

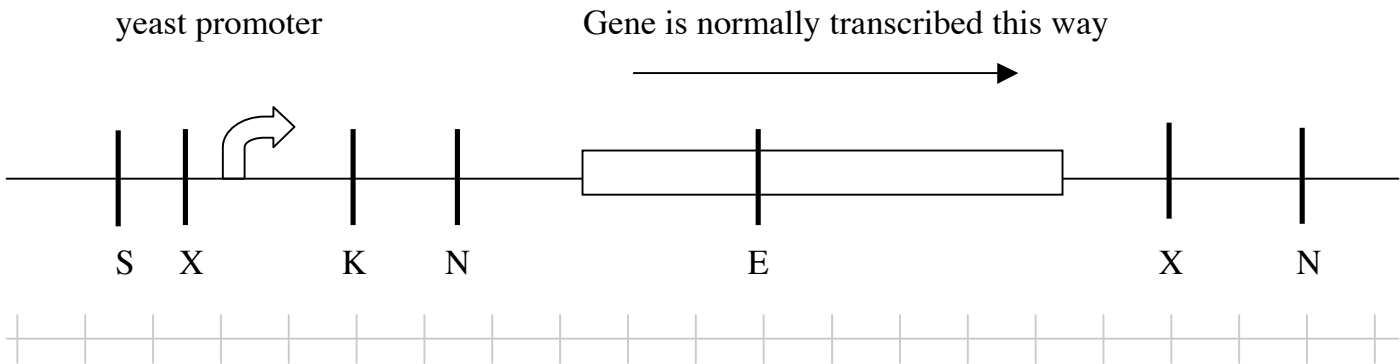
2006 7.012 Problem Set 4 KEY

Due before 5 PM on FRIDAY, October 27, 2006.

Turn answers in to the box outside of 68-120.

PLEASE WRITE YOUR ANSWERS ON THIS PRINTOUT.

1. You are studying a specific gene in yeast, and you want to express that yeast gene in *E. coli*. Your task is to design a strategy to insert the yeast gene into the bacterial plasmid. Below is a map of the area of the yeast genome surrounding the gene in which you are interested.

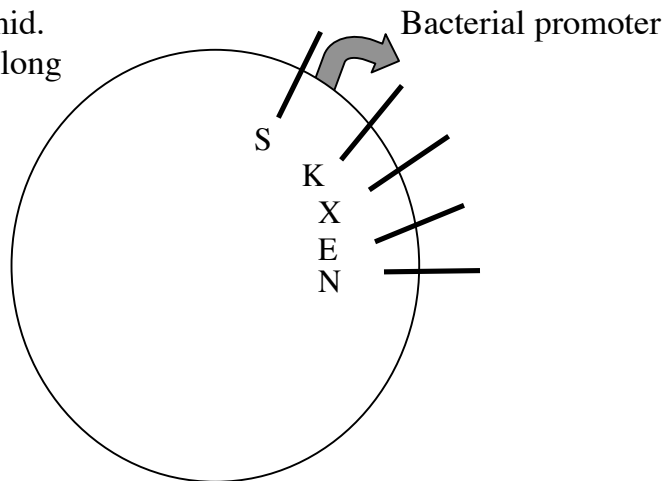


The distance between each tick mark placed on the line above is 100 bases in length.

Below are the enzymes you can use, with their specific cut sites shown 5' -XXXXXX-3'
3' -XXXXXX-5'

<i>Xba</i> I:	<i>Nde</i> I:	<i>Sal</i> I:	<i>EcoR</i> I:	<i>Kpn</i> I:
↓ TCTAGA AGATCT↑	↓ CATATG GTATAC↑	↓ GTCGAC CAGCTG↑	↓ GAATTC CTTAAG↑	↓ GGTACC CCATGG↑

This is the map of the plasmid.
The plasmid is 5,000 bases long
and the two farthest
restriction enzyme sites are
200 bases apart. The
plasmid has an ampicillin
resistance gene somewhere
on the plasmid distal from
the restriction cut sites.



