
Edited power point presentation

5 messages

corona, janis <janislcorona@lewisu.edu>
To: "Dr. Sarah E Powers - powerssa@lewisu.edu" <powerssa@lewisu.edu>


Tue, Jul 16, 2019 at 4:24 AM

Hi Professor Sarah Powers.

I made some edits to the original slide using my better judgement of what you suggested in your feedback. I am attaching it to this email. Could you please look it over and tell me what you think?

Thank,

Janis

 **Research_Presentation_UL_Risk_Genes-version3.pptx**
2945K

Powers, Dr. Sarah E. <powerssa@lewisu.edu>
To: "corona, janis I." <JanisLCorona@lewisu.edu>

Tue, Jul 16, 2019 at 11:42 AM

Janis,

In reviewing the slides you have provided, especially in reading the notes you provided on each slide, I think that it is important to correct some misconceptions you seem to have about the biological concepts. It will be very important that the information you present this week is factually accurate.

Slide 9: Here your notes seem to be with an explanation of translation, but transcription is not addressed first. I suggest that you provide an overview of how information stored at the DNA level is first transcribed into pre-messenger RNA. This pre-mRNA must then be processed into the mature mRNA form before it leaves the nucleus and moves to the cytoplasm where translation happens (and proteins are synthesized using the "blueprint" provided by the mRNA). Within your notes, you also provide a list of concepts: "cis-acting elements, promoters of genes, repressors of genes, 5' to 3' translation region, recoil by RNA binding proteins upstream, downstream, along coding region, millions of base pairs away, external stimuli effects gene expression." As provided (and considering other information present within your slides), I am not confident that you fully understand how these components fit into the process of gene expression. Here is an overview:

* On a eukaryotic chromosome, there are many different genes. When there is the need to have the information stored within a specific gene on the DNA strand expressed, the cell is able to access that gene (without accidental spread to other flanking genes) because of precise accessibility. This decision making for which gene(s) should be transcribed in a cell at any given point in time is controlled by other proteins that can take action within the nucleus. These proteins – known as transcription factors – will interact with other places on the DNA strand in order to act as activators and recruit the machinery needed to make mRNA. Sometimes these transcription factors will bind to the promoter region to trigger transcription. Other times, there will be adjacent places along the length of the chromosome (known as cis-elements, such as an enhancer region) that are responsible for regulating expression of that particular gene – these are the components that can be many base pairs away, but they are always specifically charged with regulating the same gene. In other instances, genes are kept in the inactive state because other proteins, known as repressors, will associate with the gene promoter/enhancer and make sure that the machinery needed for transcription to take place cannot access the gene.

* If the mechanism to trigger expression of a gene is reliant on transcription factors that work as

activators, when the transcription factor can bind to that region of the DNA, it will then be in an “open” state that means transcription can happen. If the mechanism to trigger expression is reliant on release from a repressor, when the repressor disengages from the DNA, it can then be in an “open” state.

- * “Open” DNA becomes less tightly associated with histones, and that means that the transcription machinery can associate with the DNA. The double helix is temporarily released from its double strand configuration, as it passes through the RNA polymerase, which reads the template half of the DNA molecule, and synthesizes the complementary pre-mRNA, building this new molecule in the 5' → 3' direction. As the RNA polymerase moves down the length of the gene, the portion of DNA it has already passed re-spools into the double stranded configuration. When RNA polymerase reaches the end of the gene region, the synthesis of the pre-mRNA is stopped (it does not flow into adjacent genes), and then the pre-mRNA can be processed into the mRNA.

- * The control of transcription (through alteration in the balance of activators and repressors) is often dependent on external cues. Thus, if the cell receives the information that it needs to alter its gene expression, an activator can be made and then go on to induce the transcription of the appropriate gene. In some instances, an activator may be responsible for activating a number of genes, but the association of the activator with the DNA is dependent upon the sequence in the DNA (relative to the gene). This decision is not dependent on chromosomal coordinates. The same concept (but in reverse) applies to repressors. If a repressor is deactivated, it will stop associating with the DNA where it has been located, via a specific DNA sequence. This applies based on function of the repressor, rather than chromosomal location. Thus, there may be two genes very close to each other on the same chromosome. If one gene needs to be transcribed, the cell will have synthesized the necessary activator, and it will associate with the specific DNA sequence (promotor or enhancer) for that gene and transcription will happen. The gene located physically next to it is not necessarily under control of that transcription factor activator, though, and so it will not be transcribed. Only if both genes have activating factors associate with the DNA will they both be actively transcribed.

