

# Target-enrichment strategies for next-generation sequencing

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

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.

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<sup>1</sup> has been used as the standard sequencing technology in many laboratories, and its acme was the completion of the human genome sequence<sup>2</sup>. 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.

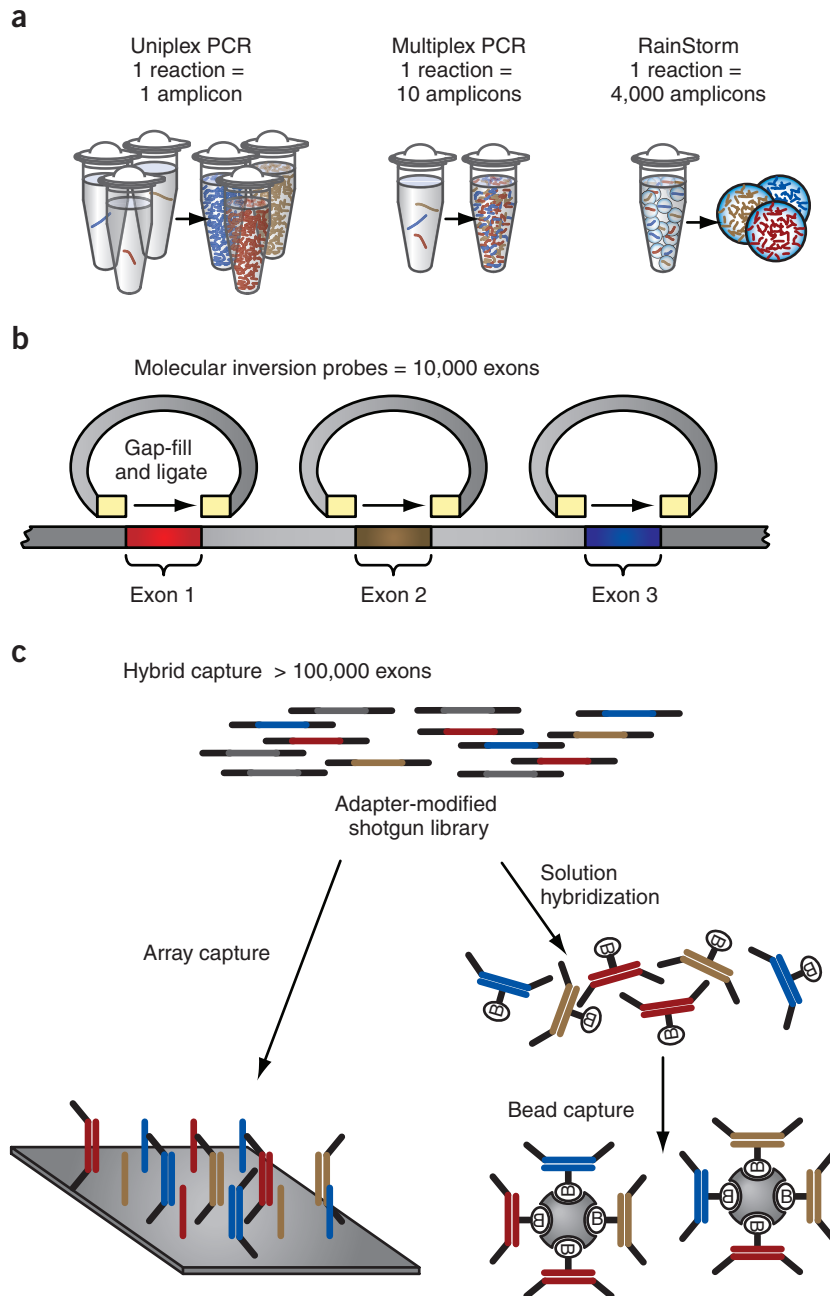
Consequently, considerable effort has been devoted to develop ‘target-enrichment’ methods, in which genomic regions are selectively captured from a DNA sample before sequencing. Resequencing the genomic regions that are retained is necessarily more time- and cost-effective, and the resulting data are 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.

<sup>1</sup>The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. <sup>2</sup>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. (djts@sanger.ac.uk).

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 pre-sequencing sample preparation technique for over 20 years<sup>5</sup>, 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<sup>6,7</sup>. Clever derivatives of multiplex PCR have been developed<sup>8–10</sup>, 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<sup>11</sup>: 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

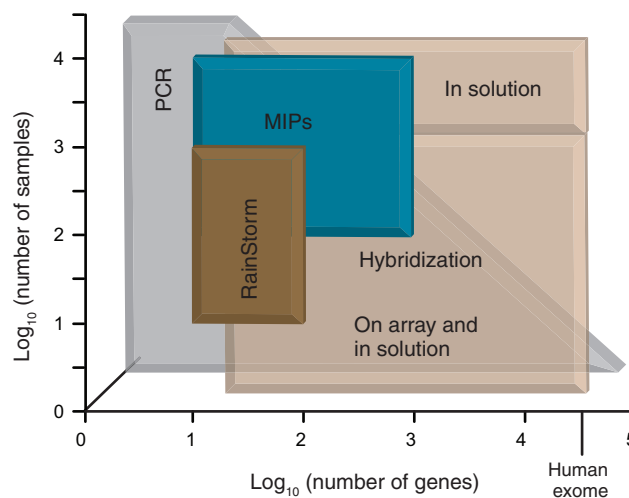
would generate an average coverage of about 30,000-fold for a 100-kb target, clearly a massive excess. For the sequencing to be economical, it is necessary to barcode and pool many samples and to sequence these pools in a single lane. Several approaches to sample barcoding have been reported<sup>12–14</sup>, 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<sup>15,16</sup>. 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<sup>14</sup>. 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<sup>17</sup>. Using Hamming codes for tag design<sup>18</sup>, 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<sup>19</sup>. 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<sup>20,21</sup>. 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<sup>22</sup>. 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.

**Table 1** | Performance of target-enrichment methods

	PCR	MIP	On-array hybrid capture	In-solution hybrid capture
<b>Cost</b>	High	<10 samples, high; >100 samples, low	Medium	<10 samples, medium; >10 samples, low
<b>Ease of use</b>	Low	High	Medium	High
<b>Mass DNA</b>	~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
<b>Sensitivity</b>	>99.5%	>98%, with stringent design constraints	98.6% of CTR <sup>a</sup>	>99.5% of CTR <sup>a</sup>
<b>Specificity</b>	93% for HapMap DNA samples, 72% for whole genome– amplified samples	>98%	Up to 70% mapping to CTR <sup>a</sup> for exons; higher for contiguous regions	Up to 80% mapping to CTR <sup>a</sup> for exons; higher for contiguous regions
<b>Uniformity</b>	80% of bases within twofold range of median	58% of CTR within tenfold coverage range; 88% within 100-fold coverage range	60% of CTR <sup>a</sup> within 0.5–1.5-fold of mean coverage (mapping quality <sup>b</sup> 30)	61% of CTR <sup>a</sup> within 0.5–1.5-fold of mean coverage (mapping quality <sup>b</sup> 30)
<b>Reproducibility</b>	Up to 100%	0.92 rank-order correlation <sup>c</sup>	For 10 <sup>7</sup> paired-end sequences, >95% reproducibility at tenfold between two samples	For 10 <sup>7</sup> paired end-sequences, >96% reproducibility at tenfold between two samples

<sup>a</sup>CTR, capture target region, that is, the regions of the desired target region to which probes could be designed after repeat masking. <sup>b</sup>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.

