Theor Chem Acc (2010) 125:637–642
DOI 10.1007/s00214-009-0684-z

REGULAR ARTICLE

Computational classification of microRNAs in next-generation sequencing data

Joshua Riback · Artemis G. Hatzigeorgiou · Martin Reczko

Received: 26 June 2009 / Accepted: 2 November 2009 / Published online: 26 November 2009

Abstract MicroRNAs (miRNAs) have been shown to play an important regulatory role in plants and animals. A large number of known and novel miRNAs can be uncovered from next-generation sequencing (NGS) experiments that measure the complement of a given cell's small RNAs under various conditions. Here, we present an algorithm based on radial basis functions for the identification of potential miRNA precursor structures. Computationally assessing features of known human miRNA precursors, such as structural linearity, normalized minimum folding energy, and nucleotide pairing frequencies, this model robustly differentiates between miRNAs and other types of non-coding RNAs. Without relying on cross species conservation, the method also identifies non-conserved precursors and achieves high sensitivity. The presented method can be used routinely for the identification of known and novel miRNAs present in NGS experiments.

Dedicated to Professor Sandor Suhai on the occasion of his 65th birthday and published as part of the Suhai Festschrift Issue.

Electronic supplementary material The online version of this article (doi:10.1007/s00214-009-0684-z) contains supplementary material, which is available to authorized users.

J. Riback · A. G. Hatzigeorgiou · M. Reczko (✉)
Institute of Molecular Oncology, Biomedical Sciences Research Center "Alexander Fleming", 16602 Varkiza, Greece
e-mail: reczko@fleming.gr; mreczko@gmail.com

A. G. Hatzigeorgiou
Department of Computer and Information Sciences, University of Pennsylvania, Philadelphia, PA, USA

M. Reczko
Synaptic Ltd., Heraklion, Greece

Keywords Next-generation sequencing · Deep sequencing · Non-coding RNA · microRNA · Gene finding

1 Introduction

Knowledge of non-coding RNAs (ncRNAs) is integral for understanding complex mechanisms occurring within the cell [1]. Types of ncRNAs include rRNAs, tRNAs, miRNAs, siRNAs, snRNAs, snoRNAs and scaRNAs; functions of the types of ncRNAs vary extensively with cellular functions, such as cell development or differentiation, and their deregulation is frequently associated with diseases such as cancer [2]. One of the most important ncRNAs are miRNAs because, while only being around 22 nucleotides long, this RNA plays an important role in plants and animals in regulating the protein levels of many genes involved in functions such as developmental timing, apoptosis, cell proliferation and differentiation, anti-viral defense and hypothetically many other cellular roles [3, 4]. Since miRNA has the ability to regulate many vital biological functions, it formulates a plethora of opportunities for miRNA research [3]. Only when scientists started to understand how miRNAs function within a cell did the importance of microRNA research become comprehendible [5].

The expression of miRNA begins in the nucleus as a long strand of primary miRNA (pri-miRNA), which, in animals, is then processed by the protein complex Drosha into a precursor miRNA (pre-miRNA) [4]. Exportin-5 then transports the pre-miRNA into the cytoplasm, where it is processed by the protein Dicer into an miRNA duplex [4]. The strand in the duplex, complementary to the final miRNA strand, is usually degraded [4]. Then, miRNAs bind to mRNAs and regulate their expression with the aid

of the RNA interference silencing complex (RISC) [4, 6]. Argonaute (AGO), one of the proteins that make up RISC, either cleaves the mRNA or represses translation, causing the inevitable degradation of the mRNA [4]. By determining the features of miRNA processing, novel miRNAs can be predicted with relatively high accuracy [2, 7, 8].

The discovery of novel ncRNAs is difficult because the methods used to isolate small RNAs can be inaccurate and inefficient. Next-generation sequencing (NGS) technologies (also called flow cell-, massively parallel- or deep-sequencing) are considerably cheaper [9] and can detect many small RNAs with a higher degree of reliability than methods such as Sanger sequencing protocols and cloning [7]. Drawbacks in the use of deep sequencing include the extensive amount of data necessary to organize from a collection, because each sequence read can align to multiple areas on the genome even without consideration of a possible nucleotide error within the read [2, 8, 10]. Sorting and compressing deep-sequencing data is necessary so that the information can be easily accessed for the researcher's purpose. An ideal use of the deep-sequencing sets is the expression profiling of known and novel miRNAs because of their similar length distribution and their relatively high abundance within the reads [2].