Slide 10:

- * Cells will take cues from many different external factors. Sometimes the external stressors you indicate may lead to a change in gene expression. Other times, access to nutrients, time of day/year, stage in development (inputs that are not necessarily stressors) will change expression.

- * It is not accurate to state “gene expression of RNA.” Gene expression is the process by which the gene at the DNA level is then converted to mRNA, and, during the translation process, tRNA and rRNA are used to synthesize the polypeptide.

- * Gene expression can be impacted at several levels:

- * Is the actively transcribed? This is a yes or no.

- * How stable is the mRNA once it is made?

- * Some mRNAs can be regulated by other types of RNA that essentially make the mRNA not available for translation

- * The duration of how long the mRNA is used for translation can be regulated by the poly-A tail length

- * How stable is the resulting protein?

- * Some proteins have a long half-life, other have a short half-life

- * Proteins can be regulated, so that even if they are present in the cell, they may not be active

Slide 11: On this slide, there are several different concepts that you have merged together – I want to make sure that you understand them all.

- * In each of the cells of our body (aside from the eggs/sperm that are used through fertilization), there are actually two copies of each of the 23 different chromosomes. We are diploid organisms, so we have two copies of each of the autosomes (numbered 1-22) and a pair of sex chromosomes (either XX or XY). That is a large amount of DNA to fit into the nucleus, so the DNA is condensed by wrapping around histones, and then these nucleosomes are condensed into chromatin fiber (this is seen in the center of the diagram).

- * When a gene needs to be accessed for expression, this wrapping process needs to be relaxed so that the transcription machinery is able to interact with the DNA molecule (this is regulated by changes in the way the DNA associates with the histones, as shown at the bottom of the diagram).

- * If a cell receives an instruction to prepare for a round of cell division, there is also a loosening process which will allow for the DNA molecule to be copied. At the conclusion of this DNA synthesis, the two exact copies of the same DNA molecule are recoiled but held together, awaiting a separation process

that makes sure that one copy of each chromosome is partitioned into each of the cells at the conclusion of cell division (the chromosome shown in the top left of the diagram represents this chromosome with the two identical copies, or sister chromatids, held together at that center of the “X”).

* When a cell is instructed to prepare for a round of cell division, it is true that there are genes, necessary for creating proteins essential for the cell division process, that have their gene expression changed (they are induced). It is only in the presence of these factors that the cell will have the ability to undergo the division process. This is not the same as gene expression in general; the process by which genes that are not involved in cell division is regulated is independent from DNA duplication.

Slide 12: For the process of gene regulation, it is true that regions of the DNA upstream or downstream of the gene (but not right at the promoter) can be important for the regulation of how that gene is expressed. Because of the size of the gene itself, this long scale regulation of the gene can be a large number of base pairs away from the transcriptional start site. However, in comparison to the large scale of the entire length of the chromosome, this is still pretty close to the coding part of the DNA itself (for instance, the gene is not at one end of the linear chromosome, with a regulatory section at the complete opposite end of the linear molecule). Thus, this regulatory aspect is distinct from the other concept you raise on this slide (linkage disequilibrium), which has more to do with genetic deviations seen across members of a population. In order to address linkage disequilibrium, it is important to cover some additional information about the way the members of a chromosome pair (the two copies of the same chromosome) behave as that information is passed from one generation to another.

Let's first look at an image that represents a single chromosome in humans (we can pretend that it is chromosome 1).

[Image result for allele]

If this is representative of chromosome 1 in you, you inherited one copy from your biological father, and the other from your biological mother. For the instance where your body needs to then generate a gamete (in human females the egg and in males the sperm), a process called meiosis is used to separate out the two members of a pair such that the resulting gamete is haploid (n), or only has one copy of each of the different chromosomes. That way, when egg and sperm come together through fertilization, the offspring will once again have the diploid ($2n$) number of chromosomes. So if you have a baby, one half of that baby's genetic material will be inherited from you and the other half from the other parent. As labeled, a gene will be located at the same physical location on each chromosome within the pair (these are called homologous chromosomes), which may or may not be exactly the same sequence (alleles are versions of genes). This is where it is then possible to assign a genotype to an individual, based on whether or not these alleles are the same.

