454 Sequencing System
Guidelines for Amplicon Experimental Design

July 2011
Table of Contents

1. General Guidelines for Amplicon Sequencing Experimental Design ......................... 5
   1.1 Basic Considerations .......................................................................................... 5
   1.2 Depth of Coverage .......................................................................................... 6
   1.3 Multiplexing ...................................................................................................... 6
      1.3.1 Multiplex PCR .................................................................................................. 6
      1.3.2 Multiplex Identifiers (MIDs) .............................................................................. 7
   1.4 Bidirectional vs. Unidirectional Sequencing ......................................................... 7
   1.5 Quantitative Results in Amplicon Sequencing ....................................................... 7

2. Five Types of Amplicon Sequencing Experimental Design ........................................... 9

3. Design 1: Basic Amplicon Sequencing ........................................................................ 11
   3.1 Library preparation .......................................................................................... 11
      3.1.1 Kit and Method .............................................................................................. 11
      3.1.2 Multiplexing .................................................................................................. 12
      3.1.3 Fusion Primer Design .................................................................................... 13
      3.1.4 Amplicon Length Considerations ................................................................. 14
      3.1.5 Fusion Primer Ordering .................................................................................. 15
   3.2 emPCR Amplification .......................................................................................... 15
      3.2.1 Kits and Methods .......................................................................................... 15
      3.2.2 Library Pooling ............................................................................................. 15
   3.3 Sequencing ......................................................................................................... 16
      3.3.1 Kit and Method .............................................................................................. 16
      3.3.2 Pooling ......................................................................................................... 16
   3.4 Data Processing .................................................................................................. 16
   3.5 Data Analysis ..................................................................................................... 17

4. Design 2: Universal Tailed Amplicon Sequencing ....................................................... 18
   4.1 Library preparation .......................................................................................... 19
      4.1.1 Kit and Method .............................................................................................. 19
      4.1.2 Multiplexing .................................................................................................. 20
      4.1.3 Fusion Primer Design .................................................................................... 21
         4.1.3.1 Adding the Universal Tails ........................................................................ 21
         4.1.3.2 Targeting the Universal Tails .................................................................... 21
      4.1.4 Amplicon Length Considerations ................................................................. 23
      4.1.5 Fusion Primer Ordering .................................................................................. 23
   4.2 emPCR Amplification .......................................................................................... 23
   4.3 Sequencing ......................................................................................................... 23
   4.4 Data Processing .................................................................................................. 23
   4.5 Data Analysis ..................................................................................................... 23

5. Design 3: Ligated Adaptors Amplicon Sequencing ....................................................... 24
   5.1 Library Preparation .......................................................................................... 25
   5.2 emPCR Amplification .......................................................................................... 26
   5.3 Sequencing ......................................................................................................... 26
   5.4 Data Processing .................................................................................................. 26
   5.5 Data Analysis ..................................................................................................... 26
6. Design 4: Long Range PCR Amplicon Sequencing ...............................................27
   6.1 Library Preparation .........................................................................................28
   6.2 emPCR Amplification ......................................................................................28
   6.3 Sequencing .....................................................................................................29
   6.4 Data Processing ..............................................................................................29
   6.5 Data Analysis .................................................................................................29

7. Design 5: One-Way Reads Amplicon Sequencing ..............................................30
   7.1 Library preparation .........................................................................................31
      7.1.1 Kit and Method .........................................................................................31
      7.1.2 Multiplexing ..............................................................................................32
      7.1.3 Fusion Primer Design ................................................................................32
      7.1.4 Amplicon Length Considerations ..............................................................33
      7.1.5 Fusion Primer Ordering .............................................................................33
   7.2 emPCR Amplification ......................................................................................33
   7.3 Sequencing ......................................................................................................34
   7.4 Data Processing ..............................................................................................34
   7.5 Data Analysis .................................................................................................34

8. Special Topics for Amplicon Sequencing ..........................................................35
   8.1 Usage of Multiplex Identifiers (MIDs) in Amplicon Sequencing (Lib-A) ..........35
      8.1.1 When To Use MIDs ..................................................................................35
      8.1.2 Sample Encoding Schemes When Using MIDs .......................................36
      8.1.3 The 454Standard MID Set ......................................................................36
   8.2 Short and Long Amplicons .............................................................................38
      8.2.1 Sequencing Short-Length Amplicon Libraries .........................................38
      8.2.2 Sequencing Long-Length Amplicon Libraries ...........................................39
   8.3 Pooling Mixed Length Amplicons ..................................................................41
   8.4 Special Amplicon Quality Filtering Strategies ..............................................42
      8.4.1 Using the Shotgun Data Processing Pipeline ...........................................43
      8.4.2 Changing Stringency by Adjusting the Full Length Error Profile ...............43
   8.5 Adaptor Removal Stringency .........................................................................46
      8.5.1 When to use this ......................................................................................46
      8.5.2 To Carry Out a Residual Adaptors QC Test .............................................47
      8.5.3 Strategies to Modify the Stringency of AMPure XP Purification ..............48
1. GENERAL GUIDELINES FOR AMPLICON SEQUENCING EXPERIMENTAL DESIGN

The processing and sequencing of Amplicon libraries is quite flexible and allows for a wide range of experimental designs. A researcher can choose from a variety of design parameters, such as the length of the Amplicons, the number of Amplicons to pool together (which affects the depth of coverage), and whether to read from the Adaptor A end, the Adaptor B end, or both. Although the setup for a given experiment will depend on the specific project goals, there are a number of general guidelines that should be considered to ensure the best possible result.

- For more details on Amplicon sequencing including nomenclature conventions specific to Amplicon sequencing used in the 454 Sequencing System documentation (e.g. Amplicon, sample, target, etc.), see the 454 Sequencing System Software Manual, Part D.

- This document describes experimental approaches that can be used with all versions of the 454 Sequencing System: the GS FLX, GS FLX+, and GS Junior Systems. As such, this document is generic; make sure to use the kits and methods appropriate for your system.

- At the time of this writing, Amplicon sequencing using the Lib-A emPCR amplification chemistry is supported in the GS FLX+ System only via its backward compatibility with the GS FLX System and the GS FLX Titanium Sequencing Kit XLR70. However, this restriction does not apply to the designs described in this document that use the Lib-L emPCR amplification chemistry.

1.1 Basic Considerations

- The optimal size range to obtain high quality (HQ) reads is 200-600 bp, including the Adaptor sequences.
  - If the context of the experiment does not allow this, see Section 8.2 for modification to the Lib-A emPCR amplification procedure that will improve results for amplicons shorter than 200 bp or between 550 and 950 bp (or longer).
  - If the experiment contains Amplicons of different sizes (e.g. range wider than 150 bp), see Section 8.3 for special considerations.

- Long range PCR products (see section 6) must be longer than 1500 bp or nebulization may be suboptimal, which could result in coverage bias.

- The highest confidence calls for variants will result from bi-directional sequencing. Including both forward and reverse reads provides excellent supporting evidence of the same sequence in two different sequence contexts, ensures that both ends are sequenced with high fidelity, and attenuates most systematic, context-dependent sequencing errors.

- A high-fidelity polymerase must be used in the Amplicon generation step. Use of a low-fidelity polymerase will result in many amplification-derived variations in the sequence. Although there are many choices of enzyme, Roche’s FastStart High Fidelity PCR System is recommended as it has been shown to result in high fidelity DNA synthesis coupled with robust amplification of a wide array of input templates.

- The appropriate level of multiplexing should be determined by considering:
  - The number of amplicons of interest
  - The desired sensitivity/depth of coverage
The number of reads that can be expected from a sequencing Run (or, for the GS FLX and GS FLX+ Systems, from a region of a sequencing gasket)

- When sequencing mixtures of multiple amplicons, care must be taken in the quantitation of individual amplicons and pooling of the amplicons prior to emPCR amplification; equimolar mixtures are typically used, but samples should be pooled so that the desired number of reads is obtained. (Note: When employing MIDs at equimolar concentrations the expected range of MID observations in the final data is within a three fold window.)
- Greater confidence in results may be achieved by running replicates of the biological material through the sequencing process and analyzing the results using statistical methods.
- Comparison of sample versus control will aid in identifying systematic errors. For example, a variant (compared to the reference sequence) that shows at the same frequency in both the sample and the control should not be considered a biologically significant difference.

1.2 Depth of Coverage

The sensitivity of variant detection is directly linked to the fold-coverage of the region (Amplicon) where the variant appears. The guidelines below are a general, though conservative, aid to help determine the level of oversampling required for a desired level of detection. This guidance accommodates a certain amount of experimental variability such as variation in quantitation, pooling, amplification / sequencing efficiencies of MID-labeled amplicons, and amplification efficiencies of long versus short amplicons.

- Heterozygote detection
  - 40x coverage will result in greater than 99.9% chance of detecting a heterozygote
- 5% variation of single base changes or multibase insertions or deletions
  - 1000x coverage will average 50 reads for a typical variant
- 1% variation of single base changes or multibase insertions or deletions
  - 5000x coverage will average 50 reads for a typical variant

1.3 Multiplexing

In many Amplicon sequencing experiments, the number of reads one can obtain in a sequencing Run vastly exceeds the number of reads needed for any given Amplicon for a given sample. To maximize the usage of the 454 Sequencing System's high throughput capabilities, such experiments would therefore benefit from pooling multiple Amplicons for sequencing. For this reason, the notion of multiplexing arises frequently in the discussion of Amplicon sequencing. Specific multiplexing considerations are addressed in the descriptions of the individual experimental designs, below.

1.3.1 Multiplex PCR

If the experiment comprises multiple Amplicons and/or samples, it is usually most straightforward to prepare each as a separate PCR reaction. However, multiplex PCR reactions can be used to generate multiple PCR products in a single reaction, to save time and effort. If multiplex PCR reactions are used to generate an Amplicon library, it is important to carefully balance the templates and primers of each amplicon so all products amplify properly and with a
similar efficiency. Note that if individual products are over- or under-represented in the multiplex PCR, they cannot be normalized for sequencing, which would lead to either too much or too little of that sequence of interest in the resulting experimental data.

1.3.2 Multiplex Identifiers (MIDs)

Multiplex Identifiers (MIDs) can be included during library preparation and used like a DNA barcode to identify Amplicons or samples. MID-containing libraries must be quantitated individually and can then be pooled for emPCR amplification and sequencing, in the desired proportional representation. For most applications, equimolar representation of all Amplicons is optimal. After the sequencing Run, the data analysis software will use the MID tags to sort the reads, per the experimental design.

**Pooling accuracy:** Accurate quantitation and pooling of multiple PCR products is crucially important to ensure that each Amplicon is adequately represented in the sequencing Run.

1.4 Bidirectional vs. Unidirectional Sequencing

Bidirectional sequencing is normally recommended for all Amplicon sequencing projects because the combination of forward and reverse reads provides higher consensus sequence and variant calling accuracy. When screening datasets for variants, therefore, one key criterion for determining whether a variant determination is valid is whether it is supported by both forward and reverse reads. Some of the considerations supporting this include:

- Reads in opposite orientations provide independent confirmation of sequence information.
- Since basecalling accuracy is highest at the beginning of reads, the beginning of reverse reads provides high accuracy sequence information of the sequences at the end of the forward reads for the same Amplicon.
- Basecalling accuracy in the 454 Sequencing System is partly dependent upon the sequence environment; since reads in opposite orientation present a different sequence context, regions that cannot be read through efficiently in one direction can often be resolved in the other direction.

However, there are situations where the sequence of a set of Amplicons is required from only one direction. For example, in experiments with an extremely wide diversity of Amplicon variants (e.g. metagenomics '16 S'-type experiments; see Section 7), the probability of encountering both forward and reverse reads from identical variant species may be low, whereas a unidirectional setup would double the number of reads from a common primer starting point which may be more useful for analytic purposes.

1.5 Quantitative Results in Amplicon Sequencing

One of the typical applications of Amplicon sequencing in the 454 Sequencing System is the detection, identification and quantitation of low-level DNA variants in a population (e.g. as low as 1%), by the “Ultra Deep Sequencing” of one or multiple target sequences of interest. For example, this can be used to sequence a collection of disease-related exons in a set of tissue samples to investigate the correlation between specific (even if rare) mutations and disease.
There are also various other circumstances where the sequencing of Amplicons can be used in the investigation of biological questions, as described in this document.

Accurate quantitation of variants is predicated upon well designed and well executed experiments. Some examples of situations that could skew quantitative results include:

- Low sampling: if the variant existed in too few copies in the initial biological material, sampling errors may be too large to allow accurate quantitation. This may happen if the variant is too rare (design problem) or if too little DNA is available (execution problem).
- Linkage between a variant being measured and another variant in the primer region: variation in the primer region could skew the representation of various fragments amplified from this primer, and therefore of any other variants that may be linked to the primer variant (design problem).
- Loss of emulsion integrity: if emulsion integrity is not maintained during emPCR amplification, many beads may amplify the same fragment, resulting in redundant reads and an overrepresentation of any variants present in this fragment (execution problem).
- In addition, the DNA sequence context can sometimes affect read quality; while this is not a problem for most types of experiments, it may be important when the objective is the accurate quantitation of rare variants, since the variant itself may constitute such a change in the DNA sequence context.

