nature genetics

Genome-wide *in situ* exon capture for selective resequencing

Emily Hodges[1,4], Zhenyu Xuan[1,2,4], Vivekanand Balija[2], Melissa Kramer[2], Michael N Molla[3], Steven W Smith[3], Christina M Middle[3], Matthew J Rodesch[3], Thomas J Albert[3], Gregory J Hannon[1] & W Richard McCombie[2]

Increasingly powerful sequencing technologies are ushering in an era of personal genome sequences and raising the possibility of using such information to guide medical decisions. Genome resequencing also promises to accelerate the identification of disease-associated mutations. Roughly 98% of the human genome is composed of repeats and intergenic or non–protein-coding sequences. Thus, it is crucial to focus resequencing on high-value genomic regions. Protein-coding exons represent one such type of high-value target. We have developed a method of using flexible, high-density microarrays to capture any desired fraction of the human genome, in this case corresponding to more than 200,000 protein-coding exons. Depending on the precise protocol, up to 55–85% of the captured fragments are associated with targeted regions and up to 98% of intended exons can be recovered. This methodology provides an adaptable route toward rapid and efficient resequencing of any sizeable, non-repeat portion of the human genome.

Creating an index of genetic contributions to human disease requires sensitive methods for exposing genomic variation at both the structural and sequence levels. At present, the latter is accomplished by extensive genomic resequencing of normal and disease-affected individuals. The discovery of common polymorphisms has greatly enhanced the mapping of disease-associated loci. However, *de novo* discovery of mutations that contribute to either inherited or sporadic disease has been limited by the low throughput and high cost of sequencing, even with massively parallel technologies. Nevertheless, recent studies have demonstrated the potential power of resequencing candidate genes to find rare variants underlying complex, disease-associated traits[1] or to profile somatic mutations that confer selective advantages in tumors[2]. In fact, among the large inventory of cancer genes, it has been noted that small-scale, single–base pair events comprise an under-represented class of identified DNA alterations, and that numerous rare variants cooperate to promote tumor growth[3]. These observations provide one of the several compelling motivations for the development of cost-effective resequencing approaches for large genomic regions, as substantial genomic territory will need to be surveyed in large numbers of tumors to obtain an understanding of how mutation drives tumor development.

The recent emergence of widely available sequencing-by-synthesis platforms has enabled searches for disease-associated mutations to be performed across both greater genomic intervals and larger numbers of individuals. These platforms vary in the length of individual sequence reads and in the numbers of reads produced. For example, the 454 system[4] performs picoscale reactions in high-density titer plates to produce hundreds of thousands of 200–300-nucleotide (nt) sequences. By contrast, the Illumina 1G system[5] generates tens of millions of ~30–50-nt reads by *in situ* synthesis on a solid surface. Accuracy, though improving, has yet to reach the level of conventional sequencing approaches. However, a somewhat higher per-read error rate can be readily compensated for by the high sequence coverage achievable with these platforms.

Before the advent of massively parallel sequencing, genome-scale analyses have largely relied on array technologies. Flexible synthesis platforms and increases in array density have allowed the deposition of large numbers of custom-designed oligonucleotides on each array. However, this approach does require a priori knowledge of the specific sequence variants to be detected. Genome-wide tiling arrays have been used for detecting copy number polymorphisms[6] and for whole-genome SNP association studies[7] as well as for mapping transcription factor binding sites by combining chromatin immunoprecipitation and microarray hybridization (ChIP-chip)[8]. Recent studies suggest that in at least some applications, deep sequencing may substitute for array hybridization. For example, ChIP followed by sequencing has been used to map genome-wide histone methylation patterns[9] and transcription factor binding sites[10,11].

Although the aforementioned studies illustrate the power of large-scale sequencing, they also highlight a limitation, in that cost-effective genome-wide analyses still require a simplification of the target population to include only a subset of the genome. Previous studies have accomplished this goal through the use of PCR-based approaches, which require the individual synthesis of large numbers of oligonucleotides and the performance of large numbers of individual

[1]Howard Hughes Medical Institute, Watson School of Biological Sciences, and [2]Watson School of Biological Sciences, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. [3]NimbleGen Systems, Inc., 1 Science Court, Madison, Wisconsin 53711, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to G.J.H. (hannon@cshl.edu) or W.R.M. (McCombie@cshl.edu).

