nature biotechnology

Microdroplet-based PCR enrichment for large-scale targeted sequencing

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

Targeted enrichment of specific loci of the human genome is a promising approach to enable sequencing-based studies of genetic variation in large populations. Here we describe an enrichment approach based on microdroplet PCR, which enables 1.5 million amplifications in parallel. We sequenced six samples enriched by microdroplet or traditional singleplex PCR using primers targeting 435 exons of 47 genes. Both methods generated similarly high-quality data: 84% of the uniquely mapping reads fell within the targeted sequences; coverage was uniform across ~90% of targeted bases; sequence variants were called with >99% accuracy; and reproducibility between samples was high (r^2 = 0.9). We scaled the microdroplet PCR to 3,976 amplicons totaling 1.49 Mb of sequence, sequenced the resulting sample with both Illumina GAII and Roche 454, and obtained data with equally high specificity and sensitivity. Our results demonstrate that microdroplet technology is well suited for processing DNA for massively parallel enrichment of specific subsets of the human genome for targeted sequencing.

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 entirety[1–4]. 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 used for enrichment of target sequences of interest. The first approach is traditional singleplex PCR, which has been used for hundreds of samples to examine large kilobase-sized contiguous intervals[5] or the exons of hundreds of genes[6,7]. 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 allow for a highly efficient multiplex reaction owing to the tethering of primer pairs by a DNA linker. However, published results show that although the captures are highly specific and represent upwards of 90% of the targets, there is >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 molecular inversion probes' linker backbone[9] (**Supplementary Discussion**).

The third approach, based on hybridization with long oligonucleotides that are either matrix-bound or in solution, captures and pulls down the target sequences[11–14]. 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 reads that map nonuniquely. 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

Microdroplet PCR Workflow

Microdroplet technology is particularly well suited for processing DNA for massively parallel amplification of sequencing targets. It involves preparation of 1.5 million separate PCR reactions from 20 µl of template solution containing just 7.5 µg of genomic DNA. Microdroplet PCR requires the following steps (**Fig. 1**): merging picoliter volume droplets of fragmented genomic DNA template with premade 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.

[1]Scripps Genomic Medicine, Scripps Translational Science Institute, The Scripps Research Institute, La Jolla, California, USA. [2]Division of Biological Sciences, University of California at San Diego, La Jolla, California, USA. [3]RainDance Technologies, Lexington, Massachusetts, USA. [4]Present addresses: Moores UCSD Cancer Center, La Jolla, California, USA (M.N.) and Affomix, Branford, Connecticut, USA (M.P.W.). Correspondence should be addressed to K.A.F. (kafrazer@ucsd.edu) or D.R.L. (dlink@raindancetech.com).

Received 15 June; accepted 1 October; published online 1 November 2009; corrected after print 11 November 2009; doi:10.1038/nbt.1583

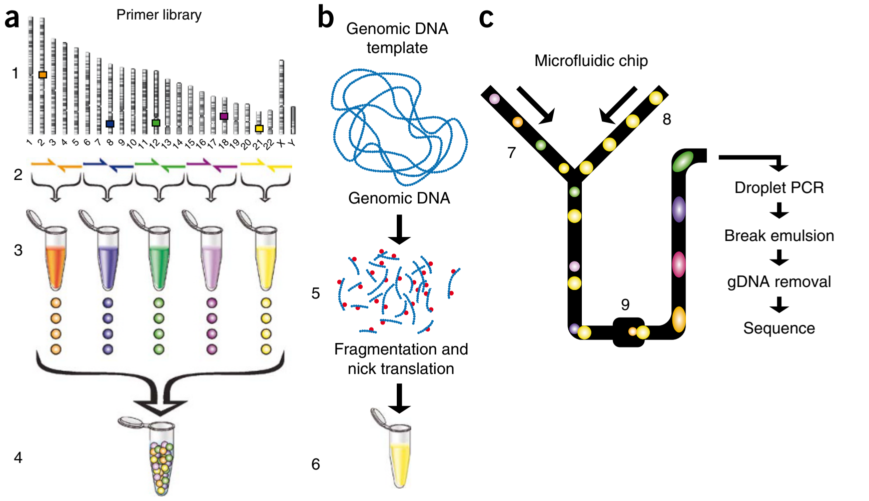


Figure 1 Microdroplet PCR workflow. (**a**) Primer library generation. (1) Identify targeted sequences of interest in the genome. (2) Design and synthesize forward and reverse primer pairs for each targeted sequence (library element). (3) Generation of primer pair droplets for each library element. A 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. (**b**) Genomic DNA template mix preparation. (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. (**c**) Droplet merge and PCR. (7) Primer library droplets are dispensed to the microfluidic chip. (8) 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 ~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.

