Human microRNAs target a functionally distinct population of genes with AT-rich 3′ UTRs

Harlan Robins* and William H. Press†‡

*Institute for Advanced Study, Einstein Dr., Princeton, NJ 08540 and
†Los Alamos National Laboratory, Los Alamos, NM 87545

‡To whom correspondence should be addressed; E-mail: wpress@lanl.gov.
Contributed by William H. Press, August 29, 2005.

While investigating microRNA targets, we have found that human genes divide into two roughly equal populations, based on the fraction of A plus T bases in their 3′ untranslated regions (3′ UTRs). Utilizing the Gene Ontology database in an novel way, we find significant functional differences between the two gene populations, with AT-rich genes implicated in transcription and translation processes, GC-rich genes in signal transduction and post-translational protein modification. Better understanding of the background distribution of nucleotides in 3′ UTRs may allow improved prediction of microRNA targeted genes in humans. We predict at least 1200 KnownGene transcripts to be regulated by microRNAs. The large majority of these microRNA targets are in the AT-rich 3′ UTR population. However, notwithstanding this preference for AT-rich targets, microRNA targets are found preferentially to be regulatory genes themselves, including both transcription factors and post-translational modifiers. These results suggest that some processes involving mRNA, of which microRNA regulation may be just one, require AT-richness of 3′ UTRs for functionality. A relationship, not simply one-to-one, between these 3′ UTR populations and large-scale genomic isochores is described.

microRNA | 3′UTR | Gene Ontology | isochore

This preprint version includes Supporting Information. The published version of this paper is at PNAS 102 (43), 15557–15562 (2005) [www.pnas.org/cgi/content/abstract/102/43/15557].

MicroRNAs (miRNAs) are short ($\sim$ 22bp), single-stranded RNA molecules that bind specific messenger RNAs (mRNAs) – their targets – and repress their translation (1, 2). Additionally, new evidence suggests that miRNAs down-regulate message levels as well as protein levels (3, 4, 5). The large majority of both known and predicted target sites on mRNA molecules are within the 3′ untranslated regions (UTRs) (6). As a necessary condition for a target site of a particular miRNA, the mRNA (usually 3′ UTR) is believed to require six continuous nucleotides that form exact Watson-Crick base pairs to positions two through seven of the miRNA, where position one is the first base of the 3′ end of the miRNA (7, 8). Applying both experimental and comparative genomics techniques, a few groups have taken advantage of this hexamer binding condition to predict that a much larger number of human genes are regulated by miRNAs than at first believed, perhaps as many as several thousand (3, 6, 9, 10, 11, 12). However, even with such a large number of regulated genes, six nucleotide binding does not provide enough specificity for a miRNA to find its intended target. It does not seem likely that additional specificity is imparted by partial binding of the miRNA to more than seven positions of the target site in humans (6, 7), although such a mechanism may operate in *C. elegans* and *D. melanogaster* (11).

We show that human miRNAs preferentially target a large, but nevertheless distinct, population of genes whose 3′ UTRs have a high proportion of A and T bases, not just near the miRNA binding site, but globally. Such genes tend also to be AT-rich in the third positions of their codons, where redundancy in the genetic code allows alternative choices of base. Since nearly half of all human genes are in this AT-rich population, the immediately implied gain in specificity is not large. However, our result is supportive of the conjecture that the additional specificity for miRNA binding lies in a global property of AT-rich target mRNAs (different from CG-rich mRNAs) not just adjacent to the target hexamer; an example would be three dimensional conformational properties (13).

As additional evidence that a gene's AT-richness is not merely an artifact, but may be a fundamental aspect of its functioning, we find that some Gene Ontology classification (14) keywords correlate highly with AT-richness; and we will show additional keyword differences, highly statistically significant, between genes that are miRNA targets and other AT-rich genes, meaning that miRNA targets are not just "typical" AT-rich genes, but a functionally distinctive subset thereof.

We have developed a variant of the method used by Lewis *et al.* (6), and similar to Krek *et al.* (9), but using a digraph background model, to predict miRNA targeted genes. The Supporting Tables include a table of predicted probabilities, by gene, of being a miRNA target, along with the set of microRNAs most likely to regulate the gene.

