Sequencing a Genome:

Inside the Washington University Genome Sequencing Center

Activity Supplement

The Golden Path

(Restriction Fragment Mapping)

Project Outline

The multimedia project *Sequencing a Genome: Inside the Washington University Genome Sequencing Center* is aimed at increasing the scientific literacy of biology students in the technology of genomic sequencing.

The following four video pieces are included on VHS cassette or CD:

- A guided tour of the Washington University Genome Sequencing Center, providing a look at the labs and offices that make up the preparation, production, and data management facilities. Includes animated explanations of the processes used to sequence genomic DNA.
- Exploration of current genomic research in pathogenic bacteria through an interview with a molecular microbiologist.
- Information about careers available at the Genome Sequencing Center presented through interviews with actual employees.
- An animated explanation of the chemistry of cycle sequencing using dideoxynucleotides.

Additional CD features include scripts of the video pieces, links to additional resources, and a glossary of terms.

As the scientific procedures presented in the video tour are complex, simple activities were specifically designed to better explain and reinforce the key concepts of restriction fragment mapping, PCR, sequencing, and electropherogram interpretation. The following paper modeling activity is an inexpensive and simple solution to presenting the "Golden Path."

Acknowledgments
This project was funded in part by a Professorship Award to Dr. Sarah C.R. Elgin from Howard Hughes Medical Institute (HHMI). Additional support was provided by the National Human Genome Research Institute (NHGRI) through its funding of the Washington University Genome Sequencing Center Outreach Program.

Editors
Sarah C.R. Elgin, Washington University Department of Biology
Susan K. Flowers, Washington University Science Outreach

Writers
Sarah C.R. Elgin, Washington University Department of Biology
Gabriella Farkas, Washington University Department of Biology
Anna Cristina Garza, biology major Washington University Arts and Sciences

Paper Model Illustrator
Gabriella Farkas, Washington University Department of Biology

Inquiries about this project may be directed to Susan Flowers, Washington University Science Outreach, Campus Box 1137, One Brookings Drive, St. Louis, Missouri 63130, flowers@wustl.edu, (314) 935-4217

THE GOLDEN PATH
TEACHER'S MANUAL

Lesson Overview

The Golden Path activity is designed to provide your students with a deeper understanding of the strategy involved in sequencing projects such as the Human Genome Project. Through readings, hands-on paper modeling, and completion of a worksheet, they will become more familiar with the mapping and selection process that determines which DNA fragments are sequenced to piece together an entire genome. They will read about the use of vectors, restriction enzymes, and agarose gel electrophoresis in the construction of a restriction map, and then model the interpretation of a gel to construct a restriction map of a section of a genome. They will then use the restriction map they have created to select the clones that contain the chunks of genomic DNA that constitute the Golden Path for the section of the genome under investigation.

Timeline

The background information should be read before the activity, assigned as homework or read as a class. The paper modeling and worksheet require one 50-minute class period to complete.

Materials

For 1-2 students:

1 set of DNA fragments, for clones #1-6
1 kb ruler
1 student manual
1 roll of transparent tape

Advance Preparation

- You will need to print and cut out one complete set of copy masters per 1-2 students. Each set of DNA fragments should be cut into their individual fragment pieces. The paper pieces will be much more durable if they are laminated before they are cut out. Lamination will ensure that you will be able to use the models year after year.
- Black print on white paper is sufficient for this activity; if desired, DNA fragments may be color-coded according to clone number.

Hints and Troubleshooting

- You may find it necessary to demonstrate the paper modeling to the class before allowing the students to work on their own. We suggest making overhead transparencies of the copy masters.
- This activity may be done by individual students or by students working in pairs, depending on their previous understanding of the material and time constraints.
- In Question 5 of Part I, the procedure tells the students to place the second paper fragment to the right of the first paper fragment. As an additional exercise, you may have your students place the second paper fragment to the *left* of the first and then complete the activity, so that they can see for themselves that this does not work. The

activity carried out to completion with the second paper fragment on the left shows pieces of DNA that apparently overlap but do not share recognition sites, which is not possible.

Additional Resources

- Other laboratory activities to bring DNA science into the genetics curriculum are available at http://www.so.wustl.edu/science_outreach/curriculum/genetics.html
- DNA Interactive, http://www.dnai.org/index.htm
- Micklos, David A., ed. DNA Science: A First Course. Second Edition. Cold Spring Harbor: Cold Spring Harbor Laboratory Press, 2003.

Additional Information

Because the information presented in the Student Manual may be difficult for the students to grasp on their own, a more in-depth explanation may prove useful, to enrich general understanding, field questions, etc. The explanation of genome sequencing provided is simplified and thus incomplete. More detailed information is provided below that may be helpful as a supplement to verbal instruction provided to the students.

- A prototype organism for genomic sequencing has been in-bred numerous times to ensure that it has a minimum number of DNA polymorphisms (places where the DNA has multiple forms) present in its genome; thus, it best represents the species as a whole.
- The recognition site of a restriction enzyme is also sometimes referred to as a restriction site.
- The following illustration may be helpful in explaining the construction of a restriction map. Three DNA fragments that have been separated from their vectors are shown, with carrots (^) to indicate the presence of restriction sites:

```
---------------^-----------^-------^---
--^----------^-------^-
---------^-------^-----
```

The restriction sites can be aligned with the following result:

```
---------------^-----------^-------^---
             --^----------^-------^-
                  ---------^-------^-----
```

As you can see, with enough overlapping fragments, a continuous map of the restriction sites can be pieced together.

