RNA Processing Dynamics of Functional Gene Sets

Shiva Velingker

Louisiana School for Math, Science, and the Arts

11/12/2014

Abstract

Currently, we understand how proteins are synthesized from DNA, but we do not fully understand the dynamics of what drives certain biological characteristics to be more fully expressed than other characteristics — that is, what drives the different products of alternative splicing. If we evaluate pre-mRNA before it becomes spliced by looking at its ratio of introns to exons, perhaps we could better understand what preferentially drives certain genes to be more expressed, thus improving our understanding of the eukaryotic RNA processing.

Genomic data was gathered from the human genome. Each gene's exonic and intronic densities were calculated based on their expression in various tissues. Their pre-mRNA fractions were computed as the ratio of the intronic density to the exonic density. These fractions were run through various gene analysis tools to show the enrichments of various gene products. The results indicate that genes with low pre-mRNA are particularly more expressed in regards to integral cellular functions, to which further experimental research must be conducted to validate this finding.

Introduction

The central dogma of molecular biology states that DNA is transcribed to RNA which is translated to proteins. Proteins are essential for cells to operate and interact with each other and the preservation of traits in organisms over time. However, proteins fold in numerous ways that make them difficult to sequence as efficiently as nucleic acids; current technology cannot yet effectively analyze the data to understand the finer dynamics of protein synthesis. Rather, it may be more efficient to look farther upstream — DNA and transcription factors — to determine shared characteristics between different tissues and organisms. Both DNA and immature RNA transcripts (pre-mRNA) contain exons and introns. Exons are the coding regions of genetic material that become covalently joined after the introns, the non-coding regions, are spliced out by spliceosomes. Pre-mRNA was originally thought to be strictly an intermediary on the way to processed mRNA (Warf 2010). It is now known that pre-mRNA actively interacts with various molecules to regulate itself and produce different mature messages (alternative splicing). This alternative splicing produces multiple proteins from a single gene and accounts for the biodiversity found in eukaryotes (Snustad 2012). This demonstrates how genome size does not necessarily correlate to genome complexity. Although various factors are involved with protein synthesis, what if other characteristics also affect the processing of pre-mRNA?

Current analyses of genes rely primarily on experimental research in which scientists isolate genes and monitor their interactions with other genes. These tests focus on a characteristic specific to sample size of the genes themselves but not the genome as a whole. In the broader perspective — that is, protein synthesis — the dynamics of *how* RNA is processed are understood, but not *why* RNA can produce different proteins. For example, are genes with high pre-mRNA fractions (ratios of intronic density to exonic density) processed differently or exhibit different fundamental characteristics when compared to genes with low pre-mRNA fractions? Data exists on all genes found in chromosomes in the human genome and can address the question of how pre-mRNA fraction levels affect gene processing (Lauder 2001). The only analysis of this pre-mRNA fraction that the investigators of this research are aware of addresses this issue with different methods and datasets (Braunschweig 2014).

The objective of this research is to utilize the resources that the Human Genome Project has to offer and determine the shared biological functions of genes with high or low pre-mRNA

fractions (Lauder 2001). This requires calculating pre-mRNA fractions for all genes within the human genome and running a filtered subset of this data through gene enrichment tools to determine if genes with high or low pre-mRNA fractions are correlated with the final gene products produced from transcription.

Methods

A script written in the Python programming language was designed to calculate the pre-mRNA fractions for each gene after getting data from genomic files. Gene Transfer Format files were obtained from Ensembl's GRCh38 genome build and contain information about all genes within a particular genome (Ensembl 2014). Additionally, the file contains information on the sources of the genes and their biotypes. For this research, the script extracted all the exons and their locations within the human genome.

The script proceeded to condense overlaps between the exons using the union exon models. Overlapping exons from the same gene were consolidated into one larger exon; while overlapping exons from different genes were discarded. An example of the consolidation is shown in Figure 1.

Figure 1: Consolidation of Exons Using the Union Exon Model

Gene Name	Exon Start	Exon End
DDX11L1	12010	12057
DDX11L1	12613	12721
DDX11L1	12613	12697
DDX11L1	13221	14409
WASH7P	14404	14501
WASH7P	15005	15038
WASH7P	15796	15947
WASH7P	16607	16765

Overlapping; same gene

Overlapping; different genes

→

Gene Name	Exon Start	Exon End
DDX11L1	12010	12057
DDX11L1	12613	12721
WASH7P	15005	15038
WASH7P	15796	15947
WASH7P	16607	16765

Introns were then inserted in-between the exons of the genes.

Bedgraph files are genomic files that contain information about the read-depth per splice of DNA per tissue. This research focused on eleven tissues obtained from the NCBI's Gene Expression Omnibus (Gene Expression Omnibus 2014). Table 1 shows the types of these tissues as well as the gender of the human they were obtained from.

