Analysis of small RNA Sequencing Data (Illumina)

miRNA Seq (Illumina) Example Dataset

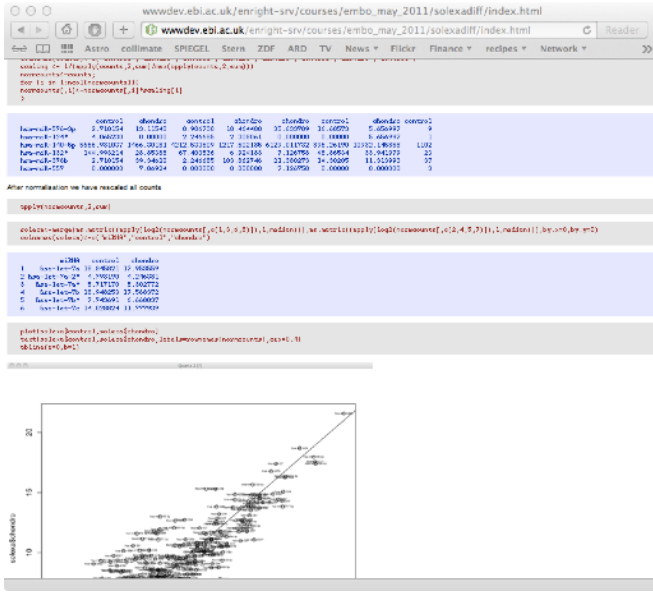
A conserved RNA regulates miRNA turnover and animal behavior through a near-perfect miRNA site

Angelo Bitetti[1,2,3,*], Allison C Mallory[1,2,3,*], Claudia Carrieri[4], Elisabetta Golini[5], Hector Carreño Gutierrez[6], Emerald Perlas[7], Yuvia A. Pérez-Rico[1,2,3,8], Glauco P. Tocchini-Valentini[5], Anton J. Enright[9], William H. J. Norton[6], Silvia Mandillo[5], Dónal O'Carroll[4] and Alena Shkumatava[1,2,3]

6 Samples from Alena
From original Paper

Neural Progenitor Cells
3 Scrambled Controls
3 wt

Not dissimilar from what you have prepared



Considerations for smallRNA seq analysis

- Single-end sequencing 50nt
- Typically more samples fail in a smallRNA seq analysis
- Replicates are essential
- Aim for 4-5 biological replicates per sample if possible
- Huge sequencing depth is not usually required per sample
- 4M reads per sample can usually be sufficient
- 20M reads per sample is ideal

- Spike-ins are sometimes used
  - Can be misleading due to sequencing biases
  - Can be hard to control quantity and accuracy

microRNA Normalisation and Differential expression

Small RNA Seq Normalisation

- Lane Depth Correction
  - Correct total miRNA mapping counts to be equal across all samples

- Statistical approaches better for normalisation and differential expression
  - DEseq
  - BaySeq
  - EdgeR

	Lane1	Lane2	Lane3		Lane1	Lane2	Lane3
	10	34	33		23.1081081	34	39.1875
	11	32	19		25.4189189	32	22.5625
	24	55	44		55.4594595	55	52.25
	11	23	21		25.4189189	23	24.9375
	11	22	21		25.4189189	22	24.9375
	7	5	6		16.1756757	5	7.125
Totals	74	171	144	Totals	171	171	171
Scaling	0.43274854	1	0.84210526				

Mitigating biases ?

- Try and ensure that each replicate of a sample has a different adapter sequence ligated.
  - i.e. Don't multiplex replicates identically.

- 5' Ligation of random nucleotides

- Spike-Ins ?

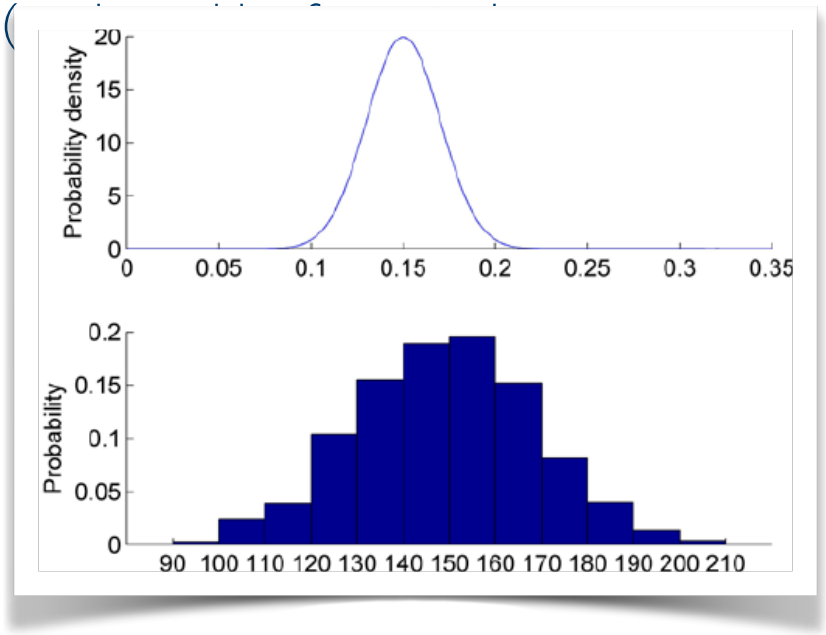
Normalisation of microRNA Count Data

- Correction via total depth normalisation not a good idea
- Count based data should always be normalised with a negative binomial approach
  - **DESeq2**
  - **EdgeR**
  - **Voom**
- Spike-ins difficult to get working correctly for smallRNA runs
- snoRNA can sometimes work to assist normalisation of microRNA samples with extreme differences (e.g. Dicer Knockdown).

