Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2010 May 01.

Published in final edited form as:

Nat Biotechnol. 2009 November ; 27(11): 1025–1031. doi:10.1038/nbt.1583.

Microdroplet-based PCR amplification for large scale targeted sequencing

Ryan Tewhey[1,2], **Jason Warner**[3], **Masakazu Nakano**[1,4], **Brian Libby**[3], **Martina Medkova**[3], **Patricia David**[3], **Steve Kotsopoulos**[3], **Michael Samuels**[3], **J. Brian Hutchison**[3], **Jonathan W. Larson**[3], **Eric J. Topol**[1], **Michael P. Weiner**[3], **Olivier Harismendy**[1,4], **Jeff Olson**[3], **Darren R. Link**[3,*], and **Kelly A. Frazer**[1,4,*]

[1] Scripps Genomic Medicine – Scripps Translational Science Institute – The Scripps Research Institute, 3344 N. Torrey Pines Court, La Jolla CA 92037

[2] Division of Biological Sciences, University of California at San Diego, La Jolla, CA 92093

[3] RainDance Technologies, 44 Hartwell Avenue, Lexington, MA 02421

Abstract

Targeted sequencing of specific loci of the human genome is a promising approach for maximizing the efficiency of second-generation sequencing technologies for population-based studies of genetic variation. Here we describe microdroplet PCR, which performs 1.5 million separate amplifications in parallel, as an approach for enriching targeted sequences in the human genome. We initially designed primers to 435 exons of 47 genes that were selected for having a broad spectrum of sequence characteristics. Using this primer set we amplified the same six samples by both microdroplet and traditional singleplex PCR and sequenced the products using the Illumina GAII demonstrating that both methods generate similarly high quality data; 84% of the uniquely mapping reads fell within the targeted sequences, uniform coverage of ~90% of the targeted bases, greater than 99% accuracy in sequence variant calls, and high reproducibility between different samples (r^2=0.9). We next scaled the microdroplet PCR to 3976 amplicons totaling 1.49 Mb of sequence, sequenced the resulting sample on both the Illumina GAII and Roche 454 platforms, and obtained data with equally high specificity and sensitivity quality. Our results demonstrate that microdroplet technology is well suited for processing DNA for massively parallel amplification of specific subsets of the human genome for targeted sequencing.

Users may view, print, copy, download and text and data- mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use: http://www.nature.com/authors/editorial_policies/license.html#terms

*Corresponding Authors: kfrazer@ucsd.edudlink@raindancetech.com.

[4]Current address – Moores UCSD Cancer Center - 3855 Health Sciences Drive #0901 La Jolla, CA 92093-0901

Author Contributions K.A.F., E.J.T., M.P.W., and D.R.L. conceived the project; K.A.F, O.H., J.O. and D.R.L. designed the experiments, J.W., M.N., R.T., B.L., M.M., P.D., S.K., M.S., J.B.H., J.W.L, and O.H. performed the experiments; R.T. and J.W. performed the data analysis; R.T., J.W., J.O., D.R.L. and K.A.F. wrote the manuscript.

Accession codes We have deposited all sequencing reads into the National Institute of Health's - Sequence Read Archive (SRA) with the accession code SRA009786.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Introduction

Technical advances in sequencing methods and instruments are rapidly transforming our ability to study both common and rare genetic variants in the human genome. Indeed, several human genomes have already been sequenced in their entirety1-4. However, for the time being, the cost for sequencing whole human genomes is prohibitive for addressing research questions in a large cohort of individuals. Three approaches are currently being utilized for enrichment of target sequences of interest. The first approach is traditional singleplex PCR which has been employed to examine in hundreds of samples large kilobase-sized contiguous intervals5 or the exons of hundreds of genes6, 7. Although traditional PCR enriches target sequences with high specificity and sensitivity, it is difficult to scale the method to match the throughput of current sequencing instruments. The second approach is based on multiplex amplification of thousands of target sequences in a single tube by array synthesized padlock “molecular inversion” probes.8-10 Molecular inversion probes (MIP) allow for a highly efficient multiplex reaction due to the tethering of primer pairs by a DNA linker. However, published results show that while the captures are highly specific and represent upwards of 90% of the targets, there is greater than 100-fold range in coverage of the targeted sequences, and 34-42% of sequence capacity is consumed by either sequencing of primer sequences or the MIPs linker backbone9 (Supplemental Discussion). The third approach is based on hybridization with long oligonucleotides that are either matrix-bound or in solution to capture and pull-down the target sequences11-14. The hybridization-based methods have good capture rates, uniform coverage of target sequences, and good reproducibility. However, the methods are known to be biased to repetitive elements, which can result in a high proportion of non-uniquely mapping reads. Additionally, sequences that are highly homologous to other sequences in the genome cannot be individually targeted. We have developed an approach, involving microdroplet-based technology, which takes advantage of the high specificity and sensitivity of PCR and allows for massively parallel singleplex amplification of complex target sequences. The discrete encapsulation of microdroplet PCR reactions prevents possible primer pair interactions allowing for highly efficient simultaneous amplification of up to 4,000 targeted sequences and greatly reduces the amount of reagents required.

Results and Discussion

Microdroplet PCR Workflow

Microdroplet technology is particularly well suited for processing DNA for massively parallel amplification of sequencing targets by enabling the routine preparation of 1.5 million separate PCR reactions from 20 microliters of template solution containing just 7.5 micrograms of genomic-DNA. Microdroplet PCR is achieved through the following steps (Fig. 1): picoliter volume droplets of fragmented genomic DNA template are merged with pre-made primer pair droplets (Primer Library), pooled thermal cycling of the resulting PCR reactions (Droplet PCR), and destabilizing the droplets to release the PCR product (Break Emulsion) for purification and sequencing.

A key component of the microdroplet PCR process is the generation of a high quality primer library. The forward and reverse primers designed to amplify a targeted genomic interval are

QIAGEN Exhibit 1051
Page 2

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

combined at a concentration of 1.1 μM per primer (Fig. 1A), and reformatted as 8 picoliter-volume primer pair droplets (encapsulation of the primer pairs in inert fluorinated carrier oil) by using a flow-focusing nozzle (Supplemental Fig. 1)15 to make a library element. Library elements are combined together such that each library element is represented by an equal number of droplets within the final primer library (Fig. 1A). The resulting primer library is mixed and quality tested for primer pair droplet size and uniformity, as well as library element representation (Supplemental Fig. 2).

The genomic DNA template mixture (Fig. 1) contains all of the components for PCR except for the primers and is prepared by fragmenting genomic DNA using DNase I to produce 2-4 kb fragments (Fig. 1B) (Online Methods). On the microfluidic chip the template mixture is made into droplets and paired with primer pair droplets. The paired template and primer droplets enter the merge area on the microfluidic chip (Fig. 1C) at a rate of ~3000 droplets per second. As the primer pair droplets (8 pL) are smaller than the template droplets (14 pL) they move faster through the channels until they contact the preceding template droplet. Field induced coalescence of these droplet pairs16 results in the two droplets merging to produce a single PCR droplet which are collected and processed as an emulsion PCR reaction (Supplemental Fig. 3).

Validation phase - targeted sequences and DNA samples

To validate the microdroplet PCR approach we selected 47 genes distributed across the genome to determine the extent to which a set of PCR primer pairs could be chosen to simultaneous amplify numerous loci with a broad spectrum of sequence compositions using a single set of reaction conditions (Supplemental Table 1). Local DNA context, such as repetitive elements, GC content and sequence variants, can affect the ability of PCR primer pairs to work. One major drawback of PCR sample preparation is that allelic imbalanced amplification (greater efficiency of amplification for one of the alleles) due to variants in primer sites is sometimes observed17, 18. This type of PCR bias results in variant base calling errors, specifically heterozygous sites are called homozygous. Although the majority of SNPs shared among all populations have been discovered19 and are taken into account during primer design, population specific and rare variants are largely unknown. To check for biases in the ability of primers designed to the reference genome to amplify genomic DNA from individuals of different ethnicities, we included three HapMap samples of European and three of African descent. We also selected ~60% of the targeted sequences from three ENCODE20 intervals that were sequenced in five of the HapMap samples in our study as part of the HapMap Consortium21. This allowed us to compare the effect of allelic imbalanced amplification in intervals of the genome in which the majority of variation is known and can be accounted for during primer design versus intervals that are less well annotated for variation.

