META-ANALYSIS OF THE GENES UBIQUITOUSLY ASSOCIATED WITH HUMAN UTERINE LEIOMYOMA DEVELOPMENT IN HEALTHY HUMANS USING THE GENE EXPRESSION OMNIBUS DATA

Janis Corona

2019

Description of Uterine Leiomyoma

What is a uterine leiomyoma (UL)?

What are the symptoms of UL?

What are risk factors of UL?

Who can get UL?

What is the treatment for UL?

How do UL develop?

UL Risk in Population Studies

TNRC6B:

UL risk gene target in Saudi Arabians, Japanese, Chinese, European Americans, and Europeans including the Icelandic and UK populations specifically

BET1L:

UL risk for Japanese, Chinese, Europeans, European Americans, and Saudi Arabians (only when TNRC6B also shows UL risk)

CYTH4:

UL risk only for African Americans

FASN:

UL risk for European Americans, Australians, Icelandic, and other Europeans

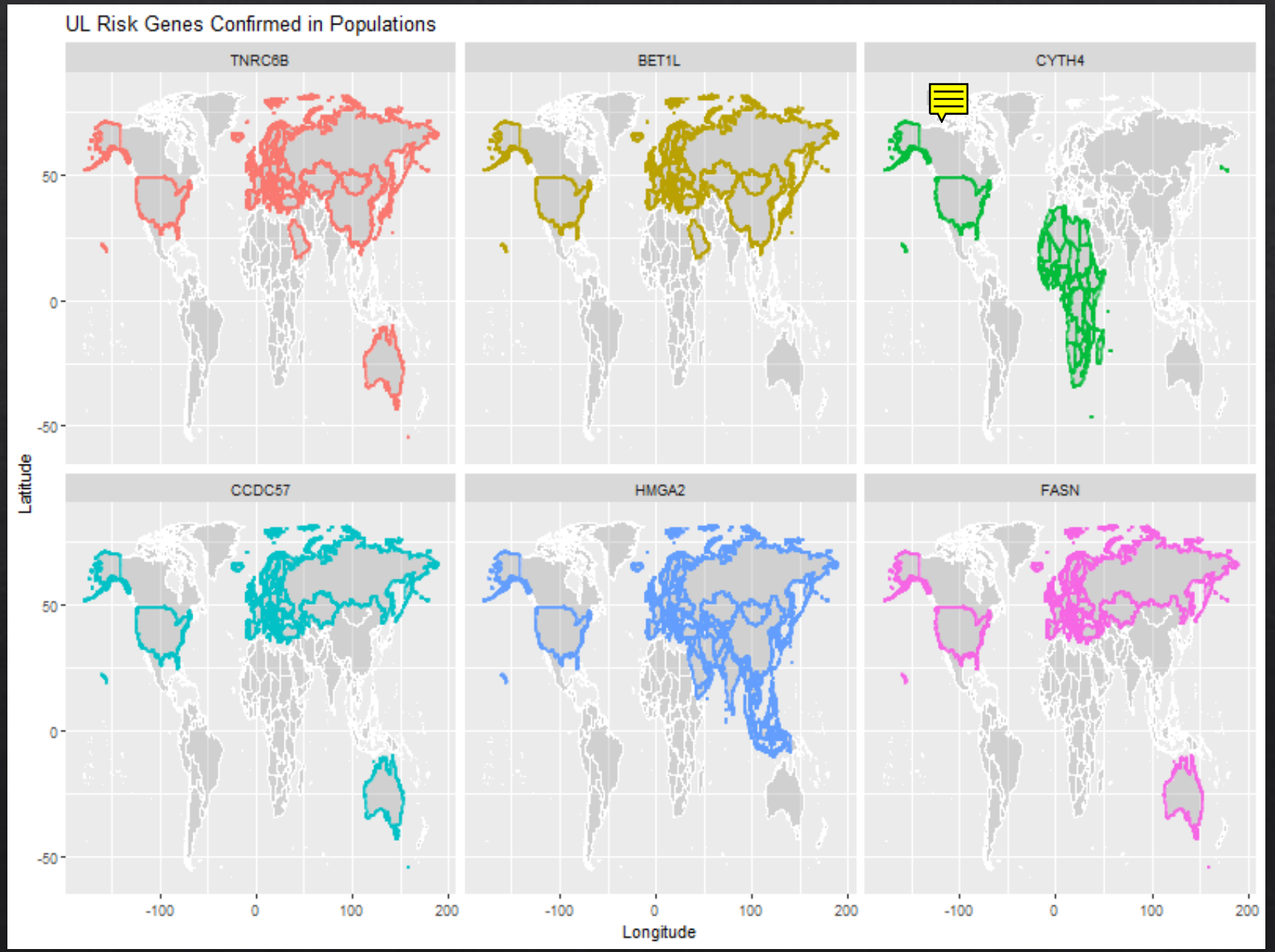
CCDC57:

UL risk for European Americans, Australians, Icelandic, and other Europeans

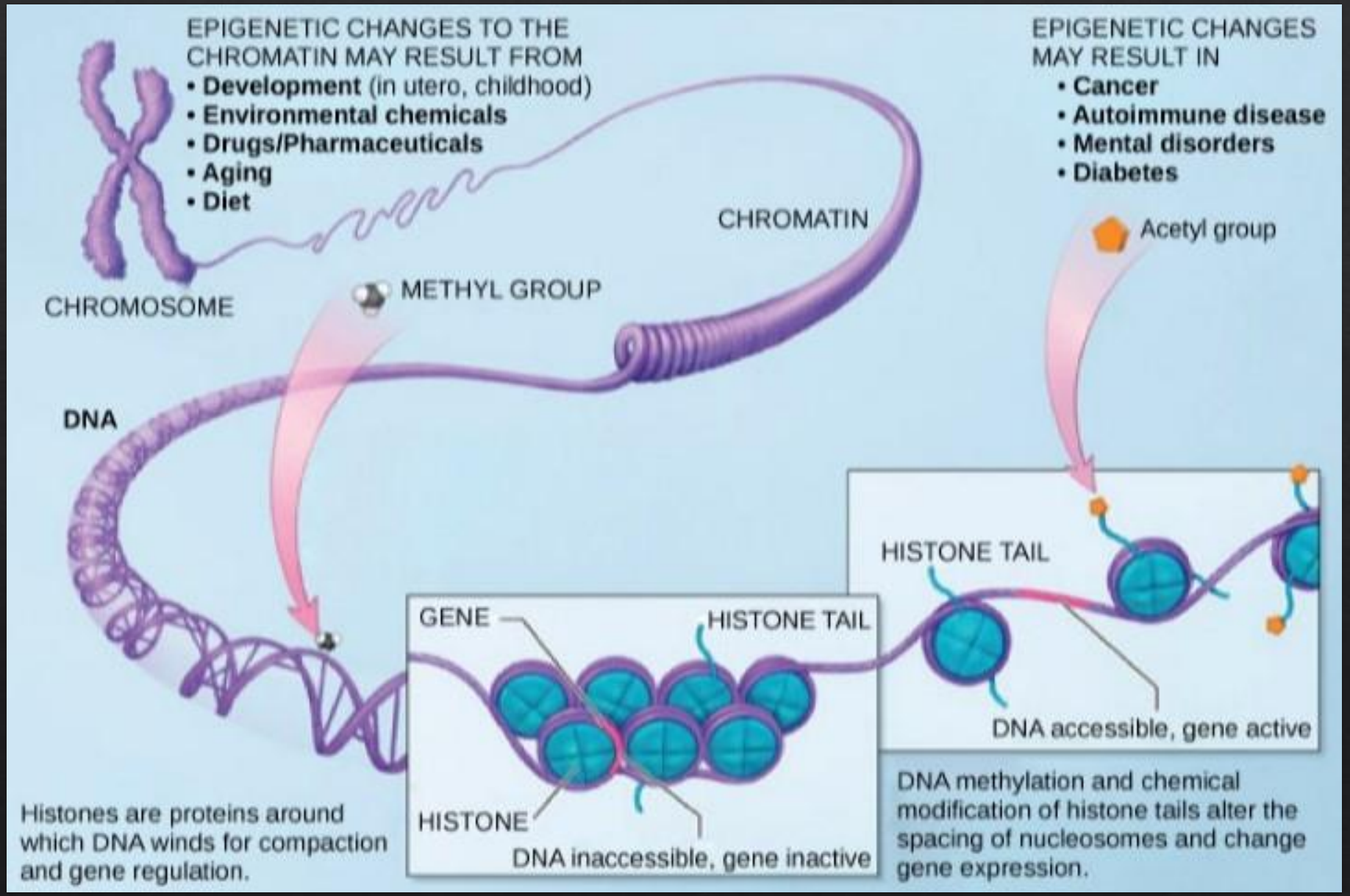
HMGA2:

