

Identification of hundreds of conserved and nonconserved human microRNAs

Isaac Bentwich^{1,2}, Amir Avniel^{1,2}, Yael Karov^{1,2}, Ranit Aharonov^{1,2}, Shlomit Gilad^{1,2}, Omer Barad¹, Adi Barzilai¹, Paz Einat¹, Uri Einav¹, Eti Meiri¹, Eilon Sharon¹, Yael Spector¹ & Zvi Bentwich¹

MicroRNAs are noncoding RNAs of ~22 nucleotides that suppress translation of target genes by binding to their mRNA and thus have a central role in gene regulation in health and disease^{1–5}. To date, 222 human microRNAs have been identified⁶, 86 by random cloning and sequencing, 43 by computational approaches and the rest as putative microRNAs homologous to microRNAs in other species. To prove our hypothesis that the total number of microRNAs may be much larger and that several have emerged only in primates, we developed an integrative approach combining bioinformatic predictions with microarray analysis and sequence-directed cloning. Here we report the use of this approach to clone and

sequence 89 new human microRNAs (nearly doubling the current number of sequenced human microRNAs), 53 of which are not conserved beyond primates. These findings suggest that the total number of human microRNAs is at least 800.

We developed microRNA discovery tools that detect microRNAs missed by existing methods, which detect only conserved hairpins. Our approach (Fig. 1a) comprises the following steps: (i) computationally scanning the entire human genome for hairpin structures; (ii) annotating all hairpins for conserved, repetitive and protein-coding regions; (iii) scoring hairpins by thermodynamic stability and structural features, using a method (PalGrade) that detects a large percentage

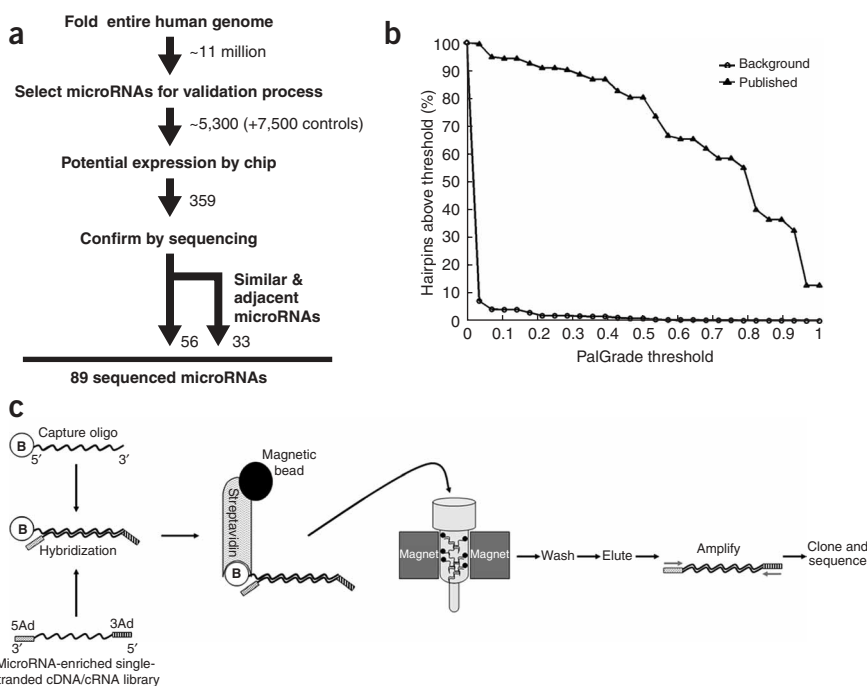


Figure 1 MicroRNA detection and validation. **(a)** The microRNA discovery approach. The initial folding of the entire human genome resulted in ~11 million hairpins. After microarray sampling, 359 microRNAs were subjected to confirmation by sequencing, identifying 89 human microRNAs that did not appear in the microRNA registry: 56 from the main pipeline and 33 additional similar and adjacent microRNAs. **(b)** Hairpin scoring algorithm performance. Percentages of known microRNA hairpins (triangles) and of all genome hairpins (circles) above or equal to different PalGrade thresholds (*x* axis). Hairpins that are not on repetitive or protein-coding regions were considered. The large separation indicates the high sensitivity and specificity (accuracy) of the scoring method. **(c)** The microRNA sequence-directed cloning method. A population of single-stranded molecules derived from the microRNA-enriched library is mixed with the biotinylated capture oligonucleotide. After hybridization, streptavidin bound to magnetic beads is added, and the mixture is loaded into a column mounted on a strong magnet. The column is then washed stringently to remove nonbound or weakly hybridized molecules. The specifically bound molecules are eluted, amplified, cloned and sequenced.