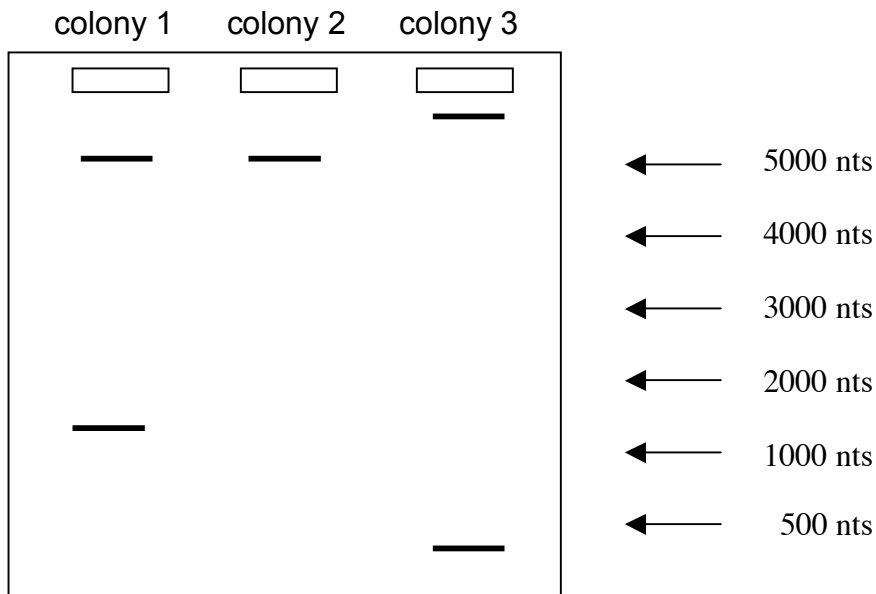
Name: KEY

(a) Which single restriction enzyme is the best choice for you use to design a way to get the insert into the vector if you can only use one single enzyme? Keep in mind that your goal is to have the yeast gene be expressed in the bacterial cells.

NdeI (N).

NdeI and XbaI are the only two enzymes that can excise the gene from the genome, as they are the only two enzymes that cut on both sides of the gene. NdeI is the only enzyme that cuts the yeast gene outside of the ORF but downstream of the yeast promoter (which you don't want as part of your insert because you want the gene to be expressed in bacteria, which means it must be under the control of a bacterial promoter).

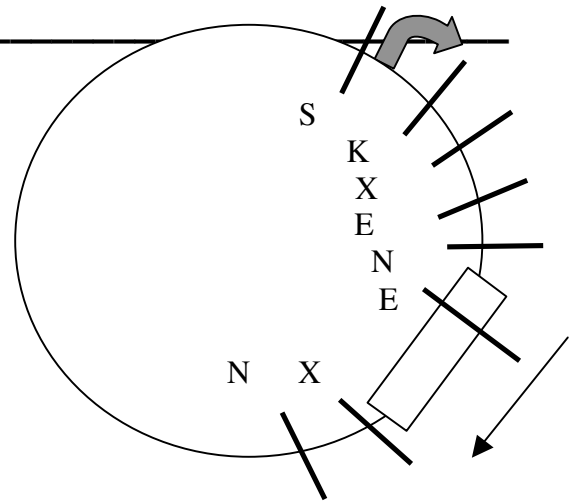
(b) You follow the cloning strategy you chose in part (a). You do the digestion of the insert and the vector and then ligate the two digestions together. You then transform the ligation into bacteria and select for ampicillin resistance. You get three colonies on your transformation plate. You isolate plasmid from each one and cut each plasmid with the enzyme XbaI. You then run your three digestions on an agarose gel and see the following patterns of bands. Describe what each plasmid actually was that was contained in each of the three colonies.



Colony 1's plasmid = **Vector with Yeast Gene in the Right Orientation**

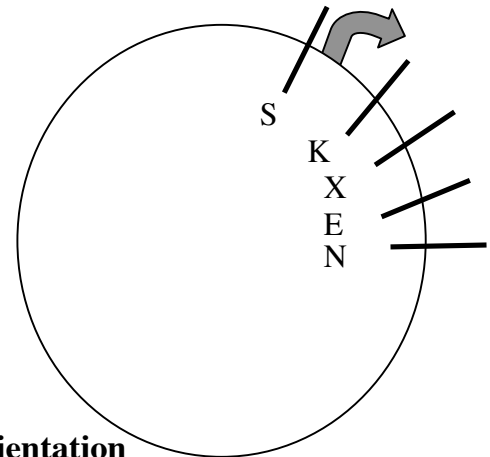
Name: KEY

When the yeast gene is ligated into the plasmid, there are 2 sites where XbaI can cleave (once in the plasmid and once after the end of the open reading frame of the yeast gene). This produces 2 bands, one is ~1250bp and the other is 5000bp.



