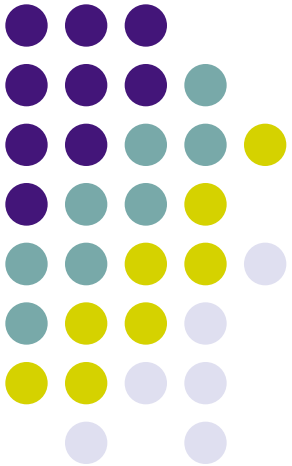
mirUtils for miRNA quantification

Anna Battenhouse

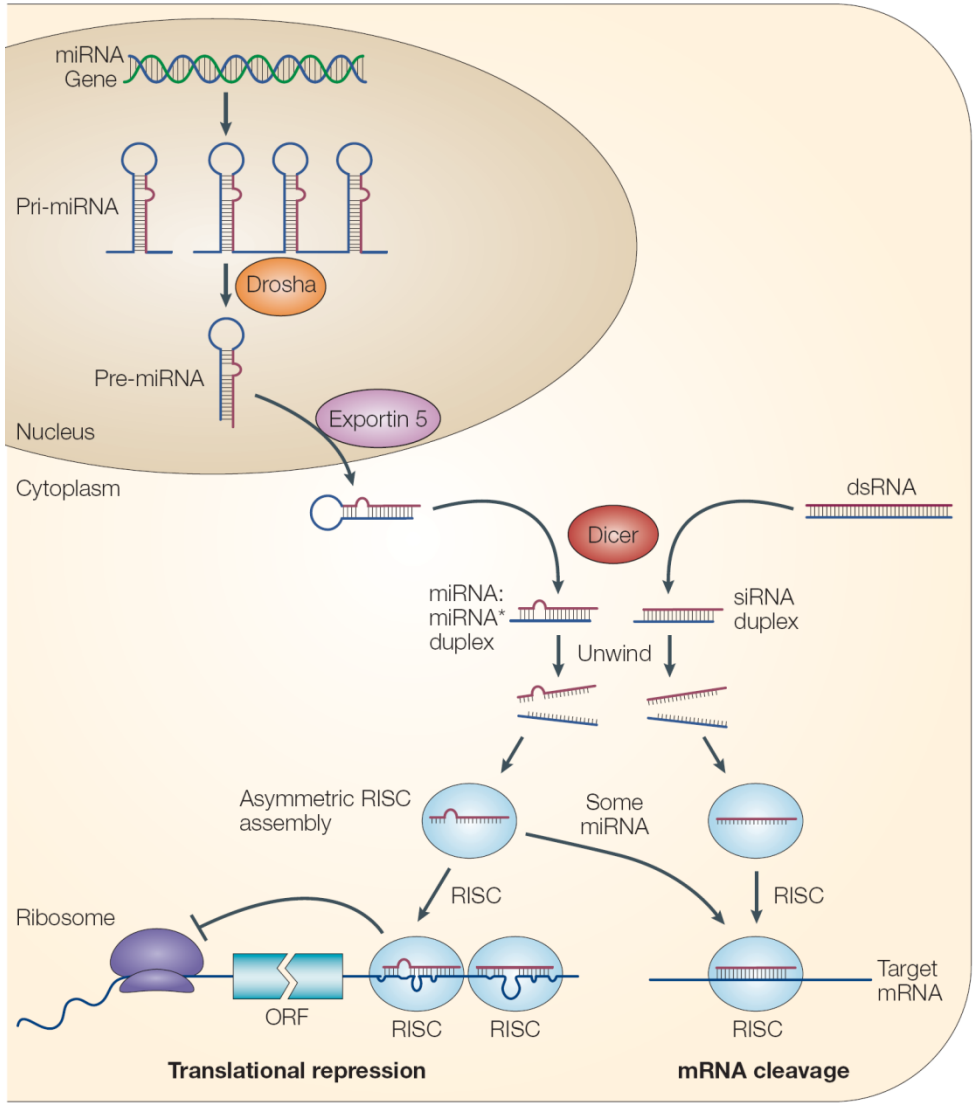
Research Associate, Iyer Lab

Nov. 13, 2014

http://mirutils.sourceforge.net

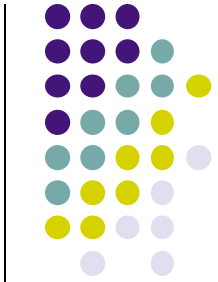


miRNA overview



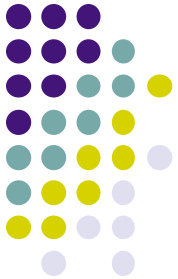
- **Primary** miRNA **transcript**
  - length, coordinates generally not known
- **Precursor** miRNA ***hairpin***
  - Drosha processing in nucleus results in 70-140 nt sequence
- **Processed *mature*** miRNA
  - Dicer cleavage in cytoplasm, RISC complex loading
  - 20-22 nt sequences function in gene silencing
  - 2 per hairpin
    - 1 generally dominant
    - other formerly designated “star”, e.g. hsa-mir-21*

miRNA quantification

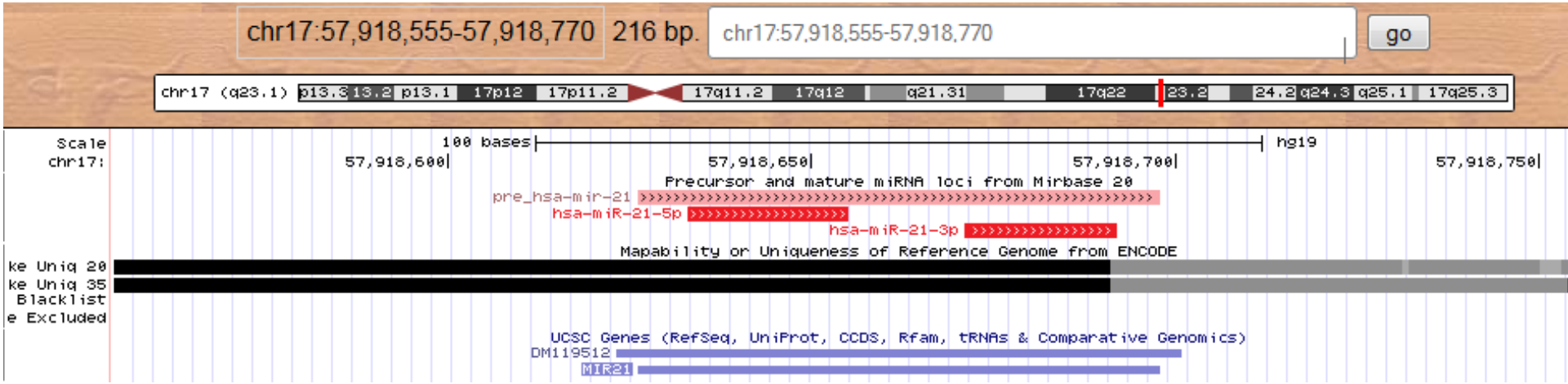


- Motivation
  - Largely driven by Nathan Abell's ENCODE small RNA analysis project
    - and quantification desires of former post-doc Adam Morris
  - Just quantifying alignment to miRBase hairpins is not enough
    - understand relationship between hairpin and mature species generated
    - quantify expression of miRNA sequences common to > 1 gene
    - need quality controls!
- Requirements
  - use taxonomy defined by miRBase resources
    - support v19 or later
  - report counts for different taxonomy levels
    - hairpins, groups & families
    - mature loci & mature sequences
  - provide quality metrics
    - distinguish between "good" and "bad" alignments to mature species
    - track alignment quality metrics such as # mismatches, mapping quality

miRBase



- Database of miRNA resources
  - 90+ organisms (*hsa*, *mmu*, *cel*, *ath*…)
  - consensus RNA sequences
    - **hairpin.fa**, **mature.fa**
  - genome GFFs for each organism
    - provide genomic locations for hairpins and their 5p/3p mature species
  - defines a taxonomy for related sequences
    - explicit (e.g. family definitions in **miFam.dat**)
    - implicit (via naming conventions in GFFs)



Community annotation

microRNA 21 also known as hsa-mir-21 or miRNA21 is a mammalian microRNA that is encoded by the MIR21 gene. MIRN21 was one of the first mammalian microRNAs identified. The mature miR-21 sequence is strongly conserved throughout evolution. The human microRNA-21 gene is located on plus strand of chromosome 17q23.2 (55273409–55273480) within a coding gene TMEM49 (also called vacuole membrane protein). Despite being located in intronic regions of a coding gene in the direction of transcription, it has its own promoter regions and forms a ~3433-nt long primary transcript of miR-21 (known as pri-miR-21) which is independently transcribed. The stem–loop recursor of miR-21(pre-miR-21) resides between nucleotides 2445 and 2516 of pri-miR-21.

