Nested Patch PCR enables highly multiplexed mutation discovery in candidate genes

Katherine Elena Varley and Robi David Mitra[1]

Department of Genetics, Center for Genome Sciences, Washington University School of Medicine, St. Louis, Missouri 63108, USA

Medical resequencing of candidate genes in individual patient samples is becoming increasingly important in the clinic and in clinical research. Medical resequencing requires the amplification and sequencing of many candidate genes in many patient samples. Here we introduce Nested Patch PCR, a novel method for highly multiplexed PCR that is very specific, can sensitively detect SNPs and mutations, and is easy to implement. This is the first method that couples multiplex PCR with sample-specific DNA barcodes and next-generation sequencing to enable highly multiplex mutation discovery in candidate genes for multiple samples in parallel. In our pilot study, we amplified exons from colon cancer and matched normal human genomic DNA. From each sample, we successfully amplified 96% (90 of 94) targeted exons from across the genome, totaling 21.6 kbp of sequence. Ninety percent of all sequencing reads were from targeted exons, demonstrating that Nested Patch PCR is highly specific. We found that the abundance of reads per exon was reproducible across samples. We reliably detected germline SNPs and discovered a colon tumor specific nonsense mutation in *APC*, a gene causally implicated in colorectal cancer. With Nested Patch PCR, candidate gene mutation discovery across multiple individual patient samples can now utilize the power of second-generation sequencing.

[Supplemental material is available online at www.genome.org.]

As the genes involved in various aspects of human physiology are elucidated, there are increasingly more candidate genes associated with disease. The application of this knowledge in the clinic and clinical research can be very powerful as we move toward personalized medicine. Examples of success include the sequencing of candidate disease loci in targeted populations, such as Ashkenazi Jews (Weinstein 2007), the sequencing of variants in drug metabolism genes to adjust dosage (Marsh and McLeod 2006), and the identification of genetic defects in cancer that make tumors more responsive to certain treatments (Marsh and McLeod 2006). However, the sequencing of many candidate genes across many individual samples necessitates the development of new technology to lower the cost and increase the throughput of medical resequencing to make clinical application more feasible.

The cost of sequencing is declining rapidly due to second-generation sequencing technologies that perform a large number of sequencing reactions in parallel while using a small amount of reagent per reaction (Metzker 2005). These technologies integrate cloning and amplification into the sequencing protocol, which is essential for achieving the greater than 100-fold cost savings over traditional methods. However, this integration results in a loss of flexibility—it is not yet feasible to sequence a subset of the human genome in a large number of samples for the same cost as sequencing the complete genome of a single individual. This is a limitation, because sequencing the complete genome of a large numbers of individuals is still cost prohibitive, and the whole genome sequence of only a few individuals does not provide enough statistical power to make correlations between genotype and phenotype. The promise of personalized medicine based on genome analysis still glows on the horizon, but the significance behind observed variability is dim without an affordable technology to drive the necessary depth of patient sampling.

Current methods for analyzing sequence variation in a subset of the human genome rely on PCR to amplify the targeted sequences (Sjoblom et al. 2006; Greenman et al. 2007; Wood et al. 2007). Efforts to multiplex PCR have been hampered by the dramatic increase in the amplification of mispriming events as more primer pairs are used (Fan et al. 2006). In addition, a large number of primer pairs often result in interprimer interactions that prevent amplification (Han et al. 2006). Therefore, separate PCRs for each region of interest are performed, a costly approach when hundreds of individual PCRs must be performed for each sample (Sjoblom et al. 2006; Greenman et al. 2007; Wood et al. 2007). Furthermore, this strategy requires a large amount of starting DNA to supply enough template for all of the individual PCR reactions; this can be a problem as DNA is often a limiting factor when working with clinical samples.

It is important to choose the appropriate strategy for sample tracking to fully harness the throughput of second-generation sequencing technologies. The sequencing capacities of these platforms are large enough so that multiple samples can be sequenced with a single instrument run. To do this, one can use a separate compartment for each sample, but this only allows for a small number of samples, and there is a reduction in the total amount of sequence generated per run. Recently, Parameswaran et al. (2007) demonstrated the power of using DNA barcodes to label samples so that they can be pooled and sequenced together on the 454 Life Sciences (Roche) GS20 Sequencer. They were able to utilize the full capacity of the instrument and still determine from which sample each read originated. To realize the full power of second-generation sequencing technologies, a multiplexing strategy should be compatible with DNA barcoding to track samples.

[1]Corresponding author.
E-mail rmitra@genetics.wustl.edu; fax (314) 362-2156.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.078204.108.

18:1844–1850 ©2008 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/08; www.genome.org

Here we present a novel method, Nested Patch PCR, that simultaneously amplifies many targeted regions from human genomic DNA. We demonstrate that Nested Patch PCR provides an effective "front-end" technology for sequencing candidate genes using second-generation sequencing. Nested Patch PCR provides an easy workflow for highly multiplexed PCR that is very specific and sensitive for identifying SNPs and mutations in individual samples. We amplified and sequenced 90 human exons from genomic DNA and discovered SNPs and a tumor-specific mutation. We demonstrate how to couple Nested Patch PCR with DNA barcodes to multiplex both the target selection and the patient samples for a single sequencing run. This method promises to be an efficient strategy for the numerous studies that perform sequencing of candidate genes in many samples. Even as the cost of whole genome sequencing declines, selectively sequencing a fraction of the genome will continue to be a fraction of the cost, enabling deeper sampling of the population.