For these reasons, quantitative data must always be considered conservatively, in context. Also, note that the ‘Basic’ experimental design (see Section 3) using fusion primers as described in the Amplicon Library Preparation Method Manual should be used to obtain the most accurate quantitative data, especially of rare variants. All other experimental designs (Section 4 through Section 7) comprise more sources of variability in the variant quantitation.
2. FIVE TYPES OF AMPLICON SEQUENCING EXPERIMENTAL DESIGN

There are five main ways to approach Amplicon sequencing experiments in the 454 Sequencing System, depending on the goal of the experiment and on the types and number of samples. Table 1 shows a brief comparative synopsis of the five design types. They are described in detail in the sections below. In brief:

- The **Basic** Amplicon sequencing design (Section 3) is the one described in the *Amplicon Library Preparation Method Manual*. It is the simplest one and is best used for experiments with relatively few Amplicons and samples.

- The **Universal Tailed Amplicon** design (Section 4) uses a slightly more complex two-step library amplification process but is more economical and greatly simplifies primer design for experiments that include a large number of amplicons and samples. The Fluidigm Access Array System uses this design.

- The **Ligated Adaptors** design (Section 5) uses the GS FLX Titanium Rapid Library Preparation Kit instead of specially designed fusion primers, and is useful for experiments where Amplicons already exist (or at least where template-specific primers already exist).

- The **Long Range PCR** design (Section 6) is a useful alternative to the Basic method or to the NimbleGen Sequence Capture approach, when a large target region (many kbp range) is to be sequenced.

- The **One-way Reads** design (Section 7) is the approach of choice for Amplicon experiments where it is most efficient to sequence from only one primer, maximizing read length, and the added accuracy that can be derived from bidirectional sequencing is not required. This may be true, for example, for applications such as metagenomics experiments monitoring 16S ribosomal genes. It is similar to the Basic design but uses the emPCR Lib-L chemistry.
## 454 Sequencing System Guidelines for Amplicon Experimental Design

### Exp. Design → Basic

<table>
<thead>
<tr>
<th>Recommended Use</th>
<th>Basic</th>
<th>Universal Tailed</th>
<th>Ligated Adaptors</th>
<th>Long Range PCR</th>
<th>One-Way Reads</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recommended Use</strong></td>
<td>• Experiments with few Amplicons or samples</td>
<td>• Experiments with a large number of amplicons and samples</td>
<td>• Experiments where Amplicons (or primers) already exist</td>
<td>• Experiments to cover large target regions (e.g. many kbp range)</td>
<td>• Experiments where read length is more useful than bidirectional accuracy (e.g. 16S)</td>
</tr>
</tbody>
</table>

### Salient Points

<table>
<thead>
<tr>
<th>Basic</th>
<th>Universal Tailed</th>
<th>Ligated Adaptors</th>
<th>Long Range PCR</th>
<th>One-Way Reads</th>
</tr>
</thead>
</table>
| • One primer pair (with MIDs) per Amplicon per sample | • Two-step amplification  
  - Target-specific step provides universal tails  
  - Universal tail-specific step provides 454 Sequencing tails (with MIDs) | • Use pre-existing Amplicons or primers  
  - Use Rapid Library Prep Method (without nebulization)  
  - Adaptors (with MIDs) are ligated non-directionally to PCR products  
  - Use Lib-L emPCR Kit | • Use long range PCR products (>1500 bp)  
  - Use Rapid Library Prep Method (including nebulization)  
  - Adaptors (with MIDs) are ligated non-directionally to nebulized PCR products  
  - Use Lib-L emPCR Kit | Similar to Basic design except fusion primers are designed with the Lib-L Adaptors A (with MIDs) and B  
  - Use Lib-L emPCR Kit  
  - Unidirectional sequencing only (from Adaptor A) |

### Advantages

<table>
<thead>
<tr>
<th>Basic</th>
<th>Universal Tailed</th>
<th>Ligated Adaptors</th>
<th>Long Range PCR</th>
<th>One-Way Reads</th>
</tr>
</thead>
</table>
| • Simplest experimental design – one step amplification goes directly into emPCR  
  - Good in-process QC  
  - Well controlled pooling  
  - Easy and robust data analysis using AVA | • Single target-specific primer pair for a given Amplicon across all samples  
  - Economical if many samples, e.g. using Fluidigm Access Array  
  - Good in-process QC  
  - Well controlled pooling | • No new 454 Sequencing System primers to design (use existing primers)  
  - Obtain reads from both strands from a single Adaptor / MID  
  - Convenient use of the Rapid Library Prep Kit  
  - Straightforward mapping (using GS Reference Mapper) | • Takes fewer Amplicons than Basic design would  
  - Simpler, quicker and requires less starting material than Sequence Capture  
  - Straightforward mapping (using GS Reference Mapper)  
  - Obtain reads from both strands from a single Adaptor / MID | Maximizes information per read  
  - Directional sequencing (but only one direction) using one emPCR kit  
  - Same simple experimental design as Basic, including good in-process QC and well controlled pooling |

### Disadvantages

<table>
<thead>
<tr>
<th>Basic</th>
<th>Universal Tailed</th>
<th>Ligated Adaptors</th>
<th>Long Range PCR</th>
<th>One-Way Reads</th>
</tr>
</thead>
</table>
| • Labor-intensive and expensive if large number of unique Amplicons or MID-labeled samples (many primer pairs)  
  - Somewhat more complex up-front design  
  - Two solution-phase PCR reactions or single but more complex PCR  
  - An additional part of the reads (the ‘Universal Tail’) is non-informative  
  - Data Analysis in AVA can be more complex | • Somewhat more complex up-front design  
  - Ligation process less efficient than PCR  
  - Becomes expensive if many samples (many Rapid Library Prep Kits)  
  - Data Analysis in AVA requires special commands | • No directionality info  
  - Ligation process less efficient than PCR  
  - Cannot use AVA (use GS Reference Mapper or third-party software)  
  - Coverage can be more variable than other designs due to nebulization | • No reverse read confirmation  
  - IDT fusion primer design tool not available  
  - Data Analysis in AVA requires special commands |

Table 1: Comparative table of the five experimental design cases for Amplicon sequencing. ¹ Requires primer design (with MIDs) in both orientations.
3. DESIGN 1: BASIC AMPLICON SEQUENCING

This is the simplest type of Amplicon sequencing design in the 454 Sequencing System. It is particularly appropriate for experiments where relatively few different Amplicons need to be sequenced in relatively few samples because each Amplicon will require the design and synthesis of a specific primer pair for each sample in which it is to be sequenced. The general workflow is shown in Figure 1.

**Library Preparation**
- Design fusion primers (one primer pair per Amplicon per sample)\(^1\)
- Order fusion primers (IDT)
- *Amplicon Library Preparation Method Manual*\(^2\)

**emPCR Amplification**
- Mix MID-labeled libraries as appropriate (typically equimolar)
- emPCR Kit (Lib-A)\(^3\)
- *emPCR Amplification Method Manual – Lib-A*\(^2,3\)

**Sequencing**
- GS Junior Titanium Sequencing Kit or GS FLX Titanium Sequencing Kit XLR70
- *Sequencing Method Manual*\(^2,4\)

**Data Processing**
- Amplicon pipeline

**Data Analysis**
- GS Amplicon Variant Analyzer (AVA), or GS Reference Mapper

Figure 1: Workflow of a ‘Basic’ Amplicon sequencing experiment.

\(^1\) May use MIDs; see Section 8.1. \(^2\) From the appropriate version of the 454 Sequencing System. \(^3\) For emPCR amplification, see the options described in Section 3.2. \(^4\) At the time of this writing, Amplicon sequencing is supported in the GS FLX+ System only via its backward compatibility with the GS FLX System and the GS FLX Titanium Sequencing Kit XLR70.

3.1 Library preparation

3.1.1 Kit and Method

The method for the preparation of a ‘Basic’ Amplicon library is described in the *Amplicon Library Preparation Method Manual*. It does not use any library preparation kit and consists of a simple PCR reaction using the DNA template of interest and a pair of special fusion primers (Figure 2). As a specific primer pair must be designed for each Amplicon for each sample, this can become complex (and onerous) for experiments comprising a large number of Amplicons or samples; in such cases, one should consider whether the ‘Universal Tailed’ Amplicon Sequencing experimental design may be more appropriate (see Section 4).

For more detail on the design of fusion primers, see Section 3.1.3.
3.1.2 Multiplexing

With many Amplicon sequencing experimental designs, a full sequencing Run with a single Amplicon would result in significant oversampling. To maximize the usage of the 454 Sequencing System’s high throughput capabilities, multiple Amplicon libraries can be pooled for emPCR amplification and sequencing (see Section 1.3; details of the pooling strategy for the ‘Basic’ experimental design are given in Section 3.2.2). However, if the multiplexing strategy involves the use of MIDs, then it must be taken into consideration when designing the fusion primers used to prepare the library.

The GS Amplicon Variant Analyzer (AVA) software application has two ways of assigning reads to the appropriate Amplicon or Sample:

- Even when pooled, all Amplicons can be identified by the AVA software by virtue of their template-specific sequences. If Amplicons are the only entities that need to be sorted, i.e. the experiment includes only one Sample or if the Samples will be sequenced in separate sequencing Runs (or, in the GS FLX or GS FLX+ Systems, in separate PTP device regions), then no particular multiplexing design needs to be used.
- However, if the experiment includes multiple Samples for which the same Amplicon(s) are to be sequenced in the same sequencing Run (or in the same PTP device region of a sequencing Run), then the Samples must be tagged by including Multiplex Identifiers (MIDs) in the design of their fusion primers. This will allow the software to assign the reads of each Amplicon to the appropriate Samples.

Although the various Amplicons of each given Sample often may not need to be tagged with separate MIDs, (per the first bullet, above), users may find it convenient to assign different MIDs to each Amplicon AND Sample nonetheless, for clarity in the design of experiments and ease of sorting reads after analysis.

The GS Reference Mapper software can also be used to analyze the results of ‘Basic’ Amplicon experiments, although only for high frequency variants (see Section 3.5). This software also can recognize MID sequences (and trim them before aligning the reads to the reference sequence), but it cannot use them for sample assignment; the GS Reference Mapper software does not address the notion of “Sample” like the AVA does.

For more detail on the design of fusion primers, see Section 3.1.3. For more detail on the usage of MIDs in Amplicon sequencing, see Section 8.1.
3.1.3 Fusion Primer Design

The primers used to generate ‘Basic’ Amplicon libraries are each composed of two or three parts, fused together, as shown in Figure 3.

![Figure 3: Components of fusion primers for ‘Basic’ Amplicon sequencing](image)

- The 5'-portion is a 25-mer whose sequence is dictated by the requirements of the 454 Sequencing System for hybridizing to the DNA Capture Beads (Lib-A), and for annealing the emPCR Amplification Primers and the Sequencing Primer; in addition, this 5'-part must always end with the sequencing key “TCAG” used for Amplicon sequencing. There exist two kinds of such primers, termed “Primer A” and “Primer B”, allowing for the directional sequencing of the target sequence from either end. The exact, required sequences are as shown in blue (Adaptor) and red (key) in Figure 3.

- The 3'-portion of each primer is designed to anneal with a specific sequence on either side of the target of interest on the DNA sample, delineating the margins of the amplicon that will be produced. This requires detailed knowledge of the target sequence, in particular the sites targeted by the primers. Typically, primer design programs are used to determine the sequence length and location optimal for PCR. The template-specific sequences are typically 20-25 nt in length, though this may vary; they are shown on green or blue-green background in Figure 3.

- Optionally, Multiplex Identifiers (MIDs) can be used to “barcode” the Amplicons or samples (see Section 8.1). When used, MIDs must be placed immediately after the sequencing key. They are shown on orange or yellow background in Figure 3.

Unidirectional sequencing:

Although bidirectional sequencing is normally recommended, there are situations where the sequence of a set of Amplicons is required from only one strand. In such cases, primers can be designed with the intention to process each of the libraries through emPCR amplification using only “A” or “B” beads. When doing this, users may want to design half of the libraries to be sequenced from Primer A and half to be sequenced from Primer B, in order to fully utilize the DNA Capture Beads provided in the emPCR Kit (Lib-A) (any version).

