Development and Performance of the VariantSEQr™ Resequencing System in High Throughput DNA Sequence Variation Studies

Bob Nutter, Jon Sorenson, Carey Gire, Pei Shen, Mary Ann Rydland, Steve Glenowski, Nathan Edwards, Manohar Furtado, Rixun Fang and Lin-Zuo Pham: Applied Biosystems, Foster City, CA 94404

Abstract

The development and performance of the high-throughput resequencing system, including the Applied Biosystems 3730 DNA Analyzer and the SeqScape® Resequencing System, is described. The 3730 DNA Analyzer includes patented technology licensed from Hitachi, Ltd. as part of a strategic partnership between Applied Biosystems and Hitachi, which is used for the VariantSEQr™ Resequencing System, a fully integrated system capable of quickly resequencing human genes in a cost-effective manner. This system consists of PCR primers of known performance, robust PCR and sequencing chemistries and the fully integrated SeqScape® v2.1 software for mutation detection and report generation. We report here results of our development of a validated process for designing primers for high-throughput amplification and resequencing of the promoter regions, exon regions, and flanking intronic regions for genes implicated in cancer and other diseases. Primer design for large scale resequencing projects has been greatly improved by our ability to correlate both unsuccessful PCR amplification and poor quality sequence results to the presence of local and global factors in the genome. Using very large datasets of PCR primer amplification results (<200,000 amplicons) and sequencing data (>10,000,000 sequence files) generated during the Applied Genome Initiative, we have developed a model that is predictive of the success rate for a given amplicon. We will present data from early test sites demonstrating the ability of the system to accurately detect variations in genes and develop genotypes to help understand the role these variations play in altered function and diseases. The validation of this system will permit the resequencing of genes from many genomes.

Figure 1

Figure 1: Diagramatic view of primers designed for a typical Resequencing Set (RSS). The darker green regions at the top represents exons and the lighter green region the promoter region or intron. The target region of the Resequencing Amplification DNA, shown in orange, has been designed to provide complete coverage of promoter regions, intron/exon junctions and all exons. The regions flanking the target regions represented in red are part of the amplicons but may not always have data of the desired high quality. Each PCR primer is marked with either the M13 Universal Forward or Reverse sequencing primer to permit robust and specific sequencing of the amplification products. While a number of software packages have the ability to design primers for amplification, it has not been possible before now to know if sequences in the genome will interfere with the generation of the high quality of sequence data necessary for resequencing projects. This uncertainty has led us to develop a means to reliably predict resequencing amplification performance without the need for every primer set to be tested in the laboratory.

Figure 2

Figure 2: Workflow of the Applied Biosystems VariantSEQr™ Resequencing System. The system has been designed to fit the workflow of a typical resequencing laboratory. An optimized protocol provides complete integration of each step from PCR amplification using PCR primers validated by a combination of laboratory and computational systems to base calling, alignment and assembly. The Collection v2.0 software with SeqScape® v2.1 software. At the end of each run, sequence files are automatically assembled and annotated using SeqScape® Resequencing Analysis. This software is an enhancement of the previously validated SeqScape®. The sequence analysis software is designed to provide complete failure of amplification. The ability to detect such failures can be correlated with the presence of CpG islands in the amplicons. These amplicons will not be sold.

Figure 3

Figure 3: Over 200,000 PCR amplicons were generated and the frequency of success was correlated to the GC content of the amplicons. The results show that GC content > 73% is linked to failure. Further analysis of the amplicons surrounding the PCR primer sites has shown other features that are correlated to PCR failure.

Figure 4

Figure 4: Classifications for Resequencing Primer Performance

- H1 - High quality high consensus sequence data for these amplicons can be expected from both the Forward and Reverse sequencing reaction. Double stranded coverage will be obtained for the target region. Consensus quality will be obtained for all individual sequencing reactions.
- H2 - High quality sequence data for these amplicons can be expected for a single orientation across the target region. Reasons for this can include homopolymer stretches, other low complexity repeats (LCR) or heterozygous insertions or deletions. High quality sequence generally will be seen for each orientation up to the homopolymer or LCR but not later. After alignment, each strand will have high quality data up to the problematic region. See diagram below:
- F1 - It is predicted that at approximately 70% of the time these amplicons will give either an H1 or H2 sequence coverage. One of the reasons for this can be high amplion GC content. However, it is not completely understood at this time what factors may lead to this result.
- F2 - Amplicon is not likely to provide at least 70% H1 or H2 sequence. This could be due to complete failure of amplification. Amplification failure has been correlated with the presence of CpG islands in the amplicons. These amplicons will not be sold.

Figure 5

Figure 5: The results of very large resequencing projects at ApInos can be expected from both the Forward and Reverse sequencing reaction. Double stranded coverage will be obtained for the target region. Consensus quality will be obtained for all individual sequencing reactions.

Conclusions

The Applied Biosystems VariantSEQr™ Resequencing System is the only fully integrated and highly automated resequencing application that enables scientists to focus on science instead of PCR primer validation and developing the required data analysis systems. The PCR primers have been validated using a combination of laboratory investigation as well as the application of an experts-constructed statistical system. Not only does this eliminate the need for time consuming and expensive PCR primer validation, it also provides templates that will provide very high quality data sequence data. Data analysis is greatly simplified by the use of SeqScape® v2.1 software. The analysis system uses geno content derived from the Celera Discovery System and provided at no additional cost. This allows resequencing projects to be automatically baseline, assembled and aligned against a reference sequence for review and report generation.