Published online 13 August 2015

doi: 10.1093/nar/gkv808

Large-scale profiling of microRNAs for The Cancer Genome Atlas

Andy Chu[1,†], Gordon Robertson[1,†], Denise Brooks[1], Andrew J. Mungall[1], Inanc Birol[1,2], Robin Coope[1], Yussanne Ma[1], Steven Jones[1,2,3] and Marco A. Marra[1,2,*]

[1]Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC, V5Z 4S6, Canada, [2]Department of Medical Genetics, University of British Columbia, Vancouver, V6H 3N1, Canada and [3]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada

Received July 7, 2015; Accepted July 28, 2015

ABSTRACT

The comprehensive multiplatform genomics data generated by The Cancer Genome Atlas (TCGA) Research Network is an enabling resource for cancer research. It includes an unprecedented amount of microRNA sequence data: ∼11 000 libraries across 33 cancer types. Combined with initiatives like the National Cancer Institute Genomics Cloud Pilots, such data resources will make intensive analysis of large-scale cancer genomics data widely accessible. To support such initiatives, and to enable comparison of TCGA microRNA data to data from other projects, we describe the process that we developed and used to generate the microRNA sequence data, from library construction through to submission of data to repositories. In the context of this process, we describe the computational pipeline that we used to characterize microRNA expression across large patient cohorts.

INTRODUCTION

MicroRNAs (miRNAs) play roles in post-transcriptional regulation, but the specific roles of many miRNAs in particular genetic contexts remain poorly understood (1,2). Profiling miRNA expression using short read sequencing offers a large dynamic range and high spatial resolution, which can help generate insights into the roles of miRNAs and their regulatory targets in cancer biology. While miRNAs can now be routinely characterized on a whole transcriptome level, clarifying the functional significance of miRNAs in tumorigenesis may benefit from comparisons across subtypes within large cohorts and across tumor types. The Cancer Genome Atlas (TCGA) Research Network has generated and made publicly available comprehensive genomic data for more than 11 000 tumor samples representing 33 cancer types. The TCGA miRNA sequencing (miRNAseq) data were generated by Canada's Michael Smith Genome Sciences Centre (GSC) at the BC Cancer Agency between 2010 and 2015. This resource includes data for both tumor and adjacent normal samples, and is the largest of its type worldwide. Going forward, data from large consortia like TCGA, along with enabling initiatives like the National Cancer Institute Genomics Cloud Pilots, will create unprecedented research opportunities. Researchers will be able to interpret miRNA data from cancer cohorts in a pan-cancer context, and teams with extensive experience in the biology of particular cancers will be able to use TCGA data to complement clinical and genomics data that they generate from their own patient cohorts.

To support such opportunities, here we describe the processes that we developed and used to generate the TCGA miRNAseq data. We expand on key details and in general provide more information than was previously available in the method descriptions in TCGA Research Network publications, thus providing a comprehensive reference document for those who wish to understand and use the miRNAseq data that we generated, or the computational pipeline. We describe the high-throughput transcriptome library construction protocol, the sequencing, and the data analysis pipelines that we developed for TCGA. We show that expression profiling results from our computational pipeline are similar to those from miRDeep2 (3) and ShortStack (4). However, unlike these and other miRNA analytical tools, our profiling software provides the processing method that was used to generate TCGA data, and so provides a mechanism for comparing miRNA sequencing data from other projects with this large dataset in a rigorous and consistent manner.

MATERIALS AND METHODS

Overview

Briefly, library construction was strand-specific and enriched for ∼22-bp miRNA mature strands. Libraries were pooled and then sequenced using Illumina technology.

[*]To whom correspondence should be addressed. Tel: +1 604 707 5800; Fax: +1 604 876 3561; Email: mmarra@bcgsc.ca

[†]These authors contributed equally to the paper as first authors.

Library-specific indices added during library construction allowed assigning reads from a sequenced pool to individual samples. During analysis, the computational pipeline demultiplexed the pools and aligned the reads to a reference genome. We classified reads as miRNAs or other small RNAs (snoRNAs, tRNAs or rRNAs) by comparing read alignments to sequence feature annotations from miRBase (5) and UCSC (6); resources like Ensembl (7) can also be used. For consistency across the five years of our involvement in TCGA projects, we used miRBase v16 annotations with GRCh37/hg19 alignments and v13 annotations with GRCh36/hg18 for the earlier cancer studies. We designed the profiling pipeline to be highly specific; it reports expression only for exact-match read alignments to miRBase miRNAs and considers neither alignments with mismatches nor novel miRNAs. The raw data are isomiR sequences. We reported expression in two forms: as raw read counts and as counts normalized to reads per million mapped reads (RPM), and we submitted both forms as isomiR and stem-loop expression data archives to the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm). Given this, archive downloads include four files per library: read counts and RPM, for stem-loops and isomiRs. Expression data for miRNA mature strands can be generated from the downloaded isomiR archives. The pipeline software includes scripts to collate sets of downloaded per-library files into a miRNAs-by-samples expression matrix for stem-loops, mature strands or isomiRs. For those who wish to apply other profiling pipelines, to identify novel miRNAs, or to do analyses that address sequence variants and untemplated additions, we also submitted to cgHub (https://cghub.ucsc.edu) BAM (8) files that contain exact-match, mismatch and unaligned sequence reads.

