Target-enrichment strategies for next-generation sequencing

Lira Mamanova[1], Alison J Coffey[1], Carol E Scott[1], Iwanka Kozarewa[1], Emily H Turner[2], Akash Kumar[2], Eleanor Howard[1], Jay Shendure[2] & Daniel J Turner[1]

We have not yet reached a point at which routine sequencing of large numbers of whole eukaryotic genomes is feasible, and so it is often necessary to select genomic regions of interest and to enrich these regions before sequencing. There are several enrichment approaches, each with unique advantages and disadvantages. Here we describe our experiences with the leading target-enrichment technologies, the optimizations that we have performed and typical results that can be obtained using each. We also provide detailed protocols for each technology so that end users can find the best compromise between sensitivity, specificity and uniformity for their particular project.

© 2010 Nature America, Inc. All rights reserved.

The ability to read the sequence of bases that comprise a polynucleotide has had an impact on biological research that is difficult to overstate. For the majority of the past 30 years, dideoxy DNA 'Sanger' sequencing[1] has been used as the standard sequencing technology in many laboratories, and its acme was the completion of the human genome sequence[2]. However, because Sanger sequencing is performed on single amplicons, its throughput is limited, and large-scale sequencing projects are expensive and laborious: the human genome sequence took hundreds of sequencing machines several years and cost several hundred million dollars.

The paradigm of DNA sequencing changed with the advent of 'next-generation' sequencing technologies (reviewed in refs. 3,4), which process hundreds of thousands to millions of DNA templates in parallel, resulting in a low cost per base of generated sequence and a throughput on the gigabase (Gb) scale. As a consequence, we can now start to define the characteristics of entire genomes and delineate differences between them. Ultimately, whole-genome sequencing of complex organisms will become routine, allowing us to gain a deeper understanding of the full spectrum of genetic variation and to define its role in phenotypic variation and the pathogenesis of complex traits.

Nevertheless, it is not yet feasible to sequence large numbers of complex genomes in their entirety because the cost and time taken are still too great. To obtain 30-fold coverage of a human genome (90 Gb in total), would currently require several sequencing runs and would cost tens of thousands of dollars. In addition to the demands such a project would place on laboratory time and funding, the primary analysis during which the captured image files are processed, as well as storage of the sequences, would place a substantial burden on a research center's informatics infrastructure.

Consequently, considerable effort has been devoted to develop 'target-enrichment' methods, in which genomic regions are selectively captured from a DNA sample before sequencing. Resequencing the genomic regions that are retained is necessarily more time- and cost-effective, and the resulting data are considerably less cumbersome to analyze. Several approaches to target enrichment have been developed (**Fig. 1**), and there are several parameters by which the performance of each can be measured, which vary from one approach to another: (i) sensitivity, or the percentage of the target bases that are represented by one or more sequence reads; (ii) specificity, or the percentage of sequences that map to the intended targets; (iii) uniformity, or the variability in sequence coverage across target regions; (iv) reproducibility, or how closely results obtained from replicate experiments correlate; (v) cost; (vi) ease of use; and (vii) amount of DNA required per experiment, or per megabase of target.

[1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [2]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. Correspondence should be addressed to J.S. (shendure@u.washington.edu) or D.J.T. (djt@sanger.ac.uk).

PUBLISHED ONLINE 28 JANUARY 2010; CORRECTED AFTER PRINT 12 APRIL 2010; DOI:10.1038/NMETH.1419

QIAGEN Exhibit 1038
Page 1

A technology that typically has a high specificity and uniformity will require less sequencing to generate adequate coverage of sequence data for the downstream analysis, making the sequencing more economical. In addition to these factors, when assessing which target-enrichment technology is the most appropriate for a particular project, thought must be given to how well matched each method is to the total size of intended target region, the number of samples (**Fig. 2**) and whether or not sample multiplexing is required to most efficiently use sequencer throughput.

Here we describe the most widely used approaches to target enrichment, our experiences with each and the optimizations that we have performed. We also provide detailed protocols, which we have developed with the aim of finding the best compromise between the parameters described above.

