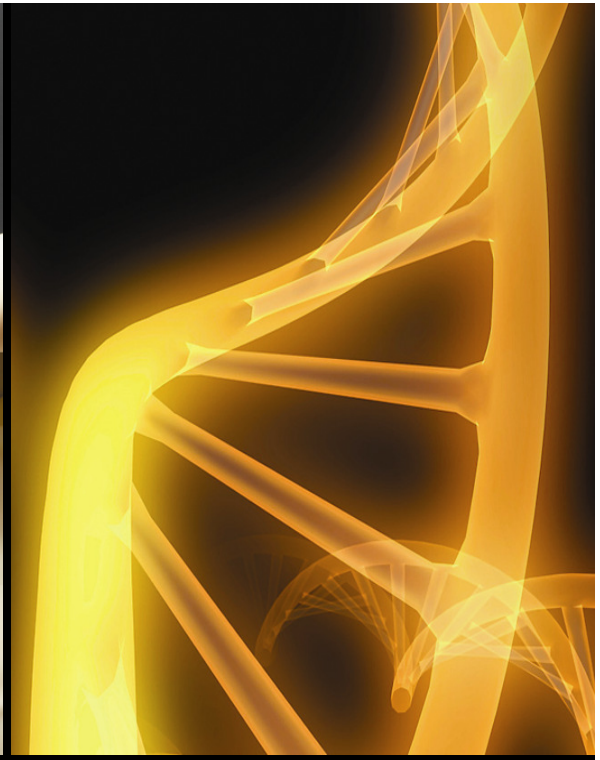
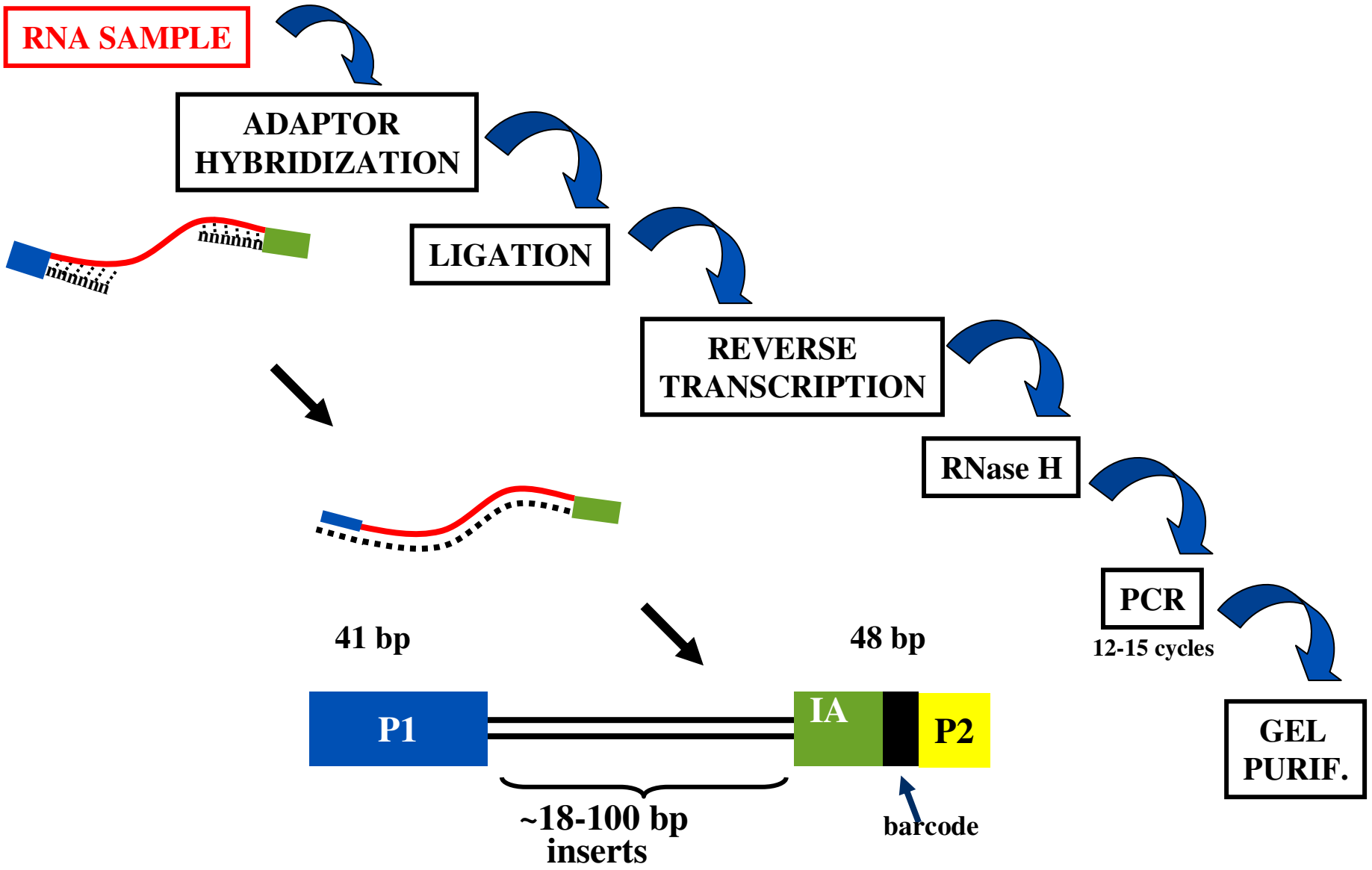