Received 1 August; accepted 15 October; published online 4 November 2007; doi:10.1038/ng.2007.42

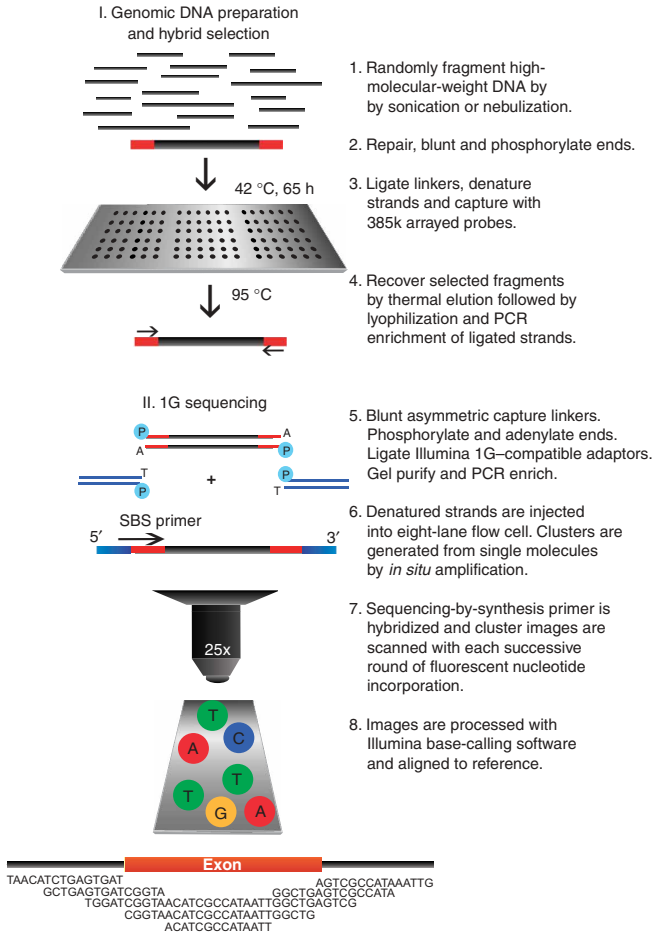

Figure 1 Array-based exon selection scheme followed by Illumina 1G sequencing. Human genomic DNA was randomly fragmented by sonication to an average size of 500 bp. Next, fragmented DNA was hybridized to exon tiling arrays, after which eluted material was ligated with Illumina 1G–compatible linkers and enriched by PCR. The enriched material was added to one lane of an eight-chamber flow cell, and sequence clusters were generated from single molecules. For each base-incorporation cycle, an image was read and a base called. The obtained sequence reads were filtered for quality and genome mapping, then aligned to the reference sequence, which in this case was a target set of exons.

amplification reactions, requiring significant investment and infrastructure. These strategies, although appropriate for subsets of genes, are not well matched in scale to the capacity of next-generation instruments to resequence numerous genetic loci of significant size in a highly parallel fashion. In fact, recent reductions in sequencing cost have made PCR-based selection of targets the most expensive and time-consuming part of large-scale resequencing projects.

Here, we present an approach that combines the individual advantages of microarrays and massively parallel sequencing to enable focused and efficient resequencing of the selected targets. We illustrate the broad potential of the strategy by applying this method for the genome-wide selection of human exons followed by Illumina 1G sequencing.

RESULTS

Overview of the approach

Our principal goal was to enable flexible selection and resequencing of discrete subsets of the genome. Conventional approaches use PCR amplification of genomic fragments to produce a substrate for sequencing; however, we sought to eliminate the burdens both of individual oligonucleotide synthesis and of large numbers of amplification reactions. As an example, a recent study of 22 breast and colorectal tumors minimally required the separate synthesis of 135,483 oligonucleotides and the performance of 3 million individual PCR reactions to obtain 465 Mb of sequence[12]. Previously, we developed methodologies for producing defined libraries of relatively short sequences by highly parallel oligonucleotide synthesis on microarrays[13]. We reasoned that the same approach could be used to produce a hybrid-selection matrix on which genomic fragments corresponding to any desired region of the genome might be captured. This is a departure from conventional array-based approaches in which DNA hybridization to a cognate probe generates a coordinate signal and the intensity is translated into biological information. Instead, we sought to recapture material from the array and use it as a substrate for sequencing. A general outline of the approach is shown in **Figure 1**.

Genome-wide exon capture

As a proof of principle, we chose to target the set of human exons corresponding to RefSeq[14] genes. Exons, both coding and noncoding, represent nearly 2%, or 55 Mb, of the euchromatic human genome (2.85 Gb) and provide an excellent target for enrichment not only because of their modest complexity but also because of their functional significance[15].

