Week 2 Progress-Janis

by Janis Corona

Submission date: 16-Jun-2019 02:34AM (UTC-0500)

Submission ID: 1144073312

File name: Week2_progress_submittal.docx (17.12K)

Word count: 678

Character count: 3367

Janis Corona

BIOL 59000: Data Science Project for Life Sciences

Week 2 Assignment: Research progress

June 15, 2019

Progress Made During Week 2

This week more work was completed using R coding to combine the Gene Expression Omnibus (GEO) data series into one data set of uterine leiomyomas (UL) and non-UL samples with other genes sharing the same chromosome location as the top genes' chromosomes. Originally there were six series to combine, but this is now five series of 121 microarray gene expression samples instead of 133. These samples consist of 70 UL samples and 51 non-UL samples. All have the six ubiquitous genes in the microarray gene expression data from GEO of TNRC6B, BET1L, CYTH4, CCDC57, FASN, and HMAG2. The excluded data series is GSE764 and it only had FASN and HMGA2. This series is still attached as a reference as a limitation for the discussion section. There are 183 genes from data that originally had between 22,000 and 55,000 genes depending on which GEO series the data was combined from. More coding was going to be done to lay down tracks of the genes like what ENSEMBL does with genes, but the operating system this research was being developed on crashed and a replacement computer was purchased the following few days.

Other progress towards this research was on making it clear that the focus of this research is not on the genotypes or single nucleotide polymorphisms (SNP)s of the ubiquitous genes found in current population studies to pose a risk of UL in females. This research is on the genes expressed in the microarray data from the five GEO data series to compare the top six ubiquitous genes with other top genes and determine how the gene expression compares across a set of UL samples to a set of non-UL samples.

Corrections from feedback of the Week 1 submittal were made in progressing through this research. The GEO direct link to the accession ID for the platform and series data obtained for this research work was placed into each GEO reference. The PMID from GEO series referenced items in the report were omitted. A run-on sentence on the first page was corrected in the proposal.

Further expansion and explaining of other items were made to the research. The relationship between estrogen and UL growth and treatment involving an estrogen inhibitor to cease UL growth was explained better than previously reported. An explanation of how the training and testing sets would be selected as partitions in the R caret package was made. The ubiquitous genes' SNPs reported as significant in studies were based on Minor Allele Frequency count and it is not known if these genes have a low or high expression in UL or non-UL data from the studies currently reviewed for this research. Although, one exception is the ubiquitous gene CYTH4 found to be expressed low in thyroids of African American females who have UL. The identifiers in the data only have sequence, chromosome, and cytoband information in one of the five data sets. The other identifiers are the alternate IDs to combine data or merge the data together by, then filter by chromosome location. This data is all human samples as only one series had rat samples, but those samples were omitted from this data collection by sample name. Possible models to build are based algorithms to train using the training data set such as linear, logistic, Bayesian, Random Forest, and K-Nearest-Neighbor. These models would each be tested on the testing data partition for accuracy in prediction.

Some more exploring with R and Bioconductor of this completed data set on UL and non-UL samples is needed to further progress towards completing research on the 'Meta-Analysis of the Ubiquitous Genes Associated with Human Uterine Leiomyoma Development in Healthy Human Tissue.' Once the data has been modeled accurately to predict a UL or non-UL sample based on gene expression, the genes will be layed out in a chromosome plot to show the top differentially expressed genes between UL and non-UL samples in unidentified race demographic samples.

Week 2 Progress-Janis

ORIGINALITY REPORT

4%	0%	0%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Lewis University Student Paper	2%
2	Submitted to University of Technology, Sydney Student Paper	2%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		

Week 2 Progress-Janis

GRADEMARK REPORT

FINAL GRADE

9/10

GENERAL COMMENTS

Instructor

PAGE 1

Comment 1

As written, it is not clear what these 183 genes represent, and under what criteria/circumstances the list was condensed from the original large set to this small subset. Please make sure that you are fully explaining this process as you work on your project. Also, please make sure that you elaborate on this moving forward. I want to make sure that I understand the steps that you are taking -- the only way I can help advise against missteps or problems in your approach is if I have all the information in order to evaluate.

Comment 2

What led you to focus on the "top six"? What do you mean by "top"? Are these the genes with the biggest fold change in expression? If so, is that in the case that the genes are more highly express in UL relative to non-UL? Or are they the biggest down-regulation in UL relative to non-UL? Keep in mind that sometimes big changes in phenotype are the result of large changes in transcription, but in other instances, significant chances in phenotype can be the consequence of minor changes in transcription rates. It will be very important to define the process by which you narrow your focus and determine if difference is meaningful.

PAGE 2

Comment 3

I'm glad that you made this edit. Can you include as part of your document next week?

Comment 4

Can you also include this portion in your next update, too?

Comment 5

I'm not sure that I understand the language that is given here. "Ubiquitous" typically is used to mean "everywhere" -- which is interpreted commonly as the same across (that is, it does not vary from individual). SNPs, though, are very specifically differences in the genetic code. Please review the word choice here and make sure to clarify in your document in the future.

Comment 6

As you work on your paper, you will need to make clear the benefits/limitations of each of these approaches, aside from the ultimate outcome that one likely performs better than the others.

Comment 7

This might be a really great way to present the information, depending on the number of genes you are ultimately presenting as differentially regulated. If there are a lot of genes, though, you may need to re-think this presentation. Ultimately, you will want to make sure that your presentation of data is able to clearly communicate information (and not overwhelm or confuse the audience).

RUBRIC: PROGRESS REPORT RUBRIC 4.50 / 5

CONTENT (50%) 4 / 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