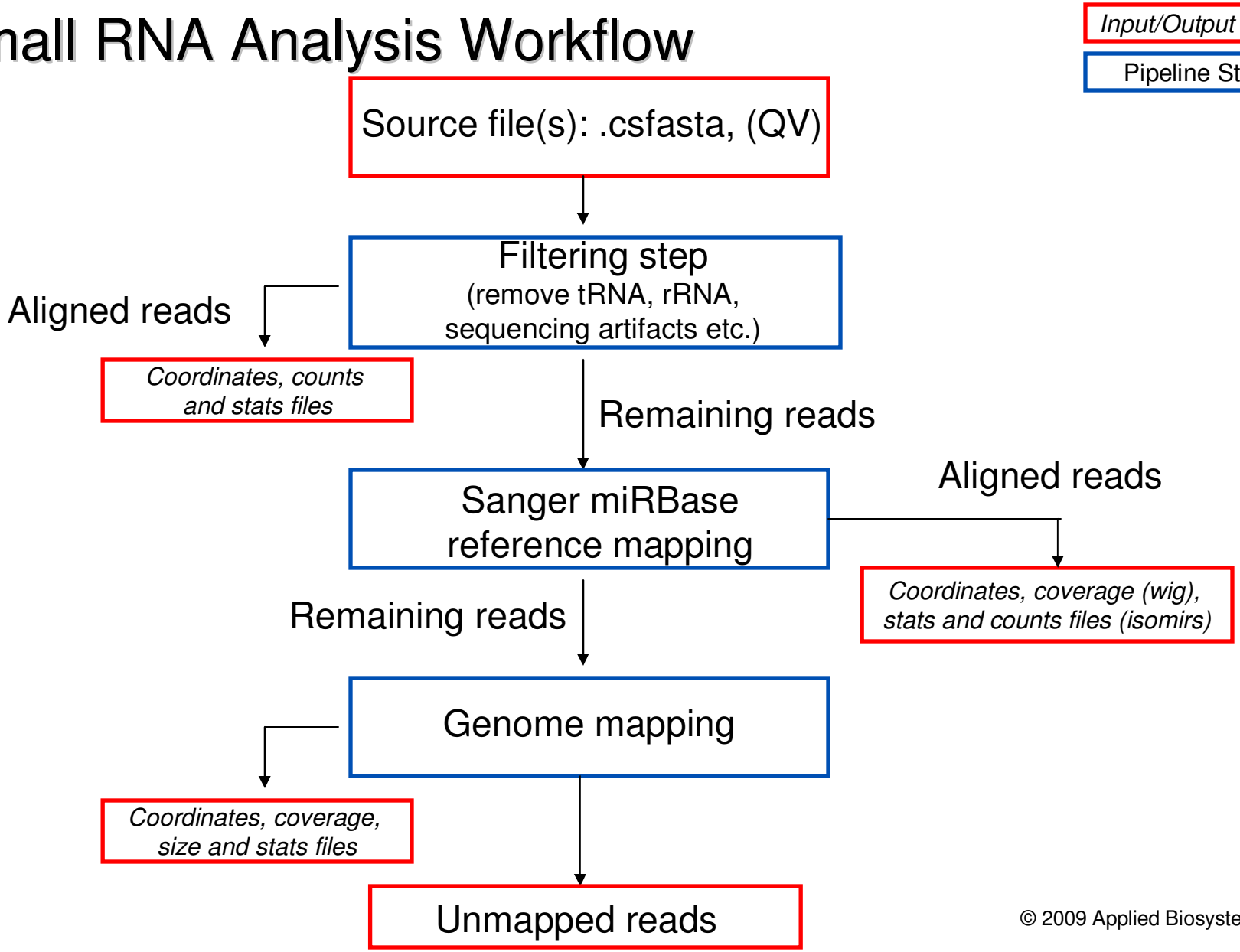


