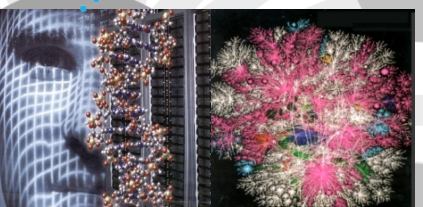
Measurement of exRNA: Quantitative Studies of Small RNA-seq

David J. Galas
Pacific Northwest Research Institute
Seattle, WA

CHI Molecular Medicine TriConference
9 March, San Francisco



Acknowledgments

Alton Etheridge, Galas lab, PNDRI
Kai Wang, ISB

Muneesh Tewari lab, U of Mich.
David Erle lab, UCSF
Louise Laurent lab, UCSD

exRNA RESEARCH PORTAL
NIH ERCC
Portal: www.exrna.org

Funding:
NATIONAL INSTITUTES OF HEALTH
PACIFIC NORTHWEST DIABETES RESEARCH INSTITUTE
uni.lu UNIVERSITÉ DU LUXEMBOURG
NSF

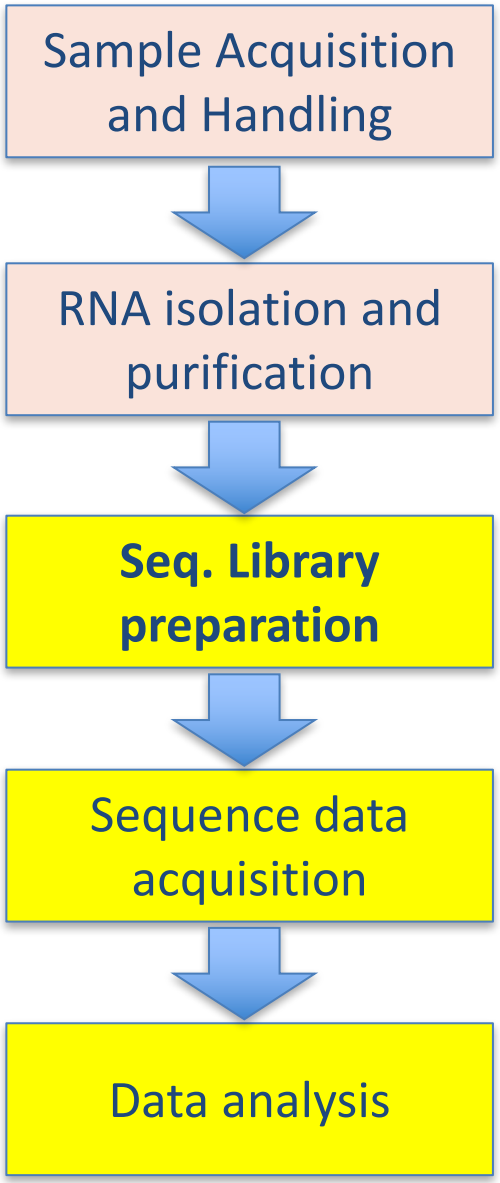
Small RNA-Seq

- Currently the most powerful method for characterizing sRNA populations,
- It plays an important role in many discovery programs, including the NIH Common Fund ERCC consortium, however…
- There are significant problems, prominent among them is sequence-specific bias
- Our Goals:
  - To understand the technology in detail
  - To modify protocols and increase its reliability and power

The RNA-seq Bias problem

- Recent studies have identified RNA-seq library construction biases as a significant problem.
- Up to several hundred-fold differences in miRNA read levels between different protocols have been reported.
- Critical objectives are to:
  - Characterize the biases
    - Quantitate – exactly how bad is it?
    - Characterize noise and reproducability)
  - Understand the source of biases, and if possible
  - Try to find a correction method
- Our focus here is on small RNAs – specifically exRNA

Sources of bias in RNA-seq Experiments



Some recent references: Bias in RNA-seq Libraries

1) Hafner M, Renwick N, Brown M, Mihailovic A, Holoch D, Lin C, Pena JT, Nusbaum JD, Morozov P, Ludwig J et al. **2011**. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA* **17**: 1697-1712.
2) Huang X, Yuan T, Tschannen M, Sun Z, Jacob H, Du M, Liang M, Dittmar RL, Liu Y, Liang M et al. **2013**. Characterization of human plasma-derived exosomal RNAs by deep sequencing. *BMC genomics* **14**: 319.
3) Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R. **2011**. Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic acids research* **39**: e141.
4) Leshkowitz D, Horn-Saban S, Parmet Y, Feldmesser E. **2013**. Differences in microRNA detection levels are technology and sequence dependent. *RNA* **19**: 527-538.
5) Raabe CA, Tang TH, Brosius J, Rozhdestvensky TS. **2014**. Biases in small RNA deep sequencing data. *Nucleic acids research* **42**: 1414-1426.
6) Sorefan K, Pais H, Hall AE, Kozomara A, Griffiths-Jones S, Moulton V, Dalmay T. **2012**. Reducing ligation bias of small RNAs in libraries for next generation sequencing. *Silence* **3**: 4.
7) Zhuang F., Lee JE, Riemondy K, Anderson EM, Yi R. **2013**. High-efficiency RNA cloning enables accurate quantification of miRNA expression by deep sequencing. *Genome biology* **14**: R109.
8) Zhuang F, Fuchs R.T., Sun, Z., Zheng, Y., Robb, G.B. **2012**. Structural bias in T4 RNA ligase-mediated 3'-adapter ligation. *Nucleic acids research* **40**: e54.
9) Baran-Gale, J, Erdos, MR, Sison, C, Young, A, Fannin, EE, Chines, PS and Sethupathy, P, Massively differential bias between two widely used Illumina library preparation methods for small RNA sequencing, bioRxiv. DOI:10.1101/001479
10) Fuchs, R.T., Sun, Z. Zhuang, F., and Robb, G.B., **2015.** Bias in Ligation-based Small RNA Sequencing Library Construction is Determined by Adaptor and RNA Structure, *PLOS ONE*, DOI:10.1371/journal.pone0126049

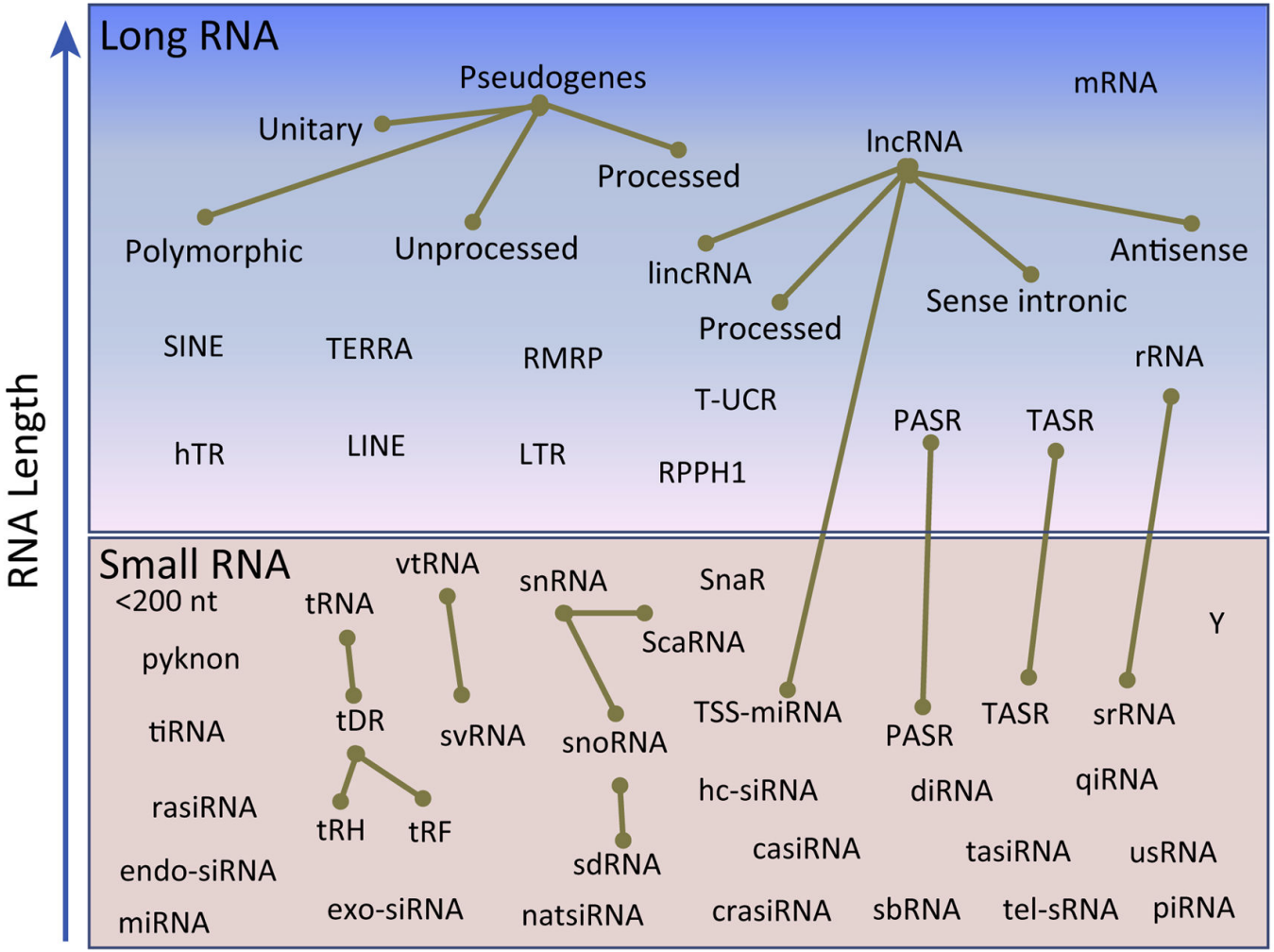
Circulating RNA

- There is wide range of extracellular RNA molecules in the blood *(e.g Wang et al., Plos One, 2010)* some for cell-cell communication
- exRNA is found in at least 12 different body fluids (*Weber et al., 2011*)
- exRNAs detected include: mRNA, rRNA, tRNA, lncRNA, Y-RNA, snoRNA, pi-RNA and miRNA etc…
- They can be taken up by other cells and change gene expression.
- Recent discovery: non-human RNA (microbial & others) is found in the blood at significant levels (*Wang et al, 2012*)

Some reference to our work:

- Wang, K., Zhang, S., Weber, J., Baxter, D., and Galas, D.J. "Mammalian cells in culture actively export specific microRNAs," Nucleic Acids Research, 38(20): 7248-7259 (2010).
- Weber, J.A., Baxter, D.H., Zhang, S., Huang, D.Y., Huang, K.H., Lee, M., Galas, D.J., and Wang, K., "The MicroRNA Spectrum in 12 Body Fluids", Clinical Chemistry, 56: 1733-1741 (2010).
- Wang, K., Li, H., Yuan, Y., Etheridge, A., Huang, D., Wilmes, P., and Galas, D.J., "The Complex Exogenous RNA Spectra in Human Plasma: an Interface with Human Gut Microbiome?", PLOS ONE, 7(12):e51009 (2012).

Some RNA in the Zoo



Vickers et al. (2015) TiBS

Notable RNAs

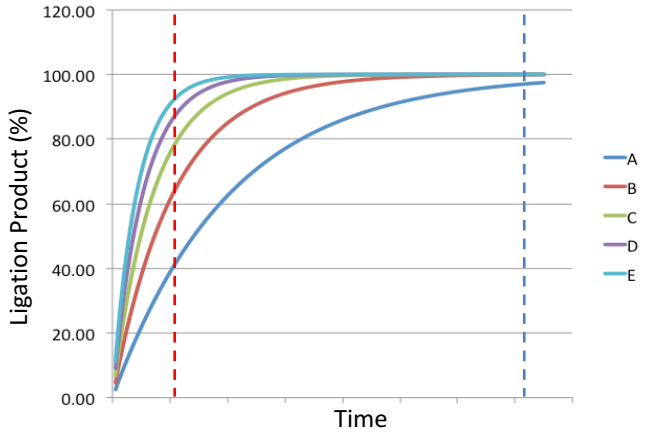
- miRNA
- tRNA derived
- snoRNA
- piRNA
- vtRNA

