Written Qualifying Exam for GMB
January 10, 2020
Woodruff Health Sciences Library Computer Lab (Room B65)

Read the entire exam before starting.

- Answer one question from each of the five (5) pairs of questions. Each question is worth a total of 20 points. The breakdown of points for each question is clearly indicated. Please ensure that your answers address each part of the question.

- Your answers should be concise (but complete) and legible.

- Start the answer for each question on a new page according to detailed directions. Do not change the font size or type.

- Label each question clearly and write your Student Letter (provided by Roberta) on the first page of each question. Clearly indicate which Question and part of a Question your answer applies to.

- The final file(s) should be saved on the USB flash drive provided.

- If writing by hand, use the notepad provided and a pen.
  Write only on one side of the page. Leave enough space in the margins so your answers are not cut off on the copy machine.
  Clearly indicate which Question and part of a Question the hand written or drawn material applies to.
  Do not detach pages from your pad; you will turn in the entire pad at the end of the exam.

- During the exam, only the word processing application may be active, with the file open in which you are typing your answers. No other programs may be used to access e-mail, internet or local files such as PowerPoint or PDFs. You are not to consult any notes or other written matter or consult with any person about your answers.

- When you are done, put all your materials (exam, notepad, paper, USB flash drive) back in the envelope provided and seal.
- Completed exams should be placed in the box by 5:00 PM. Roberta or Andrew will collect exams from the room.

The Emory University Honor Code is in effect during the duration of the exam. Please sign, date the enclosed document, and leave with your exam in the envelope.

A light breakfast (8:15 AM) and lunch (12-1 PM) will be available in the Calhoun Room, located across the library lobby. We will also have snacks and drinks available all day in the Calhoun Room.

If you have questions about the exam, call Andreas at 404-668-3753 (cell; any time before 2PM) or Ken Moberg at 404-217-7708 (cell; any time). All others questions call or text Roberta – 404-735-2129.

GOOD LUCK!

Question 1A

Cardiomyopathies are diseases characterized by mechanical and electrical dysfunction of heart muscle, with accompanying heart failure and risk of sudden death. Cardiomyopathies not due to other conditions (coronary artery disease, hypertension, viral myocarditis, alcoholism) are due to mutations in genes encoding proteins of the sarcomere, the fundamental unit of muscle contraction. Currently, at least 25 such genes are known, but these 25 gene accounts for less than half of the cases. Thus, there is a need to identify new "cardiomyopathy genes". Moreover, even for the cases in which we know the involved gene, the molecular mechanisms that lead to dysfunction of the heart muscle are unknown. It is known, however, that the vast majority of the cases are autosomal dominant.

Through exome sequencing of a large number of normal controls and cardiomyopathy patients, you have identified an amino acid variant (valine to aspartate (V->D) change) in a sarcomeric protein expressed in the heart called Kindlin2. You have a family in which the father and the son are heterozygous for this variant, and they both have cardiomyopathy. However, you would like to obtain evidence that Kindlin2 (V->D) actually causes the disease. Although *C. elegans* does not have a heart, worms have striated body wall muscle that is like human cardiac muscle and C. *elegans* is a powerful genetic system for studying the development and regulation of muscle.

Assume that all the necessary tools are available, including a specific antibody to UNC-112.

1) *C. elegans* has a single ortholog for human Kindlin2, called UNC-112, and the involved valine is conserved. Using *C. elegans*, how would you test the hypothesis that the V->D change results in muscle dysfunction?

   In your answer, explain:
   - (a) How the mutation would be generated (2 points)
   - (b) What are the two main possibilities for the phenotype you could observe (2 points)
   - (c) How you would assess muscle structure and function (2 points)

2) List two potential mechanisms for how a mutation can be dominant. (1 point).
   What genetic tests can you use to distinguish these two possibilities? (2 points)
   Explain how a missense mutation could result in one or the other mechanism. (2 points)

3) Assume that heterozygotes and homozygotes for *unc-112(V->D)* have the same muscle phenotype. Explain **two ways** in which you could conduct a suppressor screen, beginning with the homozygote, to identify loss of function mutations in other genes that suppress (improve) the muscle phenotype. (3 points)

4) Explain how you would identify the candidate suppressor gene(s) involved. (4 points)

