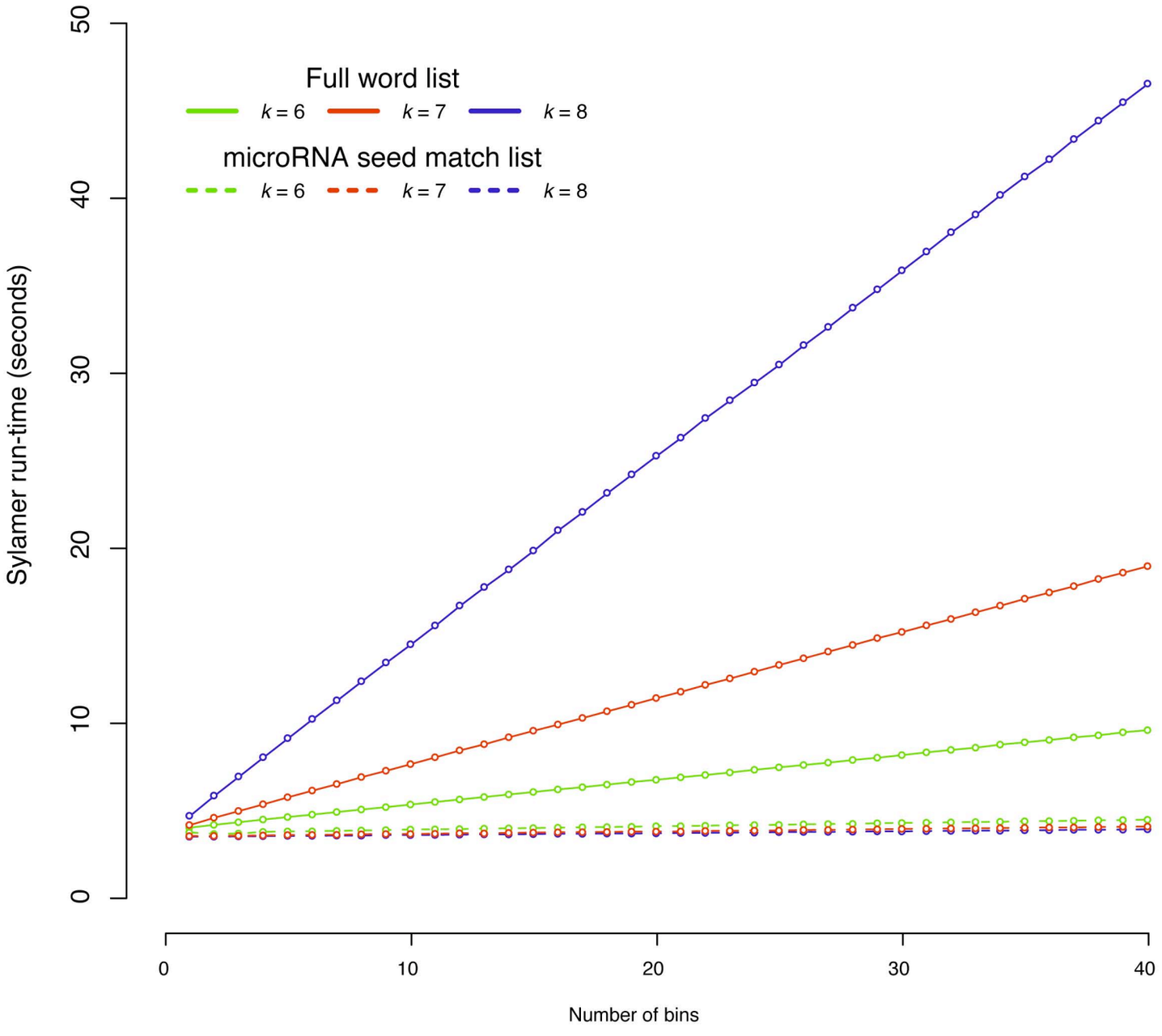
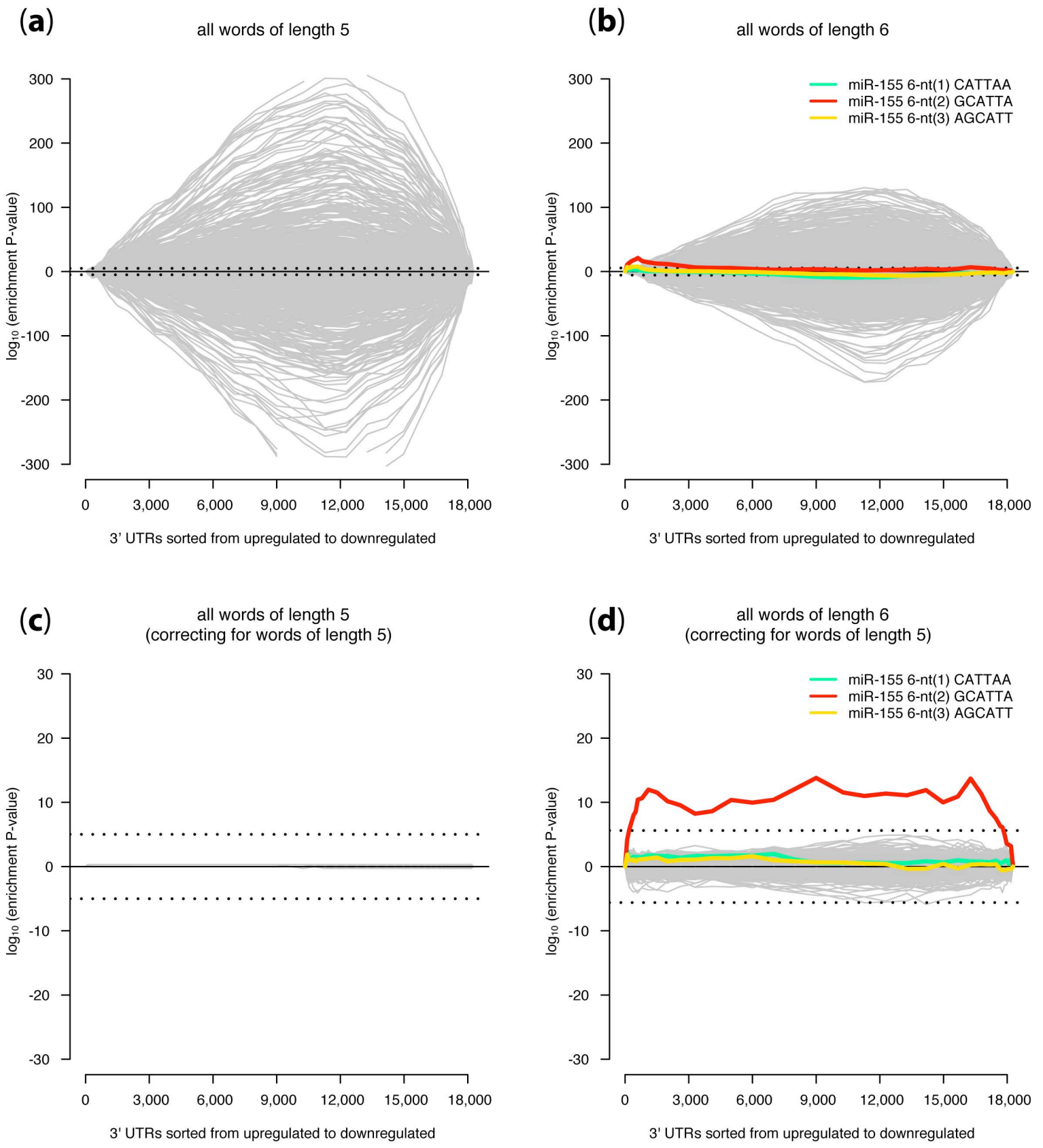
Supplementary Figure 1