[Image result for allele]

If we are considering genetic differences at a specific position along the length of the chromosome, sometimes these are within the coding region of the gene and then might be different alleles. Other times, there might be a single nucleotide polymorphism (SNP) which varies the sequence, but is not associated with a different gene product. SNPs can sometimes fall within a gene, or it is possible that they fall in a location that is between genes. If they do fall within a gene, they may be associated with an allele difference, but this is not always the case, as it has to do with the translation of the nucleotide code into the amino acid code. Remember, three letters from the nucleotide code (once it has been transcribed into the mRNA working copy) is interpreted at the level as the codon to know what amino acid should be incorporated into the amino acid chain as the protein is synthesized. There is a lot of redundancy within that “rosetta stone” of codon > amino acid (you might want to look up an example of that table as a reminder). In some cases, if the nucleotide code is changed (there is a SNP) within a codon, the new three letters might be interpreted as the information for a different amino acid at that position – that change would then likely change the way the protein works, and that would be considered a new allele. Other times, there may be a change in the three letters in the codon, but that may be translated as exactly the same amino acid. For example, the codon CCU instructs building the amino acid proline (pro) into the peptide chain. A SNP that changes the sequence so that the mRNA has a different codon could perhaps result in a new sequence of CCC (which also instructs the incorporation of proline – a new allele would not be created) or CUU (which would change the amino acid to leucine (leu) which would make a new protein sequence and likely give a new allele).

When a cell is doing its typical day to day functions (and not preparing for a round of division), the genetic content is essentially like represented in these previous images, and using transcriptional regulation processes, the cell decides what genes should be expressed at any point in time. Although we may think

about the overall context in which the gene is present along the length of the chromosome, the rationale for when/when not a gene is expressed (DNA information is turned into a protein) is not reliant upon the expression of the gene “next door.” It is important, though, to think about the orientation through which different alleles or SNPs are passed along together to the next generation – and this is dependent upon the spatial location of the DNA within the genome. To understand this part, it is important to explain cell division using mitosis (making more diploid body cells) or meiosis (making haploid reproductive cells – egg or sperm).

[Related image]

In this figure, there are two chromosomes within the homologous pair, and each chromosome represents DNA that has been duplicated in preparation for cell division. The want is to maintain order such that the resulting cells wind up with the correct assortment of chromosomes (and doesn't wind up with too many copies of chromosome 4 but none of chromosome 7, for instance) – the chromosome on the left (where the top portion looks to curve off to the left a bit) is actually one chromosome with two sister chromatids (the DNA copies) attached to each other (this would represent the paternal chromosome). The chromosome on the right (where the top bends a bit to the right) represents the two sister chromatids from the maternal chromosome. When the cell prepares to divide for the purpose of mitosis, it is every chromosome for itself – the only objective is to make sure that the two sister chromatids are successfully pulled apart so that the same amount of DNA is passed into each of the resulting cells. Although for assessment purposes chromosomes at this stage could be captured in a laboratory setting, imaged and then pictures of the chromosomes are lined up next to each other analysis software (this type of preparation known as a karyotype is useful for diagnosing chromosome abnormalities), diploid cells of the body never worry about actually pairing up homologous chromosomes prior to dividing. This is different, though in specialized cells that have the task of making eggs and sperm, and they use meiosis instead. Once the DNA is duplicated, and the sister chromatids are held together, at the start of the division process, the homologous chromosomes actually do pair up similar to what is shown in the figure above. At this time, a special process, known as crossing over, takes place. If you look at the above figure, think about labeling the portions of the chromosomes from left to right. In the left chromosome, label the left-most chromatid (the one that has “telomere” labeled) as A, the sister chromatid it is attached to as B. In the right-most chromosome, label the sister chromatid on which you have added the blue circle as C, and the sister chromatid it is attached to (that is right-most of all) as D. When these two members of the homologous pair have lined up like this, the inside chromatids from the two chromosomes (B and C) can participate in crossing over. This means that the internal chromosomes can swap arms, and so change the context of genetic information as recombinant chromosomes are generated. This image might help:

[cid:image004.jpg@01D53BDC.5C7B5C50]

