Published online May 12, 2006

Nucleic Acids Research, 2006, Vol. 34, No. 9 ***e65***
doi:10.1093/nar/gkl173

MicroRNA enrichment among short 'ultraconserved' sequences in insects

T. Tran[1], P. Havlak[2] and J. Miller[1,*]

Department of [1]Biochemistry and [2]Human Genome Sequencing Center, Baylor College of Medicine, TX, USA

Received September 28, 2005; Revised January 2, 2006; Accepted March 20, 2006

ABSTRACT

MicroRNAs are short (∼22 nt) regulatory RNA molecules that play key roles in metazoan development and have been implicated in human disease. First discovered in *Caenorhabditis elegans*, over 2500 microRNAs have been isolated in metazoans and plants; it has been estimated that there may be more than a thousand microRNA genes in the human genome alone. Motivated by the experimental observation of strong conservation of the microRNA let-7 among nearly all metazoans, we developed a novel methodology to characterize the class of such strongly conserved sequences: we identified a non-redundant set of all sequences 20 to 29 bases in length that are shared among three insects: fly, bee and mosquito. Among the few hundred sequences greater than 20 bases in length are close to 40% of the 78 confirmed fly microRNAs, along with other non-coding RNAs and coding sequence.

INTRODUCTION

Functional constraints on sequence variation lead to sequence conservation. This observation can with some care be used to infer that a sequence about which we know little, in fact has some function (1). The mouse-human genome comparison is a case in point; far more sequence is conserved between these two genomes than can be accounted for by coding genes (2). Non-coding RNA is one candidate for function that is already known to account for some of this abundance of conserved sequence (3,4).

MicroRNAs (miRNAs) represent a class of such non-coding RNA (5). Transcribed as stem–loop precursors that are eventually processed to mature sizes of 17–24 nt (6), they regulate gene expression in animals typically by binding to the 3′-untranslated region (3′-UTR) of an mRNA and suppressing translation (7,8). The set of transcripts that a given miRNA regulates are believed to be determined in a subtle manner by its sequence (8,9). MiRNAs have been shown to play critical roles in regulation of development in multi-cellular organisms (10); they have also been implicated in human diseases, such as chronic lymphocytic leukemia (11). More recently, it has been suggested that expression levels of 200-odd miRNAs may better correlated with tumor cell character than 20 000 coding transcript levels (12). Finally, *Drosophila* stem cells defective in miRNA processing are no longer capable of self-renewal (13).

Since the discovery of the first miRNA in *Caenorhabditis elegans* (10,14), it has emerged that nearly all metazoans have miRNA genes, some in large numbers. Over three hundred have been detected in humans, but many more are predicted, with estimates of their total number in human ranging to over a thousand (15). MiRNA gene sequence may simultaneously code for proteins.

MicroRNAs can be strongly and deeply conserved in evolution. Indirect evidence suggests that the second miRNA to be discovered, let-7 (16,17), is expressed in nearly all metazoans—apparently with the identity of all (or nearly all) of its 22 nt preserved exactly (18). It has been argued that this conservation is a consequence of the many distinct targets (perhaps hundreds) that it regulates (9), so that the simultaneous reciprocal mutations necessary to preserve its activity on all these targets are highly improbable (8); alternatively it has been argued it may reflect a signal for protein binding (9).

The work described here was motivated by the ubiquitous mature let-7 miRNA, which raises the questions: what constitutes the class of (short) sequences conserved exactly among diverse organisms? What functionalities, if any, are over-represented in this class? We have in this paper formulated what can loosely be thought of as an unbiased and model-free strategy to address these questions, a strategy that might plausibly qualify as the simplest conceivable incarnation of comparative genomics. In particular, we bypass the formal and technical apparatus of whole-genome alignment, and we compute an object that is defined independently of any algorithm. Among the set of three divergent insect genomes (fly, bee and

*To whom correspondence should be addressed. Tel: +1 713 798 3542; Fax: +1 713 796 9438; Email: millerj@bcm.tmc.edu

mosquito) that we study here, it turns out that confirmed mature miRNA sequences are specifically enriched within this class, and over a certain length range form a primary contribution.

Our strategy is one of direct, all-on-all, whole-genome comparison: we identify all distinct sequences above a specified minimum length that are common to a certain set of genomes—we call this collection of sequences the 'intersection' of these genomes.

We denote the lengths of these sequences by 'N' and refer to the sequences themselves as 'N-mers'. At sufficiently short lengths all possible N-mers would be contained within each of these genomes [consider 4-mers; see e.g. (19)]. Correspondingly, we would not expect to find sufficiently long sequences even in two individuals of the same species.

In the intermediate regime, estimates of the number of exactly conserved sequences common to two or more genomes that would occur by chance (coincidences) can be accurately computed based on N-mer length and genome size, under the assumption that each genome is independent and random; however, once non-randomness and evolutionary distance are taken into account, we are unaware of any reliable methods of obtaining such estimates. Two current approaches (20,21) to formulating estimates of this kind are described in Results and Discussion. In the absence of these estimates, it is difficult to anticipate either the number of such sequences or the fraction of them that are likely to arise by selection (presumably implying functionality) versus by chance alone. It may be that studies like the one described here, in which these questions are addressed empirically, are prerequisites for developing a theoretical framework for computing how many conserved N-mers are to be expected in an intersection.

Because the technical difficulty of this kind of computation increases rapidly with genome size, we chose to address these questions first within the relatively compact insect genomes rather than vertebrates so that this set of longer sequences would be smaller and more readily managed. We intersected the genomes of *Drosophila melanogaster, Apis mellifera* and *Anopheles gambiae*, and identified all those distinct N-mer sequences that are common to all three organisms, for N greater than 19. To eliminate double-counting, we tabulated a special subset of these sequences, the maximal N-mers. Our methods are novel and for this specific task out-perform whole-genome alignment, the standard practice in this subfield of comparative genomics, both in execution time and sensitivity.

In contrast to the original studies that applied the term 'ultraconserved' only to sequences conserved among three organisms of length 50 bp or greater and examined exclusively sequences in that length range (22,23), we are principally interested in sequences of length <50 bp. Furthermore, rather than relying on prior alignments of orthologous parts of the respective genomes to identify such conserved sequences as in (22,23) we perform a true whole-genome on whole-genome, all-on-all search. To maintain consistency with the definition of 'ultraconservation', we suggest that the phenomenon described here be referred to as 'microconservation'. It is worth observing that although these sequences are shorter, there are many more of them.

MATERIALS AND METHODS

Genomes

Repeat-masked (24) genomes of *D.melanogaster* (BDGP3.2.1.may.dna_rm.chunk.fa.gz), *A.mellifera* (AMEL2.0.may.dna_rm.contig.fa.gz) and *A.gambiae* (MOZ2a.may.dna_rm.chunk.fa.gz) were downloaded from the Ensembl database, www.ensembl.org. Assembled chromosomes were obtained when possible. The repeat-masked genomes range from ∼100 million to 250 million bases in length.

