Roche

Diagnostics

Genome Sequencer System

Application Note No. 5 / February 2007

Amplicon Sequencing

www.roche-applied-science.com

454 LIFE SCIENCES

Amplicon Sequencing

Corresponding author: Tom Jarvie, 454 Life Sciences Corporation, Branford, CT, USA, Email: tjarvie@454.com

Introduction

The power of the Genome Sequencer System, and its utility for sequencing amplicons, derives from the ability to sequence from single molecules within a mixture of molecules. When sequenced on this system, each amplicon molecule within the mixture is sequenced individually, allowing for the identification of rare variants and the assignment of haplotype information over the full sequenced length. Genome Sequencer technology provides "instant cloning" of hundreds of thousands of molecules via an emulsion PCR step. This is in sharp contrast to the Sanger method of direct PCR-product sequencing which results in a sequence read that is an average of all the molecules in the mixture. In order to sequence from individual molecules with the Sanger method, the molecules must first be cloned into a vector and grown in bacteria. Cloning of amplicons prior to sequencing with Sanger will increase sensitivity, but not without a large increase in time and cost.

The ability to simultaneously sequence from such a large number of clones enables a number of applications. Amplicon sub-applications that are supported by the Genome Sequencer technology include the following research fields:

- Discovery of rare somatic mutations in complex samples (*e.g.*, cancerous tumors - mixed with germline DNA) based on ultra-deep sequencing of amplicons
- Sequencing collections of exons from populations of individuals to identify diversity
- Sequencing collections of human exons from populations of individuals for the identification of rare alleles associated with disease
- Analysis of viral quasispecies present within infected populations in the context of epidemiological studies
- Evolutionary biology in populations

The multi-alignment algorithm detects a small amount of variation, so highly variable samples, such as the highly variable regions of HIV, may not align properly in the software, leading to missed variants.

Materials and Methods

Materials

Equipment:
Genome Sequencer Instrument (20 or FLX)
GS 20: Software Version 1.0.53 or
GS FLX: Software Version 1.1.01 which includes the GS Amplicon Variant Analyzer (AVA) software

Reagents from Roche Applied Science:
Sample Preparation:
GS emPCR Kit II (Amplicon A, Paired End);
GS emPCR Kit III (Amplicon B),
FastStart High Fidelity PCR System
Sequencing: GS 20 Sequencing Kit (40x75 or 70x75)*;
GS LR70 Sequencing Kit (70x75)**
GS PicoTiterPlate Kit (40x75* or 70*x75**)

* for GS 20
** for GS FLX

Methods

For complete details on the Amplicon Library preparation procedure and sequencing, please refer to the GS Guide to Amplicon Sequencing, the GS emPCR Kit User's Manual, and the Operator's Manual appropriate for your Genome Sequencer Instrument.

A detailed list of all required equipment, reagents and methods is provided in the GS Guide to Amplicon Sequencing.

Experimental Considerations

The DNA-sample preparation for Amplicon Sequencing consists of a simple PCR amplification reaction with special Fusion Primers (Figure 1). The Fusion Primers consist of a 20-25 bp target-specific sequence (3' end) and a 19 bp fixed sequence (Primer A or Primer B on the 5' end).

The 19-mer section of the primers exists in two types, "A" and "B", to match other components of the Genome Sequencer System. The design of the Fusion Primers is described in detail in the Fusion Primer Design section (page 7).

