Published online 25 September 2017 *Nucleic Acids Research, 2017, Vol. 45, No. 21* ***e177***
doi: 10.1093/nar/gkx836

Mirnovo: genome-free prediction of microRNAs from small RNA sequencing data and single-cells using decision forests

Dimitrios M. Vitsios[1], Elissavet Kentepozidou[1], Leonor Quintais[1], Elia Benito-Gutiérrez[2], Stijn van Dongen[1], Matthew P. Davis[1,*] and Anton J. Enright[1,*]

[1]European Molecular Biology Laboratory—European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [2]Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK

Received July 27, 2017; Revised September 06, 2017; Editorial Decision September 10, 2017; Accepted September 20, 2017

ABSTRACT

The discovery of microRNAs (miRNAs) remains an important problem, particularly given the growth of high-throughput sequencing, cell sorting and single cell biology. While a large number of miRNAs have already been annotated, there may well be large numbers of miRNAs that are expressed in very particular cell types and remain elusive. Sequencing allows us to quickly and accurately identify the expression of known miRNAs from small RNA-Seq data. The biogenesis of miRNAs leads to very specific characteristics observed in their sequences. In brief, miRNAs usually have a well-defined 5′ end and a more flexible 3′ end with the possibility of 3′ tailing events, such as uridylation. Previous approaches to the prediction of novel miRNAs usually involve the analysis of structural features of miRNA precursor hairpin sequences obtained from genome sequence. We surmised that it may be possible to identify miRNAs by using these biogenesis features observed directly from sequenced reads, solely or in addition to structural analysis from genome data. To this end, we have developed mirnovo, a machine learning based algorithm, which is able to identify known and novel miRNAs in animals and plants directly from small RNA-Seq data, with or without a reference genome. This method performs comparably to existing tools, however is simpler to use with reduced run time. Its performance and accuracy has been tested on multiple datasets, including species with poorly assembled genomes, RNaseIII (Drosha and/or Dicer) deficient samples and single cells (at both embryonic and adult stage).

INTRODUCTION

The identification and annotation of novel miRNAs from various species, either animals or plants, has been a challenge in the field of small non-coding RNAs for many years. Traditionally, novel miRNA prediction was based on the identification of short sequences, mapping such sequences to the genome, and searching for those loci that may produce the characteristic hairpin structure of a pre-miRNA via analysis of derived structural features. However, we sought to explore the possibility of predicting novel miRNAs with high accuracy without requiring a reference genome in the process. Our initial hypothesis is that features of microRNA (miRNA) sequences, derived from their biogenesis may be sufficient to predict miRNAs *de novo,* i.e. without using a reference genome. These 'biogenesis' features (Figure 1A) are clearly evident when one interrogates large numbers of miRNA sequencing datasets from multiple species. In order to perform genome-free feature analysis of miRNA sequences, one needs to take an input set of small RNA sequences and globally group them into clusters of related sequence. These clusters may then be multiply aligned and filtered. This alignment allows a consensus sequence to be constructed and biogenesis features to be assessed. The advantages of *de novo* discovery of miRNAs purely from sequencing data are readily apparent: (i) it does not require a reference genome, (ii) removing the genomic mapping and RNA secondary structural analysis allows for faster computation and (iii) it will produce a smaller set of novel candidate sequences, should one want to do genomic feature analysis later. To this end, we have developed a new method, *mirnovo*, which allows for prediction of novel miRNAs in animals and plants, with or without a reference genome.

*To whom correspondence should be addressed. Tel: +44 1223 494444; Fax: +44 1223 494468; Email: aje@ebi.ac.uk
Correspondence may also be addressed to Matthew P. Davis. Tel: +44 1223 494444; Fax: +44 1223 494468; Email: matdavis@ebi.ac.uk

MATERIALS AND METHODS

Input data

Mirnovo accepts as an input one gzipped (.gz) FASTQ or FASTA file for each run from either bulk or single-cell small RNA-Sequencing data. Input sequences may have already been pre-cleaned from their 3′ adapters otherwise a 3′ adapter sequence needs to be provided by the user.

Mirnovo pipeline

The 3′ adapter from input data is removed with *reaper* (27) and then cleaned sequences are de-duplicated with *tally* (27). Initial clustering of tallied sequences is performed with *vsearch* (28) using an alignment identity threshold of 0.9 by default. Clusters refinement is achieved by merging similar consensus sequences with *cd-hit* (2), based on 7-mer searches and using 0.85 as the alignment identity threshold. Multiple-sequence alignment for the refined clusters is performed with *muscle* (3). Following the miRNA prediction step, the consensus sequences of all identified known and/or novel miRNAs are mapped against the reference genome (if applicable) using bowtie2 (29) (selected parameters: *-k* 1, *-D* 20, *-R* 3, *-N* 1, *-L* 20, *-i* S,1,0.50 *–rdg* 1,1 *–rfg* 1,1). The most stable hairpins, in terms of ΔG free energy, assessed within a 90nt window around these sequences, are selected and genomic features are calculated for each hairpin candidate. Eventually, up to 5 hairpins are reported as paralog precursors for each mature miRNA in case the calculated free energies of these secondary structures are below an empirically defined threshold.

Features definition

The full set of features used for classification and prediction is described as follows:

- Twelve coverage profile features: cluster read depth, main body length, mismatches in main body, scaling rate before 5′, scaling rate after 3′, gaps before main body, mismatches in seed region, average GC content in main body, gaps after main body, average AU content after 3′, alignment identity against the potential reverse complement and average sequence length.
- Twelve sequence complexity features: A+T skew (*ats*), C+G skew (*gcs*), CpG skew (*cpg*), complexity by Wootton and Federhen (30) (*cwf*), entropy (*ce*), complexity as compression ratio using gzip (*cz*), complexity as Markov model size of N $\in$ {2,3} (*cm2*, *cm3*), Trifnov's complexity (31) with order $N \in \{2,3\}$ (*ct2*, *ct3*) and linguistic complexity with order $N \in \{2,3\}$ (*cl2*, *cl3*).
- Nine genomic features (hairpin folding retrieved using *RNAfold* from the Vienna package (32)): hairpin size estimate, mature miRNA distance from stem loop, loop size estimate, number of loops in hairpin, minimum free energy of secondary structure, 'majority' brackets in the entire folding (prevalent of the two distinguishing bracket directions, i.e. max{num['(', ')']}), miRNA bracket discrepancy (K/N, where N is the total number of brackets in the miRNA and K is the number of 'majority' brackets), miRNA bracket fraction (K/N, where N is the miRNA length and K is the number of 'majority' brackets) and number of unmatched nucleotides from the mature miRNA sequence.

Output

The results from each job contain first of all FASTA files for the predicted known and novel miRNAs (both for the mature products and their respective hairpin precursors), and for any tRNA and/or rRNA identified hits. Additionally, BED files with genomic coordinates of predicted hairpins are provided along with coverage profiles for each mature miRNA and also the secondary structures of each identified hairpin paralog. Furthermore, each job is associated with a table of performance measures with regards to the machine learning predictions. The reported measures are:

$$precision = \frac{TP}{TP+FP}; sensitivity = \frac{TP}{P}; specificity = \frac{TN}{N},$$

where

TP: is the number of predicted *known* miRNAs,

FP: is the number of predicted *novel* miRNAs,

P: is the number of all known miRNAs contained in the input data (based on the miRBase annotation),

TN: is the number of (correctly) predicted non-miRNA sequences,

N: is the number of all non-miRNA sequences contained in the input data (based on the miRBase annotation).

