EMORY UNIVERSITY

Genetics and Molecular Biology Graduate Program

GMB Written Qualifying Exam: January 15, 2021

Including Answer Suggestions

Question # 1

You are studying an RNA derived from a mouse liver whose level changes with steroid treatment and you also think it may be spliced. The lab has a tube with that purified RNA, but you don't have any other information about it.

a) (4 points) Design an experiment to test your hypothesis that this RNA has been spliced. Be specific about what reagents you use and your experimental approach.

b) (4 points) You obtain the RNA's sequence. Design an experiment to test the idea that it is a functional messenger RNA that extends what you can learn from its nucleotide sequence.

c) (12 points) Your RNA is expressed in mouse hepatocytes in culture but is absent from the cells after exposure to glucocorticoids. Your lab mate thinks this is due to a hormone-dependent repression of transcription of the gene encoding your RNA. What is an alternative to this hypothesis? Design an experiment to test your lab mate's hypothesis AND an experiment to test your alternative.

Question #1 – Answer guide

This question gets at the concept of steady state amounts of RNA and how the level of an RNA is based on its synthesis and degradation. It also asks for the use of methods needed to manipulate RNA and measure it in different contexts. The definition of mRNA in terms of its structure and function are also queried.

Possible answers:

a) (4 pts) Reverse transcribe the RNA with random primers and sequence the products (RNAseq). You can make a full length cDNA by reverse transcription once you have identified a likely 3'-end and clone it by inserting it into a plasmid to have the cDNA for use as a probe. Probe nuclear and cytoplasmic liver RNA with the cDNA (Northern blot) to see if you can identify the unspliced pre-RNA in the nucleus and the shorter spliced RNA in the cytoplasm (although it could be in the nucleus also).

   Alternatively, you can use a computational approach to retrieve related genomic sequence and compare the cDNA derived sequence (from the plasmid clone or RNAseq) to see if there is a missing intron in the genome database not seen in your cloned RNA.

b) (4 pts) If your RNA is an mRNA, prepare cytoplasmic RNA from liver cells, hybrid select it with your clone and see if it has a polyA tract by priming with oligo dT and making cDNA for RNA seq. You can test if there is a 5' cap by seeing if you can digest the mRNA with a 5' to 3 ' exonuclease (it won't chewed up if its capped). Other enzyme treatments can be applied to probe for the cap. An anti-CAP antibody can be used to test for its presence.

   A second experiment is to try to translate the original tube of RNA that was given to you or the hybrid-selected mRNA obtained from cytosol in part2, in an *in vitro* translation reaction.

That would show it is a functional mRNA and must have a cap and you can generate that protein.

Alternatively, you could isolate a polysome fraction from liver cytosol and probe it for your RNA with your cDNA clone. That would put it on ribosomes in vivo.

A less good, but possible answer, is that you see an open reading frame in the sequence and you computationally mine proteomic databases to see if the predicted protein is really made). If a database search is negative that doesn't mean its not an mRNA and, you can propose to examine hepatocyte proteins by mass spectrometry to search for the predicted protein.

c) (12 pts) An alternative hypothesis is that degradation of this mRNAs takes place after the administration of hormone. You can measure the rate of transcription change by using a methods such as nuclear runon (also referred to as GROseq) which measures pol II production of RNA. A high signal will be seen before steroid, and it will be lost over time with hormone. To measure a change in mRNA turnover, you can measure the mRNA half-life. Inhibit synthesis with amanitin or actinomycin D and do a Northern or RT-PCR to measure your RNA over time. Its rate of decay should increase in a hormone dependent manner if the reduction due to hormone is due to an increased rate of degradation.

Question #2

1. Genetically identical bacteria growing in a homogeneous solution can exhibit phenotypic heterogeneity, sometimes called “bistability”, and is seen in epigenetic processes. An example that we discussed in IBS555 is the observation that in a culture of genetically identical *Bacillus subtilis* a fraction of the bacteria in the population will develop competence for DNA uptake.

a) (14 points) Briefly describe two types of regulatory circuits that theoretically can result in bistability. You can use diagrams to supplement your explanation.

b) (6 points) Describe the concept of “noise” as related to bistability as well as its role in establishing bistability.

Question #2 – Answer guide

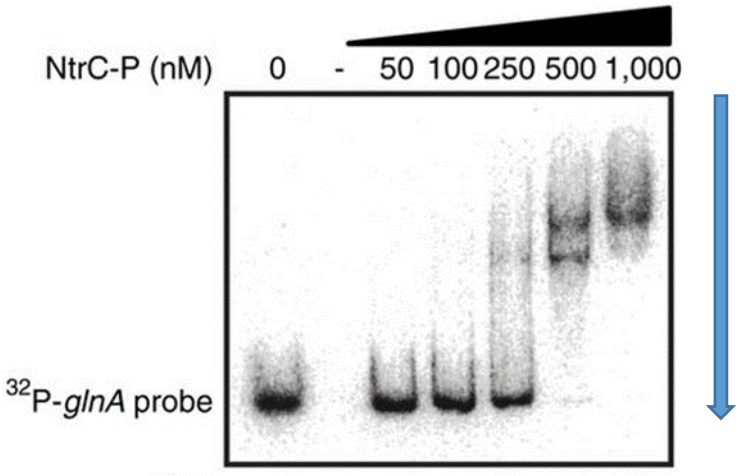
a) (7 pts) A positive feedback loop in which a positive regulator acts cooperatively to activate its own expression. Once activated the positive feedback loop tends to remain active.

(7 pts) A double negative feedback loop, for example two mutually repressing repressors that each repress the synthesis or activity of the other, such that only one them can be active in an individual cell.

b) (6 pts) Here “noise” refers to the stochastic variation in the level of expression of a gene in individual cells of a population. Higher or lower expression of this gene can tip the balance in a cell leading to activation of the positive feedback or the double negative feedback mechanism that produces the phenotypic heterogeneity in the population. This noise is essential to producing the phenotypic heterogeneity. Correctly describing the definition of intrinsic and extrinsic noise in this system is not required, but could be worth an extra point or two.

Question #3

The figure below shows a representative autoradiograph of non-denaturing gel showing the binding of increasing amounts of protein, NtrC-P, to ^{32}P-labelled 2nM DNA probes (^{32}P-glnA probes) with sequences corresponding to the upstream regions of a promoter (positions −273 to +71 relative to the translation-start site). (The lane marked "-" is empty.)



a) (4points) In the electrophoretic mobility shift assay (EMSA) shown in the figure above, explain why the DNA migrates through the gel from top to bottom, and why the addition of the NtrC-P protein slows its migration.

b) (4 points) Let's assume that the functional unit of NtrC-P is a dimer. Propose a model to explain why there are two bands in the lane containing 550nM NtrC-P, but only one band in the lane with 1000nM NtrC-P (note that one band is also visible at 250nM). If you suggest that more than one dimer is binding the DNA explain why you think the binding of two dimers is or is not likely to be cooperative.

c) (4 points) How do DNA binding proteins interact with DNA?

d) (4 points) If one adds an antibody specific to NtrC-P to the DNA binding reaction prior to loading it on the EMSA, describe two potential mutually exclusive outcomes that might be observed on the EMSA?

e) (4 points) The above experiment was done *in vitro*. How would you show that NtrC-P binds to glnA sequence inside a cell?