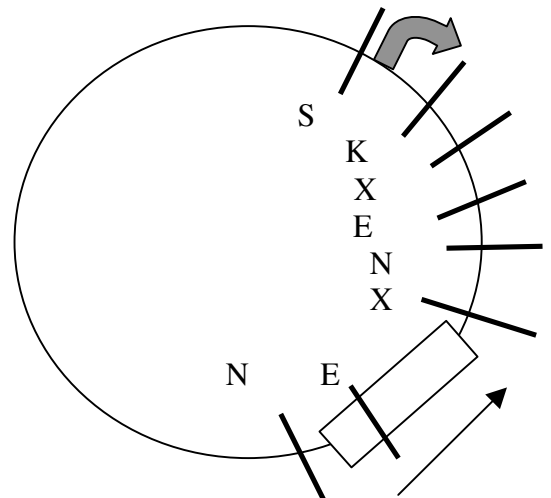
Colony 2's plasmid = Vector Alone (religated to itself)

When there is no yeast gene ligated into the plasmid, there is only one site that XbaI cuts, which produces linearized plasmid that migrates at 5000bp.



Colony 3's plasmid = Vector with Yeast Gene in the Wrong Orientation

When the insert is ligated into the plasmid in the wrong orientation, XbaI again can cut at 2 sites, one in the plasmid and one at the end of the open reading frame of the yeast gene. But now the 2 XbaI sites are closer together and give 2 bands, one that is 5850bp and the other that is ~400bp.



(c) Which colony's plasmid do you actually want to use for your studies?

You want colony #1 because you want the bacterial promoter to drive transcription of the yeast gene in the correct orientation, so that the correct strand of DNA is used as a template in transcription.

(d) Which two restriction enzymes would you use to design a way to get the insert into the vector if you had to use two enzymes simultaneously?

K (KpnI) and X (XbaI)

These are the only 2 enzymes that, after cutting, would exclude the yeast promoter, include the bacterial promoter and allow for the yeast gene to be ligated in the correct orientation following the bacterial promoter. The only two other enzymes that would cut simultaneously to give the open reading frame flanked by two different enzyme sites are X and N, and this would lead to inserting the gene in backwards into the vector.

(e) You transform your ligation planned in part (d) into bacteria and plate the bacteria on Petri plates containing ampicillin. (You actually transform six different ligation mixtures, which are described below, into six different populations of cells, and plate each transformation onto a different plate, because you want to do all of the correct controls.) The next day you come in to lab to look at how many colonies of bacteria are on each plate. You are really excited, because the number of colonies you see on each plate tells you that the entire procedure worked! Which of the three following patterns of number of colonies did you see in order to conclude that you had a successful transformation? Circle the correct pattern.

In this table, DV = digested vector. DYG = digested yeast genome.

<u>Ligation Used</u>	<u>Pattern 1</u>	<u>Pattern 2</u>	<u>Pattern 3</u>
DV only + Ligase	200	0	0
DYG only + Ligase	0	200	0
Water + Ligase	0	0	0
DV + DYG + Ligase	200	200	200
DV + DYG (NO ligase)	0	0	0
Undigested vector + Ligase	0	0	200

Pattern 3

DV only + Ligase → No colonies b/c you have digested with 2 different restriction enzymes that can't ligate together because they don't have complementary sticky ends. Thus the plasmid can't re-circularize to form a functional plasmid (remember, bacteria use circular DNA molecules – they can't replicate linear DNA molecules) and therefore there is no ampicillin resistance gene expressed and therefore the bacteria die in the presence of the antibiotic.

DYG only + Ligase → No colonies because all you transformed is the digested, linear yeast DNA. The bacteria was not given a bacterial plasmid with the ampicillin resistance gene so it will not be able to grow in the presence of the antibiotic.

Water + Ligase → No plasmid with the ampicillin resistance gene (or any DNA) was transformed into the bacteria and so it won't grow in the presence of ampicillin.

DV + DYG + Ligase → Colonies. The plasmid and yeast gene can ligate together to form a functional plasmid that will express the ampicillin resistance gene. So, bacteria that are transformed with this vector will be able to grow in the presence of ampicillin.

DV + DYG (No Ligase) → No colonies because, although you have both digested plasmid and a digested yeast gene with complementary sticky ends, you don't have any ligase around to "glue" them together to form a functional vector (remember, bacteria use circular DNA molecules – they can't replicate linear DNA molecules). So, no ampicillin resistance gene will be expressed and therefore the bacteria won't grow in the presence of antibiotic.