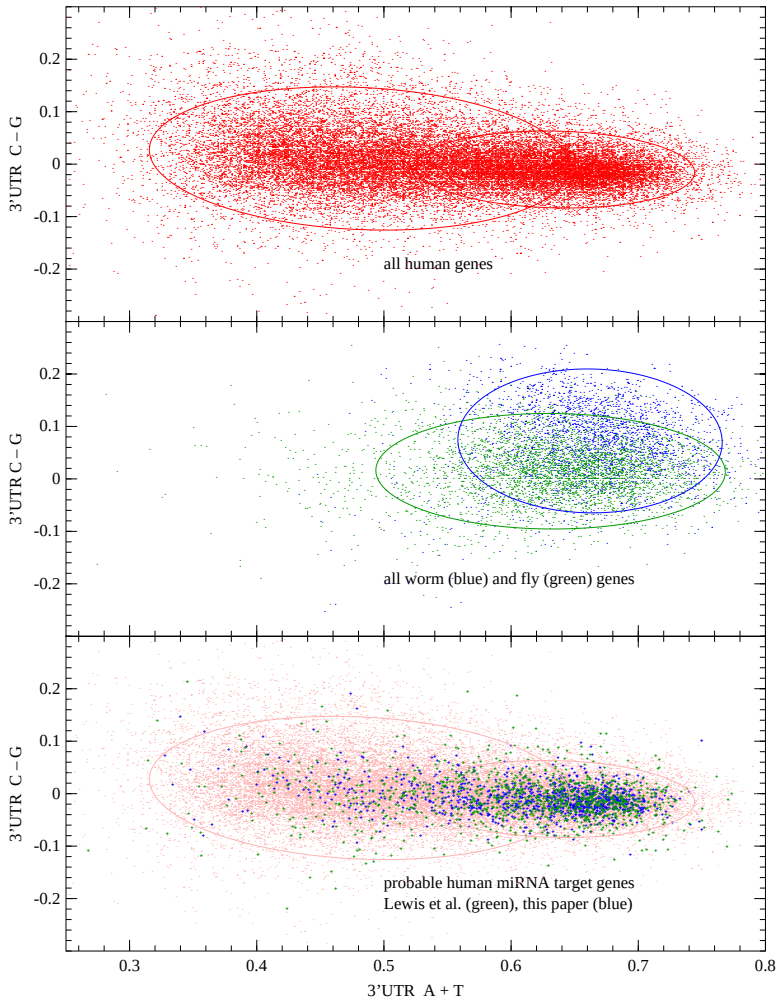

Figure 1: Composition of human 3′ UTRs with $A+T$ on the horizontal axis, $C-G$ on the vertical. Top: All human genes are plotted in red. The red ellipses are 2-σ contours of the maximum likelihood Gaussian mixture model with two components. Center: Invertebrates, including *C. elegans* and *D. melanogaster* do not evidence more than a single population when plotted on the same axes. Bottom: Same as top (light red) with probable microRNA target genes now plotted in green (Lewis *et al.* "high signal-to-noise" set) and blue (probable targets as determined by the methods of this paper). MicroRNA targets lie in the right (AT-rich) component with about 3:1 selectivity.

Composition of 3′ UTRs

If we examine the nucleotide compositions of the $\sim$ 36000 human KnownGene 3′ UTRs whose length is greater than 100 bases (so that their composition is statistically determinable to within a reasonable error), an interesting pattern emerges. If we let A, C, G, and T represent the fraction of each base in a given 3′ UTR, the pattern is best seen by plotting $A+T$ on one axis and $C-G$ on the other, as is shown in Figure 1 (top). (An animation showing all possible axes is at http://www.nr.com/bio/utrtetra.html.) One sees two populations, only partially overlapping, distinguished primarily by their mean in $A+T$ and secondarily by their dispersion in $C-G$. The ellipses in the figure are the 2-σ contours of a two-component Gaussian mixture model blindly fitted to the data (that is, with all parameters unguided by us). (See A.W. Moore's tutorial at http://www.autonlab.org/tutorials/, and also (15) for more on fitting Gaussian mixture models.) Such a model readily assigns, by the Bayes odds ratio method, a probability for each gene that it is in the AT-rich, versus AT-poor, population. For the fits shown in the Figure, and with $x \equiv A+T$, $y \equiv C-G$, the resulting assignment algorithm is

$$\begin{aligned} z_1 &= 41.3\exp(-23.7 + 99.4x - 104.5x^2 \\ &\quad + 21.50y - 36.7xy - 164.5y^2) \\ z_2 &= 118.3\exp(-79.1 + 251.2x - 199.7x^2 \\ &\quad + 40.6y - 88.9xy - 701.y^2) \\ P &= z_2/(z_2+z_1) \end{aligned} \qquad \textbf{[1]}$$

yielding the value P as the probability of being in the AT-rich population. In all the work described here, we carry forward this probability ("soft decision") rather than make a hard assignment. The model places, statistically, about 47% of genes in the AT-rich population, with a mean $A+T$ of about 0.63; 53% are in a CG-rich population with a mean $C+G = 1-(A+T)$ of about 0.53. Additional fitted parameters are given in the Supporting Text.

