Gene Expression Level and Single Nucleotide Polymorphism Rates

Asmaa A. Abdelwahab, Manar Mohamed, Abeer Shalaby

* Bioinformatics Department, Nile University

Proteins in the same species evolve at different rates so the systems evolutionary genomics field studies the factors that determine the evolutionary rates of proteins [1]. In the last thirty years, due to the lack of the sequence data, the evolutionary theory suggested that protein evolutionary rates are controlled by the density of amino acid residues in a protein under the influence of their functional importance [2,3,4]. Recently, after the genome-scale data of sequences, some genomic factors demonstrate weak but statistically significant correlations with evolutionary rates. Among the genomic factors, in unicellular organisms, the expression level is the most prominent negative correlate with evolutionary protein rate[3,4]. For instance, in yeast, the variance between paralogs after duplication is negative co-related to expression levels[3,4]. In the expression-based evolutionary analysis, due to the Genes express at different levels in various tissue types in multicellular organisms, the estimation of expression level in multicellular organisms is more complicated than for unicellular organisms. For example, some genes have a high level of expression in specific tissue types while others are approximately expressed at low levels in all tissues, indicating that the broadly expressed genes are not essentially highly expressed level genes [5].

Therefore, expression breadth mainly used in multicellular organisms which defined as the number of different tissues where a gene is significantly expressed. Broader expressed genes have more interaction partners and may produce lethal mutants. Therefore, these genes tend to evolve slowly and are less likely to gene loss across different taxonomic groups. However, genes are expected to be weakly expressed have fewer interaction partners, and these genes evolve faster and are more often lost during evolution than broader expressed genes[3,4].

Single nucleotide polymorphisms (SNPs) are valuable tools for localizing and identifying disease susceptibility genes, understanding the molecular mechanisms of mutation, and deducing the origins of modern human populations [6]. Because of their mutational history and population structure, it is believed that a subset of SNPs will capture the relevant information in the full

complement of SNPs across the genome [7]. During the last decade (SNPs) have become increasingly used because of their abundance in the genome, ease of replication in different laboratories and simplicity of analysis [8]. The existence of introns in genome is a real mystery, given the expensive energy cost for a cell to pay for copying the entire length of several introns in a gene and excising them at the exact position, controlled by big RNA and protein complexes after transcription. Introns are clearly not junk, and they provide selective advantages to cells to be evolutionarily maintained, nevertheless, it has expensive energetic costs, most completely genomes of eukaryotic cells so far carry introns in their genomes [9]. Introns occupy about 40% on average of the total length of genes, which means that most randomly occurring mutations will fall into intron regions, and do not affect protein sequences and functions. However, it is not clear how extensively and strongly this buffering effect of intron regions might have evolutionary advantages for intron retention against the pressure of removing cellular burdens. [10]

Putative functional roles of introns in various cellular processes such as splicing, mRNA transport, NMD, and expression regulation. Besides, introns may give some advantages as a mutational buffer in eukaryotic genomes protecting coding sequences from being affected by randomly occurring deleterious mutations. [11]

In general, SNP frequency is directly related to the evolutionary pressure on the target genome regions. For example, more SNPs are accumulated in repetitive sequence, introns and pseudogenes, as the evolution pressure in these regions is relatively low compared to functional gene sequence regions. Given this hypothesis, it is difficult to explain why more SNPs are observed in the sequence region close to transcriptional start site, as the region is important to the initiation of gene transcription, and sequence alteration has potentials to influence gene expression.

One possible explanation is that higher SNP frequency is related to important functions of the regions close to transcription start sites. The accumulation of SNPs in the human genome is like a snapshot of human evolutionary history in which genes, especially those with specific functions, are under continuous natural selection pressure and alteration by mutation, genetic drift and gene flow. As a result, the expression pattern of a gene may be changed. While some genes become totally inactive, others experience expression level alteration. It is possible that SNPs occurring in gene promoter regions play an important role in such scenario, so that the higher frequency of SNPs close to transcriptional start site is related to subtle alteration of gene expression which results in population diversity. [12]

Materials and methods:

Expression data

Human gene expression data for different tissue and cell types (table1) was retrieved from FANTOM database (Functional Annotation of the Mouse/Mammalian Genome) http://fantom.gsc.riken.jp/5/datafiles/latest/extra/gene_level_expression/. The gene expression is represented by TPM (Transcripts per million) values where the columns and rows represent the TPM values for the different tissues and the gene annotations respectively.