Overview of the intersection procedure

All-against-all comparisons were carried out for each pair of organisms, with no prior alignment of the genomes, to identify all distinct sequences of length N that are contained in both genomes. Because we were principally interested in length such that 4^N, the total number of possible sequences of length N, is much larger than the genome size, we employed hash table methods based on code originally developed for whole-genome assembly.

Developed for large-scale N-mer analysis as part of the Atlas genome assembly project (25), the Atlas tools enable the tabulation in memory of N-mers up to 32 bases in length. Other sequence comparison tools rely on disk storage rather than memory, yielding prohibitive access times, or are limited to small N (26).

For the first stage of the procedure, the sequence length was fixed to some minimum value K, where K was typically around 20.

In each binary intersection, the K-mers for a single genome were tabulated using a hash table that mapped each K-mer to a unique cumulative occurrence counter. Occurrences in a second genome were then tabulated in new counters in the hash table, but only for those K-mers already encountered in the first genome. Subsequently the genomes were scanned again and each K-mer common to both genomes was reported together with its locations in each genome.

The outcome is a list of contiguous sequences of length K common to both organisms, along with their locations (chromosome and position) within each genome, and the total number of times each sequence occurs in each genome. In practice, the number of bits in each word of memory must be allotted between a binary representation of the sequence and its copy number, so that copy numbers are only recorded up to some maximum value.

From these lists, the locations enable the reassembly of all contiguous sequences of lengths greater than or equal to K within each genome. These reassembled sequences are broken down into sequences of length N and length N+1 for each organism, and intersected separately. Finally, any sequence in the N-mer intersection that is contained in some sequence in the (N+1)-mer intersection is eliminated, yielding the set of 'maximal' N-mers. Once binary intersections have been computed, it is straightforward to obtain higher-order intersections. Observe that it is not the case that the set of maximal N-mers for a ternary intersection can be computed simply by intersecting two sets of maximal N-mers from any of the respective binary intersections.

The representation of the full set of N-mers by the maximal N-mers is not necessarily complete. For example, if the sequence 'XY' is contained in both organisms, then the sequence 'X' can't be maximal; however, it is possible that 'X' occurs in other contexts in both organism, e.g. as 'XA' in one organism and as 'XB' in the other organism.

Obviously if all sequence is single copy, then the maximal N-mers do represent the full N-mer set. At the lengths discussed in this paper, nearly all sequence—and in particular mature miRNA sequence—is indeed single copy. Nevertheless, multiple-copy sequence is of interest, and will be addressed in the future.

An alternative way of thinking about maximal sequences may be enlightening: For a binary intersection, at each location in a chromosome consider the shortest sequence spanning that location that appears in only one of the two genomes. Then by removing a letter from an end of this sequence, we obtain a maximal sequence.

Finally, we don't distinguish between orientations of a sequence; occurrence of any sequence is equivalent to occurrence of its reverse complement. In practice, we record only the lexically smaller of a sequence and its reverse complement, and all numbers we report are based upon this convention, unless otherwise noted.

Filtering

As discussed below, when N decreases into the 19–22 range, N-mers are increasingly enriched for 'simple' sequences—they consist entirely or in part of runs of homo-polymer, of repeating di- or tri- nucleotides, or are within a few base substitutions of such a sequence. These sequences are atypical of metazoan miRNAs, and we constructed an *ad hoc* filter to eliminate some of them.

We computed the base entropy of each N-mer, $S = \Sigma f_i \log_2 f_i$ where f_i denotes the fraction of the N-mer consisting of base i and the sum is over i = A, G, C, T. We also made a list of all periodic sequences of bases with period less than 7, and we computed the Hamming distance H of the N-mer to each of these periodic sequences, in every phase. Were any of these values of H less than 7, the N-mer was eliminated. Of the 835 distinct metazoan miRNA sequences in miRbase, 817 or 97.8%, survive this filter, which for convenience we will call a 'complexity' filter.

Data are discussed below for both unfiltered and filtered sequence sets, with filtered sets clearly identified as such. For this set of genomes, the filter did not remove from the intersections any N-mers with $N > 25$ or any confirmed miRNAs. This filter is obviously biased in favor of confirmed miRNAs from miRbase, and is introduced with identification of new miRNAs in mind; however, we believe its chief role is to eliminate 'simple' sequences that are too short for the repeat-masker algorithm to remove.

Annotations

Sequences were assigned *D.melanogaster* annotations from FlyBase (27). MiRNA sequences of *D.melanogaster, A.mellifera and A.gambiae* were obtained from miRbase release 7.0 (28) at http://microrna.sanger.ac.uk/sequences/index.shtml. Annotation features, release 4.2, (in GFF format version 3) of *D.melanogaster* and their respective DNA sequences in FASTA format were downloaded from FlyBase at http://flybase.bio.indiana.edu/. These GFF files were converted to MySQL tables using a BioPerl script (29) that can be found at (http://www.bioperl.org). For each N-mer, locations of all exact, full-length hits to the fly genome were determined with WU-BLAST 2.0 (http://blast.wustl.edu) and passed to a Perl script that searched the MySQL tables.

Annotation was performed as follows; a sequence and its reverse complement were treated separately provided it wasn't a palindrome. If a sequence is an exact match to a subsequence of an insect miRNA in miRbase, report miRNA. Then, locations of its exact hits in the fly genome are identified, and for each hit location:

(i) Scan the MySQL transcription start site, repeat region, transposable element, transposable element insertion site, protein binding site, enhancer, regulatory region and polyadenylation site tables; report all annotations encountered.
(ii) If the location is annotated as a miRNA in FlyBase, report miRNA. Next hit location.
(iii) If the location is annotated as a non-coding RNA, report the class of non-coding RNA if available from the MySQL tables; if not available report 'ncRNA.' Next hit location.
(iv) Scan the MySQL intron, 3′-UTR, 5′-UTR and coding tables; report all annotations encountered. If a hit spans an intron/coding, coding/intron, UTR/coding or coding/UTR border, report 'cborder'. Next hit location.
(v) If this location has no annotation, report 'intergenic'. Next hit location.
(vi) A single hit can have multiple annotations; consequently, the sum of the hit fractions displayed in Figure 1 may exceed unity.

Alignments

Whole-genome alignments were obtained from various sources (30–32): http://hgdownload.cse.ucsc.edu/goldenPath/ (chained and netted alignments); fly/mosquito: anoGam1/vsDm2/axtNet/*.anoGam1.dm2.net.axt.gz; fly/honeybee: apiMel2/vsDm2/melanogasterNet.axt.gz; frog/human: xenTro1/vsHg17/axtNet/all.axt.gz; frog/mouse: xenTro1/vsMm7/xenTro1.mm7.net.axt.gz; http://pipeline.lbl.gov; human/mouse/rat: /data/hg16_mm4_rn3/chr*.xmfa.gz.

