Molecular Techniques Exam I ANSWER KEY

1) **Agarose Gel Electrophoresis (10pts)**

a) Describe the factors that one must consider when running an agarose gel for analysis vs isolating DNA from an enzyme digest.

- **Percent agarose:** Matched to size of fragments to be analyzed (1% 500-10,000 bp; 1.2% 400-7000 bp; 1.5%, 200-3000 bp; 2%, 50-2000 bp.)
- **Buffer:** TAE is better for high resolution of long nucleic acid fragments (1500+ bp) but has a lower buffering capacity than TBE. TBE has a greater buffering capacity and provides sharper resolution of smaller fragments than TAE. However, borate in TBE gels interferes with enzymes and should not be used when recovery of nucleic acids is planned.
- **Melting temp.** Standard agarose gels at 35-42°C and melts at 85-95°C. It provides superior resolution to low-melting point agarose, and is better for analysis, but is not amenable to isolation of DNA by melting. Low-melting point agarose gels at 24–28°C and melts ≤65.5° and is chosen when planning to recover DNA by melting.
- **EEO:** The agarose polymer contains charged groups, especially sulfate, that retards the movement of DNA by electroendosmosis (EEO) – the flow of solvent through the gel in the opposite direction from DNA migration. Low EEO (low sulfate) agarose provides somewhat better resolution than standard agarose, but is expensive for routine use.

b) What is the depth of gel and how much buffer should be used when preparing an agarose gel?

c) The recommended thickness for agarose gel is 3–4 mm; a gel thicker than 5mm will result in fuzzy bands and higher staining background. For any buffer the depth of buffer over the gel should range from 3 to 5 mm. Too much buffer will distort bands and cause heating and partial melting of the gel. Too little buffer and the gel is likely to partially dry out. Similarly, the amount of running buffer to cover over the gel in an electrophoresis apparatus is 3–5 mm. Too much buffer decreases DNA mobility and causes band distortion.

d) What are the most common problems encountered with agarose gels?

- *Incorrect buffer (e.g. Use of TBE for LMP agarose) or improperly prepared or diluted buffer (causes smeared bands)*
- *Use of water instead of buffer (causes profound smearing)*
- *Mismatched gel percentage and fragments size (causes poor resolution)*
- *Too much DNA in sample (loss of resolution) or too little (failure to visualize bands)*
- *Too much voltage applied (causes gel to overheat, reducing resolution)*
- *Too little voltage applied (allows bands to diffuse, reducing resolution)*
- *Punctured wells (sample leaks out)*
- *Bubbles in gel lanes (distorts bands)*
- *Reversal of leads (causes sample to migrate backwards ("retrophoresis")*

We want to see most of this information in the answer. Missing or poorly written information results in loss of points.

2) Restriction Digest (10pts)

a) Create a protocol for a standard restriction digest with a total volume of 50 µl digest. Pick two enzymes from NEB and use the information from that site to create your protocol. Your plasmid is at 0.3 µg / µl.

- Since you are to perform a double-digest and have your choice of restriction enzymes, pick two that have full activity in the *same* buffer (In real life, of course, you'd have to choose enzymes whose recognition sites are actually contained within the DNA you wish to digest). For example, Xba1 (T/CTAGA) and Xho1 (C/TCGAG) both have full activity in either NEB Cut-Smart buffer, or Buffer 2.1. You can use either. If we wish to perform a standard digest of 1mg of DNA in 50µl total volume you would need:

DNA (0.3mg/µl)	3.3 µl
10X Buffer (Cut-Smart)	5.0 µl
Water	39.7 µl
XbaI (20U/µl) diluted 1:20	1.0 µl
Xho1 (5U/µl) diluted 1:5	1.0 µl

Incubate at 37◦ 1 hour

Full credit requires written description plus table. Loss of points: Failure to use compatible buffer; separate digestions; failure to specify buffer; failure to specify enzyme unit concentration; over-digestion; volume does not add up to 50 µl.

b) Now you wish to do an 'over digestion' in which you will put the maximum allowable enzyme in the digest. (Indicate how many Units will be in the digest and how you would set this digest up differently.)

- Over-digestion is 5-10 units of enzyme per mg of DNA. Any more than that and you risk star activity. To set up the reaction with a 10X over-digestion using the same enzymes, mix:

DNA (0.3mg/µl)	0.3 µl
10X Buffer (Cut-Smart)	5.0 µl
Water	39.2 µl
XbaI (20U/µl)	0.5 µl (or 1 µl of 1:2 diluted enzyme; do not try to measure less than 0.5 µl)
Xho1 (5U/µl)	2.0 µl

Incubate at 37◦ 1 hour

Loss of points: Greater than 10X over-digestion; glycerol concentration over 5% (enzymes are supplied in 50% glycerol); volume does not add up to 50 µl

c) What does "star activity" mean and why should this be a consideration in a restriction digest?

- **Star activity** is the relaxation or alteration of the specificity of restriction enzyme cleavage of DNA that can occur under reaction conditions that differ significantly from those optimal for the enzyme. The result is typically cleavage at non-canonical recognition sites, or sometimes complete loss of specificity.

- Conditions that can lead to star activity include incorrect buffer use; high (> 5% v/v) glycerol concentrations and for some enzymes (e.g. HindIII) the presence of Mg^{2+}. Avoided by use of compatible buffer and a by keeping glycerol concentration below 5%.

Loss of points: Incorrect definition; failure to define conditions that lead to it and how to avoid it.