Count Data is Different! - NGS Analysis

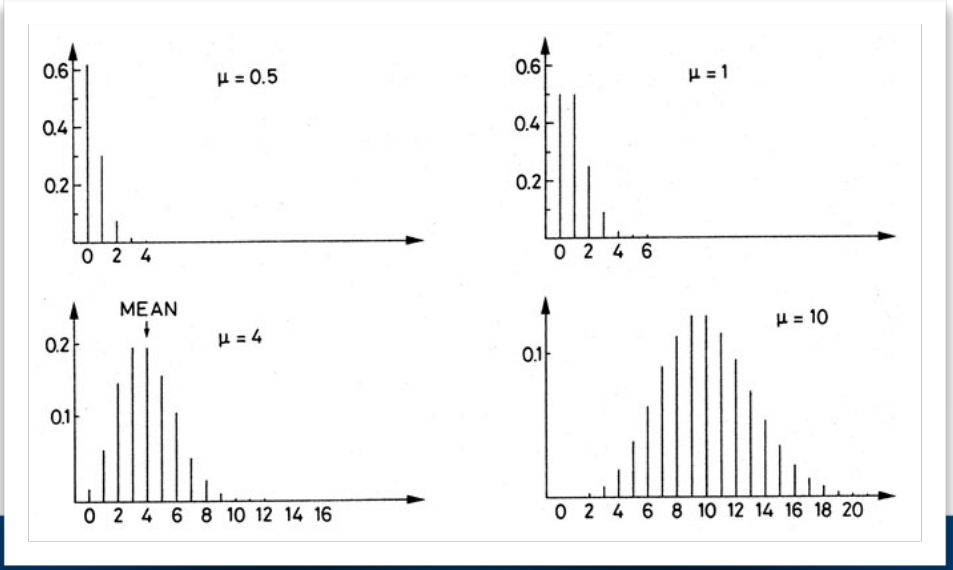
Sequencing Data Statistics

- Sequencing data produces counts not measurements
  - Forms a discrete distribution
  - Very different from continuous distributions (
    probe)
  - Positive values
  - Skewed
  - Heteroscedastic
  - Massive dynamic range
  - Large differences sample to sample



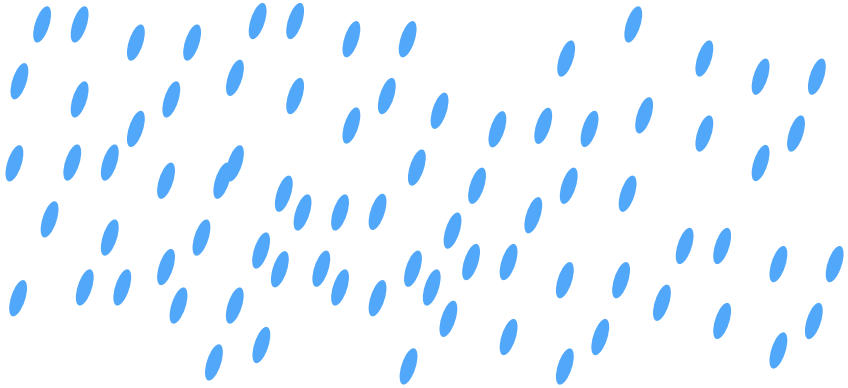
Discrete Count Data - How to model it

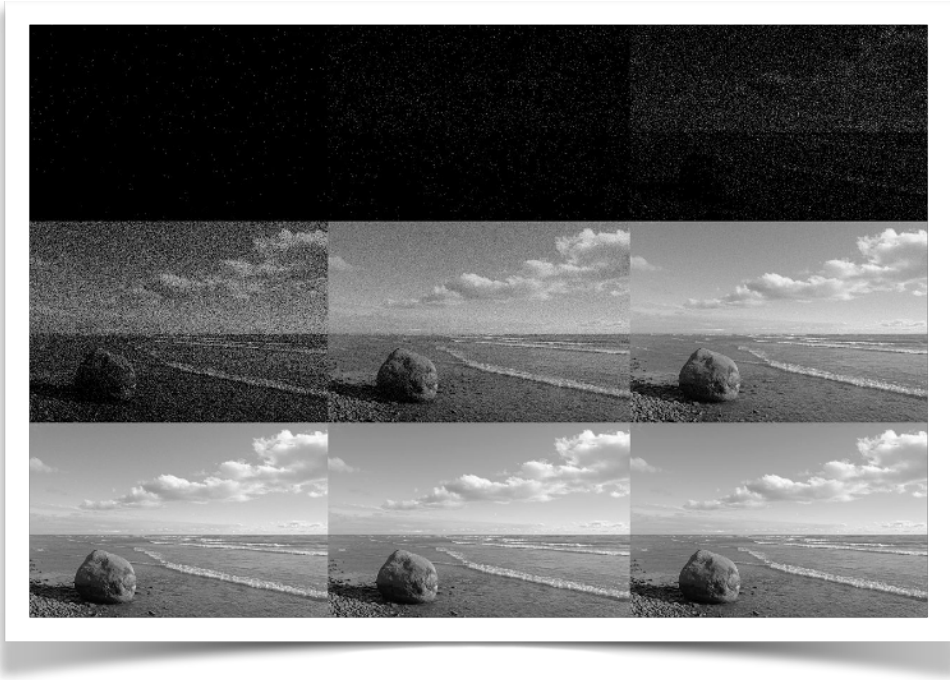
- The Poisson Distribution
- Example: A short,light rain shower with ***r*** drops per meter^2.
- What is the probability to find ***k*** drops on a paving stone of size 1 m^2?
- For poisson, the variance should equal the mean