Alternatively, to avoid this dual design (or the waste of half the emPCR Kit), the fusion primers can be designed using the “Primer A” and “Primer B” sequences of the Lib-L chemistry, and all templates will be sequenced from Primer A. This case is described in detail as the ‘One-Way Reads’ experimental design, in Section 7.
3.1.4 Amplicon Length Considerations

The Amplicon length, i.e. the distance between and including the two Adaptor sequences, should be given careful consideration. General guidelines are given in Section 1. In addition:

- Many experimental contexts require that the Amplicons be sequenced in their full length in individual reads. For example, this allows utilizing the full power of the AWA software to detect sequencing variants in the samples. In such cases, primer design must keep Amplicon lengths below the read length expectation of the 454 Sequencing System with the chemistry utilized, e.g. ~450 nt with the GS Junior or GS FLX Titanium chemistry (from key to key; see below).
- If full length bidirectional coverage is not required, longer Amplicons could be designed (i.e. the primers further apart), with reads from Primer A and Primer B covering different portions of the Amplicon. Sufficient overlap must be achieved in the middle of the Amplicon sequence between the reads in the two directions if the entire Amplicon needs to be reliably covered. It should be remembered that sequencing accuracy tends to decline toward the end of a read, so overlaps of 100 bases or more are recommended. Therefore, a good amplicon length target for this application would be about twice the expected read length minus this 100 bp read overlap. See Section 8.2.2 for a discussion of important parameters in the handling long amplicons.
- If the Amplicons are very short (e.g. <250 bp), the standard conditions for emPCR amplification may result in excess amplification that can interfere with sequencing. See Section 8.2 for more details on sequencing short amplicons and recommended modifications to the emPCR amplification procedure to address this problem.

In all cases, it is important to remember that the sequencing keys and the template-specific primers (and the MID sequences, if present) are part of the reads (included in read length) but are NOT part of the analyzable target sequence: in particular, any variants detected in the template-specific primer regions would denote errors introduced during oligonucleotide primer synthesis (which are common), not biologically relevant sequence variants.

This has direct implications for Amplicon length considerations. For example, for a read to cover an Amplicon entirely (first bullet, above), it must traverse its key (4 nt) at the proximal end and the template-specific primers (variable length; typically ~20 nt) and MID sequences (if present; 10 nt each) at both ends. This can typically add up to 50 nt or more, and must be considered during experimental design. If single reads do not need to cover the entire Amplicons (second bullet, above), then only the proximal key and template-specific primer (and MID) will yield non-informational sequence.

Also, it is important to understand that the “expected read length” for any given chemistry can only be stated approximately because the actual length achieved by a read during a sequencing Run depends on the particular sequence context that is being read. The expected read lengths are averages that assume a random, balanced template sequence, in which case, about 2.5 nt would be incorporated per nucleotide flow cycle. While the global results of a sequencing Run on a complex sample (e.g. a genomic library) are likely to approximate this average, this is not necessarily the case for non-complex Amplicon libraries. For example, an amplicon sequence identical to the repeated nucleotide flow order during a sequencing Run would incorporate 4 nt per cycle, not 2.5; and homopolymers incorporate multiple nucleotides in a single flow. Other template sequences may incorporate less than 2.5 nt per cycle. In an Amplicon library, since all the reads for a given Amplicon are similar, this sequence-specific effect on read length can have a significant impact and should be taken into account at the time of primer design.
3.1.5 Fusion Primer Ordering

454 Life Sciences, a Roche company, has entered into an exclusive agreement with Integrated DNA Technologies, Inc. (IDT) for the procurement of special primers for use in the 454 Sequencing Systems, including fusion primers for Amplicon sequencing. The IDT web site, www.idtdna.com, provides a tool to help design fusion primers (including design with MIDs; see Section 8.1) as well as all details for ordering.

3.2 emPCR Amplification

3.2.1 Kits and Methods

For the GS Junior System, the emPCR Amplification of a ‘Basic’ Amplicon library is normally done using the GS Junior Titanium emPCR Kit (Lib-A); the method is described in the emPCR Amplification Method Manual – Lib-A, GS Junior Titanium Series. If multiple sequencing Runs are needed for the experiment, one can use the GS FLX Titanium MV emPCR Kit (Lib-A) to produce up to 8 Runs worth of beads at a time, at less cost and labor; this method requires additional equipment, and is described in the emPCR Amplification Method Manual – Lib-A, Multiple Prep (MV), GS Junior Titanium Series.

For the GS FLX or the GS FLX+ Systems, use the GS FLX Titanium emPCR Kit (Lib-A) of the size appropriate for the experiment (LV, MV or SV), and the corresponding emPCR Amplification Method Manual – Lib-A, GS FLX Titanium Series. Note that at the time of this writing, Amplicon sequencing using the Lib-A emPCR amplification chemistry is supported in the GS FLX+ System only via its backward compatibility with the GS FLX System and the GS FLX Titanium Sequencing Kit XLR70.

3.2.2 Library Pooling

As mentioned in Sections 1.3 and 3.1.2, Amplicon sequencing experiments are typically conducted on a pool of Amplicon libraries rather than on individual Amplicons. When designing the pooling strategy for an experiment, however, consideration should be given to the following:

- The extent of pooling (number of Amplicon libraries to pool) should be determined by the depth of coverage needed for each Amplicon and the total number of reads expected per sequencing Run. For example, if you seek to detect rare variants to 1% frequency:
  - The guideline recommendation is to aim for 5000x Amplicon coverage, so that approximately 50 variant reads will be expected (see general guidelines in Section 1).
  - The expected average number of reads for an Amplicon sequencing Run (or PTP device region) is as follows:
    - ~70,000 reads for a GS Junior System sequencing Run
    - ~20,000 for a small region (16 region gasket) of a GS FLX System Run
    - ~75,000 for a M/S region (8 region gasket) of a GS FLX System Run
    - ~150,000 for a medium region (4 region gasket) of a GS FLX System Run
    - ~375,000 for a large region (2 region gasket) of a GS FLX System Run
  - In the GS Junior System, for example, 70,000 reads would be sufficient to sequence 14 Amplicon libraries per Run and examine variants down to the target frequency of 1% with high statistical confidence (expect approximately 5000 reads per Amplicon with close to 50 reads of a 1% variant).
If the experiment includes more Amplicons / samples than this, you can use the GS FLX System (e.g. 375,000 Amplicon reads per large region of a PTP device would allow the examination of about 75 Amplicon libraries at 1% frequency (150 Amplicon libraries per Run); or use the GS FLX Titanium MV emPCR Kit (Lib-A) to prepare up to 8 GS Junior System sequencing Runs worth of beads in a batch.

- If the GS Reference Mapper will be used to analyze the data (e.g. for common variants and heterozygote detection; see Section 3.5), pooling should be done such that each allele will be represented by about 20 ~ 100 reads.
- The pooled libraries can be either different Amplicons of the same sample, of the same Amplicon in different samples, or any combination of these. Note, however, that MIDs are necessary when combining Samples with the same Amplicons; see Section 3.1.2.
- Pooling should typically be equimolar to ensure proper coverage of all pooled libraries. Accurate quantitation of the PCR products is critical to ensure accurate pooling! See Caution in Section 1.3.2.
- Special consideration must be given before pooling when sequencing short and long Amplicons together; see Section 8.3 for more details.

### 3.3 Sequencing

#### 3.3.1 Kit and Method

There is only one kit and one method for the sequencing step in each of the versions of the 454 Sequencing System. Use the GS Junior Titanium Sequencing Kit and the method described in the *Sequencing Method Manual, GS Junior Titanium Series*; or the GS FLX Titanium Sequencing Kit XLR70 and the method described in the *Sequencing Method Manual, GS FLX Titanium Series* or the *Sequencing Method Manual, GS FLX+ Series – XLR70 Kit*, depending on your system. At the time of this writing, Amplicon sequencing is supported in the GS FLX+ System only via its backward compatibility with the GS FLX System and the GS FLX Titanium Sequencing Kit XLR70.

#### 3.3.2 Pooling

Any pooling strategy set up during library preparation (as part of primer design) and emPCR amplification carries into the sequencing step: libraries amplified together are sequenced together on the 454 Sequencing System Instrument.

### 3.4 Data Processing

Data obtained using the ‘Basic’ Amplicon Sequencing experimental design are processed using the “Amplicon” signal processing pipeline. This is normally set during the setup of the sequencing Run, in the ‘Choose Run processing type’ window of the Instrument Procedure Wizard; otherwise, make sure to select the correct pipeline when setting up signal processing. (See the *Sequencing Method Manual* or the *454 Sequencing System Software Manual, Part A.*

July 2011

16
3.5 Data Analysis

The GS Amplicon Variant Analyzer (AVA) software provides for easy and robust analysis for the identification and quantitation of known or novel sequence variants in samples sequenced using the ‘Basic’ Amplicon Sequencing experimental design. In a well designed and well executed experiment, rare variants with a prevalence of 1% or less can be analyzed. The AVA software will assign each read to the proper Amplicon and Sample (using MID information if applicable), and report on the prevalence of variants, in each Sample defined in the experiment. See the 454 Sequencing System Software Manual, Part D for details.

Alternatively, the GS Reference Mapper software (HCDiffs tab or output file) can be used instead of the AVA, when the experiment only seeks the detection of high frequency variants (e.g. > 30% of the reads for a given Amplicon in a given Sample) such as for heterozygote determination. The HCDiffs file can be screened by applying filters to identify variants with the highest confidence level.

To use the GS Reference Mapper, each allele analyzed should be observed in about 20 ~ 100 reads (i.e. low to moderate depth) because the GS Reference Mapper application is not tuned to handle very high depth of coverage like the AVA software is. If the coverage depth is greater than 100-fold, the incidence of false positives may increase.

In general, the variants in the HCDiffs file that have one or more of the following characteristics are more likely to be false positives:

- Minimal depth, e.g. fewer than 7 supporting reads
- Variant frequency less than 30%
- Skewed ratio of forward and reverse reads, e.g. less than 10% or greater than 91%
- Presence of dinucleotide repeats in the aligned region
- Proximity to more than 7 repeat units
- Proximity to homopolymer sequences longer than 6 bases

This filtering strategy has been found to eliminate most false positives while not eliminating true variants.

Note that the GS Reference Mapper cannot use MIDs to specifically assign reads to Amplicons or Samples. It can, however, selectively process reads of a data set that contain given MIDs, for proper binning. See the 454 Sequencing System Software Manual, Part C for details.
4. DESIGN 2: UNIVERSAL TAILED AMPLICON SEQUENCING

If the experiment includes a large number of samples for multiple Amplicons, the ‘Basic’ experimental design described in Section 3 could become quite complex and labor intensive because it requires a specific pair of fusion primers for each Amplicon for each sample. The ‘Universal Tailed’ Amplicon sequencing design reduces the number of primers needed by attaching, in a first round of PCR, a pair of fixed sequence tags, the “universal tails”, to either sides of all the Amplicons; this requires only as many primer pairs as there are different Amplicons in the experiment. Then, a second round of PCR is carried out, targeting the “universal tails” and adding MID-labeled Primer A and Primer B sequences to barcode the samples; this requires only as many MID-labeled primer pairs as the experiment has samples. Overall, the process requires as many primer pairs as the sum of Amplicons and samples in the experiment rather than the product of Amplicons and samples as would be the case in the ‘Basic’ design.

The general workflow of a ‘Universal Tailed’ Amplicon sequencing experiment is shown in Figure 4.

---

**Figure 4: Workflow of a ‘Universal Tailed’ Amplicon sequencing experiment**

1. **Library Preparation**
   - Design target-specific ‘Universal Tailed’ fusion primers (one primer pair per Amplicon)
   - Design ‘Universal Tailed’-directed Primers A and B fusion primers (MID-labeled; one primer pair per sample)
   - Order all fusion primers (IDT)
   - PCR round 1: add Universal Tails

2. **emPCR Amplification**
   - Mix MID-labeled libraries as appropriate (typically equimolar)
   - emPCR Kit (Lib-A)
   - emPCR Amplification Method Manual – Lib-A

3. **Sequencing**
   - GS Junior Titanium Sequencing Kit or GS FLX Titanium Sequencing Kit XLR70
   - Sequencing Method Manual

4. **Data Processing**
   - Amplicon pipeline

5. **Data Analysis**
   - GS Amplicon Variant Analyzer (AVA), or GS Reference Mapper

---

1. See Section 8.1 for the use of MIDs; 2 From the appropriate version of the 454 Sequencing System; 3 For emPCR amplification, see the options described in Section 4.2; 4 At the time of this writing, Amplicon sequencing is supported in the GS FLX+ System only via its backward compatibility with the GS FLX System and the GS FLX Titanium Sequencing Kit XLR70.
In practice, the two PCR reactions can be carried out concurrently by adding all four primers together in a single reaction. This has the obvious advantage of reducing the labor for the procedure; and by eliminating any pipetting between the two reactions, it also reduces the risk of contaminating samples with one another after the Universal Adaptors have been added, a particular delicate moment when all samples could equally react with all second-round primers. However, the single-step procedure may give rise to more spurious PCR products because the reaction mixture is more complex; and it may be less amenable to multiplex PCR reactions.

A particularly convenient implementation of the single-step version of this experimental design is the use of Fluidigm’s Access Array. The Access Array allows the user to easily set up and carry out PCR reactions for up to 48 Amplicons for each of up to 48 samples, for a total of 2304 simultaneous PCR reactions. At the end of the procedure, each sample will be collected as a pool of all of its Amplicons. For more details on Fluidigm’s Access Arrays, see www.fluidigm.com.

