THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

An integrated expression atlas of miRNAs and their promoters in human and mouse

Citation for published version:
The FANTOM Consortium & de Rie, D 2017, 'An integrated expression atlas of miRNAs and their promoters in human and mouse', *Nature Biotechnology*, vol. 35, no. 9, pp. 872-878. https://doi.org/10.1038/nbt.3947

Digital Object Identifier (DOI):
10.1038/nbt.3947

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Nature Biotechnology

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 26. Aug. 2022

Ed sum

An atlas of microRNA expression patterns and regulators is produced by deep sequencing of short RNAs in human and mouse cells.

An integrated expression atlas of miRNAs and their promoters in human and mouse

Derek de Rie[1,2], Imad Abugessaisa[1], Tanvir Alam[3], Erik Arner[1,4], Peter Arner[5], Haitham Ashoor[3], Gaby Åström[5], Magda Babina[6], Nicolas Bertin[1,4,7], A. Maxwell Burroughs[1,4,8], Ailsa J. Carlisle[9], Carsten O. Daub[1,4], Michael Detmar[10], Ruslan Deviatiiarov[1,11], Alexandre Fort[1,4], Claudia Gebhard[12,13], Daniel Goldowitz[14], Sven Guhl[6], Thomas J. Ha[14], Jayson Harshbarger[1,4], Akira Hasegawa[1,4], Kosuke Hashimoto[1,4], Meenhard Herlyn[15], Peter Heutink[16], Kelly J. Hitchens[17], Chung Chau Hon[1], Edward Huang[18,19], Yuri Ishizu[1,4], Chieko Kai[20], Takeya Kasukawa[1], Peter Klinken[21], Timo Lassmann[1,4,22], Charles-Henri Lecellier[14,23], Weonju Lee[24], Marina Lizio[1,4], Vsevolod Makeev[25,26,27], Anthony Mathelier[14], Yulia A. Medvedeva[25,28,29], Niklas Mejhert[5], Christopher J. Mungall[30], Shohei Noma[1,4], Mitsuhiro Ohshima[31], Mariko Okada-Hatakeyama[32,33], Helena Persson[34], Patrizia Rizzu[16], Filip Roudnicky[10], Pål Sætrom[35], Hiroki Sato[20], Jessica Severin[1,4], Jay W. Shin[1,4], Rolf K. Swoboda[15], Hiroshi Tarui[1,4], Hiroo Toyoda[36], Kristoffer Vitting-Seerup[37], Louise Winteringham[21], Yoko Yamaguchi[38], Kayoko Yasuzawa[1], Misako Yoneda[20], Noriko Yumoto[33], Susan Zabierowski[39], Peter G. Zhang[14], Christine A. Wells[18,19], Kim M. Summers[9,40], Hideya Kawaji[1,4,41], Albin Sandelin[37], Michael Rehli[12,13], the FANTOM

consortium, Yoshihide Hayashizaki[4,41], Piero Carninci[1,4], Alistair R. R. Forrest[#,1,4,21], Michiel J. L. de Hoon[#,1,4].

[1]Division of Genomic Technologies, RIKEN Center for Life Science Technologies, Yokohama, Japan;

[2]Centre for Integrative Bioinformatics (IBIVU), VU University Amsterdam, Amsterdam, The Netherlands;

[3]Computational Bioscience Research Center, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia;

[4]RIKEN Omics Science Center (OSC), Yokohama, Japan[§];

[5]Department of Medicine, Karolinska Institutet at Karolinska University Hospital, Huddinge, Sweden;

[6]Department of Dermatology and Allergy, Charité Campus Mitte, Universitätsmedizin Berlin, Berlin, Germany;

[7]Cancer Science Institute of Singapore, National University of Singapore, Singapore;

[8]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA;

[9]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, UK;

[10]Institute of Pharmaceutical sciences, Swiss Federal Institute of Technology (ETH) Zürich, Zürich, Switzerland;

[11]Institute of Fundamental Medicine and Biology, Kazan Federal University, Kazan, Russia;

[12]Department of Internal Medicine III, University Hospital Regensburg, Regensburg, Germany;

[13]Regensburg Centre for Interventional Immunology (RCI), Regensburg, Germany;

[14]Department of Medical Genetics, Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute, University of British Columbia, Vancouver, British Columbia, Canada;

[15]Melanoma Research Center, The Wistar Institute, Philadelphia, Pennsylvania, USA;

[16]German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany;

[17]Australian Infectious Diseases Research Centre (AID), University of Queensland, Brisbane, Queensland, Australia;

[18]The University of Melbourne Centre for Stem Cell Systems, School of Biomedical Sciences, The University of Melbourne, Victoria, Australia;

[19]Walter and Eliza Hall Institute of Medical Research, Melbourne, Victoria, Australia;

[20]Laboratory Animal Research Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan;

[21]Harry Perkins Institute of Medical Research, and the Centre for Medical Research, University of Western Australia, QEII Medical Centre, Perth, Western Australia, Australia;

[22]Telethon Kids Institute, The University of Western Australia, Subiaco, Western Australia, Australia;

[23]Institute of Molecular Genetics of Montpellier, Montpellier, France;

[24]Department of Dermatology, Kyungpook National University School of Medicine, Daegu, South Korea;

[25]Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia;

[26]Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia;

[27]Moscow Institute of Physics and Technology, Dolgoprudny, Russia;

[28]IMPPC, Institute of Predictive and Personalized Medicine of Cancer, Ctra. de Can Ruti, Badalona, Spain;

[29]Institute of Bioengineering, Research Center of Biotechnology, Moscow, Russia;

[30]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA;

[31]Department of Biochemistry, Ohu University School of Pharmaceutical Sciences, Koriyama, Japan;

[32]Laboratory of Cell Systems, Institute for Protein Research, Osaka University, Osaka, Japan;

[33]RIKEN Center for Integrative Medical Sciences, Yokohama, Japan;

[34]Division of Oncology and Pathology, Department of Clinical Sciences, Lund University, Lund, Sweden;

[35]Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway;

[36]Department of Clinical Molecular Genetics, School of Pharmacy, Tokyo University of Pharmacy and Life Sciences, Tokyo, Japan.

[37]The Bioinformatics Centre, Department of Biology, and Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Copenhagen, Denmark;

[38]Department of Biochemistry, Nihon University School of Dentistry, Tokyo, Japan;

[39]The SKI Stem Cell Research Facility, The Center for Stem Cell Biology and Developmental Biology Program, Sloan Kettering Institute, New York, New York, United States;

[40]Mater Research Institute - University of Queensland, Translational Research Institute, Brisbane, Australia;

[41]RIKEN Preventive Medicine and Diagnosis Innovation Program, Wako, Japan.

[§]RIKEN Omics Science Center ceased to exist as of April 1st 2013 due to RIKEN reorganization.

Correspondence to Michiel de Hoon (michiel.dehoon@riken.jp) and Alistair Forrest (alistair.forrest@gmail.com)

MicroRNAs (miRNAs) are short non-coding RNAs with key roles in cellular regulation. As part of the fifth edition of the Functional Annotation of Mammalian Genome (FANTOM5) project, we created an integrated expression atlas of miRNAs and their promoters by deep sequencing 492 short RNA (sRNA) libraries, with matching Cap Analysis Gene Expression (CAGE) data, from 396 human and 47 mouse RNA samples. Promoters were identified for 1,357 human and 804 mouse miRNAs and showed strong sequence conservation between species. We also found that primary and mature miRNA expression levels were correlated, allowing us to use the primary miRNA measurements as a proxy for mature miRNA levels in a total of 1,829 human and 1,029 mouse CAGE libraries. We thus provide a broad atlas of miRNA expression and promoters in primary mammalian cells, establishing a foundation for detailed analysis of miRNA expression patterns and transcriptional control regions.

MicroRNAs[1] (miRNAs) are a class of short (21-23 nt) non-coding RNAs with key roles in a wide range of biological processes including development and differentiation[2,3], immunity[4], reproduction[5], and longevity[6]. Dysregulation of miRNA expression has been implicated in numerous diseases[7], including cancer[8,9]. A detailed characterization of the expression profile of miRNAs across cell types and tissues is a fundamental requirement for understanding the function of miRNAs and their potential role in health and disease.

MicroRNAs inhibit specific mRNAs by binding to complementary sequences, usually located in the 3' UTR, leading to mRNA destabilization and a reduction in their translation output[10]. In the canonical miRNA biogenesis pathway[1,11], a primary miRNA transcript (pri-miRNA) is cleaved by the endoribonuclease Drosha in the nucleus to excise the precursor miRNA (pre-miRNA), which is exported to the cytoplasm. The pre-miRNA has a characteristic hairpin secondary structure that is recognized and cleaved in the cytoplasm by the endoribonuclease Dicer, releasing the mature miRNA.

Currently, the miRBase reference database of miRNAs[12] lists 1,881 pre-miRNAs in human; around half (54%) are produced from intergenic non-coding pri-miRNA transcripts, while the remaining 46% are excised from the introns of protein-coding transcripts. A small proportion (6%) of human mature miRNAs annotated in miRBase are located in multiple pre-miRNAs encoded in different genomic loci.

Several high-throughput approaches are available to measure the expression levels of mature miRNAs, including high-throughput qPCR, microarray, and next-generation

sequencing methods[13]. Profiling pri-miRNAs, which is more challenging due to their transient character, has been accomplished by RNAseq in cells expressing dominant-negative Drosha[14]. Additionally, since most pri-miRNAs are produced by RNA polymerase II and therefore have a 5' cap[11], they are amenable to Cap Analysis Gene Expression (CAGE) profiling[15,16], which identifies the pri-miRNA transcription start site and therefore the promoter region, while directly quantitating the pri-miRNA expression level.

Here, we analyze 492 sRNA sequencing libraries to evaluate the expression patterns of miRNAs in mammalian cells, with a particular emphasis on human primary cells. Each sRNA library was matched to a CAGE library produced from the same RNA sample, allowing us to create an integrated expression atlas of miRNAs and their promoters. The expression atlas can be accessed through a web interface at http://fantom.gsc.riken.jp/5/suppl/De_Rie_et_al_2017/). This work is part of the fifth edition of the Functional Annotation of Mammalian Genome project (FANTOM5)[17,18].

