

On the length of mature microRNAs

Dave Tang Derek de Rie
RIKEN Yokohama VU University Amsterdam

June 10, 2014

Abstract

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).

2 Methods

All code underlying this work is available at https://github.com/davetang/mirna_length. Briefly, random sequences were generated using R (version 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 1) 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].

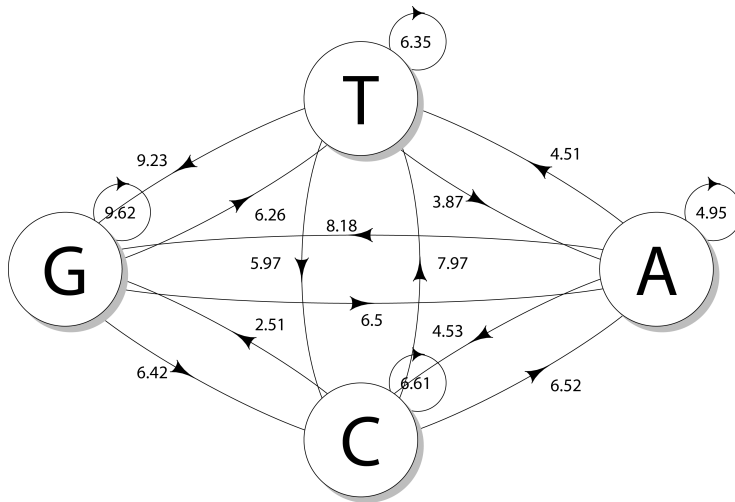


Figure 1: Transition diagram based on dinucleotide frequencies of miRBase human miRNAs.

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 2a). 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.

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.

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.

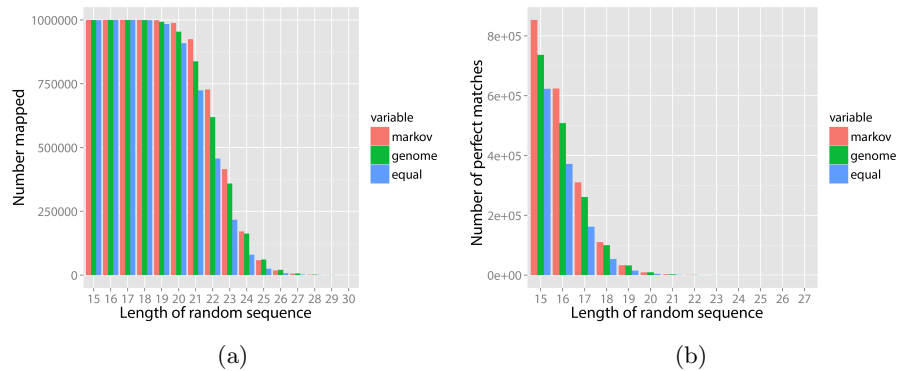


Figure 2: Mapping statistics of three sets of randomly generated sequences. (a) Number of randomly generated sequences mapped back to the human genome. (b) Number of randomly generated sequences perfectly mapped, i.e. mapped without any edits, back to the human genome.

- [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.