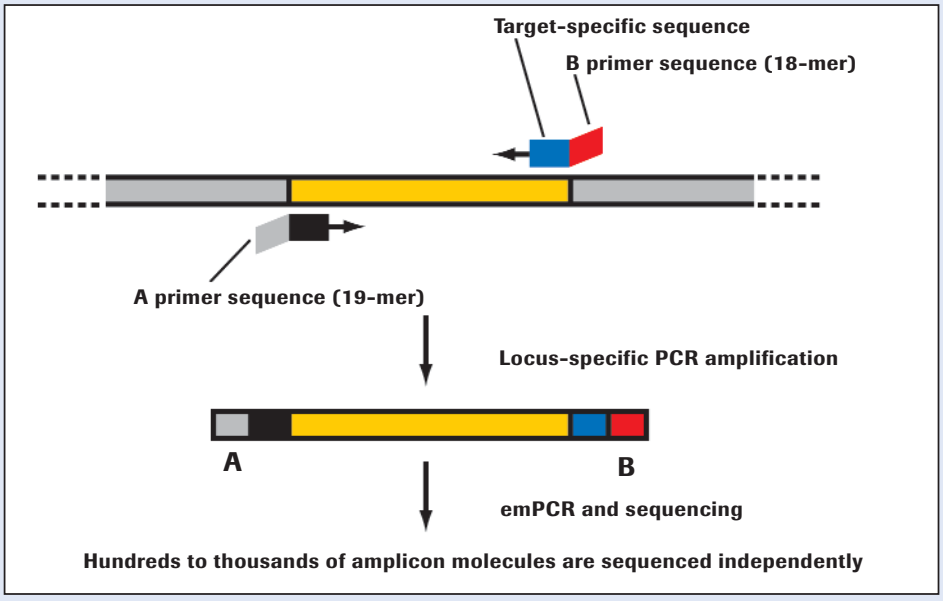

Figure 1: Schematic representation of an amplification product generated by the Amplicon Library preparation procedure. The composite primers each comprise a 20-25 bp target-specific sequence region at their 3' end and a 19 bp region (Primer A or Primer B) that will be used in subsequent clonal amplification and sequencing reactions, at their 5' end.

The Amplicon preparation procedure takes full advantage of the emulsion-based clonal amplification (emPCR) feature of the Genome Sequencer System, allowing for single-molecule sequencing without cloning the target sequences into bacteria. Two emPCR kits are available for Amplicon sequencing: one for sequencing a library from Primer A (GS emPCR Kit II) and the other for sequencing from Primer B (GS emPCR Kit III).

Amplicons for sequencing with the Genome Sequencer System should be no longer than 500 bp, as longer amplicons do not amplify well under the experimental conditions of emPCR; Fusion Primers must be designed accordingly.

Note that in a highly variable target, such as rapidly mutating viruses, it may be difficult to find appropriate sequences for the template-specific portions of the Fusion Primers, especially within the distance restriction imposed by amplification and read length considerations. The choice of appropriate PCR primers for generating the Amplicon Library is critical for a successful experimental design, as studies aimed at identifying and quantitating sequence variants will be, at best, as accurate and unbiased as the original amplification.

Some experimental designs may require the monitoring of multiple amplicons (*e.g.*, various exons of a gene of interest, various markers associated with a disease, *etc.*). For consistency and maximum control of the amplification reaction, we recommend amplifying each target separately (as its own "Amplicon Library") rather than combining multiple targets and primer pairs in a single library (multiplex PCR). However, multiple Amplicon Libraries (each comprising a single peak) can be mixed and processed together through emPCR, and sequenced together on a single PicoTiterPlate device, or even in a single PicoTiterPlate region.

Sample characteristics

The quality and quantity of the DNA sample are critical to the success of amplicon sequencing. Any contamination inherent in the starting material will be directly reflected in the quality of the output library. The amount of required input DNA (the DNA used in the initial PCR amplification with the tailed primers) will depend on the nature of the experiment and desired sensitivity level. For example, if you are searching for low-abundance sequence variants in a complex sample (such as genomic DNA), you should start with 10 – 50 ng of DNA (equivalent to ~3,000-15,000 haploid human genomes). At a minimum, the DNA sample should meet the following criteria:

- DNA should be non-degraded and contain no particulate matter
- Input DNA size should be sufficient to support amplification of the target(s)
- DNA should have an $OD_{260/280}$ ratio of 1.8 or above
- DNA should be at a concentration of 5 ng/µl or above, in TE buffer (0.5 ng/µl for cloned or PCR-generated targets)

Because DNA quantitation using OD_{260} is variable and depends on DNA purity, we recommend verifying the input DNA concentration and integrity by densitometry (*e.g.*, on a 1 – 2% agarose gel using a DNA mass ladder) or by fluorometry (*e.g.*, using the Quant-iT PicoGreen dsDNA Assay Kit [Invitrogen]).

Example of an Experimental Setup

The processing and sequencing of amplicons with the Genome Sequencer System is quite flexible and allows for a wide range of experimental design. A researcher can choose a variety of options regarding design parameters, such as the length of amplicons, the number of amplicons pooled together, the number of reads desired for a given amplicon pool, and whether to read from the A end, the 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 will ensure the best possible result.

- The highest confidence in low frequency variation will result from bi-directional reads.
- A high-fidelity polymerase must be used in the amplicon generation step. Use of a low-fidelity polymerase will result in many amplification-induced variations in the sequence. Although there are many choices of enzyme, Roche's FastStart High Fidelity PCR System has high fidelity coupled with robust amplification of a wide array of input templates.
- Greater confidence in results may be achieved by running replicates of the biological material through the sequencing process and comparing the results.
- The level of multiplexing should be determined by
  - Number of amplicons of interest
    - Not limited by the sequencing technology
  - Desired sensitivity/depth of coverage
- When sequencing mixtures of multiple amplicons, care must be taken in quantification and pooling of amplicons
  - Equimolar mixtures will generate best results
- Forward and reverse reads will eliminate most systematic, context-dependent sequencing errors
  - The ideal experiment has reads covering the amplicon forward and reverse
- Comparison of normal versus diseased or similar control samples will aid in identifying systematic errors
  - For example, a variation that shows up at 3% in the normal and the diseased is not significant
- Variations near the ends of reads may be miscalled by the software
  - The multiple aligner may not have time to recover after the variation and align the sequence properly

The number of amplicons that can be combined in an experiment, while theoretically unlimited, is primarily determined by the desired sensitivity of detection. Follow the general guidelines below when determining the level of oversampling required for a desired level of detection.

