

Gene Length and Alternative Transcription in Fruit Fly

Boris Budagyan
Bioinformatics Certificate Program
Cal State Hayward
budagyan@hotmail.com

Ann Loraine
University of Alabama, Birmingham
aloraine@ms.soph.uab.edu
loraine@loraine.net

Abstract

*Alternative transcription, in which a single gene may give rise to multiple variant mRNA forms, is widely recognized as an important source of protein diversity in complex, eukaryotic genomes. Here we show that in the *Drosophila* genome, larger genes with greater numbers of exons tend to be alternatively transcribed to a greater degree than smaller genes with fewer exons. In addition, we show that a log-normal distribution provides a good approximation for gene length distributions in *Drosophila* and that an exponential function relates the number of variants produced per gene with average exon count.*

1. Introduction

Alternative transcription, in which a single gene may give rise to multiple distinct mRNA species, is an important source of protein diversity in complex, eukaryotic genomes [1]. Little is known at a whole-genome level about the general properties of alternatively transcribed, multi-variant genes, however. Knowledge of any distinguishing properties of multi-variant genes could be useful not just for predicting when individual genes may produce as-yet undiscovered variants, but also for understanding how alternatively transcribed, multi-variant genes have evolved. To identify such properties, relationships between gene lengths, number of transcripts produced per gene, and the number of exons per gene were investigated using the *Drosophila melanogaster* genome as a model system [2].

2. Methods

A collection of manually-curated *Drosophila* genes was used in this study. The collection was downloaded as a single file in “gff” format from www.flybase.org in January, 2004. According to the source Web site, this file contained all FlyBase-annotated genes from the euchromatic region of the

Drosophila genome. A copy of this original data file, along with other resources, is available from www.loraine.net/csb2004.

Gene sizes (in genomic base pairs), gene assignments for individual mRNAs, and the number of exons per mRNA were extracted programmatically for all protein-encoding genes in the original data file. This data processing was done in Python, and statistical calculations were done using version 1.9 of the R programming language for statistics.

3. Results

Figure 1 (a) shows that gene lengths for single-variant genes approximate a log-normal distribution. This general shape was typical for all the genes, not just those which produce only one mRNA form. Because the histogram obtained from the log-transformed gene lengths resembles the familiar bell-shaped curve of the normal distribution (shown in outline), the R implementation of the Shapiro test for normality was used to test the normality of the log-transformed lengths. High values of test statistic *W* (0.97, 0.98, 0.98, and 0.99 for genes producing 1, 2, 3, and 4 transcript variants, respectively) provide good evidence for normality. In addition, quantile-quantile plots (Figure 1(b)) revealed that the bulk of the data do indeed conform to a normal curve, with deviations restricted to the distribution’s high and low ends.

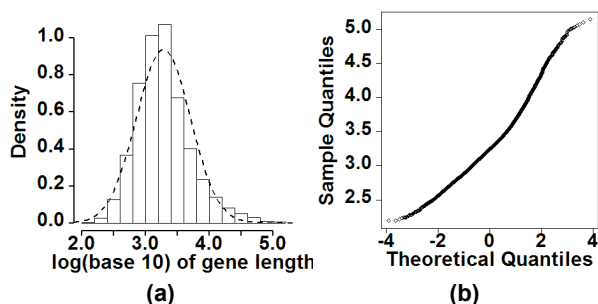


Figure 1. Length distribution (a) and Q-Q normality plot (b) for genes producing 1 transcript

Table 1. The results of calculation for genes producing different numbers of transcript variants

Number of transcripts	Genes count	Genes length mean, bp	Genes length median, bp	Mean exon count
1	10639	3587	1727	4
2	1707	8170	3877	6
3	549	11603	5727	7
4	247	16909	8580	8
5	100	19630	9796	9
6	53	17518	10899	9
7	34	16640	9026	9
8	17	30779	13852	10

The fruit fly gene set was divided into classes based on the number of transcripts produced. For each class, we computed the average and median gene length as well as the average number of exons per gene, counted as the number of exons included in the most exon-rich mRNA variant a gene produces. Table 1 summarizes our results and reveals that, as a group, multi-variant genes tend to be larger and also tend to contain more exons.

To formalize these observations, we applied a one-sided Wilcoxon rank sum test to adjacent gene classes. For each test, the null hypothesis (H_0) was that gene size or exon count for the

Table 2. Wilcoxon sum rank test results

Samples	p-values for gene length	p-values for exon count
1&2	<2.2e-16	<2.2e-16
2&3	1.25e-13	4.56e-14
3&4	7.62e-07	3.17e-3
4&5	0.107	0.019
5&6	0.249	0.476
6&7	0.789	0.415
7&8	0.069	0.428

higher-numbered class (more mRNA forms) was the same as or less than the adjacent, lower-numbered class. The alternative (H_A) or research hypothesis was that gene size for the higher-numbered class was greater. Table 2 reports p-values for each test, where the p-value represents the probability of the observed data assuming H_0 is true. At significance level 0.05, we reject H_0 for comparisons between classes 1 through 4 and conclude that larger genes do indeed, on average, produce more variants.

To characterize these relationships, transcript number per gene was plotted against median gene size (Figure 2) and average exon count (Figure 3) and a regression performed to identify functions that might best describe the relationships between variables. We find that for gene classes one through four, which account for over 98% of the total number of annotated genes in our data set, a good, relatively linear relationship exists between median gene size and the

number of variants. In addition, we find that an exponential function ($0.25e^{0.35x}$) provides a good approximation for the relationship between transcript number and average exon count for gene classes producing 1 to 8 transcript variants.

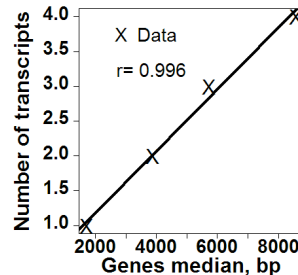


Figure 2. The relation between transcript number and median gene length

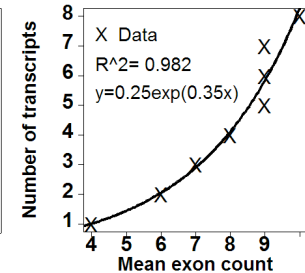


Figure 3. The relation between transcript number and mean exon count

4. Discussion and Conclusions

We show here that genes that exhibit a higher degree of alternative transcription tend to be larger and contain more exons than genes that produce fewer variants. We believe this finding has value for two main reasons.

First, this finding adds to an understanding of how alternatively transcribed genes have evolved. It has been proposed that genes acquire new variants through within-gene exon duplication [3, 4]. Our observation of the relationship between gene size and alternative transcription supports this idea. Genes occupying larger genomic regions would tend to be better, more likely targets for acquiring new 'splice-able' sequence through duplication events. In addition, genes that have already gained new variants (and new exons) would tend to be larger from having assimilated the duplicated sequence. Second, the two variables we examined - gene length and exon count - may be useful for predicting whether an individual gene produces additional, unknown variants. We plan to investigate this possibility in future work.

5. References

- [1] Loraine, et al. Journal of Bioinformatics and Computational Biology, Vol. 1, No. 2 (2003).
- [2] Misra, et al. Genome Biology, Vol. 3, No. 12 (2002).
- [3] Letunic, Human Molecular Genetics, Vol. 11, No. 13, (2002).
- [4] Kondrashov, et al. Human Molecular Genetics, Vol. 10, No. 23 (2002).