# Phage Genome Analysis Background

## Sequencing strategy

You are going to use a shotgun sequencing strategy that involves the following steps:

1) **Shearing**. Mechanically breaking the DNA into relatively small segments. You don't really know how big each genome is to start with, but it's a good bet that it is somewhere between 50,000 and 100,000bp. You'll break it into pieces that are about 2000 bp long. The method for doing this is hydrodynamic shearing in which a solution of DNA is passed though a very small hole. One reason for choosing this method is that the breaks in the DNA are at essentially random positions, such that each of the resulting pieces is of a different segment of the genome. The shearing is done with a computer-controlled system that regulates the degree of shearing. Note that there is always a range in the sizes of DNA segments that are generated, but we expect the bulk of the DNA to be within the range of 1000- 3000 bp.

2) **Repair**. When the DNA is sheared, it doesn't always shear such that both DNA strands are broken between the same base pairs. If you imagine a broken ladder, the sides are not always broken between the same two rungs. For your subsequent steps, it is important that the DNA ends don't have single-stranded extensions, since this will make it difficult to clone. You will therefore treat the DNA with enzymes that will convert the ragged ends into blunt ends.

3) **Size-fractionation**. Since there is a range of DNA sizes, and you would rather have a more narrow range of sizes, you will separate the DNA fragments on an agarose gel, which the DNA fragments move from one end to the other with smaller fragments moving faster than bigger ones. This spreads out the DNA according to size, and you can literally cut out a block of agarose that contains the desired size fraction of choice and purify the DNA from it.

4) **Cloning**. From the steps above, you now have a solution of DNA that contains DNA fragments that are all of a similar size (about 2,000bp) but are all different. You may have a microgram or so of DNA, but this represents a large number of individual molecules (about a billion). What you need to do is join together individual molecules with another DNA molecule (a vector) that will carry the DNA around. Basically, the vector has all the components necessary to ensure that the DNA is maintained and replicated in the bacteria that we going to use to prepare large amounts of that particular DNA molecule. To make these joint (recombinant) molecules, you will mix together your phage DNA fragments with a prepared form of the vector DNA and join

them together using an enzyme called DNA Ligase that sews the DNAs together. Here's a way to think about the vector DNA. It starts out as a circle and is fairly small (just a few thousand bps). This is the way that it replicates inside of cells, and the way that you isolated it. To prepare it for your experiments you will cut it with a restriction enzyme that acts like a pair of scissors and will cut the circular DNA just once. The result is still one piece of DNA, but now it is long and linear (no longer a circle), and it has two ends. Because the enzyme that you use to cut it, cuts it at a very specific location, all the DNA molecules are identical.

5) **Transformation**. After the ligation reaction, you will have a solution that contains many joined molecules. Some of these may be vector molecules joined to other vector molecules and some may be phage DNA fragments joined to other phage DNA fragments, but there will also be molecules that have one end of a vector molecule joined to one end of a phage DNA fragment and the other end joined to the other end of the same phage DNA fragment so that you again have a circle, albeit a rather larger one than with just vector alone. These are the ones you want. To recover these you need to get them into a bacterium where they will replicate and grow. You will add the DNA mixture to a sample of prepared *E. coli* cells and give them a large electric shock. In this process (called electroporation), the cells are encouraged to take up DNA molecules. However, this is overall a rather inefficient process, and only a few of the cells take up DNA. You need to have a way to differentiate between cells that have taken up DNA and those that haven't. The trick is to use a vector DNA that contains a gene that confers resistance to an antibiotic such as penicillin, and *E. coli* cells that are penicillin-sensitive. After the electroporation you will plate out the cell mixture onto agar plates that contain penicillin. The cells that have not taken up any DNA will be killed, since they are sensitive to the drug. However, cells that have taken up either vector DNA, or one of the recombinant molecules will grow into colonies, since they are now drug resistant. These colonies contain either vector DNA molecules (i.e. not containing any phage DNA insert) or a joint or recombinant molecule. A key point to recognize is that each colony recovered from this transformation is derived from the uptake of a single DNA molecule. This plasmid will then replicate within the bacteria until there are hundreds of copies of the molecule in each cell, and you can grow many, many cells easily (up to a billion cells per milliliter). But at this point, all of these will be the same (and replicas of the original single DNA molecule that was taken up). One more important point concerns how you can distinguish between those cells that have taken up a vector DNA molecule and those that took up a recombinant molecule. This is simple. The vector DNA has been designed so that the site where we cut the DNA is inside a gene that can turn the cells dark blue or black. Thus cells that have taken up vector DNA will be blue or black, whereas those that took up recombinant molecules will be white, since the addition of the insert DNA destroys the integrity of the color-forming gene. The white colonies are therefore the ones that you are after. If your procedure has worked well, about half or more of your colonies will be white.

