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 Glanowski, Nathan Edwards, Manohar Furtado, Rixun Fang and Lin-Zuo Pham; Applied Biosystems, Foster City, CA 94404

Abstract

The completion of a reference sequence for the human genome and improvements in high-throughput sequencing technology, including the Applied Biosystems 3730xl DNA Analyzer and the BigDye® Terminators v3.1 Cycle Sequencing Chemistry, have enabled the development of 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 Applera Genome Initiative, we have developed a model that is predictive of the success rate for a given amplicon. We will present the results of a comparison of the success of primer amplification and generation of high quality sequence data in the laboratory to the success predicted by our model. We will also present data from early test sites demonstrating the ability of this system to quickly and accurately detect variations in genes and develop genotypes to help understand the role these variations play in altered function and disease. The validation of this system will permit the resequencing of genes from a number of genomes.

Figure 1

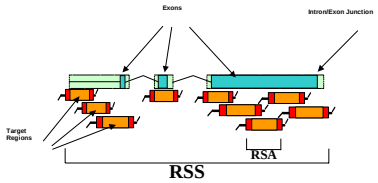


Figure 1 Diagrammatic view of primers designed for a typical Resequencing Set (RSS). The darker green regions at the top represents exons and the lighter green regions either the promoter region or introns. The target region of the Resequencing Amplicons (RSA, shown in orange) have been designed to provide 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 tailed with either the M13 Universal Forward or Reverse sequencing primer to permit robust and specific sequencing of the amplified regions. 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 lead us to develop a means to reliably predict resequencing amplicon performance without the need for every primer set to be tested in the laboratory.

Figure 2

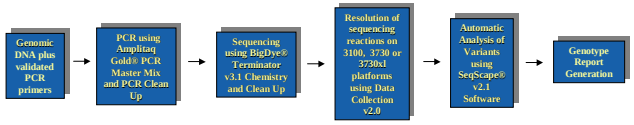


Figure 2 Workflow of the Applied Biosystems VariantSEQr™ Resequencing System. The system has been designed to fit the workflow of a typical resequencing laboratory. An optimized protocol provides complete integration of each step from PCR amplification using PCR primers validated by a combination of laboratory and computational systems to the generation of genotype reports. An important new feature is the integration of the 3100 and 3730 Data Collection v2.0 software with SeqScape® v2.1 software. At the end of each run, sequence files are are automatically basecalled and can be exported to SeqScape® v2.1 software where they are automatically trimmed, aligned and assembled against a reference sequence. Results can then be easily reviewed and reports generated at the convenience of the scientist. A Demonstration Kit, containing all components of the VariantSEQr™ Resequencing System will soon be available from your Field Applications Specialist to allow 'no risk' evaluation.

Figure 3

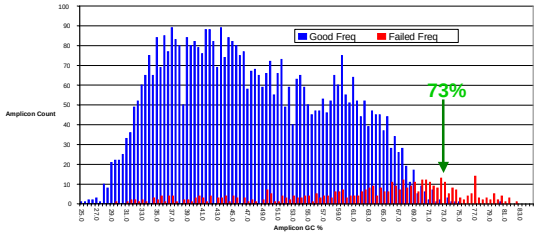


Figure 3 Over 200,000 PCR amplicons were analyzed and the frequency of success was correlated to Amplicon GC content. The results show that GC content > 73% is linked to failure. Further analysis of the sequences surrounding the PCR primer sites has shown other features that are correlated to PCR failure

Figure 4

Classifications for Resequencing Primer Performance

•H2 – High quality 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 QV>30 will be seen after assembly of the individual sequencing reactions.

•H1 – 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 after. After alignment, each strand will have high quality data up to the problematic region. See diagram below:

•M – It is predicted that at approximately 70% of the time these amplicons will give either H1 or H2 sequence coverage. One of the reasons for this can be high amplicon G+C content. However it is not completely understood at this time what other factors may lead to this result.

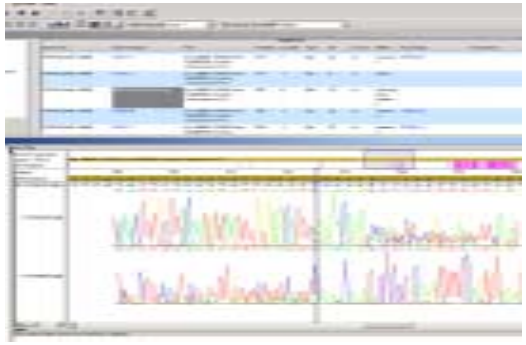
•Fail – Amplicon is not likely to provide at least 70% H1 or H2 sequence. This could be due to complete failure of amplification. Amplification failure has been correlated with the presence of CpG islands in the amplicons. These amplicons will not be sold.

Figure 5

Amplicon Value	# of Predictions	Accuracy
H2	1442	96.2%
H1	151	95.4%
M	11	63.7%
F	164	81.7%

Figure 5 The results of very large resequencing projects at Applera companies has allowed the creation of expert systems that not only permit the design of PCR primers that have maximum specificity but also searches the genomic sequence surrounding the primer sites for structures that have been correlated to poor PCR amplification and/or poor quality DNA sequence. These systems are collectively known as the PDA (Primer Design Algorithm). This figure shows the relationship between sequence data quality generated in the laboratory to the data quality predicted by the application of PDA to the designed PCR primers. There is excellent agreement between predicted and actual data quality.

Figure 6



Improved Detection of Heterozygous Indel Mutations with SeqScape® v2.1 software. Some investigators have estimated that up to 10% of all genes contain heterozygous indels of varying sizes. Only sequence based resequencing systems have the ability to detect indels of any size. This is in contrast to hybridization based resequencing systems which can not detect all indels, thus leading to unknown numbers of errors in the genotype reported. This figure shows that not only can SeqScape® v2.1 software identify indels, but it also automatically reports the sequence that was deleted or inserted.

Figure 7

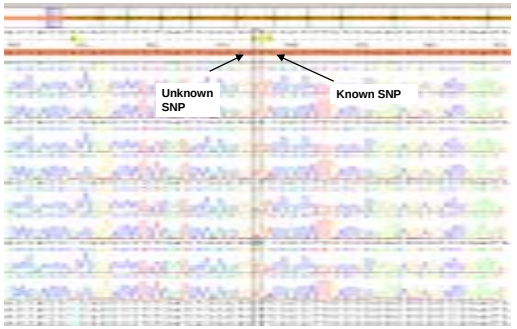


Figure 7. Discovery of Novel SNPs using the Applied BioSystems VariantSEQr™ Resequenicng System. The data for this figure was generated by an early test site. Not only where the investigators able to find all previously known SNPs in the genes they studied, but they also discovered a large number of previously unreported SNPs. Many of these unknown SNPs were located in regions, such as with this example, where they interfered with proper genotype calls using Taqman® or other probe-based SNP screening assays.

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 expert-trained computational 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 sequence data. Data analysis is greatly simplified by the use of SeqScape® v2.1 software. This analysis system uses gene content derived from the Celera Discovery System and provided at no additional cost. This allows resequencing projects to be automatically basecalled, assembled and aligned against a reference sequence for review and report generation.