So the big takeaway is that, when you inherited one copy of chromosome 1 from your biological father, it might be in the exact configuration that he inherited from one his parents (such as the solid blue chromosome), or it might have been a recombination of some genetic information from his mother plus his father (such as seen in the red/blue). The same thing would have happened with the chromosome you inherited from your biological mother. This is the genetic basis for why some people may be surprised with results from DNA analysis tests (23andMe or Ancestry). Let me apply an example to me, based on family history. The entirety of my Dad's family tree traces back to Ireland. On my maternal side, my mom's mom and her family was from Poland, and my mom's dad and his family were from Sweden. Assume that we really over simplify politics and moving borders and such that these countries and their populations are distinct (this is a gross simplification of reality, but let's pretend for this example). Based on this, I might make the prediction that 50% of my DNA would trace to Irish origins, 25% to Polish and 25% to Swedish. If there was no crossing over between homologous chromosomes when my mom's body made the egg that led to me, then this expectation would make sense. However, we know that crossing over seems to happen on average 1-3 times per homologous pair, and so the chromosomes that my mom passed to me, might not have been 50% Polish and 50% Swedish, but actually 20% Polish and 80% Swedish.

So now, if we think about different alleles or SNPs as each a genotype that should be considered individually, then this crossing over part doesn't really matter – there is a 50% chance of each version (in a heterozygote) being passed along to the next generation, and the context of that allele/SNP on the chromosome doesn't matter. For the sake of simplicity, let me limit this example to just considering SNPs. What if you really need to have 3 SNPs in a row inherited together to change the phenotype? If an individual is homozygous for these three SNPs, then even if a crossing over event were to happen between these SNPs (in the image above in the daughter cell at the bottom that is the second, if you count from the left, where the chromosome is red-blue-red) – say that SNP1 (G|G) falls in the top red region and SNP2 (A|A) and SNP3 (G|G) fall in the blue region – in a homozygote that was used to start this meiosis process,

even if crossing over did lead to this recombinant chromosome, the combo results would still have SNP1 give G, SNP2 give A, and SNP3 give G. What if we started with a heterozygote, though. For SNP1, maybe the blue chromosome has A, red has T; SNP2, blue has G, red has A; SNP3 blue has A and red has G. Then in the four daughter cells in the bottom row of the diagram, the configurations of SNPs would be:
First cell: (1) A (2) G (3) A
Second cell: (1) T (2) G (3) A
Third cell: (1) A (2) A (3) G
Fourth cell: (1) T (2) A (3) G

This illustrates the ideas of linkage. If the consequence of the way a cell functions is because of just a single location along the arm of a chromosome relative to a single gene on a different chromosome (outside of the homologous chromosome pair), you would expect no linkage (there is no association between the two entities). However, if elements (whether genes or SNPs) are located very close to each other on a single chromosome, there is an increased likelihood that versions will be inherited in the same configuration as they were inherited from the previous generation – that is, unless a crossing over event happens between these two factors on the same chromosome arm, they will be transmitted together. This clustering together is then what is known as linkage disequilibrium. If your data set were investigating differences in a population looking at SNPs located on a chromosome near a previously characterized gene associated with UL, then it would make sense to include this type of consideration in your project. However, you are working with a data set that is about gene expression. Because changes in gene expression are not typically regulated because of the more long distance distribution on the chromosome that impact linkage disequilibrium, this concept of LD does not really apply to changes in gene expression.

Thus, while it might be possible that there are similar trends in expression of genes that happen to be within a cytoband, that cannot be explained by LD. You will need to update the objective information on slide 13 accordingly.

Slide 16: This helps with understanding the initial parsing of genes (the top three levels), but the criterion for establishing majority and minority are still not clear. Please address this.

Slides 17-20: I understand that the arrows represent forward or reverse strand. If you are going to provide this information on the plot, then it needs to be given for all genes, or none. There needs to be consistency for all parts that are presented. By having a third format (no arrow) it implies that the directionality is not known.

Slide 21: What is the rationale for using 4 genes? I still do not understand the majority versus minority classification.

Slide 22: What was the rationale for this evaluation? What were you hoping to learn using this approach?

Slide 25: Why did you decide to move forward with the top 10 most differentially regulated? Why not 2? Why not 30?