3) DNA Calculations (20pts)

Monica purified the pBIT plasmid DNA by alkaline lysis followed by anion exchange chromatography and resuspended the final DNA pellet in a total volume of 500 µl TE buffer. She used the Nanodrop (as you did in lab) and obtained the following absorbance readings: *Show your math and circle the answer. Units must be correct!*

A230 = 0.16 **A260 = 2.9** **A280 = 2.2**

a) **What is the concentration of DNA in her sample? (Be sure to include the appropriate units!)**

- The extinction coefficient at 260 nM for double-stranded DNA is 50 ng-cm/µl for double-stranded DNA, therefore, DNA concentration in the tube = 2.9 x 50 or **145 ng/µl (µg/ml).**

  (Note: It does not matter if the DNA is diluted in 500 µl of TE or 5L of TE – it was measured straight out of the tube with no further dilution!)

Loss of points: Inability to multiply 2.9 x 50; incorrect units; Using a 1:500 (or other) dilution factor for undiluted sample.

b) **What volume of pBIT plasmid DNA would she need to pipet to obtain 50 ng of plasmid to use in a ligation? Explain thoroughly including dilutions that need to be done, if necessary.**

- DNA is currently at a concentration of 145 ng/µl. 50/145 = 0.34 µl – and is the volume that contains 50 ng. However, this is too small a volume to pipette accurately. So, the original DNA solution should be diluted first (e.g. if solution is diluted 1:10, then 3.4 µl will contain 50 ng of DNA).

Full credit was given, even if the initial concentration from part A was incorrect – as long as the calculations for part B were consistent. Loss of points: Incorrect calculation based on assumed starting DNA concentration; attempt to pipette absurdly small volumes of undiluted solution.

c) **She wants to set up a ligation reaction with 50 ng of vector DNA. If the vector is 4 kb and the insert is 900 bp, how many nanograms of insert DNA would she need in order to obtain a 1:3 vector to insert ratio?**

- A mole of insert weighs 900/4000 (0.225) as much as a mole of vector – hence, 3 moles of insert weighs 2700/4000 (0.675) times that of a mole of vector. Hence, for

insert and vector of these relative sizes, a 3:1 insert to vector ratio means that for a given weight of vector you will need 0.675 times that weight in insert.

0.675 x 50 ng = 33.75 ng of insert DNA.

Loss of points: Incorrect calculations, math errors.

d) **Is the DNA pure? If yes, how do you know? If not, what is the most likely contaminant?**

- Purity of nucleic acid is determined by OD 260/ OD 280 ratio. Pure DNA has a 260/280 ratio of ~1.8. OUR solution had an OD 260 of 2.9 and an OD 280 of 2.2. Hence the OD 260/280 ratio is 1.31 – clearly NOT a pure solution.

  Contamination is most likely due to protein contamination of the DNA sample.

Loss of points: Incorrect calculations; not mentioning the expected ratio for pure DNA; attribution of impurity to contamination with substances that do NOT absorb UV at 260 or 280 (e.g. ethanol), or to RNA contamination – which would increase the 260/280 ratio, not decrease it (pure RNA has a 260/280 of ~2.0)

4) Nucleotide Basics (15pts)

You were given a microfuge tube of plasmid DNA, a tube of RNA and a tube of chromosomal DNA. In a simple mistake, you have forgotten to label the tubes and didn't notice this until after you ran the sample through a mini-prep silica column (Qiagen).

a) **Which of the tubes will bind to the silica**?

- The binding of nucleic acids to silica involves reduction of the silica surface negative charge in high ionic strength (high salt) buffer, resulting in a decrease in the electrostatic repulsion between the negatively charged DNA and the negatively charged silica, and nucleic acid binding by hydrophobic interactions. The process is reversed by low ionic strength buffer (e.g. TE), allowing for elution.

- Because the RNA has been degraded by RNAseA as part of the mini-prep process, it binds weakly and is removed from the column by the wash buffer which contains a lower salt concentration than the binding buffer.

- Chromosomal DNA is denatured by the alkaline lysis step in the mini-prep. When this is neutralized in the presence of detergent (SDS), double-stranded circular plasmid re-forms and remains in solution, but chromosomal DNA forms numerous inter-strand hybrids resulting in an insoluble mass that gets trapped in the

SDS/potassium precipitate. This is then removed from the lysate in the centrifugation.

- **Hence, both plasmid DNA and RNA will initially bind, but only the plasmid DNA will be retained on the column and eluted during the final step**

b) When will the DNA or RNA elute if they bind (assume a standard procedure we used in class) and propose a simple experiment to tell which is which.

- Both will bind, but assuming RNAseA treatment, the wash buffer will elute RNA but not plasmid DNA. After the Qiagen procedure, only the tube containing plasmid DNA at the start will contain significant nucleic acid (as assessed by OD260/280) at the end.

Grading determined on what fraction of this was correctly described.

5) Bacterial Strains in Molec Bio (15 pts)

Kumar is planning to prepare plasmid DNA to clone and asked you to get a competent cell out of the freezer for him to use. You find lots of types of competent cells and have things narrowed down to a BL21 (DE3), a DH5alpha, and JM109.

a) What are the reasons for using EACH of these and which should you bring to Kumar?

- BL21(DE3) is a derivative of the B strain of E.coli that is deficient in Lon protease (cytoplasmic) and OmpT protease (outer membrane) and is used for *recombinant protein expression*. The DE3 designation means that respective strains contain integrated λDE3 phage DNA that carries the gene for T7 RNA polymerase under control of the lacUV5 promoter. This allows one to use IPTG to induce expression of the T7 RNA polymerase in order to express recombinant genes cloned downstream of a T7 promoter.

- DH5α is a derivative of strain K12 developed by Doug Hanahan (hence 'DH'). These cells can be made highly competent and is best for *general cloning & sub-cloning* where protein expression is not required. The strain supports blue/white screening and also has a number of introduced mutations (recA1 and endA1) that increase insert stability and improve the quality and yield of plasmid DNA prepared from minipreps.

- JM109 is a derivative of K-12 developed by Joachim Messing (hence 'JM') that carries the F' ('fertility') episome encoding the F pilus that is necessary for M13 bacteriophage to attach to and enter bacterial cells. JM 109 cells that are transformed with a plasmid carrying an M13 origin of replication ("phagemid"), will,

when infected by helper M13 phage, *produce single stranded DNA* suitable for Sanger sequencing.

