June, 2016

Bitter Taste Receptor Genotype/Phenotype Lab: A modular project sequence for lecture or lab

Raymond A Enke

Article

Connecting Common Genetic Polymorphisms to Protein Function: A Modular Project Sequence for Lecture or Lab[S]

Christopher E. Berndsen†,‡*
Byron H. Young†
Quinlin J. McCormick§
Raymond A. Enke‡,§*

From the †Department of Chemistry and Biochemistry, James Madison University, Virginia, ‡Center for Genome and Metagenome Studies, James Madison University, Virginia, §Department of Biology, James Madison University, Virginia

Supplementary Document 1: detailed lab protocol

Using SNPs to predict bitter-tasting ability

Our genomes are comprised of **~3.2 billion nucleotides** (nt) of double-stranded DNA. Your genome is ~99.9% identical to mine as well as everyone else in class and everyone else in the world for that matter. However, that 0.1% difference is still a significant number considering the massive size of our genome. Humans have ~3 million nt of variability compared to each other, many of these regions are **single nucleotide polymorphisms** or **SNPs**. SNPs are single nt changes in the genomic DNA sequence. Comparing the number of SNPs between individuals occurs in forensic science, paternity tasting, in determining the molecular basis of disease and finding even harmless variation in the human population. As the field of genomics expands our understanding of the similarities and differences between humans is growing. For this reason, the science of genetics is currently struggling with issues of access to genomic information, privacy, and how this information could be used to discriminate against people in getting access to insurance and other services.

Functionally, many SNPs occur in the vast non-protein coding regions of the genome and have no known affect on phenotype. Of SNPs found in protein-coding regions, some do not result in an amino acid change and therefore do not alter protein function (ie TC**T** vs TC**C**, the bolded SNP for this coding sequence in both cases codes for a serine amino acid). These SNPs are referred to as **synonymous**. SNPs resulting in an amino acid change are referred to as **non-synonymous**. Non-synonymous SNPs may or may not modulate functionality of a given protein.

This lab activity will focus on the ***TAS2R38*** gene, a gene that codes for one of the bitter taste receptor proteins. Humans can distinguish 5 basic tastes: sweet, sour, bitter, salty, and umami. These tastes are mediated by taste cells which express receptors on their cell surface which bind molecules in food. Upon forming a stable interaction with the receptor, a signaling pathway is initiated which results in a nervous system signal and ultimately the formulation of the "taste" of the food. Bitter taste is a product of approximately 30 receptors which bind various molecules. One protein receptor, TAS2R38, responds to a synthetic molecule called phenylthiocarbamide (PTC) as well as molecules found in some vegetables. Some portion of the human population can taste PTC, while others cannot. There are 3 co-occurring SNPs in the coding region of the *TAS2R38* gene that alter the ability to taste strong bitter flavors. Have you ever wondered why you hate broccoli and brussel sprouts or why you love super hoppy IPA beers? Could it be due to your dosage of the "taster" allele of the *TAS2R38* gene?

In this multi week lab exercise we will analyze our *TAS2R38* genotype and determine if your genetics is a good predictor or not of your bitter taste perception phenotype. We will also use this experiment as a vehicle for learning some basic skills in genome browsing, sequence editing and annotation, phylogenetic analysis and protein modeling.

Supplemental Document 1: Berndsen et al., 2016, BAMBED

***Informed Consent and Disclosure Statement**

Student participation in this experiment raises real-life **ethical questions** about the use of personal genetic data: What is my DNA sample being used for? Does my DNA type tell me anything about my life or health? Can my data be linked personally to me? There is consensus that a human DNA sample should be obtained only with the willing consent of a donor, who understands the purpose for which it is being collected. Thus, this experiment should be explained ahead of time and students given the **option to refrain from participating**. There is also consensus that a DNA sample be used only for the express purpose for which it is collected. Thus, student DNA samples will be **discarded after completing the experiment**.
The *TAS2R38* polymorphism has **no known relationship to disease states or sex determination**. *TAS2R38* alleles are inherited in a Mendelian fashion and can give indications about family relationships similar to knowing your ABO blood groups. Anyone wishing to not use their DNA in this experiment or to not participate in this experiment is welcome to complete a comparable alternate assignment agreed upon by the student and instructor. Please let the instructor know prior to lab if you elect this options.

Hybrid Molecular Biology Module #1: DNA Extraction, Amplification and Analysis

- DNA extraction: 30 min prep; 45 min incubation
- Agarose gel preparation: 15 min prep
- Computational analysis: 30 min
- PCR set up: 30 min

1. Extract DNA from human hairs and buccal rinses (must use gloves)

To test our genotypes, we will need some DNA from each of us. The root tips of pulled hairs and the cells that line the inside of our cheeks are very easily accessible cells that we can collect DNA from. We will use a very simple alkaline lysis technique to burst open our cells as well as denature and fragment our DNA for a PCR-based genotype assay.

PCR is extremely sensitive to contamination, all lab benches must be thoroughly wiped down with 70-95% ethanol prior to use, gloves must be worn at all time, and pre PCR pipettes must be used with filter tips

Protocol:

Before beginning, wearing gloves, clean bench top thoroughly with 70-95% ethanol

1. Wearing gloves, pluck one or several hairs from your head, arm, or any other sanitary part of your body. Ideally each member of the group will get a hair sample. Place the hair(s) onto separate unused kimwipe keeping track of who's hair is who's and which side of the hair is the root (there may be a barely visible white blob on the root end, those are cells from your scalp!). If you have no hairs to pluck or do not wish to use any of your hair for this experiment, skip step #1 - 4.

2. Cut ~1cm of the root portion of your hair(s) with sterile scissors and place them into a labeled PCR strip tube with sterile forceps. The label should have your initials and "H" for hair. FYI, 1 cm is about this much -------.
3. Add 75 uL of **Denature Buffer** to the sample (25mM NaOH and 0.2 mM EDTA). The NaOH is a base that will denature the DNA. EDTA is a chelator that absorbs metal ions in solution. This is to protect our sample against contaminating DNA nuclease enzymes that require metal ions.

4. Keep samples at room temp and proceed to step 5.

5. For buccal samples, label a 50 mL saline tube with your initials. Rinse with 10 mL of saline solution (salt water) for 45 seconds – 1 min and then spit the saline back into the plastic tube. Depending on how recently you ate, this could be gross. Ideally each group member will make a buccal sample as well. If you do not wish to use your buccal sample for this experiment, skip step #5 - 9.

6. Label a sterile 1.5 mL tube with your group#/color and initials and transfers 1 mL of your gross spit into the tube. Make sure to shake rinse tube up before the transfer to disrupt any settling that may have occurred.

7. When other groups are also ready, spin down your 1 mL buccal sample for 1 min at top speed using microfuges on back bench.

8. Immediately following the spin, carefully dump off the supernatant leaving your buccal pellet.

9. Resuspend the pellet in **75 uL Denature Buffer (DB).**

10. Create a **no DNA control** sample in one of the strip tubes by adding **75 uL Denature Buffer (DB).** Label this tube "no DNA" or "-".

11. Label your strip tube tops with #s and create a sample key in your notebook (ie sample #1 = RE hair, sample #2 = RE buccal, etc. On the back of the strip tubes put your group #/color.

10. When all groups are ready, place your closed strip tubes into the thermocycler on the back bench and start the program "Denature". This program is 98C for 45 min followed by holding at 12C.

The high heat will and NaOH base will burst the cells open spilling out the genomic DNA. This treatment will also denature and fragment DNA into smaller pieces making them readily available as template for a PCR reaction.

11. After 45 minutes heat cycle, remove your samples from the cycler and add 75 uL of **Neutralization buffer (NB)** to each tube (40 mM Tris HCl; pH 5.5). This will simply restore the pH of your prep to a neutral pH to preserve DNA.

12. Use these DNA lysates as template to set up PCRs in part 4.

2. Make a 2% gel

Cast 2% gels with with GelRed stain. Ideally several groups can load samples onto 1 gel. Note: GelRed is a nontoxic analog of ethidium bromide used to visualize DNA in a gel matrix under UV light.

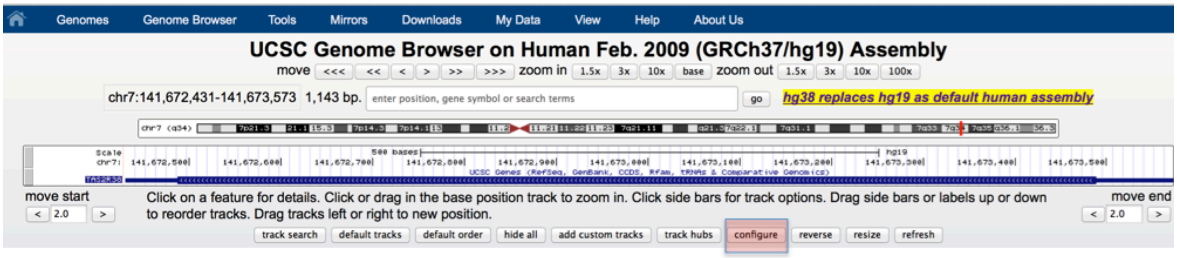
The following recipe and directions are specifically tailored for the Mupid ExU gel electrophoresis unit:

1. Mix **2 g agarose with 100 mL TBE buffer in a 500 mL flask** and swirl the flask a few times
2. Microwave the mixture for **2-3 min**, use heat gloves to swirl the flask again.
3. Let cool for 5 min and add **5 uL of GelRed** nucleic acid stain (non-toxic analog of ethidium bromide).
4. Set up a small gel casting tray with the 26 well comb
5. Use heat gloves to swirl the flask once more and pour into the casting tray with comb
6. Let stand for ~30 minutes to solidify.
7. Once solid, remove gel from the clear casting tray and wrap in plastic wrap.
8. Label gel with group info and date and store in the dark until next week. (must be in dark to protect the GelRed)

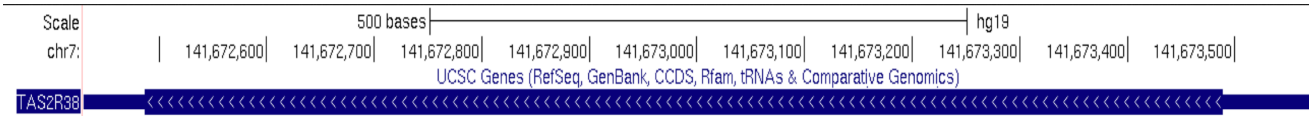
3. Obtaining *TAS2R38* genomic sequence from the UCSC Genome Browser

During the DNA extraction incubation and while waiting for your gel to solidify, obtain and annotate the human *TAS2R38* sequence using the UCSC Genome Browser and the nucleotide sequence editor A Plasmid Editor (ApE). Make a new page in your notebook to copy some information about the *TAS2R38* gene.