Of the 47 genes targeted in the validation phase, 29 are from ENCODE20 intervals, 8 are members of the Transient Receptor Potential ion (TRP) channel superfamily (included to test the ability to amplify and correctly align reads among family members with high nucleotide similarity), and the remaining 11 are candidates for deep venous thrombosis. Our primer design strategy (Online Methods) split the 435 exons of the 47 genes into 457

QIAGEN Exhibit 1051
Page 3

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

amplicons of varying sizes (119-956 bp) and GC content (24-78%) (Supplemental Tables 2 & 3). The full amplicon sequences constituted 172.2 kb and the exonic sequences comprised 75.9 kb (44%). It is important to note that the primer selection process did not attempt to minimize off-exon sequence, and thus if desired the ratio of exon:off-exon sequences in the amplicons could be increased. The primer pairs were designed and put into production in a single pass; there was no redesign of failed primer pairs.

Specificity of amplification

Using the validation phase 457 PCR primer set we amplified the six DNA samples using both traditional singleplex and microdroplet PCR. After the PCR amplification step all samples from the two methods were processed simultaneously through library creation and sequencing to ensure that any differences in the data were due to the amplification methods. Samples were labeled with a DNA barcode during the library preparation step and for each PCR method all six bar-coded samples were run on a single lane (Online Methods). The two lanes generated equivalent amounts of sequence data with each sample having an average of 1.27 million reads passing quality filters (Table 1).

To determine how well the 457 primer pairs specifically amplified the targeted sequences we calculated the percentage of filtered reads that mapped to a targeted amplicon. Considering all six samples 82% and 73% of filtered reads successfully mapped to the full amplicon sequences for the traditional and microdroplet reactions, respectively (Table 1). Considering only uniquely mapping reads 96% (231.69 Mb) of the traditional and 84% (202.15 Mb) of the microdroplet reads map back to the amplicons. After trimming primer sequences from the amplicons, which account for 12% of the total sequence, of the uniquely mapping reads 91% of the traditional and 80% of the microdroplet reads mapped to targeted sequences. Of the reads mapping to the amplicons, 55.6% and 51.1% map to the exons for the traditional and microdroplet PCR, respectively. In the microdroplet PCR reaction the ratio of input genomic DNA to amount of final PCR product is considerably greater than it is for traditional PCR (Online Methods), which may explain the ~15% difference between the two methods in reads that map off target. Overall, these data demonstrate that the microdroplet PCR enrichment method generates a high proportion of target sequence to background sequences.

Coverage uniformity and reproducibility

Uniformity of sequence coverage across targeted sequences is important because it determines the average depth to which samples have to be sequenced to have optimal sensitivity for variant calling. If the coverage differs greatly across targeted bases then one has to sequence deeply to adequately cover underrepresented bases. We initially examined how the base by base sequence coverage differs within individual amplicons. Visual inspection of coverage plots of individual targeted sequences (Fig. 2A) showed that there is typically a two to three-fold variation in the per-base sequence coverage depth across the amplicon. Interestingly, a noticeable dip in sequence coverage occurs at the transition from the 36^{th} to 37^{th} base pair in the amplicon, which is likely due to more reads (36 bp in length) starting at base one compared to bases 2 to 5. We next compared the coverage uniformity across all targeted bases for the six samples and the two PCR methods. The six samples had

QIAGEN Exhibit 1051
Page 4

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

slightly different sequence yields in the traditional and microdroplet PCR methods (Table 1) and therefore we normalized the coverage (divided the coverage of each base by the mean coverage of all targeted bases) so we could directly compare their coverage distribution plots (Fig. 3A, 3B). We first determined the percentage of targeted bases with sequence coverage between 1/5 to 5 times the mean. For traditional PCR, 92.8% of all bases fell within this range with 99.6% of the bases covered by at least one read (Fig. 3A, Supplemental Table 4). The microdroplet PCR showed similar results with 90.2% of all bases falling within 1/5 and 5 times the mean with 99.8% of all targeted bases covered by at least one read (Fig. 3B). These results indicate that for studies using microdroplet PCR amplified DNA and Illumina GA sequencing to an average depth of 25X that ~ 90% of the bases will be covered with 5 or more reads and with ~98% of all bases covered by at least a single read.

The consistency and reproducibility of coverage across targeted bases from sample to sample is of high importance for population studies. This is due to the fact that coverage is directly correlated with accuracy in base calling and to perform sequence-based association studies the same bases need to be analyzed across numerous samples. We compared the mean coverage of each amplicon between samples showing that reproducibility was excellent between samples with high correlations of amplicon coverage within a PCR method (average Lin's concordance: $r^2 = 0.88$ traditional PCR, $r^2 = 0.91$ microdroplet PCR) (Fig. 4A, B). On the other hand, coverage depth of specific amplicons between the two PCR methods varied (Fig. 4C, D), most likely due to differences in PCR chemistry and cycling conditions. Interestingly, of the 457 amplicons there was only one that failed completely (no reads) in four samples with the traditional PCR and a different amplicon that failed completely in three individuals with the microdroplet PCR. In both PCR methods >99.6% of the amplicons were successful, defined as a mean coverage greater than 5. Coverage was similar for the European and African ancestry samples and for the targeted sequences in ENCODE and non-ENCODE intervals. These data demonstrate that microdroplet PCR results in consistent and reproducible coverage of targeted sequences.

Accuracy of sequence variant calls

We evaluated the accuracy of variant bases called in the PCR amplified sequences for the six samples by comparison to HapMap genotypes for ~450 SNPs (Table 2a). There were a total of 2424 comparisons for the tradition PCR across the six samples with 22 of the comparisons being discordant for a consensus rate of 99.1% (Table 2a). The microdroplet PCR had the same consensus rate with 22 discordances out of 2390 comparisons. In both PCR methods ~ 2.5% of the HapMap variant sites were uncalled primarily due to low coverage, poor mapping quality, or the presence of neighboring variants. There was no observed difference in call rate or concordance between either the ENCODE and non-ENCODE intervals or between samples in the two ethnicity groups. Importantly, the two PCR methods were concordant with each other at all positions discordant with the HapMap. Of the 22 loci discordant with HapMap the errors fell into three classes: 1. nine are homozygous reference genotypes with HapMap homozygous alternate; 2. five are heterozygous genotypes with HapMap homozygous reference; and 3. eight are homozygous genotypes with HapMap heterozygous (Supplemental Table 5). Further examination of the discordant bases revealed most of those in class 1 are present in a single targeted sequence

QIAGEN Exhibit 1051
Page 5

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

with a highly similar homolog in the genome. Our sequencing reports the reference alleles for the target while HapMap calls the alternate alleles, which correspond to the fixed bases at those positions in the homolog. Inspection of the ENCODE traces revealed that the amplicons appeared to be amplifying both homologs as supported by all individuals being heterozygous at the differing positions (Supplemental Fig. 4). Class 2 and 3 errors most likely represent missed heterozygous calls in the HapMap and sequence data, respectively. Because allelic imbalanced amplification is a concern when performing PCR enrichment we focused on the eight Class 3 errors. Manual inspections of ENCODE sequence traces for five of these variants revealed that 4 are homozygous calls consistent with our sequence data. To determine if allelic bias during the PCR reaction was common among concordant variant calls we calculated the number of reference and alternate allele observations at heterozygote sites in the six samples. No skew was observed with 51.6% (CV 12%) and 51.1% (CV 13%) of the reads reporting the reference allele in traditional and microdroplet, respectively with no difference between ENCODE and non-ENCODE sites. Our analysis indicates that allele biased amplification results in only ~0.1% of the variants to be called incorrectly (≤ 5 out of 2390) and that >50% of the discrepancies between HapMap genotypes and variant calls in our sequence data are due to errors in the HapMap data.

