

# Week 3 Progress - UL using RStudio

*by* Janis Corona

---

**Submission date:** 22-Jun-2019 01:44AM (UTC-0500)

**Submission ID:** 1146034232

**File name:** Week\_3\_Progress.docx (24.19K)

**Word count:** 1708

**Character count:** 8819

Janis Corona

BIOL 59000: Data Science Project for Life Sciences

Week 3 Progress

June 21, 2019

1  
Meta-Analysis of the Ubiquitous Genes Associated with Human Uterine Leiomyoma Development in  
Healthy Human Tissue – Week Three Progress

During this third week in progressing towards research on the meta-analysis of gene expression microarray data on uterine leiomyoma (UL) and non-UL tissue samples of otherwise healthy **homo sapiens**, much was explored in the data using R and some solutions to feedback from the second weekly progress report were developed. Exploration was done with using the Bioconductor and R package called Gviz to draw chromosome plots of the genes being studied and compared to top gene comparisons in the **data from the Gene Expression Omnibus (GEO)**. Analysis of the compiled and combined data of only genes that were in common between the five microarray data sets was done using R base graphics and the R package called ggplot2. Corrections to the weeks one and two feedback reports were made and included in this report by referencing a review article on UL that one research article claiming that UL are estrogen dependent.

When combining the data from the five data sets in GEO that had like genes, the duplicate genes were removed. Exon information was able to be added to the gene samples but expanded the genes with markers for each exon with the same duplicate entry for each gene symbol or ENSEMBL ID, so the exon information was removed. The exon information was only needed to further test the Gviz software to see how it could produce the same chromosome type map that UCSC and ENSEMBLE have for each gene location on the reverse or forward strand of each chromosome and the neighboring genes. The exon information made a 130 gene table of 121 samples of UL and non-UL microarray data into a 7,000 gene table of many duplicate genes. Because of this large duplication of genes, the 'stacking' of genes on



the chromosomal bands was too much to plot. Originally the strand information for forward and reverse direction wasn't in the data, but the BioMart tab of ENSEMBL allowed for locating the strand direction by ENSEMBL transcript ID. When using Gviz, the chromosomes were able to be stacked with markers but not annotated with the gene. All plots and data less than 25 mb in file size have been uploaded to github to look at the progress, R script, notes, documentation, and charts. This Github online link is

<https://github.com/JanJanJan2018/Better-Cleaned-Version-UL-Research>

Using R, analysis was done on the data sets to see if there were any linear relationships between the most expressed and least expressed genes between the UL and non-UL sample. So far, there haven't been any relationships, but only a handful of genes have been compared, and the list is not the same across samples when moving from the set of UL gene expression to the set of non-UL gene expression. Scatter plots were made showing which of the five groups each sample is from and separately if the observation between two genes out of the 130 show a linear relationship. These plots have also been uploaded to the github web address at <https://github.com/JanJanJan2018/Better-Cleaned-Version-UL-Research> to view. Further exploration like plotting pairwise comparisons between the gene expression continuous values using the R package lattice and ggplot2 needs to be done on this data to zoom into the genes having dramatic changes between UL and non-UL samples. Many new fields were added, dropped, and explored to see what information it could provide to this combined set of UL and non-UL microarray gene expression data. Such as adding the ENSEMBL 'start', 'end', 'width', 'strand' field. The categorical field for showing color scatters of the observations detailing the sample set derived from. The 'exon' field was also added, but removed after discovering it isn't needed to use Gviz for gene plotting on a chromosome map.

When explaining the effects of estrogen on being declared a hormone that UL depends on to grow, this claim had to be extracted from the source of the paper the claim came from. When reviewing this article by Dvorska, Brany, Dankova, Halasova, and Visnofsky (2017), more information about the

type of hormonal treatments that shrink UL size became clear. Such as, estrogen analogues (stop gonadotropin from being produced) and estrogen antagonists (compete with androgens, progesterones, and glucocorticoids in receptor binding) have been shown to reduce the UL size in some patients, but also stop the UL from growing. To expand on this feedback in the proposal, the following change was made:

“Dvorska et al. (2017) says that androgens (hormones in the body that regulate the male body but present in all  homo sapiens) produced by the ovaries and the adrenal glands turn into  oestrogens (a family of hormones that regulate the female body) in adipose tissue, and that with every 10 kg of body fat an obese client has a 20 per cent increased risk of UL development. Dvorska also says that UL have more oestradiol receptors than neighboring myometrium tissue by 20 per cent. UL are considered oestrogen dependent because when UL patients in separate studies were given either estrogen agonists (compete with glucocorticoids, androgens, and progesterone in binding receptors) or analogues (inhibit production of gonadotropin) their UL shrank in size, but began growing once this type of treatment ended (Dvorská et al., 2017; Rafnar et al., 2018).”

The above block quote was included in the background section of the proposal to give an interesting and a more detailed explanation of estrogen dependency of UL.

Corrections and clarifications from the latest appreciated feedback was also made. When selecting the top ubiquitous genes, six of the genes that were either compared, highlighted in one of the original nine population studies, or found to be associated with UL risk were selected. The use of ubiquitous is only for those genes found in current studies to have a risk of UL. The gene expression of the gene wasn't a factor in the studies because these studies focused on SNPs and genotypes that differed within those genes. The genotypes of those GEO samples would have been selected if more than one of the GEO sets had that information, but only one data set

had the actual sequencing and genotype information attached to its UL and non-UL samples.

When using the algorithms suggested to analyze the data the reasons for using the algorithms

selected will be explained as well as the results. The linear modeling is currently being used for a

quick analysis of data trends in the microarray data of gene expression. This research will only

explore a few of the genes having relationships in gene expression changes in levels from UL and

non-UL samples and attempt to build a model that will predict if the sample is a UL or non-UL

sample once those genes showing changes in the samples is discovered. Right now, analysis is

still being done on the data to find some relationship.

## References

<sup>2</sup>  
Dvorská<sup>1</sup>, D., Braný, D., Danková, Z., Halašová, E., & Višňovský, J. (2017). Molecular and clinical treatment of uterine leiomyomas. *Tumor Biology*, 39(6). DOI: 10.1177/1010428317710226.

This article is one of the referenced articles in Rafnar et. al.'s (2018) article that further explains the estrogen hormonal agonist and antagonist treatments on UL patients. Some estrogen agonists such as Gonadotropin-releasing hormone (GnRH) can stop UL from growing in women symptomatic for UL by inhibiting the production of gonadotropin which creates a hypogonadal or under active hormonal state in UL patients. But once the treatment ends, the size of the UL can again increase. This article mentioned the GnRH antagonists that compete for receptor binding against other androgens, progesterone, and glucocorticoids in the body, with an effect that reduces the size of UL in females symptomatic with UL. This article states that in the UL myometrium, there are more oestrogen receptors making it more sensitive to oestradiol by binding 20% more in myometrial UL tissue than normal myometrium tissue. This research connects this to a fact that more obese women (having more than 30% fat in their weight content) have UL than non-obese females, because adipose tissue is where ovarian and adrenogland derived oestrogens are made.

<sup>4</sup>  
Rafnar, T., Gunnarsson, B., Stefansson, O.A., Sulem, P., Ingason, A., Frigge, M.L., ... Stefansson, K. (2018).<sup>1</sup>

Variants associating with uterine leiomyoma highlight genetic background shared by various cancers and hormone-related traits. *Nature Communications*, 9:3636. DOI:10.1038/s41467-018-05428-6

The research done in this article involved a meta-analysis of two GWAS studies of UL using Icelandic and English European females. The patients with UL are the case group and the volunteers without UL are the control group. There are two separate studies in this research.