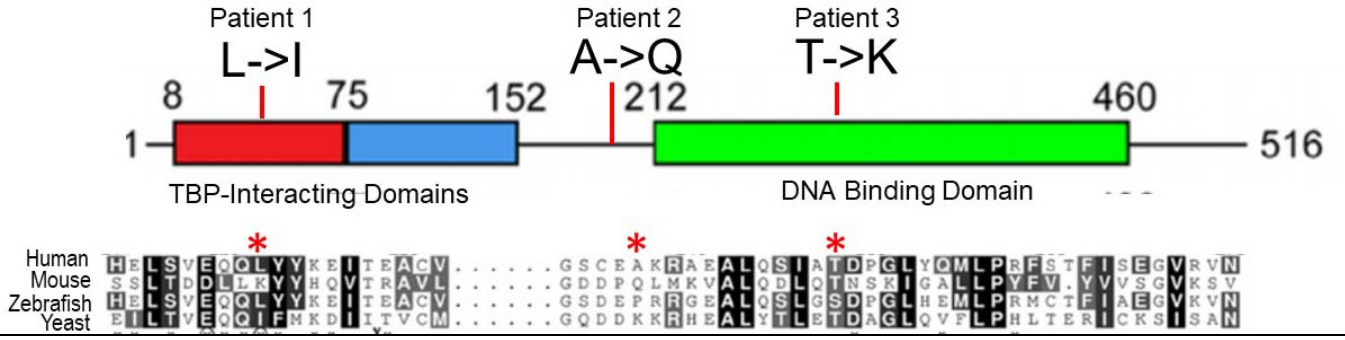
5) Once you have identified the genes, briefly explain one possibility for a mechanism of suppression. (2 points)

Question 1B

You have recently been contacted by a team of clinical investigators who has been seeking cases of an inherited form of an autosomal recessive neurodevelopmental disorder to understand the molecular basis of the disease. They have recently identified three separate families where missense mutations in a gene that is evolutionarily conserved have been linked to the disease. The patients are each homozygous for the missense mutations the researchers identified through exome sequencing. Assume for the purposes of this question that we can be confident that these missense mutations CAUSE the disease in these patients. There are homologous genes present in budding yeast, worms, flies, zebrafish, mouse, or any other organism you can think of (the Figure below shows sequence alignment for a subset of these organisms).

Surprisingly, the gene that is linked to the disease (*Disease Gene 1* or *DG1*) turns out to be ubiquitously expressed and essential in a variety of model organisms where it has been deleted. These findings make it a bit surprising that missense mutations cause such a specific neurological disorder. The gene identified in this study encodes a member of components of the transcriptional machinery termed TATA Binding Protein (TBP)-associated factors (TAF), which are proteins that associate with the TATA-binding protein in transcription initiation. Most TAF proteins have domains that bind to DNA (DNA binding domains) and regions that interaction with TBP (TBP-interacting domains).

The domain structure for the DG1 protein is shown below with the position and nature of the altered amino acids indicated in Patient 1, Patient 2, and Patient 3.



Domain structure of the DG1 protein. At the top, the domain structure of the DG1 protein is shown with the TATA Binding Protein (TBP)-Interacting Domains indicated in red and blue and the DNA Binding Domain shown in Green. The bottom shows an alignment of sequence from these domains from Human, Mouse, Zebrafish, and Yeast. The positions of the amino acid changes identified in Patient 1 (L->I), Patient 2 (A->Q) and Patient 3 (T->K) are shown above the domain structure and the exact residue altered in each Patient is indicated by the red asterisks in the sequence below that position in the domain structure. Amino acid abbreviations are in a Table following this question.

Interestingly, the patients identified show a range of clinical phenotypes. One patient has mild disease, one has moderate disease, and one has severe disease manifestations. This range of disease phenotypes suggests a genotype/phenotype correlation.

1) Based on the missense mutations identified in the *DG1* gene and illustrated as the amino acid changes they encode in the Figure, suggest which patient is most likely to have mild disease, which patient is most likely to have moderate disease, and which patient is likely to have the most severe disease phenotype. Provide the rationale for each of your suggested genotype/phenotype correlations. (6 points)

You are now interested in testing whether a change in the function of the DG1 protein due to the disease-causing amino acid changes is likely to contribute to disease pathology. You know that DG1 is implicated in transcriptional regulation. Indeed, there is a ChIP-Seq dataset that maps DG1 binding sites in cultured HeLa cells. These data show that DG1 binds to sites in the promoter of 257 genes.

To explore how the missense mutations alter DG1 function, you have opted to create a disease model. You can select any model recalling that your primary goal at this juncture is to define how the mutations impact DG1 function.

2) State which system you have chosen to use to model the *DG1* patient mutations and why you chose that model system. Include three advantages of the system that you chose and one disadvantage. (4 points)

3) Describe in detail how you would create your model to analyze the function of DG1. Be sure to include any controls you would need to generate to interpret the results you would obtain from your model. (5 points)

4) With the model you have created, describe an experimental approach to explore the function of the DG1 protein. You can suggest a hypothesis for how you think the amino acid changes might alter the function of DG1 and describe an experimental approach to test that hypothesis. Be sure to include appropriate controls that will allow you to interpret your results. (5 points)