Slides 28-29: What was the purpose of completing the bootstrap simulation? Why did you elect to use this approach, rather than another analysis? This is important for subsequent slides. You will need to provide a rationale for why you used this approach for any related results as well.

Slide 36: You need to explain why each analysis was used. Your notes seem to have clarified the names of each algorithm. The important part to present, though, is an explanation of why it was appropriate to use this method. What were you hoping to learn? Why would this model be the best approach? Part of being successful in data science is knowing what analysis approach is most appropriate. Some tests will not provide information that is biologically relevant/interpretable. You need to be able to demonstrate that you understand when it is appropriate to apply certain methods, rather than just applying a laundry list of analyses to see what happens.

Given the feedback here, I think that you will need to re-evaluate your approach to then consider whether your conclusions are still appropriate or not. I have not provided a review of the conclusion slides, as your final take away conclusions will be pending revisions made earlier in the presentation.

Sarah E. Powers, Ph.D.
 Assistant Professor
 Department of Biology
 Lewis University
 Romeoville, IL
 815.588.7079

[Quoted text hidden]

4 attachments

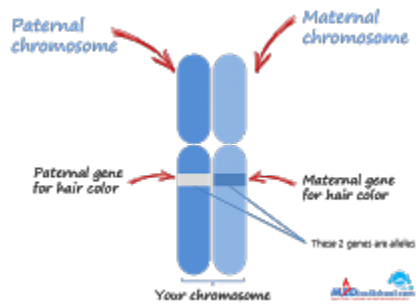


image001.png
 172K

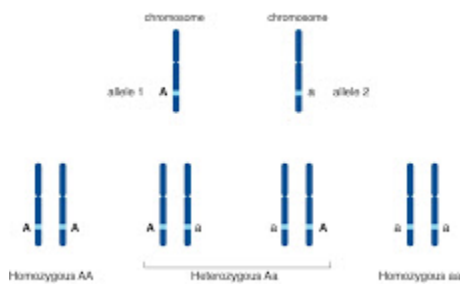


image002.jpg
 47K



image003.jpg
 43K

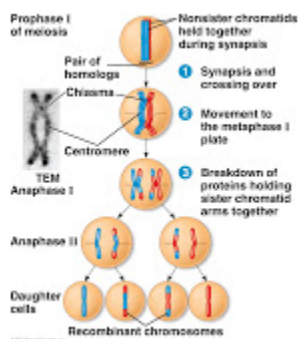


image004.jpg
 202K

Thank you for that thoughtful response and feedback of changes needed. I generalized what I know about transcription and translation. There was the same info you just wrote in the previous email in the Biology OpenStax text book that I summarized to link the LD of using the plots. I am going to remove the LD reference, genotype references or SNP references and the chromosomal locations. Although, the majority of gene expressions for the 5 least and most was a set of good predictors for LD. In data science these algorithms were used in a Coursera course, the data mining analytics course, and a statistical programming course. These genes and samples are numeric values so it makes sense to cluster similar results or to make regressions linear or non linear to determine or find a relation in gene expression. Thanks for clarifying a whole rundown of your course content of Intro to Computational biology. I think it is too long to include in those slides for my presentation.

I thought I could build a connection to how external stresses in the cell during transcription can inhibit or express genes to be duplicated in the cytoplasm through translation. As the LD only relates to recombination values for SNPs then it cannot be connected to gene location next to neighboring genes that are along those same peaks. This was the main approach for targeting these UL risk genes in six studies of the nine referenced.

In data science courses involving numeric data such as the gene expression data my research focused on examining for these genes and other gene targets that could associate with UL pathogenesis, are analyzed with regression and making a good fit of a small sample represent the population at large. Given the samples did have more than 40 each of UL and nonUL to simulate a population mean, this is why bootstrap simulations were done on those 16 genes. To show how well the Central Limit Theorem and Law of Large Numbers fits the counts of the mean samples to build a 95% confidence interval of each gene's true value of expression in UL compared to nonUL. The genes for the TOP16 fit well. But some were only slightly skewed, meaning the results could hold true in the population of all samples when comparing UL to non-UL gene expression data. The number 10 was chosen randomly for whose top ten genes showing the most change, then adding six to it of the six genes having UL risk association. Then to keep ten as a number for the majority of genes expressed on each cytoband. This was an out of the box way of grouping genes, as it would be similar to finding outliers that skew numeric data by seeing what gene expressions fall outside the normal range for changes in UL compared to non-UL samples. Data science is not an exact science and neither is producing visualizations. It is applying analysis with methods to evaluate the data to find patterns and relationships that answer questions. The research is on determining if the top genes proven UL risk genes and the top expressed genes by fold change or magnitude of differential change can be used to find gene targets to UL pathogenesis using already established algorithms in machine learning. I answered the question using data science methods. The algorithms are not able to show exact gene per gene classification in the results of R. Since the heatmaps were not informative and the dendrograms super cluttered, they were excluded.