<sup>c</sup>Rank-order correlation in capture efficiency distributions between independent samples.

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<sup>23–25</sup>. 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<sup>26–30</sup>. Single-stranded oligonucleotides, consisting of a common linker flanked by target-specific sequences<sup>31,32</sup>, 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<sup>24</sup>. 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<sup>33</sup>. 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<sup>34</sup>.

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<sup>34</sup>. 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<sup>24</sup>; 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<sup>24</sup>. 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

a range of related applications, for example, DNA methylation, RNA editing and allelic imbalance in expression<sup>34–36</sup>.

## HYBRID CAPTURE

### On-array capture

The principle of direct selection is well-established<sup>37,38</sup>: 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<sup>15,16,39</sup>. 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  $\mu$ g, though this is irrespective of whether the capture experiment is for 100 kb or an entire exome.

### In-solution capture

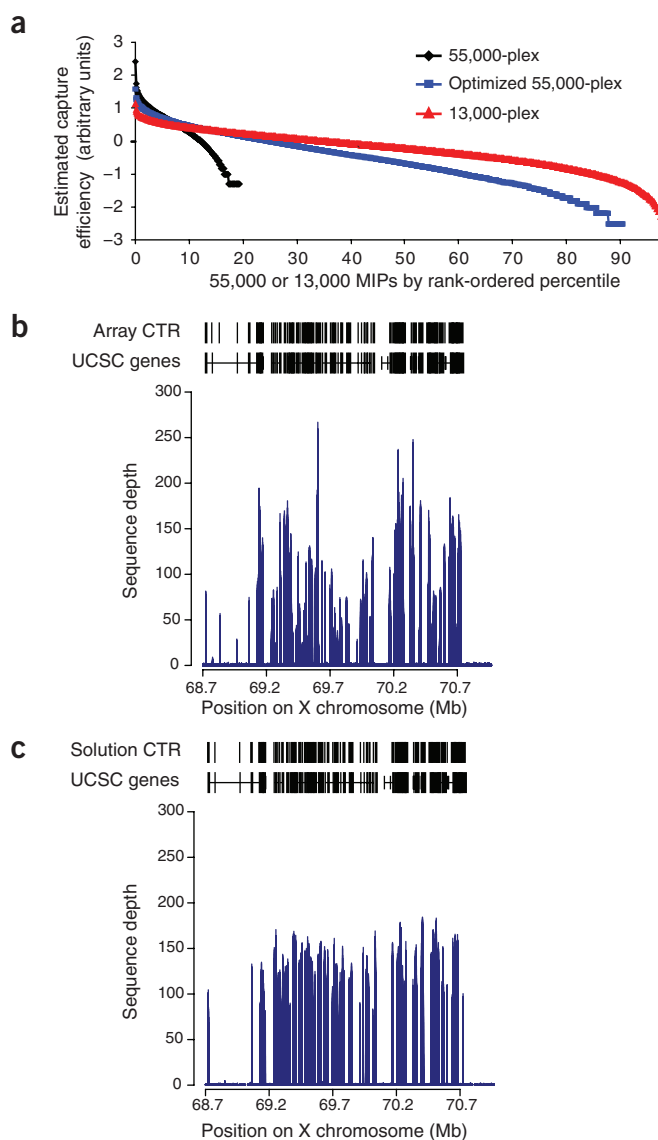
To overcome many of these disadvantages, both Agilent and NimbleGen have also developed solution-based target-enrichment protocols. The general principle is similar to array capture, in that there are specific probes designed to target regions of interest from a sequencing library, but whereas an on-array 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<sup>40</sup>. 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

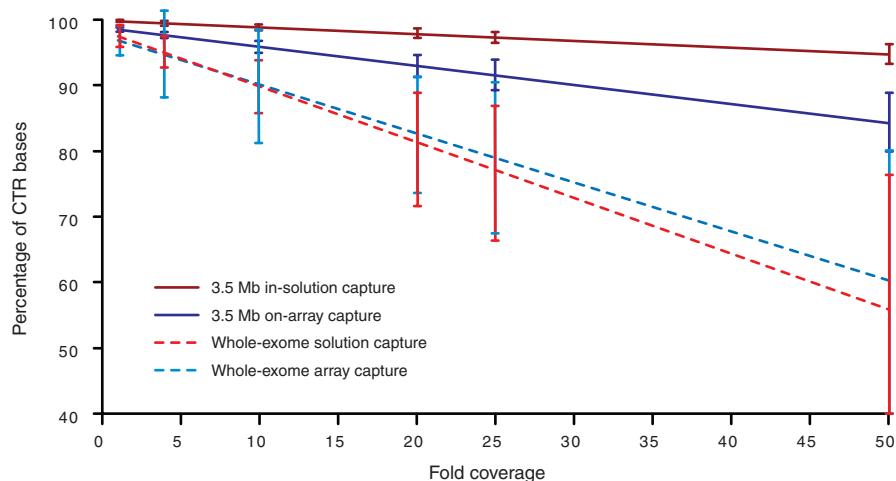
capture products is the nature of the capture probes: the NimbleGen product uses 60–90-mer DNA capture probes, whereas the Agilent one uses 150-mer RNA capture probes. We have not noticed any appreciable difference between the performance of each product.

### 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.



**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<sup>40,41</sup>. 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<sup>40</sup> (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<sup>40</sup>. 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<sup>42,43</sup>, 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<sup>43</sup>, 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  $\geq 30$  (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).