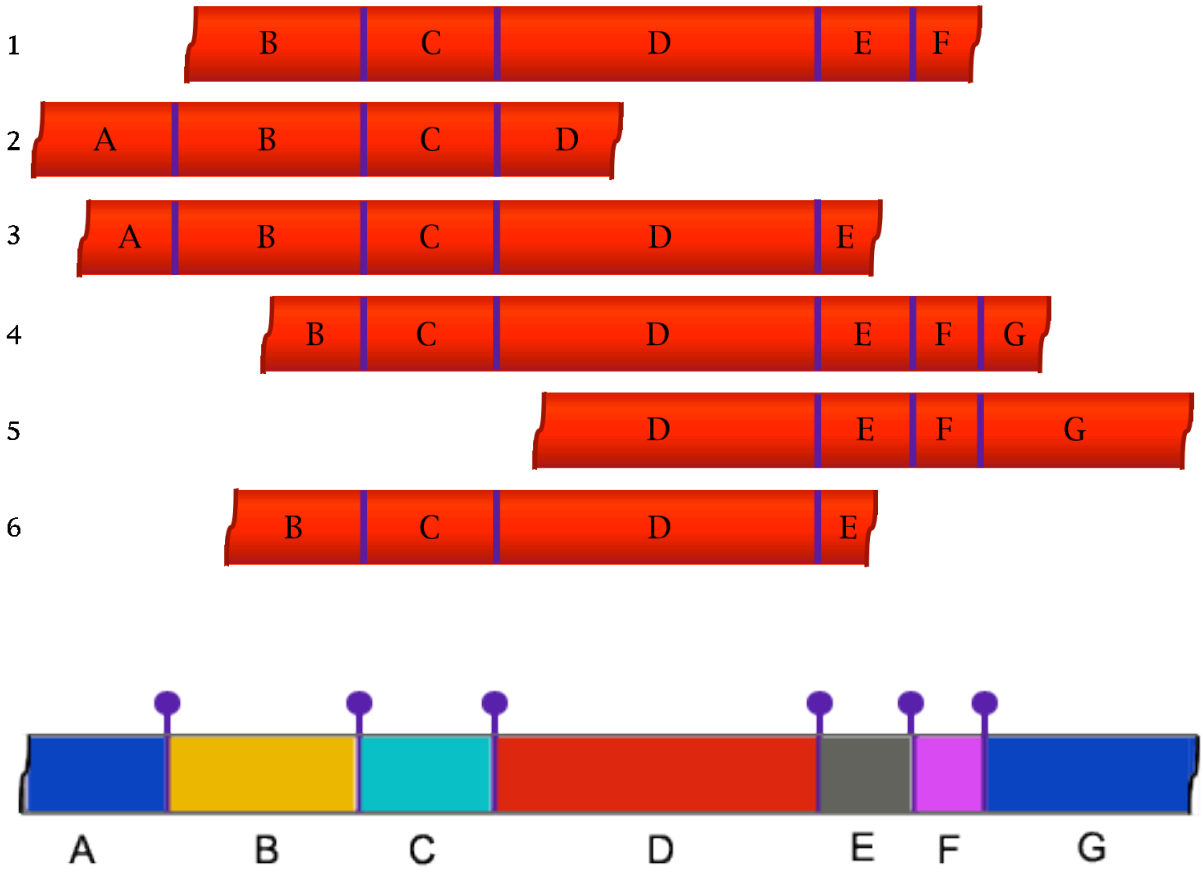
- The gel image provided does not contain the fragment that is the vector itself; though this has been omitted for the sake of simplicity, it would in reality be distinguishable on the gel as a band shared in common by all of the clones. It would not be a part of the map construction but rather would be ignored by the computer performing this process. (The possibility of *all* the clones containing a DNA fragment that is *not* vector but rather a part of the genome of interest is much lower when there are more cell lines and thus a

much longer DNA sequence under consideration, as would be the case in a realistic situation.)