4.1 Library preparation

4.1.1 Kit and Method

As for the ‘Basic’ Amplicon sequencing experimental design, the ‘Universal Tailed’ design does not use any library preparation kit. It requires two sets of fusion primer pairs, used in two rounds of PCR (successive or concurrent).

- The first round uses fusion primers targeting the template-specific sequences (defining the boundaries of the Amplicons) fused to a universal sequence that will be the target of the second round primers (Figure 5A). In order to maintain sequencing directionality, different universal tails should be designed for the Primer A and Primer B ends of the Amplicons (Univ-A and Univ-B on the Figure).
- In the second round, the Univ-A sequence is targeted by a fusion primer that is tailed by the 454 Sequencing System’s Primer A sequence plus an MID sequence to identify the sample; and same for the Univ-B sequence, with Primer B (Figure 5B).

Details on the design of the two sets of fusion primers necessary for the ‘Universal Tailed’ Amplicon sequencing experimental design are given in Section 4.1.3.

If the two rounds of PCR are performed successively, the Universal Tails are added in a first PCR reaction; and the method for the second round, targeting these Universal Tail sequences and adding the 454 Sequencing System primers, is identical to the one described in the Amplicon Library Preparation Method Manual.
4.1.2 Multiplexing

Libraries belonging to one or multiple (MID-labeled) samples can be pooled for emPCR amplification and sequencing (see Section 4.2). The general considerations for multiplexing apply, including the critical requirement for accurate quantitation and pooling of the libraries (see Section 1.3).

The GS Amplicon Variant Analyzer (AVA) software application will assign reads to the appropriate Amplicon and sample as follows:

- All Amplicons can be identified by the AVA software by virtue their template-specific sequences.
- The samples will be identified by the Multiplex Identifiers (MIDs) included in the design of their fusion primers.

Note that one should NOT assign different MIDs to each Amplicon in this case, as this would obliterate the advantage of this experimental design over the ‘Basic’ design.

As for ‘Basic’ Amplicon experiments (Section 3.5), the GS Reference Mapper software can also be used to analyze the results obtained with the ‘Universal Tailed’ design, although only for common variants (see Section 4.5), and at low to moderate depth of coverage. This software also can recognize MID sequences (and trim them before aligning the reads to the reference sequence); contrary to the ‘Basic’ design, however, the MIDs are specifically used to barcode the samples in the ‘Universal Tail’ design, so sample assignment does not need to be handled explicitly by the software.
For more detail on the design of fusion primers, see Section 4.1.3. For more detail on the usage of MIDs in Amplicon sequencing, see Section 8.1.

4.1.3 Fusion Primer Design

4.1.3.1 Adding the Universal Tails

The fusion primers for the first PCR reaction are composed of two parts, fused together (Figure 6). Neither of the two parts has a fully pre-defined sequence.

![Figure 6: Components of fusion primers for the first round of PCR for Universal Tailed Amplicon sequencing](image)

- The 5’-portion is the universal tail that will be added to all Amplicons, and will be the targets of the fusion primers for the second PCR reaction. A universal tail can be any sequence that meets the usual requirements of PCR primers regarding Tm (for annealing in the second PCR reaction), propensity to anneal spuriously with other components of the system or to self anneal (hairpin loops), etc. In order to maintain control over the directionality of the sequencing reads, you should define two such sequences, one to be associated with Primer A, termed Univ-A, and the other with Primer B, termed Univ-B, as part of the fusion primers for the second PCR reaction (see below). These are shown on [indigo] or [blue] background in Figure 6.

- The 3’-portion of each primer is designed to anneal with a specific sequence on either side of the target of interest on the DNA sample, delineating the margins of the amplicon that will be produced. This requires detailed knowledge of the target sequence, in particular the sites targeted by the primers. These template-specific sequences are typically 20-25 nt in length, though this may vary; they are shown on [green] or [blue-green] background in Figure 6.

Although M13 forward and reverse primers have been used successfully, the sequence context of the template-specific sequence could interfere if hairpin loops or other DNA structures form in the primer. As a result, the M13 sequences may not be suitable for every project.

4.1.3.2 Targeting the Universal Tails

The fusion primers for the second PCR reaction are composed of three parts, fused together (Figure 7). These are similar to the fusion primers used in the ‘Basic’ experimental design except that they target the universal tail sequences, Univ-A and Univ-B.
Figure 7: Components of fusion primers for the second round of PCR for Universal Tailed Amplicon sequencing

- The 5’-portion is a 25-mer whose sequence is dictated by the requirements of the 454 Sequencing System for binding to the DNA Capture Beads (Lib-A), and for annealing the emPCR Amplification Primers and the Sequencing Primer; in addition, this 5’-part must always end with the sequencing key “TCAG” used for Amplicon sequencing. There exist two kinds of such primers, termed “Primer A” and “Primer B”, allowing for the directional sequencing of the target sequence from either end. The exact sequences are as shown in blue (Adaptor) and red (key) in Figure 7.

- The 3’-portion of each primer is designed to anneal with the complementary strand of the universal tails added to either side of the target of interest on the DNA sample. Their sequence, therefore, must be the same as the Univ-A and Univ-B sequences (exact same sequences; not the complements nor the reverse-complements). They are marked as Univ-A and Univ-B, and shown on indigo or blue background in Figure 7.

- Multiplex Identifiers (MIDs) must be used in this experimental design, to “barcode” the samples (see Section 8.1). MIDs must be placed immediately after the sequencing key. They are shown on orange or yellow background in Figure 7.

**Unidirectional sequencing:**

Although bidirectional sequencing is normally recommended, there are situations where the sequence of a set of Amplicons is required from only one strand. In such cases, primers can be designed with the intention to process each of the Amplicons through emPCR amplification using only “A” or “B” beads. When doing this, users may want to design half of the Amplicons to be sequenced from Primer A and half to be sequenced from Primer B, in order to fully utilize the DNA Capture Beads provided in the emPCR Kit (Lib-A).

Alternatively, to avoid this dual design (or the waste of half the emPCR Kit), the fusion primers can be designed using the “Primer A” and “Primer B” sequences of the **Lib-L** chemistry (in blue below). In this case, the Amplicons will be processed using the emPCR Kit (Lib-L), and all will be sequenced from Primer A (thus, MIDs should be included only on Primer A). In this scheme, data processing will still be done using the Amplicon pipeline, and data analysis using the AVA or the GS Reference Mapper software.

**Forward primer (Primer A, Lib-L):**

5’-CCATCTCATCCTGCTCCTGCTTCCGAC TCAG-(MID) -(universal tail A (Univ-A))-3’

**Reverse primer (Primer B, Lib-L):**

5’-CCATCTCATCCTGCTCCTGCTTCCGAC TCAG-(universal tail B (Univ-B))-3’
4.1.4 Amplicon Length Considerations

The Amplicon length considerations listed for the ‘Basic’ experimental design (Section 3.1.4) also apply to the ‘Universal Tail’ design. Note that the universal tails are also part of the reads (included in read length) but are NOT part of the analyzable target sequence. Indeed, the fact that a larger portion of the reads is non-informational is one of the drawbacks of the ‘Universal Tail’ experimental design. This includes the key sequences (4 nt), the MID sequences (10 nt) and the universal tails (variable length) at both ends of the read, and the template-specific primer (variable also) at the far end; together, these can typically account for around 100 nt, a sizable fraction of the read.

Users may also consider whether the unidirectional variation of this design (see Note in the previous section) is appropriate for the experiment, as applications using this approach may not require reading through the full length of the Amplicon. In this case, primer sequences at the far end of the reads would not be considered in the analysis and longer Amplicons could be designed.

4.1.5 Fusion Primer Ordering

As described for the ‘Basic’ experimental design (Section 3.1.4), fusion primers can be designed and ordered from the IDT web site, www.idtdna.com. Note, however, that the fusion primer design tool on the IDT web site does not cover the design of universal tail sequences.

4.2 emPCR Amplification

The kits and methods for emPCR amplification and the library pooling considerations for the ‘Universal Tail’ experimental design are the same as for the ‘Basic’ design (Section 3.2).

4.3 Sequencing

The kit and method for sequencing and the library pooling considerations for the ‘Universal Tailed’ experimental design are the same as for the ‘Basic’ design (Section 3.3).

4.4 Data Processing

Data processing considerations for the ‘Universal Tailed’ experimental design are the same as for the ‘Basic’ design (Section 3.4): use the Amplicon signal processing pipeline.

4.5 Data Analysis

Data analysis considerations for the ‘Universal Tailed’ experimental design are the same as for the ‘Basic’ design (Section 3.5): the main approach is to use the GS Amplicon Variant Analyzer (AVA) software, though the GS Reference Mapper may also be used, depending on the objectives of the experiment.
5. DESIGN 3: LIGATED ADAPTORS AMPLICON SEQUENCING

This experimental design can be useful if a set of Amplicons already exists, as it allows the sequencing of pre-existing Amplicons even if they were not produced with fusion primers carrying the Primer A and Primer B sequences. Instead, the GS FLX Titanium Rapid Library Preparation Kit is used to ligate the RL Adaptors on either side of the existing Amplicons without the need to design fusion adaptors (Figure 8). Note that the task of designing the amplicon primers is not really simpler when using the ‘Ligated Adaptors’ experimental design; primer design is simply assumed to pre-exist.

In the Rapid Library preparation method, the target fragments (the Amplicons in this case) will ligate to the RL Adaptors in both orientations. Therefore, it is important to note that the information on the directionality of the reads is not available when using this experimental design, unless some known flanking sequences are present at the ends of the Amplicons. Also, the library preparation may be more complex and more expensive than the ‘Basic’ design because it requires the use of the GS FLX Titanium Rapid Library Preparation Kit (one kit for each 12 libraries); if the experiment includes a large number of Amplicons and samples, the cost of the library preparation kits should be weighed against the cost of designing and producing fusion primers and using the ‘Basic’ or the ‘Universal Tailed’ designs.

Users should also note that the RL Adaptors require the use of an emPCR Kit (Lib-L) for emPCR amplification. The use of these Adaptors has another implication, in addition to the possible loss of directionality information: the “Lib-L” Adaptors require a currently undocumented feature of the GS Amplicon Variant Analyzer (AVA) software. If you use this Amplicon sequencing experimental design, contact your Roche Technical Support Representative for help on how to use the AVA software to analyze the resulting data.

This experimental design may be particularly advantageous if many Amplicons can be produced with a simple pair of common primers, such as a library of fragments cloned into a plasmid vector providing known flanking sequences.

In such a case, however, an alternative to this experimental design would be to use the vector flanking sequences as template-specific sequences and apply the ‘Basic’ design (Section 3). This would circumvent the caveats mentioned above: directional information is ensured, the GS FLX Titanium Rapid Library Preparation Kit is not used, and the data can be readily analyzed using the AVA software.
5.1 Library Preparation

Library preparation for the ‘Ligated Adaptors’ Amplicon sequencing design is simply the Rapid Library preparation method as described in the Rapid Library Preparation Method Manual, with the following two modifications:

- Skip the nebulization step (Section 3.1 in the Method Manual)
- In Section 3.2, step 2, use 500 ng of your Amplicon, add the 9 µl of End Repair mix as described, and then complete the volume to 16 µl with 1x TE.

The Amplicons must pre-exist and be of an appropriate size for the Rapid library preparation method (but see Note, below). The RL Adaptors from the GS FLX Titanium Rapid Library Preparation Kit (or RL MID Adaptors from the GS FLX Titanium Rapid Library MID Adaptors Kit) are ligated to the ends of the Amplicons as described in the Method Manual.
Amplicon length:

- If the Amplicons are long, e.g. longer than 1000 or 1500 bp, use the ‘Long Range PCR’ experimental design, Section 6.
- If, on the other hand, the Amplicons are short, e.g. <300 bp, it may be necessary to increase the volumetric ratio of AMPure XP beads suspension to DNA solution to 1.8 : 1; this depends on the exact Amplicon length. See Section 8.5 for more details on modifying the AMPure XP beads to DNA ratio.

5.2 emPCR Amplification

Options for emPCR amplification for this experimental design are the same as for the ‘Basic’ design (Section 3.2.1) EXCEPT that it uses the Lib-L emPCR amplification chemistry rather than the Lib-A chemistry. Also, if RL MID Adaptors were used to tag different samples, these can be pooled following the usual pooling strategies that apply to multiplexed Rapid libraries, plus those described for the ‘Basic’ Amplicon sequencing experimental design (Section 3.2.2), to obtain the depth of coverage required for the experiment.

5.3 Sequencing

The kit and method for sequencing and the library pooling considerations for the ‘Ligated Adaptors’ experimental design are the same as for the ‘Basic’ design (Section 3.3).

5.4 Data Processing

Data processing considerations for the ‘Ligated Adaptors’ experimental design are the same as for the ‘Basic’ design (Section 3.4): use the Amplicon signal processing pipeline.