Results

Matched miRNA and CAGE expression profiles

In FANTOM5, a large collection of human and mouse primary cell types, cell lines, and tissues was profiled by CAGE to identify mRNA and long non-coding RNA transcription start sites and expression levels across a wide variety of biological states[17]. Here, as a complementary data set, we produced 293 sRNA sequencing libraries using FANTOM5 RNA samples from human primary cells, as well as 87 sRNA libraries from RNA samples of six time courses of stimulated human cells[18] (Table 1, S1 & S2). We also incorporated previously produced CAGE and sRNA sequencing libraries generated from human embryonic and induced pluripotent stem cells[19] (Table 1 & S1) in our analysis. In total, our sRNA sequencing data set encompassed 121 distinct human cell types. In addition, we produced 6 sRNA sequencing libraries from human tissues, and 42 sRNA libraries from mouse samples (Table 1, S1 & S2). Most sRNA libraries were produced in biological triplicate. A matching CAGE library[17-19] generated from the same RNA sample was available for 492 of the 500 sRNA libraries analyzed here (Table S3).

Establishing a robust set of miRNAs

Across the sRNA libraries, expression was confirmed for 98% (1842/1877) of human and 95% (1124/1186) of mouse pre-miRNAs annotated in release 21 of the miRBase database[12]. To assess the confidence level of annotated miRNAs, the miRBase curators defined a set of five rules evaluating their secondary structure and expression properties (Table 2), and used these rules to mark 295 human pre-miRNAs as high-confidence annotations[12]. Applying these rules to the FANTOM5 sRNA data, we found that 571 human pre-miRNAs satisfied all five high-confidence rules, 224 met

four of them, and 1076 violated two or more rules (Figure 1a). The 795 human and 502 mouse (Figure S1) pre-miRNAs satisfying at least 4 out of the 5 high-confidence rules were defined as the FANTOM5 robust set, and the remaining 1076 human and 684 mouse pre-miRNAs as the permissive set (Table S4 and Table S5). The robust set encompasses 735 human and 438 mouse mature miRNAs, and covers more than 90% of the high-confidence pre-miRNAs in miRBase (Figures S2 and S3), 90% of miRNAs well characterized in the scientific literature (Figure S4), as well as 91% (human) and 88% (mouse) of pre-miRNAs included in the manually curated MirGeneDB database[20] (Figure S5).

CAGE detects 3' cleavage products of Drosha

In zebrafish, the Drosha cleavage site at the 3' end of pre-miRNAs was recently found to be characterized by a distinctive CAGE peak[21]. We similarly observed a CAGE peak immediately downstream of the 3' end of human pre-miRNA loci in the ENCODE CAGE data[22], and a slightly wider CAGE peak starting 1 nucleotide downstream in the FANTOM5 CAGE data[17,18] (Figure 1b, c, human; Figure S6, mouse); the discrepancy between the ENCODE and FANTOM5 CAGE data is expected because of differences in the sequencer technologies employed (Figure S7). The ENCODE CAGE peak was found immediately downstream of the 3' end of the pre-miRNA locus (Figure S8) for 19 out of 25 pre-miRNAs with a full-length sequence in the FANTOM4 sRNA sequencing libraries[23], confirming that the CAGE peak marks the Drosha cleavage site. FANTOM5 and ENCODE CAGE tags at the peak were enriched in the nucleus (Figure S9), consistent with processing by Drosha. CAGE peaks were absent at the 3' end of pre-miRNA loci encoding mirtrons (Figure

S10, human; Figure S11, mouse), which are excised by the spliceosomal machinery instead of by Drosha[24].

To rule out the possibility that these CAGE tags originated from an independent transcript, we analyzed the first nucleotide of the CAGE tags at the Drosha cleavage site. Most CAGE tags originating from a transcription start site have an additional guanine as their first nucleotide, as the 7-methylguanosine cap at the 5' end of transcripts produced by RNA polymerase II can be recognized as a guanine nucleotide during reverse transcription (Figure S7). No such additional guanine nucleotides were found at the Drosha CAGE peak (Figure S12), confirming that the detected RNAs were not due to an independent transcription initiation event. The lack of guanine nucleotide enrichment also suggested that the 3' Drosha cleavage products were uncapped RNAs that were nonetheless observed to some extent in the CAGE library due to their cellular abundance. Alternatively, these RNAs may have a hypermethylated cap, as previously found for small nucleolar RNAs (snoRNAs) produced by excision from a host gene transcript[25]: no additional guanines are found as the first nucleotide of CAGE tags mapping to the 5' end of snoRNAs (Figure S12), as hypermethylation of the cap prevents base-pairing during reverse transcription.

Excluding mirtrons, about half of the robust pre-miRNAs had a significant ($P < 0.05$) Drosha CAGE peak (52%, human, Figure 1a; 64%, mouse, Figure S1; see Methods for details). This percentage decreased from 56% for human pre-miRNAs satisfying all five of the miRBase high-confidence criteria to 37% if one of the criteria was violated, while only 7% of miRNAs in the permissive set had a Drosha CAGE peak (Figure 1a). Similar results were obtained for mouse (Figure S1). The analysis of

241 Drosha CAGE peaks thus provided independent support for the stringency of the
242 selection criteria used to define the FANTOM5 robust and permissive set of miRNAs.
243
244 **Discovery of candidate novel miRNAs**
245 To discover potential miRNAs that had not been described previously, the miRDeep2
246 software[26] was applied on all unannotated sRNAs (see Methods for details). In total,
247 6,543 candidate miRNAs in human (Table S6, S7) and 1,444 in mouse (Table S8, S9)
248 were identified. Most of the candidate miRNAs were lowly expressed, with fewer
249 than 5% of them having sufficient tag counts on both arms of the pre-miRNA to
250 enable a full evaluation of the high-confidence criteria (Table 2). The 282 human and
251 34 mouse candidate miRNAs meeting at least 4 of the 5 high-confidence criteria
252 formed the robust candidate set, while the permissive candidate set consisted of the
253 remaining candidate miRNAs (Table S4 and Figure S13, human; Table S5 and Figure
254 S14, mouse). The robust candidate set comprised 279 (human) and 33 (mouse) unique
255 mature sequences, whereas the permissive candidate set provided an additional 5,826
256 (human) and 1,354 (mouse) mature sequences. Nearly 11% of robust and 5% of
257 permissive human candidate miRNAs had a significant ($P < 0.05$) Drosha CAGE
258 peak (Figure S13; Figure S14 for mouse). Validation by qPCR of a selection of robust
259 candidate miRNAs identified in monocyte and macrophage libraries confirmed their
260 expression in these cell types in multiple donors (Figure S15, Table S10).
261
262 The robust candidate set showed good concordance (127/282 or 45%) with the 3,524
263 putative miRNAs identified recently in a study of tissue- and primate-specific
264 miRNAs[27], whereas the permissive candidate set yielded a smaller overlap (352/6,261
265 or 6%). Few of these putative miRNAs[27] had a significant Drosha CAGE peak

(258/3,524 or 7%), which may be due to their low expression levels in the samples surveyed in FANTOM5.

We conclude that the vast majority of canonical, highly expressed miRNAs had already been annotated. However, our analysis also provides evidence of extensive transcription of lowly expressed short RNAs from specific genomic loci.

Expression variability of miRNAs in human primary cells

The cell type dependence of expression of individual miRNAs was evaluated by analyzing the distribution of miRNA abundance across the FANTOM5 primary cells and tissues. First, we assessed various expression normalization strategies, and found that a counts per million (cpm) normalization (i.e., dividing the tag count of each miRNA by the total number of tags mapping to miRNA loci, and multiplying by 1,000,000) yielded the best reproducibility between different donors for the same cell type, while maintaining the distinction in expression profile between different cell types (Figure S16; see Methods for details). We then created a miRNA expression table across the FANTOM5 samples for human (Table S11) and mouse (Table S12), using cpm normalization in our further analysis. Figure 2a shows a graphical overview of the human primary cells clustered by their robust miRNA expression profile using Miru[28]. An interactive heatmap of the expression data is available at http://fantom.gsc.riken.jp/5/suppl/De_Rie_et_al_2017/vis_viewer/#/heatmap.

The expression levels of miRNAs varied greatly and were highly skewed, with on average five miRNAs contributing half of the total miRNA expression in a given library (Figure 2b, human; Figure S17, mouse), whereas most known and candidate miRNAs were lowly expressed (Figure 2c, human; Figure S18, mouse). The extreme

distribution of miRNA expression across miRNAs and cell types was confirmed by qPCR (Figure S19).

Cell ontology analysis

A cell type specificity index, analogous to the previously defined tissue specificity index[29], was calculated to quantify the cell type specificity of miRNA expression across the FANTOM5 collection of primary cell types (Table S13; see Methods for details). Previously described highly cell type specific miRNAs included miR-122-5p, miR-142-5p and miR-302a-5p, which were enriched in hepatocytes, leukocytes, and pluripotent stem cells, respectively (Figure 2d). In contrast, miRNAs such as miR-100-5p and miR-29a-3p were broadly expressed but specifically depleted in particular cell types (leukocytes and pluripotent stem cells, respectively; Figure 2d). Candidate miRNAs tended to be restricted to specific cell types, with 80% of the robust candidate set and 96% of the permissive candidate set having a higher cell type specificity index than the median value for robust known miRNAs (Table S13).

We then calculated the statistical significance of expression enrichment or depletion of each miRNA (Table S13) with respect to cell ontology clusters (Table S14) defined by the FANTOM5 cell ontology annotation[30,31], which organizes FANTOM5 samples by cell type in a hierarchical framework (see Methods for details). Of miRNAs in the robust set, 54% had enriched expression in their most significant cell ontology cluster, whereas 27% were broadly expressed, with depleted expression in their most significant cell ontology cluster. The remaining 19% were lowly expressed without statistically significant enrichment or depletion in any cell ontology cluster; understanding their functionality may need profiling in further cell types or states.