A key component of the microdroplet PCR process is the generation of a high-quality primer library. Each element in the library consists of forward and reverse primers designed to amplify a targeted genomic interval. Primer pairs are combined at a concentration of 1.1 μM per primer (**Fig. 1a**) and reformatted using a flow-focusing nozzle as 8-picoliter (pl) primer pair droplets (primers are encapsulated in inert fluorinated carrier oil; **Supplementary Fig. 1**)[15]. 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 (**Supplementary 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 DNaseI 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 ~3,000 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 pairs[16] results in the two droplets merging to produce a single PCR droplet, which is collected and processed as an emulsion PCR reaction (**Supplementary Fig. 3**).

Validation phase: targeted sequences and DNA samples

To validate the microdroplet PCR approach, we selected 47 genes distributed across the genome. Genes were selected to determine the extent to which a set of PCR primer pairs could be chosen to simultaneously amplify numerous loci with a broad spectrum of sequence compositions using a single set of reaction conditions (**Supplementary 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 observed[17,18]. This type of PCR bias results in variant base-calling errors, specifically, heterozygous sites are called homozygous. Although the majority of single-nucleotide polymorphisms (SNPs) shared among all populations have been discovered[19] 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 ENCODE[20] intervals that were sequenced in five of the HapMap samples in our study as part of the HapMap Consortium[21]. 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 ENCODE[20] intervals, eight 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 amplicons of varying sizes (119–956 bp) and GC content (24–78%) (**Supplementary Tables 2** and **3**). The full amplicon sequences constituted 172.2 kb. 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 set of 457 primer pairs, we amplified the six DNA samples using both traditional singleplex and microdroplet

Table 1 Illumina GAII reads and mapping statistics

			Filtered reads[a]		Mapped reads[b] (Mb)		Uniquely mapped reads[c] (Mb)				Percent uniquely mapping to trimmed amplicons			Percent of filtered reads mapping on the exon
Set	Sample	PCR method	Number	Bases (Mb)	HG18	Full amplicons[d]	HG18	Full amplicons[d]	Primer trimmed amplicons	Exonic target	Filtered reads	Mapped reads	Uniquely mapped reads	
457	NA11832	Traditional	1496088	53.86	50.29	48.06	49.24	47.57	44.47	25.35	83.44%	89.37%	90.31%	47.06%
	NA11832	Microdroplet	1673982	60.26	56.39	44.70	52.99	44.23	41.46	21.47	69.53%	74.31%	78.24%	35.63%
	NA11992	Traditional	1213396	43.68	39.33	37.39	38.46	37.01	35.59	21.55	82.31%	91.42%	92.53%	49.34%
	NA11992	Microdroplet	1367394	49.23	43.83	30.09	39.81	29.71	28.45	16.01	58.55%	65.76%	71.47%	32.53%
	NA12006	Traditional	1256622	45.24	41.84	39.64	40.99	39.27	37.49	21.59	83.66%	90.45%	91.45%	47.72%
	NA12006	Microdroplet	1148454	41.34	37.83	30.00	35.43	29.67	28.15	15.82	68.86%	75.26%	79.46%	38.27%
	NA18505	Traditional	1222820	44.02	40.77	38.17	39.83	37.77	36.63	22.00	84.09%	90.81%	91.95%	49.97%
	NA18505	Microdroplet	1116948	40.21	36.06	30.36	34.36	30.05	29.21	15.87	73.40%	81.84%	85.01%	39.46%
	NA18517[e]	Traditional	838226	30.18	28.12	26.35	27.41	26.04	24.99	14.89	83.81%	89.93%	91.17%	49.33%
	NA18517[e]	Microdroplet	587958	21.17	18.12	13.04	16.40	12.87	12.37	6.88	59.16%	69.12%	75.41%	32.49%
	NA18489	Traditional	1429866	51.48	46.90	44.44	45.97	44.03	41.08	23.36	80.56%	88.42%	89.36%	45.37%
	NA18489	Microdroplet	1885186	67.87	63.50	56.18	60.79	55.62	51.75	27.18	77.03%	82.32%	85.14%	40.06%
	All Samples	Traditional	7457018	268.45	247.24	234.10	241.91	231.69	220.25	128.72	82.04%	89.08%	91.04%	47.95%
	All Samples	Microdroplet	7779922	280.08	255.74	204.22	239.77	202.15	191.39	103.23	68.36%	74.84%	79.82%	36.86%
3976	NA18858	Microdroplet	10603854	381.74	325.57	293.89	309.65	289.97	245.20	122.87	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; nonunique reads are randomly placed at one of the multiple locations. [c]Maq mapping score ≥20, 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 five 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.

