ERCC Data Analysis

Use Case 3:
Biomarker Potential and
Limitations of Circulating miRNA

Performed by the
Data Management and Resource Repository
(DMRR)

Williams Z., et al. (2013) Comprehensive profiling of circulating microRNA via small RNA sequencing of cDNA libraries reveals biomarker potential and limitations. PNAS 110: 4255–4260.

Use Case 3: Biomarker Potential and Limitations of Circulating miRNA

The goal of this use case is to show that the Genboree Workbench and the exceRpt small RNA-seq analysis pipeline can replicate the results of Williams et al. We have developed these pipelines because as the Extracellular RNA Communication Consortium (ERCC) begins to generate more datasets, it is vital that they be analyzed in a reproducible and comparable way.

The dataset from Williams et al is freely available in the Short Read Archive (SRA), so it makes a good practice dataset for becoming familiar with the Workbench and exceRpt.

Williams Z., et al. (2013) Comprehensive profiling of circulating microRNA via small RNA sequencing of cDNA libraries reveals biomarker potential and limitations. PNAS 110: 4255–4260.

Biological Question of Interest

This work from the Tuschl lab at Rockefeller University is mainly a technical study. They examined extracellular RNA from the bloodstream of mothers and fathers, their newborn babies, and the placenta. They were interested in quantitating how much exRNA could be found, and whether it would be possible to detect biomarker RNAs at that level. The placenta acted as a model of a tumor; their thinking was that if detecting exRNAs from placenta is feasible, the same should be true of tumors.

Williams Z., et al. (2013) Comprehensive profiling of circulating microRNA via small RNA sequencing of cDNA libraries reveals biomarker potential and limitations. PNAS 110: 4255–4260.

3

Use Case 3: Biomarker Potential and Limitations of Circulating miRNA

Biological Samples to Be Analyzed

Unique ID	Sample Name	SRA ID	Sample Description
31	C01	SRR822433	Non-pregnant woman plasma Sample #1
32	C02	SRR822461	Non-pregnant woman plasma Sample #2
33	C03	SRR822434	Non-pregnant woman serum Sample #3
34	C04	SRR822435	Non-pregnant woman serum Sample #4
35	C05	SRR822436	Non-pregnant woman serum Sample #5
18	F01	SRR822466	Father's blood from Set #1
19	F06	SRR822441	Father's blood from Set #6
20	F08	SRR822442	Father's blood from Set #8
21	F10	SRR822443	Father's blood from Set #10
22	F12	SRR822440	Father's blood from Set #12
1	M01	SRR822467	Mother's blood from Set #1
2	M02	SRR822468	Mother's blood from Set #2
3	M03	SRR822469	Mother's blood from Set #3
4	M06	SRR822445	Mother's blood from Set #6
5	M08	SRR822446	Mother's blood from Set #8
6	M10	SRR822447	Mother's blood from Set #10
7	M12	SRR822448	Mother's blood from Set #12

Unique ID	Sample Name	SRA ID	Sample Description
23	P05	SRR822452	Placenta from Set #5
24	P06	SRR822453	Placenta from Set #6
25	P07	SRR822454	Placenta from Set #7
26	P08	SRR822455	Placenta from Set #8
27	P09	SRR822456	Placenta from Set #9
28	P10	SRR822457	Placenta from Set #10
29	P11	SRR822458	Placenta from Set #11
30	P12	SRR822459	Placenta from Set #12
8	B01	SRR822462	Umbilical cord blood from Set #1
9	B02	SRR822464	Umbilical cord blood from Set #2
10	B03	SRR822464	Umbilical cord blood from Set #3
11	B04	SRR822465	Umbilical cord blood from Set #4
12	B06	SRR822437	Umbilical cord blood from Set #6
13	B07	SRR822450	Umbilical cord blood from Set #7
14	B08	SRR822438	Umbilical cord blood from Set #8
15	B10	SRR822439	Umbilical cord blood from Set #10
16	B11	SRR822451	Umbilical cord blood from Set #11
17	B12	SRR822444	Umbilical cord blood from Set #12

The dataset analyzed in this use case is available from the Short Read Archive (SRA) under Bioproject PRJNA187509.