Example of an Experimental Setup cont.

- Heterozygote detection
  – 40x
- 5% variation of single base changes and multi-base deletions
  – 1000x coverage (good statistical chance for 50 variation reads)
- 1% variation of single base changes and multi-base deletions
  – 5000x coverage (good statistical chance for 50 variation reads)
- Single-base indels will require additional depth

Empirical evidence has shown that pooled amplicons that are carefully quantitated, pooled, and subjected to emPCR are typically within a factor of two of one another when sequenced. Therefore, when sequencing from pools of amplicons, increasing the oversampling by a factor of two may be warranted.

Workflow overview

The workflow for amplicon experiments follows a few straightforward steps that are illustrated in the flow diagrams presented in Figure 2 (Generation of Template) and Figure 3 (emPCR). Figure 4 shows recommendations for pooling of amplicons in the various gasket formats for the Genome Sequencer FLX or 20 System.

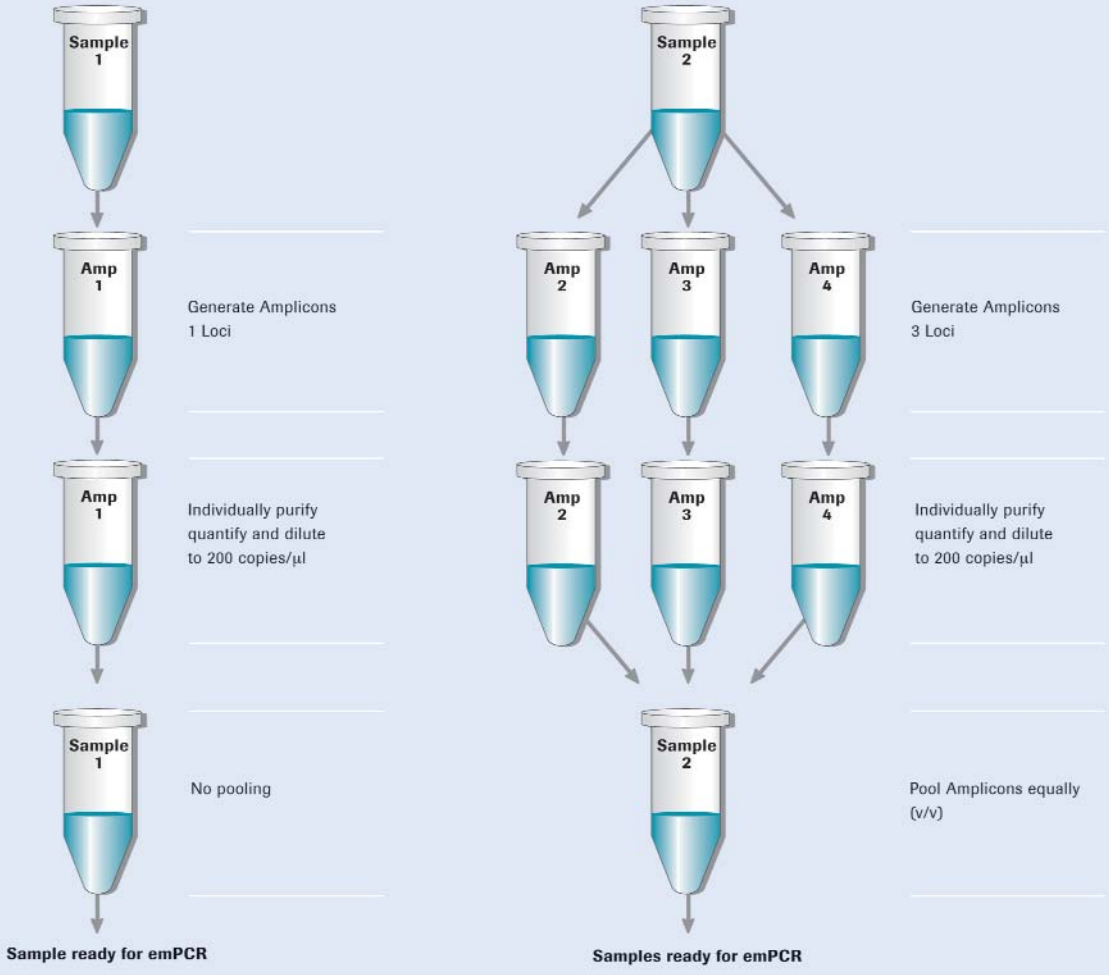

Figure 2: Generation of Template
As per the schematic below, each amplicon is amplified separately, purified, quantified, and diluted to 200,000 copies per microliter. The amplicons corresponding to a particular sample are subsequently pooled in equimolar ratios ready to be used in the emPCR reaction.