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. 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. Of these amplicon-mapped reads, after trimming primer sequences from the amplicons, which account for 12% of the total sequence, 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 threefold variation in the coverage depth per base sequence across the amplicon. A noticeable dip in sequence coverage occurs at the transition from the 36th to 37th base pair in the amplicon, which is likely due to more reads (36 bp in length) starting at base 1 compared to bases 2–5.

We next compared the coverage uniformity across all targeted bases for the six samples and the two PCR methods. The six samples had slightly different sequence yields in the traditional and microdroplet PCR methods (**Table 1**). Therefore, we normalized the coverage (that is, 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**,**b**). We first determined the percentage of targeted bases with sequence coverage from one-fifth to five times the mean. For traditional PCR, this range encompassed 92.8% of all bases, of which 99.6% were covered by at least one read (**Fig. 3a** and **Supplementary Table 4**). The microdroplet PCR showed similar results: 90.2% of all bases fell between one-fifth and five times the mean, and 99.8% of all targeted bases were 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 25×, ~90% of the bases will be covered with ≥5 reads, and ~98% of all bases will be 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 because coverage is directly correlated with accuracy in base calling, and moreover, the same bases need to be analyzed across numerous samples to perform sequence-based association studies. We compared the mean coverage of each amplicon between samples. 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

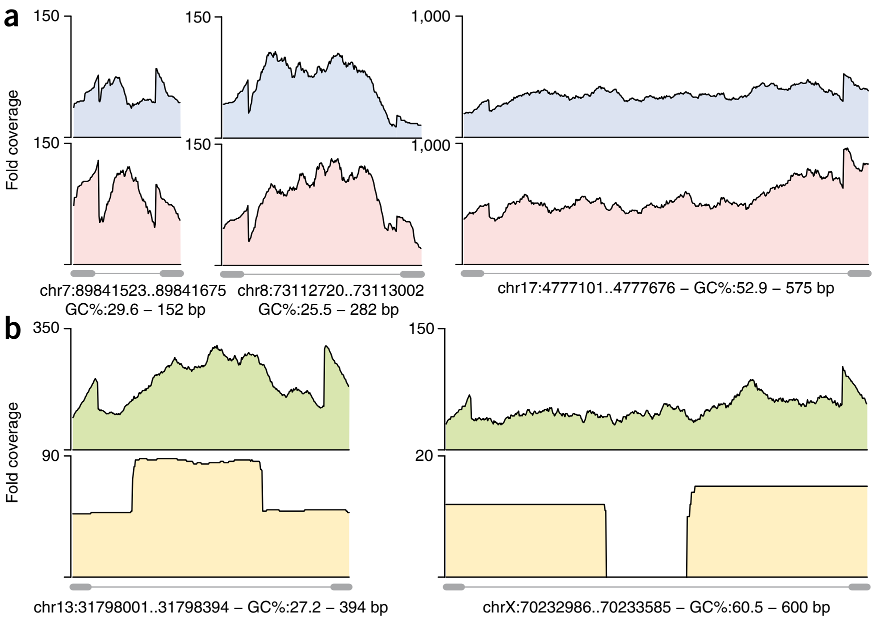


Figure 2 Coverage plots of targeted sequences. (**a**) Validation phase. Base-by-base coverage of three target sequences selected for their varying lengths and GC% amplified by microdroplet (blue) and traditional (pink) PCR. (**b**) Scale-up phase. The coverage of two targets representing an average 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 (gray dumbbells connected by line) are shown. Roche 454 end-sequencing of average-sized amplicons results in twofold higher coverage of middle bases whereas end-sequencing of larger amplicons results in middle bases having no coverage.

