locations of miRNA seeds were determined from genomic locations of 2nd and 8th bases of mature miRNAs. SNPs in dbSNP build 138 (Sherry *et al.* 2001) were downloaded from the UCSC Table browser (Karolchik *et al.* 2004). We selected the table 'snp138' from track 'ALL SNPs (138)' of the UCSC Table browser and then used the genomic locations of seeds in a 'define region' window for downloading miRNA seed SNPs. We collected SNPs +13G>A in the seed of miRNA hsa-miR-96 and +57C>T in the seed of miRNA hsa-miR-184 from a literature search. Perl codes for TargetScan (release 6.2) and C codes for miRanda (release August 2010) were downloaded for target predictions. TargetScan determines target sites based on sequence complementarity between miRNA seed and target sites. On the other hand, miRanda combines the concept of free binding energy with the sequence complementarity between miRNA sequence and target sites. For TargetScan, we did not consider the sequence conservation. The miRNA sequences were downloaded from the miRBase (release 20) and 3′-UTR sequences of all RefSeq genes were downloaded from the UCSC Table browser. For identifying seed-to-seed miRNA SNPs we compared each miRNA seed with a derived allele of a SNP against the reference seed sequences of all the miRNAs for a match.

2.2 *Computing GOSS*

We computed the similarity score GOSS for each SNP in an miRNA seed. The procedure is summarized in Figure 1. The reference targets for a SNP were the predicted targets for miRNA when the reference allele of the SNP was in the miRNA seed sequence. On the other hand, the derived targets of a SNP were the predicted targets for miRNA when the derived allele of the SNP was in the miRNA seed sequence. We defined the common targets as predicted targets for both the reference and derived alleles of a SNP. Enriched gene ontology categories for reference and derived targets were obtained from gProfileR (Reimand *et al.* 2011), which is an R interface to the gene ontology tool gProfiler (Reimand *et al.* 2007). g:Profiler performs multiple testing correction by incremental enrichment analysis to report the adjusted p-values of enriched gene ontology categories (Reimand *et al.* 2007). Similar gene ontology categories were combined by selecting the hierarchical filtering option from gProfileR. Hierarchical filtering is used to group all the similar gene ontology categories in the hierarchy of the gene ontology graph. For each group of enriched gene ontology categories, the lowest enrichment p-value is considered as the representative category for the group. Functionally-enriched gene ontology categories for reference and derived targets were then compared for their semantic similarities. We used a Wang score from Wang et al (Wang *et al.* 2007) to measure semantic similarity between a pair of GO terms. Wang scores range between 0 and 1 where 0 indicates no similarity and 1 indicates complete similarity. The Wang similarity scores for each pair of GO terms were then combined to give GOSS as an average of maximum Wang scores. To compute Wang scores and the average of maximum Wang scores, we used the GOSemSim R library (Yu *et al.* 2010).

2.3 *Preparing gene ontology graph*

We used RamiGO (Schroder *et al.* 2013), the R/bioconductor package of the AmiGO gene ontology visualization tool for preparing Graphviz supported dot format files. The dot format file includes all the enriched gene ontology categories for reference, derived, and common targets and their parents in the gene ontology graph. The node sizes were determined from the number of target genes for the enriched categories. A larger node size was used for a large number of target genes. The node colours were determined from enriched gene ontology categories from reference, common and derived target groups. Blue, green and red colour components for each node is determined from the number of disrupted, derived and common genes belongs to its gene ontology category respectively. White node colour was used for representing categories not enriched from target sets. The gene ontology graphs were prepared by Graphviz.