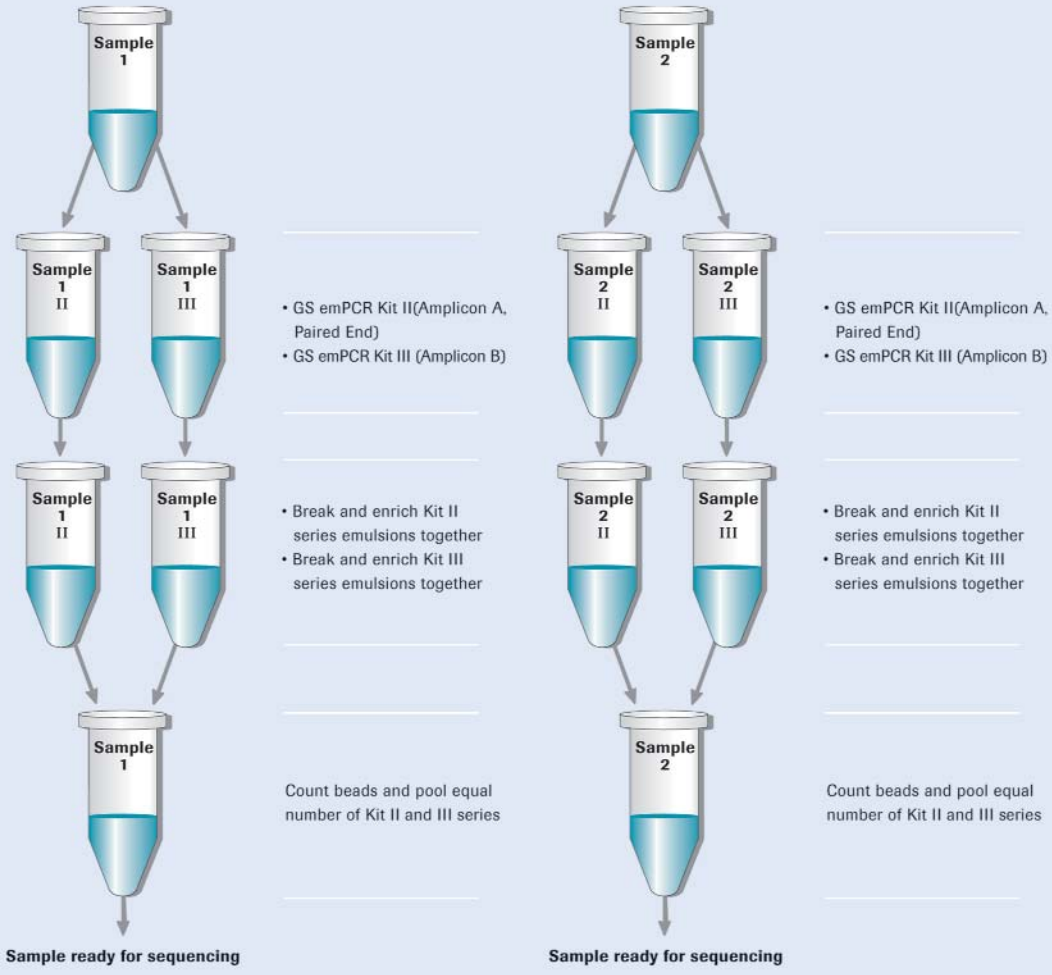

Figure 3: emPCR
Each amplicon pool is subjected to emPCR using both a GS emPCR Kit II (Amplicon A, Paired End) and a GS emPCR Kit III (Amplicon B), in separate reactions (for bi-directional sequencing). After completion of the emPCR amplification, breaking, and enrichment procedures the GS emPCR Kit II and III emulsions for each sample are pooled and processed together. The number of emPCR reactions is determined by the sequencing format/desired number of beads.