5.5 Data Analysis

Data analysis considerations for the ‘Ligated Adaptors’ experimental design are similar to those for the ‘Basic’ design (Section 3.5): use the GS Amplicon Variant Analyzer (AVA) software. However, the AVA software is not tuned by default to process Amplicons that contain the Primer A and Primer B sequences of the Rapid library Adaptors; there is also a risk that reads may be eliminated due to variability in the effectiveness of ligation, during library preparation. In addition, usage of the “Lib-L” Adaptors requires a currently undocumented feature of the GS Amplicon Variant Analyzer (AVA) software. Contact your Roche Technical Support Representative for help on how to use the AVA software to analyze the data from an experiment performed using the ‘Ligated Adaptors’ design.

The GS Reference Mapper software can be used, with the same constraints as for the ‘Basic’ experimental design (Section 3.5): the HCDiffs file output by the mapper can be used to identify heterozygotes or mutations represented by about 20 to 100 reads (i.e. low to moderate depth; the GS Reference Mapper is not tuned to handle very high depth of coverage like the AVA software is) that represent more than 30% of the reads for that Amplicon.
6. DESIGN 4: LONG RANGE PCR AMPLICON SEQUENCING

The ‘Long Range PCR’ experimental design is somewhat similar to the ‘Ligated Adaptors’ design (Section 5). It consists in the preparation of long PCR products (e.g. >1500 bp) to cover a region of interest, and their processing using the GS FLX Titanium Rapid Library Preparation Kit and method (including nebulization of the long range PCR products). Like in the ‘Ligated Adaptors’ design, the RL Adaptors (or the RL MID Adaptors) are ligated non-directionally to the ends of the sequencing templates, in this case nebulized fragments from the long range PCR product, and amplified using the emPCR Kit (Lib-L) (Figure 9).

This experimental design should be considered for the resequencing of a large target region (e.g. many kbp) from an even larger DNA source (e.g. a large genome). Other approaches that can be considered for this type of experiment include the NimbleGen Sequence Capture method and the ‘Basic’ Amplicon sequencing experimental design described in Section 3. While those methods would work well, the ‘Long Range PCR’ design may be simpler, quicker and require less starting material than the Sequence Capture method; and it would achieve the coverage of the region of interest with fewer PCR products than the ‘Basic’ design. Another advantage of this experimental design over Sequence Capture is that well-designed Amplicons can specifically target repeated genomic regions, whereas Sequence Capture cannot discriminate repeat regions and will capture similar sequences. However, the ‘Long Range PCR’ design does not provide read directionality information, and the data analysis cannot use the 454 Sequencing System’s GS Amplicon Variant Analyzer (AVA) software, so the GS Reference Mapper must be used.
6.1 Library Preparation

Library preparation for the ‘Long Range PCR’ experimental design is done in two steps:

- Prepare a set of long range PCR products covering the region of interest
- Process them using the GS FLX Titanium Rapid Library Preparation Kit and method (including the nebulization step)

Per the general guidelines for Amplicon sequencing (Section 1), the long range PCR products should be designed such that they are at least 1500 bp in length in order to nebulize well. Except for this restriction, the library preparation method is as described in the Rapid Library Preparation Method Manual, including the possible use of the GS FLX Titanium Rapid Library MID Adaptors Kit for multiplexing.

6.2 emPCR Amplification

Like the ‘Ligated Adaptors’ design, the emPCR amplification for the ‘Long Range PCR’ experimental design uses the emPCR Kit (Lib-L), but the options are otherwise the same as for the ‘Basic’ design (Section 3.2.1). Also, if RL MID Adaptors were used to tag different samples,
these can be pooled following the usual pooling strategies that apply to multiplexed Rapid libraries, plus those described for the ‘Basic’ Amplicon sequencing experimental design (Section 3.2.2), to obtain the depth of coverage required for the experiment.

6.3 Sequencing

The kit and method for sequencing and the library pooling considerations for the ‘Long Range PCR’ experimental design are the same as for the ‘Basic’ design (Section 3.3).

6.4 Data Processing

In this experimental design, the templates being sequenced are not “Amplicons” as defined in the 454 Sequencing System but rather random fragments from the initial DNA sample that were amplified by long range PCR. Therefore, contrary to the experimental designs described above, data processing for the ‘Long Range PCR’ experimental design must use the Shotgun signal processing pipeline.

6.5 Data Analysis

The nature of the library format (random fragments) and the fact that data processing is done using the Shotgun pipeline make the resulting datasets unsuitable for the GS Amplicon Variant Analyzer (AVA) software. Rather, datasets of experiments made using this ‘Long Range PCR’ experimental design must be analyzed using the GS Reference Mapper software, with the same constraints as for the ‘Basic’ experimental design (Section 3.5): the HCDiffs file output of the GS Reference Mapper can be used to identify heterozygotes or mutations represented by about 20 to 100 reads (i.e. low to moderate depth; the GS Reference Mapper is not tuned to handle very high depth of coverage like the AVA software is) that represent more than 30% of the reads for that Amplicon.
7. DESIGN 5: ONE-WAY READS AMPLICON SEQUENCING

There are situations where increasing the number and length of unidirectional reads is more informative than obtaining a mixture of forward and reverse reads on an Amplicon. An example of this is metagenomics experiments that seek to identify bacterial species in a complex sample by the sequence of their 16S ribosomal RNA: obtaining more long reads from a common primer starting point will allow to interrogate more species, whereas reverse reads would provide relatively little benefit.

The ‘One-Way Reads’ Amplicon sequencing experimental design addresses this special situation. It is similar to the ‘Basic’ design but uses fusion primers made with the ‘Lib-L’ Primer A and Primer B sequences instead of the ‘Lib-A’ ones normally used for Amplicon libraries (see Section 7.1.3). The resulting Amplicons are therefore amplified using the emPCR (Lib-L) Kit and method (Figure 10).

The fusion primers are designed such that directionality of the reads is known. However, libraries prepared and amplified using the Lib-L chemistry can be sequenced only from the Primer A end, so this design provides only unidirectional sequencing. (Bidirectional sequencing could be obtained by preparing two primer pairs for each Amplicon but this would defeat the purpose of this design.)

An alternative to this design would be to simply use the ‘Basic’ experimental design (Section 3) and design the primers with the intention to process each of the libraries through emPCR amplification using only “A” or “B” beads. When doing this, users may want to design half of the libraries to be sequenced from Primer A and half to be sequenced from Primer B, in order to fully utilize the DNA Capture Beads provided in the emPCR Kit (Lib-A).
Library Preparation
- Design fusion primers with ‘Lib-L’ Primer A and B sequences (one primer pair per Amplicon per sample)\(^1\)
- Order fusion primers (design not supported by IDT)
- *Amplicon Library Preparation Method Manual*\(^2\) (except primer sequences)

emPCR Amplification
- Mix MID-labeled libraries as appropriate (typically equimolar)
- emPCR Kit (Lib-L)\(^3\)
- *emPCR Amplification Method Manual – Lib-L*\(^2,3\)

Sequencing
- GS Junior Titanium Sequencing Kit or GS FLX Titanium Sequencing Kit XLR70
- *Sequencing Method Manual*\(^2,4\)

Data Processing
- Amplicon pipeline or Shotgun pipeline, depending on the context and goal of the experiment

Data Analysis
- GS Amplicon Variant Analyzer (AVA), with special commands; GS Reference Mapper; or third-party software

Figure 10: Workflow of a ‘One-Way Reads’ Amplicon sequencing experiment.
\(^1\) May use MIDs in Adaptor A only; see Section 8.1 for the use of MIDs; \(^2\) From the appropriate version of the 454 Sequencing System; \(^3\) For emPCR amplification, see the options described in Section 7.2; \(^4\) At the time of this writing, Amplicon sequencing is supported in the GS FLX+ System only via its backward compatibility with the GS FLX System and the GS FLX Titanium Sequencing Kit XLR70.

### 7.1 Library preparation

#### 7.1.1 Kit and Method

The method for library preparation for the ‘One-Way Reads’ experimental design is very similar to the ‘Basic’ design (Section 3.1.1). It does not use any library preparation kit and consists of a simple PCR reaction using the DNA template of interest and a pair of special fusion primers, as described in the *Amplicon Library Preparation Method Manual* (Figure 11), except that it uses fusion primers prepared with the Primer A and Primer B sequences that will anneal to Lib-L Capture Beads instead of Lib-A Capture Beads (see Section 7.1.3).

As a specific primer pair must be designed for each Amplicon for each sample, this can become complex (and onerous) for experiments comprising a large number of Amplicons or samples; in such cases, one should consider applying the concept of the ‘Universal Tailed’ Amplicon Sequencing experimental design (see Section 4); this is not described in detail with the ‘Lib-L’ primers because the main application, metagenomics 16S sequencing experiments, typically comprises only one Amplicon (defined by a single primer pair).
7.1.2 Multiplexing

Considerations for multiplexing when using the ‘One-Way Reads’ experimental design are similar to those for the ‘Basic’ design (Section 3.1.2). In brief,

- Amplicons can often be differentiated by the GS Amplicon Variant Analyzer (AVA) software by virtue their template-specific sequences (but see below). These can be pooled directly before emPCR amplification (see Section 7.2). Note that this would not apply to “metagenomics 16S” and other ultra deep-type experiments, where all the targets are similar and use a common template-specific primer pair.
- Separate samples for which the same Amplicon(s) are to be sequenced in the same sequencing Run (or in the same PTP device region of a sequencing Run in the GS FLX or the GS FLX+ Systems) must be tagged by the inclusion of Multiplex Identifiers (MIDs) in the design of their fusion primers (but, contrary to the ‘Basic’ design, only on the Primer A side).

For more detail on the design of fusion primers for this experimental design, see Section 7.1.3. For more detail on the usage of MIDs as applicable to this experimental design, see Section 7.2.

7.1.3 Fusion Primer Design

The primers used to generate Amplicon libraries for the ‘One-Way Reads’ experimental design are each composed of two or three parts, fused together as shown in Figure 12.

**Forward primer (Primer A, Lib-L):**

$$5’$$-CCATCTCATCCCTGCCTGCTTGCCGACTCAG-{MID}-{template-specific sequence}-3’

**Reverse primer (Primer B, Lib-L):**

$$5’$$-CCTATCCCTGCTGCTGCTGCTGACTCAG-{template-specific sequence}-3’

Figure 12: Components of fusion primers for ‘One-Way Reads’ Amplicon sequencing

- The 5’-portion is a 30-mer whose sequence is dictated by the requirements of the 454 Sequencing System for binding to the ‘Lib-L’ DNA Capture Beads, and for annealing the corresponding emPCR Amplification Primers and the Sequencing Primer; in addition, this 5’-part must always end with the sequencing key “TCAG” used for Amplicon sequencing. There exist two kinds of such primers, termed “Primer A” and “Primer B”,

**Figure 11:** Schematic representation of the PCR reaction components for the preparation of an Amplicon library for use in the ‘One-Way Reads’ experimental design.
allowing for directional sequencing; however, the Lib-L chemistry supports sequencing only from the Primer A end. The exact required sequences are as shown in blue (Adaptors) and red (key) in Figure 12.

- The 3'-portion of each primer is designed to anneal with a specific sequence on either side of the target of interest on the DNA sample, delineating the margins of the amplicon that will be produced. This requires detailed knowledge of the target sequence, in particular the sites targeted by the primers. These template-specific sequences are typically 20-25 nt in length, though this may vary; they are shown on green or blue-green background in Figure 12.

- Optionally, Multiplex Identifiers (MIDs) can be used to “barcode” the Amplicons or samples (see Section 8.1). When used, MIDs must be placed immediately after the sequencing key. With the Lib-L Adaptors, MIDs need only be placed on the “Primer A” side. This is shown on orange background in Figure 12.

### 7.1.4 Amplicon Length Considerations

As the objective of this experimental design is to maximize the amount of information that can be derived from the set of sequencing reads, Amplicons should be designed such that they are longer than the read length that can be expected with the sequencing chemistry utilized. This way, reads will never reach the template-specific sequence at the Primer B side, maximizing the informational content of the reads.

As in the ‘Basic’ experimental design (Section 3.1.4), the sequencing key and the template-specific primer (and the MID sequence, if present) at the proximal (Primer A) end of the reads are part of the reads (included in read length) but are NOT part of the analyzable target sequence: in particular, any variants detected in the template-specific primer regions would denote errors introduced during oligonucleotide primer synthesis (which are common), not biologically relevant sequence variants.

### 7.1.5 Fusion Primer Ordering

Fusion primers can be ordered from Integrated DNA Technologies, Inc. (IDT, www.idtdna.com). However, the fusion primer design tool available on the IDT web site does not support primers containing the Lib-L Adaptor Primer A and Primer B sequences.

### 7.2 emPCR Amplification

Like the ‘Ligated Adaptors’ (Section 5.2) and the ‘Long Range PCR’ (Section 6.2) designs, the emPCR amplification for the ‘One-Way Reads’ experimental design uses the emPCR Kit (Lib-L), but the options are otherwise the same as for the ‘Basic’ design (Section 3.2.1). Also, if MIDs were used to tag different samples (Primer A side only), these can be pooled following the usual pooling strategies that apply to multiplexed Rapid libraries, plus those described for the ‘Basic’ Amplicon sequencing experimental design (Section 3.2.2), to obtain the depth of coverage required for the experiment.
7.3 Sequencing

The kit and method for sequencing and the library pooling considerations for the ‘One-Way Reads’ experimental design are the same as for the ‘Basic’ design (Section 3.3).