UL risk for European Americans, Europeans, Australians, and Asian Americans

Map of the Population Studies Genes for UL Risk

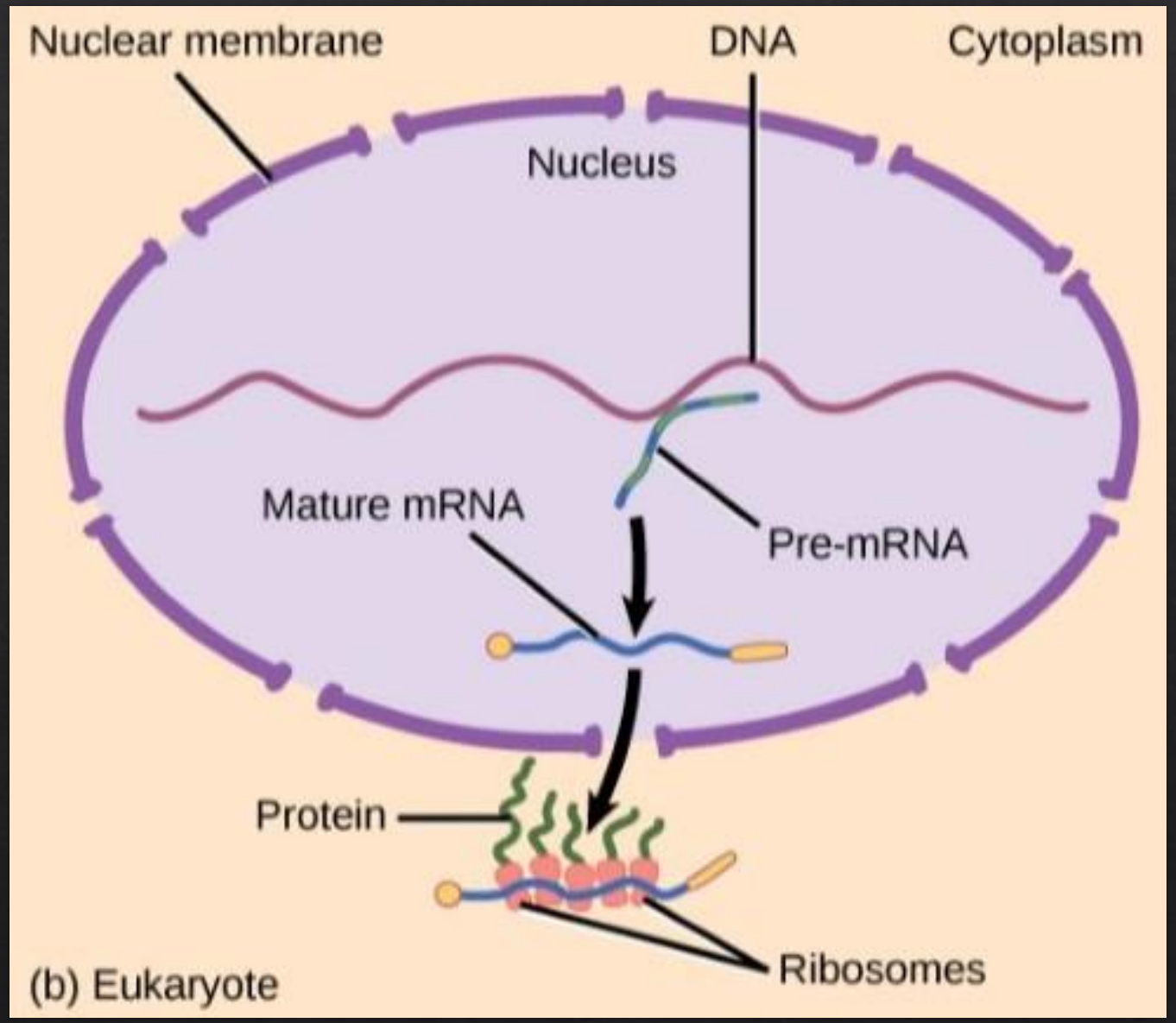


Gene Regulation: Transcription



Rye, Wise, Jurukovski, DeSai, Choi, & Avissair (2017) *Biology: OpenStax*

Gene Regulation: Transcription and Translation



Rye, Wise, Jurukovski, DeSai, Choi, & Avissair (2017) *Biology: OpenStax*

Gene Regulation: Gene Expression

What affects gene expression of RNA?

Where can the gene expression of RNA be influenced?

Objective of this Research

Meta-analysis of those six genes and/or a combination of the other top genes showing the most change in DE or fold change in UL compared to non-UL microarray samples from five GEO studies are good predictors for determining a sample as UL or non-UL. Thus, indicating the gene is a possible gene target for UL pathogenesis treatment through further research of those genes with good accuracy in prediction for UL.

Methods - Overview

- **Universal** data set of 12,173 genes from GEO's five UL risk studies using microarray gene expression data
- Subset of **130** genes from GEO that only belong to Chromosomes 11, 12, 17, or 22
- R, combining data, plotting data, machine learning, and all analysis is completed using R
- Bioinformatics, Gviz package, both use R to generate the chromosome plots of genes
- Different data sets were prepared to use machine learning on and decide which genes were better predictors.

Methods – Linkage Disequilibrium Visualizations

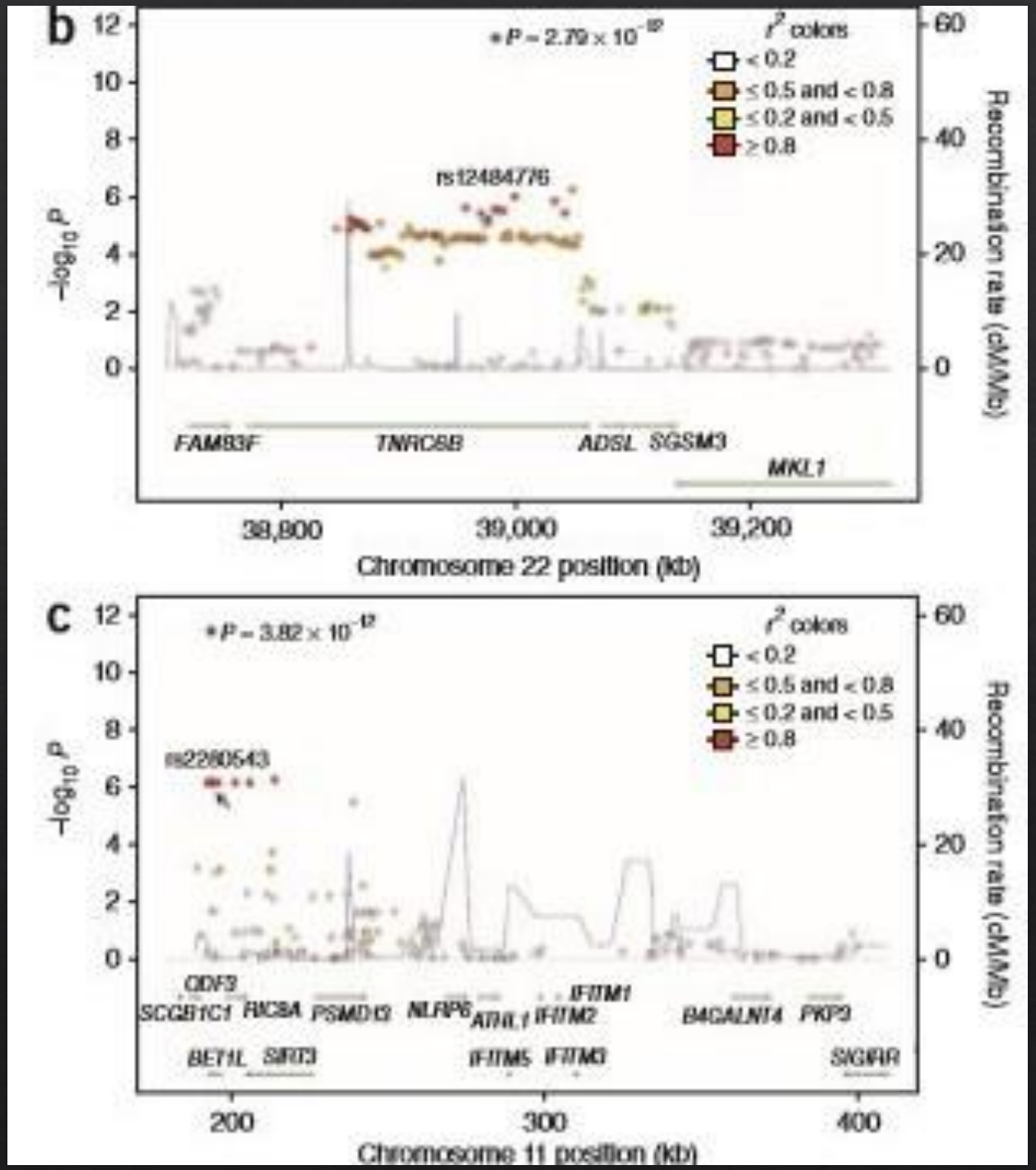
Gviz, a Bioconductor R package was used to visually see where peak links between gene regions could lie on the chromosome.

Gene Expression through Linkage Disequilibrium

What is linkage disequilibrium (LD)?

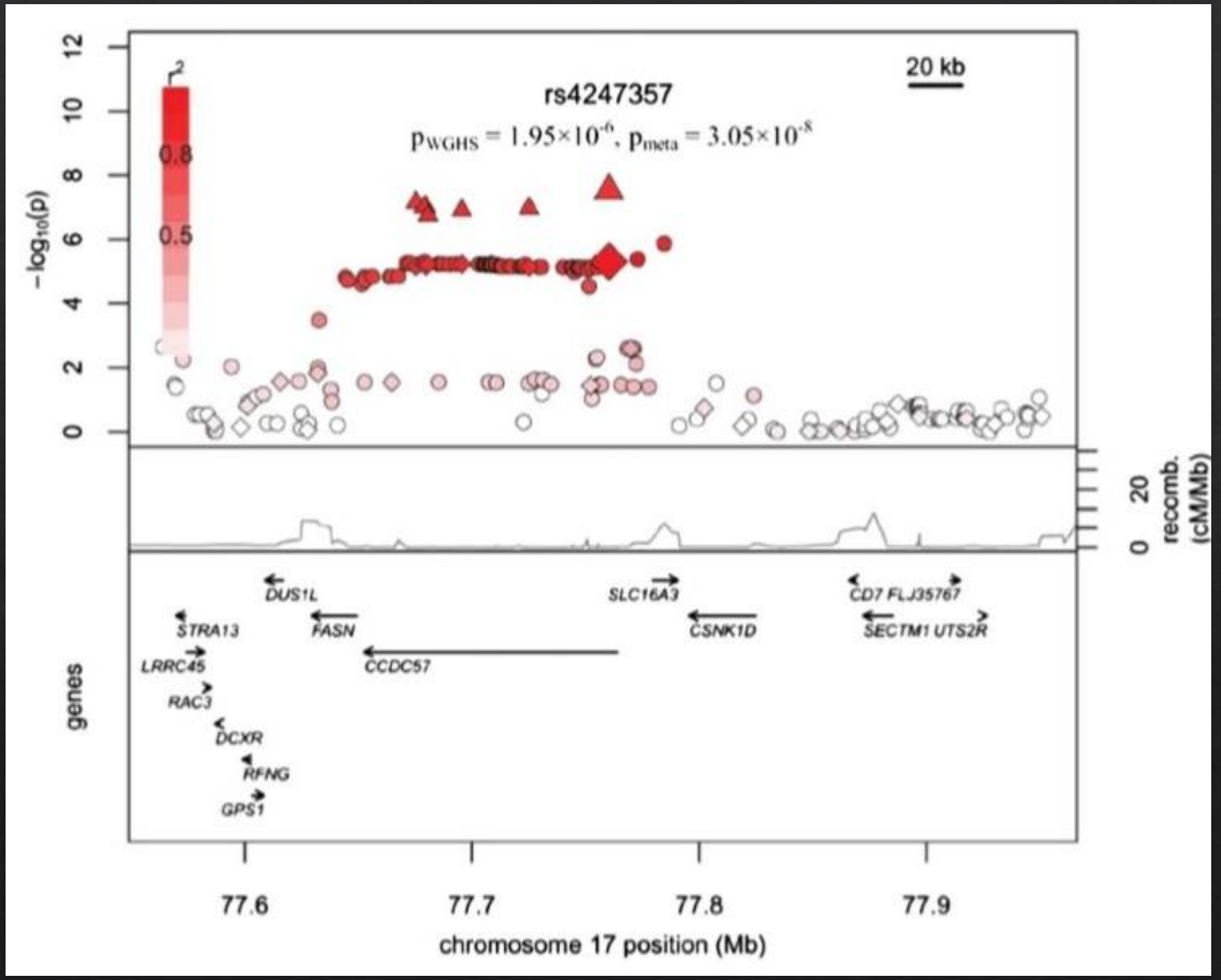
Why is LD analysis important to finding gene targets to UL risk?

