SCIENTIFIC DATA

OPEN

Data Descriptor: Small non-coding RNA transcriptome of the NCI-60 cell line panel

Erin A. Marshall[1], Adam P. Sage[1], Kevin W. Ng[1], Victor D. Martinez[1], Natalie S. Firmino[1], Kevin L. Bennewith[1] & Wan L. Lam[1]

Received: 30 June 2017
Accepted: 5 September 2017
Published: 24 October 2017

Only 3% of the transcribed human genome is translated into protein, and small non-coding RNAs from these untranslated regions have demonstrated critical roles in transcriptional and translational regulation of proteins. Here, we provide a resource that will facilitate cell line selection for gene expression studies involving sncRNAs in cancer research. As the most accessible and tractable models of tumours, cancer cell lines are widely used to study cancer development and progression. The NCI-60 panel of 59 cancer cell lines was curated to provide common models for drug screening in 9 tissue types; however, its prominence has extended to use in gene regulation, xenograft models, and beyond. Here, we present the complete small non-coding RNA (sncRNA) transcriptomes of these 59 cancer cell lines. Additionally, we examine the abundance and unique sequences of annotated microRNAs (miRNAs), PIWI-interacting RNAs (piRNAs), small nuclear RNAs (snRNAs), and small nucleolar RNAs (snoRNAs), and reveal novel unannotated microRNA sequences.

Design Type(s)	cell type comparison design • sequence analysis objective • transcription profiling identification objective
Measurement Type(s)	non-protein coding RNA sequence
Technology Type(s)	RNA sequencing
Factor Type(s)	cancer cell line
Sample Characteristic(s)	Homo sapiens • breast cancer cell line • glioblastoma cell line • glioma cell line • colonic adenocarcinoma cell line • colonic cancer cell line • lymphoblastic leukemia cell line • acute myeloid leukemia cell line • chronic myeloid leukemia cell line • multiple myeloma cell line • non-Hodgkin lymphoma cell line • melanoma cell line • mammary gland tumor cell line • non-small cell lung cancer cell line • large cell lung cancer cell line • ovary cancer cell line • prostate cancer cell line • renal cancer cell line

[1]Department of Integrative Oncology, British Columbia Cancer Research Centre, Vancouver, British Columbia, Canada V5Z 1L3. Correspondence and requests for materials should be addressed to E.A.M. (email: emarshall@bccrc.ca).

Q1

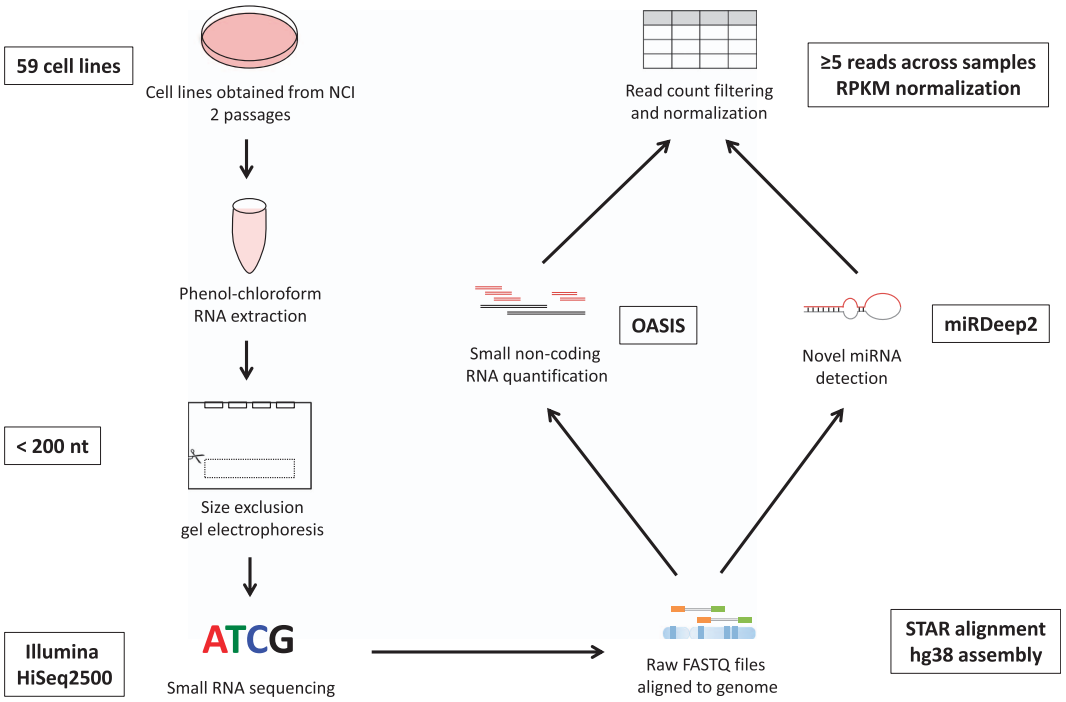

Figure 1. Experimental workflow. Graphical representation of experimental procedure used to extract, process, and analyze RNA from cell lines.

