**Biology 231 - Spring 2016**

**PCR and DNA Sequencing**

**I.** Goal of the Lab

**II.** Introduction to the vertebrate *ND3* and D-loop genes

**III.** Introduction to Restriction Digests and Agarose Gel Electrophoresis

**IV.** Introduction to DNA Extraction

**V.** Introduction to the Polymerase Chain Reaction

**VI.** Introduction to DNA Sequencing

**VII.** Lab Procedures

**Appendix 1.** DNA Sequence Variation **Appendix 2.** Introduction to Linear Regression

**I. Goal of the Lab**

 The ultimate goal of this lab is to assess molecular genetic variation in the mitochondrial *ND3* and D-loop genes of cows. To figure out how to reach this goal, it is a good idea to work backwards.

(7) Assess patterns of *ND3* and D-loop DNA sequence variation

↑

(6) Align edited *ND3* and D-loop sequences

↑

(5) Proofread and edit raw *ND3* and D-loop DNA sequence data

↑

(4) Perform DNA sequencing reactions

↑

(3) Prepare DNA sequencing templates

↑

(2) PCR amplify the *ND3* and d-loop regions

↑

(1) Extract genomic DNA from an individual cow

 We cannot perform Step 7 if the sequences are not properly aligned. We cannot perform Step 6 if there are errors in the sequences. We cannot perform Step 5 until DNA sequencing reactions are performed and the products are analyzed. We cannot perform Step 4 without suitable templates (in this case, purified PCR products). We cannot perform Step 3 without first obtaining PCR products. We cannot perform Step 2 until we have suitable genomic DNA to act as the PCR template.

**II. Introduction to the Vertebrate *ND3* and D-loop genes**

 The mitochondrial *ND3* gene encodes subunit 3 of NADH dehydrogenase, an important protein in the mitochondrial electron transport chain. This enzyme catalyzes the simultaneous oxidation of NADH and reduction of coenzyme Q by the transfer of electrons from the former to the latter.

 The following sequence was taken from Genbank file NC\_001567, the complete sequence of the *Bos taurus* mitochondrial DNA (at least for this particular cow). The cow’s circular mitochondrial chromosome is about 16.3 kilobase pairs in length. The *ND3* gene begins at position 9823 and ends at position 10168. The start codon ATA is acceptable in the modified genetic code of mammalian mitochondria. The single-letter codes for each amino acid are shown above the corresponding codons. Notice that there is no complete stop codon in the DNA sequence. A pair of adenines are added to the uracil (thymine in the DNA sequence; double-underlined at position 10168) during an mRNA editing step, thus completing the stop codon.

|  |
| --- |
|   9661 ACCACTTCGG CTTTGAAGCC GGTGCCTGAT ACTGACATTT CGTAGACGTA GTCTGACTTT 9721 TCCTCTATGT TTCTATCTAT TGATGAGGCt cctattcttt tagtattaac tagtacagct **M** N **L**  M **L** A 9781 gacttccaat cagctagttt cggtctagtc cgaaaaagaa ta**ATA**AAT**TT A**ATA**CTA**GCC **L** L **T** N **F** T **L** A  **T** L **L** V **I** I **A** F **W** L **P** Q  9841 **CTC**CTGACCA ATTTTACA**CT A**GCC**ACC**CTA **CTC**GTC**ATC**A TC**GCA**TTC**TG A**CTT**CCC**CAA **L** N  **V** Y **S** E **K** T **S** P **Y** E **C** G **F** D **P** M **G**  S  9901 **CTA**AAT**GTA**T AC**TCT**GAG**AA A**ACA**AGC**CCA **TAC**GAA**TGT**G GA**TTT**GAC**CC C**ATA**GGA**TCA **A** R **L** P **F** S **M** K **F** F **L** V **A** I **T** F **L** L **F** D  9961 **GCC**CGC**CTT**C CC**TTC**TCT**AT A**AAA**TTC**TTT **CTG**GTA**GCC**A TC**ACA**TTC**CT C**TTA**TTT**GAC **L** E **I** A **L** L **L** P **L** P **W** A **S** Q **T** A **N** L  **N** T  10021 **CTA**GAA**ATT**G CA**CTC**CTC**CT A**CCA**CTG**CCA **TGA**GCC**TCA**C AA**ACA**GCA**AA T**CTA**AAC**ACA **M** L **T** M **A** L **F** L **I** I **L** L **A** V **S** L **A** Y **E** W 10081 **ATG**CTT**ACC**A TA**GCC**CTC**TT C**CTA**ATT**ATC **CTC**CTA**GCT**G TA**AGC**CTA**GC C**TAT**GAG**TGA **T** Q **K** G **L** E **W**  T **E** 10141 **ACT**CAA**AAA**G GA**CTA**GAA**TG A**ACC**GAA**tat ggtacttagt ttaaaataaa ataaatgatt 10201 tcgactcatt agattatgat ttaattcata attaccaaAT GTCTATAGTA TACATAAACA 10261 TTATAATAGC ATTCACAGTA TCTCTTGTAG GACTACTAAT ATACCGATCC CACCTAATAT 10321 CCTCCCTTCT ATGCTTAGAA GGAATAATGC TATCCCTATT CGTTATAGCA GCCCTAACAA |
| **DNA sequence of *ND3* in *Bos taurus* (GenBank accession NC\_001567.** See text for details. |

 The upstream *COX3* gene runs from positions 8970 to 9750 (also an incomplete stop codon). The downstream *ND4L* gene runs from positions 10239 to 10535. In order to amplify the entire *ND3* gene, primers were designed against these two flanking genes. The locations of those primers correspond to the underlined sequences. Note that the sequence of the downstream primer (10310 to 10334) is actually the reverse-complement of the sequence shown. The reason for should become clear when we reach the section on PCR.

