Supporting Text

Multivariate Gaussian Fits to Base Compositions

Each human 3′ UTR is plotted as a point in the two-dimensional (A+T,C-G) plane in Fig.1. These points are fit to two normal distributions, shown as 2-σ ellipses, using a Gaussian mixture model which is described below summarizes the resulting parameters for the two distributions discussed in the main text.

Gaussian Mixture Models

To find ellipses such as those in Fig.1, we use an iterative EM ("expectation-maximization") method. Its use in this context is usually called a Gaussian mixture model.

The basic idea is to alternate between assigning observed points probabilistically to the Gaussians ("expectation") and re-estimating the parameters of the Gaussians ("maximization").

For multi-dimension Gaussians (two in our case), each point $\mathbf{x}$ and the mean μ are two-dimensional vectors. The square of the conventional standard deviation becomes now a 2×2 covariance matrix Σ. If $N(\mathbf{x}\,|\,\mu,\Sigma)$ is the multivariate Gaussian density,

$$N(\mathbf{x}\,|\,\mu,\Sigma) = \frac{1}{(2\pi)^{M/2}\det(\Sigma)^{1/2}} \exp[-\tfrac{1}{2}(\mathbf{x}-\mu)\Sigma^{-1}(\mathbf{x}-\mu)] \qquad \textbf{[1]}$$

and $P(k)$ is an estimate of the fraction of points in Gaussian k (initially a prior, but subsequently re-estimated), then we can estimate p_{nk}, the probability of point n belonging to Gaussian k, by

$$p_{nk} \equiv P(k|n) = \frac{N(\mathbf{x}_n\,|\,\mu_k,\Sigma_k)P(k)}{P(\mathbf{x}_n)} \qquad \textbf{[2]}$$

where

$$P(\mathbf{x}_n) \equiv \sum_k N(\mathbf{x}_n\,|\,\mu_k,\Sigma_k)P(k). \qquad \textbf{[3]}$$

Now that we have the p_{nk}'s, we can re-estimate the parameters of the Gaussians by

$$\begin{aligned} \hat{\mu}_k &= \sum_n p_{nk}\mathbf{x}_n \Big/ \sum_n p_{nk} \\ \hat{\Sigma}_k &= \sum_n p_{nk}(\mathbf{x}_n - \hat{\mu}_k) \otimes (\mathbf{x}_n - \hat{\mu}_k) \Big/ \sum_n p_{nk} \end{aligned} \qquad \textbf{[4]}$$

and

$$\hat{P}(k) = \frac{1}{N}\sum_n p_{nk} \qquad \textbf{[5]}$$

Now repeat the iteration to convergence. In our application, convergence is almost always very rapid. The converged result is the maximum likelihood estimate of the Gaussian parameters.

Error ellipses are drawn as follows: The locus of points $\mathbf{x}$ that are k standard deviations away from the mean μ is given by

$$k^2 = (\mathbf{x}-\mu)\Sigma^{-1}(\mathbf{x}-\mu) \qquad \textbf{[6]}$$

where Σ is the covariance matrix. If we Cholesky-decompose Σ, so that

$$\Sigma = \mathbf{L}\mathbf{L}^T \qquad \textbf{[7]}$$

then

$$\left|L^{-1}(\mathbf{x}-\mu)\right| = k \qquad \textbf{[8]}$$

Now suppose that $\mathbf{z}$ is a point on the unit circle. Then Eq. **8** implies that

$$\mathbf{x} = k\mathbf{L}\mathbf{z}+\mu \qquad \textbf{[9]}$$

is a point on the k-σ locus. Traversing the unit circle in $\mathbf{z}$ and applying the mapping Eq. **9** gives the desired ellipse.