LD Plot Showing Genes Spanning Chromosomes 11 and 22



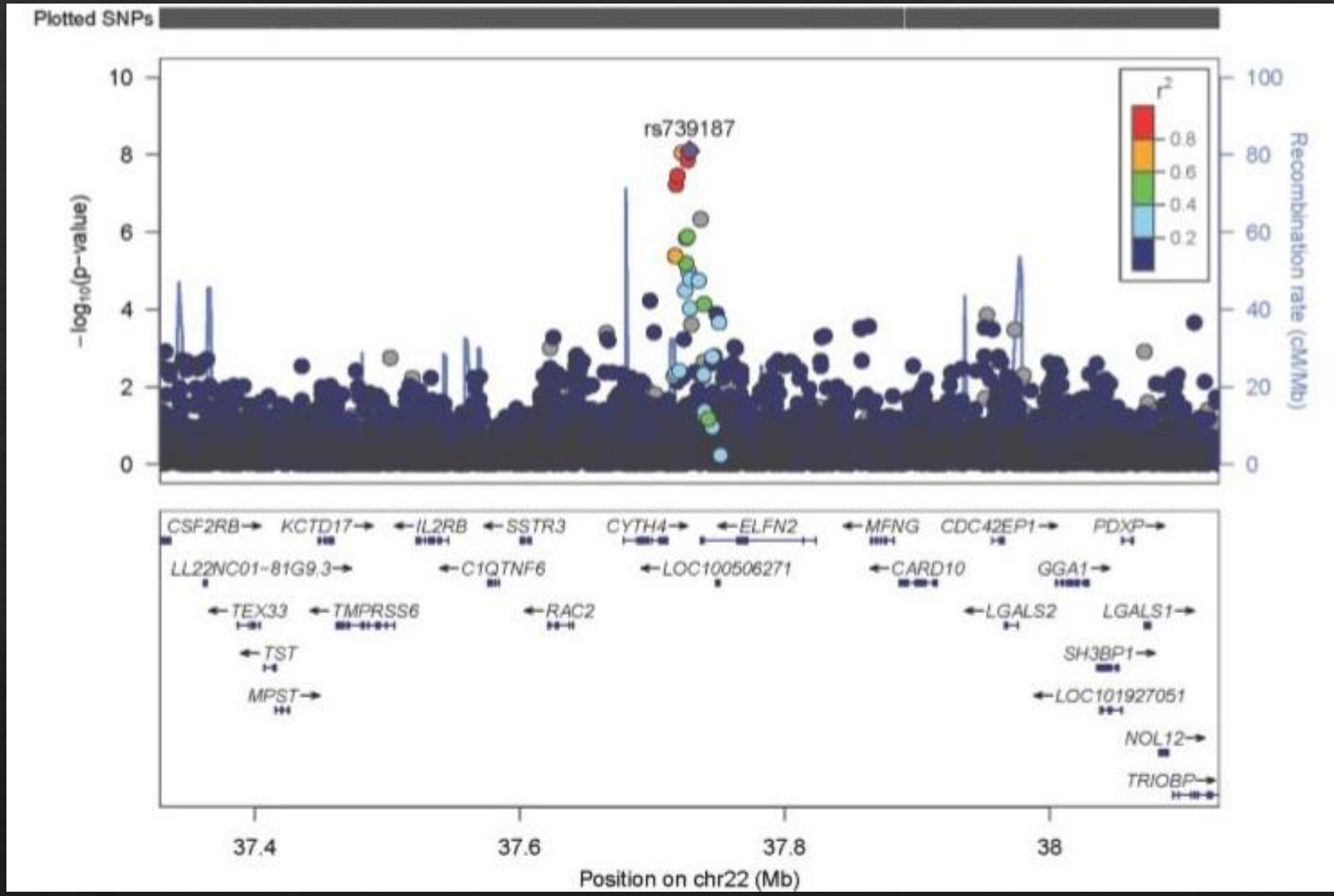
Cha et al. (2011)
Nature Genetics

LD plot for FASN along same Span as CCDC57 gene



Eggert et al. (2012)
American Journal of Human Genetics

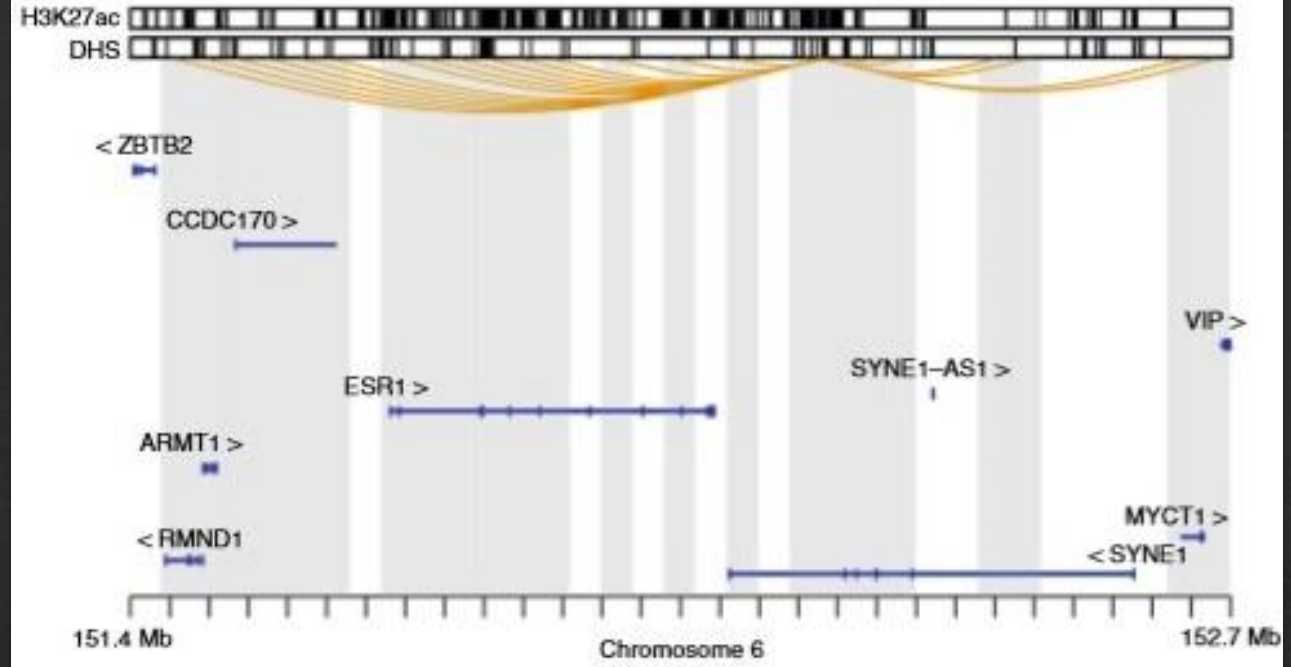
LD Plot Spanning Chromosome 22

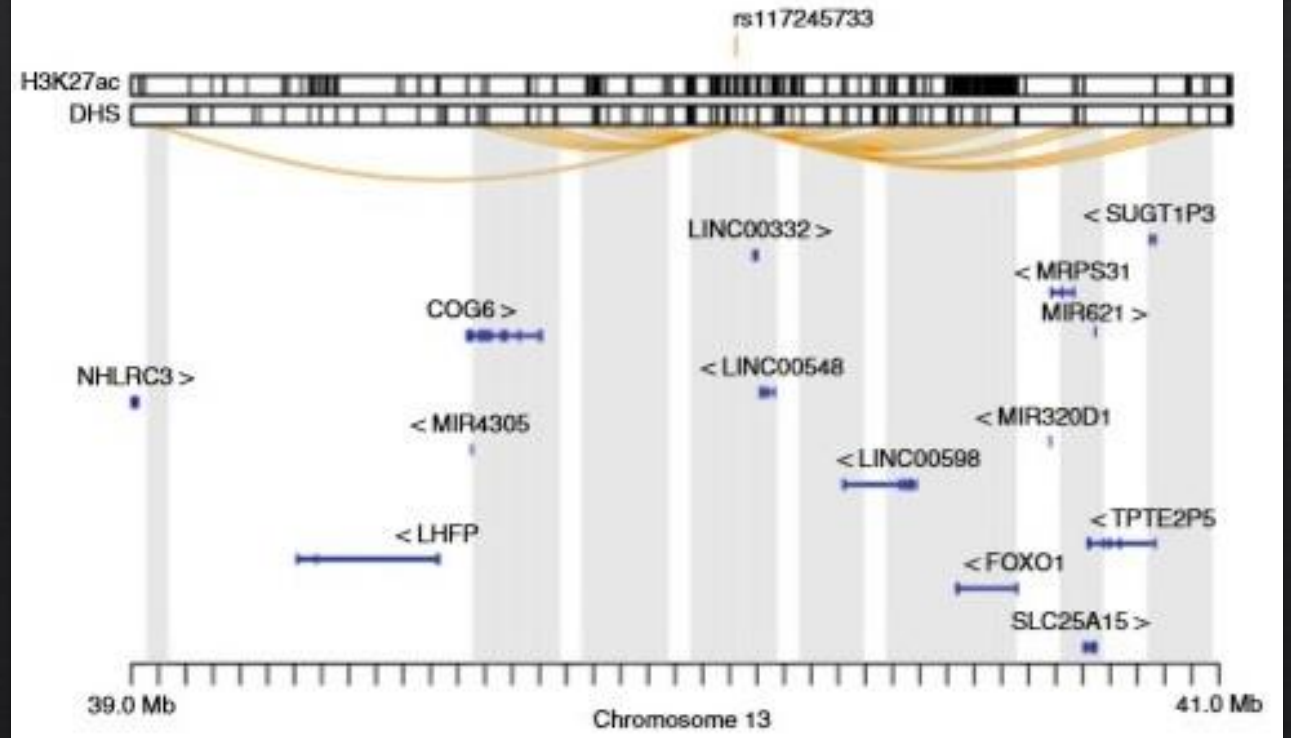


Hellwege, et al. (2017) *Human Genetics*

LD Plot Showing Chromosomes 6 and 13

Rafnar et al. (2018) *Nature Communications*





Methods – Data Sets Derived

- **The data sets that were derived to test the best machine learning results for UL prediction were as follows:**
  - The universe of all genes:
    - universe16_fold_results
    - universe16_DE_most_results
    - universe16_DE_least_results
  - The subset of 130 genes belonging only to same chromosomes (11,12,17, 22) as the six genes ubiquitous to UL risk studies:
    - TOP16_results
    - DE16_most_130_results
    - DE16_least_130_results
    - FOLD16_130_results
    - majority_10_results

Methods – Machine Learning Algorithms

- **The data sets that were derived were then each tested using seven machine learning algorithms in R, and a combined model**
  - Latent Dirichlet Allocation method of caret package in R (LDA)
  - Random Forest method of caret package in R (RF)
  - Generalized Boosted Regression Models method of caret package in R (GBM)
  - Random Forest package in R called randomForest (RF2)
  - K Nearest Neighbor (KNN) method of caret package in R