Background & Summary

The NCI-60 Human Tumour Cell Lines Screen is an initiative started by the National Institutes of Health (NIH) in the late 1980s, focusing on the development of 59 human tumour cell lines for use as an *in vitro* drug screen model[1–3] (Table 1 (available online only)). These cell lines, derived from nine solid and blood malignancies, have shown great utility both in its original purpose for therapeutic screening as well as in basic cancer research (reviewed by Shoemaker *et al.*[4]). They have since been extensively characterized for various molecular features, including karyotypic complexity[1], DNA fingerprinting[2], gene expression microarray profiling[5,6], and human leukocyte antigen typing[3]. However, the small non-coding RNA (sncRNA) transcriptomes of the NCI-60 cell lines have yet to have been reported at the sequencing level.

The advent of next-generation sequencing has revealed the large proportion of non-coding genes in the human genome, and the relevance of these non-coding species in regulating the expression of both neighbouring and distant protein-coding genes. In the context of cancer, microRNAs (miRNAs) remain the best-studied non-coding RNA species, and have been implicated in all stages of cancer: initiation, progression, and response to therapy (reviewed by Hayes *et al.*[7]). Recent advances in the bioinformatic tools used for the discovery of small non-coding RNA have considerably expanded the number of known miRNA sequences[8]. Other types of sncRNA, including PIWI-interacting RNAs (piRNAs), small nuclear RNAs (snRNAs), and small nucleolar RNAs (snoRNAs) are emerging topics in cancer biology (reviewed by Ng *et al.* and Mannoor *et al.*[9,10]). Beyond their functions in gene regulation, sncRNAs are attractive prognostic biomarkers due to their abundance and stability in various biofluids[11].

We sequenced the sncRNA transcriptomes of the 59 cell lines in the panel (Fig. 1). SncRNA profiles were generated using the OASIS analysis platform v2.0 (ref. 12). For known sncRNA species (miRNAs, piRNAs, snoRNA, snRNA, and rRNA), high quality reads were mapped to the hg38 build of the human genome and quantified based on annotations containing their specific chromosomal locations. Detection of novel miRNAs was performed using well-established prediction algorithms that assess reads for miRNA folding characteristics, among other factors that indicate the probability that the tested sequence belongs to the miRNA family of sncRNAs[13]. In total, the genomic loci of 49,961 sncRNAs were examined. Using a detection threshold of greater than or equal to 5 reads across all tissues, we detected a total of 24,621 unique sncRNAs [Data Citation 1].

We then examined the genomic distribution of the detected sncRNAs across all tissue types (Table 2, Fig. 2). Notably, sncRNAs are expressed across all chromosomes in every tissue type assessed. SncRNA loci commonly expressed among all tissues may indicate their involvement in preserved biological or cancer-relevant processes, whereas differences in expression may denote tissue specificity.

We also examined the relative frequency of detection for each sncRNA species, both in the entire NCI-60 cell line panel and in lines grouped by organ type (Fig. 3a). Beyond those annotated in miRBase (v.21), novel unannotated miRNAs were determined by integrating secondary structure formation

	Number of sncRNA							Sequencing details				
	Total	miRNA	novel miRNA	piRNA	snoRNA	snRNA	Other	Average number of reads per sample	Average contig length	Avg. quality	%GC	Average coverage
Tissue type	24,794	2,509	288	19,018	259	1,602	412	23,457,233	22.34	33.25	46.26%	28.94
Breast	11,079	1,905	183	6,909	239	1,248	403	25,310,156	22.32	33.23	45.81%	25.36
CNS	10,120	1,793	180	6,150	236	1,175	397	25,633,608	22.35	33.34	44.57%	45.08
Colon	13,985	1,977	211	8,050	232	1,276	392	25,850,349	22.22	33.2	46.15%	22.97
Leukemia	10,728	1,841	185	5,604	228	1,112	387	18,921,562	22.41	33.25	48.77%	20.23
Melanoma	9,536	1,830	179	4,694	224	1,020	383	16,116,671	22.36	32.64	46.56%	28.45
NSCLC	15,707	2,051	232	9,385	236	1,293	398	23,407,812	22.34	33.3	46.18%	22.8
Ovarian	10,422	1,916	188	6,312	238	1,167	398	27,221,870	22.4	34.12	46.56%	41.49
Prostate	6,167	1,393	121	2,589	216	771	344	35,637,053	22.3	33.54	44.44%	30.11
Renal	14,532	1,943	209	8,549	234	1,288	396	23,949,373	22.3	33	45.99%	27.05