Results

Overview of Nested Patch PCR

We designed Nested Patch PCR to be robust against the mispriming events that are typical of standard multiplex PCR. Nested Patch PCR requires four oligonucleotide hybridizations per locus, resulting in a more specific amplification than standard multiplex PCR, which requires only two hybridizations per locus. Nested Patch PCR begins with a PCR reaction containing two primers for each target (Fig. 1A). These DNA primers contain uracil substituted for thymine. The PCR is performed for a low number of cycles and serves to define the ends of the target regions. The primers are then cleaved from the amplicons by the addition of an enzyme mix containing uracil DNA glycosylase. The ends of the target regions are now internal to the PCR primer sequences (Fig. 1B). Next a second round of selection is performed. Nested Patch oligonucleotides are annealed to the target amplicons and serve as a patch between the correct amplicons and universal primers (Fig. 1C). The universal primers are then ligated to the amplicons. This reaction is highly specific because thermostable ligases are sensitive to mismatched bases near the ligation junction (Barany 1991). An added level of selectively is gained by degrading mispriming products as well as the genomic DNA with exonuclease. The selected amplicons are protected from degradation by a 3′ modification on the universal primer. The selected amplicons are then amplified together simultaneously by PCR with the universal primers (Fig. 1D). The target selection protocol is an addition-only reaction and can be performed in a single tube per sample, making it amenable to automation.

To pool and sequence multiple samples, Nested Patch PCR is first performed separately for each sample (one tube per sample). Sample-specific DNA barcodes are then incorporated into the primers used for the final universal PCR by tailing the 5′ end with sample-specific DNA sequences and 454 sequencing primers (Fig. 2). Thus, the first few bases indicate from which sample each read originated.

Nested Patch PCR and sequencing of candidate genes in colon cancer

To demonstrate the multiplexed selection and amplification of exons by Nested Patch PCR, we designed oligonucleotides for 94 exons from six genes that cause cancer when mutated in the germline (*TP53*, *APC*, *MLH1*, *RB1*, *BRCA1*, *VHL*) (Marsh and Zori 2002). These exons are located across four chromosomes, vary in length from 74 to 438 bp, and total 21.6 kbp. We performed Nested Patch PCR using genomic DNA isolated from a moderately differentiated colon adenocarcinoma and from the adjacent normal tissue in side-by-side reactions. We incorporated a 6-bp sample-specific DNA barcode, pooled the two samples, and sequenced the pool using the 454 FLX sequencer. We obtained 55,058 reads and mapped these to the human genome. We were able to map at least one read from each sample to 90 of the 94 exons (95.7%). The four exons that failed to amplify were due to imperfect primer/patch design. We were unable to amplify two of the loci in separate individual PCR reactions, indicating PCR primer failure. The other two loci failed because their Patch oligos bound to multiple locations in the genome. In the future, we will design our primers and patches to avoid this problem.

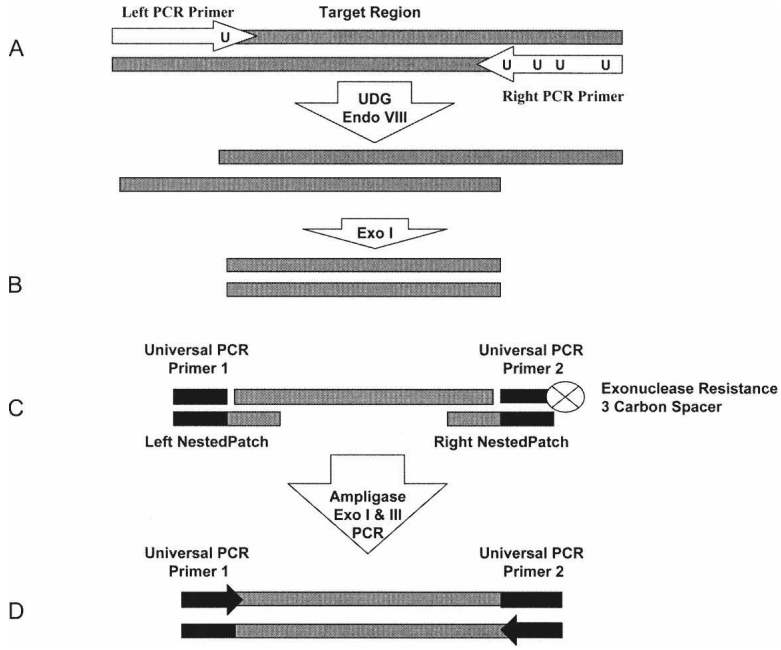

Figure 1. Schematic of Nested Patch PCR. (*A*) A PCR reaction containing primers pairs for all targets is performed on genomic DNA. The primers contain uracil substituted for thymine. The primers are then cleaved from the amplicons by the addition of heat-labile uracil DNA glycosylase, endonuclease VIII, and single-strand-specific exonuclease I. (*B*) The ends of the target regions are now internal to the PCR primers (nested). (*C*) Nested Patch oligonucleotides are annealed to the target amplicons and serve as a patch between the correct amplicons and universal primers. The universal primers are then ligated to the amplicons. The universal primer on the 3′ end of the amplicon is modified with a three carbon spacer that protects the selected amplicon from the final exonuclease reaction that degrades nonspecific products. (*D*) The selected amplicons are then amplified together simultaneously by PCR with universal primers.

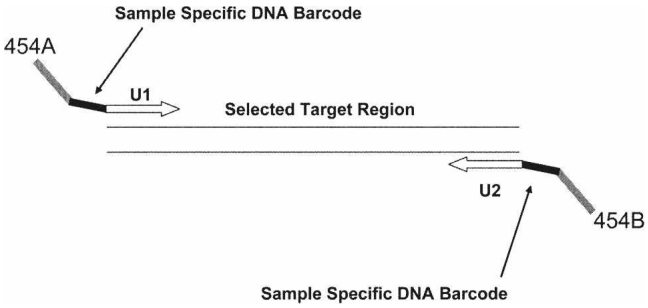

Figure 2. Schematic of sample-specific barcode PCR. Sample-specific DNA barcodes are incorporated into the primers that are used for the final universal PCR. The 5′ end of the universal primer (white) is tailed with the sequences for the 454 FLX machine (gray) and sample-specific DNA sequences (black). When sequencing from either 454A or 454B, the first few bases indicate from which sample the read originated.

