PRINT: ISSN 0972-3757 ONLINE: 2456-6360
Int J Hum Genet, 14(1): 17-22 (2014)
DOI: 10.31901/24566330.2014/14.01.03

miRNA Target Gene Identification: Sourcing miRNA Target Gene Relationships for the Analyses of TCGA Illumina MiSeq and RNA-Seq Hiseq Platform Data

James D. Doecke[1,2*], Thierry T. Chekouo[3], Francesco Stingo[3] and Kim-Anh Do[3]

[1]CSIRO Computational Informatics/Australian e-Health Research Centre, Royal Brisbane and Women's Hospital, Brisbane, QLD, Australia
[2]CSIRO Preventative Health Flagship, Parkville, Victoria, Australia
[3]MD Anderson Cancer Center, Houston, Texas, USA

KEYWORDS miRNA. Target Gene. Gene Expression. Cancer. Illumina

ABSTRACT Disruption in homeostatic levels of gene expression can lead directly to disease phenotype. miRNAs have key regulatory roles in modulating gene expression and have been shown to act as oncogenes, with their altered expression disrupting homeostatic biological mechanisms and adding to a disease phenotype. Using the Illumina MiSeq and HiSeq RNA sequencing platform data from the TCGA online resource, the objectives of the current research were to 1 Assess and quantify the existing online resource for miRNA target gene (TG) association, and 2 Define TG lists that can be used for genome-wide miRNA-mRNA-disease association analyses. Using the integration of miRNA lists from the Illumina platform and validated TG online databases, the researchers identified 307 miRNAs mapping to 3,358 validated TG's, with 9,858 miRNA-TG connections. From eight online predicted TG databases, they find 547 miRNAs that map to 18,271 unique TG's, with nearly three and a half million connections. Using the genomic location of miRNA and mRNAs assessed on the Illumina platforms, they identified 434 genes where miRNAs are co-located, and suggest that hypo/hyper methylation of these sites may play a key role in aberrant miRNA expression. In conclusion, using the Illumina miRNA and mRNA sequencing platforms, the researchers have created informative databases for the analyses of the complex interactions between miRNA and their target genes. The researchers' approaches can be applied to similar data sets for any other disease.

INTRODUCTION

miRNAs are small non-coding single stranded molecules that bind TG's to down-regulate gene expression (Iborra et al. 2012). To date, miRBase (Kozomara et al. 2011; Griffiths-Jones 2004; Griffiths-Jones et al. 2006; Griffiths-Jones et al. 2008) has approximately 1600 precursor miRNA sequences that lead to just over 2000 mature miRNA sequences in the human genome. Friedman et al. (2009) postulated that approximately 60% of the known protein coding genes may be regulated by miRNAs. In cancer, certain miRNAs are consistently identified as highly expressed and have influential effects on disease pathways. For example, miR-21 was shown overexpressed in both lung (Zhang et al. 2010) and colorectal cancer cells (Xiong et al. 2013), effectively down regulating the tumor suppressor gene *PTEN*, while the highly expressed let-7 miR was shown to down regulate oncogenes in both breast (Hu et al. 2013) and prostate cancer (Liu et al. 2012).

While there is a growing number of online databases available for TG prediction (TargetScan (Lewis et al. 2003), miRanda (John et al. 2004), miRBase (Griffiths-Jones et al. 2008), PicTar (Lall et al. 2006), PITA (Kertesz et al. 2007), DIANA-microT (Miranda et al. 2006), GeneMir (Huang et al. 2007), miRDB (Wang et al. 2008), mirDIP (Shirdel et al. 2011), only those utilized in this study) there are few validated miRNA TG databases (miRecords (Xiao et al. 2009), mirWalk (Dweep et al. 2011), miRTarBase (Vergoulis et al. 2012)). Many of these online databases store miRNA predicted TG's from multiple different prediction algorithms, and provide thousands of potential targets. A dilemma arises when a researcher would want to access the totality of predicted TG's for a given set of miRNAs. If one were to utilize the complete set of predicted TG's for all known human miRNAs, this is likely to total to the complete set of known coding genes for the human genome.