To estimate the false positive rate we analyzed the variants identified in the targeted genes in the ENCODE intervals sequenced in five of the samples as part of the HapMap project. A total of 1122 variants at 442 positions were identified in these genes, of which 62 positions were not listed as variants in the HapMap data (Supplemental Table 6). Thirty-five of these positions are listed as variant in dbSNP suggesting that they are not false positives. Of the 27 positions not listed in dbSNP, 12 have concordant calls between the traditional and microdroplet PCR for at least one sample suggesting they may be true novel variants. The remaining 15 sites were either discrepant in the traditional but not in the microdroplet PCR (1 position, 3 variant calls) or discrepant by one method and the position did not pass quality filters in the second method (14 positions, 15 variant calls). These analyses suggest that the false positive rate of the sequence data is < 1.6% (≤18 of 1122 variant calls).

Scale up phase – selection of target sequences

To efficiently perform population-based sequencing studies using next generation sequencing platforms it is important to be able to simultaneously examine large numbers of targeted sequences. In this next phase we tested the ability of the microdroplet PCR to scale-up to 3976 primer pairs, which in total amplify 1.49 Mb of sequence, of which 645 kb (43%) consists of exonic sequence. We also tested the compatibility of this enrichment process with other next generation technologies (Online Methods). To further test the robustness of the microdroplet PCR process we choose primer pairs with widely varying characteristics, including primer lengths (17 bases - 30 bases), primer Tm (55.4°C - 61.2°C), amplicon length (299 bp - 659 bp) and amplicon GC content (25.1% - 81.5%), to determine how this affected the outcome of targeted sequence amplification.

Specificity of amplification

A single HapMap sample, NA18858, was amplified using the scale-up phase 3976 amplicons by microdroplet PCR (Online Methods). The resulting amplified material was

QIAGEN Exhibit 1051
Page 6

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

then divided into two aliquots and sequenced by both the Illumina GA II (short-read platform) and Roche 454 FLX (long read platform). The Illumina run generated >10 million quality single-end reads with similar fractions, as those observed in the validation phase, of both total reads (76%) and uniquely mapped reads (79%) successfully mapped to the trimmed amplicon sequences (Table 1). The Roche 454 run generated ~350,000 quality reads with ~94% of these mapping to the full amplicons. The larger fraction of Roche reads mapping to targets is likely due to the fact that the microfluidic PCR amplified DNA is not fragmented in the Roche 454 library preparation protocol and therefore the non-specific genomic DNA carryover from the amplification step (~2 to 4 kb in length) is too long to be efficiently used in library generation. These results indicate that even when simultaneously amplifying 3976 targeted sequences microfluidic PCR generates a high ratio of target sequence to background sequence.

Coverage uniformity and accuracy of variant calls

We evaluated the ability to uniformly capture the intended target across all 3976 amplicons. Only one amplicon failed completely resulting in zero mapping reads for both the Illumina and Roche 454 platforms. After trimming primer sequences we enriched for a total of 1.35 Mb of sequence, on the Illumina platform 99.8% of the bases were covered by 1 or more reads, of which 96.6% fell within 1/5 and 5 times the mean coverage (Fig. 3C). Furthermore, 90% of all amplicons had coverage at 1/5 the mean or greater for at least 95% of the basis, for 5X coverage this increased to 98% of the amplicons. Coverage was lower for amplicons greater than 70% GC content but was unaffected by amplicon length (Supplemental Fig. 5). The proportion of reads mapping to exons is closely proportional to the fraction of exonic sequences in the amplicons. The uniformity of coverage was slightly decreased when sequenced on the Roche 454 platform, with 93.7% of the bases covered by at least 1 read and 88.4% falling within 1/5 and 5 times the mean (Fig. 3C). The lower coverage uniformity observed in the Roche 454 sequence results from sequencing the amplicon ends without shearing the input DNA in the library preparation, and therefore the middle bases of short amplicons have high coverage and the middle bases of longer amplicons receive little or no coverage (Fig. 2B).

To examine accuracy, variant bases called in the Illumina and Roche 454 sequences were compared to 2226 HapMap genotypes. The Illumina sequence had a call rate of 99.3% with 26 discordant SNPs, while the Roche 454 sequence had a call rate of 92.3% with 31 discordant SNPs. The discrepant variants overlapped at 21 sites between the two sequence platforms and largely fell into the same error classes observed in the validation phase, with the exception that several discrepant Roche 454 SNPs were present in homopolymer sequences (Supplemental Table 7) consistent with previous observations.18 The Roche 454 lower call rate is related to the low coverage bases in the middle of longer amplicons (Fig. 2B). Overall this data shows that increasing the number of amplicons in the microdroplet PCR reaction almost an order of magnitude does not affect the resulting data quality.

QIAGEN Exhibit 1051
Page 7

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Discussion

Important parameters to consider when choosing an enrichment method for targeted sequencing include: uniformity of coverage of targeted sequences; the detection rate and calling accuracy of sequence variants; the efficiency of the enrichment over background sequences; universality of the capture method (fraction of genome that can be uniquely captured); and the multiplicity of the reaction (amount of sequence that can be targeted). Compared with other enrichment methods microdroplet PCR generates substantially greater uniform coverage of targeted sequences (Supplemental Discussion and Supplemental Table 8). Importantly the high uniformity of coverage results in a higher variant detection rate: microdroplet PCR (94.5%), solution-based hybridization (64%-89%)14, Molecular Inversion Probe (75%)9. The accuracy of called variant bases is currently high across all enrichment strategies. Likewise the proportion of reads mapping to targeted sequences is similar (~65%) across the different enrichment methods9, 14. Microdroplet PCR is a universal method allowing for the unique capture of most sequences, even those highly similar to other sequences in the genome; by anchoring a primer in the divergent portion of a homologous sequence or in an adjacent unique region, almost any interval can be specifically targeted. For hybridization-based methods, repetitive elements or homologous exons cannot be individually captured. For many population studies it will be desirable to simultaneously examine large numbers of targeted sequences. We have demonstrated the ability of microdroplet PCR to enrich for 4,000 targeted sequences in a single tube per sample, and are working towards scaling this to 20,000 targets (~7.5 Mb, ~1/10th the exome) using an expanded content format with 5 sets of primers in each droplet (Online Methods) and no other changes to the workflow. The Molecular Inversion Probes method has been used to enrich for 13,000 targeted sequences with high performance9 and the solution-based hybridization approach has shown the ability to target 22,000 sequences simultaneously14. Therefore the multiplicity of the targets and the scalability of these methods enable large number of samples to be examined in population studies. Recently an array-based capture has been employed to target 26.6 Mb of coding exons of 12 individuals with similar mapping efficiency as that of solution based hybridization22. While the approach has excellent concordance and uniformity, array-based capture is difficult to scale for population studies in individual laboratories.

The microdroplet PCR process proved to be extremely efficient with greater than 99.6% of all amplicons being successful (mean coverage greater than 5) and with highly reproducible amplification of targeted sequences between samples. Allelic amplification bias resulted in surprisingly few known variants to be incorrectly called, further attesting to the robustness of the process. An improved ability to remove non-specific genomic DNA carryover (Online Methods) would increase the enrichment observed on short-read sequence platforms by ~15% to that observed on the Roche 454. The requirement for 7.5 micrograms of starting DNA used in this study limits the applicability of microdroplet PCR for samples with limited quantities. While optimization has reduced the current requirement to 2 micrograms we are working on further improvements to reduce the amount of required starting material to nano-gram quantities of DNA (Supplemental Discussion).

QIAGEN Exhibit 1051
Page 8

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Overall, our study shows that the data generated using microdroplet PCR as an enrichment method is well suited for performing sequence-based association studies.

Online Methods

A. Microdroplet Based PCR Workflow

Primer Library Generation—Preparation of a high quality primer library is achieved through several steps. Primer pairs are designed and synthesized to amplify the targeted genomic regions. The forward and reverse primers for each amplicon are combined at a concentration of 1.1 μM per primer (Fig. 1A) and reformatted as 8 picoliter primer pair droplets by using a flow-focusing nozzle (Supplemental Fig. 1) in a microfluidic chip15. Library elements are tested for droplet size and uniformity (Supplemental Fig. 2) and mixed together such that each library element is represented by an equal number of droplets within the primer library (Fig. 1A).