Discrete Count Data

- Shot Noise



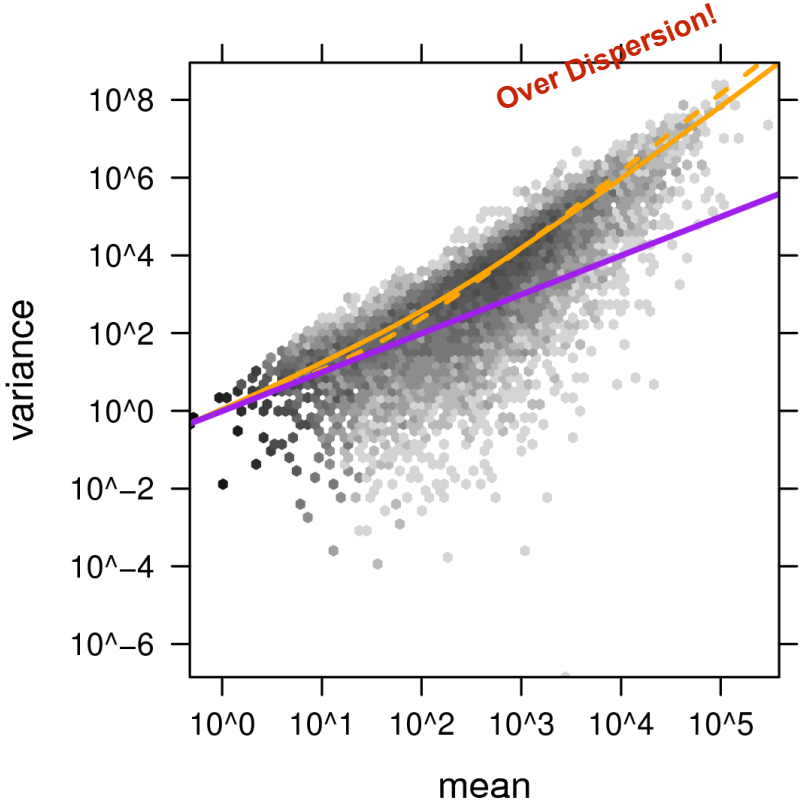


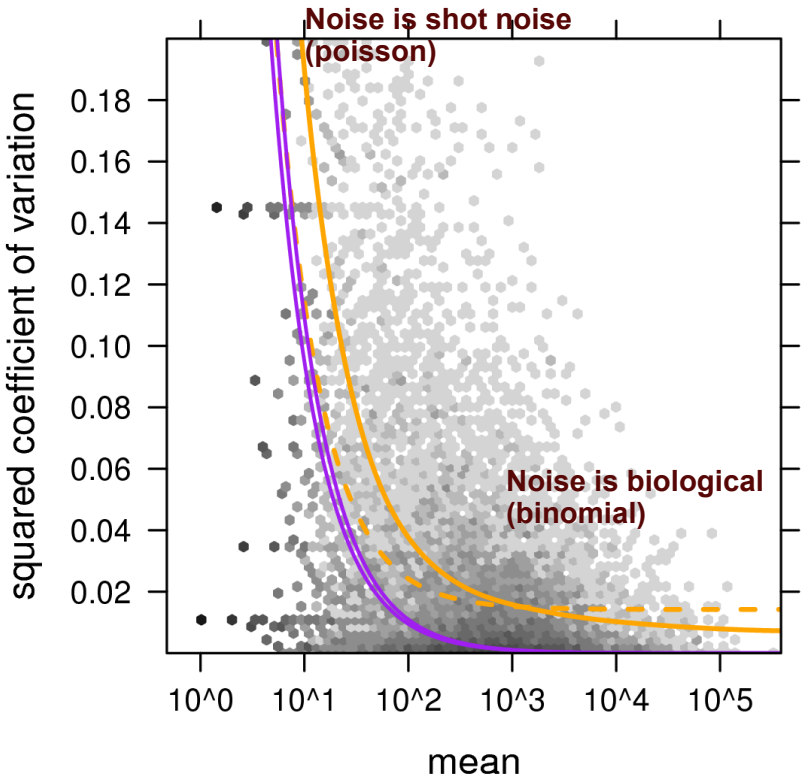

Cygnus Wall - North America Nebula - Cambridge 2014

It's Raining Genes… Hallelujah!

- Shot Noise is very different for individual genes

Sequencing Data - Variance - 2 replicates





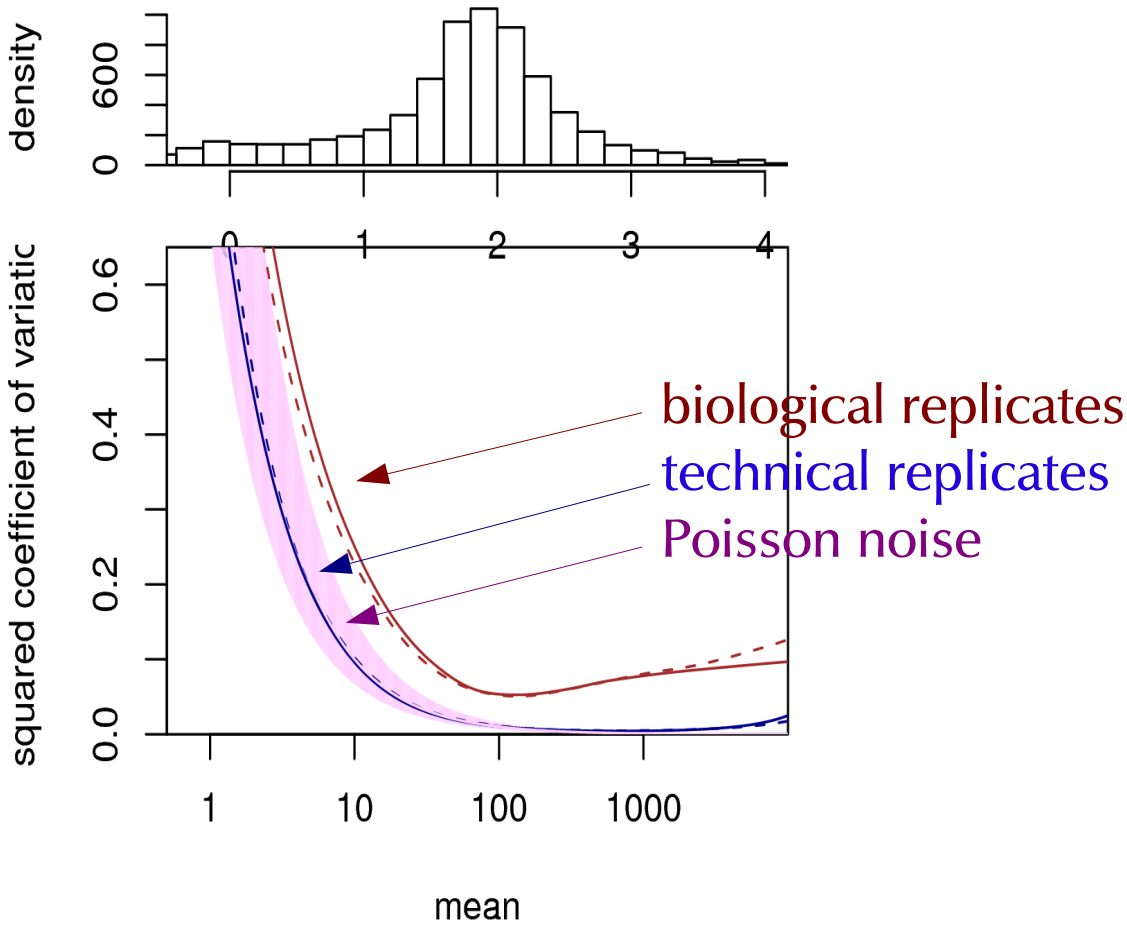
Poisson	$v = \mu$	———
Poisson + constant CV	$v = \mu + \alpha \mu^2$	- - - -
Poisson + local regression	$v = \mu + f(\mu^2)$	———

Poisson Invalidated for genes with high expression as their variance exceeds the model