¹Rosetta Genomics, 10 Plaut Street, Science Park, Rehovot 76706, Israel. ²These authors contributed equally to this work. Correspondence should be addressed to I.B. (bentwich@rosettagenomics.com).

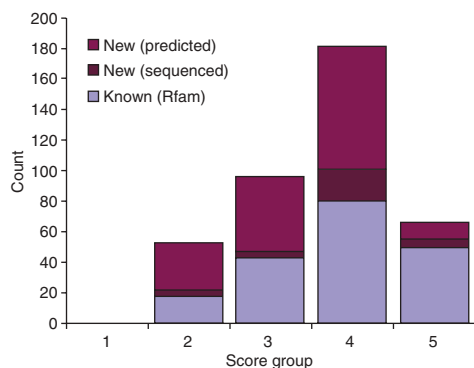


Figure 2 The number of conserved microRNAs in the human genome. Known conserved microRNAs⁶ (blue) and new conserved microRNAs (purple), including those validated by us (dashed), for each score group (PalGrade). The five score groups are composed of the following PalGrade ranges: 1, 0 (control group); 2, 0–0.25; 3, 0.25–0.55; 4, 0.55–0.93; and 5, 0.93–1. The number of new microRNAs is the number of hairpins excluding known microRNAs of same score group in the genome, multiplied by the validation success rate in a random sample. The validation success rate is the percentage cloned and sequenced from a sample taken from the group, after deliberate underestimations where success rate was below 5%, divided by the validation success rate of the known microRNAs (76%), to correct for undersampling of tissues (known, similar and adjacent microRNAs were excluded from analysis to avoid positive bias). Because 14% of the known microRNAs were not in the initial group and, hence, are in none of the score groups, these numbers should be divided by 0.86. Thus, the total projected number of conserved microRNAs is 460 (220 + 240).

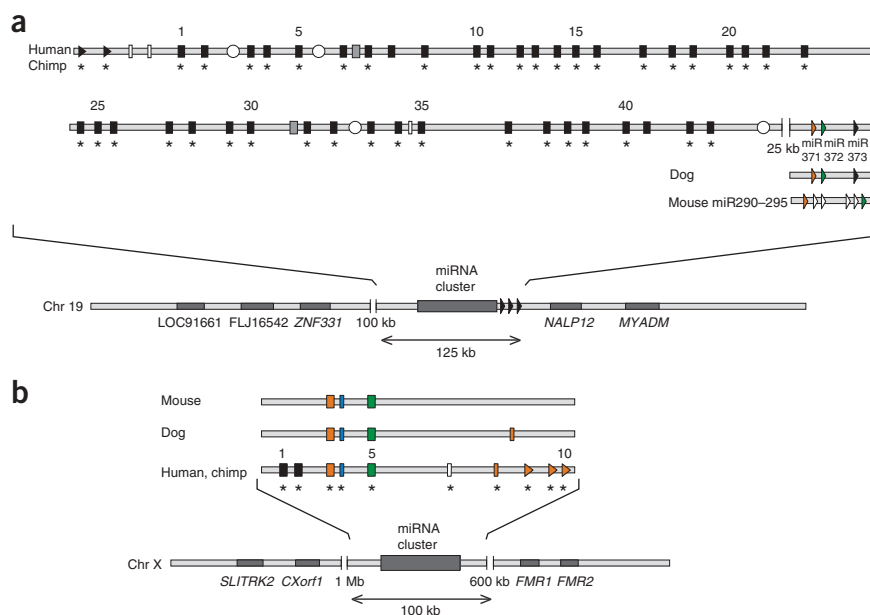
of known microRNAs while selecting a relatively small portion of all genome hairpins (**Fig. 1b**); (iv) determining the expression of computationally predicted microRNAs by a high-throughput microRNA microarray in several tissues (placenta, testis, thymus, brain and prostate); and (v) validating the sequence of predicted microRNAs that gave high signals on the microarray using a new sequence-directed cloning and sequencing method. This method uses a specific biotinylated capture oligonucleotide, designed for the predicted microRNA to be cloned, to 'fish out' the complementary sequences from the microRNA-enriched libraries, which are then amplified, cloned and sequenced (**Fig. 1c**). This high-throughput approach has enabled us to detect a substantially higher percentage of all existing microRNAs, by dealing effectively with large groups of hairpins that have a relatively high percentage of false positives.