 The mitochondrial D-loop corresponds to the origin of DNA replication. Unlike most of the vertebrate mitochondrial genome, it does not encode a protein or a functional RNA. It tends to be more variable than protein-coding regions, which has its advantages and disadvantages in the context of this project. The advantage is that there is a greater chance that we will detect genetic variation among cows (in contrast to *ND3*). The disadvantage is that the potential for genetic variation makes it more difficult to design PCR primers; ideally, the sequences corresponding to the primers will be invariant among individuals.

 To address the challenge of potential variation, mitochondrial D-loop sequences were aligned from a broad sample of cows (see S.J. Lai. et al. (2006) Genetic diversity and origin of Chinese cattle revealed by mtDNA D-loop sequence variation. *Molecular Phylogenetics and Evolution* 38: 146-154). Although there was, in fact, quite a bit of variation, there were invariant stretches of a few dozen base pairs. Primers were designed against two invariant regions, indicated by underlining in the sequence on the next page (corresponding to Genbank file AY521095.1).

|  |
| --- |
|  1 aacactatta atatagttcc ataaatacaa agagccttat cagtattaaa tttatcaaaa 61 atcccaataa ctcaacacag aatttgcacc ctaaccaaat attacaaaca ccactagcta 121 acataacacg cccatacaca gaccacagaa tgaattacct acgcaagggg taatgtacat 181 aacattaatg taataaagac ataatatgta tatagtacat taaattatat gccccatgca 241 tataagcaag tacatgacct ctatagcagt acataataca tacaattatt gactgtacat 301 agtacattat gtcaaattca ttcttgatag tatatctatt atatattcct taccattaga 361 tcacgagctt aattaccatg ccgcgtgaaa ccagcaaccc gctaggcagg gatccctctt 421 ctcgctccgg gcccataaac cgtgggggtc gctatccaat gaattttacc aggcatctgg 481 ttctttcttc agggccatct catctaaaac ggtccattct ttcctcttaa ataagacatc 541 tcgatggact aatgactaat cagcccatgc tcacacataa ctgtgctgtc atacatttgg 601 tattttttta ttttggggga tgcttggact cagctatggc cgtcaaaggc cctgacccgg 661 agcatctatt gtagctggac ttaactgcat cttgagcacc agcataatga taagcgtgga 721 cattacagtc aatggtcaca ggacataaat tatattatat atccccccct tcataaaaat 781 ttccccctta aatatctacc accactttta acagactttt ccctagatac ttatttaaat 841 ttttcacgct ttcaatactc aatttagcac tccaaacaaa gtcaatatat aaacgcaggc 901 cccccccccc c |
| **DNA sequence of a section of the *Bos taurus* mitochondrial D-loop.** See text for details. |

**III. Introduction to Restriction Digests and Agarose Gel Electrophoresis**

 DNA segments of different size can be separated by means of gel electrophoresis. The resolution of this separation (*i.e.*, how different the size the segments must be to effectively separate) depends mainly on two factors. The first is the time that the DNA is allowed to run on the gel. The second is the composition of the gel. Under the right circumstances, polyacrylamide gels can be used to separate pieces of DNA that differ in length by just one nucleotide. However, acrylamide is toxic, and usually we do not need this level of resolution. Agarose gels are safer to use, though pieces of DNA must be markedly different in size to be effectively separated.

 Gel electrophoresis is fairly straightforward in principle and in practice. Because DNA is negatively charged along its entire length (why is this?), it will be pulled toward a positively charged electrode. Regardless of its composition, a gel is essentially a dense mesh of a polymer saturated with a buffer. The DNA can be pulled through this meshwork, but there will obviously be some physical resistance. Shorter pieces will feel less resistance and, therefore, will move through the gel more quickly.

 The rate at which a piece of DNA will move through a gel is dependent not only on its size, but on the density of the gel. DNA will move more slowly through gels made with a higher concentration of agarose or acrylamide. In general, though, the rate at which DNA moves through a gel is inversely related to the logarithm of its size. [Very short and very long pieces of DNA will violate this relationship, but it works pretty well for pieces of intermediate length.] Because of this relationship, we can measure the relative distances traveled on a gel by pieces of DNA of known length, then use these measurements to generate a standard curve. The standard curve can subsequently be used to estimate the sizes of pieces of DNA of unknown length from the distances they travel on the gel.

*Digestion of λ-phage DNA with* Hin*dIII.*

 One commonly used "size standard" is λ-phage digested with the restriction enzyme *Hin*dIII. This bacteriophage has a single linear chromosome composed of double-stranded DNA. While there is some length variation in nature, λ-phage's chromosome is typically 48535 bp in length.

 It is thought that bacteria produce restriction endonucleases to protect themselves against infection by bacteriophage. These enzymes target specific sequences of double-stranded DNA, and these sequences usually show two-fold rotational symmetry. That is, they look the same if you flip them upside-down and backwards. The enzyme *Hin*dIII recognizes the sequence

5'-nnnnnn**AAGCTT**nnnnnn-3'

3'-nnnnnn**TTCGAA**nnnnnn-5'.