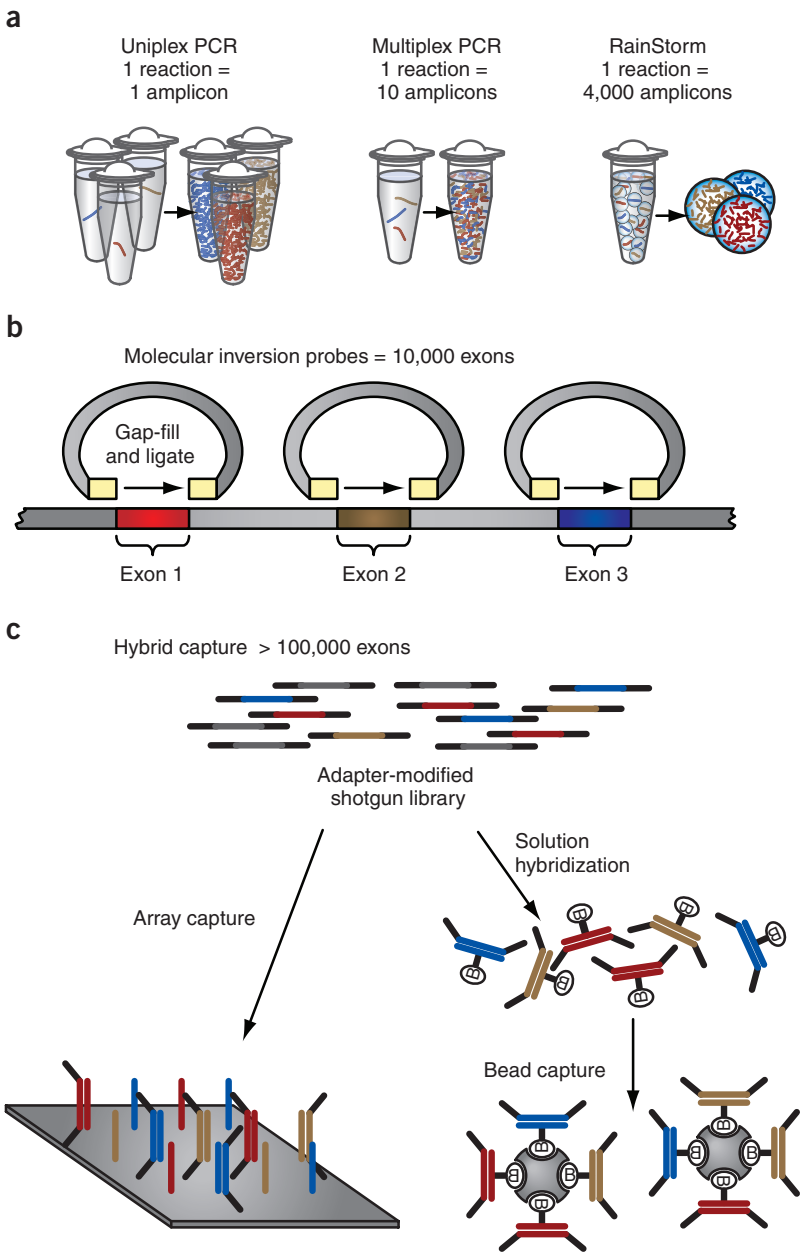

Figure 1 | Approaches to target enrichment. (**a**) In the uniplex PCR–based approach, single amplicons are generated in each reaction. In multiplexed PCR, several primer pairs are used in a single reaction, generating multiple amplicons. On the RainStorm platform, up to 4,000 primer pairs are used simultaneously in a single reaction. (**b**) In the MIP-based approach, probes consisting of a universal spacer region flanked by target-specific sequences are designed for each amplicon. These probes anneal at either side of the target region, and the gap is filled by a DNA polymerase and a ligase. Genomic DNA is digested, and the target DNA is PCR-amplified and sequenced. (**c**) In the hybrid capture–based approach, adaptor-modified genomic DNA libraries are hybridized to target-specific probes either on a microarray surface or in solution. Background DNA is washed away, and the target DNA is eluted and sequenced.

PCR

PCR has been the most widely used presequencing sample preparation technique for over 20 years[5], and it is particularly well suited to a Sanger sequencing–based approach, in which a single PCR can be used to generate a single DNA sequence and in which the sequence read length is comparable to that of a typical PCR amplicon. PCR is also potentially compatible with any next-generation sequencing platform, though to make full use of the high throughput, a large number of amplicons must be sequenced together. However, PCR is difficult to multiplex to any useful degree: the simultaneous use of many primer pairs can generate a high level of nonspecific amplification, caused by interaction between the primers, and moreover amplicons can fail to amplify[6,7]. Clever derivatives of multiplex PCR have been developed[8–10], but in practice, it is often more straightforward to perform PCRs in uniplex. Additionally, there is an upper limit to the length of amplicon that can be generated by long PCR[11]: in our experience very long PCRs tend to lack robustness, and for PCR amplification of contiguous regions, we prefer to design overlapping PCRs that are no more than 10 kilobases (kb) long. Each individual PCR must be validated and, ideally, optimized to make amplification as efficient as possible to minimize the total mass of DNA required.

After amplification, the concentration of products must be normalized before pooling to avoid sequencing one dominant PCR product above all others. There are several ways to approach normalization at this stage, but the most reliable way is to visually inspect the intensity of bands on an agarose gel, alongside a quantitative ladder. Consequently, there is an upper limit to the size of genomic target that can realistically be selected by PCR because of the workload involved. We recommend using long PCR to target regions that are up to several hundred kilobases long, as this is feasible both from the perspectives of workload and the quantity of DNA required.

By current standards, a single lane of a paired-end, 76-base sequencing run

© 2010 Nature America, Inc. All rights reserved.

QIAGEN Exhibit 1038
Page 2

would generate an average coverage of about 30,000-fold for a 100-kb target, clearly a massive excess. For the sequencing to be economical, it is necessary to barcode and pool many samples and to sequence these pools in a single lane. Several approaches to sample barcoding have been reported[12–14], but we have found ligation of barcodes to fragmented PCR amplicons to give uneven sequence coverage of different samples.

We developed a protocol for barcoding 96 samples, in which the library is prepared in 96-well plates and the barcode is included in the central region of the reverse PCR primer (**Supplementary Protocol 1**). We validated this strategy by analyzing a 25-kb region in DNA from several human populations worldwide. We sequenced 96 libraries per flowcell lane and generated 50-base paired-end sequence reads, with an additional 8 bases of sequence to generate the tag sequences. Sequence data from this study have been deposited in the European Short Read Archive. The average coverage obtained from these sequences was high: median >225-fold per lane for native DNA, and 175-fold for whole genome–amplified samples. Coverage and uniformity was poorer for whole genome–amplified samples than for genomic DNA, especially for the longest amplicon in the pool, suggesting that biases were introduced during whole-genome amplification, as has been noted previously[15,16]. However, the barcoding approach was successful, with 80% of sequenced bases covered within a twofold range of the median for the genomic samples. We called single-nucleotide polymorphisms (SNPs) at >99% of sites in approximately 98% of samples and detected 63 high-confidence SNPs; 27 of them were new and 23 were rare.

Improvements for PCR

Although this PCR-based approach was highly effective, there are several areas in which it could be improved. First, a reduction in the cost of library preparation reagents would have a major impact on the overall cost because a separate sequencing library is required for each DNA sample, making library preparation very expensive, for even a small number of lanes of sequencing. Second, improvement in the accuracy of pooling the tiled amplicons, which impacts sequence uniformity, is needed because quantifying tiled amplicons by quantitative PCR is still difficult to achieve for tens to hundreds of amplicons per sample. Third, the use of 5′-blocked primers would achieve greater sequence uniformity across amplicons[14]. Fourth, the use of a greater depth of tiling in the PCRs is another area for improvement. The failure of long PCRs has a major impact on coverage uniformity, but if every base in the target locus is covered by at least two overlapping PCRs, failure of one of these PCRs will not result in the 'loss' of that base. Finally, the use of error-correcting barcodes would allow a greater proportion of pooled sequences to be deconvoluted[17]. Using Hamming codes for tag design[18], it is possible to make tagsets in which single nucleotide-sequencing errors can be corrected, and in which two errors and single insertion-deletions can be detected unambiguously (**Supplementary Table 1**).