amplicons between the two PCR methods varied (**Fig. 4c,d**), most likely because of 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. A different amplicon 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 >5 reads. 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 2**). There were 2,424 comparisons for the traditional PCR across the six samples. Twenty-two of the comparisons were discordant for a consensus rate of 99.1% (**Table 2**). The microdroplet PCR had the same consensus rate with 22 discordances out of 2,390 comparisons. In both PCR methods ~2.5% of the HapMap variant sites were uncalled, primarily owing 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 a HapMap homozygous alternate, (2) five are heterozygous genotypes with a HapMap homozygous reference and (3) eight are homozygous reference with a HapMap heterozygous genotype (**Supplementary Table 5**). Further examination of the discordant bases revealed that most class 1 errors are present in a single targeted sequence with a highly similar homolog in the genome. Our sequencing reports the reference alleles for the target, whereas 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, an observation supported by all individuals being heterozygous at the differing positions (**Supplementary 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

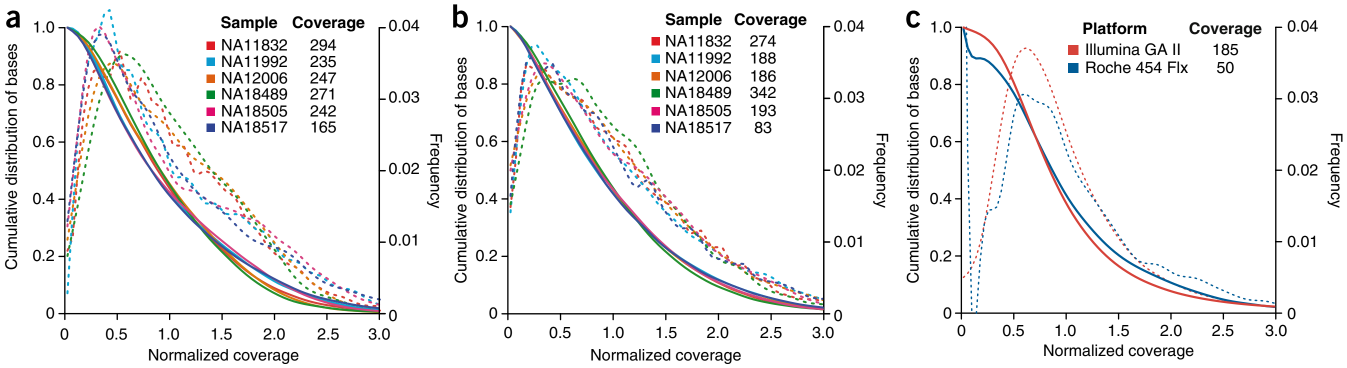
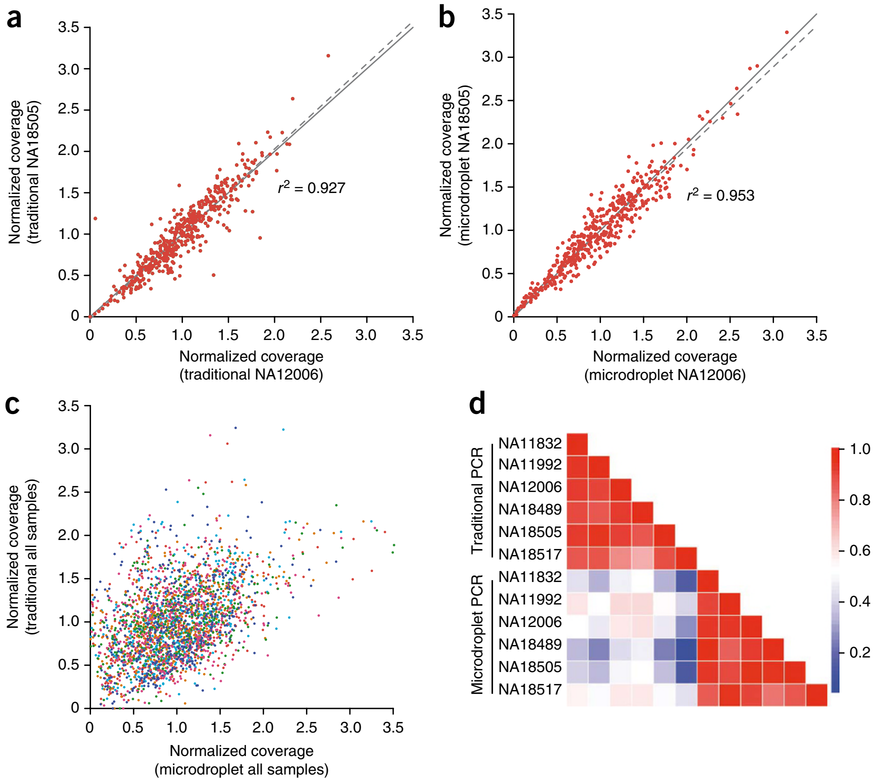