Parameters for Human Genes

The Gaussian mixture model converges to the following parameters for the human $3'$ UTR data

Population	Fraction	μ_{AT}	σ_{AT}
Low $A+T$ population	0.53	0.47	0.07
High $A+T$ population	0.47	0.63	0.05

The full means and covariances of the two fitted Gausssians are

```
 mu1 = {0.4738, 0.0125}
var1 = {{.00488, -.000544}, {-.000544, .00310}}
 mu2 = {0.6316, -.0111}
var2 = {{.00254, -.000161}, {-.000161, .000723}}
```

The probabilistic assignment of a given gene to one or the other population, via its Bayesian odds ratio using the above parameters (and inverting the 2×2 covariance matrix), is accomplished by the following algorithm,

```
z1 = 41.325*exp(-23.702 + 99.485*x - 104.50*x*x +
 21.490*y - 36.677*x*y - 164.50*y*y);
z2 = 118.28*exp(-79.114 + 251.23*x - 199.66*x*x +
 40.593*y - 88.925*x*y - 701.46*y*y);
P = z2/(z2+z1);
```

Here $x \equiv A+T$, $y \equiv C-G$, and the value P is the probability of being in the AT-rich population.

Methods: Differential Word Counts in the Gene Ontology Database

We use the gene ontology hierarchy (1) as a means of associating biologically meaningful keywords with sets of genes in a manner that also allows an estimate of the statistical significance of the association. The procedure is as follows:

- For each gene of interest, take the set of all GO id's to which it is assigned.
- Augment this set recursively by adding all GO id's that are parents of GO id's already in the set, using the GO scheme's "is_a" entries.
- Concatenate the "name" fields of the augmented set of GO's, eliminating punctuation, special symbols, etc.
- Hyphenate (and thus turn into a single "word") frequently occurring phrases (e.g., "amino acid", "programmed cell death"), and delete frequently occurring words that lack specificity (e.g., "and", "process").
- Form the set of words that remain, and assign that set to the gene.

For an individual gene, this is a very "noisy" classification scheme. However, for large sets of genes, we can look for individual words that occur more (or less) often, with high statistical significance, than in some control set of genes.

A particularly well controlled case is when we can assign every gene i a probability p_i of being in some set of interest (like having an AT-rich $3'$ UTR), and therefore a probability $1-p_i$ of not being in that set. We can then calculate for each word j a t value (deviation in standard deviations) and a P value (two-tailed tail probability) expressing the significance with which the word is associated (or negatively associated) with the set of interest.

$$\begin{aligned} n_{j+} &= \sum_i p_i \delta(i,j) \\ n_{j-} &= \sum_i (1-p_i) \delta(i,j) \\ V_{j+} &= \sum_i p_i^2 \delta(i,j) \\ V_{j-} &= \sum_i (1-p_i)^2 \delta(i,j) \\ N_+ &= \sum_i p_i \\ N_- &= \sum_i (1-p_i) \end{aligned} \qquad \textbf{[10]}$$

where $\delta(i,j)$ is 1 if word j is associated with gene i, zero otherwise. Then

$$\begin{aligned} t_j \equiv t\text{-value} &= \frac{\frac{n_{j+}}{N_+} - \frac{n_{j-}}{N_-}}{\sqrt{\frac{V_{j+}}{N_+^2} + \frac{V_{j-}}{N_-^2}}} \\ P_j \equiv p\text{-value} &= \operatorname{erfc}(|t_j|/\sqrt{2}) \end{aligned} \qquad \textbf{[11]}$$

To derive the above, consider $\delta(i,j)$ as a random variable taking the values 0 or 1,

$$\delta(i,j) = \begin{cases} 1 & \text{with probability } q_{ij} \\ 0 & \text{with probability } 1-q_{ij} \end{cases} \qquad \textbf{[12]}$$

Then, because $0^2 = 0$ and $1^2 = 1$,

$$E[\delta(i,j)] = E[\delta(i,j)^2] = q_{ij} \qquad \textbf{[13]}$$

Thus (e.g.)

