CellPress
REVIEWS

Forum

Big Strides in Cellular MicroRNA Expression

Marc K. Halushka,[1,*,@]
Bastian Fromm,[2]
Kevin J. Peterson,[3] and
Matthew N. McCall[4]

A lack of knowledge of the cellular origin of miRNAs has greatly confounded functional and biomarkers studies. Recently, three studies characterized miRNA expression patterns across >78 human cell types. These combined data expand our knowledge of miRNA expression localization and confirm that many miRNAs show cell type-specific expression patterns.

Since their discovery nearly 25 years ago, much progress has been made on understanding the roles miRNAs play in gene regulatory networks. One of the primary insights into their functionality has been the global characterization of miRNA expression at the tissue/organ level. The general observation has been that some miRNAs have conserved organ-specific expression with other miRNAs having ubiquitous expression. This has led to the idea that many miRNAs have broad expression profiles across cell types, especially in humans. However, tissues and organs are a collection of cell types. Some cell types, such as endothelial cells, red blood cells, and fibroblasts, are found in essentially all organs and the miRNAs they may exclusively express may be misinterpreted as ubiquitous [1–3]. Thus, a need for cell-specific data to clarify the localization of miRNAs in a cell-by-cell fashion is long overdue.

Recently, three groups, the fifth edition of the Functional Annotation of Mammalian Genome (FANTOM5) consortium [4], the Hemmrich-Stanisak laboratory [5], and two of us [6], published complementary papers that provide the first deep look at primary cell miRNA expression using small RNA-seq data. The FANTOM5 consortium performed small RNA-seq on 396 human samples across 118 cell types obtained from different organs [4]. The Hemmrich-Stanisak laboratory [5] performed magnetic-activated cell sorting (MACS) separation of 43 buffy coats in addition to serum and whole-blood sample collection. This yielded multiple libraries of seven hematopoietic cell types [5]. Our group grew 34 primary cell types in culture, flow sorted two cell types, then scoured the Sequence Read Archive for additional primary and cancer cell data, obtaining an additional 226 samples (126 primary cell and 100 cancer cell lines), all with a minimum of 1 million miRNA reads [6]. In combination, the three teams generated small RNA-seq data from over a thousand cell-based samples. The data were both complementary and filled gaps in each other's data sets.

A key point of the two more global surveys is that cellular microRNA expression is generally not ubiquitous. The FANTOM5 group repurposed a 'tissue specificity index' [7] into a 'cell type specificity index' in which they noted many miRNAs including miR-122-5p, miR-142-5p, and miR-302a-5p being enriched in specific cell types. Other miRNAs, such as miR-100-5p and miR-29a-3p, could be broadly expressed, but reduced in specific cell types. The McCall paper identified 13 ubiquitously expressed miRNAs, present at a threshold of 100 reads per million miRNA reads (RPM) across 46 cell types, and 94 cell-specific miRNAs, present at a level of 1000 RPM or greater in only one class of cells.

In unison, these works reported the discovery of thousands of additional 'novel miRNAs'. Interestingly, the studies of McCall and the FANTOM5 consortium in particular express their own concerns that many of their predictions do not fulfill the basic properties of genuine miRNAs, and indeed this miRNA inflation is concerning as humans only share about 450 distinct miRNAs with the macaque [8]. Applying well-understood annotation criteria reduces these thousands of novel miRNA loci to, at best, a handful of genes that are truly novel miRNAs, albeit miRNAs are evolutionary very recent, and thus probably play very minor roles (if any) in translational regulation in human cell types [8]. A serious discussion as to what to name and define these noncanonical RNAs is sorely needed.