  - Recursive Partitioning and Regression Trees (Rpart) from the rpart package in R
  - Generalized Linear Regression Model (GLM) from the MASS package in R
  - Combined Model using the 'gam' method in caret package of R on all seven algorithms above

Results: TOP16; Top 10 Differentially Expressed (DE) and the 6 Genes Ubiquitous to Current UL Risk Studies

META-ANALYSIS OF UBIQUITOUS GENES TO UL RISK,

Table 3. TOP16 with the Full Gene Name

	genes	GENE_NAME
1	ARHGDIA	Rho GDP dissociation inhibitor (GDI) alpha
2	BET1L	blocked early in transport 1 homolog (S. cerevisiae)-like
3	CANT1	calcium activated nucleotidase 1
4	CARD10	caspase recruitment domain family, member 10
5	CCDC57	coiled-coil domain containing 57
6	CYTH4	cytohesin 4
7	FASN	fatty acid synthase
8	FSCN2	fascin homolog 2, actin-bundling protein, retinal (Strongylocentrotus purpuratus)
9	GRIP1	glutamate receptor interacting protein 1
10	HMGA2	high mobility group AT-hook 2
11	IRF7	interferon regulatory factor 7
12	NOL12	nucleolar protein 12
13	RNH1	ribonuclease/angiogenin inhibitor 1
14	SLC25A10	solute carrier family 25 (mitochondrial carrier; dicarboxylate transporter), member 10
15	SLC38A10	solute carrier family 38, member 10
16	TNRC6B	trinucleotide repeat containing 6B

Results: TOP16; 10,000 Simulated Bootstrap Histograms for Each TOP16 Gene

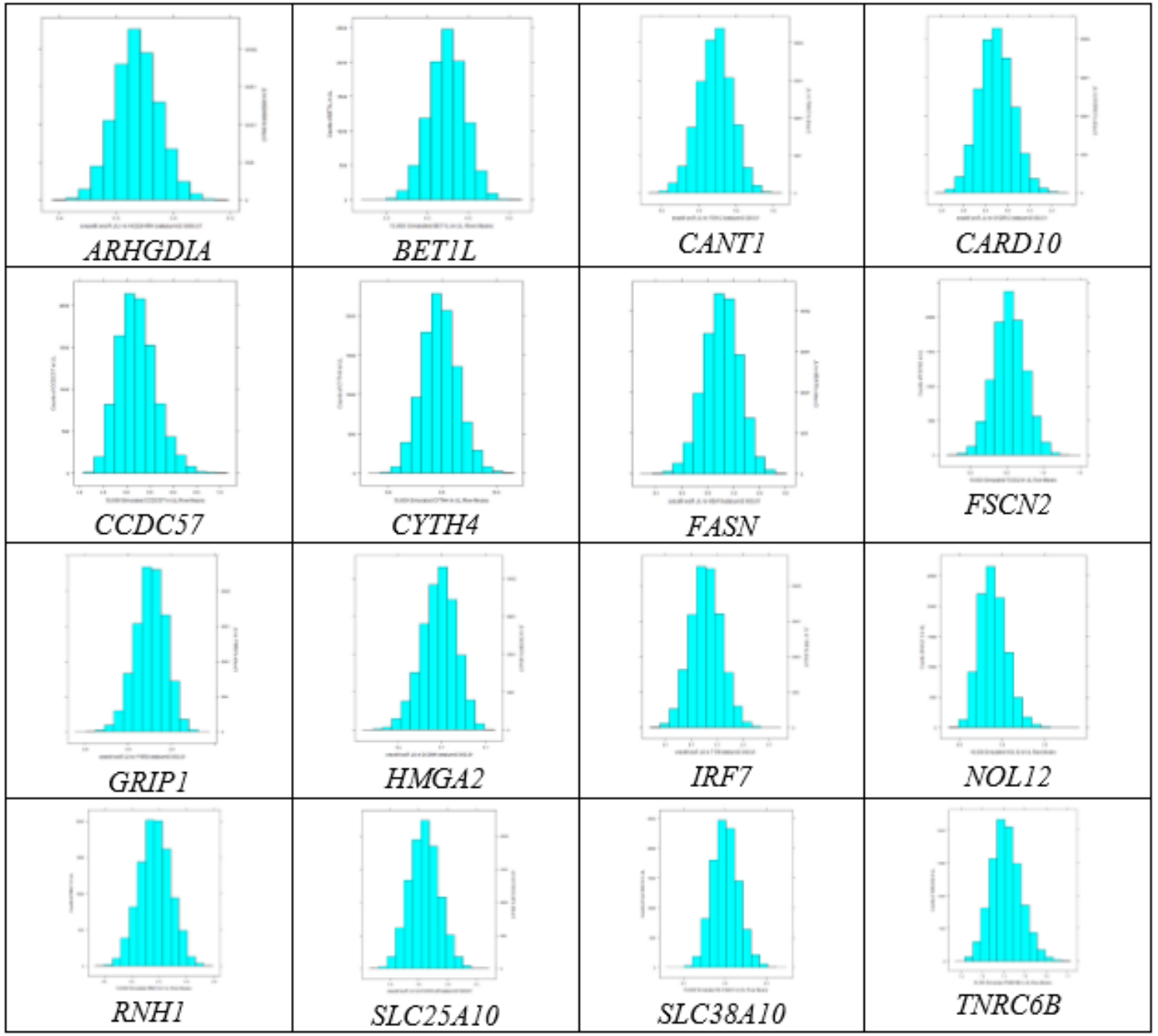


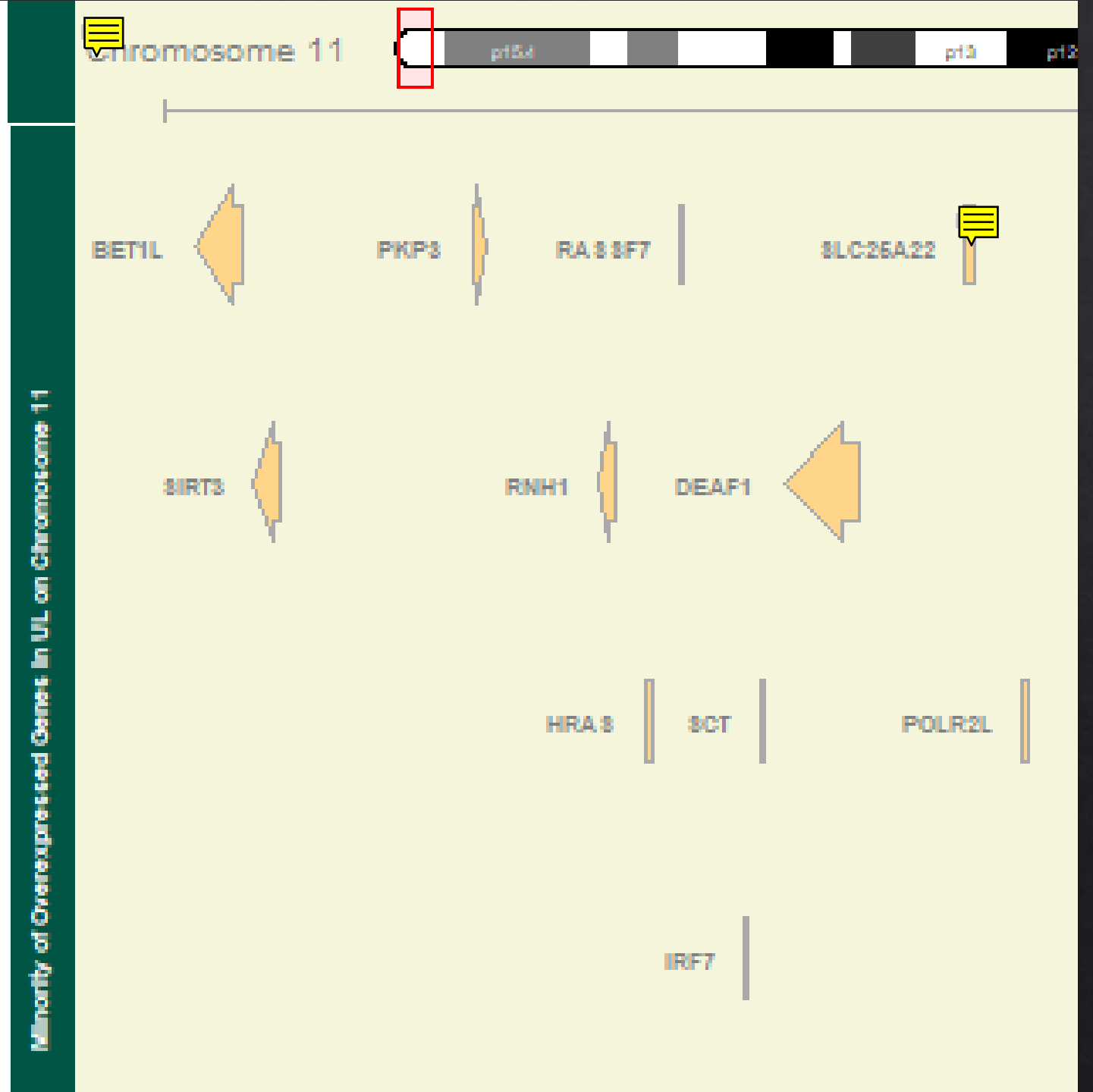
Figure 8. Histogram of TOP16 Simulated Means of 10K Samplings per Gene