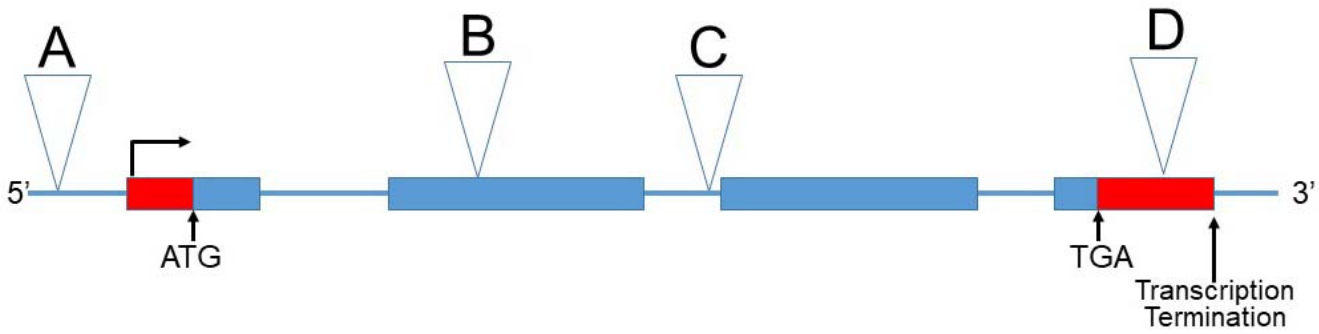
Abbreviations for amino acids

Amino acid	*Three-letter abbreviation*	*One-letter symbol*
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Asparagine or aspartic acid	Asx	B
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic acid	Glu	E
Glutamine or glutamic acid	Glx	Z
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Question 2A

In IBS515 this year, we heard many examples of how repeat expansions can lead to pathology.

1) Based on the positions of the triplet repeat expansions illustrated in the hypothetical gene illustrated below, suggest a mechanism for how a repeat expansion in each region indicated (**A**, **B**, **C**, **D**) could cause a biological change that alters cellular physiology and causes pathology. You must suggest four DIFFERENT mechanisms that reflect the function of the region of the gene affected in each scenario. You may use specific disease examples if you like, but this is not essential for full credit (2.5 points each).



2) For **one** of these examples, describe an experimental approach to test the disease mechanism that you propose. Assume that you have access to a mouse model for the disease you chose, patient data, as well as patient and control fibroblasts. Describe what results you would obtain from your experiment if your proposed mechanism were correct. Be sure to include appropriate controls so that you can interpret the data from your approach. (10 points)

Question 2B

You are studying an enzyme called ABC that has two isoforms (Isoform A and Isoform B). Isoform A, which is expressed in lung, is a 160 kDa protein that contains two distinct conserved domains. Isoform B, which is expressed in kidney cells, is an 80 kDa protein that contains one domain identical to the C-terminal domain of Isoform A.

1) Describe **three** potential explanations for how a single gene can produce two distinct protein forms. (3 points)

2) For each of the explanations suggested in (**1**), describe an experimental approach to test whether this mode of regulation occurs. Be sure to describe the results you would obtain if your explanation is correct as well as if the explanation is not correct. (12 points)

3) Provide an explanation for why it might be necessary to produce a different isoform of the same protein in two different tissues. Also include a real or hypothetical example of how different isoforms can have different functions that could contribute to differences between two tissues (e.g., kidney and lung) (5 points).

Question 3A

You have been studying a congenital limb defect, which occurs in 1 in 1000 live births, and has an unknown genetic etiology. You have whole genome sequence data from trios (parents and an affected child, individual A). In the analysis of your data, you identify a *de novo* nonsense (premature stop) mutation in *NEWGENE* in individual A.

1) You suspect that the *de novo* mutation causes the phenotype in individual A because *NEWGENE* is highly expressed in the limb bud of mouse embryos. Describe an experiment that would test your hypothesis, including appropriate controls. (5 points)

2) You are allowed by your IRB to return results to families participating in your study. Assuming you have gathered sufficient evidence to suggest that the *de novo* mutation in *NEWGENE* is the only cause of individual A's phenotype, what would you tell the parents of individual A about their risk of having another affected child? What could you say to individual A about his/her risk to have an affected child? Explain. (4 points)