For the analyses, we have used the full set of KnownGene transcripts. Some of these transcripts refer to different splice forms of the same gene. Since mRNAs regulate at the message level, this is appropriate. However, we have also verified that very similar results are obtained if one uses unique genes from the RefSeq database.

The distribution of 3′ UTRs in A plus T for organisms that are not warm-blooded vertebrates forms a single distinct population (e.g., *C. elegans* and *D. melanogaster* as shown in Figure 1, center). The two- versus one-population phenomenon is related to the existence of isochores (16, 17, 18, 19), about which we will say more below.

Methods

Use of Word Counts in the GO Database

We describe a new way of identifying statistically significant functional differences between two large populations of genes, using the Gene Ontology (GO) database (14). We will then apply the method to AT- versus CG-rich genes, and to probable microRNA targets versus all other genes.

One might think it straightforward to distinguish two large populations of genes by differences in how they are assigned to GO categories. Unfortunately, the "raw" GO data is very noisy for this purpose. Because the hierarchical GO categories are invented and populated with genes by a large community of individual investigators, they are very inhomogeneous, with breadth and depth varying widely according to the taste of the individual contributors. Also, it is not clear how one would assign a quantitative statistical significance to any differences found.

We find that a useful approach to the GO data is to assign to each gene not its GO category list, but rather the unweighted list of all biologically meaningful words (and wordlike phrases) that occur in the descriptive titles of all of the gene's GO categories. For example, we find that it is more meaningful (or at any rate less noisy) to take note of the term "nucleic acid" in one of a gene's GO categories, than to note exactly which GO category that term occurs in. As an additional set of keywords, we also include the HUGO gene name prefixes (e.g., "ZNF*," designating zinc finger genes).

If $\delta(i,j)$ has the value 1 when a word j is thus associated with a gene i, zero otherwise, and if p_i is the probability that gene i belongs to a population of interest, then we can form the probabilistic word counts for each word in that population and its complement,

$$n_{j+} = \sum_i p_i \delta(i,j) \qquad n_{j-} = \sum_i (1-p_i)\delta(i,j) \qquad \textbf{[2]}$$

Similar sums give the variance or expected error of these counts and the normalizing denominators:

$$V_{j+} = \sum_i p_i^2 \delta(i,j) \qquad V_{j-} = \sum_i (1-p_i)^2 \delta(i,j) \qquad \textbf{[3]}$$

$$N_+ = \sum_i p_i \qquad N_- = \sum_i (1-p_i) \qquad \textbf{[4]}$$

Out of these sums we can form a t-value (deviation in standard deviations) and a p-value (two-tail probability) expressing the significance with which the word is associated (or negatively associated) with the set of interest,

$$t_j \equiv t\text{-value} = \frac{n_{j+}/N_+ - n_{j-}/N_-}{\sqrt{V_{j+}/N_+^2 + V_{j-}/N_-^2}} \qquad \textbf{[5]}$$

$$P_j \equiv p\text{-value} = \mathrm{erfc}(|t_j|/\sqrt{2}) \qquad \textbf{[6]}$$

(The straightforward proof is given in the Supporting Text.)

Two criteria must be met before a difference between two gene populations can be considered as substantiated by this method. First, there must be a set of at least several words for which the p-values, as calculated above, are highly significant (less than 10^{-4}, say). This is a necessary Bonferroni constraint because the number of hypotheses – words – is large. Second, only slightly less objectively, there must be a thematic coherency among the highly significant words that makes sense biologically. This is necessary because one can readily imagine differences that, while statistically significant, are biologically uninteresting. For example, it would not be surprising to distinguish large populations of genes entered into the database by a single research group, simply by ideosyncracies in their use of nonspecific words (e.g., "process", "activity", "function").

Digraph Probability Model

Using the assumption that functional regulatory binding sites are likely to be conserved, we look at conserved hexamers in 3′ UTRs from the multiple alignment of human, mouse, rat, dog, and chicken (20) similarly to Lewis *et al.* (6). The difficult part is determining the background rate of (non-causal) conserved hexamers. It seems unwise to use a background model that is the same for every gene, at the very least because we have identified two different populations of genes. Instead, to capture the background rate at which any given hexamer should occur, we use a digraph model that is specific to each gene. This will account not only for the bias implied by variable A plus T content, but also for the known underrepresentation of CpG in the human genome and any other digraphic peculiarities of a given gene.