Predictions are accompanied with a ROCR and Precision-Recall (PR) curve, which demonstrate the performance of the machine learning method with regards to correctly identifying known and novel miRNAs, respectively. Finally, the distributions of all feature values (coverage, sequence complexity and genomic) for each class of predicted miRNA/non-miRNA sequences are visualised and made available as post-prediction QC box-plots.

Machine learning model training

The core machine learning algorithm used for training was based on Random Forests. The Random Forest implementation was provided by the *randomForest* R package (SVM and Gradient Boosting methods were also tested using the *e1071* and *gbm* R packages, respectively). In order to fine-tune our model we tested its performance independently for various numbers of randomly selected predictors (*mtry*) and numbers of trees (*ntrees*) on 65 mouse samples downloaded from ENA (31). Optimal performance was obtained for $mtry = 6$ and $ntrees = 2000$ and thus these parameters were selected for the training of each classifier (Supplementary Figure S2). Samples used for training of the animal and plant species models were also downloaded from ENA (33).

Supported species

Mirnovo can analyse datasets from any species, without requiring a genome reference or miRBase annotated miRNAs. The option '– *Not Available* –' should be used in this case in the place of the *Input species*. However, even higher accuracy can be achieved by integrating the genomic features into prediction. Thus, mirnovo has integrated genomic support for 67 species. This means that for those

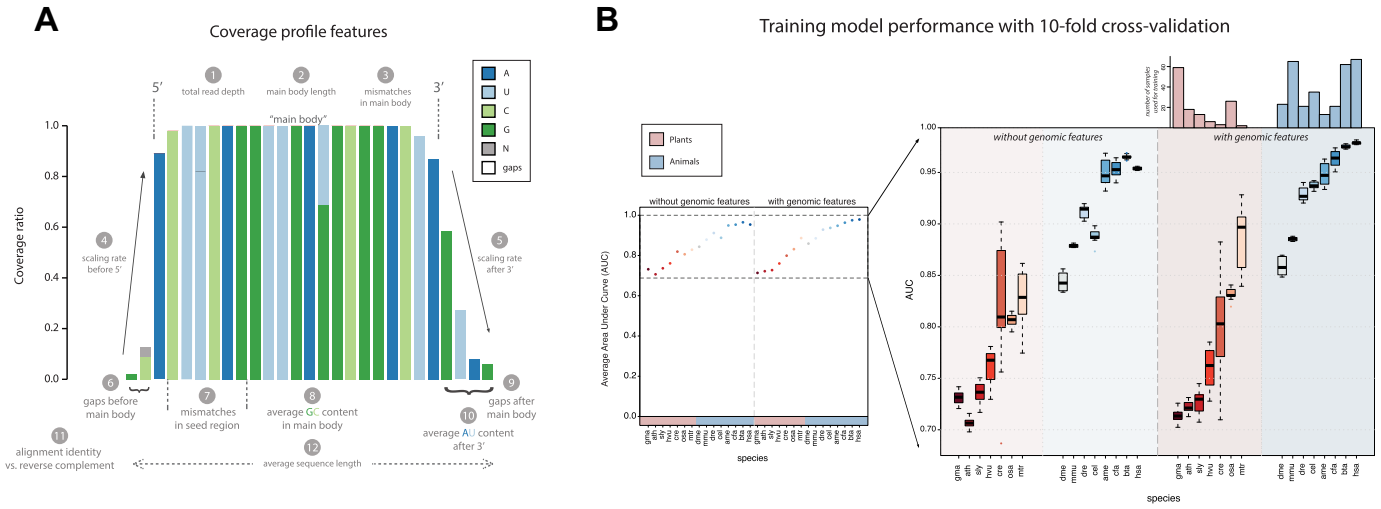

Figure 1. Mirnovo biogenesis features and performance across multiple species. (**A**) Coverage features definition for each cluster of similar sequences. (**B**) Training model performances with 10-fold cross-validation across 7 Plant and 8 Animal Species, with or without using a reference genome.

species, the full set of coverage profile, sequence complexity and genomic features can be compiled in order to identify known and predict novel miRNAs. Additionally, mirnovo supports miRNA identification and prediction for another 160 species with miRBase annotated miRNAs, but lacking genomic feature support. The command-line version of our method though allows the user to build and integrate into the identification process any custom reference genome.

Training models per species

We have trained individual models for 8 animal species (*Apis mellifera*, *Bos taurus*, *Caenorhabditis elegans*, *Canis familiaris*, *Danio rerio*, *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus)* and seven plant species (*Arabidopsis Thaliana*, *Chlamydomonas reinhardtii*, *Glycine max*, *Hordeum vulgare*, *Medicago truncatula*, *Oryza sativa japonica*, *Solanum lycopersicum*). These models offer optimised results when input files originate from one of those species. Additionally, we have created two universal models, for animals and plants respectively, that can be used generically for any species belonging to one of the two kingdoms. These models have been created by sampling data-points (refined sequence clusters) from the entire dataset of small RNA clusters from the aforementioned animal and plant species. The universal models have also been trained using 800 trees (instead of 2000) since the addition of extra trees did not improve prediction accuracy—data not shown—but only increased the file size of the produced models. Furthermore, the user is able to select which groups of features are going to be used for making miRNA predictions. The possible combinations of sets of features used for prediction are: biogenesis-only (coverage and sequence complexity), genomic-only, both biogenesis and genomic. A distinct model has been trained for each of these cases, so three different training models were eventually trained for each species to match the users' preferences in each run.

Parameter specification

MicroRNA prediction is performed by default using all 24 biogenesis features and the nine genomic features (in case the reference genome is available). However, it is also possible to completely disable genomic features, by selecting the '*Disable genomic features*' option, or use exclusively the genomic features for prediction, by checking the '*Use only genomic features for prediction*' option. Furthermore, mirnovo offers a set of three parameters in order to facilitate sequence clustering and boost correct classification of predicted miRNAs. Specifically, when analysing samples with high read depth and high sequence complexity (i.e. high number of generated clusters at the initial sequence clustering of input data with *vsearch*), we noticed that in some cases predictions contain an unexpectedly high number of novel miRNAs, sometimes even higher than the number of predicted known miRNAs (Supplementary Figure S16). In order to resolve this issue we introduced, first of all, the '*Reduce input sequence complexity*' option which allows the user to filter out unique sequences from the input file with a total read depth below a certain threshold. For instance, by using a *tally-threshold* of x3, all unique sequences from the tallied file with a depth of up to three reads are discarded from the rest of the analysis. Following the initial sequence clustering, additional filtering is possible by retaining only those clusters that have total depth equal to or greater than the *min-read-depth* value and a number of unique isoforms within the cluster at least equal to the *min-variants* parameter value. We also tested prediction performance when using pre-sampled input files, for various sampling thresholds (Supplementary Figure S17). We noticed that the high sequence complexity issue is resolved beyond a certain sampling threshold, similarly with the parameter tweaking which is supported by mirnovo. Finally, this suggests that miRNA prediction with mirnovo is in general feasible even with low sequencing depth.

Sampling test depths

Subsamples (Supplementary Figure S17) were made for reads in the SRR546155 (Accession: *PRJNA80147*) and HG00099 (18) (specifically: HG00099_5_MI_120327_3_1) FASTQ files. For each sample one of a series of probabilities was set for the sampling of input sequences (0.001, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9). Mirnovo was used to predict miRNAs in each subsample and the unsampled dataset with the default parameters, using a species-specific training set and genome. Results were classified according to miRBase annotations.

mirnovo web-server backend

Mirnovo uses one of the fastest academic networks in Europe for efficient file upload and all jobs are submitted to the EMBL-EBI high performance computing cluster. Each job process is extensively parallelised with multiple threads for calculations, processing different subsets of the data and dividing different subtasks. The job's progress is visualised in real-time through a console window at mirnovo's progress page in the browser.

mirnovo stand-alone tool

Mirnovo is available as a stand-alone package along-side the web-server version. The downloadable bundle contains all necessary scripts and binaries for execution of mirnovo, providing separate versions for either Mac OSX or Linux platforms. The only required dependencies for the local machine are: Perl (v5.24.1), Python (v2.7.10), R (v3.2.2) and bowtie2 (v2.0.6), with the recommended versions in parentheses.