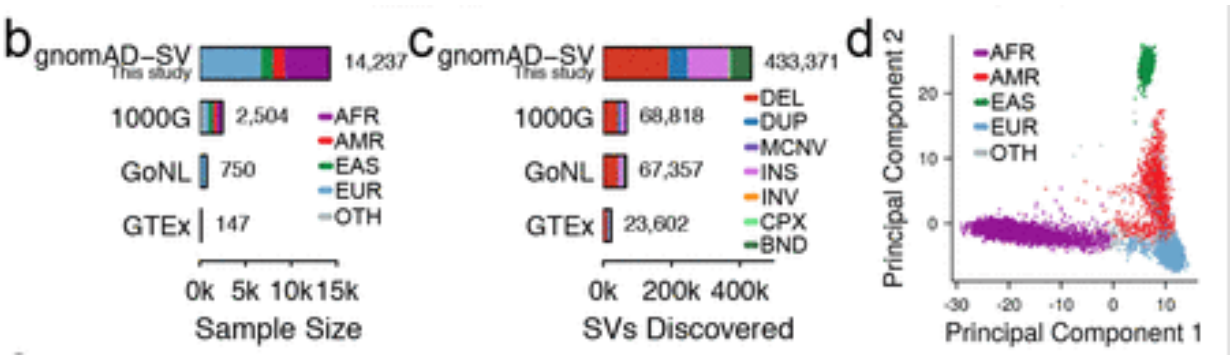
3) As you begin to write up your findings for publication, you realize that it would be helpful to have additional families with mutations in *NEWGENE* to have a compelling story. You post your gene of interest to GeneMatcher, a service that connects researchers to families with mutations in the same gene. You are surprised to find that there are five (5) additional families with a mutation in *NEWGENE*, but with a more severe phenotype than in your original family, in which multiple organ systems are affected. All the affected individuals of these five (5) new families have the same *de novo* missense in *NEWGENE*. Given this information, provide a hypothesis for how this *de novo* missense mutation can cause the observed difference in phenotype in these five (5) families vs. the mutation found in your original family? Describe an experimental strategy including relevant controls to test your hypothesis. (5 points)

4) You have also performed a genome-wide association study using an additional 500 cases with your limb defect and 500 controls. You identify a genome-wide significant association signal where the most significantly associated SNP is 10kb upstream of the transcription start site of *NEWGENE*. There are two other SNPs in complete linkage disequilibrium with the same p-values and odds ratios that are equally good candidates to be the functional SNP that causes your congenital limb phenotype. You hypothesize that one of these SNPs is a functional SNP in a regulatory element of *NEWGENE*. Based on what you know about this gene from your whole genome sequencing study, design an experiment with appropriate controls to determine which SNP is the functional SNP and to show that this region is indeed a regulatory element of *NEWGENE*. (6 points)

Question 3B

Collins *et al.* 2019 recently posted a manuscript in *bioRxiv* entitled, “An open resource of structural variation for medical and population genetics” Based on your knowledge of human genetic variation, answer the questions below.

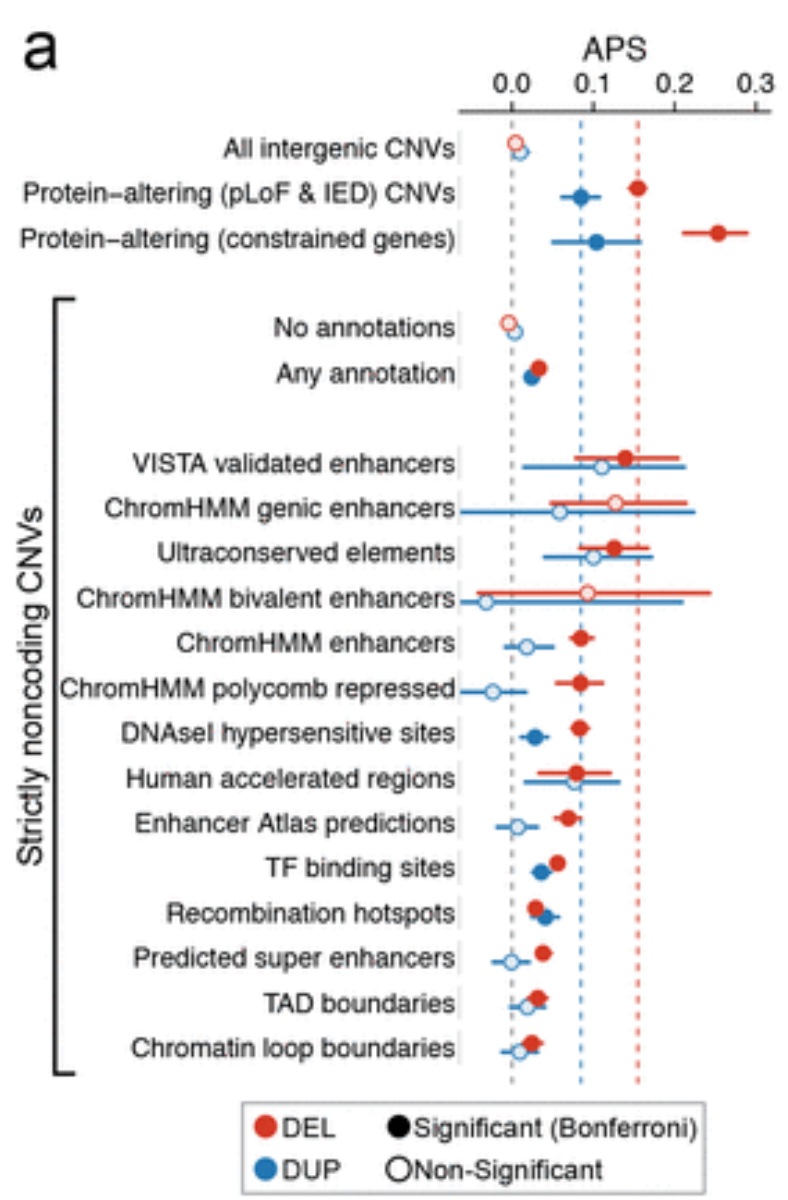
Figure 1 (parts b, c, and d) summarize the total data generated from the experiment. In total, 14,237 diverse genomes were analyzed, and 433,371 structural variants were reported. The diverse ancestries/ethnicities included African/African American (AFR), Latino (AMR), East Asian (EAS), European (EUR), Admixed or Other Populations (OTH).