RESULTS AND DISCUSSION

Results

Intersections. Starting with repeat-masked genomes of *D.melanogaster*, *A.mellifera* and *A.gambiae*, we identified all those distinct N-mer sequences that are common (shared exactly, with no mismatches or gaps) to all three organisms; we call this set of sequences a (ternary) intersection. Intersections were computed for values of N greater than 19. The raw intersections are available as Supplementary Data. To avoid double-counting, we record below in Table 1 the numbers of *maximal* N-mers only: N-mer elements of an intersection that are not subsequences of any (N+1)-mer in the same intersection. Of these N-mers, a remarkable proportion of those in the range of N from 23 to 29 include confirmed mature miRNAs

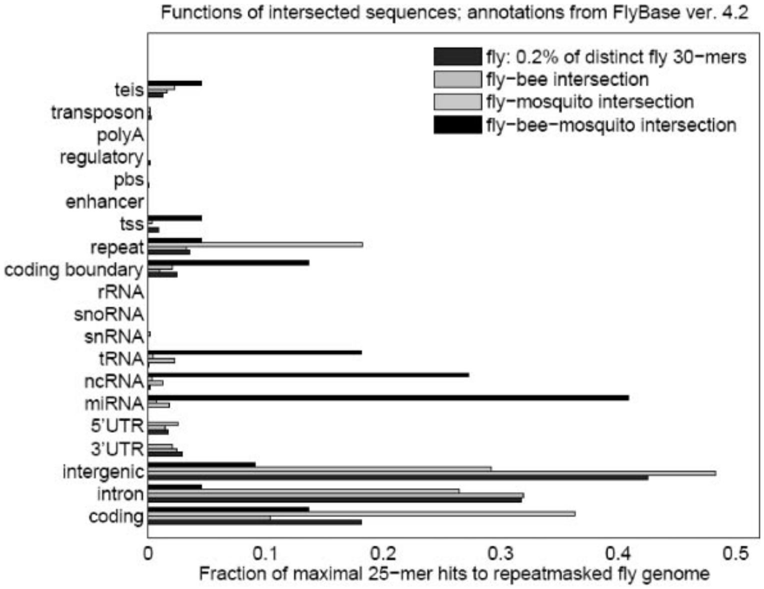

Figure 1. Fractions of unfiltered maximal 25mer hits to repeat-masked fly genome with a given annotation from FlyBase; see caption to Table 2 for details and abbreviations. Corresponding figures for all lengths are available in Supplementary Figures 1-unfiltered and 1-filtered. Annotations were counted as described in Materials and Methods: annotations. A single hit can have multiple annotations; consequently, the sum of the hit fractions displayed in the figure may exceed unity.

Table 1. Maximal N-mers in ternary insect intersections

Length N	Unfiltered N-mers	Filtered N-mers	miRNAs	cumulative miRNAs
>29	67	67	[a]2	[a]2
29	8	8	1	1
28	9	9	2	3
27	6	6	3	6
26	17	17	4	10
25	18	15	9	19
24	29	18	5	24
23	90	32	5	29
22	334	28	1	30
21	1163	88	3	33
20	5635	695	1	34
			Total: 34 of 78 in fly	

Number of unfiltered maximal N-mers [N-mers not contained in any (N+1)-mer] in the ternary intersection of repeat-masked fly, bee and mosquito genomes. Column 4 displays the number of exact, full-length hits containing mature miRNA possibly together with some flanking sequence in the precursor. Although all genomes were repeat-masked before intersection, we found that intersection strongly enriches for repetitive and low-entropy sequence, especially as N decreases. We applied a simple filter to the N-mers, detailed in the text, that increased enrichment for confirmed miRNA sequence in this regime as shown in column 3: Filtered N-mers.

[a]There were two exact hits above N = 29 to miRNA sequences containing the partners of mature miRNAs within their respective stem–loop precursors; one hit at N = 30, and one at N = 31. All hairpin precursors and the sequences of the hits to them can be found in Supplementary Tables 1-unfiltered and 1-filtered for both binary and ternary intersections.

as subsequences: except for N > 29, each entry in column 3 of Table 1 represents an exact, full-length hit to a subsequence of a miRNA precursor that contains the mature miRNA.

To estimate the scale of the enrichment, we computed the total number of distinct N-mers in repeat-masked fly genome for values of N in the range 20 to 29. This quantity works out to 10^8, within a few percent. There are roughly 78 confirmed microRNA genes in fly; the mature microRNA typically of length 19–23 bases, making up about 0.002% of the bases in the genome. A 25mer encompasses six overlapping 20mers, so that the set of confirmed mature miRNAs is comprised of 500 different (overlapping) 25mers, or about 0.0005% of the 20mers in the genome. As shown in Table 1, more than 50% of the maximal 25mers contain mature miRNAs as subsequences. For N > 22, of a total of 244 unfiltered N-mers, 29 contain mature microRNA sequences.

A simple empirically-motivated ’complexity’ filter increases the proportion of miRNAs obtained in 21 and 22mers dramatically: we retained only those N-mers with base entropy greater than 1.5 and with Hamming distance greater than six from all sequences of period less than 7. Approximately 98% of the 835 distinct mature animal miRNA sequences in miRbase satisfy this condition. Following application of the filter, the 34 mature miRNA sequences are subsumed within the ternary intersection among fewer than 200 N-mers above 20 bases and less than 30 bases in length.

Table 2 shows selected FlyBase annotations of the (unfiltered) maximal 25mers for binary and ternary fly-insect intersections, as well as for a random sample of repeat-masked fly genome; Figure 1 shows a histogram of the same data. What is perhaps most striking, aside from the enrichment for confirmed miRNAs, is the enrichment for annotated sequence (e.g. sequence labeled as other than intergenic).

Corresponding figures for all lengths are available as Supplementary Data at the NAR website, and we briefly summarize some of the features of the ternary intersection here: (i) for N > 29, 80% of the N-mers are tRNA sequences; (ii) intergenic and intronic sequence is strongly suppressed for N > 25; (iii) 5′- and 3′-UTR sequences are suppressed for N > 24; below that value they appear at the same rate as they do in repeat-masked fly genome; (iv) finally, for N < 26, binary fly/bee intersections are enriched for intergenic and intronic sequence and depleted for coding sequence.

Table 2. Annotations of unfiltered maximal 25mers

	miR	tR	snR	ncR	snoR	rR	3′	5′	cds	cborder	intron
Fly-bee-mosquito	9	4	0	6	0	0	0	0	3	3	1
Fly-bee	9	11	0	6	0	0	12	7	51	5	157
Fly-mosquito	6	4	1	3	0	0	17	21	302	17	220
Fly-sample (0.2%)	12	74	5	338	10	2	6098	3574	37573	5169	65886
	interg	repeat	tss	enh	pbs	regal	poly(A)	transpo	teis	hits	seqs
Fly-bee-mosquito	2	1	1	0	0	0	0	0	1	22	18
Fly-bee	237	16	0	0	0	0	0	1	8	491	446
Fly-mosquito	243	152	3	0	0	0	0	1	19	832	631
Fly-sample (0.2%)	88138	7293	1896	58	74	269	1	500	2660	207152	199560

Annotations from FlyBase of unfiltered maximal 25mers conserved exactly among binary and ternary intersections with fly. Annotations of a 0.2% random sample of all distinct 30mers from repeat-masked fly genome are shown for comparison. Labels: miR: microRNA; tR: transfer RNA; snR: small nuclear RNA; ncR: non-coding RNA; snoR: small nucleolar RNA; rR: ribosomal RNA; cds: protein coding; interg: intergenic; cborder: spans an coding/intron or coding/UTR boundary; repeat: repeat region; tss: transcription start site; enh: enhancer; pbs: protein binding site; regla: regulatory region; poly(A): polyadenylation site; transpo: transposable element; teis: transposable element insertion site; hits: number of hits to repeat-masked fly genome; seqs: number of sequences. Complete data for all N > 19, both filtered and unfiltered, are available as Supplementary Table 2.