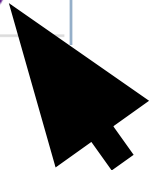http://www.ncbi.nlm.nih.gov/bioproject/PRJNA187509

Use Case 3: Data Scrubbing

To match the sample names from the article with SRA accessions from the exceRpt output, we visit the Bioproject page at the SRA:

http://www.ncbi.nlm.nih.gov/bioproject/PRJNA187509

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
SRA Experiments	38

Williams Z., et al. (2013) Comprehensive profiling of circulating microRNA via small RNA sequencing of cDNA libraries reveals biomarker potential and limitations. PNAS 110: 4255–4260.

Use Case 3: Data Scrubbing

To match the sample names from the article with SRA accessions from the exceRpt output, we visit the Bioproject page at the SRA:

http://www.ncbi.nlm.nih.gov/bioproject/PRJNA187509

mother_trio_1

1 ILLUMINA (Illumina Genome Analyzer II) run: 1.5M spots, 53.1M bases, 47.5Mb downloads

Accession: SRX262174

Williams Z., et al. (2013) Comprehensive profiling of circulating microRNA via small RNA sequencing of cDNA libraries reveals biomarker potential and limitations. PNAS 110: 4255–4260.

Use Case 3: Data Scrubbing

Study: circulating miRNA

PRJNA187509 • SRP018255 • All experiments • All runs

show Abstract

→ **Sample:** M1

SAMN01906661 • SRS402752 • All experiments • All runs

Organism: Homo sapiens

Library:

Instrument: Illumina Genome Analyzer II

Strategy: miRNA-Seq

Source: TRANSCRIPTOMIC

Selection: size fractionation

Layout: SINGLE

Spot descriptor:

1 forward

Runs: 1 run, 1.5M spots, 53.1M bases, 47.5Mb

Run	# of Spots	# of Bases	Size	Published
→ SRR822467	1,476,063	53.1M	47.5Mb	2015-07-22

Williams Z., et al. (2013) Comprehensive profiling of circulating microRNA via small RNA sequencing of cDNA libraries reveals biomarker potential and limitations. PNAS 110: 4255–4260.

Use Case 3: Data Comparison

After matching sample IDs between the two pipelines, we can import both tables into Excel and start generating comparison plots.

Williams Z., et al. (2013) Comprehensive profiling of circulating microRNA via small RNA sequencing of cDNA libraries reveals biomarker potential and limitations. PNAS 110: 4255–4260.

Results from the Article's analysis pipeline

Supplementary Table 1. Complete small RNA annotation statistics and RNA spike contribution in the study samples.