One study is on genes expressed in cancers and other benign tumors that are also expressed in UL. The other study is on the putative loci regions associated with hormone related diseases and the changes in those loci in UL. This research elucidated the relationship that hormones have on UL growth and made explicit the common genes being expressed between UL, cancer and benign tumors elsewhere in the body. The information about hormonal therapy to treat symptoms of estrogen responsive leiomyomas before hysterectomy can cause the symptoms to recur was found in this article. The TNRC6B and the BET1L genes were found to also be associated with UL in this study. But the BET1L gene found in Japanese populations and the CYTH4 gene found in African American populations were not found to be associated with UL in this study on European women. This study on Europeans confirmed one of the endometrial cancer genes associated with UL is r10917151 of CDC42/WNT4. This study found reason to exclude the other seven genes previous GWAS studies found to be associated with UL and cancer. This study also shows in a table of polygenic risk scores that there is a significant association of ULs in patients with thyroid cancer ( $R^2 = 21\%$  and  $P \text{ value} = 3.0 \times 10^{-5}$ ), endometriosis Stage III and IV ( $R^2 = 11\%$  and  $P \text{ value} = 4.1 \times 10^{-3}$ ), and kidney cancer ( $R^2 = 10\%$  and  $P \text{ value} = 2.43 \times 10^{-3}$ ). This research on Europeans, shows a connection between thyroid disfunctions and thyroid cancer that the research done on African American populations also found to be associated with UL risk.

# Week 3 Progress - UL using RStudio

## ORIGINALITY REPORT

22%

SIMILARITY INDEX

3%

INTERNET SOURCES

4%

PUBLICATIONS

23%

STUDENT PAPERS

## PRIMARY SOURCES

1

Submitted to Lewis University

Student Paper

19%

2

Submitted to Australian Catholic University

Student Paper

1%

3

Submitted to Colorado State University, Global Campus

Student Paper

1%

4

discovery.ucl.ac.uk

Internet Source

1%

Exclude quotes Off

Exclude bibliography Off

Exclude matches Off



# Week 3 Progress - UL using RStudio

---

## GRADEMARK REPORT

---

FINAL GRADE

GENERAL COMMENTS

**Instructor**

9/10

---

PAGE 1



### Comment 1

Make sure that you are using correct scientific formatting here.



### Comment 2

If I am understanding you here -- the data set wound up having information for each exon within a gene (and that's how you wound up with multiple entries per gene)? Is there not a way to incorporate the overall coordinates per gene (base #4-500, for instance)? That would then give you only a single listing per gene.

---

PAGE 2



### Comment 3

If you are locating the strand location this way, couldn't you incorporate the coordinates, too? That would give you something that is the format of chromosome#: base-position-at-start-of-gene -- base-position-at-end-of-gene.



### Comment 4

For right now, this is a good way to organize your work (and make sure you're not losing things). As you work on your paper, though, how will this information be incorporated? Have you worked on taking your process and putting it into methods? For parts that you have developed (such as your new, novel (compiled) data set), how will you share it? In the appendix section of your paper, you can write an item description and then provide a link to something -- but it will need to be a specific link to just that item (and not a larger repository of information).



### Comment 5

Here, what do you mean by "linear" relationships? In other previously completed studies comparing gene expression in (generally speaking) cancer sample versus healthy tissue, what type of comparisons have been made between the two populations? How are these data typically represented in figures?



### Comment 6

Is this a typical way that you have seen the comparison of data portrayed elsewhere?



### Comment 7

I did take a look at most (if not all) of the png files. For your next progress report, I'd like to see you format figures the way they ultimately will appear in your paper/presentation. Remember, all components need to be clearly labeled (so that the audience can, without a doubt, know what she/he is viewing). You'll also want to consider whether you are presenting information that can easily be interpreted (and demonstrate an analysis that is biologically relevant -- what is the meaning in what you are showing?).



### Comment 8

I believe this would cover the "coordinate" information I mentioned earlier -- am I incorrect?

PAGE 3

---



### Comment 9

Make sure you correctly format this scientific name.



### Comment 10

Estrogen = Oestrogen (just two different names to the same thing). American English typically uses "estrogen" -- either way, pick one spelling and use it consistently.

PAGE 4

---



### Comment 11

Will linear modeling be the most appropriate? Are you guessing that checking a single gene (and whether it goes up or down between UL and non-UL) will be a successful predictor? Or are you going to need to find a larger, overall signature (several different genes, all with a characteristic UL-non-UL pattern)? If the case is that you need more of a signature, how important is a linear relationship?

PAGE 5

---

PAGE 6

---

---

CONTENT (50%)

4 / 5

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

---

GRAMMAR (50%)

5 / 5

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