$$\begin{aligned} \operatorname{Var}(n_{j+}) &= \operatorname{Var}[\sum_i p_i \delta(i,j)] \\ &= \sum_i p_i^2 \operatorname{Var}[\delta(i,j)] \\ &= \sum_i p_i^2 \left(E[\delta(i,j)^2] - E[\delta(i,j)]^2 \right) \\ &= \sum_i p_i^2 (q_{ij} - q_{ij}^2) \approx \sum_i p_i^2 q_{ij} \\ &\approx \sum_i p_i^2 \delta(i,j) = V_{j+} \end{aligned} \qquad \textbf{[14]}$$

Here two approximations are being made. The first approximately equal sign follows from $q_{ij} \ll 1$, meaning that the probability of any single word in any single gene is small (Poisson approximation). The second approximately equal sign replaces the unknown population probability q_{ij} with what amounts to a sample (or Monte Carlo) estimator of it, namely the observed $\delta(i,j)$.

In a similar manner compute $\operatorname{Var}(n_{j-})$, and then $\operatorname{Var}(n_{j+}/N_+ - n_{j-}/N_-)$.

There is an important caution about the computed errors: They assume that a given word's probability of occurrence in each gene is independent (in the null hypothesis). But that is not quite true. For example, when there are multiple splicings, investigators often enter the same GO categories for every splicing. If all words always occurred in "clumps" of four genes, for example, the actual errors would be a factor of two larger than computed above. In reality, the effect is likely much smaller than this, but it is hard to estimate accurately.

More Accurate Prediction of miRNA Target Genes

The premise on which our prediction of miRNA target genes is based is the same as that used by Lewis *et al.* (2): If a gene has significantly more miRNA binding hexamers than expected that are conserved across multiple species, then the presumption is that it is a miRNA target. Below, when we give probabilities that individual genes are miRNA targets, we always mean this in the sense of probabilities that number of miRNA binding hexamers exceeds chance.

All sequences were downloaded from `http://genome.ucsc.edu`, including the multiple alignments. Additionally, miRNA sequences are found at `http://microrna.sanger.ac.uk`.

Target Hexamers

We use Lewis *et al.*'s (2) list of target hexamers taken from families of miRNAs conserved in human, mouse, rat, dog, and chicken. After eliminating duplications, there are 62 of them, from the population of 4,096 (4^6) possible hexamers.

```
AACCAC  AAGGGA  AATGCA  ACAATC  ACCAGC  ACGGGT
ACTGAC  ACTGCC  ACTGTG  ACTTGA  AGCAAT  AGCACA
AGCCAT  AGGTCA  AGTATT  AGTGTT  ATGTGA  ATTTCA
CACCTT  CACTAC  CACTCC  CACTGG  CACTTT  CAGGGT
CATATC  CATTCC  CCAAAG  CCGTCC  CTACCT  CTGTGA
GAACAA  GAATGT  GACACG  GAGATT  GCACTG  GCACTT
GCAGCT  GCATTA  GCCTTA  GCTGCT  GGACCA  GGTACG
GGTGCT  GTTCTC  GTTTAC  TAAGCT  TACCTC  TACTGT
TATGCA  TCAGGG  TCTTCC  TGAAGG  TGAGCC  TGCAAT
TGCACT  TGCAGT  TGCCAT  TGCCTT  TGCTGC  TGTAGC
TGTTAC  TTGCAC
```

Improving the Digraphic Model

We can test the performance of the digraphic model by seeing how well it predicts the frequency of occurrence of all hexamers except the 62 miRNA target hexamers, that is $4,096 - 62$, in $3'$ UTR conserved bases. Certain features are found to stand out immediately:

1. The predictions for hexamers that contain the digraph *CG* are not very good, even with a digraphic model. Because there are only 4 of 62 miRNA hexamers that contain CG, it is prudent simply to eliminate all hexamers containing CG from consideration. (We thus can make no predictions about genes targeted by these four miRNA hexamers.)

2. Even after accounting for digraphic probabilities, we predict low for very AT-rich hexamers, and high for very AT-poor ones. The mean ratio of actual to predicted counts, as a function of $A+T$ is as follows:

```
0 A+T: 0.276
1 A+T: 0.657
2 A+T: 0.951
3 A+T: 1.000
4 A+T: 0.999
5 A+T: 1.149
6 A+T: 1.884
```

We use these values as correction factors on our predictions. Since the vast preponderance of miRNA target hexamers have $A+T=2$, 3, or 4, the actual effect of this correction factor is not large.

