Week 4 Progress - Janis Corona

by Janis Corona

Submission date: 29-Jun-2019 11:34PM (UTC-0500)

Submission ID: 1148026471

File name: Week_4_Progress.docx (21.26K)

Word count: 791

Character count: 3799

Janis Corona

BIOL 59000: Data Science Project for Life Sciences

Week 4 Progress

June 29, 2019

Meta-Analysis of the Ubiquitous Genes Associated with Human Uterine Leiomyoma Development in Healthy Human Tissue – Week Three Progress

During this fourth week of research on uterine leiomyoma (UL) genes ubiquitous to the current research studies published on UL gene risk factors in healthy females, more analysis and visualizations were developed for the materials and methods as well as for the results section of this research. R was used to analyze the gene expression data from the five studies obtained from the Gene Expression Omnibus (GEO) and to also map out the 130 genes in common between the five studies of 121 samples consisting of 70 UL and 51 non-UL samples. Many PNG plots and additions to the R script of commands were made as the analysis and synthesis of the gene expression data unraveled. All of these additions can be obtained at https://github.com/JanJanJan2018/Better-Cleaned-Version-UL-Research for the plots, data tables and script. The script has the notes that go with the tables developed for analyzing the gene expression data.

Some of the plots made were better annotations of the genes that are expressed more (up regulated) and expressed less (down regulated) in UL tissue compared to non-UL tissue samples. Tables were made that could be added to the methods or results section showing up and down regulated genes per chromosome, as well as if those genes are up or down regulated as a majority of other genes on that chromosome that were up or down regulated. The chromosomal locations were the filtering method to shrink down the very large data table of genes from gigabytes to megabytes of data when merging the five studies together by genes in common. These five studies were selected from six studies because they all had the six genes ubiquitous to UL research of TNRC6B, BET1L, CYTH4, HMGA2, CCDC57, and FASN. An R package, lattice, was used to show a pairwise gene relationship of scatter plots on non-UL samples, but the image couldn't show any definite relationships that could be linear. When more than five genes were compared to each other at a time, the plot looked like a giant mess of tiny squares in an array having splotches in each square. These images of non-UL pairwise gene comparisons were done on 10 samples each out of some of the 51 non-UL samples and uploaded to github as Pairwise_10_Most_DE_in_nonUL.png and Pairwise_10_Least_DE_in_UL.png. The data was further filtered so that the six genes ubiquitous to the UL research studies and the top 10 genes having the highest magnitude of expression in UL compared to non-UL samples in up or down regulation is the final data set. This could also be good as a table for the methods. Some bootstrap simulations of a population mean from random sampling of the first gene in the 16 top genes table was done on FSCN2 for non-UL and UL samples to get a better approximate mean, a standard deviation, and a two-tail 95% confidence interval that is in the process of being created for all of the 16 genes. Using this data, the goal is to have the data fit a model such as linear regression to the 16 genes original output for each of the 121 UL and non-UL GEO samples in predicting 30 per cent of the samples as either UL or non-UL.

The packages used in R thus far are Gviz, lattice, ggplot2, UsingR, and dplyr. More work must be done on building the table of simulated means for each gene in the UL and non-UL samples from the simulation of 10,000 sampling observations with replacement using each gene as a vector of values for each sample. The genes from the non-UL group have 51 values or columns and 10,000 simulated observations, and the genes from the UL group have 70 values and 10,000 simulated observations.

When grouping the genes per chromosome that were up or down regulated, filtering was done within those groups to show genes as part of the majority of genes on that chromosome up or down regulated or part of the minority using a Boolean value for a created field called 'majority'. It was curious to find that three of the six genes are not part of the majority on chromosome 11 and 17, while those on chromosome 12 and 22 are up or down regulated as part of the majority. The chromosome plots were able to be annotated with the up and down regulated genes and having the gene symbol. These plots are in the github web link as PNG images having majority/minority, Down/Up, chr, and 11, 12, 17, or 22 in the file names, such as: majorityDownChr11.png or minorityUpChr17.png.

Week 4 Progress - Janis Corona

ORIGINALITY REPORT

1%	0%	0%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Grand Canyon University Student Paper	1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		

Week 4 Progress - Janis Corona

GRADEMARK REPORT

FINAL GRADE

8/10

GENERAL COMMENTS

Instructor

Janis,

It is clear that you have continued with progress on your project, and that you are using data science approaches to work with the data you've gathered. However, I am concerned about whether the analyses you are currently using are keeping you on the right track to make sure you are able to address your overall research question. As you move into week 5 and completing the draft of your paper, please make sure to make sure that your research question is very clear. Don't forget, the objective of your project is to build upon what is already known about the subject, but also make a unique contribution to the field. Therefore, what is the purpose of your work and how is this extending the current knowledge of the topic? Then, as you review your methods (and the results of your analyses), it is very important to assess:

1.) How does this analysis support the overall purpose of the study?