Scanning the entire human genome identified ~11 million hairpins, including 86% of known microRNA precursors. Of all hairpins, 434,239 passed a minimal hairpin score threshold (PalGrade score > 0) and were not located on repetitive elements or protein-coding

regions. This smaller group, the initial candidate group, retains 86% of known microRNAs, suggesting that systematic scanning would detect 86% of all microRNAs. We then divided all hairpins into conserved and nonconserved hairpins, using a criterion by which 220 of the 222 known microRNAs are conserved. From the initial candidate group we selected ~5,300 predicted microRNA sequences for high-throughput expression validation by microarray⁷. These included randomly sampled hairpins from the following groups: conserved hairpins from different PalGrade score groups (~1,500), nonconserved clustered hairpins from different PalGrade score groups (~800) and nonconserved nonclustered hairpins (~3,000). We also used a control group of ~7,500 hairpins that were not included in the initial candidate group to test various aspects of the prediction approach.

Microarray experiments in placenta, testis, thymus, brain and prostate resulted in 886 candidate microRNAs with significant signals ($P < 0.06$) of at least one of their two predicted mature microRNAs. We subjected 359 of these 886 candidate microRNAs to sequence validation using our new sequence-directed cloning and sequencing

Figure 3 Two new nonconserved microRNA clusters. **(a)** Cluster on chromosome 19 located at positions 58,861,745–58,961,404 (HG17) and comprising 54 microRNA genes grouped into four families on the basis of hairpin sequence similarity (circles, black bars, gray bars and triangles; white bars indicate nongrouped microRNAs). Of these, 43 have been cloned and sequenced (asterisks). The microRNAs of this large family are numbered (for clarity, numbers are shown only for every five microRNAs); numbers match those indicated in **Figure 4a**. The adjacent *mir-371,2,3* cluster is depicted, with its conserved hairpins found in dog and mouse. This cluster is also conserved in rat. **(b)** Cluster on chromosome X located at positions 145,967,859–146,072,859 (HG17). The cluster contains ten cloned and sequenced (asterisks) microRNAs grouped into two families on the basis of hairpin sequence similarity (wide bars and triangles; thin bars indicate nongrouped microRNAs). The microRNAs of the cluster are numbered; numbers match those indicated in **Figure 4b**. Seven precursors in the cluster have homologous hairpins in dog (convergent mapping) and three in mouse (as well as in rat), as indicated by matching colors. The white bars have not mapped to any sequence that can be folded into a hairpin. Neighboring known genes are shown for both clusters. The full list of the microRNA sequences in these clusters is given in **Supplementary Table 1** online.



