Supplementary Data
for
Biases in Illumina transcriptome sequencing caused by random hexamer priming

Kasper D. Hansen* Steven E. Brenner Sandrine Dudoit

Supplementary Methods

Data

Reads were obtained from the data sources indicated in the various publications, see Supplementary Table S3 for details.

The phi X data as well as the WT data have been deposited to the Short Read Archive (SRA) at NCBI under the accession number SRA009901.

Mapping

For each experiment, a single lane was chosen at random for analysis. The only exception is Mortazavi (1), where the available data are reads from an entire sample (pooled across several lanes).

Reads were mapped using Bowtie(15), version 0.9.8.2. The argument list includes `-v 0 -m 1 --tryhard`, ensuring that only reads mapping uniquely with zero mismatches to the reference genome are kept. Bowtie considers the “N” nucleotide a no-match to any base.

Genomes were obtained from the UCSC genome browser. The UCSC version numbers are hg18 for *H. sapiens*, sacCer1 for *S. cerevisiae*, and mm9 for *M. musculus*. The genome for ϕX 174 was obtained from NCBI (accession number NC_001422) and the genome for *S. pombe* from Sanger (files Chromosome[123].fasta dated “August 23, 2007”).

Annotations were also obtained from UCSC. The tables used were sgdGene for sacCer1, ccdsGene for hg18, and ccdsGene for mm9.

Free energies

Binding energies were computed using the program “DAN”, from the Mobyle webserver hosted at the Pasteur Institute, which provides access to the EMBOSS (16) suite of programs. Default parameter values were used, except for the flag “use rna data values” which was set to true, to ensure that binding energies for RNA-DNA duplexes were computed.

Part of the table of binding energies is provided below; we used the column DeltaG for ΔG in Figure S3.

```
Sequence     Tm    GC   DeltaG  DeltaH  DeltaS
  AAAAAA  -39.7     0   -5.570     -33   -92.0
  CCCCCC   35.2   100  -16.725     -61  -148.5
  GGGGGG   35.2   100  -16.725     -61  -148.5
  TTTTTT  -39.7     0   -5.570     -33   -92.0
...
```

Re-weighting scheme

We show in this manuscript that there is a bias in the nucleotide frequencies at the beginning of reads. We present an approach for adjusting the biased distribution at the beginning of the reads, so that it is similar to the distribution at the end of the reads, which we assume to be reflective of the nucleotide content of the transcriptome. We do this by associating a weight with each read, such that reads beginning with a heptamer over-represented in the distribution at the beginning

*To whom correspondence should be addressed.
Email: khansen@stat.berkeley.edu

relative to the end are down-weighted and vice-versa. Then, the expression level of a genomic region (e.g., exon) is obtained by adding the weights of the reads falling in that region.

Consider a set of mapped reads. Let $\hat{p}_{\text{hep}:i}$ be the observed heptamer distribution for positions i to $i+6$, so that $\hat{p}_{\text{hep}:1}$ refers to the distribution of the first heptamer of the reads (positions 1 to 7). Define a set of heptamer weights by

$$w(h) = \frac{\frac{1}{6}\sum_{i=24}^{29} \hat{p}_{\text{hep}:i}(h)}{\frac{1}{2}\left(\hat{p}_{\text{hep}:1}(h) + \hat{p}_{\text{hep}:2}(h)\right)}$$

where h denotes a heptamer (out of $4^7 = 16,384$). This set of weights is referred to as the "recommended" weights. We also consider two alternative sets of weights. One is the "naive" set of weights, which simply uses the distribution of the first heptamer in the denominator,

$$w_{\text{naive}}(h) = \frac{\hat{p}_{\text{hep}:29}(h)}{\hat{p}_{\text{hep}:1}(h)}$$

The other set of weights is used as a "control", in that it is superficially similar to the recommended set, but with adverse effects,

$$w_{\text{control}}(h) = \frac{\frac{1}{6}\sum_{i=24}^{29} \hat{p}_{\text{hep}:i}(h)}{\hat{p}_{\text{hep}:2}(h)}$$

Supplementary Figure S8 shows the marginal distributions of the three sets of weights, which are very similar and symmetric (on the log scale) around 0.

A mapped read is associated with the stranded genomic location of its 5'-end. A stranded genomic location is considered mappable, if the read associated with the location maps uniquely to the genome. If the first heptamer of the read is h, the weight associated with the read is $w(h)$. The unadjusted base-level count for a stranded genomic location is the number of reads associated with the stranded location and the re-weighted base-level count is the sum of the weights associated with these reads.

As an example, consider data from the WT experiment in *S. cerevisiae* (5) (7.8 million mapped reads). The following is an excerpt of the base-level unadjusted and re-weighted counts associated with locations on the sense strand of a highly-expressed gene (YOL086C).

strand	location l	heptamer $h(l)$	count $c(l)$	weight $w(h(l))$	re-weighted count, $c_w(l)$
…					
-1	159792	TTGGTCG	17	1.39	23.6
-1	159793	TTTGGTC	17	0.25	4.3
-1	159794	TTTTGGT	65	0.31	20.4
-1	159795	GTTTTGG	72	0.32	23.3
-1	159796	CGTTTTG	10	1.66	16.6
…					

Here, l is the genomic location, $h(l)$ is the heptamer associated with a read mapped to this location, and $c(l)$, $w(h(l))$, and $c_w(l) = c(l)w(h(l))$ are, respectively, the unadjusted base-level count, the weight, and the re-weighted base-level count.

Evaluating the re-weighting scheme

The re-weighting scheme is evaluated on a genome-wide scale using four datasets from *S. cerevisiae* (6-10 million mapped reads per dataset), with labels "WT", "IsoWT", "XRN", and "RLP" (5), and one dataset from *H. sapiens* (80 million mapped reads), labelled "MAQC" (6), see Supplementary Table S1. In all cases, a single sample was sequenced on multiple lanes, although several library preparations were pooled for the "MAQC" dataset. The data labeled "Bullard" in Figure 1 correspond to a single lane from the "MAQC" dataset and the data labelled "Lee" in Figure 1 correspond to a single lane from "IsoWT".

Based on annotation, regions of constant expression (ROCEs) were defined as genomic regions of maximal size such that all bases in the region are annotated as belonging to the same set of transcripts. (For example, two overlapping genes will be split into three ROCEs.) In general, ROCEs roughly correspond to coding sequences in *S. cerevisiae* and exons in *H. sapiens*. Furthermore, we required at least 100 mappable bases (for *S. cerevisiae*) and 50 mappable bases (for *H. sapiens*) and an average (stranded) unadjusted base-level count of at least one (on either strand). Roughly 10% of possible ROCEs are selected for each dataset, focusing on non-small, highly-expressed regions (Supplementary Table S2).