Library construction and sequencing

miRNA-seq libraries were constructed using a strand-specific, plate-based protocol developed at the GSC (Figure 1). Approximately 1 μg of either total RNA or messenger RNA-depleted RNA, containing small RNA species was used. A 3′ adapter was ligated to the RNA template using a truncated T4 RNA ligase2 (NEB Canada, cat. M0242L) with an incubation of 1 h at 22°C. This adapter was an adenylated, single-strand DNA with the sequence 5′ /5rApp/ ATCTCGTATGCCGTCTTCTGCTTGT /3ddC/, which selectively ligated miRNAs. An RNA 5′ adapter was then added, using a T4 RNA ligase (Ambion USA, cat. AM2141) and ATP, and was incubated at 37°C for 1 h. The sequence of the single strand RNA adapter is 5′GUUCAGAGUUCUACAGUCCGACGAUCUGGUCAA3′.

When ligation was completed, first strand cDNA was synthesized using Superscript II Reverse Transcriptase (Invitrogen, cat. 18064 014) and an RT primer (5′-CAAGCAGAAGACGGCATACGAGAT-3′). This was the template for the final library polymerase chain reaction (PCR), into which we introduced 6-nt index sequences that enabled libraries to be identified (i.e. demultiplexed) from a sequenced pool that contained multiple libraries. Briefly, a PCR brew mix was made with the 3′ PCR primer (5′-CAAGCAGAAGACGGCATACGAGAT-3′), Phusion Hot Start High Fidelity DNA polymerase (NEB Canada, cat. F-540L), buffer, dNTPs and dimethyl sulfoxide (DMSO). The mix was distributed evenly into a new 96-well plate. A Biomek FX (Beckman Coulter, USA) was used to transfer the PCR template (first strand cDNA) and indexed 5′ PCR primers into the brew mix plate. Each 5′ PCR primer, 5′-AATGATACGGCGACCACCGACAGNNNNNNGTTCAGAGTTCTACAGTCCGA-3′, contained a unique, fault-tolerant, 6-nt 'index' (shown here as N's), and was added to each well of the 96-well PCR brew plate. The hexamer index sequences enabled pooling samples for sequencing by supporting 'demultiplexing' the sequence data into separate sets of read sequences for each library. For TCGA, we sequenced a pool of 16 miRNA libraries in each HiSeq 2500 lane. PCR was run at 98°C for 30 s, followed by 15 cycles of 98°C for 15 s, 62°C for 30 s and 72°C for 15 s, and finally a 5 min incubation at 72°C. Quality was then checked across the whole plate using a Caliper LabChipGX DNA chip. Negative controls were added at three stages: elution buffer was added to one well when the total RNA was loaded onto the plate, water to another well just before ligating the 3′ adapter, and PCR brew mix to a final well just before PCR.

PCR products were pooled, then were size selected to remove larger cDNA fragments and smaller adapter contaminants, and to enrich constructs with ∼22-bp insert lengths (Figure 1), using an automated, 96-channel size selection robot that was developed at the GSC. After size selection, each pool was ethanol precipitated, quality checked using an Agilent Bioanalyzer DNA1000 chip and quantified using a Qubit fluorometer (Invitrogen, cat. Q32854). Each pool was then diluted to a target concentration for cluster generation and loaded into a single lane of an Illumina flow cell. Clusters were generated and lanes were then sequenced using a 30-bp read to capture the miRNA sequence and a 6-bp read to capture the index sequence.

Preprocessing and aligning reads

While the size-selected miRNAs varied in length, they were typically ∼22 bp long, so were shorter than the 30-bp read length (Figure 1). Given this, each read sequence typically extended some distance into the 3′ sequencing adapter. Because this non-biological sequence could interfere with aligning reads to the reference genome, we identified and removed any 3′ adapter sequence from each read.

Our adapter-trimming algorithm identified as long an adapter sequence as possible, allowing a number of mismatches that depended on the adapter length found. The algorithm first determined whether a read sequence should be discarded as an adapter dimer that had no cDNA insert by checking whether the 3′ adapter sequence occurred at the start of the read. For reads passing this stage, the algorithm then tried to identify an exact 15-bp match to the 3′ adapter sequence anywhere within the read sequence. If it could not, it tried again, starting from the 3′ end of the read sequence and allowing up to two mismatches. If the full 15 bp was not found, decreasing lengths of adapter were checked, down to the first eight bases, allowing one mismatch. If a match was still not found, from seven bases down to one base were