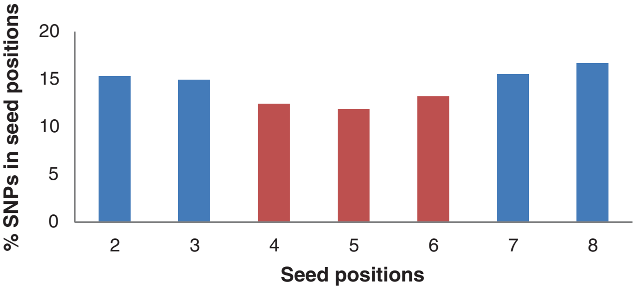

Figure 2. Positional distribution of SNPs in miRNA seeds.

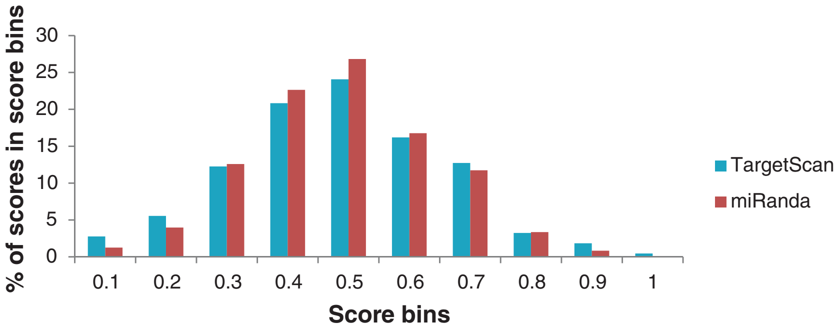

Figure 3. Distribution of GOSS in 'Biological process' for prediction of TargetScan and miRanda. An interval of 0.1 is used to break the GOSS into 10 equal size bins on x-axis. The y-axis shows percentages for number of seed SNPs in each bin.

3. Results

3.1 *Positional distribution of SNPs in miRNA seed region*

We mapped 517 SNPs to the seed regions of 414 miRNAs using the genomic coordinates of SNPs in dbSNP build 138 (Sherry *et al.* 2001) and miRNAs in miRBase release 20 (Kozomara and Griffiths-Jones, 2014). The positional distribution of the 517 SNPs (figure 2) shows that the central positions of miRNA seeds (positions 4, 5 and 6) contain fewer genetic variants than the periphery positions (2, 3, 7 and 8). We performed the Fisher exact *t*-test between the average densities of SNPs in central positions and periphery positions. The difference is statistically significant with a *p*-value of 0.002.

3.2 *Allele-dependent miRNA target prediction*

MicroRNA target recognition is usually dependent on the sequence complementarity between the miRNA seed region and the miRNA binding sites in the target genes. Most miRNA target prediction algorithms also rely on this principle for target recognition. Here we used two methods, TargetScan (Lewis *et al.* 2005) and miRanda (Enright *et al.* 2003), to predict the target genes for the reference and derived allele of each SNP. Due to the limitations of current miRNA target prediction methods, different methods may generate different target sets. Supplementary figure 1 shows the percentages of the average number of reference and derived targets from TargetScan and miRanda predictions.