Supplementary Figure 1: Sylamer Speed Analysis

Sylamer run-time for words of length 6, 7 and 8nt, plotted as a function of the number of bins, applied to all human 3'UTRs from Ensembl (release 49). The total number of non-masked bases in these sequences was 42,769,472. In all cases Markov correction is used on words of length 4, computed on the fly. The full-size lists of all possible words contained 4,096, 16,384, and 65,536 words, respectively. Each measurement was taken as the average time of three Sylamer runs applied to different rankings, resulting from randomly shuffling the transcript IDs three times. Run-time depends nearly linearly both on the number of words and the number of windows. For increasing word size, the run-time averaged over the size of the word list drops slightly, caused by larger words being less frequent. Less than 50MB of RAM was required for the analysis. Hardware used was: x64 2.0 GHz dual-core Opteron.

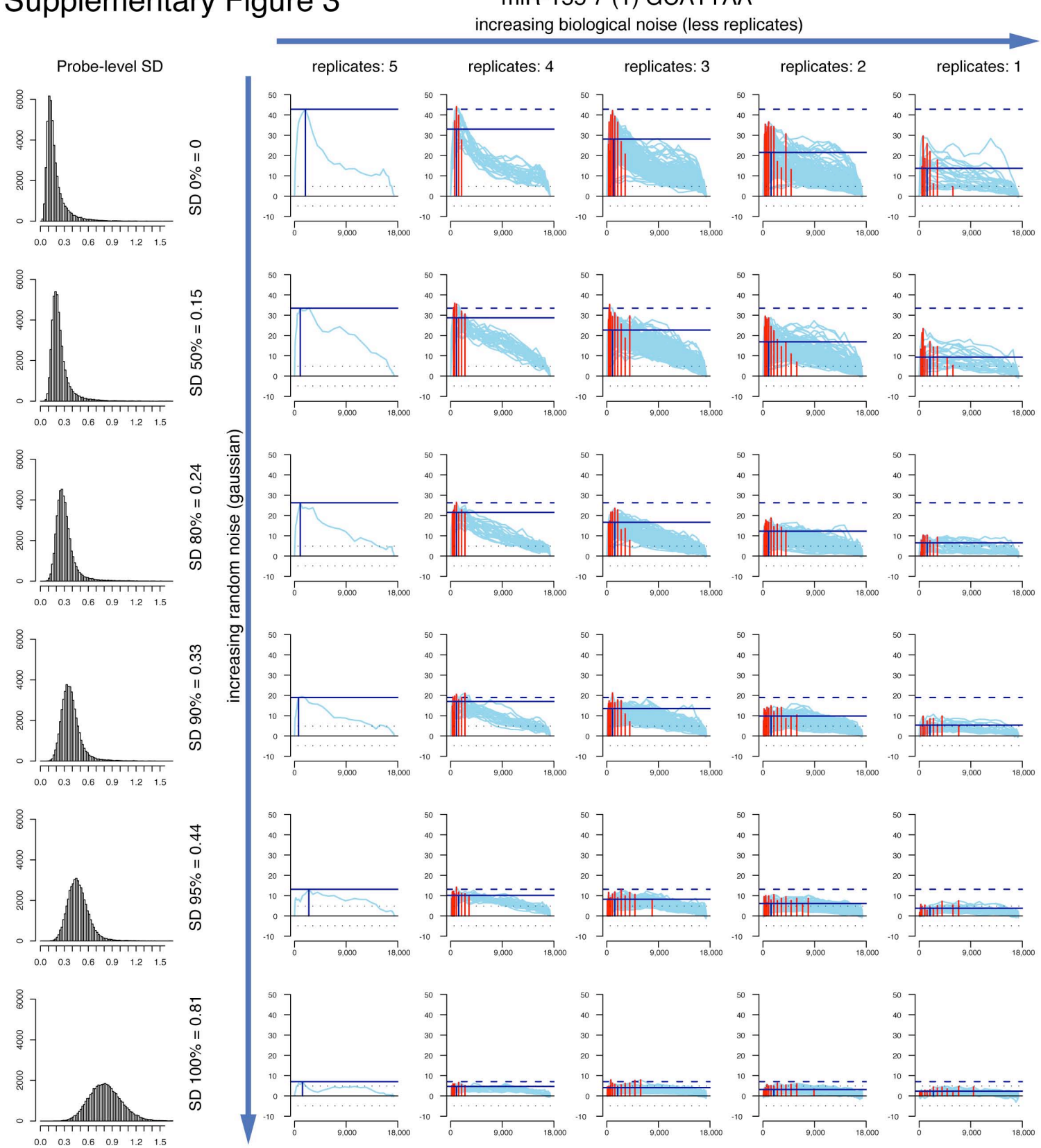
Supplementary Figure 2



Supplementary Figure 2: Compositional Biases

Sylamer enrichment plots showing large composition biases and the effect of Markov correction. (**a**) Words of length 5 are highly skewed in this ranked genelist. (**b**) This effect is still evident for words of length 6, masking out the expected miRNA seed enrichment. Markov correction using words of length 5 words (**c**) greatly ameliorates these biases allowing words of length 6 (**d**) to be correctly analyzed (other details as in Figure 1).