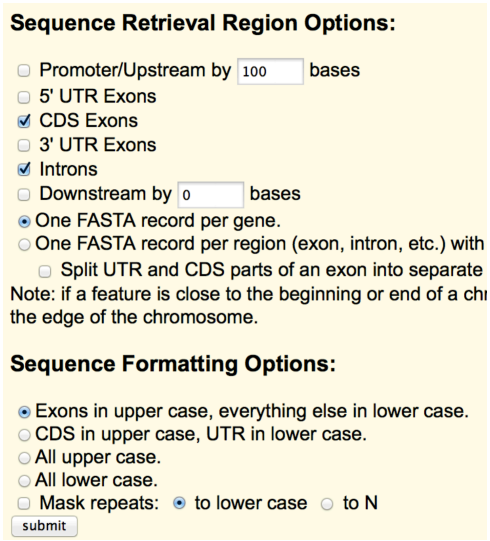
- Navigate to the human *TAS2R38* gene in the 2009 Hg19 assembly in the genome browser and hide all data except for UCSC genes in full view. This will display the *TAS2R38* gene's location in human chromosome 7



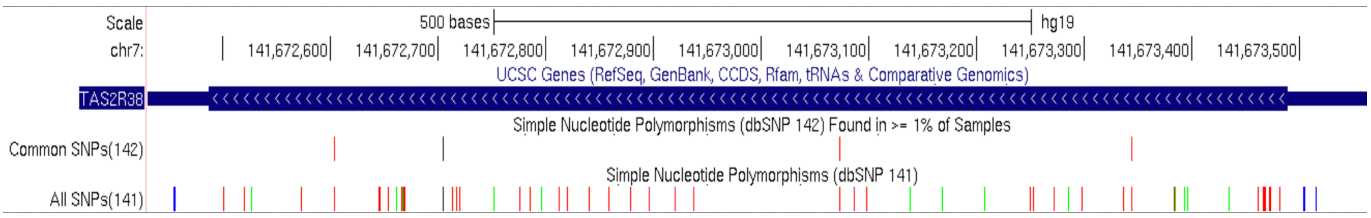
- Select the configure button below the genome viewer window to change the following display settings
  - Change the text size to 14 (will make all features larger)
  - Uncheck the "Show light blue vertical guidelines" box (to remove guidelines)
  - Hit submit to see your reformatted genome viewer window



- Next obtain the <u>DNA coding sequence only</u> (no 5' or 3' UTRs) for the *TAS2R38* gene by selecting the following options:



- Paste your output sequence into a MS Word document file saved as "human *TAS2R38* gene"
- Navigate back to the genome viewer window
- There are several options for viewing SNP data in the browser. Under the **Variation** tools, select **Common SNPs (142)** with dense view and **All SNPs (141)** with dense view.
- Change the color scheme for both data tracks by selecting track links>**Color options** so that all non-synonymous SNPs are red and all other types of SNPs are black.



These 2 tracks give you a view of all knows SNPs and SNPs that occur in >1% of the population respectively. The link for each of these tracks will tell you more about the data and how they are displayed. The color-coding of the SNPs gives you an instant readout of some potentially functional SNPs:

- How many common SNPs in the population could potentially lead to an amino acid change that alters the function of the protein coded by the *TAS2R38* gene?
- Select the track link for UCSC Genes then the link for "help on coding color" to determine the significance of color coding within the UCSC gene annotation. Based on the position of the stop codon, which strand of DNA is *TAS2R38* encoded on?

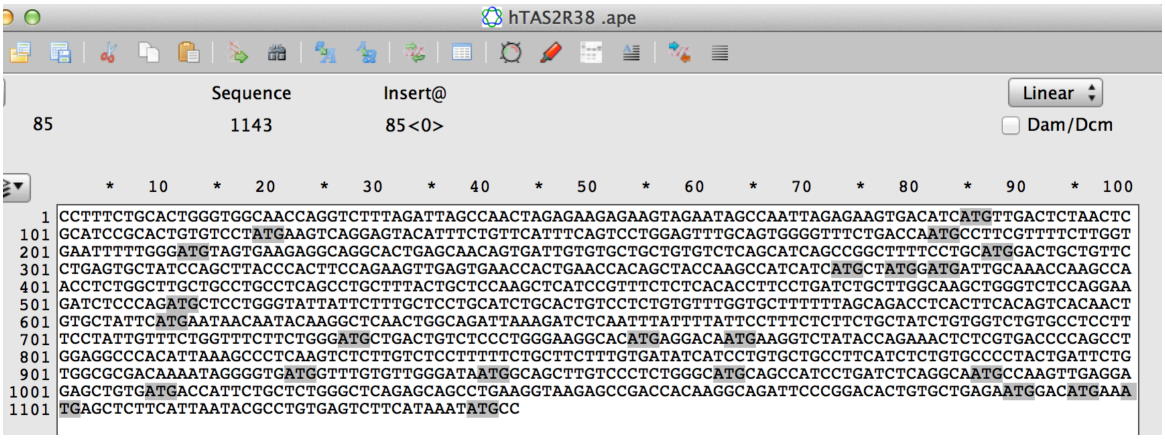
4. Annotating *TAS2R38* genomic DNA using the sequence editing software A Plasmid Editor (ApE)

ApE is a free nucleic acids sequence analysis software package developed by Wayne Davis at the University of Utah. The programed is installed on all of the #3033 lab computers but you will also need to install it on a computer that you can use outside of class. It can also be easily downloaded at http://biologylabs.utah.edu/jorgensen/wayned/ape/ and installed onto your personal computers (**note**: there are slightly different installation instructions for **Mac users**). Today we will use a primer set to **PCR amplify** a short region of your *TAS2R38* gene as the first step in determining the genotype of a common *TAS2R38* SNP. Below is a quick overview on how to use basic features in ApE.

Creating a new sequence file in ApE

Open the ApE software. The easiest way to get your sequence(s) into ApE is to copy/paste from a genome browser or elsewhere. Be sure to paste only nucleotide sequence and the FASTA text sequence tags. You'll notice the software warns you if you paste in illegal letters (ie, non-ATGCs) and will remove them. Copy/paste the human *TAS2R38* sequence into a new ApE file and save it as "hTAS2R38" to your desktop.

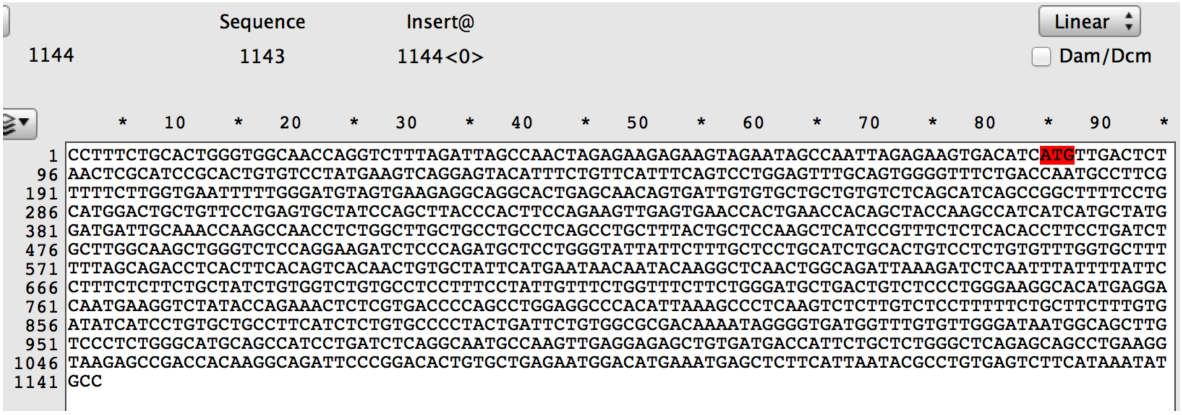
To find particular sequences**,** press "**Command F**" or click on the "**binoculars" icon** or select "**Find**" under the "**Edit**" menu. Input the sequence you're looking for (type or copy/paste) into the search field and click "**Find next**" to find the 1st occurrence of your search. Alternatively, you can select "**highlight all**" to find all occurrences of your query sequence. Use this feature to search for possible ATG start codons (**find "ATG" and "highlight all"**). These highlights are temporary and will not be saved in your file. Select **Edit>>>Clear Find Highlighted** to remove the highlights. The 1st ATG in a gene sometimes but not always codes for the starting Met amino acid codon. Place the cursor in front of the 1st ATG. Notice at the top of the sequence window there are numbers indicating the length of the entire sequence as well as the position of the cursor within your sequence. Note the position of the 1st ATG in your notebook.



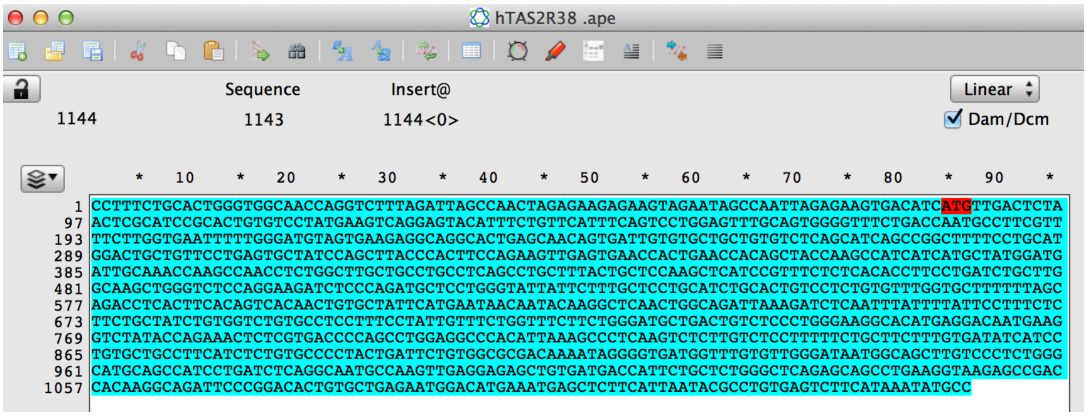
Annotating sequence features

A DNA sequence can look like nothing more than a bunch of random As Cs Gs and Ts. To make sense of them, sequences can be **annotated**, or highlight so that certain features stand out as points of reference. To annotate a portion of sequence, search for or highlight a portion of sequence with the cursor then select **Features>>>New Feature**. Give the feature a name and select a **Forward color** to highlight your sequence. (*Note, for annotating features on the reverse or

complimentary strand, such as reverse primers, select the "**Rev-Com**" option on the top right of the **edit feature** screen). These annotations can be saved to your sequence. First, highlight all ATGs again and then annotate the 1st ATG in red naming the feature "start codon". Clear the highlighted sequence leaving you with only the 1st annotated ATG in red:



As a further example, lets also annotate exons and intros. Hopefully by now you've figured out that *TAS2R38* is a single exon gene with no introns. Annotate exon 1 (the entire sequence) in light blue. Notice that your ATG annotation has been covered up by the exon 1 annotation. Fix this by selecting **Features>>>Edit Features**. Use the "Raise" and "Lower" options to order your annotations. Lower the exon 1 annotation so that the ATG annotation is on top and select OK. You should now be able to view both annotations with the ATG annotation in red sitting atop the exon 1 annotation in blue:

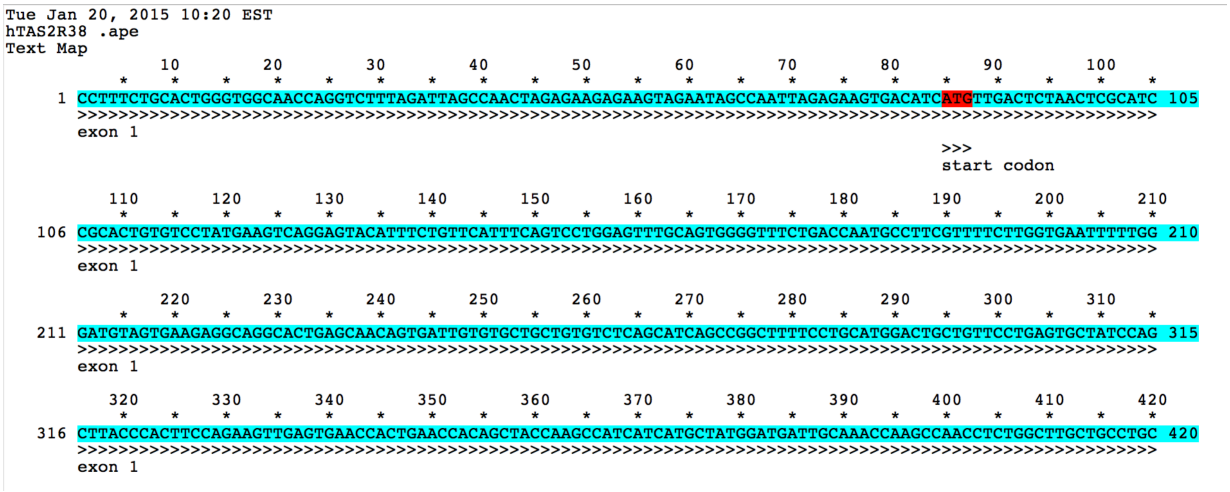


We will use this simple annotation feature quite a bit, particularly for annotating primer information into our sequences for PCR analysis.