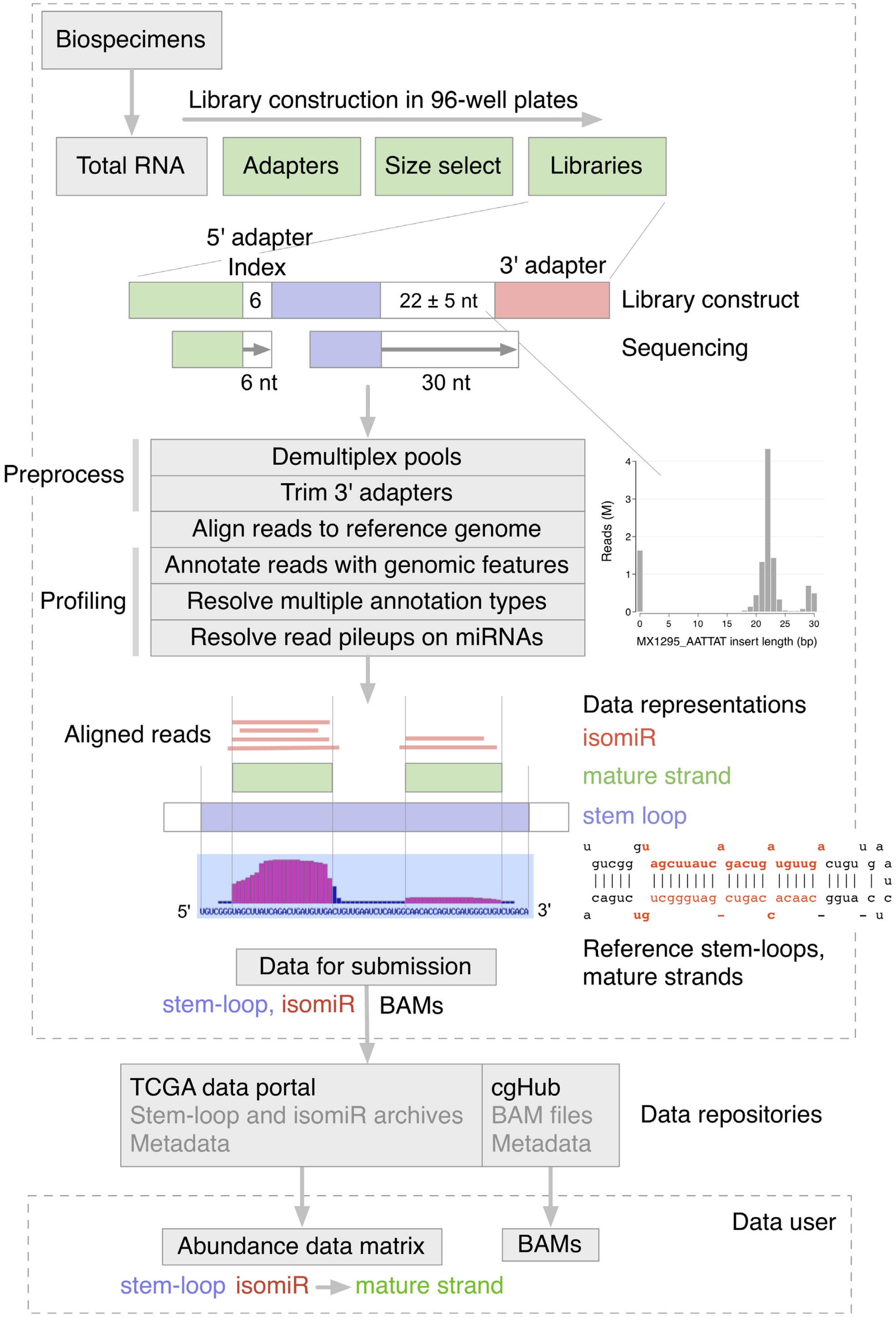

Figure 1. Process used to generate TCGA miRNA sequence data. Strand-specific library construction was performed in parallel in 96-well plates. The example reference read pileup and stemloop are hsa-mir-21 from miRBase v21.

Table 1. Annotation priorities (Pr) that are used to resolve multiple annotation type matches for a single alignment location or for multiple alignment locations for a read. See 'Profiling small RNA abundance'

Pr	Annotation type	Database
1	mature strand	miRBase
2	star strand	
3	precursor miRNA	
4	stemloop, from 1 to 6 bases outside the mature strand, between the mature and star strands	
5	'unannotated', any region other than the mature strand in miRNAs where no star strand is annotated	
6	snoRNA	UCSC small RNAs and RepeatMasker
7	tRNA	
8	rRNA	
9	snRNA	
10	scRNA	
11	srpRNA	
12	Other RNA repeats	
13	coding exons with zero annotated CDS region length	UCSC genes
14	3′ UTR	
15	5′ UTR	
16	coding exon	
17	intron	
18	LINE	UCSC RepeatMasker
19	SINE	
20	LTR	
21	Satellite	
22	RepeatMasker DNA	
23	RepeatMasker low complexity	
24	RepeatMasker simple repeat	
25	RepeatMasker other	
26	RepeatMasker unknown	

checked, with an exact match required. Finally, the algorithm would trim one base off the 3′ end of a read if this base matched the first base of the adapter. This step was based on two considerations. First, since we only used exact match alignments for expression quantification, a trimmed sequence that gave one or more exact-match alignments was preferable to a sequence that we did not use due to a one-base mismatch. Second, if only one base of the adapter was found in the 30-nt read sequence, the read was likely too long to be from a miRNA mature strand, and the effect of the trimming on its alignment should not affect the overall miRNA profiling result for the sample from which the read was derived.

Because the shortest mature miRNA in miRBase v16 is 15 bp, we discarded any trimmed read that was shorter than 15 bp. We used BWA-MEM with parameters samse -n 10 (9) to align the remaining reads to a reference genome, which, for most TCGA cancers, was GRCh37 (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/). This generated a SAM file (8) in which multiple alignments for a read were recorded in a single line using BWA's custom XA tag.