Table 1 Hybrid selection results across all human exon chips

Chip ID	Exons	Reads	Reads in exons	Exons with reads[a] (%)	Exon length[b] (bp)	Sequencing coverage (×)	Bases covered (%)	Exons with reads (+300 bp)[c] (%)	Bases covered (+300 bp) (%)	Exons with reads (+500 bp)[d] (%)	Bases covered (+500 bp) (%)
EC1	29,132	318,987	160,522	59.25	5,912,764	0.68	27.16	69.29	67.85	70.83	68.3
EC1-nWGA	*29,132*	*441,060*	*142,793*	*63.26*	*5,912,764*	*0.61*	*22.04*	*78.7*	*71.65*	*81.07*	*73.16*
EC2	30,369	658,425	363,506	70.93	6,143,349	1.5	38.2	77.73	75.98	79.11	76.02
EC3	29,041	108,073	44,121	39.62	6,098,360	0.18	12.03	53.98	48.60	57.18	51.02
EC4	27,706	162,746	81,715	46.27	5,515,092	0.37	18.32	58.71	57.71	60.77	59.3
EC5	24,421	2,507,243	1,174,468	78.3	5,083,643	5.93	53.19	83.69	82.29	85.04	82.78
EC6	26,759	203,023	82,313	47.57	5,670,844	0.37	20.17	60.64	59.22	63.09	61.38
EC7	37,062	239,783	86,556	40.39	8,285,127	0.26	13.87	52.74	48.48	54.92	51.29
Total[e]	204,490	4,198,280	1,993,201	53.77	42,709,179	1.19	25.04	64.53	61.78	66.56	63.29

[a]Without considering flanking regions. [b]Total base pairs covered by at least one exon region. [c]Extending each read by 300 bp. [d]Extending each read by 500 bp. [e]Data from non-WGA sample of EC1 (in italics) are not included.

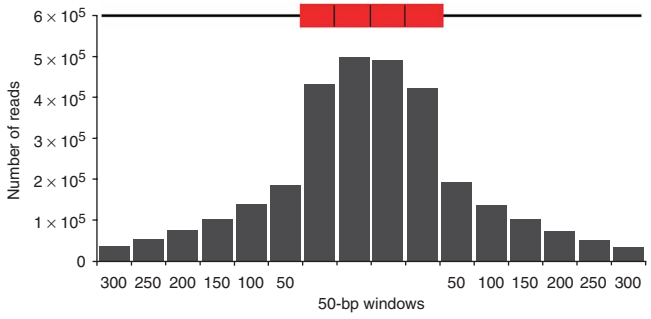

Figure 2 Read-exon distance distributions. Each exon target was split into four equally sized bins depicted by the red bar. The central bases of all reads that map within the exon bins were counted. This process was also performed for 50-bp windows both upstream (left bars) and downstream (right bars) of the exon regions.

We limited our selection to include coding exons and their adjacent splice sites, representing roughly 1% of the human genome. We designed six custom Nimblegen arrays, each containing 385,000 unique features, that tile close to 6 Mb of exonic sequence (24,000–30,000 exons per array) with overlapping 60–90-nt probes having a positional offset of 20 nt. We also generated a seventh array tiling ~37,000 exons corresponding to alternative transcripts for genes represented on the other six arrays (~8 Mb). The set of arrays tiled a total of about 44 million bases of the human genome.

Individual HapMap DNA samples were whole genome amplified and randomly fragmented by sonication to an average size of 500–600 base pairs (bp). We added universal linker sequences before selection to allow amplification, if needed, following recovery from the array. We applied 20 µg of DNA to each capture array, and we hybridized and washed under stringent conditions. Captured DNA was eluted from the array by heat denaturation.

Routinely, we amplified eluted samples through a limited number of PCR cycles using linker sequences as primers. This allowed easy optimization of DNA amounts for loading onto sequencing platforms such as 454 or Illumina 1G, where careful quantification is critical for optimal results. However, we also measured the total amount of eluted material using semiquantitative PCR, in comparison to a similarly sized fragment of known concentration. We found that 30–60 pg of input DNA was recovered, which is within the minimal range of material that can be loaded onto one chamber of the Illumina 1G flow cell. This suggests that the final amplification step, with its attendant risks of introducing artifactual mutations or of skewing the population, can be avoided if required.