Text Map Annotation View

To view and print detailed annotated sequences you will use the **Text Map View** in ApE. Select **Enzymes>>>Text Map** or by select the "**text map**" icon in the tool bar above the sequence (3rd icon from the left; see below image). Keep default setting for the configurations menu and hit **OK**.

This command gives you your sequence + annotations in a printable format. Right click to print in ApE OR save your annotated text map sequence to a Word document as a screenshot and print from Word. This is a handy trick if the computer you're printing from does not have ApE installed. Here is an example of **Text Map View**:



There are a number of other features not described in this exercise that you can explore on your own if you like.

- In your *TAS2R38* ApE sequence, annotate the F + R primers sequences that we will use for our genotyping assay on their correct strands. The primers sequences we will use are:

TAS2R38 Forward primer: 5'-CCTTCGTTTTCTTGGTGAATTTTTGGGATGTAGTGAAGAGGCGG-3'
**2nd to last G is a mismatch with TAS2R38 gene sequence*

TAS2R38 Reverse primer: 5'-AGGTTGGCTTGGTTTGCAATCATC-3'

- Please note that the 2nd to last nt (G) of the Forward primer does not match the endogenous *TAS2R38* gene sequence (its not a SNP either).
- When searching for your primers, make sure that you select the **"allow mismatches:"** and select **1** allowed mismatch to find this primer sequence.

This mismatch primer will allow us to **create a restriction enzyme site** in PCR products with a certain SNP. This is called **dCAPS PCR analysis**, which you will research a bit more outside of class.

- Annotate the F primer in blue
- Annotate the mismatched G in red naming it "G mismatch in primer"
- Annotate the R primer in green
- Make sure that you select the **"also find the Rev-com of string selection"** to highlight and annotate the R primer. Also make sure you annotate the R primer on the correct strand.
- What does dCAPS PCR analysis allow create for use in genetic analysis?
- What nt is the G mismatch in your sequence?
- What is the expected size of the PCR amplicon
- Take a screen shot of the Text Map View of your annotated sequence and copy it into your notebook. It should look like this:

When you finish annotating your sequence in ApE, begin gathering reagents to set up your PCR reactions.

5. Set up genotyping PCRs (must use gloves)

Many of you have used PCR to amplify segments of genomic DNA. This is a fundamental technique used in many facets of molecular biology, genetics, and genomics. You will use your hair and buccal DNA lysates from part 1 as template to set up *TAS2R38* genotyping PCRs. PCR is essentially a cell-free *in vitro* DNA replication reaction. **DNA polymerase** will copy a template DNA sample given the correct buffering conditions, some metal ions ($MgCl_2$), and a primer sequence that determines the specificity of what DNA pol will copy. The reaction occurs during 3 temperature-regulated steps:

1. Denaturation: **94-98C**; denatures double-stranded DNA duplexes
2. Primer annealing: **54-64C**; allows DNA primers to anneal to single-stranded DNA
3. Polymerase extension: **72C**; allows DNA pol to synthesize and extend new DNA off of a DNA/primer duplex template

Due to its thermostable properties, *Thermus aquaticus* (*Taq)* DNA Polymerase is an enzyme widely used in PCR. It can be used for many cycles of amplification including a high heat DNA denaturation step that would kill most enzymes.

Today we will amplify a short segment of the human *TAS2R38* gene that encompasses 1 of the SNPs associated with the ability to taste the bitter PTC chemical from our own DNA samples. Upon completion of a successful PCR reaction, each tube will have millions of copies of the short genomic region. We will use a restriction enzyme called H*aeIII* to differentially cut amplicons that have the tasters vs non-tasters SNP.

Setting up PCR reactions

**Use gloves for set up and take extra care not to contaminate reactions*
**Keep DNA, taq mix and other PCR reagents on ice and on cold plates at all times during set up*

Before beginning, each group needs ice, gloves, and the Bio 480 pipettes in the drawers

1. Thaw 2x Taq Mix, dH2O, human DNAs, MgCl2, and *TAS2R38* primers

2. Use an Excel spreadsheet to set up a PCR Master Mix. (Note: the 2X PCR buffer contains Taq DNA pol, pol buffer, 3 mM $MgCl_2$, and dNTPs.)

3. Each group will make 1 master mix for all their samples as well as a negative control. Calculate the amount of reagents needed to set up enough PCR reactions for all of your samples. Always calculate your master mix for 1-2 more samples than you need. Here is an example of my mix for setting up 8

PCR reactions:

PCR set up spreadsheet			
Using Sigma's 2X JumpStart ready mix; 20 uL reaction volumes (MgCl2 concentration is 1.5 mM)			
the 2X mix contains Taq polymerase, polymerase buffer, MgCl2, and dNTPs. Dilute to 1X in a 20 uL reaction with your DNA, primers, additional MgCl2, and H2O			
Individual reaction volumes		Master mix volumes	
(20 ul vol)	**master mix X**	**10**	
10	2X PCR mix	1	
1	F/R primer (10 uM)	10	
0.4	MgCl2 (25 mM)	4	(for 2.0 mM final)
6.6	H2O	66	
2	DNA	*add 2 uL of each sample to individual PCR reaction tubes	

4. Since we are all using the same primers, each group can make 1 master mix to accommodate all of the group samples. Add the reagents to your master mix one at a time keeping the mix and reagents on ice. Do not add DNA to the mix, this is the variable portion of the set up. We will add DNA last when everyone is ready to start their PCR reactions.

5. Label PCR strip tubes and after master mix is completed, gently mix by pipetting (no vortexing) and aliquot 18 uL to each of your strip tubes on a cold block.

6. Wait until the entire class is ready, then add 2 uL of your DNA or control samples to each reaction. The 2 uL DNA will make each reaction equal a 20 uL volume.

7. Cap tubes and keep on cold block until entire class is ready to start PCR thermocycler. When ready, place tubes in cycler and run "2X 58" PCR program. This program will run the following cycle:

1. 94C - 1 min (initial denaturation step)
2. 94C - 30 sec (denaturation step)
3. 58C - 30 sec (primer annealing step)
4. 72C – 1 min (DNA polymerase extension step
5. Go to step #2 X 34 more times (total of 35 cycles)
6. 72C - 3 min (final pol extension step)
7. Hold at 12C

This program take ~ 2 hours and should create millions of copies of the short *TAS2R28* gene segment from your genomic DNA template. After cycling, I will store your DNAs in the -20 freezer until next week.

Module #1 assignment (group assignment; due in lab next week):

Using today's lab and several cited references, write a mini proposal outlining your genotyping/phenotype experiment. The page limit is 2 pages double-spaced. This is a group assignment (turn in 1 hard copy per group). Here are the guidelines:

1. Thoughtful and informative title (1 pt)

2. Background information (4 pts)
-what does the reader need to know to understand the hypothesis?
3. A hypothesis statement (2 pts)
4. Methods (2 pts)
-Briefly explain how you will test your hypothesis
-How will you analyze and interpret your results
-Briefly explain CAPS and dCAPS PCR analysis genotyping and which we are using in this activity
5. Literature Cited (1 pt)
-must cite at least 2 peer-reviewed citation from the literature

Hybrid Molecular Biology Module #2: Genotype/Phenotype Determination

- RFLP analysis: 15 min prep; 10 min incubation
- Gel electrophoresis: 5 min prep; 30 min incubation; 10 min analysis
- Phenotype analysis: 20 min
- Computational analysis: 45 min

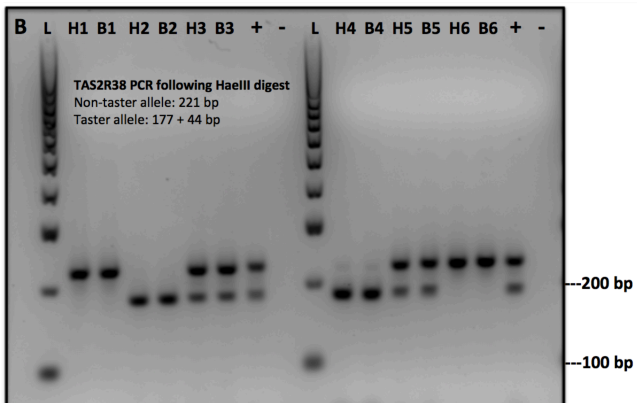
Last week you amplified a short segment of the ***TAS2R38*** gene encoding one of the bitter taste receptor proteins. This short segment contains 1 non-synonymous SNP that alters the functionality of the TAS2R38 protein and an individual's ability to taste the bitter PTC compound. By using a dCAPS primer set, the PCR fragments amplified from individuals with the "Taster" allele (T) will have a HaeIII restriction enzyme site while the amplified fragments from individuals with the "Non-taster" (Nt) alleles will not have the restriction site.

Today we will digest our PCR fragments amplified from last week and then run them on a 2% gel to determine if our genotypes are:
1. T/T homozygous
2. T/Nt heterozygous
3. Nt/Nt homozygous

We will also take a PTC bitter taste test to determine our ability to taste the bitter PTC compound. Our phenotype will fall into 1 of 3 categories:
1. Strong taster (S); extremely strong gross bitter taste
2. Weak to moderate taster (T); a not so gross weak to moderate bitter taste
3. Non-taster (N); no bitter taste sensation at all

Lastly, we will use bioinformatics analysis to gather information about the Taster and Non-taster alleles of the *TAS2R38* gene in humans and several other species.

1. HaeIII restriction digests



The dCAPS PCR we set up last week will introduce a HaeIII restriction enzyme site into the Taster allele but not the Non-taster allele. HaeIII cuts at the sequence 5'-GGCC-3'. We will be able to determine the class' genotype by digesting our PCR products and running the result on a 2% agarose gel:

Your PCR was set up in a 20 uL reaction. You will make a 10 uL restriction enzyme reaction mix and add it directly to your PCR tubes in a 30 uL final volume. You will add the following to each reaction:

- 10X enzyme buffer
- HaeIII restriction enzyme
- H2O

1. Use the "Restriction Enzyme set up" spreadsheet in the Lab Modules Tab to make a master mix for your enzyme reactions. Record your information in your notebook and make your mix. **Do not vortex your enzyme mix.** Gently pipet up and down to mix

Note: You will make a mix calculated for a 30 uL final volume (20 uL PCR reaction + 10 uL of restriction enzyme mix).

2. **Add 10 uL** of your enzyme mix to each PCR reaction and mix by gentle up and down pipetting.

3. Cap PCR tubes and incubate in 37°C incubator for **15 min**.

4. Store at room temp until ready to run on gel.

2. Bitter taste test (get a drink of water before doing the taste test)

The restriction enzyme digestion of our PCR products will determine our genotype for one of the SNPs in the *TAS2R38* gene that effects sensitivity to bitter taste. We will also determine our phenotype for bitter taste sensitivity by taking a PTC taste test.