Undigested Vector + Ligase → Colonies. The plasmid is intact because it was never cut with any restriction enzymes. So, regardless of the ligase, you have a functional plasmid that can express the ampicillin resistance gene, allowing the bacteria to grow in the presence of ampicillin.

2. You are practicing designing primers that you can use in PCR reactions. You want your primers to allow you to amplify the sequence found below.

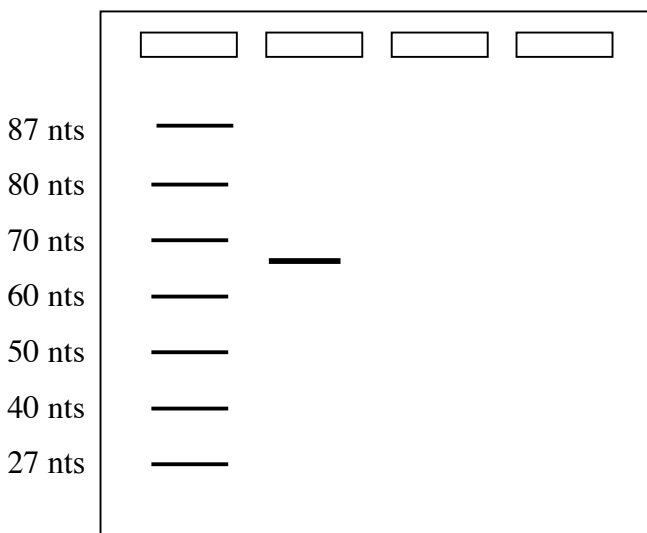
5' - ACTTCGATATGTCTAAAATACGATCGATCTGTGGGGCCTAGCTAGCTAACCAGAGACGCTACCG - 3'
 3' - TGAAGCTATAACAGATTTTATGCTAGCTAGACACCCCGGATCGATCGATTGGTCTCTGCGATGGC - 5'

Left primer should anneal to this region

Right primer should anneal to this region

Draw into the following gel lanes what size(s) of PCR products you would get if you used the following primers stated in parts (a), (b), and (c) to do a PCR reaction on the template DNA shown above.

Markers part (a) part (b) part (c)



← There should only be one band on your gel, in the 2nd lane, at 64 nts long.

Name: KEY

(a) 5'-ACTTCGATATGTCTAAAATAC-3' and 5'-CGGTAGCGTCTCTGGTTAGCT-3'

You will get a PCR product that is 64bps long. The 1st primer will anneal to the lower strand and, starting at the free 3' OH of the primer, DNA polymerase will do replication in the direction left-to-right until the end of the sequence (forming the new strand 5'→3'). The 2nd primer will anneal to the top strand with the 3' end of the primer pointing to the left, and DNA polymerase will replicate the new strand in the direction right-to-left.

(b) 5'-TGAAGCTATACAGATTTTATG-3' and 5'-GCCATCGCAGAGACCAATCGA-3'

No PCR Product. The 1st primer is complementary to the top strand but the sequence and the primer are both 5'→3' and therefore aren't antiparallel so they won't be able to anneal to each other. The 2nd primer is complementary to the bottom strand but is not antiparallel to it, so again the primer will not hybridize to the template, and DNA polymerase will have no primer to replicate from.

(c) 5'-GTATTTTAGACATATCGAAGT-3' and 5'-AGCTAACCAGAGACGCTACCG-3'

No PCR Product. The 1st primer is complementary to the top strand and will allow DNA polymerase to replicate the DNA right to left, starting from the end of the DNA sequence shown and going off the left-hand side of the page. The 2nd primer is complementary to the bottom strand @ the 3' end of the 2nd boxed region and will extend from left to right, starting from the end of the DNA sequence shown and going off the right-hand side of the page. Thus the sequence in between these primers will not be replicated in this reaction; instead, the DNA farther to the left and right than the sequences shown would be replicated.

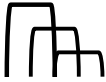
(d) You are asked to design a 15-nucleotide-long primer that could potentially hybridize to a portion of a specific mRNA that encodes the protein sequence N-Met-Ala-Tyr-Trp-Pro-C. How many different primers would you have to design in order to ensure that one of them will in fact hybridize along its full length to the mRNA?

32 different primers

There is: 1 codon for Met
4 codons for Ala
2 codons for Tyr
1 codon for Trp
4 codons for Pro

Therefore: $1 \times 4 \times 2 \times 1 \times 4 = 32$ different combinations of codons, all of which would code for the same protein sequence

3. You are a scientist who is using genomics to currently study a new bacterial species that no one has ever studied before. The following sequence is a piece of DNA within the coding region of a gene that you have recently sequenced.



