Data in Brief

Profiling miRNAs in nasopharyngeal carcinoma FFPE tissue by microarray and Next Generation Sequencing

Jin Peng a,b, Yanjun Feng a,b, Gabriel Rinaldi a,b, Paul Levine c, Samantha Easley d, Elizabeth Martinez d, Salman Hashmi e, Nader Sadeghi e, Paul J. Brindley a,b, Jason P. Mulvenna,g, Jeffrey M. Bethony a,b, Jordan L. Plieskatt a,b,⁎

a Department of Microbiology, Immunology and Tropical Medicine, School of Medicine and Health Science, George Washington University, Washington DC, USA
b Research Center for Neglected Diseases of Poverty, School of Medicine and Health Science, George Washington University, Washington DC, USA
c Department of Epidemiology and Biostatistics, The George Washington University School of Public Health and Health Services, Washington DC 20037, USA
d Department of Pathology, School of Medicine and Health Science, George Washington University, Washington DC, USA
e Medical Faculty Associates, The George Washington University, Washington DC 20037, USA
f School of Biomedical Sciences, Faculty of Medicine and Biomedical Sciences, University of Queensland, Brisbane, Australia
g Infectious Diseases Program, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia

ABSTRACT

Nasopharyngeal carcinoma (NPC) is a non-lymphomatous, squamous-cell carcinoma that occurs in the epithelial lining of the nasopharynx. Nasopharyngeal carcinoma has a geographically well-defined distribution worldwide, with the highest prevalence in China, Southeast Asia, and Northern Africa. Symptoms of nascent NPC may be unapparent or trivial, with diagnosis based on the histopathology of biopsied tissue following endoscopy of the nasopharynx. The tumor node metastasis (TNM) staging system is the benchmark for the prognosis of NPC and guides treatment strategy. However, there is a consensus that the TNM system is not sufficiently specific for the prognosis of NPC, as it does not reflect the biological heterogeneity of this tumor, making another biomarker for the detection of NPC a priority. We have previously reported on different approaches for microRNA (miRNA) biomarker discovery for Formalin Fixed Paraffin Embedded (FFPE) NPC tissue samples by both a targeted (microarray) and an untargeted (small RNA-Seq) discovery platform. Both miRNA discovery platforms produced similar results, narrowing the miRNA signature to 1–5% of the known mature human miRNAs, with untargeted (small RNA-Seq approach) having the advantage of indicating "unknown" miRNAs associated with NPC. Both miRNA profiles strongly associated with NPC, providing two potential discovery platforms for biomarker signatures for NPC. Herein, we provide a detailed description of the methods that we used to interrogate FFPE samples to discover biomarkers for NPC.

© 2014 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/3.0/).

Direct link to deposited data

Microarray deposited data can be found here: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE46172

RNA-Seq deposited data can be found here: http://www.ncbi.nlm.nih.gov/sra/?term=SRP029599

Experimental design, materials and methods

Experimental cases

Case and control tissue including sample characteristics are presented in Table 1 (and in detail in [1]). In brief, four formalin fixed paraffin-embedded (FFPE) tissues from cases of histologically confirmed non-keratinizing NPC and four FFPE cases of normal nasopharyngeal tissue...
were obtained from the biological repository in the Department of Pathology of The George Washington University Hospital, Washington, DC. Tissue sections from FFPE were reviewed by two independent pathologists (E.M. and S.E.) to confirm the diagnosis as shown in [1]. FFPE preparation, hematoxylin and eosin (H&E) staining, and representativeness images have also been previously reported [1]. It should also be noted that the SRA project submission contains four additional samples (Accession: SRX345915, SRX345913, SRX345913 and SRX345909). These samples reference a survey of serum pools from NPC positive and control individuals discussed in [1] but not further referenced herein.