Figure 3 Normalized coverage distribution plots. (**a**–**c**) The validation phase set of 457 amplicons amplified by traditional PCR (**a**) and microdroplet PCR (**b**) and the scale-up phase of 3,976 amplicons amplified by microdroplet PCR (**c**). Normalized coverage is the absolute base coverage 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 is listed.

Figure 4 Intersample reproducibility of amplicon coverage. (**a**,**b**) For the validation phase the normalized mean coverage of each amplicon is plotted for NA12006 (European) versus NA18505 (African) samples for the traditional (**a**) and microdroplet (**b**) PCR methods. (**c**) For each sample (assigned same color as in **Fig. 2**) the average normalized coverage of each amplicon is plotted for traditional versus microdroplet. (**d**) Correlation matrix for all samples depicting Lin's concordance coefficient. All samples show a high correlation among each other within a PCR method but not between the two methods.



variants revealed that four 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: 51.6% (coefficient of variation (c.v.), 12%) and 51.1% (c.v., 13%) of the reads reported the reference allele in the traditional and microdroplet PCR, 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 2,390) 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 that were sequenced in five of the samples as part of the HapMap project. A total of 1,122 variants at 442 positions were identified in these genes, of which 62 positions were not listed as variants in the HapMap data (**Supplementary Table 6**). Thirty-five of these positions are listed as variant in the National Center for Biotechnology Information SNP database, 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 (one position, three 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 1,122 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 3,976 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 chose primer pairs with widely varying characteristics, including primer lengths (17–30 bases), primer melting temperature (Tm, 55.4–61.2 °C), amplicon length (299–659 bp) and amplicon GC content (25.1–81.5%), to determine how this affected the outcome of targeted sequence amplification.

Table 2 Validation phase sequence variant detection rates and concordance

		HapMap SNPs[a]		Variant detection rate		Variant concordance		Reference
Sample	PCR Method	Total	ENCODE	Total %	ENCODE %	Total %[b]	ENCODE %	allele[c] %
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.

Specificity of amplification

A single HapMap sample, NA18858, was amplified using the scale-up phase 3,976

Table 3 Scale-up phase sequence variant detection rates and concordance

Sample	Sequencing platform	HapMap SNPs	Variant detection rate	Variant concordance	Discordant[a] SNPs	Discordant[a] Common[b]
NA18858	Illumina GAII	2,226	99.326	98.83	26	21
NA18858	454 FLX		92.273	98.49	31	