How does the use of the chromatids and allele selection you reviewed above add to my research. They were excluded because it dealt with gene expression and stresses affect gene expression throughout the phases of transcription to translation of genes. I think my summaries of how gene expression changes sounds the same as your all encompassing review and it is not that important to putting into my research if it isn't building on genes in the cytoband that could also be targets to UL pathogenesis. Am I expected to give a full summary of the entire process of transcription and translation to show why or how gene expression changes? Wouldn't I need to include reasons for why gene expression doesn't work or how SNPs are involved when you said not to include them in my research as they aren't the same as gene expression and my research didn't use SNPs. I think it is off topic to put the entire description you put in the previous email into my research presentation.

If the genes like CCDC57 and TNRC6B are on the same forward strand of the cytoplasm of chromosome 22, are you saying that none of the genes between these two that have UL risk significance can be looked at?

Thanks,

Janis

[Quoted text hidden]

Powers, Dr. Sarah E. <powerssa@lewisu.edu>
To: "corona, janis I." <JanisLCorona@lewisu.edu>

Wed, Jul 17, 2019 at 7:20 AM

Please see below for responses to what you provided in your previous email, as it seems best to respond to your comments paragraph by paragraph.

Sarah E. Powers, Ph.D.
Assistant Professor
Department of Biology
Lewis University
Romeoville, IL
815.588.7079

From: "corona, janis" <janislc corona@lewisu.edu>
Date: Tuesday, July 16, 2019 at 6:42 PM
To: "Powers, Dr. Sarah E." <powerssa@lewisu.edu>
Subject: Re: Edited power point presentation

Thank you for that thoughtful response and feedback of changes needed. I generalized what I know about transcription and translation. There was the same info you just wrote in the previous email in the Biology OpenStax text book that I summarized to link the LD of using the plots. I am going to remove the LD reference, genotype references or SNP references and the chromosomal locations.

In my last email, I provided more extensive information about the process of transcription and translation because, as I read the notes you provided on the PPT slides, there were some places where the notes were either not fully clear, or I read something that contained an inaccuracy. My attempt was to make sure that you had an explanation of the theory that was correct so that, for any content that you do included in your presentation, you have adequate information to present the biological concepts correctly.

Although, the majority of gene expressions for the 5 least and most was a set of good predictors for LD.

If you were to look at gene expression changes between UL and non-UL samples for all genes assayed, and determined the 5 genes with the largest increase in UL as well as the largest decrease in UL, and then determined the chromosomal location of these genes, then perhaps it might be true that some of these genes are located on the same chromosome as the top 6 genes that have been identified by previous research. If this were the case, then it would make sense to take the research further to determine if there are genetic variants within these genes that perhaps are inherited together more times than not, which would prompt needing to evaluate LD for these. From reading your work, though, it seems that your approach was different, as you initially focused on genes located only on the same chromosomes as the previously identified top 6, which meant that gene expression signatures from genes located on all other chromosomes were eliminated from the data set. If I have understood this procedure correctly, you need to realize that this approach has introduced bias into your analysis. While models built using these genes may still give some information about expression patterns that are associated with disease, it is different from taking a non-biased approach through which all gene expression information has been considered (there may be genes located on other chromosomes that are even better predictors for disease development).

In data science these algorithms were used in a Coursera course, the data mining analytics course, and a statistical programming course. These genes and samples are numeric values so it makes sense to cluster similar results or to make regressions linear or non linear to determine or find a relation in gene expression.

A key objective of this course and developing your own research project is to make sure that you are able to connect concepts between biology theory and data science. There are some algorithms included in your work that I agree have been used by others in order to interpret large sets of biological data. What I have wanted to emphasize is that, through your presentation and your writing, you need to explain this. Please make clear to the audience how the analysis treats the data, the power of what is generated by the analysis, and an interpretation of what your results mean, in the context of the biological data.