Refined mature miRNA quantification with Chimira (21)

Mirnovo is able to predict both hairpins and mature miRNAs, providing count data in the latter case. However, inherent sequence clustering steps (initial and refined) of the mirnovo pipeline may be imperfect in some cases and thus affect, even at a low level, the yielded expression data. Thus, in order to extract even more accurate expression data we have expanded chimira, a method that was previously published in our lab. In that case, chimira serves as a mirnovo extension, allowing the user to upload a custom set of hairpin sequences (e.g. known and/or novel hairpins predicted by mirnovo) and then align their input files against this reference set to get mature miRNA expression counts. All uploaded files are merged and sequences with an alignment identity over 0.90 are collapsed. As an additional functionality, chimira is able to generate coverage profiles of each identified mature miRNA and the secondary structure of the corresponding hairpin reference hit (using the Vienna package (32)).

GEUVADIS dataset analysis

All samples under the accession number *PRJEB3365* (PMID: 204868) were analysed. The majority of samples were run using the default mirnovo parameters (length filter: 16–28nt, *min-read-depth*: 5, *min-variants*: 1, *vsearch-id*: 0.9). The '*Reduce input sequence complexity*' option with a *tally-threshold* of $\times 3$ was used only for 2% of all datasets in order to reduce sequence complexity within the samples and thus optimise the initial sequence clustering with *vsearch*. The coverage profiles and hairpin precursors of all predicted novel miRNAs in the GEUVADIS samples are available at the following link:

http://wwwdev.ebi.ac.uk/enright-dev/mirnovo-standalone-pkg/misc/geuvadis-analysis.

mirnovo vs miRDeep2 benchmarking

miRDeep2 was always provided with the human reference genome, all known human hairpins, all known human mature miRNA sequences and also all mature miRNAs from two extra species (*D. melanogaster* and *C. briggsae*) for additional diversity. Mirnovo was tested both with and without the reference genome. With regards to the time benchmarking, mirnovo is a highly-parallelised and multi-threaded method while miRDeep2 is a serially processed method. Thus, we wanted the benchmarking to reflect the run time experienced by the end user. Both methods ran on HPC clusters consisted of 32-processor nodes equipped with the Intel(R) Xeon(R) CPU E5–2650 v2 @ 2.60 GHz CPU model. Mirnovo was run using the default number of hosts that is selected for each job (-$n = 3$) while miRDeep2 was run using -$n = 1$ (assigning -$n = 3$ hosts to miRDeep2 proved to be slightly slower -data not shown—most likely due to synchronisation latency among the hosts, and thus one host was eventually assigned for benchmarking of the miRDeep2 runs). Both methods were provided with 8GB of memory (-*M* 8192).

Analysis of moth and butterfly samples

For each sample the 3′ adapter sequence was identified using minion (27) and where possible confirmed in the relevant manuscript or database methods. Each sample was analysed using mirnovo with either default or custom set of parameters (Supplementary Table S4). The sample ids that were analysed are: *SRR035544* & *SRR035546* (GSE17965, PMID: 20199675), *SRR062599* (GSE23292, PMID: 200023292), *SRR062600* (GSE23292, PMID: 21266089), *SRR1663190* & *SRR1663191* (GSE63644, PMID: 25576364), and *SRR035545* (GSE17965, PMID: 200017965). A relevant genome was used for each sample (*Bombyx mori*: GCA_000151625.1, *Heliconius melpomene*: Hmel2 v2–0 Release_20151013, *Cameraria ohridella*: k51, *Pararge aegeria*: k51) and a *Drosophila Melanogaster* (dme) training model for miRNA predictions. To find the orthologues, novel mature miRNA sequences were compared to all miRBase sequences (v21) using swan (v17–096) requiring at least a 90% identity match (–key-value parameter).

Drosha/Dicer/XPO5-dependent analysis

Samples were normalised using the same strategy described in the original manuscript (20). Specifically, we normalised the wild-type, Drosha and XPO5 knockout samples based on the read counts of miR-320a-3p across all replicates,

since its expression is independent of Drosha. The Dicer knockout samples were respectively normalised based on the combined tRNA and rRNA levels of the WT samples, which should remain unaffected in the knockout samples as well. In order to derive expression data from all samples with reference to the hairpins that were identified and/or predicted by mirnovo, we expanded the already published method Chimira. The additional feature in Chimira allows alignment against a custom reference species that can be uploaded as a set of FASTA files by the user.

Novel miRNA prediction from single-cell RNA-Seq data

Processing of single-cell RNA-Seq data follows the same core pipeline as regular small RNA-Seq data processing. The only exception is that due to high innate noise of single-cell data, coverage and sequence complexity features are not taken into consideration at the final classification step, and thus predictions are inferred by models that have been pre-trained solely based on the genomic features. Thus, in order to make predictions from single-cell data the option '*Use only genomic features for prediction*' needs to be enabled.

RESULTS

The main methodology behind mirnovo lies in graph-based clustering of read-to-read similarities obtained from raw sequencing data (Supplementary Figure S1). Input reads get adapter cleaned, de-duplicated and clustered together into groups of highly similar sequences (see Materials and Methods). Subsequently, clusters are filtered based on the minimum number of isoform variants they contain and their overall sequencing depth. A consensus sequence is then calculated for each cluster and all clusters are aligned against Rfam (1) to identify likely rRNAs (or other contaminants). Clustering is extremely rapid (see Methods), however inconsistencies may arise. Hence, an extra refinement step has been introduced in order to merge clusters with highly similar consensus sequences using cd-hit (2). We then perform fast multiple-sequence alignment within each cluster using muscle (3). This is used to extract refined consensus sequences from merged clusters. In order to flag existing miRNAs, we compare these consensus sequences against miRBase (4). This final group of filtered, multiply aligned reads is used to compute a set of features for each cluster.

Machine learning features

We use a set of 24 biogenesis features, grouped into two categories: 12 coverage profile features (Figure 1A) and a set of 12 sequence complexity features (see Methods). Optionally, the user can also provide genomic sequence if desired, which adds 9 genomic features based on predicted RNA secondary structure from mapped consensus sequences (see Methods). Based on this feature set, mirnovo uses a machine learning classifier to identify both known and novel mature miRNAs.

Our initial hypothesis of being able to predict de novo miRNAs using these 24 genome-free features based purely on biogenesis needed to be tested comprehensively. Additionally, we wanted to directly compare miRNA prediction based on these 24 features or on the 33 features that include genomic and structural information. There are several existing tools (5–12) that address the novel miRNA prediction problem, such as miRDeep2, mirTools and miRanalyzer (Supplementary Table S1). We selected miRDeep2 as the most prominent tool (13–16), which utilises genomic and structural features, as a comparison for our approach.

Training models performance

Machine learning performance was initially assessed with 10-fold cross-validation using a labeled set of feature instances derived from 65 mouse samples. This allowed for evaluation of feature predictability and bias-free assessment of the predictive power of the classification model in a controlled dataset. We tested a range of machine learning approaches that included: Support Vector Machines (SVMs), Gradient Boosting and Random Decision Forests (Supplementary Figure S2). The most efficient approach in terms of discriminative power (based on the Area Under Curve - AUC- scores), with or without using the genomic features, turned out to be Random Decision Forests (or Random Forests). Hence, we selected this method to be integrated into mirnovo as the primary prediction algorithm. Overall, we have trained models for 8 animal and 7 plant species using 2–66 samples with labeled data in each case, making up 433 samples in total (Supplementary Tables S2 and S3).