Ideally, one would model just the conserved regions in each gene. Unfortunately, the total conserved lengths in each gene are not enough to do this. We therefore make the assumption (or approximation) that conservation probability and digraph probability are independent, and we construct the digraph model of each gene from its entire (human) sequence.

Suppose a hexamer is $abcdef$. Then, to digraph order, we can write the probability relations,

$$\begin{aligned} p(abcdef) &= p(a)p(b|a)p(c|b)p(d|c)p(e|d)p(f|e) \\ &= p(a)\frac{p(ab)}{p(a)}\frac{p(bc)}{p(b)}\frac{p(cd)}{p(d)}\frac{p(de)}{p(d)}\frac{p(ef)}{p(e)} \\ &= \frac{p(ab)p(bc)p(cd)p(de)p(ef)}{p(b)p(c)p(d)p(e)} \end{aligned} \qquad \textbf{[7]}$$

Because equation **[7]** involves the product of many terms, it is convenient to work with log-probabilities, so in abbreviated notation we have

$$\text{logprob}(abcdef) = \sum_{xy \in abcdef} \text{logprob}(xy) - \sum_{x \in bcde} \text{logprob}(x) \qquad \textbf{[8]}$$

The individual terms $p(xy)$ or $\text{logprob}(xy)$ are estimated by counting the number of times n that the digraph xy occurs in N opportunities. It is not a good idea, however, to use an estimate like $\log(n/N)$ for the log-probability, since this is divergent for $n = 0$, and biased for small n.

Estimating Log-Probabilities with Small Number Counts

We proceed by writing a Bayesian estimate for the probability of a specific value of probability, call it p_s, *given* the observed values n and N. (Note that, following usual statistical practice, commas are omitted in the following equations.) We include the possibility of having other information y associated with each gene, for example whether it is in an AT-rich or CG-rich population. Bayes theorem and elementary manipulations give

$$\begin{aligned} p(p_s|nN) &= \sum_y p(p_s y|nN) \\ &= \sum_y p(p_s|nNy)p(y|nN) \\ &= \sum_y \frac{p(nN|p_s y)p(p_s|y)}{\int dp_s p(nN|p_s y)p(p_s|y)} \frac{p(y)}{\sum_y p(y)} \end{aligned} \qquad \textbf{[9]}$$

We have replaced $p(y|nN)$ by $p(y)$ since the other information is assumed not to depend directly on nN. Using a binomial probability model for nN, and a binomial conjugate prior for $p(p_s|y)$, we get

$$\begin{aligned} p(p_s|nN) &= \sum_y \frac{p_s^n(1-p_s)^{N-n} p_s^{a_y}(1-p_s)^{b_y}}{\int dp_s p_s^n(1-p_s)^{N-n} p_s^{a_y}(1-p_s)^{b_y}} \frac{p(y)}{\sum_y p(y)} \\ &= \sum_y \frac{p_s^n(1-p_s)^{N-n} p_s^{a_y}(1-p_s)^{b_y}}{B(n+a_y+1, N-n+b_y+1)} \frac{p(y)}{\sum_y p(y)} \end{aligned} \qquad \textbf{[10]}$$

Here B denotes the beta function. In the case that y varies over $\{AT\text{-rich}, CG\text{-rich}\}$, $p(y)$ is given by the Gaussian mixture model previously discussed. Otherwise, one can simplify by assuming a single population $y = 0$ and deleting all references to $p(y)$.

The constants a_y and b_y parameterize the (conjugate) prior on p_s. Although we had initial expectations that the use of good priors could have a

beneficial effect, in practice one obtains as good or better results by taking a noninformative prior like $a_y = b_y = 1$, or any small constant.

Note that (doing an integral) we have the expectation value,

$$E(p_s) = \int_0^1 p_s \, p(p_s|nN) \, dp_s = \sum_y \frac{n + a_y + 1}{N + a_y + b_y + 2} \, p(y) \qquad \textbf{[11]}$$

and, for the case when log-probabilities are needed,

$$\begin{aligned} E(\log p_s) &= \int_0^1 \log p_s \, p(p_s|nN) \, dp_s \\ &= \sum_y [H(n + a_y) - H(N + a_y + b_y + 1)] \, p(y) \end{aligned} \qquad \textbf{[12]}$$

where $H(n)$ is the harmonic sum

$$H(n) = \sum_{k=1}^{n} \frac{1}{k} = \gamma + \psi_0(n + 1) \qquad \textbf{[13]}$$

Here the second form is valid when n is not an integer, γ is the Euler-Mascheroni constant, and ψ_0 is the digamma function. The harmonic sums play the role of logarithms, but now properly corrected for the possibility of small numbers of counts. Thus equation **[12]** is asymptotically $\approx \log(n/N)$ as we might expect, but it remains regular as n and/or N go to zero. We recommend the use of equations **[11]** and **[12]**, as appropriate, whenever small-count data is being analyzed.