Profiling small RNA abundance

The pipeline assigned aligned reads to sequence features using reference databases specified by the user (Table 1), or by corresponding flat annotation files that can represent, for example, novel miRNAs that a user wishes to profile. The output was an updated SAM file in which annotations for each alignment were stored in new XC, XD and XI tags, in the order reported by the aligner in the XA tag. The advantages of writing the intermediate results into the SAM file, rather than a separate file, derive from each library having a single text file that contains both alignment and feature annotation information, has a well-documented format, and can be readily queried by scripts or from the command line. For TCGA datasets we used a 3-bp minimum overlap between an aligned read sequence and a genomic feature annotation; this minimum can be adjusted when running the profiler. For a read that had a single mapping location, but overlapped multiple types of genomic features at that location, we reduced the set of feature types to the single highest priority type, using the heuristically determined priorities in Table 1. For a multi-mapped read, we first resolved the annotation types for each alignment as we did for a uniquely mapped read, then used the priority list again to resolve the set of annotations across the multiple alignment locations into a single annotation type for that read.

We dealt with a read that had multiple exact-match alignments to different miRNAs in one of two ways, depending on whether the read multi-mapped to mature sequences that had the same miRBase accession. A read could map to functionally identical mature miRNA strands that have identical sequences and can be expressed from different locations in the genome. For example, hsa-miR-181a-5p, whose sequence has miRBase accession ID: MIMAT0000256, occurs in both hsa-mir-181a-1 at 1q32.1 and in hsa-mir-181a-2 at 9q33.3. When we assigned a read to miR-181a-5p, we incremented the read count for this mature strand by one, while incrementing the read count by one for a random choice of one of the stem-loops hsa-mir-181a-1 or hsa-mir-181a-2. Alternatively, a read could map exactly to different miRNAs that have sequences that are similar but not identical, when the read sequence did not capture the bases

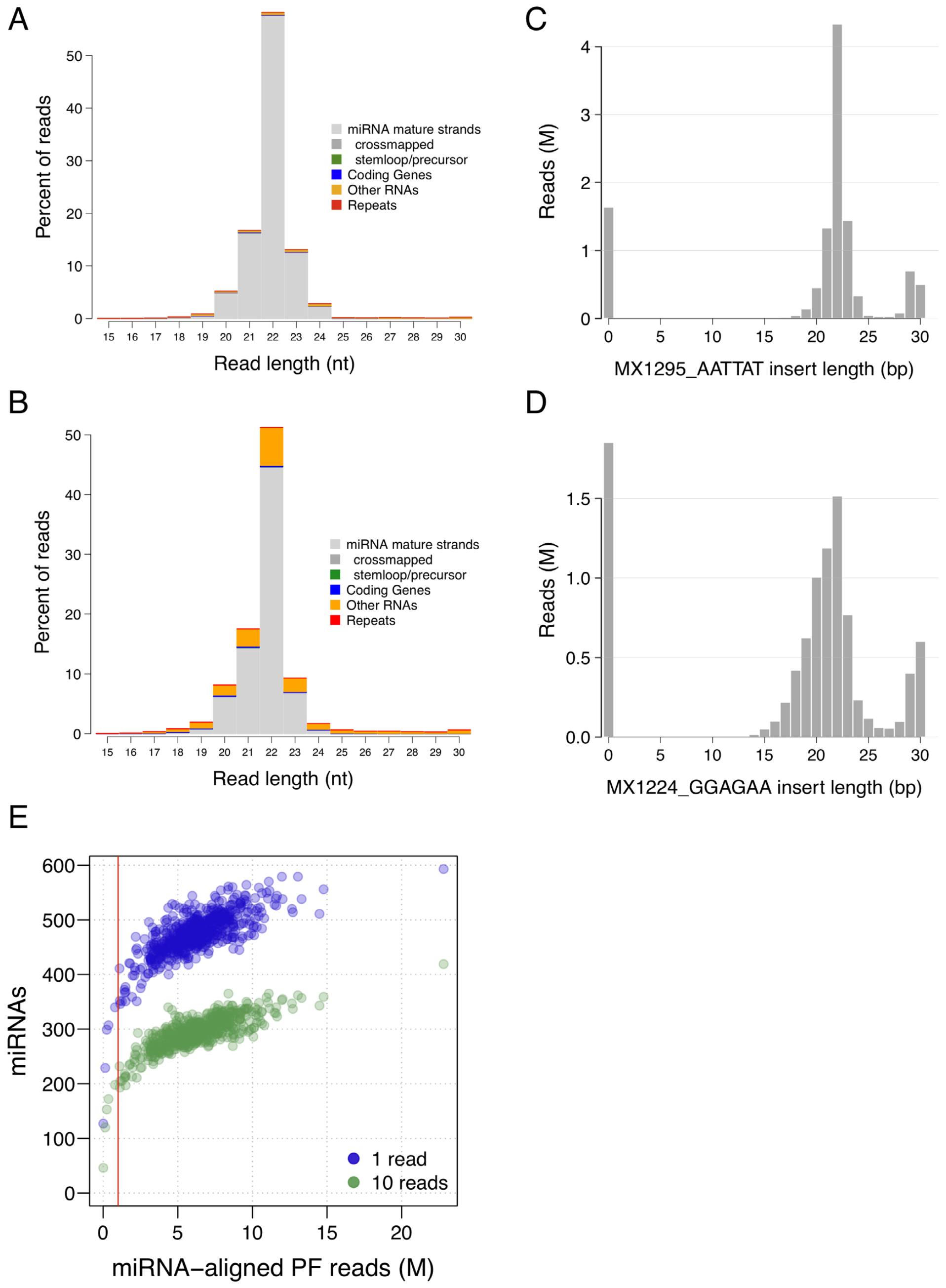

Figure 2. Example library quality graphs generated by the GSC miRNA annotation profiling pipeline for thyroid carcinoma tumor libraries (1). For an individual library: (**A** and **B**) Percentage of small RNA annotations as a function of read length for two libraries. (**C** and **D**) Distribution of read lengths after adapter trimming with (C) a preferred narrower and (D) a wider insert length distribution. (**E**) For 496 thyroid tumor libraries, the relationship of miRNA species identified with at least 1 or 10 aligned reads versus all post-filtered (PF) reads aligned to miRNAs. The vertical red line shows the TCGA-specific threshold of 1 M reads.