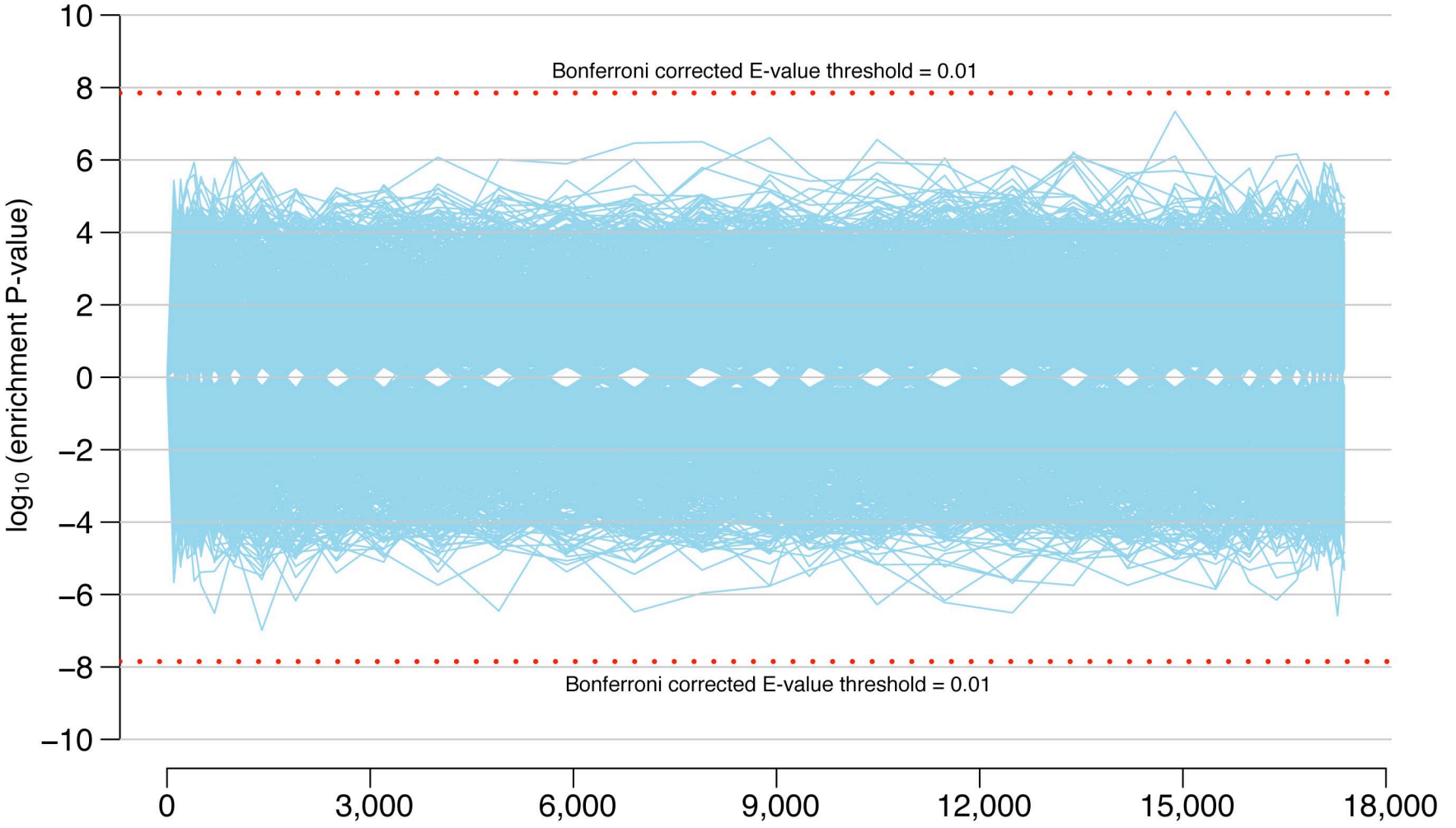
Supplementary Figure 3



Supplementary Figure 3. Effect of biological variability and random noise

This figure demonstrates the effect on the Sylamer results if noise is added by two means. First, from left to right the effect is shown if fewer replicates are used. Then, from top to bottom the effect is shown of adding increasing levels of Gaussian noise to the $\log_2$ normalized expression values. In the first column the histogram of resulting standard deviations of all the probesets, after addition of the corresponding noise level, is shown. In the first case, all five replicates are used and no noise is added. In the bottom right corner all 25 possible combinations of using a single replicate in the two conditions, with Gaussian noise added with a standard deviation such that it exceeds the standard deviation for 99% of the probesets. The full range of Gaussian levels used from top to bottom correspond to respectively the 0, 50, 80, 90, 95, and 99 percentiles of the levels of variation found in the data. Going from left to right, we use all possible combinations of a fixed number of replicates for each of the two conditions, and compute Sylamer results for all of the resulting rankings. For a discussion of these results, refer to the accompanying text.

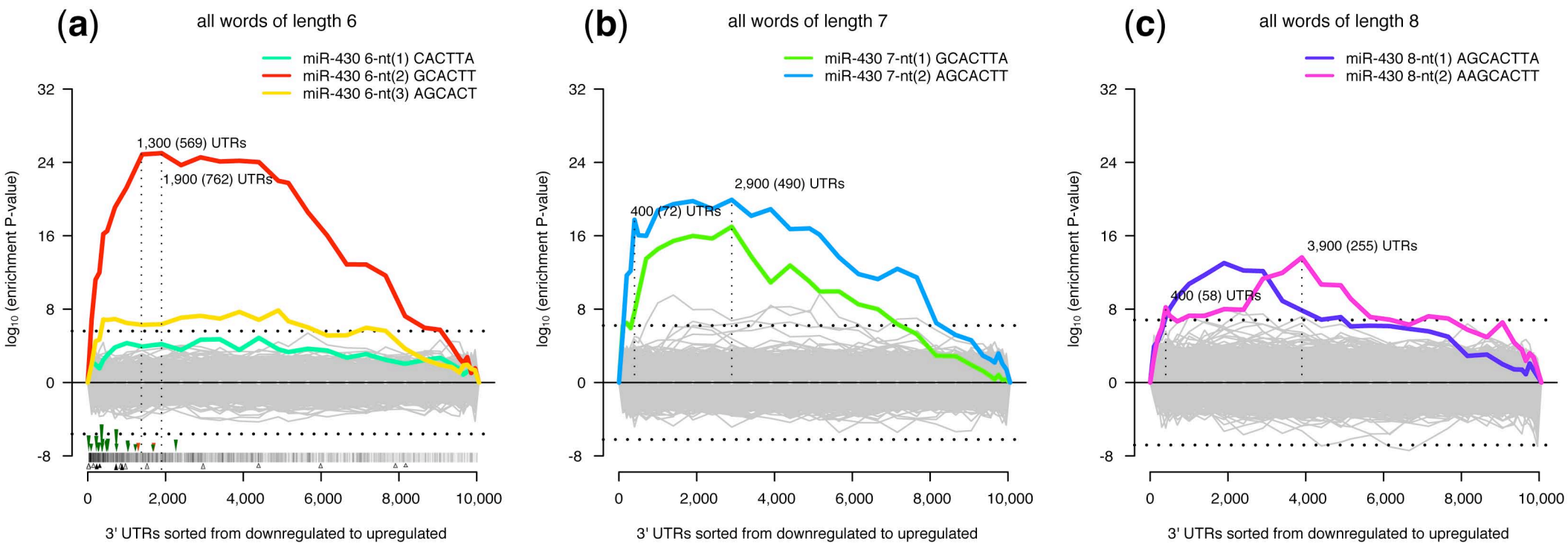
Supplementary Figure 4



Supplementary Figure 4. Permuted gene lists

In order to test the validity of the hypergeometric P-values calculated by Sylamer we randomly shuffled a gene list 1,000 times. This gene list contains the 17,384 3′ UTR sequences that we mapped to probes from the Affymetrix Mouse Genome 430 2.0 Array. To simulate a real case scenario, we ran Sylamer with the recommended parameters, but restricting the word space to the 709 miRNA 7mer seeds (taken from miRBase (ref 13) release 12). This yields a total of 709,000 distinct combinations; applying a Bonferroni correction to a P-value of 0.01 yields a modified significance cutoff of 1.41×10^{-08}. This threshold is represented in the figure as a dotted red line. It can be seen that in no case does a word become more significant than this threshold.