The 10-fold cross-validation demonstrated accuracy measures of 84.4–96.5% without a reference genome using a model built from animal species (Figure 1B). Interestingly, miRNA predictions on plant sequences still managed accuracy between 70.7% and 82.9%, despite their differences in biogenesis compared to animals (17). Inspection of the feature importance scores for the accuracy of predictions (Supplementary Figure S3) yields some of the coverage features (read depth, average sequence length of mature sequence, average GC content and average AT content after 3′ end) as the most critical ones for correct classification, in both animals and plants. Moreover, we can observe that genomic features play a more predominant role in animals than in plants, most likely because of the high variability of secondary structures of miRNA precursors in plants. Besides, this variability in plant miRNAs can be seen in the high variance of feature importance scores, in contrast with the lower variance respective animal features.

These initial results confirmed that without integrating any genome information it is still possible to reliably identify both known and novel miRNAs directly from sequencing data. Addition of the extra 9 traditional genomic features does improve accuracy, but not by as much as expected. There was a 1.65% (±1.53%) and 0.7% (±2.79%) improvement of prediction accuracy for animals and plants respectively (85.9–97.9%, 71.4–88.7% respectively). We now build a universal-animal and a universal-plant model by sampling data points (refined sequence clusters) from each respective pool of species such that they can be used uniformly by any species originating from these kingdoms (Supplementary Figure S4). Obtained accuracy for these universal models was 89.7 or 92% for animals, and 71.4 or 71.8% for plants, without and with a reference genome respectively. When using only the genomic features for

miRNA prediction, accuracy scores decrease to 58.4% and 79.3% for plants and animals, respectively (Supplementary Figure S4). This proves the essential role of the biogenesis and sequence complexity features used by our machine learning model for accurate miRNA prediction.

Large-scale miRNA predictions from GEUVADIS dataset (18)

In order to more widely assess mirnovo, we applied a large-scale benchmarking. We used all human samples from the GEUVADIS dataset (18) (derived from lymphoblastoid cell lines). The initial analysis was performed without using any human reference genome data (Supplementary Figure S5). The accuracy obtained averaged 92.14% while sensitivity and novel prediction rates were 69.07% and 34.62%, respectively. After introducing the reference genome and the extra 9 corresponding genomic features (Figure 2, Supplementary Figures S6 and S7), performance is improved (accuracy and sensitivity: 95.51% and 78.8%), while novel prediction rates fall to 18.63%. This implies that the use of genomic features boosts the prediction clarity of real miRNAs while at the same time keeping the number of false positive assignments of novel miRNAs relatively low.

We then compared the performance of mirnovo to miRDeep2, the most widely used tool for miRNA discovery. Since miRDeep2 requires genomic data, we always provided it with the human reference genome, known human hairpin and mature miRNA sequences. Mirnovo was tested both with and without the reference genome in separate runs (Figure 3a, Supplementary Figures S8 and S9). We observed that mirnovo outperforms miRDeep2 in 92% of the cases for known mature miRNAs identification and predicts more novel miRNAs in 99.9% of the cases, when provided with genomic sequence.

Benchmarking against miRDeep2

Even without a reference genome, mirnovo performs comparably with miRDeep2 in terms of predicting known miRNAs, even though miRDeep2 utilises genomic information (Figure 3A and Supplementary Figure S9). With regards to novel miRNA prediction, we observed a higher prediction rate for mirnovo, which indicates likely more false positive hits. However, mirnovo runs significantly faster than miRDeep2 (see below). When the playing fields are lebeled and both methods utilize the reference genome (Figure 3A and Supplementary Figure S8), we observe slightly improved sensitivity but a notable impact on precision. This effectively reduces the number of falsely predicted novel miRNAs (Figure 3A, Supplementary Figures S6 and S7). In summary, mirnovo predicted 2414 novel mature miRNAs (see Materials and Methods) in the GEUVADIS dataset, originating from 3173 hairpin precursors (Supplementary Data S1) including any detected paralogs. Expression of novel miRNAs was relatively balanced between samples, and similar to the expression of known miRNAs. The lengths of the majority of predicted novel miRNAs were also within the expected range of 20–23nt (Supplementary Figure S10).

Additionally, we also tested mirnovo and miRDeep2 on eight animal species (Figure 3B), using the reference genomes in both methods. Again we observed that mirnovo performed better than miRDeep2 in the majority of cases for both known and novel miRNA predictions. We then assessed the computational execution time, for the GEUVADIS benchmarking runs. Mirnovo was on average 2.5× faster than miRDeep2, with mean execution time being ∼43 min (Supplementary Figure S11), as compared to 1 h 49 min for miRDeep2. The mirnovo pipeline, being faster and more lightweight is also easier to configure and set-up as a command-line tool and also lends itself extremely well as a simple to use web based server. We certainly see advantages to running both mirnovo and miRDeep2 on sequencing data and believe that mirnovo represents a significant and useful addition to this field.

Performance in non-model organisms with poor genome-assembly

Genome-free performance of mirnovo appears to prove our initial hypothesis that biogenesis features exhibited on miRNA sequences allow accurate de novo miRNA prediction. Although the performance is slightly worse without genomic information, this enables de novo miRNA discovery in the multitudes of non-model organisms without genomes, or with low-coverage data. It also presents a tractable approach for small RNA analysis in metagenomics data. In order to test the potential of the genome-free approach in such species, we assessed mirnovo's performance on seven samples from five different moth species without fully assembled genomes (19), two of which do not have any miRBase annotation (Supplementary Table S4). Mirnovo was able to retrieve known miRNAs from all species with miRBase annotation (*B. mori, H. melpomene melpomene, H. melpomene rosina*) along with hundreds of novel miRNAs (Supplementary Data S2–S6). Additionally, mirnovo predicted 119 and 192 miRNAs from *C. ohridella* and *P. aegeria* respectively. These species did not have any miRBase annotated miRNAs (Supplementary Data S7 and S8). Among all predicted novel miRNAs, *C. ohridella* and *P. aegeria* were the species with the highest number of miRNAs aligning with paralogs from other species. This is to be expected since the other three moth species have been studied more extensively and already have miRNA entries in miRBase. A small proportion of novel miRNAs were predicted without any genomic evidence, based solely on features derived from their coverage profiles. We believe that this effectively demonstrates another strength of mirnovo for inference of miRNAs in non-model organisms and enables research on non-coding RNAs for many new species.

