Sequencing Data Report
microRNA Sequencing Discovery Service

On

G2

For

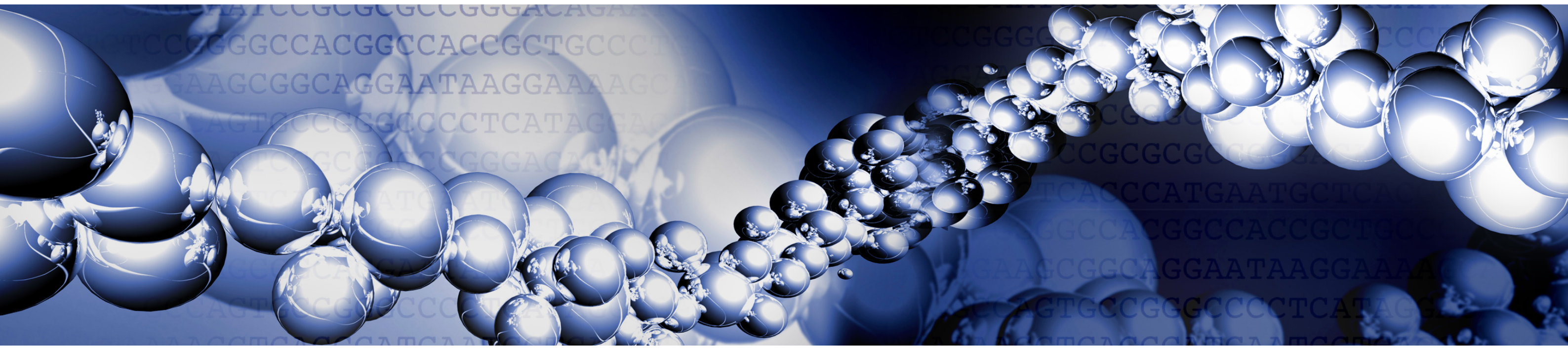
Dr. Peter Nelson

Sanders-Brown Center on Aging
University of Kentucky

Prepared by

LC Sciences, LLC

June 15, 2011



Prepared by LC Sciences, LLC | http://www.LCSciences.com | Email: support@LCSciences.com

2575 W. Bellfort, Suite 270, Houston, Texas 77054
Tel. (713)664-7087, Fax (713)664-8181

I. PROJECT INFORMATION

Table 1. Sample, service and project tracking information

A: Project information	
Customer Sample Name	G2
Sample Species	*Homo sapiens*
Sample Received Date	03/21/2011
Service Requested	microRNA Discovery Sequencing Service
LCS Project Number	4372
LCS Sample ID	G2

B: Database information		
Reference or Database Sequences	WEBlink and Information	Version or Built Date
miRNA(miRs) database	ftp://mirbase.org/pub/mirbase/CURRENT/; Specific species: hsa; Selected species: ssc, cfa, mdo, age, lla, sla, mml, mne, pbi, ggo, ppa, ptr, ppy, ssy, lca, oan, cgr, mmu, rno, bta, eca, oar	v17.0
Pre-miRNA(mirs) database	ftp://mirbase.org/pub/mirbase/CURRENT/; Specific species: hsa; Selected species: ssc, cfa, mdo, age, lla, sla, mml, mne, pbi, ggo, ppa, ptr, ppy, ssy, lca, oan, cgr, mmu, rno, bta, eca, oar	v17.0
Genome database	ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/	37.1
mRNA database	ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/RNA/	37.1
Customer database	NA	NA

II. DATA REPORT

A. Terminologies Used

Table 2. Terminologies used in data analysis

Term	**Description**
Copy Number or Count	Number of sequ seqs in the same unique seq family
Mapping	Blasting a sequence to a reference database
miRBase	A searchable database of published miRNA sequences and annotation; http://mirbase.org
mir	Pre-miRNA registered in miRBase
miR	Mature miRNAs registered in miRBase
RepBase	Prototypic sequences representing repetitive DNA from different eukaryotic species; http://www.girinst.org/repbase
RFam	Collection of many common non-coding RNA families except micro RNA; http://rfam.janelia.org
Reads	DNA sequences from reading of sequencing intruments
Sequ Seq or Reads	Raw sequencing reads generated in after image extraction and base-calling
Unique Seq	Family of sequ seq with same sequence
Selected species	A combination of species defined by user
Specific species	Species of the sample analyzed