The enzyme cuts between the two A's, so that the product of a given cut can be presented as

 5'-nnnnnn**A**-3' 5'-**AGCTT**nnnnnn-3'

 3'-nnnnnn**TTCGA**-5' 3'-**A**nnnnnn-5'.

The recognition sequence for *Hin*dIII appears seven times in the λ-phage genome, so eight pieces are produced. The first cut occurs 23150 bp from the "left" end of the chromosome. The remaining cuts occur at intervals (in order) of 2020, 2320, 9420, 560, 125 and 6560 bp. If you add these numbers up and subtract the sum from 48535, you'll find that the last site is 4380 bp from the "right" end of the chromosome. Therefore, restriction of the λ-phage genome with *Hin*dIII produces eight fragments, ranging in size from 125 to 23150 bp. The latter moves most slowly through an agarose gel, and the former moves most quickly. Given the concentration of gel we will be using, the six fragments between 500 and 104 bp in length will move through the gel at a rate inversely proportional to the logarithm of their sizes. It is those six fragments that we will use as standards to estimate the sizes of our PCR products.

**IV. Introduction to DNA Extraction**

 The length of the human haploid genome is ~3× 109 bp of DNA. The molecular weight of a base pair is about 660 g/mol, which means that 6.02 × 1023 bp weighs 660 g. The weight of one complete haploid genome is, therefore, 3 × 109 bp × (660 g / 6.02 × 1023 bp) = 3.3 × 10-12 g. Thus, 150,000 diploid cells will produce about 1 μg of double-stranded DNA (including the mitochondrial DNA), assuming perfect yield. However, that DNA will be mixed with a lot of other material, mainly protein, lipid and polysaccharide. For most molecular biology methods, we need the DNA to be fairly pure, which means we must separate it from the other macromolecules. This is usually done in four steps:

 (1) Crack open the cells.

 (2) Separate DNA from other materials.

 (3) Precipitate the DNA from aqueous solution by adding an alcohol.

 (4) Resuspend the DNA in a dilute buffer that inhibits DNA degradation.

 Lysis of mammalian cells can be accomplished by weakening cell membranes with proteinase K. There are a number of ways of purifying the DNA. In this lab, we will use a commercial kit to save time and to avoid the use of organic solvents.

**V. Introduction to the Polymerase Chain Reaction**

 PCR is a method used to amplify a segment of DNA. It is based on the general principles of DNA synthesis. Compare the basic requirements for cellular DNA synthesis and PCR:

|  |  |  |
| --- | --- | --- |
| Ingredients | Cellular DNA synthesis | PCR |
| Template for synthesis | chromosomal DNA | Unbroken DNA of sufficient length |
| Enzyme | Endogenous DNA polymerase | *Taq* DNA polymerase |
| Primers | Short pieces of RNA produced by Primase | Short pieces of artificially synthesized single-stranded DNA |
| Monomers | dNTPs | dNTPs |
| Co-factor | Mg++ | Mg++ |

|  |  |  |
| --- | --- | --- |
| Key steps | Cellular DNA synthesis | PCR |
| Separation of template strands | Helicases (with help of topoisomerases and SSBs) | Heat |
| Attachment of primerto template strand | Laid down by Primase | Annealed by cooling |
| Extension from 3' end of primer | Endogenous DNA polymerase | *Taq* DNA polymerase |

 During cellular DNA synthesis, the two strands of DNA are separated by Helicase, and kept apart by single-strand binding proteins. During PCR, the hydrogen bonds that keep the two strands of DNA together are broken by heating the reaction mix. The two strands may not become completely separated (in fact, this may be nearly impossible in the first cycle of PCR), but that's not really a problem.

Remember that all known DNA-dependent DNA polymerases require a pre-existing 3' end on which to add dNTPs. DNA-dependent RNA polymerases do not have this limitation so, in cellular DNA synthesis, the primer is produced by Primase, a type of DNA-dependent RNA polymerase. The primer automatically complements the template strand, since that is the automatic consequence of template-dependent nucleic acid synthesis. In PCR, the primers are included in the reaction mix, and they need to find their appropriate complements on the template. However, at the high temperature used to separate the template strands, the short primers will be unable to bind to the template. So the reaction is cooled to a temperate that is *just* low enough to allow the primers to bind to the template.

In the first panel of the diagram below (Round 1), a section of the original template molecule is shown in white. After denaturation the two strands have been separated by heat. Upon cooling, the primers (grey arrowed lines) anneal to the templates, and then the temperature is raised to allow *Taq* DNA polymerase to extend from the 3' ends of the primers during the elongation step. Notice that in the first round, the newly synthesized strands (shown in black) terminate with arrows; this implies that these strands, like those of the original template molecule, are longer than shown on the figure.

These three steps (denaturation, annealing, and elongation) are repeated again and again in multiple cycles. In Round 2, half of the PCR templates are the newly synthesized strands from Round 1. Notice that after extension, the corresponding products are blunt-ended on one side (circled areas), with the blunt end defined by the primer of the template strand. At the end of Round 3, two of the eight products (labeled \*) are blunt-ended on both sides. After several rounds, these double blunt-ended products − defined by the pair of primers − dominate the reaction. These are the desired PCR products.