We are not able to find any other large systematic effects that deviate from the predictions of the digraphic model.

Estimating the Number of Target Sites

Now given a background model, we have two distinguishable tasks: First, we want to estimate the number of miRNA hexamer binding sites in the genome (below). Then, since there can be, and indeed generally is, more than one miRNA binding site in each miRNA target gene, we want to estimate the number of distinct genes that are targeted. We discuss the latter task below.

We have available the counts of each hexamer in each gene (36,170 genes with $3'$ UTR >100 length), a prediction of what that count should be, based on each gene's digraph structure, and the number of available hexamer slots (that is, hexamer positions that are conserved across species) in each gene.

The 4,096 hexamers include the 58 miRNAs of interest (excluding the four that contain CG), and also our control group of 2,288 hexamers that are "nonspecial" (no CG, $2 \leq A+T \leq 4$). We can predict well the occurrence frequency of these nonspecial hexamers. Almost all the 58 miRNAs, if we didn't know that they were miRNAs, would have been considered nonspecial.

For convenience, we divide the control group up randomly into 39 groups of 58 different hexamers, each a matched control group to the 58 miRNA hexamers. (Actually we could have arbitrarily many groups by resampling, but 39 is enough to get good variance estimates.)

We will want to find estimation methods that remain well behaved and give convergent predictions even as we relax the requirement of all six conserved sites. This is because there may be sites/genes that are causal in human, but not perfectly conserved. We want to find them. Also, if an estimation method gives consistent results as the noise is increased (albeit with increasing error bars), we have more confidence that it is not subject to large systematic errors.

Method 1a: "Prediction-Free" Method

Count all miRNA hexamer occurrences in all genes. Ditto, separately, for each of the 37 control groups. Find the mean and variance of the control groups. The excess is estimated as the

difference between the miRNA count and the mean of the control groups. The error estimate is the sample variance of control groups.

The advantages of this method is that it is simple, and doesn't use the predictions at all.

The disadvantages are that it assumes that the miRNAs could have been drawn from the same distribution as the control groups. But there is only one group of miRNAs, and it is what it is! This produces a systematic error that should grow linearly with the dilution of causal miRNAs by random miRNAs as we relax the all-six-position conservation requirement. In other words the apparent excess should grow in proportion to its estimated error if the actual miRNAs happen to predict high, or shrink if they happen to predict low. (In fact we see roughly this behavior as we relax the conservation requirement.)

However, on the bright side, the systematic error should be small if the signal to noise ratio is high. This is the case, at least when we require all six conserved sites (as do Lewis *et al.* (2)).

Method 1b: "Prediction" Method

Count all miRNA hexamer occurrences in all genes, and sum the predictions of all miRNA hexamer occurrences in all genes. Now do the same for each control group. From the control groups, compute an average bias in the predictions (hopefully close to zero), and a sample variance. The excess is estimated as the miRNA occurrences minus predictions minus the bias. The error estimate is the control group sample variance.

The advantage is that this method is not subject to the particular systematic error of the prediction-free method. It lets the miRNA hexamers be what they are.

The disadvantage is that we incur an additional error due to "prediction error", although it will properly appear in the error estimated from the control groups.

The bias found is always small, around 5%, and varies little with the causal/non-causal dilution factor found for the miRNAs. This indicates that it can very likely be subtracted accurately.