Genome Sequencer 20 GUIDELINES

50%				
Format	**Full Plate**	**Half Plate**	**4 Region**	**16 Region**
# Amplicons	2000	1000	500	100
10%				
Format	**Full Plate**	**Half Plate**	**4 Region**	**16 Region**
# Amplicons	400	200	100	20
5%				
Format	**Full Plate**	**Half Plate**	**4 Region**	**16 Region**
# Amplicons	200	100	50	10
2%				
Format	**Full Plate**	**Half Plate**	**4 Region**	**16 Region**
# Amplicons	80	40	20	4

Genome Sequencer FLX GUIDELINES

50%				
Format	**Full Plate**	**Half Plate**	**4 Region**	**16 Region**
# Amplicons	4000	2000	700	120
10%				
Format	**Full Plate**	**Half Plate**	**4 Region**	**16 Region**
# Amplicons	800	400	140	24
5%				
Format	**Full Plate**	**Half Plate**	**4 Region**	**16 Region**
# Amplicons	400	200	70	12
2%				
Format	**Full Plate**	**Half Plate**	**4 Region**	**16 Region**
# Amplicons	160	80	28	4

Figure 4: Pooling of amplicons using various Genome Sequencer gasket formats. The guidelines in these tables for the number of pooled amplicons are based upon the equimolar pooling of ideal amplicons and on a model that is set to sample 50 reads of each low-frequency variant within the pool. The percentages listed are the desired sensitivity for a (rare) variant within the sample (50%, 10%, 5%, and 2%). This level of sampling (50 reads for each low-frequency variant) is a compromise between ensuring a high probability of finding and accurately quantifying the variant and pooling large numbers of amplicons. As an example of how the values in the table are calculated, consider a desired sensitivity of 10% on a full plate run on the Genome Sequencer FLX. On a full plate run, one could expect ~400,000 sequence reads; 10% of 400,000 is 40,000. If the desired number of reads for each rare variant is 50, we need to divide 40,000 by 50 to get the value of 800 pooled amplicons that is represented in the table. The actual sensitivity of the assay for a given amplicon will depend upon the relative abundance of the amplicon within the population of pooled amplicons. Additionally, certain types of variations, as discussed in the main text, require a greater depth of coverage. These guidelines are theoretical and some of the guidelines, such as the pooling of more than 400 amplicons, may not be practical.

Fusion Primer Design continued

The amplification primers are a combination of a 20-25 bp target-specific sequence and a 19 bp fixed sequence. The 19-mer serves as a priming site during clonal amplification, allows binding to the DNA Capture Beads, and serves as the primer site for the sequencing reaction. The final four bases of the 19-mer are the sequencing key "TCAG". There are two types of primers, termed "Primer A" and "Primer B", for use with the GS emPCR Kits II and III respectively, allowing the sequencing of the insert from either end. The exact sequences are as follows:

Primer A Sequence
5' GCCTCCCTCGCGCCATCAG 3'

Primer B Sequence
5' GCCTTGCCAGCCCGCTCAG 3'

The 19-mers are appended to the 5' end of the amplicon specific primers, typically 20-25 bp in length (may vary). The normal constraints of primer specificity and annealing conditions apply. Amplicons should be no longer than 500 bp because templates longer than this do not amplify well in emPCR.

Example of a Fusion Primer Pair Design
The following is a typical Fusion Primer pair design. The template of interest is HLA SNP DD14; the SNP (highlighted in red) is used as the template, and the Primer3 software package (1) is employed for primer design.

DD14

```
AGATGTAGCCCTTGAAATGTCATAAATATAGATTTTTGCTTCTGATTCAATCTGACGATCTCTG
TCTTCTAACCTATGTTCAATTCATATGGTAGTCAAAGTGAGCAAACTTGTTTCTGCAAGAGACA
AACACTGAAGCCTCAGTGGTTTAACAAAACACAGGTTTATTTTTTAGCCACGTGTAGTTCAAGG
CAGGTTGG
```

Primer Design Software Output

OLIGO	start	len	tm	gc%	any	3'	seq
LEFT PRIMER	51	24	59.06	45.83	4.00	0.00	TCTGACGATCTCTGTCTTCTAACC
RIGHT PRIMER	193	20	60.32	55.00	6.00	2.00	GCCTTGAACTACACGTGGCT
SEQUENCE SIZE:	200						

INCLUDED REGION SIZE: 200

PRODUCT SIZE: 143, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 2.00

```
1    AGATGTAGCCCTTGAAATGTCATAAATATAGATTTTTGCTTCTGATTCAATCTGACGATC
                                                  >>>>>>>>>>
61   TCTGTCTTCTAACCTATGTTCAATTCATATGGTAGTCAAAGTGAGCAAACTTGTTTCTGC
     >>>>>>>>>>>>>>
121  AAGAGACAAACACTGAAGCCTCAGTGGTTTAACAAAACACAGGTTTATTTTTTAGCCACG
                                                        <<<<<<<
181  TGTAGTTCAAGGCAGGTTGG
     <<<<<<<<<<<<<
```

Fusion Primers

Fusion primers are generated by addition of the Primer A and Primer B sequences to the 5' ends of the template-specific primers.
Primer A Left template-specific primer
Fusion Primer A: GCCTCCCTCGCGCCATCAGTCTGACGATCTCTGTCTTCTAACC
Primer B Right template-specific primer
Fusion Primer B: GCCTTGCCAGCCCGCTCAGGCCTTGAACTACACGTGGCT

Amplicon Library Preparation Procedure

Refer to GS Guide to Amplicon Sequencing Chapter 3.