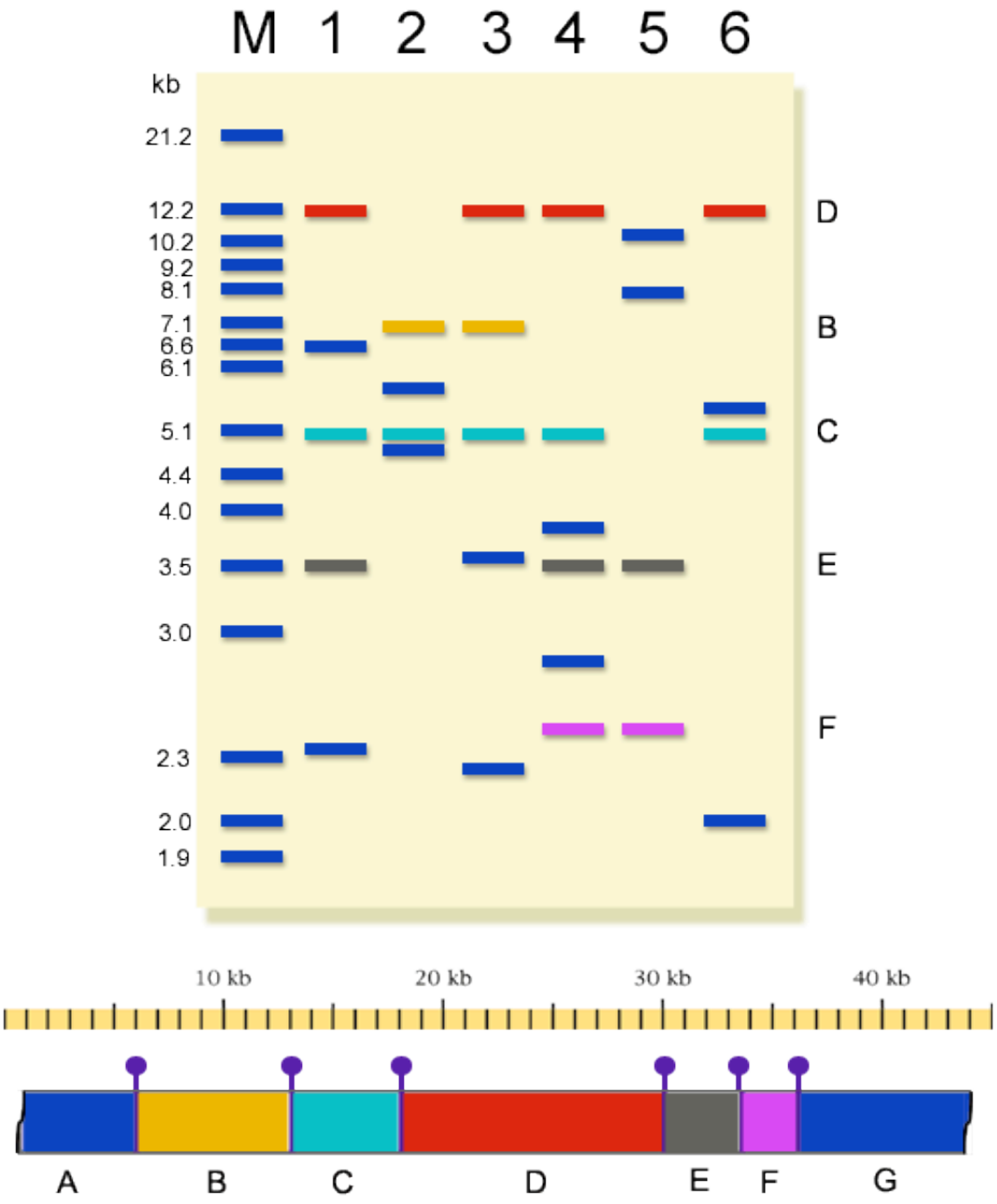
- In reality, when running a gel and interpreting the results, it is possible for two fragments to be of the same length and *not* be the same fragment (same length, different sequence). However, for the sake of simplicity, our activity will assume that each fragment is unique. The section of the genome under investigation is understood to contain little redundancy, or repetition in sequence. This means the possibility of a fragment overlapping with another fragment not identical to it in nucleotide base sequence *as well as* location is extremely low, because the entire sequence is assumed to be unique.
- The bacteria *E. coli* is the standard host organism for vectors in genome sequencing projects; that is, vectors containing genomic DNA are inserted into *E. coli*, and *E. coli* cell lines are cultured. Thus, after sequencing the genome of interest, a computer would check to make sure that the DNA under investigation was not identical with *E. coli* DNA; this would indicate an impurity in the genomic DNA sample. Sequencing centers also have other safeguards to ensure that DNA mix-ups do not take place between sequencing projects.
- The selection process to determine the Golden Path is important because every length of DNA sequenced costs a lot of money. It is important to choose fragments that have sufficient overlap – so that you are confident that the fragments truly constitute the entire continuous genome – but not *too* much overlap, because of the expense involved. Typically, BAC clones are chosen that have 50 kb of overlap (more or less, depending on the genome being sequenced and the amount of money allowed for the project). Because this activity works with a DNA sequence that is less than 50 kb in total length, the "answer" to the Golden Path is not definite; "how much is enough" is a subjective consideration, to a certain extent. It is more important to emphasize the thought process involved in finding the correct answer: the entire sequence under investigation needs to be represented, and overlap should be sufficient but minimal. This might also be a good opportunity to emphasize the fact that "right answers" are not always easy to come by in science; even a 50 kb overlap is subjective, because it has been arbitrarily chosen to be the standard – it allows for statistical confidence without wasting too much money.