Exon capture specificity

For one set of analyses, the captured genomic fragments were identified by end-sequencing using the Illumina 1G platform (results are summarized in **Table 1**). Briefly, we obtained a total of approximately 4.2 million reads, or 109 million bases of sequence, that uniquely map the human genome with at most 2 mismatches in the 26 bases obtained from each read. Among the reads generated from each chip experiment, 36–55% mapped within the specified exon boundaries, and 40–78% of the targeted exons were covered by at least one read (**Table 1**). As the regions represented on each capture array had a total average length of 6 Mb (0.2% of the genome), our results indicated an average of 237-fold enrichment of the targeted sequences.

Probing the genomic distribution of sequence reads that did not map to targeted exons revealed that the majority mapped within ~300 bases of targeted exon boundaries. This can be explained simply, as the average fragment size in the capture (~500 nt) exceeded the average size of a human exon (~200 nt). Thus, each selected fragment, by definition, contained both exonic and associated non-exonic sequences. Plotting the distribution of reads across all targeted regions produced an expected distribution (**Fig. 2**) in which approximately half of the reads mapped to exonic sequences themselves and 28% mapped to flanking sequences, with the number of sequenced ends decreasing with distance from the targeted exon. Taking this finding into account increased the specificity of the capture, from 36–55% (considering strictly 26 bp) to a range of ~55% to ~85% of each sequenced sample corresponding to targeted regions (**Table 1**, **Fig. 3**). This yielded an average enrichment of selected sequences from the genome of 323-fold.

Each capture array probed a non-overlapping set of exon sequences, with the exception of EC7. Pairwise comparisons of material recovered from all seven capture arrays established that each group of reads

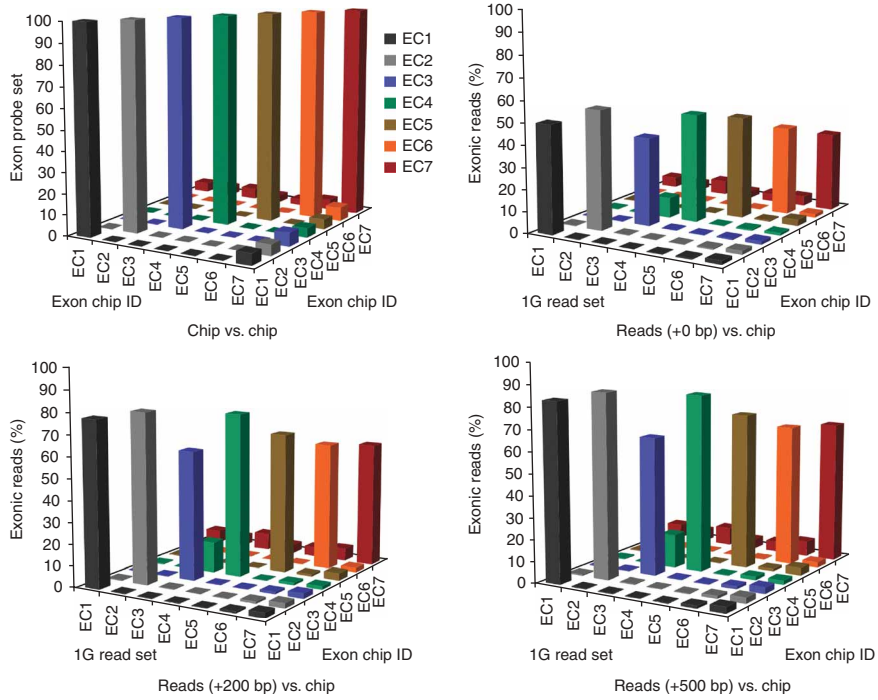

Figure 3 Pairwise comparisons of exon-capture specificity. A pairwise comparison of Illumina 1G reads obtained from each exon array capture is illustrated. The top left panel shows a comparison of probe targets between arrays, and the remaining panels compare the genomic location of mapped reads from one exon array with targets on all seven arrays. The hybrid selection specificity for each exon chip is shown as the percentage of the total number of reads obtained from each array capture. The analysis was repeated by extending the 26-bp reads with 200 and 500 bp downstream of the mapped genomic sequence.