a

```

hsa-miR-520e : --AAATCCCTCCCTTTGACGG-- 2
hsa-miR-520f : --CAATCCCTCCCTTTAGAGGTT 5
hsa-miR-520a : --AAATCCCTCCCTTTGACGTT-- 9
hsa-miR-520b : --AAATCCCTCCCTTTGACGG-- 15
hsa-miR-520c : --AAATCCCTCCCTTTGACGTT-- 18
hsa-miR-520d : --AAATCCCTCCCTTTGACGTT 24
hsa-miR-520g : ACATACTCCCTCCCTTTGACGTT-- 26
hsa-miR-520h : ACATACTCCCTCCCTTTGACGTT-- 35
hsa-miR-515-3p : --GATCCCTCCCTTTGACGCT-- 3
hsa-miR-515-3p : --GATCCCTCCCTTTGACGCT-- 6
hsa-miR-516-3p : -----TCCTCCCTCCCTTTGACGTT-- 27
hsa-miR-516-3p : -----TCCTCCCTCCCTTTGACGTT-- 32
hsa-miR-516-3p : -----TCCTCCCTCCCTTTGACGTT-- 40
hsa-miR-516-3p : -----TCCTCCCTCCCTTTGACGTT-- 42
hsa-miR-519e : --AAATCCCTCCCTTTGACGTT-- 4
hsa-miR-519d : --CAATCCCTCCCTTTGACGTT-- 22
hsa-miR-519c : --AAATCCCTCCCTTTGACGTT-- 7
hsa-miR-519b : --AAATCCCTCCCTTTGACGTT 11
hsa-miR-519a : --AAATCCCTCCCTTTGACGTT-- 38
hsa-miR-519a : --AAATCCCTCCCTTTGACGTT-- 43
hsa-miR-517a : --TCCTCCCTCCCTTTGACGTT 21
hsa-miR-517b : --TCCTCCCTCCCTTTGACGTT 25
hsa-miR-518f : --GAAATCCCTCCCTTTGACGTT-- 14
hsa-miR-518b : --CAATCCCTCCCTTTGACGTT-- 16
hsa-miR-518c : --CAATCCCTCCCTTTGACGTT-- 19
hsa-miR-518e : --AAATCCCTCCCTTTGACGTT-- 29
hsa-miR-518a : --AAATCCCTCCCTTTGACGTT-- 30
hsa-miR-518d : --CAATCCCTCCCTTTGACGTT-- 31
hsa-miR-518a : --AAATCCCTCCCTTTGACGTT-- 33
hsa-miR-524 : --GAAATCCCTCCCTTTGACGTT-- 20
hsa-miR-525* : --GAAATCCCTCCCTTTGACGTT-- 12
hsa-miR-521 : --AAATCCCTCCCTTTGACGTT-- 23
hsa-miR-521 : --AAATCCCTCCCTTTGACGTT-- 26
hsa-miR-517c : --TCCTCCCTCCCTTTGACGTT-- 34
hsa-miR-522 : --AAATCCCTCCCTTTGACGTT 37
hsa-miR-523 : --GAAATCCCTCCCTTTGACGTT-- 13

hsa-miR-526c : -----TCCTCCCTCCCTTTGACGTT 11
hsa-miR-526a : -----TCCTCCCTCCCTTTGACGTT 17
hsa-miR-520c* : -----TCCTCCCTCCCTTTGACGTT 18
hsa-miR-519c* : -----TCCTCCCTCCCTTTGACGTT 7
hsa-miR-523* : -----TCCTCCCTCCCTTTGACGTT 13
hsa-miR-518f* : -----TCCTCCCTCCCTTTGACGTT 14
hsa-miR-526a : -----TCCTCCCTCCCTTTGACGTT 28
hsa-miR-518e* : -----TCCTCCCTCCCTTTGACGTT 29
hsa-miR-518d* : -----TCCTCCCTCCCTTTGACGTT 31
hsa-miR-522* : -----TCCTCCCTCCCTTTGACGTT 37
hsa-miR-519a* : -----TCCTCCCTCCCTTTGACGTT 38
hsa-miR-517b* : -----TCCTCCCTCCCTTTGACGTT 25
hsa-miR-520a* : -----TCCTCCCTCCCTTTGACGTT 9
hsa-miR-525 : -----TCCTCCCTCCCTTTGACGTT 12
hsa-miR-515-5p : -----TCCTCCCTCCCTTTGACGTT 3
hsa-miR-519e* : ATTTCCCTCCCTTTGACGTT-- 4
hsa-miR-515-5p : -----TCCTCCCTCCCTTTGACGTT 6
hsa-miR-518a* : -----TCCTCCCTCCCTTTGACGTT 30
hsa-miR-527 : -----TCCTCCCTCCCTTTGACGTT 39
hsa-miR-518a* : -----TCCTCCCTCCCTTTGACGTT 33
hsa-miR-524* : -----TCCTCCCTCCCTTTGACGTT 20
hsa-miR-520d* : -----TCCTCCCTCCCTTTGACGTT 24
hsa-miR-516c* : -----TCCTCCCTCCCTTTGACGTT 19
hsa-miR-516-5p : -----TCCTCCCTCCCTTTGACGTT 27
hsa-miR-516-5p : -----TCCTCCCTCCCTTTGACGTT 32
hsa-miR-517a* : -----TCCTCCCTCCCTTTGACGTT 21
hsa-miR-517c* : -----TCCTCCCTCCCTTTGACGTT 34
hsa-miR-526b : -----TCCTCCCTCCCTTTGACGTT 10
hsa-miR-498 : -----TCCTCCCTCCCTTTGACGTT 1