Table 1: Origins of Tissues Samples											
Tissue	FC 1	FC 2	FC 3	FC 4	FC 5	OV 1	OV 2	PL 1	PL 2	PL 3	TS 1
Sex	M	M	M	M	M	F	F	M	M	F	M

Table 1. FC = Frontal Cortex; OV = Ovary; PL = Placenta; TS = Testis; M = Male; F = Female

After adjusting for the difference in coordinate systems between the two genomic files, the script aligned the exons with the number of coverages using Formula 1 to determine the density for each exon.

$$\frac{\sum \text{overlapping segment length} \bullet \text{coverage}}{\text{exon length}} \tag{1}$$

The same formula was used for calculating the densities of the introns. A sample output is demonstrated in Figure 2.

Figure 2: Sample Calculation Per Exon and Intron

Gene Name	Type	Density	Length
DDX11L1	Exon	17.553	47
DDX11L1	Intron	14.558	554
DDX11L1	Exon	16.076	108

Each gene's weighted exonic and intronic densities were calculated with Formula 2.

$$\frac{\sum_{n}^{exons} \text{density}_n \bullet \text{length}_n}{\sum_{n}^{exons} \text{length}_n} \tag{2}$$

Any gene with an exonic density of zero was discarded since it was unexpressed within that particular tissue. The final pre-mRNA fractions were determined using Formula 3.

$$\frac{\text{intronic density}}{\text{exonic density}} \tag{3}$$

Genes with pre-mRNA fractions greater than one were determined non-biological and were discarded. The final output file of the script included each gene's ID, name, exonic and intronic densities, and pre-mRNA fraction.

The script was originally intended to run on computer clusters since the genomic files were several gigabytes large and the script had a runtime of several hours. Equations were determined that could pre-compute the approximate runtime and memory usage of the script based on the number of exons per genome. As the functions within the script became more efficient and the intermediate data outputted was reduced, the average runtime decreased to seven minutes and memory usage to thirty megabytes.

The next stage of research required the raw data to be filtered so that it could be run through GeneTrail — a gene set analysis tool (Keller 2008). Instead of looking at gene products individually, gene set analysis tools allow large data sets to be analyzed through a process called *enrichment,* in which groups of genes are analyzed for significant over-representation of some biological characteristic (Keller 2008). Through this, the statistical significance of a group of genes can be predetermined before being experimentally validated.

Many of the genes returned from the script could have been generated from "noise," indicating that they have unreliable pre-mRNA fraction measurements or may not even be truly expressed. Further filtering was needed to reduce this noise. As previously mentioned, genes with zero exonic densities or pre-mRNA fractions greater than one were filtered out. Each tissue was additionally filtered using transcript abundance measurements from the program Cufflinks. Using Formula 4, the Fragments Per Kilobase Per Million (FPKM) could be computed for each gene within a tissue to give an estimate for the expected number of fragments per kilobase of transcript per million reads.

$$\mathrm{FPKM}_i = \frac{\text{number of mapped fragments}_i \bullet 10^9}{\sum_{j=1}^{transcripts} \text{transcript length}_j \bullet \text{number of mapped fragments}_j} \tag{4}$$

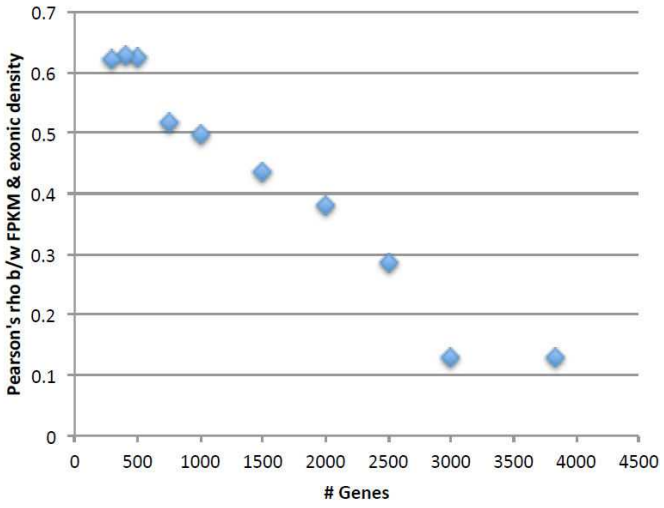
Any gene with an FPKM less than one was deemed to be unreliably expressed — perhaps as an artifact of gene model or exon filtering — and was filtered out. Table 2 shows the number of genes removed in each process of filtering through a representative tissue - Ovary Sample 1.

Table 2: Gene Filtering for Ovary Sample 1			
Argument	zero exonic density	pre-mRNA fraction > 1	FPKM < 1
Number of Genes Filtered	13,543	8,122	4,342

Table 2. Original number of genes was 62,575

These filtered genes were sorted by greatest exonic density and were tested with their pre-mRNA fractions using Pearson's chi-squared test. Figure 3 shows the calculated Rho by Pearson's test using the top X genes by exonic density genes as computed from Frontal Cortex 1.

Figure 3 - Number of Genes versus Pearson's Rho



Two thousand was determined to be the appropriate number of genes — as a tradeoff between exonic density reflecting FPKM highly while including a maximal number of genes — to test through the gene enrichment analysis tool. Pearson's Rho for each tissue with 2,000 genes is shown in Table 3.