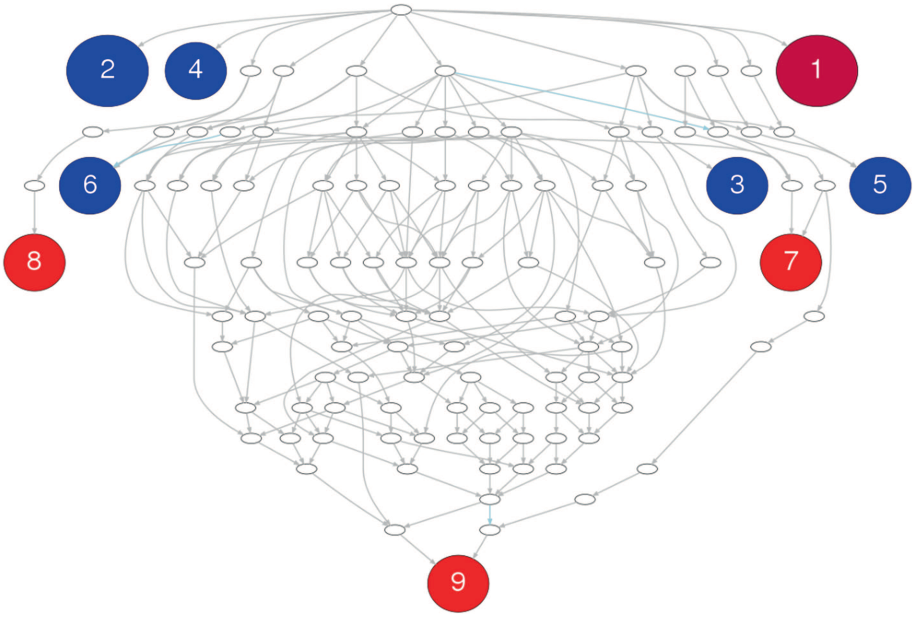

Figure 4. Gene ontology figure from the TargetScan prediction for SNP +57C>T in the fifth seed position of miRNA hsa-miR-184. The gene ontology figure shows the distribution of enriched gene ontology terms for biological processes. Nodes show the enriched gene ontology terms for sets of reference, common and derived targets. Node sizes are proportional to the number of target genes for nodes. Colour components were determined from the number of reference, common and derived target genes respectively.

Table 1. GO term descriptions for figure 4

Node number	GO ID	Name	Reference *p*-value; Number of genes	Common *p*-value; Number of genes	Derived *p*-value; Number of genes
1	GO:0023052	signaling	6e−07; 265	NE	9.08e−14; 905
2	GO:0065007	biological regulation	1.09e−08; 432	NE	NE
3	GO:0007154	cell communication	2.51e−06; 265	NE	NE
4	GO:0051179	localization	2.22e−07; 247	NE	NE
5	GO:0007275	multicellular organismal development	7.16e−11; 246	NE	NE
6	GO:0007165	signal transduction	3.08e−05; 236	NE	NE
7	GO:0051336	regulation of hydrolase activity	NE	NE	6.4e−03; 183
8	GO:0042060	wound healing	NE	NE	4.1e−04; 139
9	GO:0043087	regulation of GTPase activity	NE	NE	6.6e−03; 80

NE: not enriched.

Common targets from the two prediction algorithms were nearly 30% for both reference and derived targets.

We found that a single nucleotide variation in miRNA seeds often lead to big differences in the set of target genes, regardless of which prediction method is used. We define the overlap between the target gene set of the reference allele (R) and that of the derived allele (D) as, $\frac{\#(R \cap D)}{\#(R \cup D)}$, where # is the number of elements in a set. For example, the overlap between the reference target set and the derived target set of SNP +13G>A is 24% (TargetScan) and 37% (miRanda). The average overlap between reference target sets and derived target sets for the 517 SNPs is 23% (TargetScan) and 31% (miRanda).

3.3 *Gene Ontology-Based Similarity Score*

To quantify the functional effects of SNPs in miRNA seeds, we computed Gene Ontology-based Similarity Score (GOSS) for all 517 SNPs in miRNA seeds. Typically, GOSS ranges from 0 to 1, where 0 is for the highest functional effect and 1 is for no effect. We measured the GOSS for Gene Ontologies ‘Biological process’, ‘Molecular