It is possible to design long PCR primers for close to 100% of desired targets, but in practice, not all reactions will yield a product after amplification. This can be problematic for samples in which the integrity of the DNA is low, such as clinical specimens. Similarly, when there are SNPs in the primer annealing regions, one allele may be amplified preferentially[19]. Such difficulties can usually be overcome by optimization, primer redesign, greater tiling of amplicons or using a combination of long and short PCR.

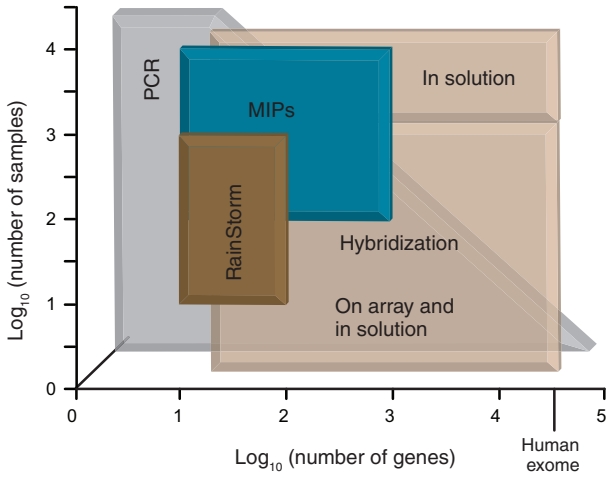

Figure 2 | Suitability of different target-enrichment strategies to different combinations of target size and sample number. Suitability was estimated from the perspective of the feasibility with which each method could be applied to the various combinations of target size and sample number, rather than the cost.

The RainStorm platform, developed by RainDance Technologies, is a convenient solution to many of the problems encountered in a standard PCR–based approach (http://www.raindancetechnologies.com/applications/next-generation-sequencing-technology.asp/). The technology uses microdroplets, similarly to emulsion PCR[20,21]. Each droplet supports an independent PCR and can be made to contain a single primer pair along with genomic DNA and other reagents. The entire population of droplets represents hundreds to thousands of distinct primer pairs and is subjected to thermal cycling, after which this emulsion is broken and products are recovered. The mixture of DNA amplicons can then be subjected to shotgun library construction and massively parallel sequencing. During the microdroplet PCR, different primer pairs cannot interact with each other, which removes one of the primary constraints on conventional multiplex PCR. The microdroplet approach also prevents direct competition of multiplex PCRs for the same reagent pool, which should improve uniformity relative to conventional multiplex PCR. The current maximum number of primer pairs that can be used is 4,000, though it is expected that the number will reach 20,000 by mid 2010 (J. Lambert, personal communication).

The proof of concept for this approach has been published recently[22]. In one experiment, the authors targeted 457 amplicons of variable size (119–956 bp) and G+C content (24–78%), totalling to 172 kb. In six samples, 84% of uniquely mapping reads aligned to targeted amplicons, and 90% of targeted bases were represented within a 25-fold abundance range. In a second experiment, they targeted 3,976 amplicons representing an aggregate target of 1.35 megabases (Mb) and observed that 79% of uniquely mapping reads aligned to targets and 97% of targeted bases were covered within a 25-fold abundance range. The specificity and uniformity of the approach compare well with those of the alternatives, and base calling demonstrated good concordance with expected HapMap genotypes. One limitation is that the approach currently has relatively high input requirements (7.5 μg per sample), but this may be reduced with optimization. In terms of the flexibility of targeting, it is reasonable to expect that this approach will have advantages and disadvantages analogous to those of conventional PCR primer design.

© 2010 Nature America, Inc. All rights reserved.

QIAGEN Exhibit 1038
Page 3

Table 1 | Performance of target-enrichment methods

	PCR	MIP	On-array hybrid capture	In-solution hybrid capture
Cost	High	<10 samples, high; >100 samples, low	Medium	<10 samples, medium; >10 samples, low
Ease of use	Low	High	Medium	High
Mass DNA	~8 µg for 1 Mb of 2× tiled, 5 kb amplicons	As little as 200 ng	10-15 µg per array for up to 30 Mb target	3 µg for up to 30 Mb target
Sensitivity	>99.5%	>98%, with stringent design constraints	98.6% of CTR[a]	>99.5% of CTR[a]
Specificity	93% for HapMap DNA samples, 72% for whole genome–amplified samples	>98%	Up to 70% mapping to CTR[a] for exons; higher for contiguous regions	Up to 80% mapping to CTR[a] for exons; higher for contiguous regions
Uniformity	80% of bases within twofold range of median	58% of CTR within tenfold coverage range; 88% within 100-fold coverage range	60% of CTR[a] within 0.5–1.5-fold of mean coverage (mapping quality[b] 30)	61% of CTR[a] within 0.5–1.5-fold of mean coverage (mapping quality[b] 30)
Reproducibility	Up to 100%	0.92 rank-order correlation[c]	For 10^7 paired-end sequences, >95% reproducibility at tenfold between two samples	For 10^7 paired end-sequences, >96% reproducibility at tenfold between two samples

[a]CTR, capture target region, that is, the regions of the desired target region to which probes could be designed after repeat masking. [b]Mapping qualities were calculated by the mapping software, MAQ, and indicate the probability that the mapping location is correct. A score of 30 or greater indicates that the quality of a read was good, and that it mapped unambiguously to that location with few mismatches. [c]Rank-order correlation in capture efficiency distributions between independent samples.