RNA isolation

Total RNA was isolated from 2 × 10 μm sections from each FFPE case using the miRNeasy FFPE kit (Qiagen) [1]. RNA concentration, purity, and integrity (RIN) were determined by spectrophotometry (Nanodrop 1000) and the Agilent 2100 Bioanalyzer using the Agilent RNA 6000 Nano and small RNA kits. Purified RNA was stored at −80 °C.

Yields of total RNA derived from FFPE were approximately 100 ng/μm with 260/280 and 260/230 ratios of ~2.0 and ~1.9, respectively. Analysis on the Agilent Bioanalyzer indicated that the samples were enriched for small RNA species with integrity (RNA Integrity Number or RIN) values of two to three. Though typically indicative of RNA degradation, the robustness of miRNAs in these FFPE tissue [2] and reports from other groups [3] that RIN values have negligible effect on miRNA results enabled us to consider this purified RNA suitable for further analysis by microarray and RNA sequencing.

Microarray, data normalization and analysis

All eight samples underwent analysis via microarray (Table 1). Total RNA isolated from each FFPE case was labeled and hybridized to an Agilent human miRNA microarray (miRBase Release 16.0) and scanned by microarray and RNA sequencing.

Table 1

| Sample ID | Type       | Sex | Age | Histological type | Clinical Stage (TNM) | Discovery Techniques | Accession Number      | File                           |
|-----------|------------|-----|-----|-------------------|----------------------|----------------------|------------------------|--------------------------------|
| 08-S-6658A | Tumor      | M   | 46  | Differentiated    | T4N2M0               | Microarray           | GSM11252481            | 1.mirNA_107_Sep09_1.txt.gz   |
| 04-S-341B  | Tumor      | F   | 80  | Differentiated    | T3N0M0               | Microarray           | GSM11252482            | 1.mirNA_107_Sep09_1.txt.gz   |
| 04-S-6103A | Tumor      | M   | 47  | Undifferentiated  | T4N2M0               | Microarray           | GSM11252483            | 1.mirNA_107_Sep09_1.txt.gz   |
| 12-S-432A  | Tumor      | M   | 49  | Undifferentiated  | T3N1M0               | Microarray           | GSM11252484            | 1.mirNA_107_Sep09_1.txt.gz   |
| 08-S-6658C | Control:   | M   | 46  | Differentiated    | T4N2M0               | Microarray           | GSM11252485            | 1.mirNA_107_Sep09_2.txt.gz   |
| 04-S-341E  | Control:   | F   | 80  | Differentiated    | T3N0M0               | Microarray           | GSM11252486            | 1.mirNA_107_Sep09_2.txt.gz   |
| 05-S-5264-F | Control:  | F   | 43  | Undifferentiated  | T4N0M0               | Microarray           | GSM11252487            | 1.mirNA_107_Sep09_2.txt.gz   |
| 11-S-11311-E | Control:  | M   | 65  | Undifferentiated  | T1N0M0               | Microarray           | GSM11252488            | 1.mirNA_107_Sep09_2.txt.gz   |

Notes: a) Denotes those from the same patient (i.e. paired NPC/Control tissue samples).
4) Quality control: The correlation coefficient value of all samples was >0.7 and therefore all the samples were used in further analysis. Further, 3D Principle Components Analysis (PCA) scores and plotting were used to determine any association among the samples (Fig. 1). It was noted that paired samples did not exhibit more significant clustering than non-paired (NPC/Control tissue) in the analysis of filtering clustering than non-paired (NPC/Control tissue) in the analysis (Fig. 1 and Hierarchical clustering [1]). In Filter by Expression, the right entity and interpretation were selected and filtered by raw data value. The lower cut-off value of the interest range was set to 20 and at least two out of eight samples had values within this range.

5) In Analysis, the condition was set as tumor versus control, tested by t-test unpaired, and an asymptotic p-value was computed without correction. The fold change cut-off was >2.0 and analyzed under conditions with tumor compared to control. Hierarchical clustering analysis of differentially expressed genes from samples was performed on both entities and the conditions by normalized intensity values using Euclidean distance metric and Centroid linkage rule.