Ninety percent of all reads (49,553 reads) mapped to one of the targeted exons. Thus, with Nested Patch PCR, we achieved a 125,000-fold enrichment from genomic DNA (90% specificity × total possible fold enrichment). When selecting a fraction of the genome this small, the total possible enrichment is 138,888-fold (3×10^9 bp genome/21.6 kbp targeted). Of the remaining 10% of reads that did not match the targeted regions, most (85%) appear to be due to concatamers of Nested Patch oligonucleotides that contain *Alu* elements. It is likely that this background could be reduced by designing oligonucleotides that do not overlap repetitive genomic elements. These results demonstrate that with Nested Patch PCR we successfully amplified 90 exons simultaneously and that the reaction is highly specific.

Abundance of exons and reproducibility

To analyze the uniformity of the sequence coverage, we graphed the number of reads obtained for each targeted exon (Fig. 3A; Supplemental Table 1). Sequence coverage ranged over 2–3 logs (base 10), with 75% (68/90) of exons having between 10 and 500 reads in both samples (50-fold abundance range). The median number of reads per exon is 145. Seventy-six percent of all exons fell within fivefold coverage of this median (29–725 reads). Exon nonuniformity did not correlate with the gene, the size of the amplicon, or the GC content of the oligos. We have not found a parameter that explains the nonuniformity. In the future, it may be possible to achieve a more uniform coverage by grouping exons that are efficiently amplified in one reaction and by grouping exons that are amplified less efficiently in a separate reaction. This strategy would require the amplification efficiency for each exon to be reproducible. To test the reproducibility, we correlated the number of reads per exon from the tumor and normal samples. The correlation was high (R^2 of 93%), indicating high reproducibility (Fig. 3B). In fact, 85% (77/90) of exons displayed at most a twofold difference in abundance between samples, and all exons were within threefold relative abundance between samples (Fig. 3C). These results demonstrate that even though the abundance varies between exons, the abundance of each exon is reproducible across different reactions and samples.

SNP and mutation discovery

We identified seven variants from the reference sequence in our samples (Table 1). We validated the SNPs and mutations

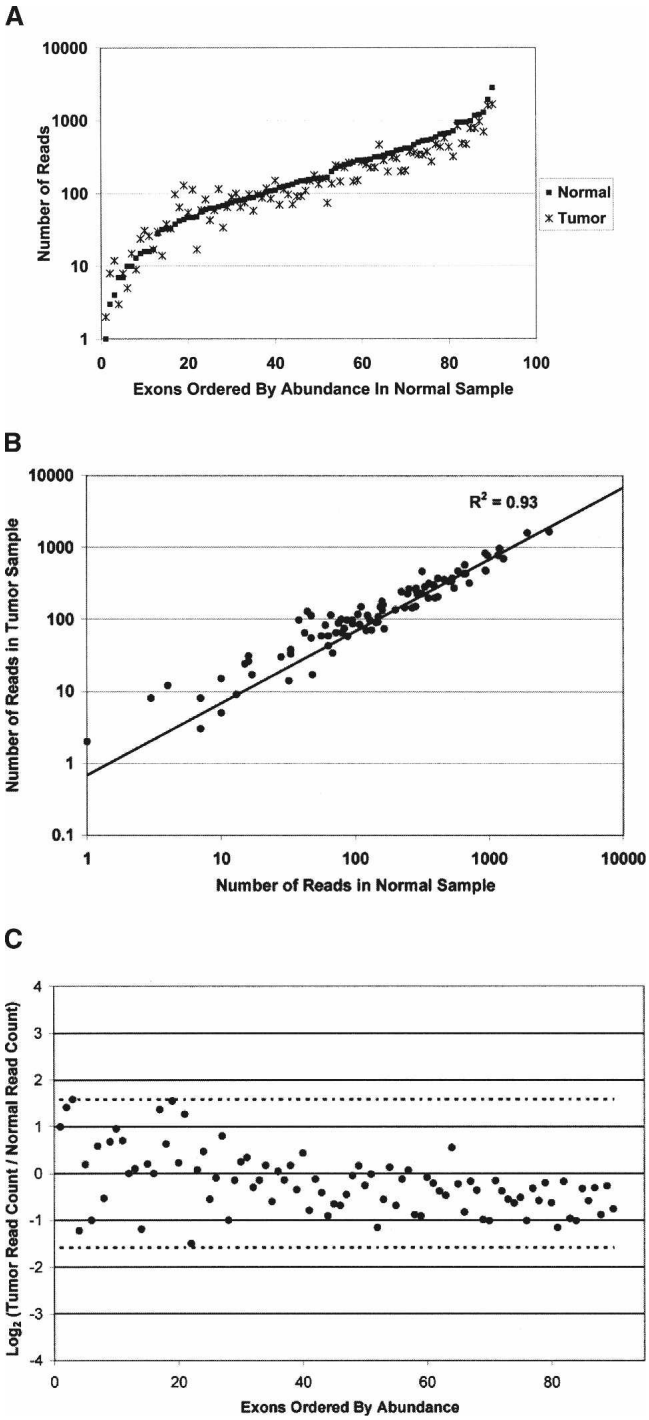

Figure 3. Quantification of the abundance and reproducibility of Nested Patch PCR per exon in each sample. (*A*) Uniform exon abundance. Graph of the number of reads obtained for each targeted exon from the colon cancer sample and adjacent normal tissue. The 90 exons for which we obtained at least one read are ordered by abundance in the normal sample on the *x*-axis. The median number of reads per exon is 145. Seventy-six percent of all exons fell within fivefold coverage of this median. All exons are within 3 $\log_{10}$ of each other. (*B*) Correlation of number of reads across samples. Each exon is depicted as a point on the graph, where the *x*-axis is the number of reads in the normal sample and the *y*-axis is the number of reads in the colon cancer sample. The correlation was high (R^2 of 93%), indicating high reproducibility across samples. (*C*) Fold difference in abundance across samples. We computed the fold change of abundance per exon between the two samples. 85% (77/90) of exons displayed a twofold or less difference in abundance between samples. One hundred percent of exons displayed a threefold or less difference in abundance between samples. Dotted line indicates threefold change [$\log_2(3)$].