316 Pluripotent stem cells were characterized by cell type specific miRNAs, whereas cell
317 type specific depletion of broadly expressed miRNAs was predominantly found in
318 leukocytes. Examples of enriched expression not reported previously included miR-
319 488-5p in neural cells, miR-506-3p in light melanocytes, and miR-205-5p in epithelial
320 cells. MiRNAs previously not reported as broadly expressed included miR-887-3p,
321 which was present in most samples but was depleted in leukocytes.
322
323 **Identification of miRNA promoters**
324 We developed an automatic pipeline to identify miRNA promoters using Gencode
325 v19 and RefSeq transcripts as candidate pri-miRNAs and the FANTOM5 CAGE data
326 as putative transcription start sites (see Methods for details). This pipeline predicted
327 promoters for 539 robust, 623 permissive, and 3,951 candidate pre-miRNAs in human
328 (Table S15), and for 358 robust, 446 permissive, and 994 candidate pre-miRNAs in
329 mouse (Table S16). Manual curation by two independent annotators confirmed the
330 selected promoter for 512 (95%) robust pre-miRNAs; the computationally selected
331 promoter was corrected for 26 pre-miRNAs and dropped for 1 pre-miRNA. Manual
332 curation furthermore identified the promoter for an additional 196, mostly intergenic,
333 pre-miRNAs, thereby generating the—to our knowledge—largest miRNA promoter
334 collection to date (Table S17, Figure S20a). Across the human robust set, an
335 associated ENCODE RAMPAGE (RNA Annotation and Mapping of Promoters for
336 the Analysis of Gene Expression[32]) 5' end was found within 300 base pairs of more
337 than 75% of the FANTOM5 curated promoters both for intergenic and intronic
338 miRNAs, outperforming the miRGen[33], Chang *et al.*[14], miRStart[34], and TSmiR[35]
339 collections of miRNA promoter annotations (Figure S20b). The median distance
340 between the FANTOM5 annotated miRNA promoter and the associated RAMPAGE

5’ end was 1 nucleotide, and was thereby closer than any of the existing miRNA promoter annotations (Figure S20c). RACE experiments confirmed that the transcripts generated at the identified promoter extended to the mature miRNA for 6 out of 7 miRNAs (Figure S21, Table S18). RNA-seq data[14] of cells expressing a dominant-negative Drosha protein provided additional evidence for the FANTOM5 annotated pri-miRNAs, with 483 out of 607 pri-miRNAs (80%) having a 5’ end within 300 base pairs of an RNAseq transcript assembly extending to the mature miRNA locus (Figure S22).

Both in human and in mouse, promoter sequences of intronic and intergenic miRNAs, like those of transcription factor genes, were highly conserved across species compared to the promoter sequences of protein-coding genes and of long non-coding RNAs (Figures 3a and S23, human; Figure S24, mouse). The distance between the transcription start site of the pri-miRNA and the mature miRNA locus was strongly conserved between human and mouse both for intergenic miRNAs and for intronic miRNAs (Figures 3b and S25). While this suggests that pri-miRNA transcripts may have some functional role beyond providing the substrate for pre-miRNA excision, there was no evidence of substantially elevated sequence conservation across species in pri-miRNAs (Figure S26).

Correlation of mature miRNA and pri-miRNA expression levels

The expression levels of mature miRNAs correlated with the CAGE expression levels of the associated promoter, with comparable correlation values for intergenic and intronic miRNAs (Figure 3c and S27; Table S19). The correlation was substantially higher for highly differentially expressed miRNAs, and exceeded correlations found

for previously published[14,33-35] miRNA promoter annotations (Figure S20d). About 11% of pri-miRNAs in human were polycistronic, containing multiple mature miRNAs with highly correlated expression levels (Figure 3d and S28). Together this suggests that miRNA expression is primarily regulated at the transcriptional level.

Using the CAGE expression level of the pri-miRNA as a proxy for the mature miRNA expression level, we extended the FANTOM5 miRNA expression atlas to the full breadth of the 1,829 (human) and 1,029 (mouse) libraries in the FANTOM5 CAGE expression compendium[17-19]. This allowed us to assess miRNA expression also in samples for which only a CAGE library was available, covering an additional 49 primary cell types, 245 cell lines, 138 tissue types, and 13 time courses in human, and an additional 48 primary cell types, 1 cell line, 234 tissue types, and 12 time courses in mouse. A cell ontology analysis was performed using the CAGE expression pattern of each human pri-miRNA (Tables S15 and S17) across 338 cell ontology clusters (Table S20) encompassing 636 CAGE libraries. This showed enriched expression of mir-202 in gonad, of mir-208a, known to be a key regulator of cardiac function[37], in heart, as well as of multiple miRNAs in brain, including mir-488, mir-556, and mir-885. Lastly, the CAGE data allowed us to measure the individual contribution of each paralog to the expression of miRNAs encoded multiple times in the human genome, providing evidence for differential regulation of paralogs in different cell types and tissues. For example, we found that mir-128-1 was expressed in most samples, while its paralog mir-128-2 was highly enriched in brain (Figure 3e).

Transcriptional regulation of miRNA expression

The accuracy of the predicted miRNA promoter regions was assessed using the Motif Activity Response Analyis (MARA) framework[38] (outlined in Figure S29 and Methods). Using this framework, the expression levels of mature miRNAs were predicted based on the presence of putative transcription factor binding sites in the identified miRNA promoter region, and compared to the expression levels of the mature miRNAs observed in the sRNA libraries. The prediction accuracy of the FANTOM5 miRNA promoter atlas outperformed those of previously published miRNA promoter annotations[14,33-35] (Figure S20e).

Discussion

MicroRNAs are key factors that contribute to cellular regulation by targeting specific transcripts for translational repression or for degradation. Advances in sequencing technology led to an increase in sequencing depth from nearly 1,300 reads per sRNA libraries in the first miRNA atlas[39] to nearly 4.4 million reads per library in FANTOM5, allowing an accurate measurement of the expression even of lowly expressed miRNAs. These lowly expressed miRNAs may be abundant in a few cells in the population sampled, or in cell types, cell lines, or cellular conditions that are not included in our sample collection. Alternatively, they may be a signature of the ongoing evolution of the human miRNA repertoire. In particular, pervasive transcription of mammalian genomes[22,40,41] generates a large number of hairpin secondary structures, which are prevalently encoded in the genome, that can act as substrates for processing by Drosha in the nucleus and Dicer in the cytoplasm. Whereas the majority of the sRNAs thus generated may be evolutionarily neutral and remain lowly expressed, some of them may provide a selective advantage, develop higher expression levels during evolution[42], and become fixed in the genome as core

miRNAs. Finally, we note that in spite of the breadth and depth of the FANTOM5 sRNA sequencing data, most sRNAs currently annotated as miRNAs failed multiple high-confidence criteria, and may belong to a different class of short non-coding RNAs, such as transcription initiation RNAs[23] or DNA damage response RNAs[43], or may be degradation products[44].

Compared to existing miRNA expression atlases[29,39], the FANTOM5 atlas covers the widest range of normal primary cells, enabling detailed analyses of miRNA expression and their contribution to establishing and maintaining cell type identity. The candidate miRNAs not reported previously were in particular highly cell type specific, and may therefore be missed in miRNA profiling studies in tissues rather than in specific cell types.

We found extensive evidence that CAGE peaks observed at the Drosha cleavage site are due to the downstream RNA fragment generated by Drosha processing of the pri-miRNA. Analysis of these CAGE tags suggested that these RNA fragments do not have a 7-methylguanosine cap, but may instead be uncapped or, alternatively, have a non-canonical cap. For polycistronic pri-miRNAs, such a cap may play a role in preventing rapid degradation of the downstream fragment, which itself may contain miRNAs.

The MARA analysis allowed us to predict miRNA expression levels based on the regulatory motifs found in the miRNA promoter region, indicating that transcriptional regulation plays a central role in governing miRNA expression levels. Comparing the promoters of miRNAs, protein-coding genes, and long non-coding RNAs showed

similar prevalences of transcription factor binding sites in proximal promoter regions (data not shown), suggesting that the basic mechanisms of transcriptional regulation are largely the same for these three classes of gene products. The identification of miRNA promoter regions as described in this work therefore paves the way for a detailed analysis of the transcriptional regulation of miRNA expression using the same computational and experimental methods that have previously proven their efficacy in the analysis of gene expression.

Acknowledgements

FANTOM5 was made possible by the following grants: Research Grant for RIKEN Omics Science Center from MEXT to Y.H.; Grant of the Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from the MEXT, Japan to Y.H.; Research Grant from MEXT to the RIKEN Center for Life Science Technologies; Research Grant to RIKEN Preventive Medicine and Diagnosis Innovation Program from MEXT to Y.H. K.V.S and A.S. were supported by the Lundbeck and Novo Nordisk Foundations. A.R.R.F. is supported by a Senior Cancer Research Fellowship from the Cancer Research Trust, funds raised by the MACA Ride to Conquer Cancer, and the Australian Research Council's Discovery Projects funding scheme (DP160101960). Y.A.M. was supported by the Russian Science Foundation, grant 15-14-30002. R.D. was supported by the Russian Science Foundation, grant 14-44-00022. We would like to thank Lucia Schwarzfischer for technical assistance and Norbert Eichner and Gunter Meister for sequencing RACE products. We would also like to thank GeNAS for data production.

Authors' contributions

P.A., G.Å., M.B., A.J.C., M.D., D.G., S.G., T.J.H., M.H., P.H., K.J.H., C.K., P.K., W.L., N.M., M.O., M.O.H., P.R., H.S., R.K.S., H.To., M.Y., N.Y., S.Z., P.G.Z., L.W., Y.Y., C.A.W., K.M.S., A.R.R.F. provided RNA samples; E.A. and C.O.D. selected samples from the FANTOM5 time courses; Y.I., S.N., and H.Ta. produced the sRNA libraries; I.A., M.L., H.K., and T.K. managed the data; D.d.R., M.J.L.d.H., K.V.S., A.M.B., T.A., H.A., A.H., T.L., H.P., C.L., A.M., V.M., M.R. carried out the bioinformatics analyses with the help of C.C.H., M.L., K.H., F.R., and J.S.; C.J.M. provided the cell ontology; K.M.S. created the Miru visualization; A.F., A.M.,

A.R.R.F., A.S., C.L., C.A.W., D.d.R., E.H., F.R., H.P., K.V.S., A.M.B., M.J.L.d.H., M.R., N.B., P.S., R.D., V.M., Y.A.M. contributed to the manual miRNA promoter annotation; K.Y. and J.W.S. performed the expression validation experiments of known miRNAs; E.H. and C.A.W. performed the validation experiments of candidate miRNAs; C.G. and M.R. performed the RACE experiments; J.H. created the web visualization tool; D.d.R., A.R.R.F. and M.J.L.d.H. wrote the manuscript with the help of E.A., A.S., A.M.B., K.M.S., K.V.S., M.R., N.B., P.C., P.S., C.A.W.; A.R.R.F. and M.J.L.d.H. designed the study; P.C. and Y.H. supervised the FANTOM5 project.

Competing financial interests

The authors declare no competing interests.