The variance should be approximately equal to the mean

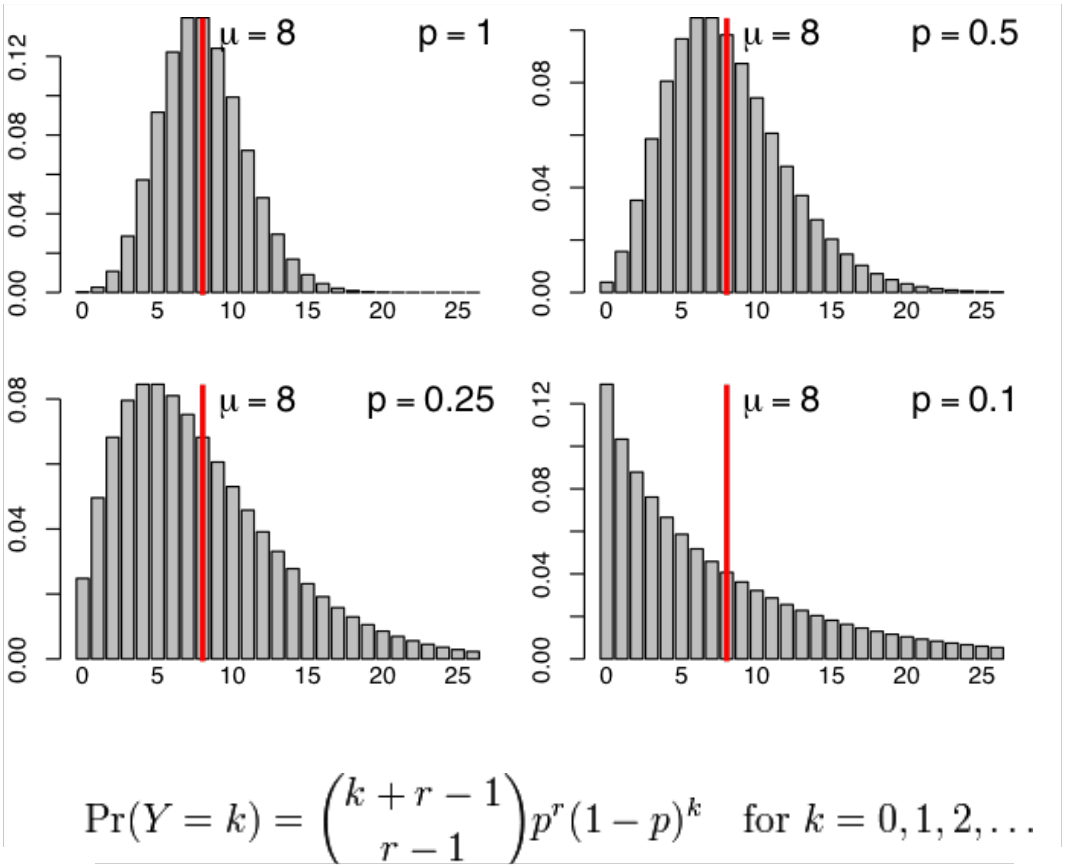
Sequencing Data - Summary

- Shot N
  - unavo
  - domi
- Technic
  - from
  - negli
- Biologi
  - unacc
  - Dom



How to resolve this issue ?

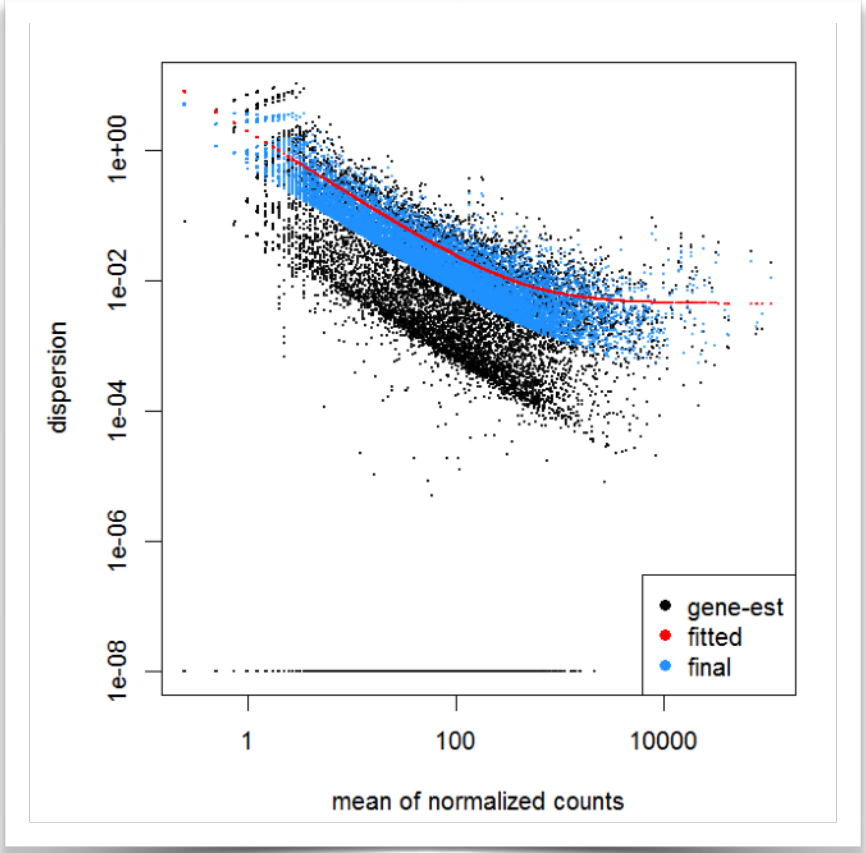
- The negative binomial distribution



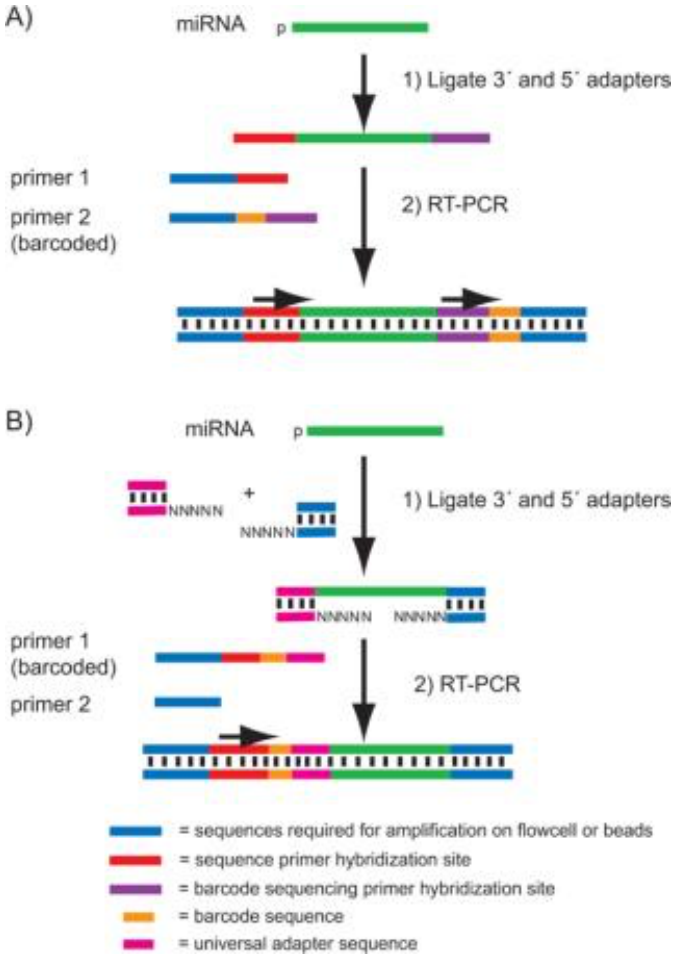
$$\Pr(Y = k) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k \quad \text{for } k = 0, 1, 2, \ldots$$