The choice of plasmid will depend on what Kumar wants to do. If he just needs to clone or subclone, DH5α is best. If he needs to express a protein in bacterial cells, BL21(DE3) is best; If he needs single-stranded DNA to do Sanger sequencing, JM109 is best.

Loss of points: Failure to recognize the different uses of these strains; failure to mention the genes/mutations that facilitate these uses

b) How is a competent cell prepared?

- Competent cells are bacterial cells that are capable of taking up exogenous plasmid DNA at appreciable frequency.

- **Chemically competent cells** are prepared by suspending a log-phase culture of *E.coli* in cold buffered $CaCl_2$ solution (protocols vary between 30-100 mM). Positively charged divalent cations such as calcium ions (Ca2+) attract both the negatively charged DNA backbone (phosphate) and the negatively charged groups in the lipopolysaccharides in the bacterial cell wall and enhances the permeability of bacterial cells to plasmid DNA.

- **Electrocompetent cells** are used with electroporation -- electrical pulses that create "pores" in the bacterial cell wall allowing genetic material like plasmids to penetrate. There is no special chemical treatment of cells required other ensuring that cultures are used in log-phase growth.

Loss of points: Failure to define "competence;" Failure to mention the role of cations in chemical competence; Providing a transformation protocol instead of a competent cell preparation protocol.

c) And, oh, by the way... what is the purpose of the 'recovery' step in transformation? When does it happen in the protocol?

- Resistance of bacteria to antibiotics is mediated by plasmid genes conferring antibiotic resistance. These genes work by encoding enzymes that degrade particular antibiotics (e.g. b-lactamase for the ampicillin resistance gene). Protein expression requires transcription and translation and is not instantaneous; hence, some period of time is required for cells to grow and express the relevant enzymes

before exposing them to antibiotic. This is the purpose of the 37° incubation with shaking after transformation and prior to plating on antibiotic-containing media.

Loss of points: Drivel about how bacteria need time to "recover" from the "shock" of transformation.

6) Software and Molec Bio (15 pts)

Using SnapGene create a plasmid map for watermelon MDH without the precursor. Use the information from our labs to create your map. Include all the pertinent information on the plasmid map.

- We are looking for a plasmid map that defines the position of the MDH insert in the MCS, locates the restriction sites upstream and downstream from the cloning site in the MCS, identifies forward and reversing sequencing primer sites, T7 or T3 promoter sites for expression, and antibiotic resistance gene.

Loss of points: Failure to locate the insert on the map; failure to identify these key features on the map; inclusion of every restriction site in the plasmid.

7) Restriction Enzyme Cloning (20pts)

You have to clone a 1.2 kb insert from a plasmid pTrouble (it is made up-don't bother looking for it) into pET30b using EcoRI and KpnI. Assume that the cut will leave the reading frame intact. Describe the workflow starting with each plasmid purified at 1 mg/ml DNA. Refer to chapter 2 in your book and your notes for reference material.

1. Double-digest 10µl (10µg) pTrouble with EcoR1 and Kpn1 using NEB buffer 1 (both enzymes have 100% activity in this buffer) in 20µl final volume using 10 -20 units (1µl) of each enzyme at 37° for 1 hour.
2. Double-digest 10µl (10µg) pET30b under the same conditions as above.
3. Add 5ul 6X loading buffer and run both on 0.7% low-melting agarose gel in TAE;100V ~30min.
4. Place gel on UV transilluminator; cut out 1.2 kb insert band from pTrouble digest and 5.4 kb plasmid band from pET30b digest; place in separate microfuge tubes.
5. Melt gel slices in 65° water bath for 5 minutes; Add 2.5 volumes TE
6. Clean using Phenol/chloroform or a silica column (Qiagen) method and precipitate DNA with ethanol.
7. Mix a 3:1-5:1 molar ratio of insert to vector; since vector is 5.4 kb and insert is 1.2 kb, 1µg of insert has 4.5 times the number of molecules as 1µg vector. Hence, an equivalent

weight is about a 4.5:1 molar ratio of insert to vector. Mix ~2µg of insert and ~2µg vector together.

8. Add 2µl 10X T4 ligase buffer and bring to 19µl with water. Add 1µl T4 ligase; Incubate RT 10 minutes.
9. Heat inactivate ligase at 65°C for 10 minutes.
10. Chill on ice and transform 2µl of the reaction into 50µl competent cells. Mix gently.
11. Place on ice 30 minutes; then heat shock at 42°C for 30 seconds
12. Add 950µl SOC; place in shaker for 60 minutes, 37°C
13. Plate on warm kanamycin plates; incubate overnight at 37°C

Loss of points: Failure to gel isolate fragments (else they will relegate); CIP treatment (unnecessary with incompatible ends); failure to clean up isolated bands; Failure to specify kanamycin as the antibiotic to be used in plates (pET30b has the Kan resistance gene).

8) <u>Classic DNA Sequencing Chemistry</u> <u>(25 pts)</u>

a. How does the chemistry of Sanger sequencing differ from Maxam-Gilbert sequencing? (10 pts.)

- Sanger sequencing relies on extension of an oligonucleotide primer in the presence of dNTPs and ddNTPs ('terminators') by DNA polymerase and a single-stranded template to be sequenced. Either the primer or the terminators can be labeled. Maxam-Gilbert sequencing relies on chemical cleavage of end-labelled, double-stranded DNA using dimethyl sulfate and hydrazine and reaction conditions (+/- formic acid; +/- NaCl) that cleave DNA after specific bases. Both techniques generate fragments that differ by single base pair lengths and are resolved by acrylamide gel electrophoresis.