Results for Excess Sites

The results are:

```
Cons.  total  Method 1a     Method 1b
Sites  mirs
----   ----   ---------     ---------
6  14098    8734 +-  437     8335 +-  394 (bias subtracted: 56)
5  24900   11389 +-  820    10421 +-  698 (bias subtracted: -17)
4  37575   13173 +- 1322    11501 +- 1053 (bias subtracted: -322)
3  52389   14561 +- 1903    12036 +- 1436 (bias subtracted: -785)
```

Here the number of inter-species conserved sites that we demand is in the first column, varying from all six down to two. The second column is the total number of miRNA hexamer matches seen in the genome . Notice that Method 1b converges nicely even as the dilution factor

becomes as large as $(68,611-12,121)/12,121 = 4.6$. Method 1a diverges systematically. This looks like the kind of systematic error that we expected, as discussed above. However, there is no reason to think that Method 1a's systematic error is large in the first line, so averaging Methods 1a and 1b there seems appropriate.

Thus, rounding, reasonable summary numbers are: $8,500 \pm 500$ excess sites conserved across all species, and $12,000 \pm 1,000$ excess sites estimated to be in human, where the quoted errors are 1-σ.

Estimating the Number of miRNA Target Genes

It is substantially harder to get an accurate estimate of the number of miRNA target genes (as opposed to binding sites). The reason is that this is actually not a well posed problem, unless we know a priori the distribution of number of causal sites per gene in target genes – which we surely don't.

The data we are given, as before, amounts to a predicted and actual number of miRNA target hexamers for each gene. One might first think of subtracting these two numbers gene by gene, and then trying to do something with the distribution of the resulting differences. The problem is that most predictions are $\ll 1$, and most counts are either 0 or 1, so the subtraction is not very meaningful.

Model-Free Bounds

In our application, "predicted" (which is the prediction of what the histogram would be without causal miRNA target sites) is known only up to some statistical errors. But we can estimate the error of the final lower and upper bounds by using in turn each control set of actual counts as "predicted", always using the miRNA counts as "observed", and looking at the dispersion of the results. This actually overestimates the error, because it also has some dispersion due to the fact that none of the control sets are in fact the (unobservable) before-state of the miRNA set. Note that we use the control sets to estimate the errors on N_{min} and N_{max}, but we use only the miRNA predicted and actual histogram to get the quoted values for N_{min} and N_{max}.

The results are:

	Lower bound	Upper bound
6	1262 +- 93	3651 +- 46
5	1006 +- 107	4518 +- 125
4	810 +- 119	5115 +- 238
3	617 +- 124	5384 +- 351

Note that as we require fewer conserved sites, hence more noise, our ability to get bounds gets not unexpectedly worse. (Also note that, because these are bounds, there is no derogatory meaning associated with their being outside each others error bars.)

Rounding for convenience and using the fact that the number of genes can only increase as we relax the conservation requirement, we can summarize as: There are at least 1,200 genes that are miRNA targets in the human genome, and at most $\approx 5,000$, independently of how causal sites are distributed among the causal genes.

Corrected Zero-Bin Method

If the predicted number of miRNA target sites were much smaller than the number of genes, so that almost all genes predict zero sites (in the "predicted" histogram), then we might simply estimate the number of casual genes by the decrease in the zero bin between "predicted" and "observed." We can call this the "uncorrected zero bin method."

In the case of requiring six conserved positions this approximation is, perhaps barely, acceptable (that is, 14,098 hits in 36,170 genes). However, as we relax the conservation requirement, it quickly goes bad. In the limit of numbers of (chance) hits much greater than number of genes there would of course be almost *no* genes in the "predicted" zero bin, and the method would estimate almost *no* miRNA target genes.

As a correction, we might guess that genes with chance occurrences of miRNA hexamers are neither more nor less likely to be miRNA target genes than genes without such chance occurrences. (This is dubious, but let us proceed a bit further.) We then get what we might call the "corrected zero-bin method":

$$N_{est} = (m_0 - n_0) \times \frac{\sum_i m_i}{m_0} \qquad \textbf{[15]}$$

Of the various methods that we describe, this method is the closest to that described in ref. (2).

Note that this method uses only m_0 from "predicted." We could get this value either (i) from our predictions by the expectated number of zero-bin occurrences, or (ii) from the average of the control groups (with the possibility of systematic error discussed previously). In either case we can use the variance of the individual control groups estimates as an error estimate.