B. Methods and Procedures

The received RNA sample was processed to generate a cDNA library which was then used to deep sequencing. The data generated were analyzed and the full data files were saved onto a DVD disc which is included in this report. Experimental procedures and analysis methods are briefly presented here and detailed descriptions are documented in Appendix I.

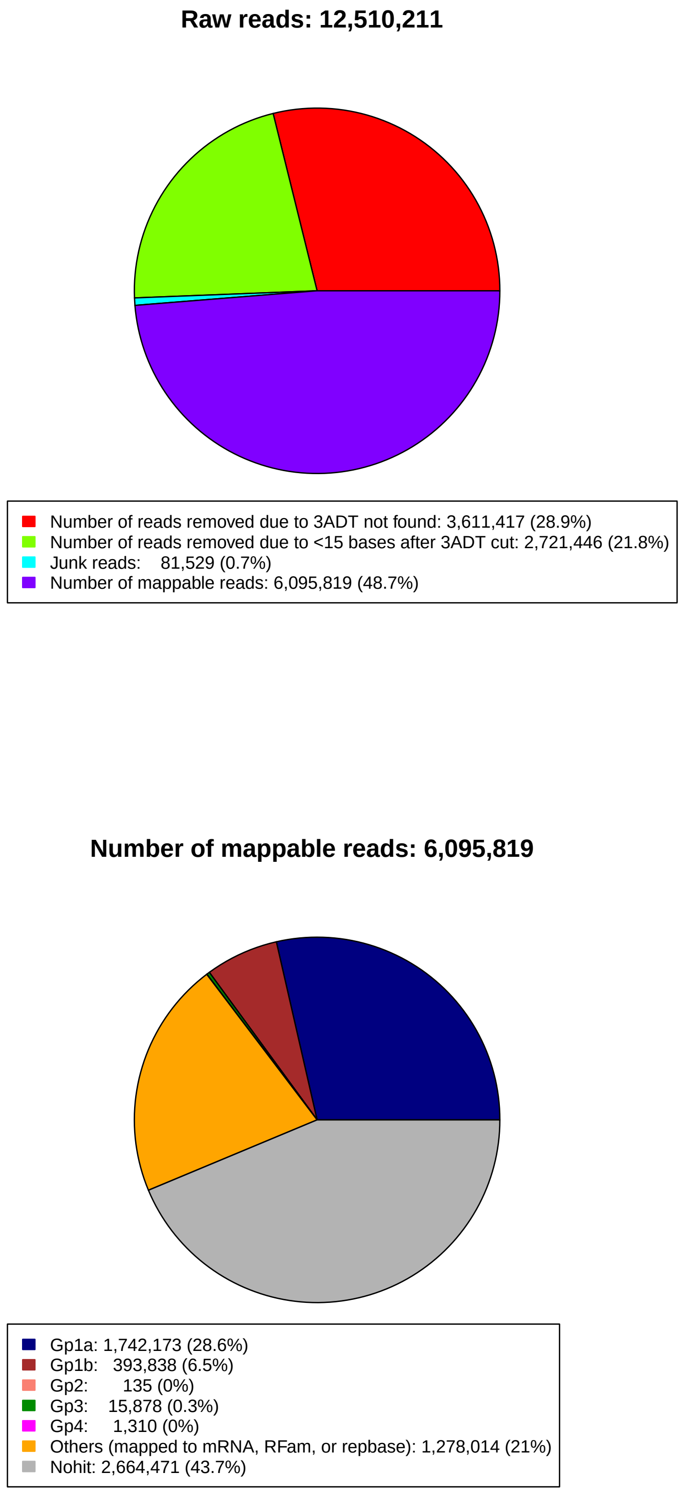


Figure 1. Pie plots of data filtering and database mapping

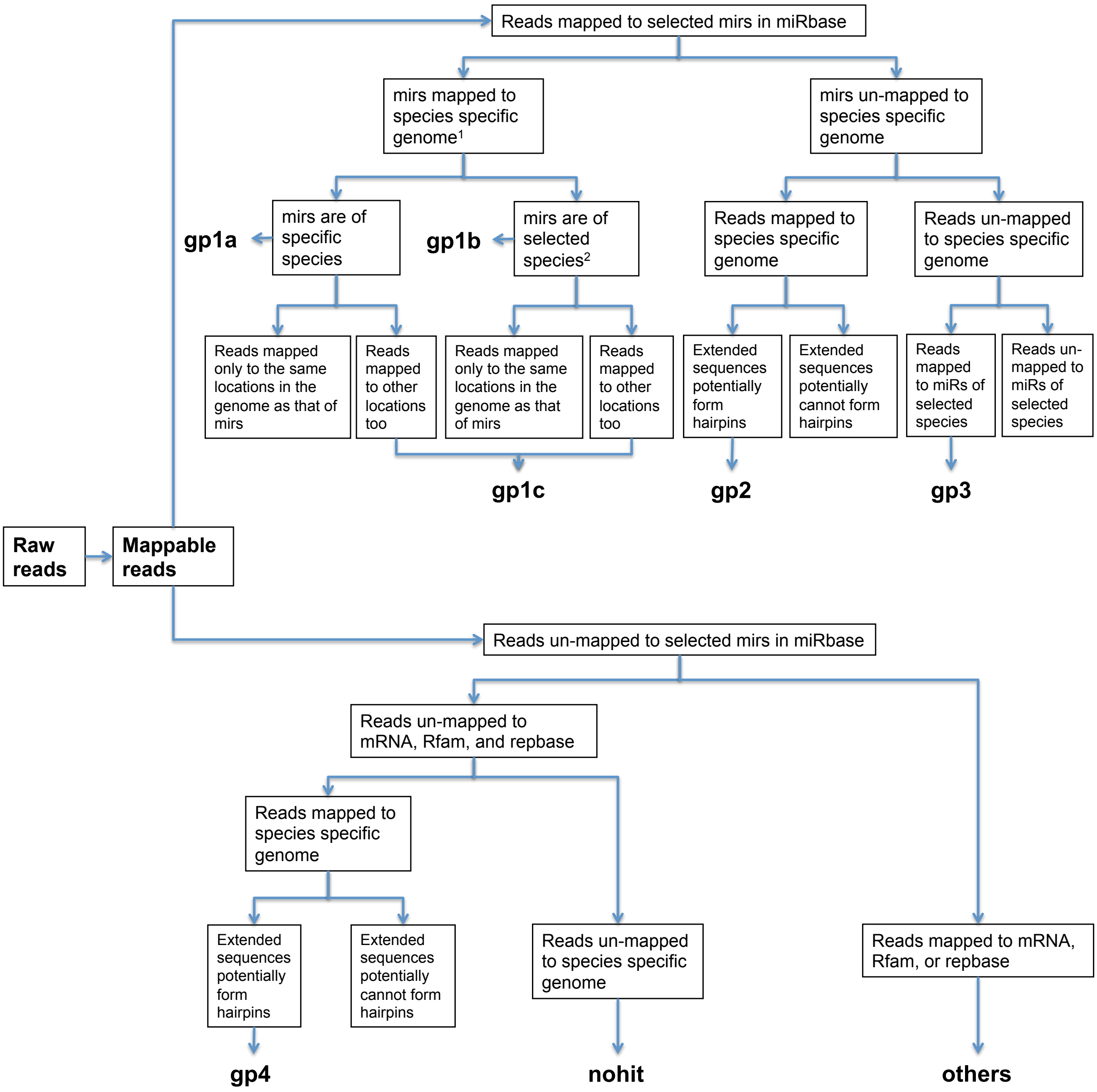

Reads mapped to selected mirs in miRbase
mirs mapped to species specific genome¹
mirs un-mapped to species specific genome
gp1a
mirs are of specific species
gp1b
mirs are of selected species²
Reads mapped to species specific genome