Our Approach

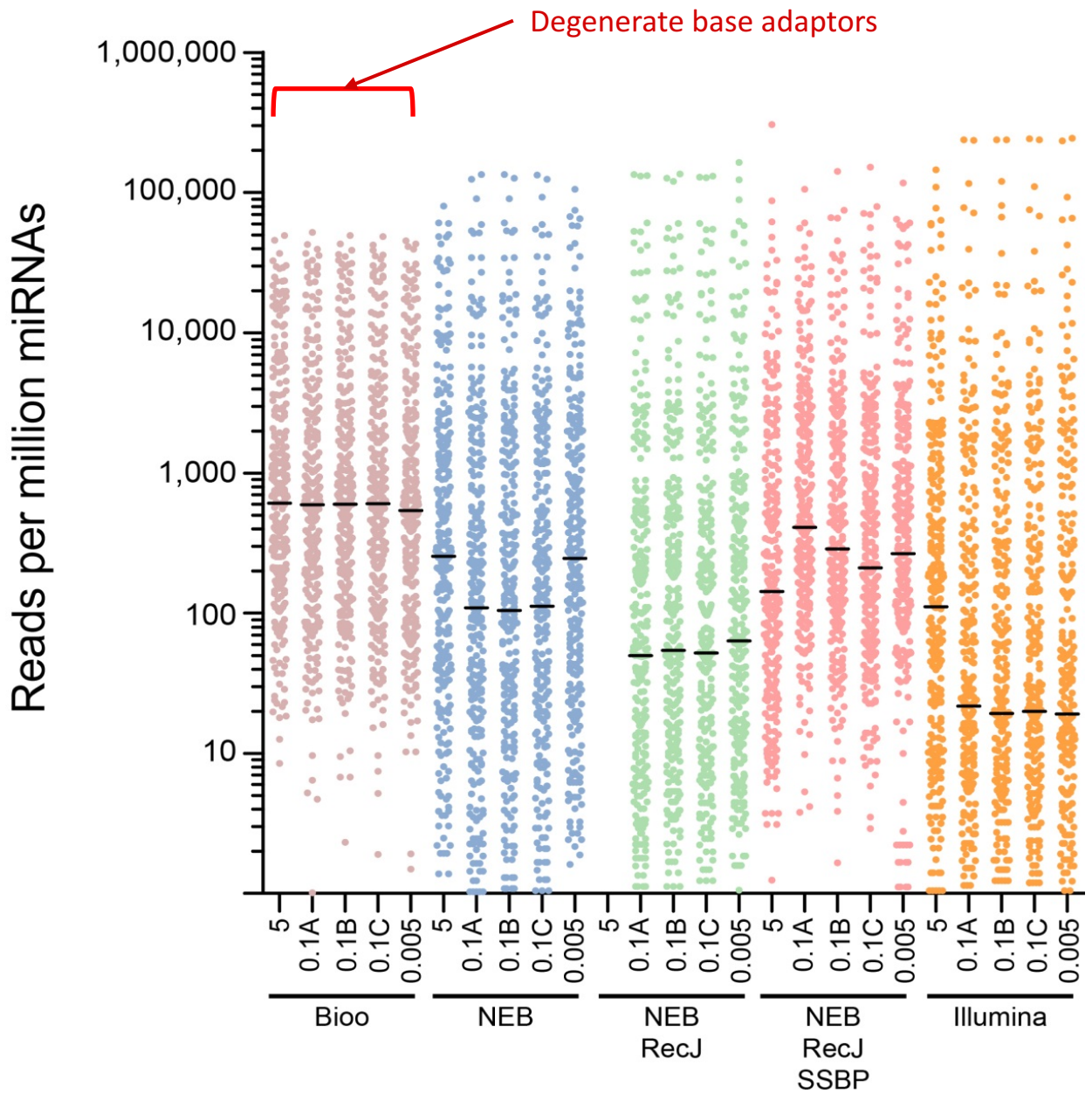
- Use synthetic miRNAs to form specific, defined populations of RNA to characterize bias quantitatively
- Examine sequence effects, reaction differences etc..
- Differing ligation reaction rates are the primary culprits (first shown by Tom Tuschl's lab)

Note that even if the reaction rates for ligation are different, pushing the reactions to completion eliminates bias

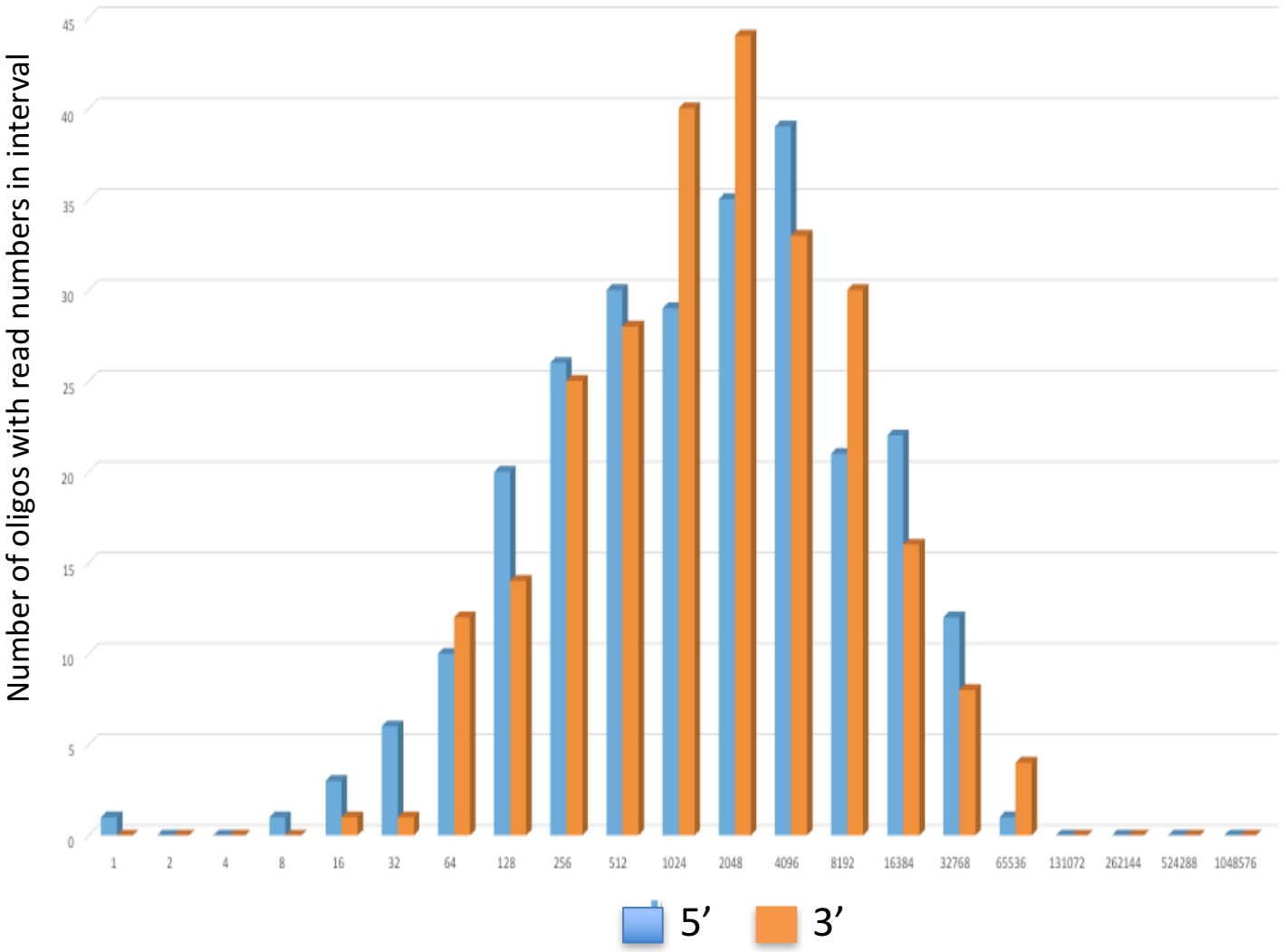
Illustration of Ligation yields for 5 oligoRNAs w/ differing rates



RNA-Seq Results for equimolar mix of 286 synthetic miRNAs

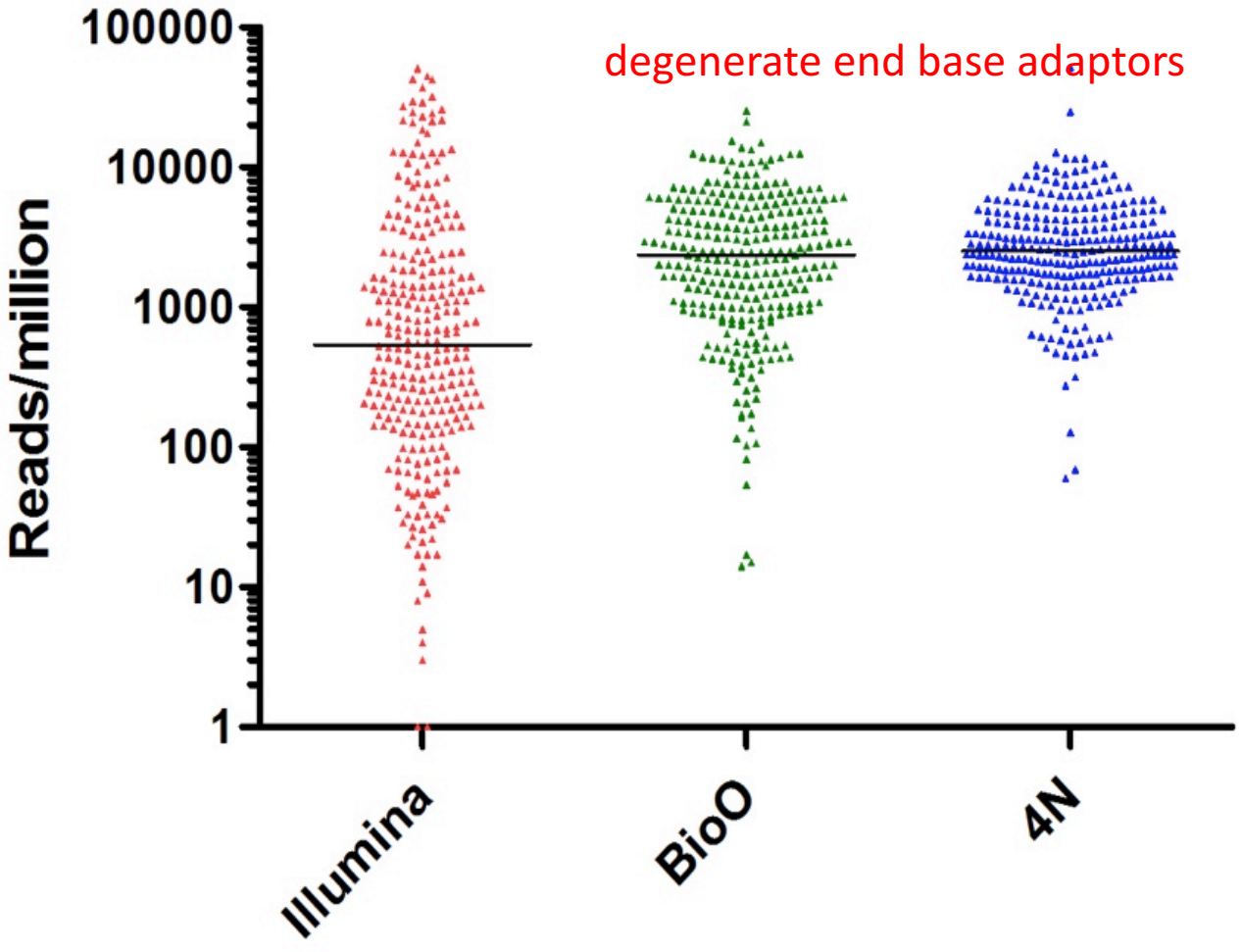


Distribution of read numbers for each end (TS)
(*approximately log normal*)



RNA-Seq on an Equimolar mix of 286 Synthetic miRs

Single libraries

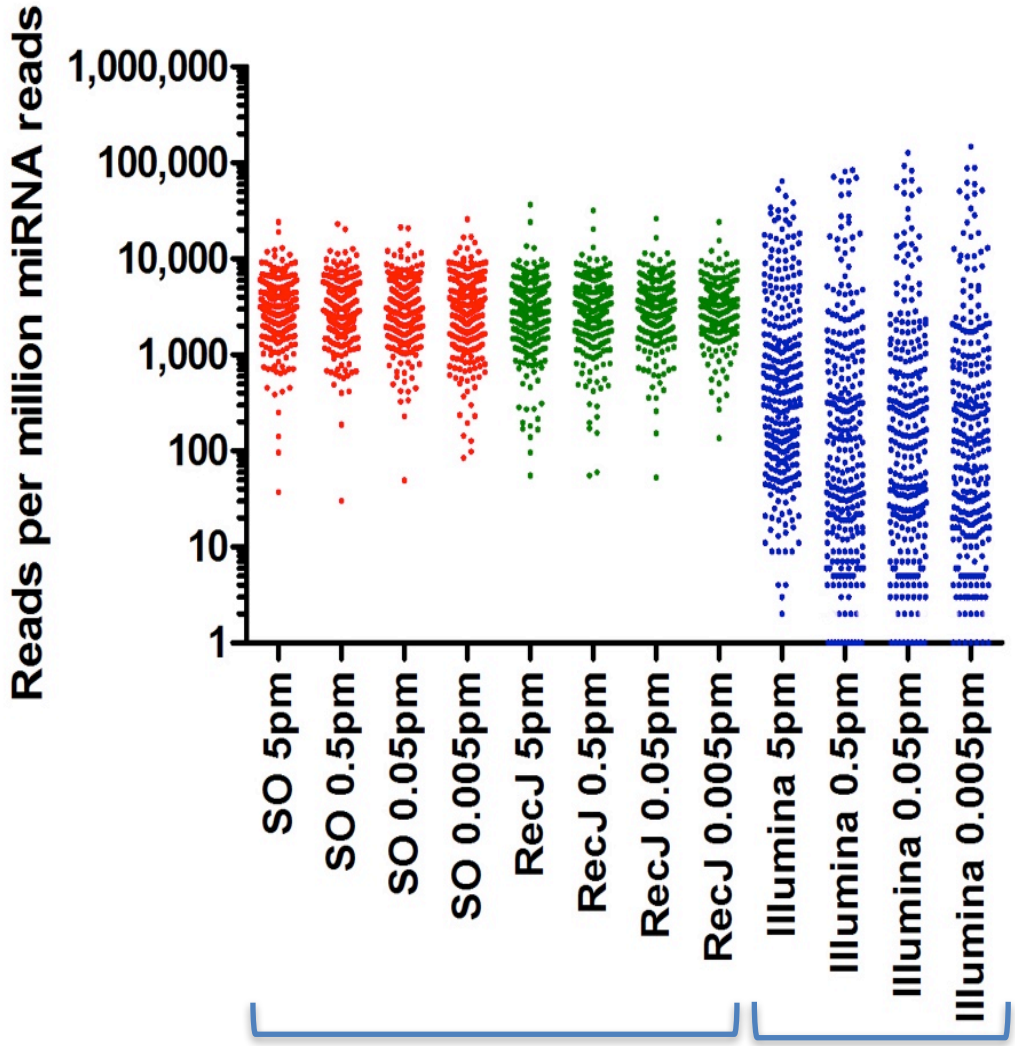


First suggestion of degenerate base ends was from Robb's group (Zhuang et al., NAR, 2012)