The results are

```
    (1)      (2)
6  1292 or 1402 +- 121
5   873 or 1046 +- 133
4   484 or  685 +- 148
```

It is discouraging, and speaks of systematic errors, that the numbers go down as we relax the conservation requirement, and that they are outside of each others' error bars.

Since, as discussed, the method should be most reliable when the dilution is small, we might expand the error bar somewhat and give a result of $1,400 \pm 150$ from this method. Lewis *et al.* (2) get a somewhat higher value ($\approx 2,000$).

Poisson Odds Ratio Method

A somewhat more rigorous method, still within the spirit of the corrected zero-bin methods, is the following.

Suppose that, in a Poisson process, we observe n counts (e.g., the number of miRNA target hexamers in a given gene). Then the Bayesian odds ratio between hypothesis H_1, that the mean is λ, and hypothesis H_2, that it is $\lambda + \delta$, is

$$\frac{p(H_2)}{p(H_1)} = \left(1 + \frac{\delta}{\lambda}\right)^n e^{-\delta} \times \frac{P(H_2)}{P(H_1)} \qquad \textbf{[16]}$$

where $P(H_1)$ and $P(H_2)$ are the respective priors on the hypotheses. Define $P_{rat} \equiv P(H_2)/P(H_1)$. We have in mind that δ is a number ≥ 1, corresponding to one or more extra (causal) hexamers per gene, and that $P_{rat} \ll 1$, corresponding to only a small fraction of genes being miRNA targets.

If $\lambda \ll 1$, then the odds ratio is, effectively,

$$\frac{p(H_2)}{p(H_1)} = \begin{cases} P_{rat} e^{-\delta} & \text{if } n = 0 \\ \infty & \text{if } n \geq 1 \end{cases} \qquad \textbf{[17]}$$

This is basically the uncorrected zero-bin method, assigning a probability ≈ 1 to genes with one or more counts, and ≈ 0 to genes with zero counts.

In the opposite limit of $\lambda \gg \delta$, we have

$$\frac{p(H_2)}{p(H_1)} \approx e^{(\frac{n}{\lambda} - 1)\delta} P_{rat} \qquad \textbf{[18]}$$

which basically assigns a gene to H_1 or H_2 in a continuous way according to whether $n > \lambda$ or the reverse, and also depending on the priors.

We now estimate the number of miRNA target genes by converting the odds ratio to a probability,

$$\hat{p}(H_2) = \frac{p(H_2)/p(H_1)}{1 + p(H_2)/p(H_1)} \qquad \textbf{[19]}$$

and then summing over all genes

$$N_{est} = \sum \hat{p}(H_2) \qquad \textbf{[20]}$$

We take $P_{rat} = 0.1$, corresponding to roughly midway between the upper and lower rigorous bounds found previously. We try all integer values of δ between 1 and 5.

The results are

delta=	1	2	3	4	5
6	1292	1378	1315	1211	1102
5	1183	1289	1264	1188	1096
4	1109	1180	1163	1103	1023
3	1086	1120	1095	1038	963
2	1068	1073	1037	978	907

The good news is that the estimates are relatively constant both as a function of the assumed δ and as we relax the conservation requirement.

The bad news is that the trend is still decreasing as we relax conservation, where we expect a constant-to-increasing trend.

The really bad news, not evident in the above table, is that the estimates are quite sensitive to the assumed prior P_{rat}. If we take 0.05 instead of 0.1, the estimates fall from $\approx 1,300$ to $\approx 1,000$. If we take 0.2 (doubless too large), they rise to $\approx 1,800$.

When a Bayesian estimate depends sensitively on the prior, it is telling us something: that the "evidence" (in the form of the predicted and actual histograms) really doesn't determine a unique answer. That is also the sense that we get from the wide gap between the rigorous upper and lower bounds, and the appearance of evident systematic errors in the zero-bin method when we relax the conservation requirement.

Summary: Number of miRNA Target Genes

We can say with confidence that the number of target genes is greater than $\approx 1,200$. We find no evidence that it is $\gtrsim 1,500$. However, the only rigorous upper bound that we get is $\approx 5,000$.