As an additional step, we tested mirnovo's performance against miReader (11) and MirPlex (12) that are also able to predict miRNAs without a reference genome at all. Both of these methods impose notable restrictions in miRNA prediction since they require that both strands of the pre-miRNA duplex are detectable in the sequencing data. However, this is not common for miRNAs since, in most of the cases, only one strand of the duplex becomes a mature miRNA while the other one gets degraded (34). Benchmarking results show that mirnovo performs either comparably or better than both of these methods (Supplementary

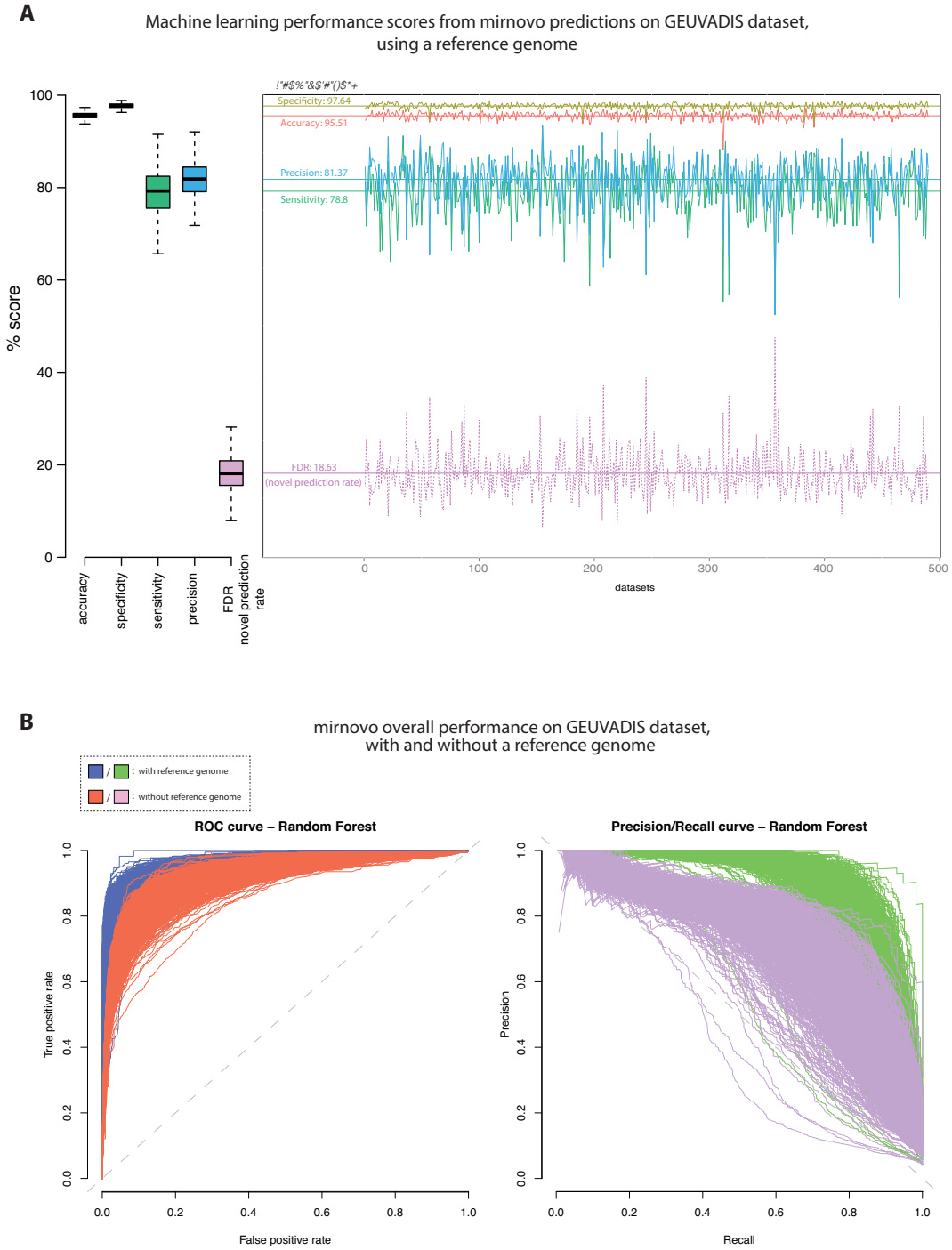

Figure 2. Mirnovo machine learning performance on a large-scale analysis involving 491 samples from the GEUVADIS (16) dataset. (**A**) The distribution of *accuracy*, *specificity*, *sensitivity*, *precision* and *FDR* (*novel prediction rate*) scores is shown across all samples of the dataset. (**B**) ROCR and Precision–Recall (PR) curves from mirnovo prediction performance across all samples from the GEUVADIS dataset, with or without using a reference genome.

Figure S18), which is expected due to mirnovo's ability to detect miRNAs using more flexible and diverse criteria.

MicroRNA prediction in RNase III-deficient cells

Novel miRNAs are predicted based on features consistent with their processing by small RNA biogenesis machinery. Hence, if they are real miRNAs, one would expect to observe their dysregulation when key miRNA biogenesis enzymes are missing or mutated. We tested this hypothesis using published experimental data from Drosha, XPO5 and Dicer knockout samples (20). These enzymes are responsible for cleavage of miRNA primary transcripts, their nuclear export and processing into functional mature miRNAs respectively. We predicted known and novel hu-

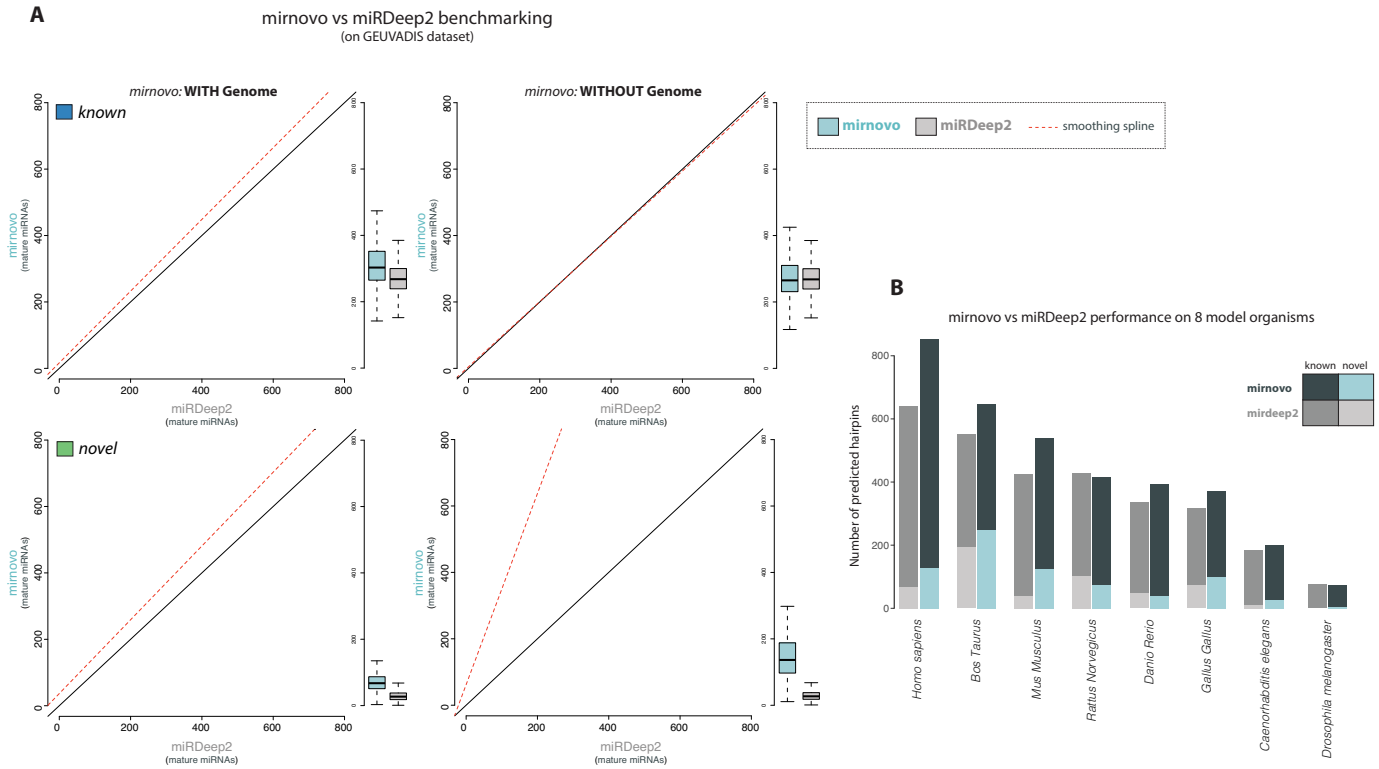

Figure 3. Benchmarking of mirnovo against miRDeep2. (**A**) mirnovo-vs-miRDeep2 prediction performance across the GEUVADIS dataset: miRDeep2 was always run using a reference genome, all known human hairpins, all known human mature miRNAs and mature miRNAs from another two species (*D. melanogaster* and *C. briggsae*). Mirnovo was run either with or without using the reference genome. (**B**) mirnovo-vs-miRDeep2 performance on identification and prediction of miRNAs from samples across eight model organisms.