Table1: representing the file of human gene expression data

	00Annotation	tpm.293SLAM%20rinderpest%20infection%2c%2000hr%2c%20biol_rep1.CNhs14406.13541-145H4	tpm.293SLAM%20rinderpest%20infection%2c%
0	C9orf152	0.000000	
1	ENST00000457273	0.000000	
2	ELMO2	3.871942	
3	RPS11	496.664564	
4	CREB3L1	0.527992	
5	PNMA1	78.846819	
6	MMP2	0.000000	
7	TMEM216	36.079460	
8	TRAF3IP2-AS1	4.575931	
9	C10orf90	0.000000	

SNPs Data

This data is considered as a VCF file that detects the locations of SNPs inside the human genome, it was downloaded from the 1000Genomes project (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.wgs.phase3_shapeit2_mvncall_integrated_v5b.20130502.sites.vcf.gz).

Features Data

The Gene transfer format (GTF) file for the human genome version 19 (GRCh37) (table3) was retrieved from the ensemble database ftp://ftp.ensembl.org/pub/release-75/gtf/homo_sapien/Homo_sapiens.GRCh37.75.gtf.gz which holds information about gene structure. It is a tab-delimited text format based on the general feature format (GFF), but contains some additional conventions specific to gene information.

Table3: representing the GTF file

	0	1	2	3	4	5	6	7	8
0	1	pseudogene	gene	11869	14412	.	+	.	gene_id "ENSG00000223972"; gene_name "DDX11L1"...
1	1	processed_transcript	transcript	11869	14409	.	+	.	gene_id "ENSG00000223972"; transcript_id "ENST...
2	1	processed_transcript	exon	11869	12227	.	+	.	gene_id "ENSG00000223972"; transcript_id "ENST...
3	1	processed_transcript	exon	12613	12721	.	+	.	gene_id "ENSG00000223972"; transcript_id "ENST...
4	1	processed_transcript	exon	13221	14409	.	+	.	gene_id "ENSG00000223972"; transcript_id "ENST...
5	1	transcribed_unprocessed_pseudogene	transcript	11872	14412	.	+	.	gene_id "ENSG00000223972"; transcript_id "ENST...
6	1	transcribed_unprocessed_pseudogene	exon	11872	12227	.	+	.	gene_id "ENSG00000223972"; transcript_id "ENST...
7	1	transcribed_unprocessed_pseudogene	exon	12613	12721	.	+	.	gene_id "ENSG00000223972"; transcript_id "ENST...
8	1	transcribed_unprocessed_pseudogene	exon	13225	14412	.	+	.	gene_id "ENSG00000223972"; transcript_id "ENST...
9	1	transcribed_unprocessed_pseudogene	transcript	11874	14409	.	+	.	gene_id "ENSG00000223972"; transcript_id "ENST...
10	1	transcribed_unprocessed_pseudogene	exon	11874	12227	.	+	.	gene_id "ENSG00000223972"; transcript_id "ENST...

Tools

Three python libraries were used; Pandas is a library for data analysis [13], Pybedtools is a python API for bedtools which are powerful tools used for genome arithmetic [14], and Mygene library for querying and retrieving gene annotations data [15].

Methods

According to many literatures many methods were used to assess the evolution rate which is related to the frequency of SNPs in genes such as expression abundance and expression breadth [16]. Three values were calculated for each gene annotation; maximum expression, median expression and expression breadth which is the count of tissues whose TPM values are greater than five (table4). Genes were ranked according to each calculated value and the top and low five percent were selected to represent the highly expressed genes and lowly expressed genes respectively.

Table 4: representing the maximum, median and breadth expression.

