Interpretation of Sequencing Chromatograms

Automated DNA Sequencers generate a four-color chromatogram showing the results of the sequencing run, as well as a computer program’s best guess at interpreting that data – a text file of sequence data. That computer program, however, does make mistakes and you need to manually double-check the interpretation of the primary data. Predictable errors occur near the beginning and again at the end of any sequencing run. Other errors can crop up in the middle, invalidating individual base calls or entire swaths of data.

This document explains how to examine the normal DNA sequencing chromatogram, describing common issues and how to interpret them. Links are also given for more complete troubleshooting of problems.

It is the responsibility of the Core clients to:

- check for mis-calls
- truncate the sequence when errors become too frequent.

The current version of this document reflect our current generation of DNA sequencers, the ABI Model 3730XL. Occasionally, information in this document will refer to older sequencers for historical purposes, and we will specifically state when this is the case.

With a little practice, you can scan a chromatogram in less than a minute and spot problems. It is not necessary to read each and every base.

1. Get a General Sense of How Clean the Sequence Is

How clear are the nucleotide peaks, in general?
You should see evenly-spaced peaks, each with only one color. Peak heights may vary 3-fold, which is normal. “Noise” (baseline) peaks may be present, but with good template and primer, they will be quite minimal. Here’s an example of an excellent sequence. Note the evenly-spaced peaks and the lack of baseline ‘noise’ (see further down for examples of higher baseline noise):
The next example has a little baseline noise, but the ‘real’ peaks are still easy to call, so there’s no problem with this sample:

Now we get to an example that has a bit too much noise. Note the multicolored peaks at 271, 273 and 279, the oddly-spaced interstitial peaks near 291 and 301, and it is impossible to determine the real nucleotide is at 310.

Noise like the above most commonly arises when the sample itself is too dim. Check the ‘Signal Intensity’ numbers for your chromatogram. They are usually available in a separate pull-up window via your chromatogram viewing program, or if you’ve printed the chromatogram, they may be printed at the top. Here are some example signal strength numbers, reflecting arbitrary “relative fluorescence units”:

Signal G:131 A:140 T:98 C:78

In our lab, we usually just look at the first number, the ‘G’ signal, for simplicity. On a 3730, this number should ideally be between 500 and 2000 for best sensitivity and low noise (although this can vary significantly from instrument to instrument, or depending on the alignment procedures used to ‘tune’ the sequencer). Signals between 50 and 100 SOMETIMES give reasonable data
(and sometimes give poor data, as depicted in the above window), but you’ll definitely see baseline noise that would never trouble you on a bright sample.

if your signal is extremely low (anything below G=50, but on some sequencers even G=100) and your peaks are uninterpretable, then you should consider the lane to be simply blank – i.e. a failed sequencing reaction. Don’t try to read baseline noise for usable data!

Note that the 3730 tends to have low baseline noise early in the chromatogram, but it may increase modestly towards the end of the run. This is normal and is a consequence of the capillary electrophoresis technology currently in use. The presence of salt in your samples exacerbates this effect.

II. Check For Mis-Called Nucleotides

Are there obvious errors in the basecalling?

Sometimes the computer will mis-call a nucleotide when a human could do better. Most often, this occurs when the basecaller calls a specific nucleotide, when the peak really was ambiguous and should have been called as ‘N’. Occasionally, the computer will call an ‘N’ when a human would be confident in making a more specific basecall. Such mis-calls can occur even in the most error-free regions of the gel. Quickly scan the gel for extremely small peaks, ‘N’ calls, and any mis-spaced peaks or nucleotides.

1. Mis-spaced peaks: One good way to detect artifacts or errors in a sequencing chromatogram is to scan through it, looking for mis-spaced peaks. At the same time, watch for mis-spaced letters in the text sequence along the top. Nucleotides that have been erroneously inserted into a sequence will often appear to be oddly spaced relative to their neighboring bases, often too close.

3730 sequencers (and probably other current-generation capillary sequencers) have predictable errors in base spacing. A common one for us is a G-A dinucleotide, which leaves a little extra space between them. Often, it’s ignored by the basecaller, as in this example at right: Note the extra space between the letters G and A (nt’s 271 and 272) corresponding to the mis-spaced peaks just below them. No harm done, in this case; the sequence is fine.
Sometimes, however, those spaces get mis-interpreted as missing nucleotides. In the example at right, note the ‘N’ called in the space between the G-A pair. That is an erroneous call; there is no missing base ‘N’ at that position. You can spot this by scanning the text sequence at the top of the window, looking for oddly-spaced letters. Of course, you may also spot this simply by looking for ‘N’ nucleotides.

The real problem comes when the basecaller attempts to interpret a gap as a real nucleotide, such as in the example at right. The typical scenario is a sequence with noticeable baseline noise, and a gap is called as if the baseline noise were a real peak. Often it’s those aforementioned G-A gaps, but not necessarily, as the example here shows. Note the real T peak (nt 58) and the real C peak (nt 60), with the G barely visible between them. Despite its size, the baseline-noise G peak was picked as if it were real. The clues to spot are (i) the oddly-spaced letters, with the G squeezed in, and (ii) the gap in the ‘real’ peaks, containing a low noise peak. This is a great example of why a weak sample, with its consequent noisy chromatogram, is untrustworthy.

2. Heterozygous (double) peaks: A single peak position within a trace may have but two peaks of different colors instead of just one. This is common when sequencing a PCR product derived from diploid genomic DNA, where polymorphic positions will show both nucleotides simultaneously. Note that the basecaller may list that base position as an ‘N’, or it may simply call the larger of the two peaks. Realize, too, that it’s easy for a human to miss these. If you want to be sure you’ve detected all of the polymorphic positions, you should be using a computer program to scan your chromatograms!
Here’s a great example of a PCR amplicon from genomic DNA, with a clear heterozygous single-nucleotide polymorphism (SNP). In this case, one allele carries a C, while the other has a T. Both peaks are present, but at roughly half the height they would show if they were homozygous. Note that the peak was called an ‘N’ by the basecaller. A comparison of text sequences would probably notify you of the presence of a SNP here.

Now we see a het that was missed by the basecaller. The text sequence simply shows a ‘C’. If all of your other sequences also had a ‘C’ here, you would never realize that you had a het SNP … unless you scanned your chromatograms. In fact, it can be difficult to go through reams of sequencing chromatograms, looking for het peaks like this. It’s fine for small projects – just look for the nested multicolor peak. For big SNP-detection projects, though, you should be using a computer program that can detect these for you. Examples are:

- **Sequencher, by GeneCodes, Inc**
- **Mutation Surveyor, by SoftGenetics, Inc**
- the **phred/phrap/polypyred/consed suite of programs**

### III. Loss of resolution later in the gel

**Even normal chromatograms stop giving accurate data after some distance:**

As the gel progresses, it loses resolution. This is normal; peaks broaden and shift, making it harder to make them out and call the bases accurately. The sequencer will continue attempting to “read” this data, but errors become more and more frequent. Here are three snapshots representing data from progressively later regions in a normal chromatogram:
This is a typical example of data from a very good sample analyzed by an ABI Model 3730XL DNA Analyzer. In this case, it is pGEM3 DNA sequenced with the T7 primer, and we are looking at a prime, high-quality portion of the sequence. Note the crisp, clean bands, well separated and with no ambiguity as to the proper basecall. You could easily call this sequence manually.
If we scroll the above chromatogram further to the right (to higher-numbered nucleotides), we see the frame depicted at left. It is evident that, here at 800 nucleotides, the sequence is still quite reliable. The peaks are broader and clearly less well-resolved, but there still is evident separation between them, and no case calls with which I would disagree. Note that the spacing between the basecall letters at top is regular, which is often a good indication of the reliability of the data. When that spacing becomes irregular, be careful!
Here, we are out at the very limit of resolution, around 900-1000 nt on a 3730XL. We get only a general sense of the sequence here; I personally would not design a primer from this sequence, for fear of wasting time on a non-functional primer. There are only a few basecalls that can be considered reliable. The G at 981 may in fact be two G's, the N could be a G or an A, and who knows how many A's there are afterwards. If you aligned the sequence with a known pGEM sequence, you might discover that it is correct, so this is what we sometimes call “useable” data, but certainly not accurate data.

General take-home conclusions: Late in the chromatogram, watch for multiple bases of any one nucleotide where there really should be only one. Watch, too, for wide peaks mis-counted by the program as two nucleotides, when it should have been just one. Wide peaks may also obscure smaller adjacent peaks (no example shown here).
The 3730 can read as far out as 1100 or 1200 nucleotides, but you should expect only 900-950 nt of really good sequence (and even then only if it was a very good sample!), and usable sequence (i.e. error-prone but informative) out to perhaps 1000-1100.

IV. Truncate the sequence when problems become too frequent for YOUR purposes:

What data quality do you need, in order to accomplish your goals?

As the gel progresses, the errors described above become more and more frequent. Where this occurs will primarily depend on the quality of your template. Our sequencers will usually read GOOD templates out to 900 or 950 nucleotides with very low error rates. Ignore the remaining data when the error rate is too high for your purposes.

An investigator trying to locate intron-exon boundaries won’t mind fairly high error rates, but one who needs publishable sequence can accept only the best sequence. One who is doing a BLAST search to identify a coding sequence can accept fairly high error rates, and still obtain a match, but if you want to spot SNPs, you can use only the highest-quality sequence.

The Core technicians do not know your needs, and therefore cannot trim the sequence data for you. We may return as much as 1100 nucleotides of sequence, but it is your responsibility to trim off that which is too error-prone for YOUR purposes.