References

1. Bartel, D.P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
2. Shenoy, A. & Blelloch, R.H. Regulation of microRNA function in somatic stem cell proliferation and differentiation. *Nat. Rev. Mol. Cell Biol.* **15**, 565–576 (2014).
3. Li, M. & Izpisua Belmonte, J.C. Roles for noncoding RNAs in cell-fate determination and regeneration. *Nat. Struct. Mol. Biol.* **22**, 2–4 (2015).
4. Mehta, A. & Baltimore, D. MicroRNAs as regulatory elements in immune system logic. *Nat. Rev. Immunol.* **16**, 279–294 (2016).
5. Hasuwa, H., Ueda, J., Ikawa, M. & Okabe, M. miR-200b and miR-429 function in mouse ovulation and are essential for female fertility. *Science* **341**, 71–73 (2013).
6. Sun, K. & Lai, E.C. Adult-specific functions of animal microRNAs. *Nat. Rev. Genet.* **14**, 535–548 (2013).
7. Mendell, J.T. & Olson, E.N. MicroRNAs in stress signaling and human disease. *Cell* **148**, 1172–1187 (2012).
8. Adams, B.D., Kasinski, A.L. & Slack, F.J. Aberrant regulation and function of microRNAs in cancer. *Curr. Biol.* **24**, R762–R776 (2014).
9. Lin, S. & Gregory, R.O. MicroRNA biogenesis pathways in cancer. *Nat. Rev. Cancer* **15**, 321–333 (2015).
10. Jonas, S. & Izaurralde, E. Towards a molecular understanding of microRNA-mediated gene silencing. *Nat. Rev. Genet.* **16**, 421–433 (2015).
11. Ha, M. & Kim, V.N. Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell. Biol.* **15**, 509–524 (2014).

12. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence miRNAs using deep sequencing data. *Nucleic Acids Res.* **42**, D68–D73 (2014).
13. Pritchard, C.C., Cheng, H.H. & Tewari, M. MicroRNA profiling: approaches and considerations. *Nat. Rev. Genet.* **13**, 358–369 (2012).
14. Chang, T.C., Pertea, M., Lee, S., Salzberg, S.L. & Mendell, J.T. Genome-wide annotation of microRNA primary transcript structures reveals novel regulatory mechanisms. *Genome Res.* **25**, 1401–1409 (2015).
15. Kanamori-Katayama, M. *et al.* Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Res.* **21**, 1150–1159 (2011).
16. Takahashi, H., Lassmann, T., Murata, M. & Carninci, P. 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat. Protoc.* **7**, 542–561 (2012).
17. Forrest, A.R.R. *et al.* A promoter level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
18. Arner, E. *et al*. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* **347**, 1010–1014 (2015).
19. Fort, A. *et al.* Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nat. Genet.* **46**, 558–566 (2014).
20. Fromm, B. *et al.* A uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. *Annu. Rev. Genet.* **49**, 213–242 (2015).
21. Nepal, C. *et al.* Transcriptional, post-transcriptional and chromatin-associated regulation of pri-miRNAs, pre-miRNAs and moRNAs. *Nucleic Acids Res.* **44**, 3070–3081 (2015).

536 22. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–
537 108 (2012).
538 23. Taft, R.J. *et al.* Tiny RNAs associated with transcription start sites in animals.
539 *Nat. Genet.* **41**, 572–578 (2009).
540 24. Westholm, J.O. & Lai, E.C. Mirtrons: microRNA biogenesis via splicing.
541 *Biochimie* **93**, 1897–1904 (2011).
542 25. Matera, A.G., Terns, R.M. & Terns, M.P. Non-coding RNAs: lessons from the
543 small nuclear and small nucleolar RNAs. *Nat. Rev. Mol. Cell Biol.* **8**, 209–220
544 (2007).
545 26. Friedländer, M.R., Mackowiak, S.D., Li, N., Chen, W. & Rajewsky, N.
546 miRDeep2 accurately identifies known and hundreds of novel miRNA genes
547 in seven animal clades. *Nucleic Acids Res.* **40**, 37–52 (2012).
548 27. Londin, E. *et al.* Analysis of 13 cell types reveals evidence for the expression
549 of numerous novel primate- and tissue-specific miRNAs. *Proc. Natl. Acad.*
550 *Sci. USA* **112**, E1106–E1115 (2015).
551 28. Freeman, T.C., *et al.* Construction, visualisation, and clustering of
552 transcription networks from microarray expression data. *PLoS Comput. Biol.*
553 **3**, 2032–2042 (2007).
554 29. Ludwig, N. *et al.* Distribution of miRNA expression across human tissues.
555 *Nucleic Acids Res.* **44**, 3865–3877 (2016).
556 30. Meehan, T.F. *et al.* Logical development of the cell ontology. *BMC*
557 *Bioinformatics* **12**, 6 (2011).
558 31. Lizio, M. *et al.* Gateways to the FANTOM5 promoter level mammalian
559 expression atlas. *Genome Biol.* **16**, 22 (2015).

32. Batut, P., Dobin, A., Plessy, C., Carninci, P. & Gingeras, T. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res.* **23**, 169–180 (2013).
33. Georgakilas, G. *et al.* DIANA-miRGen v3.0: accurate characterization of microRNA promoters and their regulators. *Nucleic Acids Res.* **44**, D190–D195 (2016).
34. Chien, C.H. *et al.* Identifying transcriptional start sites of human microRNAs based on high-throughput sequencing data. *Nucleic Acids Res.* **39**, 9345–9356 (2011).
35. Guo, Z. *et al.* Genome-wide survey of tissue-specific microRNA and transcription factor regulatory networks in 12 tissues. *Sci. Rep.* **4**, 5150 (2014).
36. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
37. Ding, J. *et al.* Trbp regulates heart function through miRNA-mediated Sox6 repression. *Nat. Genet.* **47**, 776–783 (2015).
38. Suzuki, H. *et al.* The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat. Genet.* **41**, 553–562 (2009).
39. Landgraf, P. *et al.* A mammalian miRNA expression atlas based on small RNA library sequencing. *Cell* **129**, 1401–1414 (2007).
40. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
41. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484–1488 (2007).

42. Meunier, J. *et al.* Birth and expression evolution of mammalian miRNA genes. *Genome Res.* **23**, 34–45 (2013).

43. Francia, S. *et al.* Site-specific DICER and DROSHA RNA products control the DNA-damage response. *Nature* **488**, 231–235 (2012).

44. Valen, E. *et al.* Biogenic mechanisms and utilization of small RNAs derived from human protein-coding genes. *Nat. Struct. Mol. Biol.* **18**, 1075–1082 (2011).

Figure 1: Selection of robust miRNAs and Drosha CAGE peak analysis.

(**a**) Number of miRBase high-confidence rules (Table 2) satisfied by human pre-miRNAs annotated in miRBase (n = 1,871). Pre-miRNAs with a statistically significant ($P < 0.05$) Drosha CAGE peak are shown in orange; mirtrons are shown in yellow. (**b**) Genomic locus of mir-223 in human with the total number of FANTOM5 (blue) and ENCODE (red) CAGE tags as a function of the genomic position of their 5' end, showing a Drosha CAGE peak at the 3' end of the pre-miRNA. FANTOM5 sRNA reads are shown at the bottom, colored by their read count as defined by the color bar. The exact extent of the pre-miRNA was determined from FANTOM4 full-length sequencing data[23]. (**c**) Number of CAGE tags as a function of their starting position relative to the 3' end of the pre-miRNA, averaged across human pre-miRNAs in the robust set (n = 795). The 3' end of the pre-miRNA was selected as the 3' end of the most prevalent sRNA on the 3' arm of the pre-miRNA in the FANTOM5 sRNA data, with the position indicated as zero corresponding to the first nucleotide downstream of the 3' end of the pre-miRNA.

Figure 2: Expression profile and cell ontology analysis of mature miRNAs.

(**a**) Miru[28] visualization of FANTOM5 primary cell samples based on their expression profile of robust mature miRNAs. (**b**) Number of most abundant miRNAs contributing at least 50% of the total miRNA expression in each human sRNA library in FANTOM5 (n = 420). (**c**) Reverse cumulative distribution of the maximum expression across the FANTOM5 samples of human miRNAs in the robust set (n = 735), permissive set (n = 999), and robust candidate set (n = 279). (**d**) Examples of miRNAs enriched or depleted in specific primary cell samples. Expression of miR-122-5p, miR-142-5p, and miR-302a-5p was enriched in hepatocytes, leukocytes, and

pluripotent stem cells, respectively; miR-100-5p and miR-29a-3p were broadly expressed, but depleted in leukocytes and pluripotent stem cells, respectively.

Figure 3: Analysis of the curated miRNA promoters of miRNAs in the robust set. (**a**) (left panel) Sequence conservation of the human genome, evaluated as the average phastCons[36] score, in the promoter region of non-coding pri-miRNAs (containing intergenic mature miRNAs; $n = 132$), coding pri-miRNAs (containing intronic mature miRNAs; $n = 415$), transcription factor (TF)-coding transcripts ($n = 1{,}651$), other protein-coding transcripts ($n = 15{,}350$), and long non-coding RNAs ($n = 1{,}461$). The sequence conservation of randomly selected genome regions is shown in gray. The shaded area corresponds to one standard deviation in the estimated mean phastCons score. (right panel) The average sequence conservation at promoter regions of miRNAs was higher than at the promoter regions of non-TF protein-coding genes (Mann-Whitney $P = 2 \times 10^{-16}$, two-sided) and of long non-coding RNAs (Mann-Whitney $P = 1 \times 10^{-35}$, two-sided). Error bars correspond to one standard deviation in the estimated mean phastCons score. (**b**) Distance between the transcription start site of the pri-miRNA and the 5' end of the first pre-miRNA is highly correlated between human and mouse both for coding (Spearman $r = 0.90$; $n = 78$; Student $t = 18.27$; $P = 2 \times 10^{-29}$ two-sided) and for non-coding (Spearman $r = 0.86$; $n = 27$; Student $t = 8.33$; $P = 1 \times 10^{-8}$ two-sided) pri-miRNAs, suggesting strong conservation of the genomic extent of pri-miRNAs. (**c**) Expression levels of pri-miRNAs, as measured by CAGE, and mature miRNAs, as measured by sRNA sequencing, were highly correlated both for coding (average Spearman $r = 0.25$; $n = 362$; $P = 2 \times 10^{-53}$, Mann-Whitney U test, one-sided) and non-coding (average Spearman $r = 0.27$; $n = 180$; $P = 1 \times 10^{-30}$, Mann-Whitney U test, one-sided) pri-miRNAs, compared to a background

distribution consisting of correlations between randomly paired pri-miRNAs and mature miRNAs. Correlations for polycistronic pri-miRNAs were averaged across the mature miRNAs. (**d**) Expression levels between mature miRNAs originating from the same pri-miRNA are highly correlated (average Spearman $r = 0.74$; $n = 1{,}372$; $P < 10^{-100}$, Mann-Whitney U test, one-sided), compared to a background distribution consisting of correlations between mature miRNAs originating from different pri-miRNAs. (**e**) Cell type-dependent expression of miRNA paralogs: While mir-128-1 was broadly expressed across most primary cell samples in FANTOM5, its paralog mir-128-2 was enriched in brain samples. (**c**-**d**) The box extends from the lower to the upper quartile, with the center line at the median; the whiskers indicate the full range of the data.