man hairpins from the wildtype (WT) samples. We then aligned all WT and Knockout (KO) samples using chimira (21) against the predicted known hairpins obtained from mirnovo (Figure 4A and B, Supplementary Figures S12 and S13). For this quantification step we expanded chimira's functionality, as a mirnovo extension, in order to allow for alignment against a custom set of reference hairpins, uploaded by the user. Our data verified the observed minor effect of XPO5 knockout in miRNA expression, since miRNAs are still being expressed, just in lower levels in some cases. The Dicer knockout, as expected, leads to notable decrease in miRNA expression. The absence of Drosha is verified to be the most critical one since it results in extensive depletion of the majority of miRNAs, again verifying previously results reported in the original paper. Some previously identified Dicer dependent but Drosha independent miRNAs (miRtrons) are also observed from our data.

For our novel miRNA predictions, we aligned all samples against the list of predicted novel hairpins (Figure 4C, Supplementary Figures S12 and S13). We then assessed which miRNAs were differentially expressed (fold-change > 2 and $P < 0.05$) between the WT and KO conditions and found three sets of novel miRNAs, dependent on different types of enzymes (Figure 4D and Supplementary Data S9). Overall, we find 40 novel miRNAs significantly differentially expressed both in Drosha and Dicer knockout samples (Supplementary Data S10). This implies that this set of novel miRNAs is dependent on the two most important enzymes for miRNA biogenesis (Drosha and Dicer) and thus are processed by the canonical biogenesis pathway. Additionally, we noticed that 25 novel miRNAs were dependent only on Dicer (likely miRtrons) and 33 were Drosha-only dependent (Supplementary Data S11 and S12, respectively). This finding is in accordance with previous studies (20,22–24) that miRNAs may be dependent on only one of the two key enzymes (either Drosha or Dicer) possibly originating from other structured noncoding RNAs. These results again, provide validation that mirnovo is predicting molecules likely to be processed by the canonical biogenesis machinery yet can also identify those miRNAs which are independent of one or more of the key enzymes.

MicroRNA prediction from single-cell RNA-Sequencing data

Recently, single-cell RNA sequencing has become both tractable and an extremely active topic of research. Given that some miRNAs have been shown to be extremely cell-type specific, such datasets represent an important area for novel miRNA discovery. Hence, we wished to assess the performance of mirnovo in analysis of single-cell small RNA-Seq. We initially attempted prediction using all sets of features (coverage, sequence complexity and genomic) but the extracted coverage profiles and extracted sequence complexity scores were distorting predictions due to high noise of input data. We then tried making our predictions using only genomic and we observed a clear improvement in accuracy scores, thus we followed this approach for the analysis

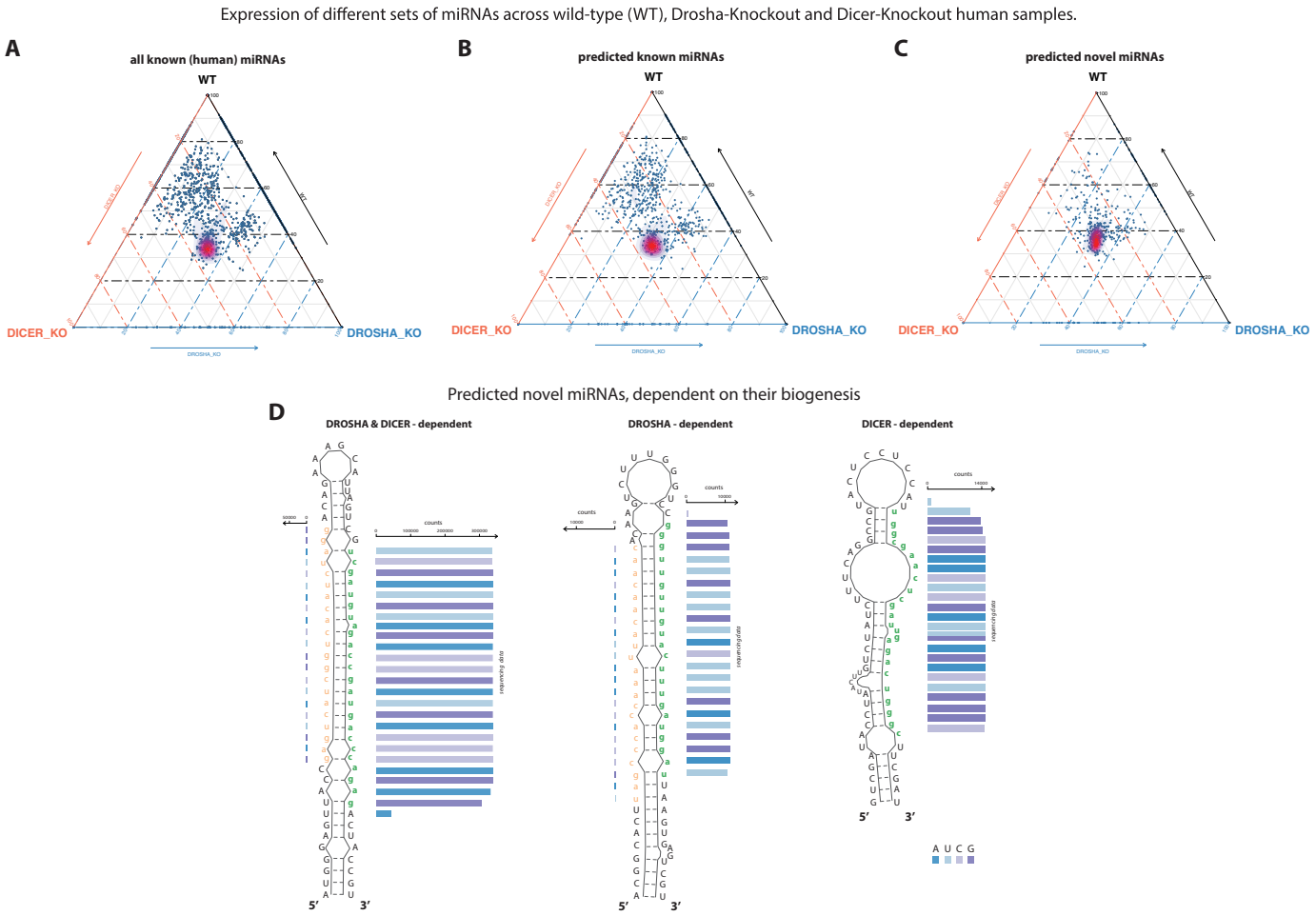

Figure 4. Dependence of novel miRNAs on RNaseIII enzymes (Drosha and Dicer) and Exportin-5. MicroRNA expression data across Wild-Type and Drosha/Dicer/XPO5 knockout conditions after alignment against: (**A**) all known human hairpins from miRBase, (**B**) all known human hairpins predicted by mirnovo in these samples, (**C**) all novel hairpins predicted from mirnovo in these samples. The ternary plots show the relative expression of each miRNA across the examined conditions. Data points (miRNAs) in the middle of the triangle are equally expressed among all samples while points proximal to a certain vertex are more highly expressed at the corresponding sample. (**D**) Predicted novel miRNAs in human cell line (HCT116) by mirnovo, dependent on both Drosha and Dicer, Drosha only or Dicer only, respectively.