Thanks for clarifying a whole rundown of your course content of Intro to Computational biology. I think it is too long to include in those slides for my presentation.

My intention was not to expect you to provide an entire overview of genetics as part of your presentation. Rather, there were places where, given your notes and the connections you were making, it seemed that you needed clarification about the concepts. Please make sure to provide adequate background overview, though, so that your Computer Science reviewer has a fundamental understanding of the type of data you worked with, and interpretations that are possible from your results.

I thought I could build a connection to how external stresses in the cell during transcription can inhibit or express genes to be duplicated in the cytoplasm through translation.

I think it makes sense to comment on the fact that external cues impact the cell, which leads to the gene being transcribed and translated. Please realize, though, that duplicating a gene means that the DNA code is copied and added to the genome (this is a different issue, known as copy number variation, which in itself can impact disease onset/progression), while changes in the amount of gene expression (as measured by how many copies of the protein are made through translation) is a different concept. Stating that “genes are duplicated in the cytoplasm through translation” is not an accurate description of changes in gene expression (it is factually incorrect, from the standpoint of biological theory).

As the LD only relates to recombination values for SNPs then it cannot be connected to gene location next to neighboring genes that are along those same peaks. This was the main approach for targeting these UL risk genes in six studies of the nine referenced.

LD does assess gene location and placement relative to neighboring genes. The way the LD is assessed through the GWAS publications is that SNPs within these genes may be associated with disease (having a certain SNP, relative to a different nucleotide at that position in the gene code, will make members of the population that carry that SNP at an increased likelihood for developing that disease).

In data science courses involving numeric data such as the gene expression data my research focused on examining for these genes and other gene targets that could associate with UL pathogenesis, are analyzed with regression and making a good fit of a small sample represent the population at large. Given the samples did have more than 40 each of UL and nonUL to simulate a population mean, this is why bootstrap simulations were done on those 16 genes. To show how well the Central Limit Theorem and Law of Large Numbers fits the counts of the mean samples to build a 95% confidence interval of each gene's true value of expression in UL compared to nonUL.

This type of explanation for why you used this analysis approach has not been explained previously – it is exactly the type of rationale that should be provided within your presentation to justify each of your analytic approaches. You cannot assume that your audience understands the utility of each analysis. One of the objectives of the presentation and your paper is for you to demonstrate that you understand these rationales, and your understanding can only be assessed if you provide it.

The genes for the TOP16 fit well. But some were only slightly skewed, meaning the results could hold true in the population of all samples when comparing UL to non-UL gene expression data. The number 10 was chosen randomly for those top ten genes showing the most change, then adding six to it of the six genes having UL risk association. Then to keep ten as a number for the majority of genes expressed on each cytoband. This was an out of the box way of grouping genes, as it would be similar to finding outliers that skew numeric data by seeing what gene expressions fall outside the normal range for changes in UL compared to non-UL samples.

Ok, considering that this is a decision that you decided to make in your analysis, you need to embrace it. As part of your explanation, it is important to clearly articulate the decisions made, and disclose your process. I am not saying that this decision is fundamentally wrong, but rather that you have not communicated your approach in the previous reports/presentation materials that I have seen.

Data science is not an exact science and neither is producing visualizations. It is applying analysis with methods to evaluate the data to find patterns and relationships that answer questions. The research is on determining if the top genes proven UL risk genes and the top expressed genes by fold change or

magnitude of differential change can be used to find gene targets to UL pathogenesis using already established algorithms in machine learning. I answered the question using data science methods. The algorithms are not able to show exact gene per gene classification in the results of R. Since the heatmaps were not informative and the dendograms super cluttered, they were excluded.