Unique ID	Sample Name	Sample Description	Total reads (incl. calib.)	Unique reads	Calibrator reads	% Calibrator reads	Human and viral miRNAs	% miRNA	rRNA	% rRNA	tRNA	% tRNA	sn/snoRNA	% sn/snoRNA	Other	% Other	read counts	% of total reads	[Plasma] fM	fmol/mcg of RNA
1	M1	Mother's blood from Set #1	1461557	242325	8180	0.6%	1144456	78.7%	33271	2.3%	34800	2.4%	2040	0.1%	238808	16.4%	1144456	78.7%	875	35.0
2	M2	Mother's blood from Set #2	924989	172060	14361	1.6%	685641	75.3%	22019	2.4%	24109	2.6%	1339	0.1%	177520	19.5%	685641	75.3%	300	11.9
3	M3	Mother's blood from Set #3	704769	156603	12846	1.8%	475326	68.7%	17339	2.5%	27495	4.0%	902	0.1%	170861	24.7%	475326	68.7%	225	9.3
4	M6	Mother's blood from Set #6	7463150	698541	584532	7.8%	5188084	75.4%	297433	4.3%	141226	2.1%	12111	0.2%	1239741	18.0%	5188084	75.4%	375	22.2
5	M8	Mother's blood from Set #8	3152642	545369	675011	21.4%	1336587	53.9%	133508	5.4%	148191	6.0%	9314	0.4%	849993	34.3%	1336587	53.9%	100	5.0
6	M10	Mother's blood from Set #10	4115354	632382	888468	21.6%	1528714	47.4%	379598	11.8%	236654	7.3%	19754	0.6%	1062013	32.9%	1528714	47.4%	75	4.3
7	M12	Mother's blood from Set #12	862603	175740	274366	31.8%	247898	42.1%	44448	7.6%	19772	3.4%	1806	0.3%	274288	46.6%	247898	42.1%	75	2.3
8	B1	Umbilical cord blood from Set #1	1355484	130111	9771	0.7%	1184912	88.1%	13889	1.0%	13845	1.0%	1791	0.1%	131276	9.8%	1184912	88.1%	1550	30.3
9	B2	Umbilical cord blood from Set #2	570482	81611	9529	1.7%	455953	81.3%	9652	1.7%	10356	1.8%	694	0.1%	84298	15.0%	455953	81.3%	500	12.0
10	B3	Umbilical cord blood from Set #3	1260465	138546	10094	0.8%	1090103	87.2%	16050	1.3%	14428	1.2%	1149	0.1%	128641	10.3%	1090103	87.2%	675	27.0
11	B4	Umbilical cord blood from Set #4	1605555	151094	15519	1.0%	1394689	87.7%	14128	0.9%	20433	1.3%	1222	0.1%	159564	10.0%	1394689	87.7%	825	22.5
12	B6	Umbilical cord blood from Set #6	11622431	390880	698958	6.0%	10228091	93.6%	52849	0.5%	30911	0.3%	14371	0.1%	597215	5.5%	10228091	93.6%	3850	36.6
13	B7	Umbilical cord blood from Set #7	8502657	112840	316070	3.7%	8064841	98.5%	15940	0.2%	13900	0.2%	7078	0.1%	83603	1.0%	8064841	98.5%	319000	255.2
14	B8	Umbilical cord blood from Set #8	13352447	1561900	911780	6.8%	4291113	34.5%	1817837	14.6%	283911	2.3%	27035	0.2%	6020760	48.4%	4291113	34.5%	380	11.8
15	B10	Umbilical cord blood from Set #10	4456395	337320	384301	8.6%	3451453	84.8%	88175	2.2%	61430	1.5%	10201	0.3%	460806	11.3%	3451453	84.8%	1101	22.5
16	B11	Umbilical cord blood from Set #11	2269555	62809	364469	16.1%	1843585	96.8%	7875	0.4%	9397	0.5%	2223	0.1%	41344	2.2%	1843585	96.8%	63250	126.5
17	B12	Umbilical cord blood from Set #12	3728380	582793	973291	26.1%	960402	34.9%	362914	13.2%	172609	6.3%	14506	0.5%	1244645	45.2%	960402	34.9%	75	2.5
18	F1	Father's blood from Set #1	1111578	88184	8075	0.7%	981143	88.9%	8749	0.8%	6357	0.6%	1292	0.1%	105959	9.6%	981143	88.9%	750	30.4
19	F6	Father's blood from Set #6	2178003	362306	383039	17.6%	1122571	62.5%	80610	4.5%	42870	2.4%	5296	0.3%	543588	30.3%	1122571	62.5%	100	7.3
20	F8	Father's blood from Set #8	3074738	514689	395819	12.9%	1407053	52.5%	171556	6.4%	51608	1.9%	5723	0.2%	1042949	38.9%	1407053	52.5%	100	8.9
21	F10	Father's blood from Set #10	4969103	699866	539790	10.9%	2483642	56.1%	459269	10.4%	309906	7.0%	19412	0.4%	1157052	26.1%	2483642	56.1%	200	11.5
22	F12	Father's blood from Set #12	1490566	280408	386307	25.9%	385004	34.9%	130696	11.8%	59727	5.4%	7298	0.7%	521508	47.2%	385004	34.9%	50	2.5
23	P5	Placenta from Set #5	2787458	127520	209373	7.5%	2413377	93.6%	26983	1.0%	42941	1.7%	3744	0.1%	90012	3.5%	2413377	93.6%	na	28.8
24	P6	Placenta from Set #6	2684368	115455	181907	6.8%	2376282	95.0%	22740	0.9%	22798	0.9%	4143	0.2%	75636	3.0%	2376282	95.0%	na	32.7
25	P7	Placenta from Set #7	6695605	206000	222412	3.3%	6194430	95.7%	40181	0.6%	61827	1.0%	10947	0.2%	165022	2.5%	6194430	95.7%	na	69.6
26	P8	Placenta from Set #8	2283764	105952	187804	8.2%	1957614	93.4%	20130	1.0%	47576	2.3%	8656	0.4%	61640	2.9%	1957614	93.4%	na	26.1
27	P9	Placenta from Set #9	2214654	95554	226597	10.2%	1879868	94.6%	16805	0.8%	27551	1.4%	6099	0.3%	57145	2.9%	1879868	94.6%	na	20.7
28	P10	Placenta from Set #10	4941885	162188	255224	5.2%	4476164	95.5%	27849	0.6%	52471	1.1%	10720	0.2%	118419	2.5%	4476164	95.5%	na	43.8
29	P11	Placenta from Set #11	3639847	128297	162793	4.5%	3319417	95.5%	26530	0.8%	34122	1.0%	10466	0.3%	86163	2.5%	3319417	95.5%	na	51.0
30	P12	Placenta from Set #12	4113247	146566	222581	5.4%	3716537	95.5%	22344	0.6%	38308	1.0%	7412	0.2%	105536	2.7%	3716537	95.5%	na	41.7
31	C1	Non-pregnant woman plasma Sample #1	4114040	476303	754314	18.3%	2223967	66.2%	173745	5.2%	94857	2.8%	9839	0.3%	857264	25.5%	2223967	66.2%	160	7.4
32	C2	Non-pregnant woman plasma Sample #2	223901	61913	3982	1.8%	149013	67.8%	8936	4.1%	6969	3.2%	698	0.3%	54303	24.7%	149013	67.8%	235	47.7
33	C3	Non-pregnant woman serum Sample #3	2045575	329169	565916	27.7%	194526	13.1%	95743	6.5%	30155	2.0%	6355	0.4%	1152866	77.9%	194526	13.1%	21	0.9
34	C4	Non-pregnant woman serum Sample #4	1305137	283307	580383	44.5%	128434	17.7%	86008	11.9%	48317	6.7%	4343	0.6%	457635	63.1%	128434	17.7%	19	0.6
35	C5	Non-pregnant woman serum Sample #5	1289020	267413	608361	47.2%	73980	10.9%	90451	13.3%	42533	6.2%	3979	0.6%	469686	69.0%	73980	10.9%	8	0.3