Pearson χ^2 goodness-of-fit statistics were calculated for each strand of each ROCE. The statistic is defined as

$$\chi^2 = \sum_{l=1}^{L} \frac{(d(l) - \lambda)^2}{\lambda}$$

where L is the number of mappable bases in the ROCE,

l indexes mappable bases, $d(l)$ is the unadjusted or re-weighted base-level count (either $d(l) = c(l)$ or $d(l) = c_w(l)$, as defined above), and $\lambda = \sum_l d(l)/L$ is the average (unadjusted or re-weighted) base-level count for the ROCE. Coefficients of variation were calculated in a similar manner, except that unmappable bases as well as bases with zero counts were excluded. (This was done to ensure that the large number of bases with zero counts did not adversely affect estimation of the standard deviation.) Anscombe residuals(17) were computed as

$$r_l^{\text{ans}} = \frac{3/2\big(d(l)^{2/3} - \lambda^{2/3}\big)}{\lambda^{1/6}}$$

These are known to approximately have a standard normal distribution, if the base-level counts $d(l)$ are independently and identically Poisson distributed.

There is one possible concern when evaluating the re-weighting scheme with the Pearson χ^2 goodness-of-fit statistic. If all weights are equal $w(h) = w$, the Pearson χ^2 goodness-of-fit statistic for the re-weighted base-level counts is exactly w times the Pearson χ^2 goodness-of-fit statistic for the unadjusted counts, implying that we report an improved fit by trivially setting the weights to be equal and less than one, $w(h) = w < 1$. This is not desirable behavior for a goodness-of-fit statistic, which is why we also consider the coefficient of variation that does not have this problem. Note, however, that the logarithm of our recommended weights are symmetrically distributed around zero (Figure S8c), showing that only around 50% of the weights are less than one.

Stranded coverage plots were made by adding the weights of reads associated with each base in each stranded ROCE (weights of one for unadjusted counts). For such a standard coverage plot, each position of the read is assigned the same weight. We have also considered position-specific weights, where each base j is associated with the heptamer h_j starting at that position and the weight w_j for base j is defined by

$$w_j(h_j) = \frac{\hat{p}_{\text{hep:29}}(h_j)}{\hat{p}_{\text{hep:}j}(h_j)}$$

This variant had little effect on the coverage plot.

Supplementary Results

Adjusting for the bias

We evaluate the re-weighting scheme on regions that are expressed highly enough to allow consideration of base-level behavior (Supplementary Methods). These ROCEs account for roughly 10% of an organism's exons or coding regions. Supplementary Figure S6 shows data from one such region in *S. cerevisiae*. The re-weighted base-level counts have fewer and smaller extreme values than the unadjusted base-level counts and the associated Anscombe residuals(17) are smaller and more symmetrically distributed around zero. This is reflected in a substantial decrease in the Pearson χ^2 goodness-of-fit statistics. However, even after re-weighting, the Pearson χ^2 goodness-of-fit statistics are still very large, indicating that the re-weighted base-level counts are far from uniformly distributed along the region. There is only a small effect on the coverage plot; while all high coverage peaks are reduced in magnitude, substantial heterogeneity still remains.

The decrease in extreme base-level counts is reflected in the reduction of the Pearson χ^2 goodness-of-fit statistics and coefficients of variation (Supplementary Figures S7 and S8). The re-weighted Pearson χ^2 goodness-of-fit statistics are roughly 50%-65% (dataset dependent) of the unadjusted statistics across the subset of highly-expressed regions.

Supplementary Figure S8 also depicts the effect of re-weighting using two alternative sets of weights. One set, denoted control weights, is based on using but the second heptamer of the read (positions 2 to 8). Unsurprisingly, this set of weights performs poorly, with a substantial increase in Pearson χ^2 goodness-of-fit statistics and coefficients of variation, despite the control weights having a marginal distribution very similar to that of the recommended weights. This shows that the performance improvement when using the recommended weights is not accidental. Also shown in Supplementary Figure S8 is the performance of naive weights, using just the first heptamer of the read. It is intriguing that the recommended weights perform much better than the naive weights. The recommended weights and the naive weights differ in two aspects. First, the recommended weights represents the distribution of heptamers at the end of the reads by averaging heptamer distributions starting at positions 24 to 29, as opposed to simply using the heptamer distribution at position 29. This is a natural variance reduction method, however the performance improvement is very modest (data not shown). The substantial performance improvement comes from estimating the heptamer distribution at the beginning of the reads by averaging the heptamer distributions starting at positions 1 and 2, something that is surprising given that these two distributions are very different. Note that the performance improvements from using the recommended set of weights instead of the naive weights are dataset dependent, with the biggest effect on “IsoWT” and “RLP” that were sequenced in the same batch. A possible, but unsatisfying explanation is that, due to end-repair, the first sequenced nucleotide may not be the first nucleotide of the primer.

There could be some concern over the fact that we apply the weights on the same dataset that was used for estimation. Specifically, since a substantial number of reads map to highl-expressed ROCEs, the weights might be optimized for these regions. To address this, we have evaluated the re-weighting scheme by estimating the weights using only reads mapping to the genome, but not to highly-expressed ROCEs. There is no discernible difference in performance when using these two estimation methods. For convenience, we recommend estimating weights using all reads mapping to the genome.

References

1. Mortazavi,A., Williams,B., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, **5**, 621–628.

2. Dohm,J.C., Lottaz,C., Borodina,T. and Himmelbauer,H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res*, **36**, e105.

3. Marioni,J., Mason,C., Mane,S., Stephens,M. and Gilad,Y. (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, **18**, 1509–1517.

4. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.

5. Lee,A., Hansen,K.D., Bullard,J., Dudoit,S. and Sherlock,G. (2008) Novel Low Abundance and Transient RNAs in Yeast Revealed by Tiling Microarrays and Ultra High-Throughput Sequencing Are Not Conserved Across Closely Related Yeast Species. *PLoS Genet*, **4**e1000299.

6. Bullard,J.H., Purdom,E.A., Hansen,K.D., and Dudoit,S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94.

7. Wang,J., Wang,W., Li,R., Li,Y., Tian,G., Goodman,L., Fan,W., Zhang,J., Li,J., Zhang,J., et al (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60–65.

8. Bentley,D.R., Balasubramanian,S., Swerdlow,H.P., Smith,G.P., Milton,J., Brown,C.G., Hall,K.P., Evers,D.J., Barnes,C.L., Bignell,H.R. et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

9. Meissner,A., Mikkelsen,T.S., Gu,H., Wernig,M,, Hanna,J., Sivachenko,A., Zhang,X., Bernstein,B.E., Nusbaum,C., Jaffe,D.B. et al (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, **454**, 766–770.

10. Chiang,D., Getz,G., Jaffe,D., O'Kelly,M., Zhao,X., Carter,S., Russ,C., Nusbaum,C., Meyerson,M. and Lander,E. (2008) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods*, **6**, 99–103.

11. Mikkelsen,T.S., Hanna,J., Zhang,X., Ku,M., Wernig,M., Schorderet,P., Bernstein,B.E., Jaenisch,R., Lander,E.S. and Meissner,A. (2008) Dissecting direct reprogramming through integrative genomic analysis. *Nature*, **454**, 49–55.