Surprisingly, there is a tremendous amount of variability in the RPM values between multiple cultures of the same cell type. Some of this may be related to the use of RPM as a normalization tool [3] or technical variation/batch effects [9]. However, it may also suggest fluctuating biology in cells. For example, in three donors of brachiocephalic smooth muscle cells from the FANTOM5 data, the RPM values for miR-100-5p ranged from 40 000 to nearly 340 000. This nearly ninefold difference in one of the most abundant miRNAs among these cells is surprising and would impact the RPM value of other miRNAs [3]. To estimate the variability in miRNA expression within cell types, it is crucial to include data that capture both biological and technical sources of variation. For example, by merging the data from these three studies, we now have access to 46 fibroblast samples from four different studies, allowing us to assess both within- and between-study variation. This is essential for downstream analyses such as tissue deconvolution with cell-specific miRNA signatures for disease studies (Figure 1).

As understanding the cell type-specific expression of miRNA signal is so useful,

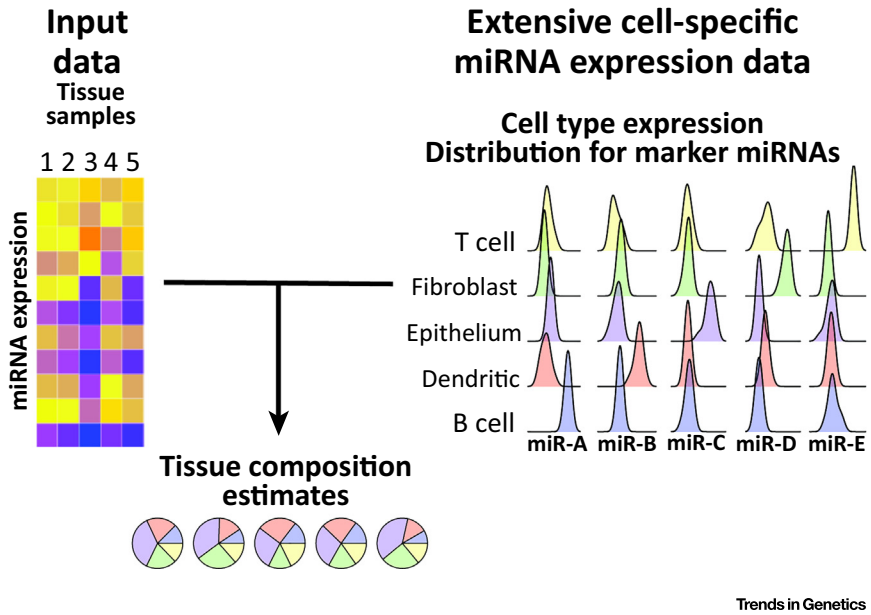

Figure 1. Robust Cell-Specific miRNA Expression Data Can Be Used to Improve the Analysis of Tissue miRNA Expression Data. This hypothetical example shows how one could use such data to estimate the composition of five tissue samples. Specifically, the input data are computationally compared with the cell-specific miRNA expression distributions of constitutive cell types to estimate the cellular composition of each tissue sample. Such deconvolution methods rely on a vast amount of cell-specific expression data to assess cell specificity. Adjusting for changes in tissue composition is important when comparing normal and diseased tissues where the composition, due to inflammation or fibrosis, is altered.

all three groups have made their data publicly available. The FANTOM5 group built an excellent Web server that is searchable by miRNA, cell type, or promoter location for both human and mouse samples (http://fantom.gsc.riken.jp/5/suppl/De_Rie_et_al_2017/vis_viewer/). The Hemmrich-Stanisak laboratory has a Web tool for miRNA comparisons across the blood cells (http://134.245.63.235/ikmb-tools/bloodmiRs/). The McCall group made their data available as an R package (http://bioconductor.org/packages/microRNAome) and combined the data from the three manuscripts (analyzing all data with miRge [10]) into a barChart track at the UCSC Genome Browser (http://www.genome.ucsc.edu/cgi-bin/hgHubConnect). These tools nicely complement a tissue-level miRNA atlas (https://ccb-web.cs.uni-saarland.de/tissueatlas/patterns) by the Keller group [7]. Finally, MirGeneDB.org has been created to serve as a repository for hand-curated, high-confident miRNA genes [8].