1) What is the purpose of the analysis summarized in Figure 1d? Provide an interpretation of this analysis. (3 points)

2) Considering material presented in the lecture and what we discussed in class, do the results from Figure 1d agree or disagree with similar analyses performed using single nucleotide variation data? Do these data suggest that these variants are common or rare? (3 points).

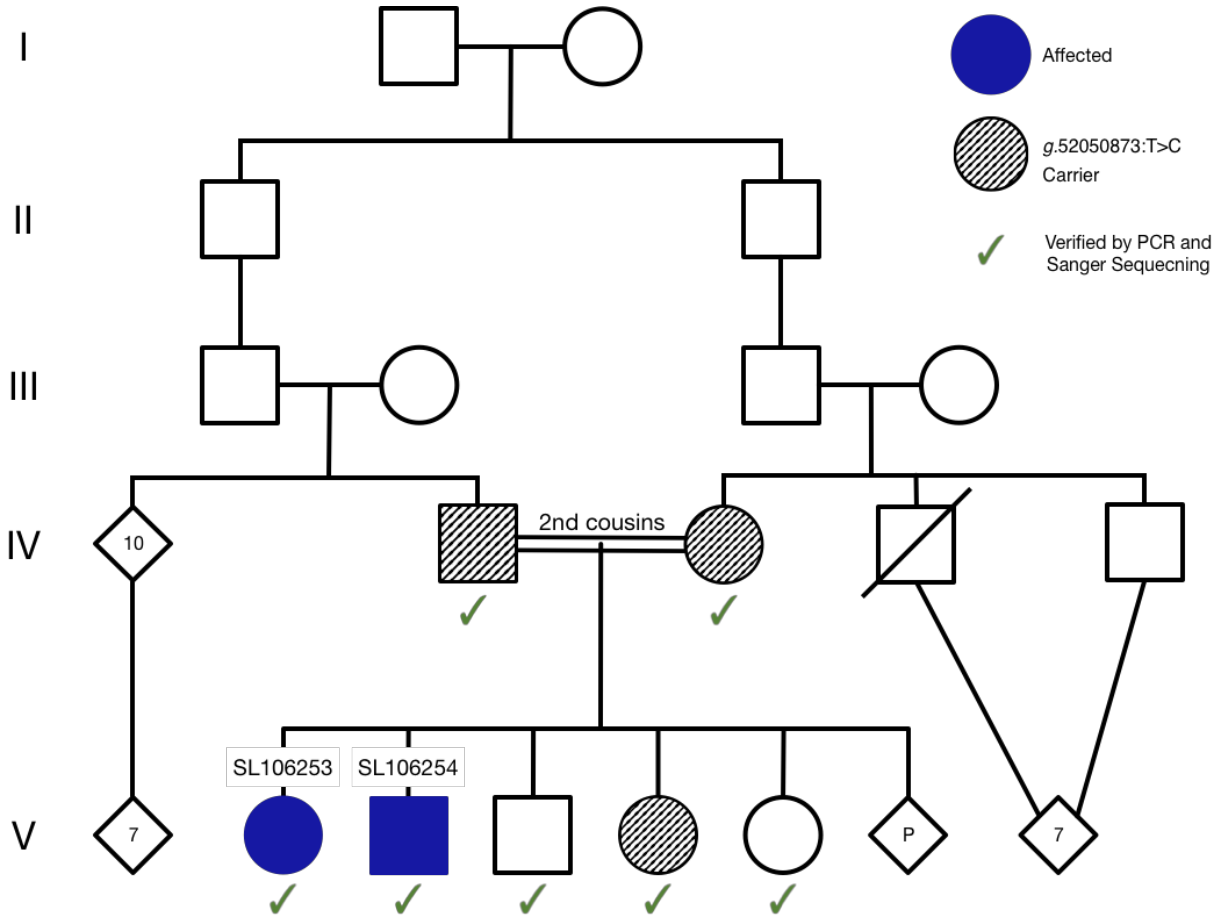
Figure a (below) is from a manuscript by Collins *et al*. APS is a measure of deleterious selection against different classes of genetic variations. The value of 0.0 indicates the expectation from the neutral theory (e.g. there are no fitness differences for the variation). Deletions (DEL), Duplications (DUP) are shown.