5' -CCCGTACGTTTACGCCGTATATATCGTCG**TAA**TCCTACG**TAG**CTCTACGAACA-3'
 3' -GGGCATGCA**AAT**GCGGCATATATAGCAGCATTAG**GAT**GCATCGA**GAT**GCTTGT-5'



(a) If you take any bacterial gene sequence, before you begin doing any sequence analysis on it, there are six potential open reading frames. Why are there six?

There are 3 reading frames on each of two strands. If the top strand looked like the mRNA, then the reading frame could start with a CCC, a CCG, or a CGT. If the bottom strand looked like the mRNA, then the reading frame could start with a TGT, a GTT, or a TTC.

(b) How many of the 6 potential open reading frames are actually open in this sequence shown above?

One. There is only 1 reading frame (the middle one on the top strand) that doesn't have a stop codon somewhere in the reading frame. And since this is sequence that's within the ORF, there shouldn't be a stop codon. We have shown all of the stop codons in bold in the sequence above. Each of the five stop codons bolded above occur in a different reading frame.

(c) You are using shotgun sequencing to determine the DNA sequence of the genome of this new bacterial species. For one strand of a 30-nucleotide long stretch of DNA, you get the following sequences out of your shotgun sequencing reaction. Assemble the entire 30-nt-long DNA sequence you are trying to sequence, and write the full sequence of the DNA. You only need to write the one strand that is shown; please make sure to label the 5' and 3' ends of that one strand.

5' -GGAGTTCCTC-3'
 5' -CGCGTTGTCACTGAC-3'
 5' -TGGGAGT-3'
 5' -TCCTCAAACGCGTTGT-3'

5'-TGGGAGTTCCTCAAACGCGTTGTCACTGAC-3'

Find the overlapping sequences in the 5'→3' direction and whatever is not overlapping with any other sequence on one strand @ the very 5' end defines the 5'

Name: KEY

end of the final sequence. Whatever region on one strand is not overlapping with any other sequence at the very 3' end of that fragment defines the 3' end of the final sequence. Doing this exercise is called "reassembling the genome" and is done by computers when the reassembly is of a whole genome sequence that is millions of base pairs long.

You put the DNA sequence that you have assembled in part (c) into a computer program that tells you that the following piece of DNA, which comes from another bacterium, is a close match to the sequence you have sequenced from your bacterium:
5' -...TGGGCATTTCTCAAGCGGGTTGTAATGGAT...-3'

This 30-nt-long sequence fragment lies in the center of a gene, and that portion of the sequence encodes for this 10-amino acid-long part of a protein:
N-...Trp-Ala-Phe-Leu-Lys-Arg-Val-Val-Met-Asp...-C

You hypothesize that the sequence you have discovered is another bacterial species' version of the same gene as this previously known gene. To measure how identical the two genes are at the DNA level and/or the two proteins are at the amino acid level, you can calculate a percentage of "identity" for each. This is the percent of nucleotides (for the gene) or the percent of amino acids (for the protein) that are identical between the two sequences.

(d) What is the % identity between the two DNA sequences?

70% Identity

Our Sequence: 5' - TGGGAGTTCCTCAAACGCGTTGTCACTGAC-3'
Other Sequence: 5' - TGGGCATTTCTCAAGCGGGTTGTAATGGAT-3'
 **** ** ***** ** ***** * **

(21 nucleotides match / 30 total nucleotides) x 100 = 70%

(e) What is the % identity between the two protein sequences?

80% Identity

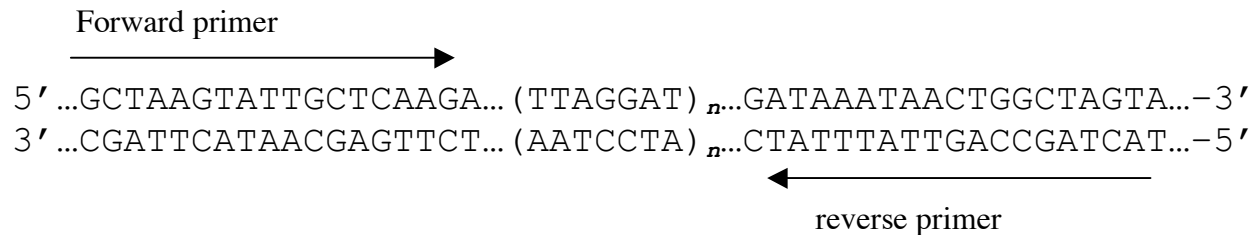
Our Protein: N-Trp-Glu-Phe-Leu-Lys-Arg-Val-Val-Thr-Asp-C
Other Protein: N-Trp-Ala-Phe-Leu-Lys-Arg-Val-Val-Met-Asp-C
 * * * * * * *