The aim of this research was to define a list of miRNA-mRNA TG connections that could be used in the analyses of the TCGA genomic resource. Specifically, the researchers define matrices of miRNA-mRNA connections using both predicted, and validated online databases using the Illumina HiSeq RNA-Seq platform. For analyses based around miRNA/RNA expression, they

defined a list of genomic locations for each miRNA that lie within an mRNA transcript region. Although the researchers base these analyses using the TCGA Renal Clear Cell Carcinoma Illumina generated data, they propose these data matrices are applicable for analyses of other cancer types where the same or similar RNA-Seq platforms are used.

METHODOLOGY

Kidney Renal Clear Cell Carcinoma (KIRC) data from 502 tissue samples was obtained from TCGA online resource (The Cancer Genome Atlas Data Portal). Of the available genomic data available for KIRC, the researchers downloaded and utilized the level three processed miRNA-Seq data generated from the miSeq Illumina platform, and mRNA-Seq data generated by the Illumina HiSeq2000 platform (Bennett 2004). For the purposes of brevity, the researchers create the acronyms *miSeq* and *HiSeq* to reference the miSeq and HiSeq Illumina platforms. Detailed information regarding data preprocessing can be found online (RNASeq Data Format). A total of 805 miRNA markers were available for analyses from the miRNA isoform platform, while 22,873 mRNA transcripts were read from the mRNA level 3 data.

To create a list of miRNA markers where the genomic location of the miRNA mapped to the genomic location of the mRNA transcripts available from the HiSeq platform, information regarding chromosome, strand, start and end nucleotide position were assessed via multivariate matching, and validated using the GenomicRanges R package. Since it is possible that miRNAs may also be created from both 5' and 3' untranslated regions, they extended the search space by 500 nucleotides both up and downstream of the transcript start and end site.

In total, eight online predicted and three online validated miRNA TG databases were assessed (Table 1). A process chart that details the matching of miRNAs to predicted and validated TG's is shown in Figure 1. The researchers created an miRNA-TG list for each of the predicted and validated online databases. Only those mRNAs that were assessed within the HiSeq platform were used for miRNA-TG mapping. The validated TG (those genes that have been experimentally verified as having interactions with miRNAs) list was created using the intersection between the totality of known TG's available for those miRNAs listed in all three validated online databases, and those assessed by the miSeq platform. Using the predicted online TG data-

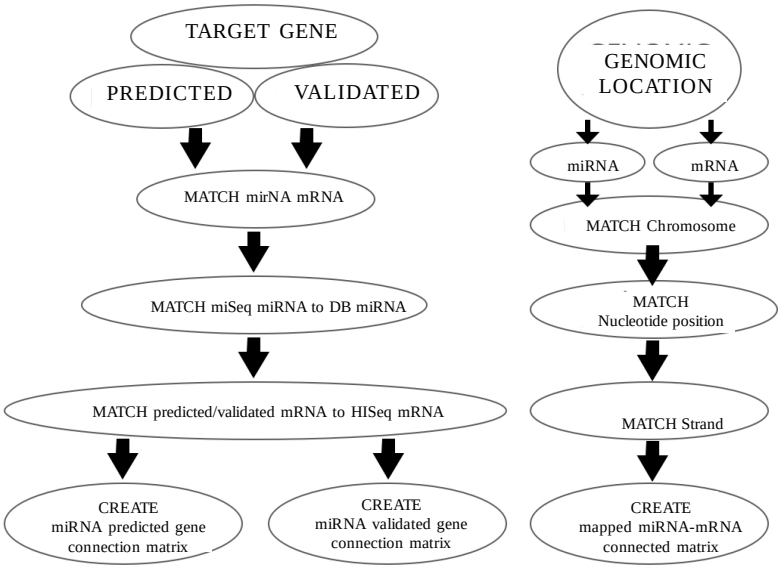

Fig. 1. miRNA-mRNA: Fine mapping construction of miRNA to both predicted and validated TG, and miRNA genomic location to mRNA transcript location for those markers analysed via the miSeq and HiSeq platforms

bases, we selected only the top ten miRNA-TG's per miRNA based upon the predicted binding score. Binding scores were assessed from each database as they related to the algorithm and scoring rules utilized.