Table 2. Average number of sncRNA species detected and sequencing coverage per tissue type.

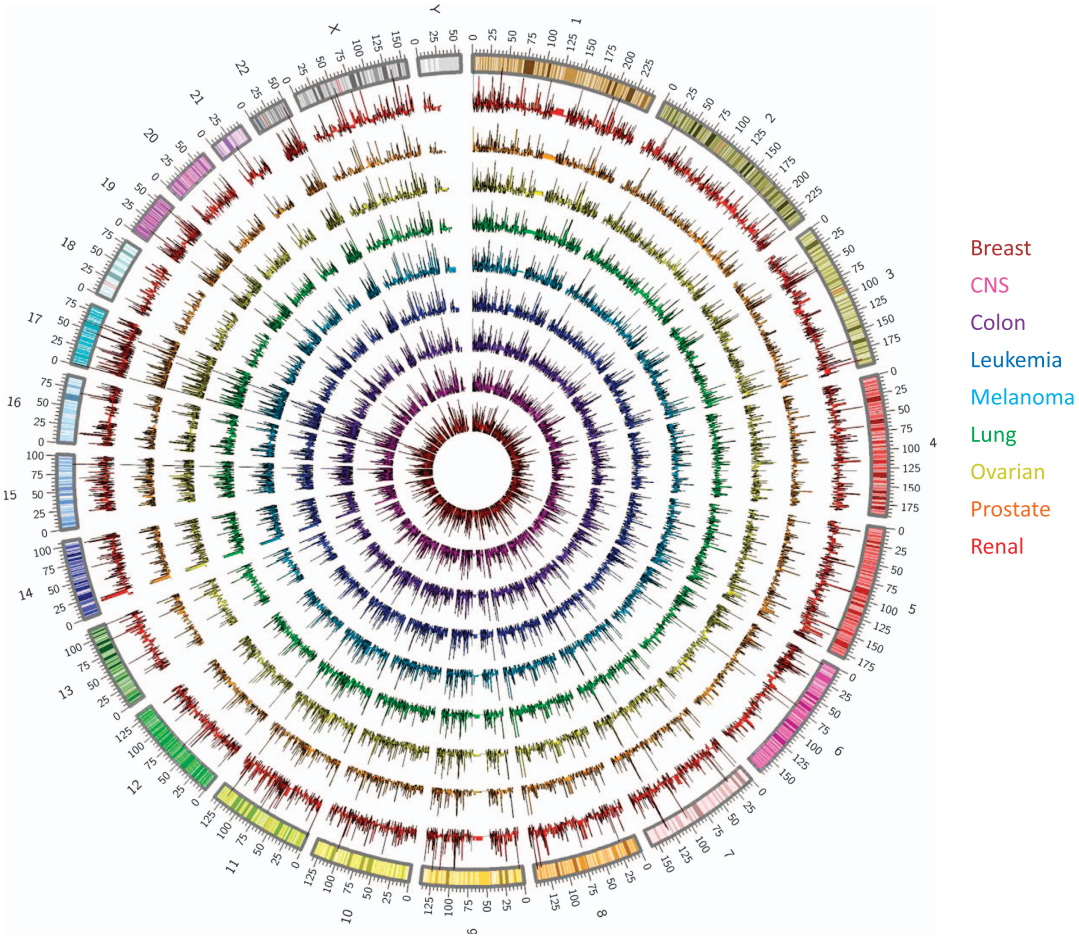

Figure 2. Genome-wide distribution of expressed small non-coding RNA by tissue type. Genomic position of sncRNAs detected (reads ≥ 5) in each tissue type in reference to the hg38 chromosome build karyotype. From inner-most ring to outer: breast (red), CNS (magenta), colon (purple), leukemia (blue), melanoma (teal), lung (green), ovarian (yellow), prostate (orange), and renal (red).

potential with free energy scoring[14]. These novel miRNAs represent an increase of approximately 10% of total miRNAs expressed across all tissue types (Fig. 3b), highlighting the constant expansion of the known non-coding transcriptome as sequencing technologies and bioinformatic tools advance.