Table 1. Mutation and SNPs discovered

Gene symbol	RefSeq ID	Exon no.	Location[a]	Reference base	Variant	Amino acid change	Fraction of reads with variant: Colon adenocarcinoma tissue		Fraction of reads with variant: Adjacent normal tissue	
APC	NM_000038	10	rs2229992	T	C	None	143/301	48%	222/468	47%
APC	NM_000038	12	rs351771	G	A	None	37/68	54%	43/79	54%
APC	**NM_000038**	**12**	**chr5:112192485**	**C**	**T**	**Arg → STOP**	**23/68**	**33%**	**3/80**	**4%**
APC	NM_000038	13	rs62626346[b]	T	C	Intronic	17/29	59%	27/50	54%
TP53	NM_000546	1	rs17883323	G	T	Intronic	41/41	100%	50/50	100%
RB1	NM_000321	11	rs185587	G	T	Intronic	79/79	100%	102/102	100%
RB1	NM_000321	24	rs3020646	C	T	Intronic	24/24	100%	18/18	100%

Mutation in bold is tumor-specific.
[a]Location is according to the March 2006 human genome assembly from the UCSC Genome Browser.
[b]Novel germline SNP.

identified by Nested Patch PCR and 454 FLX sequencing by performing individual PCR reactions from the original patient samples, cloning the amplicons, and sequencing at least eight clones per locus using standard Sanger sequencing. Five of these variants were already in dbSNP (http://www.ncbi.nlm.nih.gov/SNP/). The individual we sequenced was germline homozygous at three of these SNPs (rs17883323, rs185587, rs3020646) and was germline heterozygous at two other SNPs in the database, rs2229992 and rs351771. The C allele of the SNP rs2229992 was in 48% of reads from the tumor sample and 47% of reads from normal sample. The A allele of the SNP rs351771 was in 54% of reads from the tumor sample and 54% of reads from normal sample. The ability to detect both alleles of these known polymorphisms at near equal frequency indicates that Nested Patch PCR provides high allele sensitivity that is reproducible across samples. We also discovered a SNP in an intron of *APC* that was not yet in dbSNP (rs62626346). The individual we sequenced was heterozygous in both the tumor and normal samples at this intronic position. We discovered a novel germline SNP in the individual we sequenced in one of the most extensively surveyed genes, *APC*. This illustrates that medical resequencing of well-characterized candidate genes will yield more insight into genetic variation in individuals.

We discovered a tumor-specific nonsense mutation. It is a C-to-T substitution in the *APC* gene at chr5:112192485 that results in a codon for arginine changing to a stop codon. This is likely a significant mutation in this individual's colon tumor because it is a nonsense mutation in a gene that is already known to cause colon cancer. This mutation was in 33% of reads from the tumor sample. This mutation is adjacent to a heterozygous SNP, and we discovered that 62% of the SNP A allele reads had the nonsense mutation, and 0% of the SNP G allele reads had the nonsense mutation. This indicates that the nonsense mutation occurred on the A allele during the clonal expansion of the tumor. This mutation was previously observed in an ovarian endometrioid adenocarcinoma and is mutation ID no. 19040 in the Catalog of Somatic Mutations in Cancer (http://www.sanger.ac.uk/genetics/CGP/cosmic/).

To determine the sensitivity of the method for SNP discovery, we performed individual PCR reactions for each of our target exons, followed by direct Sanger sequencing of these PCR products. We then used both PolyPhred and manual inspection to identify variants in the sequence traces (Nickerson et al. 1997). We did not find any additional SNPs in the adjacent normal colon DNA sample beyond the six germline SNPs already identified. Thus, in this experiment, the sensitivity of the method is 100%.

Discussion

We have developed Nested Patch PCR, a novel method for multiplex PCR that is capable of amplifying 90 exons from genomic DNA simultaneously. The target selection protocol is an addition-only reaction and can be performed in a single tube per sample, making it amenable to automation. We demonstrated that Nested Patch PCR can be performed on multiple samples in parallel, which can then be labeled with sample-specific DNA barcodes and sequenced as a pool. The choice of targets and target boundaries is flexible, and a wide range of sizes can be amplified simultaneously (here, 74 bp to 438 bp). We performed Nested Patch PCR on genomic DNA and obtained sequence for 90 of the 94 exons we targeted, indicating this method is robust. This method is also highly specific: 90% of reads matched targeted exons. The number of reads per exon was highly correlated across samples indicating high reproducibility. We identified both alleles of heterozygous SNPs at near-even frequencies across samples, demonstrating that this method has the allele sensitivity necessary for variant discovery in personal genome sequencing. Furthermore, we demonstrate the applied utility of this method by discovering a colon tumor-specific mutation in an individual.

