Genomics 100 (2012) 141–148

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

MetaMirClust: Discovery of miRNA cluster patterns using a data-mining approach

Wen-Ching Chan [a,b,c], Meng-Ru Ho [d], Sung-Chou Li [d], Kuo-Wang Tsai [f], Chun-Hung Lai [e], Chun-Nan Hsu [c,g], Wen-chang Lin [a,e,*]

[a] Institute of Biomedical Informatics, National Yang-Ming University, Academia Sinica, Taipei, Taiwan, ROC
[b] Bioinformatics Program, Taiwan International Graduate Program, Academia Sinica, Taipei, Taiwan, ROC
[c] Institute of Information Sciences, Academia Sinica, Taipei, Taiwan, ROC
[d] Genomics Research Center, Academia Sinica, Taipei, Taiwan, ROC
[e] Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan, ROC
[f] Department of Medical Education and Research, Kaohsiung Veterans General Hospital, Kaohsiung, Taiwan, ROC
[g] Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA

ARTICLE INFO

Article history:
Received 5 March 2012
Accepted 15 June 2012
Available online 23 June 2012

Keywords:
MicroRNA
MiRNA cluster
Data mining
Metazoan

ABSTRACT

Recent genome-wide surveys on ncRNA have revealed that a substantial fraction of miRNA genes is likely to form clusters. However, the evolutionary and biological function implications of clustered miRNAs are still elusive. After identifying clustered miRNA genes under different maximum inter-miRNA distances (MIDs), this study intended to reveal evolution conservation patterns among these clustered miRNA genes in metazoan species using a computation algorithm. As examples, a total of 15–35% of known and predicted miRNA genes in nine selected species constitute clusters under the MIDs ranging from 1 kb to 50 kb. Intriguingly, 33 out of 37 metazoan miRNA clusters in 56 metazoan genomes are co-conserved with their up/down-stream adjacent protein-coding genes. Meanwhile, a co-expression pattern of *miR-1* and *miR-133a* in the *mir-133-1* cluster has been experimentally demonstrated. Therefore, the MetaMirClust database provides a useful bioinformatic resource for biologists to facilitate the advanced interrogations on the composition of miRNA clusters and their evolution patterns.

1. Introduction

MicroRNAs (miRNAs) are endogenous small non-coding RNA molecules of 21–23 nucleotides (nt) long in length. They play important roles in gene regulation via the RNA interference pathway [1–4]. As a wide range of investigations has been conducted on miRNA genes, a consensus scenario of miRNA biogenesis has been currently revealed. Initially, miRNA genes are transcribed from intergenic or intronic regions by RNA polymerase II [5] or III [6], and generate primary miRNA transcripts (pri-miRNAs) in the nucleus [2]. Within the same organelle, these transcripts are processed by the RNase III Drosha endonuclease, which is associated with its co-factor DGCR8 complex, into precursor forms (pre-miRNAs) about 70–90 nt [7,8]. The canonical pre-miRNAs are 70–90 nt long and fold-back to form stem-loop structures, which are characteristic secondary structures of miRNAs. Subsequently, these miRNA molecules are exported as single hairpins into the cytoplasm by the aid of Exportin 5 (XPO5) [9]. Cleaved by another RNase III Dicer enzyme, pre-miRNA hairpins are processed into double-stranded mature miRNA duplexes [10]. Preferentially, one of the single-strand mature miRNA is incorporated into the RNA-induced silencing complex (RISC) or different ribonucleoproteins (miRNPs), while the remaining strand is degraded rapidly [2,11]. Primarily depending on the degree of complementarity in sequences, the binding of miRNAs to the 3′ untranslated regions (3′ UTR) of their mRNA targets gives rise to two down-regulation mechanisms, mRNA degradation and translation repression. Up to date, a number of examples of mRNA cleavage instances have been reported in plant, and alternatively translation repression is the main mechanism observed in animal cells.