Table 3: Pearson's Rho Calculated Per 2,000 Genes											
Tissue	FC 1	FC 2	FC 3	FC 4	FC 5	OV 1	OV 2	PL 1	PL 2	PL 3	TS 1
Rho	0.38	0.25	0.24	0.48	0.26	0.23	0.61	0.30	0.44	0.58	0.14

These top 2,000 genes were extracted, sorted by pre-mRNA fraction high to low, and prepared in the file format necessary for GeneTrail. The two gene sets that GeneTrail returned, gave genome-wide significant gene products, each tested across all tissues at a False Discovery Rate (FDR) correction of 5%. Further tests were run to test the significance of the gene sets in

the tissues at a broader test – a nominal level of 5%. The gene sets returned are Pfam and AmiGO. The Pfam database contains information about protein domains, the subunits of the protein that can survive on their own, and their families. AmiGO is a gene ontology database that helps compare large data sets with gene product annotation data.

The gene products returned were then investigated through current published findings to determine the significance of the gene products in relation to their pre-mRNA fractions.

Personal Contributions of Principal Investigator

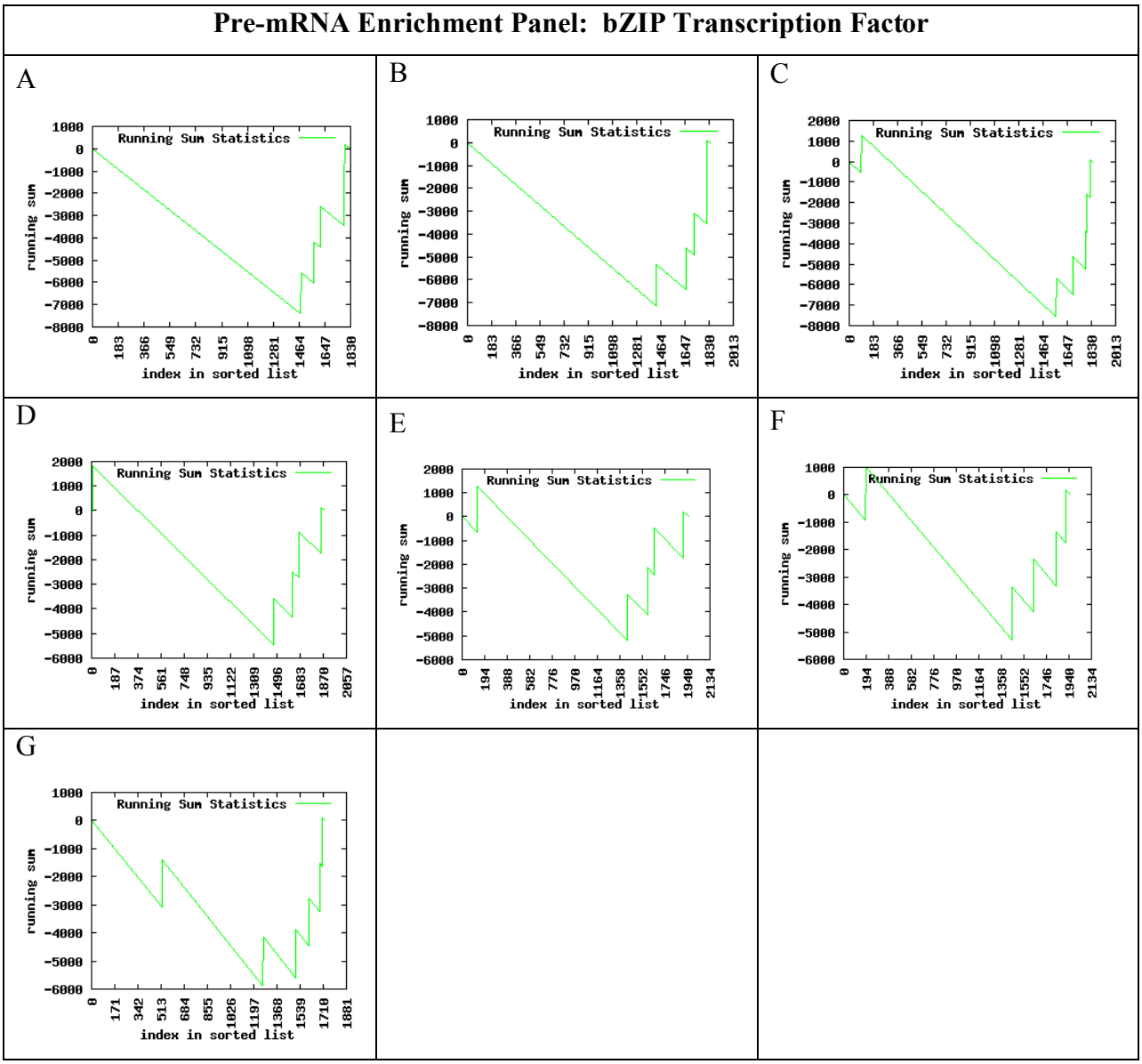
Multiple contributors were integral to the development of this research. My research mentor provided the question at hand and guided me to the appropriate resources to access. My genetics professor helped me understand the background of my research. I wrote the Python script to compute the pre-mRNA fractions, and my supercomputing teacher provided feedback for making the program more efficient for future use. I ran all the statistical tests necessary to filter the data, and I compiled all the results from GeneTrail. My research mentor then collaborated with me to interpret the results.