Reads un-mapped to species specific genome
Reads mapped only to the same locations in the genome as that of mirs
Reads mapped to other locations too
Reads mapped only to the same locations in the genome as that of mirs
Reads mapped to other locations too
Extended sequences potentially form hairpins
Extended sequences potentially cannot form hairpins
Reads mapped to miRs of selected species
Reads un-mapped to miRs of selected species
gp1c
gp2
gp3
Raw reads
Mappable reads
Reads un-mapped to selected mirs in miRbase
Reads un-mapped to mRNA, Rfam, and repbase
Reads mapped to species specific genome
Extended sequences potentially form hairpins
Extended sequences potentially cannot form hairpins
Reads un-mapped to species specific genome
Reads mapped to mRNA, Rfam, or repbase
gp4
nohit
others

Figure 2. Data analysis flowchart

[1] *Homo sapiens*

[2] Mammalian

III. DATA SUMMARIES

A. List of Data Files

Table 4. Data files delivered and programs recommended for reviewing

Folder	Data Files	Reviewing Program
1_RawData	G2_RawData.fq (zipped)	Wordpad
2_ProcessedData	G2_unique.fq	Wordpad
	G2_mappableReads.fq	Wordpad
3_MappedData	G2_gp1a_aln.txt	Wordpad
	G2_gp1a_sum.txt	Excel
	G2_gp1b_aln.txt	Wordpad
	G2_gp1b_sum.txt	Excel
	G2_gp2_aln.txt	Wordpad
	G2_gp2_sum.txt	Excel
	G2_gp3_aln.txt	Wordpad
	G2_gp3_sum.txt	Excel
	G2_gp4_aln.txt	Wordpad
	G2_gp4_sum.txt	Excel
	G2_mir_aln.txt	Wordpad
	G2_mir_sum.txt	Excel
	G2_others.txt	Excel
	G2_nohit.fq	Wordpad
4_Summary	G2_uni_miRs.txt	Excel