A substantial amount of literature has demonstrated miRNAs as crucial negative regulators in diverse physiological and developmental processes at the post-transcriptional level. In 1993 when the first miRNA *lin-4* was identified in *Caenorhabditis elegans*, the negative regulation pair between *lin-4* and its target *lin-14* was thought of as an individual case [12]. In fact, miRNAs have not gained the attention of researchers until a second similar system of *let-7* was observed in *C. elegans* [13], and then its homologous transcripts were extensively investigated in animal genomes. Thereafter, a considerable body of evidence suggests that miRNAs play important gene-regulatory roles related to organism development, cell differentiation, and tumor suppression and oncogenesis [1,12,13]. Currently, newly discovered miRNA genes either by experimental or computational approaches have steadily increased as evident by the amount of records in the miRBase registry [14]. In recent years, many studies have attempted to provide insight into the biogenesis, expression,

* Corresponding author at: Institute of Biomedical Informatics, National Yang-Ming University, Academia Sinica, Taipei, Taiwan, ROC.
E-mail addresses: wenching.chan@gmail.com (W.-C. Chan), mirrian@gmail.com (M.-R. Ho), pinus@ibms.sinica.edu.tw (S.-C. Li), tomabe33@pchome.com.tw (K.-W. Tsai), lai2213@yahoo.com (C.-H. Lai), chunnan@iis.sinica.edu.tw (C.-N. Hsu), wenlin@ibms.sinica.edu.tw (W. Lin).

0888-7543/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2012.06.007

targeting and evolution of individual miRNA genes in different species. Some well-studied examples in humans include, for instance, *mir-196* which governs the cleavage of homeobox (HOX) gene clusters [15]; *mir-375* which targets Myotrophin (MTPN), and both *mir-196* and *mir-375* are related to glucose-stimulated insulin secretion and exocytosis [16]; and *mir-143* which regulates adipocyte differentiation [17]. All these studies focused on the discovery of the biological functions of limited individual miRNA genes, not the clustered miRNAs.

Up until the present, a handful of miRNA clusters has been reported in animal genomes. To the best of our knowledge, Altuvia et al. was the first group that identified conserved regions of miRNA clusters systematically [18]. Then, Yu et al. [19] adopted the same method to enlarge the extent of conserved miRNA cluster, and thus checked the expression profile of identified human miRNA clusters. To group two or more miRNA genes in chromosomal distance within at most 3000 nt, they used available 326 human miRNA genes from the miRBase registry for clustering analysis. Finally, they identified 51 clusters composed of 148 miRNA genes and created 9 distinct paralogous clusters. Furthermore, accumulating studies have illustrated that clustered miRNA genes located on polycistronic transcripts might be expressed at similar levels and coordinately involve in an intricate regulatory network. These miRNA clusters are usually derived from polycistrons with lengths from few hundred nucleotides to almost a million base pairs [20,21]. For instance, the *mir-17* cluster and its paralogous clusters are one of the well-studied cases. In 2004, Tanzer et al. have tried to reconstruct the phylogenetic evolution of the *mir-17* cluster family mainly in nine metazoan genomes and have concluded that: (i) *mir-17* miRNA cluster consisting of six precursor miRNAs was within about 1 kb distance on chromosome 13 in humans and (ii) at least three paralogous clusters were related to the *mir-17* cluster family, which are *mir-17-92*, *mir-106-92*, and *mir-106-25*, and governed by tandem duplications [22]. Meanwhile, several studies have further demonstrated that the *mir-17* cluster family plays an important role in cell proliferation, organism development and cancer oncogenesis [23,24]. Although the entire regulatory mechanisms of clustered miRNA genes remain largely uncharacterized, it is likely that these miRNA clusters may function more efficiently in a complicated miRNA-mediated network than individual miRNAs alone [25].

Many resources were developed to investigate miRNA genes. However there is no any resource that emphasizes an efficient and comprehensive investigation of miRNA clusters. Formerly, miRNA clusters were arbitrarily defined by a fixed distance, but there was no report systematically investigating the conservation patterns of clustered miRNA genes across metazoan species. In this study, we introduced a data mining approach to efficiently discover highly conserved sets of miRNA genes upon miRNA clusters, and to facilitate researches on the conservation pattern of the clustered miRNA genes. Based on our previous homologous search of miRNA genes in animal genomes, we first performed the identification of miRNA clusters (MirClust). This identification is based on miRNA classes with respect to different maximum inter-miRNA distances (MIDs) discretely ranging from 1 kb to 50 kb. Despite the singleton miRNA classes, we utilized the FP-growth algorithm to efficiently discover the conserved co-occurrence of miRNA clusters among the miRNA clusters defined under the same MID. The FP-growth algorithm is one of highly efficient data mining methods for discovering frequent co-occurrence patterns from huge datasets such as biological sequences and gene-expression data [26]. It has been applied so far to gain insights into various bioinformatic studies [27–29]. We have now constructed a database (MetaMirClust) for interrogating the origin and conservation of miRNA clusters on a species-wide scale (http://fgfr.ibms.sinica.edu.tw/MetaMirClust/). Researchers can choose proper distances to determine a miRNA cluster for individual species or specific miRNA cluster between different MIDs for comparison. This study is the first attempt to make a feasible dataset for surveying the different recruitments of miRNA genes in homologous miRNA clusters and for comparison of the miRNA gene compositions/structures of miRNA clusters conserved in metazoan species.

