Mapping randomly generated short k-mers to genomes

Dave Tang
RIKEN Yokohama
@davetang31

Derek de Rie
VU University Amsterdam
@derekderie

June 26, 2014

Abstract

Eukaryote genomes are composed of genic and non-genic stretches of DNA and depending on the organism, the ratio between the two are extremely variable. For example, the completion of the human genome revealed that only 1-2% of the human genome is made up on protein-coding genes and the rest is composed of non-coding DNA. The portion of non-coding DNA is largely made up of repetitive elements, such as transposons and satellite repeats.

Despite their size, microRNAs (miRNAs) can have dramatic effects on human health due to their direct effect on transcripts. The misregulation of miRNAs can cause a wide spectrum of diseases, such as cancer, and as such they are under heavily investigation.

1 Introduction

MicroRNAs (miRNAs) were discovered in 1993[1] and are the most well known class of non-coding RNAs (ncRNAs). Since its discovery, *miRBase*, a widely used database of miRNA annotations, has accumulated nearly 25,000 miRNA loci in over 200 species. The flood in annotations has been paired with an extensive amount of literature published on miRNA. In the nearly 15 years since the definition of a miRNA was introduced, NCBI's PubMed has collected over 30,000 publications containing the term microRNA (Figure 1).

The characteristic biogenesis pathways of this group of ncRNA lead to the definition of the term microRNA in order to distinguish these transcripts from simliar-sized siRNA. In the canonical biogenesis pathway, mature (single stranded, $\sim$ 22-nt RNA molecules) miRNA are transcribed and subsequently processed through a variety of proteins. After transcription, primary miRNA is cleaved by the Drosha-DGCR8 complex into hairpin-shaped precursor-miRNA. These hairpins are exported into the cytosol by Exportin-5, through which the endonuclease Dicer cleaves the hairpins into $\sim$ 22-nt duplexes. After the two strands separate one is loaded into the RNA-Induced Silencing Complex (RISC),

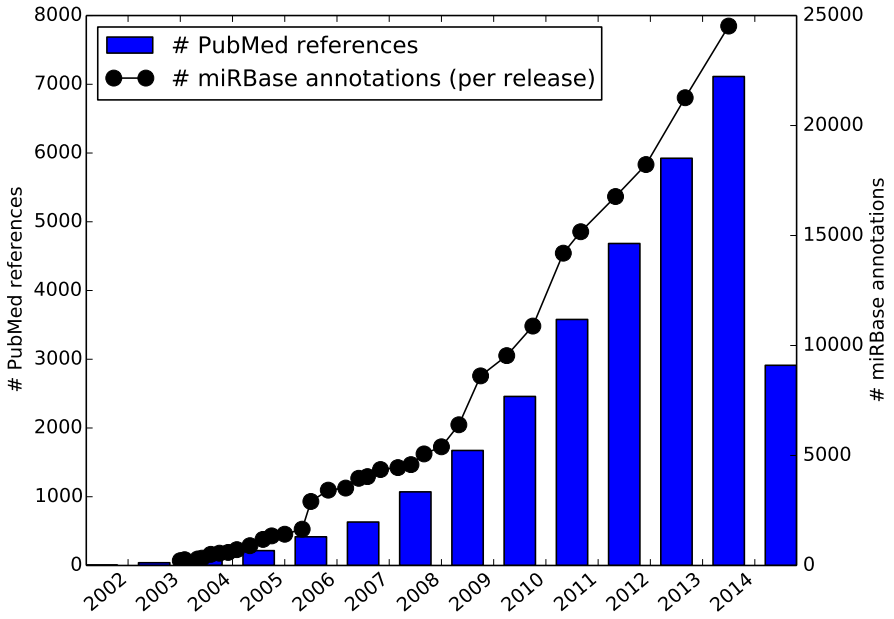


Figure 1: The miRBase miRNA annotation collection has grown with the scientific community's interest in miRNAs. (Sources: http://www.mirbase.org/, http://www.ncbi.nlm.nih.gov/pubmed)

through with miRNA exercise their regulatory functions. Among all the steps in this pathway, Exportin-5 and Dicer depend on the length of the miRNA.

The main function of miRNA is to target and suppress the translation of mRNA transcripts through either destroying the mRNA transcript or inhibiting mRNA translation. Typically, miRNA suppress gene expression through increasing the degradation rate of mRNA transcripts (referred to as mRNA destabilisation). In most cases, miRNA bind to the 3' UTR of mRNA transcripts, although exceptions have been reported. The deciding factor in the target selection has been the miRNA seed region, bases 1 through 8 at the 5' end of the mature miRNA. Imperfect binding in the seed region can be compensated by sequence complementarity in the remain gin part of the mature miRNA. In plants, perfect Watson-Crick base-pairing for the mature miRNA is required for translational inhibition.