```

b

```

hsa-miR-514 : --ATTTCACCTTCTGAGGATG-- 8
hsa-miR-514 : --ATTTCACCTTCTGAGGATG-- 9
hsa-miR-514 : --ATTTCACCTTCTGAGGATG-- 10
hsa-miR-506 : --GTAAAGGACCTTCTGAGGATG 3
hsa-miR-507 : --TTTGGACCTTCTGAGGATG 4
hsa-miR-508 : TGTATCTAGCTTCTGAGGATG 5
hsa-miR-509 : TGTATCTAGCTTCTGAGGATG 6

hsa-miR-513 : TTTCACGGAGGTTGCAATTAA-- 1
hsa-miR-513 : TTTCACGGAGGTTGCAATTAA-- 2
hsa-miR-510 : TTTCACGGAGGTTGCAATTAA 7

```

Figure 4 Multiple sequence alignments of cloned mature microRNAs from the two new nonconserved clusters. The multiple alignments were generated by the ClustalW program. Conserved nucleotides are colored as follows: black, 100%; dark gray, 80–99%; and light gray, 60–79%. The miRNA name obtained from the microRNA registry⁶ is shown to the left of each sequence. The numbers to the right of each sequence match the numbers indicated in **Figure 3**. (a) Multiple sequence alignment of cloned mature microRNAs from the highly related family in the cluster on chromosome 19 derived from the 3' stem (left column) and 5' stem (right column) of the precursors. The microRNAs are presented in groups by the 16 distinct seeds they generate (a seed is defined as nucleotides 2–8 of the mature microRNA). Mature microRNAs cloned from the other arm of precursors from which highly expressed microRNAs were cloned are marked with asterisks. (b) Multiple sequence alignment of cloned mature microRNAs from the cluster on chromosome X derived from the 3' stem (left column) and 5' stem (right column) of the precursors. The microRNAs are presented in groups by the seven distinct seeds they generate.

is the largest microRNA cluster ever reported and comprises 54 new predicted microRNAs, 43 of which we cloned and sequenced (**Fig. 3a** and **Supplementary Tables 1** and **2** online). These 54 microRNAs show similarity to the neighboring *mir-371,2,3* family (**Fig. 3a**) specifically expressed in human embryonic stem cells¹¹. Although they are highly similar, they generate 16 distinct 'seeds' (i.e., nucleotides 2–8 of the mature microRNA; **Fig. 4a**). Homology analysis showed that the cluster as a whole is conserved only in chimpanzee and possibly rhesus monkey (*Macaca mulatta*, whose genome is not yet assembled; hence, although all microRNAs are found there, the cluster structure there is not definite) and that none of its microRNA members show any homology to nonprimate genomes. Notably, the adjacent *mir-371,2,3* cluster is conserved in other mammals, although their conservation score was the lowest among the known microRNAs, and one of these microRNAs did not pass the conservation criterion (**Fig. 3a** and ref. 12).

The second cluster is located on chromosome X near the gene *FMR1* and includes ten microRNAs, which are expressed only in testis and all of which we cloned and sequenced (**Fig. 3b** and **Supplementary Table 2** online). These ten microRNAs also form a family of related sequences, which generate seven distinct seeds (**Figs. 3b** and **4b**). The cluster as a whole is conserved only in chimpanzee and possibly rhesus monkey. Seven and four of its members are conserved and clustered in dog and mouse or rat, respectively, although the seven human hairpins converge onto only four hairpins in the dog genome (**Fig. 3b**).

