Bar-coding bias in high-throughput multiplex sequencing of miRNA

Shahar Alon[1,*], Francois Vigneault[2,3,4*], Seda Eminaga[2], Danos C. Christodoulou[2], J.G. Seidman[2], George M Church[2,3] and Eli Eisenberg[5]

1 Department of Neurobiology, George S. Wise Faculty of Life Sciences, Tel-Aviv University, Israel
2 Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
3 Wyss Institute for Biologically Inspired Engineering, Boston, MA 02115, USA
4 Ragon Institute of MGH, MIT and Harvard, Massachusetts, MA 02129, USA
5 Raymond and Beverly Sackler School of Physics and Astronomy, Tel-Aviv University, Israel

* These authors contributed equally to this work.
Correspondence should be addressed to: Eli Eisenberg, elieis@post.tau.ac.il

SUPPLEMENTARY TABLES AND FIGURES

Supplementary Tables

Table S1. The number of reads after each filtering step for all the sequencing runs performed.

	Mouse normal hearts using Ligation-based protocol	Mouse diseased hearts using Ligation-based protocol	Human brain using PCR-based protocol	Mouse heart libraries using PCR-based protocol
Minimum Hamming distance between bar-codes used	2	2	3	3
Total number of reads	17,738,028	20,237,297	10,561,029	23,266,377
Reads filtered as they didn't have full (perfect) bar-code sequence. In brackets the number of reads that would have been filtered if one mismatch was allowed.	3,566,055 (3,007,116*)	4,524,554 (3,847,623*)	2,078,569 (1,304,873)	3,948,153 (3,452,339)
Low quality reads filtered	4,200,531	5,875,118	114,680	1,730,425
Reads filtered as they were too short for mature miRNA (<15 bases)	1,738,139	1,934,818	1,114,522	468,619
Reads filtered as they were too long for mature miRNA (>28 bases)	35,694	38,578	2,009,661	204,524
Total number of reads after filtering	8,197,609	7,864,229	5,243,597	16,914,656
Number of filtered reads aligned against known pre-miRNA with up to two mismatches	5,077,357	5,042,301	1,246,440	7,582,091
Number of filtered reads aligned against known pre-miRNA with up to one mismatch	4,906,950	4,881,179	1,202,408	7,462,349

* - reads with sequence that can be attributed to two different bar-codes (when allowing one mismatch) have been filtered.

Table S2. Number of reads per bar-code (in millions) in the mouse heart libraries using the ligation-based protocol.

	Bar-code 1	Bar-code 2	Bar-code 3	Bar-code 4 *	Bar-code 5	Bar-code 6	Bar-code 7	Bar-code 8	Bar-code 9	Bar-code 10 *
Mouse normal hearts	1.4	1.4	0.82	0.03	1.4	0.59	1.4	0.74	0.08	0.31
Mouse diseased hearts	1.3	1.2	0.80	0.42	1.3	0.77	1.2	0.74	0.09	0.001

* - these bar-codes were excluded from further analysis because they exhibit an exceptionally low number of reads in one of the tissues.

Table S3: Number of reads per bar-code (in millions) in the human brain library.

Bar-code 1	Bar-code 2	Bar-code 3	Bar-code 4	Bar-code 5	Bar-code 6 *	Bar-code 7	Bar-code 8	Bar-code 9	Bar-code 10	Bar-code 11 *	Bar-code 12
0.63	0.63	0.48	0.35	0.53	0.0023	0.30	0.43	0.65	0.64	0.18	0.41

* - these bar-codes were excluded from further analysis because they exhibit an exceptionally low number of reads.

Table S4. Number of reads per bar-code (in millions) in the mouse heart libraries using the PCR-based protocol.

	Bar-code 1	Bar-code 2	Bar-code 3	Bar-code 4	Bar-code 5	Bar-code 6	Bar-code 7	Bar-code 8
Condition	Mouse normal hearts Sample 1	Mouse normal hearts Sample 2	Mouse normal hearts Sample 3	Mouse diseased hearts Sample 4	Mouse diseased hearts Sample 5	Mouse diseased hearts Sample 6	Mouse normal hearts Pool of Samples 1-3	Mouse diseased hearts Pool of Samples 4-6
Number of reads	1.9	1.9	2.6	2.3	2.6	2.1	1.6	1.9

Table S5. A list of differentially expressed miRNAs between normal and diseased mouse hearts demanding a stringent cutoff of 2-fold change.

miRNA name	Fold change comparing normal hearts samples to diseased hearts samples	Fold change comparing normal hearts pool to diseased hearts pool	Higher expression in normal or in diseased mouse hearts
mmu-mir-547	3.6	4.3	Higher in diseased
mmu-mir-21	3.9	3.4	Higher in diseased
mmu-mir-376a	2.4	2.2	Higher in diseased
mmu-mir-499	2.2	2.1	Higher in normal
mmu-mir-185	2.0	2.1	Higher in normal
mmu-mir-150	2.1	2.0	Higher in normal
mmu-mir-132	2.0	2.1	Higher in diseased
mmu-mir-10b	2.1	2.0	Higher in diseased

Supplementary Figures

Figure S1: Possible use of ligation-based bar-coding data. Same as Figure 1a but comparing counts-number for normal and diseased mouse heart using the same bar-code. As long as the same bar-code is used, one finds that only 20% of the miRNAs are differentially expressed (out of the Poisson noise region). Half of these miRNAs were detected consistently using every bar-code. A similar calculation for different bar-codes but the same biological tissue resulted in an erroneous detection of 10-40% differentially expressed miRNAs due to the bar-code bias.