*Setting up a typical PCR.*

 The ingredients in a typical 50 μl PCR reaction are (i) template of varying concentration, (ii) 1 Unit of *Taq* DNA polymerase, (iii) 20 pmol of each primer and (iv) 10 nmol of each dNTP, in a buffer that has Mg++ ions at a final concentration of 1.5 mM. The following are the stock solutions we will be using:

 • *Taq* DNA polymerase in buffer to a concentration of 0.5 Units/μl

 • 10 μM ND3-F (or dloop-F) primer in 10 mM Tris/0.1 mM EDTA

 • 10 μM ND3-R (or dloop-R) primer in 10 mM Tris/0.1 mM EDTA

 • A mix of dATP, dCTP, dGTP and dTTP, each at concentrations of 2.5 mM

 • PCR 5× buffer with 7.5 mM Mg++

 For your reactions, you will use 9.6 μl of the PCR 5× buffer and 2 μl of the *Taq* polymerase in the same buffer diluted 5-fold (*i.e.*, to 1×). Together, you are adding the equivalent of 10.0 μl of the 5× PCR buffer and 1.0 Unit of the *Taq* DNA polymerase. How much of the remaining ingredients will you need?

 5× PCR buffer 9.6 μL

 Diluted *Taq* polymerase 2.0 μL

 Genomic DNA 2.0 μL

 10 μM forward primer \_\_\_ μL ⇒ 20 pmol (where p = 10-12)

 10 μM reverse primer \_\_\_ μL ⇒ 20 pmol

 2.5 mM each dNTP Mix \_\_\_ μL ⇒ 10 nmol each (where n = 10-9)

 Ultrapure H2O \_\_\_ μL calculated to bring final volume to 50.0 μl

 TOTAL VOLUME 50.0 μL

You will probably want to use the following formula:

$$x μL stock=y mol solute needed × \frac{1L stock}{z mol solute}×\frac{10^{-6}μL stock}{L stock} .$$

You're solving for *x*, but you should already know *y* and *z*. So it's a simple matter of substitution. [Notice that the middle term on the right is the inverse of the concentration stated in mol/L, while the last term on the right is just a conversion between μl and L.]

**VII. Lab procedures.**

Schedule:

Week 1. Preparation of genomic DNA and Digestion of λ-phage DNA with *Hin*dIII.

Week 2. PCR and assessment of genomic DNA.

Week 3. Purification and assessment of PCRs; DNA sequencing.

Week 4. DNA sequence analysis.

**METHODS FOR WEEK 1**

**A. Preparation of genomic DNA (using the Qiagen DNeasy Blood and Tissue kit)**

The night before lab:

**NEVER re-use pipet tips for any of the methods described for Week 1.**

1. Pour (into the sink) the isopropanol from the microcentrifuge tube containing your meat. Remove the meat, and place it onto a piece of aluminum foil. Using a razor blade, cut off a very small piece (smaller than a pencil eraser). Place a piece of weighing paper on the fine balance in the prep room, tare the balance, and weigh the piece of meat. *You want between 23 and 25 mg of meat for this prep. If you use too little, you may obtain too little DNA. If you use too much, your preparation will be too viscous, and it will clog the Qiagen column used later.*

2. Cut the meat into a few small pieces with the razor blade, and transfer these to a new 1.5 mL microcentrifuge tube.

3. Using a micropipettor, add 1 mL of **ultrapure H2O**; discard the pipet tip. Vortex 15 seconds. Pour off the H2O. Add another mL of ultrapure H2O, but do not discard the tip. Vortex 15 seconds. Pipet off the H2O. *The goal of this step is to wash out the isopropanol.*

4. Using a micropipettor, add 180 μL **Buffer ATL** from the Qiagen DNeasy kit.

5. Grind the meat in Buffer ATL with a plastic homogenizer.

6. Add 20 μL **20 mg/mL proteinase K**. Vortex briefly. Incubate in a heat block overnight at 55°C.

The day of lab:

**Note: Before completing the DNA prep, it would make sense to start the restriction digest (see Part B below).**

7. Spin at maximum speed in the microcentrifuge for 1 min.

8. Transfer the supernatant to a new 1.5 mL microcentrifuge tube.

9. Vortex 15 seconds. Add 200 μL **Buffer AL** from the Qiagen kit. Vortex briefly. Incubate at 70°C for 10 minutes.

10. Obtain and label a Qiagen column (which should come with a 2 mL collection tube).

11. Add 200 μL of 95-100% **EtOH**. Vortex briefly. Transfer the entire contents to the Qiagen column.

12. Spin at 8000 RPM for 1 min. Discard the collection tube, and transfer the column to a new collection tube.

13. Add 500 μL of **Buffer AW1** from the Qiagen kit to the column. Spin at 8000 RPM for 1 min. Discard the collection tube, and transfer the column to a new collection tube.

14. Obtain and label a 1.5 mL microcentrifuge tube.

14. Add 500 μL of **Buffer AW2** from the Qiagen kit to the column. Spin at maximum speed for 3 minutes.

15. Immediately transfer the column to the 1.5 mL tube, being very careful to keep the flow-through in the collection tube from contacting the column. *If you are not careful, and any residual EtOH from Buffer AW2 gets on the column, you will have problems later. In particular, the residual EtOH may interfere with loading samples on a gel.*

16. Add 100 μL **Buffer AE** from the Qiagen kit. Let stand at room temperature for 1 minute. Spin at 8000 RPM for 1 minute. Store at -20°C.