Primer Library Quality Control—As shown in Supplemental Figure 2 the method for determining library element representation consists of 1) measuring the total volume of primer solution that is made into droplets from each library element and 2) measuring the distribution of droplet sizes present in each library element. These two measurements are made by using a flow sensor and a stroboscopic imaging system. The time integral of the measured infusion rate on the flow sensor provides a measure of the total volume of primer solution made into droplets for each library element. The stroboscopic imaging system collects images at a rate of 10 per second and each image contains roughly 20 droplets. This sampling of roughly 200 droplets per second of the 8000 per second generated is used to measure a representative distribution of the droplet volumes for each library element. The pass criteria consist of all library elements being represented with greater than 90% of the droplets within a window between 7.0 and 9.0 pico-liters. A bar graph of the total count of droplets for each library element from the stroboscopic imaging system provides a snapshot of the uniformity achieved.

The primer library for the 3976 amplicon library was made from 11 separate 384-well plates in a process whereby an equal number of droplets were made from each well on a given plate and stored in a collection vial for that plate. The 11 collection vials were then mixed and an equal volume of emulsion was transferred from each collection vial into a pooling vial; with the exception of the last vial which had a proportionally smaller volume transferred as it contained fewer library elements. Hence, the appropriate QC metric is the droplet count for each library element that went into a given collection vial and 11 separate droplet count graphs were generated; Fig. S2d being representative. In the case of the 457 library, all library elements were collected directly into the final pooling vial and the QC on the number of droplets was appropriately conducted on all 457 library elements as shown in S2b. The histograms of droplet size distribution shown in Fig. S2c is for the pooled library containing all library elements.

Genomic DNA Fragmentation and Nick Translation—Genomic DNA is fragmented using DNase I to produce 2-4 kb fragments and nick translated to incorporate biotinylated nucleotides into the genomic DNA fragments. Fragmentation of the genomic DNA to 2 – 4

QIAGEN Exhibit 1051
Page 9

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

kb allows for optimal template size for performing PCR in droplets. Incorporation of biotinylated nucleotides during the nick translation step allows the genomic DNA to be removed during the Genomic DNA Removal step to reduce the amount of genomic DNA fragments entering into the construction of the Illumina GA II and Roche 454 sequencing libraries.

The genomic DNA samples are processed on ice: 7.8 μg of genomic DNA in 25 μL nuclease free water (Fisher, 50843418), 37.5 μL of NEBuffer 2 (NEB, B7002S), 37.5 μL of a mixture of 1 mM each dCTP, dGTP, dTTP (NEB, N0446S), 18.75 μL of 0.4 mM Biotin-14 dATP (Invitrogen, 19524-016), 211.25 μL nuclease free water (Fisher, 50843418), 7.5 μL DNA polymerase I (10 units/μL) (NEB, M0209S), 37.5 μL of DNase I (NEB, M0303S) at a concentration determined by titration to achieve a 2-4 kb size distribution. The Genomic DNA fragmentation reaction mix is incubated at 15°C for 90 minutes and the reaction stopped by adding 37.5 μL of 0.5 M EDTA, pH 8.0 (Ambion, AM9260G). The DNA is purified over a Qiagen MinElute PCR purification column (Qiagen, 28004 (28006)) using the manufacturer's recommended conditions except for eluting with 13 μL of 10 mM Tris-Cl, pH 8.5.

Genomic DNA Template Mix—9.3 μL of the purified Genomic DNA Fragmentation reaction was added to 2.5 μL 10x High-Fidelity Buffer (Invitrogen, 11304-029), 1.6 μL of $MgSO_4$ (Invitrogen, 11304-029), 0.8 μL 10 mM dNTP (New England Biolabs, NO447S/L), 4.0 μL Betaine (Sigma, B2629-50G), 4.0 μL of RDT Droplet Stabilizer (RainDance Technologies, 30-00826), 2.0 μL dimethyl sulfoxide (Sigma, D8418-50ml) and 0.8 μL 5 units/μL of Platinum High-Fidelity Taq for a final volume of 25 μL.

RDT 1000: Merge—PCR droplets are produced by combining a single droplet from the primer library with a single genomic DNA template droplet (Supplemental Fig. 3). To achieve this, 20-uL of the Genomic DNA Template Mix is delivered to a microfluidic chip where it generates 14 pL template droplets at a rate of 3000 droplets per second (Supplemental Fig. 3b). In a separate channel primer library droplets are delivered onto the same chip (Supplemental Fig. 3a) and spaced to enter the channel carrying the genomic DNA template droplets at a one-to-one ratio (Supplemental Fig. 3c). As the primer library droplets are smaller than the template droplets, they move faster and catch up to the genomic DNA template droplets so that they enter the merge area on the chip (Supplemental Fig. 3d) as droplet pairs. In the merge area, the two droplets (template and primer pair droplets) coalesce in the presence of an external electric field16 to produce a single PCR droplet. The fidelity of the microfluidic flow enables one to set-up individual PCR reactions at a rate of 3000 per second. Over the period of roughly 10 minutes the 20-uL Genomic DNA Template Mix is processed into greater than 1.5 million PCR droplets (Supplemental Fig. 3e). The collection of PCR droplets represents essentially equal numbers of individual reactions from each of library elements. The PCR droplets are collected in a single standard PCR tube (Axygen, VWR, 22234-054) for thermal cycling and further processing (Supplemental Fig. 3f). The whole process can be viewed in the attached movie (Supplemental Movie 1) depicting the delivery of the primer pair droplets and the formation of the genomic DNA template droplets that pair together in a 1:1 ratio and merge into PCR droplets which are

QIAGEN Exhibit 1051
Page 10

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

collected for downstream highly parallel amplification of over a million single-plex PCR reactions.

PCR Amplification—Samples are cycled in a Bio-Rad PTC-225 thermocycler as follows: initial denaturation at 94°C for 2 minutes; 55 cycles at 94°C for 15 seconds, 58°C for 15 seconds, and 68°C for 30 seconds; a final extension at 68°C for 10 minutes and then hold at 4°C.

Breaking Emulsion—After amplification the PCR droplets are broken to release the PCR products from each individual emulsion reaction droplet. An equal volume of RDT 1000 Droplet Destabilizer (RainDance Technologies, 40-00830) is added to the collected PCR emulsion, the tube is vortexed for 15 seconds and then spun in a microcentrifuge for 10 minutes. A Gel-Loading Aerosol-Barrier Tip (Fisher Scientific, 02-707-172) is used to remove the oil from below the aqueous phase.

Genomic DNA Removal—110 μL of streptavidin beads (NEB, S1420S) in a 1.5 mL microcentrifuge tube are spun briefly in a microcentrifuge and placed on a magnetic separation rack (NEB, S1506S) for 30 seconds, the supernatant is removed and discarded. The beads are resuspended in 110 μL of binding buffer (0.5 M NaCl, 20 mM Tris-HCl (pH 7.5), 1 mM EDTA) by vortexing, briefly spun in a microcentrifuge and placed on a magnetic separation rack for 30 seconds, the supernatant is removed and discarded. The washed beads are resuspended in 110 μL of binding buffer and 50 μL of the beads are added to the broken amplicon emulsion. The beads are vortexed and then incubated for 10 minutes rotating on a Barnstead Labquake Tube Rocker. The beads are placed in a magnetic separation rack for 30 seconds and the supernatant transferred to a new 1.5 mL microcentrifuge tube. 50 μL more of the beads are added to the supernatant followed by vortexing and placement on the magnetic separator rack for 30 seconds. The supernatant was removed to a new 1.5 mL microcentrifuge tube. The DNA is then purified over a Qiagen MinElute PCR purification column (Qiagen, 28004) using the manufacturer's recommended conditions and eluted using 11 μL of Elution Buffer (10 mM Tris-Cl, pH 8.5). 1 μL of the eluant is run on an Agilent Bionanalyzer using the DNA 1000 Kit (Agilent, 5067-1504) to determine the yield for the amplicons. Typical yields originally were 100 – 300 ng. With improved conditions the typical yields are currently 200 – 600 ng.

B. Validation Phase

Genomic DNA samples—Six HapMap-DNA samples23 from unrelated individuals of different ethnic backgrounds were obtained from the Coriell Institute for Medical Research (www.coriell.org). Three were Caucasian descent (CEPH; NA11832, NA11992, and NA12006) and three were African descent (YRI; NA18489, NA18505, and NA18517). Five of these samples are part of the extended HapMap panel genotyped on the Affymetrix 6.0 and Illumina 1M arrays.