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

amplicons by microdroplet PCR (Online Methods). The resulting amplified material was then divided into two aliquots and sequenced by both the Illumina GAII (short-read platform) and Roche 454 FLX (long-read platform). The Illumina run generated >10 million quality single-end reads with fractions similar to 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 because the microfluidic PCR-amplified DNA is not fragmented in the Roche 454 library preparation protocol, and therefore the nonspecific genomic DNA carryover from the amplification step (~2–4 kb in length) is too long to be efficiently used in library generation. These results indicate that even when simultaneously amplifying 3,976 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 3,976 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 read, of which 96.6% fell within one-fifth and five times the mean coverage (**Fig. 3c**). Furthermore, 90% of all amplicons had coverage at one-fifth the mean or greater for ≥95% of the bases. For 5× coverage, this increased to 98% of the amplicons. Coverage was lower for amplicons >70% GC content but was unaffected by amplicon length (**Supplementary 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 one read and 88.4% falling within one-fifth and five 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. 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, we compared variant bases called in the Illumina and Roche 454 sequences to 2,226 HapMap genotypes (**Table 3**). The Illumina sequence had a call rate of 99.3% with 26 discordant SNPs, whereas 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 except for several discrepant Roche 454 SNPs that were present in homopolymer sequences (**Supplementary 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 these data show that increasing the number of amplicons in the microdroplet PCR reaction by almost an order of magnitude does not affect the resulting data quality.

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 (**Supplementary Discussion** and **Supplementary 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 methods[9,14].

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. In contrast, hybridization-based methods cannot capture individual repetitive elements or homologous exons.

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. We are working toward scaling this to 20,000 targets (~7.5 Mb, ~1/10th the exome) using an expanded content format with five 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 performance[9] and the solution-based hybridization approach has shown the ability to target 22,000 sequences simultaneously[14]. Therefore, the multiplicity of the targets and the scalability of these methods enable a 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 hybridization[22]. Although 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 >99.6% of all amplicons being successful (mean coverage >5 reads) 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 nonspecific 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 μg of starting DNA used in this study limits the applicability of microdroplet PCR for samples with limited quantities. Although optimization has reduced the current requirement to 2 μg, we are working on further improvements to reduce the amount of required starting material to nanogram quantities of DNA (**Supplementary Discussion**).

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

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Accession code. National Institutes of Health Short Read Archive (SRA): SRA009786.

Note: Supplementary information is available on the Nature Biotechnology website.

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 Institutes 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.

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.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturebiotechnology/.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
2. Wheeler, D.A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
3. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
4. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
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.* **124**, 161–170 (2008).
6. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069–1075 (2008).
7. McLendon, R. *et al.* Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
8. Porreca, G.J. *et al.* Multiplex amplification of large sets of human exons. *Nat. Methods* **4**, 931–936 (2007).
9. Turner, E.H., Lee, C., Ng, S.B., Nickerson, D.A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat. Methods* **6**, 315–316 (2009).
10. Krishnakumar, S. *et al.* A comprehensive assay for targeted multiplex amplification of human DNA sequences. *Proc. Natl. Acad. Sci. USA* **105**, 9296–9301 (2008).
11. Albert, T.J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* **4**, 903–905 (2007).
12. Hodges, E. *et al.* Genome-wide *in situ* exon capture for selective resequencing. *Nat. Genet.* **39**, 1522–1527 (2007).
13. Okou, D.T. *et al.* Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods* **4**, 907–909 (2007).
14. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).
15. Anna, S.L., Bontoux, N. & Stone, H.A. Formation of dispersions using "flow focusing" in microchannels. *Appl. Phys. Lett.* **82**, 364–366 (2003).
16. Ahn, K., Agresti, J., Chong, H., Marquez, M. & Weitz, D.A. Electrocoalescence of drops synchronized by size-dependent flow in microfluidic channels. *Appl. Phys. Lett.* **88**, 264105 (2006).
17. Quinlan, A.R. & Marth, G.T. Primer-site SNPs mask mutations. *Nat. Methods* **4**, 192 (2007).
18. Harismendy, O. *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* **10**, R32 (2009).
19. Frazer, K.A., Murray, S.S., Schork, N.J. & Topol, E.J. Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10**, 241–251 (2009).
20. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
21. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
22. Ng, S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).