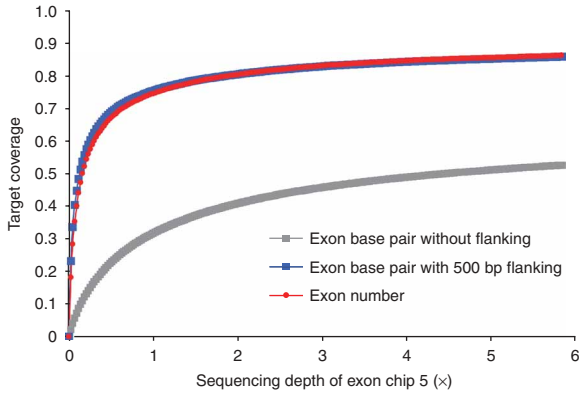

Figure 4 Exon coverage versus Illumina 1G read depth. The plotted curves illustrate the coverage calculated at the exon target level and at the base pair level. The gray square–plotted curve represents the percentage of target exon base pairs covered by 26-bp reads. The blue square–plotted curve represents the theoretical base pair coverage (in %) using 26 bp plus 500 bp flanking sequences. The red circle–plotted curve illustrates the percentage of the target exons mapped by at least one read.

corresponded almost exclusively to the expected target set, confirming the high specificity of the method (**Fig. 3**). As expected, taking into account fragment size markedly influenced the calculated percentage of exon-associated reads.

Although the specificity of array capture performance was good overall, variability did exist between arrays. However, our data suggest that the lower enrichment scores shown by some arrays were to a large extent influenced by the total number of reads, as exemplified by EC3 and EC7. Also worth noting is the observation that a small percentage (below 10%) of reads generated from EC7 overlapped with probes on other arrays, which may be explained by the composition of the alternatively spliced exons among this target set.

Exon coverage

As the recovery was highly specific, we next examined whether the content of the captured sample represented the breadth of targeted exons. Taking into account only the 26-base sequences that tag the end of each captured fragment, more than 78% of exons on array EC5, for example, were recovered. Overall, EC5 capture sequencing resulted in at least 53% of the target base pairs being covered by at least one 26-bp read. Although this coverage is relatively low, one must take two factors into account. First, we have only performed end-sequencing to identify selected fragments; thus, the sequence reads do not represent the total content of the selected sample. Second, we have not exhaustively sequenced selected populations. Considering that we tagged each ~500 base fragment with a 26-bp read allowed extrapolation of the exon coverage to include the entirety of the selected EC5 fragments, which extended exon and base pair coverage to nearly 83% (**Table 1** and **Fig. 4**).

Numerous factors seemed to influence the efficiency of generating sequence coverage of the targeted genomic regions. An obvious factor, of course, is the number of sequence reads generated from each selected sample. For optimal sequence generation on the Illumina platform, fragments of less than 250 bp in length are desirable. This was below the size range initially optimized for capture. Moreover, our analysis suggested that even increased sequencing depth reached a limit in base pair coverage that must reflect a bias in the specific exon fragments that were captured, as the ends of fragments were not randomly distributed within each exonic region. Moreover, there were reproducible differences in the performance of each capture array, with the EC5 array often performing best among those tested. This could simply reflect the larger number of sequences obtained from those samples or it could be significant that EC5 targeted the smallest genomic subset (5 Mb). In this regard, EC7, the array with the greatest target size, had one of the lowest fractions of enriched exons.

Alternative preparation of input DNA

Although whole-genome amplification (WGA) may be necessary for the use of capture approaches with many types of samples (for example, microdissected tumor material), some samples, including the HapMap sample used in this study, will be available in sufficient quantities for direct analysis. To assess whether WGA would introduce wholesale, systematic biases into the fragment populations that were captured, we applied unamplified HapMap DNA to one of the exon arrays (EC1). Illumina sequencing revealed slight differences between amplified and unamplified samples, in terms of both specificity and sequence coverage (**Table 1**). Enrichment specificity was nearly 20% higher in the WGA sample, whereas the capture sensitivity was slightly higher (4%) in the non-WGA sample. Notably, when flanking regions were considered, non-WGA reads covered a significantly higher percentage of exons and base pairs, indicating in part that non-WGA samples had a slightly higher complexity. We did note that WGA before capture, in general, biased slightly against the recovery of AT-rich exons as compared to non-WGA or PCR-amplified samples (see below), but at present we do not understand the basis of this effect.

Although captured fragments of 500–600 bases are well suited to sequencing on some platforms (for example, 454, particularly with paired-end reads), we sought to test the capture with fragment sizes that were more optimal for the Illumina 1G instrument. For this purpose, DNA from the MCF-10A cell line was first nebulized to an average size of 100–200 bp. Illumina 1G sequencing anchors were ligated to the fragments. Two independent samples were applied to EC2, and eluted material was amplified using primers corresponding to the anchors before quantification and sequencing. Overall, we found a reproducible threefold decrease in the specificity of the capture with shorter fragments (**Table 2**). However, this was more than compensated for by the increased sequencing efficiency and the broader distribution of fragment ends. In these trials, we detected 99% of the targeted exons and achieved more than 90% base pair coverage (**Table 2** and **Fig. 5a**). Plotting the distance distribution of sequenced ends in relation to the targeted exons also revealed a pattern different from what was seen with the longer fragments. We observed many fewer reads mapping in exon-adjacent regions (**Fig. 5b**). Considering