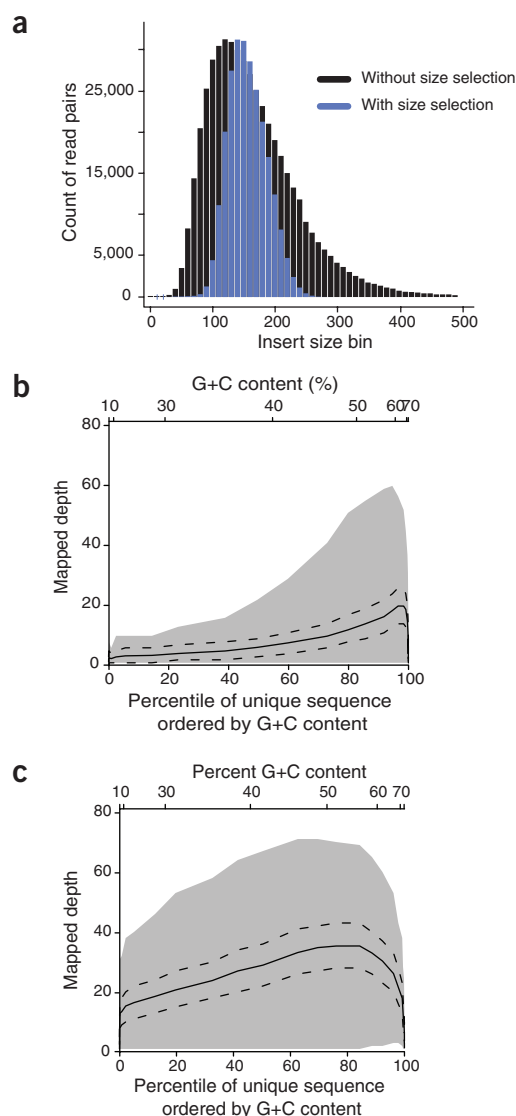
**PCR optimization.** The use of acoustic shearing and removal of the size-selection step resulted in a greater mass of DNA being available for target enrichment than when standard approaches are used, and this allowed us to investigate the effect of performing PCR amplification at different stages of 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<sup>41</sup> 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<sup>44</sup> 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<sub>0</sub>t1* DNA to the hybridization reaction is reported to increase specificity<sup>40,41</sup>. *C<sub>0</sub>t1* 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<sub>0</sub>t1* 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<sub>0</sub>t1* and library DNA buffers will contribute to the overall salt



**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<sub>0</sub>t1* 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<sup>45</sup> (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.

We will provide updates of protocols at <ftp://ftp.sanger.ac.uk/pub/pull-down/>.

**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/>.

- Sanger, F., Nicklen, S. & Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467 (1977).
- International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
- Mardis, E.R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24**, 133–141 (2008).
- Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
- 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.**
- Cho, R.J. *et al.* Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. *Nat. Genet.* **23**, 203–207 (1999).
- Wang, D.G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).
- Fredriksson, S. *et al.* Multiplex amplification of all coding sequences within 10 cancer genes by Gene-Collector. *Nucleic Acids Res.* **35**, e47 (2007).
- 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).
- Varley, K.E. & Mitra, R.D. Nested patch PCR enables highly multiplexed mutation discovery in candidate genes. *Genome Res.* **18**, 1844–1850 (2008).
- 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).
- Craig, D.W. *et al.* Identification of genetic variants using bar-coded multiplexed sequencing. *Nat. Methods* **5**, 887–893 (2008).
- Cronn, R. *et al.* Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Res.* **36**, e122 (2008).
- 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).
- Hodges, E. *et al.* Genome-wide *in situ* exon capture for selective resequencing. *Nat. Genet.* **39**, 1522–1527 (2007).
- Okou, D.T. *et al.* Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods* **4**, 907–909 (2007).
- 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).
- Hamming, R. Error detecting and error correcting codes. *Bell Syst. Tech. J.* **29**, 147–161 (1950).
- 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).
- 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).
- Tawfik, D.S. & Griffiths, A.D. Man-made cell-like compartments for molecular evolution. *Nat. Biotechnol.* **16**, 652–656 (1998).
- 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.**
- 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).
- Porreca, G.J. *et al.* Multiplex amplification of large sets of human exons. *Nat. Methods* **4**, 931–936 (2007).
- Dahl, F. *et al.* Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proc. Natl. Acad. Sci. USA* **104**, 9387–9392 (2007).
- Faruqi, A.F. *et al.* High-throughput genotyping of single nucleotide polymorphisms with rolling circle amplification. *BMC Genomics* **2**, 4 (2001).
- 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).
- Hardenbol, P. *et al.* Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat. Biotechnol.* **21**, 673–678 (2003).
- Lizardi, P.M. *et al.* Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nat. Genet.* **19**, 225–232 (1998).
- 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).
- Nilsson, M. *et al.* Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science* **265**, 2085–2088 (1994).
- 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).
- 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.**
- Deng, J. *et al.* Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat. Biotechnol.* **27**, 353–360 (2009).
- Zhang, K. *et al.* Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat. Methods* **6**, 613–618 (2009).
- Li, J.B. *et al.* Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324**, 1210–1213 (2009).
- 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).
- 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).
- 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.**
- Gnrirke, 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.**
- Hodges, E. *et al.* Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat. Protocols* **4**, 960–974 (2009).
- Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
- Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
- Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
- Quail, M.A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nat. Methods* **5**, 1005–1010 (2008).

## 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

Supplementary figures and text:

**Supplementary Table 1** Error-correcting barcode tags and PCR primers.

**Supplementary Protocol 1** PCR and 96-well library prep standard operating procedure.

**Supplementary Protocol 2** MIP standard operating procedure.

**Supplementary Protocol 3** Hybrid capture library prep standard operating procedure.

**Supplementary Protocol 4** Array capture standard operating procedure.

**Supplementary Protocol 5** Solution capture standard operating procedure.

**Supplementary Protocol 6** Hybrid capture eluate PCR standard operating procedure.

**Supplementary Table 1. Error-correcting barcode tags and PCR primers.**

Column B contains octamer barcodes, and Column D contains PCR primers that contain these barcodes.

Column C contains the reverse complement of the barcodes given in Column B. This is the sequence that is obtained from the barcodes in the sequencing reaction, and which are robust to errors.