The short sequence lengths addressed here relative to those investigated in the original ultraconservation studies (22,23), pose challenges in assessing significance that were of secondary importance for longer sequences. In particular, given a known evolutionary distance between two organisms, plausible approximate measures of neutral drift yield estimates for coincidence—the chance occurrence of an identical sequence in both organisms—that for longer sequences are minute. Of course, this conclusion relies on the assumption that the assumed neutral substitution rate is uniform (or exceeds some non-zero lower bound) over the entire genome.

Thus, larger values of N reduce the rate of coincidence, suppressing noise that could create 'false positives' within an intersection. For two random genomes (33), the rate of coincidence R can be estimated as $R \sim 1-\exp(-2G/4^{N})$ where G denotes the genome size in bases (P. Havlak, unpublished data). For $N = 20$ one obtains $R \sim 0.02\%$, which for genome sizes on the order of 10^{8} leads to ~20 000 distinct sequences; we typically obtain 10 to 20 times this number in a binary intersection at this length.

In practice, metazoan genomes that we discuss here are not random, so that this figure must be treated only as a very rough indication. We have chosen instead to stress the enrichment in annotated sequence that these intersections achieve, on the plausible expectation that a significant fraction of functional sequence within these genomes remains un-annotated. Thus, we compare (i) the fraction of annotated sequences in each intersection to (ii) the fraction of annotated sequences sampled at random from the (repeat-masked) fly genome. For $N = 25$, these data are displayed in Figure 2; the number of maximal 25mers in each intersection and the *P*-value (probability of obtaining at least the observed numbers of annotated sequences by choosing sequence randomly from the repeat-masked fly genome) are inscribed within each bar. Data for all other N may be found in Supplementary Figures 5 and 6.

In these comparisons, we have for each maximal N-mer obtained the annotation from FlyBase of every hit to the (repeat-masked) fly genome. We have adopted the following categories for convenience: 'intergenic' refers to sequence for which all hits have either (i) no annotation (e.g. between genes) or (ii) no annotation other than 'repeat'. 'Annotated' refers to sequence for which *at least* one hit has at least one annotation listed in Figure 1 other than 'repeat'.

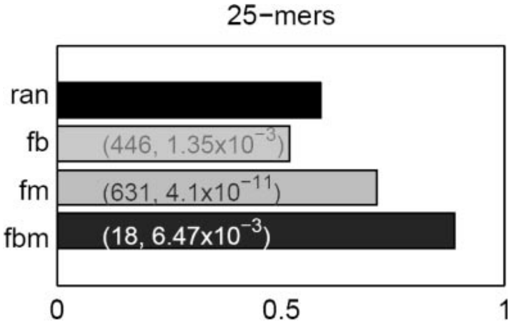

Figure 2. Fraction of unfiltered maximal 25mers with annotations from FlyBase, for a 0.2% random sample of distinct 25mers from the repeat-masked fly genome, and for binary and ternary intersections with fly genome. The number of maximal 25mers in the intersection is displayed inside the parentheses, followed by *P*-value. Corresponding figures for all other values of N are available in Supplementary Figures 2-unfiltered and 2-filtered.

The enrichments for annotated sequence in the unfiltered ternary intersections, as assessed by *P*-value, are statistically significant with a maximum *P*-value of 0.042 for 27mers (of which there are only 6), and vastly smaller for most other lengths. In the binary intersections, enrichment is obtained for lengths greater than 25; however, as lengths shorten, depletion of annotation is observed and is also statistically significant. Depletion is first seen between fly and bee, appearing only for the smallest lengths between fly and mosquito. It nevertheless holds universally that the fly-mosquito intersections are larger than the fly-bee intersections, consistent with the more recent divergence of fly and mosquito relative to bee. Note that for sufficiently short lengths, all N-mers appear in all organisms, and the annotation fractions must be equal among all intersections to within statistical error. One possible interpretation of the depletion is that much of the sequence at these lengths is functional but so far un-annotated in the fly; a more likely alternative is that at shorter lengths the intersections enrich for simple/repetitive sequence.

Whole-genome alignment. A more conventional approach to identifying highly-conserved or ultraconserved sequences and discovering miRNAs relies on performing whole-genome

Table 3. Whole-genome alignment versus N-mer intersection

Exact matches of length >22 containing mature microRNAs Genomes	Alignment	Intersection
Dm2/Am2	5	29
Dm2/Ag1	25	33
Hg17/Xt1	75	132
Mm7/Xt1	65	128
Hg16/Mm4/Rn3	207	213

For the ‘alignment’ column, we identified all (gap-free) contiguous sequence of length greater than 22 bases from UCSC whole-genome alignments (see Materials and Methods) that was both (i) aligned at corresponding positions and (ii) shared identical base sequence in all organisms contributing to the alignment. For the ‘intersection’ column, we performed N-mer intersections on the same sources, versions and assemblies of the genomes that were used for the alignments. Both sets of sequences were searched against miRbase for exact, full-length hits to mature microRNAs, and the numbers of distinct hits recorded in the table. With one or two exceptions, whenever a subsequence of an N-mer exactly matched a confirmed mature microRNA sequence, the exact matching extended into the corresponding parent hairpin up to the full N-mer length.

alignments and then searching the alignments for contiguous sequences that are both aligned and similar (22,31,32,34). A priori it is not obvious why these methods, based on large-scale sequence comparison, should necessarily be suitable for small scales and short sequence lengths. Surely, they would be successful for closely-related genomes that differ primarily by large-scale rearrangements together with a low density of substitution and short deletion/insertion events—but for more widely-divergent genomes one would expect to see indications of failure.

Binary sequence alignments for insects, human, mouse and frog were obtained from the UCSC genome browser (30) and a ternary alignment for human, mouse and rat was obtained from the VISTA genome browser (35). From these alignments were extracted all aligned and identical contiguous sequences longer than 22 bases. Our N-mer intersection method was also applied to the same sources, versions and assemblies of these genomes from which the whole-genome alignments were derived. All sequences of lengths greater than 22 from both sets were searched against mature microRNA sequence from miRbase, and the total numbers of exact hits of those sequences to confirmed mature microRNAs were recorded in Table 3 for fly/bee, fly/mosquito, human/frog, mouse/frog and human/rat/mouse alignments. With one or two exceptions, whenever a subsequence of an N-mer exactly matched a confirmed mature microRNA sequence, the exact matching extended into the corresponding parent hairpin up to the full N-mer length.