Table 2 Alternative exon captures with shorter DNA fragments

Experiment ID	Lane nos.	Reads	Reads in exons	Exons with reads (%)	Sequencing coverage (×)	Base pairs covered (%)[a]
MCF10A-1	1–4	9,508,846	2,798,622 (29.43%)	99.42	11.75	91.40–98.37
MCF10A-2	5–8	7,400,365	2,173,408 (29.36%)	99.30	7.69	89.56–98.00

[a]The base pair coverage reported here considers both the 26-bp read (minimal coverage) + 200 bp (maximal coverage).

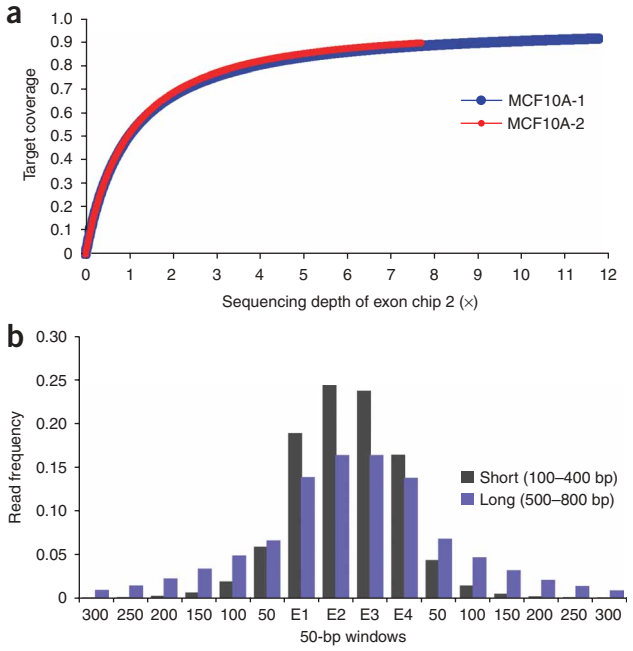

Figure 5 Effect of a variant input DNA preparation on capture efficiency and read depth. (**a**,**b**) Results from two parallel exon captures performed with DNA from MCF10A cells sheared to 100–200-bp fragments are shown. Exon coverage is plotted versus read depth for both datasets (**a**). The distribution of reads obtained from shorter DNA fragments hybridized to exon chip 2 (EC2) were compared to the distribution of reads obtained from longer input DNA (**b**).

that each of the two eluted samples was run on four lanes of a 1G flow cell, providing between 7 and 10 million reads, we estimate that longer fragments may require three times as many lanes to achieve 90% coverage on the Illumina platform.

One application of the method described here is the identification of rare polymorphisms that contribute to disease susceptibility or drug response. To probe detection of SNPs, we searched for captured regions with a coverage depth of at least three sequence reads and with a consensus base call of not more than two mismatches. As we obtained the highest sequencing depth from EC5, we focused on SNPs detected among the EC5-captured DNA. In addition, we incorporated base-quality scores generated by the Illumina 1G software package to distinguish between potential mutations and sequencing errors. The five examples provided in **Supplementary Figure 1** (online) denote previously identified SNPs with high allele frequencies within the represented HapMap population[16], in this case originating from an individual among a group of 30 trios collected from US residents of northern and western European decent. Based on available SNP genotype data for this HapMap individual, we detected 60% of all known SNPs represented on exon chip 5. This is consistent with the degree of base pair and sequence coverage within our analysis of EC5-captured material.

DISCUSSION

We have explored an *in situ* method for the selective enrichment of candidate regions of the genome destined for intensive resequencing. This approach effectively reduces sample complexity while retaining high specificity for the selected regions. Thus, the approach permits deep sequence coverage of virtually any nonrepetitive genomic region of interest. When combined with the power of massively parallel sequencing, this method is robust and efficient, requiring less time and labor than traditional approaches. Perhaps the method's biggest strength is its flexibility. It is readily scalable to address quite large or tightly focused segments of the genome because of its capacity to generate highly dense, complex arrays of probes at various tiling intervals. Moreover, the resolution and unbiased nature of the probe offset (probe frequency) are well suited to the capture of randomly fragmented DNA. Higher-resolution array selections, allowing even denser tiles, will be possible as array feature densities increase.