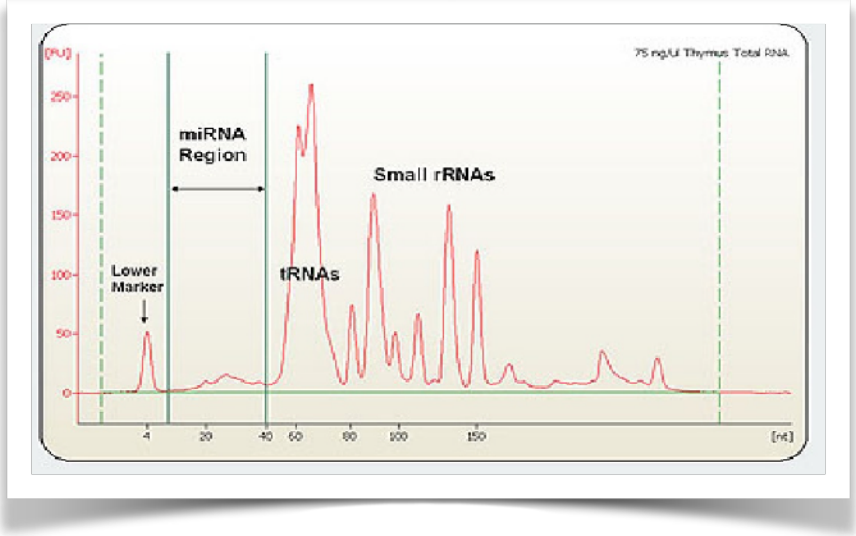
Negative Binomial Approaches

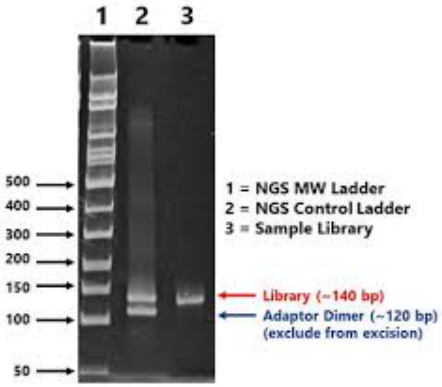
- DESeq2 - Simon Anders & Wolfgang Huber
- EdgeR - Robinson & Smyth



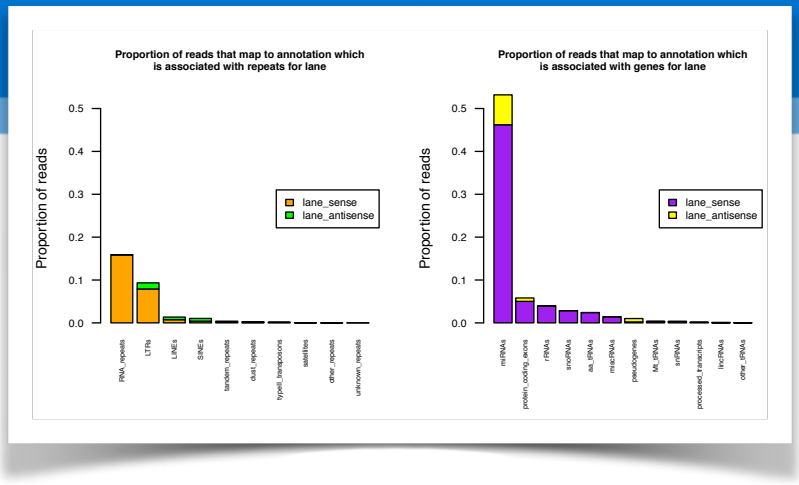
smallRNA Seq



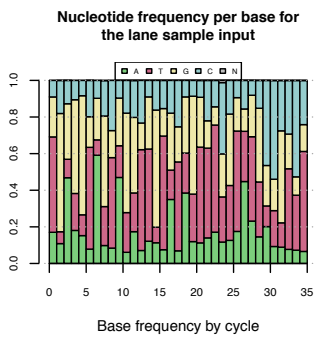


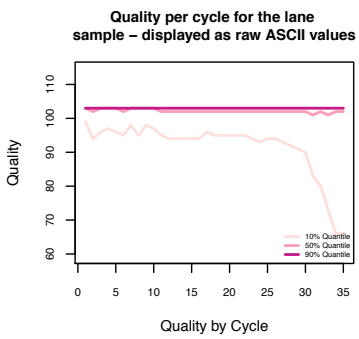


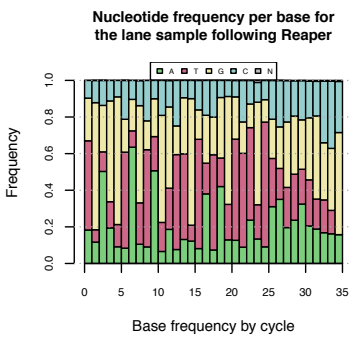
Tools for small RNAseq Analysis

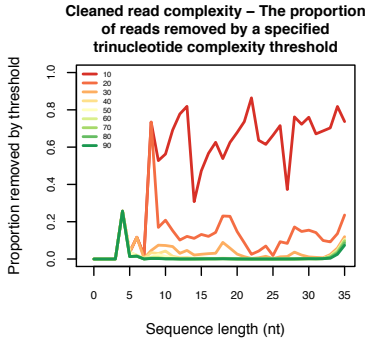


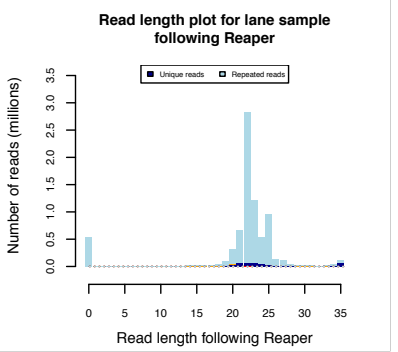
- Galaxy
- FASTX toolkit
- UEA Toolkit
- R & BioConductor
- Reaper, Tally and Kraken