Recently, several new methods have been developed for the multiplexed selection, amplification, and sequencing of genomics subsets (Bashiardes et al. 2005; Dahl et al. 2005, 2007; Albert et al. 2007; Fredriksson et al. 2007; Hodges et al. 2007; Meuzelaar et al. 2007; Okou et al. 2007; Porreca et al. 2007). Several of these methods achieve higher levels of multiplexing (Albert et al. 2007; Dahl et al. 2007; Hodges et al. 2007; Okou et al. 2007; Porreca et al. 2007), but they do not perform as well in other areas, such as the precise definition of target boundaries (Albert et al. 2007; Dahl et al. 2007; Hodges et al. 2007; Okou et al. 2007), the reproducible capture of some target regions (Porreca et al. 2007), or the fraction of reads matching target sequences (Albert et al. 2007; Hodges et al. 2007; Okou et al. 2007). In this proof-of-principle study, we did not determine the upper limit of the number of target sequences that can be amplified by Nested Patch PCR, making it difficult to directly compare our method to these technologies, particularly for applications where a high degree of multiplexing is required. However, Nested Patch PCR should prove useful for the amplification of an intermediate

number (100–1000) of candidate regions in a large number of samples. It is particularly well suited to these applications because it incorporates sample-specific DNA barcodes, allows for the precise definition of the boundaries of targeted sequences, is reproducible, is highly specific, and uniformly amplifies different alleles at a given locus. Indeed, the utility of Nested Patch PCR is best illustrated by the fact that we were able to discover a novel, cancer-specific mutation in our small pilot study.

To achieve a higher degree of multiplexing with Nested Patch PCR more oligos are needed and the cost of oligos (four oligos per target) becomes a consideration. A standard 100-nmole oligo synthesis of each primer is enough for 200,000 Nested Patch reactions. So, for the amplification of universal sets of candidate regions using Nested Patch PCR, the upfront cost of synthesizing oligos by standard methods may be economical when amortized over the hundreds of thousands of reactions that can be performed from one standard synthesis. Alternatively, for flexible design, Porreca et al. (2007) recently introduced a strategy to cost-effectively produce 55,000 oligos on a programmable microarray for use in solution-phase reactions.

Nested Patch PCR requires more sequencing than expected to cover the target regions due to the range in relative abundance of the different exons after amplification. The relative abundance of the different exons after amplification could be improved by partitioning similarly abundant targets into separate reactions, normalizing and pooling these reactions after amplification, as has been proposed elsewhere for other multiplex methods that suffer from similar or worse nonuniformity (Dahl et al. 2007; Porreca et al. 2007). This is an efficient approach for larger number of targets, so that each reaction is still highly multiplexed as long as the amplification efficiency is reproducible, as was found with Nested Patch PCR. Another approach to achieving a more uniform abundance of exons is to design multiple sets of oligos to select the same targeted region. The final abundance of a target would be the sum of reads generated by the different oligo sets, where some will be more efficient than others. This should normalize the relative abundance across all exons.

We anticipate that Nested Patch PCR will be useful for a variety of applications. Because the method is based on PCR, it will likely have the same sensitivity as PCR to detect pathogen DNA in a high background of host DNA (Elnifro et al. 2000; Akhras et al. 2007a,b) or to detect rare DNA biomarkers in peripheral samples (Fackler et al. 2006). Also, it is likely to have the sensitivity to amplify targets from degraded samples, an area for which there are no robust methods to allow for multiplexed or genome-wide amplification. Other applications that rely heavily on PCR may benefit from higher levels of multiplexing, such as the engineered assembly of many DNA fragments simultaneously in synthetic biology experiments (Reisinger et al. 2006; Forster and Church 2007).

Nested Patch PCR is a novel method for highly multiplexed PCR that performs with the high reproducibility, allele sensitivity, and specificity necessary for targeted resequencing of candidate regions to identify SNPs and mutations. This is the first method to demonstrate multiplexing of both samples and targets for next-generation sequencing by coupling multiplexed PCR and DNA barcode sample labeling. We have demonstrated the utility of Nested Patch PCR for discovering SNPs and tumor-specific mutations in individuals. This will be a useful method for selectively sequencing candidate regions in large cohorts of patients to identify variants associated with disease. Nested Patch PCR promises to improve many other methods that rely on the sensitivity of PCR and could benefit from higher multiplexing such as pathogen detection, biomarker detection in peripheral body fluids, and synthetic DNA assembly.

Methods

Design of oligonucleotides

Human exon sequence plus 150-bp flanking sequence from the March 2006 assembly was downloaded from the UCSC Genome Browser (www.genome.ucsc.edu) for the following RefSeq genes (NM_000038, NM_000546, NM_000249, NM_000321, NM_007304, NM_000551). We maintained the convention that exon numbering for each gene begins with zero throughout the analysis. Primer3 (http://frodo.wi.mit.edu/) was then used to select primers pairs flanking the exon. The design was constrained to PCR products between 50 and 500 bp, primer length 20–36 bp, and primer melting temperature (T_m) 61°C–67°C; the maximum difference in T_m between primer pairs was 5°C; and the GC content of the primer had to be between 10% and 80%. Four thousand possible primer pairs were generated per exon. Those primers pairs that ended with a T as the 3′ base were then selected. A Nested Patch oligo was then designed by extending into the sequence from the PCR primer until the T_m of Nested Patch oligo was 62°C–67°C. The selected oligos were then aligned against themselves using WUBLAST BLASTN to approximate their cross-reactivity (http://blast.wustl.edu). For each exon, the oligo sets with the fewest blastn matches to the entire set was chosen. The PCR primer sequence was substituted with a deoxyuridine in place of every deoxythymidine. The Nested Patch oligos were then concatenated with the complement universal primer sequences to result in the appropriate patch sequence. We attempted to design oligos to select and amplify all 96 exons from the chosen genes; however, two exons failed designed: The last exon of *APC* failed because of length (~6000 bp), and an exon in *RB1* failed due to the presence of *Alu* repeat elements surrounding the exon. Oligonucleotides were synthesized by SigmaGenosys (http://www.sigmaaldrich.com/Brands/Sigma_Genosys.html). Two universal primer sequences were synthesized by IDT (www.idtdna.com), including the Universal 2, which has a 5′ phosphate and a 3 carbon spacer on the 3′ end. Oligo sequences are listed in Supplemental Material 1.