Tag number	Barcode sequence	Single correcting, double & shift detecting octamers	PCR primers
1	ACAAGCTA	TAGCTTGT	CAAGCAGAAGACGGCATAACGCTAGAGATCGGTCTCGGCATTC
2	AAACATCG	CGATGTTT	CAAGCAGAAGACGGCATAACATCGGAGATCGGTCTCGGCATTC
3	ACATTGGC	GCCAATGT	CAAGCAGAAGACGGCATAACATTGGCGAGATCGGTCTCGGCATTC
4	ACCACTGT	ACAGTGGT	CAAGCAGAAGACGGCATAACCACTGTGAGATCGGTCTCGGCATTC
5	AACGTGAT	ATCACGTT	CAAGCAGAAGACGGCATAACGTGATGAGATCGGTCTCGGCATTC
6	CGCTGATC	GATCAGCG	CAAGCAGAAGACGGCATAACGCTGATCGGATCGGTCTCGGCATTC
7	CAGATCTG	CAGATCTG	CAAGCAGAAGACGGCATAACGATCTGGAGATCGGTCTCGGCATTC
8	ATGCCTAA	TTAGGCAT	CAAGCAGAAGACGGCATAATGCCTAAGAGATCGGTCTCGGCATTC
9	CTGTAGCC	GGCTACAG	CAAGCAGAAGACGGCATACTGTAGCCGAGATCGGTCTCGGCATTC
10	AGTACAAG	CTTGTAAT	CAAGCAGAAGACGGCATAAGTACAAGGAGATCGGTCTCGGCATTC
11	CATCAAGT	ACTTGATG	CAAGCAGAAGACGGCATAACATCAAGTGAAGATCGGTCTCGGCATTC
12	AGTGGTCA	TGACCACT	CAAGCAGAAGACGGCATAAGTGGTCAGAGATCGGTCTCGGCATTC
13	AACAACCA	TGGTTGTT	CAAGCAGAAGACGGCATAACAACCAAGAGATCGGTCTCGGCATTC
14	AACCGAGA	TCTCGGTT	CAAGCAGAAGACGGCATAACCGAGAGAGATCGGTCTCGGCATTC
15	AACGCTTA	TAAGCGTT	CAAGCAGAAGACGGCATAAACGCTTAGAGATCGGTCTCGGCATTC
16	AAGACGGA	TCCGTCTT	CAAGCAGAAGACGGCATAAAGACGGAGAGATCGGTCTCGGCATTC
17	AAGGTACA	TGTACCTT	CAAGCAGAAGACGGCATAAAGGTACAGAGATCGGTCTCGGCATTC
18	ACACAGAA	TTCTGTGT	CAAGCAGAAGACGGCATAACACAGAAGAGATCGGTCTCGGCATTC
19	ACAGCAGA	TCTGCTGT	CAAGCAGAAGACGGCATAACAGCAGAGAGATCGGTCTCGGCATTC
20	ACCTCCAA	TTGGAGGT	CAAGCAGAAGACGGCATAACCTCCAAGAGATCGGTCTCGGCATTC
21	ACGCTCGA	TCGAGCGT	CAAGCAGAAGACGGCATAACGCTCGAGAGATCGGTCTCGGCATTC
22	ACGTATCA	TGATACGT	CAAGCAGAAGACGGCATAACGTATCAGAGATCGGTCTCGGCATTC
23	ACTATGCA	TGCATAGT	CAAGCAGAAGACGGCATAACTATGCAGAGATCGGTCTCGGCATTC
24	AGAGTCAA	TTGACTCT	CAAGCAGAAGACGGCATAAGAGTCAAGAGATCGGTCTCGGCATTC
25	AGATCGCA	TGCGATCT	CAAGCAGAAGACGGCATAAGATCGCAGAGATCGGTCTCGGCATTC

Tag number	Barcode sequence	Single correcting, double & shift detecting octamers	PCR primers
26	AGCAGGAA	TTCCCTGCT	CAAGCAGAAGACGGCATAACGAGATAGCAGGAAGAGATCGGTCTCGGCATTC
27	AGTCACTA	TAGTGACT	CAAGCAGAAGACGGCATAACGAGATAGTCACTAGAGATCGGTCTCGGCATTC
28	ATCCTGTA	TACAGGAT	CAAGCAGAAGACGGCATAACGAGATATCCTGTAGAGATCGGTCTCGGCATTC
29	ATTGAGGA	TCCTCAAT	CAAGCAGAAGACGGCATAACGAGATATTGAGGAGAGATCGGTCTCGGCATTC
30	CAACCACA	TGTGGTTG	CAAGCAGAAGACGGCATAACGAGATCAACCACAGAGATCGGTCTCGGCATTC
31	CAAGACTA	TAGTCTTG	CAAGCAGAAGACGGCATAACGAGATCAAGACTAGAGATCGGTCTCGGCATTC
32	CAATGGAA	TTCCATTG	CAAGCAGAAGACGGCATAACGAGATCAATGGAAGAGATCGGTCTCGGCATTC
33	CACTTCGA	TCGAAGTG	CAAGCAGAAGACGGCATAACGAGATCACTTCGAGAGATCGGTCTCGGCATTC
34	CAGCGTTA	TAACGCTG	CAAGCAGAAGACGGCATAACGAGATCAGCGTTAGAGATCGGTCTCGGCATTC
35	CATACCAA	TTGGTATG	CAAGCAGAAGACGGCATAACGAGATCATAACGAAGAGATCGGTCTCGGCATTC
36	CCAGTTCA	TGAACTGG	CAAGCAGAAGACGGCATAACGAGATCCAGTTCAGAGATCGGTCTCGGCATTC
37	CCGAAGTA	TACTTCGG	CAAGCAGAAGACGGCATAACGAGATCCGAAGTAGAGATCGGTCTCGGCATTC
38	CCGTGAGA	TCTCACGG	CAAGCAGAAGACGGCATAACGAGATCCGTGAGAGAGATCGGTCTCGGCATTC
39	CCTCCTGA	TCAGGAGG	CAAGCAGAAGACGGCATAACGAGATCCTCCTGAGAGATCGGTCTCGGCATTC
40	CGAACTTA	TAAGTTCG	CAAGCAGAAGACGGCATAACGAGATCGAACTTAGAGATCGGTCTCGGCATTC
41	CGACTGGA	TCCAGTCG	CAAGCAGAAGACGGCATAACGAGATCGACTGGAGAGATCGGTCTCGGCATTC
42	CGCATACA	TGTATGCG	CAAGCAGAAGACGGCATAACGAGATCGCATACAGAGATCGGTCTCGGCATTC
43	CTCAATGA	TCATTGAG	CAAGCAGAAGACGGCATAACGAGATCTCAATGAGAGATCGGTCTCGGCATTC
44	CTGAGCCA	TGGCTCAG	CAAGCAGAAGACGGCATAACGAGATCTGAGCCAGAGATCGGTCTCGGCATTC
45	CTGGCATA	TATGCCAG	CAAGCAGAAGACGGCATAACGAGATCTGGCATAGAGATCGGTCTCGGCATTC
46	GAATCTGA	TCAGATTC	CAAGCAGAAGACGGCATAACGAGATGAATCTGAGAGATCGGTCTCGGCATTC
47	GACTAGTA	TACTAGTC	CAAGCAGAAGACGGCATAACGAGATGACTAGTAGAGATCGGTCTCGGCATTC
48	GAGCTGAA	TTCAGCTC	CAAGCAGAAGACGGCATAACGAGATGAGCTGAAGAGATCGGTCTCGGCATTC
49	GATAGACA	TGTCTATC	CAAGCAGAAGACGGCATAACGAGATGATAGACAGAGATCGGTCTCGGCATTC
50	GCCACATA	TATGTGGC	CAAGCAGAAGACGGCATAACGAGATGCCACATAGAGATCGGTCTCGGCATTC
51	GCGAGTAA	TTACTCGC	CAAGCAGAAGACGGCATAACGAGATGCGAGTAAGAGATCGGTCTCGGCATTC
52	GCTAACGA	TCGTTAGC	CAAGCAGAAGACGGCATAACGAGATGCTAACGAGAGATCGGTCTCGGCATTC
53	GCTCGGTA	TACCGAGC	CAAGCAGAAGACGGCATAACGAGATGCTCGGTAGAGATCGGTCTCGGCATTC
54	GGAGAACA	TGTTCTCC	CAAGCAGAAGACGGCATAACGAGATGGAGAACAGAGATCGGTCTCGGCATTC
55	GGTGCGAA	TTCGCACC	CAAGCAGAAGACGGCATAACGAGATGGTGCGAAGAGATCGGTCTCGGCATTC