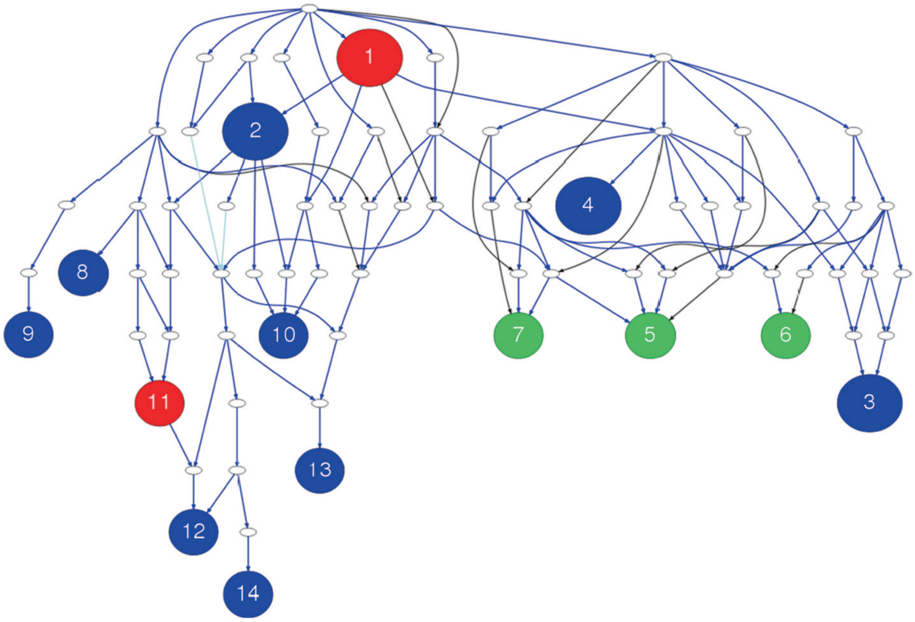

Figure 5. Gene ontology figure from TargetScan prediction for SNP +13 G>A in the fourth seed position of miRNA hsa-miR-96. The gene ontology figure shows the distribution of enriched gene ontology terms for biological processes. Nodes show the enriched gene ontology terms for sets of reference, common and derived targets. Node sizes are proportional to the number of target genes for nodes. Blue, Green and Red node colour components were determined from the number of reference, common and derived target genes respectively.

function' and 'Cellular Component' from TargetScan and miRanda predictions. Figure 3 shows the distributions of GOSS based on 'Biological process'. The average GOSS value for both TargetScan and miRanda was 0.44. From the distributions of GOSS in figure 3, we computed the GOSS percentile scores. A lower GOSS percentile score indicates larger overall differences between the functions of reference and derived target genes. Supplementary table 1 shows the read counts and number of each experiment for miRNAs from miRBase (Kozomara and Griffiths-Jones 2014). We found 44 SNPs in seeds of 40 highly expressed miRNAs. Supplementary Table S1 shows a complete list of GOSS and GOSS percentile scores.

3.4 *Visualization of functional impacts of miRNA seed SNPs using gene ontology graph*

For detailed analyses of individual SNPs or mutations, we can map the enriched functional annotations for reference, derived and common-target gene sets onto the Gene Ontology graph. For example, we considered miRNA SNPs +57C>T in hsa-miR-184 and +13G>A in hsa-miR-96. Both mutations are known to cause disease. The GOSS scores in Supplementary Table S1 for +57C>T in hsa-miR-184 are identical for both TargetScan and miRanda predictions and GOSS percentile scores are lower than 50 percentile.

The SNP +57C>T in the fifth seed position of the hsa-miR-184 seed is already known to cause for EDICT Syndrome (Iliff *et al.* 2012). The distributions of enriched gene ontology terms for +57C>T are presented in figure 4. The node descriptions for Figure 4 are listed in Table 1. For reference targets, the gene ontology biological processes 'multicellular organismal development' and 'signal transduction' were found significantly enriched with *p*-values of 7.16×10^{-11} and 3.08×10^{-5} respectively. We found that, 246 target genes from reference target sets were associated with the gene ontology category 'multicellular organismal development', but for the derived target set the gene ontology category 'multicellular organismal development' was no longer enriched from the target genes. The same is true for the 'signal transduction' function, which was enriched for 236 target genes in the reference target set.