Two Protocols and different RNA input levels (36 libraries)

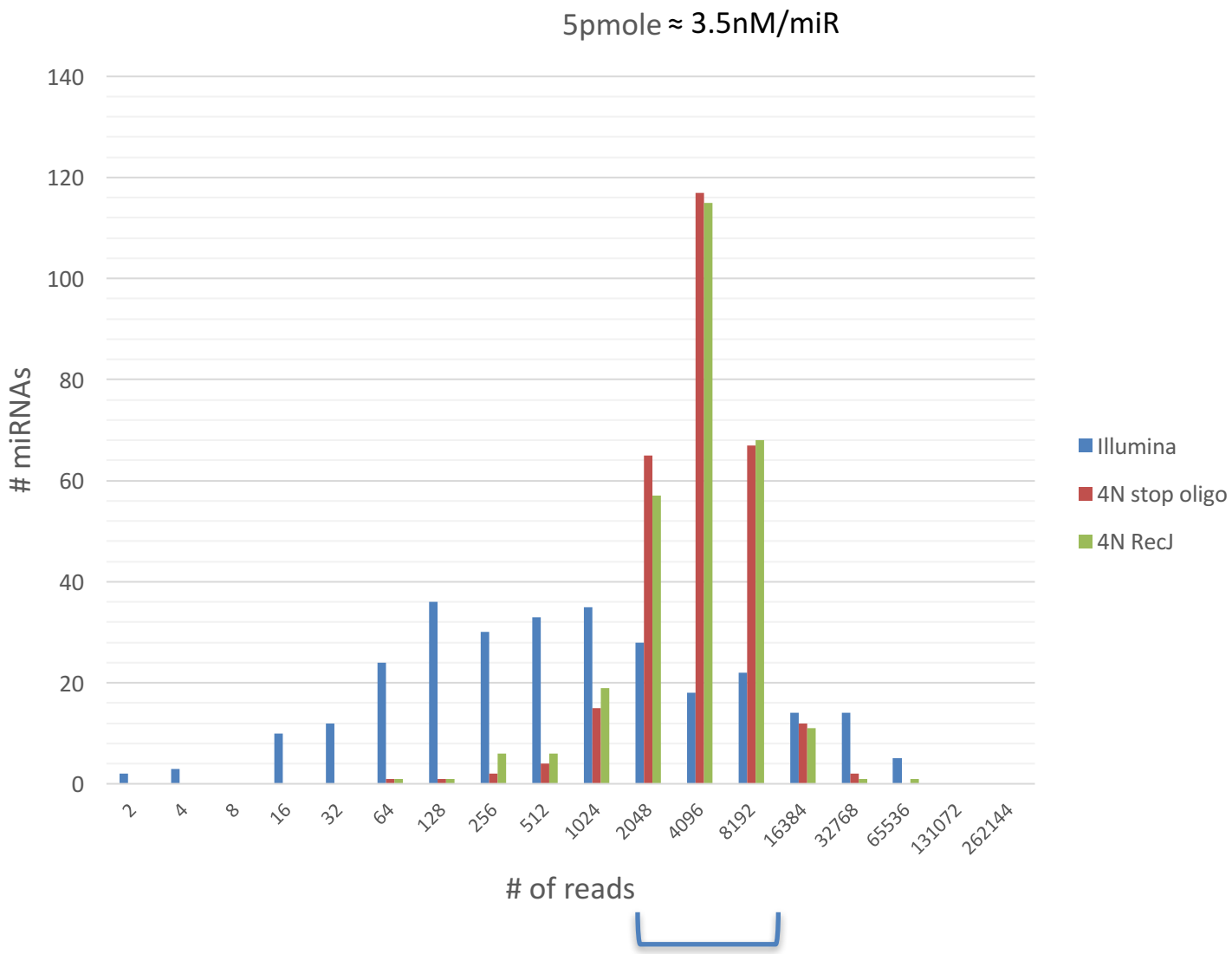
Read distributions for equimolar mix of 286 synthetic miRNAs

Three replicates of three protocols, four RNA input levels



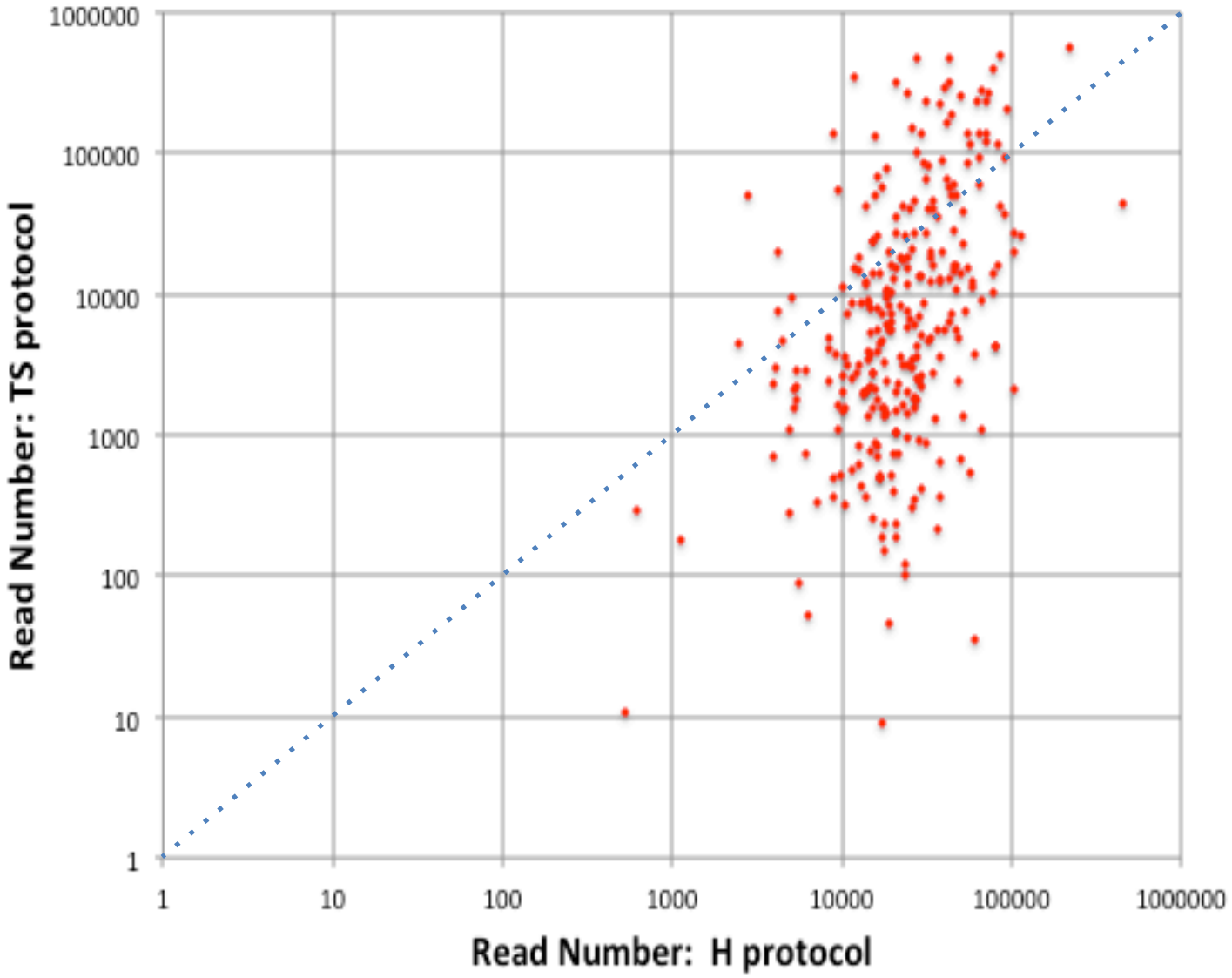
Read distributions for 286 synthetic miRNAs (approx. log normal)

Comparing protocols



2-fold around geometric mean,
but with long tails

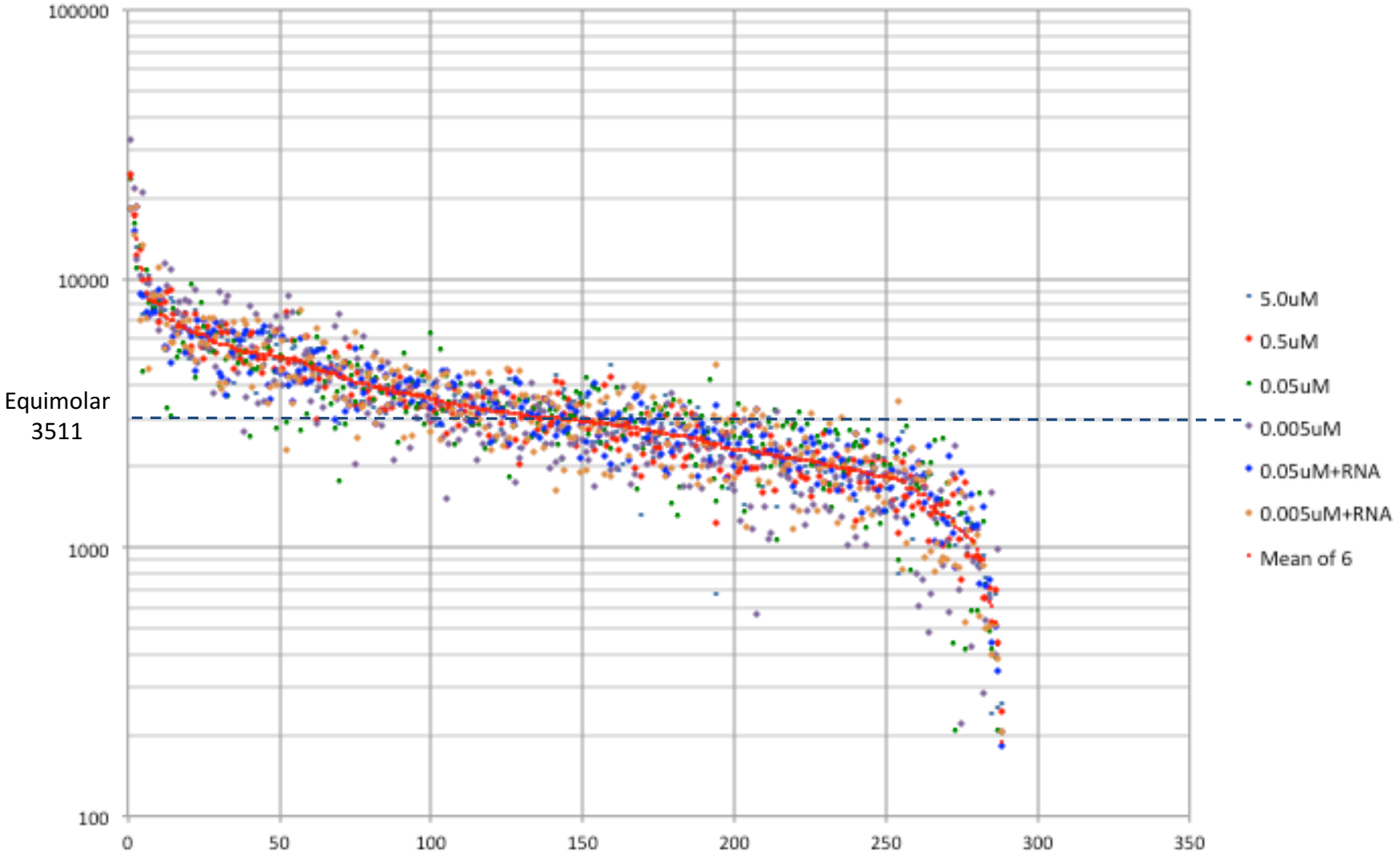
Equimolar 286 Synthetic miRs



Sequence Specificity & Variation

- There is indeed significant sequence specificity
- There is also significant variation for all miRs
- The variation, however, is not the same for different miRs
- The sequence specificity of read level and sequence specificity of variation appears unrelated