2. Results

2.1. Clustering of miRNA genes with respect to four different MIDs

The usage instruction of the MetaMirClust database is illustrated at http://fgfr.ibms.sinica.edu.tw/MetaMirClust/data/tutorial.pdf. Here, consecutive miRNA genes located in the same strand of individual chromosomes are grouped to constitute a cluster determined by pair-wise adjacent distances, or MIDs. Supplementary Table S1 presents the number of identified miRNA clusters with respect to four different MIDs in 56 metazoan genomes. Different clusters of miRNA genes may be obtained when different MIDs were employed to create a cluster. As shown in Supplementary Table S1, by increasing MIDs from 1 kb to 50 kb, the number of miRNA clusters changes from 66 to 100 in the human genome (hg19). To analyze the proportion of miRNA genes used to form a miRNA cluster, we divided the number of clustered miRNA genes to all miRNA genes for individual species. Meanwhile, we chose nine representative species, which are *C. elegans* (worm, ce6), *Drosophila melanogaster* (fly, dm3), *Danio rerio* (zebrafish, danRer6), *Gallus gallus* (chicken, galGal3), *Canis familiaris* (dog, canFam2), *Bos taurus* (cow, bosTau4), *Mus musculus* (mouse, mm9), *Rattus norvegicus* (rat, rn4), and *Homo sapiens* (human, hg19), for comparison throughout this study. Fig. 1 illustrates the proportion of clustered miRNA genes with respect to four different MIDs. Overall, an average of 15–35% of miRNA genes is densely located and form clusters in metazoan species. Distinctly, zebrafish and chicken possess the highest and the lowest frequencies of miRNA genes to form clusters, respectively. Furthermore, the dramatic increasing densities of clustered miRNA genes under the MID of 50 kb in worm and fly likely resulted from relatively small genome sizes. With the different selections of MIDs, a miRNA cluster defined by a long MID could result either by only merging singleton miRNA genes or by enlarging small miRNA clusters determined in a short MID. Fig. 1 indicates that more singleton miRNA genes are recruited to join clusters when considering a long MID. However, in order to clarify the impact of individual factors, we calculated the detailed events for each factor. By comparing miRNA clusters discovered in a short MID to those in a longer one, three scenarios are discovered: 1) forming a new miRNA cluster by merging singleton miRNA genes, 2) enlarging a small miRNA cluster by recruiting singleton miRNA genes, and 3) producing a large miRNA cluster by merging at least two small miRNA clusters. As reported in Table 1, the domination of cases in the first and second scenarios over the third scenario indicate that more singleton miRNA genes are certainly utilized to constitute miRNA clusters under a long MID, as observed in Fig. 1. Intriguingly,

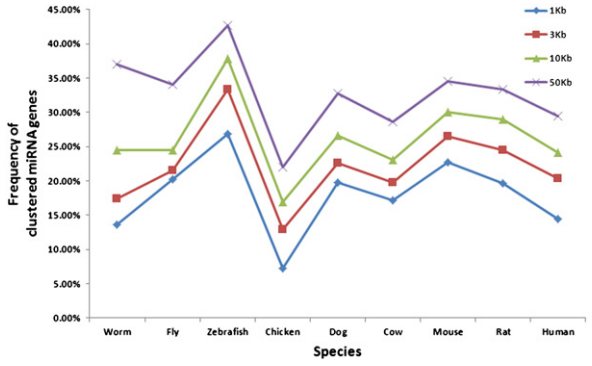

Fig. 1. Usage frequency of miRNA genes in clusters. Total miRNA genes for nine selected representative species are 265 (worm), 302 (fly), 551 (zebrafish), 751 (chicken), 681 (dog), 869 (cow), 1021 (mouse), 722 (rat), and 1265 (human). By taking the total number of miRNA genes used to generate clusters over all miRNA genes for individual species, this figure presents the proportion of clustered miRNA genes with respect to four different MIDs in nine representative species.