DNA Core Facility:
DNA Sequencing Guide

University of Missouri-Columbia
216 Life Sciences Center
Columbia, MO 65211

http://biotech.missouri.edu/dnacore/
# Table of Contents

1. Evaluating Sequencing Data ................................................................. 3

2. Template Quality .......................................................... 5
   Template Quantity .......................................................... 5
   Excess Salt ................................................................. 6
   EDTA ................................................................. 7
   Ethanol ................................................................. 7
   Multiple Templates ....................................................... 8
   Contamination by Cellular Constituents ................................. 9
   Nuclease Contamination ..................................................... 9

3. Primer Design ........................................................................... 10

4. Common Sequence Editing Problems ........................................... 12
   Insertions ........................................................................... 12
   Deletions ........................................................................... 12
   Weak G’s After A’s .......................................................... 13
   Gel Compression .............................................................. 13
   Dye Blobs ........................................................................ 14

5. Difficult Templates .............................................................. 16
   GC-rich Sequence .......................................................... 16
   Homopolymeric Regions .................................................... 17
   Repetitive Regions .......................................................... 17

6. DNA Sequencing Literature......................................................... 18
1. Evaluating Sequencing Data

The introduction of fluorescence-based DNA sequencing systems has significantly improved the accuracy and efficiency of DNA sequencing. The use of fluorescence-based systems has led to the development of software to perform gel analysis and assign bases eliminating the time consuming radioactive-based techniques. Although automation has simplified DNA sequencing allowing large amounts of data to be quickly generated and analysis performed, errors still occur in the DNA sequencing process. The researcher should not trust entirely the associated text file that supplies data in a text-only format. Many problems with sequencing results are not recognized by viewing the text file alone. The quality of sequence data should always be evaluated by studying the accompanying chromatogram to identify problem data and check base calls made by the analysis software.

The general data quality should first be determined by viewing the chromatogram file (Figure 1). Specific programs are required to view these files. Freeware programs are available and can be downloaded from the web. The following are three programs that can be used to view and analyze chromatogram files.

- **Finch TV**
  The software is available for Macintosh OSX, Windows, Linux or Solaris operating systems. All versions can be downloaded from the Geospiza web site. (http://www.geospiza.com/finchtv/index.htm)

- **Chromas**
  The software is for Windows systems and can be downloaded from the Technelysium web site. (http://www.technelysium.com.au/)

- **BioEdit**
  The software is for Windows systems and can be downloaded from the North Carolina State University web site. (*documentation is out of date and no longer maintained*)
  (http://www.mbio.ncsu.edu/BioEdit/bioedit.html)

Sequence samples in the DNA Core facility are analyzed on an ABI 3730 DNA Analyzer. Good sequence reactions can be expected to yield read lengths of at least 800 bases. Peaks should be evenly spaced with minimal background noise. A chromatogram displaying good quality sequence is illustrated in Figure 1.

![Chromatogram view of control sequence. Sequence reaction performed with pGEM plasmid (300ng) and M13 Forward primer (5 uM).](image)
Poor sequencing reactions that give weak signal strengths can be expected to yield shorter read lengths with ambiguous base calls. Often, smaller peaks beneath larger peaks are due to background noise being amplified by the analysis software. The amplification of background signal occurs when the software detects low signal strength and attempts to normalize sequence signal for the entire read length.

Background noise results in a higher percentage of N's and incorrect base calls. Notice the "noisy" nature of the poor data in Figure 2. The signal strength is weak to the point that the sequence is undetectable.

Signal strength can be used as a general indicator as to how well the sequence reaction worked. These values are found on the chromatogram info page (Figure 3) when viewing the chromatogram with Finch TV software.
The run information displayed in Figure 3 is of the chromatogram in Figure 2. Notice the “noisy” nature of the data in the chromatogram. The signal strength is low enough that background noise is also being amplified. Signal strength values of around 200 or less will most likely yield data that is unreliable. Typical signal strength values for each of the four bases are in the 1000 – 2000 range.

2. Template Quality

The quality of sequence data is directly related to the quality of the template. Several companies market plasmid purification kits. Most of these kits will work for DNA sequencing. You may need to try a couple of kits to find the method that works best in your lab. The DNA Core can provide information on plasmid preparation and purification of PCR products for sequencing.

Template Quantity

The amount of template used in a sequence reaction does affect sequence quality. The suggested quantity for plasmid DNA is 250 – 500ng. Total amount of DNA in a reaction should not exceed 1000ng. Excess template usually yields sequence data with short read lengths and excess signal in the beginning. Figure 4 displays the raw signal from two different sequence reactions. The raw signal in Figure 4a is that of a good sequence reaction that maintains a uniform signal the entire length of the gel. The raw signal in Figure 4b is that of a poor sequence reaction with excess template resulting in a rapid decline in signal strength. The 3730 instrument will amplify this problem due to the more efficient uptake of shorter fragments by electrokinetic injection.

![Fig. 4. Raw signal view as displayed in Editview 1.0.1.](image)

(a) Good sequence with uniform signal intensities. (b) Rapid decrease in signal due to excess template.

PCR products require less template than a plasmid in a sequence reaction. A rule of thumb for short PCR products, 100 – 700 bases, is to add 10ng for each 100 bases of PCR product. Larger PCR products can be treated as though they were plasmids.
Excess Salt

Salts can be introduced in a sample through poor elution from a column or a poor wash step during the precipitation method. Excess salts severely inhibit polymerase activity. Figure 5 illustrates the effect of increased salt concentration with the addition of NaCl to the reaction. Figure 5a is of pGEM control sequence from a reaction performed with no salts added to the reaction. The peaks are well resolved with good signal strength. The sequence was 98.5% accurate in the first 861 bases. Figure 5b contains sequence data of a pGEM control reaction in the presence of 20mM NaCl. The peaks are well resolved but signal strength is decreased. Read length and sequence quality is reduced with sequence being 98.5% accurate in the first 695 bases. Figure 5c contains sequence data of a pGEM control reaction in the presence of 40mM NaCl. Data shows a much more dramatic reduction in signal strength. Incorrect base calls are observed. Sequence quality is reduced with sequence 98.5% accurate in the first 640 bases. The addition of the 40mM NaCl resulted in a loss of 220 bases from the read length of the control sequence. Increased amounts of salt in the reaction beyond the 40mM concentration would eventually inhibit the reaction entirely.

Fig. 5. Chromatograms of sequence reactions containing pGEM plasmid (800ng) and M13 Forward (5uM). (a) No NaCl added to reaction (b) 20mM NaCl added to reaction (c) 40mM NaCl added to reaction
EDTA

EDTA is a potent inhibitor of the sequence reaction. Template in buffer containing 1 mM EDTA is acceptable. The final concentration in the sequencing reaction (20μl) will be 0.5 mM or less depending on the volume of template added. A final EDTA concentration greater than 1 mM will inhibit the reaction by chelating critical cations. Figure 6 illustrates the inhibiting activity of EDTA.

Fig. 6. Chromatogram of sequence reaction containing a final volume of 2.5mM EDTA. Sequence reaction performed with pGEM plasmid (800ng) and M13 Forward (5uM).

Ethanol

The presence of ethanol will inhibit polymerase activity. Ethanol contamination most commonly occurs to samples that have been precipitated with ethanol and residual ethanol remaining from an incomplete drying step. Figure 7 displays a set of reaction results that have been spiked with increasing amounts of ethanol. Small amounts of ethanol can be tolerated in the sequence reaction. The polymerase is almost entirely inhibited, however, with a final volume of 10% ethanol as shown in Figure 7c.

(a) 2.5% Ethanol
Figure 7. Series of sequencing reactions with increased amounts of ethanol. Sequence reactions performed with pGEM plasmid (800ng) and M13 Forward primer (5 uM). (a) 2.5% ethanol (b) 5% ethanol (c) 10% ethanol.

Multiple Templates

A sequence reaction that yields more than one sequence product results in multiple peaks at each nucleotide position. Figure 8 displays the appearance of two sequences superimposed over the other in a single chromatogram. Observe that the peaks are well resolved, unlike the background noise of Figure 3, but give the same ambiguous text of a failed reaction. Viewing only the text file, the wrong conclusion about the sequence reaction may be reached. The example in Figure 8 gives an additional indication that more than one product is present. At base 197 the multiple peaks end and a single peak for each base begins. It is at this point that the cloned inserts transition to plasmid sequence.

Obtaining multiple sequence data in a single sequence reaction is contributed to a variety of problems. It is essential to remember that a sequence reaction uses only one primer. Two primers will result in
multiple sequence data. Multiple templates in a submitted sample will likewise result in multiple sequence data. Two clones in a sample can occur when two colonies are inadvertently picked from an agarose plate. Contamination of a glycerol stock or growth media will yield similar results. Precautions are necessary in cloning experiments to ensure that a single clone is present. A restriction digest of the plasmid with insert is recommended to determine the presence of one insert. When sequencing from PCR products it is wise to confirm with an agarose gel that no spurious products are present. A nested primer can be used in situations where multiple products are unavoidable. A nested primer is designed to anneal to an internal site on the specific PCR product.

![Fig. 8. Chromatogram displaying sequence reaction containing two cloned inserts.](image)

**Contamination by cellular constituents**

Poor technique during template preparation will contaminate the sample with cellular constituents that will inhibit polymerase activity giving poor results.

**Nuclease contamination**

Template degradation due to nuclease contamination will result in poor sequence data.
3. Primer Design

Primer selection is critical for successful DNA sequencing. Primers to be used for sequencing must have the correct melting temperature. If the primer/DNA match is 100%, a Tm of 55-60°C (each A,T = 2°C; each G,C = 4°C) will work with cycling conditions used in the DNA Core. Primers 20 bases or longer usually ensure good hybridization. If the match is less complete, the Tm must be higher. A base mismatch on the 3’ end of the primer is more likely to fail. Avoid primers with long runs of a single base (more than three or four), as well as primers that have secondary structure or can form dimers. Primer dimer analysis and Tm calculation can be performed with various software programs. Oligo Analyzer 3.0 on the Integrated DNA Technologies web page (www.idtdna.com) is a free source for primer analysis.

Possible primer problems

- Total amount of primer in reaction should be 5 - 10 pmol. Insufficient primer concentration will result in a failed reaction.
- N-1 primer species will cause peaks to be out of frame with one another and base calling software will not be able to assign bases.
- A primer Tm lower than the annealing temperature of cycle conditions will not anneal to the template.
- Secondary structure in the primer will cause the primer to self-anneal with no primer annealing to the template.
- Poor match or no binding site for primer on template will cause the sequence reaction to fail. If a reaction fails with a standard primer, check to be sure the plasmid contains the primer-binding site and that the site was not lost due to restriction digest and cloning of an insert.
- Multiple priming sites on a template will cause peaks that overlap one another. Base calling software will not be able to assign a base.

Primer Dimer Example

Primer dimer formation is a common problem when primers are not checked for self-annealing. Figure 9 is a chromatogram displaying data collected using a primer that had not been checked for self-annealing. Most portions of the sequence is readable but of poor quality. Observe the short, intense region in the first twenty bases of the sequence. A primer dimer will itself become a template for the reaction from which extension will occur. This region is most likely sequence generated from the self-annealed primer.

![Chromatogram of sequence reaction displaying large primer dimer peaks at the beginning of the sequence.](image-url)
The first step in troubleshooting the sequence in Figure 9 was to check the primer sequence for self-annealing using Oligo Analyzer 3.0. The strongest primer dimer predicted to form is shown in Figure 10.

![Delta G -10.23 kcal/mole Base Pairs 6](image1)

The stop point of the sequence can be adjusted to analyze a specific region. Figure 11 is a chromatogram displaying only the first twenty bases of the sequencing reaction in figure 9. The sequence read from this chromatogram is 5'-ACTGTCGAAAACGCATA-3'. When compared to the predicted primer dimer structure an exact match to the sequence is found.

![Fig. 10. Predicted self-annealing structure for sequencing primer.](image2)

![Fig. 11. Chromatogram displaying shortened read length.](image3)

A second primer was designed and used with the same template. The sequence results (Figure 12) show strong, well resolved peaks with no non-specific binding.

![Fig. 12. Chromatogram displaying data from sequence reaction with new primer.](image4)
4. Common Sequence Editing Problems

Once the quality of the sequence data is determined to be satisfactory, the sequence will need to be edited. Base miscalls by the analysis software are common and should be expected. The following are examples of these and other more common base calling problems.

Insertions

The insertion of an extra base in the sequence is common near the end of a sequence run. As the resolution deteriorates the peaks broaden. The analysis software uses a set value called base spacing to locate peaks in the chromatogram. The base spacing is optimal for the middle region of collected data where resolution is best, but not optimal for the end, or beginning, where resolution is poor. The broad peaks at the end of a run can lead to a single peak being assigned as two bases by the analysis software. The chromatogram in Figure 13 illustrates base insertions that occur between bases 640 and 650. The A at position 645 is an extra base assigned to the A peak. The same is true for the T at position 648. (Also notice that the G directly under the A at position 645 has been missed! Deletions are discussed next.)

Deletions

The exclusion of a base is most common near the beginning of a sequence run, but can be found throughout the entire sequence. Resolution is poor in the beginning of the sequence with peaks sometimes overlapping. Due to the analysis software looking for bases at set intervals, a peak can be missed. Observe the missing A after the G at base 14 in Figure 14. There are two distinct green (A) peaks but the analysis software has only called one base.

---

Fig. 13. Chromatogram illustrating insertion of extra bases at position 645 and 648.

Fig. 14. Chromatogram illustrating insertion of extra bases at position 645 and 648.
Weak G's after A's

A common base miscall is a G that follows an A. The rate of incorporation of G's after A's by the enzyme is low. Compare the signal intensity of the G at base 372 and 375 with that G at base 391 in Figure 15.

![Fig. 15. Chromatogram illustrating the weak G's that follow A's at position 372 and 375.](image)

The G's after A's in Figure 15 are weak but not miscalled by the software. The G, however, can be so weak that the software is unable to assign the base. The G at position 389 and 391 in Figure 16 were assigned as an N because they were too weak causing them to be hidden by the large A peaks.

![Fig. 16. Chromatogram illustrating the weak G's that follow A's at position 372 and 375.](image)

Gel Compression

Problems due to sequence context are observed. These problems are usually associated with GC-rich sequence that is problematic due to their high melting temperature. The example in Figure 17 is a GC-rich PCR product that has a compression region between bases 220-230. The peaks are broad and unresolved with overall resolution in the 220 – 230 base region being poor.

![Fig. 17. Chromatogram illustrating region of sequence results with gel compression.](image)
The problem was solved by adding an extended denaturation step of 95°C for 5 minutes prior to loading the sample. Notice in Figure 18 that the peaks are now well defined.

![Chromatogram illustrating compression region after denaturing step.](image)

**Dye Blobs**

The presence of unincorporated fluorescently labeled dideoxyterminators in the sequence results (Figure 19) is due to their insufficient removal in the dye removal method. The resulting dye blobs, so called for their blob appearance on the array image, will interfere with the analysis of the sequence. Strong dye blobs can cause the loss of the first 60 - 120 bases of the sequence.

![Gel image containing unincorporated dideoxyterminators.](image)

Small amounts of unincorporated dideoxyterminators do sometimes carry through the removal procedure. The majority of sequence data is unaffected. Base calling by analysis software ignores the small amount of signal associated with the dye blob. A small dye blob is illustrated in the 110 – 120 base region of the chromatogram in Figure 20. As in Figure 20, small dye blobs can affect base calls but the data can usually be manually edited.
Fig. 20. Chromatogram with small dye blob between bases 110 – 120.
5. Difficult Templates

GC-Rich Sequence

DNA with a high GC-content can be difficult to sequence. Problematic sequence will begin strong but rapidly lose signal strength until there is no sequence data. Therefore, read lengths are typically shorter for these templates. Figure 21 illustrates this rapid decline in signal strength.

A problem with secondary structure is not exclusive to GC-rich template. Secondary structure caused by short regions of high GC-content is not uncommon. These regions can have secondary structure that the enzyme is unable to melt and process through. The secondary structure is observed as an abrupt stop in the sequence (Figure 22).
Homopolymeric Region

Poly A tails are difficult for the enzyme to process through. A "stutter" effect is observed in the sequence directly downstream of the poly A region. The disassociation and association of the enzyme with the template cause this effect as it processes through the poly A region. A wave appearance with the four dyes will be observed. Notice the increased number of N's directly following the poly A region in Figure 23.

![Figure 23. Chromatogram illustrating stutter effect proceeding poly A region.](image)

Other homopolymeric regions such as the run of G's shown in Figure 24 also causes problems for the polymerase. In this example, the enzyme is unable to process through the G's and disassociates from the template.

![Fig. 24. Chromatogram illustrating loss of signal due to homopolymeric region.](image)

Repetitive Region

Repetitive regions are difficult for the enzyme to process through without dissociating from the template. Usually, as observed in the example below, the signal decreases to the point that no further sequence can be obtained.

![Fig. 25. Chromatogram illustrating loss of signal due to repetitive region.](image)
Of course, not all sequencing and editing problems are shown here. These are a few of the more common problems that will arise when sequencing. Other resources are available on the web that discuss optimizing your sequence reaction. A good source to begin with is the QIAGEN Guide to Template Purification and DNA Sequencing. A good discussion of optimal conditions and the affect of contaminants on sequencing quality is found in this guide.

Questions on sequencing reactions or results can also be directed toward DNA Core Staff.

6. DNA Sequencing Literature

Sequencing Guides

The QIAGEN Guide to Template Purification and DNA Sequencing (2nd Edition)


Reference Articles