6) **DNA preparation**. Next, you need to prepare DNA from several hundred of these white colonies. You could grow up as much as you need. You know that if you grow about 1ml of cell culture until it is dense that you should be able to isolate many micrograms of DNA, which is enough for several sequencing reactions. You will grow up 'blocks' of clones, in which there are 96 wells in each block. Each well has about 1ml of broth and you will inoculate each well with a white colony. These cultures are then

grown overnight until they are saturated cultures. The BioRobot is then used to break open the cells and isolate the DNA. Normally, you will prepare anywhere from 5-10 blocks for each phage genome.

7) **DNA sequencing**. Remember that each of your DNA clones contains the same piece of vector DNA but a different segment of phage DNA. The sequencing reactions can be thought of as sequential determination of the order of insert base pairs, and you can therefore perform this from both 'sides' of the vector DNA. The reactions work by annealing a short DNA primer (typically about 17 bases long) to the vector DNA in one particular position. The sequencing reactions involve extending this primer by use of a DNA polymerase, which is the type of enzyme that makes DNA. It makes the DNA chains longer by adding one base at a time, and the base (either A, T, C, or G) is chosen as one that is complementary to the base of the template strand that is being copied. If we didn't change anything, the enzyme would just copy the template to give long newly-made DNA strands. However, these reactions coontain small amounts of each nucleotide (A, C, G, or T) altered in two ways. First, the base is altered so that it carries a chemical group that fluoresces at a particular wavelength; each base has a different dye – think of it as blue, black, green and red dyes. The second change is that these nucleotides have had their 3'OH removed, so that they stop any further lengthening of the chain – they are chain terminators. Therefore, the reaction makes a group of DNA chains of different lengths, and their precise lengths will differ by single base pairs. The base at the end of the chain will be determined by the color of the base that was added last. Since this is the only fluorescent dye in the entire chain, this chain will fluoresce that particular color. Therefore, if after you separate the chains by virtue of their length, then you see a series of different colored products, and the order of the different colors defines the order of the base pairs. To actually do this, you will add all the reagents together and run the reaction in a thermocycler. The products are then cleaned up (filtered to remove the unincorporated nucleotide dyes) and loaded into the sequencing machine. The machine takes 48 tiny reactions at a time and separates the chains in each reaction by running them through a matrix within a fine capillary. As they separate, a laser excites the dyes, and a detector recognizes which dye is fluorescing.

8) **Assembly**. You hope to obtain about 900bp or so from each primer run. Each clone can be sequenced from both ends and you will use two different primers for this, which we refer to as 'forward' and 'reverse' primers. Every clone thus gets sequenced twice, so that 5 blocks will be about 1000 sequencing reactions in total. Remember also that each of these clones has a random segment of phage DNA in it. The trick now is to take all of these pieces and put them together like a jigsaw puzzle, since many clones will represent overlapping segments. It's important to note that the sequencing reactions only give the sequence of one strand and it can be either strand. You therefore anticipate that you will sequence each base pair many times (an average of perhaps 7 or 8 times) and hopefully at least once on each strand. Note also that the sequences of both strands within a single segment of DNA are different, but they are related by the pairing rules: i.e. G pairs with C, and A pairs with T. We have computer programs that perform these assembly functions.

9) **Clean-up**. The sequencing reactions and protocols are not perfect, and the assembly often generates regions of some ambiguity. Also, there may be small areas that were not sequenced at all. To resolve these, you will make a small collection of primers that

are specific to a particular segment of phage sequence. These are then used with phage DNA as a template to sequence the weak areas.

**10) Analysis**. There are a variety of different methods of analysis once the sequence is complete. The first goal is often to identify the genes.

This step involves several pieces of data, and there are a few key facts about the way DNA is "read" that you need to know.

**THERE ARE 6 TRANSLATIONAL FRAMES TO BE CONSIDERED**

First, ***double stranded DNA***, has two antiparallel strands each with opposing 5′ and 3′ ends.

For our example

5′.....ATCGGTCAGGCTT.....3′
3′….TAGCCAGTCCGAA.....5′.

Second, ***transcription*** always occurs with the new chain growing from the 5′ to 3′ direction. (Think of the DNA strands as one-way tracks and the RNA polymerase as a vehicle that can only latch onto the track with its headlights in one direction and its tail in the other, and it can only go forward so there are two "directions".) Remember, that the RNA polymerase is making a message, a mRNA copy of the DNA that will be read ("translated') by the ribosome into the protein that is the "gene product".