Question #3 – Answer guide

a) (4 pts) The mixture of DNA and protein is subjected to electrophoresis through a polyacrylamide gel. The DNA has a negative charge; therefore, it migrates through the gel toward the anode (positive charge). Protein binding to the DNA increases the overall molecular weight, but more importantly the protein neutralizes or covers some of the negative charges on the DNA.

b) (4 pts) One hypothesis is that the lower band in the 500nM lane has one dimer bound, whereas the upper band (slower migrating band) has two dimers bound. At 1000nM all of the DNA is occupied by two dimers of NtrC-P. The lower band is also visible at 250nM, but the upper band it not. The binding of the two dimers is not likely highly cooperative, otherwise the upper band (slower migrating) that would contain the two dimers would most likely appear at lower concentrations of protein than the band that presumably is bound by a single dimer.

c) (4 pts) through hydrogen bonds or hydrophobic interactions with specific bases. Additional ionic interactions occur with the phosphate backbone of DNA, but these are not sequence specific.

d) (4pts). A. If the antibody binds to the DNA binding domain of the protein, it could eliminate interactions and there would be no gel shift. B. If the antibody binds to another part of the DNA binding protein, it could result in a supershift of the band towards the top of the gel.

e) (4pts). To see NtrC-P bind to the glnA sequence in cells, one would do a chromatin immune precipitation reaction (ChIP). Cells would be crosslinked with formaldehyde to capture all protein DNA interacitons; the DNA would be fragmented by sonication; and an antibody to NtrC-P would be used to immunoprecipitated the protein/DNA complex. After resolving the crosslinks, and purifiying the DNA, PCR for the glnA sequence would be used to demonstrate that glnA was bound by NtrC-P.

Question #4

A student in your lab conducts a genetic screen for mutations in genes that encode proteins required to maintain gut stem cells in adult *Drosophila* using ethyl methanesulfonate (EMS), a DNA alkylating agent that causes single nucleotide changes. Gut stem cell function can be tested in adult *Drosophila* by feeding dextran sodium sulfate (DSS), which causes cell damage in the gut that must be repaired by resident stem cells or else the adults die.

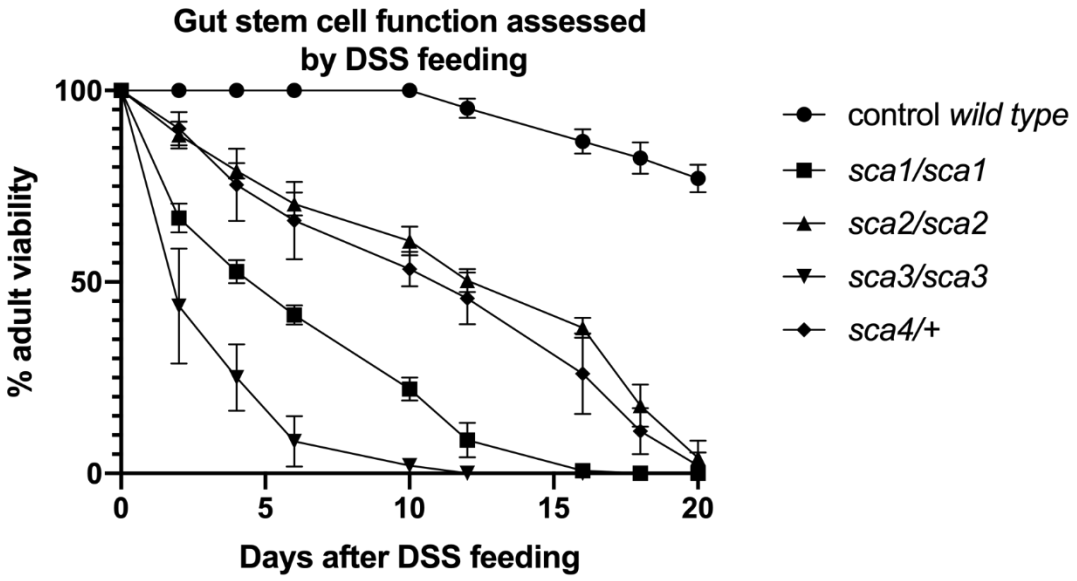
Your student gives lab meetings and reports finding four (4) different EMS alleles, that share a common phenotype: adult lethality after feeding DSS.

The student has named the alleles *stem cells absent-1, 2, 3,* and *4* (*sca1, sca2, sca3, sca4*).

sca1, sca2, sca3 only display DSS sensitivity as homozygous adults

sca4 is lethal as a homozygote but heterozygous adults show DSS sensitivity.

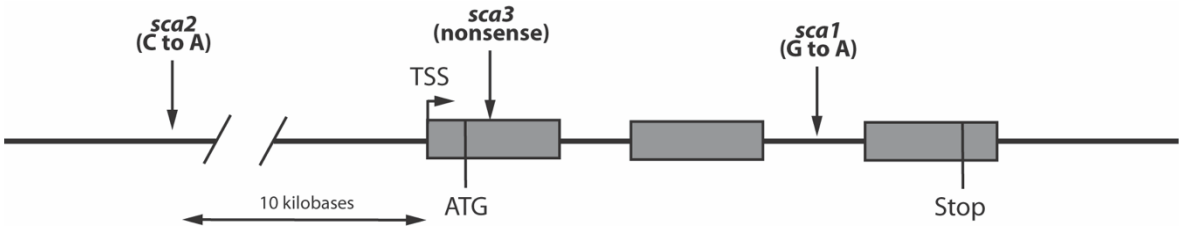
The student's data look like this:



a) (4 points) Your student presents a hypothesis about the relative strength **and** nature of the *sca1-4* alleles. You agree with their assessment. What is it?

b) (2 points) Propose one genetic approach to testing whether these alleles represent mutations in *the same gene*. What would you tell your student about the limitations of this approach as it relates to this group of alleles?

Using a combination of genetics and sequencing data, you are able to map the mutations in *sca1-3* to a relatively small region of **chromosome 2** containing a single predicted protein-coding gene (see below; location of base changes in *sca1-3* are indicated).



c) (6 points) Propose a hypothesis for how each allele shown above disrupts gene expression **and** propose experiments to test each of these hypotheses. Assume you have access to necessary reagents (e.g. RNAi, antibodies etc). Explain how the outcome of the proposed experiment would tell you that your hypothesis is correct.

After some hard work, the student maps *sca4* to chromosome 3. Screening a collection of deficiencies that collectively cover the entire *Drosophila* genome identifies a single chromosomal deficiency (Df) on chromosome 4 that interacts genetically with *sca4*:

sca4/+ = sensitivity to DSS feeding in adults
Df/+ = wild type sensitivity to DSS feeding in adults
sca4/+;*Df*/+ = wild-type sensitivity to DSS feeding in adults

d) (4 pts) What is your hypothesis about why the *Df* modifies the *sca4*/+ phenotype?

e) (4 pts) Propose one genetic experiment to test the hypothesis that the gene affected by the *sca1-3* alleles is in the same pathway as the gene affected by *sca4*.