**B. Digestion of λ-phage DNA with *Hin*dIII**

**NOTE: this should be started before the remaining steps in the**

 **DNA prep, in order to use lab time more effectively.**

**NEVER place a used pipet tip into a stock vial of enzyme,**

**10× endonuclease buffer, λ-phage DNA or ultra-pure H2O.**

1. Using a micropipettor and sterile pipet tips, add 2.0 μL **10× endonuclease buffer**, 1.5 μL **λ-phage DNA** and 14.5 μL **ultrapure H2O** to a sterile 1.5 mL microcentrifuge tube. Mix by gently pipetting up and down a few times, *being careful not to produce any bubbles*.

2. Add 2.0 μL ***Hin*dIII**. Mix well by pipetting up and down a few times. Close the tube tightly.

3. Incubate at 37°C for 60 minutes. Then transfer the tube to 75°C for 5 minutes.

4. Add 4.0 μL **6× stop/load buffer** and mix well by pipetting.

4. Store the **λ-*Hin*dIII digest** at -20°C.

**METHODS FOR WEEK 2**

**A. PCR**

1. Label the upper side (but not the cap!) of two 0.2 mL round-topped PCR tubes.

2. One student should prepare the following *ND3* master mix for the class:

• 200 μL **5× PCR Buffer (7.5 mM MgCl**2**)**

• 80 μL **dNTP mix**

 • 40 μL **primer ND3-F (10 μM solution)**

• 40 μL **primer ND3-R (10 μM solution)**

• 4 μL ***Taq* polymerase**

• 600 μL **ultrapure H2O**. Mix gently by pipetting up and down using the same pipet tip used to add H2O. *Avoid producing bubbles*.

A second master mix should be prepared for the D-loop. The only difference will be the substitution of **dloop-F** and **dloop-R** primers in place of the ND3 primers.

3. Transfer 2 μl of your genomic DNA prep to each of two 0.2 mL PCR tubes. *Be sure that the drop is at the bottom of the tube.*

4. Add 48 μl of the appropriate PCR master mix to each 0.2 mL PCR tube. Mix by pipetting gently 3-4 times.

5. Place the tubes in one of the wells in the **thermal cycler**. Once all tubes are loaded, the instructor will start the unit. You will not be required to wait for PCR to be completed.

*A thermal cycler rapidly shifts the temperature of a reaction mixture. The reactions will first be heated to 94°C for 10 minutes to allow the genomic DNA to denature well. This will be followed by 30 cycles of (i) 30 seconds at 94°C, (ii) 30 seconds at 54°C and (iii) 1 minute at 72°C. After the 30 cycles, which should take about two hours, we will allow a final extension period of 7 minutes at 72°C. This allows the polymerase to complete the synthesis of the final strands. The thermal cycler will then cool the samples to 4°C overnight. They will then be stored at -20°C.*

*In your final report, you will want to include the 5'→3' sequences of the primers in your Materials and Methods section. You can infer these sequences from an earlier section of the handout.*

**B. Assessment of genomic DNA**

1. Prepare a 0.6% agarose TAE gel. Weigh 0.24 g **agarose**, and transfer this to a 125 mL Erlenmeyer flask. Add 40 mL **1× TAE**. Your instructor will show you how to melt the agarose using the microwave oven. Once the gel is sufficiently cool, it can be poured. [Don't forget the comb!] Once the gel has solidified (~20 minutes), overlay it with 1× TAE, such that the TAE is ~2 mm above the surface of the gel. *Each gel will be used by four students*. Wait for the gel to solidify before continuing to Step 2.

2. Using a micropipettor, transfer 3 μL **6× stop/load buffer** to a small sheet of parafilm. *Avoid getting this on your clothes, since it will stain permanently.*

3. Add 10 μL of your **genomic DNA prep** to this drop and mix well by gently pipetting. Return the genomic DNA prep to -20°C.

4. Wearing gloves, carefully load 5 μL of your **λ-*Hin*dIII digest** into one well on the gel and carefully load the entire drop containing your genomic DNA into an adjacent well. Once other students have finished adding their samples, cover the electrophoresis apparatus and ***wait for assistance*** to turn on the power supply. You should run then gel at approximately 100 volts for about 60 minutes.

5. Turn off the power supply. Wearing gloves, remove the cover and gently remove the tray holding the gel. Gently drain the buffer back into the gel box.

6. Wearing gloves, transfer the gel to a tray containing a solution of **ethidium bromide** (EtBr). **[NOTE: ETHIDIUM BROMIDE IS A SUSPECTED MUTAGEN.]** The gel will be gently shaken on a platform shaker for about 20 minutes, which should be adequate for staining.

7. Wearing gloves and **UV-shielding goggles**, transfer the stained gel (it won't look any different) to the UV light box. Turn on the UV light, photograph the gel, and turn off the UV light. Print a copy of the gel photograph. *This step will confirm (we hope) that both your digest and genomic DNA prep were successful.*

**METHODS FOR WEEK 3**

**A. Purification and assessment of PCRs**

1. Prepare a 0.6% agarose TAE gel as you did in Week 2. This gel will be used by four students.

2. Pipet 2 μL **6× stop/load buffer** to a small sheet of parafilm. Pipet 5 μL of your PCR from Week 2 to this drop; mix gently by pipetting. Return the PCR to -20°C.