3) Compare and contrast the strength of selection against deletions and duplications using the data presented in Figure a. (3 points)

4) Imagine you are performing a genetic analysis of a rare Mendelian disorder. You find a new mutation in an ultraconserved element and a mutation in a chromatin loop boundary that both segregate with the disease. Based on Figure a, which would you functionally pursue first? Explain your reasoning? (3 points)

The next questions (parts 5, 6 and 7) are based on the pedigree shown below. The disorder is very rare with a prevalence of 1 per 1 million in the general population. The prevalence does not vary with ancestry. The goal of this experiment is to identify the disease-causing mutation.



5) What genetic mode of inheritance is most consistent with this pedigree? (2 points)

6) Assuming you have access to human whole genome sequencing and can obtain biospecimens from the individuals in this pedigree, what is the most efficient (with respect to time and cost) experimental design to identify the disease-causing mutation? (2 points)

7) Assume you have identified 10 variants that could be the causative mutation. Now you want to narrow down those that you will follow up with functional studies. What aspects of human genomic variation can you use to help narrow the search? (4 points)

Question 4A

In collaboration with a pharmaceutical company, you are studying a new chemical compound that has been identified as a potential anti-cancer drug that decreases the proliferation of cells in culture. The molecular mechanism of the compound is unclear. You are tasked with starting to define *how* this drug alters cell physiology. To begin this task, you decided to compare the complete proteomes of cancer cells that have either been treated with drug at a concentration that you know to be active, or have been left untreated.

You harvest protein samples from cancer cells treated with drug for three hours (+drug 3hrs) or, as a control, you treat the cells in the same manner but without drug (-drug 3hrs). To ensure rigor and reproducibility in your results, you collect biological triplicates of both conditions. You then use mass spectrometry (MS) to analyze the proteome of these sets of samples.

From your MS datasets, you determine that 247 proteins show a statistically significant decrease in the +drug 3hrs samples while 36 proteins show a statistically significant increase in the +drug 3hrs samples.

1) Based on the experimental setup described above, briefly describe how you would have arrived at the numbers of proteins increased and decreased. You do not have to go provide detail of how mass spectrometry works - just provide the overall experimental setup and what is actually measured in MS, and describe how biological triplicates increases your analytical power. (2 points)

2) Keeping in mind that your goal is to understand the mechanism of your drug's effect on the cells, you decide to begin a follow-up analysis of EITHER the set of 247 proteins that show a decrease in steady-state levels, OR the set of 36 proteins that show an increase in steady-state levels. Which set of proteins (those that decrease or those that increase) would you choose, and why? Provide rationale for your choice. (3 points)

Having selected one of the groups (increased or decreased) to analyze further, you need to consider experimental approaches that will provide some insight into why the drug changes steady-state levels of a large group of proteins. Assume you have ready access to antibodies or any other materials that you would need for your proposed follow up experiments.

3) Regardless of which set of proteins you choose to pursue for your follow-up studies (increased or decreased), you cannot experimentally analyze *all* of them at once, so you need to narrow your set down to a number that is feasible to analyze at the bench (e.g. ~10). How would you use your MS data to define the 10 most interesting potential targets to pursue, validate, and follow-up? (2 points)

4) Describe one experimental approach to validate the results of the MS experiments for the 10 specific proteins that you have chosen to pursue. Assume that the MS analysis consumed all the samples that you previously prepared. Describe the experimental set-up, the analysis, as well as any critical controls. (3 points)

5) To narrow the drug's mechanism of action, describe one experimental approach that tests how the drug changes levels of your 10 validated candidates. Include the experimental set-up and essential controls require to interpret your results. Keep in mind the central dogma of gene expression as you consider your experimental design for this analysis. (3 points)

6) One of your collaborators from the pharmaceutical company shows you preliminary data suggesting that the drug target (i.e. the protein it interacts with) could be a sequence-specific DNA-binding protein called TxFac which is a known transcriptional activator. Your colleague suggests that the drug might alter the preferred DNA binding site for TxFac. Describe an experimental approach to test your colleague's suggestion that the drug alters the DNA specificity of TxFac binding in your cell culture system. Be sure to describe the experimental approach and include appropriate controls to interpret your results. (4 points)

7) Assuming that your data supports your pharma colleague's hypothesis, propose a model for how this Drug could decrease cell proliferation. Your model should incorporate the results of your studies thus far. (2 points)

8) In light of your successful collaboration with pharma, your colleagues suggest moving the drug into mouse cancer models as a step toward human trials. However, all of your studies of the drug have thus far employed a cultured cancer cell line, and you are concerned about side effects. Suggest a critical control you would like to perform before considering moving the candidate drug into in vivo studies, which are both time-consuming and expensive. (1 point)

Question 4B

A recent paper investigated the role of the non-coding *Ifng-as1* gene on expression of the nearby coding gene interferon-γ (gamma) (*Ifng*) in Th1 helper T cells. Use the data from the figure below to answer the following questions.