Tag number	Barcode sequence	Single correcting, double & shift detecting octamers	PCR primers
56	GTACGCAA	TTGCGTAC	CAAGCAGAAGACGGCATAACGAGATGTACGCAAGAGATCGGTCTCGGCATTC
57	GTCGTAGA	TCTACGAC	CAAGCAGAAGACGGCATAACGAGATGTCGTAGAGAGATCGGTCTCGGCATTC
58	GTCTGTCA	TGACAGAC	CAAGCAGAAGACGGCATAACGAGATGTCGTAGAGAGATCGGTCTCGGCATTC
59	GTGTTCTA	TAGAACAC	CAAGCAGAAGACGGCATAACGAGATGTGTTCTAGAGATCGGTCTCGGCATTC
60	TAGGATGA	TCATCCTA	CAAGCAGAAGACGGCATAACGAGATTAGGATGAGAGATCGGTCTCGGCATTC
61	TATCAGCA	TGCTGATA	CAAGCAGAAGACGGCATAACGAGATTATCAGCAGAGATCGGTCTCGGCATTC
62	TCCGTCTA	TAGACGGA	CAAGCAGAAGACGGCATAACGAGATTCCGTCTAGAGATCGGTCTCGGCATTC
63	TCTTCACA	TGTGAAGA	CAAGCAGAAGACGGCATAACGAGATTCTTCACAGAGATCGGTCTCGGCATTC
64	TGAAGAGA	TCTCTTCA	CAAGCAGAAGACGGCATAACGAGATTGAAGAGAGAGATCGGTCTCGGCATTC
65	TGGAACAA	TTGTTCCA	CAAGCAGAAGACGGCATAACGAGATTGGAACAAGAGATCGGTCTCGGCATTC
66	TGGCTTCA	TGAAGCCA	CAAGCAGAAGACGGCATAACGAGATTGGCTTCAGAGATCGGTCTCGGCATTC
67	TGGTGGTA	TACCACCA	CAAGCAGAAGACGGCATAACGAGATTGGTGGTAGAGATCGGTCTCGGCATTC
68	TTCACGCA	TGCGTGAA	CAAGCAGAAGACGGCATAACGAGATTTCAAGCAGAGATCGGTCTCGGCATTC
69	AACTCACC	GGTGAGTT	CAAGCAGAAGACGGCATAACGAGATAACTCACCAGAGATCGGTCTCGGCATTC
70	AAGAGATC	GATCTCTT	CAAGCAGAAGACGGCATAACGAGATAAAGAGATCGAGATCGGTCTCGGCATTC
71	AAGGACAC	GTGTCCTT	CAAGCAGAAGACGGCATAACGAGATAAAGGACACGAGATCGGTCTCGGCATTC
72	AATCCGTC	GACGGATT	CAAGCAGAAGACGGCATAACGAGATAATCCGTTCGAGATCGGTCTCGGCATTC
73	AATGTTGC	GCAACATT	CAAGCAGAAGACGGCATAACGAGATAATGTTGCGAGATCGGTCTCGGCATTC
74	ACACGACC	GGTCGTGT	CAAGCAGAAGACGGCATAACGAGATACACGACCGAGATCGGTCTCGGCATTC
75	ACAGATTC	GAATCTGT	CAAGCAGAAGACGGCATAACGAGATACAGATTCGAGATCGGTCTCGGCATTC
76	AGATGTAC	GTACATCT	CAAGCAGAAGACGGCATAACGAGATAGATGTACGAGATCGGTCTCGGCATTC
77	AGCACCTC	GAGGTGCT	CAAGCAGAAGACGGCATAACGAGATAGCACCTCGAGATCGGTCTCGGCATTC
78	AGCCATGC	GCATGGCT	CAAGCAGAAGACGGCATAACGAGATAGCCATGCGAGATCGGTCTCGGCATTC
79	AGGCTAAC	GTTAGCCT	CAAGCAGAAGACGGCATAACGAGATAGGCTAACGAGATCGGTCTCGGCATTC
80	ATAGCGAC	GTCGCTAT	CAAGCAGAAGACGGCATAACGAGATATAGCGACGAGATCGGTCTCGGCATTC
81	ATCATTCC	GGAATGAT	CAAGCAGAAGACGGCATAACGAGATATCATTCCGAGATCGGTCTCGGCATTC
82	ATTGGCTC	GAGCCAAT	CAAGCAGAAGACGGCATAACGAGATATTGGCTCGAGATCGGTCTCGGCATTC
83	CAAGGAGC	GCTCCTTG	CAAGCAGAAGACGGCATAACGAGATCAAGGAGCGAGATCGGTCTCGGCATTC
84	CACCTTAC	GTAAGGTG	CAAGCAGAAGACGGCATAACGAGATCACCTTACGAGATCGGTCTCGGCATTC
85	CCATCCTC	GAGGATGG	CAAGCAGAAGACGGCATAACGAGATCCATCCTCGAGATCGGTCTCGGCATTC