RESULTS

Using the standard mRNA transcript nucleotide start and end boundaries provided with Illumina HiSeq platform the researchers identified 424 unique miRNAs that lie within the mRNA transcript boundaries of 383 unique genes with a total of 452 miRNA-mRNA location connections. Of these 424 miRNAs, there were 409 miRNAs that were located within single genes, and 15 miRNAs that were located within two genes, however in all cases the nucleotide positions for each of these genes overlapped.

For those miRNA-mRNA transcript mappings where the miRNA was located within only one gene (409), we identified that 32 genes had more than one miRNA located within gene boundaries. Specifically, of the 409 location mappings, 323 genes contained only one miRNA, 22 genes contained two miRNAs, six genes contained three miRNAs, one gene contained four miRNAs, two genes contained six miRNAs, and lastly one gene (CLCN5) contained eight miRNAs. Using the extended nucleotide search space (500bp both upstream and downstream), we captured an extra ten location matches to a total of 434 miRNAs and a total of 466 miRNA-mRNA location connections.

The number of miRNAs from each online predicted TG database that matched the 805 miRNAs from the miSeq platform KIRC data ranged from 65 (GeneMir) through to 537 (TargetScan). Not taking into consideration the mRNA transcripts available from the HiSeq platform, the PITA database had by far the largest number of predicted gene targets (~4.1 million), while GeneMir had the lowest (6,387). For the purpose of dimension reduction, the researchers reduced (by algorithm score) the number of predicted TG's per miRNA to the top ten predicted genes per data base. After taking the intersection of unique genes per miRNA per database, they created a matrix of 547 miRNAs, 7,371 unique genes and 20,630 miRNA-mRNA connections. Thus 258 miRNAs assessed with the miSeq platform had no predicted TG's from the eight online databases. Complete numbers for miRNAs and TG's identified irrespective of Illumina platform and dimension reduction are shown in Table 2.

Using the three validated TG databases, the researchers identified variable numbers of target genes per miRNA for each database. miRecords had a median number of validated TG's per miRNA of 2, miRTarBase had a median of 4.5, and mirWalk had a median of 8 validated TG's. In total the researchers identified 307 miRNAs within the three online validated TG databases, with 3,358 unique validated TG's, and 9,858 connections. The median number of miRNA-mRNA connections across all three databases was 8.

Assessing the totality of predicted TG's for all possible miRNAs from the eight online databases, the researchers identified 1,615 unique miRNAs that predicted to bind to 23,783 unique genes, with a total of 9,513,366 connections. Reducing this to only those from the Illumina HiSeq and miSeq platforms, the researchers found 547 miRNAs, 18,271 genes and 3,444,143 connections. Dimension reduction by score algorithm reduced the number of genes to 7,371

Table 1: miRNA target gene databases

Data base predicted	*Web resource*
PICTAR 2012	http://dorina.mdc-berlin.de/rbp_browser/hg19.html
mirDIP v1.1.2	http://ophid.utoronto.ca/mirDIP/search.jsp
PITA v4	http://genie.weizmann.ac.il/pubs/mir07/mir07_data.html
DIANA micro T v3.0	http://diana.cslab.ece.ntua.gr/microT/
miRANDA 2010	http://www.microrna.org/microrna/getDownloads.do
GeneMir	http://www.psi.toronto.edu/genmir/
miRDB v4.0	http://mirdb.org/miRDB/download.html
Target Scan v6.2	http://www.targetscan.org/cgi-bin/targetscan/data_download.cgi?db=vert_61
Validated	
miRecords	http://mirecords.umn.edu/miRecords/download.php
mirWalk[>>]	http://www.umm.uni-heidelberg.de/apps/zmf/mirwalk/mirnatargetpub.php
miRTarBase	http://mirtarbase.mbc.nctu.edu.tw/php/browse.php