Table 1.

Human sRNA data sets analyzed in this study.

Origin	Data collection	Number of samples	Number of cell types	
Primary cells	FANTOM5	293	118	119
	Fort *et al.*[19]	6	3	
ES cells	Fort *et al.*[19]	6		1
iPS cells	Fort *et al.*[19]	6		1
Tissues	FANTOM5	6		4 tissues
Time courses	FANTOM5	87		6 time courses
Total number of sequenced reads: 1,519,621,910				

660 **Table 2.**

661 The miRBase high-confidence rules[12]. As a meaningful evaluation of the second,
662 third, and fourth rule relies on accurate knowledge of the position and extent of the
663 mature miRNA on both strands of the pre-miRNA, we evaluated these three rules
664 only if the first rule was satisfied.

1.	≧ 10 tags on each arm of the pre-miRNA, or ≧ 100 tags on one arm of the pre-miRNA, with ≧ 5 tags on the other arm
2.	≧ 50% of the tags on each arm of the pre-miRNA have the same 5' end
3.	0–4 nt overhang at the mature 3' end on each arm
4.	≧ 60% of nucleotides of the mature sequence on each arm are base-paired
5.	$\Delta G < -0.2$ kcal/mole/nucleotide

665

666

Methods

Samples and library preparation

Short RNA libraries were prepared following the Illumina TruSeq Small RNA Sample Preparation protocol (catalog number RS-200-0012, RS-200-0024, RS-200-0036, RS-200-0048) using the same RNA samples from which CAGE libraries were produced previously[17,18], as well as one additional RNA sample without a matching CAGE library. RNA samples not previously described are listed in Table S2. TruSeq Small RNA Sample Prep Index Sequences were used as bar codes to allow pooling of multiple samples in one library. The short RNA libraries were sequenced using the Illumina HiSeq2000 sequencer in single-read, 50 base mode. The metadata of all FANTOM5 RNA samples, including those used for sRNA sequencing, are available in the FANTOM5 Semantic catalog of Samples, Transcription initiation And Regulators[45] (SSTAR; http://fantom.gsc.riken.jp/5/sstar). SSTAR sample pages also provide links to the FANTOM5 miRNA expression atlas web interface.

Data processing

We extracted the short RNA sequences from the raw sequences using in-house scripts. We removed linker artifact sequences using TagDust[46] version 1.13, ribosomal sequences using rRNAdust[17] version 1.00, and filtered against mature tRNAs, ribosomal RNA, and 7SL RNA using global alignment. We mapped the remaining sequences using the Burrows-Wheeler Alignment (bwa) tool[47] version 0.5.9-r16 to genome assembly hg19 (human) or mm9 (mouse), including chromosome Y if the donor was known to be male. Table S3 shows the number of short RNA sequences mapped to the genome for each sample. Two samples had fewer than 100,000 mapped tags and were discarded from the further analysis.

692

693 ***Short RNA annotation and filtering***

694 We used release 21 of the miRBase database[12], lifted over to genome assembly hg19
695 (human) or mm9 (mouse), as our reference set of known miRNAs. Four pre-miRNAs
696 in human that could not be lifted over to genome assembly hg19 and an additional six
697 human pre-miRNAs that were lifted over to unplaced chromosomes were excluded
698 from the analysis. We annotated all mapped short RNA reads mapping to genomic
699 loci for ribosomal RNA, tRNAs, the RNA component 7SL of the signal recognition
700 particle, small nuclear RNAs, small nucleolar RNAs, small Cajal body-specific RNA,
701 small cytoplasmic RNAs, and piRNAs. We corrected for cross-mapping as described
702 previously[48], discarding all mappings to unannotated loci if the short RNA sequence
703 could be mapped to an annotated locus instead.

704

705 ***Drosha CAGE peak analysis***

706 We calculated the total number of CAGE tags starting at each genomic position across
707 all 1,885 (human) and 1,202 (mouse) FANTOM5 CAGE libraries[17,18], as available at
708 http://fantom.gsc.riken.jp/5/datafiles/latest/basic/, as well as all 145 human ENCODE
709 CAGE data[22], which we downloaded from
710 http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeRikenCage/.
711 We defined the 3' end of the pre-miRNA as the 3' nucleotide of the mature miRNA
712 on the 3' arm of the pre-miRNA; the expected Drosha cleavage site is immediately
713 downstream of this nucleotide. For each pre-miRNA in the robust set, we calculated,
714 for each position with respect to the expected Drosha cleavage site, the total number
715 of CAGE tags in the CAGE libraries. We normalized by dividing by the sum over the
716 positions to obtain the CAGE profile with respect to the expected Drosha cleavage

site for each pre-miRNA. We then summed the CAGE profiles across the pre-miRNAs to obtain the average CAGE profile with respect to the expected Drosha cleavage site. Based on this profile, we selected a 9-basepair window between -2 and +7 base pairs with respect to the expected Drosha cleavage site for the FANTOM5 CAGE data, and an 8-basepair window between -2 and +6 base pairs for the ENCODE CAGE data, as the Drosha CAGE peak window for a given pre-miRNA.

For each pre-miRNA, we counted the number of CAGE tags with a 5' end within this window, as well as the number of CAGE tags with a 5' end anywhere between the pre-miRNA boundaries. Since CAGE tags tend to occur in clusters on the genome, we expect the distribution of the CAGE tag counts to be heavily overdispersed compared to the Poisson distribution. We therefore used the negative binomial distribution instead, with the dispersion parameter r estimated by fitting the distribution to the number of CAGE tags in any 8- or 9-basepair window on the human or mouse genome. This resulted in a dispersion of 1.856943 and 1.616542 for the FANTOM5 human and mouse CAGE data, respectively (using a 9-basepair window), and 0.325001 for the ENCODE CAGE data (using a 8-basepair window). Using these dispersion values, we calculated the statistical significance of the FANTOM5 and ENCODE CAGE peaks given the number k of CAGE tags within the window, the number K of CAGE tags within the pre-miRNA, the window size w, as well as the genomic extent L of the pre-miRNA as $I_p(k, r)$, where I is the regularized incomplete beta function and $p = \mu / (r + \mu)$, with $\mu = w K / L$ the expected number of tags at the Drosha CAGE peak under the null hypothesis.

For human, we calculated an overall statistical significance value by combining the FANTOM5 and ENCODE statistical significance into a single P-value using Fisher's method.

742 The relative occurrence of CAGE tags in different subcellular fractionations (Figure
743 S9) and the bias in the first nucleotide of CAGE tags (Figure S12) were evaluated for
744 pre-miRNAs in the robust set with a statistically significant Drosha CAGE peak.
745
746 ***Identification of candidate novel miRNA***
747 Candidate novel miRNAs were identified using miRDeep2[26], resulting in 7,461
748 (human) and 2,034 (mouse) predicted pre-miRNAs, including 918 (human) and 590
749 (mouse) known pre-miRNAs. To avoid predicted miRNAs from failing the miRBase
750 high-confidence rules due to flaws in the predicted secondary structure, we repeated
751 the secondary structure calculation for each predicted miRNA by applying RNAfold[49]
752 version 2.1.2 on the sequence of the precursor miRNA while constraining the
753 structure by allowing nucleotides in each arm of the hairpin to only base-pair to
754 nucleotides in the other arm of the hairpin. In the comparison of the candidate novel
755 miRNAs with the 3,524 tissue- and primate-specific miRNAs published recently[27], we
756 required the pre-miRNAs to overlap by at least 80%.
757
758 ***Validation of candidate novel miRNA expression by qPCR***
759 Fresh buffy coat was obtained from the Red Cross following approval from the human
760 research ethics committee of The University of Melbourne (ethics ID 1646608.1) and
761 material supply agreement with Red Cross (16-05VIC-21).
762 Peripheral blood mononuclear cells (PBMC) were isolated from buffy coat using
763 Ficoll hypaque (GE Healthcare, Uppsala, Sweden) as described previously[50]. CD14+
764 human monocytes were isolated from PBMC using human CD14+ magnetic beads
765 (Milteny Biotec, Sydney, New South Wales, Australia). CD14+ monocytes were
766 differentiated to macrophages in complete RPMI1640 media supplemented with 10%

767 fetal calf serum and 100 ng ml^{-1} human macrophage colony-stimulating factor (M-
768 CSF) (PeproTech, Rehovot, Israel) for 5 days. Suspended cells were removed and
769 adherent cells were washed with PBS before macrophages were collected.
770 MicroRNAs were isolated from monocytes and macrophages using mirVana miRNA
771 Isolation Kit (Life Technologies, Melbourne, Victoria, Australia) following the
772 manufacturer's protocol. Briefly, cells were lysed in lysis buffer followed by phenol
773 extraction, and miRNAs were isolated from the phenol aqueous phase using a spin
774 column followed by elution in RNase-free water. Following manufacturer's protocol,
775 cDNA synthesis was performed using miScript PCR Starter Kit (Qiagen, Hilden,
776 Germany) by ligating a poly(A) tail to the miRNA followed by reverse transcription
777 in the presence of universal tag. Samples without reverse transcriptase but with all
778 other components were included and used as negative controls.
779 Forward primers specific to the candidate novel miRNAs were designed using
780 miRprimer[51] (Table S10). Real-time PCR was performed using miScript PCR Starter
781 Kit (Qiagen, Hilden, Germany) and following the manufacturer's protocol. The PCR
782 reaction was set up with the custom-made forward primers and the universal reverse
783 primer supplied with the kit. No-template controls and cDNA samples without reverse
784 transcriptase were included as negative controls. Thermal cycling was performed as
785 suggested by the manufacturer's protocol.
786 The expression levels of a wide range of miRNAs have been analyzed using our
787 miRNA PCR assay in order to evaluate the sensitivity of the assay and determine the
788 confidence of our results. Short RNAs commonly used as a reference, including
789 RNU6 and let-7a-5p (ref. 52), showed relatively high expression levels. Other
790 miRNAs that are highly conserved in metazoans or known to be expressed in myeloid
791 cells, including miR-191-5p (ref. 53), miR-15a-5p (ref. 54), miR-206 (ref. 55), miR-

335-5p (ref. 56) and miR-339-3p (ref. 56), were included and used as positive controls, and showed moderate expression levels. Expression levels of miRNAs reported to be cell markers for other cell types and assumed to be lowly expressed in myeloid cells, including miR-153-3p (ref. 57) and miR-345-5p (ref. 58), were also analyzed in order to determine the detection limit of the assay. Our results demonstrate that the miRNA PCR assay could specifically detect the presence of the target miRNAs, and measure a wide spectrum of expression levels. The expression levels of the selected candidate novel miRNAs fell within the detection spectrum of our miRNA PCR assay, proving the reliability of our results.