Sensitivity to bitter taste can be assessed using paper strips laced with the bitter compound PhenylThioCarbamide (PTC). Make sure you take a drink of water prior to taking the taste test (PTC is more difficult to taste in a dry mouth). Place the entire strip on your tongue working it to the back where your bitter taste buds are. Score your bitter taste phenotype on the following scale:

Strong Taster (ST) = intense bitter taste
Weak Taster (WT) = weak to moderate bitter taste
Non-Taster (NT) = no bitter taste at all

Record your bitter taste phenotype in the experiment Google Spreadsheet as well as your notebook:
tinyurl.com/ppx4t

3. Gel electrophoresis of dCAPS PCR/restriction digest

When the restriction digest incubation is finished, use the 2% agarose that your table cast last week to analyze your digested PCR products. 2 groups at each table will run their PCRs on the same gel. Record the order that you will load your samples and transfer this information to your notebook. For example:

1	2	3	4	5	6	7	8	9
100 bp marker	Initials hair	Initials buccal	Initials hair	Initials buccal	Initials hair	Initials buccal	negative control	

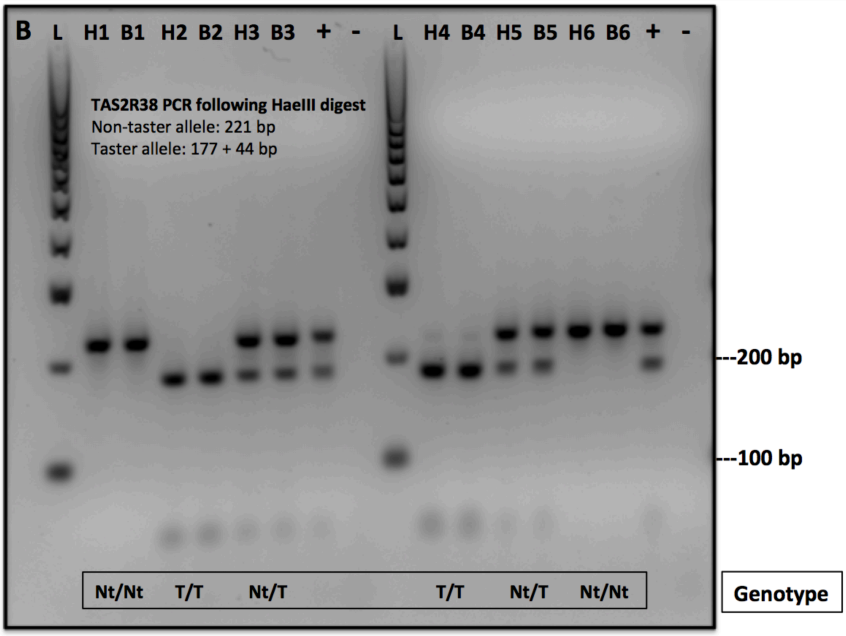
To run the gel:

- Fill the gel tank with ~400 mL of 1X TBE buffer and place the solidified gel (in the translucent tray) into the tank. Make sure the wells of the gel face up.
- Make sure the gel is completely submerged, if not add more buffer as needed.

- Add **5 uL of 6X gel dye** to each of your digested PCR products.
- Add **5 uL of 100 bp ladder** to lane #1 on the gel
- Load **10 uL of each digested PCR sample** starting with lane 2.
- Load **5 uL of 100 bp ladder** between the 2 sets of group's samples
- Once all lanes are loaded, put the lid on the gel box, set it to 135 V and run the gel for 30 min.

The **GelRed stain** added to the gel will fluoresce under ultraviolet (UV) light only when it is bound to DNA (or RNA) similar to ethidium bromide. We will image our gels using a UV transilluminator attached to a digital camera.

- Photo gel under UV light and save/export image for labeling and documentation into your lab notebook. Use the below example image for your gel annotation including primer name and group color/number as well as lane numbering and indication of your negative H20 control:



- Rinse the gel box, casting trays, and combs with tap water
- Record formatted gel image and experiment info in your lab notebook

Determine your TAS2R38 genotype by assessing the bands on the gel. Record individual genotypes in your notebook as well as on the the experiment Google Spreadsheet:
tinyurl.com/ppx4tpp

Compile and correlate the class genotype and phenotype data using the table below. Make a representative table in Excel and enter this data into your notebook.

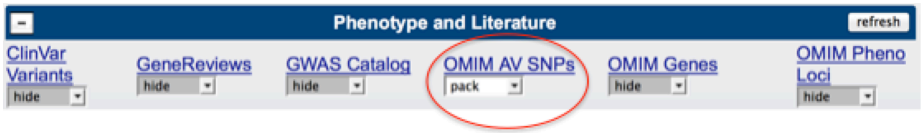
	Phenotype		
Genotype	**Strong taster**	**Weak taster**	**Non-taster**
TT (homozygous)			
Tt (heterozygous)			
tt (homozygous			

Discussion questions to be addressed in your lab report:

1. According to your class results, how well does *TAS2R38* genotype predict PTC-tasting phenotype? What does this tell you about classical dominant/recessive inheritance?
2. How does the *Hae*III enzyme discriminate between the C-G polymorphism in the *TAS2R38* gene?
3. The forward primer used in this experiment incorporates part of the *Hae*III recognition site, GGCC. How is this different from the sequence of the human *TAS2R38* gene? What characteristic of the PCR reaction allows the primer sequence to "override" the natural gene sequence?
4. What ethical issues are raised by human DNA typing experiments?

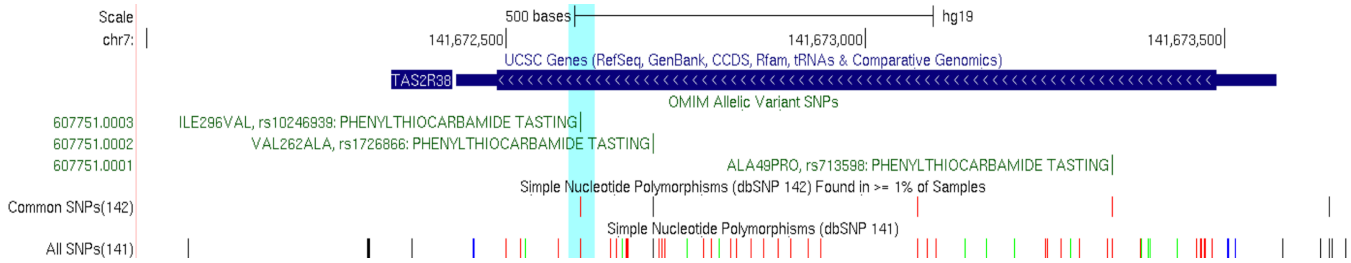
4. OMIM database

The Online Mendelian Inheritance in Man (OMIM) database is a collection of SNPs and mutations associated with human disease or altered phenotype. OMIM can be accessed directly through the NCBI (National Center for Biotechnology Information) homepage. There is also an OMIM SNPs data track in the UCSC Genome Browser under the Phenotype and Literature menu.



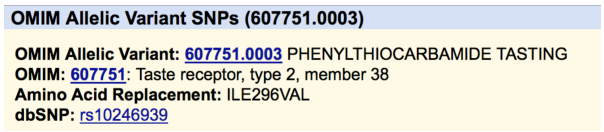
Last week you explored the All SNPs and Common SNPs associated with the *TAS2R38* gene. Another useful data track are **OMIM SNPs** under the **Phenotype and Literature** tools.

- Select **OMIM AV SNPs** with full view



Select any of the OMIM SNPs in the browser window to navigate to a **SNP index page**. Here information about each SNP can be gathered such as the specific amino acid change associated with the SNP. There are also links to the OMIM database to gather additional information.

- Select the OMIM Allelic Variant link to identify the exact nucleotide position of SNPs associated with the *TAS2R38* gene
- Select the OMIM link to collect general information about the *TAS2R38* gene and OMIM SNPs.



Use information in the OMIM database to help construct your **Introduction, Results, and Discussion** sections of your lab report. Some specific information to include in your **Results** section:

- How many common SNPs are found in the *TAS2R38* gene? How many of these result in a phenotypic change (ie how many are in the OMIM database)?
- What are the positions and sequence changes of the OMIM SNPs?
- What is the phenotypic change associated with these SNPs?
- Annotate all OMIM SNPs in red in your ApE sequence indicating which genotype is associated with tasting and none tasting (use the table below in section 5 for help).

5. Sequence Alignments in ApE

The 3 the *TAS2R38* OMIM SNPs are usually inherited as a unit, or a **haplotype** with the C-G-G haplotype correlating most strongly with bitter tasting ability:

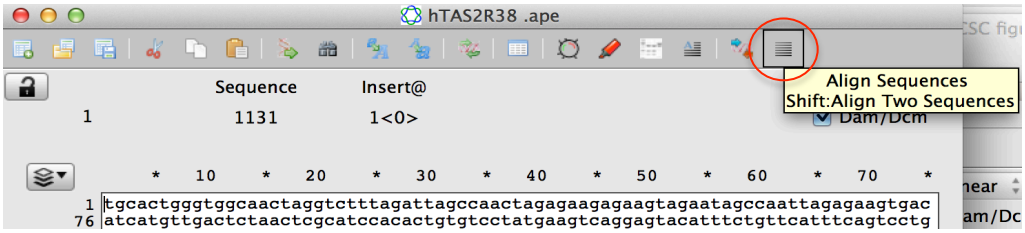
Nucleotide Position	Taster		Non-taster	
	Codon	Amino Acid	Codon	Amino Acid
145	**C**CA	proline	**G**CA	alanine
785	G**C**T	alanine	G**T**T	valine
886	**G**TC	valine	**A**TC	isoleucine

Sequences can be compared or **aligned** in order to understand the relationship among conserved nucleic acids and proteins found in within the same organism or in different organisms. We will use ApE to compare the human Taster and Non-taster allele to each other and then to some other primate species.

- Open your annotated human *TAS2R38* sequence
- Is the consensus sequence provided in the browser representative of a Taster or Non-taster allele?
- Create 2 new ApE sequence files named "human taster allele" and "human non-taster allele". Make sure that the 3 OMIM SNPs are correctly changed for each respective sequence

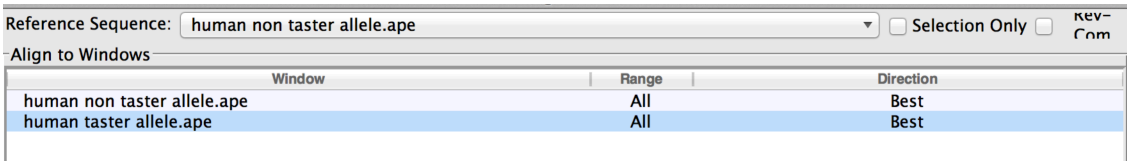
Sequence alignments can be performed in ApE.

- **Tools>>>Align Sequences** or by selecting the Align Sequences icon in the ApE toolbar

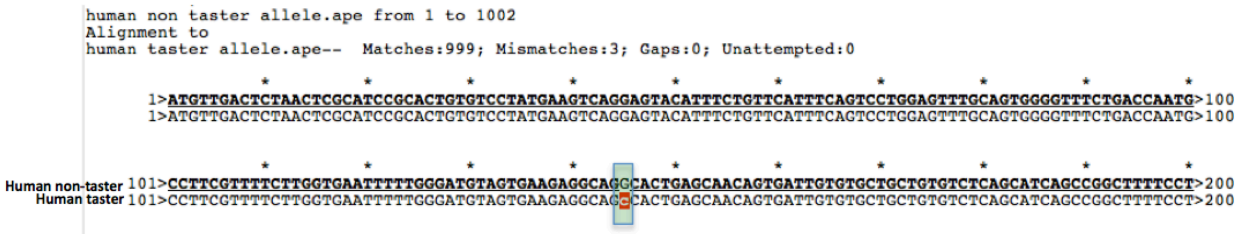


- Select "human non-taster allele" for the reference sequence
- Select "human taster allele" in the align to Window and select OK

These commands will tell ApE to use the non-taster sequence as the reference to compare the taster sequence to.



The resulting output will have the reference sequence on top in bold (non-taster) and the aligned sequence(s) below it. Differences between the sequence are highlighted in red. The below image is showing the 1st 200 nt of the alignment including SNP #145.



Note: The sequence label and the green box around SNP #145 in MS PowerPoint.

In order to investigate the evolution of the *TAS2R38* gene you will align sequences from several other primates with the human alleles.