(8 amino acid matches / 10 amino acids total) x 100 = 80%

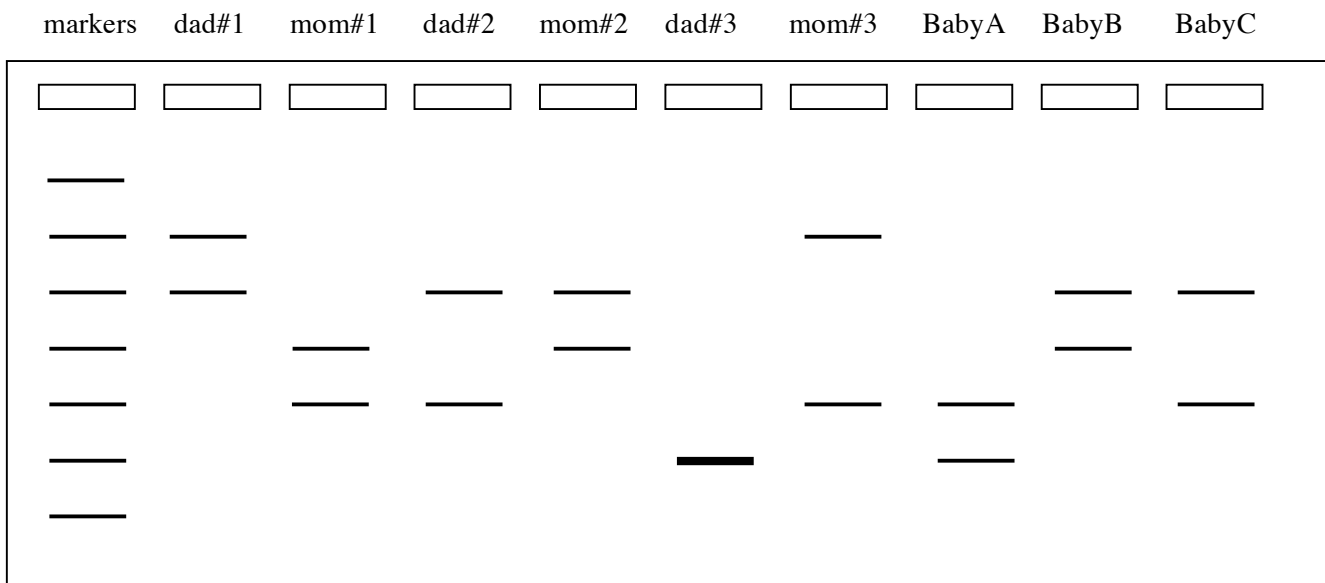
Name: KEY

4. One of the most common applications of using recombinant DNA techniques like PCR, restriction enzyme digests, and agarose gels is to test samples of human cells such as blood cells to identify people for forensic analysis or paternity testing. This problem is designed to show you how this type of analysis, called DNA fingerprinting, can be used to determine paternity. There are three babies (Baby A, Baby B and Baby C) in a maternity ward, and three sets of confused and worried parents. (Father and Mother #1 are a couple, as are Father and Mother #2, and Father and Mother #3.) This problem will show you how to figure out which baby goes with each set of parents.

As we have spoken about in class, most of the human genome (>95%) is not genes. Most of the DNA sequence differences between humans are found in these non-coding regions. Some of these non-coding regions are just series of DNA sequence repeats found over and over again. Different humans can have different numbers of repeats at these regions (i.e. “*n*” shown below can vary). The way you assay how many repeats someone has is by doing a PCR reaction using his/her DNA as a template. The primers are designed to the sequences flanking the repeated region. For instance, take the DNA sequence below. Say it is found somewhere on chromosome #15. Different humans differ by the value of “*n*” (the number of TTAGGAT repeats). You do each PCR reaction and load each one into a separate well of an agarose gel, and then run the gel.



You obtain the following results.



Name: KEY

(a) Why is it that having more repeats leads to a band that is higher in the gel?

The more repeats you have, the longer your DNA fragment is, and longer DNA fragments migrate more slowly in gels. The larger a DNA fragment is, the harder it is for that fragment to move through the pores in the gel, and so it migrates slower than the smaller fragments with fewer repeats.