3. Wearing gloves, two students should load 5 μL of their **λ-*Hin*dIII digests** into the first and last wells. Each student using the gel should then load ~7.0 μl of each PCR into the remaining wells. The layout of the gel should be:



4. Cover the gel apparatus and run the gel at ~100 Volts for 60 minutes.

Continue to PCR clean-up using the Qiagen QIAquick PCR Purification kit:

5. For each PCR (*i.e.*, *ND3* and D-loop), transfer what remains of one PCR (~45 μL) to a 1.5 mL microcentrifuge tube, and transfer what remains of the other PCR to a separate 1.5 mL tube. To each of these 1.5 mL tubes, add 5 volumes (5 × 45 μL = 225 μL) of **Buffer PB** from the Qiagen kit. Mix by pipetting. ***Steps 6 through 15 are performed in parallel for both PCRs; that is, you are doing these steps twice.***

6. Obtain and label a Qiagen column. Place this in a 2 mL collection tube.

7. Transfer the entire PCR and Buffer PB mixture (~270 μL) to the column.

8. Spin at maximum speed for 1 minute. Discard the flow-through (the material in the collection tube), but do not discard the collection tube.

9. Add 750 μL of **Buffer PE** from the Qiagen kit to the column.

10. Label a 1.5 mL microcentrifuge tube with the name of the PCR and your initials .

11. Spin the column at maximum speed for 1 minute. Discard the flow-through and spin for another minute. Immediately transfer the column to the labeled microcentrifuge tube.

12. Apply 30 μL of **Buffer EB** from the Qiagen kit to the center of the column membrane, making sure to avoid touching the membrane with the pipet tip. Let stand at room temperature for 1 minute.

13. Spin the column at maximum speed for 1 minute.

14. Discard the column. *Approximately 28 μL of cleaned PCR will have transferred to the 1.5 mL tube.*

15. Use the Nanodrop spectrophotometer to estimate the concentration of your purified PCR product. Be prepared to dilute the purified PCR product in Buffer EB if its concentration exceeds 40 ng/μL; we need the final concentration to fall between 20 and 40 ng/μL. **This is very important; DNA sequencing is very sensitive to template concentration. You don't want to waste five weeks of work by using the wrong amount of template!**

Finish electrophoresis:

16. Turn off the power supply. Wearing gloves, remove the cover and gently remove the tray holding the gel. Gently drain the buffer back into the gel box.

17. Wearing gloves, transfer the gel to a tray containing a solution of **ethidium bromide** (EtBr). **[NOTE: ETHIDIUM BROMIDE IS A SUSPECTED MUTAGEN.]** The gel will be gently shaken on a platform shaker for about 20 minutes, which should be adequate for staining.

18. Wearing gloves and **UV-shielding goggles**, transfer the stained gel (it won't look any different) to the UV light box. Turn on the UV light, photograph the gel, and turn off the UV light. Print a copy of the gel photograph. *This step will let you see the condition of your PCR.*

 **At this stage, we will decide which PCR products will be sequenced. This decision should be based on yield and on the quality of the PCR products. We will sequence 8 PCRs per gene per lab section.**

**B. DNA sequencing reactions**

1. Four cycle-sequencing master mixes will be prepared by volunteers. Each will be composed of 36 μL **DTCS reaction mix**, 9 μL **2.5 μM primer**, 9 μL**10× reaction buffer** and 121.5 μL **ultrapure H2O**. You will then prepare the following reactions using the cleaned PCRs that were diluted to the appropriate concentration:

 Forward *ND3* Rx Reverse *ND3* Rx

 0.5 μL cleaned *ND3* PCR 0.5 μL cleaned *ND3* PCR

 19.5 μL ND3-F master mix 19.5 μL ND3-R master mix

 Forward D-loop Rx Reverse D-loop Rx

 0.5 μL cleaned D-loop PCR 0.5 μL cleaned D-loop PCR

 19.5 μL dloop-F master mix 19.5 μL dloop-R master mix

 *Note: we are technically performing "half-reactions," which is standard for automated capillary sequencers. The DTCS reaction mix contains 2.5× reaction buffer, Taq polymerase and nucleotides. The additional 10×reaction buffer is needed to ensure proper buffer composition for the half-reaction.*

2. Load the tubes into the thermal cycler. *You will not be required to wait for cycle-sequencing to be completed.* The cycle-sequencing entails thirty cycles of (a) 96°C for 20 seconds, (b) 50°C for 20 seconds and (c) 60°C for 4 minutes. The reaction is then cooled to 4°C.

**METHODS FOR WEEK 4**

**A. Size estimation of PCR products**

You will be using the method of linear regression to generate the formula for the line that relates distance traveled on your gel to size of a DNA fragment. This formula will be based on six of the restriction digest fragments, whose sizes you know. Once you have this formula, you will use it to estimate the sizes of your two PCR products, based on the distances that they traveled on the gel.

1. For the following procedure, use as much of the graph paper as possible. Measure the distances traveled (in mm, measured to a resolution of 0.5 mm) by the six λ-*Hin*dIII restriction fragments between 500 and 10,000 bp in length. Plot these distances (on the *y*-axis) against **log10** of their size in bp (on the *x*-axis). This requires choosing appropriate scales for the two axes. Remember, it is not necessary for the axes to meet at the point (0,0).

2. Draw an eye-fitted line through the points.

3. Calculate the slope of the regression line through the six points used for the eye-fitted line, worksheet at the end of this handout; the worksheet uses the "machine" formulas for variance and covariance (see Appendix 2 for introduction to the standard and machine formulas).

4. Calculate the *y*-intercept (*c*) and write the equation of the regression line ($y = bx + c$) on your graph. By definition, the regression line goes through the point ($\overbar{y},\overbar{x}$). Substitute these into the formula$ \overbar{y} = b\overbar{x} + c$, and solve for *c*.