Column group headers: Samples | Extraction statistics sample | Small RNA annotation categories* excl. calib. (miRNA reads based on mapping to mirna precursor) | miRNA

* other category includes viral miRNAs, piRNAs, non-prototypical doubtful miRNAs, 19-24 oligonucleotide marker used during cDNA library preparation, miscellaneous RNA, repeat masker, sequences that mapped to the genome but have no annotation, sequences that did not map to the genome and have no annotation

Williams Z., et al. (2013) Comprehensive profiling of circulating microRNA via small RNA sequencing of cDNA libraries reveals biomarker potential and limitations. PNAS 110: 4255–4260.

Use Case 3: Biomarker Potential and Limitations of Circulating miRNA

We can compare the number of reads mapped to various categories of RNA in Table S1 from the article with the same data from the exceRpt read mapping summary.

Categories in common are total input reads, microRNA, ribosomal RNA, and transfer RNA.

Williams Z., et al. (2013) Comprehensive profiling of circulating microRNA via small RNA sequencing of cDNA libraries reveals biomarker potential and limitations. PNAS 110: 4255–4260.

Use Case 3: exceRpt Pipeline Results - Number of Input Reads
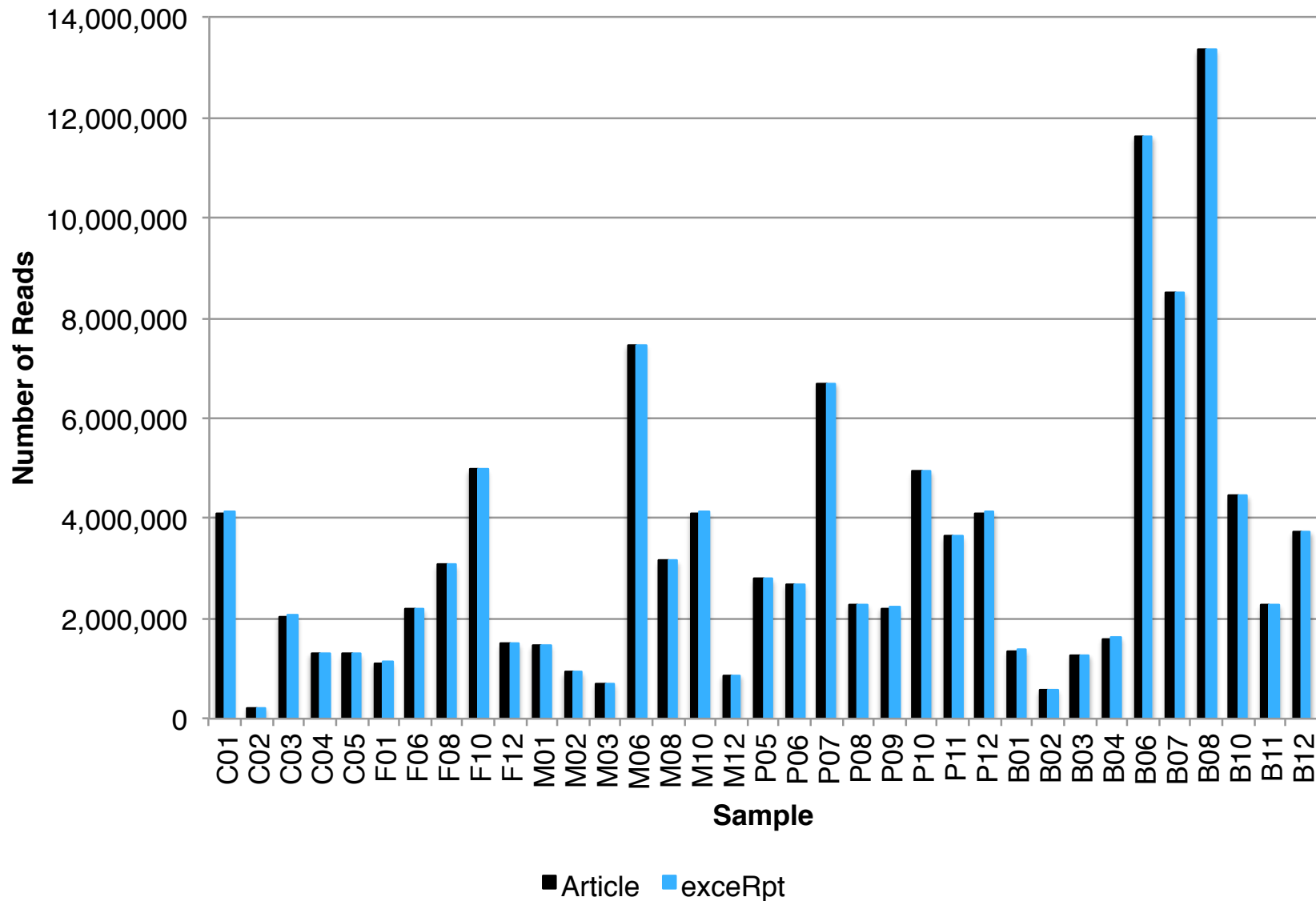
Use Case 3: exceRpt Pipeline Results – MicroRNA
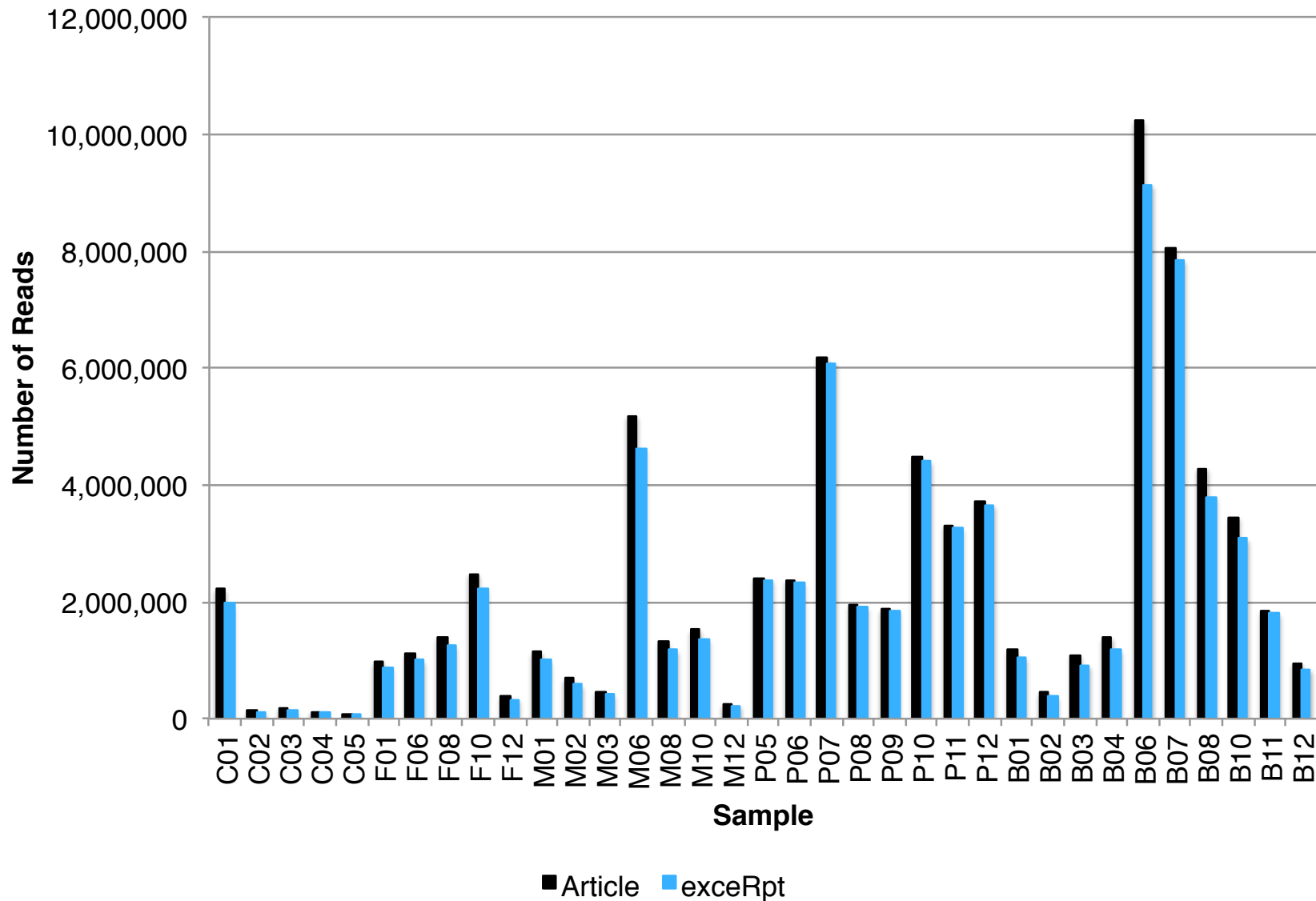
Use Case 3: exceRpt Pipeline Results - Ribosomal RNA
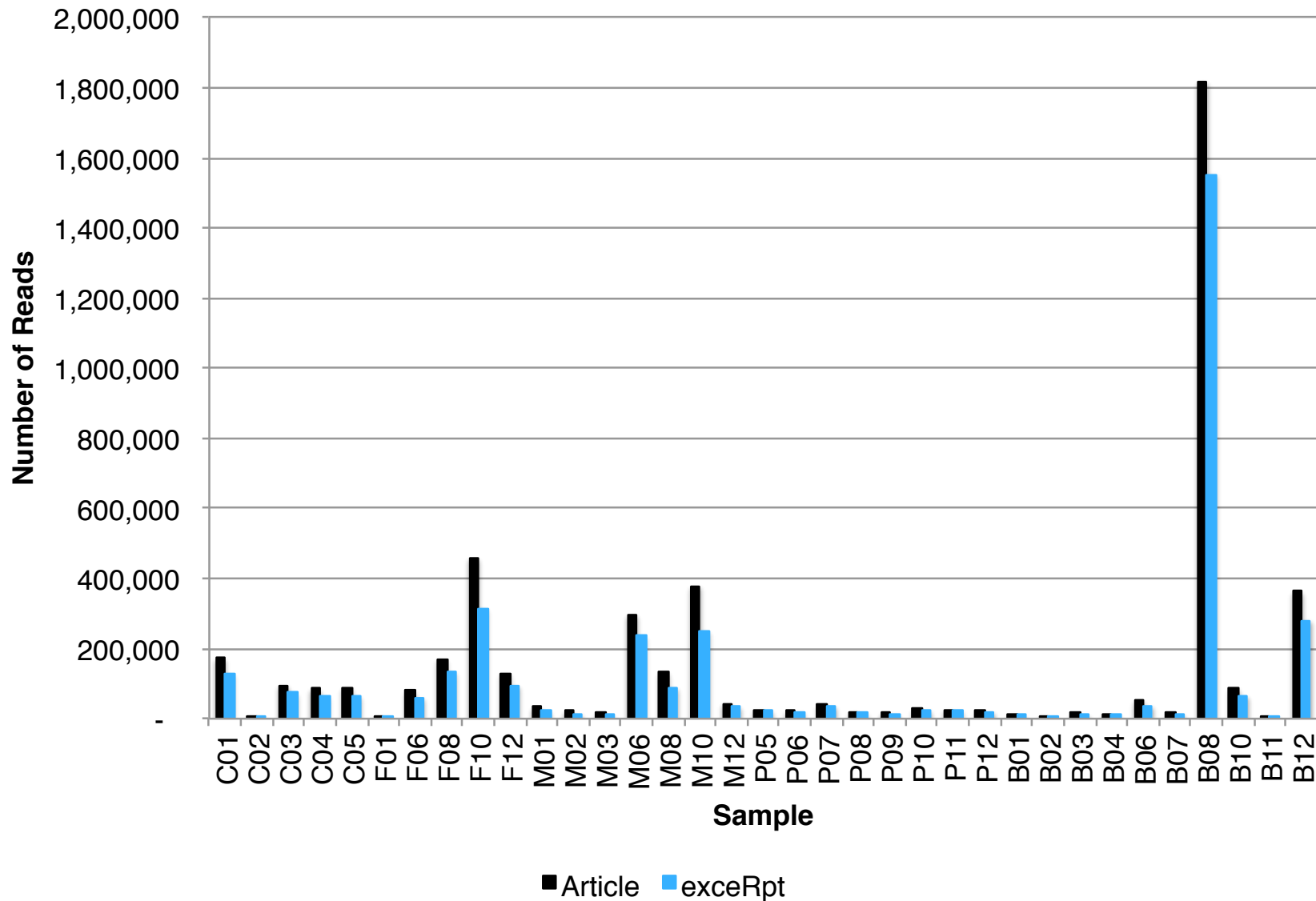
Use Case 3: exceRpt Pipeline Results - Transfer RNA
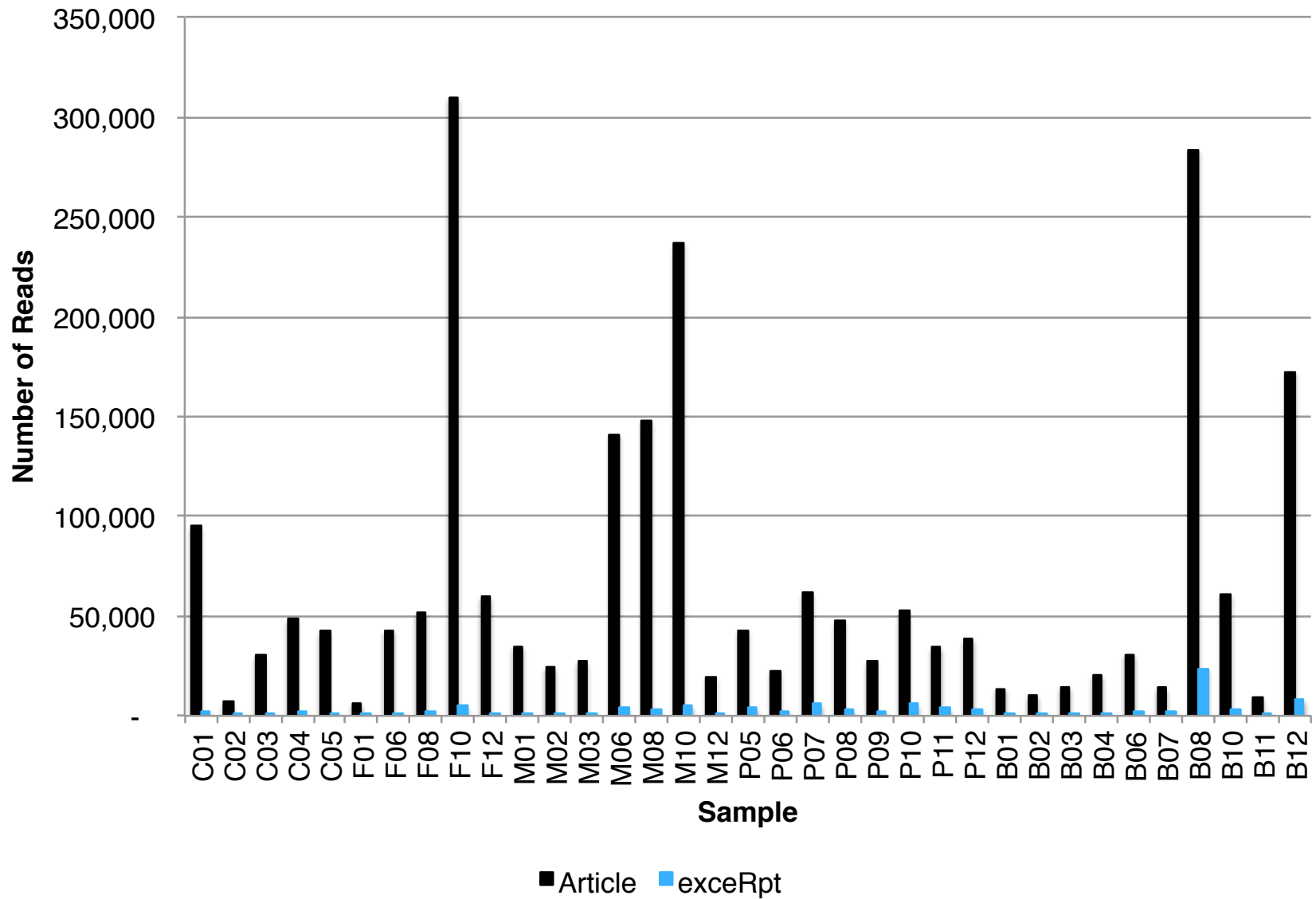
Use Case 3: Summary