Additional Example of Restriction Fragment Mapping

Solution for Fragment Ordering and Restriction Map Worksheet



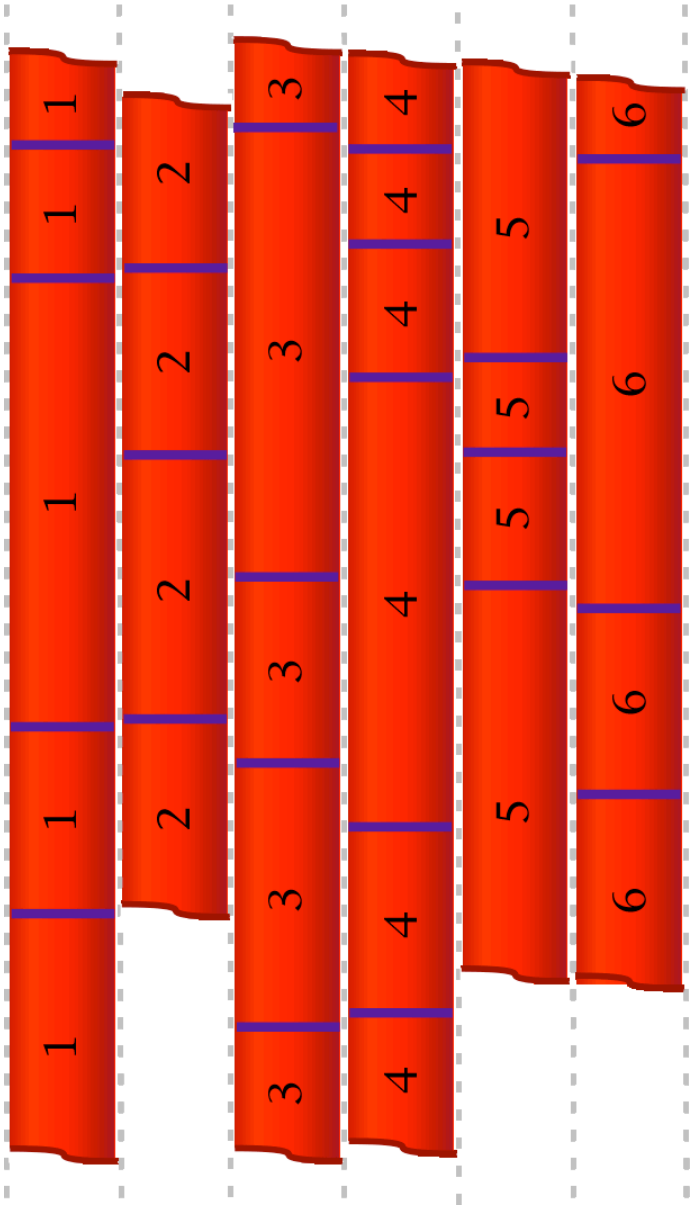
Comparison of Gel Image and Restriction Map

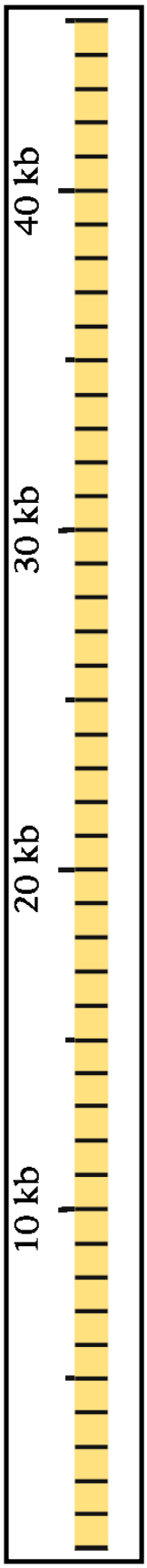


Answers to Student Discussion Questions

1. Which clones would you choose to make up the Golden Path? Give your criteria for choosing this. Are there other possibilities? If so, what are they?
   There is only one possible answer, cell lines 2 and 4. These two clones allow for sufficient overlap without overlapping too much. Because we are not considering the end fragments, the amount of overlap that they allow for is not relevant. The size of the end fragments is also not relevant, because regardless of the extent of the end fragments, they will not be sequenced; they contain a mixture of DNA of interest and vector DNA, as explained in the Student's Manual.
2. What is the importance of choosing a Golden Path? (Why would you not just sequence all of the fragments from all of the clones?)
   A Golden Path is chosen to increase the efficiency of genomic sequencing and ensure confidence in the sequence under investigation while limiting the expense involved.
3. Define restriction enzyme and explain what a restriction enzyme does to DNA.
   A restriction enzyme is a nuclease, an enzyme that cuts the bonds between adjacent nucleotides in a piece of DNA at specific points in a sequence known as recognition sites. Restriction enzymes cleave DNA into many smaller fragments of DNA, according to the locations of recognition sites within the DNA.
4. How can a restriction enzyme help construct a map of a particular section of DNA?
   A restriction enzyme always cuts a piece of DNA at its recognition site, a specific target nucleotide sequence. Different chunks of DNA containing the same recognition sites will be cut at the same locations within their sequences, to produce the same fragments of DNA. These can be lined up and together constitute the entire original piece of DNA, marked with the locations of the recognition sites, thus forming a map.
5. The gel image that you used in this activity was simplified and missing a common band in all six of the clone lanes. Why would all six lanes have a common band?
   The common band would be from vector DNA present in all of the clones.

COPY MASTER
Each group of 1-2 students needs one set of 28 fragments and one ruler.





THE GOLDEN PATH
STUDENT'S MANUAL

Background Information

A **genome** is all of the DNA in a particular organism; **genomic sequencing** refers to the determination of the exact nucleotide base sequence of the entire genome. The **Human Genome Project**, which you have probably heard of, is an undertaking to determine the entire human genome sequence; data from this project will allow us to better understand how our bodies work, and how they sometimes don't. Scientists and doctors will be able to study and treat certain diseases, genetic disorders, and even cancers more effectively, because of data from the Human Genome Project.