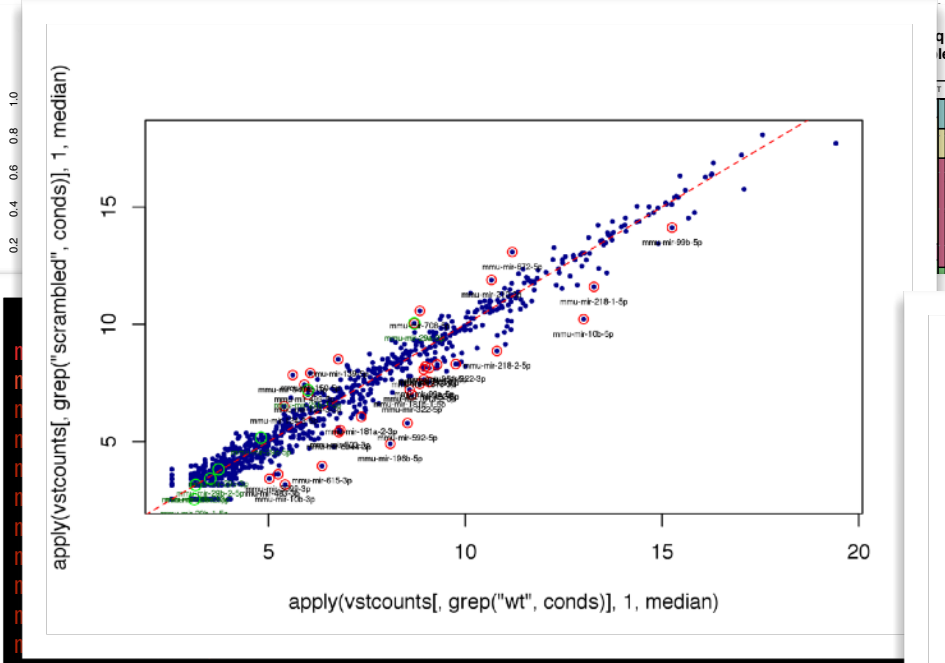


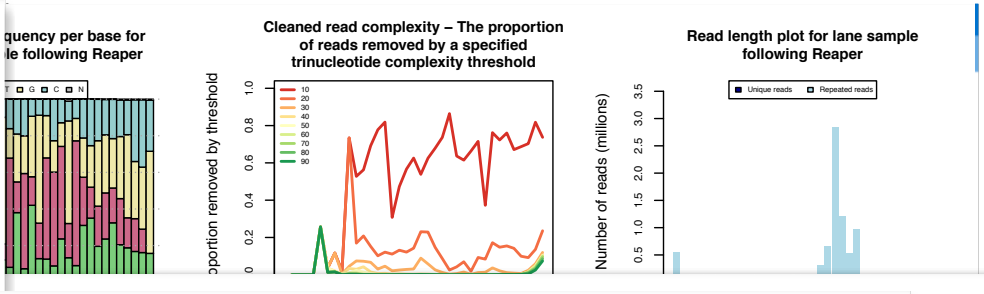


Downstream Analysis

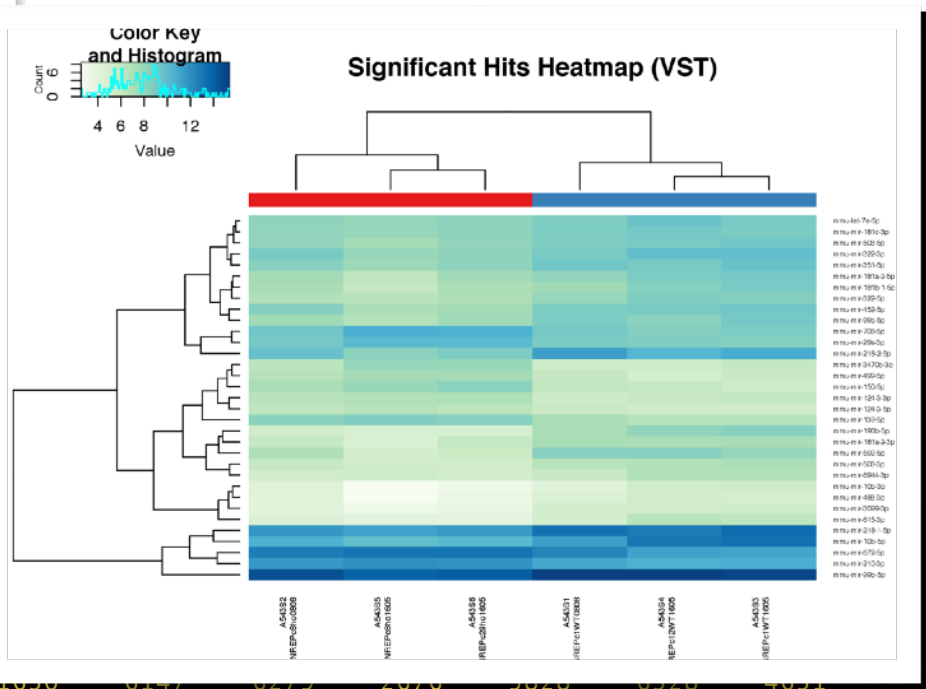
Analysis of microRNA Expression Data

Workflow: Raw miRNA NGS to Results





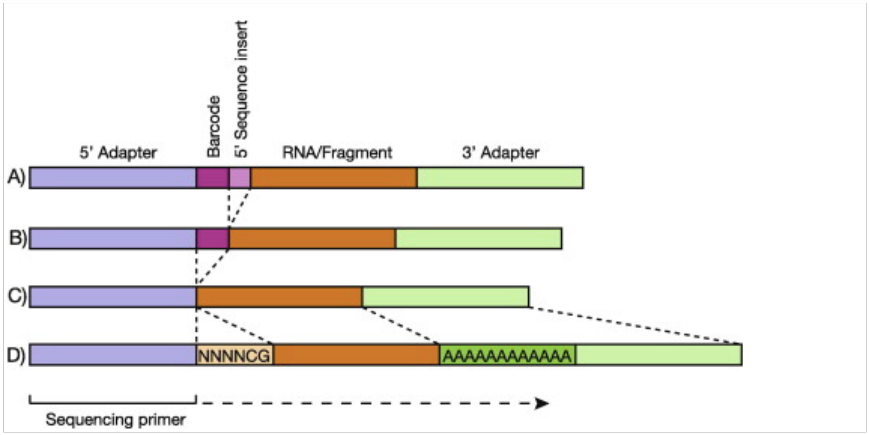
mmu-mir-342-3p	627	607	767	670	591
mmu-mir-26a-2-5p	5656	6673	4293	5680	5627
mmu-mir-5109-3p	509	533	989	641	647
mmu-let-7a-2-5p	2374	3183	2259	2434	2849
mmu-mir-30e-5p	4385	4553	3438	4604	4269
mmu-mir-150-5p	4437	6858	1210	5429	4831
mmu-mir-5099-3p	4162	5294	6956	6017	4928
mmu-mir-186-5p	2670	2818	2888	2601	2505
mmu-mir-148a-3p	3214	3768	5707	3413	4515
mmu-mir-181a-2-5p	1178	1709	2342	1401	1971
mmu-mir-25-3p	2419	2531	2060	2670	2536