	Annotation	max_expr	median_expr	expr_breadth
0	C9orf152	90.000000	90.0	3
1	ENST00000457273	2.349140	0.0	0
2	ELMO2	1695.000000	1695.0	3
3	RPS11	3314.932418	1826.0	3
4	CREB3L1	1021.000000	1021.0	3
5	PNMA1	1780.000000	1780.0	3
6	MMP2	8337.690244	1224.0	3
7	TMEM216	1553.000000	1553.0	3
8	TRAF3IP2-AS1	483.000000	483.0	3
9	C10orf90	1212.386209	121.0	3
10	ENST00000435872	7.363069	3.0	1

To get the file representing the coordinates of the previously selected genes, conversion from gene annotations to Ensembl identifiers was needed. Two bed files were obtained for exons and introns separately. Then, the coordinates of these genes were selected from each file according to these Ensembl identifiers.

Bedtools were used to determine the count of SNPs in the highly and lowly expressed genes by intersecting each bed file containing the coordinates of selected genes with the vcf file containing the coordinates of SNPs in each gene of the human genome 19.

Finally, the frequency of SNPs in the highly and lowly expressed genes inside the human genome is calculated by dividing the count of SNPs over the length of nucleotides in genes.

Results

Gene expression is studied and found that the frequency of SNPs is significantly lower in genes that are particularly highly/broadly expressed than in genes that are particularly lowly/narrowly expressed, but the difference is simple.

Gene expression in exons

To obtain highly gene expression data for different tissue and cell types. Rank genes based on maximum expression (0.030), median expression (0.029) and expression breadth (0.028).
To obtain lowly gene expression data for different tissue and cell types. Rank genes based on maximum expression (0.032), median expression (0.031) and expression breadth (0.031).

Gene expression is of critical importance to many fundamental biological processes, including species divergence [17], protein evolution[18] ,and adaptation to microenvironment[19]. In multicellular organisms, complexity of gene expression is often summarized by two measures: first, how many transcripts are generated per locus (referred to as 'gene expression level') and second, how broadly each transcript is found in different tissues (referred to as 'gene expression breadth'). Together, levels and breadths of gene expression shape the diversity of organismal transcriptomes and eventually facilitate the development and the maintenance of complex biological systems.

Expression gene in intron

Introns can increase gene expression without functioning as a binding site for transcription factors. Introns can increase transcript levels by affecting the rate of transcription, nuclear export, and transcript stability. Moreover, introns can also increase the efficiency of mRNA translation. Absence of introns or gene length alone does *not* predict gene expression during fast cell cycles. First, not *all* short genes are expressed during early embryogenesis and, second, introns in some of the expressed short genes might feedback positively to facilitate rapid transcription. For rapid expression, the best genes are short with a few introns and a short first exon [20,21]. To obtain highly gene expression data for different tissue and cell types. Rank genes based on maximum expression (0.29), median expression (0.33) and expression breadth (0.26). To obtain lowly gene expression data for different tissue and cell types. Rank genes based on maximum expression (0.63), median expression (0.51) and expression breadth (0.60).

Discussion

Two hypotheses were tested in this project ,the main ;correlate the presence of single nucleotide polymorphism with exons which have the higher level of gene expression , that was slightly supported with our result .the second ; correlate the presence of SNPs within the introns which have less gene expression level and that was strongly supported with our result ,and both was determined at three different expression regions.at 2012 a previous study for [Dingox] used exons with single-nucleotide polymorphisms (SNPs) which often used as genetic makers[22]. But at 2018 study for [long G oa] used gene regulatory network (introns) to identify the noncoding risk variant [23] both of them used in Genome-wide association (GWA) studies which were currently one of the most powerful tools in identifying disease-associated genes or variants. According to our results we conclude that for exons: SNPs frequency inversely proportional gene expression mainly according to breadth expression, while for introns: SNPs frequency inversely proportional gene expression mainly according to max expression.