Show Wikipedia entry | View @ Wikipedia | Edit Wikipedia entry

Stem-loop

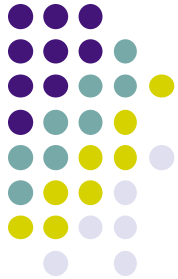
```
u     gu        a     a     a    u a
 gucgg  agcuuauc gacug uguug cugu g a
 |||||  |||||||| ||||| ||||| |||| | u
 caguc  ucggguag cugac acaac ggua c c
a     ug        -     c     -    - u
```

Get sequence

Deep sequencing

100456 reads, 2.36e+04 reads per million, 76 experiments

miRBase GFFs



- Organism GFFs provide genomic coordinates for precursor hairpin & mature miRNA species
  - *and* define group & mature sequence taxonomies
    - v20 *hsa* – hg19 coordinates
    - v21 *hsa* – hg38 coordinates

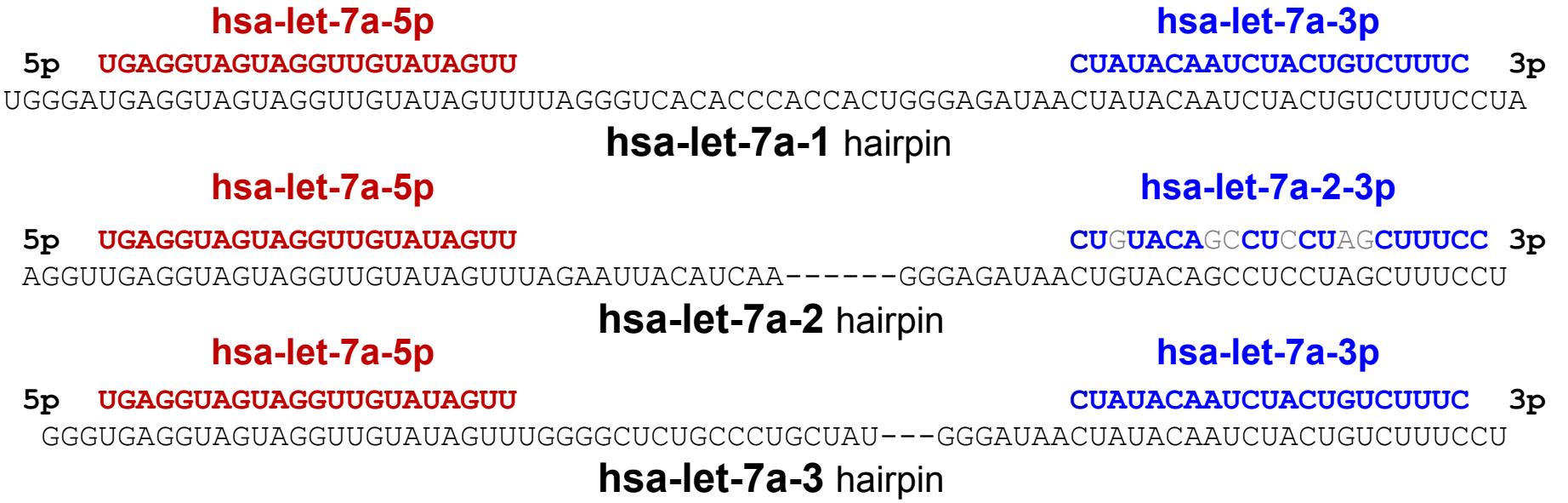
```
chr9  hairpin  94175957  94176036 + ID=MI0000060;Alias=MI0000060;Name=hsa-let-7a-1
chr9  mature   94175962  94175983 + ID=MIMAT0000062_2;Alias=MIMAT0000062;Name=hsa-let-7a-5p;Derives_from=MI0000060
chr9  mature   94176013  94176033 + ID=MIMAT0004481_1;Alias=MIMAT0004481;Name=hsa-let-7a-3p;Derives_from=MI0000060

chr11 hairpin 122146522 122146593 - ID=MI0000061;Alias=MI0000061;Name=hsa-let-7a-2
chr11 mature  122146568 122146589 - ID=MIMAT0000062;Alias=MIMAT0000062;Name=hsa-let-7a-5p;Derives_from=MI0000061
chr11 mature  122146523 122146544 - ID=MIMAT0010195;Alias=MIMAT0010195;Name=hsa-let-7a-2-3p;Derives_from=MI0000061

chr22 hairpin  46112749  46112822 + ID=MI0000062;Alias=MI0000062;Name=hsa-let-7a-3
chr22 mature   46112752  46112773 + ID=MIMAT0000062_1;Alias=MIMAT0000062;Name=hsa-let-7a-5p;Derives_from=MI0000062
chr22 mature   46112800  46112820 + ID=MIMAT0004481;Alias=MIMAT0004481;Name=hsa-let-7a-3p;Derives_from=MI0000062
```

miRBase groups

- hairpin **groups**
  - miRNA hairpin precursors with closely related mature sequences
    - hairpin name ends in -1, -2, -3
  - can be considered as one group: “hsa-let-7a[3]”
- mature miRNAs
  - these 3 hsa-let-7a hairpins have 6 **mature loci**
    - but only three distinct **mature sequences** (one 5p and two 3p)



mirBase families

- miRNA hairpin **families**
  - significant sequence homology
    - especially in the seed regions
  - defined largely by common targets
    - **miFam.dat** file
  - family name: “let-7[12]”

AC	MIPF00000	2
ID	let-7	
MI	MI0000060	hsa-let-7a-1
MI	MI0000061	hsa-let-7a-2
MI	MI0000062	hsa-let-7a-3
MI	MI0000063	hsa-let-7b
MI	MI0000064	hsa-let-7c
MI	MI0000065	hsa-let-7d
MI	MI0000066	hsa-let-7e
MI	MI0000067	hsa-let-7f-1
MI	MI0000068	hsa-let-7f-2
MI	MI0000100	hsa-mir-98
MI	MI0000433	hsa-let-7g
MI	MI0000434	hsa-let-7i

hsa-let-7a-1

```
5p   UGAGGUAGUAGGUUGUAUAGUU                                  CUAUACAAUCUACUGUCUUUCCU 3p
UGGGAUGAGGUAGUAGGUUGUAUAGUUUUAGGGUCACA---CCCACCACUGGGAGAUAACUAUACAAUCUACUGUCUUUCCUA
```

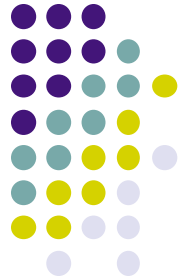
hsa-let-7b

```
5p   UGAGGUAGUAGGUUGUGUGGUUUCAGGGCAGUGAUGUUGCCCCUCGGAAGAUAACUAUACAACCUACUGCCUUCCCUG
```

hsa-mir-98

```
5p   UGAGGUAGUAAGUUGUAUUGUU                                  CUAUACAACUUACUACUUUCCCU 3p
[…]GGUGAGGUAGUAAGUUGUAUUGUUGUGGGGUAG[...]GCCCCAAUUAGAAGAUAACUAUACAACUUACUACUUUCCCUG[…]
```

mirUtils tool suite



- Provides a set of tools to support quantitative analysis of miRBase-aligned miRNA sequences

mirUtils mbaseRefFa [options] <organism(s)>

- Make a cDNA fasta file for specified organism(s)

mirUtils mbaseMirInfo [options] <organism(s)>

- Write miRBase metadata information in searchable format

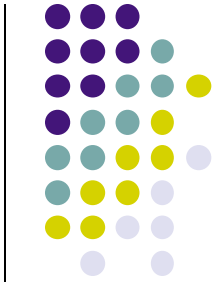
mirUtils mbaseMirStats [options] <bam file(s)>

