Selective enrichment and massively parallel sequencing of medically relevant genomic regions using long-range PCR

Adam N. Harris, John Bishop, Isabel Cisneros, Qingli Mi, Lushen Li, Eric C. Olivares, Kyusung Park, Peter Welch, and Rob Bennett.
Life Technologies • 5791 Van Allen Way • Carlsbad, California 92008 • USA

Abstract

The high capacity of current generation massively parallel sequencing technologies enables medical resequencing of gene families or candidate genomic regions implicated in disease states across large sets of patient samples. To increase sample multiplexing, enrichment of these target sequences by PCR has been employed. Often, PCR requires several rounds of optimization, and its cost is an issue for regions significantly larger than 1 Mbp. Using a PCR formulation designed to maximize success rate of long PCR, we observe robust amplification of up to 20 kbp per amplicon.

We describe the use of this PCR enrichment method to demonstrate amplification and sequencing on the SOLiD™ platform for specific target regions. We additionally explore the combination of this approach with other cost-saving measures including ligation-blocked primers to reduce over-sampling of amplicon ends and a simple amplicon normalization solution.

Introduction

The SequalPrep™ Long PCR Kit with dNTPs offers a convenient, factory-proof sample preparation that balances success rate, yield, and fidelity. Little to no optimization is needed for amplification of targets (including GC rich targets) up to 20 kb in length, thus significantly reducing the required number of PCR reactions. In combination with the SequalPrep™ Normalization Plate, this tool is ideal for medical re-sequencing of large chromosome regions.

We tested the effective of the SequalPrep™ system by amplifying the ~190kb Cystic Fibrosis Transmembrane Receptor (CFTR) gene with two complete sets of primer pairs for a total of 42 amplicons. With no optimization, 88% of the amplicons could be generated on the first attempt. An additional 5% of the amplicons were produced following the recommended optimization regimen. We found coverage to be more uniform when employing 5’ amine-blocked primers and the SequalPrep™ Normalization Plates.

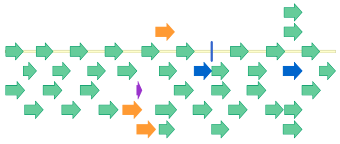
Materials & Methods

CFTR PCR. Primer pairs for ~9-11 kb amplicons spanning the entire human CFTR gene in two sets were designed using Vector NTI. Primers were synthesized by Life Technologies with and without a 5’-amine modification to test suppression of overrepresented amplicon ends. PCR with SequalPrep™ Long was performed on 100 ng of K562 human gDNA per reaction and the recommended starting point of 0.5x Enhancer A. Touchdown PCR was performed on an Applied Biosystems 9700 thermocycler with annealing temperatures within 5°C of the predicted Tm for each primer pair and an initial extension time of 11 minutes. Aliquots of each PCR reaction were checked on E-Gel® 48 1% agarose gels; PCR reactions with a band centered at the expected size were scored as successful. Primer pairs which failed to yield product of the expected size were used in subsequent PCR reactions at 1x Enhancer A, and 0.5x and 1x Enhancer B. Successful PCR reactions were combined to produce a non-normalized pool of CFTR amplicons. Separately, SequalPrep™ Long Normalization 96 well plates were used to normalize DNA yields from 11 ul of each of the successful PCR reactions according to the kit instructions and pooled.

SOLiD™ sequencing. Four amplicons pools (non-normalized and normalized CFTR amplicons each with both blocked and non-blocked primers) were separately processed to generate libraries using Invitrogen’s SOLiD™ Fragment Library Construction Kit according to the SOLiD™ manual, using a Covaris S2 for shearing and a pre-amplification size-selection step with the included E-Gel® SizeSelect™ 2% gels. Libraries were amplified for 11 cycles and then quantitated by SOLiD™ library-specific TaqMan™ qPCR. Emulsion PCR was performed followed by 35 cycles of sequencing on 4 octet spots (one half of one flow cell) on a SOLiD™ instrument using V2 chemistry. An average of > 7 million reads were mapped from each spot with 3 or fewer mismatches to a Chr7q reference.