The agreement of our intersections between genome versions obtained from Ensembl and UCSC for insects is reassuring. The repeat-masking applied to Ensemble-curated genomes apparently masks more sequence than the repeat-masking applied to UCSC-curated genomes, so that the specificity for microRNAs drops considerably in intersections performed with the latter; however, our objective here is to illustrate the loss of sensitivity exhibited by the whole-genome alignments, particularly for the more distantly-related genomes, such as fly/bee—by a factor of nearly six—and human/frog and mouse/frog—by a factor of around 2. It is worth observing that (i) alignments of honeybee and frog genomes currently available at UCSC are based upon relatively early versions whose assemblies may be incomplete; (ii) most computational comparative-genomics-based microRNA discovery methods rely on whole-genome alignment, so that the current list of confirmed microRNAs is likely to be heavily biased toward those revealed by whole-genome alignment.

The role of microRNAs here is chiefly to provide a case-study. A comprehensive analysis of the recovery of known functional elements from whole-genome alignment versus N-mer intersection is certainly merited.

Discussion

Comparative genomics. Reviewed in (36), comparative genomics has a long and distinguished history in the context of both coding and non-coding sequence. As orthologous sequence from a greater diversity of organisms has become available, first in targeted regions (37) and then on the whole-genome scale, multi-species comparisons have proliferated (38) and increasingly focused on potential regulatory sequences (39). Comparative methods have been aimed at microRNA discovery since the early days of the miRNA field (40,41), with steady gains in breadth and sophistication (42), often involving homology searches followed by secondary structure characterization to establish that the flanking sequence can form a precursor-like stem–loop, and conservation profiling of flanking sequence (15,43). A recent review is contained in (44). As highly-expressed miRNAs are more readily isolated experimentally, searches for potential miRNAs expressed at low copy number, in specific tissues, in small sub-populations of cells, or only transiently in time, have relied increasingly on these computational tools (44).

Comparative methodologies for identifying potential regulatory elements based on sequence conservation have multiple parameters at their disposal: the number and choice of organisms, a window size over which sequences from different genomes are compared, and a threshold for identity (1). As these parameters are explored, it is becoming evident that the specificity, sensitivity and type of regulatory elements that can be identified depend strongly on these parameters (45). This observation is anticipated, as the effective rate of substitution (or mutation) can depend strongly on the functionality of the sequence. One consequence is that the ‘proper’ selection of appropriate species for comparison, window size, and threshold—and the consequent sensitivities and specificities—may remain, for the foreseeable future, issues that can only be resolved empirically. They may not be definitively resolved until the entire range of functional elements in genomes has been elucidated.

Microconservation. ‘Ultraconserved’ sequence was recently defined by the criteria of 100% identity, throughout windows of 50 bases or more, along orthologous sequences that have been previously aligned on the whole-genome scale, among certain groups of multiple organisms (22,23). Although other definitions are possible (46), we believe that this particular formulation of ultraconservation represents a natural starting point for a systematic and comprehensive characterization of the parameter space of comparative sequence methods.

Contiguous sequence of 50 bases or more that is exactly conserved among multiple genomes occurs at frequencies well beyond those expected on neutral drift (22,23). As mentioned above, the pioneering studies of ultraconserved sequence were based on intersections of orthologous regions of the human, mouse and rat genomes. These sequences often turned out also to be highly-conserved among other organisms, such as cow and dog. They tended to cluster near regulatory genes, and enhancer sequences were especially well-represented among them (23,46).

More recently, binary intersections of the *D.melanogaster* and *Drosophila pseudoobscura* genomes were reported (22) again based on UCSC whole-genome alignments (47). They revealed 17 miRNA sequences among over 10^6 sequences contained in that intersection between 50 and 113 bases in length; 2 of these were also contained in *A.gambiae* (22).

In contrast, we focus here entirely on sequences of length less then 50 bases, and on ternary intersections of *D.melanogaster, A.gambiae* and *A.mellifera*. Our intersections represent comparisons of every (repeat-masked) sequence in one genome against every sequence in another genome; they are truly all-against-all. No whole-genome alignment or prior screen for orthology is involved. Finally, we have found it useful to eliminate redundancy by identifying maximal sequences of each length. The specificity for miRNAs achieved here is evidently several orders of magnitude greater than observed in (22). Thus, the methodology described here is both novel and effective; to distinguish this short sequence regime from the ultraconservation regime, we term the phenomenon described here 'microconservation.'

MicroRNA discovery. In comparing this methodology to other computational approaches to miRNA discovery, it is worth stressing the assumptions about miRNA genes that are prerequisites for these other computational approaches. For example, their essential steps involve screening candidate precursor sequences for a stereotypical stem–loop secondary structure, the hallmark of a miRNA precursor; subsequently, an elaborate set of requirements, the 'Ambros criteria,' are often applied that place conditions on, among other things, the number of paired bases in the stem, the size of the loop and the length and placement of overhangs (48).

In contrast, our algorithm places what might be regarded as the simplest conceivable requirement on conservation, and isolates all sequences that satisfy this requirement down to lengths at which noise in the form of simple sequence dominates the intersection.

Had these whole-genome sequences been available 10 years ago, our method would have identified the same set of sequences that we exhibit here, whereas other miRNA discovery methods, because they rely on examples provided by known miRNAs, would not have been applicable. In a plausible sense, our algorithm is 'pristine.'

Obviously, many more of these sequences would not have been annotated 10 years ago, including all except one or two of the miRNAs. The explicit identification here of an exhaustive set of exactly conserved sequences raises, for the as yet un-annotated sequences among them, the challenges of identifying their (possibly novel) functions, if functional, or of revising our expectations for sequence conservation in the absence of selective pressure, if non-functional.

In view of our own rather limited set of assumptions, we find it remarkable that from Table 1 one finds, for example, that of the entire set of 154 distinct unfiltered N-mers with length greater than 23 bases contained in the ternary intersection, 24 of them (almost 16%) are confirmed fly miRNAs, constituting over 30% of all known fly miRNAs. It may be worth observing that it is from a practical point of view irrelevant whether sequence in mosquito or bee orthologous to confirmed fly miRNAs actually constitutes functional miRNA or not; the fact of the enrichment is significant irregardless of *why* it occurs. Of course, the success of such a naïve procedure rests on the shoulders of the genome projects. Naturally, an overall loss of specificity would be expected as the minimum length is reduced, so that in practice it would be natural to validate the functionality of the longest sequences in an intersection first—those least likely to have occurred by chance—and work systematically downwards.

Whole-genome alignment and significance of exact matches. Whole-genome alignment. The state-of-the-art in estimating genome-wide the significance of sequences highly-conserved among multiple species is represented by two recent papers describing distinct approaches to this task (20,21). Both approaches rely on whole-genome alignment as a starting point, and incorporate phylogenetic information into the assessment of conservation.