- Generate miRNA statistics reports from miRBase-aligned bam file(s)

mirUtils filterAligns [options] <bam file(s)>

- Extract 'good fit' alignments from miRBase-aligned bam for further analysis

mirUtils mbaseMirStats



mirUtils mbaseMirStats [options] <bam file(s)>

--organism miRBase organism prefix (hsa)

--version miRBase version (v21) mirUtils bundle includes all of miRBase v19, v20, v21

--min-overlap minimum base overlap between alignment and mature locus to be counted as "only", a.k.a "good fit" (13)

--margin maximum distance before annotated start or after annotated end of mature locus to be counted as "good fit" (5)

--cluster-distance inter-hairpin distance used to define clusters (10000)

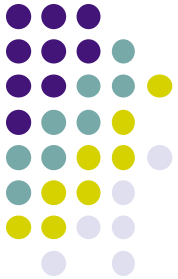
--bam-flags flags and options to pass to samtools view when reading BAM file ('-F 0x4')

--bam-locs contig names to pass to samtools view when reading BAM

--out-prefix prefix for output files (default based on BAM name)

--cmb-prefix prefix for combined output files when multiple BAM are processed

mbaseMirStats reports

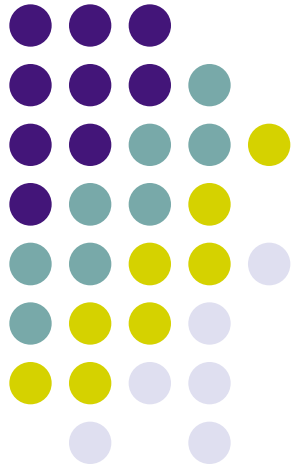


mirUtils mbaseMirStats [options] <bam file(s)>

- creates a set of report files for each miRBase-aligned bam
  - ***per-hairpin-location*** counts
    - **<prefix>.coverage**, **<prefix>.starts**
  - miRNA ***hairpin*** related statistics
    - **<prefix>. hairpin.hist**
    - **<prefix>. group.hist** ← usually want to use this one
    - **<prefix>. family.hist**
    - **<prefix>. cluster.hist**, **.cluster+.hist**, **.cluster-.hist**
  - ***mature*** miRNA statistics
    - **<prefix>. mature.hist**
    - **<prefix>. matseq.hist** ← usually want to use this one
  - ***metadata*** summaries (of hairpin & mature taxonomies)
    - **<organism>_ <version>_cluster<distance>.hpInfo**, **.matInfo**

per-hairpin-base counts

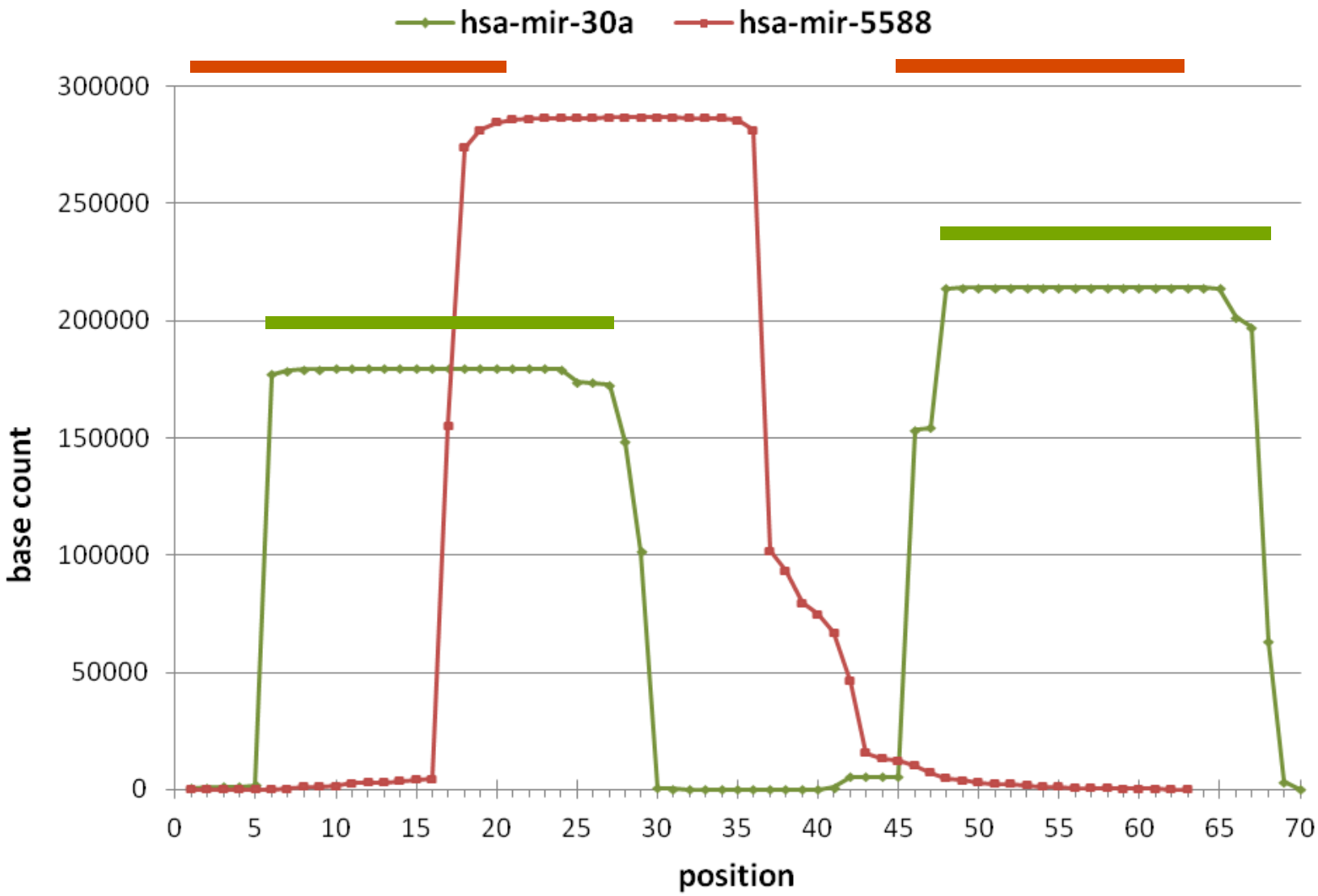
- coverage
  - count of all aligned bases at each position
- starts
  - count of alignments starting at each position