- Use the genome browser to make new ApE sequence files for the *TAS2R38* gene from chimpanzee, baboon, and gorilla
- Name the files "Chimp *TAS2R38",* "Baboon *TAS2R38",* "Gorilla *TAS2R38"*
- Create a sequence alignment in ApE using the human non-taster allele as the reference sequence and the chimp, baboon, gorilla, and human taster allele as the alignment sequences.

Take a screenshot of the resulting alignment and paste in into a PowerPoint file. Add in sequence labels and boxes around SNPs #145, #785, and #886 as in the example above.

Questions to address in your lab report **Discussion** section:

1. Based on the sequence alignment, is it likely that chimps, baboons, and gorillas are PTC tasters or non-tasters?
2. Did the non-taster allele likely arise before or after the human lineage split from other primate species?

Protein modeling analysis of TAS2R38 structure and function

The structure-function relationship in biochemistry is well founded but the structure side of the equation is missing for many human proteins. Protein structure prediction is playing an increasing role in the how scientists develop hypotheses and theories on the structural basis for functional defects. This is due to both the increase in computational power and the low price to do experiments. When many types of prediction are combined this type of analysis can be useful however, it is important understand that the results are predictions of what are possible.

Persons with the *TAS2R38* taster allele encode with P49, A262, and V296 in their TAS2R38 receptor can taste PTC, while person with the A49, V262, and I296 in their TAS2R38 receptor are less sensitive or cannot taste PTC. The biochemical connection between the protein sequence and the ability to taste PTC however is not known. In this activity your group will hypothesize a basis for PTC tasting and provide evidence to support your hypothesis using protein structural modeling, docking studies, structural prediction, and statistics. Other forms of analysis are possible and encouraged but should be cleared with your instructor to avoid futile work. Your group may consult with your instructor at any time if you reach a point where you are unsure how to proceed.

While we will have some in-class time to work on various aspects of the project, much of this project will require work outside of class. Some of the servers will take days to provide data, so please do not save this for the last minute!

Instructor note*: Protein modeling activities are designed for open-ended student exploration of databases listed in each module. We have provided one specific protocol for each protein modeling module as an example activity.*

Computational Module #1: Protein Sequence Analysis and Prediction

Some features of protein structure can be predicted based on the primary sequence. These features include ligand/substrate binding sites, cellular localization sequences, and secondary structure. Use several of the below resources to identify SNP

consequences to the primary amino acid protein sequence, catalog information on SNP changes to the TAS2R38 protein, and identify sequence motifs and localization features of the protein.

Protein Sequence Analysis		
Program/Website Name	*Description*	*Website*
JPred4	Predicts secondary structure from alignments with homologs	http://www.compbio.dundee.ac.uk/jpred/
PSI-Pred	Comprehensive secondary structure prediction and motif/domain search. Also predicts transmembrane sections.	http://bioinf.cs.ucl.ac.uk/psipred/
ExPASy	A large server of bioinformatics program for structure and function prediction. Recommended for use.	http://www.expasy.org/
ConSurf	Predicts functional regions based on sequence alignment or structure	http://consurf.tau.ac.il/verify.php
Translate	Converts DNA sequence to Protein sequence	http://ca.expasy.org/tools/dna.html
Domain Mapping of Disease Mutations	Database of known mutations that result in a polymorphism or disease	http://bioinf.umbc.edu/dmdm/
BioMuta	Database of known mutations that result in a polymorphism or disease	https://hive.biochemistry.gwu.edu/tools/biomuta/
UniProt	Database of protein sequence and functional information. A good starting point for students to obtain a broad overview of the protein structure and function	http://www.uniprot.org/

Example Activity: Identifying domains and motifs from a protein sequence using the ExPASy Bioinformatics portal

Protein structure is defined by the primary sequence of amino acids. Therefore the amino acid sequence of a protein can provide clues about structural domains and motifs which may suggest function. While sequence based methods are not a substitute for functional assays and determining the structure of the protein directly, it is a starting place for building a hypothesis to test with future experiments.

Protein sequence analysis can proceed through several methods. As a first pass, a simple query of sequence databases using programs like BLAST can be conducted to identify structure based on similarity to known proteins. Alternatively, searching within the sequence for patterns that match canonical motifs or domains such as the CoA binding motif found in acetyltransferases or the calcium binding motif EF-hand. The best strategy is often a combination of both methods. In this activity you will learn to navigate NCBI BLAST and the ExPASy Bioinformatics portal to predict functional domains. You will also use UniProt Knowledge Base as a resource for gathering information on a protein sequence.

Most programs require a sequence in FASTA format. FASTA format contains a line of information preceded by a ">". Any information on this line is a description for the user and the program will largely ignore this information except to refer to the sequence. Below the ">" is the amino acid sequence in one letter code, typically in all uppercase.

Example of FASTA:
>Example protein information
MPRTRERIMQQERFGETE

To obtain your sequence, navigate to the UniProt website: http://www.uniprot.org/
Uniprot is database that aggregates information from several sources to provide an introduction to proteins. In the search box, type in TAS2R38 and begin the database search. The results show all the deposited protein sequences that match the defined criteria. Select the proper page for TAS2R38 from the list of search results (be sure to verify organism, gene name, and protein length). The page on this protein shows a wide variety of information from proposed function and the reliability of that prediction to sequence and structure. Click on sequence to view. Click the FASTA button to output the sequence selecting to view it in a new tab for toggling between the Uniprot page and the sequence page easily. In a 3rd tab, navigate to NCBI protein BLAST: http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins

Copy the FASTA sequence into the Query sequence box at the top left. Search the nonredundant protein sequences database and set the algorithm to blastp. Click BLAST (Basic Local Alignment Search Tool) to start the search. This can take from several minutes up to an hour depending on server load and the complexity of the sequence. When BLAST finishes, note the conserved domains suggested by BLAST and compare to those under the Family & Domains header on the UniProt. Do similar results appear? On the BLAST page, scroll down to the Descriptions box and look at the proposed functions/protein names listed. Are there any trends or consistent terms?

Scroll down further to see the sequence alignments for each protein and your sequence. Do the alignments cover the entire sequence or just a small section? What do you think the TAS2R38 protein does based on the domain information from UniProt and BLAST? Why? What is your predicted biochemical basis for the phenotype change in taster vs non-tasters based on these observations?

Computational Module #2: Structural Modeling

For proteins with an unknown structure, a homology model is useful to see how the amino acids may orient in 3D space. Homology models use known structures of similar sequence to generate a proposed tertiary structure for a protein of unknown structure. Use the below homology modeling resources to predict and compare the structure of taster and non-taster TAS2R38 protein folds.

Homology Modeling		
Program/Website Name	*Description*	*Website*
SwissModel	Online homology modeling	http://swissmodel.expasy.org/
I-TASSER	Structure and function prediction as well as ligand binding prediction. Also has many additional web-based programs for structure and function prediction.	http://zhanglab.ccmb.med.umich.edu/I-TASSER/
PHYRE2	Web-based protein structure prediction interfaced for beginner users	http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index

Example Activity: Homology modeling using Phyre^2

Homology modeling can be applied to investigate protein structure for experimental protein that have no previously described structure. This analysis uses sequence homology to template proteins with known structure in order to synthesize structure predictions for uncharacterized sequences. Homology models are often built using multiple templates in order to improve its accuracy. This technique is valid because proteins of similar sequence and function often have similar structures (though this is not always the case). General chemistry principles (i.e. electrostatics, Van der Waals) also used to improve structure prediction. This exercise utilizes the Phyre^2 online homology model generating server in order to build a homology model for the TAS2R38 protein.

To begin, obtain the amino acid sequence of the human TAS2R38 protein in FASTA format using either the UCSC Genome Browser or Uniprot. To begin building a homology model navigate to the Phyre^2 (Protein Homology/analogy Recognition Engine V 2.0): http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index

Enter your email address and job description in the respective boxes. Paste the copied TAS2R38 amino acid sequence into the appropriate box, choose normal modeling mode, and click “phyre search”. Allow for 2-6 hours run time for homology modeling depending on server traffic. Once your model is complete, you will be sent an email containing a link to the webpage containing your model and templates used to build it. Click on the cartoon representation of the model in order to download the .pdb file for the model (it will be called “final.casp.pdb”). In the lower half of the webpage you will find a list of the templates used in the modeling. In the “Template” column of this list, the middle four letters of the alphanumeric code listed are the PDB code (i.e. 1P0H).

Navigate to the RCSB protein data bank (PDB) at http://www.rcsb.org/pdb/home/home.do and input the PDB codes for the template models into the search bar. The RCSB search results will provide structural and functional data for each template used. These will provide information about potential function of the modeled protein. Investigate the structures and functions of these templates and keep these in mind when analyzing your model.

In order to analyze your homology model, you will need to look at its structure in three dimensions. To do so, open the .pdb file of your TAS2R38 homology model in PyMOL or another .pdb file-viewer program. Use these modeling data to develop hypotheses as to how the structure of your protein dictates its function. If using PyMOL, commands such as “show cartoon”, “show surf”, “show mesh”, and the operations on the right hand side of the window will aid you in this

investigation. Since TAS2R38 may have a function similar to that of the templates used, investigate the functions of those proteins using Uniprot and try to draw conclusions based on structural similarities.

Computational Module #3: Substrate/Ligand Docking

Modulation of protein/ligand interactions provides an obvious way for functional differences to translate into altered phenotype. Use the below protein visualization databases to identify and compare ligand binding pockets of taster and non-taster TAS2R38 proteins.

Structure Analysis		
Program/Website Name	*Description*	*Website*
PopMuSiC	Predicts protein stability changes due to point mutations	http://soft.dezyme.com/home
Whiscy	Predicts protein interfaces based on structure and a sequence alignment	http://nmr.chem.uu.nl/Software/whiscy/index.html
Dali Server	Identifies structural homologs by searching the Protein Data Bank	http://ekhidna.biocenter.helsinki.fi/dali_server/start
MolProbity	Structure quality analysis.	http://molprobity.biochem.duke.edu/
CABS-flex	Protein flexibility prediction based on structure	http://biocomp.chem.uw.edu.pl/CABSflex/index.php
PDBePISA	Protein-protein interface prediction	http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html
WEBnm@	Indicates protein flexilbity via normal mode analysis	http://apps.cbu.uib.no/webnma/home
FlexServ	Predicts protein flexiblity based on structure	http://mmb.irbbarcelona.org/FlexServ/input.php
RosettaBackrub	Determines change in protein structure based on mutation	https://kortemmelab.ucsf.edu/backrub/cgi-bin/rosettaweb.py?query=index
PDB2PQR	Useful for relative pKA prediction	http://nbcr-222.ucsd.edu/pdb2pqr_2.0.0/
COACH	Protein ligand bindign site prediction	http://zhanglab.ccmb.med.umich.edu/COACH/

Example Activity: Identifying function and substrate/ligand binding using COACH/Dali server

Determining 3D structure is a helpful initial step in biochemical analysis of proteins. However, characterizing functionality of structure will yield further insights into a proteins biochemistry. Experimentally determining the relationship between structure and function for a single protein may require years of *in vitro* investigation for which a multitude of techniques are required. Sometimes, enzymes have the ability to catalyze reactions using several different substrates, leading to even more functional complexity. Clues to functionality can be gleaned from comparing unknown or predicted structures to with previously characterized function. This method requires an experimental structure or a structural model that is high quality and reliable to be most effective. To identify functional motifs or domains, the structure in question is aligned to the structures of a reference group. The lower the Root Mean Square Deviation (or RMSD), the better the structures match. This method may also suggest how the substrate or ligand interacts with the enzyme or protein, thus linking the protein structure to function and metabolic pathway involvement.

To begin, calculate the energy minimized structure of the TAS2R38 protein by navigating to either the Dali Server or the COACH server (links below) and upload your structure to the server along with your email address for results output. Allow 1-24 hours run time for search and comparison analysis.