Applied Biosystems

Small RNA Pipeline

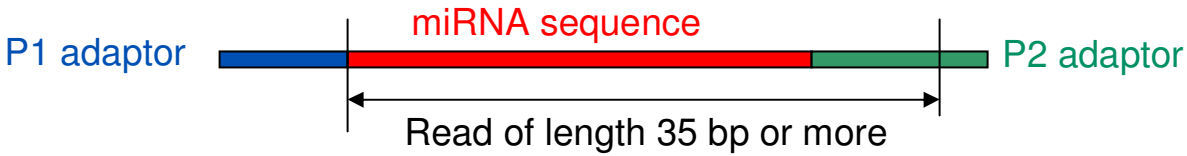


Small RNA Analysis Workflow



Data Analysis Issues

- Small RNA and transcript fragments are generally smaller than the read length Therefore, the reads the SOLiD system is producing contain small RNA fragment + P2 adaptor sequence.



- The current matching tools used on the SOLiD instrument are designed for reads of equal length.

Mapping Approach – Seeding Step

• Map first 18 colors of the read to the reference with mapreads. This is the ‘seeding step’. The initial seed length of 18 is a definable parameter.

masked positions

READ (color space)

T13300022121022213 2……

133000221210222132……

REFERENCE (color space)

• Extend mapping seed and test all adaptor start points. The one giving the fewest number of mismatches is considered to be the beginning of the adaptor

Mapping Approach – Extension Step

Read: T133000221210213222031223302010303
Genome: T133300221210213222031220122021121
Adaptor: C33020103031312312

Start	Alignment	MM
19	Read: T133000221210213222031223302010303 3 Genome: T133300221210213220330201030313112 3	14
20	Read: T133000221210213222031223302010303 3 Genome: T133300221210213222330201030313112	12
	⋮	
24	Read: T133000221210213222031223302010303 3 Genome: T133300221210213222031230201030313	10
25	Read: T133000221210213222031223302010303 3 Genome: T133300221210213222031203020103031	3
26	Read: T133000221210213222031223302010303 3 Genome: T133300221210213222031220302010303	11

Remarks

- Making use of the full length reads allows for more accurate template-adapter junction identification
- The filtering and miRBase reference files used are relatively small compared to human genome, this reduces computational time
- The stepwise approach generates a higher number of “uniquely” mapped reads than a single step using only the human genome