of single-cell data. This proves to be another useful feature of mirnovo, since the user is always able to switch off certain sets of features in order to make their predictions based on the specific requirements, quality or noise of input data. We re-analysed 204 such samples from HEK cells, naive human embryonic stem cells (hESCs) and primed hESCs (25). Naive embryonic stem cells are obtained from pre-implantation embryos while primed ones are obtained from post-implantation embryos (26). Mirnovo predicted 4747 novel hairpin candidates overall from these samples, 356 of which have also been predicted from the GEUVADIS dataset on human lymphoblastoid cell lines. These initial findings require further filtering based on mature miRNAs expression, in order to account for noise due to single-cell data. Specifically, we found that 3135 and 361 miRNAs were expressed above median and average expression of all novel miRNAs, respectively. We then aligned all samples against the predicted set of hairpins using chimira, and obtained mature miRNA expression data for each cell sample. Novel miRNA expression is quite balanced between both types of ESCs, with a small group of miRNAs being down-regulated during the transition from naive to primed ESCs (Figure 5A and Supplementary Figure S14). On the other hand, both types of embryonic stem cells show notable differential expression compared to differentiated HEK cells. Interestingly, highly similar expression patterns can be observed with regards to known miRNA expression across these cell types (Figure 5B and Supplementary Figure S14).

We observed that novel miRNA expression varies across different states of pluripotency and/or development in *Homo sapiens*, with a more significant difference observed between embryonic stem cells versus fully differentiated cell types. We performed hierarchical clustering for all cells based on their novel miRNAs expression. We identified five major groups of cells with similar novel miRNA signature (Figure 5C). Two of those groups were exclusively comprised of HEK cells, two groups were primarily populated by ESC naive and ESC primed cells respectively. Finally, the last group consisted of cells from all three cell types. Hierarchical clustering of these cells based on known miRNA

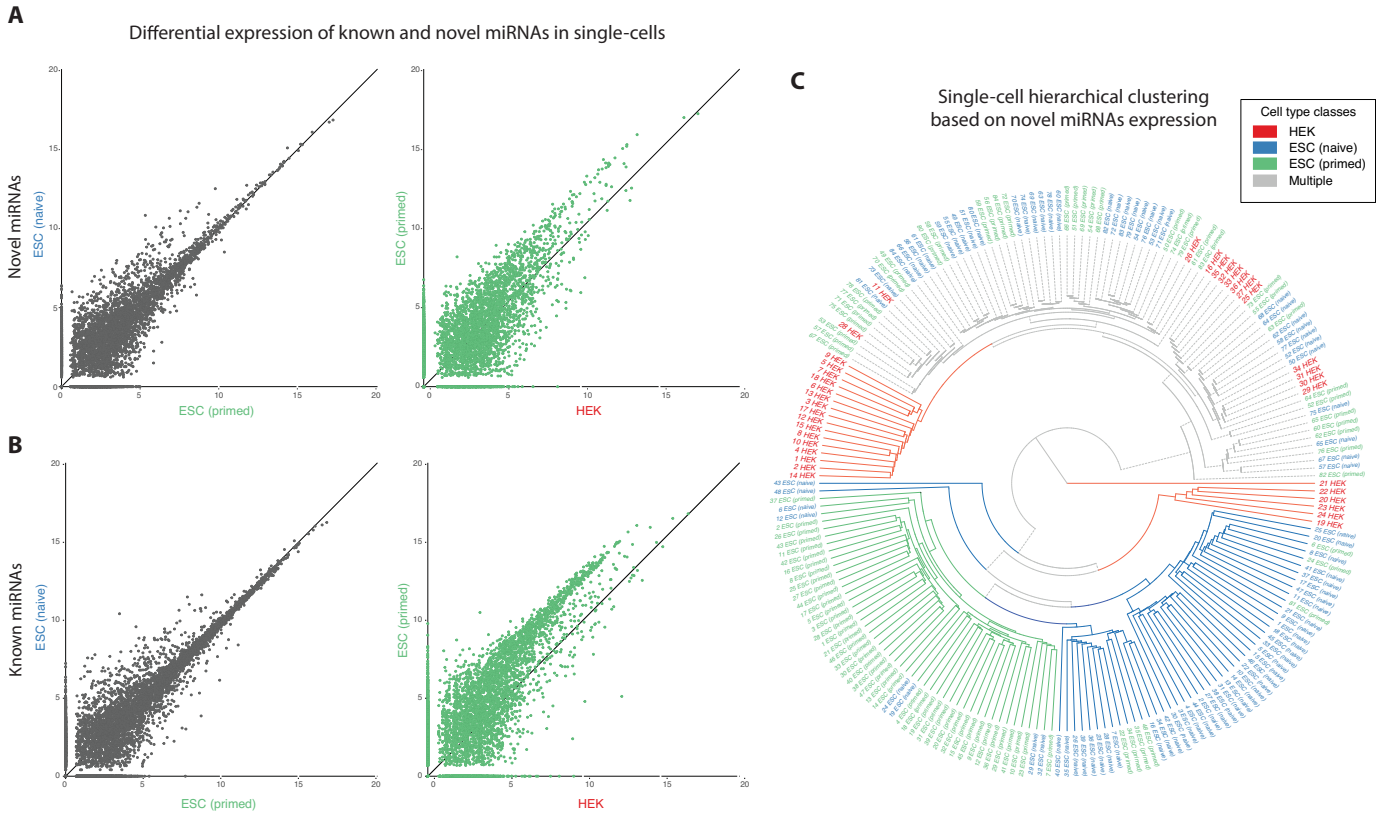

Figure 5. Novel miRNA predictions from single-cell RNA-Sequencing data involving three cell-types: naive ESCs, primed ESCs and HEK cells. (**A**) Normalized expression of *novel* miRNAs, predicted by mirnovo and quantified by chimira, across the three sample conditions in pairwise plots. (**B**) Normalized expression of *known* miRNAs, predicted by mirnovo and quantified by chimira, across the three sample conditions in pairwise plots. (**C**) Hierarchical clustering of naïve, primed human ESCs and HEK cells based on novel miRNA expression.

expression also yield similar grouping of the samples based on their cell type (Supplementary Figure S15). This finding illustrates that individual cells may contain a unique novel miRNA signature that is characteristic for the cell type of origin while other cells may show a lower degree of differentiation and thus retain a more generic miRNA expression profile, regardless of cell type.

CONCLUSION

We have demonstrated that machine learning based, genome-free discovery of miRNAs is possible from small RNA sequencing in animal and plant species. Our approach has similar levels of accuracy to the most widely used previously published tool, which utilises genomic information (miRDeep2). Additionally, our approach exceeds miRDeep2's performance when genome information is available and does so at a significantly lower computational cost. This approach has been extensively validated using multiple species, training sets and 10-fold cross validation. Besides, our method has been validated using large-scale datasets and miRNA biogenesis mutant datasets that elucidate potential novel miRNA biogenesis pathways, based on their dependency on different types of RNaseIII enzymes. We have also demonstrated the possibility of discovering novel miRNA candidates from single-cell data, despite their inherent noise, and thus further enable the discovery of novel miRNA molecules associated with very particular cell types and/or conditions.