Results

Each tissue showed statistical significance with several gene products, but three gene products in particular showed strong replication across all tissues. Each major gene product and the tissues that they were expressed in is shown below. Any enrichment with an FDR significance of 5% is noted. Below each gene product is a listing of the genes which drove the enrichment score.

Database: Pfam							
Gene Product: bZIP Transcription Factor							
Tissue	FC 1*	FC 2*	OV 1	PL 1	PL 2	PL 3	TS 1
Number of Positives	5	5	6	5	5	5	6
P-Value	2.6e-4	5.3e-4	1.2e-3	2.0e-2	3.7e-2	3.1e-2	1.1e-2
pre-mRNA enrichment panel	A	B	C	D	E	F	G

*Tissues with FDR significance of 5%



Panel 3. Plots show enrichment score for corresponding tissues for pre-mRNA fractions

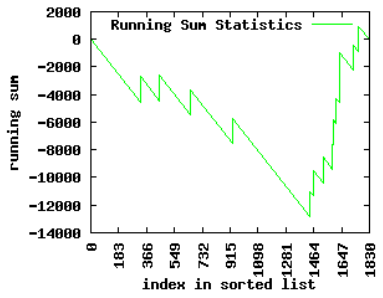
Genes Driving Enrichment of bZIP Transcription Factor	
Gene Name	**Gene Function**
ATF3	Cellular stress response
CREB3	Cellular proliferation
FOS	Cell proliferation, differentiation, transformation, apoptotic cell death
JDP2	UV-induced apoptosis, cell differentiation, tumorigenesis, antitumogeneris

Data obtained from National Center for Biotechnology Information (Gene Expression Omnibus 2014)

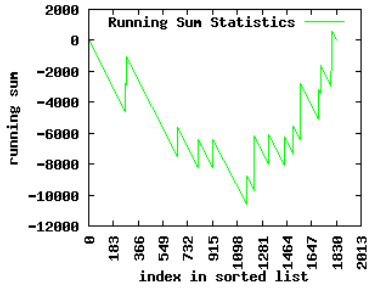
Database: Pfam										
Gene Product: Fibronectin Type III Domain										
Tissue	FC 1*	FC 2	FC 3*	FC 4	FC 5	OV 2	PL 1	PL 2	PL 3	TS 1
Number of Positives	14	17	13	18	19	20	19	16	16	14
P-Value	4.2e-4	1.4e-2	4.1e-4	9.0e-3	1.5e-2	3.0e-2	2.2e-2	1.5e-3	1.0e-3	5.3e-3
pre-mRNA enrichment panel	H	I	J	K	L	M	N	O	P	Q