Usage

- Modify configuration file appropriately
  - Set of name <tab> value pairs that define parameters
- Generate shell scripts for job processing
  - RNA_matching_analysis_pipeline.pl
    - -r = .csfasta file
    - -c = configuration file
    - -corona = path to RNA pipeline package
    - -o = output directory
- Submit jobs to cluster
  - submit_scripts_to_PBS.pl
    - -j = JOB_LIST.txt
  - Also supports LSF and SGE schedulers

Output

- Four main output directories
  - filter – created from the filtering step
  - miRBase – created from the miRBase mapping step
  - genome – created from genome mapping step
  - scripts – contains shell scripts executed by pipeline
- Files
  - Statistics (.stats / _extend.stats) – mapping statistics
  - Counts (_extended.gff) – gff file containing mapping locations and number of hits
  - Mapping (_extend.ma) – mapping files for “seed” and “extension” steps
  - WIG (.wig) – coverage file for visualization with UCSC browser

Example Output Files: _extend.stats

- One file generated for each step of the pipeline

Filter		miRBase		Genome	
24268128 total beads found		24268128 total beads found		24268128 total beads found	
Total Beads		Total Beads		Total Beads	
0MM 816695 (3.37%)		0MM 527727 (2.17%)		0MM 185286 (0.76%)	
1MM 2133894 (8.79%)	2950589 (12.16%)	1MM 202812 (0.84%)	730539 (3.01%)	1MM 378107 (1.56%)	563393 (2.32%)
2MM 1475858 (6.08%)	4426447 (18.24%)	2MM 484957 (2.00%)	1215496 (5.01%)	2MM 620037 (2.55%)	1183430 (4.88%)
		3MM 315935 (1.30%)	1531431 (6.31%)	3MM 539843 (2.22%)	1723273 (7.10%)
		4MM 344281 (1.42%)	1875712 (7.73%)	4MM 503041 (2.07%)	2226314 (9.17%)
Uniquely Placed Beads		5MM 245494 (1.01%)	2121206 (8.74%)	5MM 490340 (2.02%)	2716654 (11.19%)
0MM 682791 (2.81%)		6MM 176314 (0.73%)	2297520 (9.47%)	6MM 583802 (2.41%)	3300456 (13.60%)
1MM 2060934 (8.49%)	2743725 (11.31%)				
2MM 1410194 (5.81%)	4153919 (17.12%)				
		Uniquely Placed Beads		Uniquely Placed Beads	
		0MM 426916 (1.76%)		0MM 24832 (0.10%)	
		1MM 168790 (0.70%)	595706 (2.45%)	1MM 59153 (0.24%)	83985 (0.35%)
		2MM 378447 (1.56%)	974153 (4.01%)	2MM 243840 (1.00%)	327825 (1.35%)
		3MM 255404 (1.05%)	1229557 (5.07%)	3MM 256960 (1.06%)	584785 (2.41%)
		4MM 260539 (1.07%)	1490096 (6.14%)	4MM 208291 (0.86%)	793076 (3.27%)
		5MM 195405 (0.81%)	1685501 (6.95%)	5MM 206302 (0.85%)	999378 (4.12%)
		6MM 142566 (0.59%)	1828067 (7.53%)	6MM 370979 (1.53%)	1370357 (5.65%)