12. Wilhelm,B., Marguerat,S., Watt,S., Schubert,F., Wood,V., Goodhead,I., Penkett,C., Rogers,J. and Bahler,J. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239–1243.

13. Bloom,J.S., Khan,Z., Kruglyak,L., Singh,M. and Caudy,A.A. (2009) Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics*, **10**, 221.

14. Nagalakshmi,U., Wang,Z., Waern,K., Shou,C., Raha,D., Gerstein,M. and Snyder,M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.

15. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**, R25.

16. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet*, **16**, 276–277.

17. McCullagh,P. and Nelder,J.A. (1990) *Generalized Linear Models*. CRC Press, Bacon Raton, FL.

18. Mamanova, L., Andrews,R.M., James,K.D., Sheridan,E.M., Ellis,P.D., Langford,C.F., Ost,T.W.B., Collins,J.E. and Turner,D.J. (2010) FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nat Methods*, **7**, 130–132.

Supplementary Tables and Figures

Experiment	Category	Organism	Read Length	Total Number of Reads	Number of Mapped Reads	Percent Mapped
Bentley (8)	DNA	*H. sapiens*	37	4,255,242	2,419,677	57
Bloom (13)	RNA, other	*S. cerevisiae*	32	5,254,616	2,203,211	42
Bullard (6)	RNA	*H. sapiens*	35	14,032,875	6,357,627	45
Chiang (10)	DNA	*H. sapiens*	36	4,610,681	1,415,113	31
Lee (5)	RNA	*S. cerevisiae*	36	11,147,684	2,425,485	22
Mamanova (18)	-	37	6,246,372	1,541,669	25	
Marioni (3)	RNA	*H. sapiens*	32	13,017,169	2,947,601	23
Meissner (9)	DNA	*M. musculus*	36	9,469,890	3,137,481	33
Mikkelsen (11)	DNA	*M. musculus*	36	6,389,720	511,408	8
Mortazavi (1)	RNA	*M. musculus*	25	31,116,663	8,144,763	26
Nagalakshmi, RH (14)	RNA, other	*S. cerevisiae*	35	6,219,951	394,358	6
Nagalakshmi, DT (14)	RNA, other	*S. cerevisiae*	35	3,444,654	378,593	11
phi X	DNA	ϕX	45	6,133,524	4,188,163	68
Wang, heart (7)	RNA	*H. sapiens*	32	20,169,301	7,423,983	37
Wang, brain (7)	RNA	*H. sapiens*	36	10,112,968	3,097,183	31
Wang, DNA (4)	DNA	*H. sapiens*	23	2,701,647	1,433,185	53
Wilhelm (12)	RNA, other	*S. pombe*	38	2,697,239	1,133,432	42

Table S1. Numerical summaries of the different experiments. Number of total and mapped reads for each experiment, as well as read length. The following shorthand notation is used in the “Category” column, “RNA” for “RNA-Seq”, “DNA” for “DNA-Seq”, and “RNA, other” for “RNA-Seq, other protocols”. Most datasets corresponds to a single lane worth of data, although there are exceptions. Note that some of the input files have been “purity filtered” according to standard settings in the Illumina pipeline and some have not. Purity filtering typically removes around 50% of the reads and is not specified in the data files. Hence, comparisons of “Percent Mapped” between different data sources are difficult.

Name	Organism	Lanes	Read Length	Total Number of Reads	Number of Mapped Reads	Percent Mapped	Number of ROCEs
MAQC	*H. sapiens*	2x7	35	183,797,505	80,454,187	44	27,596
IsoWT	*S. cerevisiae*	1x4	36	45,101,647	9,611,129	21	662
WT	*S. cerevisiae*	1x4	36	46,087,723	7,832,287	17	552
XRN	*S. cerevisiae*	1x4	36	46,716,590	5,945,829	13	459
RLP	*S. cerevisiae*	1x4	36	45,346,044	9,939,379	22	711

Table S2. Numerical summaries of datasets used to evaluate the re-weighting scheme. The dataset "MAQC" is from (6), while the datasets "IsoWT", "WT", "XRN", and "RLP" all are from (5). The notation "XxY" in the "Lanes" column indicates the sample was run on "X" flow-cells and "Y" lanes (per flow-cell). The number of ROCEs is the number of highly-expressed regions in Figures S6- S8.

Experiment	File	Source
Bentley	`200x36x36-071113_EAS56_0053-s_1_2.fastq`	SRA
Bloom	`RMg9.fastq`	Caudy Lab
Bullard	`FL1_B_4_export.txt`	SRA
Chiang	`SRR002793.fastq`	SRA

IsoWT	`SRR003157.fastq`, `SRR003158.fastq`
	`SRR003159.fastq`,`SRR003160.fastq`
		SRA

Lee	`SRR003159.fastq`	SRA
Mamanova	`ERR007689_1.fastq`	ENA
Marioni	`s_1_eland_result.txt`	SRA
MAQC	SRX16369, SRX16370 SRX16371, SRX16372	SRA
Meissner	`205CY.7.all.fastq`	Broad
Mikkelsen	`13530.2.all.fastq`	Broad
Mortazavi	`mm9Brain1.comb.eland2`	Wold Lab
Nagalakshmi, RH	`SRR002058.fastq`	SRA
Nagalakshmi, DT	`SRR002062.fastq`	SRA
phi X	SRA009901	SRA (this study)

RLP	`SRR003161.fastq`, `SRR003162.fastq`
	`SRR003163.fastq`, `SRR003164.fastq`
		SRA

Wang, heart	`GSM325478_heart_HCT170_hg18realign.txt`	GEO
Wang, brain	`GSM325490_brain_s1368_realign.txt`	GEO
Wang, DNA	`070706_S80_FC6083_L1_YHDASA.fq`	BGI
Wilhelm	`run30_s7_Spombe_cDNA.fastq`	AE
WT	SRA009901	SRA (this study)

XRN	`SRR003169.fastq`, `SRR003170.fastq`
	`SRR003171.fastq`, `SRR003172.fastq`
		SRA

Table S3. Data sources. For each experiment, we note what input file was used in this manuscript and from where it was obtained. “GEO” is the gene expression omnibus, “SRA” is the short read archive, “AE” is Array Express, “ENA” is the European Nucleotide Archive, “BGI” is the Beijing Genomics Institute website[1], “Wold Lab” is the Wold Lab website[2], “Caudy Lab” is the Caudy lab website[3], “Broad” is the Broad Institute website[4], and “SRA (this study)” means the data was submitted to SRA as part of this manuscript.