Whole-genome alignment-based methods can identify conserved sequences that our N-mer intersection method in its current form can't. For example, a subset of those sequences differing between two organisms only by mismatches and indels are identified by whole-genome alignment; however, there is no reason to expect whole-genome alignment to reveal a greater proportion of imperfect matches than it does, as suggested by Table 3, of perfect matches. In addition, whole-genome alignment methods can in principle indicate the significance of stretches of highly-conserved sequence as short as a few bases, once again beyond the capability of implementations of N-mer intersection described here.

Nevertheless, whole-genome alignment-based methods have their shortcomings, some of which are candidly discussed in (21). To that discussion, we would add four virtues of N-mer intersection over whole-genome alignment:

(i) Whole-genome alignment is defined algorithmically in (20,21). It is unclear how to specify an algorithm-independent formulation for these whole-genome alignment procedures. In contrast, N-mer intersection is not algorithmically defined—rather it is defined as 'the set of sequences common to multiple genomes'. It is not necessary to specify how this set of sequences is computed—the set exists independently of any algorithm. This property is not only conceptually desirable, but also has practical impact, as discussed below.

(ii) An immediate consequence of (1) is that in contrast to whole-genome alignment, N-mer intersection does not depend on the order in which the organisms are considered. This order-dependence is a well-known and rather disturbing aspect of whole-genome alignment (20) that is certain to complicate its statistical characterization. In principle, for *M* genomes there exist *M!* multiple whole-genome alignments, but only a unique N-mer intersection. Which of these *M!* alternatives is the best, and how do we recognize them?

(iii) A number of differing methods for whole-genome alignment have been proposed over the years. They are all defined algorithmically, and their application involves choices of parameters that make their application an art. We are confident that countless variations on the whole-genome alignments that we obtained from UCSC to produce Table 3 exist or could be produced that might well improve (or degrade) their ability to reveal the relatively short exact matches characteristic of mature miRNAs. We are not certain it would be a useful exercise to study these variations, as in the absence of a definition of what a whole-genome alignment is supposed to achieve (e.g. optimization of an objective function) it is unclear how to judge which of these alignments are superior. N-mer intersection involves a minimum of choices and assumptions, and produces a unique well-defined outcome. For a given set of genome sequences it can be accomplished once and for all.

(iv) The computational demands of whole-genome alignment are intense. The alignment of mouse and human using BLASTZ required 481 c.p.u.-days, even though it was tuned to ignore the shortest matches (2). Tools such as LAGAN and MAVID can produce highly detailed alignments, but when run at the genome scale they are jump-started, presumably for efficiency, with homology maps pre-computed by separate tools (31,49). Like MAVID, Shuffle-LAGAN uses a model for rearrangements to produce a more detailed alignment. The total time needed to align human and mouse with this tool is 25 c.p.u.-days for Shuffle-LAGAN using the CHAOS local alignment tool (50). Reliance on a rearrangement model to find short matches within a longer region of homology carries attendant biases and risks: the characterization and development of these models remains a field of active research, one that may undergo significant refinement as increasing numbers of whole genomes become available. In contrast, our methods reveal short matches with no need to model complicated order and orientation changes within longer homologies.

In contrast, an exact intersection of two mammalian genomes with our method takes 2 c.p.u.-hours and is guaranteed to catch all exact matches of the targeted length, which can in turn be readily combined to identify all longer exact matches.

We would suggest that whole-genome alignment and N-mer intersection should be viewed for the moment as complementary, rather than competing, approaches to whole-genome analysis. They have distinct strengths and weaknesses. Whole-genome alignment is surely the approach of choice for genomes that differ only by large-scale rearrangement and a sufficiently low density of substitutions, insertions and deletions. Under those conditions, one might expect the outcome of the alignment procedure to be order-independent and more-or-less unique; N-mer intersection would be futile. As the genomes diverge further, uniqueness will be lost and distinct but equally valid alignments will ensue depending on details of the algorithm and the parameter settings. In that regime, N-mer intersection can identify matches that are lost to whole-genome alignment, such as those exhibited in Table 3.

Whole-genome alignments constitute a major scientific achievement that will continue to have significant impact on our understanding of molecular, cellular, genome and systems biology. The observation that they possess certain fundamental and practical limitations should not be seen as detracting from this achievement.

Estimating significance of exact matches. The incorporation of phylogeny into significance estimates clearly represents an important step forward. We discuss the methods of Cooper *et al.* (20) and Siepel *et al.* (21) separately.

Cooper et al. (20). In (20) a multiple alignment of 29 mammalian genomes was constructed and analyzed as follows: (a) whole genomes are chopped into pieces and aligned with the algorithms LAGAN and shuffle-LAGAN (51). (b) Gapped positions are removed together with anomalously long stretches of perfect matches and exceptionally poor matches. (c) A position-by-position maximum-likelihood optimization of expected substitution rate is performed upon the remaining sequences. (d) At each position, 'rejected substitutions' are computed as the log-ratio of expected over observed substitution rates (logRS). (e) The logRS are summed over runs of positions with sufficiently high logRS, and if the sum exceeds a specified threshold, the run is classified as a potential 'highly-conserved' sequence. (f) Random spatial permutations of the rate ratios are taken (the 'null model') and step (e) is applied to this null model. (g) The total number of bases obtained from (e) and (f) are compared to assess significance.

The whole-genome alignment plays a critical role in identifying positions with differing logRS, and thus in enabling significance to be estimated in (g). No estimate is made of what proportion of sequences has been lost in the process of multiple alignment so that it is unclear how reliable and estimate is yielded of the number of highly-conserved sequences expected by chance. Futhermore it is not obvious how to apply (f) to our N-mer intersections, where by definition 100% conservation is required at each base of the N-mer.

Finally, the validity of step (f) (the 'null model') depends on the absence of significant positional correlations along the sequence. The presence of strong positional correlations in non-coding DNA has been recognized for almost fifteen years (52–54); we have demonstrated elsewhere (Miller,J. and Havlak,P., personal communication, 2005; Salerno,W., Havlak,P. and Miller,J., manuscript submitted) their decisive contribution to the length distribution of (non-repetitive) highly-conserved sequence from N-mer intersections and whole-genome alignments, including the data described in (20). In short, the 'null model' of (20) is inappropriate (except possibly for coding sequence) and we anticipate that the path to reliable estimation of significance will involve, at a minimum, reformulation of (f).

Siepel et al. (21). Phylogenetic hidden Markov models (55) have proven themselves elegant and stunningly effective tools for quantifying the conservation of multiply-aligned sequence. Siepel *et al.* (21) apply their phylo-HMM subsequent to whole-genome alignment, and their computations are therefore subject to some of the same reservations expressed above. More disturbing is the bias towards coding exons that they have explicitly built into their models and parameters. Although this bias may be appropriate for unicellular organisms, we have observed that our human/dog/frog N-mer intersections (Miller,J. and Havlak,P., personal communication, 2005; Tran,T. Havlak,P. and Miller,J., manuscript submitted: Functional element enrichment in human/dog/frog intersections) while strongly enriching for (known) non-coding functional elements, are almost neutral with respect to coding sequence. Furthermore the strong positional correlations mentioned above are absent in coding exons, yet these correlations dominate the length distribution of ultraconserved sequences from vertebrate genomes (Miller,J. and Havlak,P., personal communication, 2005). Finally, as the authors observe, positional correlations are not accounted for in their computations, and they assume a 'null model' similar in spirit to that of (20).