Nested Patch PCR

Genomic DNA from a moderately differentiated colon adenocarcinoma primary tumor and adjacent normal tissue from an 81-yr-old male was obtained from Biochain (www.biochain.com), catalog no. D8235090-PP-10. Targets were initially amplified by PCR containing 1 µg human genomic DNA, 50 nM each of 94 forward PCR primers, 50 nM each of 94 reverse PCR primers, 5 U of AmpliTaq Polyermase Stoffel Fragment (Applied Biosystems), 200 µM each dNTP, 2 mM $MgCl_2$, 20 mM Tris-HCl (pH 8.4), and 50 mM KCl in a total volume of 10 µL. This reaction was incubated for 2 min at 94°C followed by (30 sec at 94°C, 30 sec at 56°C, 6 min at 72°C) × 10 cycles and was then held at 4°C.

Primers were cleaved from the amplicons by the addition of 1 U of heat labile uracil-DNA glycosylase (USB), 10 U of endonuclease VIII (NEB), and 10 U of exonulcease I (USB). This mix was incubated for 2 h at 37°C followed by heat inactivation for 20 min at 95°C and was held at 4°C. To remove the unincorporated nucleotide from the mix, 0.05 U of Apyrase (NEB) was added to the reaction and incubated for 30 min at 30°C.

Nested Patch driven ligation of the universal primers to correct amplicons is performed by addition of more reactants to the initial tube to result in the following final concentrations: 20 nM

each Nested Patch oligo, 40 nM universal primer 1, 40 nM universal primer 2 with 5′ phosphate and 3′ three carbon spacer, 5 U of Ampligase (Epicentre), and 1× Ampligase reaction buffer (Epicentre) in a total volume of 25 µL. This reaction was incubated for 15 min at 95°C followed by (30 sec at 94°C, 2 min at 65°C, 1 min at 55°C, 5 min at 60°C) for 100 cycles and was held at 4°C.

Incorrect products, template genomic DNA and excess primer were degraded by the addition of 10 U of exonuclease I (USB) and 200 U of exonuclease III (Epicentre). This mix was incubated for 2 h at 37°C followed by heat inactivation for 20 min at 95°C and was held at 4°C.

Sample-specific DNA barcode universal PCR

Each selection reaction was purified using a Qiaquick Spin Column (Qiagen), and the final elution was performed with 30 µL of EB. For the PCR, we added reagents to the elution to result in these final concentrations in 50 µL: 0.5 µM each tailed universal primer (see below), 10 U of Platinum Taq polymerase (Invitrogen), 0.5 mM each dNTP, 2 mM $MgCl_2$, 0.5 M betaine, 20 mM Tris-HCl (pH 8.4), and 50 mM KCl. This reaction was incubated at 93°C for 2 min followed by (30 sec at 93°C, 6 min at 60°C) for 27 cycles and was held at 4°C. The universal PCR used primers tailed with 454 Life Sciences A or B oligo at the 5′ end, followed by a sample-specific DNA sequence and ending at the 3′ end with the same universal primer sequence ligated to the amplicons in the Nested Patch PCR procedure. The PCR product smear between the expected sizes was confirmed by running on a 3% Metaphor Agarose gel (Lonza). The reactions were then purified on a Qiaquick Spin Column (Qiagen). The eluted DNA was quantified on the Nanodrop (www.nanodrop.com), and the same quantity of DNA was pooled together from the two separate samples. This pooled sample was submitted to Cogenics Inc. (www.cogenics.com) for sequencing on the 454 Life Sciences (Roche) FLX machine.

Sequencing data analysis

We obtained 55,068 sequencing reads from Cogenics Inc. (www.cogenics.com). To determine which sequences matched our targets, we aligned the reads against a database of reference target sequences for each target using BLASTN (http://blast.wustl.edu). We then determined how many reads matched significantly to each exon ($P < 0.02$). We identified whether each read came from the tumor sample or the normal sample based on the first six bases of sequence, which was the sample-specific DNA barcode. To determine the reproducibility of the method, we computed the relative fold change of read counts per exon between the colon cancer and adjacent normal samples (higher read count/lower read count). We then determined how many reads did not match targeted sequence and aligned them to a database of Nested Patch oligo sequence to identify what fraction was due to primer artifacts. For each exon, we used CLUSTALW to generate a multiple sequence alignments of all of the reads against the reference sequence (Larkin et al. 2007). The majority of the differences from the reference sequence were indels adjacent to homopolymers, which is a known error-prone feature for 454 sequencing (Ronaghi et al. 1998). To filter these out, we examined all the positions that did not match the reference sequence but were in greater than 30% of the reads. We then used the UCSC Genome browser to determine whether these variants were in dbSNP (www.ncbi.nlm.nih.gov/projects/SNP/index.html) and whether they disrupted a codon. To determine if the tumor specific mutation we identified had been previously reported, we searched the Catalog of Somatic Mutations in Cancer (www.sanger.ac.uk/genetics/CGP/cosmic/).