Third, ***translation*** occurs when the ribosome reads the message RNA (the words or "codons" are three bases long, each three specifying one amino acid, which is then added to the next amino acid specified by the next three bases,). A protein is a chain of amino acids whose identity was determined by the order of bases of the gene's DNA. Now if you think of a chain of bases, for our example

5′……AUCGGUCAGGCUU……3′, it could be broken into triplets of
AUC-GGU-CAG-GCU-U.. or
UCG-GUC-AGG-CUU-… or
CGG-UCA-GGC-UU..-…

The protein depends on where the ribosome starts "reading". These are called the "frames" of translation. For each strand of the DNA, there are three frames.

Remember there is another strand that would be 3′…..UAGCCAGUCCGAA……5′. It is almost always written 5′ to 3′ so it would be 5′…..AAGCCUGACCGAU…….3′

And the three frames would be
AAG-CCU-GAC-CGA-U….
AGC-CTUG-ACC-GAU…
GCC-UGA-CCG-A

Together these six frames describe six different sequences of amino acids, so each piece of DNA potentially has the information for six amino acid sequences. The way nature works in phages is that only one of these six frames is actually part of a gene. (In other words,

genes don't really significantly overlap. One part of DNA is generally part of only one gene.)

Interesting exceptions to some of this are known, but can be saved for another discussion!!

Some of the codons say "START" and some of the codons say "STOP" to the ribosome, and that determines where the message reading begins and ends. In other words, what frame is used to start the translation depends on where the "START" signal is (preceded by another signal called a Ribosome Binding Site (RBS) or Shine-Delgarno sequence that tells the ribosome to latch on), and then that frame is translated into a protein that ends when the ribosome encounters a "STOP".

The beauty of it is that some really smart hard working phage scientists in the last forty years worked out the details of this code!!!!! You just open a book (or a computer) and the DNA sequence can be translated immediately into all six frames. Your job is to then look at the DNA where the computer tells you there are "START" and "STOP" and figure out if the stuff between these signals is a gene.

You have several tools to help you figure this out.

The first fact is that phages like to conserve fuel on their superhighway. So, there is not a lot of "junk" DNA that is not part of a gene. In fact, there is darn little such DNA (unlike your own cells, but that is another topic). So, the fact that most of the DNA is "coding" and the earlier fact that genes don't really significantly overlap tells you that the genes should be generally much right next to each other along the whole phage genome.

Second, once you know the DNA sequence, the computer can tell you instantly what sequence of amino acids the DNA in between would tell the ribosome (via the message RNA of course) to string into a protein. Maybe you think, "Big deal what does

*Methionine-alanine-guanine-cytosine-alanine-lysine-etc.etc.etc.* mean to me???

You can take that sequence and "BLAST" it. This means to enter it as a search query into Genbank. What the BLAST program does is to take this sequence and compare it to all possible sequences that have been identified in the billions of bases of DNA sequences that have entered into the public database. These sequences are from all sorts of organisms, from phages, bacteria, plants, animals, etc. Anytime you have a significant match to something that is in the database, it is a good bet that, indeed, you have the right frame (this START……..STOP region is a bona fide gene, encoding a protein product with similarity to something in the database). You can record it and move on to the next one!

But, if you remember that one reason why we actually do all this is because phages have NEW and DIFFERENT gene products, that means that a lot of phage genes, when BLASTED, come up with "No Database Matches". How do you identify these genes? A lot of it is based on how we know the phages typically pack their genes rightly, so you call genes that make sense in this pattern. More importantly the "coding potential" of the gene sequence you want to call is evaluated.

Coding potential comes from some more of the basic facts from molecular biology. Remember that there are 4 bases. This means there are 64 possible codons for the

ribosome to read. The ribosome, however, has only twenty amino acids (words) in its language to put into the growing protein (plus three that say STOP). What this means is that some of the 61 codons that code for amino acids have to mean the same word (the code is degenerate: more than one codon specifies certain amino acids.) Different organisms have different preferences for the triplets that they use for certain amino acids, and by studying enough of any one organisms's genes and gene products, these preferences can be determined. Then, a stretch of DNA can be scanned for its "coding potential," (are the preferred codons used? Then it has high "coding potential"! Is the stretch full of rarely used codons? Then it has low coding potential.) Again, thanks be to those who worked all this out! There are programs that can graphically output the coding potential of all six frames of the whole phage genome (along with marking the STARTs and STOPs.

You can take all of these tools, (six-phase-translation, coding potential output, BLAST results, general knowledge of how phages pack their genomes with genes) and call all the genes.

The final and highest level of the analysis, which really never ends, is comparing these genes to other genes and the overall organization of the phage to other phages and to bacterial sequences and to try to understand how Nature is assembling these jewels. These are the product of natural selection acting on a whole slew of phages in the world. You can find the results of successful mixes and matches. The unsuccessful ones never propagate to wind up in your dirt samples!