2.) What is the biological relevance of the result? How do you interpret the outcome of the analysis?

3.) How does this observation extend what was already known? And with this information that has been observed, what might be the next logical extension (what else would you want to know)?

PAGE 1

Comment 1

Thank you for making this distinction. It will be important to make the directionality of expression changes clear within your written and oral reports.

Comment 2

This is an interesting way to think about changes in gene expression (relative to coordinates)

This is an interesting way to think about changes in gene expression (relative to coordinates). Based on what you've learned, do you think there is any relationship between the physical location of a gene within the genome (and more specifically, along the length of a chromosome) and its expression profile? If you haven't thought about this extensively, you may want to look into some basics about how eukaryotic genes are regulated (mechanisms) as well as chromosome territories/placement of parts of the chromosome within the nucleus, in relationship to expression. Especially for an audience that may not be well versed in the biology, it will likely be helpful to include some commentary about this within your report -- in the written explanation, it will likely fit into either the results or discussion section.

PAGE 2

Comment 3

A note about formatting convention: often, the name of a gene and its corresponding gene product (often a protein) will be the same. In order to distinguish between the gene and protein, differential formatting of the name is often used. For the sake of your report/presentation, please use the NCBI Style Guide approach, which gives the gene name in italics and the protein name in regular font. For more information about NCBI conventions for formatting, see Chapter 5 of the NCBI Style Guide here: https://www.ncbi.nlm.nih.gov/books/NBK995/

Comment 4

In the published articles that have dealt with similar analyses, is this the type of investigation they carried out? What evidence do you have that looking for a linear relationship between a large number of genes will be useful? For example, check out this article and look at figure 5B: https://science.sciencemag.org/content/298/5597/1395.long

Here, each dot represents a single gene and the coordinates give information about how much it is expressed in WT and KO (not that different than the instance of non-UL versus UL). There are some genes that fall along the line -- but those are the "not interesting" ones, as that means the expression is exactly the same in the two different sample types. The more important genes (such as those highlighted in red) are the ones that are away from the line. Therefore, unless I am misunderstanding what you're working on, I don't know that looking for a linear relationship will actually reveal something "cool" or of interest.

Comment 5

Rather than doing this type of pairwise comparison, have you used the data for generate a heat map? This type of analysis will often cluster together cohorts of genes that take on similar patterns of change (such as all those that are highly up regulated from non-UL to UL, those that are highly down regulated from non-UL to UL, etc.).

Comment 6

Is this similar to a format that you've seen in some of the published works you've read in preparing for this project? I'm not familiar with this approach to present expression data.

Comment 7

Typically, the methods section should be explanatory about how you completed an analysis, but not show any outcomes of that analysis. Plan on incorporating the outcomes as part of the "results" section of your paper.

Comment 8

It seems here that you've placed a good amount of emphasis on genes that have already been identified as important -- but remember that your research needs to be generating something novel. Are you using these as a way to validate your approach (on the way to generating something that is an addition to the knowledge base)?

Comment 9

What is the utility in performing this type of analysis? Remember, it will be important to remind your audience why you have used this approach, in the context of the biological relevance. What are you able to learn? How will this be helpful for understanding more about UL?

Comment 10

What is the significance in this outcome? It will be very important to make sure that you communicate what can be learned from this observation, in the context of what it means for the biology/behavior of these cells. It can be interesting to find certain outcomes in sets of data, but unless there is an interpretation for what that means "in real life" it will be difficult to comprehend why the audience should "care" about this outcome.

RUBRIC: PROGRESS REPORT RUBRIC 4 / 5

CONTENT (50%) 3 / 5

5 (5)	Exemplary demonstration of project progression; provided materials greatly augment the status of ongoing research
4 (4)	Very good demonstration of project progression; provided materials show ongoing research work
3 (3)	Adequate demonstration of project progression; some ideas posed or resources shared, but not fully connected to previous work
2 (2)	Limited demonstration project progression; is not clear that significant gains have been made
1 (1)	Inadequate demonstration of project progression; update does not demonstrate time has been devoted to working on project

GRAMMAR (50%) 5 / 5

5 (5)	Less than 3 spelling, grammatical or mechanical errors
4 (4)	No more than 5 spelling, grammatical and/or mechanical errors
3 (3)	Fewer than 8 spelling, grammatical and/or mechanical errors
2 (2)	Less than 10 spelling, grammatical and/or mechanical errors
1 (1)	More than 10 spelling, grammatical and/or mechanical errors