Figure S2: Illumina multiplexing kit is incompatible with miRNAs sequencing protocol. In an attempt to use Illumina indexing kit (http://www.illumina.com/products/multiplexing_sample_preparation_oligonucleotide_kit.ilmn), which introduces bar-codes during the PCR step, we adenylated (rApp) the appropriate compatible primer sequence (Illumina standard paired-ends oligonucleotide), so that it could be used for microRNA capture and sequencing (see Vigneault *et al*., 2008 for details regarding adenylation and microRNA ligation efficiency assay). Ligation of these oligonucleotides to miRNAs (lane 3) results in poor ligation efficiency and creation of ligation artifacts rendering post-library processing impractical compared to the use of known positive control oligonucleotides such as rApp Illumina v1.5 microRNA adapter (lane 2 – incompatible with multiplexing) as well as rApp IDT miRNA cloning linker-1 (lane 4, http://www.idtdna.com/catalog/smallRNAcloning/Page1.aspx). Our own no-bias bar-code PCR compatible oligonucleotide (lane 5 – compatible with multiplexing) was able to ligate properly to miRNA without ligation artifacts.

Figure S3: The PCR-based protocol for bar-code introduction allows reliable comparison between biological samples. For each miRNA, the average of the counts in three different samples of normal mouse heart was compared to the counts in a pooled sample. As expected for bias-free bar-codes, 99% of all points fall inside the Poisson noise region.

Figure S1

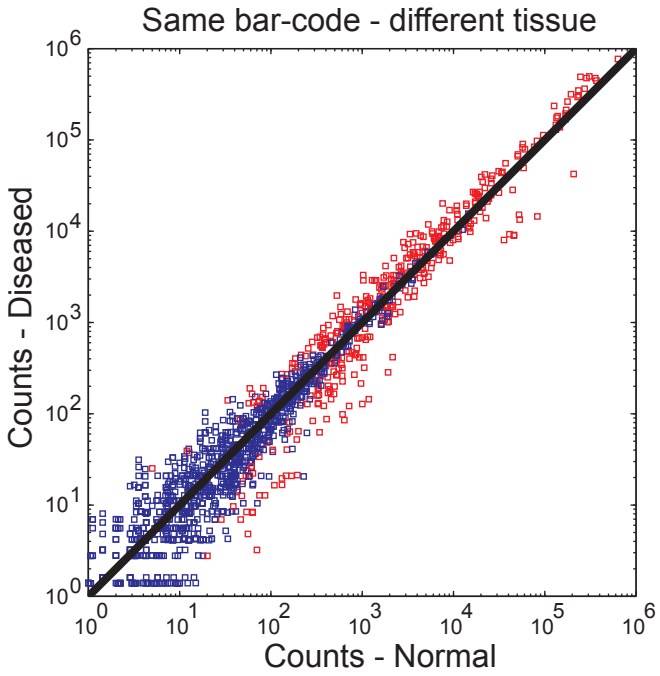

Same bar-code - different tissue
Counts - Diseased
Counts - Normal
10^0
10^1
10^2
10^3
10^4
10^5
10^6