Evaluation of miRNA expression normalization strategies

We counted the number of short RNA sequences with a length between 18 and 25 nucleotides overlapping the mature miRNA loci in each of the primary cell samples. We then applied the following normalization strategies:

- CPM (counts per million): Divide the count by the sum of counts for mature miRNAs in the robust set, and multiply by 1,000,000;
- TMM (trimmed mean of M values): Apply the "calcNormFactors" function in edgeR[59] with method "TMM" to the table of counts;
- RLE (relative log expression): Apply the "calcNormFactors" function in edgeR[59] with method "RLE" to the table of counts;
- DESeq (effective library size): Apply the "estimateSizeFactorsForMatrix" function in DESeq[60] to the table of counts;
- UQ (upper quantile normalization): Divide the count by the sum of the counts of the top-25% most abundant miRNAs in each sample;

- UD (upper decile normalization): Divide the count by the sum of the counts of the top-10% most abundant miRNAs in each sample.

To evaluate each normalization strategy, we divided the primary cell samples in FANTOM5 into groups (n = 96) of independent donors of the same cell type. For each cell type group, we calculated the variance for each miRNA across the donors. To find the error between different cell types, we first calculated the average expression for each miRNA across donors in each cell type group, and then calculated the difference in the average expression between each pair of cell type groups ($n = \frac{1}{2} \times 96 \times 95 = 4{,}560$) for each miRNA. To evaluate the total error, we calculated the mean squared error across miRNAs for each cell type group, as well as the mean squared error across miRNAs for each pair of cell type groups, and took the square root of each to find the root mean square (RMS) with cell type groups and between cell type groups (Figure S16a). We averaged the RMS error over the n = 96 cell type groups, and over the n = 4,560 pairs of cell type groups, and calculated the ratio of the average RMS error within cell types to the average RMS error between cell types (Figure S16b). To evaluate the standard error (Figure S16c), we calculated the mean square error across cell type groups for each miRNA, as well as the mean squared error across pairs of cell type groups for each miRNA, took the square root, and plotted the resulting RMS value for each miRNA against its mean expression level. We then used linear regression to calculate the slope of the RMS error within each cell type and between different cell types as a function of the miRNA expression level. Dividing these two slopes yielded the ratio in RMS error within cell types and between different cell types, normalized by miRNA expression level (Figure S16d).

Clustering and visualisation of miRNA expression patterns

MicroRNAs were clustered based on their expression patterns using the network visualisation and analysis tool Miru[28] (http://kajeka.com/miru/miru-about/). The Pearson correlation was calculated for each pair of miRNAs. A modified Fruchterman-Rheingold algorithm was used to lay out the network graph in 3-dimensional space, in which 502 nodes representing miRNAs were connected by 3,369 weighted, undirected edges representing correlations of at least 0.6 between expression patterns. Areas of high connectively and correlation, representing groups of miRNAs with similar expression profiles, were identified using the Markov clustering algorithm (MCL) with an MCL inflation value of 2.2. Clusters were manually annotated based on the cell type or tissue of greatest expression. All nodes in one cluster and the label describing the cluster are shown in the same color. The smallest labeled cluster contains six nodes; for clarity, smaller clusters have not been labeled but can be identified by groups of nodes of the same color.

Validation of miRNA expression quantitation by qPCR

Expression of selected miRNAs was measured using the TaqMan® MicroRNA Assay (Applied Biosystems) according to its protocol. RNA samples 11544, 11624, 11705 (CD19+ B cells, donor1, 2, and 3), 11269, 11346, 11418 (dermal fibroblast donor1, 2, and 3), 12626 (H9 embryonic stem cells), and 11523, 11603, 11684 (hepatocyte donor1, 2, and 3) (Table S3) were used after confirmation of the RNA quality by measuring the RIN value using a TapeStation and the 260/280 and 260/230 ratios using NanoDrop. The Ct values obtained were normalized against the Ct value of small nucleolar RNA SNORD48.

Cell type specificity index

Following the definition of the tissue specificity index (TSI) [29], we define the cell type specificity index of miRNA *j* as

$$\text{index}_j = \frac{1}{N-1}\sum_{i=1}^{N}\left(1-\frac{x_{j,i}}{\max_{i'} x_{j,i'}}\right)$$

where N is the number of primary cell types in FANTOM5, and $x_{j,i}$ is the expression in counts-per-million of miRNA *j* in cell type *i*, averaged over independent donors.

Guide strand selection

For each pre-miRNA, we designated the hairpin arm with the highest expression level (in counts-per-million) in any of the FANTOM5 samples as the guide strand, and refer to the opposite arm as the passenger strand.

Cell ontology analysis

We used the FANTOM5 cell ontology[30,31] to create cell ontology clusters (Tables S14 and S20). We performed a likelihood-ratio test comparing the expression data between the samples in each cell ontology cluster and the background, consisting of all other samples listed in Tables S14 and S20, modeling the tag counts by a negative binomial distribution. For each miRNA, we selected the three cell ontology terms for which the expression in the cell ontology cluster compared to the background was statistically most significantly higher, and the three cell ontology terms for which the expression in the cell ontology cluster compared to the background was statistically most significantly lower. The *P*-values listed in Tables S13, S15, and S17 for each miRNA for specific cell ontology clusters were not corrected for multiple testing.

Identification of miRNA promoters

Candidate pri-miRNAs consisted of transcripts annotated in Gencode[61] v19 (human) or vM5 (mouse) or in the NCBI Entrez Gene database[62]. For each pre-miRNA, we selected all candidate pri- miRNAs with a transcription start site upstream of the pre-miRNA and a 3' end downstream of the 5' end of the pre-miRNA, and defined all FANTOM5 permissive CAGE peaks[17] within the genomic region from 500 bp upstream of the 5' end of the pri-miRNA to the 5' end of the pre-miRNA as the set of candidate promoters associated with the pre-miRNA. We averaged the expression level (in tags-per-million) of each candidate promoter across all FANTOM5 CAGE samples, and selected the candidate promoter with the highest average expression level as the (computationally predicted) promoter of the miRNA. Each human miRNA in the robust set was manually curated by two annotators.

Validation of miRNA promoters by RAMPAGE

We downloaded all 212 BAM files containing ENCODE RAMPAGE sequencing data mapped to human genome assembly hg19 that were not marked as "low read depth" or "low replicate concordance". We retained the 5' end positions of RAMPAGE transcripts with a 3' end within 1,000 basepairs of a pre-miRNA locus, discarding 5' end positions supported by fewer than 5 RAMPAGE transcripts, and associated the remaining 5' end positions with the pre-miRNA as putative transcription start sites.

Validation of miRNA promoters by RACE

We mixed 4.0 µl 5X First-Strand Buffer, 0.5 µl DTT (100 mM; Invitrogen, catalog number 70726), 1.0 µl dNTP Mix (20 mM), spun briefly in a microcentrifuge, and kept at room temperature. We combined 1.0-10.0 µl with 1 µg total RNA from monocytes, macrophages, and dendritic cells, 1.0 µl Random Primer Mix (N-15) (20

μM), and 0-9 μl sterile water to reach a total volume of 11.0 μl in separate microcentrifuge tubes, mixed the contents and spun the tubes briefly. We incubated these tubes at 72 °C for 3 minutes, and then cooled to 42 °C for 2 minutes. After cooling, we spun the tubes for 10 seconds at 14,000 g to collect the contents at the bottom. Next, we added 1.0 μl of Smarter oligo (20 μM) per reaction, and mixed well by vortexing and spun the tube briefly in a microcentrifuge. We then added 0.5 μl RNase Inhibitor (40 U/μl; Invitrogen RNaseOUT™, catalog number 10777019) and 2.0 μl SMARTScribe Reverse Transcriptase (100 U; Clontech, catalog number 639537) to the buffer mix, and mixed these reagents at room temperature. Next, we added 8.0 μl of the master mix to the RNA solution, mixed the contents of the tubes by gently pipetting, and spun the tubes briefly. We incubated the tubes at 42 °C for 90 minutes and heated the tubes at 70 °C for 10 minutes in a hot-lid thermal cycler. We then added 90 μl Tricine-EDTA buffer to each tube.

We prepared the master mix for the first PCR by combining 2.5 μl of the cDNA solution, 5.0 μl 10X Advantage 2 PCR buffer (Clontech, catalog number 639207), 1.0 μl dNTP Mix (10mM each) 50X Advantage 2 Polymerase Mix (Clontech), 1.0 μl of the smarterRACE_forward primer at 10 pmol/μl, 1.0 μl of the miRNA-specific outer primer (Table S18) at 10 pmol/μl, and added PCR-grade water to reach a volume of 50 μl. We ran a 2-step PCR program consisting of 1 minute at 95 °C, 25 cycles of 30 seconds at 95 °C followed by 70 seconds at 68 °C, 7 minutes at 68 °C, and finishing at 8 °C. We diluted 5 μl of the primary PCR product into 245 μl of Tricine-EDTA buffer.