Identifying microRNA Target Genes

The digraph model and the observed number of conserved sites gives, for each gene, the expected number of conserved microRNA binding hexamers that should occur by chance and an error estimate (as described in the Supporting Text). We can compare this to the number actually observed and thus assign a probability that any excess is causal, which we take to be the probability that the gene is an actual microRNA target. Our methodology for this is not conceptually different from Lewis *et al.* (6) and is detailed in the Supporting Text. Of interest here, however, is a novel method that we have used to get model-free bounds for the total number of targeted genes.

Consider two histograms, "predicted" and "observed", each giving the number of genes that contain i conserved microRNA binding sites. Each histogram has the same total number of genes. The idea is that "observed" is obtained from "predicted" by pushing some genes to the right in the histogram, that is, by adding – never subtracting – some causal conserved binding sites to

the chance ones in that gene. Note that we are not using the correspondence gene-by-gene, since it is very noisy, but only the resulting histograms, which, because the number of genes is large, have good signal-to-noise.

Can we say anything about how many genes have been pushed to the right, without knowing anything about the distribution of how far each gene was pushed? Yes. In fact, we can get both lower and upper bounds.

Let the numbers in bin i be m_i for "predicted", and n_i for "observed", $i = 0, 1, 2, \ldots$. Since the histograms have the same area (number of genes), the sum of the positive binwise differences must equal the sum of negative binwise differences. That is,

$$\sum_i \max(0, n_i - m_i) = \sum_i \max(0, m_i - n_i) \qquad \textbf{[14]}$$

The way to move the smallest number of genes is to take them strictly from bins with $m_i > n_i$ and move them strictly to bins with $n_i > m_i$. If one does this starting from the right, then one can always achieve this by moving genes in the positive direction. A lower bound on the number of target genes is thus

$$N_{min} = \sum_i \max(0, n_i - m_i) \qquad \textbf{[15]}$$

One might at first think that the upper bound is just the number of extra counts in "observed", spreading them out maximally with one new count per gene. This would give

$$N_{max} = \sum_i i(n_i - m_i) \qquad \text{(Wrong!)} \qquad \textbf{[16]}$$

The problem is that one can't always do this construction moving genes strictly to the right. The actual bound is often substantially lower and thus more meaningful.

The bound is achieved by working from the right and building up the desired n_i distribution taking genes from the closest bin of m_i that has any left to donate. That way, one never "wastes" a possible gene move by leaving a gene in place that could otherwise have been moved. (This is a little bit like Chinese checkers, but where one wants to *avoid* jumping one's marbles.) An explicit formula for the result is,

$$N_{max} = \sum_{i=0}^{\infty} \min\left(m_i, \sum_{j=i+1}^{\infty} n_i - m_i\right) \qquad \textbf{[17]}$$

In fact it is easy to show that equation **[16]** is obtained if the first argument in the min is never used, that is, if one always has enough genes to move at each stage.

Table 1: GO words most associated with AT-rich 3′ UTR genes

word or phrase	t-value	p-value	n_{j+}	n_{j-}
`nucleic acid`	8.75	0.000000	2297	1789
`nucleus`	7.11	0.000000	1722	1365
`transition metal`	6.80	0.000000	1095	824
`zinc`	6.65	0.000000	998	746
`bound`	5.99	0.000000	2398	2042
`ZNF*`	5.87	0.000000	119	49
`RNA`	5.53	0.000000	613	448
`organelle`	5.30	0.000000	2489	2169
`cellular component`	4.63	0.000004	3244	2927
`binding`	4.45	0.000009	4405	4054
`mRNA`	4.25	0.000022	102	53
`metal`	4.11	0.000039	1631	1429
`cycle`	4.07	0.000046	394	296
`DNA`	3.99	0.000067	1324	1149
`nucleobase`	3.71	0.000205	1468	1297

To give a sense of how much better equation [**17**] is than equation [**16**]: For a typical histogram in this study, equation [**17**] yields an upper bound of 3650 (genes), while equation [**16**] would yield a much less restrictive bound of 8400. Equation [**15**] gives a lower bound of 1260. (In the results section, below, we give values that include an additional allowance for statistical error, as described in the Supporting Text.)