Tag number	Barcode sequence	Single correcting, double & shift detecting octamers	PCR primers
86	CCGACAAC	GTTGTCCG	CAAGCAGAAGACGGCATAACGAGATCCGACAACGAGATCGGTCTCGGCATTC
87	CCTAATCC	GGATTAGG	CAAGCAGAAGACGGCATAACGAGATCCTAATCCGAGATCGGTCTCGGCATTC
88	CCTCTATC	GATAGAGG	CAAGCAGAAGACGGCATAACGAGATCCTCTATCGAGATCGGTCTCGGCATTC
89	CGACACAC	GTGTGTCG	CAAGCAGAAGACGGCATAACGAGATCGACACACGAGATCGGTCTCGGCATTC
90	CGGATTGC	GCAATCCG	CAAGCAGAAGACGGCATAACGAGATCGGATTGCGAGATCGGTCTCGGCATTC
91	CTAAGGTC	GACCTTAG	CAAGCAGAAGACGGCATAACGAGATCTAAGGTCGAGATCGGTCTCGGCATTC
92	GAACAGGC	GCCTGTTC	CAAGCAGAAGACGGCATAACGAGATGAACAGGCAGATCGGTCTCGGCATTC
93	GACAGTGC	GCACTGTC	CAAGCAGAAGACGGCATAACGAGATGACAGTGCAGATCGGTCTCGGCATTC
94	GAGTTAGC	GCTAACTC	CAAGCAGAAGACGGCATAACGAGATGAGTTAGCGAGATCGGTCTCGGCATTC
95	GATGAATC	GATTCATC	CAAGCAGAAGACGGCATAACGAGATGATGAATCGAGATCGGTCTCGGCATTC
96	GCCAAGAC	GTCTTGCC	CAAGCAGAAGACGGCATAACGAGATGCCAAGACGAGATCGGTCTCGGCATTC

**Supplementary Protocol 1****PCR and 96-well library prep standard operating procedure****Adapter preparation**

Custom adapters are required. These should be HPLC purified. Adapters are phosphorylated and annealed together, as described<sup>1</sup>.

```
Ind_t 5'   ACACTCTTCCCTACACGACGCTCTTCCGATC*T   3'
Ind_b 5'   GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTC 3'
```

\* indicates phosphorothioate

**Sample shearing**

1. Take 1 µg of pooled long PCR products and make volume up to 75 µl with water.
2. Shear to approximately 200 bp by acoustic shearing: transfer to a 6 mm × 16 mm AFA fiber vial (Covaris cat. no. 520031).
3. Seal the tube with an 8 mm crimp seal cap (Covaris cat no. 520028) and crimping tool.
4. Shear with a Covaris, using the settings:

Duty cycle	20 %
Intensity	5
Cycle/burst	200
Time	150 sec

5. Transfer samples to 96-well PCR plate, gently spin and keep frozen until used.

**End-repair**

*This protocol converts the overhangs resulting from fragmentation into blunt ends, using T4 DNA polymerase and E. coli DNA polymerase I Klenow fragment. The 3' to 5' exonuclease activity of these enzymes removes 3' overhangs and the polymerase activity fills in the 5' overhangs.*

1. Mix the reagents below (all NEB) in a 15ml Falcon Tube, and decant into a reagent reservoir.

	1X	104X
10x T4 DNA ligase buffer with 10 mM ATP	10 $\mu$ l	1040 $\mu$ l
10 $\mu$ M dNTP mix	4 $\mu$ l	416 $\mu$ l
3 U / $\mu$ l T4 DNA polymerase	5 $\mu$ l	520 $\mu$ l
5 U / $\mu$ l Klenow DNA polymerase	1 $\mu$ l	104 $\mu$ l
10 U / $\mu$ l T4 PNK	5 $\mu$ l	520 $\mu$ l

2. Add 25  $\mu$ l of master mix to each sample using an electronic pipette. Cover the plate with transparent cover, vortex briefly and gently spin down.
3. Incubate plate for 30 min at 20 °C in a thermocycler.
4. While incubating, prepare SPRI beads for the reaction cleanup: allow SPRI beads to come to room temperature for at least 30 minutes. Mix well, and ensure that the beads appear homogeneous and consistent in colour.
5. Add 180  $\mu$ l of SPRI beads per 100  $\mu$ l of end-repaired DNA in a 1.5 ml Lo-Bind tube.
6. Vortex and leave at room temperature for 5 minutes.
7. Place tubes in a magnetic rack.
8. Leave for 5 minutes or until sample is clear.
9. Carefully remove the clear solution from the tubes and discard.
10. Dispense 700  $\mu$ l of 70 % ethanol into each tube while in the magnetic rack taking care not to disturb the magnetic beads. Aspirate and discard ethanol.
11. Repeat the ethanol wash once again (total of two washes).
12. Dry the samples on a heat block (keep the lid of the tube open) at 37 °C for 5 to 10 minutes or until the residual ethanol has evaporated.
13. Add 32  $\mu$ l of molecular biology grade water, vortex and incubate at room temperature for 2 minutes.
14. Place tubes into the magnetic rack and leave for 2-3 minutes or until sample is clear.
15. Carefully remove the water and retain in a new 1.5 ml Lo-Bind tube.
16. Centrifuge the eluates at 13,000 rpm in a bench top centrifuge for 10 minutes.
17. Transfer eluates to a 96-well plate leaving behind any precipitated beads.

18. Proceed immediately with A-tailing.

### A-tailing

*This protocol adds an 'A' base to the 3' end of the blunt phosphorylated DNA fragments, using the polymerase activity of Klenow fragment (3' to 5' exo minus). This prepares the DNA fragments for ligation to the adapters, which have a single 'T' base overhang at their 3' end.*

1. Mix the reagents below (all NEB) in a 15ml Falcon Tube, and decant into a reagent reservoir.

	1X	104X
10x NEB buffer 2	5 $\mu$ l	520 $\mu$ l
1 mM dATP	10 $\mu$ l	1040 $\mu$ l
5 U / $\mu$ l Klenow fragment (3' to 5' exo <sup>-</sup> )	3 $\mu$ l	312 $\mu$ l

2. Add 18 $\mu$ l to each sample using an electronic pipette. Cover the plate with transparent cover, vortex and gently spin.
3. Incubate plate for 30 min at 37 °C in a thermocycler. Clean using SPRI beads, in the 96-well reaction plate (see End-Repair protocol above. Use 90  $\mu$ l beads for each 50 $\mu$ l reaction, and elute in a mixture of 10  $\mu$ l EB + 8.5  $\mu$ l water).
4. Proceed immediately with ligation.

### Ligation

*This protocol ligates adapters to the ends of the DNA fragments.*

1. Mix the reagents below in a 15ml Falcon Tube, and decant into a reagent reservoir.

	1X	104X
2x Quick DNA ligase buffer (NEB)	25 $\mu$ l	2.6 ml
Adapter oligo mix (see above)	1.5 $\mu$ l	156 $\mu$ l

2. Add 26.5  $\mu$ l to each sample using an electronic pipette. Cover the plate with transparent cover, vortex and gently spin.
3. Pipette 5  $\mu$ l Quick Ligase (NEB) into each well using a manual pipette. Mix by pipetting, do not cover, and after pipetting in the last column leave for 15 min at room temperature (20 °C).