© 2010 Nature America, Inc. All rights reserved.

Even with an efficient, automated PCR pipeline, it is not feasible to use conventional PCR to target genomic regions that are several megabases in size because of the high cost of primers and reagents and the DNA input requirements, particularly in large sample sets (**Fig. 2**). Similarly, there is a limit to the maximum target size that can be selected using the RainStorm platform (2–3 Mb), and its sample throughput is limited to approximately 8 per workday (**Fig. 2**). Consequently, for very large target regions such as the approximately 30 Mb human exome, or to select moderately sized regions in very large numbers of samples, other approaches to target enrichment should be used.

MOLECULAR INVERSION PROBES

Various enzymatic methods for targeted amplification are compatible with extensive multiplexing based on target circularization[23–25]. One approach in the latter category relies on the use of molecular inversion probes (MIPs), which initially had been developed for multiplex target detection and SNP genotyping[26–30]. Single-stranded oligonucleotides, consisting of a common linker flanked by target-specific sequences[31,32], anneal to their target sequence and become circularized by a ligase. Uncircularized species are digested by exonucleases to reduce background, and circularized species are PCR amplified via primers directed at the common linker. To adapt this method to perform exon capture in combination with next-generation sequencing, a DNA polymerase can be used to 'gap-fill' between target-specific MIP sequences designed to flank a full or partial exon, before ligase-driven circularization, thereby capturing a copy of the intervening sequence[24]. The assay initially demonstrated low uniformity, largely owing to inefficiencies in the capture reaction itself, but more recently an optimized, simplified protocol for MIP-based exon capture has been reported[33]. This revised protocol (**Supplementary Protocol 2**) retains the high specificity of MIP capture, with >98% of mapped reads aligning to a targeted exon but additionally, uniformity is markedly improved, with 58% of targeted bases in 13,000 targets captured to within a tenfold range and 88% to within a 100-fold range (**Fig. 3a** and **Table 1**).

The improved capture uniformity resolves the issue of stochastic allelic bias that plagued the initial proof of concept, showing that accurate genotypes can be derived from massively parallel sequencing of MIP capture products. Furthermore, MIP amplification products can be directly sequenced on a next-generation sequencing platform to interrogate variation in targeted sequences, thereby bypassing the need for shotgun library construction.

Our current view is that the approach of MIP-based capture followed by direct sequencing may be most relevant for projects involving relatively small numbers of targets but large numbers of samples (**Fig. 2**). This is based on the following characteristics. (i) Gap-fill reactions and PCRs take place in aqueous solution, in small volumes, so they are easy to scale to large numbers of samples on 96-well plates; no mechanical shearing, gel-based size purification, ligation or A-tailing is required. (ii) Sample-identifying barcodes can be nested in one of the primers used in post-capture amplification, allowing products from multiple samples to be pooled and sequenced in a single lane. (iii) As with PCR, capture is performed directly on genomic DNA rather than after conversion to a shotgun library, reducing input requirements to as low as 200 ng[34].

The main disadvantages of using MIPs for target enrichment are, first, that capture uniformity, though markedly improved, compares poorly with the most recent reports on capture by hybridization and is the foremost challenge for the approach. To help circumvent this, MIPs can potentially be grouped into sets based on similar capture efficiencies because biases tend to be systematically reproducible[34]. Also, modeling of the causes of nonuniformity can be fed back to MIP design algorithms. Second, MIP oligonucleotides can be costly and difficult to obtain in large numbers to cover large target sets. To mitigate the high cost of column-based oligonucleotide synthesis, thousands of oligos can be obtained by synthesis and release from programmable microarrays (Agilent[24]; LC Sciences). Provided that these are designed in an amplifiable format, they can potentially be used to generate MIP probes to support thousands of samples. Alternatively, one can undertake column-based synthesis of individual MIPs followed by pooling. Although the initial cost for this can be high, sufficient material is obtained to support an extraordinarily large number of capture reactions[24]. The availability of individual probes would also facilitate empirical repooling to improve capture uniformity (J.S.; unpublished data). Finally, it is worth noting that MIPs offer flexibility to address

QIAGEN Exhibit 1038
Page 4

a range of related applications, for example, DNA methylation, RNA editing and allelic imbalance in expression[34–36].

HYBRID CAPTURE

On-array capture

The principle of direct selection is well-established[37,38]: a shotgun fragment library is hybridized to an immobilized probe, nonspecific hybrids are removed by washing and targeted DNA is eluted. Roche NimbleGen and their collaborators were the first to adapt the technology to be compatible with next-generation sequencing[15,16,39]. In the original format, library DNA is hybridized to a single microarray containing 385,000 isothermal probes (the HD1 NimbleGen array), ranging from 60 to 90 bases in length, and with a total capture size of around 4–5 Mb. More recently, the HD2 array has been made available, with 2.1 million probes per array and the ability to capture up to 34 Mb on a single array (**Fig. 2**). The technology was originally designed to be used with the Roche 454 sequencer, but many groups, including ours, expended a considerable amount of effort to modify and optimize protocols for use with the Illumina Genome Analyzer. Agilent's Capture Arrays and comparative genomic hybridization (CGH) arrays are perhaps the most direct competitor to NimbleGen's HD1 arrays, though Agilent's Capture Arrays contain only 244,000 probes on the surface (10^6 for CGH arrays). We found that the performance of both NimbleGen's and Agilent's arrays is similar (**Table 1**).

There are clear advantages to on-array target enrichment of large regions over PCR-based approaches: it is far quicker and less laborious than PCR. But there are also drawbacks: working with microarray slides requires expensive hardware, such as a hybridization station. Additionally, the limit to the number of arrays that a single person can realistically perform each day is approximately 24. As arrays that are hybridized at the same time must also be eluted together, studies with very large numbers of samples are unfeasible. Finally, to have enough DNA library for a target-enrichment experiment, it is necessary to start library preparation with a relatively large amount of DNA, around 10–15 μg, though this is irrespective of whether the capture experiment is for 100 kb or an entire exome.