Pre-mRNA Enrichment Panel: Fibronectin Type III Domain

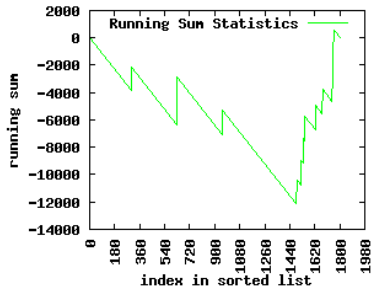
H



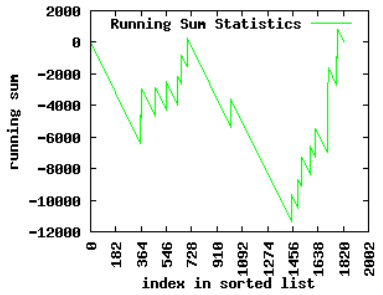
I



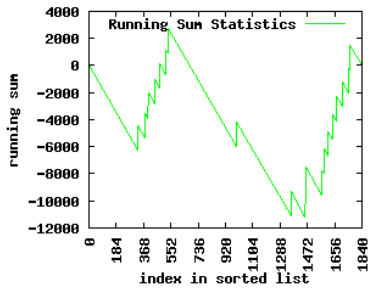
J



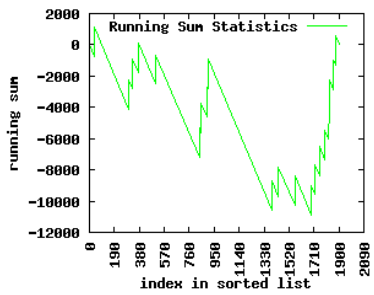
K

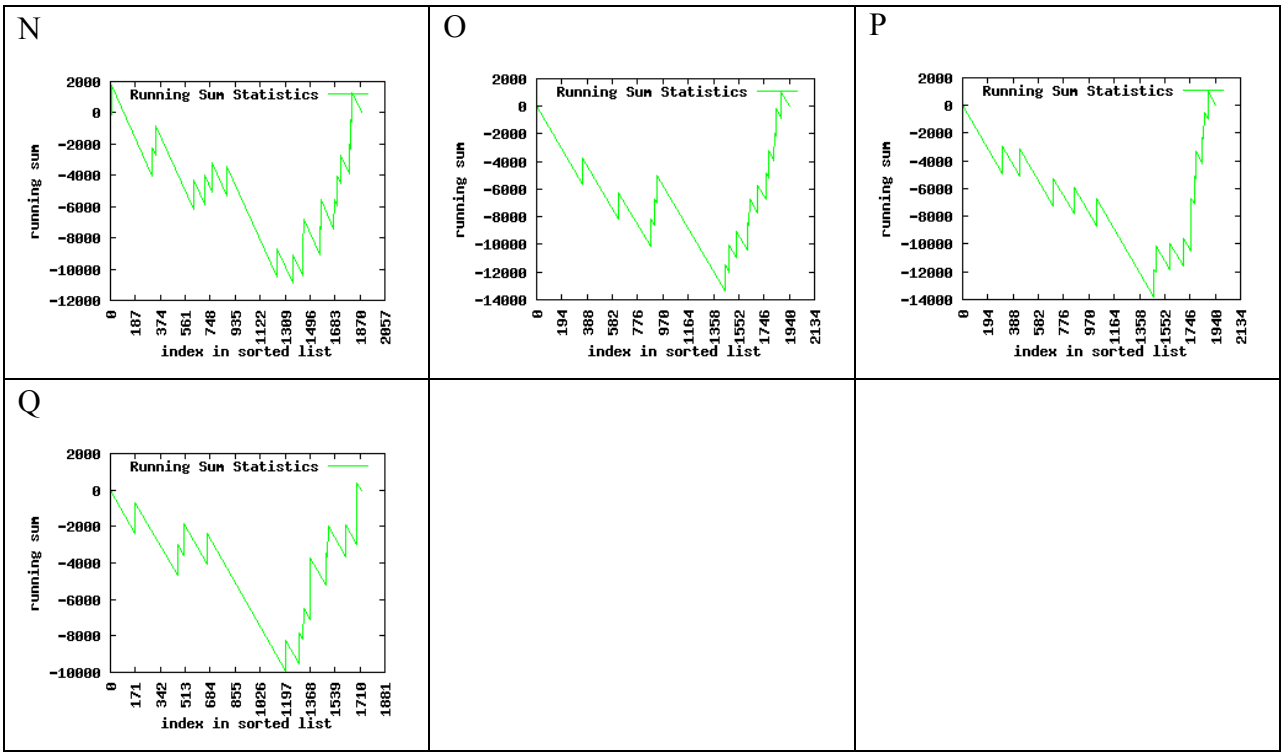


L



M





Panel 3. Plots show enrichment score for corresponding tissues for pre-mRNA fractions

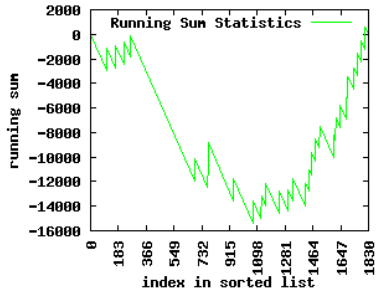
Genes Driving Enrichment of Fibronectin Type III Domain	
Gene Name	**Gene Function**
INSR	Insulin receptor with key role in regulation of glucose homeostasis
LIFR	Receptor complex for cell differentiation and proliferation
PTPRD	Cell growth, differentiation, mitotic cycle, oncogenic transformation
PTPRM	Cell growth, differentiation, mitotic cycle, oncogenic transformation

Data obtained from National Center for Biotechnology Information (Gene Expression Omnibus 2014)

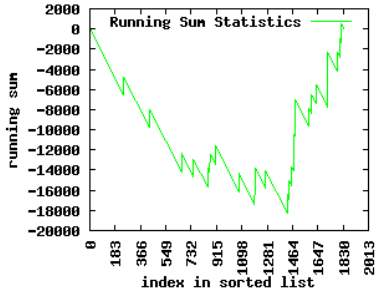
Database: AmiGO									
Gene Product: Negative Regulation of Transcription from RNA Polymerase II Promoter									
Tissue	FC 1	FC 2	FC 3	FC 4	FC 5	PL 1	PL 2*	PL 3*	TS 1
Number of Positives	28	27	31	29	32	36	26	26	29
P-Value	4.6e-3	3.8e-4	1.2e-1	2.7e-3	7.3e-4	1.2e-2	9.4e-6	9.2e-5	9.5e-3
pre-mRNA enrichment panel	R	S	T	U	V	W	X	Y	Z