Both clusters differ significantly from known microRNA clusters, all of which are found as a whole in all other mammals (except *mir-371,2,3*, which exists as a whole in dog and partially in rat and mouse; **Fig. 3a**). In addition, none of the microRNAs in the two new clusters has a conservation score passing the criterion discussed above and most cannot be mapped at all to nonprimate genomes. The high cluster-member similarity in both clusters, which are fully conserved only in primates; the convergent mapping of members of the second cluster; and the fact that many of the microRNAs in the first cluster are embedded in long (400–700 nucleotides) sequences that are repeated along the cluster suggest that both clusters evolved through

method (60 of 90 conserved hairpins from different PalGrade score groups; 50 of 72 nonconserved clustered hairpins from different PalGrade score groups; 59 of 161 randomly selected from nonconserved nonclustered hairpins selected from different PalGrade score groups; and 190 of 563 randomly selected from the control group). In some cases, the cloning and sequencing method resulted in sequencing of similar microRNAs that were slightly different in sequence from the microRNA originally sought. We also carried out sequence validation on 69 bioinformatically predicted microRNAs, which were not present on the microarray but are located adjacent to microRNAs that were successfully sequenced, resulting in more new sequences called adjacent microRNAs.

Using this approach, we successfully cloned and sequenced 89 human microRNA genes that do not appear in the microRNA registry⁶ (version 5.1; **Supplementary Table 1** and **Supplementary Figs. 1** and **2** online). Of these, 56 were found through the method's main pipeline (i.e., they were part of the original samples), and 33 are either similar or adjacent microRNAs (these 33 were ignored in calculating success rates). Only 1 of the 89 emerged from the large control group, supporting the distinction between the initial candidate group and the remaining >10 million genomic hairpins (**Fig. 2**). Thirty-two of the 36 conserved microRNAs discovered and sequenced by our approach also score highly on MirScan⁸ (**Supplementary Table 1** online). Bioinformatic predictions of conserved microRNAs published after we obtained our results include 32 of these 36 conserved microRNAs (12 appear in the 958 predictions of Berezikov *et al.*⁹, 10 in the 129 predictions of Xie *et al.*¹⁰ and 10 in both; of these, 8 were validated by northern-blot or primer extension analysis).

Fifty three of the new microRNAs that we found and sequenced are located in two large nonconserved clusters (24 of these were in the original sample and 29 were found by searching for adjacent microRNAs; **Fig. 3** and **Supplementary Methods** online). One of the clusters, located on chromosome 19 and expressed only in placenta,

duplication and mutation events unique to primates. Therefore, we report a new class of microRNAs, nonconserved clustered microRNAs, that are not detected using other microRNA detection approaches and were not taken into consideration in previous estimates of the total number of microRNAs in the genome.

The 89 cloned and sequenced microRNAs that we report bring the total number of human microRNAs to 311, well above a previously stated upper bound of 255 (ref. 4). Moreover, our method allows an estimation of the total number of both conserved and nonconserved microRNA classes. We computed the percentage of true microRNAs comprised by each class using two independent methods: (i) using only informatic data, based on probabilistic arguments, and (ii) calculating the validation success rate in samples from each score group, ignoring similar and adjacent microRNAs to avoid positive bias, and then multiplying the validation success rate by the number of hairpins in the genome belonging to that group (**Supplementary Methods** online). The validation success rate strongly correlated with PalGrade.

The expected number of conserved microRNAs is at least 442 using the probabilistic approach, and at least 460 on the basis of validation success rates, using deliberate underestimations of success rates to account for noisy statistics and correcting for the 14% known microRNAs that were not in the initial group (**Fig. 2** and **Supplementary Methods** online). Estimating the total number of nonconserved microRNAs is more difficult, because this group is new and, hence, less well characterized. We therefore focused on nonconserved hairpins present in a cluster and, as above, obtained two independent estimates for this subgroup. On the basis of validation success rates, these are estimated to be at least 159 nonconserved hairpins. This is probably an underestimate, because our accounting for undersampling of tissues was based on known microRNAs and nonconserved clusters have much higher tissue specificity (**Supplementary Table 2** online). Using the probabilistic approach, we estimate that there are at least 336 nonconserved clustered human microRNAs (**Supplementary Methods** online).