[1] `http://yh.genomics.org.cn/rawdataDownload.jsp`

[2] `http://woldlab.caltech.edu/html/rnaseq`

[3] `http://genomics.princeton.edu/caudylab/yeast_cDNA_sequencing`

[4] `ftp://ftp.broad.mit.edu/pub/papers/chipseq`

Protocol	Group	Experiments
mRNA fragmentation, followed by random hexamer priming	RNA-Seq	Bullard, Lee, Marioni, Mortazavi
Random hexamer priming, no fragmentation	RNA-Seq	Wang (heart, brain)
Random hexamer priming, followed by cDNA fragmentation using DNase I	RNA-Seq, other protocols	Nagalakshmi, RH
Oligo-dT priming, followed by cDNA fragmentation using DNase I	RNA-Seq, other protocols	Nagalakshmi, DT
Oligo-dT priming, followed by cDNA fragmentation by nebulization	RNA-Seq, other protocols	Wilhelm
Oligo-dT priming, followed by cDNA fragmentation by sonication	RNA-Seq, other protocols	Bloom

Table S4. Overview of RNA-Seq protocols. All experiments used poly-A selection to enrich for mRNA, except for “Bloom” that employed ribominus to deplete the sample of rRNA.

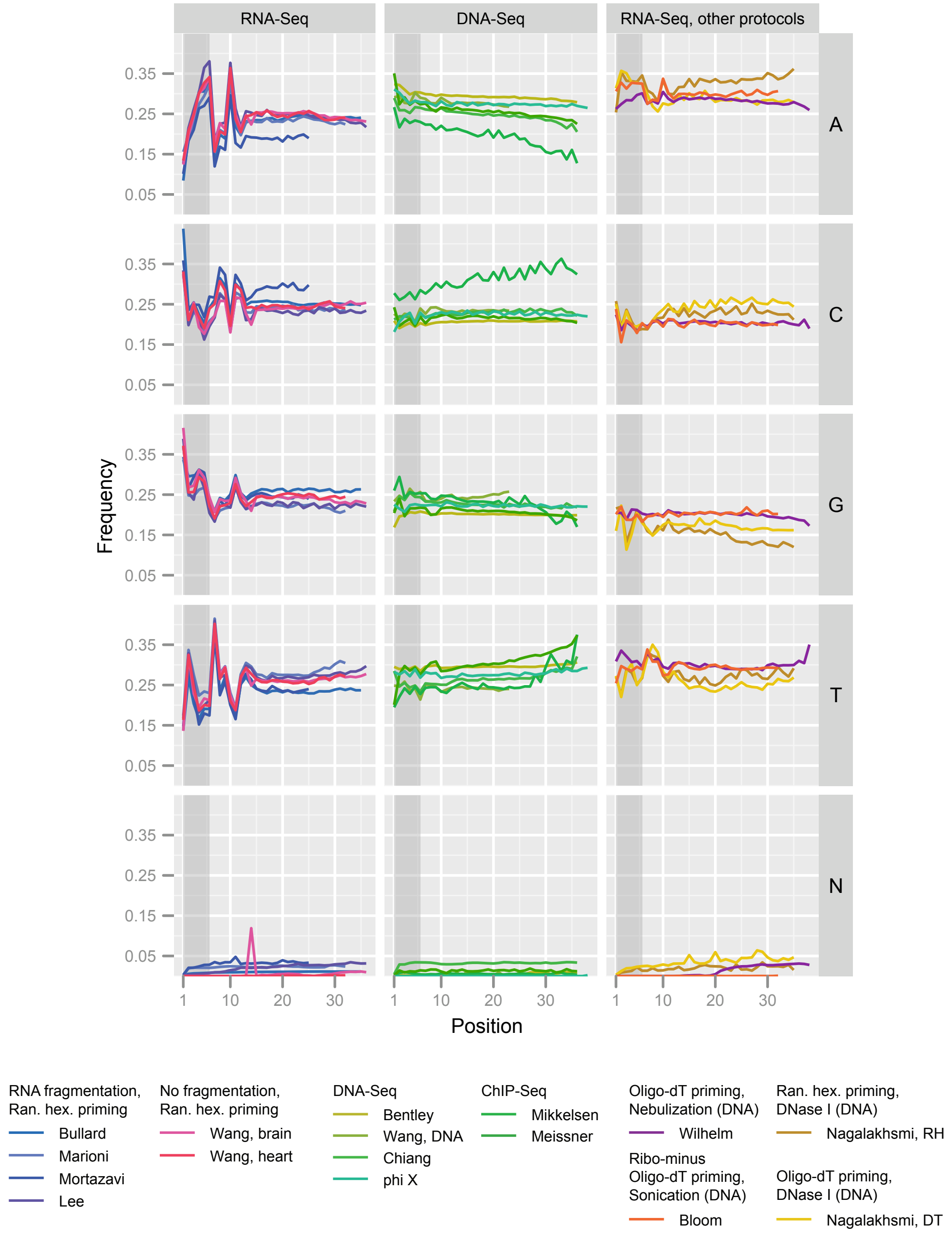

Figure S1. Nucleotide frequency vs. position for all reads. See Supplementary Table S1 for the total number of reads. This figure is comparable to Figure 1. Since the reads have not been mapped, they cannot be extended upstream and may include an “N” call.