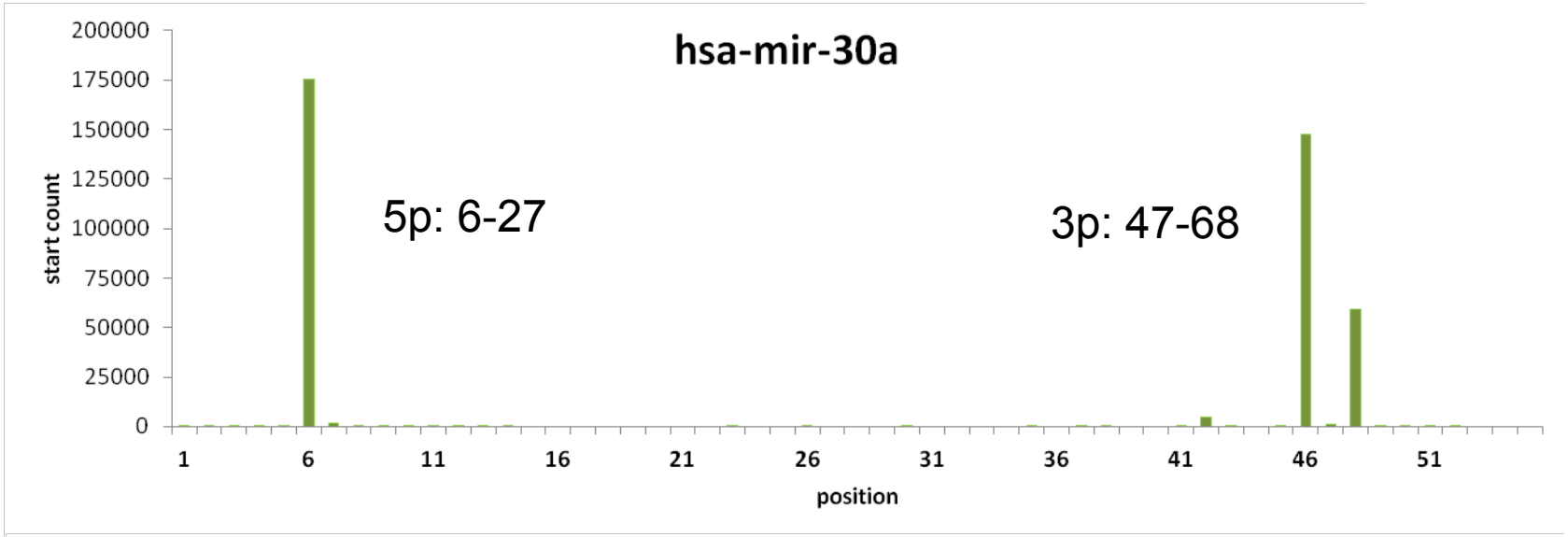
a549_cmb.coverage

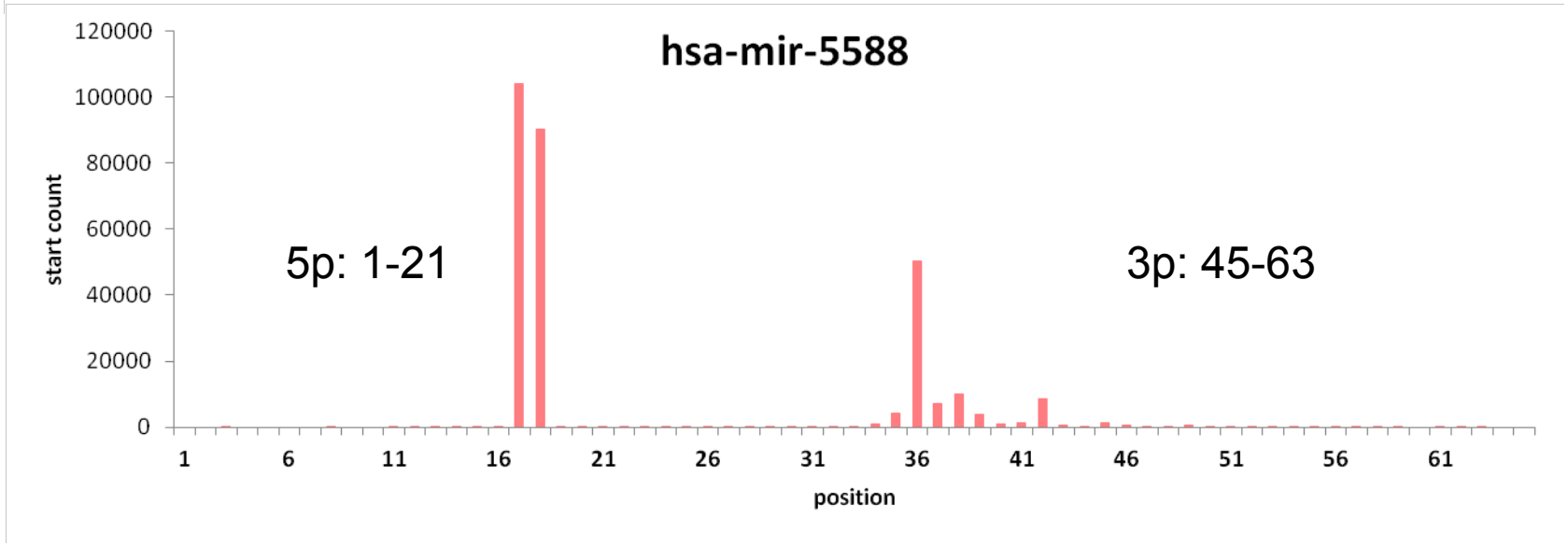
					annotation					positions					
hairpin	rank	reads	bases	strand	5pPos1	5pPos2	3pPos1	3pPos2	length	1	2	3	4	5	...
hsa-mir-21	1	13773459	326470489	+	8	29	46	66	72	525	17189	92320	93714	94121	
hsa-mir-20a	2	1578057	35973838	+	8	30	44	65	71		129	49091	49527	49696	
hsa-mir-424	3	1439026	28879620	-	11	32	48	68	98			5	5	6	
hsa-mir-27b	4	1233215	26629821	+	19	40	61	81	97	19	20	20	20	20	
hsa-mir-16-2	5	1203818	26667271	+	10	31	53	74	81				19	24665	
hsa-mir-16-1	6	1201110	26606466	-	14	35	56	77	89						
hsa-let-7f-1	7	842079	18851757	+	7	28	63	84	87	7715	33372	33527	33684	35452	
hsa-mir-17	8	829105	18781829	+	14	36	51	72	84						
hsa-mir-27a	9	812546	17081852	-	10	31	51	71	78						
hsa-mir-18a	10	717194	16299892	+	6	28	47	69	71	780	835	1109	2339	4707	
hsa-mir-106b	11	665872	13972057	-	12	32	52	73	82						
hsa-mir-93	12	613231	14055510	-	11	33	50	71	80	1	1	1	3	5	
hsa-let-7a-3	13	556056	12486047	+	4	25	52	72	74	55	1015	553651	555016	555244	
hsa-mir-194-1	14	530499	11710470	-	15	36			85					2	
hsa-let-7i	15	441410	9483940	+	6	27	62	83	84	376	415	477	1691	4521	
hsa-mir-30a	16	393602	8831295	-	6	27	47	68	71	774	974	1098	1230	1798	
hsa-mir-34a	17	390555	8640819	-	22	43	64	85	110	1	6	7	11	14	
hsa-mir-151a	18	386713	8147529	-	11	31	47	67	90				19	320	
hsa-mir-137	19	333886	7839037	-			59	81	102	1	1	2	3	3	
hsa-mir-3074	20	302194	6604284	-	12	32	50	71	81	6	7	9	12	13	
hsa-mir-23a	21	299384	6323180	-	9	30	45	65	73					2	
hsa-mir-5588	22	286679	6141821	-	1	21	45	63	63	17	49	92	107	155	
hsa-mir-138-1	23	267003	6111900	+	23	45	63	84	99	37	42	52	65	69	
hsa-mir-26a-1	24	263525	6277346	+	10	31	49	70	77		26	26	708	1672	