Consistent with the number of annotated loci in the human genome, piRNAs represent the largest proportion of sncRNA species expressed, followed by miRNA and snRNA (Fig. 3a). Of note, an appreciable number of tissue-specific piRNA sequences across all tissues analyzed increased the relative fraction of piRNAs for all tissues expressed (Fig. 3c,d). Thus, as parts of the small non-coding RNA transcriptome are significantly understudied, we provide this resource to the research community for studying sncRNA-related genetic and epigenetic regulation in cancer using the NCI-60 cell models.

Methods

Cell line and sequencing information

Cell line doubling times were obtained directly from the National Institutes of Health NCI (https://dtp.cancer.gov/discovery_development/nci-60/cell_list.htm), and year-of-origin information refers to data of first publication containing the cell line (Table 1 (available online only)). Cell lines were obtained directly from the National Cancer Institute (NCI), were thawed and passaged twice precisely before total RNA was manually extracted using phenol-chloroform protocols from all cell lines using Trizol reagent (Invitrogen, CA, USA). 5,000 ng of extracted RNA per sample was used for sequencing input. Sequencing was performed in accordance with The Cancer Genome Atlas miRNA sequencing protocol (described by Chu *et al.*[15]). Briefly, after ligation to adaptors, 15 cycles of PCR was performed for amplification (98 °C-15 s, 62 °C-30 s and 72 °C-15 s), followed by 5 min at 72 °C. Small RNA exclusion was performed using gel extraction on a 3% MetaPhor Agarose gel (Lonza Inc., Basel, Switzerland), selecting species shorter that 200 nucleotides in order to enrich for targets optimized at 22 nucleotides in length, and was subsequently ethanol-precipitated. Library quality was confirmed by analysis on the Agilent Bioanalyzer DNA1000 chip (Agilent Technologies). Small non-coding RNA sequencing was performed on the Illumina HiSeq2500 platform at the Michael Smith Genome Sciences Centre at the BC Cancer Research Centre, with 8 multiplexed libraries per sequencing lane (Table 3 (available online only), Fig. 1)[15,16]. Data resulting from small non-coding RNA sequencing can be found on the *Sequence Read Archive* [Data Citation 2].

Pre-processing and small non-coding RNA species detection

Small-RNA sequencing data was analyzed according to published protocols[17]. In order to extract