In-solution capture

To overcome many of these disadvantages, both Agilent and NimbleGen have also developed solution-based target-enrichment protocols. The general principle is similar to array capture, in that there are specific probes designed to target regions of interest from a sequencing library, but whereas an on-array target enrichment uses a vast excess of DNA library over probes, solution capture has an excess of probes over template, which drives the hybridization reaction further to completion using a smaller quantity of sequencing library[40]. In our experiments to test the performance of array versus solution capture, we observed that for smaller target sizes (~3.5 Mb), the uniformity and specificity of sequences obtained from a solution capture experiment tend to be slightly higher than that of array capture (**Fig. 3b,c**). Thus in the 3.5-Mb range, solution capture yields superior sequence coverage of the target regions from a similar yield of sequences. However, for whole-exome captures, both solution and array appear to perform equivalently (**Fig. 4**).

In-solution target enrichment can be performed in 96-well plates, using a thermal cycler, so it is more readily scalable than on-array enrichment and does not require specialized equipment (**Fig. 2**). The principal difference between the Agilent and NimbleGen solution capture products is the nature of the capture probes: the NimbleGen product uses 60–90-mer DNA capture probes, whereas the Agilent one uses 150-mer RNA capture probes. We have not noticed any appreciable difference between the performance of each product.

Library preparation for hybrid capture

Our aim has been to establish a robust production pipeline that can support both on-array and in-solution target enrichment. The manufacturers' workflows for these approaches are very similar and several general principles apply, which allowed us to produce a standard library preparation protocol for both approaches (**Supplementary Protocol 3**).

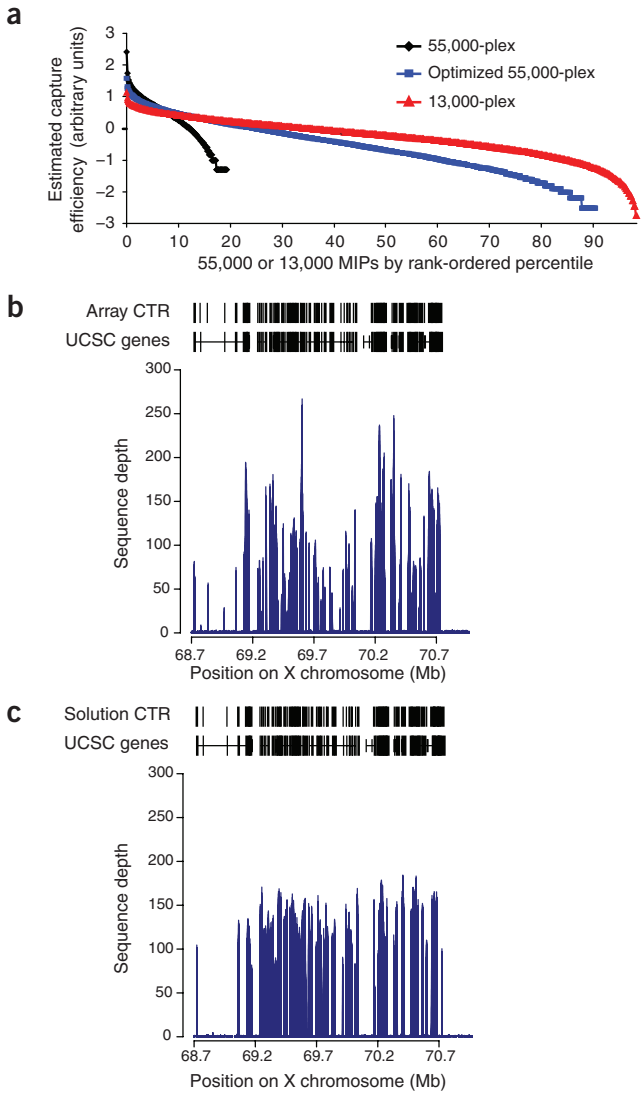

Figure 3 | Uniformity of approaches to target enrichment. (**a**) Capture efficiency obtained with the MIP-based approach, showing improvements in uniformity for optimized protocols and reduced target size. Image was adapted from ref. 33. (**b,c**) A region of human chromosome X, detailing the regions to which capture probes could be designed: the capture target region (CTR) for array capture (**b**) and for solution capture (**c**). Below the CTR are the UCSC genes, taken from the UCSC genome browser. Below this sequence depth obtained from a single lane of Illumina sequences, for a 3.5 Mb capture experiment, is shown.

© 2010 Nature America, Inc. All rights reserved.

QIAGEN Exhibit 1038
Page 5

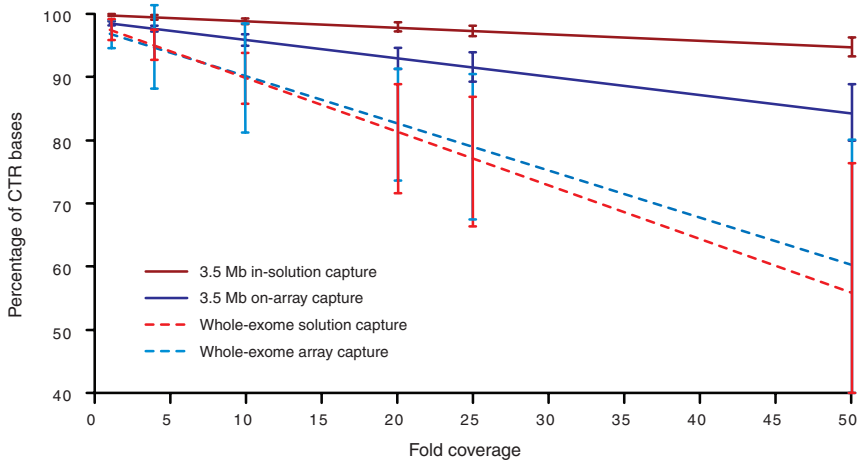

Figure 4 | Coverage plot for array and solution hybrid capture, for 3.5 Mb of exonic target and whole human exome. Values were taken from five independent array and solution experiments, using the same CTR, with each capture using a different DNA sample, and each yielding roughly 10^7 mappable sequences per lane. One lane of sequencing was used for 3.5 Mb captures, whereas two or three lanes were used for the whole exome. Error bars, s.d. ($n = 5$).