When searching for potential miRNAs from within the genome, the tradeoff between more false positives, if miRNA features are not strict enough, or eliminating potential miRNAs, if miRNA features are too firm [2, 7, 8], has to be balanced. Sequence conservation and structural information of previously found miRNA precursors can be exploited for the purpose of finding novel miRNAs. Algorithms to this end heavily rely on conservation information and exclude a large number of non-conserved miRNAs [10]. More recent methods thus focus on the ab initio prediction of miRNAs, without relying on comparative genomics. The assessment of thermodynamic properties of the miRNA precursor is fairly common because of the relative stability of miRNAs when compared to most other ncRNA [1, 10, 11]. Using thermodynamics alone, however, can either create too many potential pre-miRNA, producing low specificity (ratio of correctly identified non-miRNA RNAs (the negatives) over all negatives), or result in low sensitivity when strict cutoffs are applied. Additional features such as GC content, number of paired bases, negative normalized minimum folding energy (nnMFE) and linearity [10–12] are used to increase the specificity. Some miRNA gene finders are restricted to the analysis of NGS data, comparable to the presented approach. The first such approach, miRDeep [7], uses a model that reflects the enrichment of sequence fragments according to the cleavage of the pre-miRNAs with Dicer. The method used in Burnside et al. [13] relies only on structural features of the predicted precursor hairpins.

In the following, we present a pre-miRNA prediction algorithm tailored for the elimination of ncRNAs that are not miRNAs in NGS data, without the use of conservation information.

2 Results

2.1 Assessing potential features and miRNA scoring

Predicting potential pre-miRNAs using NGS data has different requirements than prediction based on genomic sequences alone, because roughly 60% of the NGS reads can be mapped to miRNAs [2]. This entails a new class of prediction algorithms that are less reliant on typical miRNA features to reach higher prediction sensitivities.

The most commonly used features are normalized free energy, paired/unpaired bases ratio, GC content, hairpin structure linearity and evolutionary conservation [10–12]. We exclude conservation from this analysis to enable the prediction of novel non-conserved mature miRNAs. The most informative features and their definitions are:

GC content is defined as the percentage of guanine or cytosine nucleotides in the given sequence.

Negative normalized minimum folding energy (nnMFE) is calculated as the negative minimum free energy (MFE) as obtained from RNAfold program of the Vienna package [14] divided by the length of the folded sequence. As larger structures can obtain lower MFEs, this normalization compensates for the correlation between the free energy and the sequence length (Pearson's $R^2 = 0.15$).

Ratios of A to U (ratio of G to C) are characterized by the number of adenine (guanine) divided by the number of uracil (cytosine) nucleotides in the potential precursor.

Unbound nucleotide percentage is the number of unpaired nucleotides normalized by the length of the potential precursor.

Linearity can be described as how relatively straight the hairpin structure of the precursor is and determined by the amount of loops, which form against the direction of the hairpin. As an approximation, linearity is calculated as follows. All unpaired nucleotides are removed from the structure, all opening hydrogen bonds in the first half of the resulting structure are counted and this count is divided by the total length of the first half of this structure, representing the linearity score.

To determine the importance of features to find novel pre-miRNAs, we assess their ability to discriminate miRNAs among other ncRNAs, especially rRNAs and tRNAs. A feature significance is defined as the ratio of the number of known miRNAs to the number of all ncRNAs including miRNAs, where both numbers count the cases in

which the value of the feature is within the 95% percentile around the mean of that feature for known human miRNAs.