(b) Why is it that some people show only one band?

They are homozygotes at this region. This means that on both of their chromosome #15's (maternal and paternal), they have the same number of repeats. (For example: Dad #3 has 25 repeats on both versions of his chromosome 15).

(c) Why is it that some people show two bands?

They are heterozygotes at this region. This means that they have a different number of repeats on each version of their chromosome 15. (For example: Mom #3 has 45 repeats on one version of chromosome 15 and 30 repeats on the other).

(d) Given the data so far, for which of the three babies (only Baby A, only Baby B, or only Baby C, all of them, or none of them) can you already conclusively tell who the parents are?

Baby A. Baby A has a band @ $n=25$ and there is only one set of parents that could have given a chromosome 15 with $n=25$ to their offspring, couple #3. This is because Dad #3 is the only parent with $n=25$ @ chromosome 15 (both homologs of chromosome 15 have 25 repeats).

Other parts of the non-coding regions in our genome are not genes but they are also not regions of repeats. Humans can vary by DNA sequence at these sites, instead of varying by number of repeats in a row. For instance, take the DNA sequence below. Say it is found somewhere on chromosome #7. Different humans differ by which basepair is found at the position marked in bold below; some people have a T-A basepair, whereas others have an A-T basepair at this bolded position. It just so happens that one version of this site can be cleaved by a restriction enzyme that recognizes the sequence 5'-TTGCAA-3' and cuts between the two Ts. The way you assay which sequence someone has at this region is by doing a PCR reaction using his/her DNA as a template. The primers are designed to the sequences flanking the site of variable sequence; each primer is about 20 nucleotides long. You then treat the PCR product with the restriction enzyme and run the products of the digestion reaction on an agarose gel.

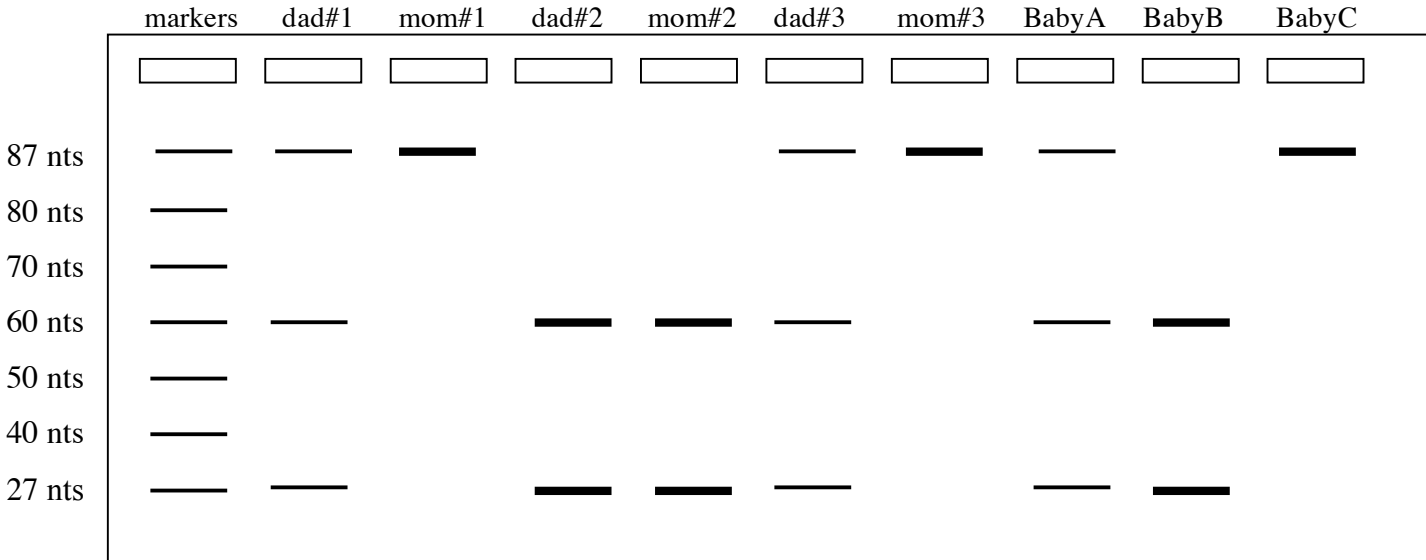
Forward primer

5' -...GATATCTT**G**CAAGTCCATCCTGCATGCACATGCTGATACGCGCAACGGT...-3'

3' -...CTATAG**A**ACGTTTCAGGTAGGACGTACGTGTACGACTATGCGCGTTGCCA...-5'

Reverse primer

You obtain the following results.