Source: Web resource for online databases

Table 2: miRNA-mRNA transcript frequencies

	Total				*Illumina only*		*Illumina only: Reduced*	
Data base predicted	*N miRNA*	*N genes*	*N connections*[t]	*N miRNA*[*]	*N genes*[^]	*N connections*	*N genes*	*N connections*
PICTAR 2012	1,144	9,248	238,898	183	9,043	49,196	912	1830
mirDIP v1.1.2	81	2,205	10,153	81	2,104	9,739	508	810
PITA v4	677	16,942	4,095,751	379	13,853	1,842,181	2,827	3790
DIANA micro T v3.0	555 [+]	17,229 [+]	1,878,979	301	16,432	751,330	1,400	3010
miRANDA 2010	249	19,281	1,097,064	172	16,738	470,040	1,543	1720
GeneMir	114	890	6,387	65	702	3,328	247	650
miRDB v4.0	1,376	16,415 [&]	1,172,337	345	13,767	115,605	2,379	3450
Target Scan v6.2	1,537	15,031	1,013,797	537	12,030	202,724	1,938	5370
Total unique validated	1,615	23,783	9,513,366	547	18,271	3,444,143	7,371	20,630
miRecords	273	1,049	5,236	135	384	528		
mirWalk[>>]	266	3,050	9,829	266	2,922	6,983		
miRTarBase	323	1,959	3,597	223	1,433	2,347		
Total unique	498	3,945	18,662	307	3,358	9,858		

[*] only those miRNAs are included that match the Illumina MiSeq2000 miRNA-Seq platform
[^] maximum number of unique TG's are included that match the Illumina HiSeq2000 RNA-Seq platform
[t] complete number of connections identified per miRNA database
[+] only those available miRNA and genes that map to the ENSEMBL identifiers through the R biomaRt package
[&] only those available miRNA and genes that map to the NCBI reference identifiers through the R biomaRt package
[>>] number of miRNAs with validated TG's derived from the website match to the Illumina list of 805 miRNA ids loaded

with 20,630 possible connections for testing, a more manageable number for future statistical analyses.

DISCUSSION

The current research was aimed towards defining target sets of genes for analyses with the miRNA markers available from miSeq Illumina platform. The researchers defined lists of validated and predicted TG's that align with a biologically directed approach to the co-analyses of miRNA and mRNA expression, and a list of genomic locations for each miRNA marker. Using the 805 miRNAs assessed by the Illumina miSeq platform, and three online validated TG databases, they identified 307 miRNAs matching 3,358 genes with 9,858 connections. Investigating the eight online predicted TG databases, they identified 547 with predicted TG's, with a total of 23,783 connections. Assessing the number of miRNAs that reside within gene boundaries, the researchers find 434 miRNA (54%) within the genes assessed within the Illumina HiSeq platform. The researchers provide upon request three databases, 1 miRNA gene co-location, 2 miRNA predicted TG connections, and 3 miRNA validated TG connections. A database with an increased number predicted TG's per miRNA is also available.

The list of miRNA-validated TG's will be of interest to many different diseases and may serve as an excellent starting point for those wishing to compare a list of previously identified candidate miRNA-mRNA associations. Utilizing the secondary list of miRNA-predicted TG associations will ultimately lead to an increase in the number of validated miRNA-mRNA associations. Within this research, the researchers minimized the list of predicted TG's per miRNA by using only the top ten after ranking the TG's by score.

Compared to the total number of miRNAs that have mature miRNA sequence available from miRBase 2042 (miRBase), the number of miRNAs with validated TG's amounts to quite a small number, ultimately leaving a large proportion of TG validation research for consideration. The Illumina platform assessed 805 mature miRNAs, of which the researchers' identified 547 that have predicted TG's. They acknowledge that the remaining 258 miRNAs do have gene targets, however these data are yet to be published online. The complete number of predicted TG's from the eight different online predicted TG databases 23,783 was close to the hypothesized

number of known protein coding genes 20-25,000 (Stein 2004). Using this list for miRNA-mRNA expression analyses would be no different to analyzing the complete genome of protein coding genes with each miRNA.