Results

GO Database Word Counts

Table 1 lists the 15 top words (or wordlike phrases) that are positively associated with the AT-rich 3′ UTR population of genes, while Table 2 is the corresponding list that is positively associated with the CG-rich 3′ UTR population (that is, negatively associated with the AT-rich population). As shown by the listed t- and p-values, all of the associations are highly significant. Note from the values of n_{j+} and n_{j-} (the probabilistic word counts), however, that the word frequencies differ by at most $\sim 25\%$ in the two populations. Virtually all biologically meaningful words occur, to a greater or lesser extent, in both populations. However, having large numbers of genes allows us to extract signal with high significance even from these modest differences.

It is striking that each of the two lists evidence a clear thematic coherency,

Table 2: GO words most associated with CG-rich 3′ UTR genes

word or phrase	t-value	p-value	n_{j+}	n_{j-}
`receptor`	−5.43	0.000000	852	1085
`signal transduction`	−5.16	0.000000	968	1204
`signaling cascade`	−5.13	0.000000	349	494
`transducer`	−4.88	0.000001	880	1093
`communication`	−4.80	0.000002	1172	1413
`signal`	−4.56	0.000005	902	1102
`transmembrane`	−4.37	0.000012	381	506
`filament`	−4.31	0.000016	86	150
`cell`	−3.83	0.000129	1840	2081
`channel`	−3.77	0.000159	151	222
`immune`	−3.62	0.000291	217	296
`pore`	−3.39	0.000708	162	227
`defense`	−3.30	0.000961	237	311
`structural`	−3.22	0.001281	241	314
`development`	−3.21	0.001300	518	625

and that the two lists are thematically very different. Genes with AT-rich 3′ UTRs are preferentially associated with transcription and translation events, especially nucleic-acid and nucleic-acid binding processes (e.g., zinc finger motifs). These functions are evolutionary old. By contrast, the high GC population is associated with functions coupled to sensing and responding to the external environment. These include signal-transduction pathways and membrane transport. A unifying theme of the high-GC population is that its functions tend towards post-translational protein modification and signaling interactions, as opposed to transcriptional regulation.

Although the evidence is only indirect, the strong association of AT-rich 3′ UTRs with genes that are implicated in RNA and mRNA processing supports the same conjecture as for miRNA target specificity. That is, some aspect of AT-richness in the 3′ UTR is necessary for at least some processes involving mRNA, of which regulation by miRNAs may be just one.

microRNA Target Genes

By the method of equations **[15]** and **[17]**, we find among the ~ 36000 KnownGenes a solid lower bound of 1200 miRNA targets, and an upper bound of about 5000. This method, however, does not identify which specific genes are likely to be targets. To accomplish this, and also to get a most probable to-

tal count of targets (between the two bounds), we use a Poisson odds-ratio method, as described in the Supporting Text. However, this most probable value is model dependent and rather less well determined. We get about 1400 ± 150, but we consider this value as likely subject to uncontrolled systematic errors. Lewis *et al.* have identified a set of "high signal-to-noise" likely miRNA target genes (6). Although there is significant overlap, our set of most probable target genes is different in detail from this set. We believe that our use of a digraphic probability model, specific to each gene examined, ought to give superior predictions. However, a final verdict on this claim must await experimental evidence. (The Supporting Tables includes a table giving our predictions by gene.)

Figure 1 (bottom) is identical to Figure 1 (top), with the Lewis *et al.* likely targets now plotted in green. The association with the AT-rich population, in both $A + T$ mean and $C - G$ dispersion, is immediately apparent, and easy to substantiate statistically ($p < 10^{-10}$). Genes which we predict to be microRNA targets with $> 50\%$ probability are plotted in blue in the Figure. Using these probabilities, we can substantiate that $\sim 75\%$ of miRNA target genes are in the AT-rich population, about a 3:1 selectivity. However, there is no trend towards fewer targets in the CG-rich population as miRNA target probability goes to 1, indicating that the $\sim 25\%$ minority of miRNA targets that are CG-rich are in fact genuine, though atypical.

We also find weak, but statistically significant, associations between the population of genes with AT-rich $3'$ UTRs and those genes identified by the microarray analysis of Lim *et al.* as being targets of two specific miRNAs, miR-1 ($N = 82$, $p < 0.001$) and miR-124 ($N = 152$, $p < 0.01$).

We can perform the same GO keyword analysis as before on the population of (probabilistically known) miRNA targets. Knowing that miRNA targets lie strongly preferentially in the AT-rich population, we might expect such an analysis to yield an associated word list much like Table 1. The actual result, shown in Table 3, is unexpected and much more interesting. Comparing the two tables, it is striking that the multiple words that associated AT-rich genes with nucleic acid processes are completely absent from the miRNA preferential word list. Instead, the list is dominated by the word "regulation" and its closely related concepts. This is statistically strong evidence that miRNA targets are themselves preferentially (though by no means exclusively) regulators.