Our results suggest that the world of human microRNAs is larger than initially believed and is not limited to conserved sequences. We estimate that the number of conserved microRNAs is ~400–500 (in accordance with findings of recent studies that used different approaches^{9,10}) and that the total number of human microRNAs, including nonconserved clustered microRNAs, is at least 800. These findings support the notion that microRNAs have a central role in the regulation of protein translation throughout the human genome. Our results further indicate that a substantial portion of microRNAs are primate-specific. The fact that the primate-specific clusters described here are specifically expressed in developmental tissues supports the notion that microRNAs may have a key role in the evolutionary process and in the evolved complexity of higher mammals (see also ref. 13).

METHODS

Identifying and scoring candidate microRNA precursors. Step 1: Extracting hairpin structures from the entire genome. We folded the entire human genome using the Vienna package¹⁴ in windows of 1,000 nucleotides with an overlap of 150 nucleotides. All hairpin structures that have at least six base pairs, are at least 55 nucleotides long and have a loop not longer than 20 nucleotides were extracted from the minimum free energy fold of the window (excluding overlapping hairpins). Of the 222 known microRNAs, 14% are missed by this step either because of hairpins that do not fit with our definition (e.g., have a loop that is too large or more than one end-loop) or because of the massive folding in overlapping windows.

Step 2: Assigning each hairpin a stability score. A hairpin is energetically stable if it tends to appear in many folding configurations, which is indicated by the similarity of the minimum free energy graph and the partition function

graph provided by the Vienna package for that hairpin. The stability score of the hairpin is 1 minus the mean absolute difference between the two graphs in the hairpin region. The difference is calculated after normalizing the mfe graph so that the mean difference between the two graphs is zero in this region. Thus, scores closer to 1 indicate higher stability.

Step 3: Scoring hairpins. We compared features of known human microRNA precursors with features of a background set of 10,000 randomly selected hairpins found in non-protein-coding regions to identify features that characterize real microRNA precursors relative to background. We used structural features including hairpin length, loop length, stability score, free energy per nucleotide, number of matching base pairs and bulge size, and sequence features including sequence repetitiveness, regular and inverted internal repeats and free energy per nucleotide composition. We constructed an optimal predictor by finding the combination of features that best distinguished between true microRNA precursors and the background set (**Supplementary Methods** online).

Determining hairpin conservation. We divided hairpins into conserved and nonconserved hairpins using the University of California Santa Cruz phastCons^{15,16} data. These data contain a measure of evolutionary conservation for each nucleotide in the human genome against the genomes of chimpanzee, mouse, rat, dog, chicken, pufferfish and zebrafish, which is based on a phylogenetic hidden Markov model using best-in-genome pairwise alignment for each species (based on BlastZ), followed by multiZ alignment of the eight genomes^{15,16}. We defined a hairpin as conserved if the average phastCons conservation score over the seven species in any 15-nucleotide sequence in the hairpin stem is at least 0.9 (see also ref. 9).

Microarray high-throughput validation. We carried out microarray experiments designed to detect expression of mature microRNAs as previously described⁷. The microarray contains two probes per candidate microRNA gene, one for each predicted mature microRNA. Raw signals vary from a minimal signal of ~400 to a saturated signal of ~65,000. We considered probes with a signal above 2,500 to be positive but not necessarily reliable. To determine the reliability of the signal, we designed a group of 3,000 randomly chosen 35-mers from the human genome and added it to the microarray probes. We observed high correlation ($R^2 = 0.53$) between the probe's maximal signal over the tested tissues and the probe's cytosine (C) content. Only 6% of the background probes with C content below 35% had signals above 2,500 in at least one tissue, whereas more than 70% of the background probes with C content above 35% fulfilled the same condition. Thus, only candidate mature microRNAs with C content below 35% and a signal above 2,500 in at least one tissue passed the microarray high-throughput filtering. Those microRNAs that passed the filtering had a *P* value of 0.06.