The Dali Server searches for structural homologs and aligns the structure reporting a RMSD value and a Z score. In this context, a Z score is the number of standard deviations above the average or expected value the alignment is and gives an idea of the quality of the alignment. A Z score greater than 2 is required to be useful and a value greater than 5 (meaning 5 standard deviations above the average) is near certainty of some structural match. There is also information on the % identical amino acids and the length of the aligned section which is informative on whether the whole protein or just as single domain matches the reference structure.

Dali Server: http://ekhidna.biocenter.helsinki.fi/dali_server/start

The COACH server predicts protein-ligand binding sites and using structural alignment but also uses multiple alignment of different structures to suggest functional residues in the binding site. The results will show a C-score which is a

confidence value in the results and a value closer to 1 is better. There is also information on the number of structures in the alignment, which can indicate how well known the ligand binding site is, and a list of the predicted functional amino acids.

COACH server: http://zhanglab.ccmb.med.umich.edu/COACH/

From both servers after completion, you can download the reference PDB from which the functional information is derived and in the case of COACH you can download a ligand bound structure. When the server returns your results, look at the results in total. Did the servers find a good match to your protein? Are there enzyme descriptions or consistent phrases that show up? What are they?

Download the top two hits from each server and align them to your TAS2R38 model in pyMol and record the RMSD value that pyMol returns to you. A perfect match in RMSD is 0, while a poor match is one where the RMSD value is >3 Å, however a high RMSD value does not mean there are not regions of local similarity. A visual comparison is always helpful! How well do the top hits match your model? Can you see the similarity? Does the similarity cover the entire protein?

To identify binding sites, some structures will have substrates/ligands bound. Observe if there is any match in the ligand/substrate binding sites. If the top 2 hits from each server do not have a ligand bound, go down the list until you find one that does. Download this structure and align it to your model in pyMol. Does the ligand "fit" into the aligned sites? (It will not be perfect, so look for how bumps could fit into holes or nearby holes!)

If there is a reasonable match, you should be able to identify amino acids in the binding pocket that could help bind the substrate/ligand. Download the literature file from Pubmed or Scopus that describes the reference structure. Read how the authors describe the binding pocket and determine if the key/essential contacts described for the reference structure can be found in your model. Draw how the ligand might bind in your model. Are the interactions shown in your model favorable interactions?

This initial suggestion of how the ligand/substrate binds to your model can be refined to improve the fit using energy minimization. In pyMol, align the ligand/substrate bound reference structure to your model. Then hide (but not delete) your model and separate the ligand/substrate from the protein. When this is done, combine your model and the free ligand into a single .pdb file and close pyMol. Re-open this new .pdb file in pyMol and make sure the ligand is bound to your model as you expect. Make sure the reference structure is completely removed. The next steps require your model and the ligand bound in the approximate binding site.

Upload your combined model-ligand .pdb file to the energy minimization servers from the MolProbity activity. When the results are returned, align your starting model with ligand to the minimized model with ligand in pyMol. Were there differences between the two models? Does this minimized binding site match the binding site described in the literature better? Did the interactions between substrate/ligand and your model improve? Would you change your drawing of how the ligand/substrate binds to the model?

Computational Module #4: Molecular Dynamics

A common step following generation of a homology model or docking a ligand is to perform molecular dynamics to allow the protein/protein-ligand complex to reach a stable conformation and increase accuracy of pK_A predictions, interaction energies, etc. Use the below databases to calculate molecular dynamics simulations of taster and non-taster TAS2R38 proteins.

Molecular Dynamics		
Program/Website Name	*Description*	*Website*
CHARMMing	CHARMM based molecular dynamics on the web	http://www.charmming.org/charmming/
MDWeb	Web-based molecular dynamics	http://mmb.irbbarcelona.org/MDWeb/index.php
YASARA Energy Minimization server	YASARA based molecule minimization	http://www.yasara.org/minimizationserver.htm
Non-web based programs		
YASARA Dynamics	Interactive molecular dynamics program	www.yasara.org
GROMACS		http://www.gromacs.org/
NAMD		http://www.ks.uiuc.edu/Research/namd/

Example Activity: Analysis of protein structure using the Molprobity server

Structural models, whether derived from experimental methods such as NMR and X-ray crystallography or from homology modeling, are useful only if they are realistic. Therefore, it is important to know how well the model compares to reality from a chemical perspective such as bond angles, clashes between atoms, and overall geometry. A structure that is chemically accurate is more likely to be an effective tool for the prediction of function and mechanism. You will use the MolProbity server to analyze the quality of your homology model from the homology modeling server (ex. Phyre[2]) activity and identify areas where the structure is not as good. MolProbity analyzes a given structure to identify clashes, hydrogen bonding, van der Waals contacts, geometry, and looks for "unusual" structural features of the model. Most protein structures have a known set of chemical parameters that are conserved despite the structural diversity of proteins.

Navigate to the MolProbity server athttp://molprobity.biochem.duke.edu/ and upload your TAS2R38 homology model file in .pdb format. Select Analyze All-Atom Contacts and Geometry. On the next page choose the outputs you want to see. Select Clashes, Hydrogen Bonds, van der Waals contacts and Geometry evaluation under Universal. Under Protein, select Ramachandran plots, Rotamer evaluation, Cβ deviations, and cis-peptide evaluations. Click the "Run" button at the bottom of the screen and wait for the server to run (<2 minutes). The output page will show the summary analysis of the structure. Note any boxes that are yellow or red, as these indicate less than ideal statistics. Click Multi-criterion chart to see the structure evaluation at a single residue level. Amino acids that fall outside the ideal will be highlighted in red.

Look at your structure in PyMol and initially focus on at the troublesome residues. Can you see the issue at that site? Compare your observations to amino acid positions that are ideal, preferably of the same type of amino acid. How do they differ? Looking back at your model prediction statistics, do the "bad" amino acids occur in areas that were tougher to model or for which there was less template information? If this were an experimentally derived structure, how would decide if "bad" areas were due to bad data fitting versus an unusual structure?

Homology modeling may not produce the most stable interactions between amino acids. Energy minimization uses a simple molecular dynamics simulation to lower the energy of the molecule in an attempt to produce the most optimal structure. Sometimes, structural model issues can be "fixed" through this method. Now submit your structure to one (or more) of the energy minimizing servers below:

http://www.yasara.org/minimizationserver.htm
http://csb.stanford.edu/kobamin/about.html
http://zhanglab.ccmb.med.umich.edu/ModRefiner/

Be sure to submit the *TAS2R38* protein structure as a .pdb file and it may take several hours to finish minimizing the structure. After you obtain the minimized structure, submit the structure to MolProbity as you did before and compare the

quality statistics with the non-minimized structure. Did energy minimization help the Molprobity Score? Which parameters got better? Did any parameters get worse?

Computational Module #5: Flexibility Analysis

Protein dynamics and flexibility are underappreciated affects of genetic mutation on protein functionality. Homology models are snapshots of conformations the protein *could* adopt, however, these are likely not the only conformations adopted by the protein. Determination of protein flexibility can come from molecular dynamics simulations like those described above which often report RMSF values or B factors, or from analysis of segments of the sequence/structure and predicting mobility based on historical trends for the amino acids. This latter method does not require a structure and the results are often consistent with more involved methods of analysis. However, this method does not perform well with multimeric proteins due to the lack of information on the spatial organization of the interface contained in the protein sequence. Use the below protein structure databases to compare protein flexibility in the taster and non-taster TAS2R38 proteins.

Structure Analysis		
Program/Website Name	*Description*	*Website*
PopMuSiC	Predicts protein stability changes due to point mutations	http://soft.dezyme.com/home
Whiscy	Predicts protein interfaces based on structure and a sequence alignment	http://nmr.chem.uu.nl/Software/whiscy/index.html
Dali Server	Identifies structural homologs by searching the Protein Data Bank	http://ekhidna.biocenter.helsinki.fi/dali_server/start
MolProbity	Structure quality analysis.	http://molprobity.biochem.duke.edu/
CABS-flex	Protein flexibility prediction based on structure	http://biocomp.chem.uw.edu.pl/CABSflex/index.php
PDBePISA	Protein-protein interface prediction	http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html
WEBnm@	Indicates protein flexilbity via normal mode analysis	http://apps.cbu.uib.no/webnma/home
FlexServ	Predicts protein flexiblity based on structure	http://mmb.irbbarcelona.org/FlexServ/input.php
RosettaBackrub	Determines change in protein structure based on mutation	https://kortemmelab.ucsf.edu/backrub/cgi-bin/rosettaweb.py?query=index
PDB2PQR	Useful for relative pKA prediction	http://nbcr-222.ucsd.edu/pdb2pqr_2.0.0/
COACH	Protein ligand bindign site prediction	http://zhanglab.ccmb.med.umich.edu/COACH/

Example Activity: Analysis of protein flexibility using the CABS-Flex server

Protein flexibility is often critical for protein function to exclude solvent from reactions after substrate binding (ex. Hexokinase), to help regulate substrate binding (ex. Ubiquitin Hydrolase) or to diversify the active site surfaces during a complex reaction (ex. Ubiquitin activating enzyme). Therefore, changes to flexibility can have significant effects on functionality. It is well documented that certain amino acids stabilize certain structures while others are destabilizing. Therefore, the sequence alone can cause unstructured areas in the protein, which may become more flexible. Further analysis using molecular dynamics simulations of the protein can account for the three dimensional packing of amino acids and suggest how entire motifs or domains may move during the functional cycle of the enzyme. This exercise uses the CABS-Flex server to predict the flexible regions of TAS2R38 in the substrate/ligand bound and apo (no substrates bound) form. You will then compare the results to determine how substrate/ligand binding affects protein flexibility.

You will need your energy minimized model structure and your model bound to a substrate in .PDB file format. The CABS-flex server performs 10 ns of molecular dynamics in explicit solvent (meaning waters are present and simulated) and then clusters the structures from the simulation together to identify flexible regions. The results are presented in a movie form as well as producing structures representing the clusters for comparison to the original structure. Navigate to the CABS-flex server at http://biocomp.chem.uw.edu.pl/CABSflex/index.php. Click submit a new job. In the box on the right side, provide a project name and upload your apo structure of the protein. Click submit and then click the link on the next page to go to the results page. In a new tab, repeat the above for the ligand/substrate bound structure. This server

will take 4 to 12 hours to process the data, so do not wait until just before class to perform this experiment. You will need to save your link to access the data if you cannot leave the tab open while the job runs. Once the experiment for the apo model is complete, you can watch the movie of TAS2R38 and see the simulated flexibility. Click on the Models tab and look at the residue fluctuation profile and cluster alignment. Is the protein uniformly flexible or are there regions of higher flexibility? What structural features are apparently more flexible?

Make sure you download the fluctuation data in .TXT format and the multimodel files. Repeat the analysis for the substrate/ligand bound model and download the files. Import the fluctuation data for each model into Excel (or another appropriate plotting program) and plot Fluctuation vs. Residue Index for the apo and substrate/ligand bound protein on the same plot as a scatter plot. Are there regions of the ligand/bound structure that are significantly more or less flexible compared to the apo structure? Compare the flexible regions using the cluster structures in pyMol and correlate the above plot to the structures. Do the regions with flexibility differences noted above correspond to areas near the substrate/ligand binding site?

Potential for Investigation of Other Loci

Here we use the example of the polymorphic *TAS2R38* gene modulating taste perception of the bitter compound PTC. However, the modules described in this study can be applied to a broad range of polymorphic loci associated with phenotypic change offering a high degree of versatility to instructors. Projects can easily be developed to meet the specific needs of each class based on factors such as learning objectives, student interest, and available resources. Naturally, the sensitive implications of human genotyping call for the use of discretion in the selection process. In particular, variants associated with clinical conditions and other phenotypes with sensitive implications should be avoided. In an effort to demonstrate the broad range of loci with potential for investigation, a variety of literature and tools have been utilized to curate five suitable loci taking into account the aforementioned considerations (Figure S1-S5; 34, 35, 36, 37, 38). These example loci include SNPs as well as variable number tandem repeats (VNTRs), and represent intronic, exonic, and intergenic regions.