The feature significances are shown for each ncRNA type found in Ensembl in Fig. 1. In these comparisons, a significantly higher value of the linearity of miRNAs compared to all other ncRNAs is already evident. The class with the largest feature significance differences compared to miRNA is the small nuclear RNAs (snRNAs) that show both relatively low energetic stability and the lowest structural linearity. This can be ascribed to the known occurrence of larger unpaired segments in these structures. Linearity had the highest ratio of 54%, confirming its importance found also in earlier studies [11]. Interestingly, the GC content showed low importance (15%) for the discrimination between miRNAs and other ncRNAs, in contrast to previous studies [10, 11] that contrast miRNAs with other hairpin structures as negative examples. This exemplifies the importance of the classification scheme used in the following.

2.2 Pre-miRNA similarity analysis

A discretized radial basis classifier using the miRNA features, linearity, negatively normalized MFE and normalized unbound nucleotides, is used to identify miRNA. Analysis of these three features shows (Fig. 2) that the histograms of known miRBase pre-miRNAs are normally distributed (two sided for nnMFE and unbound nucleotides and one sided for linearity). A discretized quadratic fit to the distributions of these features is used as the radial basis function independently for each feature. The radial basis centers are approximated at 0.44 for the nnMFE, 1 for linearity and 0.28 for normalized unbound nucleotides. The pre-miRNA similarity score (miSA score) for a potential miRNA is defined as the sum of the scores for each of the three features according to Table 1. To account for the higher significance of the linearity feature described above, a maximal score of 7 is used for this feature, while the other features may contribute a maximal score of 6.

2.3 miRNA precursor prediction performance and tests on NGS data

When comparing the predicted miSA scores for all types of known human ncRNAs, the scores for miRNAs are significantly higher than for any other ncRNA class (Fig. 3). The ncRNA class having scores closest to the miRNA family is the small cytoplasmic RNAs (scRNA) that shows both high stability and a large fraction of bound nucleotides similar to miRNAs. This is due to their typical structure with a large amount of helical regions, but the branched arrangement of these helices leads to a low linearity score. The high specificity discriminating against all other ncRNAs for scores above 10 is also evident from the ROC curve in Fig. 4. These classifications are obtained from all ncRNAs present in the NGS data set. At this cutoff, a sensitivity of 63% is achieved with a false positive rate of 3%. The area under the ROC curve (AUC) sums up to 0.89. The scores for the 134 known miRNAs and additional 2535 potential miRNAs obtained from scoring all hairpins on the human genome with lengths between 40 and 150 nucleotides and having overlapping reads in the NGS data are available as supplementary material 1.

3 Conclusions

The rapidly increasing use of NGS for unbiased measurement of small RNA expression levels produces large amounts of data in which both known and novel ncRNA have to be mined. The method presented here for pre-miRNA prediction in this type of data is designed to obtain

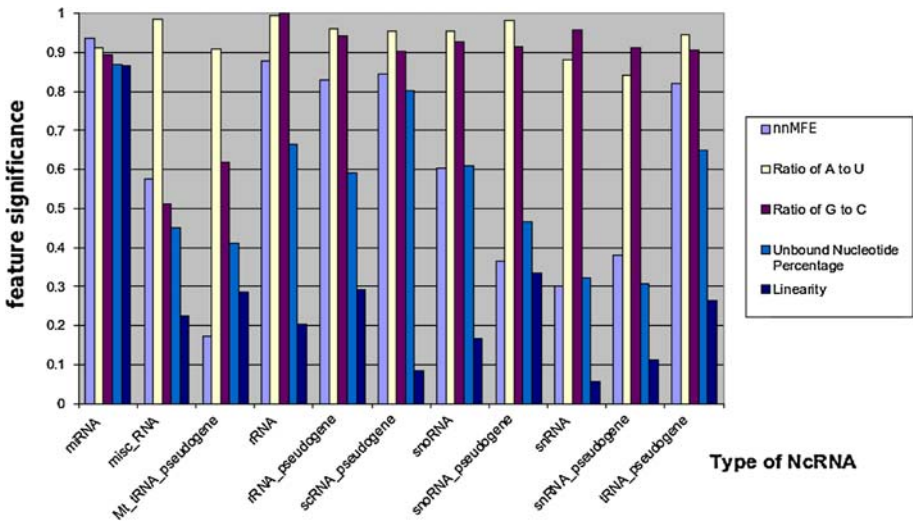

Fig. 1 Assessing potential features. The feature significance, defined as the ratio of the number of miRNAs to the total number of ncRNAs, which are in the 95% percentile of the features distribution on mirBase miRNAs, is shown for each type of Ensembl-determined ncRNA (see Methods) and for the features: nnMFE, ratio of A to U, ratio of G to C, unbound nucleotide percentage and linearity