Limitations: inexact matches. In performing exact intersections among multiple genomes we have sacrificed sensitivity for specificity. For example, the mosquito *A.gambiae* let-7 miRNA differs by one internal base substitution from the identical mature let-7 miRNAs of fly and bee (28); the same substitution is present in *Aedes aegypti* (Anzola,J. and Elsik,C. personal communication). Supplementary Table 4 shows all fly mature microRNA sequences from miRbase together with their confirmed orthologs from mosquito and bee; these data are also summarized there together with the positions of the base substitutions. Of the 78 fly miRNAs, 37 have no (as yet confirmed) orthologs in mosquito or bee; 14 are shared exactly among all three organisms; 29 among two organisms only. A total of 15 of them differ by one base from an ortholog; four by two bases; two by three bases; and one by four bases (obviously these sets are not mutually exclusive). Thus such internal base substitutions certainly account for many of the confirmed insect miRNAs absent from our intersections, perhaps for some fly miRNAs not presently known to have orthologs in bee or mosquito, and possibly for some novel ones as well; we are developing multiple strategies to recover them.

Conclusions. We believe this study is of interest for several reasons, both fundamental and practical:

(i) The characterization of miRNAs as a primary component of the maximal 23 to 29mers may reflect distinctive evolutionary forces.
(ii) Other constituents of these sets, both un-annotated and previously annotated, are likely to be functional and merit further characterization. For example, it is possible that sequences annotated as coding are in some cases exactly conserved because they also have thus-far unidentified functions, perhaps as regulatory sequence, non-coding RNA or miRNA target. 'Simple' sequence and/or sequence annotated only as repeat may also be worth further study.
(iii) The methodology developed here may prove useful in identifying novel functional sequence elements and non-coding RNAs, as well as miRNAs in other organisms whose complement of miRNAs remains to be fully elucidated.
(iv) Since the method does not rely on prior orthology or whole-genome alignment, it can be applied to unassembled genomes.
(v) This study represents a prototype for exhaustive, all-against-all comparisons in which the stringency of conservation is systematically reduced to allow inexact matching ('indels') and approximate intersection.
(vi) It raises the possibility of novel approaches to eukaryotic phylogeny (56), e.g. 'N-mer phylogenies'.
(vii) The method may offer the potential to improve the recovery of microconserved sequence in whole-genome alignment.

Each of these threads is being pursued in our laboratory; in particular, lists of maximal N-mers for binary and ternary intersections of all publicly available whole-genome sequences are currently under production and we anticipate they will eventually be available on a public website. Finally, we remark that miRNAs are also significantly enriched in ternary fish intersections, and intersections of higher vertebrate genomes (Tran,T. Havlak,P. and Miller,J., manuscript submitted).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

The authors are grateful to J. Reid, W. Salerno, D. Graur and A. McCullough for advice and discussion; to the Honeybee consortium for permission to use their unpublished sequence assemblies from the Ensembl website; to K. Worley for advice on annotation. P.H. was supported in part through the grant, 'Large-scale sequencing at BCM-HGSC' from NIH-NHGRI (1 U01 HG003273). The project was conceived in discussions between J.M. and Yuriy Fofanov of the Department of Computer Science, University of Houston, and we are extremely grateful for his suggestions and advice. Funding to pay the Open Access publication charges for this article was provided by the authors.

Conflict of interest statement. None declared.

REFERENCES

1. Stone,E.A., Cooper,G.M. and Sidow,A. (2005) Trade-offs in detecting evolutionarily constrained sequence by comparative genomics. *Annu. Rev. Genomics. Hum. Genet.*, **6**, 143–164.
2. Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D. and Miller,W. (2003) Human-mouse alignments with BLASTZ. *Genome Res.*, **13**, 103–107.
3. Mattick,J.S. (2004) The hidden genetic program of complex organisms. *Sci. Am.*, **291**, 60–67.
4. Mattick,J.S. (2005) The functional genomics of noncoding RNA. *Science*, **309**, 1527–1528.
5. Mattick,J.S. and Makunin,I.V. (2005) Small regulatory RNAs in mammals. *Hum. Mol. Genet.*, **14**, R121–R132.
6. Cullen,B.R. (2004) Transcription and processing of human microRNA precursors. *Mol. Cell*, **16**, 861–865.
7. Ambros,V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350–355.
8. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
9. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
10. Ambros,V. and Horvitz,H.R. (1984) Heterochronic mutants of the nematode *Caenorhabditis elegans*. *Science*, **226**, 409–416.