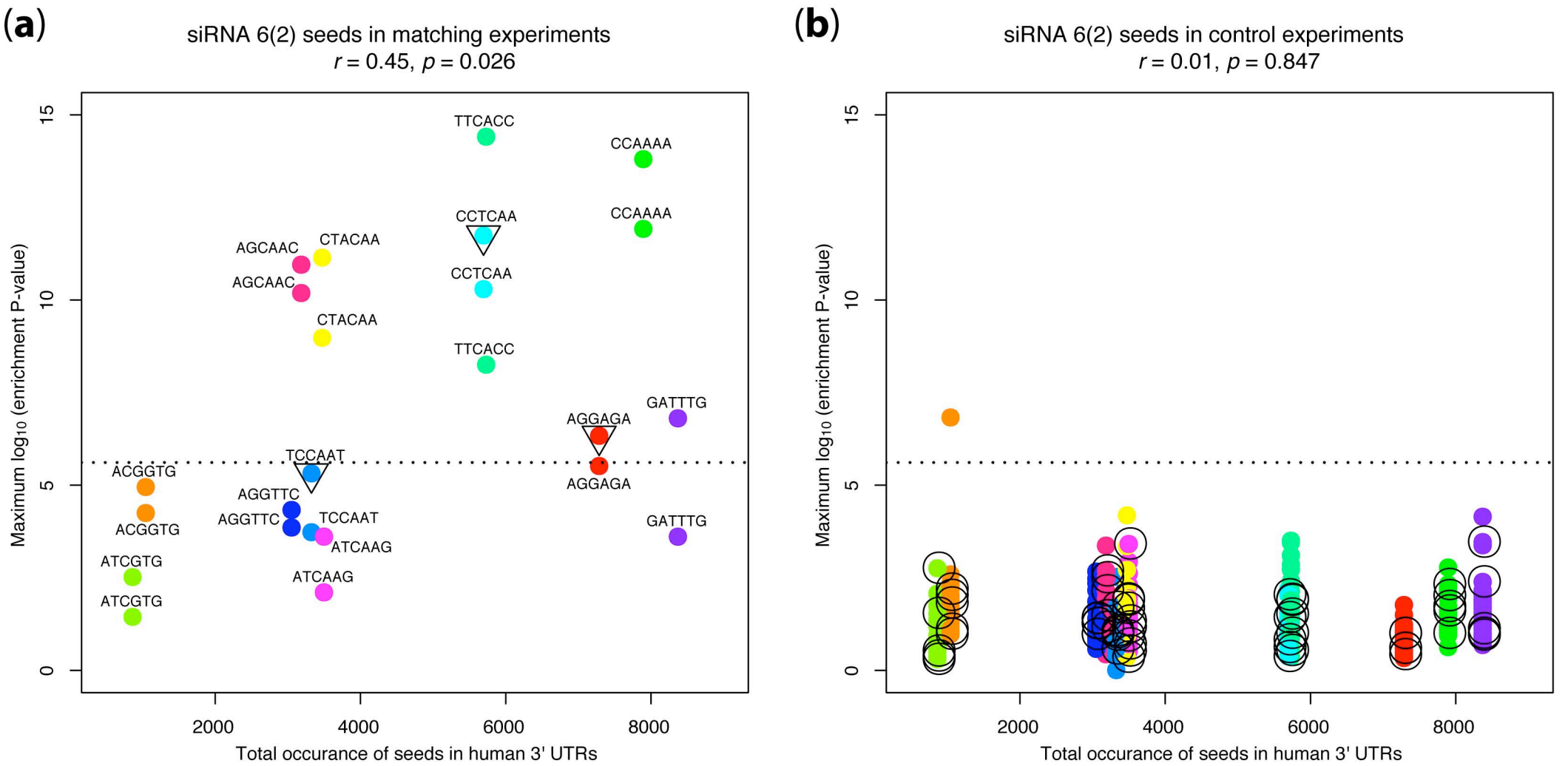
Supplementary Figure 5



Supplementary Figure 5: Zebrafish miR-430 induction experiment

Sylamer enrichment landscape plots for the miR-430 Zebrafish injection experiment. Plots are shown for all (**a**) 6nt, (**b**) 7nt and (**c**) 8nt words. In each case, the *x*-axis represents the sorted genelist from downregulated (left) to upregulated (right). Grey lines show the profiles of words unrelated to the seed region of miR-430, while colored lines represent the words that are complementary to the seed-region of miR-430 (other details as in Fig.1). In addition this figure illustrates the positions of previously validated targets within the ranked genelist. Green triangles represent previously validated targets with 6mer seed matches. Red triangles indicate targets that validated but which did not possess a 6mer match. Triangles below represent those genes not showing reporter assay activity. In all cases size of the triangle reflects activity level from the reporter assay. The horizontal bar shows a vector representation of the genelist showing the frequency (white = low, black = high) of 6mer matches across the binned genelist.

Supplementary Figure 6



Supplementary Figure 6: siRNA seed analysis

Effect of total number of seed matches in human 3' UTRs on Sylamer enrichment. (**a**) The maximum enrichment score of each siRNA 6-nt seed was obtained for both replicate transfection experiments. A small but significant positive correlation is observed for the total occurrence count of the seed matches and this score. Each siRNA is represented by a different color and the enriched word is shown. For three cases siRNAs were transfected at half concentration (triangles) but this does not show an observable effect. (**b**) Control Sylamer runs show that there is no correlation between the total occurrence count of the seed matches and the maximum enrichment scores obtained in all the experiments in which the corresponding siRNA was not transfected. Real control experiments (mock transfections) are highlighted with black circles. The dotted line represents an E-value significance threshold of 0.01.

Supplementary Discussion

Interpreting the landscape plot

The curves generated by Sylamer in the landscape plot describe, for each word, the enrichment or depletion for that word across the ranked gene list. It is important to note that Sylamer does not attempt to explain the expression data in terms of these biases, and does not assume any particular quantitative relationship between the two. It simply describes the biases found in a robust manner, associating a P-value with the event of finding a given number of sites in a set of sequences, each of a certain length. For random rankings of the gene list one typically finds maximum enrichment peaks consistent with the number of words tested (**Supplementary Fig. 4**). For biological experiments, coordinated patterns of enrichment or depletion can occur for a large number of words simultaneously. This is invariably due to composition biases, and we strongly recommend the use of Markov correction to remove such biases. A strong enrichment or depletion signal is found as a sharp peak on the positive Y-axis or a deep trough on the negative Y-axis near one of the ends of the gene ranking. It is useful to establish the general principle of the relationship between peak and trough locations and enrichment and depletion patterns. A number of cases can be distinguished, assuming that the genes are ordered from upregulated (on the left) to downregulated (on the right).