5. Using the equation for the line, calculate *y* for two arbitrary values of *x* (*e.g.*, for *x* = 3 and for *x* = 4), and mark these points on the graph. You can now draw the regression line through the two points. *How does this compare to your eye-fitted line?*

If your regression line does not seem to fit the points well, it probably means you have made an arithmetic error. A properly calculated regression line should go below some points and above others. If the fit of the regression line to your points is not *very* good, you'll need to fix this before continuing on to Step 13.

6. Using the regression equation, estimate the size of your PCR product. [Note: you will have to rearrange the algebraic formula to solve for *x*.] ***It would be worthwhile to compare these to the sizes predicted by the locations of the primers in the published sequences.***

**B. Analysis of DNA sequence variation**

1. In the DNA sequence alignments, identify all variable positions. For the protein-coding gene, determine if the polymorphism is synonymous or nonsynonymous. Be sure to use the vertebrate mitochondrial genetic code, which differs a little from the "universal" genetic code.

2. Calculate diversity among cows.

3. Calculate divergence between cows and bison.

**Appendix 1. Introduction to DNA Sequence Variation.**

 When population geneticists discuss genetic variation, they usually use the term *polymorphism* to refer to variation within a group (*e.g.*, a species or a population). The term *divergence* is used to refer to variation among groups. In this lab, you are dealing mainly with polymorphisms (DNA sequence variation among cows). However, you are also considering divergence from an outgroup (*i.e..*, buffalo). Before we can discuss the nature of genetic variation, it is essential to understand how it arises.

 Of course, variation requires mutation. When a mutation occurs at a particular nucleotide position in a chromosome, it has two possible fates. The first, which occurs most often, is extinction. Since the mutation begins at a frequency of 1/*N* (where *N* is the number of chromosomes in the population), there is a 1-1/*N* probability that it will drift to extinction. If the mutation is harmful, it is more likely to be lost. If it is beneficial, it has a decreased probability of becoming extinct, but this probability can still be quite high.

 Consider a chromosome with the following sequence:

 **5’-AAGCAGCGA-3’**

 **3’-TTCGTCGCT-5’**

If a mutation occurs at the fourth position of the upper strand, there is an initial mismatch:

 **5’-AAGTAGCGA-3’**

 **3’-TTCGTCGCT-5’**

 **^**

However, this will ultimately be resolved, either by repair or by replication. Some of the time, it will be resolved such that the base pair is changed:

 **5’-AAGCAGCGA-3’ → 5’-AAGTAGCGA-3’**

 **3’-TTCGTCGCT-5’ 3’-TTCATCGCT-5’**

The C:G and T:A base pairs that correspond to this nucleotide position are said to be **orthologous**. That is, they are descended from a common ancestral base pair by replication. [Orthology is a subset of homology.]

 **It is essential in the study of DNA sequence variation to compare orthologous nucleotides.** Therefore, a good deal of effort can go into aligning DNA sequences. If we focus on the upper strands of the above sequences, the alignment would be:

 **5’-AAGCAGCGA-3’**

 **5’-AAGTAGCGA-3’**

 The following is a section of an alignment of seven copies of the *period* gene. The first six are from *Drosophila melanogaster*; the seventh is from a close relative, *D. simulans*.

 Dm1 ct**G**agcggcagccactcctccggcagcagtggctatgg**G**ggcaagccctcgacgcaggcc

 Dm2 ct**G**agcggcagccactcctccggcagcagtggctatgg**G**ggcaagccctcgacgcaggcc

 Dm3 ct**C**agcggcagccactcctccggcagcagtggctatgg**G**ggcaagccctcgacgcaggcc

 Dm4 ct**G**agcggcagccactcctccggcagcagtggctatgg**G**ggcaagccctcgacgcaggcc

 Dm5 ct**G**agcggcagccactcctccggcagcagtggctatgg**G**ggcaagccctcgacgcaggcc

 Dm6 ct**G**agcggcagccactcctccggcagcagtggctatgg**G**ggcaagccctcgacgcaggcc

 Ds ct**G**agcggcagccactcctccggcagcagtggctatgg**C**ggcaagccctcgacgcaggcc

 ^ ^

In this stretch of aligned sequence, there are two variable positions. The first represents a polymorphism within *D. melanogaster*. The second represents divergence between *D. melanogaster* and *D. simulans*. It should be obvious that the sites are orthologous, because most of the surrounding sequence is invariant. If these other sites were not orthologous, it would be an incredible coincidence that they would be identical!

 In addition to categorizing variation as polymorphism or divergence, the nature of the variation is also important to us. Do the nucleotide changes occur in protein-coding regions? If so, do they change the encoded amino acid? It turns out that, in the *Drosophila* sequences above, both changes are synonymous. That is, they occur in codons of a protein-coding region, but the amino acids represented by the codons are unchanged. CTG and CTC both encode leucine, while GGG and GGC both encode glycine. [A note of caution: you are analyzing a mitochondrial protein-coding gene, and the genetic code differs slightly from the "universal" genetic code.]

 In this lab, you will only be aligning the protein-coding sequence of the *ND3* gene. You will certainly obtain sequence to DNA that flanks this gene, since the PCR product is about twice the size of the gene, but you will “trim” the flanking sequence before doing any analysis.