Our results raise a number of points for consideration when designing this type of genomic capture strategy. The most critical parameters are those that introduce biases in fragment capture, as these will greatly affect the average depth of sequence coverage. One possible source of intrinsic bias is the size of the contiguous targeted region. This is particularly true in the case of exons, as the region tiled by probes is often smaller than the average length of the sheared genomic fragments. Indeed, our data indicate that exons 800 bp or more in length are selectively captured, as indicated by higher numbers of sequence reads (data not shown). Balancing the number of probes according to target length may help to compensate for this tendency. Of course, such compensation could generate other types of biases, particularly if individual probes have not been validated for capture. Another source of potential bias is the base composition of exons, as this could affect hybridization efficiency. We did find that the captured exons from the WGA samples have similar base compositions, whereas the unrepresented exons are significantly AT rich (58% versus 50%, $P < 2 \times 10^{-16}$). This effect, however, did not occur in the non-WGA samples (49.14% and 50.66% AT for nonamplified and PCR-amplified samples, respectively), suggesting that it arises more from sample preparation than from differences in hybridization or elution during capture.

Even though exons are a rather special case, the information gained from these studies will also be applicable to the capture of other genomic regions. Given the capabilities of the new generation of sequencers and the likely benefit of sequencing complete genic regions rather than just exons, this is a natural progression of the approach described herein. All capture protocols are likely, by necessity, to eliminate repeat sequences from probe sets. The average inter-repeat distance in the human genome is roughly 400 bases, producing larger contiguously tiled targets, on average, than do exons. However, such a target is still smaller than the sheared fragment sizes that seem to provide the most highly specific capture.

Overall, the methodologies that we present allow a sensible approach to disease-focused resequencing projects, and we hope that they will expand the capacity of individual investigators or small consortia to efficiently detect new disease-causing mutations.

METHODS

Exon array design and capture. Primary sequence data from all human exons was extracted from Build 36, version 2, of the NCBI's genome annotation. All exons found to be shorter than 135 bases were extended in both the 5′ and 3′ directions to include a minimum of 135 bases of genomic sequence. Overlapping microarray probes (>60 bases) were designed to span each target region, with a probe positioned every 20 bases for the forward strand of the genome. A set of seven arrays (Nimblegen) was created to capture all sequences, with each array containing 385,000 probes.

To avoid nonspecific binding of genomic elements to capture arrays, highly repetitive elements were excluded from probe selection through a method that uses a strategy similar to the WindowMasker program to identify these regions[17]. The process compares the set of probes against a precomputed frequency count histogram of all possible 15-mer probes in the human genome.

For each probe, the frequency counts of the 15-mers comprising the probe are then used to calculate the average 15-mer frequency count of the probe. The higher the average 15-mer frequency count, the more likely the probe is to lie within a repetitive region of the genome. Only probes with an average 15-mer frequency count less than 100 were used. This method results in better coverage of the genome, as compared to the conventionally applied RepeatMasker, while still avoiding highly repetitive regions.

Purified genomic DNA originating from the US National Institute of General Medical Sciences Human Genetic Cell Repository (HapMap sample ID NA12762) was purchased from the Coriell Institute. DNA was whole genome amplified (Qiagen) and randomly fragmented by sonication, treated with the Klenow fragment of DNA polymerase I (NEB) to generate blunt ends, and then phosphorylated with polynucleotide kinase (NEB). Adaptor oligonucleotides (linkers 1 and 2; **Supplementary Table 1** online) were annealed and ligated to the fragment ends. Alternatively, MCF10A DNA was nebulized and preligated with 1G adaptors (see below).

Ligated samples were hybridized to capture arrays in the presence of 1× NimbleGen hybridization buffer (NimbleGen) for approximately 65 h at 42 °C with active mixing using a MAUI hybridization station (NimbleGen). After hybridization, arrays were stringently washed three times for 5 min each with Stringent Wash Buffer (NimbleGen) and rinsed with Wash Buffers 1, 2 and 3 (NimbleGen). Captured DNA fragments were immediately eluted twice, with 250 ml of water each time, at 95 °C. Single-stranded samples were lyophilized and resuspended for amplification using the primers complementary to previously ligated linkers.