Small RNA sequencing

Small RNA sequencing was performed on five of the same samples used in microarray analysis (Table 1) of the three samples used in the previous analysis (control samples 341E and 11311E and tumor sample 341B) omitted due to the exhaustion of total RNA purified from the small tissue areas available for the study. Total RNA derived from the FFPE was subjected to Ribo-Zero Pretreatment using Ribo-Zero rRNA Removal Kit (Epicentre) as described by the manufacturer and in [1]. Library preparation and sequencing have been described in further details in [1]. Briefly, Illumina libraries were constructed from 1 μg of total RNA using the TruSeq Small RNA Sample Kit (Illumina). Libraries were subjected to quality control prior to sequencing using an Agilent 2100 BioAnalyzer and concentration determination using PicoGreen (Invitrogen). The Illumina Genome Analyzer IIX was used to perform the sequencing by Expression Analysis, A Quintiles Company (Durham, NC).

![Fig. 1. 3D principal components analysis (PCA).](image)
Sequencing processing: alignment, mapping and annotation

Initial processing was performed using both FastqMcf and FastQC both of which can be accessed at http://code.google.com/p/ea-utils/wiki/FastqMcf and http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

After adaptor removal and quality filtering, ~28 million reads were aligned to the human (UCSC hg19) and Human herpes virus 4 (Epstein–Barr virus or EBV) genome (NCBI NC_007605.1) and miRNA counts generated for each sample [1]. Both miRDeep 2.0.0.5 [5] and miRExpress 2.0 [6] were used to generate counts, and each provided comparable results, with over 50% of the reads mapping to miRNAs in either the human or EBV genomes (Table 3). Identification of known miRNAs was based on miRBase Release 19 [7], with an alignment identity of 1%, a tolerance range of 4, and a similarity threshold of 0.8 [1].

In total, using miRDeep and miRExpress, 984 and 847 human and EBV miRNAs were identified, respectively, with a count per million greater than one in at least two of the samples.

Using EdgeR [8], a binomial distribution was used to compare the independent analyses from miRDeep and miRExpress [1]. The biological coefficient of variation (BCV) was used to estimate the variability across the dataset and plotted via the plotBCV function (Fig. 2A), with a common dispersion of 67% indicating a relatively high dispersion of gene expression levels. Given that this was an observational study on independent NPC cases using NPC tumors of different histological grades, such a value would not be considered atypical. Using the function plotsmear in EdgeR, log-fold changes were plotted against log-cpm (Fig. 2B). Using EdgeR, 99 dysregulated miRNAs were identified in NPC tumor tissue versus control tissue samples.

Table 3

| Sample ID | Type    | Total reads | miRNA mapped | Unmapped reads | %mapped |
|-----------|---------|-------------|--------------|----------------|---------|
| 08-S-6658A* | Tumor   | 5,609,415.00 | 2,569,276.00 | 3,040,139.00 | 46%     |
| 04-S-6103A | Tumor   | 4,338,399.00 | 2,258,744.00 | 2,079,655.00 | 52%     |
| 12-S-432A  | Tumor   | 5,354,631.00 | 3,236,216.00 | 2,118,415.00 | 60%     |
| 08-S-6658C* | Control | 6,265,846.00 | 3,755,819.00 | 2,510,027.00 | 60%     |
| 05-S-5264-F| Control | 6,533,346.00 | 4,262,886.00 | 2,270,460.00 | 65%     |
| Total      |         | 28,101,637.00 | 16,082,941.00 | 12,018,696.00 | 57%     |

* Denotes from the same patient (i.e. paired NPC/Control tissue samples).

Fig. 2. RNA-sequencing output of five NPC FFPE samples. (A) Biological coefficient of variation reported against average log CPM. Red common trend line indicates the BCV of 67%. (B) logFC reported versus average log CPM. Log fold change of two is indicated by bracketed blue lines. Red dots indicate human miRNAs identified as significant (p value < 0.05).