Consider a word W and its associated curve in the Sylamer landscape plot. For a peak occurring on the positive Y-axis, W is *overrepresented* in the 3'UTRs for the genes to the left of that peak, and W is *underrepresented* in the genes to the right. The P-values associated with these two events are identical, which is a generic property of the hypergeometric distribution. If we suppose the maximum deviation is found as a trough on the negative Y-axis, then W is *underrepresented* in the genes to the left of that trough and W is *overrepresented* in the genes to the right of that trough. If W is the complement of a microRNA seed (seed match) and a significant peak is found on the positive Y-axis, it means that the set of genes to the left of the peak is overrepresented for that word, likely as a result of an underlying biological trait of the system being examined. This is consistent with an experiment in which the microRNA was knocked out or knocked down, where the genes that it would normally suppress are going up. If W is the seed match of a microRNA and a trough is found on the negative Y-axis, it means that the set of genes to the *right* of the trough is overrepresented for that word. This is consistent with an experiment where a microRNA is overexpressed, reinjected in a null model, or possibly occurring as a nonnative microRNA (e.g. with a site mutation in the seed). In this case the microRNA suppresses genes, which are found in the downregulated part of the genelist. In this case, if we reverse the gene ranking, the signal will now show as a peak on the positive Y-axis, with the genes on the left of the peak downregulated and enriched for the seed match. Another condition may arise where a ranking is constructed by comparing mRNA expression for a particular tissue against for example average expression in a range of other tissues, and genes are again ranked from upregulated to downregulated. In this case genes that are upregulated in the tissue of interest are likely to avoid regulation by a microRNA that is highly expressed in that tissue. This effect has been demonstrated to exist [1, 2], and is revealed as a depletion signal for the seed match in the upregulated genes. In the Sylamer landscape plot this will show as a trough on the negative Y-axis on the *left* side of the gene ranking.

Threshold selection

The main distinction is whether a clear peak is present near the beginning or end of the ranking, or whether the signal is stretched across a large part of the gene list. In the latter case, we do not propose to use the peak location in order to infer a threshold for the candidate genelist. Instead, the landscape plot can be used as qualitative support for the hypothesis that the relevant microRNA is involved in the pathways affected by the experiment. Next consider the case where a clear enrichment peak is found near the beginning of the ranked genelist. Results for hexamers, heptamers, and octamers should be compared. In the ideal scenario, the signals for these results, if all present, should approximately agree in the shape of the curves and the location of their peaks. The peak closest to the start of the ranking can be chosen as a conservative threshold. This strategy also applies to cases where multiple peaks are present across the landscape, but a significant one is found near the beginning of the ranking (eg. **Supplementary Fig. 2d**).

One must also consider the experiment from which the gene ranking is derived. If a sufficient number of biological replicates are used, the ranking as well as the peak location should be fairly stable (as discussed in **Analysis of noise effects**), and the user may construct a candidate gene list by considering genes containing matches for the selected microRNA in the initial segment delimited by the peak. Alternatively, the gene ranking could be derived from either a single replicate or biological replicates of which a significant number were of lower quality (or for which the biological source was inherently more variable, e.g. whole tissue or embryonic sample). In this case, if the peak is not clearly and consistently defined, in the face of uncertainty about the sampling errors and biological variability, we propose the peak is only taken as an upper bound for any candidate gene list. One should be cautious in handling this candidate gene list, and be prepared for a lower success rate in doing follow-up experiments, and ideally, consider adding biological replicates or focusing on a more controlled experimental design.

Analysis of noise effects

As shown (**Fig. 1**), the ranking resulting from analysing an experiment with five biological replicates exhibits a clearly defined peak for the word complementary to the miR-155 seed. Below we demonstrate that the robustness of the results obtained by Sylamer is directly related to, and simply dependent on, the robustness and reliability of the microarray measurements themselves. We also demonstrate, in the example studied, that by combining biological replicates we obtain an enrichment signal, consistent with the biological traits of the system, that is much stronger than the average enrichment signal obtained by using combinations of fewer replicates. This strongly suggests that using more replicates indeed reduces the effect of biological variability and sampling errors and that the final result reflects more accurately the underlying biology.

We took two approaches to study the effect of noise in the system, the first showing the effect of biological variability, and the second showing the effect of adding Gaussian noise to the log2 normalized expression values. We also study the combined effect of the two different types of noise. It can firstly be observed that each biological replicate represents a measurement of a noisy system, where the measuring tool itself (the microarray) is also prone to sampling errors. Hence we show the effect of using fewer or no biological replicates in constructing the ranked genelist (**Supplementary Fig. 3)**. To this end, we obtained a ranking based on the fold-change resulting from comparing each of the mutant replicates with each of the wildtype replicates, obtaining 25 different rankings in total. For each of these rankings we ran Sylamer for the seed match of miR-155 (GCATTAA). For the purpose of this analysis we omit all other words as the seed match peak is standing out as the most significant by far. We then re-applied this procedure by changing the number of replicates used. The two extreme cases are either using all replicates or using only a single replicate. Other possibilities are using 10 combinations of two replicates for each of the two conditions, leading to 100 different rankings, using all 10 possible combinations of three replicates (100 rankings) and using all 5 possible combinations of four replicates (25 rankings). The results (**Supplementary Fig. 3)** are ordered from left to right by first showing the original result using all replicates, and subsequently decrementing the number of replicates used. In the second approach, we study what happens when we add artificial noise to the system under study. To this end, we computed the standard deviation of all the probesets across the ten arrays. We selected cutoff values at which respectively 50%, 80%, 90%, 95% and 99% of the probesets have a standard deviation that is less than the cutoff value. These selected SD values were 0.15, 0.24, 0.33, 0.44 and 0.81 respectively. These values were then used as a new standard deviation to generate Gaussian noise with. The top row (**Supplementary Fig. 3)** describes the case where no noise was added. The subsequent rows represent the cases where respectively the 50, 80, 90, 95, and 99 percent cutoff values were used. The first column of each row shows the histogram of the resulting SD for all probesets after applying the corresponding level of Gaussian noise. Accordingly, (**Supplementary Fig. 3)**, going from left to right the effect of using fewer biological replicates is shown, and going from top to bottom the effect of adding an increasing level of Gaussian noise to the underlying expression data is shown. It can be clearly seen that as noise increases in either direction the spread in location of the peaks is increasing, and the height of the peaks is decreasing. The first is to be expected, as different subsets of replicates will have, on average, more replicates in common for larger subsets. In the case where we combine four out of five biological replicates, two different combinations in the same condition will share exactly three replicates. If we combine three out five replicates, two different combinations in the same condition will share either one or two replicates. Hence it is to be expected that the shapes of the curves converge as we use more replicates. However, the fact that the height

(i.e. the significance level) strictly and considerably increases as more replicates are combined or as less noise is added, demonstrates that the Sylamer results become more stable and clearly defined as the effect of noise in the system is reduced. At the same time, it is clear from the experiments done on fewer replicates or those with more noise, that the Sylamer plots reveal a distinguishable seed match signal for miR-155 in nearly all of the cases.

