An Expression Atlas of small RNAs generated using the SOLiD™ Sequencing System

Scott Kuersten[1], Jian Gu[1], Catalin Barbacioru[2], Xiaohui Wang, Nicole Cloonan[3], Brooke Gardiner[3], Kristi Lea[1], Sheila Heater[1], Melissa Barker[2], Laura Chapman[1], Thomas Gulham[2], Linda Wong[2], Sean Grimmond[3]. Applied Biosystems, [1]2130 Woodward St., Austin, TX and [2]800 Lincoln Centre Dr., Foster City, CA. [3]Institute for Molecular Bioscience, The University of Queensland, St Lucia QLD 4072, Australia.

Introduction

We have developed a novel and simple ligation-based method (commercialized as the SOLiD™ Small RNA Expression Kit, or SREK) to capture and amplify small noncoding RNAs for sequencing using the SOLiD™ Sequencing System to generate 25-35 base sequence tags. The sensitivity and precision provided by the analysis of these sequences promises to change and improve our understanding of the complexity of miRNAs structure and function.

We barcoded and sequenced small RNA libraries from ten different human tissues to saturating levels of detection; generating up to 200 million total mappable sequence tags in a single run. Comparing both independent sequencing runs and independent libraries indicates the system is highly reproducible and capable of up to 6 logs of dynamic range of detection. Fold-change expression of miRNAs between SOLiD and Taqman® qPCR datasets demonstrate correlation values of up to 0.94, indicating the system is a valid and accurate profiling tool. We detect a far greater repertoire of miRNA variants or 'isomirs' than previously observed suggesting a much broader range of mRNA targets for regulation.

We have also developed a Support Vector Machine to predict novel miRNAs contained within the SOLiD datasets and have identified many potentially novel sequence tags. To a subset of these custom TaqMan® miRNA assays were designed to validate by real-time PCR analysis and demonstrate both the presence and profile of >50% of the novel sequences.

This human miRNA expression atlas provides a unique opportunity to understand the sequence complexity and identity of small noncoding RNAs present in human tissues.

Table 1: ***A human small RNA tissue atlas.*** **A summary of the number of tags and distribution of reads obtained from ten distinct human tissue samples obtained with the equivalent of 1.5 instrument runs. The csfasta files from the SOLiD Sequencing System were mapped using the RNA2MAP tool (www.solidsoftwaretools.com) for small RNAs and the number of tags that map to known miRNAs, tRNAs/rRNAs/repetitive elements or to the genome are indicated. The relative proportion of reads that map to known miRNAs differ significantly between tissues, ranging from ~20-63% of the mappable data. This may reflect the level of miRNA regulation or complexity required in these tissue types.**

Tissue	Barcode	slide	length	miRNA	% miRNA	genome (uniq)	genome (mult)	rRNA, tRNA, rep	Empty	mapped	Total tag counts	
Placenta	N/A	N/A	35	36,019,875	63.0%	4,049,050	12,141,623	4,464,562	467,867	57,142,977	86,298,208	
Heart	BC1	1	25	5,570,533	50.5%	386,416	3,549,755	2,130,880	2,956	11,640,540	15,125,795	23,072,667
		2	35	3,554,298		362,202	898,018	1,627,742	432	6,442,692	7,946,872	
Brain	BC2	1	25	5,515,954	49.7%	422,372	3,151,055	2,488,133	3,976	11,581,490	14,636,256	23,397,311
		2	35	3,763,415		421,896	965,000	1,922,879	733	7,073,923	8,761,055	
Liver	BC3	1	25	1,466,692	23.2%	512,530	2,751,481	2,389,409	51,228	7,171,340	10,613,941	17,773,089
		2	35	1,066,724		285,380	537,050	1,830,431	12,346	3,731,931	7,159,148	
Testes	BC4	1	25	2,187,097	33.4%	276,376	2,586,102	1,657,420	4,861	6,711,856	8,577,871	14,134,716
		2	35	1,465,256		667,769	731,261	1,347,745	218	4,212,249	5,556,845	
Spleen	BC5	1	25	2,035,620	29.3%	806,692	3,205,382	1,990,267	145,184	8,183,145	13,207,827	21,610,493
		2	35	1,451,102		269,685	448,741	1,544,771	9,871	3,724,170	8,402,666	
Kidney	BC6	1	25	4,089,418	48.2%	327,842	2,695,267	1,774,155	4,176	8,890,858	11,493,615	18,134,969
		2	35	2,712,018		331,484	774,607	1,415,604	586	5,234,299	6,641,354	
Thymus	BC7	1	25	4,106,372	47.2%	501,144	2,583,754	2,064,285	20,813	9,276,368	12,495,092	20,185,960
		2	35	2,752,681		284,508	596,810	1,606,851	1,082	5,241,932	7,690,868	
Lung	BC8	1	25	4,061,447	62.6%	264,283	1,490,522	886,328	42,281	6,744,861	8,819,784	14,724,200
		2	35	2,901,002		250,256	550,856	670,192	3,104	4,375,410	5,904,416	
Ovary	BC9	1	25	1,094,096	21.1%	334,655	4,211,053	130,894	175,221	5,945,919	12,958,296	20,983,757
		2	35	750,448		666,118	1,274,042	97,602	26,201	2,814,411	8,025,461	
				86,564,048		11,420,658	45,142,379	32,040,150	973,136	176,140,371	260,315,370	Totals