What is also surprising, in view of the results of Tables 1 and 2, is that miRNA target preferences include both transcription factors and also post-translational regulators, the latter evidenced in words such as "protein modification", "phosphorylation", "kinase", "signaling cascade", and so forth. The dominant theme of regulation is also seen in a set of words including and related to "development", including "morphogenesis" and "neurogenesis".

In other words, within the population of genes with AT-rich $3'$ UTRs that

Table 3: GO words most associated with probable microRNA target genes

word or phrase	t-value	p-value	n_{j+}	n_{j-}
transcription regulator	5.86	0.000000	134	1114
transcription factor	5.86	0.000000	129	1068
regulation	5.56	0.000000	315	3215
regulation of transcription	5.36	0.000000	205	1970
development	4.69	0.000003	140	1326
protein modification	4.65	0.000003	192	1897
serine/threonine kinase	4.42	0.000010	68	521
nucleus	4.42	0.000010	319	3477
phosphorylation	4.30	0.000017	90	766
signal transduction	4.09	0.000043	231	2449
promoter	4.07	0.000048	46	347
phosphate	4.04	0.000052	133	1286
signaling cascade	4.02	0.000058	99	908
morphogenesis	3.96	0.000075	66	567
kinase	3.88	0.000106	133	1311
phosphotransferase	3.88	0.000106	105	977
DNA	3.82	0.000132	251	2752
cell	3.71	0.000155	30	205
intracellular	3.72	0.000202	557	6573
neurogenesis	3.71	0.000205	30	205

miRNAs preferentially target, miRNAs tend to regulate other regulatory genes, even when the regulated processes are post-translational and uncharacteristic of the AT-rich population generally. In particular, keywords like "signaling cascade" and "signal transduction" are among those strongly *positively* associated with miRNA targets, even though they are strongly *negatively* associated with AT-rich genes generally.

Since a smaller fraction ($\sim 25\%$) of miRNA targets are genes with GC-rich, rather than AT-rich, 3′ UTRs, one may wonder whether those miRNA targets associated with post-translational processes are associated with that fraction. The answer is no: keyword analysis of miRNA targets that are AT-rich (the majority), versus those that are CG-rich (the minority) show no significant differences. (By way of example, "protein modification" happens paradoxically to be the top word associated with AT-rich miRNA targets, while 3 of the 5 top words associated with CG-rich miRNA targets refer to transcription.)

Discussion

So-called isochores (16, 17, 18, 19, 21) are long, megabase-scale regions of CG-richness that are found in the genomes of warm-blooded vertebrates, including human, and absent in lower organisms. Isochores span intron, exon, and intergene regions indiscriminately, as distinct from the comparatively tiny ($\sim$ 1000 base) scale of the individual 3′ UTRs discussed here. Although we defer a detailed discussion of the relationship between these very different scaled phenomena to another paper, we need here to remark on the obvious question as to whether our two populations of genes (characterized only by their 3′ UTRs) are located in CG-rich isochores, versus the complementary AT-rich isochores, in the genome. In other words, have we simply rediscovered a previously-known phenomenon?

Interestingly, the answer is both yes and no. Analysis shows that, with a high degree of selectivity, AT-rich isochores do contain only genes with AT-rich 3′ UTRs. CG-rich isochores, however, contain an apparently random mixture of genes with CG-rich and AT-rich 3′ UTRs. Although this result sheds no new light, per se, on the (evolutionarily recent) origin of isochores, its relevance to this paper is that it does add support to the idea that an AT-rich 3′ UTR is necessary for some functionally distinct subset of genes. Such genes would naturally resist the evolutionary trend that formed the CG-isochores (whatever it may have been (21)), resulting in the mixture of genes that is seen in CG-rich isochores.

Given the observation that there are genes with AT-rich 3′ UTRs in both AT-rich and CG-rich isochores, it is also natural to ask whether one or the other set is dominantly responsible for the strong functional signal demonstrated in Table 1. The answer is that virtually *all* of the functional signal comes from

those AT-rich 3′ UTR genes in CG-rich isochores. If AT-richness of the 3′ UTR is indeed functionally necessary for some genes, the most likely candidates for experimental verification should therefore be sought in CG-rich isochores.

More speculatively, the evidence seems to indicate that, with respect to evolutionary pressure towards CG-richness, AT isochores were "never challenged", as opposed to "challenged and resisted". That is, AT isochores appear to include populations of AT-rich genes with functionalities that, had they been in a CG isochore, could have become CG-rich without difficulty (Table 2). Conversely, CG isochores include a functionally distinct population of AT-rich genes (Table 1) that seem to have strongly resisted such conversion.