Fragment size. Fragment size, obtained by shearing or other fragmentation approaches, has a large influence over the outcome of a target-enrichment experiment, with shorter fragments invariably being captured with higher specificity than longer ones[40,41]. This is not necessarily surprising, given that a longer fragment will contain a higher proportion of off-target sequence, and the effect is especially apparent for exons, whose mean length is relatively short: 164 bp[40] (for example, a 100-bp exon that is part of a 200-bp fragment will be 50% off target just because the captured fragment is larger). However, in our experiments comparing hybrid-capture protocols, the decrease in specificity with increasing template size that we observed was more pronounced than could be accounted for by just the inclusion of off-target portions of longer template sequences and presumably reflects the increase in potential for cross-hybridization between longer fragments themselves.

We assume that there is also a lower size limit to fragments for efficient capture, but in practice the minimum fragment size is determined by the length one would wish to sequence. Longer reads would be expected to map to the reference sequence with lower ambiguity than shorter reads and can help to reduce overrepresentation toward the end of capture probes[40]. For target enrichment of human DNA, we typically generate 76-base paired-end reads, and consequently, it is useful to generate fragments that are around 200 bp to avoid overlap between reads 1 and 2 (**Supplementary Protocol 3**).

Target enrichment sample preparation protocols include a size-selection step to generate a narrow fragment size range, as this is assumed to assist with read mapping. However, this step is not compatible with a high-throughput workflow because it is too labor-intensive, and, in any case, many read-mapping software packages first align each read and then pair the reads[42,43], requiring only a maximum allowed insert size. A score or mapping quality is then assigned to the reads to indicate the probability that the reads are assigned to the incorrect location. Therefore, we investigated the effect of omitting this gel-based size-selection step by performing sequence-capture experiments on libraries prepared with and without this step. Using acoustic shearing, we could generate a sufficiently narrow fragment-size distribution that the size-selection step can be omitted (**Fig. 5a**), and when mapped using MAQ[43], we found little difference between libraries made with or without a size-selection step. In a single experiment, the percentage of mapped reads with score ≥30 (indicating that the base quality of the reads is good and that the read maps unambiguously to the selected location with few mismatches) was just over 1% lower for a library made without size selection compared to the same library after size selection (89.9 and 88.7%, respectively).

PCR optimization. The use of acoustic shearing and removal of the size-selection step resulted in a greater mass of DNA being available for target enrichment than when standard approaches are used, and this allowed us to investigate the effect of performing PCR amplification at different stages of the target-enrichment process.

We noted a negative influence of PCR amplification on the uniformity of enrichment in both on-array and in-solution methods: performing 18 cycles of PCR amplification of libraries both before and after hybridization can introduce severe bias toward neutral G+C content in the resulting sequences (**Fig. 5b**). Avoiding the PCR step altogether before hybridization greatly improved the situation (**Fig. 5c**), so it is desirable to keep PCR amplification to a minimum and only perform it after hybridization.

However, an amplification-free library preparation tends to lack robustness, especially with samples of lower integrity, such as clinical specimens, compared to intact DNA. In these cases, we recommend around six cycles of amplification before hybridization and the use of blocking adapters[41] to avoid a reduction in specificity caused by random concatenation of libraries, so-called 'daisy-chaining'. If no PCR is performed before hybridization, there is no need to use blocking adapters if sequencing is performed on Illumina's Genome Analyzer because the pre-PCR adapters are partially noncomplementary[44] and are thus not problematic in this way. We recommend that hybrid capture be performed following the manufacturers' standard protocols (**Supplementary Protocols 4**,**5**) and that 14–18 cycles of PCR be performed on the samples eluted after hybridization (**Supplementary Protocol 6**).

Prehybridization cleanup. Addition of commercial preparations of C_0t1 DNA to the hybridization reaction is reported to increase specificity[40,41]. C_0t1 DNA comprises short fragments (50–300 bp) of human placental DNA that is enriched for repetitive sequences. Thus it is capable of hybridizing to repetitive sequences in the library DNA, rendering them inert during target enrichment. C_0t1 DNA is generally added in a 5–20-fold excess over the input library. We have observed little difference in performance within this range but typically use a fivefold excess for on-array target enrichment and a 20-fold excess in solution.

Salt concentration is an important factor in determining the specificity and efficiency of hybridization. Any salts in the C_0t1 and library DNA buffers will contribute to the overall salt

© 2010 Nature America, Inc. All rights reserved.