ONLINE METHODS

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-pl primer pair droplets by using a flow-focusing nozzle (**Supplementary Fig. 1**) in a microfluidic chip[15]. Library elements are tested for droplet size and uniformity (**Supplementary 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. The method for determining library element representation consists of measuring the total volume of primer solution that is made into droplets from each library element and measuring the distribution of droplet sizes present in each library element (**Supplementary Fig. 2**). 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 8,000 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 >90% of the droplets between 7.0 and 9.0 pl. 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 3,976-amplicon library was made from 11 separate 384-well plates in a process whereby an equal number of droplets was 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 quality-control metric is the droplet count for each library element that went into a given collection vial and 11 separate droplet count graphs were generated; **Supplementary Fig. 2d** being representative. In the case of the 457 library, all library elements were collected directly into the final pooling vial and the quality control on the number of droplets was appropriately conducted on all 457 library elements (**Supplementary Fig. 2b**). The histogram of droplet-size distribution (**Supplementary Fig. 2c**) 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 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 number of genomic DNA fragments entering into the construction of the Illumina GAII 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), 37.5 μl of NEBuffer 2 (New England Biolabs (NEB)), 37.5 μl of a mixture of 1 mM each dCTP, dGTP, dTTP (NEB), 18.75 μl of 0.4 mM Biotin-14 dATP (Invitrogen), 211.25 μl nuclease-free water, 7.5 μl DNA polymerase I (10 units/μl) (NEB), 37.5 μl of DNase I (NEB) at a concentration determined by titration to achieve a 2- to 4-kb size distribution. The genomic DNA fragmentation reaction mix is incubated at 15 °C for 90 min and the reaction stopped by adding 37.5 μl of 0.5 M EDTA, pH 8.0 (Ambion). The DNA is purified over a Qiagen MinElute PCR purification column (Qiagen) 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 10× high-fidelity buffer (Invitrogen), 1.6 μl of $MgSO_4$ (Invitrogen), 0.8 μl 10 mM dNTP (NEB), 4.0 μl Betaine (Sigma), 4.0 μl of RDT Droplet Stabilizer (RainDance Technologies), 2.0 μl dimethyl sulfoxide (Sigma) 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 (**Supplementary Fig. 3**). To achieve this, 20 μl of the Genomic DNA Template Mix is delivered to a microfluidic chip where it generates 14-pl template droplets at a rate of 3,000 droplets s^{-1} (**Supplementary Fig. 3b**). In a separate channel primer library droplets are delivered onto the same chip (**Supplementary Fig. 3a**) and spaced to enter the channel carrying the genomic DNA template droplets at a one-to-one ratio (**Supplementary 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 (**Supplementary 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 field[16] to produce a single PCR droplet. The fidelity of the microfluidic flow enables one to set up individual PCR reactions at a rate of 3,000 per second. Over the period of roughly 10 min the 20-μl genomic DNA template mix is processed into greater than 1.5 million PCR droplets (**Supplementary Fig. 3e**). The collection of PCR droplets represents essentially equal numbers of individual reactions from each library element. The PCR droplets are collected in a single standard PCR tube (Axygen) for thermal cycling and further processing (**Supplementary Fig. 3f**). The whole process can be viewed in **Supplementary 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 collected for downstream highly parallel amplification of over a million singleplex PCR reactions.

PCR amplification. Samples are cycled in a Bio-Rad PTC-225 thermocycler as follows: initial denaturation at 94 °C for 2 min; 55 cycles at 94 °C for 15 s, 58 °C for 15 s, and 68 °C for 30 s; a final extension at 68 °C for 10 min 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) is added to the collected PCR emulsion, the tube is vortexed for 15 s and then spun in a microcentrifuge for 10 min. A Gel-Loading Aerosol-Barrier Tip (Fisher Scientific) 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 s; 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 s; 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 min rotating on a Barnstead Labquake Tube Rocker. The beads are placed in a magnetic separation rack for 30 s 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 s. 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) 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 Bioanalyzer 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.