(e) Why is it that some people only have one band?

They are homozygotes for the allele that cannot be cut. This means that, on both of their homologs of chromosome 7, they have an A instead of a T on the top strand, so the DNA can't be recognized and cleaved by the restriction enzyme, so there is only 1 band at 87nts.

(f) Why is it that some people show two bands?

They are homozygotes for the allele that can be cut. This mean that, on both of their homologs of chromosome 7, they have T's on the top strand, which are recognized and cleaved by the restriction enzyme, producing 2 bands (at 27 and 60 nts).

(g) Why is it that some people show three bands?

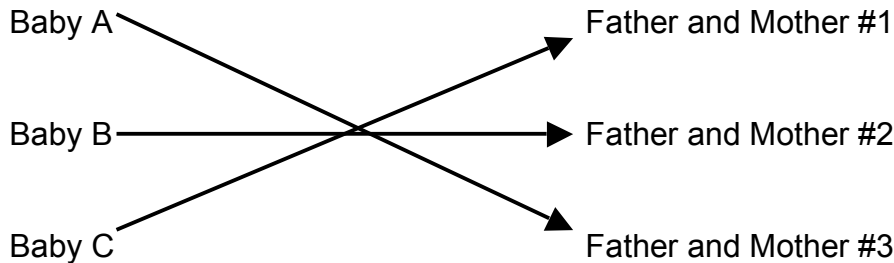
They are heterozygotes, and possess one allele that can be cut and one allele that cannot be cut. This means that, on one homolog of their chromosome 7, they have a T on the top strand that's recognized and cleaved by the restriction enzyme, giving 2 bands at 27 and 60 nts. On their other homolog of chromosome 7, they have an A on the top strand, and the restriction enzyme doesn't recognize or cleave this strand, giving the larger band at 87nts.

Name: KEY

(h) Match the babies up with their actual parents.

BABY

PARENTS



Baby A: We already know that couple #3 had Baby A. See section 3d.

Baby B: Any child of couple #1 must have allele that does not get cut, because the mother has two alleles that do not get cut, so she will give one to every child she has.

Baby B has two alleles, both of which can get cut. Thus Baby B cannot be couple #1's child. (Nor can it be couple #3's child, because couple #3 had BabyA). This only leaves couple #2 as an option.

Baby C: Baby C is from couple #1 by process of elimination.

(i) Now go back and look at the first gel drawn in the problem, and answer the question: which number of repeats did Baby C inherit from his mother?

30 repeats

Since we've determined that dad and mom #1 are Baby C's parents, then it inherits 40 repeats from dad and 30 repeats from mom. The 40-repeat allele is the only one Baby C could have inherited from dad, since he doesn't have a 30-repeat allele to give to any of his children. The 30-repeat allele is the only one Baby C could have inherited from mom, since she doesn't have a 40-repeat allele to give to any of her children.

5. Where in a eukaryotic cell do you think you would find each of the following proteins residing when it is actively performing its function? Be as specific as you can in terms of subcellular location.

(a) DNA polymerase

In the nucleus of cells, associated with DNA if the cell is actively replicating

(b) RNA polymerase

In the nucleus, associated with DNA, as transcription is actively occurring.

(c) the ribosomal proteins

Name: KEY _____

In the cytoplasm where translation takes place, complexed together with other ribosomal proteins and rRNA, and associated with mRNA as translation is actively occurring.

(d) DNA ligase

In the nucleus, ligating lagging strand DNA fragments together during DNA replication

(e) helicase

In the nucleus, associated with DNA (unwinding the helix) during DNA replication.

(f) an activator protein

In the nucleus bound to DNA to aid RNA polymerase in binding to the promoter of specific genes.

(g) a repressor protein

In the nucleus bound to DNA to block RNA polymerase from binding to the promoter of specific genes.

(h) an enzyme in the glycolysis pathway

In the cytoplasm where glycolysis occurs.

(i) an enzyme in the Krebs/TCA cycle

Inside the mitochondria (in the mitochondrial matrix) where the Krebs's Cycle occurs.

(j) a protein that allows ions to pass in and out of the cell

Inside the plasma membrane.

(k) enzymes that splice mRNAs

In the nucleus where splicing occurs, associated with unprocessed pre-mRNAs.

(l) a protein that forms a channel through which mRNAs can be exported into the cytoplasm

Inside the nuclear membrane