Existing motif discovery methods

There is a large and varied body of research related to the discovery of motifs in biological sequences[3-10]. A significant amount of this work has been devoted to discovering the binding site specificity of Transcription Factors and other nucleotide-binding proteins. These binding sites reflect binding properties between peptide chains and nucleotides, and a common way of representing this is by a Position Specific Scoring Matrix (PSSM) reflecting the frequency of occurrence of each nucleotide at each position of known or predicted binding sites. In the miRNA scenario, the interaction of interest is between two RNA molecules, and it has been shown that a region of 6-8 consecutive nucleotides of perfect complementarity at the 5' end of the miRNA (seed region) is critical. Hence, simple models that use occurrences or frequencies of un-gapped k-mers are the most appropriate. These models have the advantage that they can be used in an exhaustive manner. In genome wide experiments, it can be inconvenient to classify sequences into discrete categories. For example, in microarray experiments resulting in a fold-change value for each gene, one would have to use an arbitrary cut-off to classify genes as changed and unchanged. A more natural approach would be to take these fold-change values into account or the ranking of the genes according to these values. We first compare Sylamer against oligo-analysis, a method that is generally applicable to analyze k-mer frequency enrichment or depletion in sequences. We then compare Sylamer with two approaches that have been used to discover miRNA signals by analysing un-gapped k-mers in the context of a ranked gene list.

Oligo-analysis[3, 11]

Oligo-analysis is available on the web (http://rsat.ulb.ac.be/rsat/) and the programs are available on request from the authors.

Oligo-analysis computes a binomial P-value for each word comparing the observed frequency in a single set of unranked sequences against an expected background frequency. The background model can be freely chosen, and can be based on the observed frequencies in the genome or on expected frequencies computed using a Markov model derived from an arbitrary set of sequences. Sylamer improves on this by including the ability to scan down a list of ranked sequences, calculating significance values at each step. The binomial distribution assumes that the universe is infinitely large. For our purpose, where the 3' UTRs of interest can be a large fraction of the genome, this assumption is no longer valid, and the hypergeometric distribution is the natural choice. Finally, oligo-analysis is currently implemented in Perl, and hence Sylamer can be faster by several orders by magnitude, even though it uses more sophisticated statistics that are computationally more demanding.

Farh et al. [1]**.**

This method is not available and would require a significant amount of coding and testing to implement.

The approach by Farh et al. was designed to find microRNA effects when comparing expression data from different tissues, in particular focused on finding depletion instead of enrichment. The method consists in splitting the ranked gene list into bins of fixed size and calculating for each one the difference between observed and expected occurrences for each miRNA seed word. The expected occurrences are calculated using the observed occurrences of trinucleotides and 3'UTR length (a second order Markov chain), and are corrected for the deviation between the estimated occurrence and observed occurrence of each seed word in the complete set of 3' UTRs. A cumulative sum of these differences is calculated across all the bins, and the maximum of the absolute values is recorded. In order to estimate a P-value for this signal, the authors used a non-parametric

Kolmogorov-Smirnoff test comparing the observed value with that obtained when using 1,000 control cohorts of genes. For empirical P-values below 0.02 the authors fit the asymptotic KS statistic tail probability to the tail of the empirical distribution to derive an approximate P-value. In their paper, the authors tested 73 miRNA families (with distinct 7mer seed sequence) and the rankings according to relative expression in 61 mouse tissues. Among other things, they showed that genes that are specifically expressed in tissues where a miRNA is also specifically expressed tend to be depleted for the target sequence of that miRNA, supporting the anti-target hypothesis[12].

Sylamer improves on the approach by Farh, et al. in a number of ways. The hypergeometric is a simple yet robust statistical framework that allows us to test for depletion or enrichment of miRNA seed words. The use of an exact test statistic means that Sylamer does not require computationally expensive simulations and can potentially return more accurate results. To prove that the hypergeometric model is valid for this kind of analysis, we permuted the list of all mouse 3'UTR sequences (corresponding to the Affymetrix Mouse 430 2.0 array) 1,000 times and plotted the enrichment landscape for all 709 distinct miRNA 7mer seed words (**Supplementary Fig. 4**) taken from miRBase[13]. We can see that the P-values fall within the expected range given the number of tests performed (an E-value threshold of 0.01 is shown with the dotted red lines). So the use of the hypergeometric distribution allows us to perform a fast analysis that is directly comparable between different words. Sylamer also allows great flexibility in the options of how miRNA targets are counted (number of genes targeted, or number of target sites among their 3' UTRs), and it can calculate the expected number of targets in multiple ways, including Markov models of any level. Sylamer is not restricted to a particular word size, although for miRNA analysis the most sensible choices are 6, 7 and 8. For datasets that have a very drastic composition bias (**Supplementary Fig. 2**) it can be useful to visualize the distribution of smaller words, to understand the underlying biases before interpreting the results of the miRNA seed words. Sylamer also includes modes to reverse the genelist, permute it, or remove a certain number of genes from either end of it, among many other options. In conclusion, Sylamer is a complete package specially tailored for the type of problems that we have encountered when analysing miRNA datasets that is available for anyone to use.

REDUCE[14]

The original REDUCE algorithm is available as a web server only, restricted to analyse selected sets of promoter sequences, and thus not directly applicable to the miRNA field.

REDUCE is a motif-based regression method for microarray analysis, and has been re-implemented and applied for the purpose of detecting microRNA signals in expression data [2]. It assumes a model where increase and decrease of expression levels are explained by the presence of regulatory motifs in a linear model. We believe this assumption is stronger than required for our purposes, and note that in the Sylamer approach and context no model is needed. REDUCE deals with one word at a time, removing its effect from the data. This precludes an exhaustive approach, as one may reasonably expect the linear model to break down once large amounts of words are incorporated. In our case, the exhaustive approach effectively works as an additional safeguard (showing the user the full background of every test conducted) and is facilitated by the fact that the motifs of interest are simple k-mers rather than more general sequence motifs. We expect the repeated fitting and solving of a linear model to be computationally costly. Also, REDUCE does not yield P-values by a rank-cutoff so that its results cannot be used to select the most enriched or depleted sequences. As far as we are aware the issue of composition biases is not addressed. A later method, by Foat et al. is MatrixREDUCE[5]. Based on a statistical mechanical model of the physical interaction between a transcription factor and DNA, the MatrixREDUCE algorithm uses genome-wide occupancy data for a transcription factor and associated nucleotide sequences to discover the sequence-specific binding affinity of the transcription factor modeled as Position Specific Affinity Matrix. The improvements, although noteworthy, are not relevant for our purpose. Sood et al. applied a REDUCE-like algorithm to correlate UTR motifs with changes in mRNA levels upon miRNA overexpression or knockdown[2] . This method has not been made available, and issues such as the effect of composition biases and scalability (speed) were not addressed in their paper.