Example Output Files: .counts

- gff tab-delimited format

ChrName	Dot	miRNA	Start	End	Dot	Strand	Counts	Attributes
1	.	miRNA	1092368	1092387	.	+	1	ACC="MI0000342"; ID=hsa-mir-200b";
1	.	miRNA	1092369	1092389	.	+	1	ACC="MI0000342"; ID=hsa-mir-200b";
1	.	miRNA	1092404	1092427	.	+	2	ACC="MI0000342"; ID=hsa-mir-200b";
1	.	miRNA	1092404	1092424	.	+	19	ACC="MI0000342"; ID=hsa-mir-200b";
1	.	miRNA	1092404	1092425	.	+	22	ACC="MI0000342"; ID=hsa-mir-200b";
1	.	miRNA	1092404	1092423	.	+	1	ACC="MI0000342"; ID=hsa-mir-200b";
1	.	miRNA	1092404	1092428	.	+	2	ACC="MI0000342"; ID=hsa-mir-200b";
1	.	miRNA	1092404	1092426	.	+	10	ACC="MI0000342"; ID=hsa-mir-200b";
1	.	miRNA	1092405	1092422	.	+	2	ACC="MI0000342"; ID=hsa-mir-200b";
1	.	miRNA	1092405	1092423	.	+	1	ACC="MI0000342"; ID=hsa-mir-200b";
1	.	miRNA	1092405	1092427	.	+	1	ACC="MI0000342"; ID=hsa-mir-200b";
1	.	miRNA	1092405	1092426	.	+	1	ACC="MI0000342"; ID=hsa-mir-200b";
1	.	miRNA	1092405	1092425	.	+	1	ACC="MI0000342"; ID=hsa-mir-200b";