Question #4 – Answer guide

a) (2 pts) *Based on the data, sca1,2,3 behave like recessive loss-of-function alleles. The sca4 allele has a dominant phenotype, and therefore is likely to be gain-of-function (e.g. dominant-active or dominant-negative). Another conceivable, but less likely answer, is that sca4 is l-o-f and haploinsufficient.*

(2 pts) *The first three alleles are in a series: sca3>sca1>sca2 from strongest to weakest. The dominant allele sca4 gives a phenotype that is somewhat equivalent to the sca2/sca2 homozygotes, but because its dominant it is hard to be confident of where it fits in the sca3>sca1>sca2 allelic series.*

b) *(2 pts) The three recessive alleles could be tested by classic genetic complementation ie. creating trans-hets and testing whether they show DSS sensitivity. This approach cannot be applied to sca4, since it has a dominant phenotype.*

If the student really thinks this through, they might realize that they could test whether sca1/2/3 alleles enhance the sca4 phenotype; this could provide insight into how these alleles, and thus their encoded products, interact, but it would not tell them if the alleles all affect the same gene.

c) 2pts for each. *There are many possibilities here, but the general idea is that:*

(2 pts) *sca2 is likely a regulatory allele and probably affects a distant enhancer that promotes sca mRNA transcription. A test of this would assess steady state mRNA levels by qPCR, Northern etc. A student might also suggest using EMSA to see if TF binding to the enhancer is lost in sca2 mutants, but this would not test an effect on mRNA and would not get full credit. Controls vs wt are needed**.***

(2 pts) *sca3 truncates the encoded protein. Do a western blot with an antibody that recognizes the N-term and see of the band gets smaller vs wt*

(2 pts) *sca1 likely alters splicing of the 2nd intron leading to retained intron/alt 5' splice site (really up to the student to chose). This could be tested by RT-PCR with flanking primers located in the adjoining exons, or by a Northern blot. Controls vs wt are needed.*

d) *(4 pts) The ability of the Df, as a heterozygote, to* ***suppress*** *the DSS sensitivity of sca4/+ indicates that a gene in the Df region is required for the dominant effect of sca4. This genetic requirement data is consistent with a model in which the unidentified Df gene acts in the same pathway as sca4.*

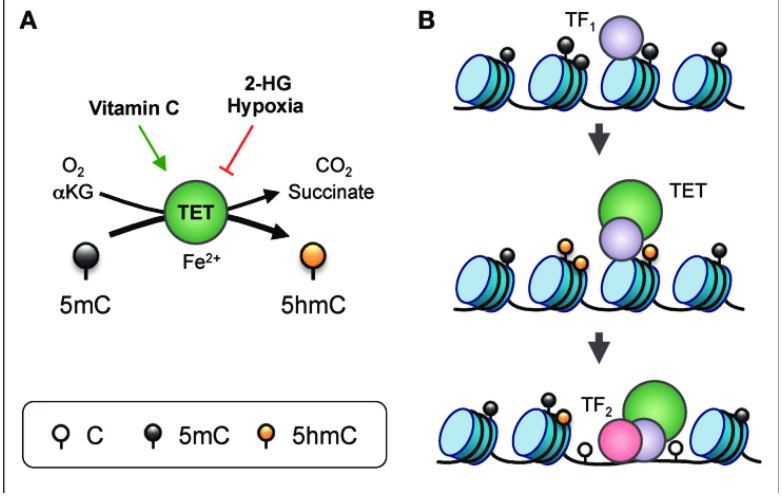
e) *(4 pts) There are two options here:*

1) One option is to use the sca1/2/3 alleles and the Df to replicate the genetic scheme they were given in the preceeding question. This could link sca1/2/3 to the Df, and by extension to sca4. As an example: is DSS resistance restored in sca2/sca2;Df/+ vs sca2/sca2?

2) An alternative answer could be to test whether sca1/2/3 alleles act as dominant enhancers of the sca4/+ DSS phenotype. As an example: is DSS resistance enhanced/made worse in sca2/+;sca4/+ vs sca4/+?

Question #5

Methylation of the fifth position of cytosine is one of the best-studied and most mechanistically understood epigenetic modifications and is well conserved among most plant, animal and fungal models. DNA methylation involves the chemical covalent methylation of the 5-carbon position of cytosine (5mC) by a group of DNA methyltransferases. 5mC plays fundamental roles in regulating gene expression, genomic stability and X-chromosomal inactivation. Interestingly, 5mC, once thought to be a permanent and irreversible epigenetic mark, can be actively and dynamically regulated. Active DNA demethylation is carried out by the TET (ten-eleven translocation) methylcytosine dioxygenases, which progressively oxidize 5mC to 5-hydroxymethylcytosine (5hmC) and downstream derivatives.



Gene regulation by TET proteins. (A) Enzymatic activity of TET. TET proteins, with the co-factors Fe 2+ and α-ketoglutarate (αKG), use oxygen to oxidize 5mC into 5hmC, generating CO 2 and succinate as by-products. The enzymatic activity of TET can be modulated by additional factors. For instance, vitamin C (ascorbate) can enhance TET activity, potentially via reduction of the iron ion. On the other hand, the "oncometabolite" 2-hydroxyglutarate (2-HG), generated in acute myeloid leukemia and glioblastoma by recurrent dominant-active mutants of isocitrate dehydrogenase 1 or 2 (IDH1/2), inhibits TET activity. Furthermore, lack of oxygen in hypoxia also inhibits TET function. (B) Model of TET-mediated enhancer regulation. Prior to the commissioning of an enhancer, pioneer transcription factor (indicated as TF-1) binds to nucleosomal DNA and recruits TET which oxidizes the surrounding 5mC into 5hmC (and/or other oxi-mCs), facilitating DNA demethylation. TET proteins and, TF-1 promote enhancer accessibility by recruiting nucleosome remodeling complexes, thus allowing binding of secondary

a) (7 points) As shown in Figure A, the activity of TET protein can be stimulated or inhibited by Vitamin C or 2-hydroxyglutarate (2-HG), respectively, resulting in a genome-wide increase or decrease of 5hmC levels. Design an experiment to precisely determine the genomic loci that show 5hmC increase or decrease in response to Vitamin C or 2-HG treatment in a cultured human cell line.

b) (6 points) As shown in Figure B, prior to the commissioning of an enhancer, a pioneer transcription factor (indicated as TF-1) binds to nucleosomal DNA and recruits TET which oxidizes the surrounding 5mC into 5hmC (and/or other oxi-mCs), facilitating DNA demethylation. Through TET-immunoprecipitation, coupled with mass spectrometry analysis, you now know the identity of TF-1 (Figure B, purple). Design an experiment to determine the global binding sites of TF-1.

c) (7 points) Design an experiment to test whether TF-1 is a prerequisite for the TET proteins to bind to the enhancer regions indicated in Figure B. In other words, would loss-of-TF1 impair the TET proteins binding to their targets?