Hairpin coverage



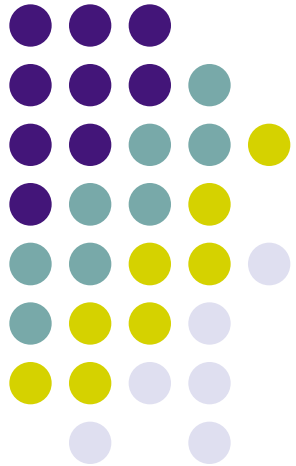
Hairpin starts



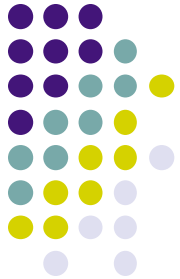


hairpin & mature statistics reports

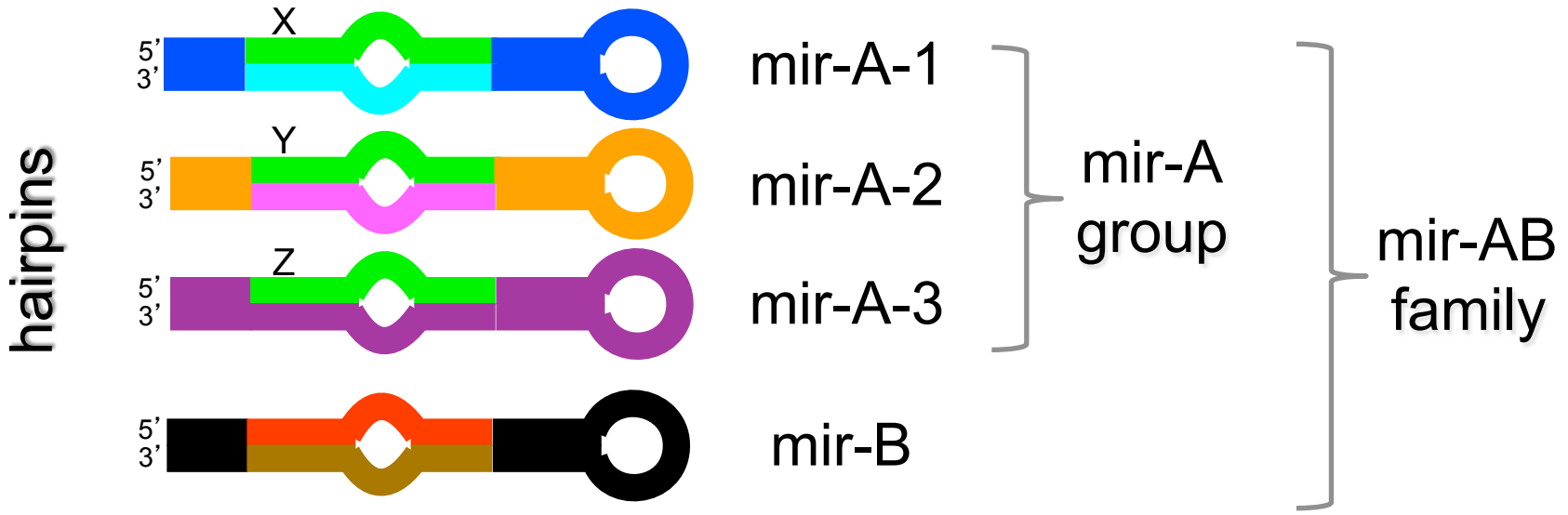
- hairpin – 6 taxonomy levels
  - hairpin locus, group, family
  - cluster+, cluster-, cluster
- mature miRNA – 2 levels
  - mature locus, mature sequence

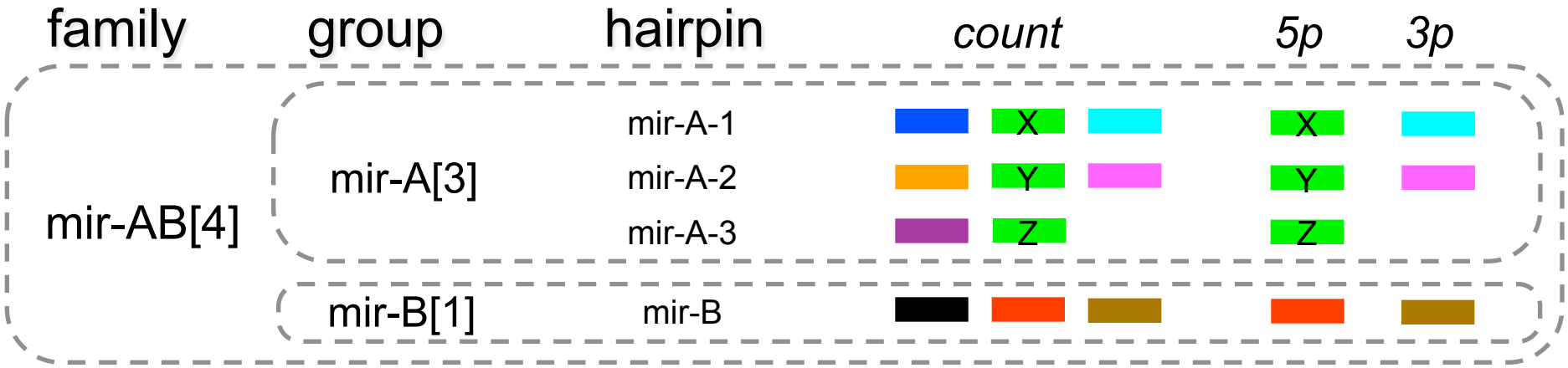


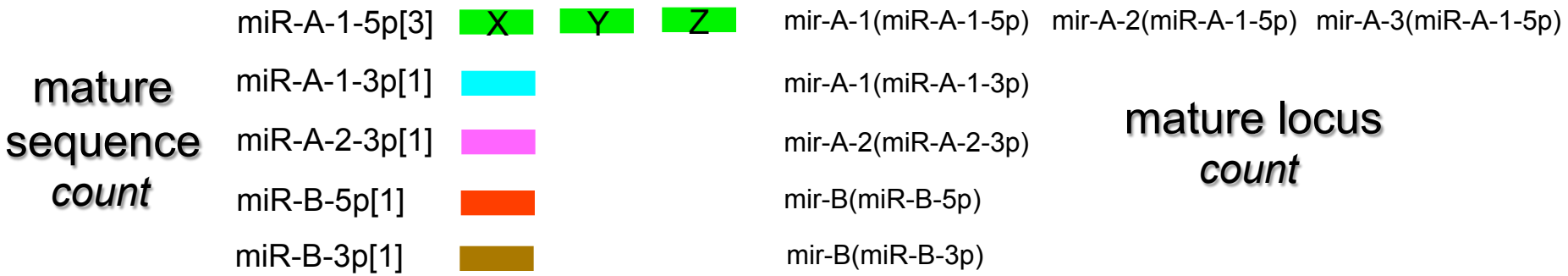
How mirUtils counts



- All mirUtils reports are based on ***counts*** of individual alignment records from the BAM/SAM input
  - basic counts are at individual hairpin locus and mature locus level
  - higer taxonomy levels simply ***sum*** the counts for alignments in the taxonomy set
- Power of mirUtils reports comes from:
  - careful definition of the taxonomy reporting set membership
  - careful recording of alignment features that apply to the full precursor hairpin or mature miRNA locus

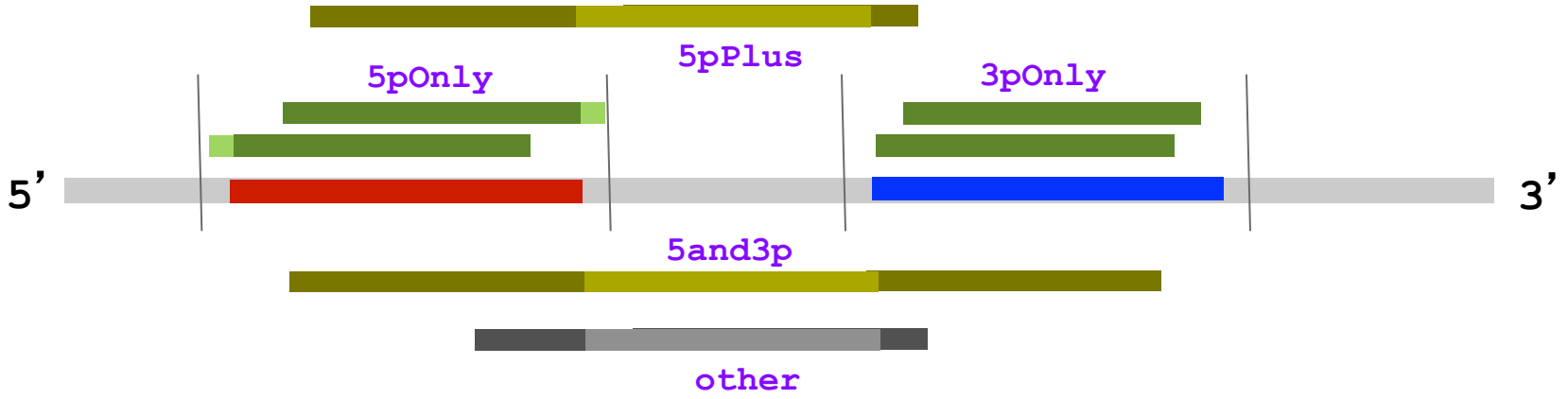

hairpins
X
Y
Z
5'
3'
mir-A-1
mir-A-2
mir-A-3
mir-B
mir-A group
mir-AB family


family
group
hairpin
count
5p
3p
mir-AB[4]
mir-A[3]
mir-A-1
mir-A-2
mir-A-3
mir-B[1]
mir-B
X
Y
Z


mature sequence count
miR-A-1-5p[3]
miR-A-1-3p[1]
miR-A-2-3p[1]
miR-B-5p[1]
miR-B-3p[1]
X
Y
Z
mir-A-1(miR-A-1-5p)
mir-A-2(miR-A-1-5p)
mir-A-3(miR-A-1-5p)
mir-A-1(miR-A-1-3p)
mir-A-2(miR-A-2-3p)
mir-B(miR-B-5p)
mir-B(miR-B-3p)
mature locus count