SNP and mutation validation by cloning and Sanger sequencing

We validated the variants from the reference sequence identified by Nested Patch PCR and 454 FLX sequencing by performing individual PCRs for each variant locus, cloning the amplicons into *Escherichia coli*, and sequencing 12 clones for each variant as follows. The PCR for each locus in each sample was performed in a total volume of 50 µL. The reaction contained 1× PCR buffer—$MgCl_2$ (Invitrogen), 10 U of Platinum Taq polymerase (Invitrogen), 0.5 mM each dNTP, 0.5 M betaine, 0.5 µM forward primer, 0.5 µM reverse primer, and 100 ng genomic DNA from either the colon tumor or the adjacent normal tissue (Biochain, catalog no. D8235090-PP-10). This reaction was incubated for 2 min at 93°C, followed by (30 sec at 93°C, 6 min at 55°C) × 30 cycles and was held at 4°C. One fifth of the PCR reaction was verified by electrophoresis on a 2% agarose gel. To clone the PCR products, we ligated them into the pGEM-T Easy Vector using Rapid Ligation Buffer according to the manufacturer's instructions (Promega). We then transformed the ligated vector into GC10 competent cells (Gene Choice) and grew them overnight on LB-agar (Luria-Broth) plates containing standard concentrations of carbenicillin, X-gal, and IPTG. After overnight growth, colonies were picked from the plates and added to 50-µL colony PCR reactions containing 1× PCR reaction buffer (Sigma), 2 U of Jumpstart Taq polymerase (Sigma), 0.2 mM each dNTP, 0.5 µM M13 forward primer (5′-CGCCAGGGTTTTCCCAGTCACGAC-3′), 0.5 µM M13 reverse primer (5′-TCACACAGGAAACAGCTATGAC-3′), and 0.01% Tween. The reaction was incubated for 10 min at 94°C, followed by (1 min 30 sec at 94°C, 1 min at 55°C, 1 min at 72°C) × 35 cycles and was held at 4°C. These reactions were then treated with 10 µL of ExoSAP to degrade the remaining primers and nucleotides by adding 0.2 U of Exonuclease I (USB) and 0.2 U of shrimp alkaline phosphatase (SAP) (Promega) in 1× SAP buffer (Promega), incubating for 30 min at 37°C and then 30 min at 80°C. The Sanger sequencing/cycle sequencing reactions were 20 µL and contained 1.5 µL of Exo-SAP-treated colony PCR, 1 µL of Big Dye Terminator v3.1 RR-100 Mix (Applied Biosystems), 2 mM $MgCl_2$, and 0.16 µM M13 forward primer. They were incubated for 1 min at 96°C, followed by (10 sec at 96°C, 5 sec at 50°C, 4 min at 60°C) × 24 cycles and were held at 4°C. The reactions were ethanol precipitated with sodium acetate and submitted to the Washington University Genome Sequencing Center to load on the ABI 3730 (Applied Biosystems). Trace files were analyzed using *phred* (Ewing and Green 1998; Ewing et al. 1998), and the resulting sequencing reads were aligned to the reference sequence using BLAT on the UCSC Genome Browser (Kent 2002; Kent et al. 2002).

SNP sensitivity analysis by Sanger sequencing

We determined if our method failed to detect any SNPs present in the sample by performing individual PCRs and direct Sanger sequencing of the PCR products for each exon on gDNA from the same samples as the novel method described herein. The PCR for each locus in each sample was performed in a total volume of 50 µL. The reaction contained 1× PCR buffer—$MgCl_2$ (Invitrogen), 5 U of Platinum Taq polymerase (Invitrogen), 0.5 mM each dNTP, 0.5 M betaine, 0.5 µM locus-specific forward primer, 0.5 µM locus-specific reverse primer, and 20 ng genomic DNA from the adjacent normal tissue (Biochain, catalog no. D8235090-PP-10). This reaction was incubated for 2 min at 93°C, followed by (30 sec at 93°C, 6 min at 55°C) × 30 cycles and was held at 4°C. One fifth of the PCR reaction was verified by electrophoresis on a 2% agarose gel. These reactions were then treated with 10 µL of ExoSAP to degrade the remaining primers and nucleotides by

adding 0.2 U of Exonuclease I (USB) and 0.2 U of SAP (Promega) in 1× SAP buffer (Promega), incubating for 30 min at 37°C and then 30 min at 80°C. The Sanger sequencing/cycle sequencing reactions were 20 µL and contained 1.5 µL of Exo-SAP–treated individual exon PCR, 1 µL of Big Dye Terminator v3.1 RR-100 Mix (Applied Biosystems), 2 mM $MgCl_2$, and 0.16 µM forward or reverse PCR primer. They were incubated for 1 min at 96°C, followed by (10 sec at 96°C, 5 sec at 50°C, 4 min at 60°C) × 24 cycles and were held at 4°C. The reactions were ethanol precipitated with sodium acetate and submitted to the Washington University Genome Sequencing Center to load on the ABI 3730 (Applied Biosystems). Trace files from both forward and reverse reads were analyzed for SNPs using PolyPhred and manual inspection (Nickerson et al. 1997).

Acknowledgments

We thank Jason Gertz and Lee Tessler for suggestions, discussion, and critical reading of the manuscript. We thank Melissa Fuller, Barak Cohen, Todd Druley, German Leparc, Michael Brooks, and Yue Yun for helpful discussion. This work was supported by the Genome Analysis Training Program (T32 HG000045) and a Center for Excellence in Genome Sciences grant from the National Human Genome Research Institute (5P50HG003170-03).

References

Akhras, M.S., Thiyagarajan, S., Villablanca, A.C., Davis, R.W., Nyren, P., and Pourmand, N. 2007a. PathogenMip assay: A multiplex pathogen detection assay. *PLoS ONE* **2:** e223. doi: 10.1371/journal.pone.0000223.

Akhras, M.S., Unemo, M., Thiyagarajan, S., Nyren, P., Davis, R.W., Fire, A.Z., and Pourmand, N. 2007b. Connector inversion probe technology: A powerful one-primer multiplex DNA amplification system for numerous scientific applications. *PLoS ONE* **2:** e915. doi: 10.1371/journal.pone.0000915.

Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J., et al. 2007. Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* **4:** 903–905.

Barany, F. 1991. Genetic disease detection and DNA amplification using cloned thermostable ligase. *Proc. Natl. Acad. Sci.* **88:** 189–193.

Bashiardes, S., Veile, R., Helms, C., Mardis, E.R., Bowcock, A.M., and Lovett, M. 2005. Direct genomic selection. *Nat. Methods* **2:** 63–69.

Dahl, F., Gullberg, M., Stenberg, J., Landegren, U., and Nilsson, M. 2005. Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic Acids Res.* **33:** e71. doi: 10.1093/nar/gni070.

Dahl, F., Stenberg, J., Fredriksson, S., Welch, K., Zhang, M., Nilsson, M., Bicknell, D., Bodmer, W.F., Davis, R.W., and Ji, H. 2007. Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proc. Natl. Acad. Sci.* **104:** 9387–9392.

Elnifro, E.M., Ashshi, A.M., Cooper, R.J., and Klapper, P.E. 2000. Multiplex PCR: Optimization and application in diagnostic virology. *Clin. Microbiol. Rev.* **13:** 559–570.

Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using *phred*. II. Error probabilities. *Genome Res.* **8:** 186–194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using *phred*. I. Accuracy assessment. *Genome Res.* **8:** 175–185.

Fackler, M.J., Malone, K., Zhang, Z., Schilling, E., Garrett-Mayer, E., Swift-Scanlan, T., Lange, J., Nayar, R., Davidson, N.E., Khan, S.A., et al. 2006. Quantitative multiplex methylation-specific PCR analysis doubles detection of tumor cells in breast ductal fluid. *Clin. Cancer Res.* **12:** 3306–3310.

Fan, J.B., Chee, M.S., and Gunderson, K.L. 2006. Highly parallel genomic assays. *Nat. Rev.* **7:** 632–644.

Forster, A.C. and Church, G.M. 2007. Synthetic biology projects in vitro. *Genome Res.* **17:** 1–6.

Fredriksson, S., Baner, J., Dahl, F., Chu, A., Ji, H., Welch, K., and Davis, R.W. 2007. Multiplex amplification of all coding sequences within 10 cancer genes by Gene-Collector. *Nucleic Acids Res.* **35:** e47. doi: 10.1093/nar/gkm078.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., et al. 2007. Patterns of somatic mutation in human cancer genomes. *Nature* **446:** 153–158.

Han, J., Swan, D.C., Smith, S.J., Lum, S.H., Sefers, S.E., Unger, E.R., and Tang, Y.W. 2006. Simultaneous amplification and identification of 25 human papillomavirus types with Templex technology. *J. Clin. Microbiol.* **44:** 4157–4162.

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J., et al. 2007. Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* **39:** 1522–1527.

Kent, W.J. 2002. BLAT—The BLAST-like alignment tool. *Genome Res.* **12:** 656–664.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res.* **12:** 996–1006.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23:** 2947–2948.

Marsh, S. and McLeod, H.L. 2006. Pharmacogenomics: From bedside to clinical practice. *Hum. Mol. Genet.* **15:** R89–R93.

Marsh, D. and Zori, R. 2002. Genetic insights into familial cancers—Update and recent discoveries. *Cancer Lett.* **181:** 125–164.

Metzker, M.L. 2005. Emerging technologies in DNA sequencing. *Genome Res.* **15:** 1767–1776.

Meuzelaar, L.S., Lancaster, O., Pasche, J.P., Kopal, G., and Brookes, A.J. 2007. MegaPlex PCR: A strategy for multiplex amplification. *Nat. Methods* **4:** 835–837.

Nickerson, D.A., Tobe, V.O., and Taylor, S.L. 1997. PolyPhred: Automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.* **25:** 2745–2751.

Okou, D.T., Steinberg, K.M., Middle, C., Cutler, D.J., Albert, T.J., and Zwick, M.E. 2007. Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods* **4:** 907–909.

Parameswaran, P., Jalili, R., Tao, L., Shokralla, S., Gharizadeh, B., Ronaghi, M., and Fire, A.Z. 2007. A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Res.* **35:** e130. doi: 10.1093/nar/gkm760.

Porreca, G.J., Zhang, K., Li, J.B., Xie, B., Austin, D., Vassallo, S.L., LeProust, E.M., Peck, B.J., Emig, C.J., Dahl, F., et al. 2007. Multiplex amplification of large sets of human exons. *Nat. Methods* **4:** 931–936.

Reisinger, S.J., Patel, K.G., and Santi, D.V. 2006. Total synthesis of multi-kilobase DNA sequences from oligonucleotides. *Nat. Protoc.* **1:** 2596–2603.

Ronaghi, M., Uhlen, M., and Nyren, P. 1998. A sequencing method based on real-time pyrophosphate. *Science* **281:** 363–365.

Sjoblom, T., Jones, S., Wood, L.D., Parsons, D.W., Lin, J., Barber, T.D., Mandelker, D., Leary, R.J., Ptak, J., Silliman, N., et al. 2006. The consensus coding sequences of human breast and colorectal cancers. *Science* **314:** 268–274.

Weinstein, L.B. 2007. Selected genetic disorders affecting Ashkenazi Jewish families. *Fam. Community Health* **30:** 50–62.

Wood, L.D., Parsons, D.W., Jones, S., Lin, J., Sjoblom, T., Leary, R.J., Shen, D., Boca, S.M., Barber, T., Ptak, J., et al. 2007. The genomic landscapes of human breast and colorectal cancers. *Science* **318:** 1108–1113.

Received March 11, 2008; accepted in revised form July 29, 2008.