Question #5 – Answer guide

a) (7 pts) The hmC modification can be detected by coupling biotin to it as an affinity tag. A viral enzyme, β-glucosyltransferase (β-GT), can catalyze the transfer of a glucose moiety from uridine diphosphoglucose (UDP-Glu) to the hydroxyl group of 5-hmC, yielding β-glucosyl-5-hydroxymethyl-cytosine (5-gmC) in duplex DNA. Like unmodified glucose, azide-modified glucose is well tolerated by βGT and can be efficiently transferred to 5hmC. Biotin can be subsequently coupled to the azido group. Relying on the extremely tight and specific binding between biotin and streptavidin, which has virtually no modification density bias, we can in principle label every 5hmC and perform selective pull-down for genome-wide profiling or loci-specific analysis of 5hmC distribution through sequencing (Song et al, Nat Biotechnol 2011 Jan;29(1):68-72. 2012). This procedure should be carried out on control, vitamin C, and 2-HG treated cells.

An alternative, is the use of antibody-based hydroxymethyl-DNA immunoprecipitation-sequencing (hMeDIP) coupled with high-throughput sequencing.

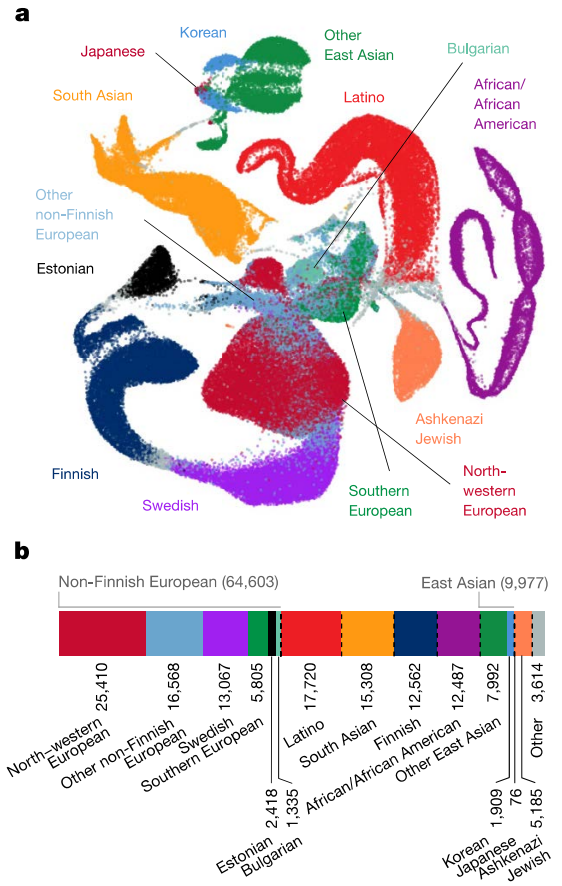
b) (6 pts) Perform a TF-1 ChIP-seq experiment to determine the genome-wide TF-1 binding sites. The student can assume they have an antibody to TF-1 or can express an epitope tagged TF-1. ChIP-Seq typically starts with crosslinking of DNA-protein complexes. The chromatin samples are then fragmented and treated with an exonuclease to digest unbound oligonucleotides. DNA-bound protein is immunoprecipitated using a specific antibody. The bound DNA is then coprecipitated, purified, and sequenced. The CUT&RUN and or CUT&TAG affinity-purification methods are also an acceptable alternative to ChIP-seq.

c) (7 pts) In order to determine the hierarchical orders of TF-1 and TET proteins, we can deplete the TF-1 by siRNA or CRISPR-knockout and perform the TET protein ChIP-seq in the presence and absence of TF-1. A parallel control would be to deplete TET proteins and perform ChIP-seq of TF-1 but is not required. If TF-1 is a prerequisite for TET proteins to bind (or responsible to recruit TET proteins), one would expect the drastic reduction of TET protein binding peaks in the ChIP-seq datasets.

Question #6

Karczweski et al 2020 recently published an article in Nature entitled “The mutational constraint spectrum quantified from variation in 141,456 humans.” Based on your knowledge of human genetic variation, answer the questions below.

Figure 1 (parts a and b) summarize the aggregation of 141,456 exome and genome sequences.

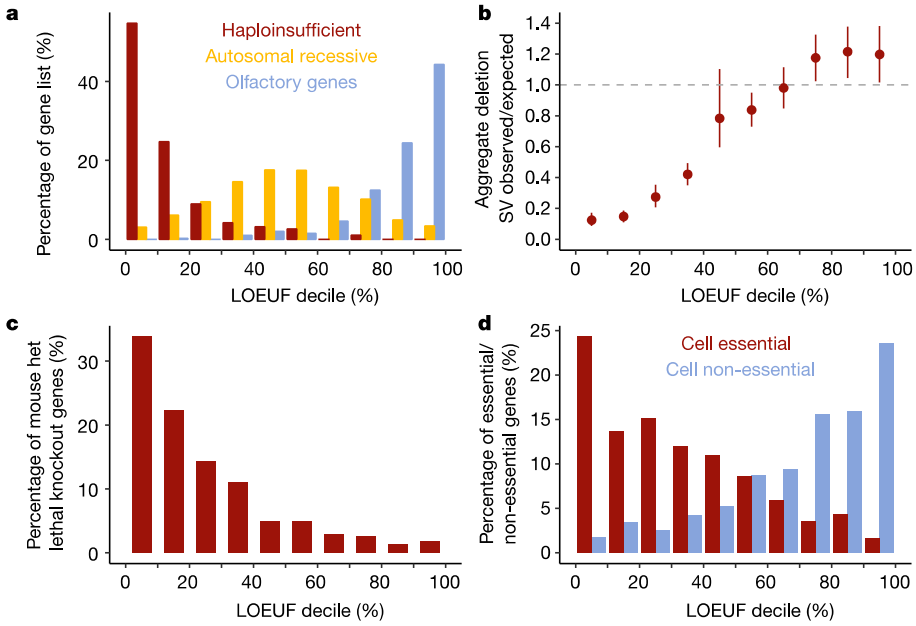


A. (5 points) What is the purpose of the analysis summarized in Figure 1a? Provide an interpretation of this figure.

B. (5 points) In what ways do the data summarized in Figure 1b represent genetic variation in humans? In what ways do the data summarized in Figure 1b fail to adequately represent genetic variation in humans?

C. (10 points) Figure 3 (below) from the Karczweski et al 2020 paper present attributes of loss of function variants observed among the sequenced exomes and genomes. The loss-of-function observed/expected upper bound fraction (LOEUF) enables a quantitative assessment of constraint among genes and genetic variants by

comparing observed numbers of loss of function variants against that expected for a given gene. In Figure 3b, SV stands for structural variants.



For each section of the figure (a, b, c, d) define the terminology and provide a written interpretation for the data shown.

Question #6 – Answer guide

A) (5 pts) Purpose: The purpose of Figure 1 a is to infer the relationship among the exomes and genomes sequenced in this study. It is called a Uniform Manifold Approximation and Projection (UMAP).