LCT / MCM6: **Lactase Persistence**

Investigation of this locus explores hypolactasia, or the lactose intolerance phenotype. Lactase persistence is a commonly occurring autosomal recessive condition characterized by a decline in the presence and activity of lactase-phlorizin hydrolase, an enzyme encoded for by the *LCT* gene and responsible for the conversion of lactose to glucose and galactose in adult liver cells. A single SNP (rs4988234) located 13.9 kb upstream of *LCT*, in intron 13 of the neighboring *MCM6* gene, provides an opportunity for students to explore the profound effects non-coding variants and long range cis-regulatory elements can have on important genes (Figure S1). This SNP, in combination with a SNP located 22 kb upstream of *LCT* in intron 9 of *MCM6*, comprises a haplotype experimentally shown to predict lactose persistence in individuals of European descent with 77% accuracy [1]. Individuals homozygous for the C allele will likely be unable to metabolize lactose as an adult, while those heterozygous or homozygous C are likely to metabolize lactose throughout their life. It has been suggested that the lactose-tolerant allele was strongly selected for among populations of European dairy farmers ~7,500 years ago, allowing students to explore a recent evolutionary event.

A number of studies provide evidence that the -13,910 SNP functions is an important location in a 3 kb lactase enhancer region [2, 3]. A 2003 study demonstrated this in vitro when the transfection of the intestinal cell culture lactose persistence-associated (T) variant and lactose non-persistence-associated (C) variant resulted in 2.8-fold and 2.2-fold induction of the lactase promoter region, respectively [2]. Furthermore, DNA affinity purification has shown the transcription factor OCT1 to bind more strongly to the T variant than the A variant when co-expressed with the HNF1 transcription factor, and a number of other transcription factor binding sites known to be expressed in intestines have been

located in the vicinity of the *LCT* -13910 region [3]. Classes studying non-coding variants, enhancers, and long-range controlling elements would benefit greatly from the exploration of this locus. Potential protein modeling exercises could investigate the biochemical nature of the transcription factor OCT1's interactions with the binding site affected by this variation and predict other variations to this site might affect this interaction.

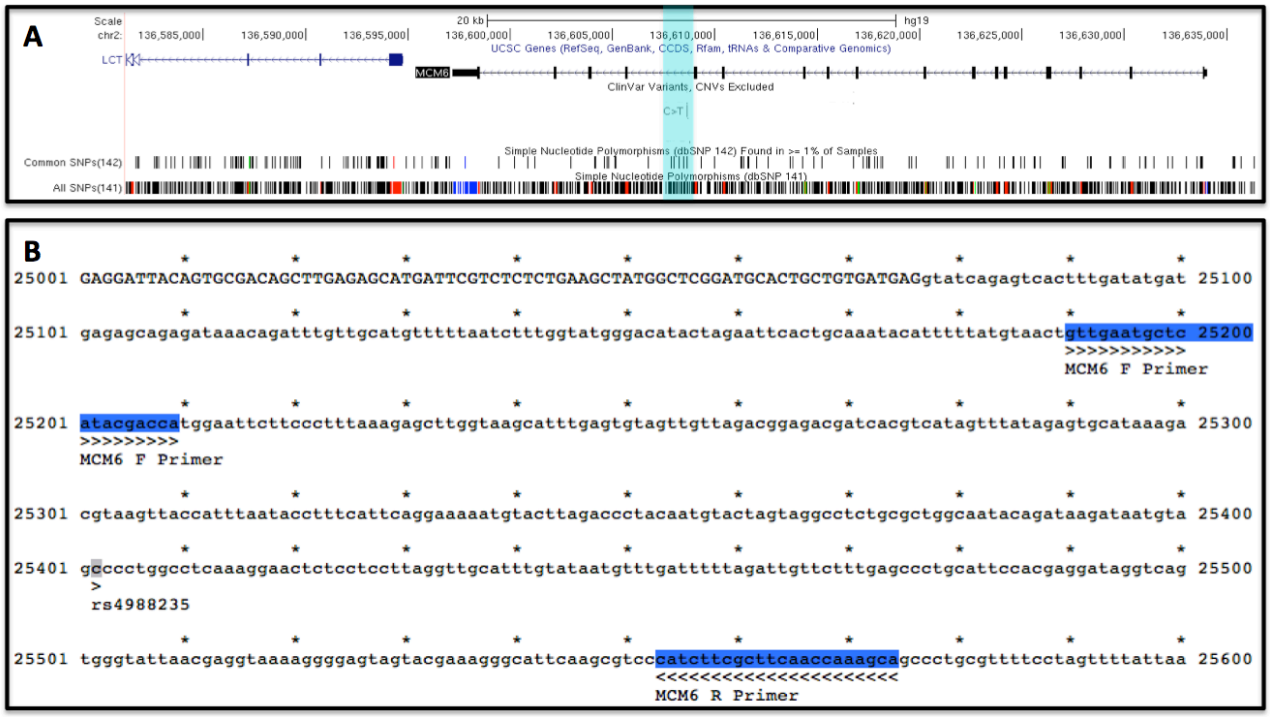


Figure S1. Computational analysis of the human *MCM6* gene. Gene visualization and genetic variation data tracks obtained from the UCSC Genome Browser (A). Gene sequence exported from the browser and annotated in the ApE sequence editing software (B).

CYP1A2: Caffeine Metabolism

The *CYP1A2* gene encodes for cytochrome P450 family protein CYP1A2, an enzyme responsible for metabolizing a number of drugs in the human liver including the most widely used drug in the world, caffeine, of which it selectively catalyzes the N-3 demethylation of into parazanthine [4,5]. CYP1A2 accounts for over 95% of the primary metabolism of caffeine, which has led to its use as a probe drug for investigating CYP1A2 activity in humans [6,7]. A single SNP (rs762551) located in intron 1 of *CYP1A2* has been identified as a predictor of how quickly an individual metabolizes caffeine after ingestion (Figure S2). This phenotype is of considerable relevance to college students, for many of whom caffeine intake is a daily occurrence. Students are encouraged to develop a hypothesis for their own phenotype (fast metabolizer or slow metabolizer) based on their experiences with caffeine before isolating their own DNA and analyzing genotype results against their hypotheses.

Classes studying the biochemical processes of metabolism pathways or interested in in topics such as pharmacogenetics and pharmacogenomics are great candidates to perform this investigation. Protein modeling exercises can explore a variety of enzyme-catalyzed metabolic reactions and interactions, including the aforementioned first step in caffeine metabolism, cytochrome P450-catalyzed N-3 demethylation of caffeine to parazanthine, and subsequent metabolic reactions. This investigation can also be used to educate students regarding the theories and potential of personalized medicine by illustrating that the same drug can be metabolized differently by different individuals, and how this difference in metabolism can be predicted through genotyping.

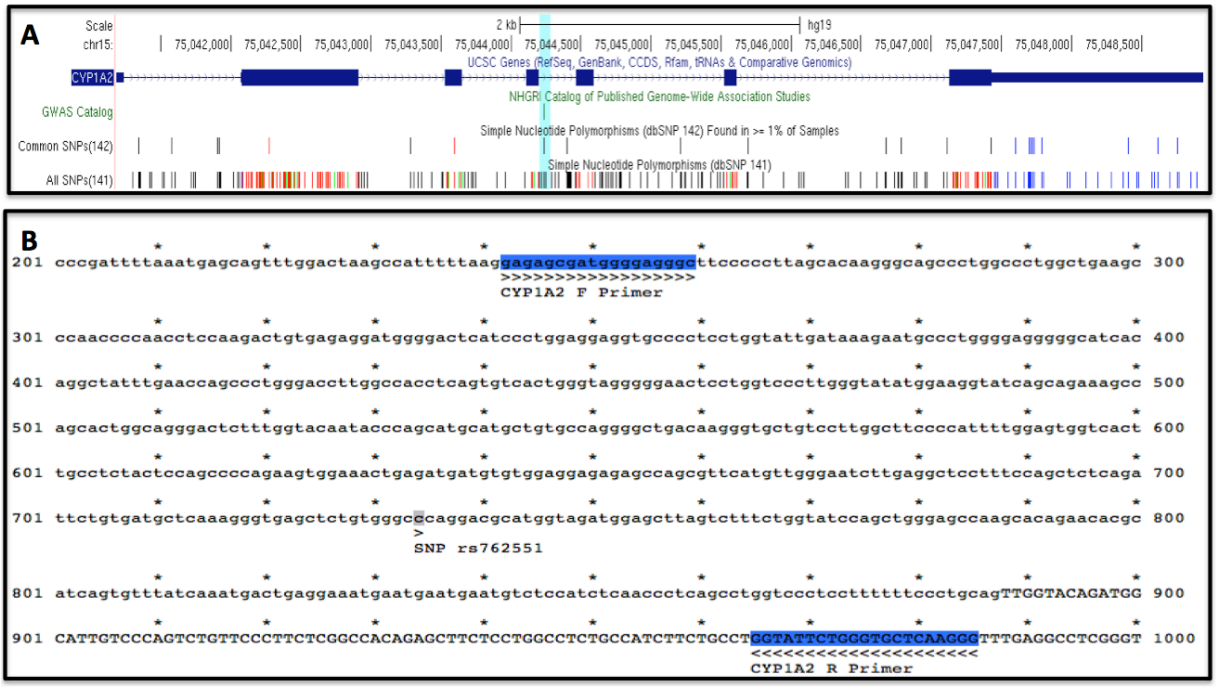


Figure S2. Computational analysis of the human *CYP1A2* gene. Gene visualization and genetic variation data tracks obtained from the UCSC Genome Browser (A). Gene sequence exported from the browser and annotated in the ApE sequence editing software (B).

PER3: Diurnal Preference

The *PER3* gene encodes the Period Circadian Protein Homolog 3 and is one of a number of genes influencing circadian rhythms. Study of this locus investigates the impact of a 54 bp variable number tandem repeat (VNTR) (rs57875989) located in exon 18 of *PER3* on the diurnal preference phenotype- whether an individual is a "morning person" or an "evening person" (Figure S3). Individuals with four copies of this repeat are likely to have a preference for night time, while those with 5 copies tend to prefer the morning, and therefore wake up and go to sleep earlier than the "evening" people. The VNTR variation allows for genotyping to be tested via a simple PCR and gel electrophoresis, as genotype will be denoted by the length of the amplified fragment. Student's phenotypes can be tested using a number morning-evening questionnaires (MEQs) freely available online. Relatively little is known about the exact function of Period Circadian Protein Homolog 3 except that it is an important factor in the regulation of sleep timing, and is suspected to perform this regulation through transcriptional regulation of genes involved in sleep homeostasis [8]. The VNTR occurs in a putative phosphorylation domain; a potential topic of investigation for protein modeling exercises.