Pre-mRNA Enrichment Panel: Negative Regulation from RNA Polymerase II Promoter

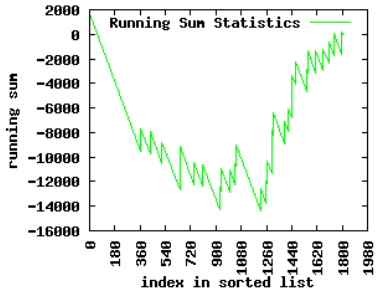
R



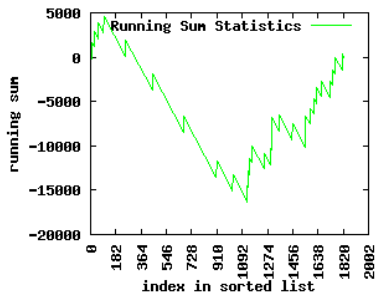
S



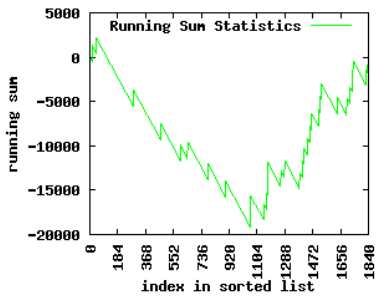
T



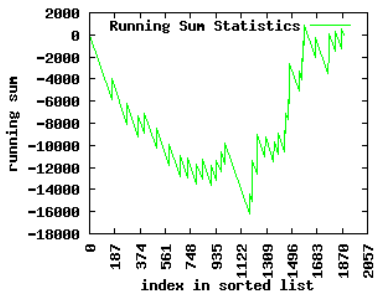
U

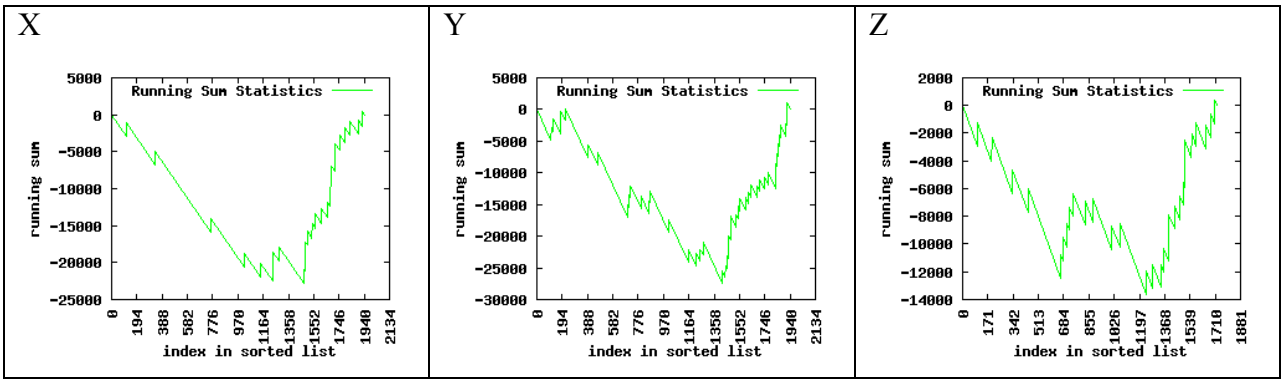


V



W





Panel 3. Plots show enrichment score for corresponding tissues for pre-mRNA fractions

Genes Driving Enrichment of Negative Regulation from RNA Polymerase II Promoter	
Gene Name	**Gene Function**
AES	Neurogenesis
CTNNBIP1	Prevents interaction between CTNNB1 and T-Cell Transcription Factor (TCF)
HIVEP1	Transcriptional regulation of viral and cellular genes
JARID2	Transcriptional repressor
JAZF1	Transcriptional repressor
JDP2	UV-induced apoptosis, cell differentiation, tumorigenesis, antitumogeneris
MBD3	Mediates association of metastasis-associated protein with core histone
NFIC	Cellular transcription factors and replication factors for adenovirus DNA replication
NFX1	Regulation of duration of inflammatory response
NIPBL	Facilitates enhancer-promoter communication and developmental regulation
PIAS4	protein inhibitor of activated STAT
SMARCA2	Alter chromatin structure for transcription regulation
SMARCA4	Alter chromatin structure for transcription regulation
ZBTB7A	Key role in repressing T-cell instructive Notch signal
ZFP161	Transcriptional repressor of MYC and thymidine kinase promoters

Data obtained from National Center for Biotechnology Information (Gene Expression Omnibus 2014)

Discussion

The objective of this research was to determine the shared biological functions of genes with high or low pre-mRNA fractions. The gene products produced from GeneTrail were globally significant in at least one tissue and broadly replicated at a nominal p-value of 0.05. This indicates that these results are likely to be real. Each enrichment plot generally showed a largely negative running sum, which indicates that the pre-mRNA fractions of the genes that were enriched were low.

The bZIP domain contains transcription factors that are integral to various functions within the human body, such as cancer development and hormone synthesis (Vlahopoulos 2008; Manna 2002). The bZIP domain is also highly involved in cell cycle regulation — bringing viruses out of latency and killing human epithelial cells (Cayrol 1996, Whitfield 1995). As shown in the "Genes Driving Enrichment" for bZIP, these transcription factors that drove the gene enrichment have low pre-mRNA fractions. According to Formula 3, this means that these factors have very dense exons. This suggests that the factors have genes that are very exon-dense and are very well spliced in relation to the other genes.