Interpretation: These data show that human genetic variation is fundamentally continuous with groupings that corresponds to continental origin of ancestors.

B) (5 pts) The exome and genome data summarized in Figure 1b were obtained from prior funded studies and not generated for this specific study.

Represents Genetic Variation in Humans: The dataset includes exomes and genomes from many individuals with diverse ancestry. While most genetic variation (~ 85%) is shared among all humans irrespective of the continental origin of their ancestors, some genetic variation exhibits structure that reflects geographic origins (see Figure 1a).

Fails to Represent Genetic Variation in Humans: While there are many different groups represented in Figure 1b, the data show a pronounced overrepresentation of individuals of European ancestry and an underrepresentation of individuals from

Africa and Asia. A more complete picture of human genetic variation would need to include exomes and genomes from individuals of African and Asian ancestory.

C) (10 pts)

C-a. (3 points) Haploinsufficent genes are those genes that result in a phenotype when one copy of a gene is inactivated. These are dominant loci. Autosomal recessive loci require loss of both copies of a gene to result in a phenotype. Olfactory genes include numerous pseudogenes and are known to be under little evolutionary constraint in human populations. Figure (a) shows that Haploinsufficient genetic loci have the fewest loss of function (LoF) alleles in the human samples surveyed. This is because loss of one allele results in a dominant phenotype that is strongly selected against (purifying selection). Autosomal recessive alleles have more LoF alleles. Olfactory genes have the most LoF alleles since they are not undergoing purifying selection.

C-b. (2 points) The data suggest that genes included in most structural variations (SV) have fewer loss of function alleles than that expected by chance. So most SV are subject to purifying selection.

C-c. (3 points) Lethal knockout genes in the mouse have fewer LoF alleles in human than that expected by chance. This suggests that important functions of these genes are conserved between mice and humans.

C-d. (2 points) Cell essential genes are those needed to the viability of cells. Cell non-essential genes can be lost without affected cell viability. Cell essential genes have fewer LoF variants than expected by chance. This patter differs from that seen for cell non-essential genes which tend to have more LoF variants.

Question #7

You are given the following observed genotype table for marker A/a for studying a complex disease, in a cohort of 1000 samples:

	AA	Aa	aa
Controls	100	100	120
Cases	260	360	60

a) (4 points) Please conduct a Chi-square test for testing **Hardy-Weinberg Equilibrium** and state your conclusion about if this marker follows Hardy-Weinberg Equilibrium, with significance level 0.05. (Chi-square table is provided below)

b) (4 points) Given the genotype information provided above, please state the assumptions of an **Allelic Association Test** and determine if these assumptions are met for this marker and disease.

If the assumptions are met, apply an **Allelic Association Test** with significance level 0.05 to test the association of this marker with the disease state in this cohort.

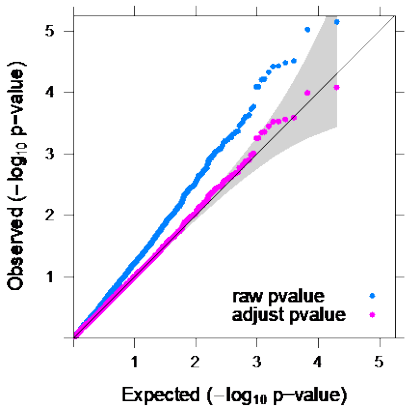
If the assumptions for an Allelic Association Test are not met, explain your reasoning.

c) (4 points) Given the information provided above, what is the **odds ratio** for having the disease allele A under the Allelic Association Test model? What does the odds ratio value mean? Please explain.

d) (4 points) If the median p-value of your Genome Wide Association Study (GWAS) results is 0.4, what is the **genomic control factor** of your GWAS results?

e) (4 points) The graph below shows a q-q plot for your GWAS with raw p values presented as blue dots, and adjusted p values presented as pink dots.

What are the potential issues and causes for the distribution seen for the raw p values (blue data points)?

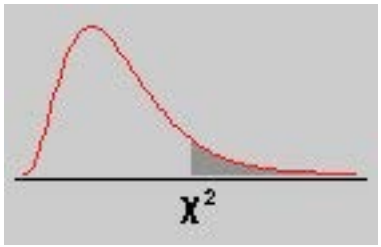


How would you **adjust for this issue by using a genomic control factor** to make your GWAS results look like what is illustrated by pink dots in the graph?

P-value table of χ^2 statistic

The p-values with respect to corresponding Chi-square Statistic values and degrees of freedom (df) are shown as follows:

	p-value									
	0.9	0.7	0.5	0.45	0.4	0.3	0.1	0.05	0.01	0.005
df=1	0.016	0.148	0.455	0.571	0.708	1.074	2.706	3.841	6.635	7.879
df=2	0.211	0.713	1.386	1.597	1.833	2.408	4.605	5.991	9.21	10.597
df=3	0.584	1.424	2.366	2.643	2.946	3.665	6.251	7.815	11.345	12.838
df=4	1.064	2.195	3.357	3.687	4.045	4.878	7.779	9.488	13.277	14.86



Question #7 – Answer guide

You are given the following observed genotype table for marker A/a for studying a complex disease, in a cohort of 1000 samples:

	AA	Aa	aa
Controls	100	100	120
Cases	260	360	60

a) **(4 pts)** Please conduct a Chi-square test for testing **Hardy-Weinberg Equilibrium** and state your conclusion about if this marker follows Hardy-Weinberg Equilibrium, with significance level 0.05.

First, calculate the expected number of genotypes AA, Aa, aa based on the allele frequency

$$p_A = \frac{360 * 2 + 460}{2000} = 0.59; \quad p_a = 1 - p_A = 0.41$$

and total sample size 1000:

$$E_{AA} = p_A^2 * 1000 = 348.1\ ;\ E_{Aa} = 2 * p_A * p_a * 1000 = 483.8;$$

$$E_{aa} = p_a * p_a * 1000 = 168.1$$

The Chi-square test statistic for testing HWE is given by

$$X^2 = \frac{(360-348.1)^2}{348.1} + \frac{(460-483.8)^2}{483.8} + \frac{(180-168.1)^2}{168.1} = 2.42$$

Based on the given Chi-square test statistic table, we can obtain $0.05 < p_{value} < 0.3$. We will not reject the null hypothesis and claim that this marker follows HWE.

b) **(4 pts)** Given the genotype information provided above, please state the assumptions of **Allelic Association Test** and determine if these assumptions are met for this marker. If yes, conduct test the association of this marker by **Allelic Association Test** with significance level 0.05. If not, explain why.

The assumptions for Allelic Association Test are additive disease model and HWE for the test marker. Yes, this marker and disease do meet these requirements, because of the HWE test in a) and being a complex disease.