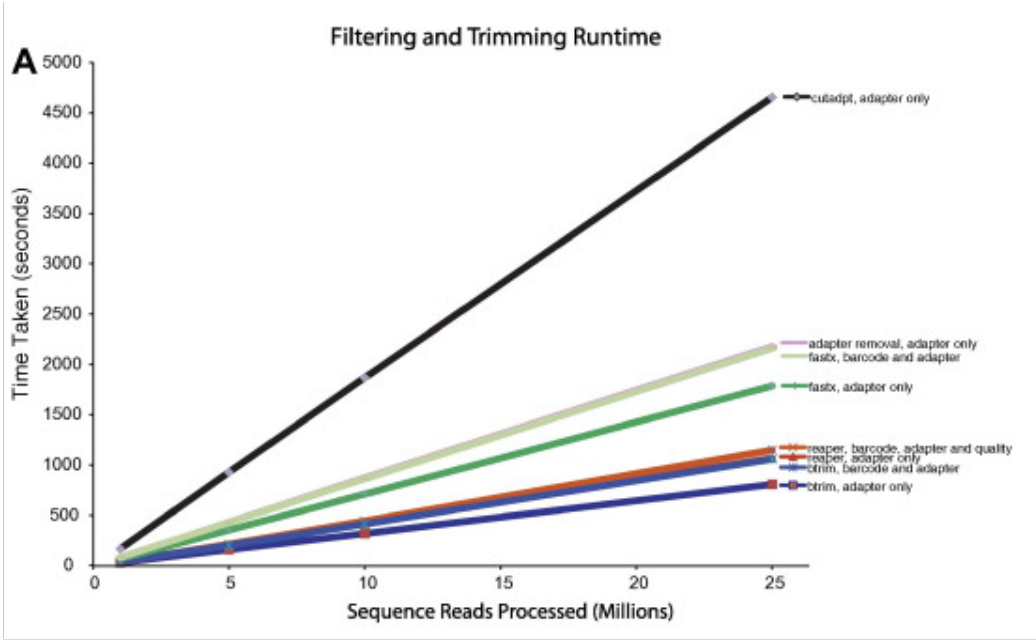


UNIVERSITY OF CAMBRIDGE

The Kraken Suite of Tools for smallRNA Analysis

- **Minion**
  - De bruijn graph 3' assembly analysis for small RNA reads
  - Can identify adapter sequences and large-scale contaminants
- **Reaper**
  - Extremely fast and accurate adapter finder and trimmer
  - Also cleans
    - Low-complexity tracts
    - PolyA, PolyN
    - Low-scoring tracts
  - Deals with complex read geometries and random seqs
- **Tally**
  - Collapses redundant reads together rapidly





UNIVERSITY OF CAMBRIDGE

Kraken: A set of tools for quality control and analysis of high-throughput sequence data
Davis, Matthew, van Dongen Stijn, Abreu-Goodger Cei, Bartonicek Nenad, and Enright Anton J.
Methods. (2013)

Turning Processed Reads into Counts on microRNAs

```
@HS24_10147:1:1101:1067:1989#0
GCCCGCGCTAGCTCAGTCCCTAGAGGATCCCAGTGCAATTGTCCCCTCCCN
+
BCCFF
@HS24
TCGCT
+
BBCFF
```

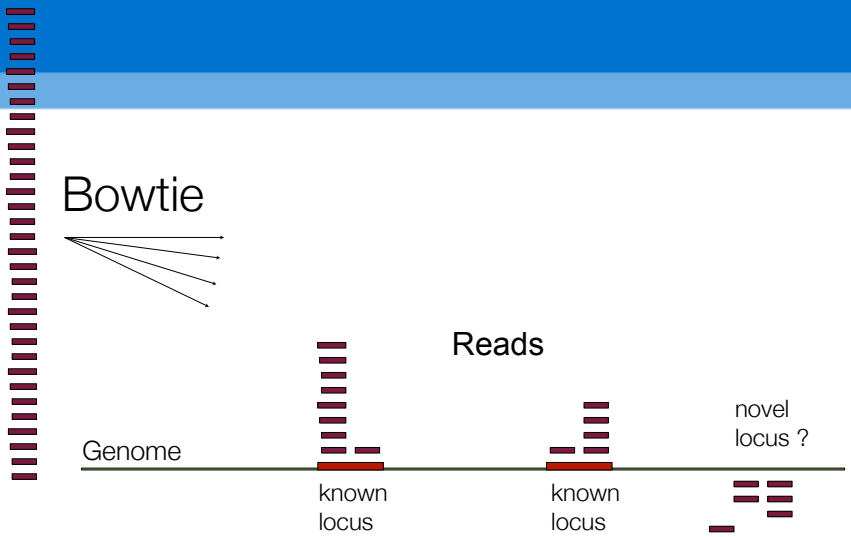
```
>1_t11_w33_x157998
GTTTCCGTAGTGTAGTGGTTATCACGTTCGCCT
>2_t11_w32_x80579
GCATTGGTGGTTCAGTGGTAGAATTCTCGCCT
>3_t10_w36_x27074
```