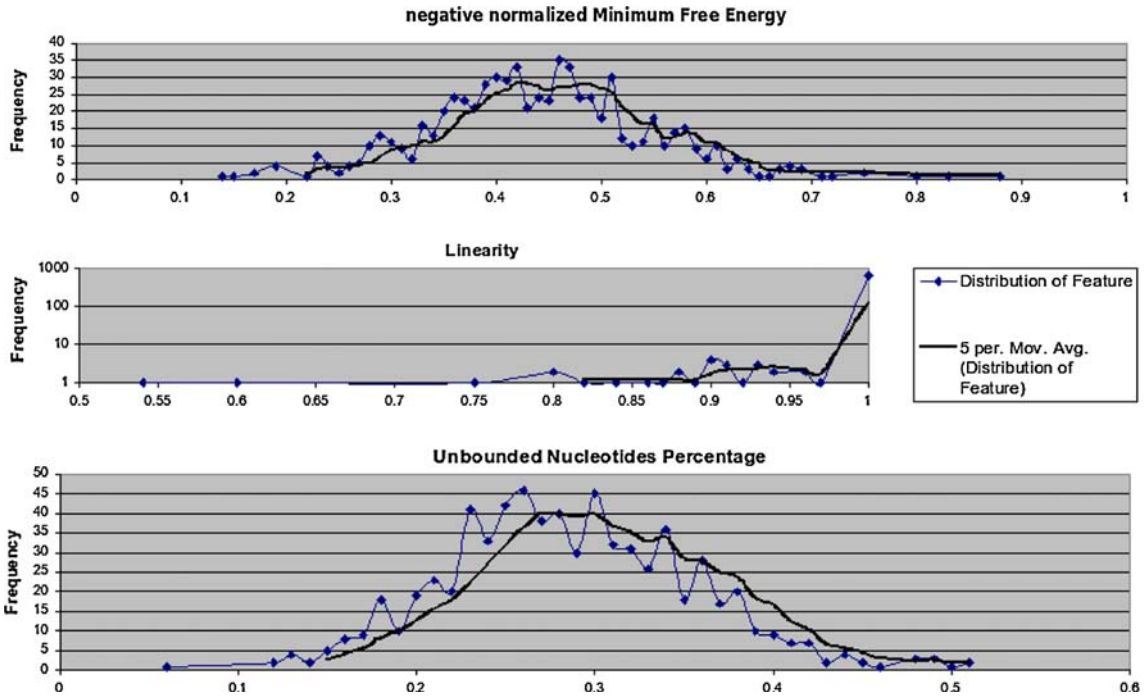

Fig. 2 Analyzing motifs of precursor features. The three features used for classification are calculated for all human miRBase pre-miRNAs. An enrichment of linear structures can be observed and this feature, representing the relative straightness of the hairpin, is most significant for the classification

Table 1 Precursor analysis scoring system

Percentage of pre-miRNA	Difference from mean	Value added to score	nnMFE
0.988	0.27	1	
0.940	0.19	2	STD
0.842	0.14	3	0.104
0.748	0.11	4	
0.570	0.07	5	
0.363	0.04	6	
Percentage of pre-miRNA	Difference from mean	Value added to score	Linearity
0.996	0.20	1	
0.991	0.16	2	STD
0.960	0.00	7	0.032
Percentage of pre-miRNA	Difference from mean	Value added to score	Unbound
0.990	0.2	1	
0.951	0.13	2	STD
0.882	0.1	3	0.069
0.788	0.08	4	
0.581	0.05	5	
0.393	0.03	6	