As much as these cell-specific miRNA expression data will clear up confusion in the literature, they also point to some gaps in our collective understanding of miRNA expressions. For example, among the sampled cell types, miR-1 is only found in skeletal and cardiac myocytes. However, a consistent, low miR-1-3p signal can be detected in organs including the prostate, breast, and colon, which are devoid of skeletal muscle cells. This could suggest that the appropriate miR-1-3p-expressing cell is not present in the combined atlases. Another interpretation is that miR-1 is expressed in a cell type that was covered, but that the miR-1 expression was lost in culture. Support for the later contention comes from the Levonen group [11] who compared the miRNA RNA-seq expression signatures of human umbilical vein endothelial cells from the time of isolation through three passages. They specifically noted a loss of miR-126-3p from about 130 000 transcripts per million (TPM) down to about 55 000 TPM in passaging. Conversely, miRNAs of the approximately 17–92 cluster, thought to be involved in cellular proliferation, were threefold to sixfold higher in passage 3 cells [11]. miR-126 is abundant and cell enriched in endothelial cells. Its lower levels in culture may indicate that many cell-restricted miRNAs that have specific tasks in matured cells are reduced while cells are actively proliferating and in a less mature phenotypic state in cell culture. Indeed, from our data [6], the levels of miR-200c, an epithelial-specific miRNA, are higher in bladder (about 60 000 RPM) than in urothelial cells (about 5000 RPM; the epithelial cell of the bladder wall). No other cell in the bladder can account for the discrepancy, supporting this idea of reduced maturity-related miRNA expression.

Unlike the cell culture data of the FANTOM5 and McCall reports, the Hemmrich-Stanisak miRNA expression data were built upon MACS-sorted cells [5]. These data would not have the same miRNA alterations observed from cell culture and should hew closer to accurate *in vivo* data. However, isolation methods (flow sorting vs. positive/negative bead selection) greatly impact miRNA expression variability [12]. In our experience, the extended time periods in buffers from similar activities (flow sorting) reduce the miRNA yield relative to other small RNAs. Indeed, for many cell types, the relative abundance of miRNAs in their samples was low. For example, of 40 neutrophil samples with at least 1 million total reads, only three had >500 000 miRNA reads (median 96 000).

Thus, while the curtain has been pulled back on cellular miRNA expression by

the evidence of these papers, another curtain appears behind it. To pull this second curtain back, we must develop strategies to isolate cells directly from tissues. What this successful technology will be is not yet clear, but without this information, our understanding of the roles miRNAs play in cellular differentiation and homeostasis remains incomplete. Accurate miRNA expression estimates are even more important as we learn about the importance of the relative abundance of miRNAs to their targets and sponges [13].

All together, these insights and resources greatly advance miRNA research.

Acknowledgments

M.K.H. and M.N.M. are supported by grant 1R01HL137811 from the National Institutes of Health. M.K.H. is also supported by an American Heart Association Grant-in-Aid (17GRNT33670405). M.N.M. is also supported by the University of Rochester CTSA award number UL1 TR002001 from the National Center for Advancing Translational Sciences of the National Institutes of Health. B.F. is supported by the South-Eastern Norway Regional Health Authority (Grant No. 2014041). K.J.P. is supported by NASA-Ames.

[1]Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA
[2]Department of Tumor Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, N-0424 Oslo, Norway
[3]Department of Biological Sciences, Dartmouth College, Hanover, NH 03755, USA
[4]Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY 14642, USA
@Twitter: @Marc_Halushka

*Correspondence: mhalush1@jhmi.edu (M.K. Halushka).
URL: http://labs.pathology.jhu.edu/halushka/.