Although we found clear evidence that the number of causal target sites was larger in human than in the conserved intersection of all five species, we find no such evidence for the number of target genes. It may well be that essentially all the target genes are conserved among the species, even if there are target sites on those genes that are missing, or added, in one species or another.

Use RefSeq Instead of KnownGenes

Everything up to now has used the (hg-17) KnownGenes set of genes. These include many multiple splicings, many of which have overlapping or identical 3' UTRs. While they are all valid genes (make unique proteins), one might worry that because these genes may come in "groups", some of the statistics above, especially error bars, might be erroneous.

Therefore, we repeat the whole analysis using RefSeq genes, which correspond to a single splicing. Instead of 36,170 genes with 3' UTR lengths > 100, we now have just 21,542 ($\approx 60\%$ as many).

Results corresponding the Results for Excess Sites Method are

```
Cons.  total  Method 1a     Method 1b
Sites  mirs
----   ----   ---------     ---------
6  13198    8584 +-  420     8225 +-  387 (bias subtracted: 52)
5  22136   11161 +-  769    10324 +-  658 (bias subtracted: 62)
4  32505   12961 +- 1216    11532 +-  948 (bias subtracted: 384)
3  45227   14305 +- 1903    12129 +- 1436 (bias subtracted: 825)
```

These values are essentially identical to the previous, even much closer than the indicated error bars. (Of course, they are mostly the same genes, so the errors are not independent.)

Results corresponding to the Model-Free Bounds are:

```
      lower bound    upper bound
6     1142 +-  90    3101 +-  90
5      892 +-  92    3716 +-  97
4      707 +-  93    4190 +- 116
3      570 +- 100    4541 +- 195
```

These seem typically $\approx 10\%$ lower (both upper and lower bounds) than previously, but the errors are also on that order; and these are only bounds, not values. It is not clear whether there are fewer miRNA targets in RefSeq or not, but in any case the difference with KnownGenes is much less than the ratio of the number of genes.

Results corresponding to the Corrected Zero-Bin Method are:

```
    (1)     (2)
6  1207 or 1326 +- 108
5   813 or  984 +- 112
4   492 or  676 +- 119
```

Again, a decrease of $\lesssim 10\%$ from before.

Results corresponding to the Poisson Odds Ratio Method are:

```
delta= 1     2     3     4     5
6  1133  1258  1229  1155  1067
5   991  1127  1137  1092  1025
4   914  1009  1024   994   942
3   883   938   940   909   861
2   860   887   877   843   797
```

Ditto, ditto, $\lesssim 10\%$.

Overall, the summary conclusions from above remain valid, with perhaps a 10% downward correction (much smaller than the 40% decrease in raw number of genes).

Assigning a Probability That a Gene is a miRNA Target

Even though the overall normalization of the Poisson odds ratio method is dependent on the assumed prior, we find that the *relative* probabilities of being a miRNA target are quite stable as we adjust the prior. With this caution about the overall normalization of the probabilities, we can use the Poisson odds ratio method to assign probabilities to individual genes.

The accompanying spreadsheet lists all genes whose indicated probability of being a miRNA target is greater than 10%. At the top of the list, where the indicated probability saturates at values very close to 1, we expect that a large fraction of the listed genes are microRNA targets. At the bottom of the list, we would expect on the order of 1 listed gene in 10 to be a target.

The spreadsheet also lists, for each gene, the three most over-represented conserved miRNA hexamers (relative to the digraph model predictions). In those cases where the gene is indeed a miRNA target, we would expect that the targeting miRNA is one of these three, most likely the first. However, these predictions of specific miRNA families are very noisy and should be taken as having considerable uncertainty.

1. Harris, M. A., Clark, J., Irel, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., *et al.* (2004) *Nucleic Acids Res* **32**, D258-D261.

2. Lewis, B. P., Burge, C. B. & Bartel, D. P. (2005) *Cell* **120**, 15-20.