Results

Figure 1 – Amplicon Design and Success



ref NC 000007.12 NC 000007:116907000-117096155

Figure 1. Schematic of 42 overlapping ~9-11 kb amplicons (all arrows except purple) spanning the entire CFTR gene. 37 amplicons (88%, green arrows) were successfully generated using the recommended starting conditions for SequalPrep™ Long PCR (0.5x Enhancer A). Two more amplicons (blue arrows) were successfully generated by using 1x Enhancer B during PCR. Three amplicons (orange arrows) failed to amplify after four PCR attempts, for a final success rate of 39/42 (93%). A ~3kb amplicon (purple arrow) was subsequently generated using Platinum® Taq High Fidelity to close a gap in the coverage.

Figure 2 – Coverage on Human Chromosome 7q

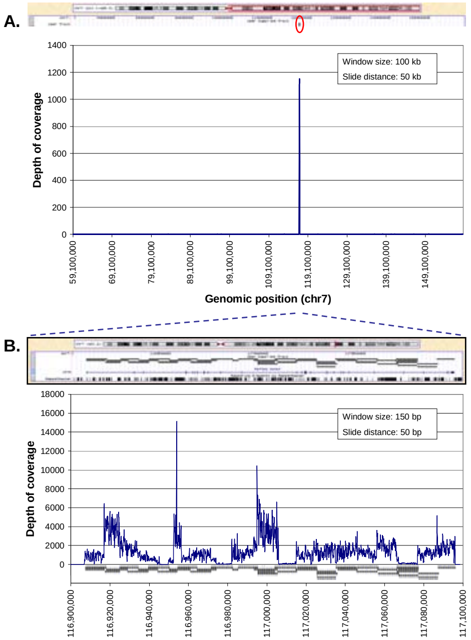


Figure 2. The vast majority of sequence coverage mapped to chr7q was derived from the ~190kb CFTR region. **A)** UCSC Genome Browser schematic of chromosome 7 with the amplicon region (red elipse) displayed in the context of the 7q arm. Depth of coverage along 7q is plotted as average base coverage over a 100 kb sliding region in 50 kb increments. **B)** A zoomed in view of the CFTR region showing amplicons and repetitve elements. Coverage is now shown at a finer level using 150 bp windows sliding 50bp at a time. Coverage is higher (with some notable exceptions) where multiple amplicons overlap and near the ends of some amplicons.