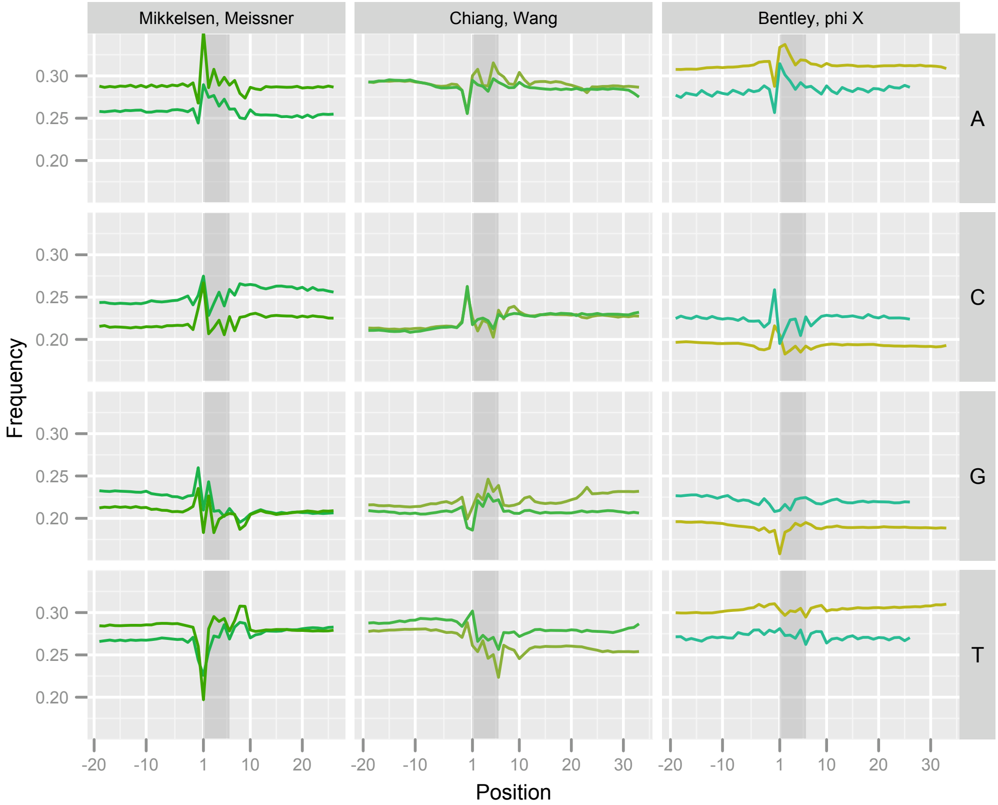

Figure S2. Nucleotide frequency vs. position for mapped DNA-Seq reads by fragmentation method. This figure is a more careful look at the "DNA-Seq" data from Figure 1, where the experiments have been split by fragmentation method. "Mikkelsen" and "Meissner" refer to ChIP-Seq experiments performed by the same group at the Broad Institute. The experimental protocol describes the fragmentation as using either a "Branson 250 Sonifier or a Diagenode Bioruptor" to a size range of 200-700bp followed by ChIP. The "Chiang" experiment "sheared the DNA according to Illumina's protocols" and was grouped with the "Wang" experiment based on visual assessment of their similarity. Both "Wang" and "Bentley" used fragmentation by nebulization; for "Wang" the fragmentation lasted 9 minutes, while for "Bentley" it lasted 6 minutes. "phi X" is a control lane and was grouped with "Bentley" based on visual assessment. The two ChIP-Seq experiments show a pattern distinctly different from the DNA-Seq experiments. There is some similarity between the 9-minute and 6-minute nebulization. The patterns extend upstream of the first base of the read, consistent with a fragmentation effect. The conclusion is that the employed fragmentation method does create a distinct pattern, but much smaller in magnitude than that caused by the RNA-Seq random hexamer priming. Legend as in Supplementary Figure S1.

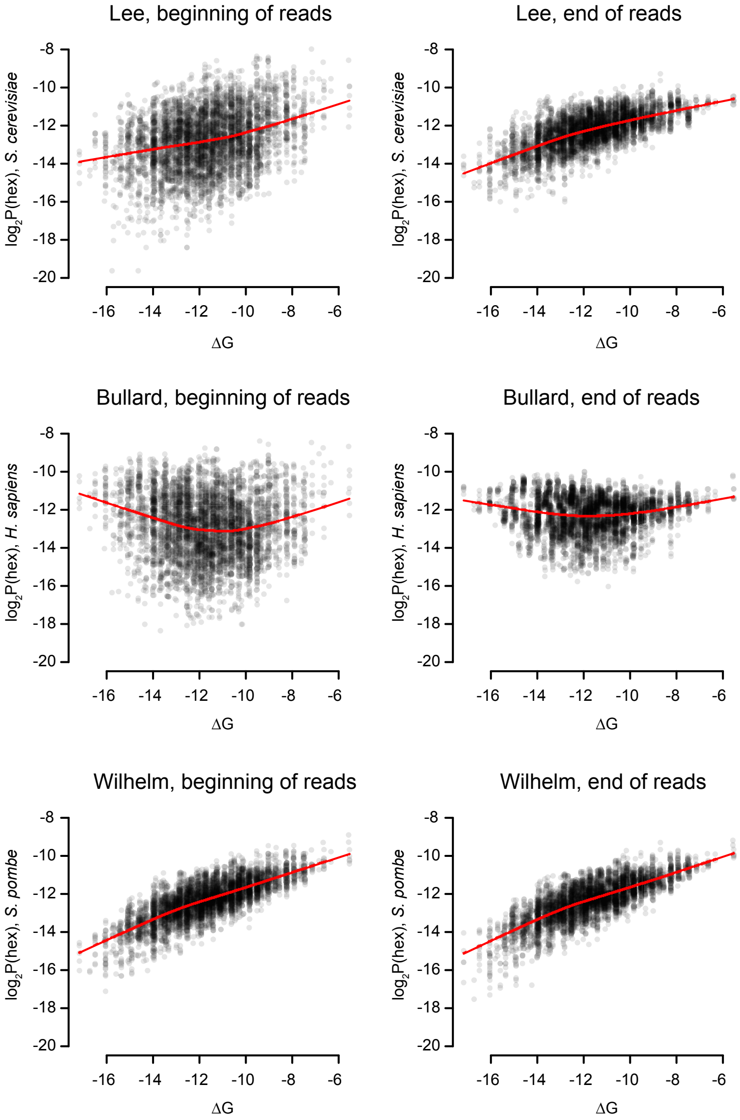

Figure S3. Hexamer frequencies vs. binding energies. For each hexamer, the logarithm (base 2) of its frequency is plotted against its binding energy (see Supplementary Data, Methods). Depicted are an experiment in *S. cerevisiae* (5) ("Lee"), an experiment in *H. sapiens* (6) ("Bullard"), and an experiment in *S. pombe* (12) ("Wilhelm"), with hexamer frequencies computed either at the beginning of the reads (positions 1 to 6) or at the end (positions 25 to 30). The red line is a lowess smoother (a robust local regression). The "Wilhelm" experiment, which used oligo-dT primers, and the hexamer distributions at the end of the reads serve as controls. The high correlation between hexamer frequency and binding energy for the "Wilhelm" and "Lee" experiments is a feature of the two organisms' transcriptomes and not of the use of random hexamers for priming. Indeed, it appears that the use of random hexamers for priming in the *S. cerevisiae* experiment leads to worse correlation. The distributions depicted in the figure are representative; experiments in *D. melanogaster* and *M. musculus* show the same relationship between hexamer frequencies and binding energies as the "Bullard" experiment.