Most people went into far more detail than this, but the essential difference is chemical cleavage of double-stranded DNA vs. enzymatic extension/termination of a primer from a single-stranded template.

b. How does the early radioactive Sanger method differ from the modern approach using four colored dyes in a dideoxy sequencing method (be specific and more detailed than the notes alone)? (15 pts.)

- The original Sanger technique employed a 5' radiolabeled primer that anneals to a single-stranded template (usually generated by Phage M13 cloning) and is extended by DNA polymerase in the presence of a mixture of dNTPs and ddNTPs. The reactions were carried out in four separate tubes, each containing a different dideoxynucleotide terminator. This gave rise to a series of fragments in each reaction tube of varying lengths tall of which terminate at the complementary base in the template corresponding to the

terminator used in the reaction. The fragments are resolved in four adjacent lanes on an acrylamide gel and the sequence is read by determining which reaction resulted in chain termination for every fragment size.

- Modern Sanger sequencing is performed by automated Sequencers using a single reaction in which each of the four terminators is labeled with a different fluorescent dye. This enables the fragments to be resolved in a single lane of a slab (or capillary) gel, with the sequence determined by the order of the colors that pass by a laser fluorescence detector. Single-stranded template is generated by simply melting a double stranded fragment to be sequenced at high temperature and using a thermostable polymerase. This is often repeated in several cycles of melting, primer annealing and extension in a process called "cycle sequencing."

  Loss of points: Mention of radiolabeled terminators (it was the primers that were labeled in original technique); Failure to describe the four tube vs. one tube method; failure to recognize radioactive vs. fluorescent tags

9) NextGen Sequencing (40pts)

a) How does NextGen sequencing differ from a standard modern four colored sequencing in throughput? (5)

- Modern capillary Sanger four-color sequencers can analyze up to 384 samples per run with typical read lengths of 500-1000 bp per sample. Next-Gen sequencers can simultaneously read from 25 to 400 million sequences, with typical read lengths of 150-300 bases

b) Describe either Illumina or another mechanism (step by step) for a NextGen sequencing method. (15)

- Illumina sequencing can be divided into three phases: Library generation, Solid-phase bridge amplification and actual sequencing.

- **Library generation:** To prepare a library for sequencing, DNA to be analyzed is first sheared into fragments, blunt-ended with T4 polymerase, and ligated to specialized adaptor molecules. These adaptor molecules contain complementary sequences to form a double-stranded portion capable of ligating to the DNA, as well as non-complementary portions that contain short index sequences and PCR/sequencing priming sites. When amplified by PCR using the primers complementary to these sites, a double stranded product is obtained in which the 5' and 3' ends of the molecule contain different, but identifiable sequences. The PCR-amplified DNA is then resolved on an agarose gel to isolate fragments between 200-1000 base pairs in length.

- **Solid-phase bridge amplification:** The amplified DNA is added to a 'flow cell' which contains a lawn of two different oligonucleotides complementary to each of the two different priming sites on either end of each DNA fragment. The DNA is melted apart and the single-stranded fragments are 'captured' by the complementary oligonucleotides on the surface of the flow cells. Non-bound DNA is washed away, and the primers are extended by polymerase and dNTPs. This creates a double-stranded fragment that has the5' end of one of it strands covalently bound to the surface of the flow cell. The DNA is then melted again, and the non-covalently bound strand is washed away.

- This remaining single-stranded DNA then folds over and the 3' end is captured by the alternative complementary oligo adjacent to it on the flow cell, to form a bridge. This second oligo is then extended by DNA polymerase, after which the DNA is melted apart and extended by DNA polymerase as before. The process is repeated several times forming local "clusters" of identical and reverse complementary DNA molecules attached to the solid phase flow cell at their 5' ends.

- **Sequencing:** Sequencing is performed using primers complementary to the original adaptor sequences extended using DNA polymerase and fluorescent ddNTPs. The key innovation is the use of reversible terminators in which the fluorescent dye is cleaved and the ddNTP is converted into a dNTP in a single reaction. Each cycle therefore extends the primer by one base and causes a cluster to fluoresce in a single color depending on whether an A, G, T or C has been added. The reaction is then “reversed” and the next cycle is performed. A digital camera records the color of each cluster after every cycle and obtains the sequences of each. Clusters with the same index sequence (and hence derived from the same PCR reaction) are pooled together for analysis and contig generation.

Grading based on fraction of above information included in answer, absence of incorrect information, and quality of explanation.