	ko3	ko4	ko6	wt1	wt2	wt5	exoko3	exoko4	exoko6	exowt2	exowt3	exowt5
mmu-mir-21a-5p	33561	26530	38601	34692	32023	29922	66165	54907	20833	56207	66980	33955
mmu-mir-92a-1-3p	28479	29132	26069	37165	36005	24342	46675	41345	16045	49316	55999	31826
mmu-mir-27b-3p	29042	23071	47952	23737	25318	28089	50006	38837	24160	39562	40224	30662
mmu-mir-191-5p	12427	11341	17954	13162	14266	12682	27299	31861	13093	38412	33249	20480
mmu-mir-142-5p	25394	25189	23562	28321	25120	16983	24762	22825	7703	22614	24194	12249
mmu-mir-182-5p	11649	10632	22673	10975	13652	17014	19423	17306	7473	18978	19913	11553
mmu-mir-16-2-5p	9108	10131	7848	8892	8200	6021	15566	20059	5009	12214	13723	9118
mmu-let-7f-2-5p	8067	10065	9573	9193	9740	7885	17642	16180	3971	16347	19757	7533
mmu-mir-92a-2-3p	7579	8116	7360	8248	7350	5298	15695	15029	6080	12634	14693	9517
mmu-let-7i-5p	6779	6875	6660	6448	6025	5110	13050	12705	4014	9939	11520	5813
mmu-mir-22-3p	4009	3224	10072	3418	3199	6876	12495	10807	5810	8814	10434	9580
mmu-mir-342-3p	627	607	767	670	591	581	6404	4218	2507	11565	9414	5303
mmu-mir-26a-2-5p	5656	6673	4293	5680	5627	3160	9024	10408	1889	8836	9367	3577
mmu-mir-5109-3p	509	533	989	641	647	750	3964	3432	1583	3604	3459	8811
mmu-let-7a-2-5p	2374	3183	2259	2434	2849	1831	7530	7570	1765	7058	8507	3296
mmu-mir-30e-5p	4385	4553	3438	4604	4269	2656	7653	7930	2277	6561	6912	4121
mmu-mir-150-5p	4437	6858	1210	5429	4831	986	6483	6694	768	5442	6969	1413
mmu-mir-5099-3p	4162	5294	6956	6017	4928	5860	5766	4972	2184	4623	5696	2644
mmu-mir-186-5p	2670	2818	2888	2601	2505	1950	6329	6524	1965	5874	5847	3680
mmu-mir-148a-3p	3214	3768	5707	3413	4515	4453	5320	6501	3769	5812	5436	5619
mmu-mir-181a-2-5p	1178	1709	2342	1401	1971	1785	5358	6478	1676	5687	5467	2561
mmu-mir-25-3p	2419	2531	2060	2670	2536	1630	6147	6279	2676	5828	6328	4631

Mapping small RNA Reads

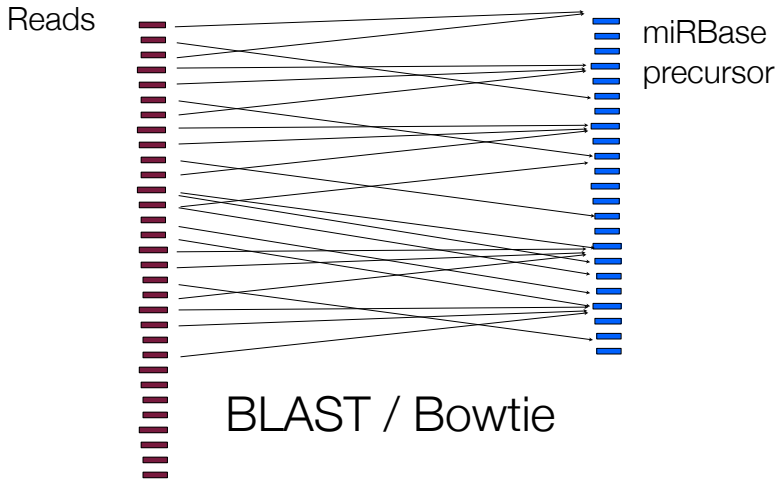
- **Genome Based Approach**

- Map all reads to the genome with an aligner tool
  - Bowtie: http://bowtie-bio.sourceforge.net
  - Select reads that overlap known miRNA locii
  - Problems resolving depth across loci
  - Runtime: 4-12 hours

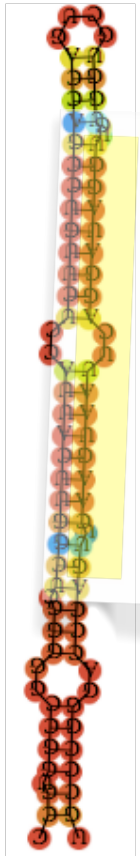


- **Read versus MicroRNA Approach**

- Compare reads directly against miRBase sequences
  - Fast and accurate
  - look for >95% identity and no more than 1-2 mismatches
  - Runtime: 10-15 minutes



MicroRNA Read Level Analysis



mmu-mir-26b-5p

Depth: **5741** reads

Aligned Reads on Precursor

```
UGCCCGGGACCCAGUUCAAGUAAUUCAGGAUAGGUUGUGGUGCUGACCAGCCUGUUCUCCAUUACUUGGCUCGGGGGCCGGUGCC
.((((((..((((((.(((((((((..(((((((((..((((....))))))))))))))))).))))))))))).))))..)))). (-42.70)
                UUCAAGUAAUUCAGGAUAGGUU     103_t9_w22 Depth:3837 Modification: no_modification
                UUCAAGUAAUUCAGGAUAGGU      481_t10_w21 Depth:698 Modification: no_modification
                UUCAAGUAAUUCAGGAUAGGUUU    771_t9_w23 Depth:360 Modification: _nont_3p_U
                UUCAAGUAAUUCAGGAUAGGUA     896_t11_w22 Depth:291 Modification: _nont_3p_A
                UUCAAGUAAUUCAGGAUAGGUUA    1300_t9_w23 Depth:175 Modification: _nont_3p_A
                UUCAAGUAAUUCAGGAUAGG       1565_t10_w20 Depth:140 Modification: no_modification
                UUCAAGUAAUUCAGGAUAGGUAA    3127_t11_w23 Depth:58 Modification: _nont_3p_AA
                UUCAAGUAAUUCAGGAUAGGUAU    3388_t10_w23 Depth:52 Modification: _nont_3p_AU
                UUCAAGUAAUUCAGGAUAGGUUG    5767_t9_w23 Depth:26 Modification: no_modification
                UUCAAGUAAUUCAGGAUAGGUUUU   8212_t10_w24 Depth:17 Modification: _nont_3p_UU
                UUCAAGUAAUUCAGGAUAGGCU     8687_t9_w22 Depth:16 Modification: _nont_3p_CU
                UUCAAGUAAUUCAGGAUAGGA      9287_t11_w21 Depth:14 Modification: _nont_3p_A
               CUUCAAGUAAUUCAGGAUAGGUU     10438_t9_w23 Depth:13 Modification: _nont_5p_C
               NUCAAGUAAUUCAGGAUAGGUU      11348_t8_w22 Depth:11 Modification: _nont_5p_N
                  AAGUAAUUCAGGAUAGGUU      13435_t6_w19 Depth:9 Modification: no_modification
                UUCAAGUAAUUCAGGAUAGGUUGU   13682_t8_w24 Depth:9 Modification: no_modification
                UUCAAGUAAUUCAGGAUAG        15177_t9_w19 Depth:8 Modification: no_modification
                UUCAAGUAAUUCAGGAUAGGUUC    15767_t12_w23 Depth:7 Modification: _nont_3p_C
```

Distribution across Precursor