Acknowledgments

The authors thank Arnold Levine, Gerald Joyce, Curt Callan, Richard Padgett, David Haussler, and Hagar Barak for reading various drafts and making numerous useful suggestions. John Kern provided important statistical insight. This work was supported in part by the Shelby White and Leon Levy Initiatives Fund.

References

1. Bartel, D. P. (2004) *Cell* **116**, 281–297.

2. Ambros, V. (2004) *Nature* **431**, 350–355.

3. Lim, L. P., Lau, N. C., Garrett-Engele, P., Grimson, A., Schelter, J. M., Castle, J., Bartel, D. P., Linsley, P. S., & Johnson, J. M. (2005) *Nature* **433**, 769–773.

4. Liu, J., Valencia-Sanchez, M. A., Hannon, G. J., & Parker, R. (2005) *Nat Cell Biol* **7**, 719–723.

5. Sen, G. L. & Blau, H. M. (2005) *Nat Cell Biol* **7**, 633–636.

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

7. Doench, J. G. & Sharp, P. A. (2004) *Genes Dev* **18**, 504–511.

8. Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P., & Burge, C. B. (2003) *Cell* **115**, 787–798.

9. Krek, A., Grun, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., da Piedade, I., Gunsalus, K. C., Stoffel, M., & Rajewsky, N. (2005) *Nat Genet* **37**, 495–500.

10. John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., & Marks, D. S. (2004) *PLoS Biol* **2**, 1862–1879.

11. Brennecke, J., Stark, A., Russell, R. B., & Cohen, S. M. (2005) *PLoS* **3**, 404–418.

12. Grun, D., Wang, Y., Langenberger, D., Gunsalus, K. C., & Rajewsky, N. (2005) *PLoS Comp Biol* **1**, 51–66.

13. Robins, H., Li, Y., & Padgett, R. W. (2005) *Proc Natl Acad Sci USA* **102**, 4006–4009.

14. Harris, M. A., Clark, J., Irel, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., & White, R. (2004) *Nucleic Acids Res* **32**, D258–D261.

15. McLachlan, G. & Peel, D. (2000) *Finite Mixture Models*. (Wiley).

16. Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., & Rodier, F. (1985) *Science* **228**, 953–958.

17. Bernardi, G. (2000) *Gene* **241**, 3–17.

18. Cohen, N., Dagan, T., Stone, L., & Graur, D. (2005) *Mol Biol Evol* **22**, 1260–1272.

19. Vinogradov, A. E. (2003) *Nucleic Acids Res* **31**, 5212–20.

20. Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y. T., Roskin, K. M., Schwartz, M., Sugnet, C. W., Thomas, D. J., Weber, R. J., Haussler, D., & Kent, W. J. (2003) *Nucleic Acids Res* **31**, 51–54.

21. Eyre-Walker, A. & Hurst, L. D. (2001) *Nat Rev Gen* **2**, 549–555.

Supporting Animation

Link: `http://www.nr.com/bio/utrtetra.html`

Caption: The nucleotide content of the 3′ untranslated regions of a representative random sample of human genes is here plotted. Each point in the interior of the tetrahedron represents one gene. Each vertex of the tetrahedron represents a content of 100% of the respectively labeled nucleotide; interior points are linear combinations of all four nucleotides, with fractions summing to unity.

Supporting Tables (Spreadsheets)

Link to Table 4: `http://www.nr.com/bio/PNAS20051025Table4.xls`

Link to Table 5: `http://www.nr.com/bio/PNAS20051025Table5.xls`

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 $\boldsymbol{\mu}$ are two-dimensional vectors. The square of the conventional standard deviation becomes now a 2×2 covariance matrix $\boldsymbol{\Sigma}$. If $N(\mathbf{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate Gaussian density,

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

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

$$p_{nk} \equiv P(k|n) = \frac{N(\mathbf{x}_n \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) P(k)}{P(\mathbf{x}_n)} \qquad [2]$$

where

$$P(\mathbf{x}_n) \equiv \sum_k N(\mathbf{x}_n \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) P(k). \qquad [3]$$

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

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

and

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

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 $\boldsymbol{\mu}$ is given by

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

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

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

then

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

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

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

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.)

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

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 $\mathrm{Var}(n_{j-})$, and then $\mathrm{Var}(n_{j+}/N_+ - n_{j-}/N_-)$.

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 (Supporting Table 4). 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 "non-special" (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.

Supporting Table 5 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.