**Sanger-Sequencing of PCR Products Explained**

See “**Science Aid Center**” on Core website for additional information as indicated below.

Determining the best approach to sequencing DNA templates requires an understanding of several issues in the context of the goal(s) of your sequencing efforts. However, please note that the issues covered below are interrelated and sometimes operate in opposite ‘directions’; thus, you need to decide what aspects are of primary importance to your specific needs. Finally, for a comprehensive “Sanger-Sequencing Guide”, please consult the Science Aid Center on the LSU Genomics Core website.

# Most importantly, why are you sequencing your templates?

This really is the most critical question of all... so, again, **why are** you bothering to sequence your DNA in the first place? Here are a few examples of potential ‘goals’ for Sanger-Sequencing: determine basic identity of template (e.g., what gene?); identify rare SNPs or high-frequency SNPs; verify product inserted into a vector; and, determine precise junction of a DNA construct. Also, sometimes you need sequence data for just part of the template vs. the entire sequence. It should be obvious how different goals might require different approaches (at least, it should be once you read the rest of this document); however, the goal can also dramatically affect how good your sequencing results need to be. Of course, you should always aim for high quality results; however, the dirty reality is that they won’t always be pretty! As such, you also need to recognize when the actual quality achieved is ‘good enough’ for your specific goal... or you will waste a lot of valuable money and time repeating your sequencing efforts.

# Can the PCR product be directly sequenced (vs. cloned first)?

Most of the time, PCR products can be sequenced directly and this will typically be much less expensive and faster than cloning and sequencing. However, sometimes cloning first is either a requirement or at least beneficial. For instance, when your PCR cannot be optimized to generate a single product (even by gel extraction), direct sequencing will generate messy or non-usable data. Also, when templates contain ‘difficult-to-sequence’ motifs (e.g., homopolymers, microsatellites, high GC content), sequencing with cloned templates might help push through such trouble-spots.

# Pros-&-Cons of sequencing PCR products directly vs. by cloning?

Due to somatic mutations, PCR products from genomic DNA preps will usually be a mix of slightly different sequences (i.e., various point mutations). Typically, each individual SNP will be represented at a very low level compared to the ‘consensus’ nucleotide at that position. In that scenario, if you directly sequence the PCR product, the resulting read will be the overall consensus... not the SNPs. By contrast, if you clone the PCR product and sequence a single clone, the data will fix that fragment’s specific SNPs (if any exist)... not the consensus sequence. Depending on the goal of your sequencing, this difference may or may not matter. For instance, if the goal is a consensus sequence, then direct sequencing would be better as you would need to sequence at least 8-10 clones from each PCR product in order to have >95% confidence that you can derive the consensus sequence. By contrast, if you are interested in ‘rare’ SNPs, then cloning would be required as rare SNPs would be drowned out by the consensus nucleotides from direct sequencing. At the same time, if you are interested in relatively high frequency SNPs, these may be detected more easily by direct sequencing because they would appear as multiple peaks (usually dual) in one position with intensities that are roughly in proportion to the allele frequency in the sample.

In terms of costs, cloning and sequencing PCR products is usually much more expensive and time consuming than sequencing the PCR products directly... even if sequencing just one clone per PCR product. If heterozygous sequences are of interest, costs increase by another ~10-fold as at least 8-10 clones (of each PCR product) must be sequenced to essentially guarantee that heterozygous samples can be detected. On the other hand, sequencing pure clones will produce extremely clean basecalls and may improve results with ‘difficult-to-sequence’ templates; thus, under some circumstances, cloning first might be essential or even be less expensive in the long run than direct sequencing.

How essential is it to remove the original PCR primers?

Success with DNA Sequencing requires that there be only 1 primer per reaction; the presence of more than 1 primer per reaction will result in data that are messy (at best) or non-usable (most likely). This is true whether the additional primers were deliberately added or were residual from a prior PCR. In fact, even if the concentration of the desired primer is ~100-fold higher (in the sequencing reaction), residual opposite PCR primer will result in extremely messy basecalling. Thus, when directly sequencing PCR products, you must ensure that the original primers have essentially been eliminated. There are multiple approaches to accomplish this, as detailed in the Science Aid Center. By contrast, if cloning the PCR products first, residual primers from the original PCR will not be an issue.