that differ between these miRNAs. For such a case, we reported the read as cross-mapped (10). Supplementary Table S1 shows an example from a TCGA head-and-neck tumor library in which a 17-nt isomiR read had exact-match alignments to 22-nt 3p mature strands of both hsa-mir-30a at 6q13 and hsa-mir-30e at 1p34.2, whose sequences differ only at position 18. When a read cross-mapped, we incremented the read count by one for each miRNA to which it cross-mapped. We preserved all of its miRNA annotations and discarded all of its non-miRNA annotations, which ensured that the SAM file retained all annotation information about ambiguously mapped miRNAs, allowing the ambiguity to be addressed in downstream analysis.

For each library, after aligned reads had been assigned to genomic feature annotations, the read counts for each retained annotation were summed to generate expression profiles. Reads with alignments to multiple miRNA genes were flagged as cross-mapped, and the pipeline then wrote a crossmapping report to support downstream filtering, produced graphs illustrating library content (Figure 2A and B), and generated an expression profile. When counting read annotations, reads were filtered by a set of adjustable thresholds. By default we retained reads that failed the Illumina base-calling chastity filter and excluded from expression quantification those with more than three alignments, and with mismatches or soft-clipped CIGAR strings (8); however, we retained all reads in the BAM files that we submitted to cgHub.

Quality control in computational processing

The pipeline includes quality control steps before and after read alignment. After adapter trimming, a size distribution report was generated for the distribution of read lengths, which was assessed manually to ensure that size selection was successful (Figures 1, 2C and D). A narrow size distribution centered at 22 bp was typical of a good quality library, and skewed or wide distributions indicated a poor quality library or problems with size selection of library constructs. When profiling was completed, a summary report was generated that gave quality metrics that included the total number of reads aligned to miRNAs, the number of miRNA species covered by 1 and 10 reads, and a count of the total reads classified as each annotation type. To ensure that a minimum number of reads were obtained for each sample, we set a project-specific threshold from the relationship between the total number of reads that align to miRNAs and the number of miRNA species annotated. For example, for HiSeq TCGA datasets, we used a threshold of 1 M miRNA-aligned reads (Figure 2E). Finally, we used an in-house tool based on iterative principal components analysis to screen samples for lane-wise batch effects (http://www.bcgsc.ca/platform/bioinfo/software/bliss).

Consensus clustering of miRNA expression profiles

We identified potential subtypes using unsupervised Nonnegative Matrix Factorization (NMF) consensus clustering, using the CRAN R package (11). First, we removed records for miRs that had been retired by miRBase (MIRNA.dead). We ranked mature strands by variance of expression across tumor samples, then created the NMF input by selecting records for the most-variant 25% of 5p and 3p strands, corresponding to 303 v16 miRBase strands. We processed this data subset with NMF's default Brunet algorithm, surveying a range of candidate clustering solutions. We selected a preferred result by considering the profiles of the cophenetic coefficient and average silhouette widths of consensus membership matrices (12), and associations with clinical and molecular covariates, and with survival, seeking the most informative solution for the overall analysis.

We generated a miR abundance heatmap for the selected clustering solution by identifying miRs that were differentially abundant between the unsupervised clusters with a SAMseq multiclass analysis in R, using the samr v2.0 R package, with a read-count input matrix and an FDR threshold of 0.05 (13). We used the pheatmap v0.7.7 R package to display the row-scaled $\log_{10}$(RPM + 1) abundance matrix of the 40–60 miRs that had the largest SAMseq scores and a mean tumor expression greater than 25 RPM. RPM filtering addresses potential sponge effects from competitive endogeneous RNAs (ceRNAs), which can result in weakly abundant miRs being less influential (14). Leaving samples (i.e. columns) in NMF order, we row-scaled the data matrix and clustered rows with a Euclidean distance metric. While we used (v13 or v16) miRBase MIMAT accession IDs in the isomiR data that we submitted to data repositories, for heatmaps we translated MIMATs into 5 and 3p mature strand names using a current version of miRBase. Finally, we used R to characterize associations between clusters and categorical clinical and molecular covariates by Fisher exact or Chi-square tests, and between clusters and real-valued covariates (e.g. tumor purity) by Kruskal–Wallis tests.

Comparing pipeline outputs to miRDeep2 and ShortStack

To demonstrate that results from our miRNA profiling pipeline are comparable to those from alternative pipelines, we processed FASTQ files for aligned reads for five TCGA bladder cancer miRNA-seq libraries (15) with miRDeep2 (3) and ShortStack (4). miRDeep2 and ShortStack annotate and quantify reference-aligned small RNA-seq data, and can also report putative novel miRNAs. Because we designed the TCGA pipeline to report only known miRNAs, we included only known miRNAs in the pipeline comparisons. We ran both miRDeep2 and ShortStack using default settings (Supplementary Tables S2 and S3).