Supplementary Methods

Algorithm

Sylamer takes as input a file containing one gene identifier per line, and a file in FASTA format containing sequences for these genes. The identifiers in the rank file and the FASTA file should be exact matches. Sylamer will order the sequences in the FASTA file according to the ordering of the genes in the rank file, and in the usual mode of operation sequences that are present in the FASTA file for which the identifier is absent in the rank file are ignored, as are identifiers in the rank file for which no sequence is found in the FASTA file. Sylamer accepts a size argument specifying the size of the words to be analyzed. By default all possible DNA/RNA words (composed of ACG[T/U]) of that size are analysed, but it is possible to specify a smaller set of words by supplying them in a word file. When the number of words is large, the issue of multiple testing should be considered (see **Multiple testing correction**). Sylamer tests the ranked genelist using multiple cutoffs, at each cutoff asking whether a particular word is more or less abundant in the top of the list then expected when compared to the rest of the list. By default this is done until the cutoff includes the entire set of sequences to be analyzed. Cutoffs are constructed using a *stepsize*, determining the granularity of measurements along the ranked gene list. The *n-th* cutoff taken thus defines the initial *n * stepsize* sequences. Such a set of sequences is called a leading bin (of the ordered sequences). The Sylamer signficance values (see below) are passed to a drawing program. This can be an R script (an example is included with Sylamer) or a wrapper program. We also provide jSylamer, a Java interface to Sylamer. A significance curve is now constructed for each word separately, plotting significance scores on the Y-axis for each cutoff on the X-axis. The X-axis thus tracks the number of genes in the top of the list. Signficance is calculated using hypergeometric (default setting) or binomial statistics (see **Statistical model**). At each cutoff, for each word, this yields two P-values, one for depletion of the word, and one for enrichment. The smallest is taken and then negative log transformed if the word is enriched in the top of the list, or log transformed if the word is depleted. The transformed value is plotted on the Y-axis at the relevant cutoff position on the X-axis. All the plotted points for a single word are connected to form a curve.

For larger word sizes (i.e. 8-15) restricted word lists speed up Sylamer significantly. Running Sylamer on 40 bins takes less than a minute using the full word lists (**Supplementary Fig. 1**), but takes approximately 5 seconds using word lists consisting of microRNA seed matches only. Tracking and plotting all possible words has the advantage that sequence composition biases are revealed. These biases can be corrected, and low-complexity sequences are removed by filtering (see **Correction of composition biases** and **Low-complexity and redundant sequences**).

Statistical model

The hypergeometric distribution is a discrete probability distribution that describes the number of successes in a sequence of *n* draws from a finite population without replacement[15]. The binomial distribution is similar but corresponds to drawing *with* replacement. In our case, for a particular miRNA, each *draw* is a word obtained from a set of sequences and a *success* is where that word is complementary to the miRNA seed. Sylamer computes P-values for both the hypergeometric and binomial cumulative density functions. Binomial and hypergeometric results are similar for small sets, but binomial P-values become meaningless (tending to 1) for larger sets. The hypergeometric model allows no replacement and, as such, represents a real partitioning of the genelist into two sets, above and below the cutoff. An important property arises in the context of ranked genelists. The P-value associated with the over- or under-representation of a word in a leading bin (i.e. the set of sequences falling below a given rank cutoff), is equal to the P-value associated with the under- or over-representation of that word in the complement bin. This has the considerable advantage of imparting symmetry to the results. At any cutoff, every P-value represents the over-representation of a word in one of the sets and its under-representation in the complement set. The genelist does not need to be analyzed in reverse order, as required with the binomial distribution. A plot can simply be rotated 180 degrees to represent the results of the reversed ranking.

The cumulative hypergeometric P-value requires four input parameters: population size (N), sample size (S), number of instances of a given type (T) in the population (N_T), and the number of instances of type T in the sample (S_T). The hypergeometric cumulative over-representation P-value represents the probability of obtaining at least S_T instances of type T when drawing S instances from a population of size N without replacement. Similarly, the cumulative under-representation P-value represents the probability of drawing at most S_T instances. The population size N is chosen as the total number of words present in the FASTA file. Only those words that do not contain masked bases are counted. The sample size S is the total number of words present in the leading bin of size $b \cdot stepsize$, for each bin b. Each different word of length K represents a different type T. For this type we define the counts S_T and N_T as respectively, the number of occurrences of the word in the sample and the universe. Sylamer uses the GNU Scientific Library[16] to compute cumulative hypergeometric P-values. This library provides highly efficient and optimized routines for scientific computing, available without restrictions on use or modification.

Multiple testing correction

Sylamer generates a P-value for word occurrences for all DNA/RNA words of a given length K, for each bin in the ranked sequence universe. Assuming b bins and word length K this produces $b4^K$ P-values in total. Additionally, a user may test multiple word lengths. Simply looking at the best P-values would be overly simplistic. Firstly, it can be noticed that findings for different word lengths should be expected to corroborate one another. Different word lengths should thus be used to create a bigger picture of an overall finding. A good example in this respect is miRNA seed match analysis where one uses words of length 6, 7, and 8. Secondly, a typical result takes the form of a significant incline or decline towards one of the ends of the ranked sequence universe. The number of bins does not play a large role in this respect, and merely determines the granularity of measurement across the ranked universe. Other results, where the extreme is found towards the middle of the universe, may also occur. In each case, the result should be underpinned by a biological explanation or hypothesis (see **Interpreting the landscape plots** for further discussion). Thirdly, a large number of words are always being tested. In the examples shown, the words corresponding to miRNA seed matches are clearly significant in relationship to the background. This can be validated by multiplying the resulting P-values by the number of words tested (Bonferroni correction), but equally important, the plot for the pertinent miRNA seed matches show a clear separation from the plots for other words of the same length. Plots obtained from the Sylamer GUI will automatically show a Bonferroni multiple-testing threshold. It is strongly advised that peaks observed below this peak are not used for validation or target selection.

Correction of composition biases

Composition biases in a section of the ranked sequences (e.g. %GC, di- or tri-nucleotide content) may cause word occurrence biases. When analyzing an experiment with Sylamer, one must ensure that enrichment values observed are specific to the phenomenon of interest (for miRNAs, words of length 6-8). If smaller words show larger enrichment scores, it is likely that the effect of interest is being masked. An experiment comparing Th2 cells between miR-155 knockout and wild-type mice[17] exhibited strong compositional biases (**Supplementary Fig. 2**). A general trend with a large fraction of words showing abnormally high enrichment scores is observed. These biases are much higher for words of length 5 than for 6, and thus are unlikely to be caused by miRNA effects. Sylamer provides a correction step where population word counts are replaced by expected word counts based on bin composition bias. The expected word count for a given word is parameterized on a smaller word size, and computed as an expected frequency multiplied by the population size. Expected frequency is computed using counts of sub-words of the smaller size. The expected word count is derived from occurrence counts of smaller constituent words, relative to the current bin[3]. This approach is known as incorporating a higher order Markov model[18]. As the bin size grows, expected word counts based on this model increasingly deviate from true universe occurrences. This generally means that log P-values inflate rapidly. In order to dampen this behavior, expected word counts are modulated by a factor that incorporates both the globally expected word count and the true universe count[1]. This combined approach succeeds in reducing bias effects and is recommended when running Sylamer (**Supplementary Fig. 2d**). We strongly suggest employing this correction so that composition biases are removed to the largest possible extent.