References:

1. Margoliash, E. 1963. “PRIMARY STRUCTURE AND EVOLUTION OF CYTOCHROME C.” *Proceedings of the National Academy of Sciences* 50 (4): 672–79. https://doi.org/10.1073/pnas.50.4.672.
2. Zuckerkandl, Emile. 1976. “Evolutionary Processes and Evolutionary Noise at the Molecular Level.” *Journal of Molecular Evolution* 7 (4): 269–311. https://doi.org/10.1007/BF01743626.
3. Koonin, Eugene V, and Yuri I Wolf. 2006. “Evolutionary Systems Biology: Links between Gene Evolution and Function.” *Current Opinion in Biotechnology* 17 (5): 481–87. https://doi.org/10.1016/j.copbio.2006.08.003.
4. Wolf, Yuri I., Liran Carmel, and Eugene V Koonin. 2006. “Correlations between Quantitative Measures of Genome Evolution, Expression and Function,” no. January: 0–12. https://doi.org/10.1007/0-387-36747-0.
5. S.G., Park, and Choi S.S. 2010. “Expression Breadth and Expression Abundance Behave Differently in Correlations with Evolutionary Rates.” *BMC Evolutionary Biology* 10: 241. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed9&NEWS=N&AN=20691101.

6. Zhao, Z., Boerwinkle, E., 2002. Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. Genome Res. 12, 1679–1686. Assessment of population structure by single nucleotide polymorphisms (SNPs) in goat breeds.
7. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J., Lander, E.S., 2001. High-resolution haplotype structure in the human genome. Nat. G
8. Pariset L, Cappuccio I, Ajmone Marsan P, Dunner S, Luikart G, England PR, Obexer-Ruff G, Peter C, Marletta D, Pilla F, Valentini A, ECONOGENE Consortium.J Chromatogr B Analyt Technol Biomed Life Sci. 2006 Mar 20; 833(1):117-20.
9. Deutsch M, Long M. Intron-exon structures of eukaryotic model organisms. Nucleic Acids Res. 1999;27:3219–3228. [PMC free article] [PubMed]
10. Jo BS, Choi SS. Introns: The Functional Benefits of Introns in Genomes. *Genomics Inform*. 2015;13(4):112-8.
11. Simpson AG, MacQuarrie EK, Roger AJ. Eukaryotic evolution: early origin of canonical introns. Nature. 2002;419:270. [PubMed]
12. Guo Y, Jamison DC. The distribution of SNPs in human gene regulatory regions. *BMC Genomics*. 2005;6:140. Published 2005 Oct 6. doi:10.1186/1471-2164-6-140
13. Mckinney, Wes. (2011). pandas: a Foundational Python Library for Data Analysis and Statistics. Python High Performance Science Computer.
14. Dale RK, Pedersen BS, Quinlan AR. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics. 2011;27(24):3423-4.
15. Xin J, Mark A, Afrasiabi C, Tsueng G, Juchler M, Gopal N, Stupp GS, Putman TE, Ainscough BJ, Griffith OL, Torkamani A, Whetzel PL, Mungall CJ, Mooney SD, Su AI, Wu C (2016) High-performance web services for querying gene and variant annotation. Genome Biology 17(1):1-7 [link].
16. Park, S., & Choi, S. (2010). Expression breadth and expression abundance behave differently in correlations with evolutionary rates. BMC Evolutionary Biology, 10(1), 241. doi:10.1186/1471-2148-10-241
17. King M.-C., Wilson A.C. Evolution at two levels in humans and chimpanzees. Science. 1975;188:107–116.
18. Drummond D.A., Raval A., Wilke C.O. A single determinant dominates the rate of yeast protein evolution. Mol. Biol. Evol. 2006;23:327–337

19. Lopez-Maury L., Marguerat S., Bahler J. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. Nat. Rev. Genet. 2008;9:583–593
20. Bieberstein NI, Carrillo Oesterreich F, Straube K, Neugebauer KM. First exon length controls active chromatin signatures and transcription. Cell Rep. 2012;2:62–8
21. Heyn P, Kircher M, Dahl A, Kelso J, et al. The earliest transcribed zygotic genes are short, newly evolved, and different across species. Cell Rep. 2014;6:285–92.
22. Xiao Dong, Tingyan Zhong, Tao Xu, Yunting Xia, Biqing Li, Chao Li, Liyun Yuan, Guohui Ding, Yixue Li,Evaluating coverage of exons by HapMap SNPs,Genomics,Volume 101, Issue 1,2013,Pages 20-23,ISSN 0888-7543,https://doi.org/10.1016/j.ygeno.2012.09.003.
23. *Gao, Long et al. "Identifying noncoding risk variants using disease-relevant gene regulatory networks." Nature Communications (2018).*