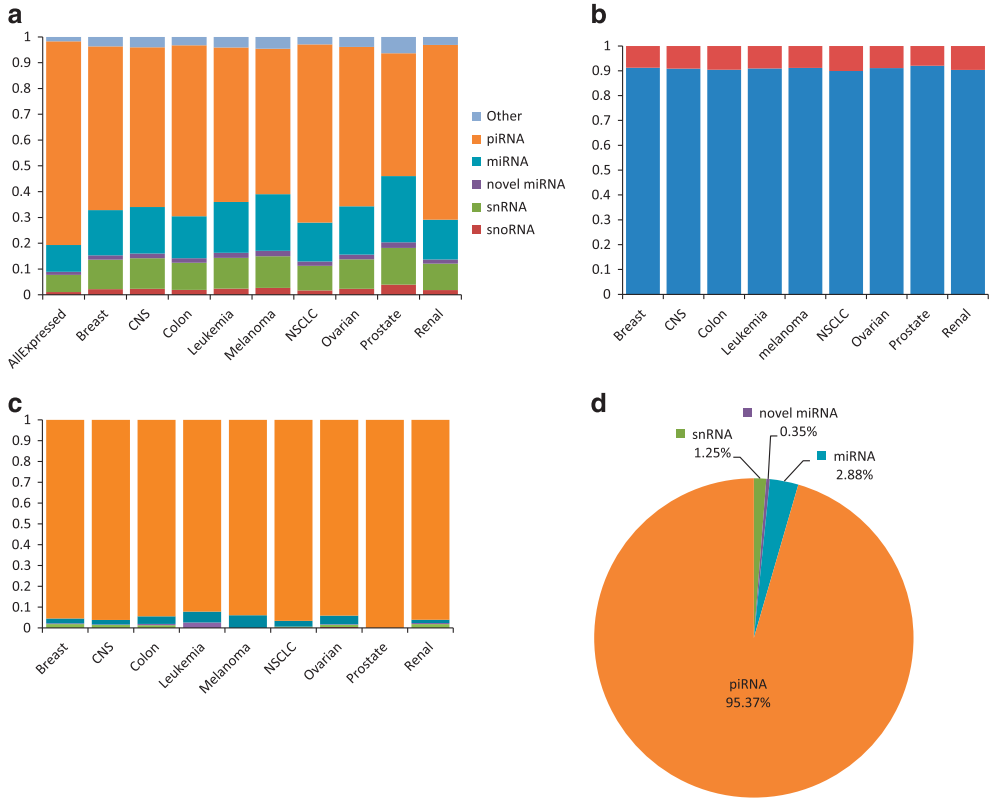

Figure 3. sncRNA distribution by tissue type. (**a**) Relative fraction of sncRNA species detected per tissue type. (**b**) Average fraction of currently annotated (blue) and novel unannotated (red) miRNA per tissue type. (**c**) Relative fraction of tissue-specific unique sncRNA sequences detected per tissue type. (**d**) Fraction of tissue-specific unique sncRNA species.

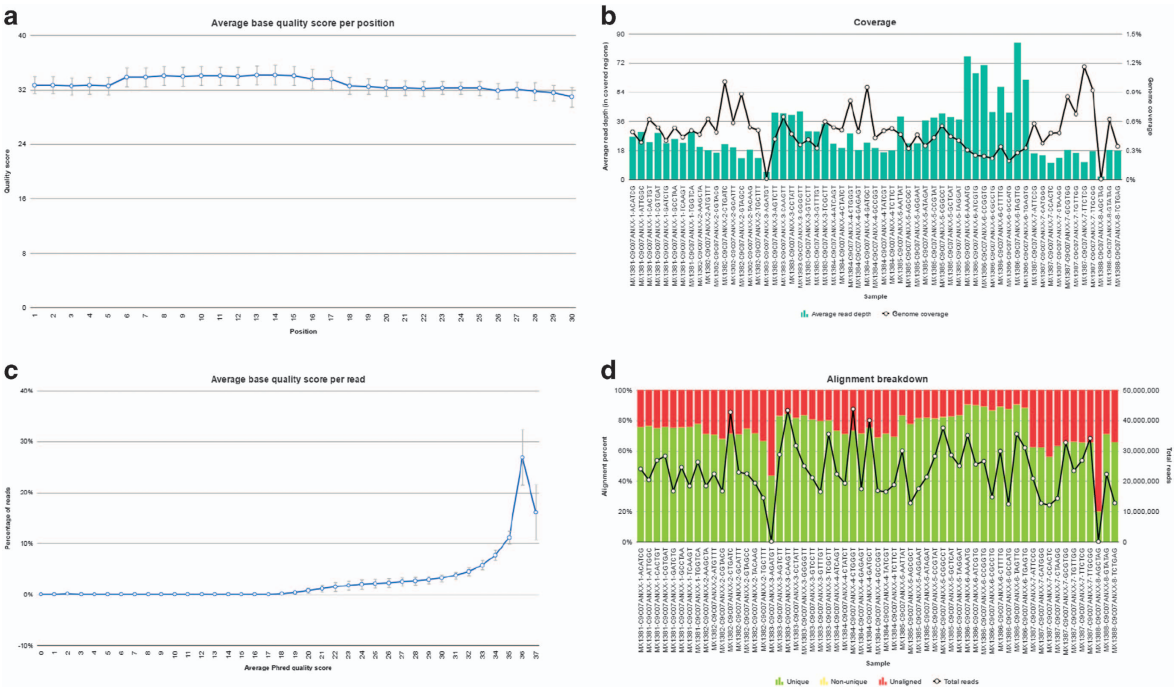

Figure 4. Sequencing and mapping quality. (**a**) Phred quality score per sncRNA base position. (**b**) Genome-wide read depth (column) and genome coverage (line) per sample. (**c**) Fraction of sequencing reads per Phred score. (**d**) Percentage of total reads aligned (unique: green, unaligned: red).