As you can imagine, an entire genome is a lot of DNA. The genome must first be broken into pieces – basically, it is divided into random chunks of DNA. These pieces are then each inserted into a **vector**. In molecular biology, the word vector usually refers to a double-stranded circular DNA molecule that has the ability to carry foreign DNA sequences into a host cell. Once in the host cell, the vector can be replicated, or copied, and thus the foreign gene sequences carried in the vector will also be replicated. Essentially, vectors are vehicles used to move specific DNA sequences from one organism to another organism.

When sequencing a genome as large as a human genome, the vector used most often is a **bacterial artificial chromosome**, or a **BAC**. A BAC that carries a fragment of genomic DNA is then inserted into a bacterial cell, so that as the bacterial cell grows and divides, the number of BAC copies increases also. Each **cell line** – that is, each population of cells descending from one original bacterial cell – carries a different BAC, and so it also carries a different chunk of the original genome. Such a cell line is called a **clone**, a group of genetically identical cells derived by asexual division from a common ancestor.

A BAC clone by itself can't tell you much of anything. In fact, the genomic DNA fragment it contains is too big to be sequenced and must be fragmented and sub-cloned before being sequenced. But, we know that all of the clones together contain the entire original genome. Each contains a different DNA fragment, and all of the fragments fit together in a particular order. The question is, how can you figure out this order? What is needed is some sort of map that pieces together the separate genomic fragments into one whole genome sequence. To help create this map, each BAC clone is treated with **restriction enzymes**. Restriction enzymes are endonucleases, special enzymes that cut the bonds between adjacent nucleotides in a double-stranded piece of DNA, resulting in separate double-stranded DNA fragments. When DNA has been treated with a restriction enzyme, it is often referred to as having been **digested**. Restriction enzymes recognize very specific target nucleotide sequences, known as **recognition sites**. When a restriction enzyme recognizes one of these sites, it cleaves the DNA chain at a specific place in the DNA sequence. For example, EcoRI is a restriction enzyme that recognizes the following double-stranded sequence:

----GAATTC----
----CTTAAG----

It cuts the bonds in the sugar-phosphate backbone of the DNA between the G and the A on the top strand, and between the A and the G on the bottom strand, resulting in the following:

```
----G          AATTC----
----CTTAA            G----
```

The ends of these DNA fragments are now referred to as being "sticky" or "overhanging" because at each end, one strand is slightly longer than the other. However, not all restriction enzymes can create "sticky" ends when they cut DNA; other restriction enzymes create "blunt" ends after they have cut the DNA. This means that both strands are exactly the same length after being cut. For example, the restriction enzyme SmaI recognizes the following double-stranded sequence:

```
----GGGCCC----
----CCCGGG----
```

It generates the following blunt ends:

```
----GGG          CCC----
----CCC          GGG----
```

Restriction enzymes are used to help get DNA sequences into vectors. They are also used to separate the vector DNA from the DNA of interest *and* to separate that DNA into smaller pieces.

Once the DNA has been treated with restriction enzymes, the resulting fragments can be separated by **agarose gel electrophoresis**. This is a method that uses an electric current to pull DNA molecules through a gel-like substance called **agarose**. Small DNA molecules move through the gel faster than large ones. Thus, this procedure allows you to see how many pieces were produced when a particular chunk of DNA was cut by a restriction enzyme, and allows you to estimate the relative sizes of the resulting fragments. The size of a DNA fragment refers to the number of nucleotide bases it contains; that is, fragments are distinguishable on the basis of their lengths.

To recap a little, each BAC clone is treated with restriction enzyme. The fragments for a particular BAC clone are then put into a **lane** on the gel. After electrophoresis, each fragment from the BAC clone can be seen on the gel as a distinct **band**. (Fragments of the same size travel the same distance and are found in the same band.) The relative size of a fragment in a particular band can be determined from a gel because, as mentioned above, bigger DNA fragments do not travel as far and are thus closer to the starting end; smaller fragments travel faster and are further towards the bottom of the gel.

The actual size of a fragment, in terms of the number of nucleotide bases in contains, can be determined because a set of **DNA size markers** is always run on the gel, in its own lane and at the same time as any clones. This marker is often referred to as a **ladder** because of its ladder-like appearance; it consists of DNA fragments of known size and can be used as a reference standard. Because two DNA fragments of the same size will travel the same

distance down the gel, a fragment that lines up horizontally with a fragment from the ladder must be the same size.

The set of bands that results in a lane for a particular BAC clone is sometimes referred to as its **fingerprint**; it is a unique pattern that no other clone possesses, because no other clone contains the exact same fragments of DNA. However, different clones can have *overlapping* banding patterns (sharing some fragments), which shows that they contain overlapping segments of the genome. This means that although two cell lines contain two different chunks of the genome, it is possible for their chunks to overlap in certain regions. At these locations of overlap, they have identical DNA sequences; so when a restriction enzyme recognition site is present in one of those chunks, it is necessarily present in the other chunk as well. This means these regions of overlap will fragment in the same way when the two separate cell lines are treated with a restriction enzyme. Thus, when two cell lines possess the same fragment of DNA as part of their bigger chunk of DNA, their fingerprints on a gel have bands that line up horizontally, just like a band can line up with a band from the ladder.