Low-complexity and redundant sequences

Low-complexity sequences such as monomer or dimer repeats can cause drastic biases in counts, particularly for words that are part of the repeat. Sylamer will by default collapse any monomer or dimer repeat into a single occurrence. For more complex patterns, we suggest using DUST[19] (also Tatusov & Lipman; *unpublished*) prior to Sylamer analysis. In higher organisms, many alternative transcripts contain identical or overlapping 3′ UTR sequences. Paralogous genes can also encode for transcripts with very similar 3′ UTRs. If these sequences end up at similar positions in the genelist, it can lead to inflated estimations of significance of any word contained (or depleted) in them. To avoid this, we suggest masking repetitive and redundant sequences by using, for example, the RSAT purge-sequence interface (http://rsat.ulb.ac.be/rsat/ and ref 4) to Vmatch (http://www.vmatch.de/).

Microarray Data Analysis

Microarray datasets used were: *E-TABM-232* (ArrayExpress[20] Accession) for the miR-155 experiment[17], *GSE4201* (GEO[21] Accession) for the miR-430 experiment[22] and *E-MEXP-1402* (ArrayExpress Accession) for the RNAi experiment[23]. For the first two, Affymetrix CEL files were obtained directly from ArrayExpress and GEO. Both were processed using R & Bioconductor[24]. The data were background corrected and normalized using RMA[25]. Differential expression analysis was performed using the limma package[26], from which moderated t-statistics were obtained. For the RNAi experiment, fold changes for each siRNA transfection experiment were obtained directly from ArrayExpress. The ranked genelists in each case are available (**Supplementary Data 1,2 and 3** online).

3′UTR sequences

For Affymetrix datasets 3′UTRs were obtained by mapping probeset IDs to RefSeq transcript identifiers, using annotation files provided by Affymetrix. Where a probeset did not map directly to a RefSeq sequence, or did not have a 3′UTR we took (where available) the 3′UTR sequence from the longest annotated Ensembl transcript. The sequence files used are available (**Supplementary Data 1,2 and 3** online). Sequences and identifiers for genes tested in the miR-430 reporter assays were obtained from the literature[22].

Implementation and Availability

Sylamer is implemented in C, and optimized for DNA and RNA sequences. Counts are by default tracked in arrays covering the full set of words of a given length, using an unsigned integer map of a word as offset into this array. For smaller word sizes this is very fast (**Supplementary Fig. 1**) as the full array fits into CPU cache. When a word list is specified integer hash arrays are used to limit the size of the count arrays, thus optimizing the CPU cache hit rate. This is typically useful for word sizes ≥8. If the counts for many words are tracked, computing the hypergeometric and binomial P-values becomes the main bottleneck. The GNU Scientific Library (GSL)[16] provides fast C routines for computing these. The figures shown herein were obtained by plotting Sylamer output using R. A script for producing enrichment plots is packaged with Sylamer together with a JAVA Graphical User Interface.

Sylamer and related programs, documentation and data are available at:
http://www.ebi.ac.uk/enright/sylamer/

Supplementary References

1. Farh, K.K. et al. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science* **310**, 1817-1821 (2005).

2. Sood, P., Krek, A., Zavolan, M., Macino, G. & Rajewsky, N. Cell-type-specific signatures of microRNAs on target mRNA expression. *Proc Natl Acad Sci U S A* **103**, 2746-2751 (2006).
3. van Helden, J. Regulatory sequence analysis tools. *Nucleic acids research* **31**, 3593-3596 (2003).
4. Bailey, T.L., Williams, N., Misleh, C. & Li, W.W. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic acids research* **34**, W369-373 (2006).
5. Foat, B.C., Morozov, A.V. & Bussemaker, H.J. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* **22**, e141-149 (2006).
6. Eden, E., Lipson, D., Yogev, S. & Yakhini, Z. Discovering motifs in ranked lists of DNA sequences. *PLoS Comput Biol* **3**, e39 (2007).
7. Tanay, A. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome research* **16**, 962-972 (2006).
8. Flicek, P. et al. Ensembl 2008. *Nucleic acids research* **36**, D707-714 (2008).
9. Berger, M.F. et al. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology* **24**, 1429-1435 (2006).
10. Tompa, M. et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nature biotechnology* **23**, 137-144 (2005).
11. van Helden, J., Andre, B. & Collado-Vides, J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* **281**, 827-842 (1998).
12. Bartel, D.P. & Chen, C.Z. Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. *Nature reviews* **5**, 396-400 (2004).
13. Griffiths-Jones, S., Saini, H.K., van Dongen, S. & Enright, A.J. miRBase: tools for microRNA genomics. *Nucleic acids research* **36**, D154-158 (2008).
14. Bussemaker, H.J., Li, H. & Siggia, E.D. Regulatory element detection using correlation with expression. *Nature genetics* **27**, 167-171 (2001).
15. Freedman, D., Pisani, R. & Purves, R. Statistics, Edn. 3rd ed., International student ed. / David Freedman, Robert Pisani, Roger Purves. (W.W. Norton, New York ; London; 1998).
16. Galassi, M. et al. Gnu Scientific Library: Reference Manual, Edn. 2nd ed. (Network Theory Ltd., Bristol; 2003).
17. Rodriguez, A. et al. Requirement of bic/microRNA-155 for normal immune function. *Science* **316**, 608-611 (2007).
18. Phillips, G.J., Arnold, J. & Ivarie, R. Mono- through hexanucleotide composition of the Escherichia coli genome: a Markov chain analysis. *Nucleic acids research* **15**, 2611-2626 (1987).
19. Hancock, J.M. & Armstrong, J.S. SIMPLE34: an improved and enhanced implementation for VAX and Sun computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences. *Comput Appl Biosci* **10**, 67-70 (1994).
20. Parkinson, H. et al. ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic acids research* **35**, D747-750 (2007).
21. Barrett, T. & Edgar, R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol* **411**, 352-369 (2006).
22. Giraldez, A.J. et al. Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science* **312**, 75-79 (2006).
23. Birmingham, A. et al. 3' UTR seed matches, but not overall identity, are associated with RNAi off-targets. *Nat Methods* **3**, 199-204 (2006).
24. Gentleman, R.C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80 (2004).
25. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193 (2003).
26. Smyth, G.K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, Article3 (2004).