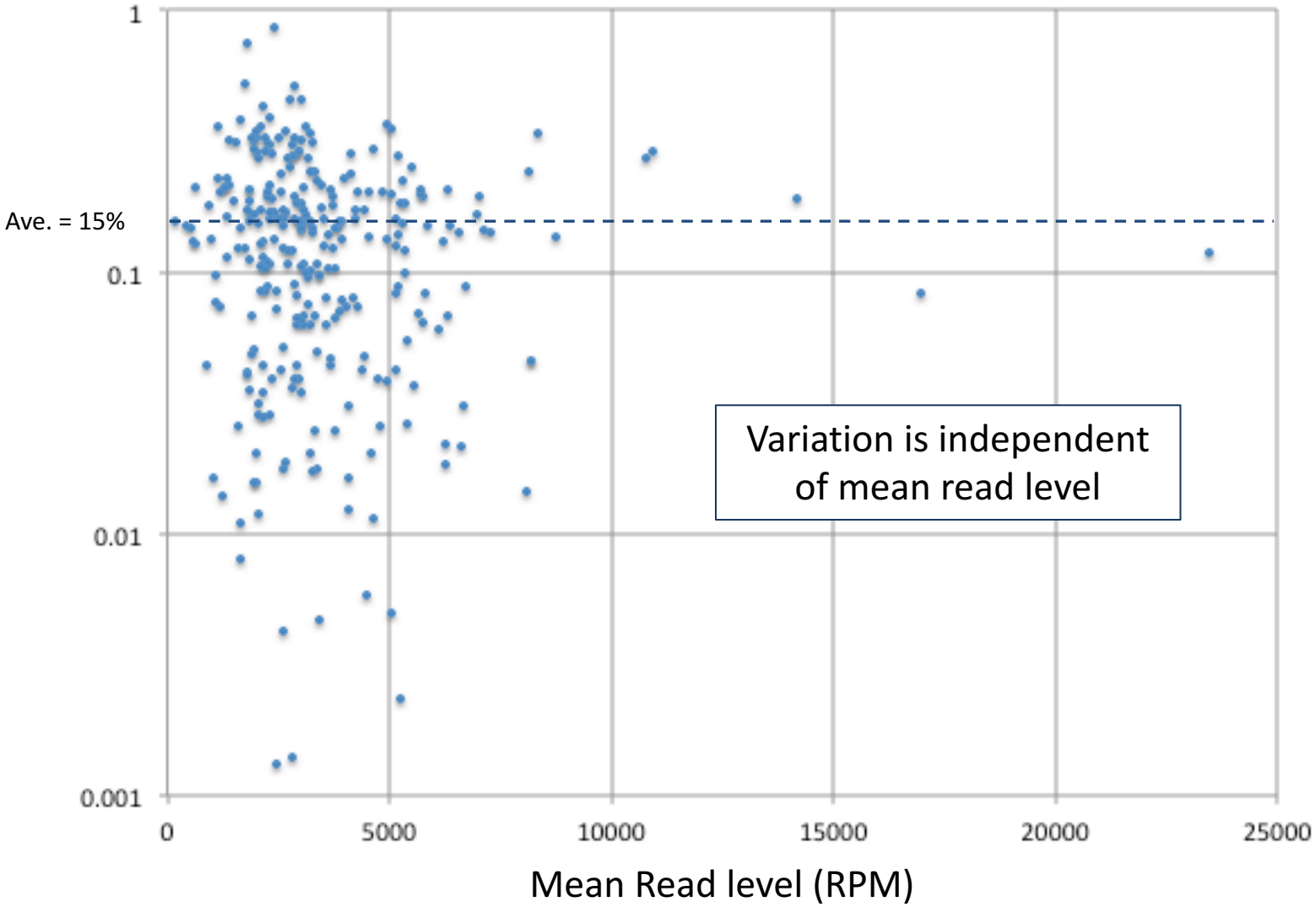
Replicates of 286 synthetic miRs at different input levels (RPM)



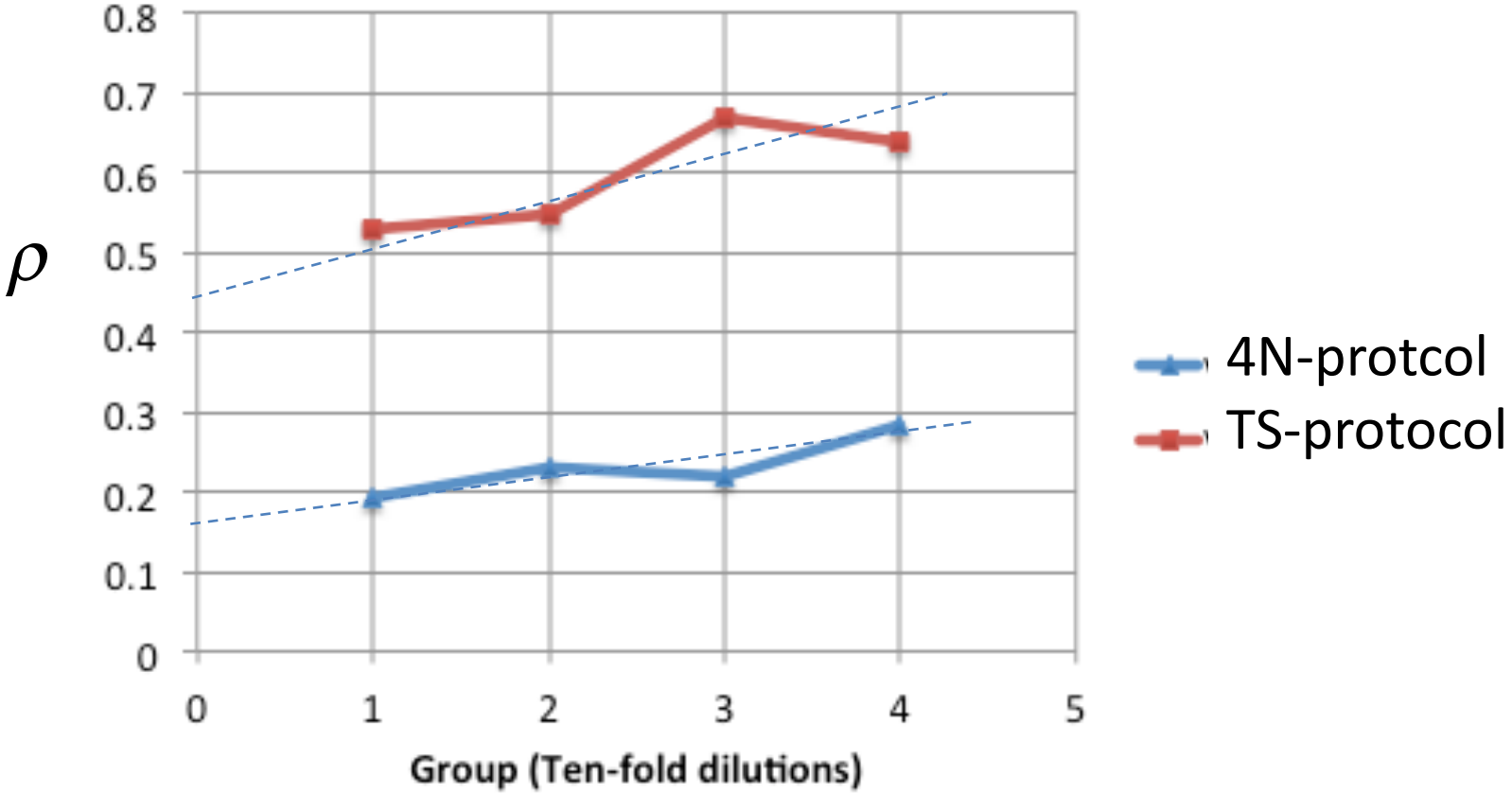
Protocol 4N-H

Stnd. Dev. / Mean as a Function of the Mean (4N)

286 Synthetic miRs, RPM (6 replicates, dilutions)



Reproduceability Error

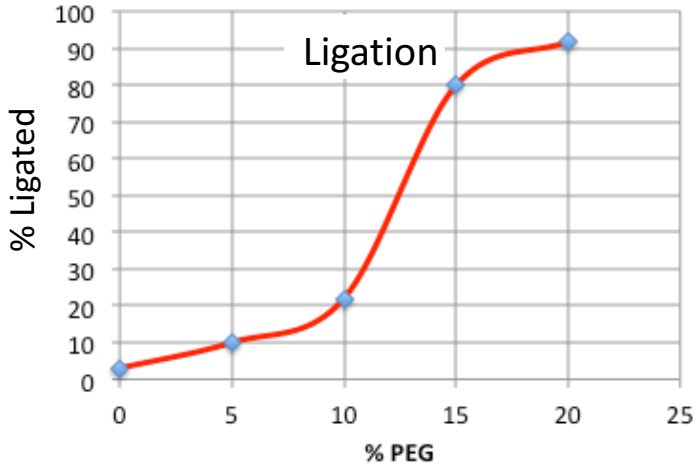


Measure, ρ, is average normalized difference $\rho = \left\langle \frac{2\sqrt{(r_E - r_D)^2}}{r_E + r_D} \right\rangle$

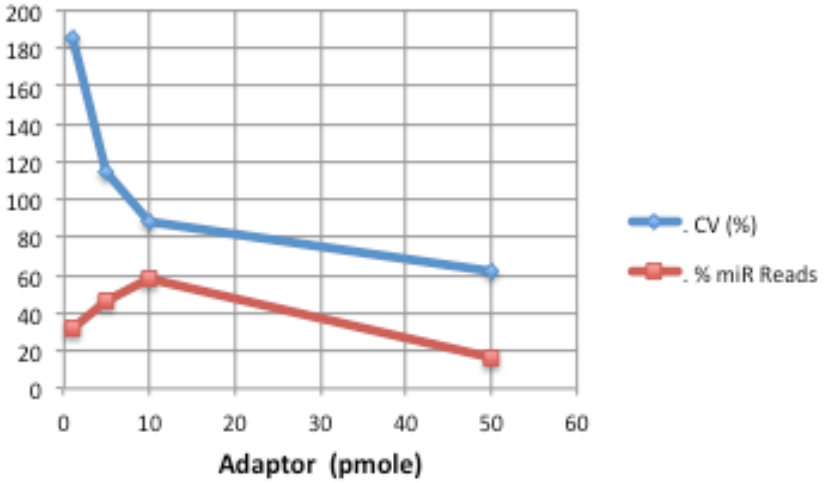
Effect of PEG and Adaptor Conc. on Ligations & Libraries

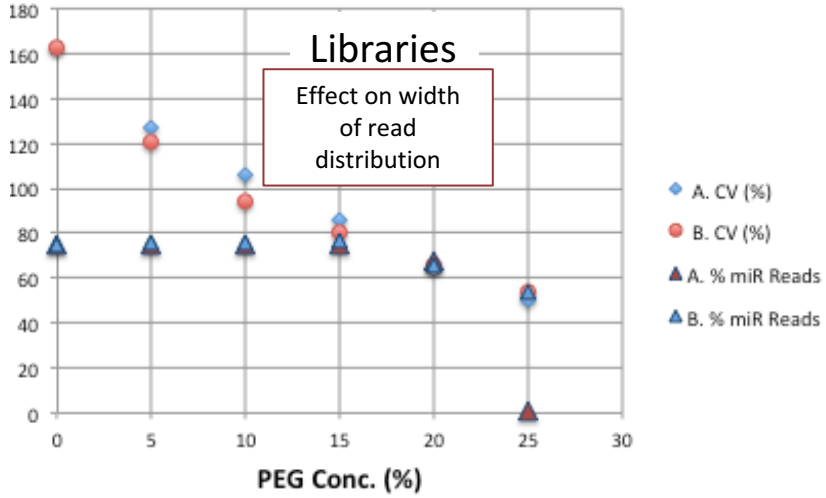
driving the reaction forward

PEG



Adaptor





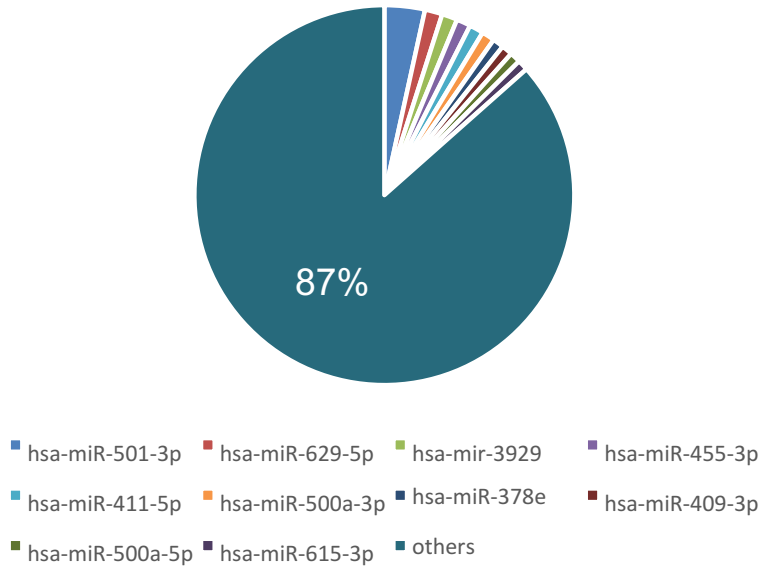
Optimization

Adjusting these two parameters:

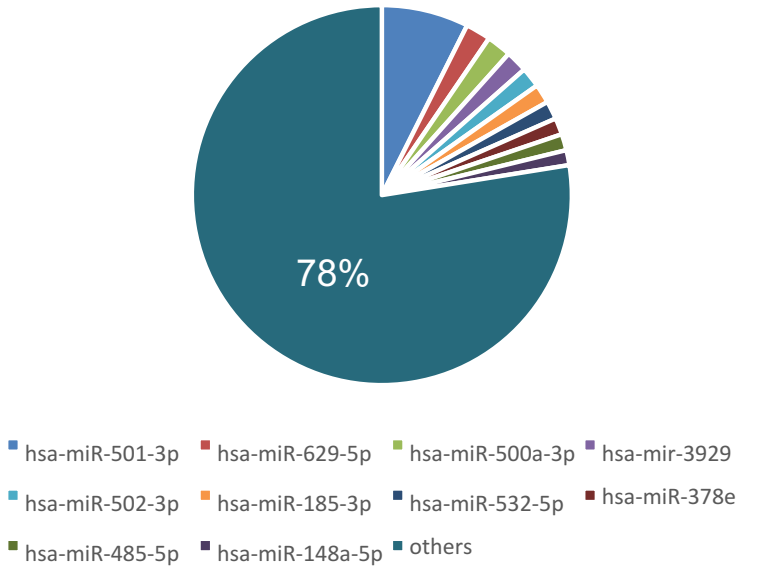
- To increase miR reads relative to adaptor dimers
- To decrease the bias (drive reaction)
- Protocol optimized for plasma will be available on the NIH consortium Portal: www.exrna.org

Synthetic libraries, 4N protocol

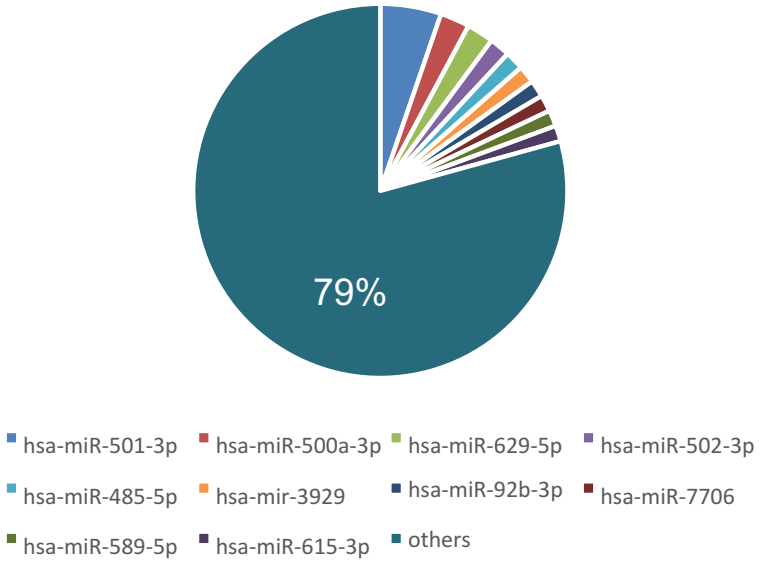
Hi PEG/Hi adapter



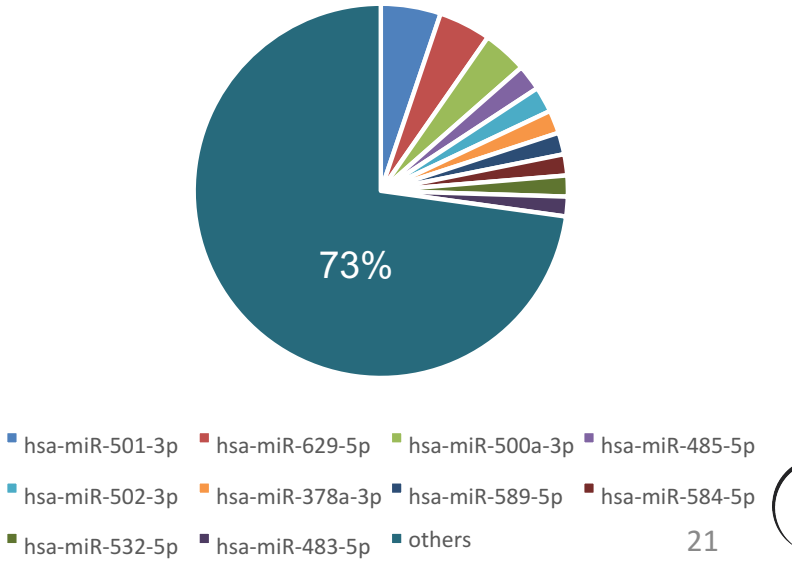
Hi PEG/Lo adapter



Lo PEG/Hi adapter

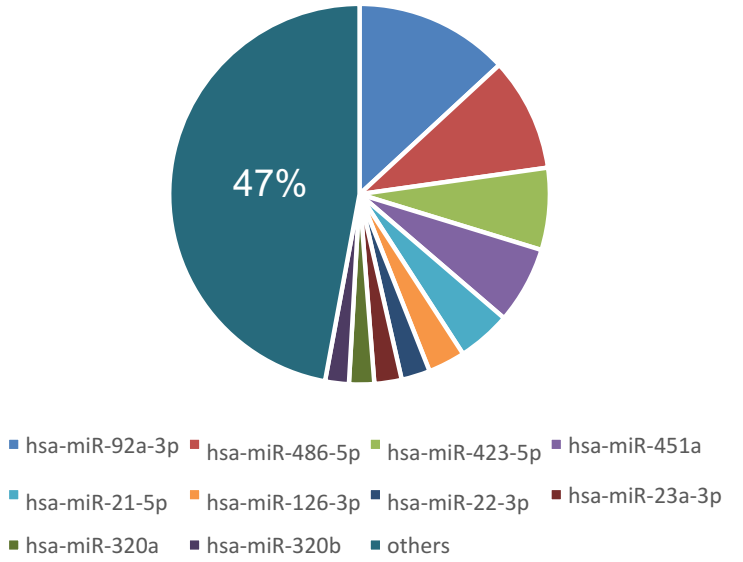


Lo PEG/Lo adapter

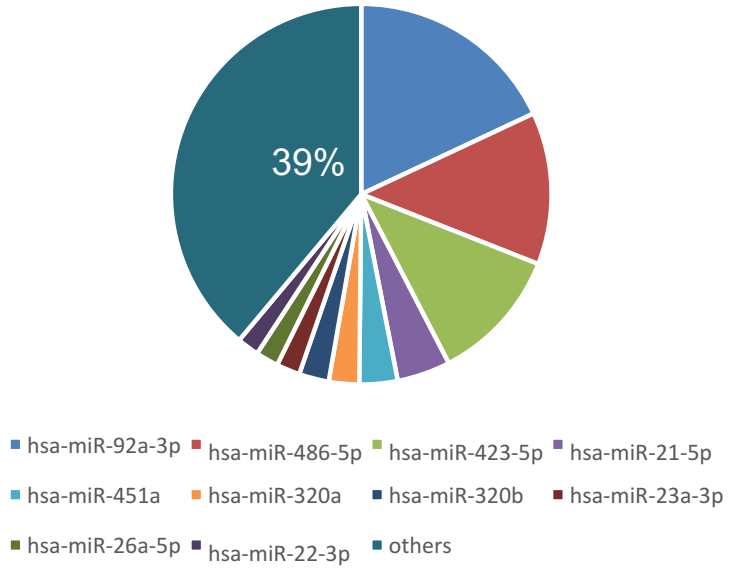


Plasma libraries, 4N protocol

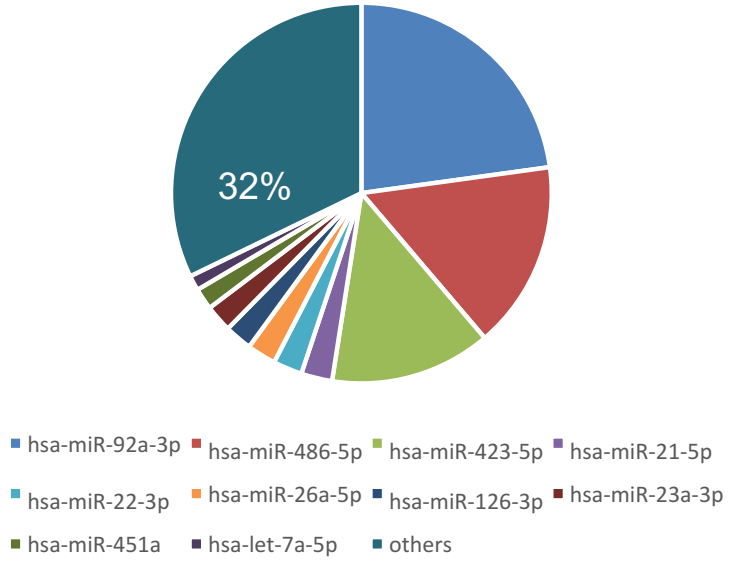
Hi PEG/Hi adapter



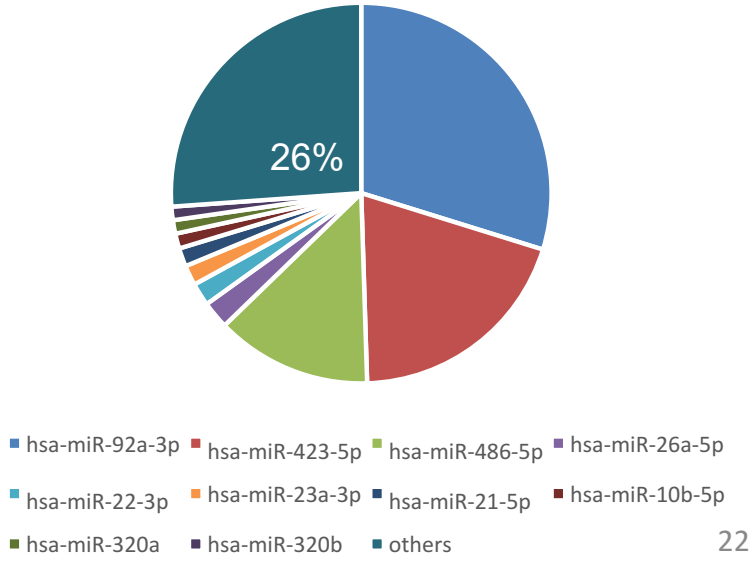
Hi PEG/Lo adapter



Lo PEG/Hi adapter



Lo PEG/Lo adapter



miRxplore (962 synthetic miRNAs library correlations

	BioO A	BioO B	Erle A	Erle B	Galas A	Galas B
Bioo A	1	0.99	0.38	0.38	0.37	0.38
Bioo B		1	0.40	0.40	0.39	0.40
Erle A			1	1.00	0.45	0.47
Erle B				1	0.45	0.47
Galas A					1	0.99
Galas B						1

Summary

- All libraries made with 4N adapters
- Correlation between replicates is good, correlation between protocols is not.