information for the sncRNA species of interest, unaligned reads (in FASTQ format) were trimmed for adaptors (Cutadapt v1.7.1) and based on sequencing quality ('trim bases' from Partek Flow v6.0.17.0614) to reach a Phred quality score ≥20 (Fig. 4a–d). FASTQ files were then aligned using the Spliced Transcripts Alignment to a Reference (STAR v2.4.1d) aligner to the human genome (hg38)[18]. Quantification algorithms (featureCounts v1.4.6 (ref. 19) were applied using chromosomal location annotations for known miRNA (Mirbase v.21 (ref. 20), piRNA (piRNAbank v.2 (ref. 21), snoRNA (Ensembl v.84 (ref. 22), and snRNA (Ensembl v.84 (ref. 22) locations[12]. Detection of novel miRNA is performed using the miRDeep2 algorithm (v2.0.0.5), which considers the relative free energy of miRNAs and their random folding *P*-values[13]. Chromosomal position of expressed small RNAs was plotted against and hg38 karyotype obtained from UCSC Genome Browser (Fig. 2). According to OASIS sncRNA software recommendation (v2.0), sncRNA species were considered expressed if the total reads across all samples considered summed to ≥5 reads[12]. Data resulting from species quantification can be found in Data Citation 1.

Normalization and quantification

Raw reads were scaled/normalized using reads per kilobase exon per million mapped reads (RPKM) method[23], and expression correlation matrices were created using Pearson scores with unsupervised hierarchal clustering performed using one-minus-Pearson correlation scores (Fig. 5). For validation of sncRNA expression, we then correlated miRNA species present in two published microarray cohorts of the NCI-60 cell lines. For the 50 (of the 59) cell lines also present in the Sanger Cell Line Database[24] (http://www.cancerrxgene.org/translation/CellLine), raw reads from each unique sequence were correlated with expression of the sequence previously detected by microarray by rank-normalized Spearman`s correlations (Table 4 (available online only)), and performed a similar analysis against all cell lines present in the cohort described by Sokilde *et al.*[5].

Data Records

Raw unaligned sequencing reads (in FASTQ file format) are available through the *Sequence Read Archive* (Data Citation 2). Raw sequencing file names (in FASTQ format) are listed in Table 3 (available online only). A summary of raw sequencing reads for each detected small RNA species are available at through *Figshare* (Data Citation 1).

Technical Validation

High-throughput sequencing allows for direct in-depth analyses of the human genome, recently revealing a critical role for the expression of the non-coding transcriptome in both genetic and epigenetic regulatory processes.