Results: Linkage Disequilibrium Visualizations

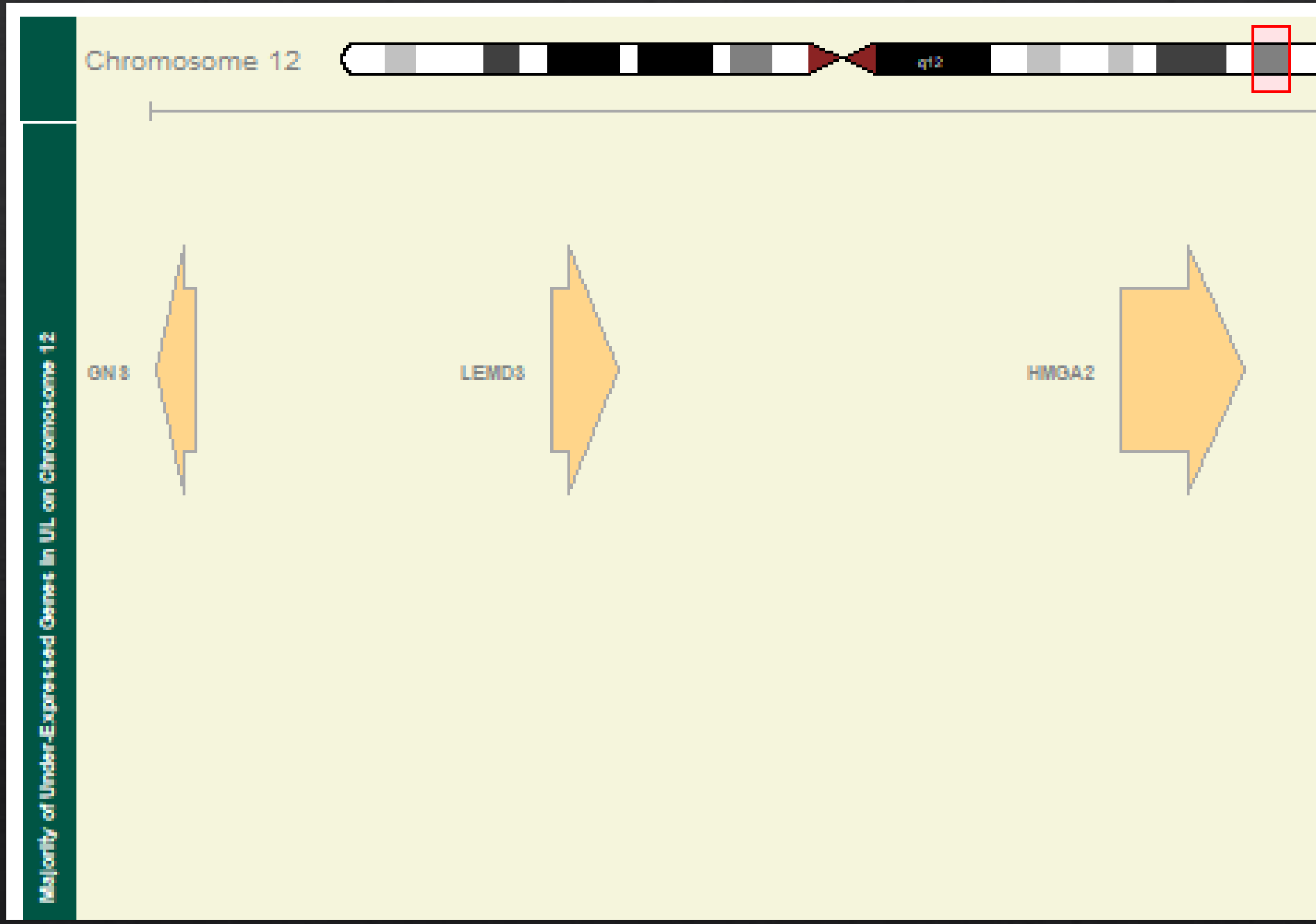
Gviz, a Bioconductor R package was used to visually see where peak links between gene regions could lie on the chromosome.

The next four slides are graphic in nature to where these genes live in the chromosomal cytobands and the links between neighboring genes.

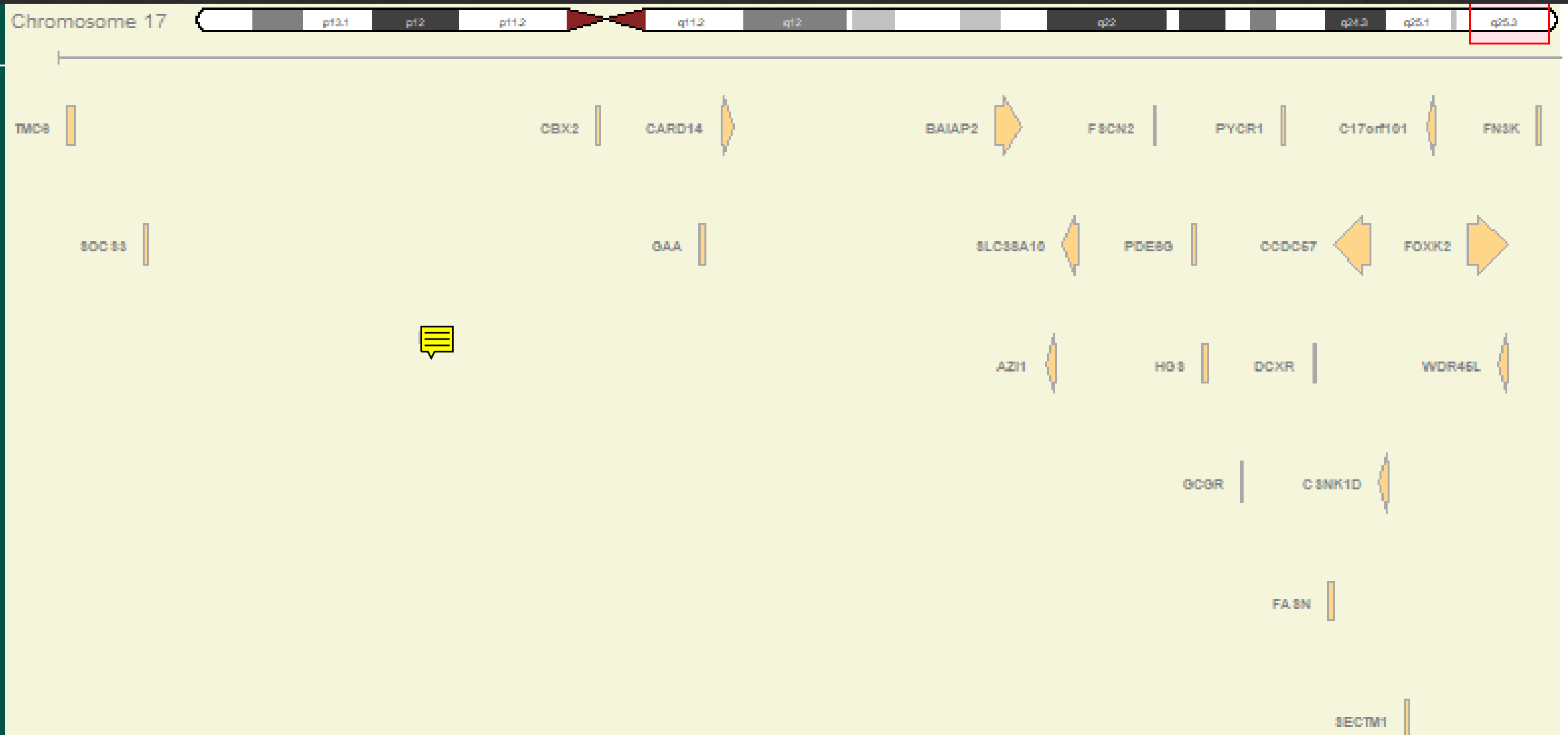
Gviz Map of Genes Over-expressed as a Minority of Genes on Cytoband p15.5 of Chromosome 11



Gviz Map of Genes Over-expressed as a Majority of Genes on Cytoband q14.3 of Chromosome 12