4. Clean using AMPure SPRI beads as described above, eluting in 20  $\mu$ l EB.
5. Run an DNA 1000 chip on an Agilent BioAnalyzer 2100 on a small selection of samples. You should detect smear between 300-1000 bp.

### Indexing Enrichment PCR

Indexes can be added to the central region of the reverse PCR primer. The general sequence is:

5' CAAGCAGAAGACGGCATAACGAGAT-INDEX-GAGATCGGTCTCGGCATTC 3'

where INDEX represents an oligonucleotide sequence that is used to identify the sample. Indexes should be selected so that they are maximally different from one another. We typically use 8-base indexes. Indexed primers must be PAGE purified. For an example of a 96-plex set of error-correcting barcodes, and primer sequences, see **Supplementary Table 1**.

The common forward primer (HPLC purified) has the sequence:

5' AATGATACGGCGACCACCGAGATCTACACTCTTCCCTACACGACGCTCTTCCGATCT 3'

1. Dilute indexed primers to 10  $\mu$ M with water.
2. Pipette 2  $\mu$ l of each library into a well of a 96-well plate.
3. Add 6  $\mu$ l of 10  $\mu$ M indexed primer to each well.
4. Prepare a master mix of all other reagents:

	1X	52X
10X Pfx buffer (Invitrogen)	10 $\mu$ l	520 $\mu$ l
2.5 mM dNTPs	20 $\mu$ l	1040 $\mu$ l
50 mM MgSO <sub>4</sub>	4 $\mu$ l	208 $\mu$ l
10 $\mu$ M Forward primer	6 $\mu$ l	312 $\mu$ l
Platinum Pfx Polymerase (Invitrogen)	1 $\mu$ l	52 $\mu$ l
Water	51 $\mu$ l	2652 $\mu$ l

5. Mix, and add 92  $\mu$ l master mix to each well. The PCR is performed using the following program :

94 °C for 2 min

94 °C for 15 sec

68 °C for 45 sec x 12 cycles

4°C indefinitely

6. Clean PCR reactions SPRI beads, as described above, eluting in 18 µl EB.
7. Run an DNA 1000 chip on an Agilent BioAnalyzer 2100 on a small selection of samples. You should detect smear between 300-1000 bp.

### Normalization

*Normalize PCR products by qPCR before pooling. The primers used for the qPCR are locus-specific, designed to amplify one of the long PCR products in the original pool. In this way, adapter dimers are not problematic. As a concentration standard, the most concentrated library from the 6 random libraries quantified by Agilent Bioanalyzer 2100 in the preceding step, is used.*

1. Dilute 2µl of the chosen concentration standard 10x, 100x and 1,000x with EB buffer.
2. Dilute all other libraries 80x in EB.
3. Perform qPCRs in duplicate, using a 2x SybrGreen master mix.
4. Pool products in equimolar ratios.

### Size selection

1. Load sample pools on a 2% agarose gel in 1X TBE, and are electrophoresed against a Low Molecular Weight ladder (NEB), at 5V cm<sup>-1</sup> for approximately 1 hour.
2. From each lane, cut a 250-450 bp gel slice is cut and extract the DNA.
3. Quantify pooled libraries by SYBRGreen qPCR<sup>2</sup>.
4. Sequence using standard primers for reads 1 and 2. Indexes are sequenced after read 1, using a custom primer.

Ind\_seq 5' AAGAGCGGTTTCAGCAGGAATGCCGAGACCGATCTC 3'

**References**

1. Kozarewa I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* **6**, 291-295 (2009).
2. Quail M.A., Swerdlow H. & Turner D.J. Improved protocols for the illumina genome analyzer sequencing system. *Curr Protoc Hum Genet* **Chapter 18**, Unit 18 12 (2009).

**Supplementary Protocol 2****Molecular Inversion Probe standard operating procedure****Generate Probes**

*Amplify off-array oligonucleotides (MIP precursors) using PCR: (2.5 hrs)*

1. Dissolve array-derived MIP precursor oligonucleotides (mixture of 100-mers obtained from Agilent) to a final concentration of 100 nM in Tris-EDTA buffer with a pH of 8 and 0.1 % Tween.
2. Prepare the following 400  $\mu$ l PCR mix in a 1.5 ml centrifuge tube. Mix and spin down:

Reagent	Volume ( $\mu$ l)	Final Concentration
2x iProof HF PCR master mix (Biorad)	200	1x
Oligo_Fwd_Amp Primer (100 $\mu$ M)	2	500 nM
Oligo_Rev_Amp Primer (100 $\mu$ M)	2	500 nM
SYBRGreen I 100 x (Invitrogen)**	1	0.2X
Template (100 nM in 0.1 % Tween)	1	250 pM
Water	194	

Split into 8 x 50  $\mu$ l reactions in 0.2 ml PCR tubes. One PCR preparation can be expected to yield around 1.5  $\mu$ g of amplified DNA. \*\*Use SYBR Green when using a real-time thermocycling instrument.

3. Use the following PCR cycling program, ideally on a real-time thermocycling instrument such as the Biorad MJ Mini.

98 °C for 30 seconds

98 °C for 10 seconds

60 °C for 30 seconds

x 25 cycles

72 °C for 30 seconds (read plate)

4 °C indefinitely

\*\*We typically stop our reactions after 20 rounds of PCR. When using RT-PCR, stop the reaction slightly before the fluorescence curve plateaus to avoid over-amplification.

4. Combine and clean up PCR reactions on one column using the QIAquick PCR purification kit following the manufacturer's instructions. Elute with 90  $\mu$ l elution buffer.
5. Use a Qubit High Sensitivity dsDNA Assay Kit to quantify 1  $\mu$ l of the amplified DNA.
6. Analyze 1  $\mu$ l amplified DNA on a 6 % TBE PAGE gel (Invitrogen) to verify amplification. Product should appear as a single band at 110 bp, as the primers add an additional 10 bp.

*Digest PCR product with nicking restriction endonucleases to generate 70-mer MIPs (7.5 hrs):*

7. Add 10  $\mu$ l of NEB-2 (10x) and 5  $\mu$ l of Nt.AlwI (10 U /  $\mu$ l; NEB) to 85  $\mu$ l of PCR product (total volume of 100  $\mu$ l)
8. Mix and split to two tubes of 50  $\mu$ l each. Incubate at 37 °C for 3 hours, followed by 80 °C for 20 minutes in a thermocycler
9. Let the temperature drop to 65 °C for at least 1 minute. Add 2.5  $\mu$ l of Nb.BsrDI (2 U /  $\mu$ l; NEB) to each of the 50  $\mu$ l reactions
10. Leave at 65 °C for 3 hours, followed by 80°C for 20 minutes
11. Purify two 50  $\mu$ l digestion reactions on one column using reagents from the QIAquick Nucleotide Removal Kit. Elute each column in 30  $\mu$ l elution buffer. We have observed yields of 80-90 % for this step.