MicroRNA sequence-directed cloning and sequencing. We prepared microRNA enriched libraries as previously described¹⁷ using suitable adaptors. We used RT-PCR amplification with an excess of the reverse primer (1:50 ratio) to produce a cDNA library. We then hybridized biotinylated capture oligonucleotides (22–30 nucleotides long, with biotin at the 5' end) to an aliquot (5 μ l) of the library in TEN buffer. We then added μ MACS Streptavidin Microbeads and incubated the reaction for 2 min at the hybridization temperature. We then loaded the mixture onto a magnetized μ MACS Streptavidin Kit column and eluted the hybridized single-stranded library molecules by adding 150 μ l of water preheated to 80 °C. We recovered the single-stranded cDNA library molecules, amplified them by PCR, ligated them into a pTZ57R/T vector and transformed the ligation products into JM109 bacteria. We identified and sequenced positive colonies (**Supplementary Methods** online).

Determining cluster homology. We compared microRNA precursors with all assembled genomes in the University of California Santa Cruz genome browser (BlastZ analysis¹⁶). A cluster was considered fully conserved if all its microRNA precursors have homologs that are also clustered. We looked for homologs of individual microRNA cluster members using Blast analysis against the whole-genome sequence data in the National Center for Biotechnology Information Trace databases.

Databases and accession numbers. We submitted new microRNA sequences to the microRNA Registry⁶ (Rfam miRNA names are given in **Supplementary Table 1** online). The GEO accession number for microarray data is GSE2708.

Note: Supplementary information is available on the *Nature Genetics* website.

ACKNOWLEDGMENTS

We thank the members of the Rosetta Genomics team for their dedication and contribution.

AUTHORS' CONTRIBUTIONS

I.B., A.A., Y.K. and R.A. conceived and designed the methodology of detecting microRNAs and directed the work of the other authors. O.B. designed microarray algorithms, and S.G. conceived the sequencing method. Other authors made significant noninventive contributions: O.B., A.B., P.E., U.E., E.S. and Y.S. developed bioinformatics and algorithmic elements; E.M. and S.G. prepared RNA for microarray analysis and sequenced microRNAs; and Z.B. provided scientific vision and guidance.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Genetics* website for details).

Received 24 January; accepted 31 May 2005

Published online at <http://www.nature.com/naturegenetics/>

1. Ambros, V., Lee, R.C., Lavanway, A., Williams, P.T. & Jewell, D. MicroRNAs and other tiny endogenous RNAs in *C. elegans*. *Curr. Biol.* **13**, 807–818 (2003).

2. Bartel, D.P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
3. Johnston, R.J. & Hobert, O. A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature* **426**, 845–849 (2003).
4. Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B. & Bartel, D.P. Vertebrate microRNA genes. *Science* **299**, 1540 (2003).
5. Poy, M.N. *et al.* A pancreatic islet-specific microRNA regulates insulin secretion. *Nature* **432**, 226–230 (2004).
6. Griffiths-Jones, S. The microRNA registry. *Nucleic Acids Res.* **32**, D109–D111 (2004).
7. Barad, O. *et al.* MicroRNA expression detected by oligonucleotide microarrays: System establishment and expression profiling in human tissues. *Genome Res.* **14**, 2486–2494 (2004).
8. Lim, L.P. *et al.* The microRNAs of *Caenorhabditis elegans*. *Genes Dev.* **17**, 991–1008 (2003).
9. Berezikov, E. *et al.* Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120**, 21–24 (2005).
10. Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**, 338–345 (2005).
11. Suh, M. *et al.* Human embryonic stem cells express a unique set of microRNAs. *Dev. Biol.* **270**, 488–498 (2004).
12. Houbaviy, H.B., Murray, M.F. & Sharp, P.A. Embryonic stem cell-specific microRNAs. *Dev. Cell* **5**, 351–358 (2003).
13. Bentwich, I. A postulated role for microRNA in cellular differentiation. *FASEB J.* **19**, 875–879 (2005).
14. Hofacker, I.L. Vienna RNA secondary structure server. *Nucleic Acids Res.* **31**, 3429–3431 (2003).
15. Siepel, A. & Haussler, D. Combining phylogenetic and hidden Markov models in biosequence analysis. *J. Comput. Biol.* **11**, 413–428 (2004).
16. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
17. Elbashir, S.M., Lendeckel, W. & Tuschl, T. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.* **15**, 188–200 (2001).