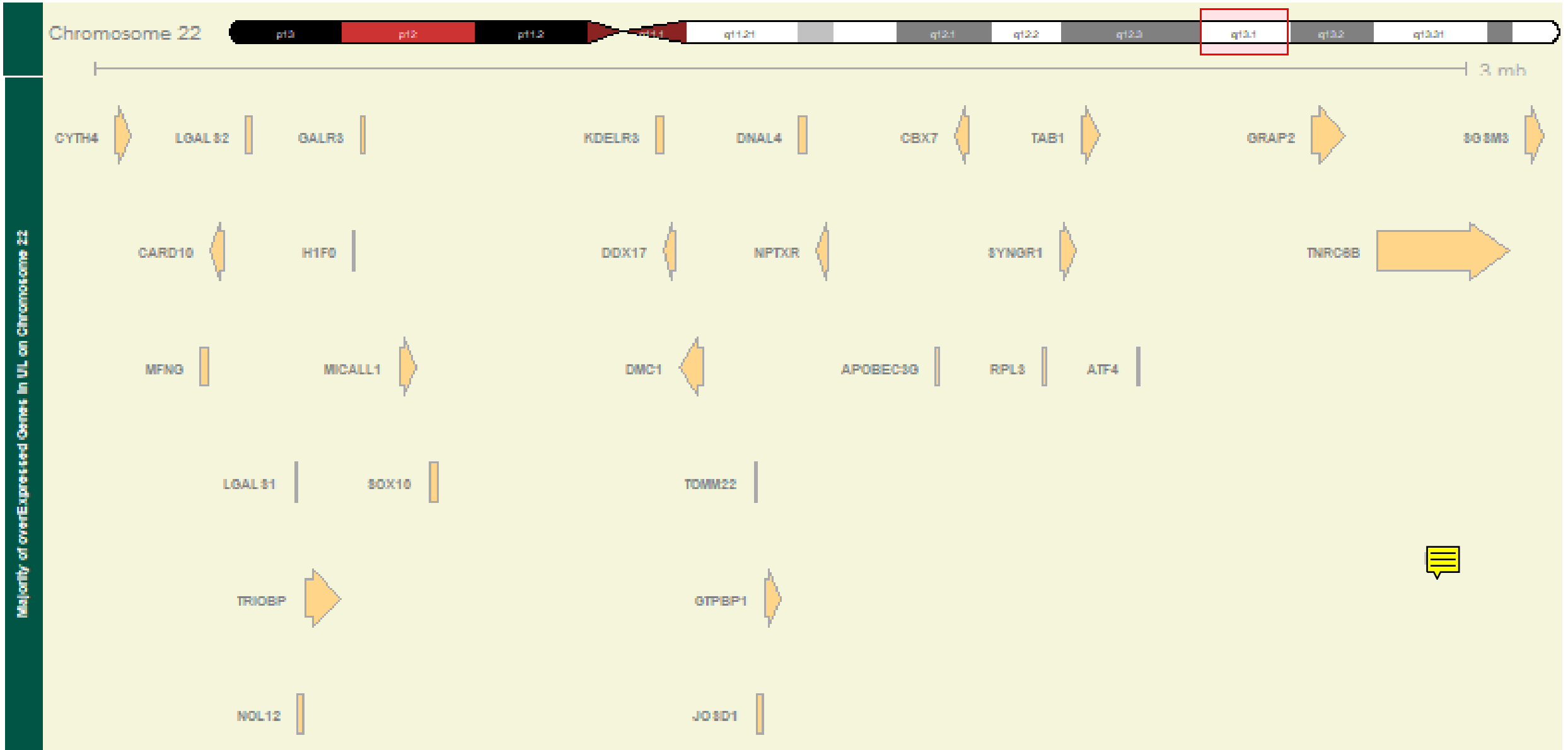


Gviz Map of Genes Under-expressed as a Minority of Genes on Cytoband q25.3 of Chromosome 17

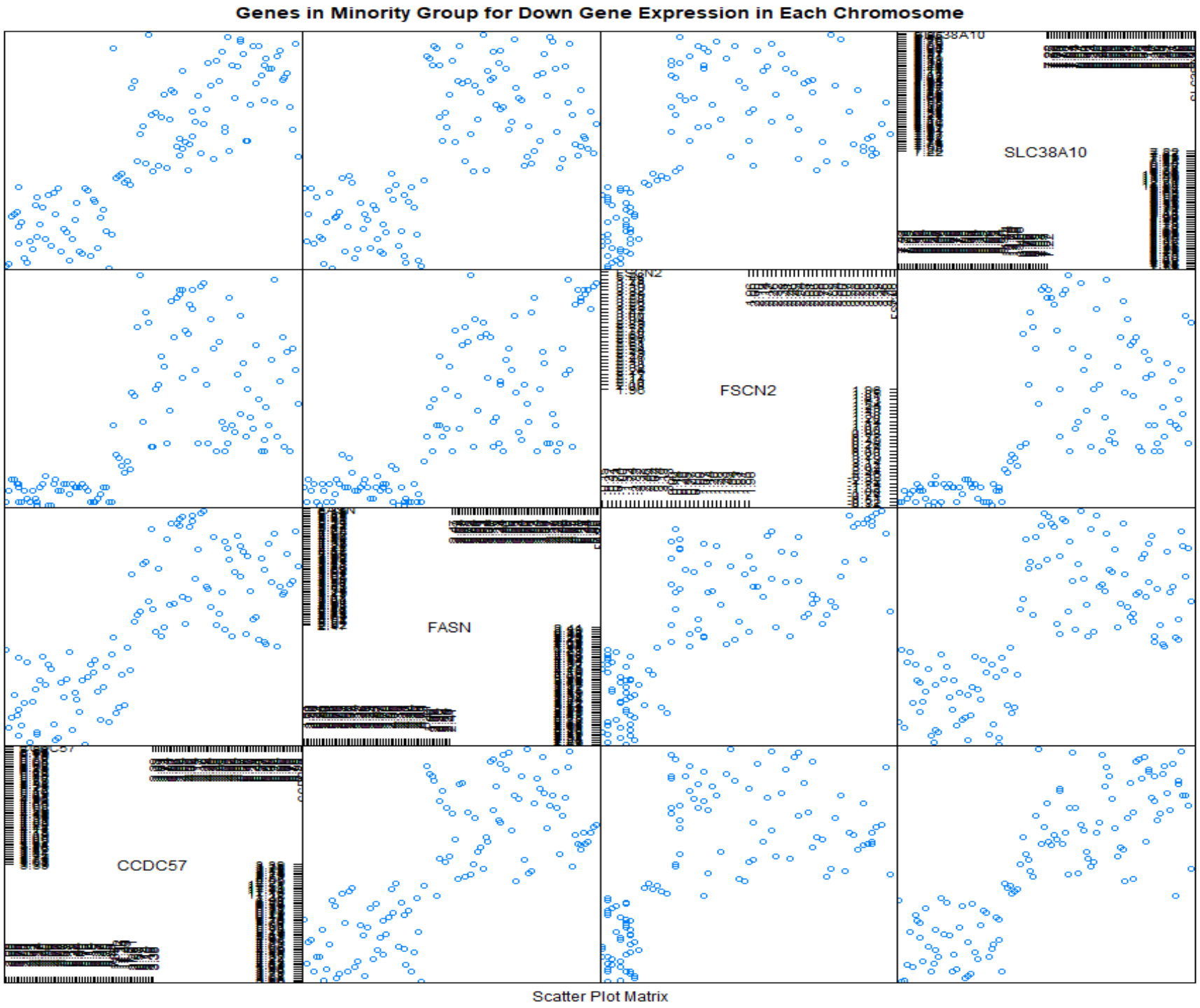


Minority of UnderExpressed Genes in UL on Chromosome 17

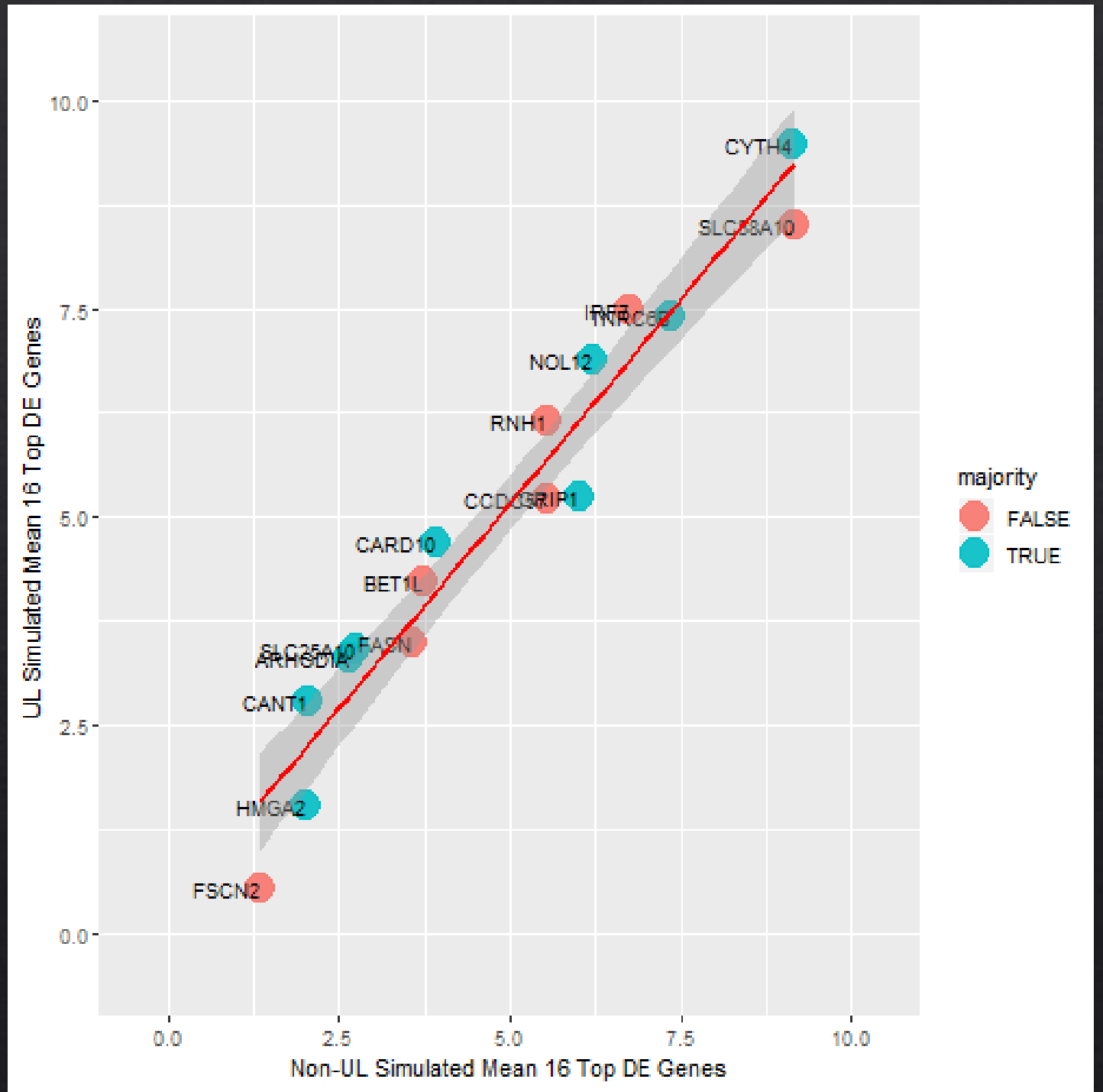
Gviz Map of Genes Under-expressed as a Minority of Genes on Cytoband q25.3 of Chromosome 22