Results

Sequencing was performed with the GS Sequencing Kits listed on page 2. The sequencing results are shown in the accompanying figures, presented in two views. The first view (as shown in Figures 6 and 8) is the variation histogram plot. The variation histogram plot shows the reference sequence on the horizontal axis, the percentage at which a given variation is present on the left vertical axis, and the depth of coverage (the number of independent, clonal sequence reads at a given position in the reference sequence) on the right vertical axis. In this view, variations from the reference sequence are represented as a 'histogram bar'. Only positions in the individual clonal reads that disagree with the reference sequence are shown. The individual sequence reads are shown below the histogram in a gapped, multiple alignment.

The second view (as shown in Figures 7 and 9) is the tri-flowgram view. The top frame in the tri-flowgram view is an idealized flowgram of the reference sequence. The middle frame is the flowgram of an individual read. The bottom frame is the difference between the individual read and reference flowgrams. If a variation results in a flow cycle shift, a gap (represented in grey) is introduced into the flowgram in order to make the rest of the flowgram align. The tri-flowgram view has the advantage of allowing a direct observation of the signal in the individual sequence reads. Comparing the flowgram of an individual read with a variation to a flowgram of the reference sequence provides confirmation of the variation.

When interpreting the results, one needs to be cognizant of the strengths and limitations of Amplicon sequencing assays. The following is a list of criteria, presented in order from the most to least convincing, for the identification of a valid variation.

1. A gap (cycle shift) is introduced in order to make the rest of the flowgram align.
2. No gap is introduced, but the magnitudes of two neighboring flows change in opposite directions.
3. The magnitude of a flow has changed, but it is not upstream or downstream of a large-magnitude monomer repeat of the same flow type.
4. The magnitude of a flow has changed.
5. Differences are seen at the left or right edges of the Aligned Flowgram.

Basic magnitude changes that lead to single base insertions or deletions are more convincing if the signal distribution is evenly distributed around the called value, rather than consistently above or below. Even if a difference looks genuine, one should ask the question "Is it biologically relevant?" (*e.g.*, might it have been caused by PCR artifact?).

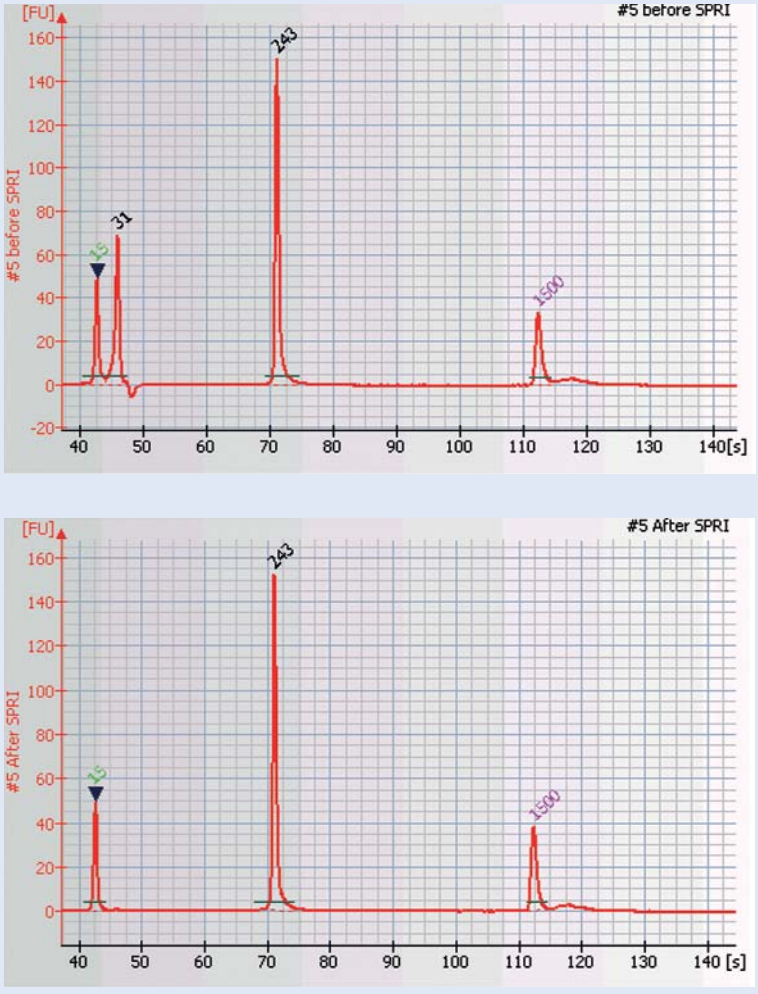

Figure 5: (A) **DNA 1000 LabChip profile** of an Amplicon library comprised of a single amplicon of 243 bp, showing a sizable adaptor dimer peak at 31 bp. (B) The same library after an additional purification on Ampure SPRI beads. If unintended amplification products are observed, such as primer dimers, optimization of the PCR conditions should be considered, or the SPRI bead purification repeated: these contaminating products must be removed or they will interfere with sequencing. The peaks at 15 and 1500 bp are internal markers.