Fig. 3. Top common human miRNAs illustrated [11] as detected in corresponding independent analyses from both microarray and RNA-Seq. A total of eight common miRNAs were highlighted across both methods under the statistical cut-offs previously described [1].
Comparison of datasets

Both targeted (microarray) and untargeted (small RNA-Seq) approaches were extensively compared in our previously published manuscript [1]. While only eight dysregulated human miRNAs were identified in both the microarray and RNA-Seq analysis (Fig. 3) as statistically significant, the overall datasets were comparable. All but three miRNAs identified by microarray as significantly dysregulated were also identified by RNA-Seq, albeit not as significant [1]. In addition, these miRNAs also showed a similar dysregulation: i.e. if identified as up-regulated by microarray, they were also identified as up-regulated by small RNA-Seq [1].

Acknowledgments

The Agilent microarrays were hybridized and scanned at the Genomics and Epigenomics Shared Resource, Georgetown University Medical Center with the assistance of David Goerlitz.

This research was partially supported by awards R01CA155297 (JMB, JPM, and PJB) from the National Cancer Institute, P50AI098639 (BS, JMB, and PJB) from the National Institute of Allergy and Infectious Disease, fellowship support (JPM) and research support (JMB and JPM) under GNT1051627 from the National Health and Medical Research Council of Australia, and research support from the Dr. Cyrus And Myrtle Katzen Cancer Research Center at the George Washington University (PJB and JMB) (RG00982). The contents are solely the responsibility of the authors and do not represent the official views of NIAID, NCI, the Katzen Cancer Research Center of the George Washington University, or the NHMRC of Australia.

References

[1] J.L. Plieskatt, G. Rinaldi, Y. Feng, P.J. Levine, S. Easley, E. Martinez, S. Hashmi, N. Sadeghi, P.J. Brindley, J.M. Bethony, J.P. Mulvenna. Methods and matrices: approaches to identifying miRNAs for Nasopharyngeal carcinoma. J. Transl. Med. 12 (2014) 3.
[2] Y. Xi, G. Nakajima, E. Gavin, C.G. Morris, K. Kudo, K. Hayashi, J. Ju, Systematic analysis of microRNA expression of RNA extracted from fresh frozen and formalin-fixed paraffin-embedded samples. RNA 13 (2007) 1668–1674.
[3] M. Jung, A. Schaefer, I. Steiner, C. Kempkenstefen, C. Stephan, A. Erbersdobler, K. Jung, Robust microRNA stability in degraded RNA preparations from human tissue and cell samples. Clin. Chem. 56 (2010) 998–1006.
[4] Agilent GeneSpring user manual, edition: May 2013 http://genespring-support.com/files/gs_12_6/GeneSpring-manual.pdf.
[5] M.R. Friedlander, W. Chen, C. Adamidi, J. Maaskola, R. Einspanier, S. Knespel, N. Rajewsky, Discovering microRNAs from deep sequencing data using mirDeep. Nat. Biotechnol. 26 (2008) 407–415.
[6] W.C. Wang, F.M. Lin, W.C. Chang, K.Y. Lin, H.D. Huang, Lin NS: miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. BMC Bioinforma. 10 (2009) 328.
[7] A. Kozomara, S. Griffiths-Jones, miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 39 (2011) D152–D157.
[8] M.D. Robinson, G.K. Smyth, Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 23 (2007) 2881–2887.
[9] S. Edge, D. Byrd, C. Compton, A. Fritz, F. Greene, A. Trotti, AJCC Cancer Staging Manual. 7th edition Springer, New York, NY, 2010.
[10] J. Chan, B. Pilch, T. Kuo, A. Lee, Nasopharyngeal Carcinoma. IARC Press, Lyon, France, 2005.
[11] J.C. Oliveros, VENNY. An interactive tool for comparing lists with Venn Diagrams. http://bioinfogp.cnb.csic.es/tools/venny/index.html2007.