QIAGEN Exhibit 1038
Page 6

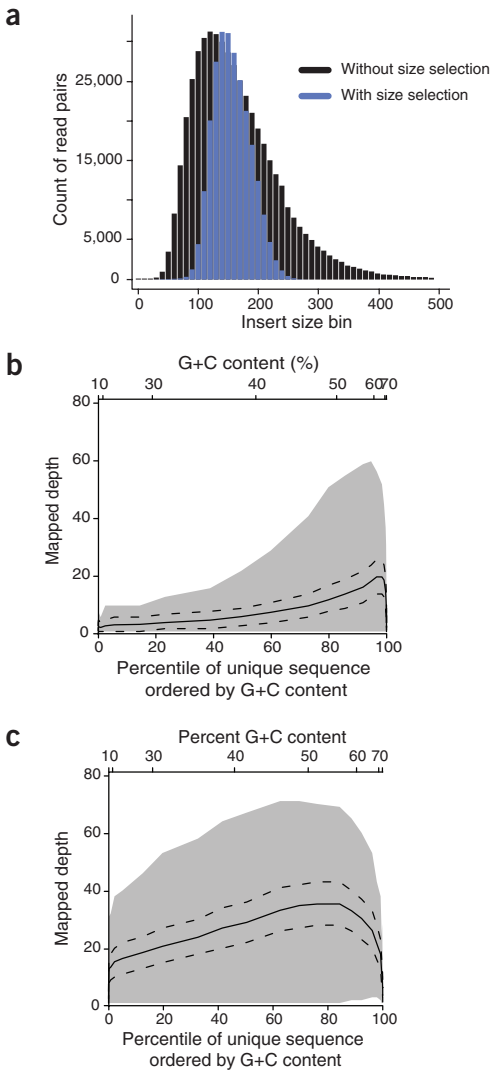

Figure 5 | Library prep optimizations for hybrid capture. (**a**) Distribution of insert sizes derived from mapped sequence data for solution capture performed with and without agarose gel–based size selection. (**b**,**c**) G+C content plot showing mapped sequence data for a 3.5 Mb array capture, in which PCR was performed before and after hybridization (**b**), or in which PCR was performed only after hybridization (**c**). The solid black line indicates the mean value, the dotted lines at either side indicate the s.d., and the shaded area shows the distribution of reads with the indicated G+C content.

concentration in the hybridization buffer, and so we prefer to desalt both the C_0t1 and library DNA before hybridization. A convenient way to achieve this is using solid-phase reversible immobilization (SPRI) beads. These are paramagnetic beads to which nucleic acids can bind reversibly, and captured DNA can be eluted in water[45] (**Supplementary Protocol 3**).

Improvements for hybrid capture

Using **Supplementary Protocols 3** and **6**, we have been able to obtain robust, reproducible target-enrichment results, both on array and in solution, allowing us to transfer target enrichment into a production environment. The turnaround time for synthesis of custom capture arrays (1 or 2 weeks) is typically shorter than for solution probes (~4 weeks), though this is likely to improve as solution probes become more established as a commercial product. We also found in-solution target enrichment to have an equivalent or slightly better performance than on-array enrichment (**Figs. 3b**,**c**, **4** and **Table 1**), and that the former was the more attractive option for high-throughput target enrichment.

Array and solution hybridization are sensitive to sample base composition, and sequences at the extremes of high A+T or G+C content can be lost through poor annealing and secondary structure, respectively. Although not a major issue for human exonic DNA, this sensitivity could be more problematic for other genomes. Another consideration is that it is seldom possible to capture all of a desired target region in a hybrid capture experiment: targets are generally subjected to repeat masking before probe design to avoid capture of homologous repetitive elements. For exonic targets, <5–15% of the primary target region can be lost in this way, leaving a region to which probes could be designed after repeat masking, or target capture region, that constitutes 85% to >95% of the primary target region. For contiguous regions, the percentage of primary target region that is represented in the capture target region is generally lower (~50–65%), but this is highly variable between regions.

CONCLUSIONS

Inevitably, there is always the temptation to quantitatively compare approaches to target enrichment. The specificity of PCR will almost certainly always exceed that of hybrid capture, and its uniformity may never be matched by either hybrid capture or MIPs. But specificity and uniformity are not everything: the chief advantage of these alternative methods is their ability to capture large target regions in a single experiment, more rapidly and conveniently than PCR. To capture the entire 30 Mb human exome, for example, would require at least 6,000 separate PCRs, each of which would need to be optimized, the products would need to be normalized, and a total of around 120 µg of genomic DNA would be required for the experiment. The same could be performed in a single hybrid capture experiment, taking a single day for the library preparation and about two additional days for the hybridization and elution, and requiring as little as 3 µg of DNA.

Target enrichment can be a highly effective way of reducing sequencing costs and saving sequencing time, and has the power to bring the field of genomics into smaller laboratories, as well as being an invaluable tool for the detection of disease-causing variants. Conversely, target enrichment increases sample preparation cost and time. Assuming that the throughput of next-generation runs and our ability to analyze large numbers of whole-genome datasets both continue to increase, and the cost per base of sequence continues to decrease, there will come a point at which it is no longer economical to perform target enrichment of single samples, compared to whole-genome sequencing. The cost of performing target enrichment by hybridization can be reduced by pooling samples before hybridization, though in our experience results from capturing pooled samples on arrays have been poorer than in solution. This is presumably a reflection of the difference in the probe:sample ratio for the two capture methods.

The logical extension of sample pooling is to perform multiplexed target enrichments in which samples are barcoded before capture. We expect this to have many applications in the future, and the technical details of this are currently being worked out.

© 2010 Nature America, Inc. All rights reserved.

QIAGEN Exhibit 1038
Page 7

We will provide updates of protocols at ftp://ftp.sanger.ac.uk/pub/pulldown/.

Accession codes. European Short Read Archive: ERA000184.

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

ACKNOWLEDGMENTS

We thank D. MacArthur, Q. Ayub and C. Tyler-Smith for their work on long PCR and subsequent analyses, P. Akan, A. Palotie, P. Tarpey, H. Arbury and M. Humphries for their work on hybrid capture and E. Sheridan for critical reading of the standard operating procedures. This work was supported by the Wellcome Trust grant WT079643 and by US National Institutes of Health National Human Genome Research Institute grants 5R21HG004749 and 5R01HL094976.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

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

© 2010 Nature America, Inc. All rights reserved.