A similar conclusion can be made for the Fibronectin Type III domain. Its genes controlled other regulation receptors that are required for basic cellular functions — such as the cytoskeletal structure of the cell (Genetics Home Reference). It, too, showed a low pre-mRNA fractions in all of its enrichments.

The genes from the gene ontology results from AmiGO show several characteristics of inhibiting or repressing the RNA processing. With low pre-mRNA fractions, this suggests that these genes encode proteins that generally slow or inhibit transcription from the Polymerase II promotor, which is the major route of transcript for producing mRNAs.

These results indicate that genes with low pre-mRNA fractions are highly expressed in the cell with regards to the cell cycle. They are so well-spliced in the cell, probably due to the fact that they are integral to the cell's function. These results give a starting point to investigate, in a more general way, the potential causal mechanisms underlying especially low pre-mRNA fractions. Future scientists can use this, along with experimental data, as a basis to figure out what different groups of genes that have low pre-mRNA fractions, and in doing so, they can start to unravel the causal drivers of these correlations to further investigate fundamentals of RNA

processing.

Conclusions

Previously, little was known about how pre-mRNA fractions of genes affected RNA processing. The strong replication across most tissues suggests that this data is real and that pre-mRNA fractions indeed have an effect on RNA processing. However, further research must be conducted in order to validate these results and determine the finer details of how pre-mRNA fractions affect protein synthesis.

The next stage of this research would require experimental research to validate the pre-mRNA fractions. This can be accomplished using custom designed qRT-PCR primers which would enable measurements of exons and introns as they are generated through each cycle of PCR process . This would help to confirm the initial data gathered that was run through GeneTrail. To validate the gene product results, the domains from the gene products could be inserted into genes to monitor their effects. Looking at the bZIP domain specifically, the bZIP domain could added into a gene with an already high pre-mRNA fraction to see if it caused a lower pre-mRNA fraction or if its regulatory elements in the transcription factor genes caused the low pre-mRNA fraction. Since this research looked at a whole gene set, one could evaluate the domains with several genes to know which result generalizes instead of over-generalizing from one result.

Future research could also evaluate the pre-mRNA fraction levels across other genomes to determine evolutionary conservation across species in regards to RNA processing mechanisms. If these results could be replicated in other species, such as mice, the case would be strengthened that there exists functional reasons for these genes to have been spliced in the ways that they were. With more time, the investigators of this research would attempt to see if these results could be replicated across other genomes and evaluate the data's implications for RNA processing as a whole.

References

Braunschweig, Ulrich, Nuno L. Barbosa-Morais, and Qun Pan. "Widespread Intron Retention in Mammals Functionally Tunes Transcriptomes." *Cold Spring Harbor Laboratory Press* (2014): n. pag. *Genome Research*. Web.

Brawand, David, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, Angélica Liechti, Ayinuer Aximu-Petri, Martin Kircher, Frank W. Albert, Ulrich Zeller, Philipp Khaitovich, Frank Grützner, Sven Bergmann, Rasmus Nielsen, Svante Pääbo, and Henrik Kaessmann. "The Evolution of Gene Expression Levels in Mammalian Organs." *Nature* 478.7369 (2011): 343-48. Web.

C, Cayrol, and Flemington EK. "The Epstein-Barr Virus BZIP Transcription Factor Zta Causes G0/G1 Cell Cycle Arrest through Induction of Cyclin-dependent Kinase Inhibitors." *EMBO Journal* (1996): n. pag. *US National Library of Medicine*. Web. 15 Oct. 2014.

Ensembl. "Human Genome." *Ensembl Genome Browser 76: Homo Sapiens*. N.p., n.d. Web. June 2014.

Gene Expression Omnibus. "Functional Genomics Studies." *National Center for Biotechnology Information*. U.S. National Library of Medicine, n.d. Web. 30 Sept. 2014.

Genetics Home Reference. "Fibronectin Type III Domain Containing Gene Family." *Genetics Home Reference*. NCBI, n.d. Web.

Keller, A., Backes, C., Al-Awadhi, M., Gerasch, A., Kuentzer, J., Kohlbacher, O., Kaufmann, M., and Lenhof, H.P. GeneTrailExpress: a web-based pipeline for the statistical evaluation of microarray experiments. BMC Bioinformatics 2008, 9:552.