Target exons—In total, 435 exons from 47 genes (Supplemental Table 1) were targeted, of which 197 exons from 28 genes are in the ENCODE intervals sequenced in our study samples as part of the HapMap project and were selected to serve as reference standards.

QIAGEN Exhibit 1051
Page 11

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The remaining 19 genes (238 exons) included 8 members of the TRP channel superfamily (147 exons) to test for the ability to amplify and correctly align reads among gene family members with high nucleotide similarity.

Primer Design and Synthesis—The sequences for the 435 exons were masked for both repetitive elements (RepeatMask, UCSC Genome Browser) and known SNPs (dbSNP128). Exons larger than 600 bp were split into smaller elements to yield 457 targets for primer design. A five stage approach utilizing Primer3 (http://primer3.sourceforge.net/) for design was employed with stage one using the most stringent design parameters and the relaxing of these parameters at each subsequent stage (Supplemental Table 2). At the different stages the parameters altered include the amount of flanking sequence around the targeted region that Primer3 was allowed to use for design (200 & 300 bp), Tm (59.5-60.5 & 57.5-62.5), and removal of repeat masking of either the 5′ or 3′ flanking sequences. Primer length (18-27 nucleotides) was kept constant at all five stages. The majority (424) of the amplicons designed in the first stage (200 bp flanking sequence, Tm = 59.5-60.5, repeat masking of both 5′ and 3′ flanking sequence). The remaining 33 amplicons were designed within the next 4 stages. The 457 amplicons had a mean size of 400 bp ranging from 120 to 957 bp (Supplemental Table 3). The primer sets were obtained from Integrated DNA Technologies, Inc. with 5′-blocked end modification (Amino Modifier C6) as described.24 The 5′ blocked PCR primers were used to prevent their ligation to linkers in the library creation stage and thus over-representation in sequence coverage of the amplicon ends.

Traditional PCR amplification—We performed traditional PCR amplification of the 457 amplicons in individual reactions as follows: 5 ng of genomic DNA was amplified using 1.25 uM of each primer, 0.5 ul Titanium Taq DNA Polymerase (Clontech), 0.5 mM dNTPs, 32 mM Tricine, 4% DMSO, 60 mM Trizma, 3.2 mM $MgCl_2$, 17 mM $(NH_4)_2SO_4$ and 0.6X MasterAmp PCR Enhancer (EPICENTRE Technologies) per reaction, in a volume of 12 ul. The reactions were performed using a GeneAmp PCR System 9700 thermocycler (Applied Biosystems): initial denaturation at 96°C for 5 min; 55 cycles at 96°C for 2 sec and 60°C for 2 min; and a final extension at 50°C for 15 min. Following traditional PCR, the 457 amplicons generated from each genomic DNA sample were pooled in equal volume amounts. The 55 cycles of PCR were to exhaust the primers in each reaction in order to generate a uniform amount of each amplicon regardless of the efficiency of the PCR primers.

Library preparation for Illumina GA—The Illumina GA libraries for both the microdroplet and traditional PCR samples were made in parallel according to the manufacturer's instructions except for the following modifications. The fragmentation step was performed by shearing ~130-300 ng of the pooled PCR amplicons using Adaptive Focused Acoustics (Covaris S1, Covaris, Inc.) with the following conditions: Duty cycle, 20%; Intensity, 8; Cycles per burst, 200; Time, 6 minutes. This resulted in fragmented amplicons an average size of ~100 bp ranging from 50 to 175 bp. Paired-end adaptors used in ligation contained a different 4-bp barcode for each HapMap sample (NA11832-CAGT, NA11992-CTCT, NA12006-CCCT, NA18489-CACT , NA18505-CTGT, NA18515-CCGT) allow multiplexing of all six samples into a single lane24. The barcodes were sequenced as

QIAGEN Exhibit 1051
Page 12

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

bases 1-4 of the sequencing chemistry, reads were binned by barcode and trimmed to start at base 5 for downstream analysis. Because we have observed slight biases in sequencing efficiency between barcodes the same barcode for each sample was used for both microdroplet and traditional PCR. Following size selection, libraries were enriched by 12 (NA18517 for both microdroplet and traditional methods) or 15 (all other samples) cycles of PCR amplification using 10 ul of ligated product per library. Post enrichment the samples were quantified by Picogreen in quadruplicate and normalized to a 10 nM concentration. The six indexed libraries generated from the microdroplet PCR were combined together into one pool and the six libraries generated by traditional PCR were combined together into a second pool. 2.3 pM of each pool was then loaded into two separate lanes of the flow cell and sequenced using Illumina paired end cluster generation kit v1 and SBS kit v.2 for 40 cycles on each of the two reads.

C. Scale-up Phase

Genomic DNA Sample—A single HapMap-DNA sample23 of Yoruban descent (NA18858) was obtained from the Coriell Institute for Medical Research (www.coriell.org).

Target Exons, Primer Design and Synthesis—3976 PCR primer pairs were selected from a published set designed to amplify the coding regions of 13,062 genes25. The PCR primers were a subset of those published chosen to investigate how varying primer lengths (17 bases - 30 bases), primer Tm (55.4°C - 61.2°C), amplicon length (299 bp - 659 bp) and amplicon GC content (25.1% - 81.5%) affected the outcome of the targeted sequence enrichment and sequencing process. As described25 primer pairs were designed using Primer3 (http://primer3.sourceforge.net/), were required to be at least 50 bp outside the exon boundary, and avoided the presence of known SNPs in the five most 3′ bases. Primers were synthesized at Integrated DNA Technologies, Inc.

Library Preparation for Illumina GA and Roche 454—In the scale-up phase a single HapMap sample, NA18858, was amplified using 3976 amplicons by the microdroplet-based PCR workflow. The resulting amplified material was then divided into two aliquots and sequenced by both the Illumina GAII and Roche 454 platforms.

Illumina GA Library Preparation—Prior to the standard protocol for library generation on the Illumina GAII, the following steps were performed with the amplicons obtained from the sequence enrichment procedure. The amplicons were chloroform extracted and ethanol precipitated. Using the New England BioLabs Quick Blunting kit (E1201L) the fragments were blunt ended and phosphorylated. The fragments were then concatenated by ligation overnight at 25°C using the New England BioLabs Quick Ligation kit (M2200L). The concatenated amplicons were fragmented to 400 bp by Adaptive Focused Acoustics (Covaris S1, Covaris, Inc.) using the following settings: Mode, Frequency sweeping; Number of cycles, 1; Bath temperature limit, 20 degrees Celsius; Total process time, 3 minutes; Number of treatment, 4; Total time per treatment, 45 seconds; Duty cycle, 10%; Intensity, 5; Cycles per burst, 200. The sheared DNA was purified using a Qiagen QIAquick PCR purification column (28106) at which point the sample entered the standard Illumina GA library preparation protocol at the blunt ending and phosphorylation step, which follows

QIAGEN Exhibit 1051
Page 13

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

genomic DNA fragmentation. All subsequent steps were according to the standard library preparation protocol for sequencing on the Illumina GAII. Following library preparation the sample was sequenced on an Illumina GAII using two lanes of a flow cell.

Roche 454 FLX Library Preparation—Library generation for 454 FLX sequencing was carried out using the manufacturer's standard protocol. The ends of the PCR amplicons generated in the sequence enrichment procedure were made blunt and phosphorylated using the standard 454 protocols. The 454 sequencing adaptors were directly ligated onto the ends of the amplicons. The amplicons with the ligated 454 adaptors were then processed through the remainder of the library preparation protocol according to 454 FLX standard procedures for Low Molecular Weight DNA samples. The sample was sequenced using one half of a PicoTiterPlate on the Roche 454 FLX instrument.

D. Ratio of input genomic DNA to amount of final PCR product

In the microdroplet PCR reaction the ratio of input genomic DNA to amount of final PCR product is considerably greater than it is for traditional PCR, which may explain the ~15% difference between the two methods in reads that map off target. It is likely that a greater amount of genomic DNA enters into the Illumina GA libraries for the microdroplet PCR versus the traditional PCR method.

In the microdroplet PCR method 7.8 ug of DNA is input into the nick translation reaction. After nick translation and purification ~4.5 ug of DNA is input into the template buffer for droplet generation and merging with the library droplets. The amplicon yield for the validation phase library with 457 primers ranged from 128 ng to 300 ng.per sample for all amplicons combined. The amplicon yield for the scale up phase library with 3976 primers was 200 ng for all amplicons combined.