Each DNA fragment from a restriction digest is defined by two recognition sites – one on either end of the fragment. The recognition sites of many different fragments from many different cell lines (each containing a different chunk of the genome) can be aligned and the information from this alignment compiled in such a way that a map of the entire genome (or the section of the genome under investigation) can be created. This is known as a **restriction map**, an illustration of the DNA molecule that shows the positions of all recognition sites (for a particular restriction enzyme) in that molecule. (The use of recognition sites to identify overlapping DNA fragments will become clearer as you work through the activity.) Information regarding the sizes of the DNA fragments of each chunk of genomic DNA is fed into a computer, which then does the work of piecing together the restriction map. It is key to remember that the creation of a map is possible because a restriction enzyme *always* cuts at the same place in a DNA sequence; remember that fragments of the same size possess the same number of bases and, in our case, are identical pieces of DNA.

Once the order of the fragmented genomic DNA is determined, you sequence all those fragments to get the genomic sequence, right? Not exactly. Remember, this is still a lot of DNA and the chunks of DNA that have been pieced together can have significant overlap. Rather than sequence all of them, it is more efficient and less costly to choose the smallest number necessary to cover the whole genome and only sequence those chunks. This collection of genomic sequence fragments is known as the **Golden Path**.

When choosing the Golden Path, one wants a good amount of overlap, in order to be confident of the map created. However, the necessary amount of overlap is somewhat subjective, depending on the organism being studied and the amount of money available. A computer using the restriction map information will decide upon the Golden Path, selecting the cell lines (that contain the appropriate overlapping chunks of DNA) to send forward in the genomic sequencing process. Thus, the entire genome, broken up into random pieces, has been ordered once again and its sequence determined.

Activity Overview

In this activity, you will play the role of a computer in Mapping Core, the facility at the Washington University Genome Sequencing Center responsible for map construction of BAC or fosmid clones (each carrying a part of the genome to be sequenced) and determination of the Golden Path. A prototype organism has been selected for genome sequencing, and a fosmid library has been prepared.

In our activity, we will use **fosmids** as vectors. BACs and fosmids perform the same function and differ mainly in the quantity of DNA they are able to reliably carry; that is, they have different "carrying capacities." A fosmid is a vector system suitable for cloning genomic inserts approximately 40,000 bases, or 40 kilobases (kb) in size. A BAC, on the other hand, is capable of carrying a fragment of DNA 50-250 kb in size. Thus, fosmids are often used when sequencing smaller genomes and, for our purposes, provide a simpler illustration of this complex process; all of the concepts of vectors and BACs previously discussed will still be applicable.

So, the genome under investigation has been broken up into chunks, and individual chunks have been inserted into fosmids, which have in turn been inserted into bacterial cells, and cell lines (clones) of these have been cultured. For this activity, six clones, each representing a section of the genome, are under investigation. (Remember, in map construction the DNA chunks carried by the clones must overlap at least a minimal amount, so that the "path" chosen for sequencing is continuous.) Each of these six clones was treated with the same restriction enzyme, cutting the DNA into pieces wherever there was a recognition site. The DNA chunk of interest was digested into smaller DNA fragments, *and* the DNA chunk was separated from the vector itself.

The fragments resulting from digestion were then subjected to gel electrophoresis, and an image was obtained from the resulting gel. In Part I of this activity, you will be the computer; you will match the sizes of the fragments seen on the gel, and then interpret this information to create a restriction map on the basis of the fragment overlaps. You will receive paper fragments representing the fragments of various sizes created by the restriction enzyme treatment of six different fosmids; they are labeled with their corresponding cell line. (A fragment labeled '1' belongs to the chunk of genome found in cell line #1, and likewise with fragments from the fosmids of the cell lines #2 through #6.) You will use these DNA pieces and a kb ruler (indicating the number of kilobases, kb) to interpret the gel and order the fragments of the given section of the genome onto the fragment ordering worksheet provided. In Part II, you will draw the restriction map constructed from the gel image and use it to decide what the "Golden Path" is for this DNA sequence.

Materials

For 1-2 students:

- 1 set of DNA fragments, for clones #1-6
- 1 kb ruler
- 1 student manual
- 1 roll of transparent tape