B. Length Distribution of Reads after 3ADT cut

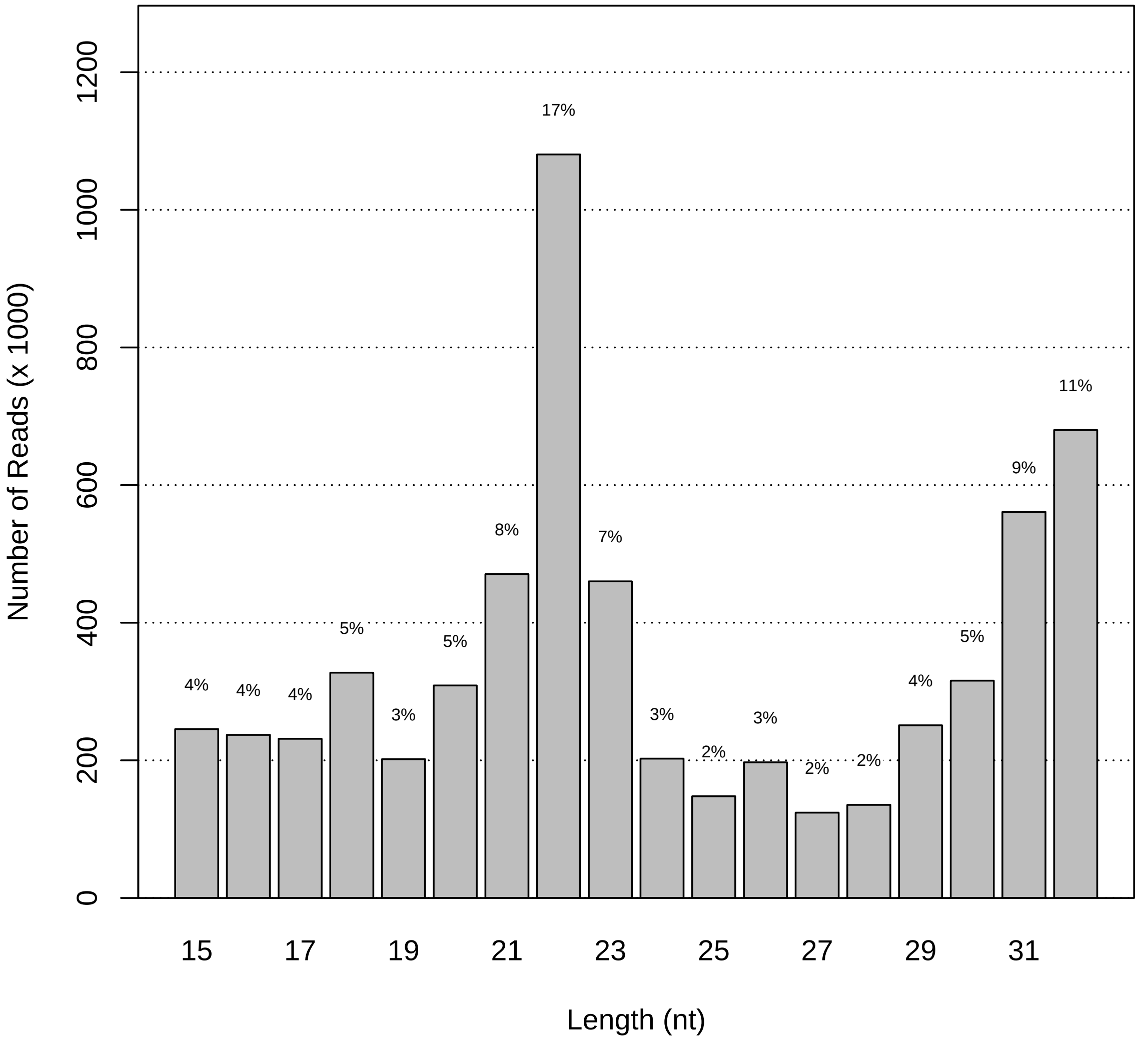


Figure 3. Length distribution of reads after 3ADT cut

Table 5. Length distribution of reads after 3ADT cut

Length	#SequSeq	%Total
15	245,414	4%
16	237,045	3.8%
17	231,260	3.7%
18	327,351	5.3%
19	201,740	3.3%
20	308,794	5%
21	470,625	7.6%
22	1,080,496	17.5%
23	460,137	7.4%
24	202,489	3.3%
25	147,791	2.4%
26	197,065	3.2%
27	124,010	2%
28	135,324	2.2%
29	250,878	4.1%
30	315,826	5.1%
31	561,145	9.1%
32	679,958	11%
Total	6,177,348	100%