c) The following series of DNA sequences were obtained from 60 base reads on an Illumina NexSeq instrument in a shotgun sequencing experiment. Sequences were obtained in both directions. They were generated in a large run containing numerous other genes, but these were all identifiable by a common index sequence originally contained in the primers. Using BLAST and any other tools you care to, define the forward and reverse primer sequences and the index sequence, and assemble and align these reads into a contig. Finally, identify the gene from which they were derived. (20)

```
5′CAGCTATGACAGGTCAGGCCACGAAGGGTCCTCCTCCTCAGCCATCTCCTGGTCGTCTTC
5′TCGTTCGACCAGGTCAGGAGTGAACTGCAGCACTTAGCCACCTCCCGCCTGCCAAGCTTC
5′CAGCTATGACAGGTCAGGTCGTCTTCGTAGGAGGGTGGGTTTAGCCTGGGTGTTTTGTGC
5′CAGCTATGACAGGTCATTGTGCTGCACGATTCTGGAGTAGTACGTGTTGAGGCAGAAGAC
5′CAGCTATGACAGGTCACAGAAGACGTCGGCCGTCGCCCTGATGAACCTCTTCTCTTCCTC
5′CAGCTATGACAGGTCACTTCCTCAGTGAAGCTTGGCAGGCGGGAGGTGGCTAAGTGCTGC
5′CAGCTATGACAGGTCAAAGTGCTGCAGTTCACTCCTGTTCCCCACTTTCCACTTCATGGT
5′CAGCTATGACAGGTCACATTTCAGGCATATGACCCAGGGAATGTTGGAAGATTCTTTAAG
5′CAGCTATGACAGGTCACTGGTGTCAGGATAGTCTCCGTTTCTAAAAATGGGGTGACAAAC
5′CAGCTATGACAGGTCACAAACCAGCCCAGGGAGAACTGCAGCATTCGGTCAGCGGCTTCC
5′TCGTTCGACCAGGTCAAGATGTGGAAGCCGGCTGACCGAATGCTGCAGTTCTCCCTGGGC
5′TCGTTCGACCAGGTCACCCATTTTTAGAAACGGAGACTATCCTGACACCATGAAGTGGAA
5′CAGCTATGACAGGTCAGAACCCTCACACACCTTTCAAAGCCTCATTGATGTAGGTTTTGT
5′CAGCTATGACAGGTCAGCGTCCCCCAGGGCGCAGCTCTGTTCATTGCCGTGGAAGGCCAC
5′CAGCTATGACAGGTCATGGTTTTGTGGTTTCCTTCGTTCCTCCGTGTTCGGTTGGTCGCC
5′TCGTTCGACCAGGTCATCTGCCTCAACACGTACTACTCCAGAATCGTGCAGCACAAAACA
5′TCGTTCGACCAGGTCACAGGAGATGGCTGAGGAGGAGGACCCTTCGTGGCCTTCCACGGC
5′TCGTTCGACCAGGTCAGCTGCTGAACTGGATCAAGGAAGAGTATGGTGACATCCCCATTT
5′TCGTTCGACCAGGTCACGAACACGGAGGATACTGATAGGATATTTTACCACAAAACCTAC
5′CAGCTATGACAGGTCATGGTTCATAAATCATCCCGGCAAAACATTTCAGGCATAGACCCA
5′CAGCTATGACAGGTCAGGAAGATTTCTTTAAGGGGAACTGAACCCTCACACACCTTTCAA
```

There is no substitute for actually LOOKING at the sequences. Are there stretches of sequence IN COMMON? These must be either index sequences or primer sequences. If we insert a space between the relevant stretches, these become apparent:

```
5′CAGCTATGA CAGGTCA GGCCACGAAGGGTCCTCCTCCTCAGCCATCTCCTGGTCGTCTTC
5′TCGTTCGAC CAGGTCA GGAGTGAACTGCAGCACTTAGCCACCTCCCGCCTGCCAAGCTTC
5′CAGCTATGA CAGGTCA GGTCGTCTTCGTAGGAGGGTGGGTTTAGCCTGGGTGTTTTGTGC
5′CAGCTATGA CAGGTCA TTGTGCTGCACGATTCTGGAGTAGTACGTGTTGAGGCAGAAGAC
5′CAGCTATGA CAGGTCA CAGAAGACGTCGGCCGTCGCCCTGATGAACCTCTTCTCTTCCTC
5′CAGCTATGA CAGGTCA CTTCCTCAGTGAAGCTTGGCAGGCGGGAGGTGGCTAAGTGCTGC
5′CAGCTATGA CAGGTCA AAGTGCTGCAGTTCACTCCTGTTCCCCACTTTCCACTTCATGGT
5′CAGCTATGA CAGGTCA CATTTCAGGCATATGACCCAGGGAATGTTGGAAGATTCTTTAAG
5′CAGCTATGA CAGGTCA CTGGTGTCAGGATAGTCTCCGTTTCTAAAAATGGGGTGACAAAC
5′CAGCTATGA CAGGTCA CAAACCAGCCCAGGGAGAACTGCAGCATTCGGTCAGCGGCTTCC
5′TCGTTCGAC CAGGTCA AGATGTGGAAGCCGGCTGACCGAATGCTGCAGTTCTCCCTGGGC
5′TCGTTCGAC CAGGTCA CCCATTTTTAGAAACGGAGACTATCCTGACACCATGAAGTGGAA
5′CAGCTATGA CAGGTCA GAACCCTCACACACCTTTCAAAGCCTCATTGATGTAGGTTTTGT
5′CAGCTATGA CAGGTCA GCGTCCCCCAGGGCGCAGCTCTGTTCATTGCCGTGGAAGGCCAC
5′CAGCTATGA CAGGTCA TGGTTTTGTGGTTTCCTTCGTTCCTCCGTGTTCGGTTGGTCGCC
5′TCGTTCGAC CAGGTCA TCTGCCTCAACACGTACTACTCCAGAATCGTGCAGCACAAAACA
5′TCGTTCGAC CAGGTCA CAGGAGATGGCTGAGGAGGAGGACCCTTCGTGGCCTTCCACGGC
5′TCGTTCGAC CAGGTCA GCTGCTGAACTGGATCAAGGAAGAGTATGGTGACATCCCCATTT
5′TCGTTCGAC CAGGTCA CGAACACGGAGGATACTGATAGGATATTTTACCACAAAACCTAC
5′CAGCTATGA CAGGTCA TGGTTCATAAATCATCCCGGCAAAACATTTCAGGCATAGACCCA
5′CAGCTATGA CAGGTCA GGAAGATTTCTTTAAGGGGAACTGAACCCTCACACACCTTTCAA
```

First, notice the yellow highlighted sequence ‘CAGGTCA’ is present in EVERY read. This must be an INDEX sequence and is what allows thousands of different samples and genes to be analyzed simultaneously in the same run of the sequencer -- because all sequences with the same index are derived from the same PCR reaction and can be pooled together.