Hypothesis: sequence-specific bias

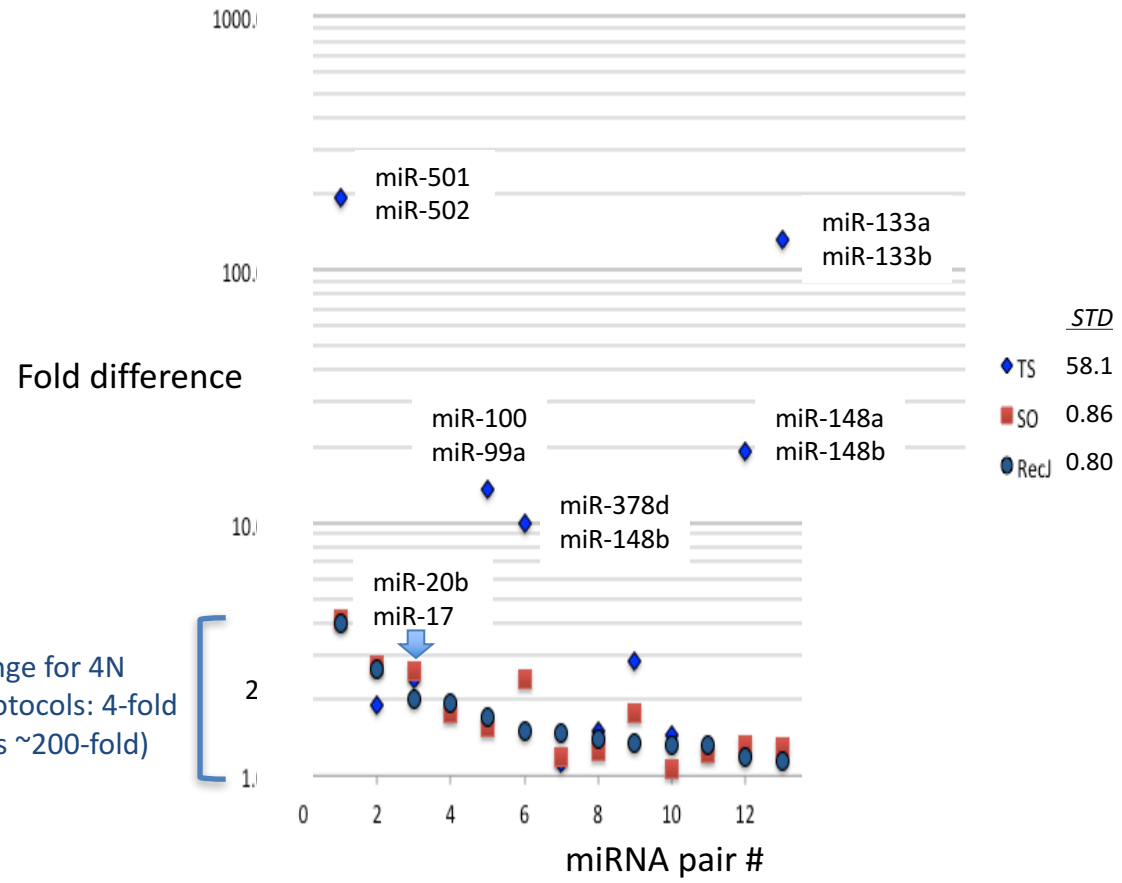
- The ligation step seems to be the critical one for small RNA bias (based on several studies)
- Thus we might guess that the end sequences of the RNA are the most important.
- Characterizing the bias for specific end sequences might be able to provide a correction factor for each protocol

What causes the sequence-specific bias?

Single base effect in miR's

Fold differences for 13 pairs of miRNAs among the 286 with single base differences

*Each point is an average from **12 libraries** (3 replicates at 4 RNA input levels)*



miRNA Name	Sequence
hsa-miR-100-5p	AACCCGTAGATCCGAACTTGTG
hsa-miR-99a-5p	AACCCGTAGATCCGATCTTGTG
hsa-miR-501-3p	AATGCACCCGGGCAAGGATTCT
hsa-miR-502-3p	AATGCACCTGGGCAAGGATTCA
hsa-miR-378b	ACTGGACTTGGAGGCAGAA
hsa-miR-378d	ACTGGACTTGGAGTCAGAAA
hsa-miR-107	AGCAGCATTGTACAGGGCTATCA
hsa-miR-103a-3p	AGCAGCATTGTACAGGGCTATGA
hsa-miR-23b-3p	ATCACATTGCCAGGGATTACC
hsa-miR-23a-3p	ATCACATTGCCAGGGATTTCC
hsa-miR-20b-5p	CAAAGTGCTCATAGTGCAGGTAG
hsa-miR-17-5p	CAAAGTGCTTACAGTGCAGGTAG
hsa-miR-301a-3p	CAGTGCAATAGTATTGTCAAAGC
hsa-miR-301b	CAGTGCAATGATATTGTCAAAGC
hsa-miR-130b-3p	CAGTGCAATGATGAAAGGGCAT
hsa-miR-130a-3p	CAGTGCAATGTTAAAAGGGCAT
hsa-miR-18a-5p	TAAGGTGCATCTAGTGCAGATAG
hsa-miR-18b-5p	TAAGGTGCATCTAGTGCAGTTAG
hsa-miR-10b-5p	TACCCTGTAGAACCGAATTTGTG
hsa-miR-10a-5p	TACCCTGTAGATCCGAATTTGTG
hsa-miR-148a-3p	TCAGTGCACTACAGAACTTTGT
hsa-miR-148b-3p	TCAGTGCATCACAGAACTTTGT
hsa-miR-19b-3p	TGTGCAAATCCATGCAAAACTGA
hsa-miR-19a-3p	TGTGCAAATCTATGCAAAACTGA
hsa-miR-133b	TTTGGTCCCCTTCAACCAGCTA
hsa-miR-133a-3p	TTTGGTCCCCTTCAACCAGCTG

Experiments with Degenerate base RNA populations

Some Libraries with Degenerate Bases

Library	Sequence	Number of OligoRN's
L1	NNNNAUGGCUGACGUACGUNNNN	$4^8 \cong 65000$
L2	NNNNUUCGUGCGAUCUAGGNNNN	$4^8 \cong 65000$
L5	UUGNAUGNCUGNCGUNCGUNACG	$4^5 = 1024$
L6	UUGAAUGGCNNNNNUACGUGACG	$4^5 = 1024$
L7	NNNNGCUAGCGUUCAGGUCNNNN	$4^8 \cong 65000$
L8	NNNNCAACCAUCGAGCUAANNNN	$4^8 \cong 65000$

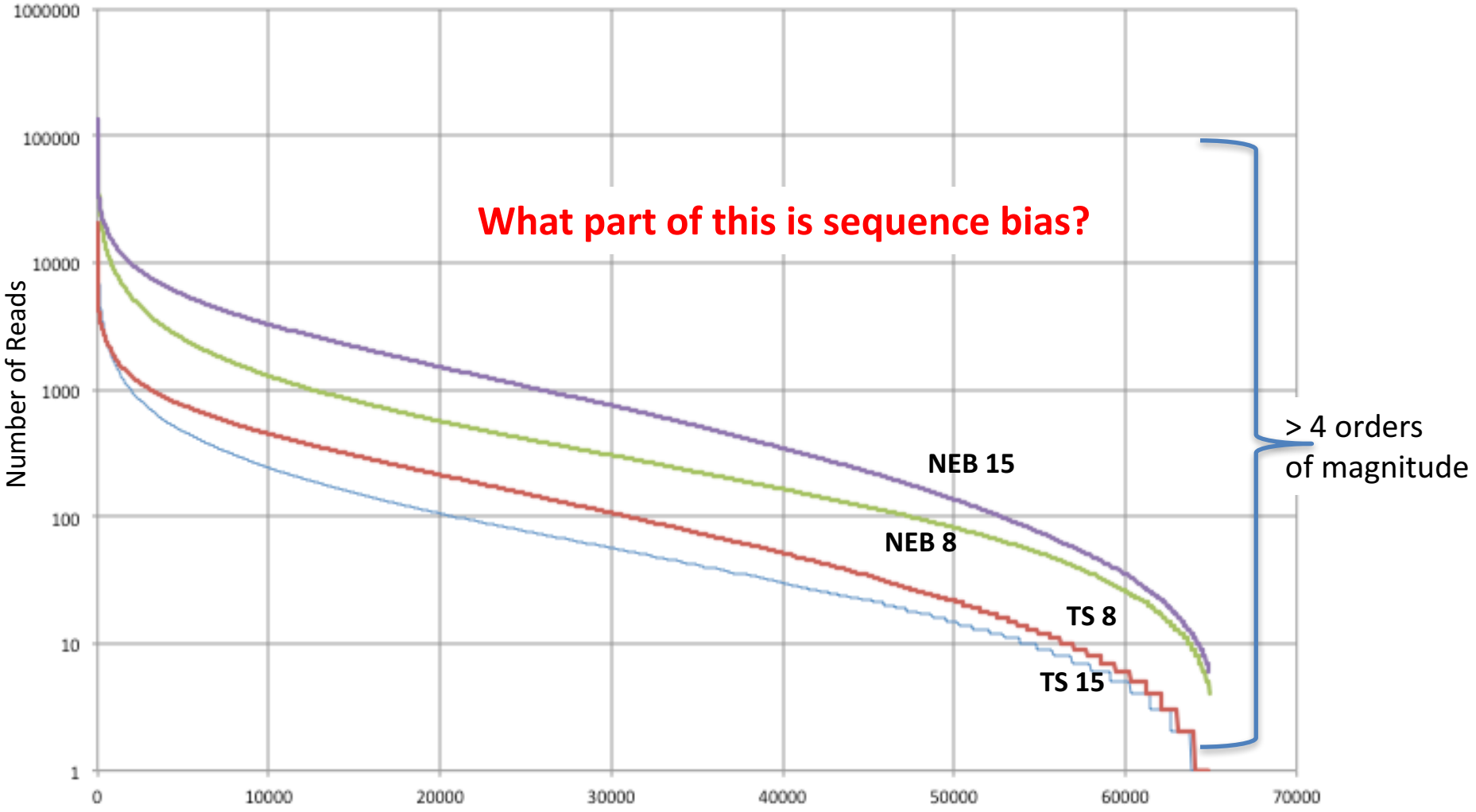
Systematic Approach to Bias

Ribo-oligonucleotide Libraries with degenerate bases

- A range of synthetic RNA populations with fixed and degenerate bases have been made and sequenced.
- With several RNA-seq library protocols
  - Illumina True-seq, NEB, BioO and our 4N protocol
- In the course of this work we discovered:
  - How to assess the synthesis bias and
  - How to separate out the sequence-specific library bias

Wide range of read numbers over the ~65,000 Oligonucleotides

L1 core, Protocols: TS 15, TS 8, NEB 15, NEB 8

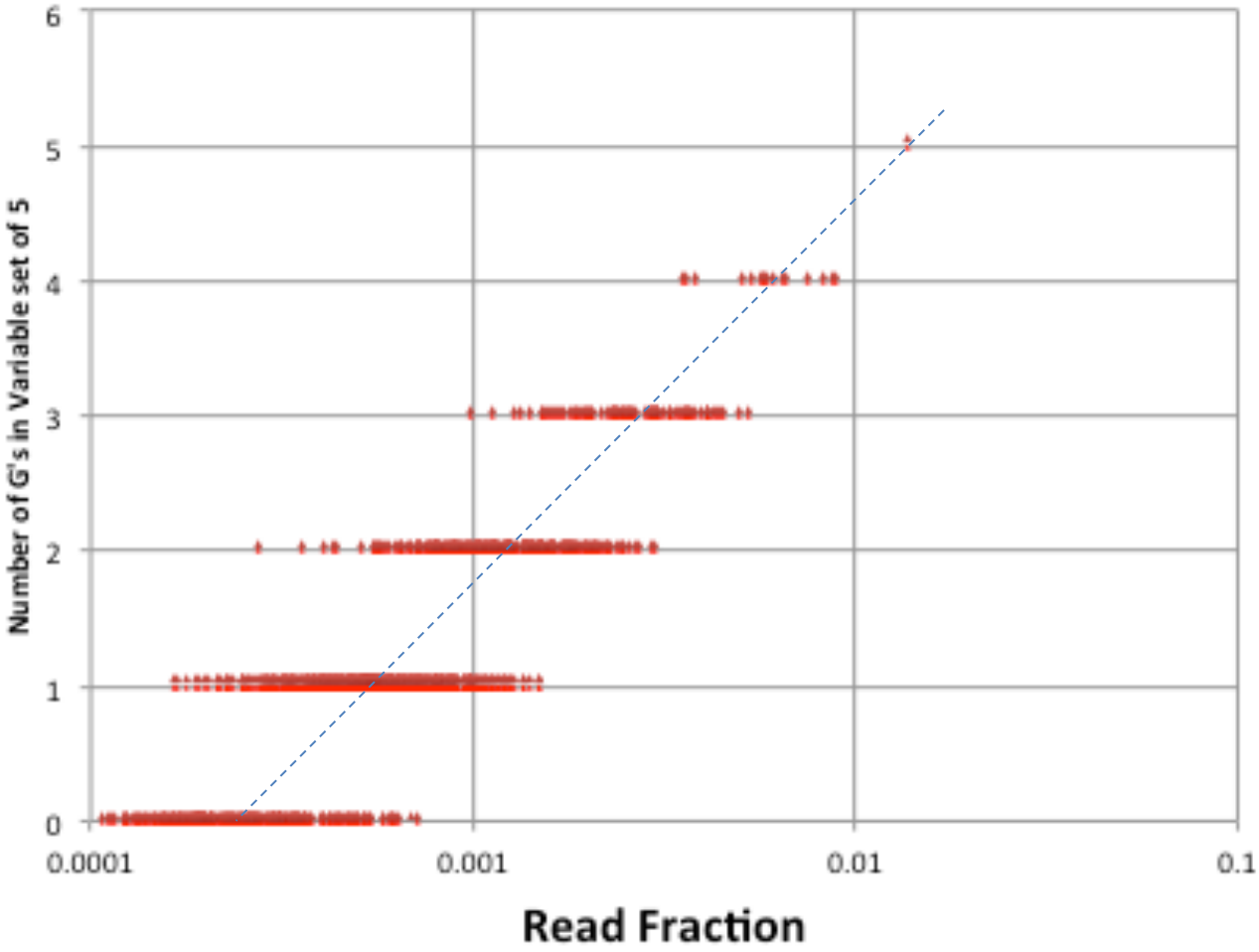


TS = Illumina True seq protocol
NEB = New England Biolab protocol
Number = number of PCR cycles