I completely understand that any type of science is a process of trial and error, and I agree that there is an overall objective to look for patterns and relationships within data that can hopefully answer questions. Along the way, there will likely be many instances where the attempt simply does not work. I am not asking you to perform miracles. Instead, what I need you to do is provide a story about the process you have taken through your research. For example, you should always start with the clear question – what are you attempting to address? With that in mind, you select an analysis approach. How does this algorithm, for instance, work, and once results are generated, how will the outcomes potentially answer your question? This piece demonstrates the comprehension that you understand both the data science (what the analysis does) and the biology (what do the results mean). Then, present the results. Sometimes the outcome is messy or flat out does not give interpretable results – and that's ok. It is fine to still work with less than awesome outcomes, as it gives you an opportunity to also evaluate limitations of the approach and contemplate what would be needed to make the approach more viable. So, thinking about what you have done, let's say that you wanted to include one of the analyses with the heatmap. Many other researchers have published heatmaps in which it is in fact not possible to have exact gene names listed (honestly, full labels for genes usually are only reported if the list of compared genes is really small). With that as a limitation, say you decide to include the heatmap as a representation of this type of clustering approach. It is perfectly acceptable to eliminate the text labels that are not readable, and instead focus on whether or not it is possible to see patterns within the image itself. If there is a portion of the analysis that does look interesting, you can find an alternate way to report the specific genes that fall into that portion of the grid. The associated dendogram might also be cluttered (and that's not the end of the world). By including this outcome, though, you are still adding to the knowledge base; the interpretation might be that, given the genes searched, there is NOT an expression signature that clearly can stratify UL from non-UL. As long as you can explain that interpretation, it makes sense to include, as it helps provide a rationale for why you selected a different subsequent approach. And this is how the story gets built – you present a first step in the analysis, and use the outcome of that as a starting point for the next analysis – it justifies why you needed to attempt to answer the question in a different way. In science, there almost never will be a case where the outcome of a study is definitive – that's part of the beauty of the game. Questions that are being investigated today are only possible because they are built on observations made previously. And results from an experiment today, will inform the experiment of tomorrow. As the researcher, your job is to explain where you started, report the attempts you have made, interpret the outcomes (whether awesome or less than awesome), and then contemplate how it would make sense to move forward, given what you know now.

How does the use of the chromatids and allele selection you reviewed above add to my research.

The information I provided about chromatids and alleles is not directly needed for you to consider questions about gene expression. It would be relevant if you were working with a data set that included SNPs. The only reason that I provided that information was an attempt to explain why it is not appropriate to include LD in your contemplation of gene expression data. It seemed unfair for me to make a blanket statement about why LD and chromosome location of genes was problematic without providing an explanation.

They were excluded because it dealt with gene expression and stresses affect gene expression throughout the phases of transcription to translation of genes. I think my summaries of how gene expression changes sounds the same as your all encompassing review and it is not that important to putting into my research if it isn't building on genes in the cytoband that could also be targets to UL pathogenesis. Am I expected to give a full summary of the entire process of transcription and translation to show why or how gene expression changes?

You do not need to spend the entire time of your presentation explaining transcription and translation. Given that your data set is focused on gene expression, I think it would be appropriate to give an overview of fundamental concepts such that all human genes from the genome are contained in each cell of our body, and yet that repository does not consistently access and express all of those. By giving an overview of the concept that gene expression can be modified, it will help to explain why it was worthwhile use the data you selected to try to identify differences between UL and non-UL.

Wouldn't I need to include reasons for why gene expression doesn't work or how SNPs are involved when you said not to include them in my research as they aren't the same as gene expression and my research didn't use SNPs. I think it is off topic to put the entire description you put in the previous email into my research presentation.

My intention by providing that information in my previous email was not to write a script for your presentation. Rather, I was attempting to make sure you had adequate information to understand why the LD and SNP information was off topic, as previous emails from you indicated you did not understand my previous comments about concern that gene expression and SNPs were still intermingled.

If the genes like *CCDC57* and *TNRC6B* are on the same forward strand of the cytoplasm of chromosome 22, are you saying that none of the genes between these two that have UL risk significance can be looked at?

I am not saying that genes should be excluded based on their location on chromosome 22. My point is that, if you are assessing gene bearing on disease development based on gene expression changes, there is no set rule to state that gene expression is dependent on chromosome location. When considering whether or not gene expression has an impact on disease, this concept is fully independent of positionality within the genome.

Thanks,

Janis

[Quoted text hidden]

[Quoted text hidden]

 **winmail.dat**
48K

corona, janis <janis@corona@lewisu.edu>
To: "Powers, Dr. Sarah E." <powerssa@lewisu.edu>

Wed, Jul 17, 2019 at 10:46 AM

Thank you for breaking down that last email. I think it makes sense.

[Quoted text hidden]