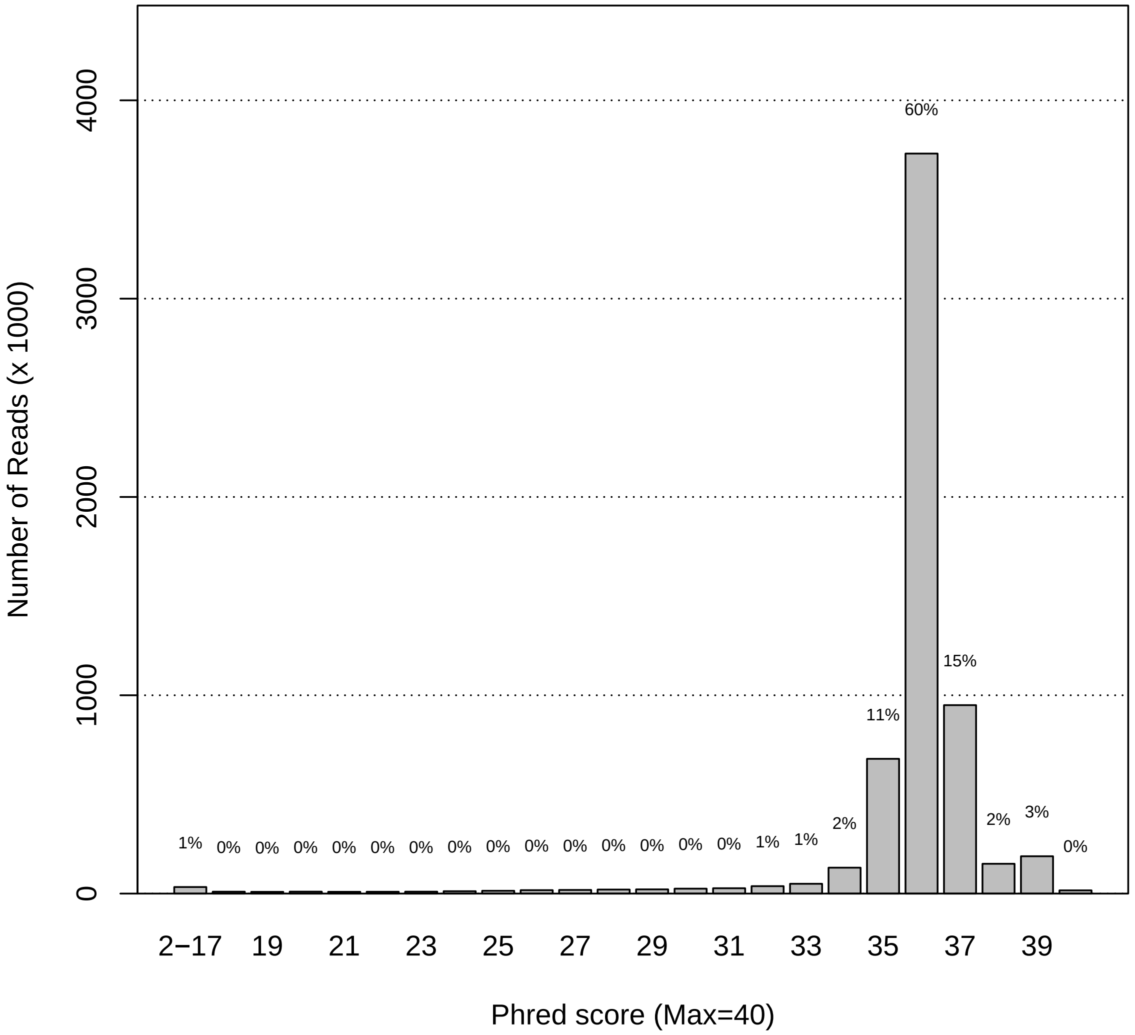


Figure 4. Histogram of the average phred score[1] per base in a read after 3ADT cut

[1] Phred score larger than 30 stands for probability of incorrect base calls less than 1 in 1,000 (above 99.9% accuracy) in one sequencing read.

C. Results Summary

Table 6. A summary of standard data analysis results

	#Seqseq	%Mappable SequSeq
Raw	12,510,211	
Total mappable reads	6,095,819	100%
Group 1a	1,742,173	28.6%
Group 1b	393,838	6.5%
Group 1c	1,058,650	17.4%
Group 2	135	0%
Group 3	15,878	0.3%
Group 4	1,310	0%
Mapped to mRNA	545,346	8.9%
Mapped to other RNAs (RFam: rRNA, tRNA, snRNA, snoRNA and others)	1,252,583	20.5%
Mapped to Repbase	35,325	0.6%
Mapped to custom database if applicable	0	0%
Nohit	2,664,471	43.7%

Table 7. Known and predicted miRs

	Group	#Unique miRs
Known miRs		
of specific species[1]	Group 1a	951
of selected species[2], but novel to specific species	Group 1b	145
of specific and selected species, but with new genome[1] locations	Group 1c	589
Predicted miRs		
Mapped to known mirs of selected species and genome; within hairpins	Group 2	23
Mapped to known miRs of selected species but unmapped to genome	Group 3	162
UmMapped to known miRs but mapped to genome and within hairpins	Group 4	383
Overall (Unique miRs)		1,442

[1] *Homo sapiens*

[2] Mammalian