We prepared the master mix for the second PCR by combining 5.0 μl of the product of the first PCR after dilution with 5 μl of the 10X Advantage 2 PCR buffer, 1.0 μl dNTP Mix (10 mM), 1.0 μl of 50X Advantage 2 Polymerase Mix (Clontech), 2.0 μl of

940 the Nextera_i7 primer, 2.0 µl of the miRNA-specific inner primer (Table S18), and 34
941 µl of PCR-grade water. We ran a 2-step PCR program consisting of 1 minute at 95 °C,
942 20 cycles of 30 seconds at 95 °C followed by 70 seconds at 68 °C, 7 minutes at 68 °C,
943 and finishing at 8 °C. We diluted 5 µl of the PCR product into 245 µl of Tricine-
944 EDTA buffer.
945 We prepared the master mix for the third PCR by combining 5.0 µl of the PCR
946 product of the second PCR with 5.0 µl of 10X Advantage 2 PCR buffer, 1.0 µl dNTP
947 mix (10 mM), 1.0 µl of 40X Advantage 2 Polymerase Mix (Clontech), 2.0 µl of the
948 Nextera_i7 primer, 2.0 µl of the Nextera_i5 primer, and 34 µl of PCR-grade water.
949 We purified by AMPure at a 1.8 ratio, checked 2 µl of the second PCR product on a
950 TapeStation, kept the libraries at -20 °C until sequencing, and pooled the PCR
951 products, each with a different barcode combination before paired-end sequencing on
952 a MiSeq sequencer (Illumina).
953 We mapped the sequencing data to the human genome using Blat[63], merged each pair
954 into a single mapped transcript, and retained transcripts that overlap an inner primer.
955 The histograms in Figure S21 show the position of the 5' end of these transcripts.
956
957 ***Promoter sequence conservation analysis***
958 We previously compiled a list of transcription factors in human and mouse[17]. Protein-
959 coding genes and lncRNAs consisted of all other genes annotated in the NCBI Entrez
960 Gene database[62] as protein coding or miscRNA, respectively. For each gene in these
961 three categories, we selected the associated p1 CAGE peak as defined previously[17] as
962 the gene promoter, and discarded all genes without an associated CAGE peak. We
963 then found the phastCons conservation score[36], obtained from the UCSC Genome
964 Browser database[64], for the alignment of 99 vertebrate organisms against the human

genome hg19, as a function of position relative to the transcription start site for each gene and miRNA, and averaged these scores for each category at each position.

Construction of the FANTOM5 miRNA expression atlas of miRNAs

CAGE tag start site (CTSS) files[31], excluding universal and whole body RNA samples, were downloaded from http://fantom.gsc.riken.jp/5/datafiles/latest/basic/. CAGE tag counts for technical replicates of the same RNA sample were summed for each genomic position. CAGE libraries published by Fort *et al.*[19] were downloaded from DDBJ, accession DRA000914. The number of CAGE tags at each genome position were counted to generate CTSS files, and pri-miRNA expression tables were generated by summing the CAGE tags under each promoter, calculating the total number of tags mapped to the genome, and using this number to normalize to tags per million (tpm).

Mature miRNA expression tables were generated by counting the number of sRNA tags to each miRNA locus, calculating the total number of tags mapping to the robust miRNAs, and using this number to normalize to counts per million (cpm).

The CAGE and sRNA expression tables are available for download at the miRNA expression viewer at

http://fantom.gsc.riken.jp/5/suppl/De_Rie_et_al_2017/

To generate the heatmap, we averaged the cpm-normalized expression values of each miRNA across donors for each cell type, and converted the expression profile of each miRNA to Z-scores by subtracting the mean and dividing by the standard deviation across cell types. The heatmap was sorted both for cell types and for miRNAs by performed centroid-linkage hierarchical clustering, using the Pearson correlation as the similarity measure.

990

991 ***Motif activity response analysis (MARA)***

992 The genome-wide predictions of transcription factor binding sites were produced as
993 described previously[18]. Briefly, we downloaded the whole-genome alignment of the
994 human genome hg19 against 99 other vertebrate genomes, and of the mouse genome
995 mm9 against 29 other vertebrate genomes, from the UCSC Genome Browser
996 database[64], and extracted the multiple alignments of human, macaque, mouse, rat,
997 cow, horse, dog, opossum, and chicken. We divided the genome into segments and
998 realigned each segment using T-Coffee[65], and generated genome-wide transcription
999 factor binding site (TFBS) predictions using MotEvo[66] for the SwissRegulon set of
1000 position-weight matrix motifs[67] (Figure S29a). We then counted the number of
1001 predicted TFBSs for each motif in the –300 to +100 base pair base proximal promoter
1002 regions of genes in the NCBI Entrez Gene database[62], excluding all miRNA
1003 promoters (Figure S29b). Next, we used MARA[38] to decompose the FANTOM5
1004 CAGE expression profiles of these promoters in terms of their associated motifs,
1005 yielding the activity profile of each motif across the FANTOM primary samples
1006 (Figure S29c). We then counted the number of TFBSs for each motif in the –300 to +
1007 100 base pair base proximal promoter region of each miRNA (Figure S29d), and
1008 predicted the miRNA expression level by calculating the weighted sum of the
1009 activities for motifs found (Figure S29e). We compared the predicted expression
1010 levels to the expression levels of the mature miRNA observed in the FANTOM5
1011 sRNA sequencing data (Figure S29f) and calculated their correlation (Figure S29g) as
1012 a measure of the accuracy of the miRNA promoter identification. Following the
1013 MARA procedure[38], we normalized the cpm expression values of miRNAs by adding
1014 0.5, taking the base-2 logarithm, subtracting the mean across samples, and finally

subtracting the mean across miRNAs. We defined strongly differentially expressed miRNAs, included in Figure S20d and S20e, as those that had a standard deviation in expression, after normalization, across samples larger than 2.

Data availability

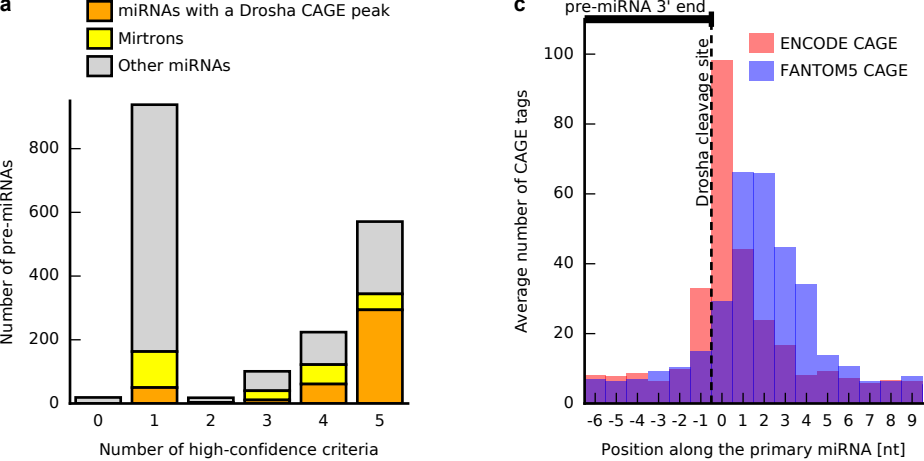
Raw sequencing data of the sRNA libraries are available at the DNA Data Bank of Japan (DDBJ; http://www.ddbj.nig.ac.jp/) under accession numbers DRA001101, DRA002711, DRA003804, and DRA003807, and for the RACE experiments at the NCBI Gene Expression Omnibus (NCBI GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE98695.

Supplemental references

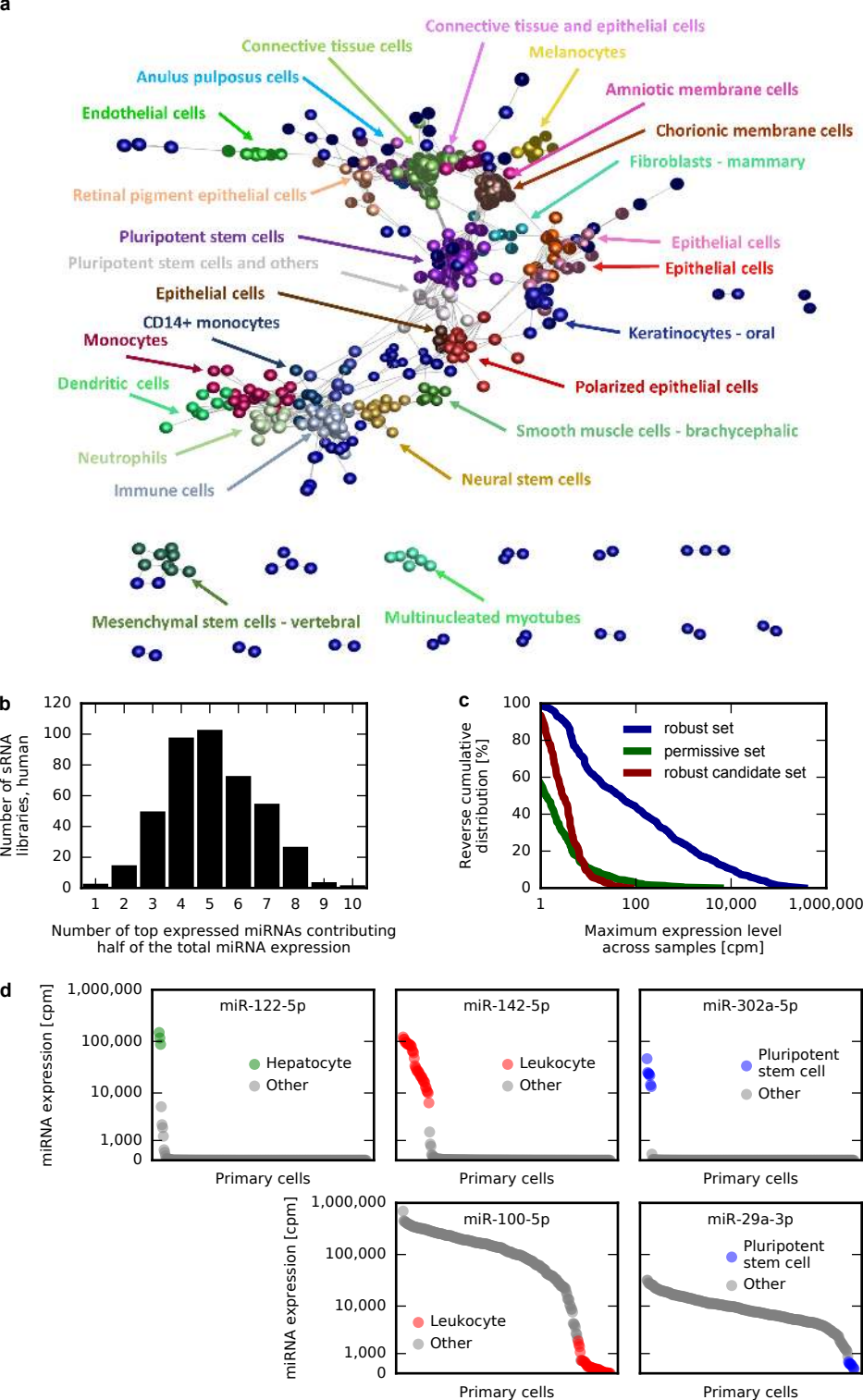
45. Abugessaisa, I., *et al*. FANTOM5 transcriptome catalog of cellular states based on Semantic MediaWiki. *Database (Oxford)* **2016**, baw105 (2016).
46. Lassmann, T., Hayashizaki, Y. & Daub, C.O. TagDust—a program to eliminate artifacts from next generation sequencing data. *Bioinformatics* **25**, 2839–2840 (2009).
47. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
48. De Hoon, M.J.L. *et al.* Cross-mapping and the identification of editing sites in mature microRNAs in high-throughput sequencing libraries. *Genome Res.* **20**, 257–264 (2010).
49. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).