“Good fit” overlap & margin

- **count**
  - total number of reads aligned to hairpin
- **5pOnly** / **3pOnly** (**“good fit”**)
  - # reads aligned to 5p/3p mature locus with at least **--min-overlap (13)** bases of overlap and within **--margin (5)** bases of start & end
- **5and3p**
  - # reads aligned to both 5p/3p mature loci with at least **--min-overlap** bases of overlap (suggests un-processed transcript)
- **5pPlus** / **3pPlus**
  - # reads aligned to 5p/3p mature locus with at least **--min-overlap** bases of overlap that do not minimally overlap the 3p/5p also (may be partially processed transcript)
- **5pOnly** + **5pPlus** + **3pOnly** + **3pPlus** + <other> = **count**



a549_cmb.hairpin.hist

name	rank	count	dup	oppStrand	mm0	mm1	mm2	mm3p	indel	mq0	mq1-19	mq20-29	mq30p	5pOnly	5pPlus	3pOnly	3pPlus	5and3p
hsa-mir-21	1	13773459	13773281	34	13559436	210810	2411	802	2947	85	13561	1104870	12654943	13667873	20163	79114	709	0
hsa-mir-20a	2	1578057	1577970	3	1560725	16913	363	56	49324	17	1458368	108337	11335	1565145	929	11810	84	0
hsa-mir-424	3	1439026	1438930	10	1344363	94043	582	38	25	11	1557	1413471	23987	1429912	97	8928	56	0
hsa-mir-27b	4	1233215	1233127	5	1211916	20223	251	825	215	16106	1177684	9609	29816	32145	169	1196690	3865	0
hsa-mir-16-2	5	1203818	1203751	2	1189360	12048	232	2178	24728	175085	1026607	680	1446	1200422	2705	675	9	0
hsa-mir-16-1	6	1201110	1201050	3	1146694	52606	866	944	261	174705	1025891	376	138	1198211	2859	29	3	0
hsa-let-7f-1	7	842079	842010	12	796331	44940	745	63	90	2888	833920	730	4541	833928	8044	64	9	0
hsa-mir-17	8	829105	829013	2	805295	23301	387	122	316	1094	479543	82920	265548	480259	552	347868	220	0
hsa-mir-27a	9	812546	812466	3	796991	15304	207	44	94	15788	790346	6054	358	2614	24	808830	1043	0
hsa-mir-18a	10	717194	717132	2	707958	8328	463	445	844	10795	701515	4402	482	715618	971	472	0	0
hsa-mir-106b	11	665872	665801	9	660504	5260	85	23	46	24	12123	575963	77762	663193	168	2237	4	0
hsa-mir-93	12	613231	613150	8	605405	7617	179	30	826	1	1054	90875	521301	611349	509	1141	24	0
hsa-let-7a-3	13	556056	556009	9	548021	7873	132	30	86	3378	551659	680	339	555489	44	505	7	0
hsa-mir-194-1	14	530499	530439	68	515976	14337	169	17	107	10613	518800	452	634	527665	2762	0	0	0
hsa-let-7i	15	441410	441309	28	435525	5716	117	52	423	230	248317	190298	2565	440739	169	402	36	0
hsa-mir-30a	16	393602	393525	0	383622	9748	142	90	901	250	382126	11124	102	179359	163	213270	793	0
hsa-mir-34a	17	390555	390466	61	386205	4102	64	184	119	3	599	73722	316231	389076	399	559	15	0
hsa-mir-151a	18	386713	386629	3	378618	7880	162	53	394	6352	99355	257122	23884	104738	330	280326	1279	0
hsa-mir-137	19	333886	333840	8	330789	2981	72	44	84	1	386	32371	301128	0	0	333781	79	0
hsa-mir-3074	20	302194	302086	302124	298580	3533	60	21	7788	199828	102312	32	22	301968	72	128	0	0
hsa-mir-23a	21	299384	299325	5	295700	3595	71	18	113	1425	291737	5999	223	108	3	299141	103	0
hsa-mir-5588	22	286679	286317	91669	135822	65829	47593	37435	2948	59220	184605	42852	2	3	1397	0	545	0
hsa-mir-138-1	23	267003	266814	52	263098	3367	115	423	211	32540	232329	1578	556	264657	321	76	113	0
hsa-mir-26a-1	24	263525	263492	3	259977	3494	46	8	108	1255	168474	85055	8741	262785	717	4	0	0
hsa-mir-224	25	243132	243078	3	239851	2882	306	93	65	1	285	32827	210019	241387	854	688	23	0
hsa-mir-138-2	26	237448	237393	12	233993	3244	119	92	193	32950	204464	20	14	237358	83	0	1	0
hsa-mir-376c	27	212641	212579	61	210651	1929	13	48	39	1	5061	203999	3580	1	6	212567	12	0
hsa-let-7e	28	210604	210565	16	207009	3449	100	46	346	20	201156	9229	199	210520	38	34	2	0
hsa-mir-335	29	191476	191387	24	190085	1257	34	100	9	2	250	67334	123890	191178	107	155	21	0

a549_cmb.mature.hist

- only “good fit” alignments are included here (5pOnly or 3pOnly)
- mismatch / indel counts are within mature locus only

name	rank	count	dup	oppStrand	mm0	mm1	mm2	mm3p	indel	mq0	mq1-19	mq20-29	mq30p
hsa-mir-21(hsa-miR-21-5p)	1	13667873	13667746	33	13550916	115700	1157	100	21	78	13285	1086148	12568362
hsa-mir-20a(hsa-miR-20a-5p)	2	1565145	1565086	1	1549190	15649	297	9	48553	17	1458219	105015	1894
hsa-mir-151a(hsa-miR-151a-3p)	21	280326	280290	0	277943	2361	20	2	11	3	736	256897	22690
hsa-mir-138-1(hsa-miR-138-5p)	22	264657	264619	1	261252	3318	78	9	90	32493	232112	11	41
hsa-mir-26a-1(hsa-miR-26a-5p)	23	262785	262760	1	260059	2681	40	5	1	1255	168095	84874	8561
hsa-mir-224(hsa-miR-224-5p)	24	241387	241356	0	239235	2123	28	1	4	1	259	31722	209405
hsa-mir-138-2(hsa-miR-138-5p)	25	237358	237313	5	234189	3069	90	10	134	32943	204392	9	14
hsa-mir-30a(hsa-miR-30a-3p)	26	213270	213247	0	211525	1726	17	2	22	191	209472	3579	28
hsa-mir-376c(hsa-miR-376c-3p)	27	212567	212536	0	210671	1894	2	0	1	1	5061	203931	3574
hsa-let-7e(hsa-let-7e-5p)	28	210520	210489	12	207096	3350	72	2	333	18	201136	9193	173
hsa-mir-335(hsa-miR-335-5p)	29	191178	191143	0	189928	1222	23	5	1	0	242	67210	123726
hsa-let-7g(hsa-let-7g-5p)	30	184726	184688	0	182523	2169	33	1	0	283	118823	64954	666
hsa-mir-30a(hsa-miR-30a-5p)	31	179359	179315	0	176935	2381	41	2	769	59	172350	6895	55
hsa-mir-26b(hsa-miR-26b-5p)	32	164709	164667	0	163311	1382	16	0	0	88	113681	50550	390
hsa-mir-24-2(hsa-miR-24-3p)	33	158233	158170	0	156474	1726	30	3	9034	99443	58760	7	23
hsa-mir-24-1(hsa-miR-24-3p)	34	153773	153698	2	151956	1776	35	6	3953	100663	53094	0	16
hsa-mir-3189(hsa-miR-3189-5p)	1696	3	2	0	3	0	0	0	0	0	0	0	3
hsa-mir-5588(hsa-miR-5588-5p)	1697	3	3	3	0	2	1	0	0	0	0	3	0
hsa-mir-564(hsa-miR-564(5p))	1698	3	1	0	2	0	1	0	0	0	0	3	0