Procedure – Part I

1. Locate all of the materials listed above. Place the gel image off to one side, where you can still see and read it clearly. Place the fragment ordering worksheet to the other side, also within easy reach. Lay out all of the fragments in front of you, face up, and separate them according to their cell line number. Each set of fragments represents the genomic DNA chunk carried in that particular fosmid.
2. Now look at the gel image. First look at each individual lane, labeled at the top with the clone number. In each lane, there are two bands not shared with any other lane. (**Note**: Lane M is the ladder, not a clone; if a lane shares a band with Lane M, that tells you nothing except the size of the fragment in that band.) These two pieces must be the ends of the DNA fragment, because all of the other fragments must be present in at least two clones for there to be overlap. Look at the end fragments of each clone and measure them, to confirm that they correspond to the two unique bands. These fragments will only be used to designate the end of the DNA chunk of a clone; they will not aid the map construction otherwise.
3. After looking down each lane, look across the lanes. Which band is the most common in the six clones? (Think about how you can tell a fragment is present in more than one cell line.) Find this size paper fragment in each of the cell lines it is present in, and measure it with the kb ruler, to ensure that it is the same fragment in all of the clones.
4. When you are confident of this, line up these paper fragments in the middle of the fragment ordering worksheet and tape them down, being careful to keep them aligned vertically. It is important to note that these fragments align perfectly because they contain the same recognition sites at both ends; that is, their nucleotide sequences are exactly the same at their ends.
5. Look again at the gel image and find the next most common band. Find this size paper fragment in each of the cell lines in which it is present. Compare the fosmids represented by the first size paper fragment with those represented by the second; note that the second is present in one fewer cell line. Because one fosmid does not contain this fragment, its genomic DNA chunk must end before this next recognition site. Do you understand why?
6. Place this second fragment to the right of the first fragment taped to the fragment ordering worksheet. **Note**: In reality, you would not know whether to place the second fragment on the right or the left of the first fragment; trial and error would reveal to you that it is only possible to place it on the right. For the sake of time, you have been given this additional information.
7. Because you know that the genomic DNA chunk in one fosmid ends, as described in Step 4, you also know that any remaining pieces of the DNA chunk present in this fosmid necessarily extend in the other direction. (You can put the end fragment of this fosmid in place to help you remember not to add fragments to that side anymore.) Find this remaining piece, determine its size using the kb ruler, and use its size to locate it on the gel image as well as to determine which other cell line(s) also contain this fragment.
8. Place this piece onto the fragment ordering worksheet, being careful to align it across cell lines.

9. Just as in Step 4, you know that if the fragments of a cell line previously overlapped with the fragments of another and no longer do, the chunk of genomic DNA present in that cell line must have ended before the next recognition site.
10. At this point, it should be clear which cell lines have only the ends of their DNA fragments remaining to be added. For those that have other remaining pieces, you should be able to use the gel image to determine how those pieces align across genomic DNA chunks, which in turn should help you decide their ordering. Always remember that if a recognition site is not evident from a fragment size, that fragment must have ended before that site. **If two sequences align, they necessarily have the exact same nucleotide sequence in their regions of alignment. Thus, if a sequence overlaps with a recognition site in another sequence, it must also contain that recognition site.**
11. When you have finished, you should now have the DNA from the six clones aligned on the fragment ordering worksheet, such that it is apparent where these six clones are identical in their fragments and where they differ.

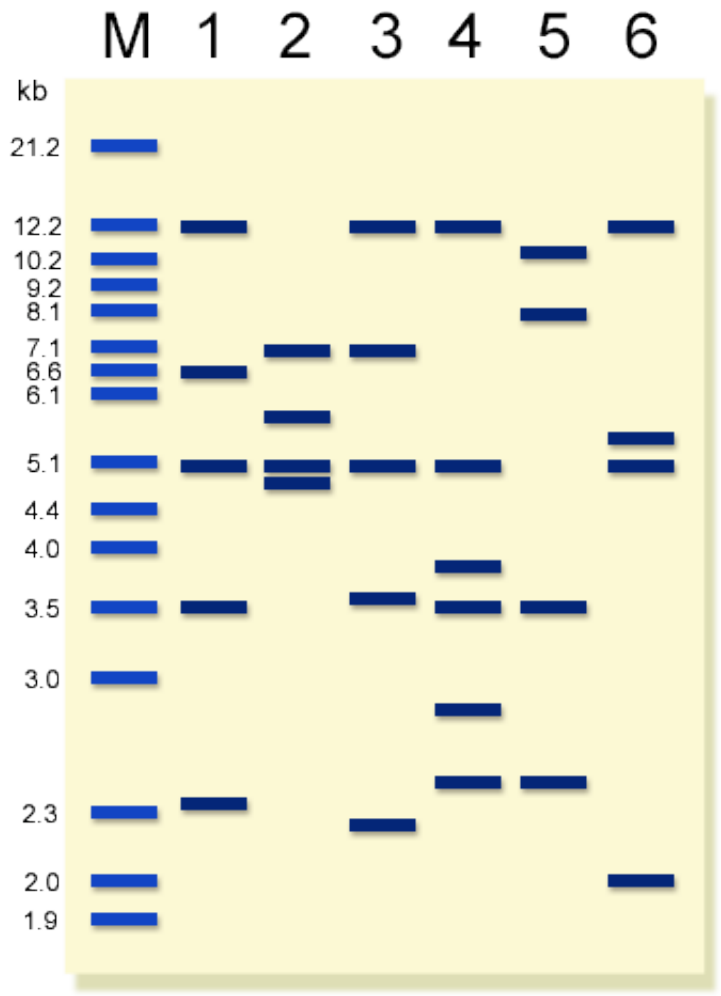
Procedure – Part II

1. Look at your completed fragment ordering worksheet. At the bottom, in the space provided, draw in the restriction map for the total genomic sequence, including all of the recognition sites. (This should show the locations of all of the shared DNA fragments.)
2. To determine the Golden Path of this given section of the genome, choose the minimum number of fosmids that spans the entire section with the least amount of overlap. If you have added the end fragments of each clone, ignore them (or remove them) when deciding the Golden Path. Remember, the DNA in these end fragments is a mixture of DNA from the organism of interest and vector DNA and thus will not be used in genome sequencing.
3. When you are finished answering the discussion questions, remove the fragments from the fragment ordering worksheet and remove any tape from them. Return the kb ruler and the set of DNA fragments in their envelope to your teacher.