Effect of Number of Variable G's: H

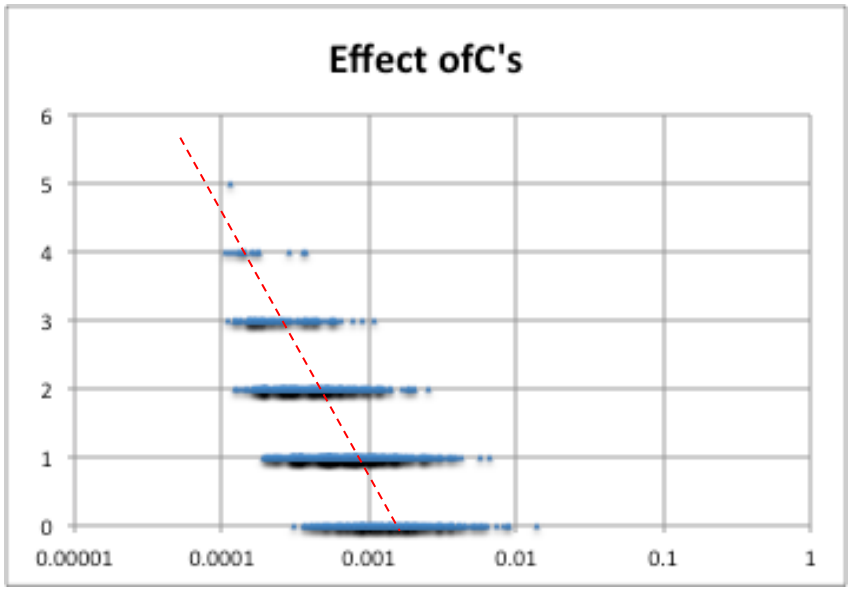
TTGNATGNCTGNCGTNCGTNACG (L5)

H protocol: 4N adaptors



Read fraction median as a function of number of G's, *n*:

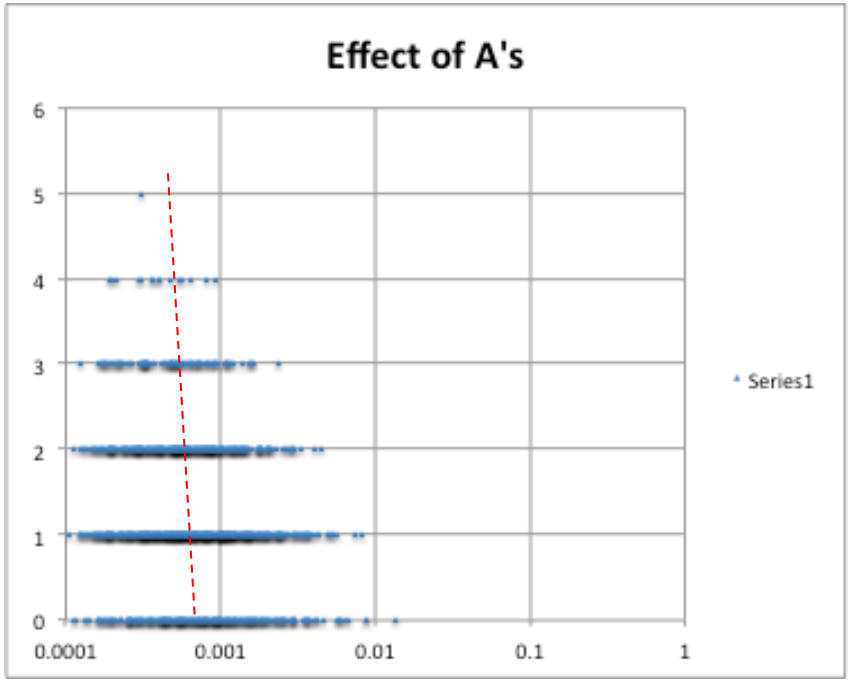
$$f(n) \cong 0.0003\ e^{0.757\,n}$$

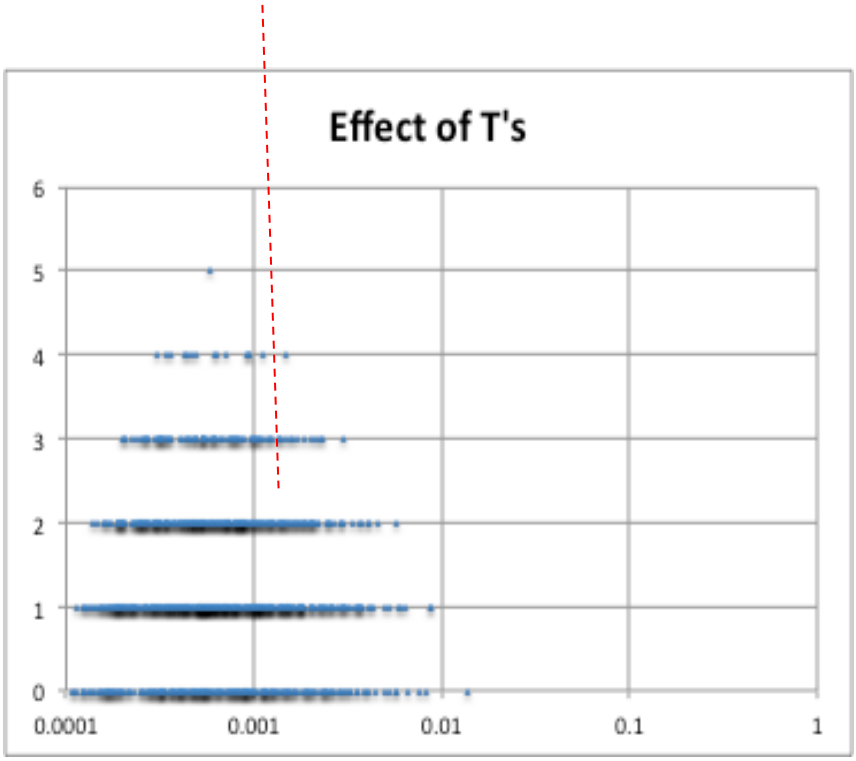


Internal “RANDOM” bases

TTGNATGNCTGNCGTNCGTNACG

C has the opposite effect of G, but
A & T have almost no effect





Synthesis bias

- In principle variations could be due to any combination of Synthesis and other Sequence bias
- Can we distinguish synthesis bias in our experimental libraries? Hypothesis:
    1. Context is not important in synthesis bias, and sequence-specific bias is mostly context,
    2. the fractional composition profile of reads can then be predicted from insertion bias
    3. This can be directly tested as follows
    4. For *n* degenerate bases there are *n*+1 quantities predicted by a single parameter, so a total of 4(*n*+1) measured quantities and 4 parameters (*base content measured, base synthesis probability*)
    5. In addition there are 4 other measured quantities, the base content of the degenerate bases

Inference of synthesis bias results

Hypothesis: Composition of collective degenerate bases doesn't matter. Sequence bias comes from order "only".

Evidence points to:
Synthesis bias as source of major differences in composition of degenerate bases

What is the **base composition profile** of the sequence reads? We can measure this!

Can we explain the profile with a few parameters? YES (strong fit to a simple model)

This means there is **no overall context effect** within the reads. What are the parameters? Equal to the overall base composition of the reads (tight fit)!!

Then there is no selection bias of reads from the synthetic population (sequencing bias): Thus,

$$\textbf{Composition}_{seq} = \textbf{Composition}_{syn}$$

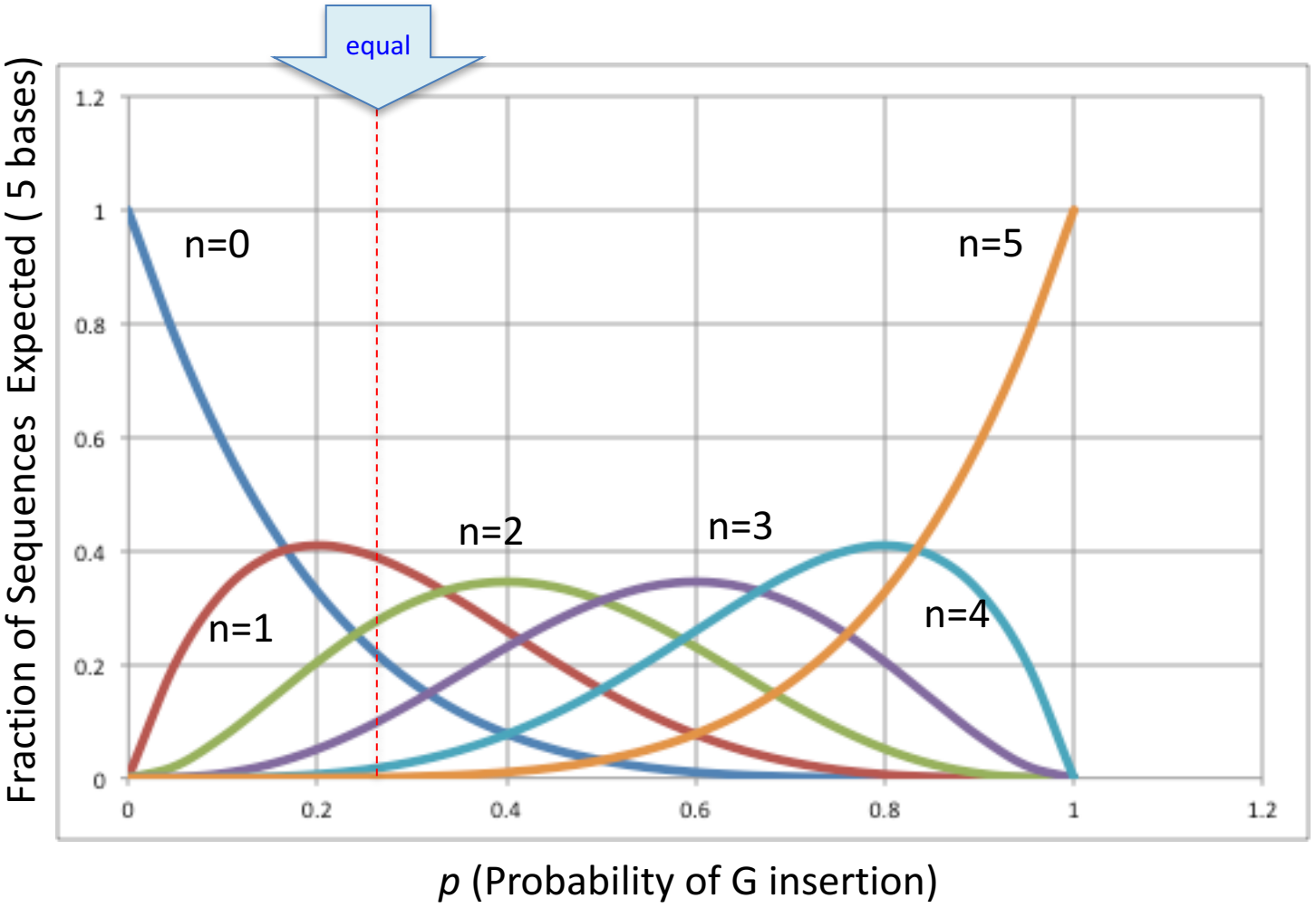
This implies that the base composition is equal to the **synthetic base insertion frequency**

We can then directly infer the composition of the synthetic RNA population from the read population

Sequence specificity is caused by the order of bases among the specific composition groups of oligos