7.4 Data Processing

Data processing considerations for the ‘One-Way Reads’ experimental design depend on the experimental goals. In some cases, use of the Amplicon signal processing pipeline as with the ‘Basic’ design (Section 3.4) may be appropriate (perhaps after applying custom filtering parameters). By contrast, in a 16S-type experiment where the number of reads is most important, it can be advantageous to select the Shotgun data processing scheme, or a modification of the Amplicon data processing scheme that allows for 3’ trimming of the reads: if a read suffers from poor quality near its end, these options will trim the read rather than discarding the read altogether as the default Amplicon pipeline would. Thus, these will tend to provide more reads, though with somewhat shorter average read length. See Section 8.4 for more details on these options.

7.5 Data Analysis

Data analysis considerations for the ‘One-Way Reads’ experimental design are the same as those for the ‘Ligated Adaptors’ design (Section 5.5): use the GS Amplicon Variant Analyzer (AVA) software. However, the AVA software is not tuned by default to process Amplicons that contain the ‘Lib-L’ Primer A and Primer B sequences. Also, usage of the “Lib-L” Adaptors requires a currently undocumented feature of the GS Amplicon Variant Analyzer (AVA) software. Contact your Roche Technical Support Representative for help on how to use the AVA software to analyze the data from an experiment performed using the ‘One-Way Adaptors’ design.

For metagenomics 16S-type experiments, the reads will typically be “blasted” against a database to identify which species were present. Other tools are also available on the internet, such as MG-RAST, which accept 454 Sequencing System data and can perform extensive analyses of these data sets.
8. SPECIAL TOPICS FOR AMPLICON SEQUENCING

8.1 Usage of Multiplex Identifiers (MIDs) in Amplicon Sequencing (Lib-A)

The information in this section is related primarily to the GS Amplicon Variant Analyzer (AVA) software. This software is described in detail in the *454 Sequencing System Software Manual, Part D*.

8.1.1 When To Use MIDs

The GS Amplicon Variant Analyzer (AVA) software provides a number of mechanisms for demultiplexing reads, allowing multiple Amplicons from the same or different Samples to be sequenced simultaneously on a PTP device (or, for the GS FLX or GS FLX+ Systems, within a region of a PTP device). The simplest demultiplexing method exploits the template-specific primer regions of the Adaptors used to prepare the library to identify the Amplicons. The Amplicon library preparation method places these sequences at the beginning of the reads, just after the sequencing key (which is part of Primers A and B). If an experiment calls for measuring multiple distinct Amplicons from the same Sample, those Amplicons may be pooled and sequenced together, and the Project can be set up such that reads of the various Amplicons are associated with the appropriate Sample by virtue of their known template-specific primer sequences.

But with the large number of sequencing reads that can be obtained in a single PTP device (or region thereof) in the 454 Sequencing System, the situation is common whereby a single region would produce a vast excess of reads compared to what is necessary for any given Amplicon library (Sample). If the experiment includes multiple Samples, the obvious economical solution would be to load multiple Samples in each region such that each Sample will be covered at the appropriate depth, in a single sequencing Run. However, if the same Amplicon, or set of Amplicons, are to be sequenced in several Samples (a very common situation), then the common Amplicons’ template-specific primer sequences would not allow to assign each read to the proper Sample.

In such cases, the experimental setup must include Multiplex Identifiers (MIDs). MIDs are short, recognizable sequence tags that can be added to the design of the fusion primers used for library preparation; they are placed between the sequencing key and the template-specific primer (see, for example, Figure 3). Multiple Amplicon libraries (the Project’s Samples) can be prepared that include the same Amplicon target sequences (with the same template-specific primers), each labeled with different MID tags.

Certain Amplicon sequencing experiments can also be analyzed using the GS Reference Mapper software (see Sections 3 to 7). This software also can recognize MID sequences, but it does not address the notion of “Sample” like the AVA does, so it cannot assign individual reads to samples; it can, however, process selectively the reads of a dataset that contain only one (or more) given MIDs.
8.1.2 Sample Encoding Schemes When Using MIDs

Contrary to the situation with library types that are amplified with the emPCR Kit (Lib-L), where an MID sequence can only be used on Adaptor A (reads are not produced from the Adaptor B end), Amplicon libraries prepared using the emPCR Kit (Lib-A) can be constructed with MIDs at either or both ends of the reads. This provides for considerable flexibility in the design of MID-tagged Amplicon libraries. Indeed, the AVA software recognizes four ways to encode Samples using MIDs, through the use of “Multiplexers”. This is described in detail in the 454 Sequencing System Software Manual, Part D. In brief the encoding types are:

- **Primer 1 MID**: This encoding provides an MID signature only on the end of the read that contains the template-specific primer defined as “Primer 1” in the Project. This will be at the beginning of the “forward” reads, or at the end of “reverse” (complemented) reads. For the reverse reads to be assigned to their Samples, therefore, the Amplicons encoded using the “Primer 1 MID” scheme must be shorter than the read length of the sequencing Run.

- **Primer 2 MID**: This encoding is the same as Primer 1 MID encoding, except that the MID appears at the “Primer 2” end of the Amplicons.

- **Both**: This encoding provides MIDs at both ends of the Amplicons and requires that read length be sufficient to read through to the distal MID, in both orientations. The paired combination of MIDs located on the Primer 1 and Primer 2 sides is used to assign reads to their proper Sample, as defined by the Multiplexer, such that a small number of MID tags can encode a much larger number of Samples.

- **Either**: This encoding also provides MIDs at both ends of the Amplicons, but assigns the reads to their proper Sample on the basis of only the proximal MID on the read, whatever its orientation. This allows for proper assignment of both forward and reverse reads even if the Amplicon is longer than the read length provided by the sequencing Run script. Note that even if full read-through to the distal end of the read is possible, only the proximal MID will be used for Sample assignment (and any contradiction between the MIDs seen at the two ends will be assumed to be the effect of sequencing artifacts at the distal end of the read).

Selecting the proper encoding: It is crucially important to select the encoding method that truly corresponds to the way the libraries were prepared. For example, if a library was prepared with the ‘Either’ chemistry in mind, it may be tempting to use a ‘Primer 1 MID’ or ‘Primer2 MID’ encoded Multiplexer since the distal MID gets discounted in favor of the proximal MID, in ‘Either’ encoding. However, the AVA software needs to know whether or not MIDs are expected to be found at both ends: without that knowledge, the trimmer might get a suboptimal alignment of the distal primer, which in certain cases could drop valid reads out of the analysis.

8.1.3 The 454Standard MID Set

While any sequence can be used as MID, we have designed a set of 14 ten-base MID sequences which are carefully engineered to avoid mis-assignment of reads and are tolerant to several types of errors, including those often introduced during primer synthesis. This is called the 454Standard MID set, and their use is strongly recommended when designing MID-tagged Amplicon libraries that will be analyzed using the AVA software.

The MIDs of the 454Standard MID set have been especially selected for the following qualities:
• Their sequences are as divergent as possible and include only single nucleotide homopolymers, minimizing the risk that eventual sequencing errors would make any of these MIDs look like one of the other MIDs of the Group (which would cause the assignment of the read to a wrong Sample). Indeed, no fewer than six changes (insertions, deletions, and/or substitutions) separate any two MIDs of the 454Standard MID Group.

• All MIDs of the Group are of the same length (10-mers, in this case). The AVA software requires that all the MIDs used for a given end of the Amplicons (Primer A/Primer 1 or Primer B/Primer 2) have the same length, so they are on equal footing for error distance calculation purposes. It is permitted, however, to use a different MID length on each side of the Amplicons being handled by a Multiplexer. For instance, custom 5 bp MIDs might be used on the Primer 1 side and 10 bp MIDs might be used on the Primer 2 side.

• None start with a “G” nucleotide, the last base of the sequencing key, to ensure clear reading of the MID tag.

• All MIDs are read within 4 or 5 nucleotide flow cycles, to minimize usage of Run flows to read the MID tags and leave as much as possible for the sample sequence.

Note that the AVA software does not constrain the user to only these 14 MIDs: any other set of sequences can be designed for use as multiplexing tags, incorporated between the sequencing key and the template-specific primer of the Adaptors used to prepare the Amplicon libraries, and defined as MIDs in the Amplicon Project. This flexibility can be useful, for example, if the user prefers to use shorter MIDs (leaving more nucleotide cycles of the sequencing Run for the target sequence); or if Amplicon libraries already exist that have intrinsic sequences that can be used for demultiplexing; or if the experiment requires the multiplexing of more Samples than can be differentiated with the 14 MIDs of the 454Standard MID Group (with the “Both” encoding, the 14 MIDs of the 454Standard MID set can tag 196 different libraries). If you design your own MIDs, it is recommended that you keep in mind the 4 criteria listed above, for best results.

For your convenience, these 14 MID sequences come pre-loaded in the GS Amplicon Variant Analyzer software (AVA), as an MID Group named “454Standard”. Their exact sequences are shown in Table 2.

<table>
<thead>
<tr>
<th>ID</th>
<th>MID Sequence</th>
<th>ID</th>
<th>MID Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>MID1</td>
<td>ACGAGTGCGT</td>
<td>MID8</td>
<td>CTCGCGTGTC</td>
</tr>
<tr>
<td>MID2</td>
<td>ACGCTCGACA</td>
<td>MID9</td>
<td>TAGATCAGC</td>
</tr>
<tr>
<td>MID3</td>
<td>AGACGCACCTC</td>
<td>MID10</td>
<td>TCTCTATGCGC</td>
</tr>
<tr>
<td>MID4</td>
<td>AGCACTGTAG</td>
<td>MID11</td>
<td>TGATACGTCT</td>
</tr>
<tr>
<td>MID5</td>
<td>ATCAGACACG</td>
<td>MID12</td>
<td>TACTGAGCTA</td>
</tr>
<tr>
<td>MID6</td>
<td>ATATCAGCGG</td>
<td>MID13</td>
<td>CATAGTAGTG</td>
</tr>
<tr>
<td>MID7</td>
<td>CGTGTCTCTTA</td>
<td>MID14</td>
<td>CGAGAGATAC</td>
</tr>
</tbody>
</table>

Table 2: Sequences of the 14 MIDs of the 454Standard MID set

Note that sequences labeled MID1 through MID12 in this set are DIFFERENT from the sequences labeled similarly (MID RL001 through MID RL012) and available as full double-stranded Adaptors in the GS FLX Titanium Rapid Library MID Adaptors Kit. These MID-tagged Adaptors are used to multiplex Rapid libraries or other libraries prepared using the emPCR Kit (Lib-L), including Amplicon libraries prepared using the ‘Ligated Adaptors’, ‘Long Range PCR’ and ‘One-Way Reads’ experimental designs described in this document (see Section 5, 6, and 7).
For experiments that require more than 14 MIDs, Roche has designed an extended second-tier set of 151 MIDs. These are described in TCB No. 005-2009.

8.2 Short and Long Amplicons

The conditions of emPCR amplification described in the *emPCR Amplification Method Manual – Lib-A* are optimized to successfully amplify targets between 200 and 600 bp in length. If the specifics of the experiment require the production of Amplicons whose length is outside this range, the modifications to the emPCR amplification procedures described below can be used to improve the results. If the experiment requires the pooling of amplicons of various lengths, see Section 8.3.

8.2.1 Sequencing Short-Length Amplicon Libraries

If an Amplicon library is very short (e.g. Amplicons <200 bp), the normal conditions for emPCR amplification may result in excess amplification. An excess of sequencing templates on the DNA beads would cause an increased consumption of nucleotides during the sequencing Run which may result in incomplete extension events; also, the increased signal may result in light scattering into nearby wells (well crosstalk) and cause their elimination due to signal processing filtering. To prevent this, the emPCR amplification procedure can be modified for short Amplicon libraries such that the amount of Amp Primer (A or B) is reduced by 75% in the Live Amp Mix (and this volume replaced with Molecular Biology Grade Water).

The exact volume adjustments will depend on the specifics of the experiment. For the GS Junior System, for example, the volume of Live Amp Mix to prepare is different between Lib-A and Lib-L preparations. For the Lib-A method, the amount of Amp Primer in the mix should be reduced from 40 μl to 10 μl (see the *emPCR Amplification Method Manual – Lib-A, GS Junior Titanium Series*); for the Lib-L method, e.g. as used for the ‘Ligated Adaptors’ experimental design, the amount of Amp Primer in the mix should be reduced from 80 μl to 20 μl (i.e. use the conditions prescribed for Paired End libraries; see the *emPCR Amplification Method Manual – Lib-L, GS Junior Titanium Series*). In all cases, the volume reduction in the mix should be compensated with molecular biology grade water. Overall, the composition of the Live Amp Mixes for Amplicon sequencing experiments using the GS Junior System where the Amplicons are <250 bp in length should be as shown in Table 3.