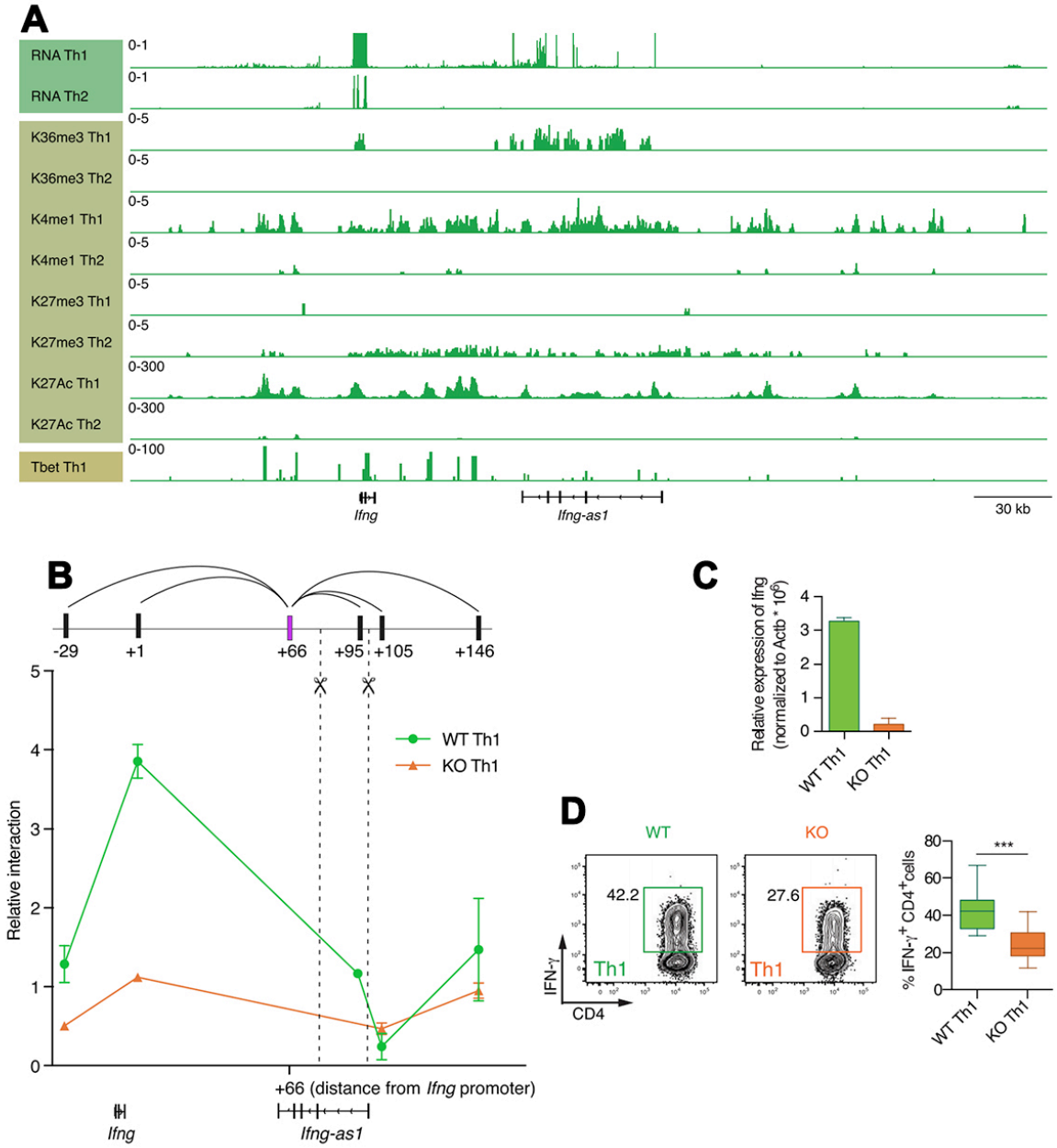


Figure 1: (A) Shows a genome track showing the *Ifng* and *Ifng-as1* genes and results from chromatin-IP sequencing ("ChIP-seq") experiments in Th1 and Th2 cells. Exons (thick line) and introns (thin line) are indicated. **(B)** Shows 3C ('chromosome conformation capture') data gathered in Th1 cells at selected sites within the region shown in **A**. Distances from *Ifng* transcription start site (TSS) are shown in kb (+66 = 66kb downstream). **(C)** *Ifng* mRNA expression in Th1 cells quantified by qPCR. **(D)** INF-γ protein expression in Th1 cells quantified by fluorescence activated cell sorting (FACS) with antibodies to the CD4 protein (which marks Th1 cells) and to INF-γ protein. Box plot shows an average of 12 FACS profiles.