1. Sanger, F., Nicklen, S. & Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467 (1977).
2. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
3. Mardis, E.R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24**, 133–141 (2008).
4. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
5. Saiki, R.K. *et al.* Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**, 487–491 (1988).
This paper was the first description of PCR, which, coupled to electrophoretic sequencing, is the primary conventional method for targeted variation analysis.
6. Cho, R.J. *et al.* Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. *Nat. Genet.* **23**, 203–207 (1999).
7. Wang, D.G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).
8. Fredriksson, S. *et al.* Multiplex amplification of all coding sequences within 10 cancer genes by Gene-Collector. *Nucleic Acids Res.* **35**, e47 (2007).
9. Meuzelaar, L.S., Lancaster, O., Pasche, J.P., Kopal, G. & Brookes, A.J. MegaPlex PCR: a strategy for multiplex amplification. *Nat. Methods* **4**, 835–837 (2007).
10. Varley, K.E. & Mitra, R.D. Nested patch PCR enables highly multiplexed mutation discovery in candidate genes. *Genome Res.* **18**, 1844–1850 (2008).
11. Barnes, W.M. PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. *Proc. Natl. Acad. Sci. USA* **91**, 2216–2220 (1994).
12. Craig, D.W. *et al.* Identification of genetic variants using bar-coded multiplexed sequencing. *Nat. Methods* **5**, 887–893 (2008).
13. Cronn, R. *et al.* Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Res.* **36**, e122 (2008).
14. 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).
15. Hodges, E. *et al.* Genome-wide *in situ* exon capture for selective resequencing. *Nat. Genet.* **39**, 1522–1527 (2007).
16. Okou, D.T. *et al.* Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods* **4**, 907–909 (2007).
17. Hamady, M., Walker, J.J., Harris, J.K., Gold, N.J. & Knight, R. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat. Methods* **5**, 235–237 (2008).
18. Hamming, R. Error detecting and error correcting codes. *Bell Syst. Tech. J.* **29**, 147–161 (1950).
19. Ikegawa, S., Mabuchi, A., Ogawa, M. & Ikeda, T. Allele-specific PCR amplification due to sequence identity between a PCR primer and an amplicon: is direct sequencing so reliable? *Hum. Genet.* **110**, 606–608 (2002).
20. Dressman, D., Yan, H., Traverso, G., Kinzler, K.W. & Vogelstein, B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl. Acad. Sci. USA* **100**, 8817–8822 (2003).
21. Tawfik, D.S. & Griffiths, A.D. Man-made cell-like compartments for molecular evolution. *Nat. Biotechnol.* **16**, 652–656 (1998).
22. Tewhey, R. *et al.* Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nat. Biotechnol.* **27**, 1025–1031 (2009).
This paper describes the performance of the RainDance technology, which facilitates multiplex PCR by compartmentalizing primer pairs in distinct microdroplet populations that are then mixed and thermocycled in aggregate.
23. Dahl, F., Gullberg, M., Stenberg, J., Landegren, U. & Nilsson, M. Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic Acids Res.* **33**, e71 (2005).
24. Porreca, G.J. *et al.* Multiplex amplification of large sets of human exons. *Nat. Methods* **4**, 931–936 (2007).
25. Dahl, F. *et al.* Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proc. Natl. Acad. Sci. USA* **104**, 9387–9392 (2007).
26. Faruqi, A.F. *et al.* High-throughput genotyping of single nucleotide polymorphisms with rolling circle amplification. *BMC Genomics* **2**, 4 (2001).
27. Antson, D.O., Isaksson, A., Landegren, U. & Nilsson, M. PCR-generated padlock probes detect single nucleotide variation in genomic DNA. *Nucleic Acids Res.* **28**, E58 (2000).
28. Hardenbol, P. *et al.* Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat. Biotechnol.* **21**, 673–678 (2003).
29. Lizardi, P.M. *et al.* Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nat. Genet.* **19**, 225–232 (1998).
30. Hardenbol, P. *et al.* Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay. *Genome Res.* **15**, 269–275 (2005).
31. Nilsson, M. *et al.* Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science* **265**, 2085–2088 (1994).
32. Landegren, U. *et al.* Molecular tools for a molecular medicine: analyzing genes, transcripts and proteins using padlock and proximity probes. *J. Mol. Recognit.* **17**, 194–197 (2004).
33. 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).
This paper demonstrates a substantially optimized protocol for using molecular inversion probes for exon capture that also enables library-free integration of multiplex capture and next-generation sequencing.
34. Deng, J. *et al.* Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat. Biotechnol.* **27**, 353–360 (2009).
35. Zhang, K. *et al.* Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat. Methods* **6**, 613–618 (2009).
36. Li, J.B. *et al.* Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324**, 1210–1213 (2009).
37. Lovett, M., Kere, J. & Hinton, L.M. Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. *Proc. Natl. Acad. Sci. USA* **88**, 9628–9632 (1991).
38. Parimoo, S., Patanjali, S.R., Shukla, H., Chaplin, D.D. & Weissman, S.M. cDNA selection: efficient PCR approach for the selection of cDNAs encoded in large chromosomal DNA fragments. *Proc. Natl. Acad. Sci. USA* **88**, 9623–9627 (1991).
39. Albert, T.J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* **4**, 903–905 (2007).
This was one of three papers that described solid-phase, hybridization-based enrichment of targeted sequences in shotgun DNA libraries using programmable microarrays.
40. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).
This paper was the first description of the method, now commercialized by Agilent, for solution-phase hybridization-based capture using complex libraries of RNA 'bait' to capture from a shotgun DNA 'pond' library.
41. Hodges, E. *et al.* Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat. Protocols* **4**, 960–974 (2009).
42. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
43. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
44. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
45. Quail, M.A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nat. Methods* **5**, 1005–1010 (2008).

QIAGEN Exhibit 1038
Page 8

Erratum: Target-enrichment strategies for next-generation sequencing

Lira Mamanova, Alison J Coffey, Carol E Scott, Iwanka Kozarewa, Emily H Turner, Akash Kumar, Eleanor Howard, Jay Shendure & Daniel J Turner

Nat. Methods **7, 111–118 (2010); published online 28 January 2010; corrected after print 12 April 2010.**

In the version of this article initially published, the publication date was incorrectly designated as 28 January 2009 instead of 28 January 2010. The error has been corrected in the HTML and PDF versions of the article.

© 2010 Nature America, Inc. All rights reserved.

QIAGEN Exhibit 1038
Page 9