 You will also be aligning the D-loop sequences. Since sequencing reactions differ in their efficiency, an *ad hoc* decision will be made as to the region that will be compared across samples. Remember, this is not a protein-coding region.

 While there are many ways to quantify polymorphism and divergence, we will use the most straightforward: average pairwise variation. If there are *k* cow sequences in the sample, then there are *k*(*k*-1)/2 ways to pair and compare these sequences. Population geneticists use the symbol π to represent the average proportion of pairwise differences per site among sequences (a.k.a., "average nucleotide diversity"). For the stretch of the *period* gene shown above, the following triangular matrix of pairwise variation can be constructed:

 Dm1 Dm2 Dm3 Dm4 Dm5

 Dm2 0/60

 Dm3 1/60 1/60

 Dm4 0/60 0/60 1/60

 Dm5 0/60 0/60 1/60 0/60

 Dm6 0/60 0/60 1/60 0/60 0/60

If we sum these (5/60) and divide by the number of comparisons (15), we calculate an average pairwise variation of π = 0.0056.

 With respect to divergence, there are *k1* × *k2* pairwise comparisons, where *k1* and *k2* are the sample sizes of the respective species. In lab, *k1* would be the number of cow sequences and *k2* would be 1 (since there will be only one buffalo sequence). The symbol *D* is used by population geneticists to represent average divergence. For the stretch of the *period* gene shown above, the following simple matrix of pairwise divergence can be constructed:

 Dm1 Dm2 Dm3 Dm4 Dm5 Dm6

 Ds 1/60 1/60 2/60 1/60 1/60 1/60

If we sum these (7/60) and divide by the number of comparisons (6), we calculate an average pairwise divergence of *D* = 0.0194.

 ***In lab, you will be shown a shortcut for both of the calculations. However, you should understand why this shortcut works.***

**Appendix 2. Introduction to Linear Regression**

 The purpose of linear regression is to find the slope of a line that best fits a series of three or more points that are assumed to lie along a straight line. Often you can do this by graphing the points and drawing an eye-fitted line. However, we want to avoid the subjectivity associated with eye-fitting. Linear regression will allow you to calculate the slope (*b*) and *y*-intercept (*c*) of a line with the formula $y'=bx+c$, where $y'$ is the predicted value of *y* that corresponds to a given value of *x*. This line is a compromise that lies closest to the points − where "closeness" is defined by the squared vertical distance of each point from the line.

The slope of the best-fit line, *b*, is calculated as the covariance of *x* and *y* divided by the variance of *x*. For a small number of points, both covariance and variance can be calculated using the standard formulas:

$$COV\_{xy}=\frac{\sum\_{i=1}^{N}\left(x\_{i}-\overbar{x}\right)\left(y\_{i}-\overbar{y}\right)}{N-1} =\frac{\sum\_{}^{}\left(x-\overbar{x}\right)\left(y-\overbar{y}\right)}{N-1},$$

where *N* is the number of points. The variance of *x* is calculated as

$$s\_{x}^{2}=\frac{\sum\_{i=1}^{N}\left(x\_{i}-\overbar{x}\right)^{2}}{N-1}=\frac{\sum\_{}^{}\left(x-\overbar{x}\right)^{2}}{N-1} .$$

Notice that, to use these formulas directly, you would first have to calculate the means of *x* and *y*. However, there is an alternative method − one that does not require calculating the means. This method, called the **"machine" method**, is the one used by computers and calculators. All you need to know for the machine method are (i) the sum of the *x* values, (ii) the sum of the *y* values, (iii) the sum of *xy* for each point, and (iv) the sum of the squared *x* values. Every time you add a point to the data set, *x*, *y*, *xy* and *x*2 are simply added by the computer or calculator to the cumulative sums.

$$COV\_{xy}=\frac{\sum\_{i=1}^{N}x\_{i}y\_{i}-\frac{\left(\sum\_{i=1}^{N}x\right)\left(\sum\_{i=1}^{N}y\right)}{N}}{N-1}=\frac{\sum\_{}^{}xy-\frac{\sum\_{}^{}x\sum\_{}^{}y}{N}}{N-1} ,$$

$$s\_{x}^{2}=\frac{\sum\_{i=1}^{N}x\_{i}^{2}-\frac{\left(\sum\_{i=1}^{N}x\right)^{2}}{N}}{N-1} =\frac{\sum\_{}^{}x^{2}-\frac{\left(\sum\_{}^{}x\right)^{2}}{N}}{N-1}$$

WORKSHEET

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Fragment length (bp) | log10 fragmentlength (*x*i) | distance migrated (*y*i) | *x*i2 | *x*i*y*i |
| 9,420 |  |  |  |  |
| 6,560 |  |  |  |  |
| 4,380 |  |  |  |  |
| 2,320 |  |  |  |  |
| 2,020 |  |  |  |  |
| 560 |  |  |  |  |
| Sum | *A* | *C* | *E* | *F* |
| Mean | *B* | *D* |  |  |

Since there are 6 points on the line, *N* = 6. If you are missing a point (sometimes the 560 bp band is difficult to see), then adjust *N* accordingly!

(1) Calculate the numerator

for the slope of the line:

[*F* - (*A*×C)/*N*] / (*N*-1)

(2) Calculate the denominator

for the slope of the line:

(*E* - *A*2/*N*) / (*N*-1):

(3) Calculate the

slope (*b*):

 (4) Calculate the

 *y*-intercept (*c*):

(5) Write out the equation (*y* = *bx* + *c*):