Probability of G composition in Synthesis

5 degenerate bases ➡ 6 quantities



$$\Pr obability(n/5) = \binom{5}{n} p^n (1-p)^{5-n}$$

Measured quantities

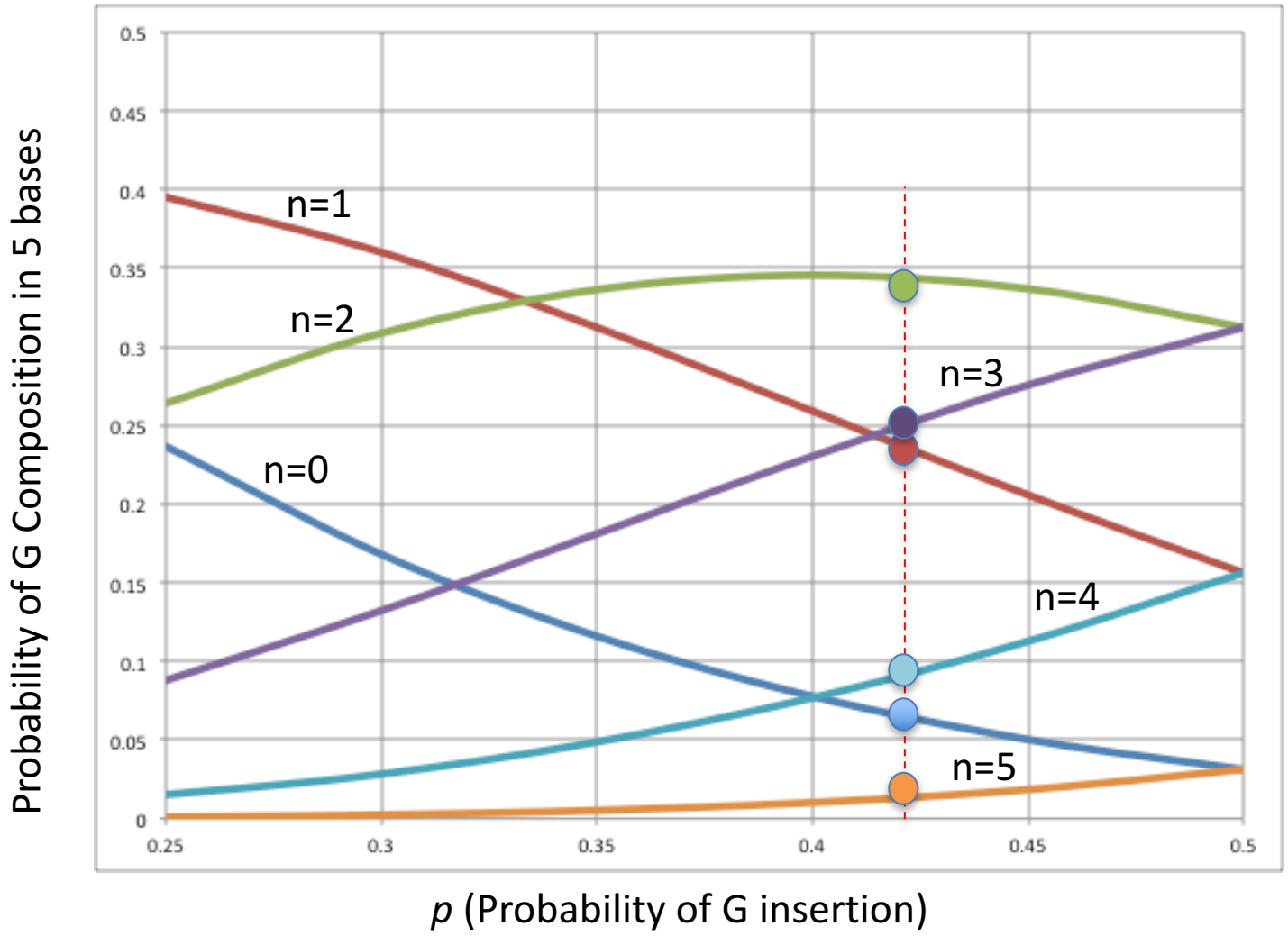
- For L5 there are 6 quantities measured for G (or any other base) composition.
- If we assume the independent insertion frequency model there is then one parameter to fit these 6 quantities
- All these equations must be satisfied simultaneously

$$f(G=0)=(1-p)^5$$
$$f(G=1)=5p(1-p)^4$$
$$f(G=2)=10p^2(1-p)^3$$
$$f(G=3)=10p^3(1-p)^2$$
$$f(G=4)=5p^4(1-p)$$
$$f(G=5)=p^5$$

The measured quantity, which is the fractional base content of the degenerate bases, $f(G)$ should be equal to p

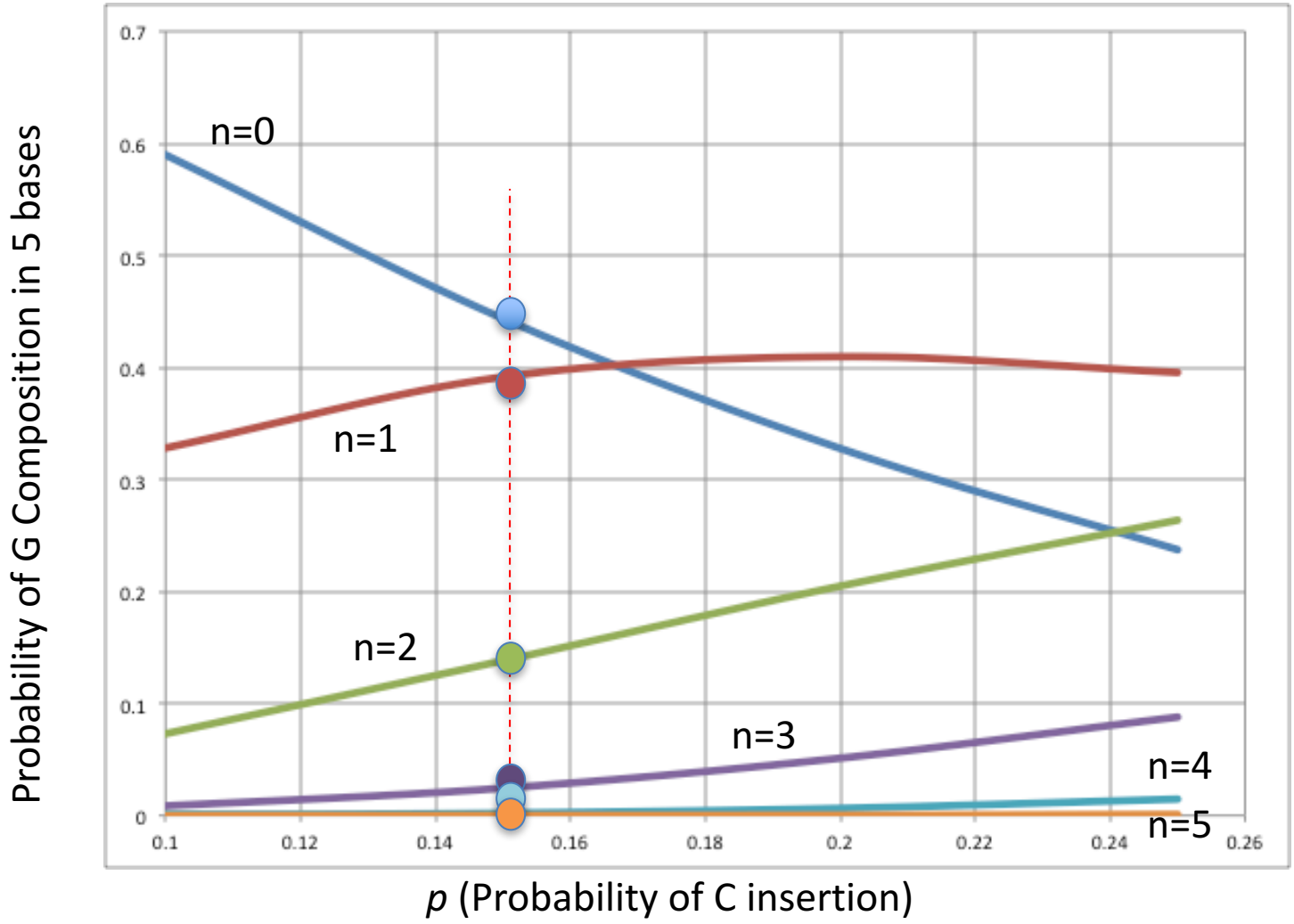
L5: Comparison with Data: G

TTGNATGNCTGNCGTNCGTNACG



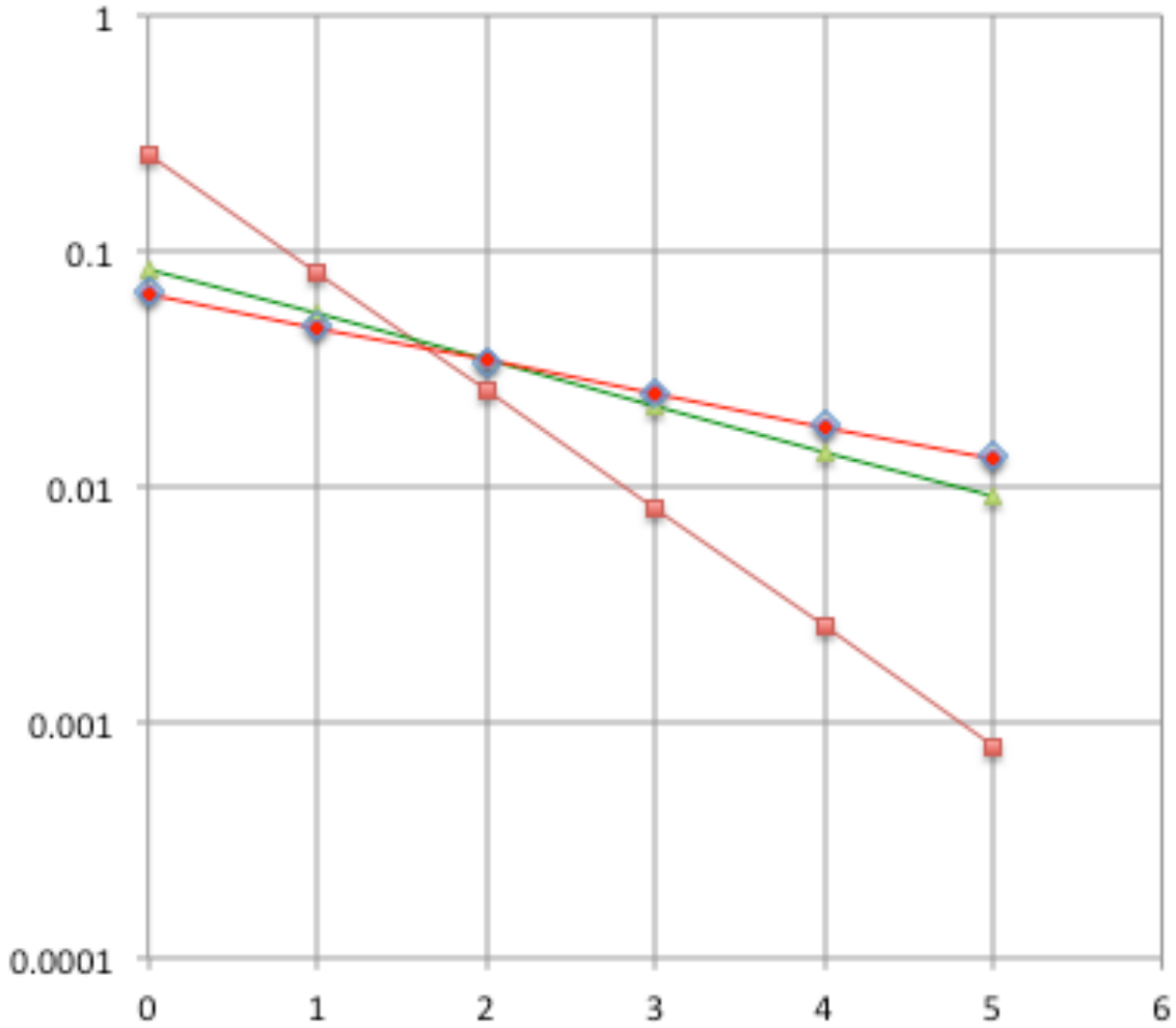
Data is explained by an insertion rate of G's of **~ 0.42,**
and ***f*(G), the fraction of degenerate bases that are G = ~0.42**

L5: Comparison with Data: C



Data is explained by an insertion rate of C's of **0.15,** and ***f*(G)=0.15**

L5: Fraction of reads for oligos with specific number of G's (simple transform)



$f(G=n)=B(5,n)p^n(1-p)^{5-n}$

B(5,n) = binomial coefficient

◇ Fraction G: Composition measured

p=0.25

p=0.4

p=0.42

That *f/B* is a straight line in this semi-log plot confirms **independent insertion model**

The RNA populations synthesized

Library	Sequence
L1	NNNNAUGGCUGACGUACGUNNNN
L2	NNNNUUCGUGCGAUCUAGGNNNN
L5	UUGNAUGNCUGNCGUNCGUNACG
L6	UUGAAUGGCNNNNNUACGUGACG
L7	NNNNGCUAGCGUUCAGGUCNNNN
L8	NNNNCAACCAUCGAGCUAANNNN

Base content in the degenerate bases

Library	% A	% C	% G	% T
L1	26	15	34	25
L2	27	15	33	25
L5	20	15	42	23
L6	22	16	39	23
L7	23	18	37	22
L8	23	17	41	19
Avg.	**23.5**	**16.0**	**37.7**	**22.8**

So we can characterize the input populations in detail

Summary

- We have a good library protocol for exRNA analysis
- We can now quantitatively characterize large synthetic RNA-populations to a high degree
- Large synthetic RNA populations are powerful tools for dissecting sequence-specific bias.
- New computational approaches probing sequence features that are the sources of bias detect patterns
- Goal is optimized protocols and bias correction to improve quantitation of sRNAS-seq.
- Much more to be done. *A work in progress!*

Acknowledgments

Alton Etheridge, Galas lab, PNDRI
Kai Wang, ISB

Muneesh Tewari lab, U of Mich.
David Erle lab, UCSF

exRNA RESEARCH PORTAL
NIH ERCC
Portal: www.exrna.org

Funding: