Exploiting the FLX+ chemistry to sequence long amplicons

Summary
Roche recently launched the GS FLX+ System, an instrument “capable of generating extra-long sequencing reads up to 1,000 bp in length”. Although the FLX+ chemistry offered some gain in read length compared to its predecessor Titanium chemistry, it was not until the availability of a software upgrade in the Fall of 2012 that reads of up to 1000 bases and beyond could be consistently obtained in our hands. Presumably, the gain in read length with the FLX+ chemistry requires a higher signal-to-noise ratio. While this is useful for sequencing high complexity libraries to longer read lengths, the lack of sequence complexity in amplicons leads to signal saturation. Even though Roche does not officially support the use of FLX+ chemistry for the sequencing of amplicons (as of May 2013), we will evaluate in this technical note various approaches to overcome this signal saturation limitation and report on the sequencing of amplicons of 800 bp over their full length with high quality.

As you will be able to appreciate, we have successfully sequenced amplicons using the FLX+ chemistry. A combination of factors allowed for this achievement on amplicon targets of approximately 800 bp and although we have yet to demonstrate it, based on the read lengths obtained with high complexity libraries we suspect that the upper limit, provided that the quality of the PCR products to be sequenced is adequate, is greater than 1000 bp.

Background
The McGill University and Génome Québec Innovation Centre was among a few selected sites to participate in an early access program for the new GS FLX software version 2.8 in the Fall of 2012. Perhaps the most significant addition in this new version of the software is the availability of an acyclic nucleotide flow pattern. A significant increase in read length can be achieved with this new software (Figure 1) with high complexity libraries such as genomic DNA libraries (also referred to as Rapid Libraries) or transcriptome libraries. Compared to FLX+ sequencing with v2.6 software, an increase of 50% to 100% in read length could be achieved with v2.8 software, depending on which sequencing runs and libraries are compared, given that the variability in read length with the v2.6 software was quite high in our hands.

Despite the fact that the FLX+ chemistry has been available for more than one year, sequencing of amplicons using the FLX+ chemistry is still not supported. Attempts to sequence amplicons on FLX+ chemistry using version 2.6 software were not successful; despite experimenting with several parameters, the raw read lengths were consistently excessively long and of low quality (data not shown), presumably because the signal intensity was too strong and the v2.6 software misinterpreted such strong signals as homopolymers. We describe below how we were able to successfully sequence relatively long amplicons using the FLX+ chemistry over their entire length.

Figure 1. Examples of read length distributions of four typical high complexity libraries (all of bacterial genomes) sequenced using software 2.6 versus 2.8.
The traces show the typical range of read length distributions obtained with the two versions of the software using the same sequencing kits. The version 2.8 software yields not only longer but more consistent read length distributions.
Exploiting the FLX+ chemistry to sequence long amplicons

**Strategy**

The strategy employed to increase the nucleotide complexity of the amplicon pool was to stagger the primers such that the resultant reads would not exhibit a synchronized incorporation signal during the nucleotide flow cycles. Figure 2 shows two approaches that can be used to generate a staggered Fusion primer design (Lib-L chemistry shown). The approach in Panel A involves the truncation of the multiplex identifier (MID), usually 10 bases in length, by 0, 1, or 2 nucleotides whereas the approach in Panel B involves the insertion of 0, 1, or 2 random nucleotides between the MID and the target-specific sequence, preferably the inserted nucleotides would be specific degenerate nucleotides so as to avoid extending the target-specific sequence and thereby affect the Tm. Both staggering approaches A and B should generate equally diverse nucleotide complexities, but option B is preferred as it employs the standard, validated MIDs and thereby avoids possible complications with deconvolution of the barcoded reads. The approach used in this study is that shown in Panel A.

**Figure 2.** Schematic representation of the Fusion primer design with two alternative staggering approaches. Staggering can be achieved by either truncation of the MID sequence (panel A) or by insertion of nucleotides between the MID and template specific primer sequence (panel B). Sequences shown are for Lib-L chemistry.
Exploiting the FLX+ chemistry to sequence long amplicons

A nested PCR strategy was used to generate the amplicons sequenced in this study. The first round of PCR was performed to amplify an approximately 1.2 kb fragment that included the target sequences of interest. Products of the first PCR round were diluted to 1:10 and used as template in the subsequent nested PCR. Staggered Fusion primers employing the MID truncation strategy (Panel A) were used in the second round of PCR to amplify an 800 bp target sequence of interest (excluding adapter sequences). The amplicons were agarose-gel purified using the PureLink Quick Gel Extraction kit (Invitrogen, Life Technologies). The amplicons were quantified using the Quant-iT PicoGreen dsDNA assay. Equal amounts of each PCR product were pooled. The amplicon pools were subjected to a final purification using a stringent DNA:bead ratio to deplete any trace amounts of short fragments. The Bioanalyzer traces of the resulting amplicon pools are shown in Figure 3. A very clean and sharp peak migrating at approximately 910 bp is seen for amplicons prepared using both the HPLC-purified and non-HPLC-purified Fusion primers. A high complexity library (Rapid Library) was used as a positive control in this study.

Figure 3. Bioanalyzer High Sensitivity chip analysis of the amplicon pools used in this study. Note the absence of any low molecular weight fragments and a single peak of ~900 bp.
Exploiting the FLX+ chemistry to sequence long amplicons

Analysis

Read length distributions

The read length distributions (post-filtering) are shown in Figure 4. The positive control Rapid Library (Trace F) gave the expected read length distribution which extends beyond 1000 bases. All amplicon sequencing conditions that were tested in this study yielded a sharp peak at just over 800 bases as expected for the amplicon. The difference between the read lengths obtained at the sequencing stage (Figure 4) and the profiles seen in the Bioanalyzer QC profile (Figures 3) is due largely to the trimming of the adapter sequence; there is very little trimming of the 3’ ends of the reads due to poor quality base calls. This result suggests that it is feasible to sequence amplicon targets (excluding adapters) beyond ~800 bp.

![Read length distributions](image)

Sequencing run metrics

Several sequencing run metrics are summarized in Table 1. The amplicons prepared using non-HPLC purified Fusion primers yield reads that are 10-15 bases shorter on average and roughly one Phred quality score unit lower on average as compared to their HPLC-purified counterparts. Also, although a similar number of enriched beads were loaded for amplicons prepared with both HPLC-purified and non-HPLC-purified Fusion primers, the latter seem to generate relatively fewer signal-positive wells (Raw Wells), which in turn translates to a fewer number of reads in the final output.

<table>
<thead>
<tr>
<th>Sequecing library</th>
<th>Beads loaded per 1/8 PTP region</th>
<th>Raw Wells per 1/8 PTP region</th>
<th>Raw Wells per Bead Loaded</th>
<th>% Reads Key Pass</th>
<th>% Reads Passed Filter</th>
<th>Number Reads PF</th>
<th>Number Bases PF</th>
<th>Average length (bp)</th>
<th>Median length (bp)</th>
<th>Average Base Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amplicons (HPLC-purified primers)</td>
<td>340,000</td>
<td>160,901</td>
<td>0.47</td>
<td>98.0%</td>
<td>86.7%</td>
<td>129,521</td>
<td>90,132,206</td>
<td>696</td>
<td>790</td>
<td>33.7</td>
</tr>
<tr>
<td>Amplicons (non-HPLC-purified primers)</td>
<td>305,091</td>
<td>94,024</td>
<td>0.31</td>
<td>97.4%</td>
<td>89.6%</td>
<td>79,370</td>
<td>53,600,598</td>
<td>675</td>
<td>773</td>
<td>32.5</td>
</tr>
<tr>
<td>RL +ve control</td>
<td>340,000</td>
<td>153,059</td>
<td>0.45</td>
<td>97.7%</td>
<td>69.5%</td>
<td>100,437</td>
<td>77,985,789</td>
<td>776</td>
<td>829</td>
<td>30.7</td>
</tr>
</tbody>
</table>

Figure 4. Read length distributions.

Note the distribution of read lengths for the positive control (RL) library which has a peak at approximately 1000 bases and contains reads that extend up to ~1100 bases. The amplicons prepared with both HPLC-purified and non-HPLC-purified Fusion primers both show a single sharp peak at ~800 bases.
Exploiting the FLX+ chemistry to sequence long amplicons

Sequence quality

The average Phred quality score over the first 810 bases of the sequencing datasets was calculated using the FASTX-Toolkit. A comparison of the quality score profiles obtained for amplicons prepared using HPLC-purified or non-HPLC purified Fusion primers is shown in Figure 5 along with the positive control Rapid Library. The quality score profile of the positive control Rapid Library shows a relatively smooth curve with a progressive, gradual decay in quality with increasing read length. Generally speaking, the quality score profiles of amplicons follow a similar overall trend, but exhibits sharp and reproducible fluctuations. Closer analysis reveals that these fluctuations generally coincide with regions of lower complexity (data not shown). These fluctuations are not unique to amplicon sequencing with the FLX+ chemistry as the same type of fluctuations are observed with amplicon sequencing using Titanium chemistry (data not shown).

A significant effect on base quality is seen when using non-HPLC purified Fusion primers, whereby a premature deterioration in base quality can be seen as early as 200-300 bases into the run relative to HPLC-purified Fusion primers. Moreover, non-HPLC purified Fusion primers seem to give rise to more pronounced localized drops in base quality, which drops to a Phred score of 20 at one point with non-HPLC purified material. The HPLC-purified material also shows a similar decrease (at base 420 in Figure 5), but is slightly better, with a Phred score of 21. On average, the gap between quality scores is 2-3 Phred units beyond 200-300 bases into the run, in favor of the amplicons prepared using the HPLC-purified Fusion primers.

**Figure 5.** Phred quality score profile comparisons of FLX+ amplicon sequencing with HPLC-purified vs non-HPLC-purified Fusion primers. The quality score is roughly equivalent over the first 200 or so bases for the amplicons prepared using HPLC-purified or non-HPLC purified Fusion primers. Beyond 200 bases, the amplicons prepared using non-HPLC purified Fusion primers show decay in quality as compared to their HPLC-purified counterparts.
Exploiting the FLX+ chemistry to sequence long amplicons

Conclusion

We have successfully sequenced amplicons using the FLX+ chemistry. A combination of factors allowed for this achievement on amplicon targets of approximately 800 bp, but the results suggest that sequencing of longer amplicons may also be possible. Although we have yet to demonstrate it, based on the read lengths obtained with high complexity libraries we suspect that the upper limit, provided that the quality of the PCR products to be sequenced is adequate, is greater than 1000 bp. The introduction of an acyclic flow of nucleotides in the version 2.8 software is a major advancement and is in large part responsible for the favorable results obtained herein. The hurdle of signal saturation with amplicons (low complexity libraries) was overcome by employing a staggering approach which offset the synchrony of base incorporations. The use of HPLC-purified Fusion primers (or equivalent) is highly recommended as this seems to increase the number and quality of reads that are obtained. It is important to note that the effect of using non-HPLC-purified Fusion primers may be more pronounced depending on the efficiency of synthesis, i.e. vendor from which the primers are purchased.

We would also like to emphasize the importance of having clean PCR products; all precautions should be taken to optimize PCR conditions such that a single, sharp band/peak is obtained and that clean-up be performed on the resulting PCR product to remove primers and short fragments. A quick QC that can be performed on amplicons (or amplicon pools) before submitting them for sequencing is to take a dilute aliquot and subject it to 15-20 cycles of PCR and running the resulting material on a Bioanalyzer or a high percentage agarose gel to see if small fragments amplify. If detected, such small fragments should be depleted as they will be preferentially represented in the final sequencing data.

We compared amplicons prepared using HPLC-purified and non-HPLC-purified Fusion primers because the use of non-HPLC-purified primers increases the likelihood of amplifying non-specific PCR products. Non-specific PCR products hinder sequencing particularly when they are of low molecular weight as they tend to preferentially amplify and hence produce an even stronger signal. The effect of using non-HPLC-purified Fusion primers may vary depending on their length, sequence context, but most importantly on their purity. The latter is a function of the cumulative coupling efficiency of synthesis and can vary significantly from one oligonucleotide vendor to another.

The implication of these results is that larger amplicons can now be sequenced in one contiguous read. Until such time as Roche supports the sequencing of amplicons using the FLX+ chemistry, we recommend that amplicons employ a staggering approach to benefit from the longer read lengths offered by the FLX+ chemistry.

Acknowledgements

We thank Génome Canada and Génome Québec for financial support to the Genorem project. We thank Darina Frieder and Lisa Tremblay for their helpful advice and Roche Diagnostics Canada providing some sequencing kits used to perform this study.

Alfredo Staffa¹, Saad Eldin Hassan², Geneviève Geneau¹, Patrick Willett¹, Jonathan Roux¹, Gary Leveque¹, Alexandre Montpetit¹, Yves Terrat², and Marc St-Arnaud²

¹ McGill University and Génome Québec Innovation Centre; Montréal, Québec, Canada.
² Institut de Recherche en Biologie Végétale, Université de Montréal, Montréal, Québec, Canada.