GEL IMAGE

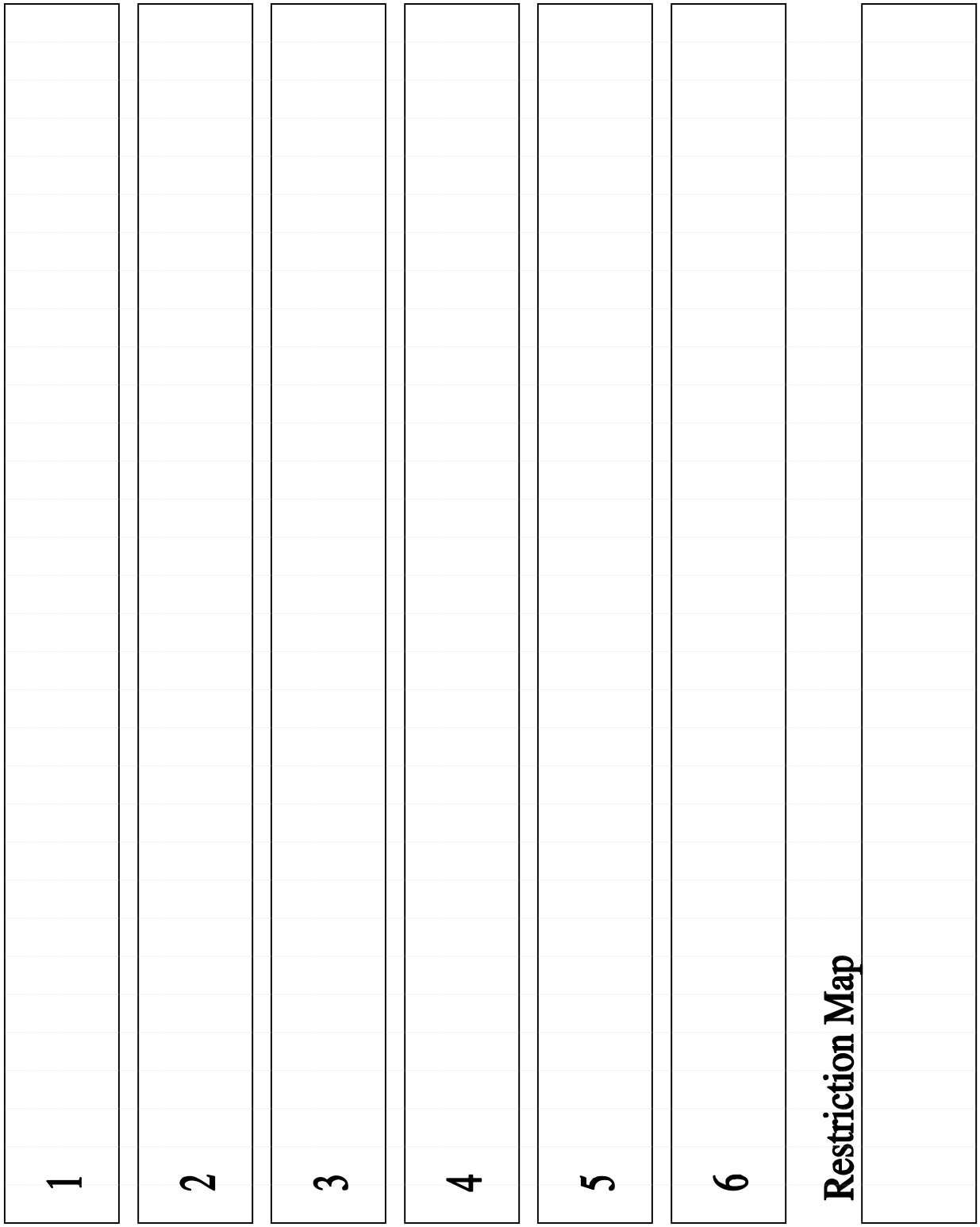
Lane M contains DNA size markers. The bands in this lane are DNA fragments of known sizes. The bands in this lane can be used to estimate the sizes of the fragments in lanes 1 through 6.

Lanes 1 through 6 contain DNA fragments from separate clones. Each clone carries a fosmid vector containing a different section of genomic DNA from the prototype organism; the six clones selected have overlapping DNA. The fosmid DNA was isolated from the cells and then treated with a single restriction enzyme.



Name(s) _____ Class Hour _____ Date _____

Fragment Ordering and Restriction Map Worksheet



Name(s) _____ Class Hour _____ Date _____

Discussion Questions

1. Which clones would you choose to make up the Golden Path? Give your criteria for choosing this. Are there other possibilities? If so, what are they?

2. What is the importance of choosing a Golden Path? (Why would you not just sequence all of the fosmids from all of the cell lines?)

3. Define "restriction enzyme" and explain what a restriction enzyme does to DNA.

4. How can a restriction enzyme help construct a map of a particular section of DNA?

5. The gel image that you used in this activity was simplified and actually missing a common band in all six of the clone lanes. Why would all six lanes have a common band? (That is, what is the source of the DNA in this band?)