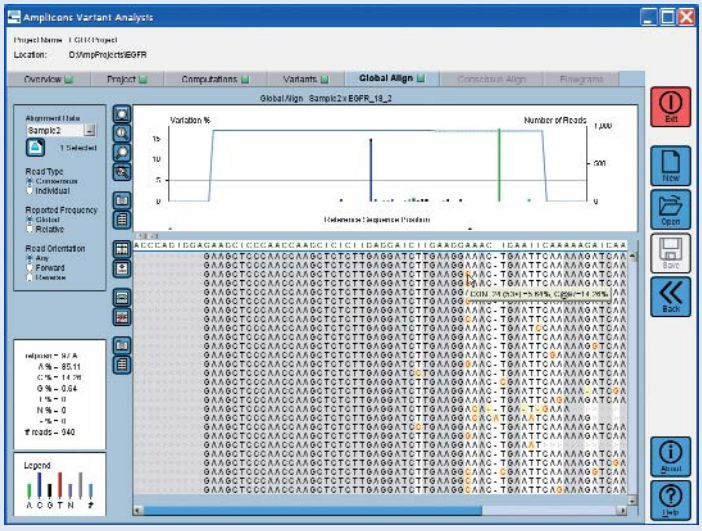

Figure 6: Variation Histogram Plot of two single-base-change variations. The sample is EGFR exon 18 from a non-small cell lung carcinoma research sample. This particular amplicon was read to a depth of 940 clonal reads. The Variation Histogram clearly reveals two variations; an A to C change (blue bar at approximately 14%) and a G to A change (green bar at approximately 17%). The corresponding reads are shown in the gapped, multiple alignment below the histogram. (Sample courtesy of Drs. Roman Thomas and Matthew Meyerson, Dana-Farber Cancer Institute, Harvard Medical School, and The Broad Institute of MIT and Harvard).

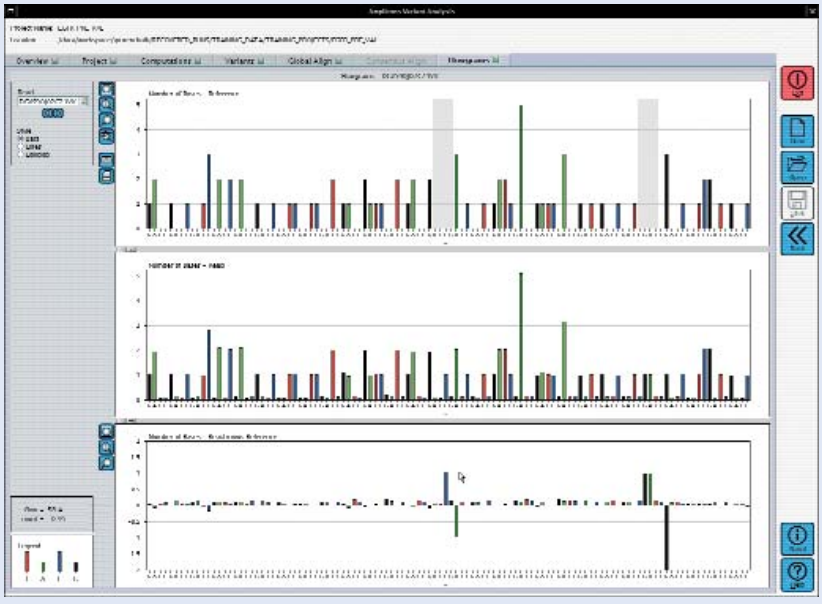

Figure 7: Tri-flowgram view associated with the Variation Histogram Plot in Figure 6. The top frame in the tri-flowgram view is an idealized flowgram of the reference sequence. The middle frame is the flowgram of an individual read. The bottom frame is the difference between the individual read flowgram and reference flowgram. Noticeable in this case is that both mutations cause a flow cycle shift (represented in grey), adding confidence to the called mutations. Inspection of individual sequencing reads (flowgrams) reveals that the two mutations are located on the same allele.

For clarity, the local sequence around the variations is presented above the reference and sample sequence flowgrams. Note that the difference flowgram represents single base changes as a 'zero sum' of insertions and deletions. (Sample courtesy of Drs. Roman Thomas and Matthew Meyerson, Dana-Farber Cancer Institute, Harvard Medical School, and The Broad Institute of MIT and Harvard).