Assume additive disease model, we will conduct a Chi-square test based on the following contingency table

	A	a	Row Total
Controls	300	340	640
Cases	880	480	1360
Column Total	1180	820	2000

Similarly, first calculate the expected number of genotypes per cell in the contingency table based on the genotype frequency:

$$E_{A\ in\ controls} = 640 * (p_A) = 377.6$$

$$E_{a\ in\ controls} = 640 * (p_a) = 262.4$$

$$E_{A\ in\ cases} = 1360 * (p_A) = 802.4$$

$$E_{a\ in\ cases} = 1360 * (p_a) = 557.6$$

Then the Chi-square test statistic is given by

$$X^2 = \frac{(300-377.6)^2}{377.6} + \frac{(340-262.4)^2}{262.4} + \frac{(880-802.4)^2}{802.4} + \frac{(480-557.6)^2}{557.6} = 57.2$$

The p-value will be < 0.05 based on the given Chi-square test statistic table. Thus, we will reject the null hypothesis and claim there is significant association between marker A/a and the disease of interest.

c) **(4 pts)** Given the information provided above, what is the **odds ratio** for having the disease allele A under the Allelic Association Test model? What does the odds ratio value mean?

The odds ratio is given by

$$\frac{880 * 340}{300 * 480} = 2.07$$

This means that caring the disease allele A will have 2.07 folds higher risk than carrying allele a.

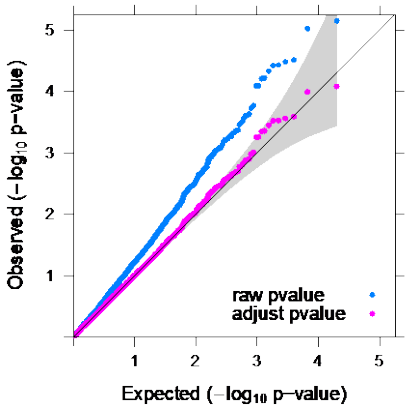
d) **(4 pts)** If the median p-value of your GWAS results is 0.4, what is the **genomic control factor** of your GWAS results?

From the given chi-square test statistic table, p-value 0.4 corresponds to a chi-square test statistic value 0.708, while the expected median p-value 0.5 corresponds to a chi-square test statistic value 0.455. The genomic control factor is given by

$$\frac{0.708}{0.455} = 1.55$$

e) **(4 pts)** If the qqplot of the p-values of your GWAS results are shown as the blue dots in the following plot, what are the potential issues and causes? How would you **adjust for this issue by using genomic control factor** to make your GWAS results look like the one shown in red dots?

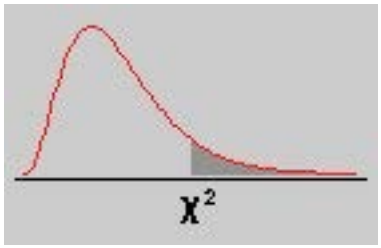
The qqplot in blue dots shows potential inflated false positives that may due to population stratification or batch effects. By using genomic control factor, one will first convert all p-values to their corresponding chi-square test statistic values. Second, one will scale the chi-square test statistic values by dividing by the genomic control factor. Last, one will convert the scaled chi-square test statistic values back to p-values. The adjusted p-values are expected to look like the one shown in red dots.



P-value table of χ^2 statistic

The p-values with respect to corresponding Chi-square Statistic values and degrees of freedom (df) are shown as follows:

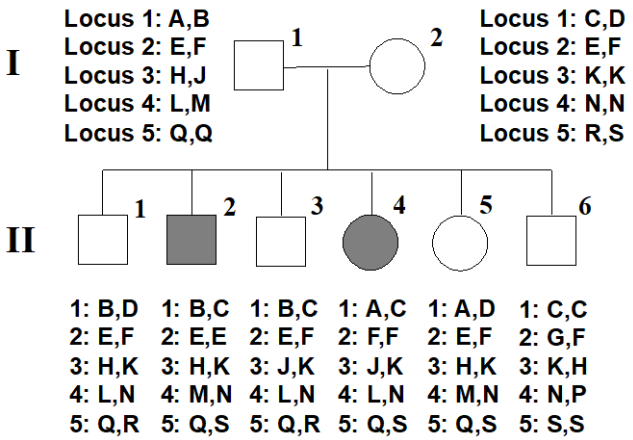
	p-value									
	0.9	0.7	0.5	0.45	0.4	0.3	0.1	0.05	0.01	0.005
df=1	0.016	0.148	0.455	0.571	0.708	1.074	2.706	3.841	6.635	7.879
df=2	0.211	0.713	1.386	1.597	1.833	2.408	4.605	5.991	9.21	10.597
df=3	0.584	1.424	2.366	2.643	2.946	3.665	6.251	7.815	11.345	12.838
df=4	1.064	2.195	3.357	3.687	4.045	4.878	7.779	9.488	13.277	14.86



Question #8

You are the resident geneticist in a small rural hospital in Tahiti. You drew the pedigree on the right after speaking with a family that includes 2 children (II-2 and II-4) affected by a condition you recognize as Strawberry Disease (SBD). SBD is an extremely rare *autosomal recessive* condition characterized by the occurrence of red spots on the arms and legs in the neonatal period. Assume no other known history of SBD in this family, no new mutations, and full penetrance.

SBD is clearly genetic and has been mapped to chromosome 3, but the causal gene has yet to be defined, preventing direct genetic testing. There are also no established enzymatic or metabolic biomarkers of SBD that can be used for diagnosis or carrier testing. But you have found this family, and have genotyped all 8 members to define their alleles present (coded as letters in the figure) at each of 5 polymorphic marker loci (1, 2, 3, 4, and 5) on chromosome 3. You remember from a genetics class in graduate school that when children in a family are affected by an autosomal recessive disorder like SBD with no new mutations, that means both biological parents are obligate carriers.



You hypothesize the genetic defect that causes SBD is tightly linked to one of the 5 loci you genotyped in the family, and this is your big chance to test for possible linkage. To be clear, *you do not believe the alleles present at any of the 5 loci you genotyped directly cause SBD; only that the causal mutation is tightly linked to one of these loci*.

a) (6 points) Look carefully at the alleles present at each locus in all 8 people depicted in the pedigree and track which parent transmitted which allele to each child. Do you see anything problematic? Describe what you see and offer 2 DIFFERENT reasonable explanations. How, given appropriate consent, you could test to distinguish between these possibilities?

b) (8 points) Back to looking for linkage to SBD. From the pedigree, do you see evidence consistent with any of the 5 loci (1, 2, 3, 4, or 5) being tightly linked with causal mutations for SBD in this family? Explain your conclusions and reasoning, taking both affected and unaffected children into account.

c) (6 points) Building on your answers from 1a and 1b, and considering the genotypes determined for the 4 unaffected children in the pedigree, can you draw any conclusions about the possible SBD carrier status of each unaffected child? Explain your reasoning for each child.