Moreover, we observed a higher degree of consistency in predicting novel miRNAs in animals than in plants, in terms of the features with the most discriminative power, which complies with the more complex miRNA biogenesis mechanisms present in plants. However, miRNA predictions on plants still managed high levels of accuracy and thus mirnovo can serve additionally as a formidable and easy-to-use resource for researchers of the plants community.

Our method, mirnovo, is simple to install as a command-line tool and may also be used as a user-friendly web-based method. Given the quality of results obtained without genome data, we believe this method could have an important role for miRNA discovery in non-model organisms. We believe that mirnovo represents a significant new contribution to the miRNA field and in particular the prediction of novel miRNAs.

DATA AVAILABILITY

Our method is available as both a web-application (http://wwwdev.ebi.ac.uk/enright-dev/mirnovo) and a stand-alone tool (https://github.com/dvitsios/mirnovo). The coverage profiles and hairpin precursors of all predicted novel miR-

NAs in the GEUVADIS samples are available at the following link:
http://wwwdev.ebi.ac.uk/enright-dev/mirnovo-standalone-pkg/misc/geuvadis-analysis.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank members of the Enright lab (EMBL) for interesting and useful discussions and support.

FUNDING

EMBL core funding; MRC methodology research fellowship [MR/L012367/1 to M.P.D.]. Funding for open access charge: EMBL Core funding.
Conflict of interest statement. None declared.

REFERENCES

1. Nawrocki,E.P., Burge,S.W., Bateman,A., Daub,J., Eberhardt,R.Y., Eddy,S.R., Floden,E.W., Gardner,P.P., Jones,T.A., Tate,J. *et al.* (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, **43**, D130–D137.
2. Limin,F., Beifang,N., Zhengwei,Z., Sitao,W. and Weizhong,L. (2012). CD-HIT: accelerated for clustering the next generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
3. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res.*, **32**, 1792–1797.
4. Griffiths-Jones,S., Saini,H.K., van Dongen,S. and Enright,A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.
5. Friedländer,M.R., Mackowiak,S.D., Li,N., Chen,W. and Rajewsky,N. (2012) miRDeep2 accurately indentifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.*, **40**, 37–52.
6. Hackenberg,M., Rodríguez-Ezpeleta,N. and Aransay,A.M. (2011) miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Res.*, **39**, W132–W138.
7. Wu,J., Liu,Q., Wang,X., Zheng,J., Wang,T., You,M., Sheng Sun,Z. and Shi,Q. (2013) mirTools 2.0 for non-coding RNA discovery, profiling, and functional annotation based on high-throughput sequencing. *RNA Biol.*, **10**, 1087–1092.
8. Hendrix,D., Levine,M. and Shi,W. (2010) miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data. *Genome Biol.*, **11**, R39.
9. Yang,X. and Li,L. (2011) miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics*, **27**, 2614–2615.
10. Mathelier,A. and Carbone,A. (2010) MIReNA finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. *Bioinformatics*, **26**, 2226–2234.
11. Jha,A. and Shankar,R. (2013) miReader: discovering novel mirnas in species without sequenced genome. *PLoS One*, **8**, e66857.
12. Mapleson,D., Moxon,S., Dalmay,T. and Moulton,V. (2013) MirPlex: a tool for identifying miRNAs in high-throughput sRNA datasets without a genome. *J. Exp. Zool. B Mol. Dev. Evol.*, **320**, 47–56.
13. Lomate,P.R., Mahajan,N.S., Kale,S.M., Gupta,V.S. and Giri,A.P. (2014) Identification and expression profiling of Helicoverpa armigera microRNAs and their possible role in the regulation of digestive protease genes. *Insect Biochem. Mol. Biol.*, **54**, 129–137.
14. Friedländer,M.R., Lizano,E., Houben,A.J., Bezdan,D., Báñez-Coronel,M., Kudla,G., Mateu-Huertas,E., Kagerbauer,B., González,J., Chen,K.C. *et al.* (2014) Evidence for the biogenesis of more than 1,000 novel human microRNAs. *Genome Biol.*, **15**, R57.
15. Dhahbi,J.M., Atamna,H., Boffelli,D., Magis,W., Spindler,S.R. and Martin,D.I. (2011) Deep sequencing reveals novel microRNAs and regulation of microRNA expression during cell senescence. *PLoS One*, **6**, e20509.
16. Murakami,Y., Tanahashi,T., Okada,R., Toyoda,H., Kumada,T., Enomoto,M., Tamori,A., Kawada,N., Taguchi,Y.H. and Azuma,T. (2014) Comparison of hepatocellular carcinoma miRNA expression profiling as evaluated by next generation sequencing and microarray. *PLoS One*, **9**, e106314.
17. Chen,X. (2005) MicroRNA biogenesis and function in plants. *FEBS Lett.*, **579**, 5923–5931.
18. Lappalainen,T., Sammeth,M., Friedländer,M.R., 't Hoen,P.A., Monlong,J., Rivas,M.A., Gonzàlez-Porta,M., Kurbatova,N., Griebel,T., Ferreira,P.G. *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**, 506–511.
19. Quah,S., Hui,J.H. and Holland,P.W. (2015) A burst of miRNA innovation in the early evolution of butterflies and moths. *Mol. Biol. Evol.*, **32**, 1161–1174.
20. Kim,Y.K., Kim,B. and Kim,V.N. (2016) Re-evaluation of the roles of DROSHA, Export in 5, and DICER in microRNA biogenesis. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E1881–E1889.
21. Vitsios,D.M. and Enright,A.J. (2015) Chimira: analysis of small RNA sequencing data and microRNA modifications. *Bioinformatics*, **31**, 3365–3367.
22. Ruby,J.G., Jan,C.H. and Bartel,D.P. (2007) Intronic microRNA precursors that bypass drosha processing. *Nature*, **448**, 83–86.
23. Cheloufi,S., Dos Santos,C.O., Chong,M.M. and Hannon,G.J. (2010) A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. *Nature*, **465**, 584–589.
24. Cifuentes,D., Xue,H., Taylor,D.W., Patnode,H., Mishima,Y., Cheloufi,S., Ma,E., Mane,S., Hannon,G.J., Lawson,N.D. *et al.* (2010) A Novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity. *Science*, **328**, 1694–1698.
25. Faridani,O.R., Abdullayev,I., Hagemann-Jensen,M., Schell,J.P., Lanner,F. and Sandberg,R. (2016) Single-cell sequencing of the small-RNA transcriptome. *Nat. Biotechnol.*, **34**, 1264–1266.
26. Nichols,J. and Smith,A. (2009) Naive and primed pluripotent states. *Cell Stem Cell*, **4**, 487–492.
27. Davis,M.P., van Dongen,S., Abreu-Goodger,C., Bartonicek,N. and Enright,A.J. (2013) Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods*, **63**, 41–49.
28. Rognes,T., Flouri,T., Nichols,B., Quince,C. and Mahé,F. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, **4**, e2584.
29. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
30. Wootton,J.C. and Federhen,S. (1993) Statistics of local complexity in amino-acid-sequences and sequence databases. *Comput. Chem.*, **17**, 149–163.
31. Trifonov,E.N. (1990) In: Sarma,RH and Sarma,MH (eds). *Making Sense of the Human Genome*. Structure & Methods Adenine Press, Albany, Vol. **1**, pp. 69–77.
32. Lorenz,R., Bernhart,S.H., Höner Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
33. Leinonen,R., Akhtar,R., Birney,E., Bower,L., Cerdeno-Tárraga,A., Cheng,Y., Cleland,I., Faruque,N., Goodgame,N., Gibson,R. *et al.* (2011) The European nucleotide archive. *Nucleic Acids Res.*, **39**, D28–D31.
34. Ha,M. and Kim,N. (2014) Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.*, **15**, 509–524.