11. Cimmino,A., Calin,G.A., Fabbri,M., Iorio,M.V., Ferracin,M., Shimizu,M., Wojcik,S.E., Aqeilan,R.I., Zupo,S., Dono,M. *et al.* (2005) miR-15 and miR-16 induce apoptosis by targeting BCL2. *Proc. Natl Acad. Sci. USA*, **102**, 13944–13949.
12. Lu,J., Getz,G., Miska,E.A., Alvarez-Saavedra,E., Lamb,J., Peck,D., Sweet-Cordero,A., Ebert,B.L., Mak,R.H., Ferrando,A.A. *et al.* (2005) MicroRNA expression profiles classify human cancers. *Nature*, **435**, 834–838.
13. Hatfield,S.D., Shcherbata,H.R., Fischer,K.A., Nakahara,K., Carthew,R.W. and Ruohola-Baker,H. (2005) Stem cell division is regulated by the microRNA pathway. *Nature*, **435**, 974–978.
14. Lee,R.C., Feinbaum,R.L. and Ambros,V. (1993) The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.
15. Berezikov,E., Guryev,V., van de Belt,J., Wienholds,E., Plasterk,R.H. and Cuppen,E. (2005) Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, **120**, 21–24.
16. Reinhart,B.J., Slack,F.J., Basson,M., Pasquinelli,A.E., Bettinger,J.C., Rougvie,A.E., Horvitz,H.R. and Ruvkun,G. (2000) The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature*, **403**, 901–906.
17. Pasquinelli,A.E., Reinhart,B.J., Slack,F., Martindale,M.Q., Kuroda,M.I., Maller,B., Hayward,D.C., Ball,E.E., Degnan,B., Muller,P. *et al.* (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, **408**, 86–89.
18. Pasquinelli,A.E., McCoy,A., Jimenez,E., Salo,E., Ruvkun,G., Martindale,M.Q. and Baguna,J. (2003) Expression of the 22 nucleotide let-7 heterochronic RNA throughout the Metazoa: a role in life history evolution? *Evol. Dev.*, **5**, 372–378.
19. Chen,Y.-H., Nyeo,S-L. and Yeh,C-Y. (2005) Model for the distributions of k-mers in DNA sequences. *Phys. Rev. E*, **72**, 011908.
20. Cooper,G.M., Stone,E.A., Asimenos,G., Green,E.D., Batzoglou,S. and Sidow,A. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15**, 901–913.
21. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
22. Glazov,E.A., Pheasant,M., McGraw,E.A., Bejerano,G. and Mattick,J.S. (2005) Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. *Genome Res.*, **15**, 800–808.
23. Hillier,L.W., Miller,W., Birney,E., Warren,W., Hardison,R.C., Ponting,C.P., Bork,P., Burt,D.W., Groenen,M.A., Delany,M.E. *et al.* (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**, 695–716.
24. Smit,A.F.A., Hubley,R. and Green,P. RepeatMasker Open-3.0.
25. Havlak,P., Chen,R., Durbin,K.J., Egan,A., Ren,Y., Song,X.Z., Weinstock,G.M. and Gibbs,R.A. (2004) The Atlas genome assembly system. *Genome Res.*, **14**, 721–732.
26. Ning,Z., Cox,A.J. and Mullikin,J.C. (2001) SSAHA: a fast search method for large DNA databases. *Genome Res.*, **11**, 1725–1729.
27. Drysdale,R.A. and Crosby,M.A. (2005) FlyBase: genes and gene models. *Nucleic Acids Res.*, **33**, D390–D395.
28. Griffiths-Jones,S. (2004) The microRNA registry. *Nucleic Acids Res.*, **32**, D109–D111.
29. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
30. 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. *et al.* (2003) The UCSC genome browser database. *Nucleic Acids Res.*, **31**, 51–54.
31. Brudno,M., Poliakov,A., Salamov,A., Cooper,G.M., Sidow,A., Rubin,E.M., Solovyev,V., Batzoglou,S. and Dubchak,I. (2004) Automated whole-genome multiple alignment of rat, mouse, and human. *Genome Res.*, **14**, 685–692.
32. Brudno,M., Do,C.B., Cooper,G.M., Kim,M.F., Davydov,E., Green,E.D., Sidow,A. and Batzoglou,S. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.*, **13**, 721–731.
33. Fofanov,Y., Luo,Y., Katili,C., Wang,J., Belosludtsev,Y., Powdrill,T., Belapurkar,C., Fofanov,V., Li,T.B., Chumakov,S. *et al.* (2004) How independent are the appearances of n-mers in different genomes? *Bioinformatics*, **20**, 2421–2428.
34. Bejerano,G., Pheasant,M., Makunin,I., Stephen,S., Kent,W.J., Mattick,J.S. and Haussler,D. (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321–1325.
35. Frazer,K.A., Pachter,L., Poliakov,A., Rubin,E.M. and Dubchak,I. (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res.*, **32**, W273–W279.
36. Miller,W., Makova,K.D., Nekrutenko,A. and Hardison,R.C. (2004) Comparative genomics. *Annu. Rev. Genomics Hum. Genet.*, **5**, 15–56.
37. Thomas,J.W., Touchman,J.W., Blakesley,R.W., Bouffard,G.G., Beckstrom-Sternberg,S.M., Margulies,E.H., Blanchette,M., Siepel,A.C., Thomas,P.J., McDowell,J.C. *et al.* (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, **424**, 788–793.
38. Kolbe,D., Taylor,J., Elnitski,L., Eswara,P., Li,J., Miller,W., Hardison,R. and Chiaromonte,F. (2004) Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome Res.*, **14**, 700–707.
39. Elnitski,L., Hardison,R.C., Li,J., Yang,S., Kolbe,D., Eswara,P., O'Connor,M.J., Schwartz,S., Miller,W. and Chiaromonte,F. (2003) Distinguishing regulatory DNA from neutral sites. *Genome Res.*, **13**, 64–72.
40. Lai,E.C., Tomancak,P., Williams,R.W. and Rubin,G.M. (2003) Computational identification of *Drosophila* microRNA genes. *Genome Biol.*, **4**, R42.
41. Grad,Y., Aach,J., Hayes,G.D., Reinhart,B.J., Church,G.M., Ruvkun,G. and Kim,J. (2003) Computational and experimental identification of *C. elegans* microRNAs. *Mol. Cell*, **11**, 1253–1263.
42. Weber,M.J. (2005) New human and mouse microRNA genes found by homology search. *FEBS. J.*, **272**, 59–73.
43. Legendre,M., Lambert,A. and Gautheret,D. (2005) Profile-based detection of microRNA precursors in animal genomes. *Bioinformatics*, **27**, 841–845.
44. Brown,J.R. and Sanseau,P. (2005) A computational view of microRNAs and their targets. *Drug. Discov. Today*, **10**, 595–601.
45. Boffelli,D., Nobrega,M.A. and Rubin,E.M. (2004) Comparative genomics at the vertebrate extremes. *Nature Rev. Genet.*, **5**, 456–465.
46. Sandelin,A., Bailey,P., Bruce,S., Engstrom,P.G., Klos,J.M., Wasserman,W.W., Ericson,J. and Lenhard,B. (2004) Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics*, **5**, 99.
47. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2004) The UCSC table browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
48. Ambros,V., Bartel,B., Bartel,D.P., Burge,C.B., Carrington,J.C., Chen,X., Dreyfuss,G., Eddy,S.R., Griffiths-Jones,S., Marshall,M. *et al.* (2003) A uniform system for microRNA annotation. *RNA.*, **9**, 277–279.
49. Bray,N. and Pachter,L. (2004) MAVID: constrained ancestral alignment of multiple sequences. *Genome Res.*, **14**, 693–699.
50. Brudno,M., Malde,S., Poliakov,A., Do,C.B., Couronne,O., Dubchak,I. and Batzoglou,S. (2003) Glocal alignment: finding rearrangements during alignment. *Bioinformatics*, **19**, i54–i62.
51. Brudno,M., Do,C., Cooper,G.M., Kim,M.F., Davydov,E., Green,E.D., Sidow,A. and Batzoglou,S. (2003) NISC Comparative Sequencing Program LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genomic Res.*, **13**, 721–731.
52. Voss,R.F. (1992) Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Phys. Rev. Lett.*, **68**, 3805–3808.
53. Li,W. and Kaneko,K. (1992) Long-range correlation and partial 1/f-alpha spectrum in a noncoding DNA-sequence. *Europhys. Lett.*, **17**, 655–660.
54. Peng,C.K., Buldyrev,S.V., Goldberger,A.L., Havlin,S., Sciortino,F., Simons,M. and Stanley,H.E. (1992) Long-range correlations in nucleotide sequences. *Nature*, **356**, 168–170.
55. Siepel,A. and Haussler,D. (2004) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.*, **21**, 468–488.
56. Zdobnov,E.M., von Mering,C., Letunic,I. and Bork,P. (2005) Consistency of genome-based methods in measuring Metazoan evolution. *FEBS Lett.*, **579**, 3355–3361.