Question #8 – Answer guide

a) (6 pts) Everything looks fine for the first 5 children -- meaning you see the expected Mendelian transmission of alleles with each child carrying one allele at each locus that could have been contributed by each parent. However, there is a problem explaining the alleles seen in child II-6. Specifically, at each locus you see one allele that could have been contributed by Mom, but at most of the loci you see a second allele that could not have been contributed by Dad. You might hypothesize something rare like uniparental disomy, or even new mutation at a marker locus, but you would need to invoke new mutation at more than one locus to explain what you see, basically ruling out that option, and at some loci the child has an allele not present in Mom or Dad, ruling out uniparental disomy. The most likely explanations remaining are non-paternity or sample mix-up, and with appropriate consent you could distinguish between these options by genotyping the child, and both Mom and Dad, at a sufficient number of other polymorphic loci in the genome. In short, you would be looking to see if Mom could have contributed one allele at every locus genotyped in your new test (testing if Mom is the biological mother) and also if Dad could have contributed one allele at every locus genotyped in your new test (testing if Dad is the biological father). If sample mix-up, you should find loci where neither Mom nor Dad could have contributed some of the alleles you see in the child, and if non-paternity then at every locus you should see an allele in the child that could have been contributed by Mom, but at multiple loci you should see alleles that could not have been contributed by Dad.

b) (8 pts) To decide if the causal mutations in this family appear tightly linked to 1 of the 5 loci tested (1, 2, 3, 4, and 5) look at the alleles for each locus inherited by the 2 affected and 3 unaffected children in this family. We are told SBD has full penetrance and no new mutations, so at least we don't need to worry about those potential complications.

With 2 affected children, we know that both parents (I-1 and I-2) must be carriers, and neither is affected, which means that at the locus linked to SBD each parent must carry one "at risk" allele and one "not at risk" allele. If we track transmission of the alleles these parents gave to each of their children, we can start ruling out tight linkage to SBD for some loci:

- locus 1 can't be tightly linked to the SBD causal mutation because children II-2 and II-3 both inherited exactly the same alleles from both parents, yet one is affected and the other isn't, so these 2 alleles can't bring the mutation to one child and not to the other. Note -- we can only use this reasoning because we are looking for TIGHT linkage, meaning no recombination expected between the marker locus and the causal mutation.
- locus 2 also can't be tightly linked to SBD in this family because of children II-2 and II-4: these children are both affected yet they inherited different alleles from the parents at locus 2
- locus 3 also can't be tightly linked to the SBD mutation in this family because of children II-2 and II-4 -- as above they are both affected yet one inherited allele H from Dad at this locus while the other inherited allele J
- locus 4 also can't be tightly linked because of what we see in children II-2 and II-4, namely, we know they inherited different alleles from Dad at this marker locus, yet both are affected

and Dad is not, so both must have inherited Dad's SBD mutation, which tells us that mutation can't be lightly linked to marker locus 4.

- now look at locus 5. Both affected children, II-2 and II-4, inherited what might be exactly the same alleles from their parents (Q, S). I say "might be" because we know for sure they inherited the same allele from Mom (allele S), and since Dad has 2 copies of allele Q at locus 5 (one inherited from his mother and one from his father -- we can't distinguish from the information provided) we really can't know for sure if he transmitted the same exact allele to both II-2 and II-4 at this locus or if he might have transmitted one of his allele Q's at this locus to one child and the other allele Q at this locus to the other child. But at least it MIGHT be the same allele.

c) (6 pts) Look at the marker locus 5 allele combinations in each unaffected child and for each allele ask what is the likelihood that allele brought along a tightly linked SBD mutation. From the reasoning in 1a and 1b above, we believe Mom's SBD mutation is tightly linked and in cis with her allele S at locus 5. We also believe Dad carries an SBD mutation tightly linked and in cis with one of his allele Q's at locus 5, but for now we cannot know which one -- so it's a 50/50 chance for each one.

II-1: 50% chance carrier since she did not inherit her Mom's at-risk allele (S) at locus 5, so whether she inherited her Dad's at-risk allele Q or his "other" allele Q at locus 5 is not something we have data to figure out now -- so it's a 50/50 shot.

II-3: same answer as for II-1

II-5: this child likely is a carrier if our assumptions about tight linkage of the SBD mutation with locus 5 in this family are correct. Reasoning: child inherited Mom's at-risk allele (S) but doesn't have SBD so must have not inherited Dad's at-risk allele.

II-6: If Mom is the biological mother then this child likely is a carrier if our assumptions about tight linkage of the SBD mutation with locus 5 in this family are correct. Reasoning: child inherited Mom's at-risk allele at locus 5 (S). If Dad is not the biological father then tracking his alleles doesn't add information here. If this child's genotype information represents sample mix-up then we can't make assumptions about the allele at locus 5 that looks like it could have come from Mom, either, so we would have no choice but to cite the population carrier risk for this child -- or get a new sample from this child and have it retested.

Question #9

You have conducted an EMS (ethyl methanesulfonate) mutagenesis screen to isolate mutants of *Arabidopsis thaliana* that show premature flowering compared to wild type plants. While wild-type plants normally begin flowering at 3 weeks after germination under long day conditions, one recessive mutant from your screen begins to flower after only 10 days.

a) (5 points) Describe a genetic mapping approach to identify the mutated gene responsible for this early-flowering phenotype in your EMS mutant.

b) (5 points) You find that a single point mutation in a gene encoding a SWI/SNF-type ATPase (called *SWC4*) is responsible for early flowering in your mutant. You also observe that the mutants do not accumulate any detectable *SWC4* transcript. List three possible locations for this point mutation in the *SWC4* gene, and explain how each could lead to the lack of transcript accumulation. Two of these mutated sites should be outside the transcribed region of the gene and one should be within the transcribed region.

c) (5 points) FLOWERING LOCUS C (FLC) is a repressor of flowering that is normally expressed at high levels during vegetative growth, when the gene is in a euchromatic state. The *FLC* gene is later silenced by the polycomb silencing system, converting it to a facultative heterochromatin state, thus allowing flowering to begin. You hypothesize that SWC4 normally prevents silencing of *FLC* and that your *swc4* mutation results in premature polycomb silencing of *FLC*, thereby causing early flowering. Describe the predictions of this hypothesis with regard to the chromatin state and expression of *FLC* in 1-week-old wild type and mutant plants, and explain how you would test these predictions.

d) (5 points) *FLC* is normally silenced to allow flowering, but the gene must be reactivated during embryogenesis in order to repress flowering after seed germination. Given that *FLC* silencing occurs by polycomb-induced facultative heterochromatin formation, propose a model for how this could be reversed by a SWI/SNF remodeler to reactivate the gene during embryogenesis. Also describe an experiment to test this model.

Question #9 – Answer guide

a) (5 pts) The mutant could be outcrossed to a different (polymorphic) wild-type strain, and then this plant allowed to self-fertilize to generate a large number of progeny segregating for polymorphisms that differ between the two wild type strains as well as the early flowering mutation. Phenotyping of these progeny for early flowering and genotyping for polymorphic markers across the entire genome (either by classical marker-based methods or whole-genome sequencing) would allow one to associate the flowering phenotype with homozygosity in the region carrying the early flowering mutation. This would identify the region of the genome carrying the mutation. With sufficiently dense markers, one could narrow this region down to a single gene.

It is also possible to uses the other EMS-induced mutations across the genome as the polymorphic markers, rather than outcrossing to a polymorphic wild-type strain.

These are not the only ways, and any feasible approach will be acceptable

b) (5 pts)

Outside the transcribed region

1. Mutation in the core promoter: This could disrupt binding of the basal transcription machinery or RNA polymerase itself, thereby causing the gene not to be transcribed.