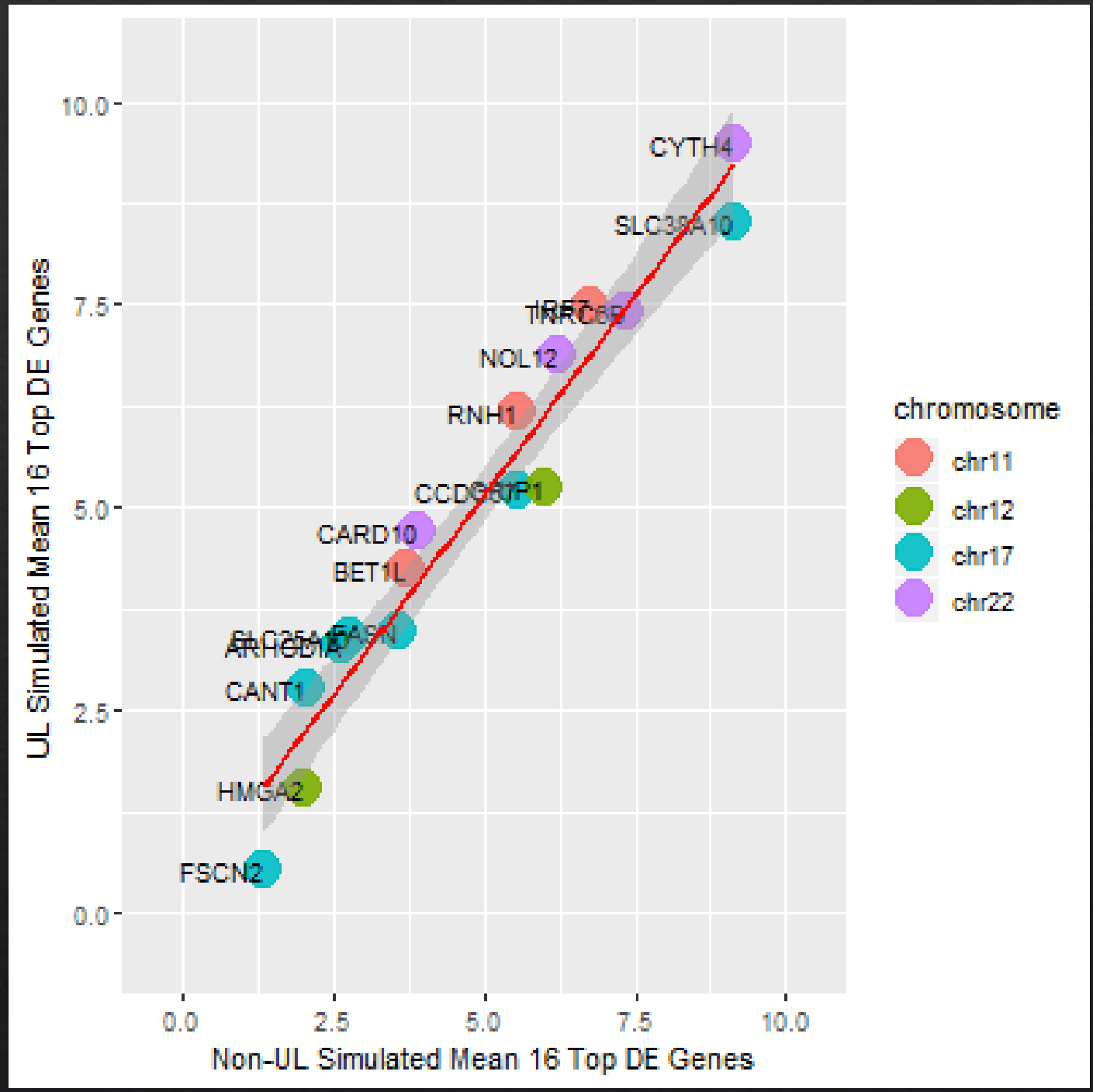
Results: A Lattice Pairwise Comparison of Chromosome 17 Minority of Genes Expressed Least



Results: Plot in ggplot2 of TOP16 Genes as Majority or Not When Comparing UL to Non-UL Simulated Means



Results: Plot Using ggplot2 showing TOP16 Genes and the Chromosome Each Lives Comparing Simulated UL and Non-UL Means of Each Gene



Results: Machine Learning TOP16 Outcomes

Table5. Table of the Combined Model

	predRF	*predRF2*	*predlda*	*predGbm*	*predKNN*	*predRPART*	*predGLM*	*CombinedPredictions2*	*TYPE*
gsm1667145	UL	UL	nonUL	UL	UL	UL	nonUL	nonUL	nonUL
gsm336254	nonUL	nonUL	nonUL	nonUL	UL	nonUL	nonUL	nonUL	nonUL
gsm336258	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL
gsm336260	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL
gsm336270	nonUL	UL	nonUL	nonUL	nonUL	UL	nonUL	nonUL	nonUL
gsm336273	nonUL	nonUL	nonUL	nonUL	nonUL	UL	nonUL	nonUL	nonUL
gsm336276	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL
gsm52662	nonUL	nonUL	nonUL	nonUL	nonUL	UL	nonUL	nonUL	nonUL
gsm52663	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL
gsm52665	UL	UL	nonUL	UL	nonUL	UL	nonUL	nonUL	nonUL
gsm52667	UL	UL	nonUL	UL	nonUL	UL	nonUL	nonUL	nonUL
gsm52669	UL	nonUL	nonUL	nonUL	nonUL	UL	nonUL	UL	nonUL
gsm9099	UL	UL	nonUL	UL	nonUL	UL	nonUL	nonUL	nonUL
gsm569425	nonUL	nonUL	nonUL	nonUL	nonUL	UL	nonUL	nonUL	nonUL
gsm569427	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL
gsm336202ul	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL	UL
gsm336208ul	UL	UL	UL	UL	UL	UL	UL	UL	UL
gsm336209ul	UL	UL	UL	UL	UL	UL	UL	UL	UL
gsm336214ul	UL	UL	UL	UL	UL	UL	UL	UL	UL
gsm336215ul	UL	UL	UL	UL	UL	UL	UL	UL	UL
gsm336218ul	UL	UL	UL	UL	UL	UL	UL	UL	UL
gsm336220ul	nonUL	nonUL	UL	UL	UL	nonUL	UL	UL	UL
gsm336229ul	UL	UL	UL	UL	UL	nonUL	UL	UL	UL
gsm336232ul	UL	UL	UL	UL	UL	nonUL	UL	UL	UL
gsm336234ul	UL	UL	UL	UL	nonUL	UL	UL	UL	UL
gsm336238ul	UL	UL	nonUL	nonUL	UL	UL	nonUL	UL	UL
gsm336239ul	UL	UL	nonUL	nonUL	UL	UL	nonUL	UL	UL
gsm336240ul	UL	UL	UL	UL	UL	UL	UL	UL	UL
gsm336241ul	nonUL	nonUL	nonUL	UL	nonUL	nonUL	UL	UL	UL
gsm336245ul	nonUL	nonUL	UL	nonUL	UL	nonUL	nonUL	UL	UL
gsm336248ul	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL	nonUL	UL
gsm38689ul	UL	nonUL	nonUL	UL	nonUL	UL	nonUL	nonUL	UL
gsm38692ul	nonUL	nonUL	nonUL	nonUL	nonUL	UL	nonUL	nonUL	UL
gsm9094ul	UL	UL	nonUL	nonUL	nonUL	UL	nonUL	UL	UL
gsm569429ul	UL	UL	nonUL	UL	nonUL	UL	nonUL	nonUL	UL
results	0.69	0.66	0.74	0.69	0.71	0.54	0.74	0.83	100

Results: Machine Learning All Data Set Outcomes

Table 6. Various Data Set Results from Machine Learning Algorithms

	predRF	predRF2	predlda	predGbm	predKNN	predRPART	predGLM	CombinedPredictions2	TYPE
TOP16_results	0.69	0.66	0.74	0.69	0.71	0.54	0.74	0.83	100
DE16_most_130_results	0.69	0.67	0.78	0.67	0.92	0.61	0.72	0.92	100
DE16_least_130_results	0.47	0.53	0.42	0.47	0.50	0.58	0.33	0.72	100
FOLD16_130_results	0.64	0.69	0.67	0.72	0.69	0.58	0.72	0.78	100
majority_10_results	0.72	0.69	0.81	0.58	0.81	0.67	0.81	0.83	100
universe16_fold_results	0.86	0.86	0.86	0.86	0.86	0.86	0.72	0.92	100
universe16_DE_most_results	0.67	0.64	0.81	0.69	0.72	0.53	0.78	0.86	100
universe16_DE_least_results	0.31	0.33	0.33	0.47	0.56	0.58	0.42	0.75	100