hairpin loci

name	rank	count	oppStrand	mm0	mm1	mm2	mm3p	indel	mq0	mq1-19	mq20-29	mq30p	5pOnly	5pPlus	3pOnly	3pPlus
hsa-mir-27b	4	1233215	5	1211916	20223	251	825	215	16106	1177684	9609	29816	32145	169	1196690	3865
hsa-mir-16-2	5	1203818	2	1189360	12048	232	2178	24728	175085	1026607	680	1446	1200422	2705	675	9
hsa-mir-16-1	6	1201110	3	1146694	52606	866	944	261	174705	1025891	376	138	1198211	2859	29	3
hsa-let-7f-1	7	842079	12	796331	44940	745	63	90	2888	833920	730	4541	833928	8044	64	9
hsa-let-7a-3	13	556056	9	548021	7873	132	30	86	3378	551659	680	339	555489	44	505	7
hsa-mir-194-1	14	530499	68	515976	14337	169	17	107	10613	518800	452	634	527665	2762	0	0
hsa-let-7i	15	441410	28	435525	5716	117	52	423	230	248317	190298	2565	440739	169	402	36
hsa-mir-744	95	22392	2	21485	881	20	6	15	1	38	4763	17590	22307	54	16	1
hsa-mir-194-2	96	22086	69	19294	2733	44	15	5	10371	11470	154	91	21689	242	79	0
hsa-mir-181a-2	97	21934	52	20265	1630	29	10	15	1932	17510	1787	705	19039	413	2313	1

hairpin groups

name	rank	count	oppStrand	mm0	mm1	mm2	mm3p	indel	mq0	mq1-19	mq20-29	mq30p	5pOnly	5pPlus	3pOnly	3pPlus
hsa-mir-21[1]	1	13773459	34	13559436	210810	2411	802	2947	85	13561	1104870	12654943	13667873	20163	79114	709
hsa-mir-16[2]	2	2404928	5	2336054	64654	1098	3122	24989	349790	2052498	1056	1584	2398633	5564	704	12
hsa-mir-20a[1]	3	1578057	3	1560725	16913	363	56	49324	17	1458368	108337	11335	1565145	929	11810	84
hsa-mir-424[1]	4	1439026	10	1344363	94043	582	38	25	11	1557	1413471	23987	1429912	97	8928	56
hsa-mir-27b[1]	5	1233215	5	1211916	20223	251	825	215	16106	1177684	9609	29816	32145	169	1196690	3865
hsa-let-7f[2]	6	853488	14	807525	45130	755	78	91	5731	841953	946	4858	845039	8134	268	11
hsa-mir-17[1]	7	829105	2	805295	23301	387	122	316	1094	479543	82920	265548	480259	552	347868	220
hsa-mir-27a[1]	8	812546	3	796991	15304	207	44	94	15788	790346	6054	358	2614	24	808830	1043
hsa-mir-18a[1]	9	717194	2	707958	8328	463	445	844	10795	701515	4402	482	715618	971	472	0
hsa-mir-106b[1]	10	665872	9	660504	5260	85	23	46	24	12123	575963	77762	663193	168	2237	4
hsa-mir-93[1]	11	613231	8	605405	7617	179	30	826	1	1054	90875	521301	611349	509	1141	24
hsa-let-7a[3]	12	565984	18	557319	8283	209	173	104	6569	557874	1004	537	564672	233	1005	16
hsa-mir-194[2]	13	552585	137	535270	17070	213	32	112	20984	530270	606	725	549354	3004	79	0
hsa-mir-138[2]	14	504451	64	497091	6611	234	515	404	65490	436793	1598	570	502015	404	76	114
hsa-let-7i[1]	15	441410	28	435525	5716	117	52	423	230	248317	190298	2565	440739	169	402	36
hsa-mir-30a[1]	16	393602	0	383622	9748	142	90	901	250	382126	11124	102	179359	163	213270	793

mature loci

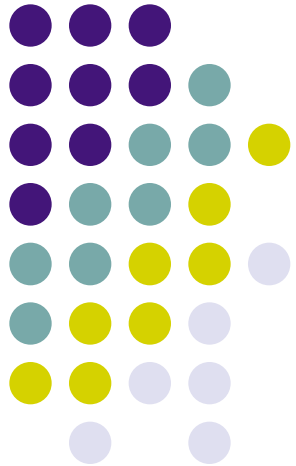
name	rank	count	dup	oppStrand	mm0	mm1	mm2	mm3p	indel	mq0	mq1-19	mq20-29	mq30p
hsa-mir-21(hsa-miR-21-5p)	1	13667873	13667746	33	13550916	115700	1157	100	21	78	13285	1086148	12568362
hsa-mir-20a(hsa-miR-20a-5p)	2	1565145	1565086	1	1549190	15649	297	9	48553	17	1458219	105015	1894
hsa-mir-424(hsa-miR-424-5p)	3	1429912	1429857	0	1335547	93805	555	5	5	8	1528	1404670	23706
hsa-mir-16-2(hsa-miR-16-5p)	4	1200422	1200369	0	1189668	10661	91	2	24396	174411	1024626	100	1285
hsa-mir-16-1(hsa-miR-16-5p)	5	1198211	1198157	1	1187692	10417	100	2	5	174019	1024058	10	124
hsa-mir-27b(hsa-miR-27b-3p)	6	1196690	1196641	4	1181444	15079	165	2	65	16089	1175335	5220	46
hsa-let-7f-1(hsa-let-7f-5p)	7	833928	833884	6	821340	12351	227	10	1	2887	828403	375	2263
hsa-mir-224(hsa-miR-224-3p)	305	688	675	1	604	83	1	0	0	0	0	561	127
hsa-mir-16-2(hsa-miR-16-2-3p)	306	675	666	2	672	2	0	1	0	0	269	277	129
hsa-mir-502(hsa-miR-502-5p)	307	675	667	0	659	16	0	0	0	0	11	197	467
hsa-mir-5094(hsa-miR-5094(5p))	817	29	25	0	27	2	0	0	0	0	0	18	11
hsa-mir-16-1(hsa-miR-16-1-3p)	818	29	26	0	28	1	0	0	0	0	0	25	4
hsa-mir-486-1(hsa-miR-486-3p)	819	29	24	21	29	0	0	0	0	26	3	0	0

mature sequences

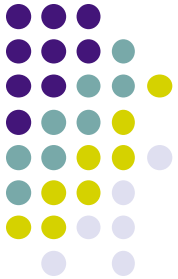
name	rank	count	dup	oppStrand	mm0	mm1	mm2	mm3p	indel	mq0	mq1-19	mq20-29	mq30p
hsa-miR-21-5p[1]	1	13667873	13667746	33	13550916	115700	1157	100	21	78	13285	1086148	12568362
hsa-miR-16-5p[2]	2	2398633	2398526	1	2377360	21078	191	4	24401	348430	2048684	110	1409
hsa-miR-20a-5p[1]	3	1565145	1565086	1	1549190	15649	297	9	48553	17	1458219	105015	1894
hsa-miR-378d[2]	278	679	660	0	644	33	1	1	5	555	120	4	0
hsa-miR-16-2-3p[1]	279	675	666	2	672	2	0	1	0	0	269	277	129
hsa-miR-502-5p[1]	280	675	667	0	659	16	0	0	0	0	11	197	467
hsa-miR-5094[1]	769	29	25	0	27	2	0	0	0	0	0	18	11
hsa-miR-16-1-3p[1]	770	29	26	0	28	1	0	0	0	0	0	25	4
hsa-miR-6739-5p[1]	771	29	10	3	1	8	9	11	0	1	2	26	0

metadata reports

- relate different taxonomy levels
  - miRNA **hairpin** metadata
  - **mature** miRNA metadata



hsa_v21_cluster10000.hpInfo



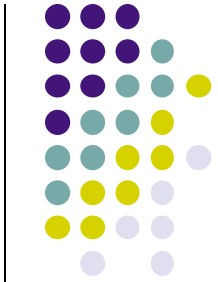
- Summarizes miRBase hairpin metadata
  - maps each hairpin to its group, family, cluster, etc.