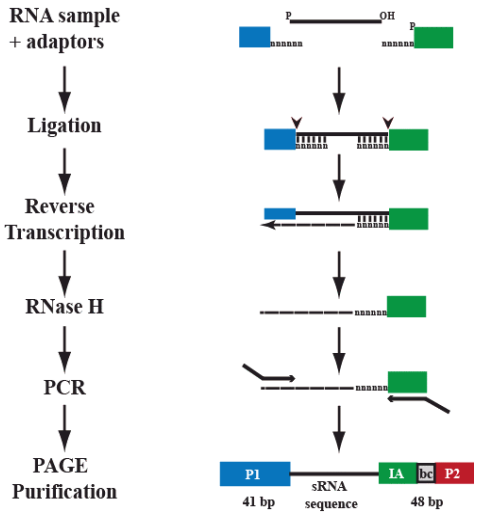

Figure 1: ***Method for library construction using the SOLiD™ Small RNA Expression Kit (SREK):*** **FlashPAGE™ Fractionator purified small RNAs (~18-40 nt, 100 ng) from ten human tissues were combined with a mixture of specific 5' and 3' adaptors and ligated in a single step. The ligated templates are converted to cDNA, digested with RNase H and then amplified by PCR. Different tissue samples were amplified with PCR primers containing a unique 6 nt barcode (bc) between the 3' internal adaptor (IA) and P2. Amplified products were then purified, quantitated and equal amounts pooled, subjected to emulsion PCR and sequenced using standard SOLiD sequencing chemistry.**

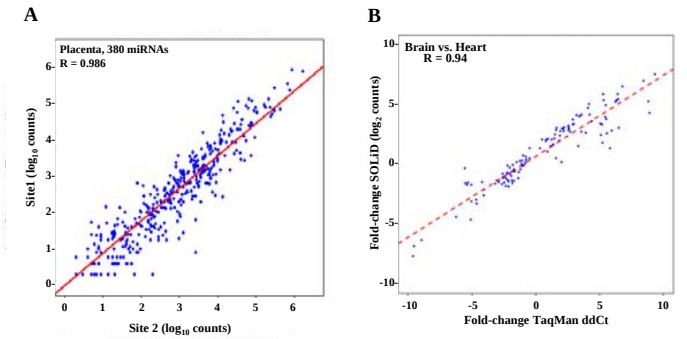

Figure 2: ***Reproducibility and correlation with qPCR suggest SREK and SOLiD are valid approaches for profiling miRNAs.*** **A) Two independent placenta library preps using either total RNA or FlashPAGE™ purified small RNA were constructed and sequenced independently on two different instruments and tag counts of 380 known miRNAs were compared. The data demonstrate that both the library and instrument performance are highly reproducible. B) Quantitative fold-chance comparison between SOLiD and qPCR were compared using Brain and Heart datasets. A correlation value of 0.94 suggest strong concordance between the two methods.**

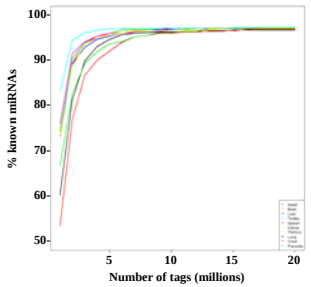

Figure 3: ***The number of tags required to saturate detection of miRNAs in a sample..*** **The level of complexity within the ten tissue samples were calculated by clustering and identifying the number of non-redundant miRBase tags captured at 1 million read intervals. Using a minimum filter of at least 3 reads establishing detection, approximately 8-10 million tags are needed to detect the maximum number (~95%) of known miRNAs in each tissue. This suggest a surprisingly high degree of miRNA complexity in all tissues.**

Figure 4: ***Expression of miRNA variants or isomirs are widespread and highlight the diversity of small RNAs and their potential targets.*** **The expression profile of the canonical (blue) and all isomirs originating from the miR-214 precursor miRNA (red) found in testes. In general, 3' isomirs are more common and diverse. However the sequence differences seen with some 5' isomirs significantly change the seed region of the miRNA, known to significantly contribute to mRNA target recognition. This suggests that isomirs generated from the same precursor miRNA may have different mRNA targets thus expanding the range and complexity of miRNA function. Also, this example shows the canonical miRNA for miR-214 is not detected in this tissue and the most prevalent miRNAs are from the so-called 'star' form. The role this diversity plays in gene regulation is not understood at this time.**

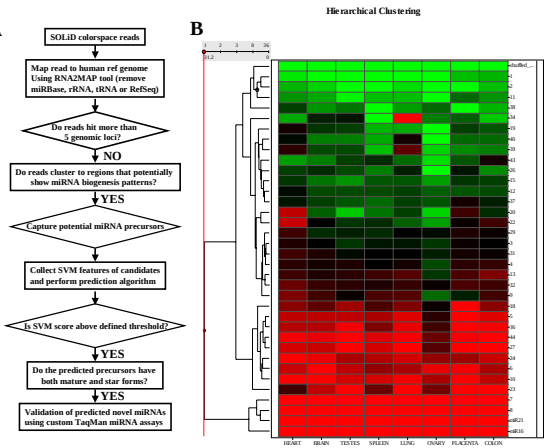

Figure 5: ***An SVM algorithm to search for novel miRNAs and validation using custom TaqMan® assays.*** **A) General overview of the SVM approach. B) Forty custom TaqMan® miRNA assays were designed and tested against eight tissue samples and 33 novel RNAs were validated. Their normalized expression levels were clustered; red = high, black=medium & green = low expression levels.**

Conclusions:

- Novel and simple library construction.
- Barcoding allows multiplexing and lowers cost per sample.
- Libraries are highly reproducible.
- Sequencing data is validated byTaqman.
- Isomir diversity is high and can impact biological function.
- Many potential novel small RNAs to discover.