2 Methods

All code underlying this work is available at `https://github.com/davetang/mirna_length`. Briefly, random sequences were generated using R (version

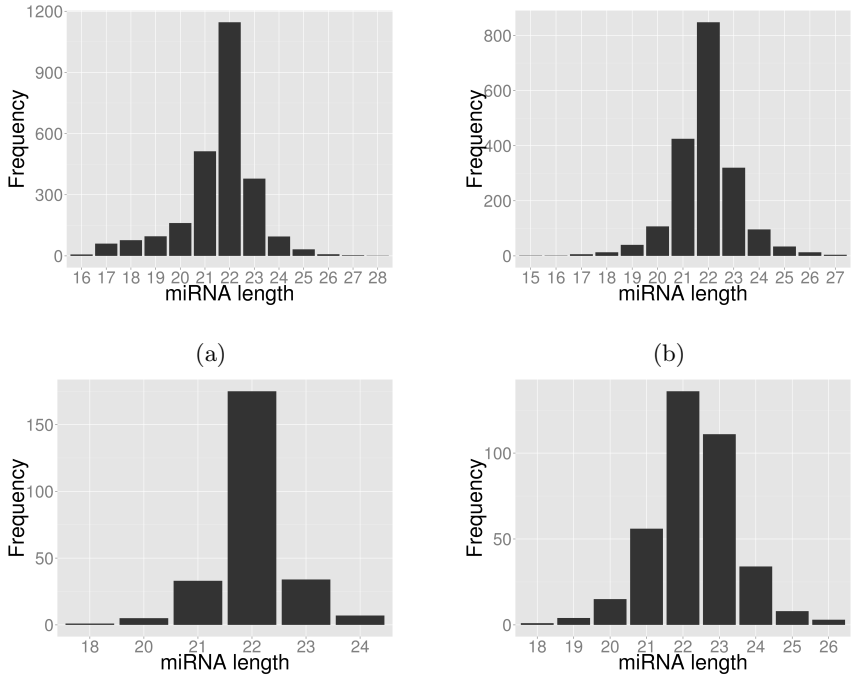


Figure 2: Length of miRNAs from miRBase for (a) human, (b) mouse, (c) zebrafish, and (d) nematode.

3.1.0) and the R Bioconductor package Biostrings[2]. We generated three sets of one million random sequences that ranged from 15 to 30 base pairs in length, totalling 48 million sequences. The first two sets of sequences were generated based on the multinomial sequence model where each nucleotide in the sequence is independent and identically distributed based on a probability. The first set of randomly generated sequences used an equal probability for each nucleotide ($p_a = 0.25$, $p_c = 0.25$, $p_g = 0.25$, $p_t = 0.25$) and the second set used probabilities based on the nucleotide frequency observed in the human genome (hg38): $p_a = 0.29$, $p_c = 0.20$, $p_g = 0.21$, $p_t = 0.30$. The third set of randomly generated sequences used a Markov chain model, where the next sequence depends on the previous sequence. Transitions probabilities (Figure 3) were derived from the dinucleotide frequencies observed from a set of mature human miRNAs downloaded from miRBase[3]. The probability of the first base was derived from the frequency of nucleotides at the first base of human mature miRNAs. The alignment of the sequences was performed using BWA[4] (version 0.7.9a-r786) using aln/samse and summaries of the mapping were created using Perl scripts. The bar plots were created in R using ggplot2[5] and reshape2[6].