The SNP +13G>A in seed of miRNA 'hsa-miR-96' was specifically linked to the pathogenesis of progressive hearing loss in earlier investigations (Mencia *et al.* 2009). Figure 5 shows the gene ontology graph from the enriched biological process category for the mutation +13G>A. Node descriptions for figure 5 are listed in table 2. We found that the biological process 'neurotrophin TRK receptor signaling pathway' was significantly enriched for reference target genes with a p-value of 5.34×10^{-6}. The gene ontology term 'neurotrophin TRK receptor signaling pathway' was also

Table 2. GO term descriptions for figure 5

Node number	GO ID	Name	Reference *p*-value; Number of genes	Common *p*-value; Number of genes	Derived *p*-value; Number of genes
1	GO:0009987	cellular process		NE	7.25e−15; 1865
2	GO:0044763	single-organism cellular process	1.14e−18; 1814	NE	NE
3	GO:0006464	cellular protein modification process	1.73e−10; 532	NE	NE
4	GO:0006793	phosphorus metabolic process	2.18e−06; 495	NE	NE
5	GO:0019219	regulation of nucleobase containin compound metabolic process	NE	2.35e−04; 207	NE
6	GO:0010468	regulation of gene expression	NE	6.07e−04; 205	NE
7	GO:0031326	regulation of cellular biosynthetic process	NE	7.32e−04; 202	NE
8	GO:1901698	response to nitrogen compound	8.92e−05; 151	NE	NE
9	GO:0042060	wound healing	1.95e−04; 148	NE	NE
10	GO:0030036	actin cytoskeleton organization	4.3e−07; 120	NE	NE
11	GO:0071363	cellular response to growth factor stimulus	NULL	NE	1.38e−04; 118
12	GO:0048011	neurotrophin TRK receptor signaling pathway	5.34e−06; 80	NE	NE
13	GO:0038093	Fc receptor signaling pathway	2.58e−05; 71	NE	NE
14	GO:0007173	epidermal growth factor receptor signaling pathway	9.58e−04; 61	NE	NE

NE: not enriched.

enriched from the miRanda target prediction with a *p*-value of 3.44×10^{-6}. Another gene ontology term 'epidermal growth factor receptor signaling pathway' was also found to be enriched for reference targets with a *p*-value of 9.584×10^{-4} and 2.99×10^{-4} from TargetScan and miRanda respectively.

4. Discussion

For the present study, we performed comprehensive functional analyses of SNPs in miRNA seeds. Here, we report 517 SNPs in miRNA seeds, a number which is significantly more than previously known number of SNPs in miRNA seeds. The large number of SNPs in miRNA seeds allowed us to look at the distribution of SNPs in different seed positions. Interestingly, we found that the SNP densities at the middle of the miRNA seeds were significantly lower than at the two ends of the seeds. Lower SNP density is known to be associated with higher selective pressure. The distribution of SNPs in these miRNA seeds raises the possibility of a greater importance of middle seed base positions 4 to 6 in miRNA function.

Here we have presented a computational pipeline for quantifying the overall functional impacts of miRNA seed SNPs in the form of semantic similarity score GOSS. The GOSS score measures the semantic similarity between enriched gene-ontology categories for reference and derived target sets. The target sets were predicted by both TargetScan and miRanda. TargetScan predicts a target based on the sequence complementarity while miRanda predicts a target based on free binding energy. From the use of TargetScan and miRanda separately we were able to incorporate the most commonly used target prediction criteria's in our results (Peterson *et al.* 2014).

SNP +13G>A in miR-96 is already known to cause progressive hearing loss, but the details of the pathway are yet to be elucidated. In our detailed analyses of the functional enrichment of target gene sets, we found that the biological processes 'Neurotrophin TRK receptor signaling pathway' and 'epidermal growth factor receptor signaling pathway' were significantly enriched for the reference targets of miRNA hsa-miR-96. In contrast, they were not enriched for derived targets of hsa-miR-96. The 'Neurotrophin TRK receptor signaling pathway' is known to control the degeneration of spiral ganglion neurons (SGNs) which are important for progressive hearing loss (Sato *et al.* 2006). Another gene ontology term, 'epidermal growth factor receptor signaling pathway' (Furness *et al.* 2013) is also related to progressive hearing loss. Although we have restricted the scope of our analyses to miRNA seed SNPs only, the computational pipeline we have presented is not limited to SNPs in miRNA seeds. The same could be applied for understanding the functional changes cause by somatic mutations in miRNA seeds. The gene ontology-based detailed analyses of functional changes for somatic mutations in miRNA seed may prove very useful for understanding the disease-causing pathways.