In the traditional PCR, 5 ng of DNA is input into each of the 457 reaction mixtures (2.3 ug total). The yield ranged from 271ug to 413 ug per sample for all amplicons combined.

E. Data Analysis

Illumina GAII - Mapping, Amplicon Efficiency and Coverage Uniformity—All samples sequenced on the Illumina GAII were processed through the Illumina pipeline v1.3 using default parameters. High quality filtered reads were mapped to the entire human genome (hg18) reference sequence (NCBI Build 36.1), using Maq v0.7126 default parameters except for insert size (-a 150) and mismatches (-n 3). Uniquely mapping reads were considered reads with a mapping quality score of 20 or greater corresponding to a 1% chance of being incorrectly mapped. The coverage uniformity analysis calculated coverage at the predicted amplicons after trimming primer sequences. Mean amplicon coverage and total coverage plots were produced using custom scripts and the statistics package R. None of the 457 amplicons in the validation phase failed (mean coverage less than 1) across all six samples. For traditional PCR there were 4 failures of a single amplicon across the 6 samples, this amplicon performed well for microdroplet-based PCR. Likewise a different amplicon failed for 3 of the 6 microdroplet-based amplified samples but performed well in the traditional PCR.

QIAGEN Exhibit 1051
Page 14

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Illumina GAII - Variant Detection and Quality Assessment—Mapped reads from the coverage analysis were used for variant detection using default parameters of the Maq SNP calling algorithm except for the following parameters (-w 10, -N 2, -W 2). Variants were additionally filtered using the MAQ variant consensus score. This value was empirically derived for the validation and scale-up phases by plotting the variant call rate against the consensus score and selecting the highest score just prior to a drop in call rate. The cutoff for the validation phase (paired end reads) was set at 30 while the scale up phase (single end reads) was filtered at a score of 50. The need to use a higher Maq quality cutoff for single versus paired end reads is well characterized 6.

A combined set of both HapMap Phase II and III genotypes were used for comparisons, the forward strand calls were downloaded from the HapMap website (http://www.hapmap.org). Concordance was calculated by the number of single nucleotide variants called the same between the HapMap and sequence data divided by the number of single nucleotide variants in the HapMap data for which the sequences data had a MAQ variant consensus greater than the derived threshold as well as coverage of 5 reads or more at the base. Bases with a consensus score less than the threshold or fewer than 5 reads covering the site were masked as no calls and are counted against the variant detection rate score but not the concordance rate. Calls were considered discordant regardless of the type of discordance, for example, an AB to AA error affected the concordance score the same as an AA to BB error.

Roche 454 – Mapping and Variant Detection and Quality Assessment—Data from the sequencing run on the 454 FLX was processed using 454′s standard data analysis software. For the amplicon amplification efficiency and coverage uniformity analyses high quality filtered reads were mapped to the predicted amplicon sequences, which were extracted from the hg18 reference sequence (NCBI Build 36.1), using the default parameters of the command line version of the Newbler module GS Mapper (version 2.0.00.20-64).

Prior to performing variant detection analysis the high quality filtered reads were mapped to the entire human genome reference sequence (hg18). Reads that aligned to one of the predicted amplicon sequences in the coverage analysis but matched to a homologous region when mapped to the entire human genome were removed prior to variant detection. Variants were detected using Atlas-SNP (version was 0.9.9.2) (http://code.google.com/p/atlas-snp/) with default parameters. Variant probability scores, which measure the probability of a substitution error at the SNP site were generated using the atlas_snp_evaluate.rb script with an estimated substitution error rate of 0.0008 and an estimated SNP rate of 0.001. Variant sites were called at positions containing 5 or more reads, sites with fewer than 5 reads were marked as no calls. Heterozygote positions were called when the alternate allele was present in 20-80% of the reads and contained an atlas probability score of greater than 1×10^{-5}. Probable heterozygotes with a probability score of less than 1×10^{-5} were deemed unreliable and marked as a no call site. Variants were then compared to the HapMap Phase II and III genotypes and variant detection rate and concordance rate were calculated in the same manner described above for the Illumina GAII sequence data.

QIAGEN Exhibit 1051
Page 15

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank X. Wang, K. Post (STSI), O. Iartchouk (Partners HealthCare Center for Personalized Genetic Medicine, K. Makowski (Agencourt Bioscience Corporation) for excellent technical assistance, N. Schork (STSI) for helpful conversations, N. Hafez (seqWise) for assistance with data analysis, and the US National Institute of Health (CTSA grant 1U54RR025204-01; Innovative Technologies for Molecular Analysis of Cancer grant 1R21CA125693-01) and Japan Foundation for Aging and Health (MN fellowship) for support for this effort.

References

1. Levy S, et al. The diploid genome sequence of an individual human. PLoS Biol. 2007; 5:e254. [PubMed: 17803354]
2. Wheeler DA, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008; 452:872–876. [PubMed: 18421352]
3. Wang J, et al. The diploid genome sequence of an Asian individual. Nature. 2008; 456:60–65. [PubMed: 18987735]
4. Bentley DR, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008; 456:53–59. [PubMed: 18987734]
5. Yeager M, et al. Comprehensive resequence analysis of a 136 kb region of human chromosome 8q24 associated with prostate and colon cancers. Hum Genet. 2008; 124:161–170. [PubMed: 18704501]
6. Ding L, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008; 455:1069–1075. [PubMed: 18948947]
7. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; 455:1061–1068. [PubMed: 18772890]
8. Porreca GJ, et al. Multiplex amplification of large sets of human exons. Nat Methods. 2007; 4:931–936. [PubMed: 17934468]
9. Turner EH, Lee C, Ng SB, Nickerson DA, Shendure J. Massively parallel exon capture and library–free resequencing across 16 genomes. Nat Methods. 2009:1–2.
10. Krishnakumar S, et al. A comprehensive assay for targeted multiplex amplification of human DNA sequences. Proc Natl Acad Sci U S A. 2008; 105:9296–9301. [PubMed: 18599465]
11. Albert TJ, et al. Direct selection of human genomic loci by microarray hybridization. Nat Methods. 2007; 4:903–905. [PubMed: 17934467]
12. Hodges E, et al. Genome–wide in situ exon capture for selective resequencing. Nat Genet. 2007; 39:1522–1527. [PubMed: 17982454]
13. Okou DT, et al. Microarray–based genomic selection for high–throughput resequencing. Nat Methods. 2007; 4:907–909. [PubMed: 17934469]
14. Gnirke A, et al. Solution hybrid selection with ultra–long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 2009; 27:182–189. [PubMed: 19182786]
15. Anna SL, Bontoux N, Stone HA. Formation of dispersions using “flow focusing” in microchannels. Applied Physics Letters. 2003; 82:364–366.
16. Ahn K, Agresti J, Chong H, Marquez M, Weitz DA. Electrocoalescence of drops synchronized by size–dependent flow in microfluidic channels. Applied Physics Letters. 2006; 88
17. Quinlan AR, Marth GT. Primer–site SNPs mask mutations. Nat Methods. 2007; 4:192. [PubMed: 17327845]
18. Harismendy O, et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. 2009; 10:R32. [PubMed: 19327155]
19. Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nat Rev Genet. 2009; 10:241–251. [PubMed: 19293820]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

20. Birney E, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447:799–816. [PubMed: 17571346]
21. Sachidanandam R, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001; 409:928–933. [PubMed: 11237013]
22. Ng SB, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009
23. Altshuler D, et al. A haplotype map of the human genome. Nature. 2005; 437:1299–1320. [PubMed: 16255080]
24. Harismendy O, Frazer K. Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR–PCR using Illumina GA sequencing–by–synthesis technology. Biotechniques. 2009; 46:229–231. [PubMed: 19317667]
25. Sjoblom T, et al. The consensus coding sequences of human breast and colorectal cancers. Science. 2006; 314:268–274. [PubMed: 16959974]
26. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008; 18:1851–1858. [PubMed: 18714091]