This table defines the scoring of potential pre-miRNA. Three groups of scores are defined for the three features used. The 'percentage of pre-miRNA' indicates the fraction of miRBase human miRNAs where the feature of the group is within the range defined in 'difference from mean'. The used means for the scores are 0.44 for nnMFE, 1 for linearity, and 0.28 for normalized unbound nucleotides. From each feature only one value is added toward the accumulative score. (STD is the standard deviation)

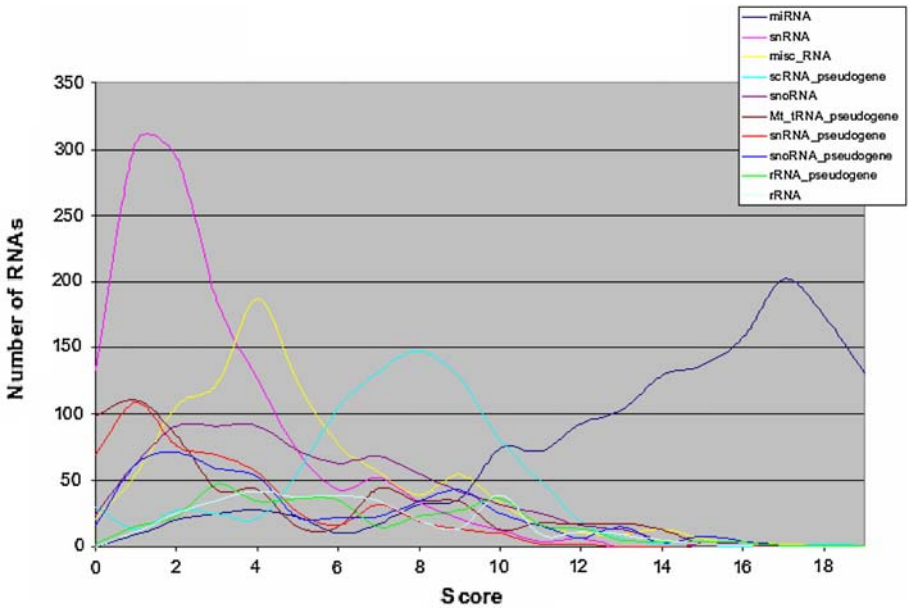

Fig. 3 Pre-miRNA similarity scores for NcRNA. Using data from all annotated ncRNAs, a histogram of the scores from the Precursor Analysis Scoring System (miSA scores) demonstrates the ability to classify miRNAs versus other ncRNAs

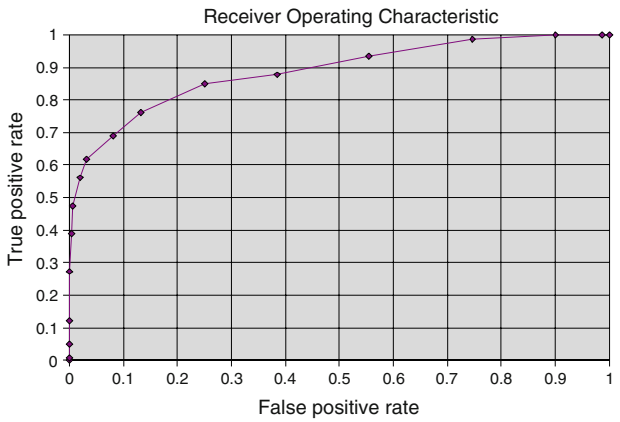

Fig. 4 ROC analysis for detecting miRNAs at different thresholds. Using data from Human HeLa Cells NGS data set for only small RNAs [7], the sensitivity (true positive rate) for detecting miRNAs and the specificity (false positive rate) is shown for different cutoffs of the miSA scores. The threshold values from 17 to 0 are indicated with *dots*

high sensitivity by avoiding the requirement of any evolutionary conservation. The feature selection using all types of known ncRNAs results in a reliable discrimination between miRNAs and all other classes of ncRNA. This aspect has not been addressed adequately in the related work by [7], where all known other ncRNAs are removed before scoring. However, as the set of ncRNAs is still incomplete, many unknown non-miRNA ncRNAs might be classified as false positives. The sensitivity of the method presented here is likely to increase novel miRNA retrieval. If the number of potential novel miRNAs has to be reduced further, a subsequent cross species conservation analysis may be performed.