1) Based on the data in Figure 1A, do you hypothesize that the *Ifng* gene is expressed more highly in Th1 or Th2 cells? Why? Explain all data in Figure 1A that support your answer. (4 points)

The investigators created a knockout of *Ifng-as1* by deleting the genomic region between the two dotted lines in Figure 1B. The effect of this deletion on *Ifng* is shown in Figures C and D.

2) Describe one experimental approach that the researchers could have used to create the deletion of the *Ifng-as1* gene. Be sure to include sufficient detail on how the deletion will be made. (3 points)

3) Based on the data in Figure 1B, what is the normal state of local chromatin interactions in the genomic interval and how does this change in cells with the *Ifng-as1* deletion? (4 points)

4) What effect does the *Ifng-as1* deletion have on expression of the *Ifng* gene? (3 points)

5) Propose a model that explains the normal role of the *Ifng-as1* gene in controlling *Ifng* transcription, taking into account *all* the data in Figure 1. (6 points)

Question 5A

The abstract of a published research paper states:

Recent studies reveal that epigenetic regulation, such as histone methylation and acetylation, plays a critical role in determining cell fate. In particular, the expression of key developmental genes tends to be regulated by trimethylation of histone H3 lysine 4 (H3K4me3) and lysine 27 (H3K27me3). Myoblasts are primary muscle stem cells that can proliferate, differentiate, and fuse into mature muscle fibers. Myoblast differentiation is tightly regulated by the receptor activator of nuclear factor κB ligand (NKL) and a transcription factor called "nuclear factor–activated T cell (NFAT) c1". We found that NKL-induced NFATc1 expression is associated with the demethylation of H3K27me3. The Jumonji domain containing-3 protein, a H3K27 demethylase, is induced in muscle stem cell–derived myoblasts in response to NKL stimulation and may play a critical role in the demethylation of H3K27me3 in the NFATc1 gene.

Based on their results, the authors make two primary conclusions in this abstract and they propose a model.

1) Clearly state the two conclusions and describe the proposed model. (4 points)

2) Assuming that you have all of the necessary reagents at your disposal to support each conclusion, describe the experiment(s), including appropriate controls, required to justify each of the two conclusions (5 points each).

3) Describe the results that would support the authors' proposed model. (2 points)

4) Finally, briefly describe the next experiment that you would perform to test the authors' proposed model. Describe the controls that would be required. (4 points)

Question 5B

You just read an exciting paper in *Science* entitled “Evidence for Network Evolution in an *Arabidopsis* Interactome Map”. In the abstract the authors state that, “We describe a proteome-wide protein-protein interaction map for the interactome network of the plant *Arabidopsis thaliana* containing about 6200 highly reliable interactions between about 2700 proteins. A global organization of plant biological processes emerges from community analyses of the resulting network, together with large numbers of novel hypothetical functional links between proteins and pathways.”

1) Describe in detail two experimental approaches that these researchers could have used to generate such an extensive protein-protein interaction map. (4 points).

2) Describe an experiment that you could use to determine the validity of a select few of the protein interactions that were defined in the paper. This experiment must use a different approach from either one proposed in your answer to Question 1. (4 points).

3) In examining the data in this paper, you see that a putative transcription factor of unknown function (Factor X) has been determined to physically interact with three different components of a hormone signaling pathway. It was previously shown that disruption of this signaling pathway causes defects in photomorphogenesis (yellow plants with underdeveloped leaves) while overactivity of the pathway results in embryonic lethality. Explain how you would determine whether Factor X is involved in this hormone pathway (2 points) and, assuming it is, whether it plays a positive or negative role in the pathway. (2 points)

Having read up on the literature about this hormone signaling pathway, you’ve become particularly interested in plant embryonic development. Over lunch, a colleague who is an expert in plant embryogenesis tells you that, by her estimate, less than half of the genes that are essential for proper embryo development have been identified. You know that:

a. *Arabidopsis* is a self-fertilizing, diploid plant.
b. Each plant produces hundreds of embryos, which develop into seeds.
c. Normal embryogenesis gives rise to a healthy seed, but an embryo that is homozygous for a lethal mutation will result in a shriveled seed that does not germinate.

4) Describe a forward genetic approach that you could use to identify new genes that are essential for embryo development. (4 points)

5) Having cloned one such gene, describe two types of analyses that you would use to determine the role of this gene in embryogenesis. (4 points)