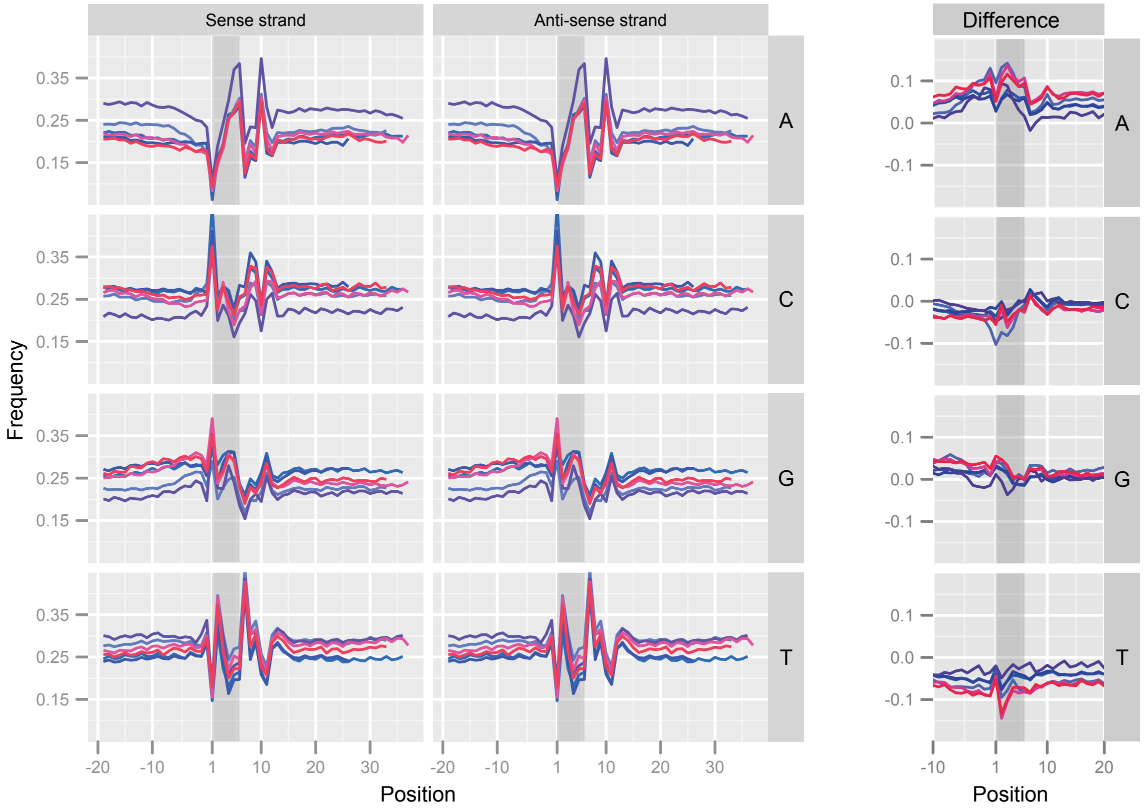

Figure S4. Nucleotide frequency vs. position for stringently mapped, stranded RNA-Seq reads. As Figure 3, but for all four nucleotides.

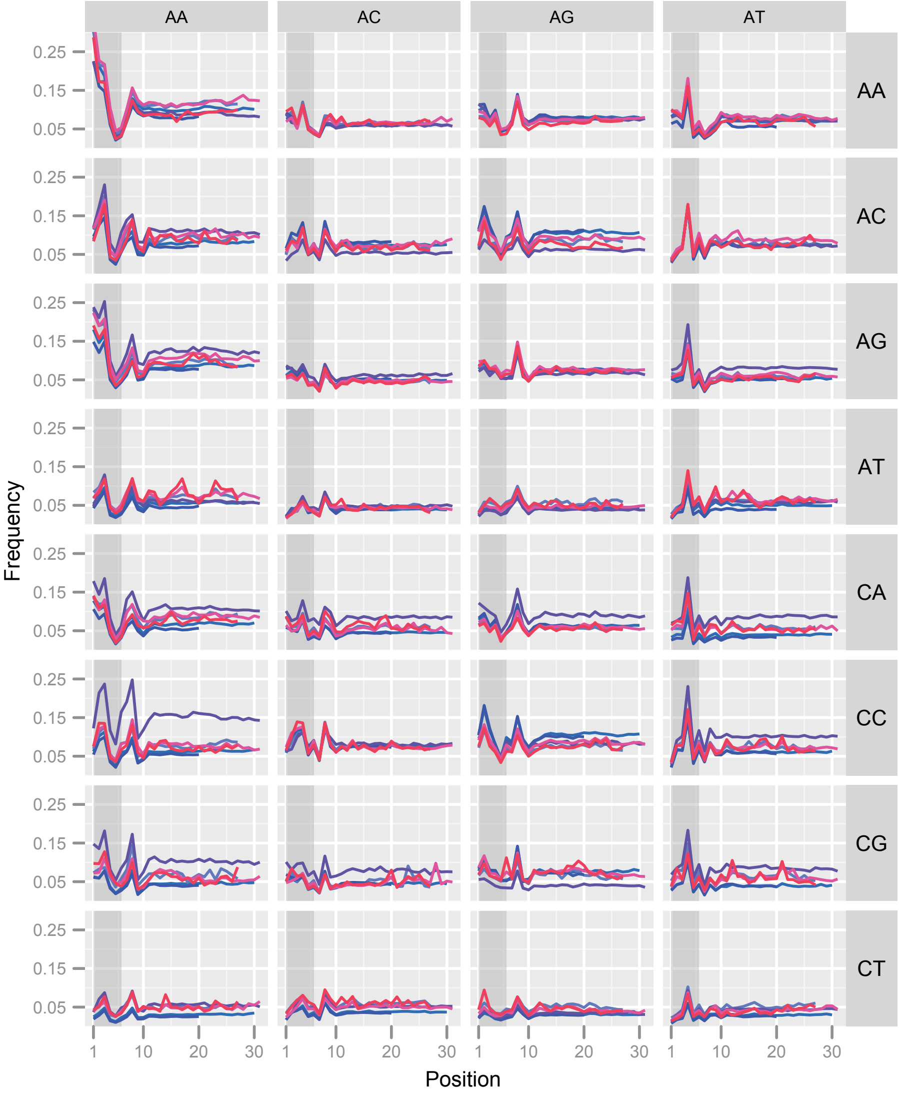

Figure S5. Dinucleotide transition probabilities vs. position for stringently mapped RNA-Seq reads. Transition probabilities are defined as probabilities of a downstream dinucleotide conditional on the upstream dinucleotide. The row label indicates the "From" dinucleotide, while the column label indicates the "To" dinucleotide. For example, the upper right-hand corner shows the frequency of "AT" conditional on observing an "AA", at each position in the read. This figure only shows a subset of the entire set of $4^2 \times 4^2$ = 256 transition probabilities. Note that the scale of the y-axis is different from other figures in this manuscript since there are 16 dinucleotides and only 4 nucleotides. The position indicates the 5' nucleotide of the four nucleotides corresponding to the conditional distributions. In this figure, we examine whether the transition probabilities show a pattern and how far it extends. Since the pattern extends to position 10, we conclude that the nucleotide frequency pattern seen in Figure 1 is not solely caused by a bias in positions 1 to 6 combined with serial correlation of nucleotides in the transcriptome. Legend as in Supplementary Figure S1.