Sequencing. Eluted DNA was prepared for 1G sequencing by simultaneously blunting, repairing and phosphorylating ends using a mixture of T4 DNA polymerase, DNA polymerase I Klenow fragment and T4 polynucleotide kinase according to the manufacturer's instructions (Illumina). The repaired and phosphorylated fragments were then subjected to 3′ adenylation with Klenow exo$^-$ fragment (Illumina). After each step, the DNA was recovered using the QIAquick PCR Purification kit (Qiagen) according to the manufacturer's recommendations. Illumina 1G–compatible adaptors were added by rapid ligation to the adenylated fragments, and the ligated fragments were gel purified (on Qiagen purification columns). A minimal PCR amplification step of 18 cycles was performed using Phusion polymerase PCR mix (Finnzymes) and adaptor-compatible primers 1.1 and 2.1 (Illumina). Following amplification, the DNA fragments were purified on Qiagen purification columns. The DNA was quantified using the Nanodrop 7500 and diluted to a working concentration of 10 nM. Cluster generation was performed for samples representing each exon-capture chip in individual lanes of the Illumina 1G flow cell. A custom-designed primer (**Supplementary Table 1**) was hybridized to the prepared flow cell and 36 cycles of base incorporation were carried out on the Illumina 1G analyzer.

Read mapping and coverage analysis. The ELAND program provided with the Illumina 1G software package was used to map all 26-bp reads to the human genome, allowing at most two mismatches. Only reads that mapped uniquely in the genome were retained for further analysis. The mapped genomic locations were compared with the expected exon regions to calculate the coverage and specificity based on their overlapping by at least one base pair. The coverage was defined in two ways: (i) at the target level, by the number of target exon regions with at least one associated read, and (ii) at the base pair level, by the number of base pairs in all exon regions covered by reads. The specificity was calculated as the percentage of reads associated with the specific target exon regions out of the total number of reads that uniquely map the genome. To calculate the theoretical coverage and specificity based on DNA fragment size, the 26-bp reads were extended with 100, 200, 300, 400 or 500 bp of genomic sequence flanking either the 3′ end (for plus-strand reads) or the 5′ end (for minus-strand reads).

Read-exon distance distribution. For all mapped reads, the genomic distance between a mapped read and its closest corresponding exon target was calculated. Each exon was divided into four segments of equal length in base pairs, and each segment was deposited into one of four corresponding bins for all exons. Genomic segments upstream and downstream of exon regions were also deposited into 50-bp bins. For each bin, Illumina 1G reads were counted if their central bases mapped to any sequence therein, and the distribution was plotted.

SNP detection. For each genome alignment, we searched for single-nucleotide discrepancies between the reference genome and the mapped Illumina 1G read. As a strict criterion, nucleotide variations were regarded as reliable SNPs if the quality of the base call achieved a maximum score of 40 (equal to a 0.01% error rate). The base call quality score is an internal measurement of the Illumina 1G pipeline that is comparable to a Phred score. Identified SNPs were verified by comparison with SNP126, a database listing all known polymorphisms from the HapMap project.

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

ACKNOWLEDGMENTS

The authors thank M.Q. Zhang and A. Smith for their help in the read mapping and analysis, J. Silva for providing the MCF10A cell line DNA, and M. Rooks, S. McCarthy and members of the McCombie and Hannon laboratories for helpful discussion. G.J.H. is an Investigator of the Howard Hughes Medical Institute and is supported in part by a kind gift from Kathryn W. Davis and major support from the Stanley Foundation. Purchase of instrumentation and this work were supported in part by grants from the US National Science Foundation and National Institutes of Health (M.Q. Zhang, G.J.H. and W.R.M.).

COMPETING INTERESTS STATEMENT

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

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

1. Topol, E.J. & Frazer, K.A. The resequencing imperative. *Nat. Genet.* **39**, 439–440 (2007).
2. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).
3. Futreal, P.A., Wooster, R. & Stratton, M.R. Somatic mutations in human cancer: insights from resequencing the protein kinase gene family. *Cold Spring Harb. Symp. Quant. Biol.* **70**, 43–49 (2005).
4. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
5. Bentley, D.R. Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.* **16**, 545–552 (2006).
6. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
7. Gunderson, K.L., Steemers, F.J., Lee, G., Mendoza, L.G. & Chee, M.S. A genome-wide scalable SNP genotyping assay using microarray technology. *Nat. Genet.* **37**, 549–554 (2005).
8. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
9. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
10. Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
11. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657 (2007).
12. Sjoblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006).
13. Cleary, M.A. *et al.* Production of complex nucleic acid libraries using highly parallel *in situ* oligonucleotide synthesis. *Nat. Methods* **1**, 241–248 (2004).
14. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61–D65 (2007).
15. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
16. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
17. Morgulis, A., Gertz, E.M., Schaffer, A.A. & Agarwala, R. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics* **22**, 134–141 (2006).