Acknowledgements

The authors gratefully acknowledge the editorial assistance of Dr Martha M. Howe provided through the Writing Assistance Program of the College of Graduate Health Sciences. This work was partly supported by The University of Tennessee Center for Integrative and Translational Genomics.

References

Bao L *et al*. 2007 PolymiRTS database: linking polymorphisms in microRNA target sites with complex traits. *Nucleic Acids Res.* **35** D51–D54

Bartel DP 2004 MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116** 281–297

Bhattacharya A, Ziebarth JD and Cui Y 2013 SomamiR: a database for somatic mutations impacting microRNA function in cancer. *Nucleic Acids Res.* **41** D977–D982

Bhattacharya A, Ziebarth JD and Cui Y 2014 PolymiRTS Database 3.0: linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways. *Nucleic Acids Res.* **42** D86–D91

Enright AJ *et al*. 2003 MicroRNA targets in drosophila. *Genome Biol.* **5** R1

Furness DN *et al*. 2013 Progressive hearing loss and gradual deterioration of sensory hair bundles in the ears of mice lacking the actin-binding protein Eps8L2. *Proc. Natl. Acad. Sci. USA* **110** 13898–13903

Hill CG *et al*. 2014 Functional and evolutionary significance of human MicroRNA seed region mutations. *PLoS One* **9** e115241

Iliff BW, Riazuddin SA and Gottsch JD 2012 A single-base substitution in the seed region of miR-184 causes EDICT syndrome. *Invest. Ophthalmol. Vis. Sci.* **53** 348–353

Jevsinek Skok D *et al*. 2013 Genome-wide in silico screening for microRNA genetic variability in livestock species. *Anim. Genet.* **44** 669–677

Karolchik D *et al*. 2004 The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32** D493–D496

Kozomara A and Griffiths-Jones S 2014 miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42** D68–D73

Lewis BP, Burge CB and Bartel DP 2005 Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are MicroRNA targets. *Cell* **120** 15–20

Mencia A *et al*. 2009 Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. *Nat. Genet.* **41** 609–613

Peterson SM *et al*. 2014 Common features of microRNA target prediction tools. *Front. Genet.* **5** 23

Reimand J, Arak T and Vilo J 2011 g:Profiler–a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Res.* **39** W307–W315

Reimand J *et al*. 2007 g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* **35** W193–W200

Sato T *et al*. 2006 Progressive hearing loss in mice carrying a mutation in the p75 gene. *Brain Res.* **1091** 224–234

Schroder MS *et al*. 2013 RamiGO: an R/Bioconductor package providing an AmiGO visualize interface. *Bioinformatics* **29** 666–668

Sherry ST *et al*. 2001 dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29** 308–311

Siomi H and Siomi MC 2010 Posttranscriptional regulation of microRNA biogenesis in animals. *Mol. Cell.* **38** 323–332

Wang JZ *et al*. 2007 A new method to measure the semantic similarity of GO terms. *Bioinformatics* **23** 1274–1281

Wheeler BM *et al*. 2009 The deep evolution of metazoan microRNAs. *Evol. Dev.* **11** 50–68

Yu G *et al*. 2010 GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26** 976–978

Ziebarth JD et al 2012 PolymiRTS Database 2.0: linking polymorphisms in microRNA target sites with human diseases and complex traits. *Nucleic Acids Res.* **40** D216–D221