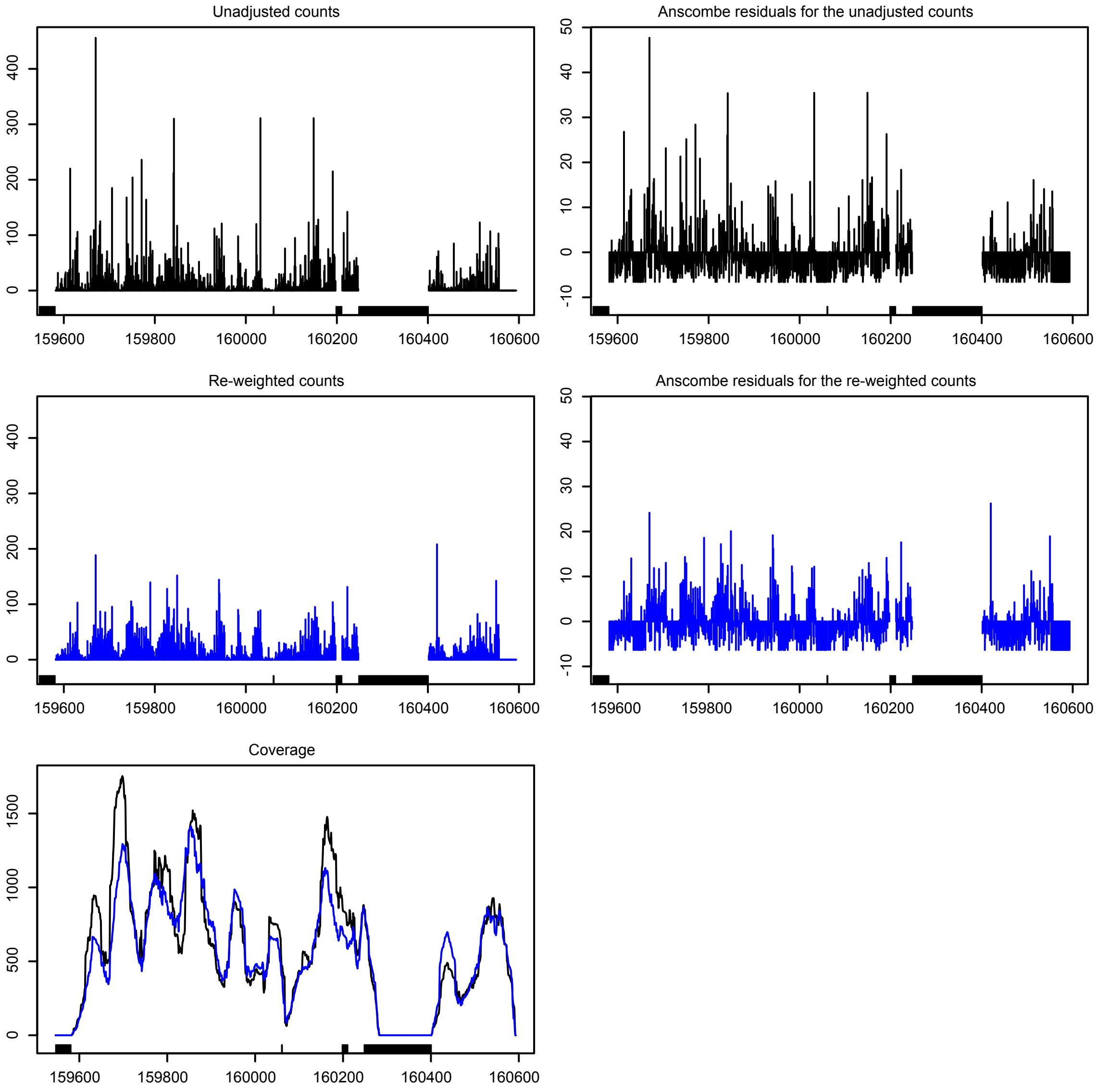

Figure S6. The effect of re-weighting on a single gene. The gene shown is the sense strand of YOL086C for the "WT" experiment in *S. cerevisiae*(5). The plots of unadjusted and re-weighted counts show the base-level counts starting at each location. The Anscombe residuals(17) are designed to have an approximately standard normal distribution if the base-level counts are Poisson distributed. The ticks on the x-axis indicate unmappable bases. The base-level counts have fewer and smaller extreme values using the re-weighting scheme, also reflected in the Anscombe residuals that become more symmetric around zero. There is less effect of the re-weighting on the coverage plot, although the magnitudes of the coverage peaks are reduced. The Pearson χ^2 goodness-of-fit statistic for this region is reduced from 63,022 to 30,620 and the coefficient of variation from 1.68 to 1.27. See Supplementary Data for details.

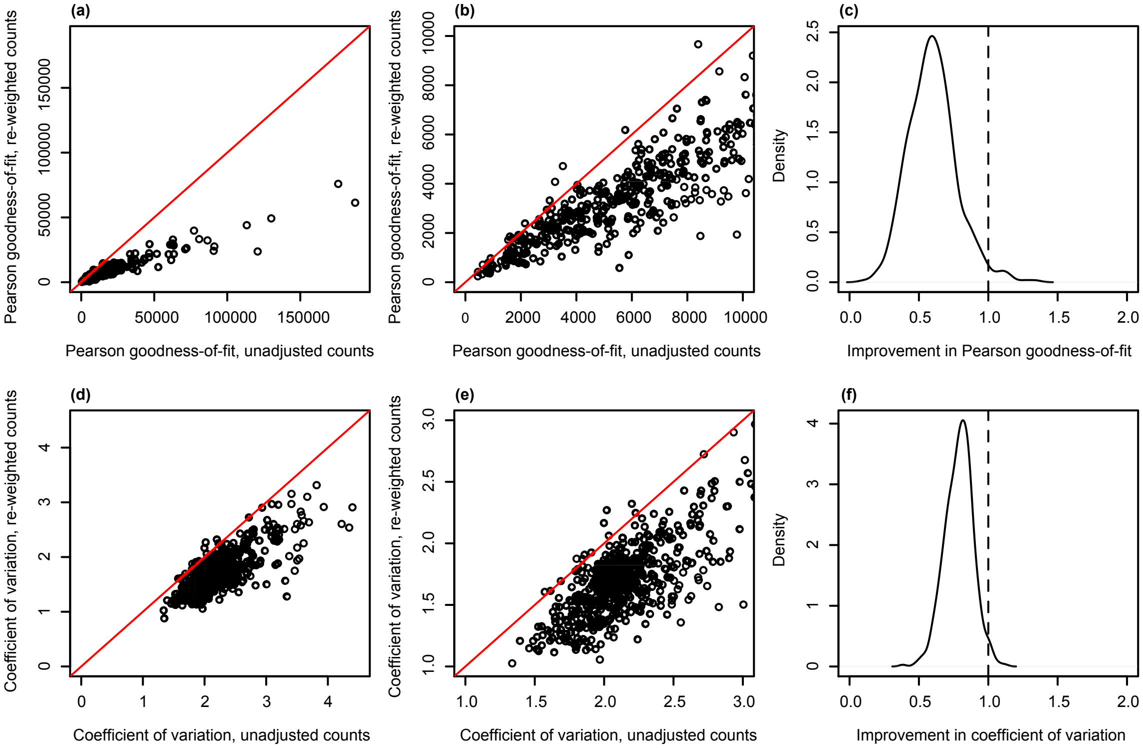

Figure S7. The effect of re-weighting on the Pearson χ^2 goodness-of-fit statistic and coefficient of variation. The data are base-level counts of the sense strand of 552 highly-expressed, non-small regions of constant expression for the "WT" experiment in *S. cerevisiae*. (a) Pearson χ^2 goodness-of-fit statistics for the re-weighted vs. unadjusted base-level counts. (b) A close-up of (a). (c) A density estimate of the distribution of ratios between the re-weighted and the unadjusted Pearson χ^2 goodness-of-fit statistics (values less than one represent improvement due to re-weighting). (d)-(f) As (a)-(c), but for the coefficient of variation. The results for the anti-sense strand are similar, with a slightly larger improvement.