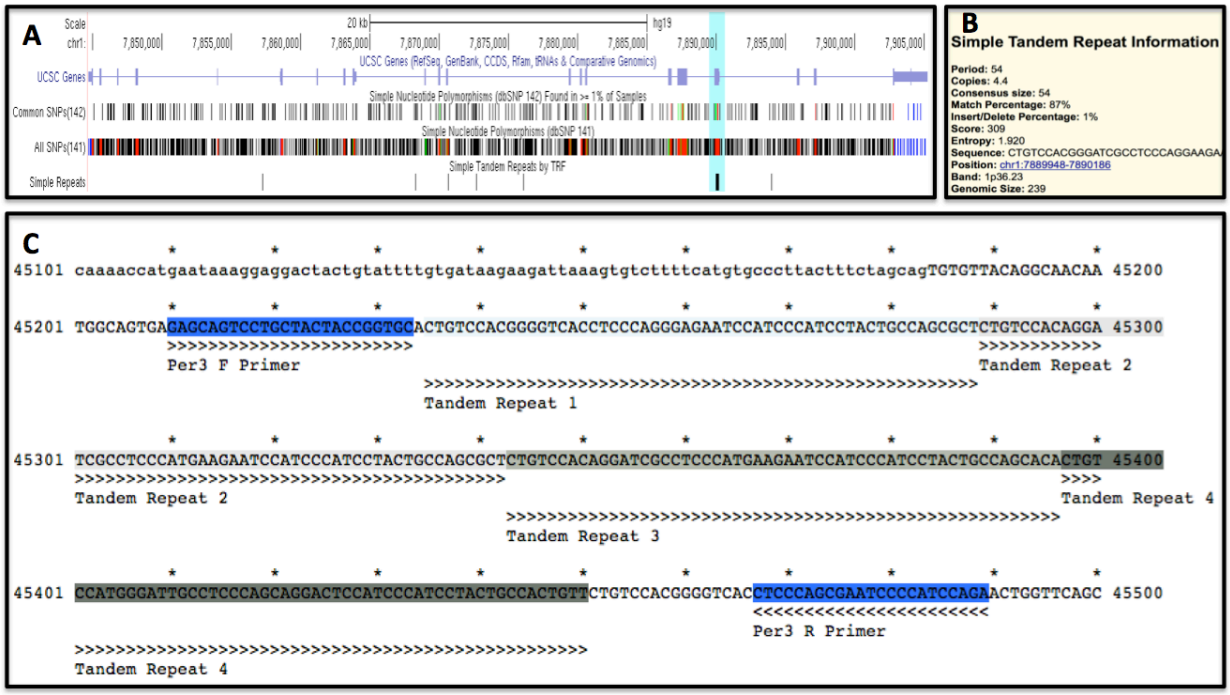


Figure S3. Computational analysis of the human *PER3* gene. Gene visualization and genetic variation data tracks obtained from the UCSC Genome Browser (A). Data specific to the VNTR of interest obtained from its respective entry in the 'Simple Repeat' UCSC Genome Browser track (B). Gene sequence exported from the browser and annotated in the ApE sequence editing software (C).

Olfactory Cluster*:* Asparagus Anosmia

The inability to smell the sulfurous odor of urine after recent asparagus ingestion has been linked to a single SNP (rs4481887) located in an inter-genic region of a 50-gene olfactory gene cluster on chromosome 1, with closest proximity to *OR2M7*, an olfactory receptor gene (Figure S4). This discovery came about through a genome-wide association study, and the biological causes of this association have not been well characterized. This is partly due to the fact that the metabolites responsible for the distinct odor in urine have not been identified, which renders researchers unable to test the response of neighboring olfactory genes to the odorous asparagus metabolites [9]. This does however provide a lab topic that many undergraduate students may consider humorous and therefore interesting to investigate.

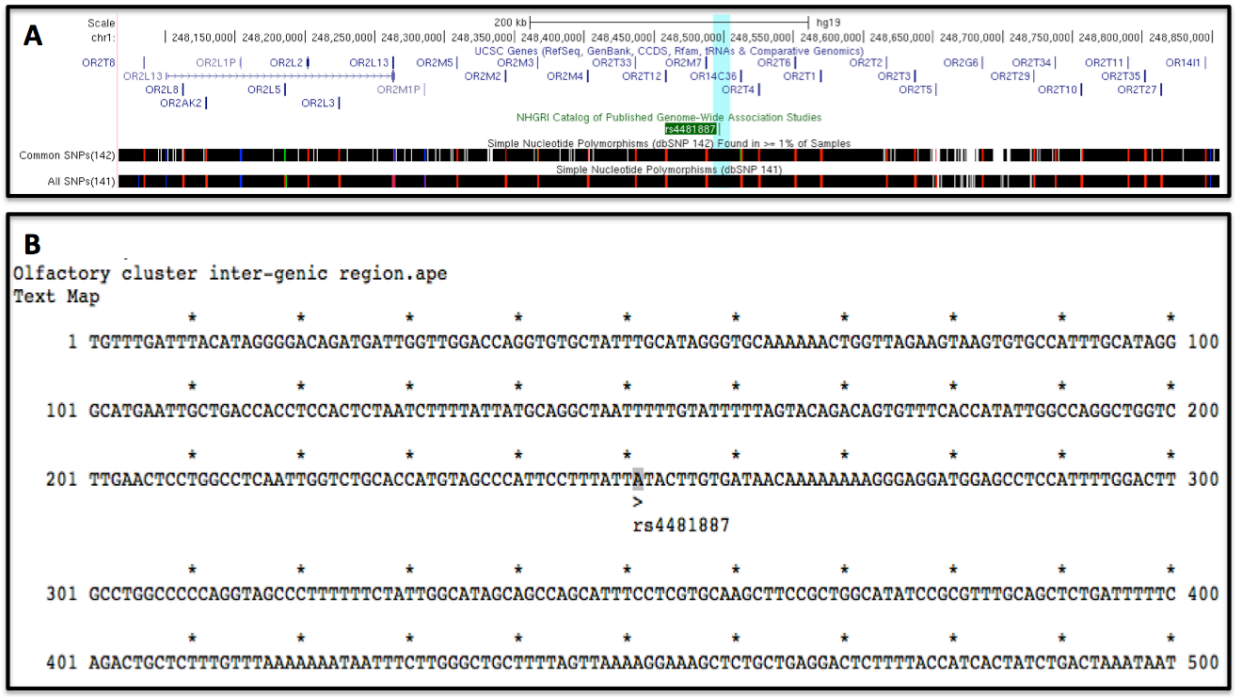


Figure S4. Computational analysis of an olfactory gene cluster on chromosome 1 . Gene visualization and genetic variation data tracks obtained from the UCSC Genome Browser (A). Gene sequence exported from the browser and annotated in the ApE sequence editing software (B).

SLC18A2: Level of Spirituality

This investigation was inspired by an argument made in Dean Hamer's book *The God Gene: How Faith is Hardwired into Our Genes* that a single SNP in the human *VMAT2* gene (now known as *SLC18A2*) plays a small role in determining an individual's level of spirituality. *SLC18A2* encodes a protein responsible for transporting monoamine neurotransmitters into vesicles in the brain, and therefore could theoretically alter brain activity, a notion that has been supported by evidence associating variations in *SLC18A2* with substance abuse and Parkinson's disease [8].

Unfortunately, the exact location of the SNP Hamer refers to was not shared, nor was it ever published in scientific literature, and therefore classes have had to predict the SNP he refers to using locational and allelic frequency clues. A previous undergraduate investigation identified the SNP rs363371 as the likeliest candidate for Hamer's SNP of interest [Figure S5; 9]. Furthermore, little is known about the impact this SNP has on the functionality of *SLC18A2*, except that its intronic location eludes to a potential regulatory impact. Investigation of this site has not produced results heavily supporting this association, which renders the validity of the association questionable.

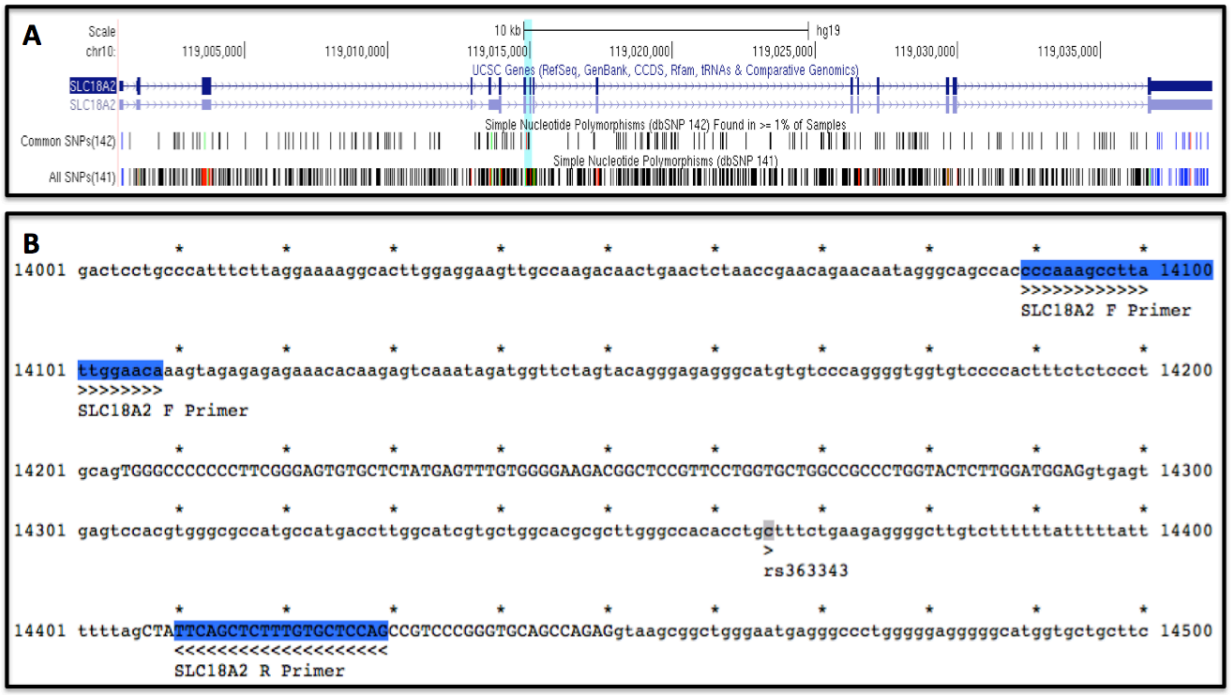


Figure S5. Computational analysis of the human *SLC18A2* gene. Gene visualization and genetic variation data tracks obtained from the UCSC Genome Browser (A). Gene sequence exported from the browser and annotated in the ApE sequence editing software (B).

References

[1] Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., Rhodes, M., Reich, D. E., Hirschhorn, J. N. Genetic signatures of strong recent positive selection at the lactase gene. Am. J. Hum. Genet. 74: 1111-1120, 2004.

[2] Olds, L. C., Sibley, E. Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. Hum. Molec. Genet. 12: 2333-2340, 2003.

[3] Lewinsky, R. H., Jensen, T. G. K., Moller, J., Stensballe, A., Olsen, J., Troelsen, J. T. T-13910 DNA variant associated with lactase persistence interacts with Oct-1 and stimulates lactase promoter activity in vitro. Hum. Molec. Genetics. 14: 3945-3953, 2005.

[4] Donavan JL, DeVane, CL. A primer on acaffeine pharmacology and its drug intereactions inclinical psychopharmocology. Phychopharmocol Bull, 2001;35:30-48.

[5] Kot M, Daniel WA. Caffeine as a marker substrate for testing cytochrome P450 activity in human and rat. Pharmacol Rep. 2008;60:789–97.

[6] Thorn C. F., Aklillu E., McDonagh E. M., Klein T. E., Altman R. B., (2012). PharmGKB summary: caffeine pathway, Pharmacogenetics and genomics, Article: 22293536@PubMed.

[7] Kalow, W. and Tang, B.-K. (1993), The use of caffeine for enzyme assays: A critical appraisal. Clinical Pharmacology & Therapeutics; 53: 503-514. Doi: 10.1038/clpt.1993.63.

[8] Silveira, L. A. (2008). Experimenting with spirituality: Analyzing the god gene in a nonmajors laboratory course. CBE Life Sciences Education,7(1), 132-145. doi:10.1187/cbe.07-05-0029.

[9] Lin Z, Walther D, Yu XY, Li S, Drgon T, Uhl GR. SLC18A2 promoter haplotypes and identification of a novel protective factor against alcoholism. Hum Mol Genet. 2005;14:1393–1404. doi: 10.1093/hmg/ddi148.