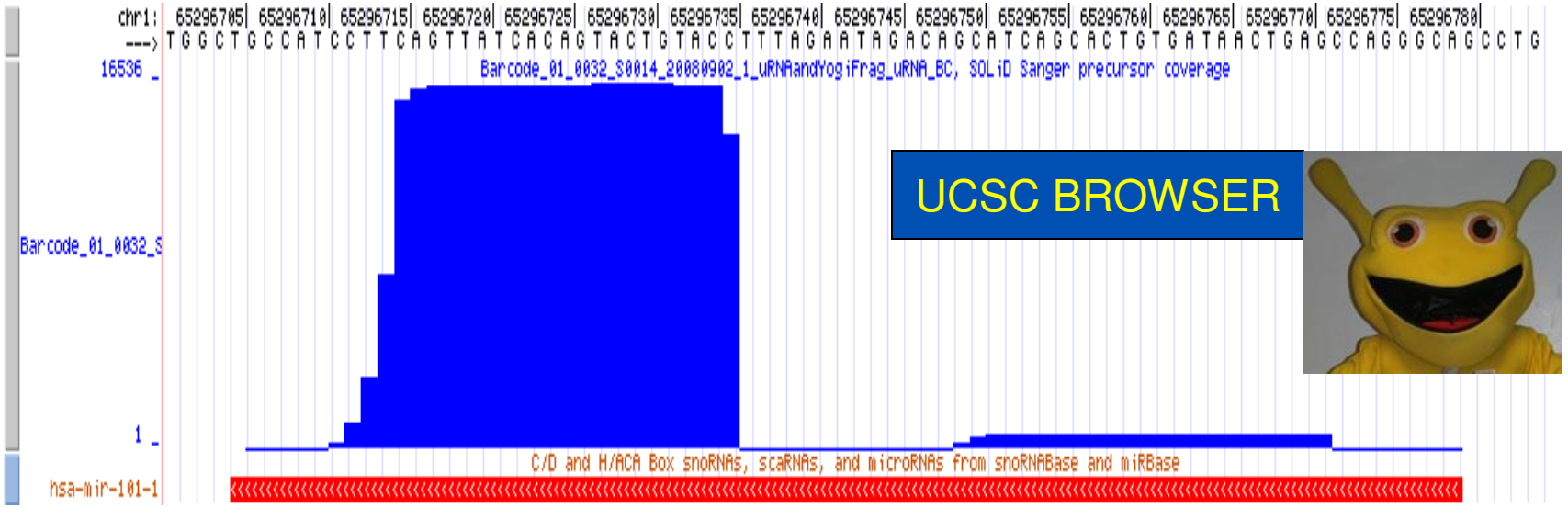
Example Output Files: _extend.ma

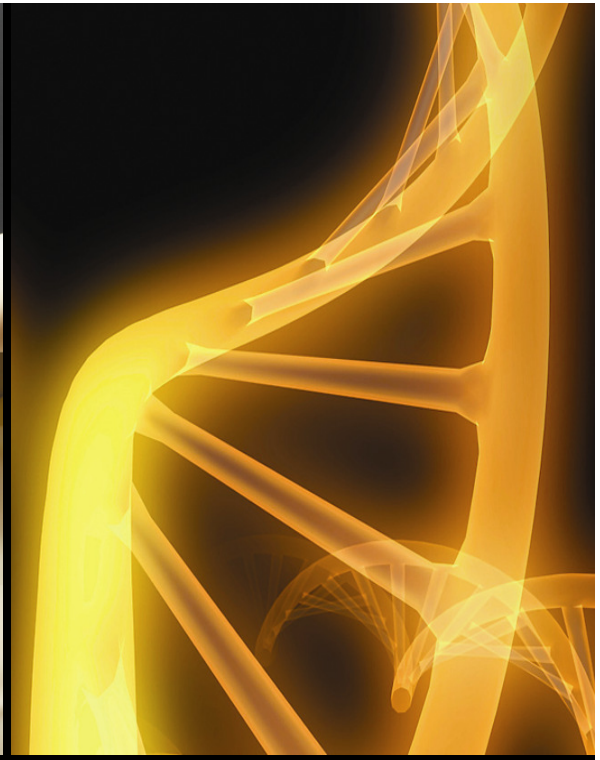
- FASTA format (>bead_ID,[A]_[B][C].[D].[E])
  - A = sequence reference
  - B = strand (nothing if plus strand)
  - C = reference coordinate of map event
  - D = number of mismatches
  - E = alignment size

```
>28_11_1618_F3,2_-182417417.6.23
T10201232002201202032002223020123030
>28_12_947_F3,1_116291255.6.22,1_-182837655.6.22,9_-52744104.6.22
T00212020221232322120003302000323 03
>28_13_272_F3,1_554814.6.28,2_68790944.6.28,5_-95213628.5.28,
T02130201212323003301323300003000201
>28_16_1797_F3,7_-20426471.6.21
T30100333223103211221130302012300131
```

Example Output Files: .wig

```
browser position chr1:1092368-222511396
browser hide all
track type=wiggle_0 name="SOLiD miRBase coverage" description="SOLiD Sanger precursor coverage"
visibility=full color=0,0,255 yLineMark=0 yLineOnOff=on priority=10
variableStep chrom=chr1 span=1
1092368 1
1092369 2
1092370 2
1092371 2
1092372 2
1092373 2
1092374 2
1092375 2
1092376 2
1092377 2
1092378 2
```





Applied Biosystems

Appendix

Configuration File

Parameter	Description	Example
RUN_FILTERING	Whether or not filtering step to be performed. If it is set to 'true' a reference file needs to be provided in filter_step_reference_file field	true
RUN_MIRBASE_MATCHING	Whether or not data needs to be matched against known precursor sequences. If it is set to 'true' a reference gff file needs to be provided in miRBase_step_gff_reference_file	true
RUN_GENOME_MATCHING	Whether or not data needs to be mapped to the genome.	true
tag_length	Read length	35
pattern	A sequence of '0' and '1', same length as the read length, which specifies which positions to be used for matching: '0' for masking and '1' for using.	111…111
adaptor	P2 adaptor sequence (provided)	CGCC…
filter_step_reference_file	Path to a user provided multi fasta format file containing (base) sequences that need to be removed from the reads file. This sequences can be primers/adaptor sequences used in library preparation, ribosomal RNA, tRNA. An example of such file is located in RNA_pipeline/data/ human_filter_reference.fasta	human_filter_reference.fasta
filter_step_number_of_bases_to_use	The number of bases to be used when reads are aligned against filter reference sequence.	25
filter_step_number_of_errors	The number of errors allowed when aligning reads to filter reference sequences.	2