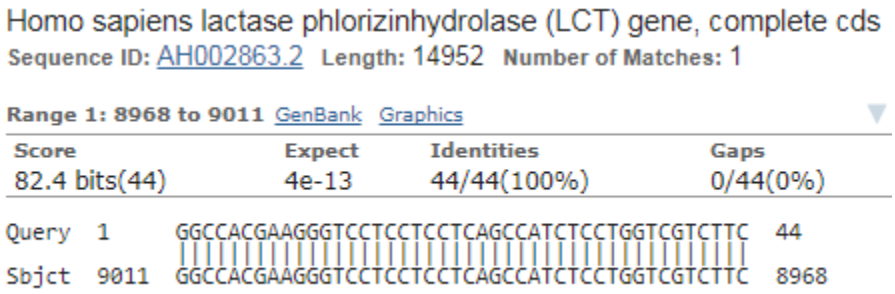
Next, notice that the 5' end of each read has one of two sequences, here designated in blue or green. These must represent sequences that distinguish the forward and reverse primers – and, like the index sequence, are part of the adaptors that were initially ligated to the DNA fragments and amplified by PCR. It is only the sequences that are un-

highlighted that are part of the gene of interest and must be assembled into a contig. Moreover, sequences that start with the green vs blue priming sequences must be in opposite orientations relative to the intact gene.

To align the sequences, start with any one – say, the first, and do a blastn search against the determined (non-adaptor) part of the sequence:

GGCCACGAAGGGTCCTCCTCCTCAGCCATCTCCTGGTCGTCTTC

Choose any 100% match – in this case, I chose the first one:



This search tells you three things:

First, the sequence comes from the human lactase gene.

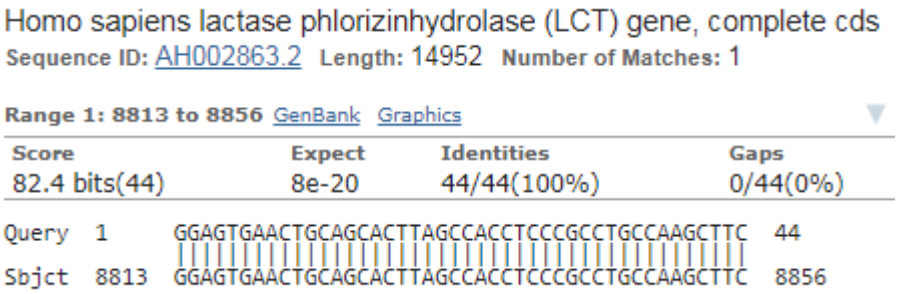
Second, the 'match' is from base 9011 to 8968 in the LCT gene. Because the numbers are going down instead of up, this tells you that you are dealing with the REVERSE strand – and therefore all the sequences with green-labeled primer sites must be derived from the REVERSE primer, and all the blue ones from the forward primer.

Finally, you are provided with a sequence ID (accession number) for this gene: AH002863.2. THIS is the parent sequence you should search against for all subsequent BLAST searches for the other sequences.

The next thing to is to BLAST search all the subsequent sequences by using the "Align two sequences" function and pasting the query sequence into the top box, and the Accession number into the bottom box. If we do that for the next sequence on the list,

GGAGTGAACTGCAGCACTTAGCCACCTCCCGCCTGCCAAGCTTC

Here we get:



Notice that this sequence was from a BLUE primer – the forward primer, and sure enough, now the match numbers go UP – from 8813 to 8856. By repeating this process for all 21 sequences, searching each against the LCT gene, you can figure out their relative positions:

5′CAGCTATGA CAGGTCA GGCCACGAAGGGTCCTCCTCCTCAGCCATCTCCTGGTCGTCTTC 9011-8968
5′TCGTTCGAC CAGGTCA GGAGTGAACTGCAGCACTTAGCCACCTCCCGCCTGCCAAGCTTC 8813-8856
5′CAGCTATGA CAGGTCA GGTCGTCTTCGTAGGAGGGTGGGTTTAGCCTGGGTGTTTTGTGC 8977-8934
5′CAGCTATGA CAGGTCA TTGTGCTGCACGATTCTGGAGTAGTACGTGTTGAGGCAGAAGAC 8939-8896
5′CAGCTATGA CAGGTCA CAGAAGACGTCGGCCGTCGCCCTGATGAACCTCTTCTCTTCCTC 8903-8860
5′CAGCTATGA CAGGTCA CTTCCTCAGTGAAGCTTGGCAGGCGGGAGGTGGCTAAGTGCTGC 8866-8823
5′CAGCTATGA CAGGTCA AAGTGCTGCAGTTCACTCCTGTTCCCCACTTTCCACTTCATGGT 8831-8788
5′CAGCTATGA CAGGTCA CATTTCAGGCATATGACCCAGGGAATGTTGGAAGATTCTTTAAG 9259-9215
5′CAGCTATGA CAGGTCA CTGGTGTCAGGATAGTCTCCGTTTCTAAAAATGGGGTGACAAAC 8791-8748
5′CAGCTATGA CAGGTCA CAAACCAGCCCAGGGAGAACTGCAGCATTCGGTCAGCGGCTTCC 8752-8710
5′TCGTTCGAC CAGGTCA AGATGTGGAAGCCGGCTGACCGAATGCTGCAGTTCTCCCTGGGC 8703-8745
5′TCGTTCGAC CAGGTCA CCCATTTTTAGAAACGGAGACTATCCTGACACCATGAAGTGGAA 8758-8801
5′CAGCTATGA CAGGTCA GAACCCTCACACACCTTTCAAAGCCTCATTGATGTAGGTTTTGT 9207-9164
5′CAGCTATGA CAGGTCA GCGTCCCCCAGGGCGCAGCTCTGTTCATTGCCGTGGAAGGCCAC 9049-9006
5′CAGCTATGA CAGGTCA TGGTTTTGTGGTTTCCTTCGTTCCTCCGTGTTCGGTTGGTCGCC 9143-9123
5′TCGTTCGAC CAGGTCA TCTGCCTCAACACGTACTACTCCAGAATCGTGCAGCACAAAACA 8900-8943
5′TCGTTCGAC CAGGTCA CAGGAGATGGCTGAGGAGGAGGACCCTTCGTGGCCTTCCACGGC 8977-9020
5′TCGTTCGAC CAGGTCA GCTGCTGAACTGGATCAAGGAAGAGTATGGTGACATCCCCATTT 9054-9097
5′TCGTTCGAC CAGGTCA CGAACACGGAGGATACTGATAGGATATTTTACCACAAAACCTAC 9131-9174
5′CAGCTATGA CAGGTCA TGGTTCATAAATCATCCCGGCAAAACATTTCAGGCATAGACCCA 9285-9240
5′CAGCTATGA CAGGTCA GGAAGATTTCTTTAAGGGGAACTGAACCCTCACACACCTTTCAA 9230-9187

The next thing to do is to bring all the sequences into the same orientation by reversing complementing all the sequences derived from the reverse primer. This can be done by hand, or by any number of online tools. Be sure to reverse the numbers as well.