RESULTS

Comparison with miRDeep2 and ShortStack

Working with FASTQ files for aligned reads for five TCGA bladder cancer datasets, we first assessed run times. Because we were quantifying only known miRNA, we constrained ShortStack to use only miRBase annotations, which reduced its run times. miRDeep2 ran in the shortest time, while ShortStack had comparable runtimes to the our TCGA pipeline (Table 2). Averaging over these five libraries, 85% of the GSC's total processing time involved BWA-MEM read alignments; post-alignment processing time was 15% of the total, with four of the five libraries requiring <0.5 h for this stage. All three tools had run times that were practical for large scale profiling, given the compute cluster resources that would typically be available to a group doing such work.

No pipeline consistently gave the smallest or largest percentage of sequence reads mapped to miRNA annotations in all libraries; our TCGA datasets reported the largest percentage in two of the libraries and miRDeep2 in three (Table 3). The miR expression profiles produced by our pipeline were highly correlated with the profiles produced by miRDeep2 (Pearson $r = 0.996$) and ShortStack ($r = 0.988$) (Table 4).

Our miRNA species were highly concordant with those from the other two pipelines (Figure 3, Table 5); across the five test libraries, an average of 99.5% of the miRs that our profiler annotated as having nonzero read support were concordant with miRs annotated by the other two pipelines. Corresponding averages for ShortStack and miRDeep2 were 98.1 and 95.6%. We noted that the miRs detected by only one or two of the methods were largely those with low abundance (Tables 6 and 7), which are less

Table 2. Average run times (h:min:s) for the five TCGA bladder cancer libraries

TCGA barcode	GSC	miRDeep2	ShortStack
TCGA-G2-A2EL-01A-12R-A18B-13	3:00:38 (23)	0:49:44	2:59:49
TCGA-G2-A2EK-01A-22R-A18B-13	3:44:41 (63)	0:39:24	2:57:15
TCGA-FD-A3B3–01A-12R-A205–13	3:07:41 (27)	0:29:57	2:58:43
TCGA-CF-A3MI-01A-11R-A20E-13	2:49:32 (15)	0:35:44	3:00:05
TCGA-DK-A3IS-01A-21R-A21E-13	3:16:38 (25)	0:49:45	2:59:06

The GSC and miRDeep2 pipelines were run with default settings. ShortStack was run with defaults but was constrained to miRBase v16 annotations. For GSC, numbers in parentheses are annotation times, i.e. processing time following read alignment, in minutes.

Table 3. Percentage of reads annotated as miRNA in the five TCGA libraries

TCGA barcode	GSC	miRDeep2	ShortStack
TCGA-G2-A2EL-01A-12R-A18B-13	31.0	**37.8**	29.5
TCGA-G2-A2EK-01A-22R-A18B-13	32.6	**48.8**	44.3
TCGA-FD-A3B3–01A-12R-A205–13	52.5	**60.4**	56.2
TCGA-CF-A3MI-01A-11R-A20E-13	**40.0**	35.7	32.0
TCGA-DK-A3IS-01A-21R-A21E-13	**58.2**	54.6	46.3

Bold text marks the maximum percentage for the three methods for a library.

Table 4. Pearson correlation coefficients for normalized miR abundance profiles generated by the three annotation methods for five TCGA libraries

	All miRs			miRs with ≥10 reads		
TCGA barcode	GSC MD	GSC SS	MD SS	GSC MD	GSC SSS	MD SS
TCGA-G2-A2EL-01A-12R-A18B-13	0.998	0.975	0.981	0.998	0.975	0.981
TCGA-G2-A2EK-01A-22R-A18B-13	0.998	0.997	1.000	0.998	0.997	1.000
TCGA-FD-A3B3–01A-12R-A205–13	0.998	0.997	1.000	0.998	0.997	1.000
TCGA-CF-A3MI-01A-11R-A20E-13	0.998	0.996	0.999	0.998	0.996	0.999
TCGA-DK-A3IS-01A-21R-A21E-13	0.990	0.976	0.994	0.989	0.975	0.994

Per-sample miR profiles were normalized to reads per 1 M miR-annotated reads (RPM). MD: miRDeep2. SS: ShortStack

Table 5. Number of miRNA species to which reads were mapped using miRBase v16 (of 1222 possible mature miRs)

	GSC		miRDeep2		ShortStack	
TCGA barcode	Total	>10 reads	Total	>10 reads	Total	>10 reads
TCGA-G2-A2EL-01A-12R-A18B-13	648	429	719	485	687	461
TCGA-G2-A2EK-01A-22R-A18B-13	638	421	699	466	680	437
TCGA-FD-A3B3–01A-12R-A205–13	612	365	676	404	658	385
TCGA-CF-A3MI-01A-11R-A20E-13	575	323	629	363	614	348
TCGA-DK-A3IS-01A-21R-A21E-13	650	426	697	473	689	445

In a library with 1 M miR-aligned reads, 10 reads corresponds to 1e-5 RPM.