For the GS FLX and the GS FLX+ Systems, the volumes vary not only between the Lib-L and Lib-A setups (as for the GS Junior System) but also between the LV, MV and SV protocols, and some of the protocols even specify different volumes of Live Amp Mix to prepare for different PTP device region sizes. In those cases, refer to the corresponding Method Manuals and apply the same 75% reduction in Amp Primer and volume replacement with molecular biology grade water, for the preparation of the Live Amp Mixes. However, note that as of this writing, Amplicon sequencing is supported in the GS FLX+ System only via its backward compatibility with the GS FLX System and the GS FLX Titanium Sequencing Kit XLR70.
A: For the Lib-A method (GS Junior System)

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Volume (μl)</th>
<th>Live Amp Mix A</th>
<th>Volume (μl)</th>
<th>Live Amp Mix B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive</td>
<td>260</td>
<td></td>
<td>Additive</td>
<td>260</td>
</tr>
<tr>
<td>Amp Mix</td>
<td>135</td>
<td></td>
<td>Amp Mix</td>
<td>135</td>
</tr>
<tr>
<td>Amp Primer A</td>
<td>10</td>
<td></td>
<td>Amp Primer B</td>
<td>10</td>
</tr>
<tr>
<td>Enzyme Mix</td>
<td>35</td>
<td></td>
<td>Enzyme Mix</td>
<td>35</td>
</tr>
<tr>
<td>PPIase</td>
<td>1</td>
<td></td>
<td>PPIase</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>676</strong></td>
<td></td>
<td><strong>Total:</strong></td>
<td><strong>676</strong></td>
</tr>
</tbody>
</table>

B: For the Lib-L method (GS Junior System)

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Volume (μl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mol. Bio. Grade Water</td>
<td>470</td>
</tr>
<tr>
<td>Additive</td>
<td>515</td>
</tr>
<tr>
<td>Amp Mix</td>
<td>270</td>
</tr>
<tr>
<td>Amp Primer</td>
<td>20</td>
</tr>
<tr>
<td>Enzyme Mix</td>
<td>70</td>
</tr>
<tr>
<td>PPIase</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>1347</strong></td>
</tr>
</tbody>
</table>

Table 3: Composition of the Live Amp Mixes for the sequencing of short Amplicons using the GS Junior System, showing reduced amount of Amp Primer and correspondingly increased molecular biology grade water. A: Live Amp Mixes A and B for the Lib-A emPCR amplification method; B: Live Amp Mix for the Lib-L emPCR amplification method.

8.2.2 Sequencing Long-Length Amplicon Libraries

If, on the contrary, an Amplicon library is long (e.g. Amplicons >550 bp), the standard emPCR amplification conditions may not allow for efficient full extension of the PCR primers during each cycle, resulting in very low yield. To prevent this, the emPCR amplification procedure can be modified for long Amplicon libraries by adjusting the composition of the Live Amplification Mix as well as the thermocycling conditions as follows:

- increase the amount of 5x Amplification Mix by about 10%;
- increase the amount of Amplification Primer (A or B) by about 30%;
- reduce the volume of Molecular Biology Grade Water by the corresponding amounts; and
- modify the thermocycling steps as follows:
  - 1x 4 minutes at 94°C
  - 50x 30 seconds at 94°C, 10 minutes at 60°C
  - 1x 10°C on hold
The exact volume adjustments will depend on the specifics of the experiment. For the GS FLX and the GS FLX+ Systems, the volumes vary not only between the Lib-L and Lib-A setups but also between the LV, MV and SV protocols, and some of the protocols even specify different volumes of Live Amplification Mix to prepare for different PTP device region sizes. For the Lib-A method using the SV protocol, for example, the amount of 5x Amplification Mix reagent in the mix for 16 emulsions should be increased from 780 µl to 860 µl; and the amount Amplification Primer should be increased from 230 µl to 300 µl (see the emPCR Amplification Method Manual – Lib-A SV, GS FLX Titanium Series); for the Lib-L method, e.g. as used for the 'Ligated Adaptors' experimental design, the amount of 5x Amplification Mix reagent in the mix for 16 emulsions in the SV protocol should be increased from 840 µl to 925 µl; and the amount of Amplification Primer should be increased from 300 µl to 390 µl; see the emPCR Amplification Method Manual – Lib-L SV, GS FLX Titanium Series). In all cases, the volume increase in the mix should be compensated with molecular biology grade water.

Overall, the composition of the Live Amplification Mixes for Amplicon sequencing experiments using the SV protocol in the GS FLX or GS FLX+ Systems where the Amplicons are >550 bp in length should be as shown in Table 4. For other cases, refer to the corresponding Method Manuals and apply the same percentage reagent increase and volume replacement with molecular biology grade water, for the preparation of the Live Amplification Mixes; and modification of the PCR conditions.

Note, however, that while this should in theory apply to all versions of the 454 Sequencing System, there is not enough reagents in the GS Junior Titanium emPCR Amplification Kits (single prep procedure). Therefore, when applied to the GS Junior System, this modified procedure for the amplification of long amplicons is supported only when the GS FLX Titanium MV emPCR Kit is used (multiple prep procedure). Note also that as of this writing, Amplicon sequencing is supported in the GS FLX+ System only via its backward compatibility with the GS FLX System and the GS FLX Titanium Sequencing Kit XLR70.
A: For the Lib-A method, SV protocol, 16 emulsions (GS FLX or GS FLX+ System)

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Volume (μl)</th>
<th>Reagent</th>
<th>Volume (μl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive</td>
<td>1500</td>
<td>Additive</td>
<td>1500</td>
</tr>
<tr>
<td>5x Amplification Mix</td>
<td>860</td>
<td>5x Amplification Mix</td>
<td>860</td>
</tr>
<tr>
<td>Amplification Primer A</td>
<td>300</td>
<td>Amplification Primer B</td>
<td>300</td>
</tr>
<tr>
<td>Enzyme Mix</td>
<td>200</td>
<td>Enzyme Mix</td>
<td>200</td>
</tr>
<tr>
<td>PPiase</td>
<td>5</td>
<td>PPiase</td>
<td>5</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>3915</strong></td>
<td><strong>Total:</strong></td>
<td><strong>3915</strong></td>
</tr>
</tbody>
</table>

Table 4: Composition of the Live Amplification Mixes for the sequencing of long Amplicons using the GS FLX or GS FLX+ Systems, showing increased amounts of 5x Amplification Mix and Amplification Primer and correspondingly decreased amounts of molecular biology grade water. A: Live Amplification Mixes A and B for the Lib-A emPCR amplification method; B: Live Amplification Mix for the Lib-L emPCR amplification method. Make sure to also use the adjusted thermocycling conditions when using this modified procedure. Note that Amplicon sequencing is currently supported in the GS FLX+ System only via its backward compatibility with the GS FLX System and the GS FLX Titanium Sequencing Kit XLR70, and that this modified procedure for the emPCR amplification of long amplicons is supported in the GS Junior System only when the GS FLX Titanium MV emPCR Kit is used (multiple prep procedure).

B: For the Lib-L method, SV protocol, 16 emulsions (GS FLX or GS FLX+ System)

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Volume (μl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mol. Bio. Grade Water</td>
<td>785</td>
</tr>
<tr>
<td>Additive</td>
<td>1440</td>
</tr>
<tr>
<td>5x Amplification Mix</td>
<td>925</td>
</tr>
<tr>
<td>Amplification Primer</td>
<td>390</td>
</tr>
<tr>
<td>Enzyme Mix</td>
<td>200</td>
</tr>
<tr>
<td>PPiase</td>
<td>8</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>3748</strong></td>
</tr>
</tbody>
</table>

8.3 Pooling Mixed Length Amplicons

It is well documented that PCR (both in solution and solid-phase) favors the amplification of smaller fragments over longer ones, in a complex mixture of different length templates. This differential amplification may need to be taken into consideration in order to maximize the sequencing yield and sample preparation efficiency when preparing pooled Amplicons spanning a significant length range, e.g. wider than 150 bp (whether or not MIDs are used to tag them).

Pooling mixed-length Amplicons through emPCR amplification and sequencing can have negative effects in the following two ways:

- Although the emPCR amplification process is less prone to length bias than ordinary PCR, the DNA library beads derived from Amplicons of different length can still result in a different amount of amplified target and, consequently, in a wider signal distribution, potentially lowering the sequencing Run yield.
- If two templates of different sizes are present in any given emulsion droplet, the shorter fragment will amplify preferentially over the longer one. This will result in either an increase in mixed reads (which will be discarded by the data processing filters), or an over representation of the shorter fragments in the sequencing results; in both cases, this would introduce a potential sequencing bias and a reduction in Run yield.

Four approaches can be taken to minimize or avoid these biases:

1. **Amplicon Design:** If possible, design (or redesign) Amplicons closer in length, e.g. within approximately 150 bp of each other. Amplicons with such a narrow range of lengths can be efficiently processed together in a single emPCR emulsion, following the normal emPCR amplification methods.

2. **Weighted Amplicon Library Mixture:** Amplicon libraries of different lengths can be combined at different ratios during emulsion set-up, to counterbalance the effect of length on amplification efficiency. Though the optimal balance will have to be determined experimentally, a good starting point would be to add four times more of the longer Amplicons than the shorter ones.

3. **Separate Libraries and emPCR Amplifications:** Amplicon libraries of different lengths can be emulsified and amplified separately (or in separate pools where each pool contains only Amplicons of similar lengths), and then pooled for emulsion breaking, enrichment, and sequencing. This can be particularly useful when the Amplicons (or pools of Amplicons) are to be processed using different emPCR amplification conditions, e.g. some using the modified procedures for short or long Amplicons described in section 8.2. Although this will still result in a wider signal distribution than if all Amplicons sequenced together were of similar lengths (first bullet, above), it will ameliorate the second mechanism (subpopulation of mixed reads) that can reduce throughput and introduce a representativity bias.

4. **Fully Separate Library Preparation and Sequencing:** All Amplicon libraries of an experiment can be prepared and completely processed through emPCR amplification individually, and then sequenced on separate PTP devices or in separate PTP regions. Again, this allows the use of different emPCR amplification conditions for different libraries, e.g. if some should use the modified procedures for short or long Amplicons described in section 8.2. This will tend to reduce the effects of both mechanisms known to lower the throughput (bullets, above), but is more onerous and labor-intensive.

### 8.4 Special Amplicon Quality Filtering Strategies

The default signal processing for Amplicons using the current 454 Sequencing System software contains several assumptions that result in very high stringency (high quality) results. In particular, reads that fall below a certain signal quality threshold before reaching the 3’-end of the amplicon will not be considered as candidate High Quality reads and will be rejected. These assumptions were specifically designed for applications employing bidirectional, overlapped Amplicon sequencing (up to ~400 bp long). However, other Amplicon experimental designs may benefit from different processing conditions than these defaults.

Guidance for certain such cases is provided in the sub-sections below. The two conceptual strategies they exploit can be summarized as:

- Allowing for trimming of 3’ bases
- Specifying an acceptable “error” profile along a longer portion of the reads
Note that these strategies only affect the filtering of the High Quality Reads and do not change the algorithm for base-calling. Refer to Section 1 of the 454 Sequencing System Software Manual, Part B, for background information on data processing pipelines.

8.4.1 Allowing For Trimming of 3’ Bases

8.4.1.1 Use the Shotgun Data Processing Pipeline

The user can choose the Shotgun signal processing pipeline for Amplicon experimental designs. Given that sequencing accuracy decays with read length, this can be useful because the upper level assumption with the Shotgun pipeline is that it allows for trimming of the bases from the 3’ end of the reads until an error rate of ~1% is observed. This results in reads of varying lengths based on the quality of the sample, design, and sequencing, whereas the more stringent Amplicon pipeline may have just rejected many of those reads altogether.

This alternative method should be considered for experiments comprising long (>400 bp) complex Amplicon pools (e.g. >500 unique Amplicons). Complex pools of Amplicons benefit from this scheme because it circumvents the application of a clustering algorithm specific to the Amplicon signal processing pipeline but not used by the Shotgun pipeline. The clustering algorithm improves base-calling accuracy when few unique targets are generating signal in unison, but complex Amplicon pools are more akin to the context of Shotgun sequencing.

Note, however, that data sets produced by the Shotgun pipeline are not ideal for analysis using the GS Amplicon Variant Analyzer software (AVA) and therefore the strategy outlined in Section 8.4.2 may be better if AVA will be used.

8.4.1.2 Modify the Amplicon Data Processing Pipeline

This hybrid method uses the Amplicon data processing pipeline with modified settings to allow for a shotgun-like trimming at the 3’ end of the reads. This may be preferred over the simple use of the Shotgun pipeline for experiments that lack the complexity of a large, highly diverse pool of amplicons. Smaller pools of Amplicons processed in this way will still gain the benefit of the amplicon-specific clustering algorithm and of the automatic trimming of amplicon adapters performed by Amplicon signal processing; and will also be subject to 3’ read-trimming. Again, this will result in reads of varying lengths based on the quality of the sample, design, and sequencing, whereas the more stringent Amplicon pipeline may have just rejected many of those reads altogether.