4 Materials and methods

4.1 Data sets

4.1.1 NGS data

The NGS data set of Human HeLa Cells for small RNAs used by (77) is used. (Data Set available at: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10829). This set contains more than 320,000 sequences of 35 nucleotides length each.

4.1.2 MirBase

All miRNA sets used for training are from the human sequences in miRBase 12.0 [15]. A total of 695 known human pre-miRNA and corresponding mature miRNAs are used.

4.1.3 Ensembl

The annotation of noncoding RNAs is extracted from Ensembl release 50 and includes both predicted and known ncRNAs [16].

4.2 Short sequence genomic alignments

4.2.1 MegaBLAST

The NGS data was aligned to the human genome sequences (NCBI Build 36.1) using the MegaBLAST tool [17] to allow for a small number of mismatches, insertions and

deletions in the same way as that described in the miRDeep approach [7].

4.3 RNA structure prediction

4.3.1 Vienna package

All RNA secondary structure and energy (MFE) predictions for the potential precursors are determined by the Vienna Package 1.7.2 [14]. In accordance with most other miRNA gene finding approaches, only the energetically best structure is considered.

Acknowledgments We would like to acknowledge all anonymous reviewers for their helpful comments and suggestions.

References

1. Kawaji H, Nakamura M, Takahashi Y, Sandelin A, Katayama S, Fukuda S, Daub CO, Kai C, Kawai J, Yasuda J, Carninci P, Hayashizaki Y (2008) Hidden layers of human small RNAs. BMC Genomics 9:57–58
2. Kawaji H, Hayashizaki Y (2008) Exploration of small RNAs. PLOS Genet 4:e22
3. Enright AJ, Jogn B, Gaul U, Tuschl T, Sander S, Marks DS (2003) MicroRNA targets in *Drosophila*. Genome Biol 5:R1
4. Wang Y, Stricker HM, Gou D, Liu L (2007) MicroRNA: past and present. Front Biosci 12:2316–2329
5. Maziere P, Enright A (2007) Prediction of microRNA targets. Drug Discov Today 12(11–12):452–458
6. Brennecke J, Stark A, Russell RB, Cohen SM (2005) Principles of microRNA-target recognition. PLoS Biol 3(3):e85
7. Friedlander MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N (2008) Discovering microRNAs from deep-sequencing data using miRDeep. Nat Biotechnol 26:407–415
8. Stark A, Kheradpour P, Parts L, Brennecke J, Hodges E, Hannon GJ, Kellis M (2007) Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes. Genome Res 17:1865–1879
9. Holt RA, Jones SJ (2008) The new paradigm of flow cell sequencing. Genome Res 18:839–846
10. Sheng Y, Engström PG, Lenhard B (2007) Mammalian microRNA prediction through a support vector machine model of sequence and structure. PLoS One 2:e946
11. Szafranski K, Megraw M, Reczko M, Hatzigeorgiou GH (2006) Support vector machine for predicting microRNA hairpins. In: Proceedings of the 2006 international conference on bioinformatics and computational biology, pp 270–276
12. Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z (2007) MiPred:classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Res 35(Web Server issue): W339–W344
13. Burnside J, Ouyang M, Anderson A, Bernberg E, Lu C, Meyers B, Green P, Markis M, Isaacs G, Huang E, Morgan R (2008) Deep sequencing of chicken microRNAs. BMC Genomics 9:185
14. Hofacker IL (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31(13):3429–3431
15. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. Nucleic Acids Res 36:D154–D158
16. Hubbard JP, Aken BL, Beal1 K, Ballester1 B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Overduin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E (2007) Ensembl 2007. Nucleic Acids Res 33 (Database issue):D447–D453
17. Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA (2008) Database indexing for production MegaBLAST searches. Bioinformatics 15:1757–1764