Figure S2

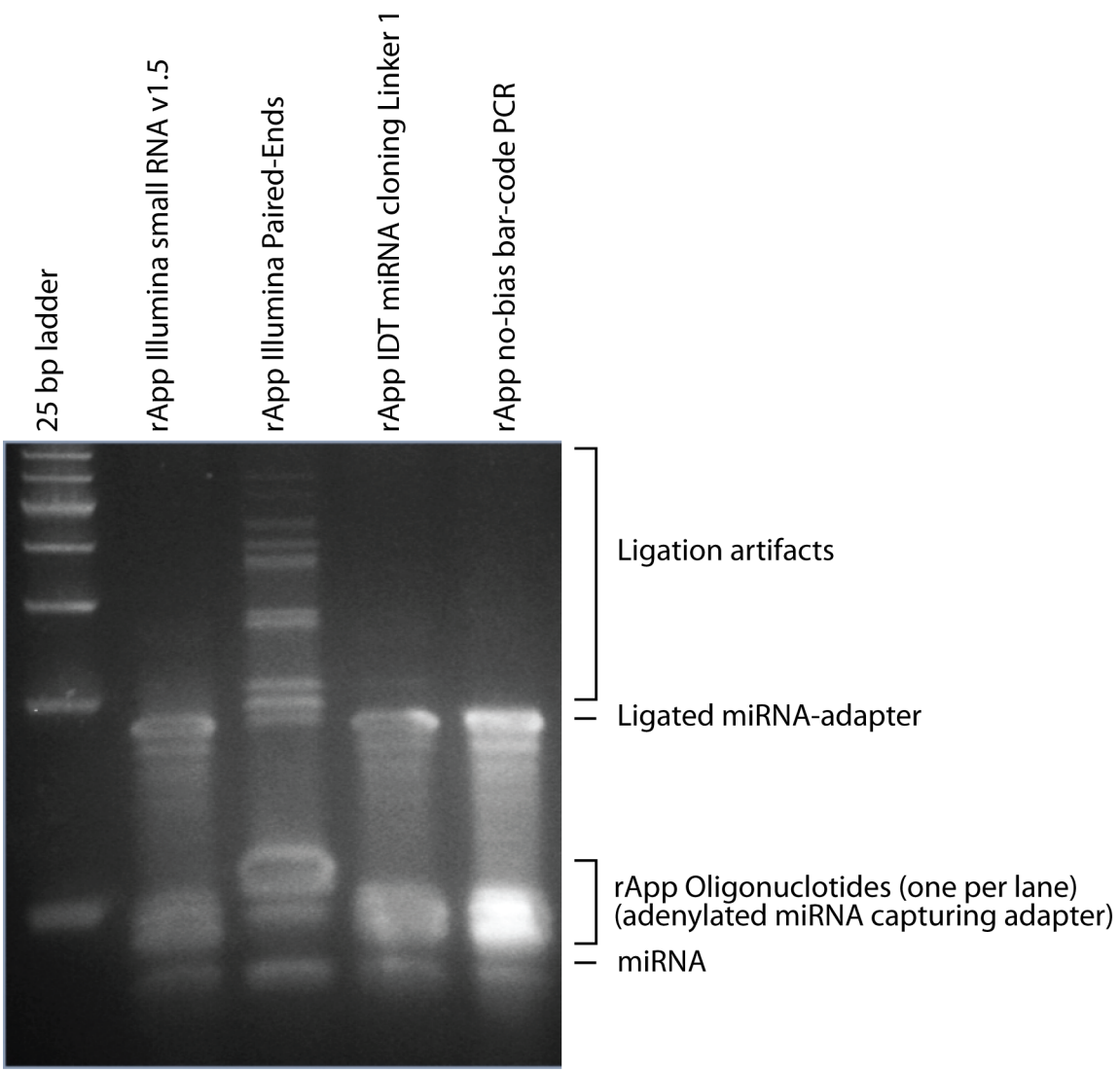
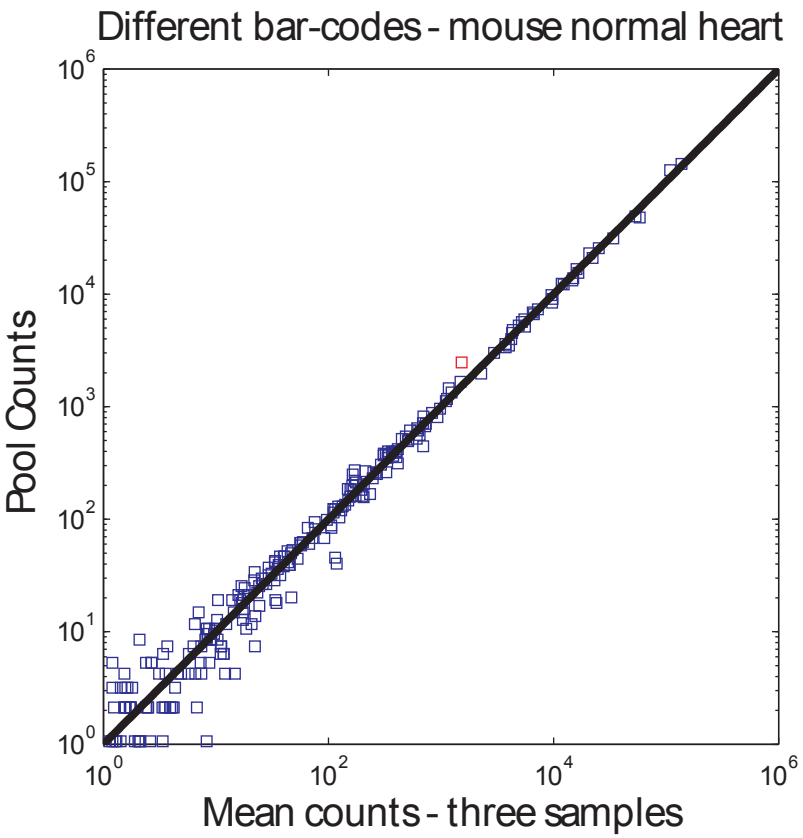

25 bp ladder
rApp Illumina small RNA v1.5
rApp Illumina Paired-Ends
rApp IDT miRNA cloning Linker 1
rApp no-bias bar-code PCR
Ligation artifacts
Ligated miRNA-adapter
rApp Oligonuclotides (one per lane)
(adenylated miRNA capturing adapter)
miRNA
lane number: 1 2 3 4 5

Figure S3


Different bar-codes - mouse normal heart
Pool Counts
Mean counts - three samples
10^0
10^1
10^2
10^3
10^4
10^5
10^6