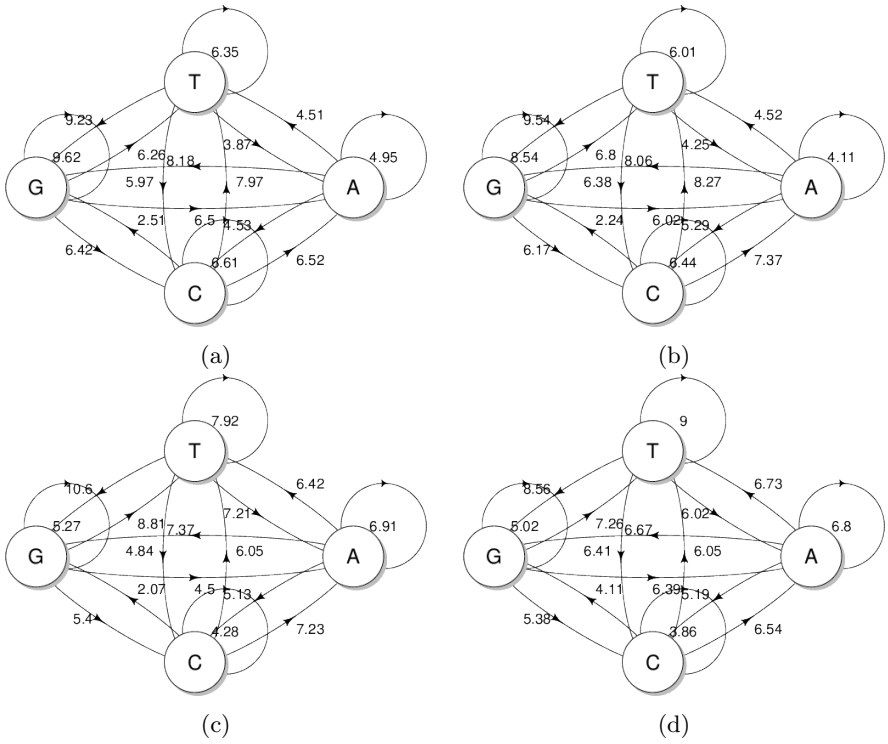


Figure 3: Transition diagrams for (a) human, (b) mouse, (c) zebrafish, and (d) nematode based on dinucleotide frequencies from miRBase mature miRNAs.

3 Results

The human genome is mainly composed of repetitive elements, many of which are transposable elements and thus its sequence composition is not a mosaic of random sequences. With this in mind, we would expect that randomly generated sequences would not map back to the human genome. However, it is not known at what length, random sequences will not map back. We investigated this by generating random sequences ranging from 15 base pairs to 30 base pairs under three different models (Section 2) and mapped these sequences to the human genome (Figure 4a). At length 18, almost all randomly generated sequences could be mapped to the genome. The number of possible DNA sequences of n length is 4^n and thus a 18 base pair sequence has 68,719,476,736 possible sequences.

Hello, here is some text without a meaning. This text should show, how a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like ≫Huardest gefburn≪. Kjift – Never mind! A blind text like this gives you information about the selected font, how the letters are

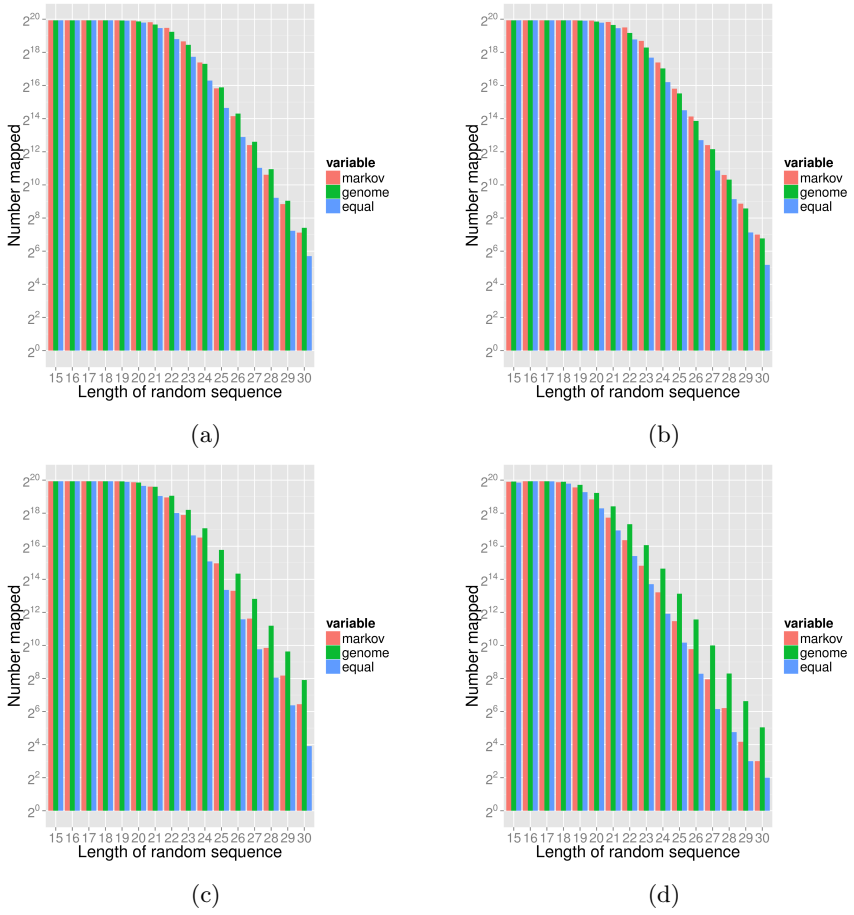


Figure 4

written and the impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for a special contents, but the length of words should match to the language.