An important aspect of this research was the identification of miRNAs that lie within known protein coding gene boundaries. According to Lutter et al. (2010), approximately 37% of known miRNAs are thought to reside within intronic regions of protein coding genes. The current research identified 424 miRNAs out of a possible 805 ~53% that are located within the strict gene boundaries provided by Illumina. This was increased to 434 miRNAs ~54% including 500bp of 5' and 3' UTR regions. It is widely known that many cancerous tissues have gene expression dis-regulated by either hypo- or hyper-methylated CPG sites in and around the gene. For example, hyper-methylation of the MGMT O6-Methyl Guanine Methyltransferase promoter was associated with inactivation of the MGMT gene (Farzanehfar et al. 2013), while hypo-methylation of the long interspersed nuclear element-1 LINE-1 was shown to activate normally methylation-silenced proto-oncogenes in colorectal cancer (Hur et al. 2013). A combination of DNA methylation and aberrant 3' miRNA regulation has also recently been associated with an overall reduction in MGMT expression, showing that together aberrant methylation and miRNA expression can adversely affect disease status (Kreth et al. 2013). Thus the current research has identified a list of genes where the aberrant miRNA expression may be mapped to specific areas of DNA methylation across gene location.

CONCLUSION

The miRNA-mRNA TG and genomic location lists identified in this research will be extremely valuable for all researchers who aim to assess the biological relationship between miRNA expression, gene expression and disease. The researchers expect that these lists will need to be updated as the Illumina miSeq and HiSeq platforms are updated with greater numbers of targets. They advocate that these lists should be used across multiple different cancers and disease types, with the ultimate aim of increasing the online resource of validated miRNA-TG and disease associated relationships.

RECOMMENDATIONS

The researchers recommend the data lists provided with this paper be utilized for further exploration of Illumina sequencing platform data. The combination of miRNA and mRNA expression data may elucidate target gene associations integral to cancer progression pathways and future treatment targets.

ACKNOWLEDGEMENTS

The researchers would like to acknowledge all the research participants contributing to the TCGA resource for providing high quality data for analyses. K-A Do was partially supported by the M. D. Anderson Cancer Center Prostate SPORE (P50 CA140388 02), both K-A Do and F. Stingo were partially supported by the Cancer Center Support Grant (CCSG) (P30 CA016672), Thierry Chekouo was fully funded by MD Anderson internal funds.

NOTES

Manuscript Abbreviations:

miRNA: microRNA, TG: target gene, TCGA: The Cancer Genome Atlas, miSeq: Illumina mRNA-Seq platform, HiSeq: Illumina HiSeq2000 platform

Databases provided by the MD Anderson Cancer Center Biostatistics Department: http://www. mdanderson. org/education-and-research/departments-programs-and-labs/departments-and-divisions/biostatistics/index.html

REFERENCES

Bennett S 2004. Solexa Ltd. *Pharmacogentics,* 55: 433-438.

Dweep H, Sticht C, Pandey P, Gretz N 2011. miRWalk database: Prediction of possible miRNA binding sites by "walking" the genes of three genomes. *J Biomed Inform,* 445: 839-847.

Farzanehfar M, Vossoughinia H, Jabini R, Tavassoli A, Saadatnia H, Khorashad AK, Ahadi M, Afzalaghaee M, Ghayoor Karimiani E, Mirzaei F, Ayatollahi H 2013. Evaluation of methylation of MGMT O6-Methyl Guanine Methyltransferase gene promoter in sporadic colorectal cancer. *DNA Cell Biol,* 327: 371-377.

Friedman RC, Farh KK, Burge CB, Bartel DP 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*, 19(1): 92–105.

Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ 2006. miRBase: Microrna sequences, targets and gene nomenclature. *Nucleic Acids Res,* 34 Database Issue: D140-D144.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ 2008. miRBase: Tools for microRNA genomics. *Nucleic Acids Res,* 36 Suppl. 1: D154-158.

Griffiths-Jones S 2004. The microRNA registry. *Nucleic Acids Res,* 32 Database Issue: D109-D111

Hu X, Guo J, Zheng L, Li C, Zheng TM, Tanyi JL, Liang S, Benedetto C, Mitidieri M, Katsaros D, Zhao X, Zhang Y, Huang Q, Zhang L 2013. The heterochronic microRNA let-7 inhibits cell motility by regulating the genes in the actin cytoskeleton pathway in breast cancer. *Mol Cancer Res,* 113: 240-250.