QIAGEN Exhibit 1051
Page 17

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

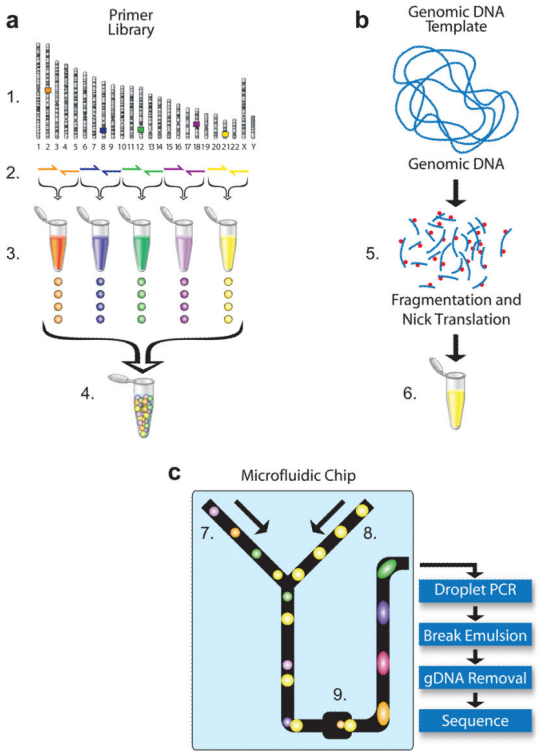


Figure 1. Microdroplet PCR workflow

Primer Library Generation (A): (1) Identify targeted sequences of interest in the genome. (2) Design and synthesize forward and reverse primer pairs for each targeted sequence (library element). (3) Generation of primer pair droplets for each library element. A microfluidic chip is used to encapsulate the aqueous PCR primers in inert fluorinated carrier oil with a block-copolymer surfactant to generate the equivalent of a picoliter scale test tube compatible with standard molecular biology. (4) Primer library, primer pair droplets of library elements are mixed together so that each library element has an equal representation. *Genomic DNA Template Mix Preparation (B):* (5) Genomic DNA is biotinylated (red dots), fragmented into 2 to 4 kb fragments and purified. (6) Purified genomic DNA is mixed together with all of the components of the PCR reaction (DNA polymerase, dNTPs, and buffer) except for the PCR primers. *Droplet Merge and PCR (C):* (7) Primer Library droplets are dispensed to the microfluidic chip (8) while the Genomic DNA Template is delivered as an aqueous solution and template droplets are formed within the microfluidic chip. The primer pair droplets and template droplets are then paired together in a 1:1 ratio. (9) Paired droplets flow through the channel of the microfluidic chip to pass through a merge area where an electric field induces the two discrete droplets to coalesce into a single PCR droplet. The roughly 1.5 million PCR droplets are collected into a single 0.2 ml PCR tube. The collection of PCR droplets (PCR Library) is processed in a standard thermal cycler for targeted amplification, followed by breaking the emulsion of PCR droplets to release the PCR amplicons into solution for genomic DNA (gDNA) removal, purification and sequencing.

QIAGEN Exhibit 1051
Page 18

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

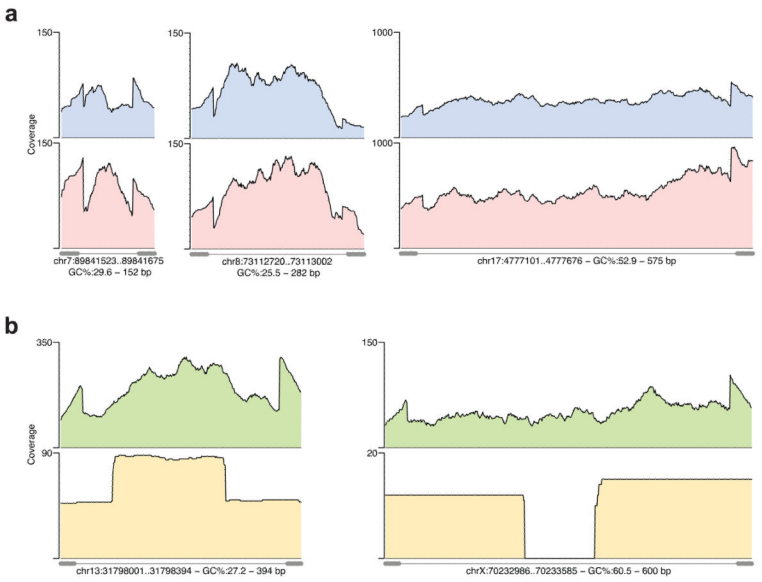


Figure 2. Coverage plots of targeted sequences

For the validation phase (A) base by base coverage of three target sequences selected for their varying lengths and GC% amplified by microdroplet (blue) and traditional (red) PCR. For the scale-up phase (B) the coverage of two targets representing an average and maximum amplicon length sequenced by Illumina GA (green) and Roche 454 (yellow) is shown. At the bottom of each plot the PCR primer positions (grey dumbbells connected by line) are shown. Roche 454 end sequencing of average sized amplicons results in 2-fold higher coverage of middle bases whereas end sequencing of larger amplicons results in middle bases having no coverage.

QIAGEN Exhibit 1051
Page 19

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

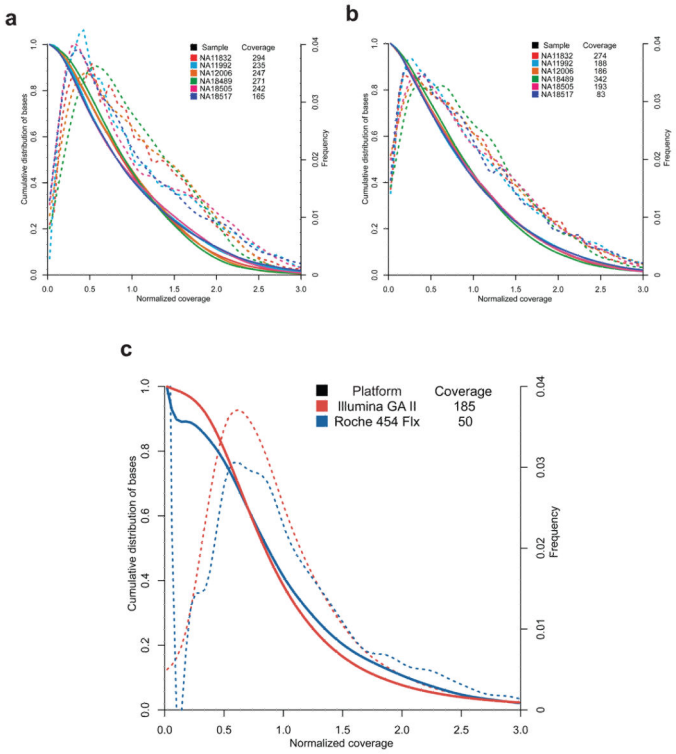


Figure 3. Normalized Coverage Distribution Plots

The validation phase 457 amplicons amplified by traditional PCR (A) and microdroplet PCR (B) and the scale-up phase 3976 amplicons amplified by microdroplet PCR (C). Normalized coverage is the absolute base coverage divided by the mean coverage of bases for the indicated sample. Each colored line represents either one of the six samples (A & B) or one of two sequencing platforms (C). The solid colored lines represent the cumulative distribution (left axis) for each sample. The colored dashed lines indicate a skewed normal distribution (right axis) for each sample. For each sample the mean coverage across all bases are listed.