Note also that, similarly to data sets generated by the Shotgun pipeline, data sets produced in this way are not ideal for analysis using the GS Amplicon Variant Analyzer software (AVA) and therefore the strategy outlined in Section 8.4.2 may be better if AVA will be used.

To activate 3’ read trimming for the Amplicon pipeline, do the following:

1. Open a command line terminal and change to the run result directory for the run of interest – either the D_...signalProcessingAmplicons or D_...fullProcessingAmplicons directory, as appropriate.
2. Execute the following command:

```
gsRunProcessor --template=filterOnlyAmplicons > MyAmpliconsConfig.xml
```
3. Edit the `MyAmpliconsConfig.xml` file with a text editor such as `nedit`:

```
nedit MyAmpliconsConfig.xml
```

4. Under the `<qualityFilter>` section, the parameter that controls 3’ trimming is called `vfScanAllFlows`. Below is the section of the `MyAmpliconsConfig` file that defines the values for this parameter:

```
<vfScanAllFlows>tiOnly</vfScanAllFlows>
```

5. Change the value to “false”:

```
<vfScanAllFlows>false</vfScanAllFlows>
```

6. Save the file and exit.

7. Change to the parent directory (one directory up) of the current `D_...` folder:

```
cd ..
```

8. Rerun the quality filter to start the reprocessing on the previous base-called results:

```
runAnalysisFilter --pipe=/Path/to/MyAmpliconsConfig.xml /Path/to/D_folder
```

### 8.4.2 Specifying an Acceptable “Error” Profile Along a Longer Portion of the Reads

This strategy should be considered when fewer than ~500 unique but very long Amplicons (e.g. >550 bp) are sequenced together on a PTP device (or, for the GS FLX or GS FLX+ Systems, in a single region of a PTP device). It uses the Amplicon signal processing assumptions (clustering) but allows the user to set a constant allowable “error” profile across a longer portion of the reads (including their complete length). To accomplish this, edit the XML file according to the directions below:

1. Open a command line terminal and change to the run result directory for the run of interest – either the `D_...signalProcessingAmplicons` or `D_...fullProcessingAmplicons` directory as appropriate.
2. Execute the following command:

```
gsRunProcessor --template=filterOnlyAmplicons > MyAmpliconsConfig.xml
```

3. Edit the `MyAmpliconsConfig.xml` file with a text editor such as `nedit`:

```
nedit MyAmpliconsConfig.xml
```

4. Under the `<qualityFilter>` section, there are two parameters that can affect the quality and quantity of output data, `vfTrimBackScaleFactor` and `vfScanLimit`. Below is the section of the `MyAmpliconsConfig` file that defines the values for these parameters:

```
<vfScanLimit>X</vfScanLimit>
<vfTrimBackScaleFactor>Y</vfTrimBackScaleFactor>
```
5. If the data set contains short reads or if the experimental design would bring the expectation that the 3’ end of the reads will be of lower quality, the valley filter can be constrained such that the end of the read is not counted against the valley filter. The parameter, \texttt{vfScanLimit}, adjusts the final flow over which the valley filter operates. Decreasing this parameter reduces the scope of the valley filter and will allow more reads with high quality at their 5’ end, but may have low quality after the flow representing the \texttt{vfScanLimit} is reached. Adjust quality filter stringency by changing this parameter (highlighted \texttt{Z}). The units for this parameter are in flows so setting it at the total number of flows in the sequencing Run will apply the filter along the entire length of all reads. The effects of this parameter are summarized in Table 5.

<table>
<thead>
<tr>
<th>\texttt{vfScanLimit}</th>
<th>Quality Filter Stringency</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textbf{\downarrow} DECREASE</td>
<td>\textbf{\downarrow} LOWER (more reads, some may have lower quality 3’ ends)</td>
</tr>
<tr>
<td>\textbf{\uparrow} INCREASE</td>
<td>\textbf{\uparrow} HIGHER (fewer reads, more consistent quality across the read length)</td>
</tr>
</tbody>
</table>

Table 5: Effect of modifications to the \texttt{vfScanLimit} parameter on the stringency of read filtering

6. The second parameter, \texttt{vfTrimBackScaleFactor} adjusts the stringency of the valley filter scoring system. Increasing this number penalizes ambiguous flows more severely, reducing the amount of data being generated, but increasing its overall quality. Note that the word \texttt{vfTrimBackScaleFactor} appears twice in the template file, once for runs performed on GS Junior Instruments and once for others. For datasets acquired on the GS Junior Instrument, change the \texttt{vfTrimBackScaleFactor} by changing this parameter inside the \texttt{<if>} block (highlighted \texttt{Z}); for datasets acquired on another 454 Sequencing System Instrument, change the \texttt{vfTrimBackScaleFactor} by changing this parameter on the line above the \texttt{<if>} block (highlighted \texttt{D}). A change of ±0.5 would be considered a moderate increase or decrease, while a change of ±1.0 would be considered a significant increase or decrease. The effects of this parameter are summarized in Table 6.

<table>
<thead>
<tr>
<th>\texttt{vfTrimBackScaleFactor}</th>
<th>Quality Filter Stringency</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textbf{\downarrow} DECREASE</td>
<td>\textbf{\downarrow} LOWER (more, lower quality reads)</td>
</tr>
<tr>
<td>\textbf{\uparrow} INCREASE</td>
<td>\textbf{\uparrow} HIGHER (fewer, higher quality reads)</td>
</tr>
</tbody>
</table>

Table 6: Effect of modifications to the \texttt{vfTrimBackScaleFactor} parameter on the stringency of read filtering

7. Save the file and exit.
8. Change to the parent directory (one directory up) of the current \texttt{D...} folder:

\texttt{cd ..}

9. Rerun the quality filter to start the reprocessing on the previous base-called results:

\texttt{runAnalysisFilter --pipe=/Path/to/MyAmpliconsConfig.xml /Path/to/D_folder}
8.5 Adaptor Removal Stringency

It is important to remove any excess Adaptors from the library before emPCR amplification, for several reasons:

- Remaining Adaptors will bind to the Capture Beads during the preparation of the emPCR reaction, competing against library fragments.
- Adaptors may form dimers which would be very good (short) templates for amplification and could overwhelm the emPCR reaction and could end up representing a sizable fraction of the reads produced by the sequencing Run.
- Reads resulting from Adaptor dimers tend to generate high signals, resulting in increased crosstalk between wells, during the sequencing reaction.
- If Adaptor dimers are present in the emPCR reaction, they will be included in the quantitation of the library; those quantitation results will lead the user to include less of the library than intended in the sequencing reaction, which will skew the quantitative aspect of the Amplicon sequencing experiment.
- Skewed quantitation of the Amplicon libraries could affect not only the absolute quantitative aspect of the experiment, but also the relative representation of any pooled libraries, if the extent of the problem varies between libraries.

In all library preparation methods of the 454 Sequencing System, small molecular species such as free Adaptors and Adaptor dimers, are removed using AMPure XP beads (Beckman Coulter). In this procedure, the size cut-off is determined by the volume ratio of AMPure XP bead suspension to DNA solution: a lower ratio results in a higher size cut-off, i.e. a more stringent removal of the smaller size molecules.

In the Amplicon library preparation method described in the Amplicon Library Preparation Method Manual, this ratio is set to 1.6 : 1 (72 μl of AMPure XP beads to 45 μl of DNA solution). This was determined as a compromise to minimize the risk of removing Amplicons designed in the lower acceptable size range (around 200 bp; see Section 1.1). With proper technique (e.g. careful removal of ALL the supernatant after pelleting the AMPure XP beads), this should work properly in most cases; but if residual Adaptors are a problem in your experiment, modifying the AMPure XP purification step of the library preparation method, as described below, may prove helpful.

8.5.1 When to use this

If the Amplicon sequencing experiment contains only Amplicon near the "optimal" size for the GS FLX Titanium or GS Junior Titanium chemistry of 400 bp, the compromise for shorter Amplicons described above is unnecessary, and more stringent Adaptor removal could be used, for more safety. Also, users may simply desire to be more conservative, or their experience may have shown that the default AMPure XP procedure described in the Method Manual occasionally leads to excess residual Adaptors in their experiments.

The main symptom of the presence of Adaptor dimers in a library is observed only in the final sequencing results, as the presence of a large number of very short reads going straight from the forward Adaptor sequence into the reverse Adaptor sequence. To avoid the possible waste of a full sequencing experiment, users may elect to carry out a quality control (QC) test on the Amplicon libraries prior to emPCR amplification (Section 8.5.2); if residual Adaptors are observed in the QC test, they can be removed using one of the alternative AMPure XP purification strategies described in Section 8.5.3, before proceeding with the rest of the
8.5.2 To Carry Out a Residual Adaptors QC Test

There are many ways to test a library for the presence of small fragments (Adaptors). Two recommended methods for which the required equipment is specified in the Tables of Materials document are:

- Use of the Agilent Bioanalyzer 2100
- Use of the FlashGel System

In either case, the QC test consists in looking for small molecular species, roughly in the 60 to 120 bp range. This assumes that each Adaptor is composed of:

<table>
<thead>
<tr>
<th>Component</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primer A or Primer B sequence</td>
<td>29 bp</td>
</tr>
<tr>
<td>MID sequence</td>
<td>10 bp</td>
</tr>
<tr>
<td>Template-specific sequence</td>
<td>20-25 bp</td>
</tr>
<tr>
<td><strong>Total Adaptor length</strong></td>
<td>~60 bp</td>
</tr>
</tbody>
</table>

…and that the Adaptors may dimerize. If a peak in this size range is observed in the QC test, one (or more) of the three strategies described in Section 8.5.3, below, should be carried out to remove residual primers.

This QC test can be carried out “in line”, during the preparation of the library, right after the AMPure XP purification step. Alternatively, a systematic optimization of the AMPure XP beads to DNA ratio can be done upfront by titrating the amount of AMPure XP beads and carrying out the QC test on each test point. This upfront optimization work may be appropriate for primer sets that are to be used across many experiments.

A general approach to carry out such an AMPure XP titration QC test would be as follows:

1. Prepare several reactions (e.g. four to six) per the “Amplicon Preparation (PCR)” step of the Amplicon library preparation method (Section 3.1 of the Amplicon Library Preparation Method Manual).
2. Carry out the “Library Purification” step (Section 3.2 of the Method Manual) on the different reactions with decreasing ratios of AMPure beads to DNA solution, e.g. 1.6 : 1, 1.4 : 1, 1.2 : 1, and 1.0 : 1 (72, 63, 54, and 45 μl of AMPure beads suspension for each tube of 45 μl of Amplicon DNA solution).
3. Analyze the purified Amplicon libraries from each condition using the Bioanalyzer or a FlashGel.
4. The optimal ratio of AMPure beads to DNA solution is the lowest ratio (i.e. the most stringent to remove smaller contaminants) that does not remove any of the desired length Amplicons.
The actual size cut-off for any given AMPure beads to DNA ratio is subjected to some lot to lot variability. In library preparation methods that use the Rapid library preparation method, this variability is addressed by the use of the Sizing Solution, available in the kit. Since there is no kit for the preparation of Amplicon libraries and the Sizing Solution is thus not available in these cases, it is advisable to carry out a calibration of any new lot of AMPure XP beads. A more complete procedure for the removal of short fragments, including the AMPure XP beads calibration procedure, is provided in TCB No. 2011-002.

8.5.3 Strategies to Modify the Stringency of AMPure XP Purification

There are three main strategies to modify the stringency of AMPure XP purification, for optimal removal of residual Adaptors. These strategies can also be used in combination.

1. Decrease the ratio AMPure XP beads to Amplicon DNA library, to 1 : 1 (i.e. add 45 μl of AMPure XP beads to 45 μl of DNA). This will remove a distribution of fragments from around 300 bp and smaller and will be very effective in removing free Adaptors and Adaptor dimers. Obviously, this strategy should NOT be applied if the Amplicons of the experiment are 300 bp in length or shorter, but it would work quite well for Amplicons designed to be in the optimal size range (~ 400 bp).

2. Perform a second round of AMPure XP purification (both rounds back to back, before the emPCR amplification). This strategy is the most conservative. However, it adds labor and expense, especially if the experiment comprises a large number of Amplicons, and can result in a significant reduction in the library preparation yield. It is also unnecessary when the sample does not contain residual Adaptors after the first AMPure XP purification.

3. For experiments that require the pooling of many Amplicon libraries (most Amplicon sequencing experiments), perform the usual AMPure XP purification step on the individual libraries (per the Method Manual), and a second round of AMPure XP purification on the pool, prior to the emPCR amplification. For large Amplicon pools, this strategy would substantially reduce expense and effort compared to the previous one. However, it should be noted that the samples cannot be un-pooled: if one Amplicon library contained a high amount of residual Adaptors, these Adaptors will contaminate the entire pool.