Lander, Eric S., Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William Fitzhugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie Levine, Paul Mcewan, Kevin Mckernan, James Meldrim, Jill P. Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee, Nigel Carter, Alan Coulson, Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew Jones, Christine Lloyd, Amanda Mcmurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C. Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Shownkeen, Sarah Sims, Robert H. Waterston, Richard K. Wilson, Ladeana W. Hillier, John D. Mcpherson, Marco A. Marra, Elaine R. Mardis, Lucinda A. Fulton, Asif T. Chinwalla, Kymberlie H. Pepin, Warren R. Gish, Stephanie L. Chissoe, Michael C. Wendl, Kim D. Delehaunty, Tracie L. Miner, Andrew Delehaunty, Jason B. Kramer, Lisa L. Cook, Robert S. Fulton, Douglas L. Johnson, Patrick J. Minx, Sandra W. Clifton, Trevor Hawkins, Elbert Branscomb, Paul Predki, Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan-Fang Cheng, Anne Olsen, Susan Lucas, Christopher Elkin, Edward Uberbacher, Marvin Frazier, Richard A. Gibbs, Donna M. Muzny, Steven E. Scherer, John B. Bouck, Erica J. Sodergren, Kim C. Worley, Catherine M. Rives, James H. Gorrell, Michael L. Metzker, Susan L. Naylor, Raju S. Kucherlapati, David L. Nelson, George M. Weinstock, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Todd Taylor, Jean Weissenbach, Roland Heilig, William Saurin, Francois Artiguenave, Philippe Brottier, Thomas Bruls, Eric Pelletier, Catherine Robert, Patrick Wincker, Douglas R. Smith, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Hong Mei Lee, Joann Dubois, André Rosenthal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, Andreas Rump, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Jun Gu, Leroy Hood, Lee Rowen, Anup Madan, Shizen Qin, Ronald W. Davis, Nancy A.

Federspiel, A. Pia Abola, Michael J. Proctor, Richard M. Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood, David R. Cox, Maynard V. Olson, Rajinder Kaul, Christopher Raymond, Nobuyoshi Shimizu, Kazuhiko Kawasaki, Shinsei Minoshima, Glen A. Evans, Maria Athanasiou, Roger Schultz, Bruce A. Roe, Feng Chen, Huaqin Pan, Juliane Ramser, Hans Lehrach, Richard Reinhardt, W. Richard Mccombie, Melissa De La Bastide, Neilay Dedhia, Helmut Blöcker, Klaus Hornischer, Gabriele Nordsiek, Richa Agarwala, L. Aravind, Jeffrey A. Bailey, Alex Bateman, Serafim Batzoglou, Ewan Birney, Peer Bork, Daniel G. Brown, Christopher B. Burge, Lorenzo Cerutti, Hsiu-Chuan Chen, Deanna Church, Michele Clamp, Richard R. Copley, Tobias Doerks, Sean R. Eddy, Evan E. Eichler, Terrence S. Furey, James Galagan, James G. R. Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L. Steven Johnson, Thomas A. Jones, Simon Kasif, Arek Kaspryzk, Scot Kennedy, W. James Kent, Paul Kitts, Eugene V. Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M. Lowe, Aoife Mclysaght, Tarjei Mikkelsen, John V. Moran, Nicola Mulder, Victor J. Pollara, Chris P. Ponting, Greg Schuler, Jörg Schultz, Guy Slater, Arian F. A. Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Yuri I. Wolf, Kenneth H. Wolfe, Shiaw-Pyng Yang, Ru-Fang Yeh, Francis Collins, Mark S. Guyer, Jane Peterson, Adam Felsenfeld, Kris A. Wetterstrand, Aristides Patrinos, and Michael J. Morgan. "Initial Sequencing and Analysis of the Human Genome." *Nature* 409.6822 (2001): 860-921. Web.

Licatalosi, Donny D., and Robert B. Darnell. "RNA Processing and Its Regulation: Global Insights into Biological Networks." *Nature Reviews Genetics* 11.1 (2010): 75-87. Web.

Manna PR, Dyson MT, Eubank DW, Clark BJ, Lalli E, Sassone-Corsi P, Zeleznik AJ, Stocco DM (January 2002). "Regulation of steroidogenesis and the steroidogenic acute regulatory protein by a member of the cAMP response-element binding protein family". Mol. Endocrinol. 16 (2002): 184–99

Necsulea, Anamaria, Magali Soumillon, Maria Warnefors, Angélica Liechti, Tasman Daish, Ulrich Zeller, Julie C. Baker, Frank Grützner, and Henrik Kaessmann. "The Evolution of LncRNA Repertoires and Expression Patterns in Tetrapods." *Nature* (2014): n. pag. Web.

Snustad, D. Peter, and Michael J. Simmons. *Principles of Genetics*. 6th ed. Chichester: John Wiley & Sons, 2012. Print.

Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R. Kelley, Harold Pimentel, Steven L. Salzberg, John L. Rinn, and Lior Pachter. "Differential Gene and Transcript Expression Analysis of RNA-seq Experiments with TopHat and Cufflinks." *Nature Protocols* 7.3 (2012): 562-78. Web.

Vlahopoulos SA, Logotheti S, Mikas D, Giarika A, Gorgoulis V, Zoumpourlis V (April 2008). "The role of ATF-2 in oncogenesis". BioEssays 30 (2008): 314–27.

Warf, M. Bryan, and J. Andrew Berglund. "Role of RNA Structure in Regulating Pre-mRNA Splicing." *Trends in Biochemical Sciences* 35.3 (2010): 169-78. Web.

Whitfield, James F., and James F. Whitfield. *Calcium in Cell Cycles and Cancer*. Boca Raton, FL: CRC, 1995. Print.