2. Mutation in an enhancer element: This could prevent binding of an activating transcription factor, resulting in the gene never being activated.

Inside the transcribed region

There are numerous possibilities here. It could be a mutation in an intronic enhancer (like #2), a mutation in a region of the RNA that would cause destabilization (gene is transcribed, but transcript is degraded rapidly), or a mutation that causes a premature stop codon in an exon and results in nonsense-mediated decay of the transcript.

There are other possibilities and any feasible answers will be acceptable

c) (5 pts) The hypothesis predicts that SWC4 is bound to the FLC gene in wild-type plants, that FLC will be expressed, and that the gene will be in a euchromatic state, carrying histone modifications such as H3K4me3. In contrast, in the mutant, no SWC4 will be bound to FLC, the gene will not be expressed, and chromatin will be devoid of euchromatic marks such as H3K4me3. Polycomb silencing marks (H3K27me3) will also be present in the mutant.

The predictions regarding expression can be tested by RT-PCR or RNA-seq on 1-week-old wild-type and mutant plants, and the chromatin predictions can be tested by ChIP-qPCR or ChIP-seq using antibodies against SWC4 and euchromatic and heterochromatic modifications.

it is not necessary to cite the specific histone modifications for full credit, as long the student notes that there are differences in these modifications between euchromatin and heterochromatin. They could also invoke nucleosome density, compaction, and/or chromatin accessibility as hallmark differences between euchromatin and heterochromatin

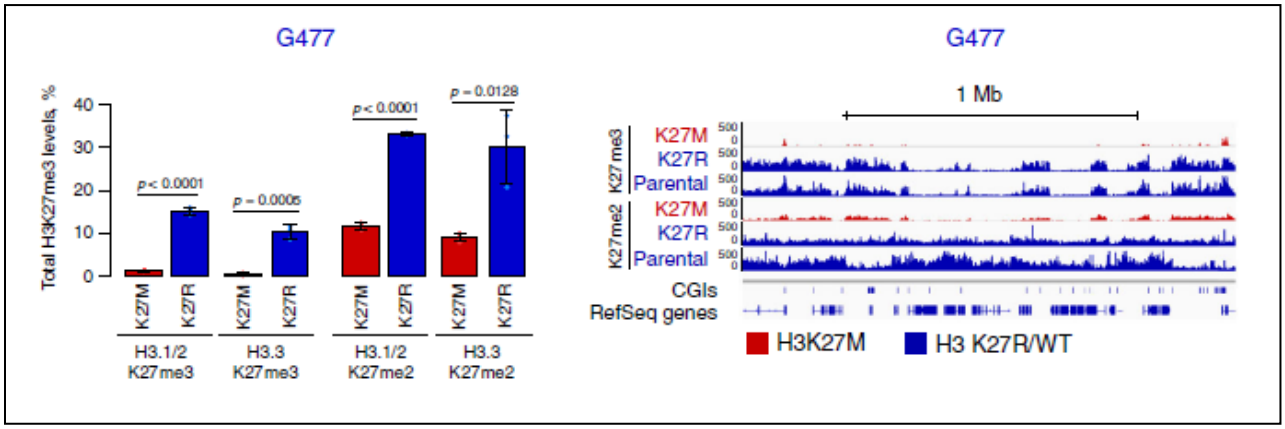
d) (5 pts) A plausible model is that the SWI/SNF remodeler (SWC4) is recruited to FLC during embryogenesis and ejects the polycomb complex (PRC2) from the locus. This would then allow the removal of polycomb marks and addition of euchromatic marks to activate the gene again.

One way to test this model would be to examine FLC expression and chromatin state (as in part 3) before, during, and after embryogenesis in wild-type plants and mutants for the remodeler.

Other models are possible but they must be feasible based on the biology of Arabidopsis and our current understanding of the interactions between remodelers and polycomb. In that case the experiments should fully test the model proposed.

Question #10

After taking IBS515 this year, you have developed an interest in biological methylation and have been reading up on oncohistones as you have gained an appreciation for how critical methylation is in regulating the function of histones. You are interested in determining the mechanism by which mutations in histone genes cause cancer.



a) (5 points) Design an experiment that would produce the results shown for G477 cells in this Figure. Describe the experimental approach, include appropriate controls, and a brief rationale for what question this experiment is designed to answer.

b) (4 points) When analyzing the results shown, you notice that unlike H3K27M, the H3K27R mutation does not exhibit many genomic differences in H3K27me2 or H3K27me3 in comparison to the parental cells. Present a feasible model to explain this result, including a discussion about the relevant properties of the amino acids.

c) (2 points) Based on the above experiment, you conclude that H3K27M reduces the genomic spread of H3K27me2 and H3K27me3. How do you propose that this change in histone H3 modification could alter gene expression to promote an oncogenic phenotype?

d) (5 points) Design an experiment to test your hypothesis in (c). Include appropriate controls and experimental detail as well as describing the results that would support or refute your hypothesis.

e) (4 points) Would you hypothesize oncohistones function in *cis*, or in *cis* and in *trans*, to exert changes to the genome and why? Explain how *cis* and *trans* regulation differ as part of your answer.

Question #10 – Answer guide

a) (5 pts) These results could be obtained from an experiment designed to address whether the H3K27M oncohistone affects genome-wide and locus-specific H3K27me2 and H3K27me3 (The paper this figure is derived from addresses whether H3K27M expression negatively impacts H3K27me2/3 spreading on chromatin). Approach = H3K27me3/H3K27me3 ChIP-seq. Neg control = parental cells that do not express the H3K27M oncohistone. Second Neg control = expression of a non-oncogenic histone mutation (H3K27R).

b) (4 pts) H3K27R is not an oncohistone. Discuss R group charges and structural similarities and differences between the parental amino acid (K) with the non-oncogenic mutation (R) and the oncogenic mutation (M).

c) (2 pts) Many possible theoretical explanations. H3K27M expression could inhibit the ability of chromatin modifying enzymes like the PRC2 complex to associate with chromatin and methylate it (H3K27me3). Because H3K27me3 is a repressive chromatin modification, by reducing H3K27me3, chromatin is no longer in a repressive state, leading to aberrant gene expression of genomic loci that fail to be silenced.

d) (5 pts) Depends on what students propose for (3). Answer will be graded based on ability to test hypothesis proposed in (3).

e) (4 pts)

Cis regulation = regulation of the chromatin that directly contains the H3K27M-mutant histone.

Trans regulation = regulation of chromatin that does not contain the H3K27M-mutant histone. Trans regulation could indicate the “other” histone within the same nucleosome (since nucleosomes are hetero-octamers and contain two H3 histones) OR could indicate different nucleosomes than the H3K27M-containing nucleosome.

Of note, there are 15 unique genes that encode H3 and oncogenic H3 mutation usually occurs in a single allele of a single H3 gene. This results in oncohistone-associated genomic changes occurring in cis and in trans, since most nucleosomes will NOT contain a (or two!) copies of the oncogenic H3 (e.g., H3K27M). This explains why known oncohistones are dominant mutations.