Conclusions

- Five studies from GEO using microarray gene expression data was analyzed to see if the six genes ubiquitous to current UL risk studies make good predictors of UL
- Eight data sets were built and used to test whether the six genes were good predictors of UL or if the machine learning algorithms were better on the most expressed genes in the samples in fold change or DE
- Results showed 83 per cent accuracy on the six genes and 10 most DE genes in TOP16
- Results also showed 92 per cent accuracy for the most expressed genes in DE or fold change out of all genes the five studies had in common
- Results also showed the least expressed genes out of all genes in DE and fold change made the worst predictors
- This could mean there is a need to keep looking at the most expressed genes and to keep the six genes ubiquitous to UL risk studies as gene targets for UL pathogenesis

Conclusions

- Results also showed 92 per cent accuracy for the most expressed genes in DE or fold change out of all genes the five studies had in common
- Results also showed the least expressed genes out of all genes in DE and fold change made the worst predictors
- This could mean there is a need to keep looking at the most expressed genes and to keep the six genes ubiquitous to UL risk studies as gene targets for UL pathogenesis

Conclusions

- Implications of this study:
  - This could mean there is a need to keep looking at the most expressed genes and to keep the six genes ubiquitous to UL risk studies as gene targets for UL pathogenesis

- Limitations of this study:
  - Only one data set had exon information to screen for genotypes and gene variants, so this method of finding the gene targets in UL by genotype variations was not used

- Future extensions to this study:
  - Test the most differentially expressed genes and highest fold change genes and see if any linkage or cytoband locus of genes could indicate UL pathogenesis other UL risk studies didn't find

Remaining Questions

- It is apparent that gene expression data was able to predict up to 92 per cent accuracy of a sample being UL or not.
  - What genes can be included along other chromosomes to indicate UL pathogenesis?
  - What specific role does each gene play in UL pathogenesis if any when expressed less or more in UL compared to non-UL samples?

References

- Aissani, B., Zhang, K., and Wiener, H. (2015). Evaluation of GWAS candidate susceptibility loci for uterine leiomyoma in the multi-ethnic NIEHS uterine fibroid study. *Frontiers in Genetics*, 6, 241. DOI:10.3389/fgene.2015.00241
- Bondagji, N., Morad, F., Al-Nefaei, A., Khan, I., Elango, R., Abdullah, L., ..., Shaik, N. (2017). Replication of GWAS loci revealed the moderate effect of TNRC6B locus on susceptibility of Saudi women to develop uterine leiomyomas. *Journal of Obstetrics and Gynaecology*, 43(2):330-338. DOI:10.1111/jog.13217
- Cha, P, Takahashi, A., Hosono, N., Low, S., Kamatani, N., Kubo, M., & Nakamura, Y. (2011) A genome-wide association study identifies three loci associated with susceptibility to uterine fibroids. *Nature Genetics* 43(5).
- Dvorská1, 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.

References

- Edwards, T., Hartmann, K., & Edwards, D. (2013). Variants in BET1L and TNRC6B associate with increasing fibroid volume and fibroid type among European Americans. *Human Genetics*, 132(12). DOI:10.1007/s00439-013-1340-1
- Eggert, S., Huyck, K., Somasundaram, P., Kavalla, R., Stewart, E., Lu, A., … Morton, C. (2012) Genome-wide linkage and association analyses implicate FASN in predisposition to uterine leiomyomata. *American Journal of Human Genetics*, 91(4), 621–628. DOI: 10.1016/j.ajhg.2012.08.009
- Hellwege, et al. (2017) A multi-stage genome-wide association study of uterine fibroids in African Americans. *Human Genetics*, 136(10), 1363–1373. DOI:10.1007/s00439-017-1836-1
- Hodge, J.C., Kim, T., Dreyfuss, J.M., Somasundaram, P., Christacos, N.C., Rouselle, M., … Morton, C.C. (2012). Expression profiling of uterine leiomyomata cytogenetic subgroups reveals distinct signatures in matched myometrium: transcriptional profiling of the t (12;14) and evidence in support of predisposing genetic heterogeneity. *Human Molecular Genetics*, 21, 102312–2329. DOI:10.1093/hmg/dds051

References

- Liu, B., Wang, T., Jiang, J., Li, M., Ma, W., Wu, H., & Zhou, Q. (2018). Association of BET1L and TNRC6B with uterine leiomyoma risk and its relevant clinical features in Han Chinese population. *Scientific Reports*, 8,7401. DOI:10.1038/s41598-018-25792-z
- Rafnar et al. (2018) 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
- Rye, Wise, Jurukovski, DeSai, Choi, & Avissair (2017) 'Eukaryotic Epigenetic Gene Regulation.' *Biology: OpenStax*. Retrieved from: https://cnx.org/contents/jVCgr5SL@15.43:5cz8bfb2@10/16-3-Eukaryotic-Epigenetic-Gene-Regulation

Comment Summary

Page 2

1. The questions that you have presented here represent important information to share with your audience. However, this format seems to suggest that you plan to provide a lot of information along with just this single slide, and there is not much to help keep the audience engaged while you are speaking. Please revise this -- I suggest that you spread this information over several slides, and provide supporting images to help the audience follow what you communicate orally. For instance, the image labeled "fibroid locations" on this site (https://www.mayoclinic.org/diseases-conditions/uterine-fibroids/symptoms-causes/syc-20354288) would make it much easier for someone with limited knowledge of the anatomy to understand where the tumors can be located.

Page 3

2. On the next slide you have a nice image to help represent the geographic part of this information -- rather than describing everything here and then flipping to the image, I suggest you reorder a bit. Change the image from the next slide to only show the top three (you can change by cropping), and then along with those maps, list the genes -- but for the text, perhaps summarize the function of the gene product, rather than re-writing info about location that is conveyed through the map. Then on the second slide, give the bottom row of the maps and information about those genes.

Page 4

3. What is the source for this image? Also, make sure that the aspect ratio has not been changed -- it looks like it might have been stretched up and down a bit.

Page 5

4. Remember, it is important to explain the more general concept before giving more specific information. On the following slide, you included an image to help explain the general concepts of transcription and translation. Since that is the most general idea, it would make sense to present that first. Two slides after this, you pose some questions about how transcription is altered, to which this slide seems to be at least a partial answer. Thus, fix the ordering, likely to have the transcription/translation overview, then the text slide that poses questions, and then this slide last.

Page 8

5. As written, this seems to be a summary of your conclusion, rather than an open ended question. Here, make clear:
   - What is the overall question you wanted to address (that had not currently been tackled in the literature)?
   - How do you propose to address?
   - What can be done with this new information?

   I also suggest that the ideas are conveyed with phrases, rather than a block of text. Avoid using abbreviations (such as DE) that are not easy to interpret.

Page 9

6. Please do not explain all methods at once -- the reality is that if you provide a full listing of the procedural information disconnected from the data itself, your audience will have forgotten the necessary information by the time you display your results. Rather, please group your relevant information by subject. For example, first explain information relevant to the data -- make clear the sets you started with as well as the working set(s) you generated. Then explain the first analysis method, followed by presentation of related results. Next, explain the second analysis approach, followed by the related results. Repeat as necessary.

Page 11

7. This is important background information. You should share this with your audience before you state your particular research question.

Page 12

8. This is helpful to include with your background information about LD. Reorganize to before stating your research objective.

   Also, make sure that the resolution of the image is not blurry.

Page 13

9. Move to background section.

Page 14

10. Move to background section.

Page 15

11. Move to background section.

Page 17

12. Rather than listing all together, state each and show relevant results.

Page 18

13. Please update the format of this table, such that you take the contents and format them into a nice table (rather than just taking an image capture of the output from your analysis).
14. Immediately prior to showing these results, explain the methods needed to generate this list of differentially expressed genes.

Page 19

15. After you have presented the TOP16, make sure to (at least verbally) provide a transition to explain why this analysis was performed next (give a rationale for why you completed this analysis). Follow the rationale with an explanation of the methods used to make this happen on a slide immediately before this one, and then present the results, explaining their meaning.

Page 20

16. After the presentation of the previous data:
    1) Make clear conclusions from that analysis.
    2) Explain the rationale for next completing the LD analysis.
    3) Provide the methods for this analysis
    4) Provide the results
17. In your explanation, it will be important to verbally link together this information, but you do not need to provide this in text on a slide.

Page 21

18. Make sure that the resolution of this image is adequate.
19. Why do some have arrows while others do not?
20. When will you explain this concept of "minority of genes"?
21. What is the relationship between this information and understanding changes in gene expression?

Page 22

22. What do you mean by "majority"? You will need to make sure that this concept is explained to your audience.

Page 23

23. Why do some components have arrows (presumably marking the orientation on plus or minus strand) while others do not?

Page 24

24. Like other similar previous slides, make sure to address minority/majority; resolution of the image; arrow versus non-arrow components.

Page 25

25. Before showing these results, you need to make clear:
    1) The rationale for why this analysis was performed.
    2) The method used to complete this analysis
    Then
    3) Show the results
    4) Explain how knowing this information contributes to your overall objective (what was the big take-away? what were you able to learn about UL?)

Page 26

26. Before giving the results, immediately prior:
    1) What was the rationale for this analysis?
    2) What methods were used?
    Then
    3) What does this presentation of the data actually tell you?
    4) How does this help reach your objective/provide a new insight about UL?

Page 27

27. See comments on previous slide.

Page 28

28. Make sure to explain methods immediately before showing these results. You will need to make clear what the different terms all mean (especially the different rows, as labeled in the first column).

Page 29

29. Explain methods immediately before results. How is this table different from the previous? What new information do you know at this point, and how does that help with your overall objective?

Page 30

30. Here, you have not made any comments about chromosome location. Therefore, what was the utility of

including that information in your presentation?

Page 32

31. Your study has been defined as using differences in gene expression to learn more about UL, correct? Since that is the case, you should not really give any information about SNPs (genotype information).
    Do not provide "limitations" corresponding to analysis you did not perform -- instead, you must be a critical evaluator of the work you DID perform.
32. Didn't you filter your data set to only focus on chromosomes that contained genes previously implicated? Why did you do this? Why didn't you look at ALL genes?

Page 33

33. You seem to have assumed that chromosome location has some impact on gene expression -- this is not inherently the case. Can you justify this assumption?