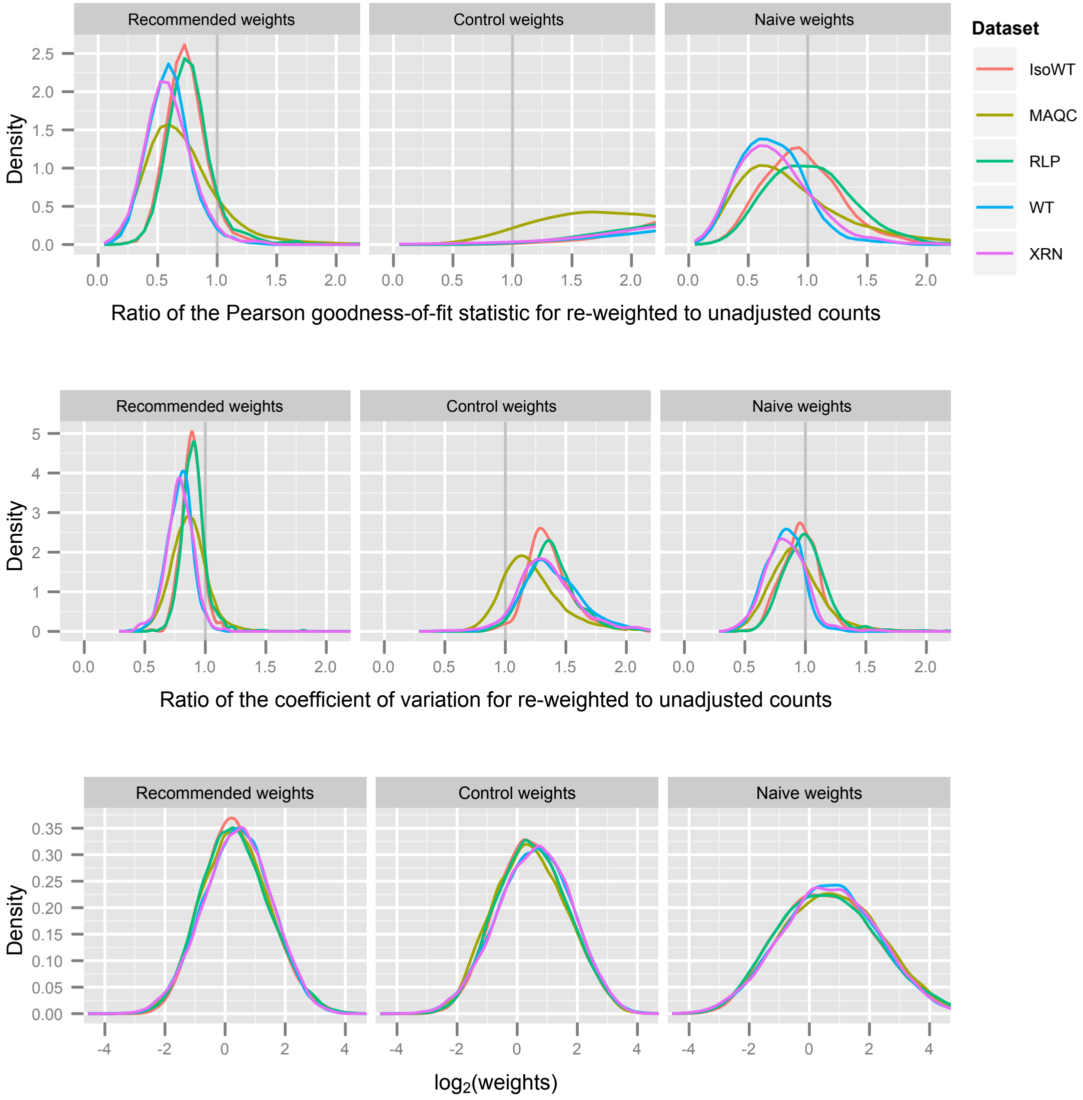

Figure S8. The effect of re-weighting on several datasets, for different re-weighting schemes. For each re-weighting scheme, each dataset, and the sense strand of highly-expressed, non-small regions of constant expression, the effect of re-weighting on the Pearson χ^2 goodness-of-fit statistic and the coefficient of variation (values less than one are improvements for re-weighting) is assessed. Compare to Figure S7c and S7f. The data for the anti-sense strand show similar but slightly better improvements. Also shown is the marginal distribution of the logarithm (base 2) of the weights. Three sets of weights are shown, all described in the Supplementary Methods. "Recommended weights" refer to the set of weights we found to perform best, "control weights" are used as an example of weights which perform poorly, and "naive weights" are a simple example of the re-weighting idea.

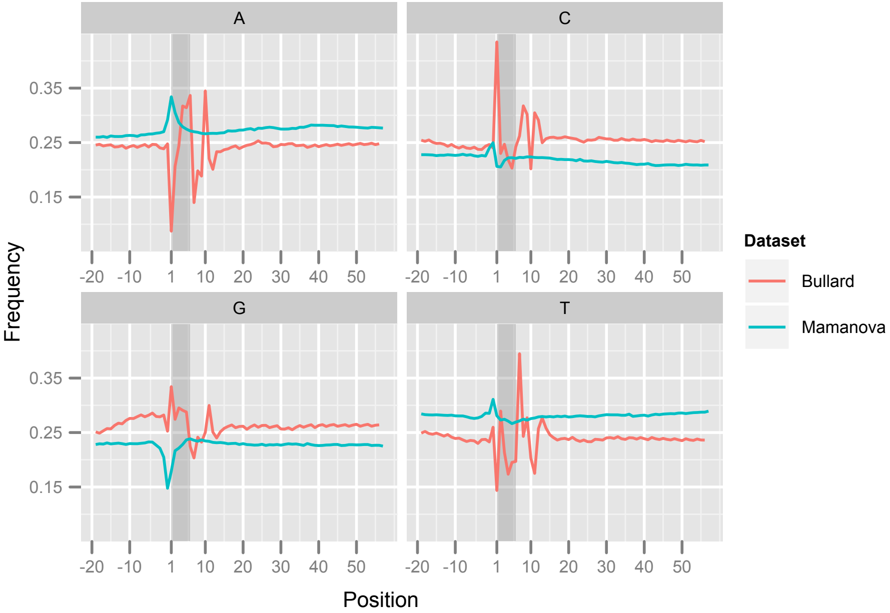

Figure S9. Data from Mamanova et al.(18), generated using the FRT-Seq protocol. This is the first read from a paired-end experiment. Also depicted is a randomly primed RNA-Seq dataset, also from *H. sapiens* (this dataset is also depicted in Figure 1). While the two datasets are from the same organism, they are from different commercially available RNA. As in Figure 1, the reads have been extended 20 bases upstream and downstream using the genome. We see very little pattern in the FRT-Seq data, although there is some nucleotide bias near the first base of the read (with most of it upstream from actual read). Note that the lack of independent datasets generated using the same protocol makes it hard to infer whether this is a general phenomena.