# Do you need Sequence within ~30-bp of the 3’-ends of primers?

Typically, a 50-cm capillary array with POP7 polymer can generate >800-bp of high quality sequence... with decent quality of >850-950 bp (low quality sequence might >1000 bp). By contrast, under the absolute best of circumstances, you cannot obtain reliable sequence data any closer than ~5-10 bp out from the 3’-end of your sequencing primer. This is due in part to the difficulty of recovering such small DNA fragments and in part to the inability of the polymer (used in capillary electrophoresis) to resolve such short fragments at the 1-nt level; typically, the ‘lost sequence’ is on the order of 10-30+ nt. Thus, if you sequence in only one direction or if the PCR product is longer than ~850-900 bp, you can expect to lose some data near the primers. By contrast, if the PCR product is under ~800-850 bp and sequencing involved both forward and reverse reactions, the final contig will typically have the full sequence. Alternatively, if the PCR product is only ~700-750 bp and you clone it first, sequencing with the vector primers (typically located ~70-120 nt from the MCS) will usually give you the complete insert sequence.

# Do you need sequence data for the entire length of the PCR template?

Typically, most PCR products are under ~1500-bp; thus, by using both forward and reverse primers (in separate sequencing reactions), you can combine the sequence data to create a contig with more than ~100-bp of high-quality overlap in the center. However, if you do not need the entire sequence (e.g., you are interested in only the first ~800-bp from one primer), it might be possible to sequence in only one direction. For instance, you might simply want to verify that the PCR product is not a non-specific product of the same length as your intended target; in that case, up to 800-bp might be sufficient.

# Is the PCR product (DNA template) longer than ~1,500-bp?

If you don’t need the sequence from the ‘center’ of the PCR product, even just the forward and reverse primers might provide sufficient data from such long templates. However, if you need high quality data in the overlapping portions of each fragment of the final contig, you will need to ‘primer walk’. In most cases, the best results will be generated by four sequencing reactions per template in which the non-overlapping sections of each read are <700-bp: forward primer; reverse primer; and, a pair of internal primers oriented towards each other with each internal primer no more than ~350-bp from the center of the template. Of course, longer templates will require additional internal primers.

# Does the DNA template contain “difficult-to-sequence” motifs?

Some of the more problematic motifs include microsatellites, high GC-content, and homopolymers. For direct sequencing of PCR products, homopolymers are particularly troublesome when they are either long (>7 nt) or occur in series; this is due to strand-slippage during the original PCR which effectively creates a set of non-synchronous templates following the homopolymer (i.e., the PCR generates several different lengths of homopolymer). For single homopolymers within a template, readable sequence will typically exist following the homopolymer if it is shorter than 8 nucleotides; if the homopolymer is 8-10 nts long, chances of readable sequence following the homopolymer are considerably reduced; and, if the homopolymer is >10 nts long, the data following the homopolymer will be ‘trash’ when direct sequencing the PCR product. When multiple homopolymers in series are involved, even shorter homopolymers can cause similar problems after the second or third homopolymer. However, if the exact length of a single long homopolymer is not of great consequence (and if the template is <800 bp on each side of the homopolymer), a usable contig can still be generated from direct sequencing of the template in both the forward and reverse directions. Alternatively, because cloning essentially separates each template produced during PCR, all products from a cloned reaction will have the same length homopolymer and it is possible to sequence through as many as ~35-45 nt when sequencing cloned templates. Cloning might even be essential with templates having a series of homopolymers, although sequencing of multiple clones will be required to determine the true homopolymer length(s).

Microsatellites and high GC-content regions (or even AT-content, to a lesser degree) also have adverse effects on DNA sequencing. However, rather than abruptly transitioning into ‘trash’ following the motif (as with longer homopolymers), these motifs tend to dramatically reduce signal intensity of the read following the motif. Sometimes, better results can be generated by cloning the templates first... although this is not as likely as when dealing with homopolymers. Typically, the best approach to such templates is to follow recommendations for ‘difficult-to-sequence’ templates in the Science Aid Center.

# For cloned DNA, are original PCR primers or vector primers better?

Generally speaking, primers used for PCR will also work in sequencing reactions. However, because PCR is an ‘exponential’ process, a low-function primer might generate PCR products, but be unable to generate sufficient signal in a sequencing reaction (which is ‘arithmetic’). Otherwise, the choice of original vs. vector primer depends primarily on your ‘insert’ length and the amount of sequence data required. First, recognize that vector primers are usually located about 70-120 bp away from the MCS, whereas your original PCR primers are located at the ends of your insert. Thus, if you use an original PCR primer, the resulting sequence will be missing some early nts (typically, up to 10-30 bp) after the primer; on the other hand, that same sequence will extend farther into the insert than would the associated vector primer. By contrast, a vector primer will produce sequence from the vector into the insert, thereby capturing all nts after the original PCR primer... but, it won’t reach as far across the insert. The extent to which a primer will reach into the insert becomes critical when the insert template length approaches ~1500-bp, as the quality of basecalling typically begins to drop off after ~800-bp; thus, long templates might not have enough high-quality overlap (to generate a contig) from using the vector primers. Finally, placement of the primer relative to a difficult-to-sequence motif can be critical; results can sometimes be improved by using a primer that is farther back from the motif (if the other primer is <10-50 bp from the motif) or closer to the motif (if the other primer is far from the motif, e.g., >600-700 bp).

# How much DNA template should you use?

In general, as long as there was a visible band on a gel, it is hard to have too little DNA to generate good sequence data; e.g., for positive controls, I use as little as 25 ng of pGEM3z(+) plasmid DNA. Instead, the typical problem is using too much DNA... or, at least too much DNA volume. DNA volume is important when the sample is not sufficiently purified; if the sample contains anything that can adversely affect the sequencing reaction, more volume = more adverse effects... and either less sequence data or lower-quality data (or both). Frankly, mid/low-range DNA mass values tend to give the best overall results (read length & signal strength); by contrast, higher ranges (e.g., >600-700 ng for vectors; >15-20 ng for PCR products) can greatly shorten read lengths, especially for reactions using minimal amounts of BigDye (i.e., <0.5 ul BD/10-ul reaction). Please note that these mass values are based on assumption that the vector-insert is ~5000-bp and the PCR products are ~500-bp; because ‘mass’ is being used as a proxy for ‘copy number’, scale these values for vectors or PCR products that differ significantly from those lengths. Again, please see the Science Aid Center for further details.

How much BigDye should you use?

The original BigDye protocols used massive amounts of the BigDye (e.g., 8 ul in a 20 ul reaction); however, you can generate excellent sequence data with much less BigDye as long as you replace the BigDye volume with the appropriate sequencing buffer mix. For routine use, a highly robust and cost-effective protocol is to use 0.5-ul BigDye in a 10-ul reaction; difficult templates might require more BigDye while short templates can be sequenced with less BigDye – see the Science Aid Center and Notes on the Core’s Bulk Sequencing Request form. Please note that reactions become increasingly more sensitive to contaminates and excess DNA as BigDye volume drops below 0.5-ul/10-ul reaction... especially if you don’t vigorously vortex the BigDye before each use.

# Finally, can you trust your DNA sequencing results?

Many people have the mistaken impression that the text file of their sequencing data contains, for lack of a better word, ‘truth’. Well, it does not! The \*.seq file is just a text file that contains the output of basecalling algorithms that made their best guesses about the flashes of light generated as the sample moved through the DNA Sequencer’s capillary array. There are reasons too numerous to list here for why those algorithms might have gotten the wrong results, particularly with regard to specific nucleotide positions and at each end of the sequence. At a bare minimum, you should always examine at least the electropherograms so that you can get an overall sense of the data quality as well as see places where the basecalling might be problematic. However, please note that the analyzed (vs. raw) electropherograms have been ‘massaged’ to look pretty and tend to convey an impression of better quality than really exists. Thus, ideally, you should also have the capacity to examine the raw traces so that you can see the actual quality of peak resolution as well as the signal intensity. This topic is covered in more detail on the Science Aid Center.