Figure 3 – Amine Blocking Amplicon Ends

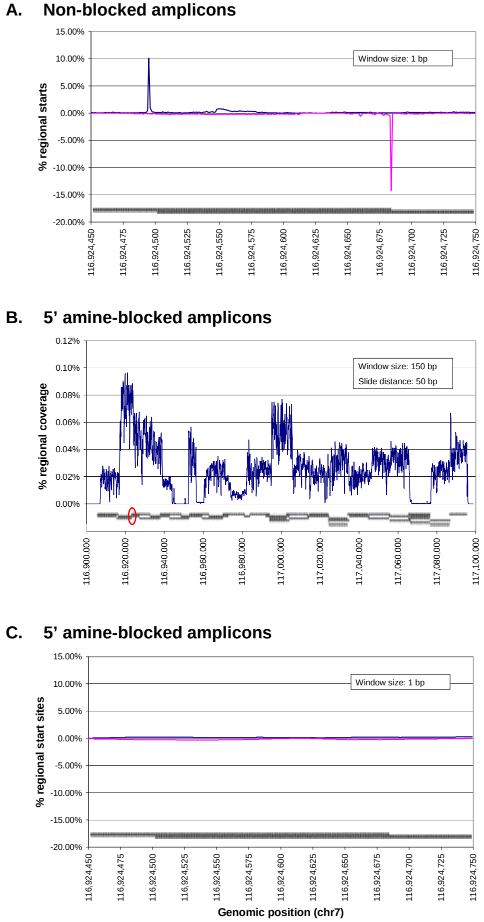


Figure 3. Amplicon end overrepresentation can be blocked by amine-modified primers. To normalize between samples with different overall sequencing coverage, values on the Y axes are displayed as a percentage of the number of start points or depth of coverage for the region shown. **A)** Sequence read start points in a 300 bp region with three overlapping amplicons and two amplicon ends (schematic at bottom of chart). Large spikes in the forward (blue line, positive values) and reverse (magenta line, negative values) directions are evident where the ends of the amplicons are overrepresented as fragment ends in the sequencing library. **B)** Overall CFTR coverage from pooled amplicons generated with 5’ amine-blocked primers which prevent the amplicon ends from being ligated to library adaptors. Coverage is more even than for amplicons made with non-blocked (Fig. 2B) primers. The 300 bp region from (A) is indicated by the red elipse. **C)** The same 300 bp region as in (A) but with blocked primers shows no spikes in start sites at the amplicon ends.

Figure 4 – Automatic Normalization Plates

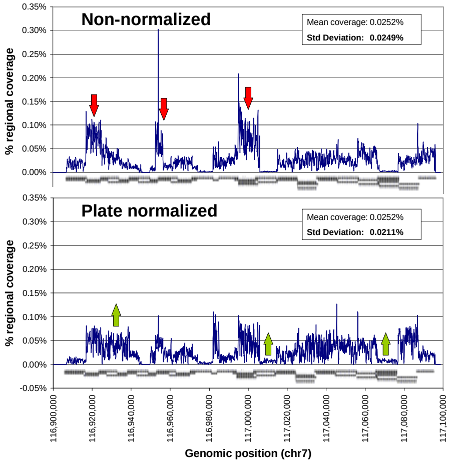


Figure 4. Comparison of sequencing coverage from amplicons (non-blocked) pooled without regard to yield or automatically normalized using a SequalPrep™ Normalization Plate prior to pooling. Coverage as a percent of total coverage across the region is displayed for 150 bp windows sliding in 50 bp increments. Plate normalized samples show more even coverage as a whole (reduced standard deviation). The most overrepresented sequences in the original amplicons are suppressed (red arrows), while regions of low representation are enhanced (green arrows). Although several regions which were apparently at very low initial concentrations could not be completely rescued, an increase in coverage for those regions is evident after normalization.

Discussion

While second generation DNA sequencing technology has dramatically reduced the cost of genome resequencing, it is still more cost effective to use long range PCR to first amplify a region of interest (up to several Mbp) above the remaining genome background when performing medical resequencing of multiple patient samples. We have shown that a single spot of an octet slide on a SOLiD™ instrument running V2 chemistry produces excess coverage (> 7M mapped 35 bp reads) of a 190 kb region, allowing 16 physically separate samples to be sequenced in a single run on both of the instrument’s flow cells. Because the SOLiD™ 3 instrument promises to deliver significantly higher throughput, it will be possible to increase both the region size and the number of samples several fold by employing DNA barcodes to index pooled samples.

The SequalPrep™ Long PCR kits were designed to reduce the reagent costs, labor, and time normally required to optimize long range PCR. We show here that 39 of 42 targets ~9-11 kb in size were successfully amplified on the first try. In addition, using 5’ amine-blocked primers reduced waste of sequencing capacity from oversampling of the amplicon ends, while plate-based normalization suppressed overrepresented and enhanced underrepresented amplicons. Together with the massively parallel sequencing afforded by second generation sequencers, these tools will enable cost-effective medical resequencing on a large scale.

Reference

1. Pritchard, J. K. & Cox, N. J. (2002) Hum. Mol. Gen. 11:2417

For additional product information on the SequalPrep™ System and E-Gel® SizeSelect 2%, go to www.invitrogen.com/ngs