4 Discussion

Mammalian miRNAs are able to recognise their target miRNA by as little as 6-8 nucleotides; this region is known as the seed region, which lies at the 5' end of a miRNA. However despite this, most miRNAs are 22-23 base pairs in length.

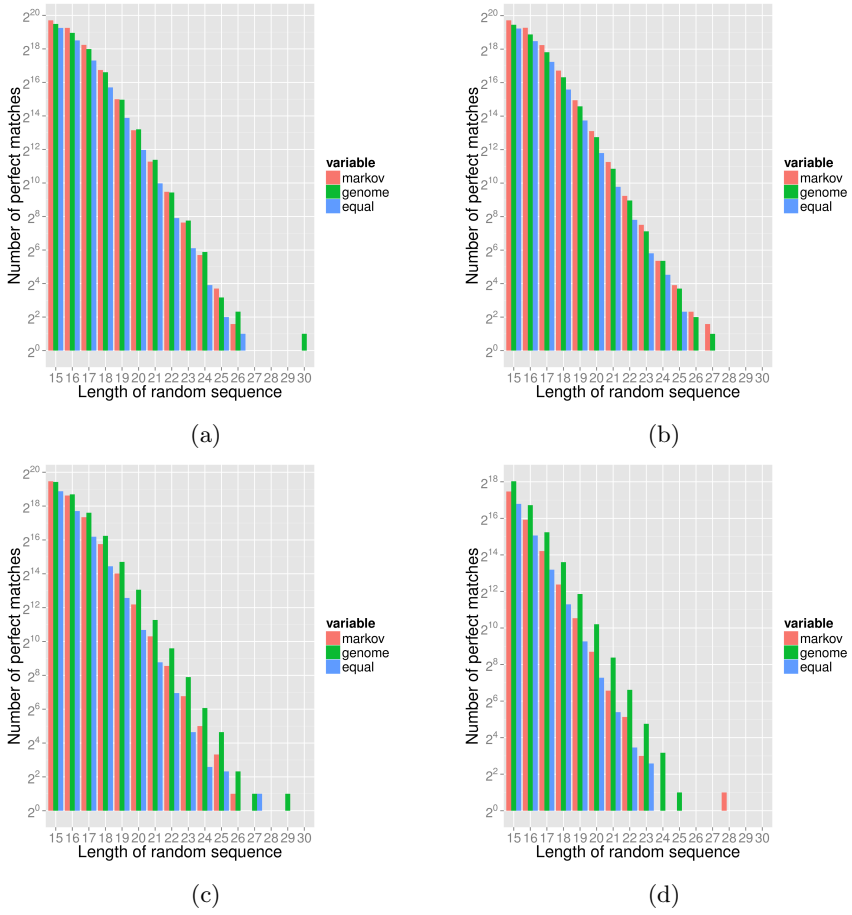
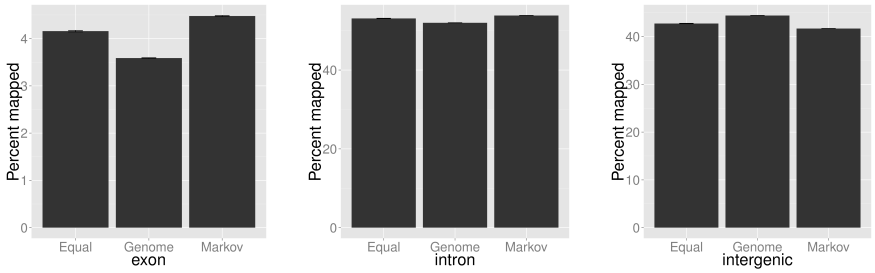


Figure 5

5 Authors' contributions

DD wrote the section on miRNAs, suggested analyses, and discussed the results. DT did everything else. All authors read and approved the final manuscript.



(a) (b) (c)

Figure 6

References

[1] R. C. Lee, R. L. Feinbaum, and V. Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, Dec 1993.

[2] H. Pages, P. Aboyoun, R. Gentleman, and S. DebRoy. *Biostrings: String objects representing biological sequences, and matching algorithms*. R package version 2.30.1.

[3] A. Kozomara and S. Griffiths-Jones. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, 39(Database issue):D152–157, Jan 2011.

[4] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, Jul 2009.

[5] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

[6] Hadley Wickham. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12):1–20, 2007.