GAAGACGACCAGGAGATGGCTGAGGAGGAGGACCCTTCGTGGCC 8968-9011
GGAGTGAACTGCAGCACTTAGCCACCTCCCGCCTGCCAAGCTTC 8813-8856
GCACAAAACACCCAGGCTAAACCCACCCTCCTACGAAGACGACC 8934-8977
GTCTTCTGCCTCAACACGTACTACTCCAGAATCGTGCAGCACAA 8896-8939
GAGGAAGAGAAGAGGTTCATCAGGGCGACGGCCGACGTCTTCTG 8860-8903
GCAGCACTTAGCCACCTCCCGCCTGCCAAGCTTCACTGAGGAAG 8823-8866
ACCATGAAGTGGAAAGTGGGGAACAGGAGTGAACTGCAGCACTT 8788-8831
CTTAAAGAATCTTCCAACATTCCCTGGGTCATATGCCTGAAATG 9215-9259
GTTTGTCACCCCATTTTTAGAAACGGAGACTATCCTGACACCAG 8748-8791
GGAAGCCGCTGACCGAATGCTGCAGTTCTCCCTGGGCTGGTTTG 8710-8752
AGATGTGGAAGCCGGCTGACCGAATGCTGCAGTTCTCCCTGGGC 8703-8745
CCCATTTTTAGAAACGGAGACTATCCTGACACCATGAAGTGGAA 8758-8801
ACAAAACCTACATCAATGAGGCTTTGAAAGGTGTGTGAGGGTTC 9164-9207
GTGGCCTTCCACGGCAATGAACAGAGCTGCGCCCTGGGGGACGC 9006-9049
GGCGACCAACCGAACACGGAGGAACGAAGGAAACCACAAAACCA 9123-9143
TCTGCCTCAACACGTACTACTCCAGAATCGTGCAGCACAAAACA 8900-8943
CAGGAGATGGCTGAGGAGGAGGACCCTTCGTGGCCTTCCACGGC 8977-9020
GCTGCTGAACTGGATCAAGGAAGAGTATGGTGACATCCCCATTT 9054-9097

```
CGAACACGGAGGATACTGATAGGATATTTTACCACAAAACCTAC 9131-9174
TGGGTCTATGCCTGAAATGTTTTGCCGGGATGATTTATGAACCA 9240-9285
TTGAAAGGTGTGTGAGGGTTCAGTTCCCCTTAAAGAAATCTTCC 9187-9230
```

Then, use the numbers to order the sequences:

```
AGATGTGGAAGCCGGCTGACCGAATGCTGCAGTTCTCCCTGGGC 8703-8745
GGAAGCCGCTGACCGAATGCTGCAGTTCTCCCTGGGCTGGTTTG 8710-8752
GTTTGTCACCCCATTTTTAGAAACGGAGACTATCCTGACACCAG 8748-8791
CCCATTTTTAGAAACGGAGACTATCCTGACACCATGAAGTGGAA 8758-8801
ACCATGAAGTGGAAAGTGGGGAACAGGAGTGAACTGCAGCACTT 8788-8831
GGAGTGAACTGCAGCACTTAGCCACCTCCCGCCTGCCAAGCTTC 8813-8856
GCAGCACTTAGCCACCTCCCGCCTGCCAAGCTTCACTGAGGAAG 8823-8866
GAGGAAGAGAAGAGGTTCATCAGGGCGACGGCCGACGTCTTCTG 8860-8903
GTCTTCTGCCTCAACACGTACTACTCCAGAATCGTGCAGCACAA 8896-8939
TCTGCCTCAACACGTACTACTCCAGAATCGTGCAGCACAAAACA 8900-8943
GCACAAAACACCCAGGCTAAACCCACCCTCCTACGAAGACGACC 8934-8977
GAAGACGACCAGGAGATGGCTGAGGAGGAGGACCCTTCGTGGCC 8968-9011
CAGGAGATGGCTGAGGAGGAGGACCCTTCGTGGCCTTCCACGGC 8977-9020
GTGGCCTTCCACGGCAATGAACAGAGCTGCGCCCTGGGGGACGC 9006-9049
GCTGCTGAACTGGATCAAGGAAGAGTATGGTGACATCCCCATTT 9054-9097
GGCGACCAACCGAACACGGAGGAACGAAGGAAACCACAAAACCA 9123-9143
CGAACACGGAGGATACTGATAGGATATTTTACCACAAAACCTAC 9131-9174
ACAAAACCTACATCAATGAGGCTTTGAAAGGTGTGTGAGGGTTC 9164-9207
TTGAAAGGTGTGTGAGGGTTCAGTTCCCCTTAAAGAAATCTTCC 9187-9230
CTTAAAGAATCTTCCAACATTCCCTGGGTCATATGCCTGAAATG 9215-9259
TGGGTCTATGCCTGAAATGTTTTGCCGGGATGATTTATGAACCA 9240-9285
```