Table 6. The number and normalized abundance (RPM) for miRNA species reported by the GSC, MiRDeep2 and Shortstack profiling pipelines, using default settings

		GSC RPM		MiRDeep2 RPM		Shortstack RPM	
Venn Membership	Total miRs	Median (mean)	Max	Median (mean)	Max	Median (mean)	Max
All pipelines	785	3.0 (1045)	223 316	2.1 (1044)	345 035	3.18 (1274)	378 133
GSC and MD	12	0.05 (0.09)	0.38	0.02 (0.03)	0.14		
GSC and SS	18	0.14 (0.57)	7.1			0.17 (0.62)	7.4
MD and SS	49			0.17 (3.96)	113.9	0.07 (0.77)	13.7
Only SS	19					0.04 (3.01)	42.4
Only MD	30			0.06 (0.16)	1.1		
Only GSC	4	0.04 (0.04)	0.08				
Grand Total	917						

See Figure 3, compare to Table 7. MD: miRDeep2. SS: ShortStack

Table 7. The number and normalized abundance (RPM) for miRNA species reported by GSC, MiRDeep2 and Shortstack profiling pipelines, relaxing settings for GSC to allow up to one base mismatch and any number of mapping locations

		GSC RPM		MiRDeep2 RPM		ShortStack RPM	
Venn membership	Total miRs	Median (mean)	Max	Median (mean)	Max	Median (mean)	Max
All pipelines	833	2.85 (1201)	273 827	3.44 (1201)	354 265	2.91 (1201)	378 133
GSC and MD	25	0.17 (0.27)	1.90	0.17 (0.42)	3.07		
GSC and SS	25	0.17 (0.40)	4.60			0.32 (0.62)	7.4
MD and SS	1			0.79 (0.79)	0.79	0.22 (0.22)	0.22
Only SS	12					0.21 (4.85)	42.4
Only MD	17			0.28 (0.35)	1.1		
Only GSC	38	0.14 (0.15)	0.34				
Grand Total	951						

Table 8. Average RPM of miRs found by both miRDeep and Shortstack in all five samples and were not counted as annotated by the GSC

MIMAT	miR name	miRBase sequence	miRDeep2 Mean RPM	Shortstack Mean RPM
MIMAT0005911	miR-1260	AUCCCACCUCUGCCACCA	17.34	1.86
MIMAT0005927	miR-1274a	GUCCCUGUUCAGGCGCCA	0.27	0.09
MIMAT0005946	miR-1280	UCCCACCGCUGCCACCC	0.61	0.14
MIMAT0005450	miR-518e-5p	CUCUAGAGGGAAGCGCUUUCUG	0.43	0.06
MIMAT0002831	miR-519c-5p	CUCUAGAGGGAAGCGCUUUCUG	0.43	0.06
MIMAT0004984	miR-941	CACCCGGCUGUGUGCACAUGUGC	0.92	0.09

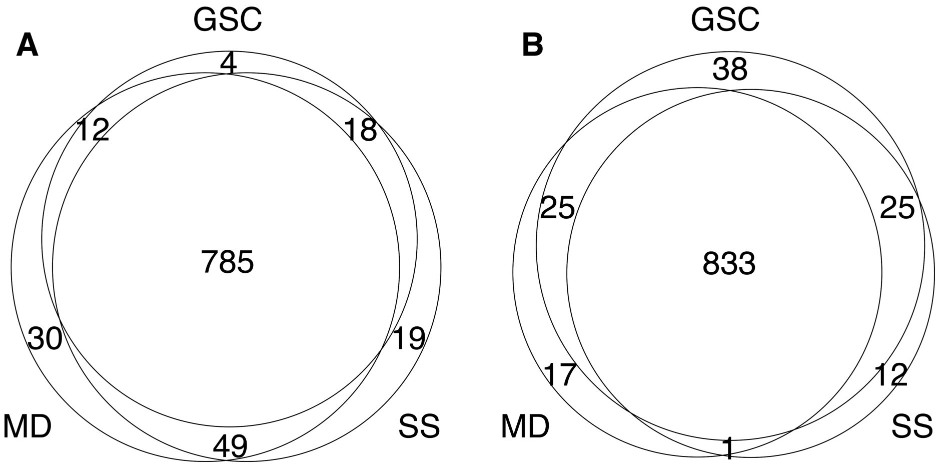

Figure 3. Venn diagram of miRNAs detected at any level of read coverage, in at least one of the five bladder cancer libraries, by the GSC, miRDeep2 (MD) and ShortStack (SS). (**A**) All methods were run with default settings. (**B**) The GSC pipeline was run with lowered quality settings that allowed multimapping and one mismatch. miRs detected by only one or two methods had low RPMs. See Tables 6 and 7.

likely to be biologically influential than more abundant miRs (14). Only six miRs, and all with relatively low abundance, were identified by both miRDeep2 and ShortStack in all of the five samples that were not reported as expressed in our TCGA results (Table 8). While our profiler annotated reads for all of these miRs, we did not retain them for the archives submitted to the TCGA repositories because the alignments did not pass our filters, which required an exact read alignment to a known miRNA, with three or fewer multiple mappings to the reference genome. When we relaxed these filters to allow any number of ambiguous mappings with up to one base mismatch, our profiler mapped reads to each of these miRs in at least one sample.