QIAGEN Exhibit 1051
Page 20

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

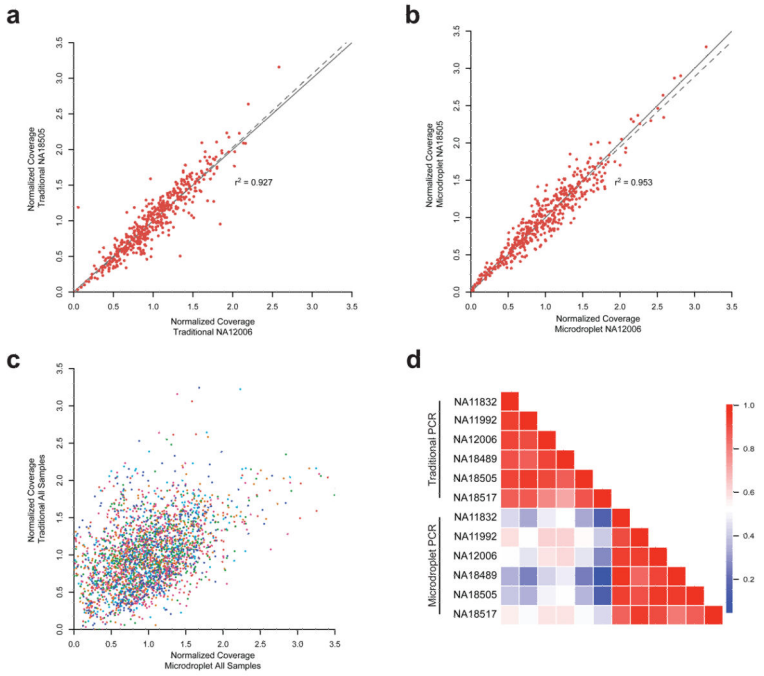


Figure 4. Inter-sample reproducibility of amplicon coverage

For the validation phase the normalized mean coverage of each amplicon is plotted for NA12006 (Caucasian) versus NA18505 (African) samples for the traditional (A) and microdroplet (B) PCR methods. For each sample (assigned same color as in Figure 2) the average normalized coverage of each amplicon is plotted for traditional versus microdroplet (C). Correlation matrix for all samples depicting Lin's concordance coefficient (D). All samples show a high correlation among each other within a PCR method but not between the two methods.

QIAGEN Exhibit 1051
Page 21

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Illumina GAII reads and mapping statistics

			Filtered Reads[a]		Mapped Reads[b]		Uniquely Mapped Reads[c]				% Uniquely Mapping to Trimmed Amplicons			
Set	**Sample**	**PCR Method**	**Number**	**Bases**	**HG18**	**Full Amplicons[d]**	**HG18**	**Full Amplicons[d]**	**Primer Trimmed Amplicons**	**Exonic Target**	**Filtered Reads**	**Mapped Reads**	**Uniquely Mapped Reads**	**% of Filtered Reads Mapping on the Exon**
457	NA11832	Traditional	1496088	53.86 Mb	50.29 Mb	48.06 Mb	49.24 Mb	47.57 Mb	44.47 Mb	25.35 Mb	83.44%	89.37%	90.31%	47.06%
	NA11832	Microdroplet	1673982	60.26 Mb	56.39 Mb	44.70 Mb	52.99 Mb	44.23 Mb	41.46 Mb	21.47 Mb	69.53%	74.31%	78.24%	35.63%
	NA11992	Traditional	1213396	43.68 Mb	39.33 Mb	37.39 Mb	38.46 Mb	37.01 Mb	35 59 Mb	21.55 Mb	82.31%	91.42%	92.53%	49.34%
	NA11992	Microdroplet	1367394	49.23 Mb	43.83 Mb	30.09 Mb	39.81 Mb	29.71 Mb	28.45 Mb	16.01 Mb	58.55%	65.76%	71.47%	32.53%
	NA12006	Traditional	1256622	45.24 Mb	41.84 Mb	39.64 Mb	40.99 Mb	39.27 Mb	37.49 Mb	21.59 Mb	83.66%	90.45%	91.45%	47.72%
	NA12006	Microdroplet	1148454	41.34 Mb	37.83 Mb	30.00 Mb	35.43 Mb	29.67 Mb	28.15 Mb	15.82 Mb	68.86%	75.26%	79.46%	38.27%
	NA18505	Traditional	1222820	44.02 Mb	40.77 Mb	38.17 Mb	39.83 Mb	37.77 Mb	36.63 Mb	22.00 Mb	84.09%	90.81%	91.95%	49.97%
	NA18505	Microdroplet	1116948	40.21 Mb	36.06 Mb	30.36 Mb	34.36 Mb	30.05 Mb	29.21 Mb	15.87 Mb	73.40%	81.84%	85.01%	39.46%
	NA18517[e]	Traditional	838226	30.18 Mb	28.12 Mb	26.35 Mb	27.41 Mb	26.04 Mb	24.99 Mb	14.89 Mb	83.81%	89.93%	91.17%	49.33%
	NA18517[e]	Microdroplet	587958	21.17 Mb	18.12 Mb	13.04 Mb	16.40 Mb	12.87 Mb	12 37 Mb	6.88 Mb	59.16%	69.12%	75.41%	32.49%
	NA18489	Traditional	1429866	51.48 Mb	46.90 Mb	44.44 Mb	45.97 Mb	44.03 Mb	41.08 Mb	23.36 Mb	80.56%	88.42%	89.36%	45.37%
	NA18489	Microdroplet	1885186	67.87 Mb	63.50 Mb	56.18 Mb	60.79 Mb	55.62 Mb	51.75 Mb	27.18 Mb	77.03%	82.32%	85.14%	40.06%
	All Samples	Traditional	7457018	268.45 Mb	247.24 Mb	234.10 Mb	241.91 Mb	231.69 Mb	220.25 Mb	128.72 Mb	82.04%	89.08%	91.04%	47.95%
	All Samples	Microdroplet	7779922	280.08 Mb	255.74 Mb	204.22 Mb	239.77 Mb	202.15 Mb	191.39 Mb	103.23 Mb	68.36%	74.84%	79.82%	36.86%
3976	NA18858	Microdroplet	10603854	381.74 Mb	325.57 Mb	293.89 Mb	309.65 Mb	289.97 Mb	245.20 Mb	122.87 Mb	64.23%	76.30%	79.19%	32.19%

[a] - High quality reads from Illumina Pipeline 1.3

[b] - Number of bases from filtered reads that were mapped by Maq, non-unique reads are randomly placed at one of the multiple locations

[c] - Maq mapping score of 20 or greater, corresponding to a 1% chance of being incorrectly mapped.

[d] -Amount of sequence mapping to amplicons including primer sequences

[e] - NA18517 had fewer reads than the other 5 samples for both traditional and microdroplet PCR. This was likely a technical issue with both library preparations that is independent of PCR method and DNA quality.

Nat Biotechnol. Author manuscript; available in PMC 2010 May 01.

QIAGEN Exhibit 1051
Page 22

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2a

Validation phase sequence variant detection rates and concordance

		HapMap SNPs[a]		Variant Detection Rate		Variant Concordance		Reference Allele[c]
Sample	**PCR Method**	**Total**	**ENCODE**	**Total %**	**ENCODE %**	**Total %[b]**	**ENCODE %**	**%**
NA11832	Traditional	459	279	98.47	97.85	99.67	98.53	51.08
NA11832	Microdroplet			97.39	97.13	98.66	98.52	51.11
NA11992	Traditional	460	279	99.35	99.28	99.13	98.92	51.55
NA11992	Microdroplet			98.48	98.57	99.33	98.91	51.64
NA12006	Traditional	461	280	99.13	99.29	99.34	99.28	51.63
NA12006	Microdroplet			97.61	97.86	98.85	99.27	50.02
NA18505	Traditional	444	265	98.87	98.49	99.09	98.85	53.40
NA18505	Microdroplet			97.52	98.11	98.85	98.46	52.62
NA18517	Traditional	439	262	97.95	97.33	99.07	99.21	50.40
NA18517	Microdroplet			95.67	95.80	99.29	99.20	49.69
NA18489	Traditional	189	99	100	100	99.47	100	51.67
NA18489	Microdroplet			98.94	98.99	99.47	100	51.66

[a]- Number of called genotypes in 457 amplicons (Total) and the number mapping to the 234 amplicons in the ENCODE intervals (HapMap release 27),

[b]- The number of genotypes that match between sequence and Hapmap divided by total comparisons

[c]- The average across all concordant heterozygote sites, the number of observations of the reference allele in comparison to the alternate allele

Nat Biotechnol. Author manuscript; available in PMC 2010 May 01.

QIAGEN Exhibit 1051
Page 23

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2b

Scale-up phase sequence variant detection rates and concordance

Sample	Sequencing Platform	HapMap SNPs	Variant Detection Rate	Variant Concordance	Discordant[a]	
					SNPs	Common
NA18858	Illumina GAII	2226	99.326	98.83	26	21
NA18858	454 Flx		92.273	98.49	31	

[a] - Number of SNPs with discordant genotypes between HapMap and Illumina and 454 sequence data and the number of discordant SNPs in common between both platforms.

QIAGEN Exhibit 1051
Page 24