Validation phase: genomic DNA samples. Six HapMap-DNA samples[23] from unrelated individuals of different ethnic backgrounds were obtained from the Coriell Institute for Medical Research (http://www.coriell.org/). Three were of European descent (CEPH; NA11832, NA11992 and NA12006) and three were of African descent (YRI; NA18489, NA18505 and

doi:10.1038/nbt.1583

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 (**Supplementary 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. 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 >600 bp were split into smaller elements to yield 457 targets for primer design. A five-stage approach using 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 (**Supplementary 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 and 300 bp), Tm (59.5–60.5 °C and 57.5–62.5 °C) 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 °C, repeat masking of both 5′ and 3′ flanking sequence). The remaining 33 amplicons were designed within the next four stages. The 457 amplicons had a mean size of 400 bp ranging from 120 to 957 bp (**Supplementary Table 3**). The primer sets were obtained from Integrated DNA Technologies 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 overrepresentation 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 µM of each primer, 0.5 µl 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 0.6× MasterAmp PCR Enhancer (EPICENTRE Technologies) per reaction, in a volume of 12 µl. 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 s 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 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) with the following conditions: duty cycle, 20%; intensity, 8; cycles per burst, 200; time, 6 min. This resulted in fragmented amplicons with 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 lane[24]. The barcodes were sequenced as 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 µl of ligated product per library. After enrichment the samples were quantified by Picogreen (Invitrogen) 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. We then loaded 2.3 pM of each pool into two separate lanes of the flow cell and sequenced them using Illumina paired-end cluster generation kit v1 and SBS kit v2 for 40 cycles on each of the 2 reads.

Scale-up phase: genomic DNA sample. A single HapMap-DNA sample[23] of Yoruban descent (NA18858) was obtained from the Coriell Institute for Medical Research.

Target exons, primer design and synthesis. We selected 3,976 PCR primer pairs from a published set designed to amplify the coding regions of 13,062 genes[25]. The PCR primers were a subset of those published chosen to investigate how varying primer lengths (17–30 bases), primer Tm (55.4–61.2 °C), amplicon length (299–659 bp) and amplicon GC content (25.1–81.5%) affected the outcome of the targeted sequence enrichment and sequencing process. As described[25], primer pairs were designed using Primer3 and were required to be at least 50 bp outside the exon boundary and to not have known SNPs in the five most 3′ bases. Primers were synthesized at Integrated DNA Technologies.

Library preparation for Illumina GA and Roche 454. In the scale-up phase a single HapMap sample, NA18858, was amplified using 3,976 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 NEB Quick Blunting kit the fragments were blunt ended and phosphorylated. The fragments were then concatenated by ligation overnight at 25 °C using the NEB Quick Ligation kit. The concatenated amplicons were fragmented to 400 bp by Adaptive Focused Acoustics using the following settings: mode, frequency sweeping; number of cycles, 1; bath temperature limit, 20 °C; total process time, 3 min; number of treatment, 4; total time per treatment, 45 s; 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 genomic DNA fragmentation. All subsequent steps were done 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.

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 7.5 µg of DNA is input into the nick-translation reaction. After nick translation and purification, ~4.5 µg 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 128–300 ng per sample for all amplicons combined. The amplicon yield for the scale-up phase library with 3,976 primers was 200 ng for all amplicons combined.

doi:10.1038/nbt.1583

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

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 v0.71 (ref. 26) default parameters except for insert size (-a 150) and mismatches (-n 3). Uniquely mapping reads were considered reads with a mapping quality score ≥20 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 <1) across all six samples. For traditional PCR there were four failures of a single amplicon across the six samples; this amplicon performed well for microdroplet-based PCR. Likewise a different amplicon failed for three of the six microdroplet-based amplified samples but performed well in the traditional PCR.

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 before a drop in call rate. The cutoff for the validation phase (paired-end reads) was set at 30. 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 were 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 before 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 <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.

23. The International HapMap Consortium A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
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* **46**, 229–231 (2009).
25. Sjoblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006).
26. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).

doi:10.1038/nbt.1583

Corrigendum: Microdroplet-based PCR enrichment for large-scale targeted sequencing

Ryan Tewhey, Jason B Warner, Masakazu Nakano, Brian Libby, Martina Medkova, Patricia H David, Steve K Kotsopoulos, Michael L Samuels, J Brian Hutchison, Jonathan W Larson, Eric J Topol, Michael P Weiner, Olivier Harismendy, Jeff Olson, Darren R Link & Kelly A Frazer

Nat. Biotechnol. **11, 1025–1031 (2009); published online 01 November 2009; corrected after print 11 November 2009**

In the version of this article initially published, the email address for K.A.F. should have been kafrazer@ucsd.edu. The error has been corrected in the HTML and PDF versions of the article.