The exceRpt analysis pipeline recapitulates the analysis from Williams et al. except for significant differences in the number of reads mapped to transfer RNA.

Use Case 3: Summary

We used exceRpt version 2.8.8; the latest version is 3.1.9.
At the Genboree Workbench there is a version history.

v2.2.8

- We moved the alignment against endogenous repetitive elements (RE) to occur after the main smallRNA alignments performed by sRNABench. This is because we noticed that the RE library was able to 'compete' for reads that would be better annotated/interpreted as coming from tRNAs, piRNAs, or other transcripts. This competitive alignment did not ever affect miRNAs as these are always aligned to before other annotated RNAs, but we expect that this update will faithfully capture reads aligning to repetitive small-RNAs, especially tRNAs, piRNAs, and snoRNAs.

- exceRpt still aligns to REs as a final step before aligning to exogenous sequences as this is critical to remove highly repetitive endogenous sequences that might otherwise be confused as exogenous sequences.

16

http://genboree.org/theCommons/projects/exrna-tools-may2014/wiki/Version_Updates

Acknowledgements

Use Case 3: References

1. Williams Z., et al. **(2013)** Comprehensive profiling of circulating microRNA via small RNA sequencing of cDNA libraries reveals biomarker potential and limitations. **PNAS** 110: 4255–4260.
2. Farazi T.A., et al. **(2012)** Bioinformatic analysis of barcoded cDNA libraries for small RNA profiling by next-generation sequencing. **Methods** 58: 171-187.
3. **ExRNA Portal Software Resources (http://exrna.org/resources/software)**

Useful Links

- **exRNA Portal Software Resources – http://exrna.org/resources/software**

- **exRNA Atlas – http://genboree.org/exRNA-atlas/**

- **Genboree Workbench – http://genboree.org/java-bin/workbench.jsp**

- **Data Coordination Center Wiki –**
  http://genboree.org/theCommons/projects/exrna-mads/wiki

- **exRNA Data Analysis Tools Wiki –**
  http://genboree.org/theCommons/projects/exrna-tools-may2014/wiki

- **Use Case Tutorials – exRNA Portal Data Resource**
  http://exrna.org/resources/data/