DISCUSSION

To support resource-intensive analysis of comprehensive cancer genomics data, and to enable the comparison of microRNA sequencing data from other projects to TCGA data, we have described the process that we developed and used to generate miRNA sequencing datasets for TCGA, and the computational pipeline with which we characterized miRNA expression. In order that readers have access to a more detailed description of the nature of the TCGA miRNA data, which emphasizes mature strands and isomiRs, we have expanded on material that has been published in Methods sections in TCGA marker papers, and describe the pipeline in the context of the overall data generation process. We designed the process for high-throughput processing across large cohorts, specificity for known miRNAs, and for stability and comparability across a multiyear project. Since we initially implemented the pipeline in 2009, we have developed library and computational processes, and automation, increasing our processing capacity by 10-fold, from several hundred samples per year to several thousand per year. We have used the production system to analyze human and mouse sequence data in projects involving as few as three samples, to projects as large as TCGA, for which, to date, we processed 11 506 miRNA samples and submitted sequence data for 11 259 samples across 33 cancer types. We have also analyzed 4049 miRNA sequencing libraries in other projects. The TCGA data that we generated have contributed to a number of publications (15–25), and are contributing to over ten TCGA projects that are underway at the time of writing. While we focus here on applying the pipeline to large-scale cancer datasets in the context of TCGA, the modular pipeline can run on any SAM-format read alignment file generated for any species that has a reference genome and annotated miRNAs, accessing annotations from a database or from user-created text files. The profiling software is freely available to enable rigorous and consistent comparisons to TCGA data.

AVAILABILITY

The profiling pipeline runs on Linux and Unix-like systems and is available at: http://www.bcgsc.ca/platform/bioinfo/software/mirna-profiling, and the adapter trimming software at: http://www.bcgsc.ca/platform/bioinfo/software/adapter-trimming-for-small-rna-sequencing.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Drs Ryan Morin and Martin Hirst for their input during the development of the profiling algorithm, Dr Jeremy Parker for reviewing miRNA profiles, Yisu Li for insightful discussions, The Cancer Genome Atlas Research Network for their collaborative contributions, and Payal Sipahimalani for project management. We are grateful to the BC Cancer Foundation and their donors for their support. MAM is a Scholar of the Canada Research Chairs Program.

FUNDING

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number U24CA143866. This work was also supported by Genome Canada and Genome BC through the Genomics Platform and Science and Technology Innovation Centre at the Genome Sciences Centre. The content is solely the responsibility of the authors. Funding for open access charge: NIH.

Conflict of interest statement. None declared.

REFERENCES

1. Gurtan,A.M. and Sharp,P.A. (2013) The role of miRNAs in regulating gene expression networks. *J. Mol. Biol.*, **425**, 3582–3600.
2. Hausser,J. and Zavolan,M. (2014) Identification and consequences of miRNA-target interactions–beyond repression of gene expression. *Nat. Rev. Genet.*, **15**, 599–612.
3. Friedlander,M.R., Chen,W., Adamidi,C., Maaskola,J., Einspanier,R., Knespel,S. and Rajewsky,N. (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.*, **26**, 407–415.
4. Axtell,M.J. (2013) ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA*, **19**, 740–751.
5. Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68–D73.
6. Rosenbloom,K.R., Armstrong,J., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Dreszer,T.R., Fujita,P.A., Guruvadoo,L., Haeussler,M. *et al.* (2015) The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.*, **43**, D670–D681.
7. Cunningham,F., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2015) Ensembl 2015. *Nucleic Acids Res.*, **43**, D662–D669.
8. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
9. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
10. de Hoon,M.J., Taft,R.J., Hashimoto,T., Kanamori-Katayama,M., Kawaji,H., Kawano,M., Kishima,M., Lassmann,T., Faulkner,G.J., Mattick,J.S. *et al.* (2010) Cross-mapping and the identification of editing sites in mature microRNAs in high-throughput sequencing libraries. *Genome Res.*, **20**, 257–264.
11. Gaujoux,R. and Seoighe,C. (2010) A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*, **11**, 367.
12. Lovmar,L., Ahlford,A., Jonsson,M. and Syvanen,A.C. (2005) Silhouette scores for assessment of SNP genotype clusters. *BMC Genomics*, **6**, 35.
13. Li,J. and Tibshirani,R. (2013) Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat. Methods Med. Res.*, **22**, 519–536.
14. Mullokandov,G., Baccarini,A., Ruzo,A., Jayaprakash,A.D., Tung,N., Israelow,B., Evans,M.J., Sachidanandam,R. and Brown,B.D. (2012) High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries. *Nat. Methods*, **9**, 840–846.
15. Cancer Genome Atlas Research Network. (2014) Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, **507**, 315–322.
16. Cancer Genome Atlas Network. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.
17. Cancer Genome Atlas Network. (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**, 330–337.
18. Cancer Genome Atlas Network. (2015) Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, **517**, 576–582.
19. Cancer Genome Atlas Research Network. (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, **489**, 519–525.
20. Cancer Genome Atlas Research Network. (2013) Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.*, **368**, 2059–2074.
21. Cancer Genome Atlas Research Network. (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, **499**, 43–49.
22. Cancer Genome Atlas Research Network. (2014) Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, **511**, 543–550.
23. Cancer Genome Atlas Research Network. (2014) Integrated genomic characterization of papillary thyroid carcinoma. *Cell*, **159**, 676–690.
24. Cancer Genome Atlas Research Network. (2014) Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, **513**, 202–209.
25. Cancer Genome Atlas Research Network, Kandoth,C., Schultz,N., Cherniack,A.D., Akbani,R., Liu,Y., Shen,H., Robertson,A.G., Pashtan,I., Shen,R. *et al.* (2013) Integrated genomic characterization of endometrial carcinoma. *Nature*, **497**, 67–73.