chrom	strand	start	end	length	hpid	hairpin	group	family	cluster
chr3	+	160404588	160404685	98	MI0000438	**hsa-mir-15b**	hsa-mir-15b[1]	mir-15[5]	cluster(chr3:160404588-160404825)[2]
chr3	+	160404745	160404825	81	MI0000115	**hsa-mir-16-2**	hsa-mir-16[2]	mir-15[5]	cluster(chr3:160404588-160404825)[2]
chr13	-	50048973	50049061	89	MI0000070	**hsa-mir-16-1**	hsa-mir-16[2]	mir-15[5]	cluster(chr13:50048973-50049201)[2]
chr13	-	50049119	50049201	83	MI0000069	**hsa-mir-15a**	hsa-mir-15a[1]	mir-15[5]	cluster(chr13:50048973-50049201)[2]
chr17	-	7017615	7017701	87	MI0000489	**hsa-mir-195**	hsa-mir-195[1]	mir-15[5]	cluster(chr17:7017615-7018022)[2]
chr1	-	220117853	220117962	110	MI0000291	**hsa-mir-215**	hsa-mir-215[1]	mir-192[2]	cluster(chr1:220117853-220118241)[2]
chr1	-	220118157	220118241	85	MI0000488	**hsa-mir-194-1**	hsa-mir-194[2]	mir-194[2]	cluster(chr1:220117853-220118241)[2]
chr1	-	220200538	220200619	82	MI0006442	**hsa-mir-664a**	hsa-mir-664a[1]	mir-664[2]	cluster(chr1:220200538-220200619)[1]
chr11	-	64891137	64891246	110	MI0000234	**hsa-mir-192**	hsa-mir-192[1]	mir-192[2]	cluster(chr11:64891137-64902455)[4]
chr11	-	64891355	64891439	85	MI0000732	**hsa-mir-194-2**	hsa-mir-194[2]	mir-194[2]	cluster(chr11:64891137-64902455)[4]
chr11	-	64898363	64898437	75	MI0022595	**hsa-mir-6750**	hsa-mir-6750[1]	hsa-mir-6750[unk]	cluster(chr11:64891137-64902455)[4]
chr11	-	64902387	64902455	69	MI0022594	**hsa-mir-6749**	hsa-mir-6749[1]	hsa-mir-6749[unk]	cluster(chr11:64891137-64902455)[4]

a549_cmb.family.hist

name	rank	count	oppStrand	mm0	mm1	mm2	mm3p	indel	mq0	mq1-19	mq20-29	mq30p	5pOnly	5pPlus	3pOnly	3pPlus
mir-21[1]	1	13773459	34	13559436	210810	2411	802	2947	85	13561	1104870	12654943	13667873	20163	79114	709
mir-17[8]	2	4418353	33	4354406	61730	1509	708	51365	23874	2655518	862533	876428	4050425	3153	363535	332
mir-15[5]	3	2623951	12	2547565	71949	1227	3210	25078	349794	2069445	187803	16909	2615599	5730	2402	44
let-7[12]	4	2344816	356	2276198	66780	1335	503	1010	13975	1986923	315370	28548	2332769	9179	1954	116
mir-27[2]	5	2045761	8	2008907	35527	458	869	309	31894	1968030	15663	30174	34759	193	2005520	4908
mir-322[1]	6	1439026	10	1344363	94043	582	38	25	11	1557	1413471	23987	1429912	97	8928	56
mir-194[2]	7	552585	137	535270	17070	213	32	112	20984	530270	606	725	549354	3004	79	0
mir-30[6]	8	535205	33	523236	11596	217	156	1077	571	507837	25703	1094	229321	316	304212	890
mir-28[3]	9	520700	28	507192	12160	710	638	1105	12640	118526	360907	28627	200355	342	316590	1903
mir-138[2]	10	504451	64	497091	6611	234	515	404	65490	436793	1598	570	502015	404	76	114
mir-26[3]	11	432223	29	426859	5260	78	26	220	2505	284598	135901	9219	431058	1047	53	5
mir-23[2]	12	425156	35	419784	5161	148	63	158	2878	410688	11298	292	114	21	424613	340
mir-34[3]	13	410346	73	405497	4567	76	206	124	4	1018	77295	332029	408776	448	582	24
mir-103[5]	14	390582	264449	239843	147369	3055	315	346	109895	280606	78	3	262259	2140	117236	8812
mir-137[1]	15	333886	8	330789	2981	72	44	84	1	386	32371	301128	0	0	333781	79
mir-24[2]	16	312543	48	308174	4245	92	32	13083	200159	112163	139	82	316	12	312006	147
mir-368[4]	17	307536	106	290806	16619	33	78	46	45428	53842	204588	3678	118	10	307245	40
mir-3074[1]	18	302194	302124	298580	3533	60	21	7788	199828	102312	32	22	301968	72	128	0
hsa-mir-5588[unk]	19	286679	91669	135822	65829	47593	37435	2948	59220	184605	42852	2	3	1397	0	545
mir-224[1]	20	243132	3	239851	2882	306	93	65	1	285	32827	210019	241387	854	688	23
mir-130[4]	21	193187	35	184182	8833	120	52	94	9	6295	64507	122376	131	21	190578	1847
mir-335[1]	22	191476	24	190085	1257	34	100	9	2	250	67334	123890	191178	107	155	21
mir-10[8]	23	184592	382	169433	14539	294	326	91	255	169624	9683	5030	179669	382	3974	37
mir-29[4]	24	169474	488	167387	1642	117	328	1070	6571	145606	12419	4878	3002	76	163446	165
mir-450[3]	25	159351	75	157978	1050	60	263	65	103389	49163	5769	1030	155084	323	3903	12
mir-31[1]	26	142120	22	139258	2712	127	23	36	1	169	31787	110163	140232	151	1691	30
mir-378[5]	27	128323	379	126502	1428	137	256	186	3760	91375	32198	990	1396	2	126490	30
mir-503[1]	28	117607	8	111647	5836	106	18	169	0	136	76853	40618	117092	133	40	99
hsa-mir-3908[unk]	29	115604	88227	4821	58192	8532	44059	3336	24565	30905	60133	1	1	4	0	0
mir-221[2]	30	111054	21	109320	1683	30	21	27	2	136	26745	84171	106565	863	3176	89
mir-191[1]	31	100248	7	98490	1708	35	15	35	1	133	22449	77665	100101	30	54	2

Limitations



- miRBase is extensive but not definitive
  - naming conventions are inconsistent
    - e.g. “groups” for plant vs non-plant species
    - how well do implied groups represent sequence similarity?
- miRBase annotation quality is variable
  - best for extensively studied organisms (human/mouse)
    - mature sequence relationship to loci often absent
- library & alignment limitations
  - small size selection or very short reads
  - inherent ambiguity of alignment process

Thank You!

Anna Battenhouse

Research Associate, Iyer Lab

Nov. 13, 2014

http://mirutils.sourceforge.net