Configuration File

miRBase_step_gff_reference_file	File in gff format that contains genomic locations of known miRNAs. For common species it can be downloaded from http://microrna.sanger.ac.uk/sequences/ftp.shtml	hsa.gff
MAKE_PRECURSOR_FASTA_REFERENCE	Whether or not to generate a fasta format file containing known precursor (base) sequences. Needs to be generated only the first time miRBase_step_gff_reference_file is used	yes/no
miRBase_step_reference_fasta_file	If MAKE_PRECURSOR_FASTA_REFERENCE is set to 'yes', that the location where the file containing known precursor (base) sequences need to be saved. If the file exists that the path to it.	/path/has.fasta
miRBase_step_reference_extenssion	How many bases need to be used in extension step (tag_length)	35
miRBase_step_seeds_number_of_bases_to_use	Number of bases to use when generating initial seeds locations	18
miRBase_step_seeds_number_of_errors	Number of mismatches allowed when generating seeds locations	3
miRBase_step_extension_max_number_of_errors	Number of errors allowed in full length of the read after extension step	6

Configuration File

miRBase_step_output_read_type	The type of the reads to use when generating, counts and wiggle files. Options are: all, in which case all matching locations are used; unique, only uniquely matched reads are used; random: for reads that match in multiple locations only a randomly chosen matching location is used.	all
miRBase_step_output_counts	Whether or not to generate a file (gff format) containing counts (both strands) of reads matching precursor sequences	yes
miRBase_step_output_wiggle	Whether or not to generate a coverage file (wiggle format, both strands) of precursor sequences	yes
miRBase_step_output_gff_read	Whether or not to convert matching file (extension ma) to gff format	no
genome_step_reference_fasta_file	Path to genome (base) sequence file. The names of the entries of this file should coincide with the ones used in miRBase_step_gff_reference_file	Human.fasta
genome_step_seeds_number_of_bases_to_use	Number of bases to use when generating initial seeds locations	20
genome_step_seeds_number_of_errors	Number of errors allowed when generating initial seeds locations	2
genome_step_maximum_number_of_hits_per_tag	Number of matching locations retained when generating initial seeds locations	5

Configuration File

genome_step_extension_max_number_of_errors	Number of errors allowed in full length of the read after extension step	5
genome_step_output_read_type	The type of the reads to use when generating, coverage (wiggle) files (one/chromosome). Options are the same as in miRBase step.	all
genome_step_output_wiggle	Whether or not to generate (chromosomes) coverage files (wiggle format)	yes
genome_step_output_wiggle_coverage_cut	Threshold for base coverage to output in coverage files.	10
NAME_OF_QUEUE	The name that is displayed in jobs q	
FOLDER_FOR_TEMPORARY_FILES_ON_COMPUTE_NODES	Path to where the intermediate files will be generated	/scratch
NUMBER_OF_PROCESSORS_ALLOCATED_PER_NODE	The number of processors allocated for each job	1
MAX_MEMORY_PER_JOB_IN_BYTES		2e9

Run miRNA analysis

perl /share/apps/small_RNA_0.5.0/bin/RNA_matching_analysis_pipeline.pl -r /state/partition1/home/trainer06/miRNARun/humanMiRNA.csfasta -c /state/partition1/home/trainer06/miRNARun/small_rna_config_file_example.txt -corona /share/apps/small_RNA_0.5.0 -o /state/partition1/home/trainer06/miRNARun/output >step1.log 2>&1

perl /share/apps/small_RNA_0.5.0/bin/submit_scripts_to_SGE.pl -j /state/partition1/home/trainer06/miRNARun/output/ JOB_LIST.txt >step2.log 2>&1 &