https://doi.org/10.1016/j.tig.2017.12.015

References

1. Kent, O.A. *et al.* (2014) Lessons from miR-143/145: the importance of cell-type localization of miRNAs. *Nucleic Acids Res.* 42, 7528–7538
2. McCall, M.N. *et al.* (2011) MicroRNA profiling of diverse endothelial cell types. *BMC Med. Genom.* 4, 78
3. Witwer, K.W. and Halushka, M.K. (2016) Toward the promise of microRNAs – enhancing reproducibility and rigor in microRNA research. *RNA Biol.* 13, 1103–1116
4. de Rie, D. *et al.* (2017) An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nat. Biotechnol.* 35, 872–878
5. Juzenas, S. *et al.* (2017) A comprehensive, cell specific microRNA catalogue of human peripheral blood. *Nucleic Acids Res.* 45, 9290–9301
6. McCall, M.N. *et al.* (2017) Toward the human cellular microRNAome. *Genome Res.* 27, 1769–1781
7. Ludwig, N. *et al.* (2016) Distribution of miRNA expression across human tissues. *Nucleic Acids Res.* 44, 3865–3877
8. Fromm, B. *et al.* (2015) A uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. *Annu. Rev. Genet.* 49, 213–242
9. Pritchard, C.C. *et al.* (2012) MicroRNA profiling: approaches and considerations. *Nat. Rev. Genet.* 13, 358–369
10. Baras, A.S. *et al.* (2015) miRge – a multiplexed method of processing small RNA-seq data to determine microRNA entropy. *PLoS One* 10, e0143066
11. Kuosmanen, S.M. *et al.* (2017) MicroRNA profiling reveals distinct profiles for tissue-derived and cultured endothelial cells. *Sci. Rep.* 7, 10943
12. Schwarz, E.C. *et al.* (2016) Deep characterization of blood cell miRNomes by NGS. *Cell. Mol. Life Sci.* 73, 3169–3181
13. Pinzon, N. *et al.* (2017) microRNA target prediction programs predict many false positives. *Genome Res.* 27, 234–245

Forum

The Definition of Open Reading Frame Revisited

Patricia Sieber,[1] Matthias Platzer,[2] and Stefan Schuster[1,*]

The term open reading frame (ORF) is of central importance to gene finding. Surprisingly, at least three definitions are in use. We discuss several molecular biological and bioinformatics aspects, and we recommend using the definition in which an ORF is bounded by stop codons.

Open reading frame (ORF) is a basic term in molecular genetics and bioinformatics. The detection of ORFs is an important step in finding protein-coding genes in genomic sequences, including analyses based on highly fragmented draft (meta) genome assemblies [1–3]. ORFs can be detected by simple *in silico* analysis, while proving that a sequence is really protein-coding requires more effort. Surprisingly, in many textbooks not much attention is spent on defining the term ORF, apparently taking its meaning for granted. Moreover, the given definitions are often not perfectly clear-cut. For example, the standard textbook *Genes VII* by Lewin [4] states on p. 26: 'A reading frame that consists exclusively of triplets representing amino acids is called an open reading frame or ORF. A sequence that is translated into protein has a reading frame that starts with a special initiation codon (AUG) and that extends through a series of triplets representing amino acids until it ends at one of the three types of termination codon'. The first sentence defines an ORF as bounded by stop codons (stop/stop definition) whereas the second sentence may be (mis)understood as beginning with a start codon (start/stop definition). Currently at least three definitions are in use, which differ in the location of the ORF boundaries [5] (Box 1).

Before going into detail, it is worth recalling the different meanings of the term 'definition' itself. A 'lexical definition' reports the most common usage of a term [6,7]. It is the definition likely to be found in a dictionary and can change over time. An 'operational definition' focuses on a specific objective or application and may differ from the lexical definition [6]. In our case, the main objective is gene finding using bioinformatics software. The question arises of how the term ORF deviates from that of a coding DNA sequence (CDS). A CDS means a nucleotide sequence that is eventually translated into a protein [8]. This implies that the CDS of a particular protein is bounded by translation start and stop codons. In some cases the term ORF is considered equivalent to that of CDS [9]. Other authors describe an ORF as a potential protein-coding sequence which can be determined by sequence features alone [8]. Note that there is a difference between the