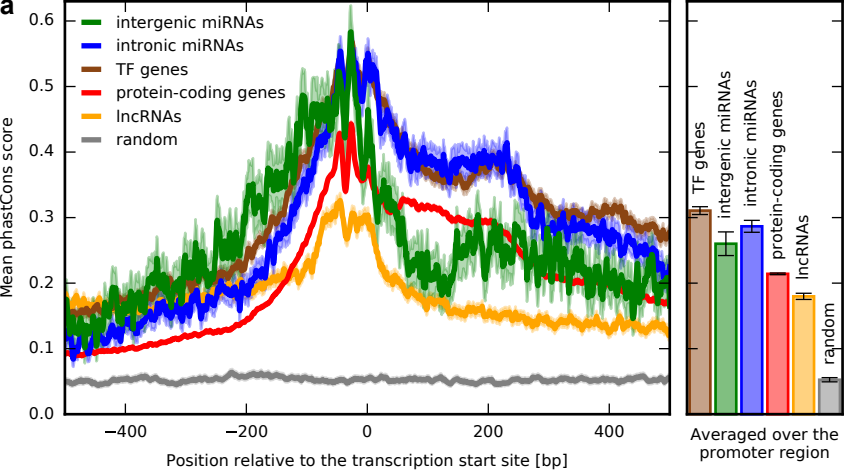
50. Vijayan, D., Radford, K.J., Beckhouse, A.G., Ashman, R.B. & Wells, C.A. Mincle polarizes human monocyte and neutrophil responses to *Candida albicans*. *Immunol. Cell Biol.* **90**, 889–895 (2012).
51. Busk, P.K. A tool for design of primers for microRNA-specific quantitative RT-qPCR. *BMC Bioinformatics* **15**, 29 (2014).
52. Schwarzenbach, H., Da Silva, A.M., Calin, G. & Pantel, K. Data normalization strategies for microRNA quantification. *Clin. Chem.* **61**, 1333–1342 (2015).
53. Nagpal, N. & Kulshreshtha, R. miR-191: an emerging player in disease biology. *Front. Genet.* **5**, 99 (2014).
54. Moon, H.G., Yang, J., Zheng, Y. & Jin, Y. miR-15a/16 regulates macrophage phagocytosis after bacterial infection. *J. Immunol*. **193**, 4558–4567 (2014).
55. Vinod, M., *et al.* miR-206 controls LXRα expression and promotes LXR-mediated cholesterol efflux in macrophages. *Biochim. Biophys. Acta* **1841**, 827–835 (2014).
56. Cobos Jiménez, V., *et al.* Next-generation sequencing of microRNAs in primary human polarized macrophages. *Genom. Data* **2**, 181–183 (2014).
57. Zhang, L., *et al.* miR-153 supports colorectal cancer progression via pleiotropic effects that enhance invasion and chemotherapeutic resistance. *Cancer Res*. **73**, 6435–6447 (2013).
58. Srivastava, S.K., *et al.* MicroRNA-345 induces apoptosis in pancreatic cancer cells through potentiation of caspase-dependent and -independent pathways. *Br. J. Cancer* **113**, 660–668 (2015).

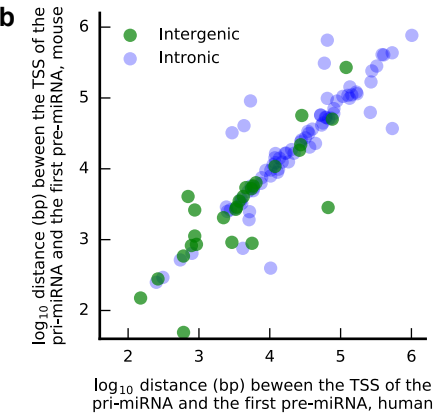
59. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
60. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
61. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
62. Brown, G.R. *et al*. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.* **43**, D36–D42 (2015).
63. Kent, W.J. BLAT—The BLAST-like Alignment Tool. *Genome Res.* **12**, 656–664 (2002).
64. Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.* **42**, D764–D770 (2014).
65. Notredame, C., Higgins, D.G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).
66. Arnold, P., Erb, I., Pachkov, M., Molina, N. & Van Nimwegen, E. MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. *Bioinformatics* **28**, 487–494 (2012).
67. Pachkov, M., Balwierz, P.J., Arnold, P., Ozonov, E. & Van Nimwegen, E. SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Res.* **41**, D214–D220 (2013).

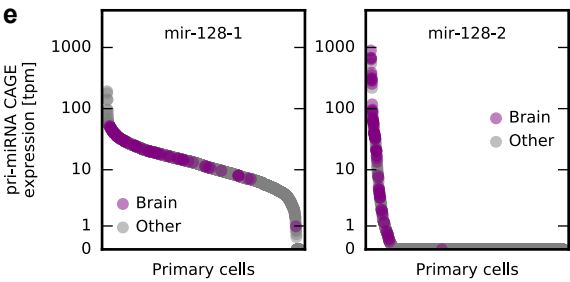

a
miRNAs with a Drosha CAGE peak
Mirtrons
Other miRNAs
Number of pre-miRNAs
800
600
400
200
0
0
1
2
3
4
5
Number of high-confidence criteria
c
pre-miRNA 3' end
Drosha cleavage site
ENCODE CAGE
FANTOM5 CAGE
Average number of CAGE tags
100
80
60
40
20
0
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9
Position along the primary miRNA [nt]

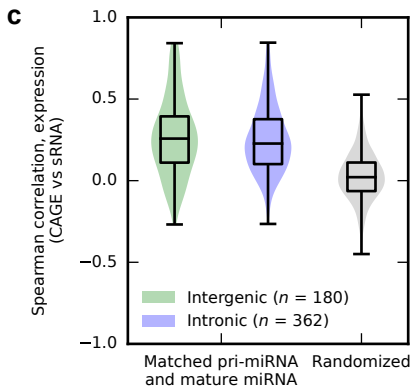

b
chrX
65238750
65238800
ACGCTCCGTGTATTTGACAAGCTGAGTTGGACACTCCATGTGGTAGAGTGTCAGTTTGTCAAATACCCCAAGTGCGGC
pri-mir-223
miRBase release 20
miR-223-5p (passenger strand)
miR-223-3p (mature miRNA)
FANTOM4 full-length pre-miRNAs
pre-mir-223
Number of CAGE tags:
ENCODE CAGE (maximum: 6848 tags)
5000
Forward
0
Reverse
5000
Forward strand
Reverse strand
FANTOM5 CAGE (maximum: 403 tags)
250
Forward
0
Reverse
250
Forward strand
Reverse strand
Drosha cleavage site
Number of sRNA reads:
FANTOM5 sRNA (forward strand only)
8 million
6 million
4 million
2 million
0

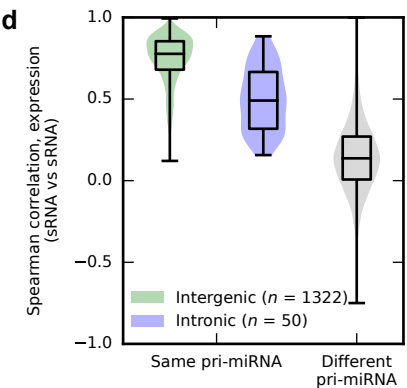

a
Connective tissue and epithelial cells
Connective tissue cells
Melanocytes
Anulus pulposus cells
Amniotic membrane cells
Endothelial cells
Chorionic membrane cells
Fibroblasts - mammary
Retinal pigment epithelial cells
Pluripotent stem cells
Epithelial cells
Pluripotent stem cells and others
Epithelial cells
Epithelial cells
CD14+ monocytes
Keratinocytes - oral
Monocytes
Dendritic cells
Polarized epithelial cells
Smooth muscle cells - brachycephalic
Neutrophils
Neural stem cells
Immune cells
Mesenchymal stem cells - vertebral
Multinucleated myotubes
b
Number of sRNA libraries, human
Number of top expressed miRNAs contributing half of the total miRNA expression
c
Reverse cumulative distribution [%]
robust set
permissive set
robust candidate set
Maximum expression level across samples [cpm]
d
miRNA expression [cpm]
miR-122-5p
Hepatocyte
Other
miR-142-5p
Leukocyte
Other
miR-302a-5p
Pluripotent stem cell
Other
miR-100-5p
Leukocyte
Other
miR-29a-3p
Pluripotent stem cell
Other
Primary cells


a
intergenic miRNAs
intronic miRNAs
TF genes
protein-coding genes
lncRNAs
random
Mean phastCons score
Position relative to the transcription start site [bp]
TF genes
intergenic miRNAs
intronic miRNAs
protein-coding genes
lncRNAs
random
Averaged over the promoter region


b
Intergenic
Intronic
log10 distance (bp) beween the TSS of the pri-miRNA and the first pre-miRNA, mouse
log10 distance (bp) beween the TSS of the pri-miRNA and the first pre-miRNA, human


e
mir-128-1
mir-128-2
pri-miRNA CAGE expression [tpm]
Brain
Other
Primary cells


c
Spearman correlation, expression (CAGE vs sRNA)
Intergenic (n = 180)
Intronic (n = 362)
Matched pri-miRNA and mature miRNA
Randomized


d
Spearman correlation, expression (sRNA vs sRNA)
Intergenic (n = 1322)
Intronic (n = 50)
Same pri-miRNA
Different pri-miRNA