Finally, align the overlaps into a contig:

```
AGATGTGGAAGCCGGCTGACCGAATGCTGCAGTTCTCCCTGGGC
      GGAAGCCGCTGACCGAATGCTGCAGTTCTCCCTGGGCTGGTTTG
                                              GTTTGTCACCCCATTTTTAGAAACGGAGACTATCCTGACACCAG

GTTTGTCACCCCATTTTTAGAAACGGAGACTATCCTGACACCAG
         CCCATTTTTAGAAACGGAGACTATCCTGACACCATGAAGTGGAA
                                      ACCATGAAGTGGAAAGTGGGGAACAGGAGTGAACTGCAGCACTT

ACCATGAAGTGGAAAGTGGGGAACAGGAGTGAACTGCAGCACTT
                          GGAGTGAACTGCAGCACTTAGCCACCTCCCGCCTGCCAAGCTTC
                                    GCAGCACTTAGCCACCTCCCGCCTGCCAAGCTTCACTGAGGAAG

GCAGCACTTAGCCACCTCCCGCCTGCCAAGCTTCACTGAGGAAG
                                       GAGGAAGAGAAGAGGTTCATCAGGGCGACGGCCGACGTCTTCTG

GAGGAAGAGAAGAGGTTCATCAGGGCGACGGCCGACGTCTTCTG
                                     GTCTTCTGCCTCAACACGTACTACTCCAGAATCGTGCAGCACAA
                                         TCTGCCTCAACACGTACTACTCCAGAATCGTGCAGCACAAAACA

TCTGCCTCAACACGTACTACTCCAGAATCGTGCAGCACAAAACA
                                    GCACAAAACACCCAGGCTAAACCCACCCTCCTACGAAGACGACC

GCACAAAACACCCAGGCTAAACCCACCCTCCTACGAAGACGACC
                                    GAAGACGACCAGGAGATGGCTGAGGAGGAGGACCCTTCGTGGCC
                                             CAGGAGATGGCTGAGGAGGAGGACCCTTCGTGGCCTTCCACGGC

CAGGAGATGGCTGAGGAGGAGGACCCTTCGTGGCCTTCCACGGC
                               GTGGCCTTCCACGGCAATGAACAGAGCTGCGCCCTGGGGGACGC


GCTGCTGAACTGGATCAAGGAAGAGTATGGTGACATCCCCATTT

GGCGACCAACCGAACACGGAGGAACGAAGGAAACCACAAAACCA
          CGAACACGGAGGATACTGATAGGATATTTTACCACAAAACCTAC
                                           ACAAAACCTACATCAATGAGGCTTTGAAAGGTGTGTGAGGGTTC

TTGAAAGGTGTGTGAGGGTTCAGTTCCCCTTAAAGAAATCTTCC
                              CTTAAAGAATCTTCCAACATTCCCTGGGTCATATGCCTGAAATG
```

TGGGTC-TATGCCTGAAATGTTTTGCCGGGATGATTTATGAACC

Notice that some of the sequences are not perfect matches to either the reference gene, or to other sequences in the contig. For example, the last sequence has missed an "A" in the base call and a gap needed to be inserted to align it to another fragment. There is also a gap where sequences don't overlap. This is typical – and is why this kind of sequencing only works when there is a reference sequence to compare it to – as now exists in GenBank. In any case, all of what we just did here is what the SOFTWARE of the sequencer does – including the alignment to the reference sequence.

Grading for part C:

18-20: Correct gene ID, identification of index sequence, forward and reverse primers and assembly into a contig showing overlaps.

16-18: Correct gene ID, identification of index sequence, forward and reverse primers but no contig assembly.

14-16: Correct gene ID, but failure to either identify index sequence or forward and reverse primers and no contig assembly.

12-14: Correct gene ID, but failure to identify both index sequence and forward and reverse primers and no contig assembly.

10:12 Gene ID only

0-10: No attempt at answer, or incorrect across the board.

10) Sequence Alignment (10pts)

A question that has long intrigued taxonomists and vertebrate biologists is the evolutionary relationship of the hippopotamus to other vertebrates. Although they are sometimes called "river hogs," hippos have often been put in the same taxa as whales.

As the molecular biologist on the team, yours will be a critical piece of evidence in the controversy! So weigh in: Using the alpha chain of hemoglobin as a reference protein, perform pair-wise alignments between hippo, whale and pig orthologs. Is hippo hemoglobin more related to that of the pig or of the whale? Does the answer change if you alter the substitution matrix from the default BLOSUM62?

Hemoglobin alpha chain Gen-Bank accession numbers: Whale P18971; Hippo: P19015; Pig: P01973 (Hint: Accession numbers can be copied or pasted into BLAST query boxes just like FASTA sequences, and the software will retrieve the sequences from Gen-Bank.)

- Hippo hemoglobin is more related to pig than whale, with 85% amino acid identity to pig vs. 82% identity to whale. Percent identical residues is, of course, not changed by

altering the substitution matrix; however, percent similarity IS changed – because the scoring for particular amino acid substitutions differs from one matrix to the next. As shown in the chart below, percent similarity between hippo and pig ranges from 89-96 percent, while amino acid similarity between hippo and whale ranges between 87 – 90% depending on the matrix used. However, in all cases where comparisons are made using the same substitution matrix, hippo hemoglobin alpha chain is slightly more similar to that of a pig than to a whale.

	Pig (Identity)	Pig (Similarity)	Whale(Identity)	Whale (Similarity)
BLOSUM45	85	90	82	89
BLOSUM50	85	90	82	89
BLOSUM62	85	90	82	89
BLOSUM80	85	90	82	87
BLOSUM90	85	89	82	87
PAM30	85	87	82	85
PAM70	85	90	82	90
PAM250	85	96	82	95

Grading:

9-10: All of the above information

7-8: Correct analysis, no chart of similarity by matrix.

5-6: Confusing Identity with similarity; Using E-values inappropriately

4-5 Coming up with greater similarity to whale than pig.

0-4 No answer at all