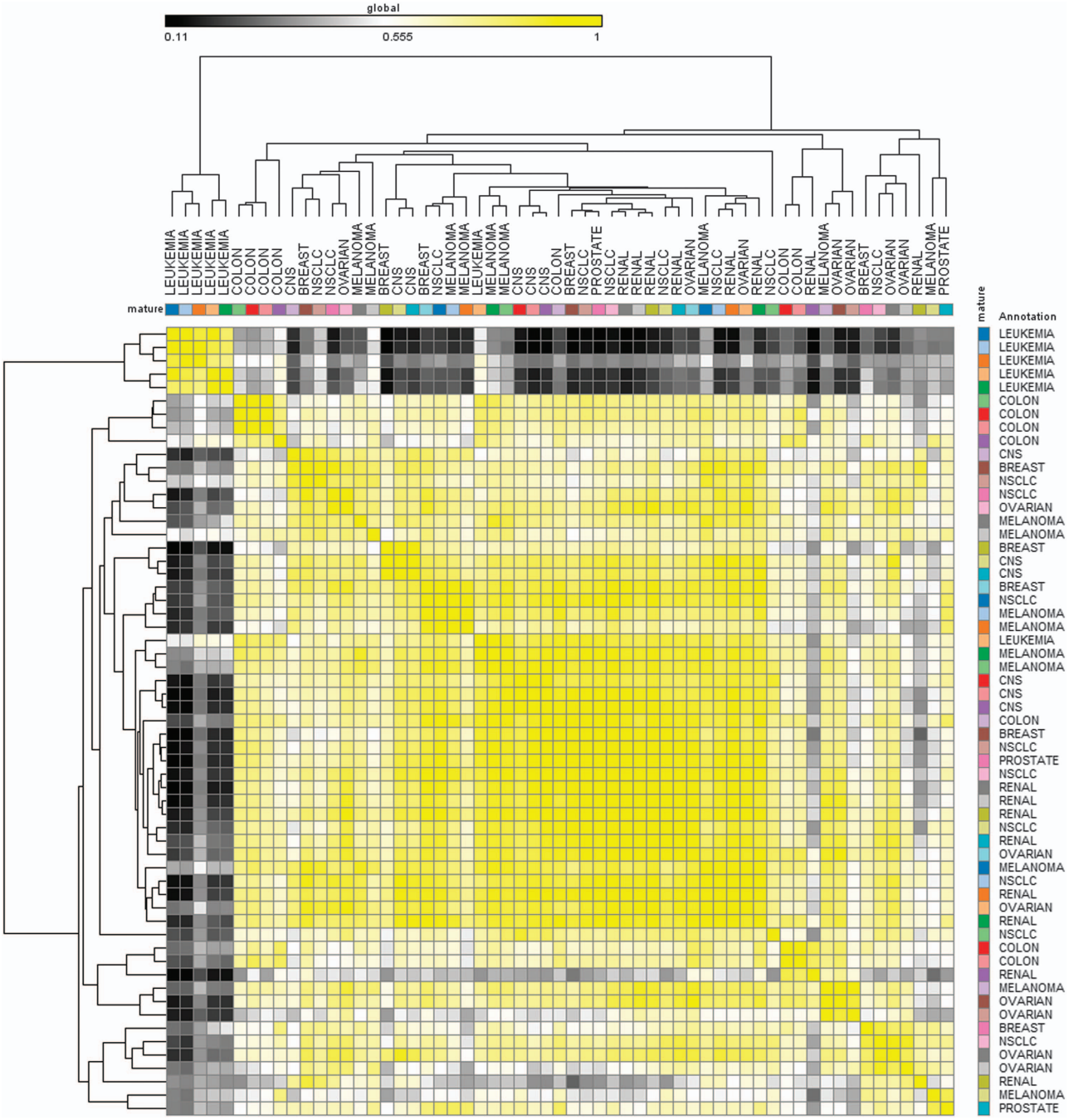

Figure 5. Cell line similarity matrix of small non-coding RNA expression profiles of NCI-60 cell lines by tissue type.

Sequencing quality control

We examined only high-confidence reads from miRNA sequencing. Samples were sequenced to an average depth of 22.34 ± 0.14 (mean $\pm$ s.d.; Table 3 (available online only), Fig. 4b). In order to assure only the calling of high-quality sequencing reads, we filtered detected reads to only to include Phred scores ≥ 20. On average, samples had a Phred score of 33.24 ± 1.28 (Table 3 (available online only), Fig. 4c). Additionally, reads for each sample had an average percent GC content of $46.26 \pm 1.6\%$ (Table 3) (available online only). Unsupervised hierarchical clustering and similarity (one-minus-Spearman correlation) of normalized reads revealed relative similarity of sncRNA expression profiles across all cell lines and tissue types analyzed (Fig. 5).

miRNA detection validation

In order to validate the detection of the sncRNA species in these cell lines, we correlated the raw reads per miRNA detected with corresponding miRNA detected by microarray[24,25]. This analysis was performed for the 50 NCI-60 cell lines present in the Sanger Cell Line miRNA Normalized Data from the Broad Institute (http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi; File name: Sanger_miR_data1.pn.cn.matlab2.res). Using Spearman's Rank-Order correlation, we analyzed the correlation of this RMA-normalized miRNA expression to reads obtained from sequencing this cell line panel. Expression of miRNAs in all lines analyzed correlated significantly between sequencing and microarray analysis (Table 4 (available online only); *P*-values < 0.0001, $r_{mean} = 0.67$). Similarly, we correlated sequencing-detected miRNA expression against a complete NCI-60 microarray cohort described by Sokilde *et al.*[5]. In this study, profiling was performed on the LNA-enhanced mercury Dx 9.2 microarray platform, and data was log_2-normalized after pre-processing (Table 4; *P*-value range < 0.0001–0.0647, $r_{mean} = 0.28$). Microarray data from multiple platforms was compared to sequencing data presented here in order to de-emphasize platform bias and illustrate the need for comprehensive profiling when considering small RNA expression[26].