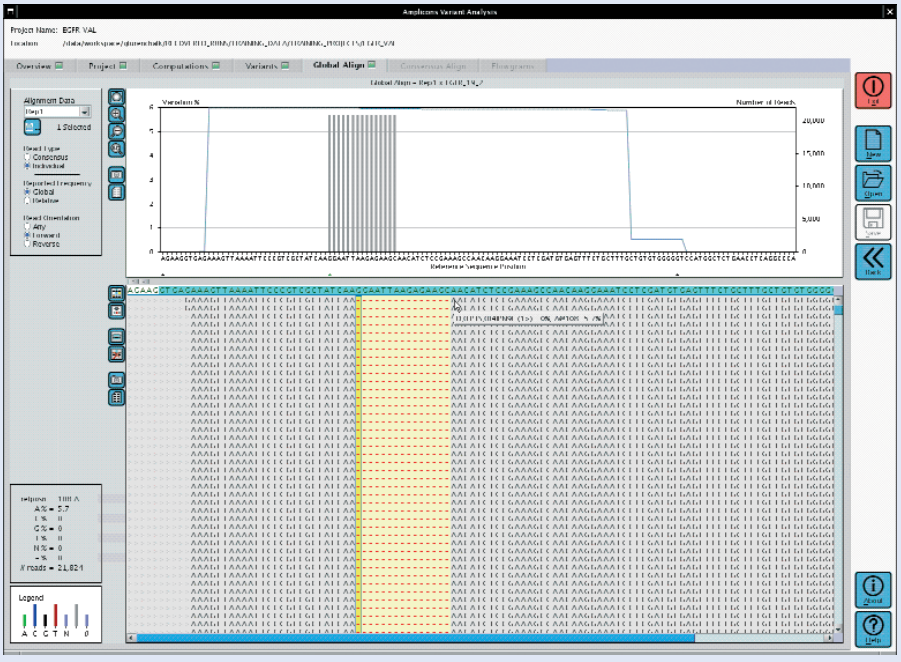

Figure 8: Variation Histogram Plot of EGFR exon 19 from a non- small cell lung carcinoma research sample. 454 Amplicon Sequencing uncovered a 15 bp deletion at ~6% abundance (indicated in light blue). The deletion is also readily-seen in the gapped alignment of the FASTA reads. (Sample courtesy of Drs. Roman Thomas and Matthew Meyerson, Dana-Farber Cancer Institute, Harvard Medical School, and The Broad Institute of MIT and Harvard).

Figure 9: Tri-flowgram view associated with the Variation Histogram Plot in Figure 8. For clarity, the local sequence around the variations is presented above the reference and sample sequence flowgrams. The 15 bp deletion causes multiple, consecutive flow cycle shifts (represented as the large grey region in the individual read flowgram). The changes in the difference flowgram, a 17 bp deletion and a 2 bp insertion, add up the 15 bp deletion. (Sample courtesy of Drs. Roman Thomas and Matthew Meyerson, Dana-Farber Cancer Institute, Harvard Medical School, and The Broad Institute of MIT and Harvard).

References

1. Rozen, S., Skaletsky, H. (2000) “Primer3 on the WWW for general users and for biologist programmers”. In: Krawetz, S., Misener, S. (eds) Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ, pp 365-386.
2. Thomas, R.K., *et al.*, “Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing”. Nature Medicine 2006 July, 12(7):852-5.
3. Sogin, M.L. *et al.*," Microbial diversity in the deep sea and the underexplored "rare biosphere," Proc Natl Acad Sci U S A. 2006,103:12115-12120.
4. Sogin, M.L. *et al.*,"Rapid Sequencing Procedure Enables Most Exact Assessment of Number and Diversity of Marine Species," Biochemica 2007 No. 1, 4-6.
5. Turcotte, C., Jarvie, T.(2006) "Detection and Quantitation of Methylation Patterns using Amplicon Sequencing," Application Note No.3, Roche Applied Science, www.genome-sequencing.com

NOTICE TO PURCHASER

RESTRICTION ON USE: Purchaser is only authorized to use the Genome Sequencer Instrument with PicoTiterPlate devices supplied by 454 Life Sciences Corporation and in conformity with the procedures contained in the Operator’s Manual.

Trademarks

454, GENOME SEQUENCER, PICOTITERPLATE, emPCR, and ULTRA DEEP SEQUENCING are trademarks of 454 Life Sciences Corporation, Branford, CT, USA.

Other brands or product names are trademarks of their respective holders. FASTSTART is a trademark of Roche.

For more information, visit
www.genome-sequencing.com

04976240001 ① 0207

454 LIFE SCIENCES

Roche Diagnostics

Roche Diagnostics GmbH
Roche Applied Science
68298 Mannheim
Germany
www.roche-applied-science.com