Huang JC, Babak T, Corson TW, Chua G, Khan S, Gallie BL, Hughes TR, Blencowe BJ, Frey BJ, Morris QD 2007. Using expression profiling data to identify human microRNA targets. *Nat Methods,* 412: 1045-1049.

Hur K, Cejas P, Feliu J, Moreno-Rubio J, Burgos E, Boland CR, Goel A 2013. Hypomethylation of long interspersed nuclear element-1 LINE-1 leads to activation of proto-oncogenes in human colorectal cancer metastasis. *Gut,* Epub ahead of print.

Iborra M, Bernuzzi F, Invernizzi P, Danese S 2012. MicroRNAs in autoimmunity and inflammatory bowel disease: Crucial regulators in immune response. *Autoimmun Rev,* 115: 305-314.

John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS 2004. Human microRNA targets. *PLoS Biol,* 211: e363.

Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E 2007. The role of site accessibility in microRNA target recognition. *Nat Genet,* 3910: 1278-1284.

Kozomara A, Griffiths-Jones S 2011. miRBase: Integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res,* 39 Database Issue: D152-D157.

Kreth S, Limbeck E, Hinske LC, Schütz SV, Thon N, Hoefig K, Egensperger R, Kreth FW 2013. In human glioblastomas transcript elongation by alternative polyadenylation and miRNA targeting is a potent mechanism of MGMT silencing. *Acta Neuropathol,* 1255: 671-681.

Lall S, Grün D, Krek A, Chen K, Wang YL, Dewey CN, Sood P, Colombo T, Bray N, Macmenamin P, Kao HL, Gunsalus KC, Pachter L, Piano F, Rajewsky N 2006. A genome-wide map of conserved microRNA targets in C. elegans. *Curr Biol,* 165: 460-471.

Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB 2003. Prediction of mammalian microRNA targets. *Cell,* 1157: 787-798.

Liu C, Kelnar K, Vlassov AV, Brown D, Wang J, Tang DG 2012. Distinct microRNA expression profiles in prostate cancer stem/progenitor cells and tumor-suppressive functions of let-7. *Cancer Res,* 72(13): 3393-3404

Lutter D, Marr C, Krumsiek J, Lang EW, Theis FJ 2010. Intronic microRNAs support their host genes by mediating synergistic and antagonistic regulatory effects. *BMC Genomics,* 11: 224.

Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, Thomson AM, Lim B, Rigoutsos I 2006. A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes. *Cell,* 1266: 1203-1217.

miRBase, From <http: //www.mirbase.org> (Retrieved on October 2013).

RNASeq Data Format, National Cancer Institute, NCI Wiki 2013. From <https: //wiki.nci.nih.gov/display/TCGA/RNASeq+Data+Format+Specification> (Retrieved on October 2013).

Shirdel EA, Xie W, Mak TW, Jurisica I 2011. NAViGaTing the micronome: Using multiple microrna prediction databases to identify signalling Pathway-associated microRNAs. *PLoS ONE,* 62: e17429.

Stein L D 2004. Human genome: End of the beginning. *Nature,* 431: 915-916.

The Cancer Genome Atlas Data Portal. National Human Genome Research Institute. From <tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp> (Retrieved on October 2013).

Vergoulis T, Vlachos IS, Alexiou P, Georgakilas G, Maragkakis M, Reczko M, Gerangelos S, Koziris N, Dalamagas T, Hatzigeorgiou AG 2012. TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res,* 40 Database issue: D222-229.

Wang X 2008. miRDB: A microRNA target prediction and functional annotation database with a wiki interface. *RNA,* 146: 1012-1017.

Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T 2009. miRecords: An integrated resource for microRNA-target interactions. *Nucleic Acids Res,* 37 Suppl. 1: D105-110.

Xiong B, Cheng Y, Ma L, Zhang C 2013. MiR-21 regulates biological behavior through the PTEN/PI-3 K/Akt signaling pathway in human colorectal cancer cells. *Int J Oncol,* 421: 219-228.

Zhang JG, Wang JJ, Zhao F, Liu Q, Jiang K, Yang GH 2010. MicroRNA-21 miR-21 represses tumor suppressor PTEN and promotes growth and invasion in non-small cell lung cancer NSCLC. *Clin Chim Acta,* 411(11-12): 846-852.