References

1. Lorenzi, P. L. *et al.* DNA fingerprinting of the NCI-60 cell line panel. *Molecular cancer therapeutics* **8,** 713–724 (2009).
2. Roschke, A. V. *et al.* Karyotypic complexity of the NCI-60 drug-screening panel. *Cancer research* **63,** 8634–8647 (2003).
3. Adams, S. *et al.* HLA class I and II genotype of the NCI-60 cell lines. *Journal of translational medicine* **3,** 11 (2005).
4. Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. *Nature reviews Cancer* **6,** 813–823 (2006).
5. Sokilde, R. *et al.* Global microRNA analysis of the NCI-60 cancer cell panel. *Molecular cancer therapeutics* **10,** 375–384 (2011).
6. Kohn, K. W. *et al.* Gene expression profiles of the NCI-60 human tumor cell lines define molecular interaction networks governing cell migration processes. *PLoS ONE* **7,** e35716 (2012).
7. Hayes, J., Peruzzi, P. P. & Lawler, S. MicroRNAs in cancer: biomarkers, functions and therapy. *Trends in molecular medicine* **20,** 460–469 (2014).
8. Londin, E. *et al.* Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs. *Proceedings of the National Academy of Sciences of the United States of America* **112,** E1106–E1115 (2015).
9. Ng, K. W. *et al.* Piwi-interacting RNAs in cancer: emerging functions and clinical utility. *Molecular cancer* **15,** 5 (2016).
10. Mannoor, K., Liao, J. & Jiang, F. Small nucleolar RNAs in cancer. *Biochimica et biophysica acta* **1826,** 121–128 (2012).
11. Majem, B., Rigau, M., Reventos, J. & Wong, D. T. Non-coding RNAs in saliva: emerging biomarkers for molecular diagnostics. *International journal of molecular sciences* **16,** 8676–8698 (2015).
12. Capece, V. *et al.* Oasis: online analysis of small RNA deep sequencing data. *Bioinformatics* **31,** 2205–2207 (2015).
13. Friedlander, M. R. *et al.* Discovering microRNAs from deep sequencing data using miRDeep. *Nature biotechnology* **26,** 407–415 (2008).
14. Friedlander, M. R., Mackowiak, S. D., Li, N., Chen, W. & Rajewsky, N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic acids research* **40,** 37–52 (2012).
15. Chu, A. *et al.* Large-scale profiling of microRNAs for The Cancer Genome Atlas. *Nucleic acids research* **44,** e3 (2016).
16. Becker-Santos, D. D. *et al.* Developmental transcription factor NFIB is a putative target of oncofetal miRNAs and is associated with tumour aggressiveness in lung adenocarcinoma. *The Journal of pathology* **240,** 161–172 (2016).
17. Martinez, V. D. *et al.* Unique somatic and malignant expression patterns implicate PIWI-interacting RNAs in cancer-type specific biology. *Scientific reports* **5,** 10423 (2015).
18. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
19. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923–930 (2014).
20. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research* **42,** D68–D73 (2014).
21. Sai Lakshmi, S. & Agrawal, S. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic acids research* **36,** D173–D177 (2008).
22. Aken, B. L. *et al.* The Ensembl gene annotation system. *Database* **2016,** baw093–baw093 (2016).
23. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5,** 621–628 (2008).
24. Yang, W. *et al.* Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research* **41,** D955–D961 (2013).
25. Enfield, K. S. *et al.* MicroRNA gene dosage alterations and drug response in lung cancer. *Journal of biomedicine & biotechnology* **2011,** 474632 (2011).
26. Patnaik, S. K. *et al.* Expression of microRNAs in the NCI-60 cancer cell-lines. *PLoS ONE* **7,** e49918 (2012).

Data Citations

1. Marshall, E. A. *et al. figshare* https://doi.org/10.6084/m9.figshare.c.3811156 (2017).
2. *NCBI Sequence Read Archive* SRP109305 (2017).

Acknowledgements

This work was supported by grants from the Canadian Institutes for Health Research (CIHR FRN-143345), the National Institute of Health (NIH-1R01HD089713-01) and scholarships from CIHR and the University of British Columbia. The authors would like to thank May Zhang, Miwa Suzuki and Emma Conway for assistance with cell culture and RNA harvesting. The authors would also like to thank Monica Fuss for assistance in data analysis.

Author Contributions

E.A.M. and V.D.M. were responsible for project design and manuscript construction. V.D.M. supervised data generation. E.A.M., A.P.S., K.W.N., V.D.M. and contributed to data analysis and manuscript

preparation. N.S.F. contributed to data generation. K.L.B. and W.L.L. are the principal investigators who initiated this project. All authors contributed to the editing of the manuscript.

Additional Information

Tables 1, 3 and 4 are only available in the online version of this paper.

Competing interests: The authors declare no competing financial interests.

How to cite this article: Marshall, E. A. *et al.* Small non-coding RNA transcriptome of the NCI-60 cell line panel. *Sci. Data* 4:170157 doi: 10.1038/sdata.2017.157 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.