*Quantify usable probe using a denaturing gel (2 hrs):*

12. Accurate quantification of usable MIP inside the digested probe mix is important as it determines how much probe mix to add to the capture reaction.
13. Prepare two-fold dilutions of a NEB 100 bp DNA ladder (we used dilutions from 500 ng to 62 ng).
14. Mix 2x TBE-Urea sample buffer (Invitrogen) with 1  $\mu$ l digested probe and the dilutions made above.
15. Denature DNA by heating to 95 °C for 5 minutes and immediately transferring to ice.
16. Run samples on a precast 6 % TBE-urea denaturing PAGE gel (Invitrogen) for 1 hr at 160 V.

- Quantify the amount of usable MIP in the digested mixture by comparing the intensity of ladder dilutions with the intensity of the 70 bp band. Use this MIP concentration when determining the volume of probe mix to add to a capture reaction.

### Capture Reaction

Note: We have found that it is no longer necessary to gel-purify the single stranded 70 bp MIP from the digested probe mix in a capture reaction. Instead, we use a blocking oligo to limit hybridization of the undigested strand of the MIP precursor (still a 100-mer) to the active 70-mer MIP in the capture reaction, which could potentially interfere with MIP hybridization to genomic DNA targets.

*Hybridize probes to genomic DNA (37 hrs):*

- For each sample to capture, add the following reagents in a 0.2 ml PCR tube. The final capture reaction volume is 25  $\mu$ l. Because there is no size selection of the 70 bp MIP, the volume of probe mix to add is based on the concentration of usable MIP.

Reagent	Volume ( $\mu$ l) per sample	Final Concentration in reaction
750 ng genomic DNA*	3	30 ng / $\mu$ l*
10 x Ampligase buffer (Epicentre)	2.5	1x
40 ng (2 pmol) of MIP**	3	1.6 ng / $\mu$ l**
Blocking Oligonucleotide (100 $\mu$ M)	0.1	0.4 $\mu$ M
Water	16.4	

\* Additionally, prepare a blank capture reaction containing MIP probe but no gDNA to detect cross contamination.

\*\* For a reaction targeting 55,000 regions in the genome. We currently aim for a ratio of MIPs to genomic DNA of 100:1 (i.e. 100 copies of each MIP in the mix for each genomic equivalent). The concentration of MIPs can be adjusted accordingly depending on complexity of the targeting reaction.

- Denature at 95 °C for 10 minutes.
- Incubate at 60 °C for at least 36 hours to hybridize MIPs to gDNA.

*Circularize captured exons: (1 day)*

- Prepare a mix of ligase and polymerase enzymes to add to each capture reaction:

Reagent	Volume ( $\mu$ l)	Final

	per sample	Concentration in capture reaction
10 x Ampligase buffer (Epicentre)	0.45	1x
10 U / $\mu$ l Stoffel** (Applied Biosystems)	2	0.8 U / $\mu$ l
100 U / $\mu$ l Ampligase** (Epicentre)	1	4 U / $\mu$ l
0.25 mM dNTP**	1.25	12 $\mu$ M

Prepare this mix on ice, and keep cold before adding 4.7  $\mu$ l into the capture reaction.

- Incubate at 60 °C for an additional 24 hours to allow for gap-fill and ligation to circularize captured regions.

*Exonuclease select for circularized product: (1hr)*

- Prepare a mix of exonucleases to add to each capture reaction in order to remove uncaptured gDNA, excess probe and blocking oligonucleotide:

Reagent	Volume ( $\mu$ l) per sample	Final Concentration in reaction
Exo I 20 U / $\mu$ l	2	1.7 U / $\mu$ l
Exo III 100 U / $\mu$ l	2	8.3 U / $\mu$ l

- Reduce the temperature of the capture reaction to 37 °C and allow it to incubate for at least one minute before adding 4  $\mu$ l of exonuclease mix.
- Incubate for 15 minutes at 37 °C.
- Inactivate exonuclease enzymes by heating reaction at 95 °C for 2 minutes.
- Use 5  $\mu$ l of the reaction product as the template for PCR. There is no need to purify the reaction product before PCR.

#### Amplify and Verify Captured Product:

- Prepare the following PCR mix in a 1.5 ml centrifuge tube. Mix and spin down.

Reagent	Volume ( $\mu$ l) per sample	Final Concentration in capture reaction
iProof PCR master mix (2x)	25	1x
SLXA_Paired_End_CP2_Fwd (100 $\mu$ M)	0.25	500 nM
SLXA_Paired_End_CP2_Rev (100 $\mu$ M)	0.25	500 nM

SYBR Green I 100X (Invitrogen)	0.25	0.5X
Water	19.25	

- Add 45  $\mu$ l of master mix to 5  $\mu$ l of each sample to obtain a total reaction volume of 50  $\mu$ l. Mix gently and spin down.
- Amplify on a RT-PCR machine using the following conditions:

98 °C for 30 seconds

98 °C for 10 seconds

60 °C for 30 seconds x 25 cycles

72 °C for 60 seconds (read plate)

\*\*When using RT-PCR, stop the reaction slightly before the fluorescence curve plateaus to avoid overamplification.

- Purify each sample reaction on one column QIAquick PCR Purification column following the manufacturer's instructions. Elute each column in 30  $\mu$ l EB buffer.
- Use a Qubit HS dsDNA Assay Kit to quantify 1  $\mu$ l of the amplified DNA.
- Analyze 2  $\mu$ l of the amplified DNA on a PAGE gel as described in step 6 to validate that the amplified product is of the expected size range. Currently, we aim for a uniform gap-fill size of 112 bp for all targets. Assuming 20 bp targeting arms, paired-end 76 bp reads enable full coverage of this gap-fill size. At this step, we consequently expect a tight band centered at 245 bp (49 bp primer (SLXA\_Paired\_End\_CP2\_Fwd) + 20 bp targeting arm + 112 bp gap-fill + 20 bp targeting arm + 44 bp primer (SLXA\_Paired\_End\_CP2\_Rev) = 245 bp).
- Samples are now ready for analysis using the Illumina Genome Analyzer. Use PE\_Capture\_Sequencing and PE\_Rev\_Capture\_Sequencing primers for sequencing

### Timeline: (4 days)

#### Generate Probes (Steps 1-17): 12 hrs

Amplify off-array oligonucleotides using PCR: (2.5 hrs)

Digest oligonucleotides: (7.5 hrs)

Quantify usable probe using a denaturing gel: (2 hrs)

#### Capture Exons (Steps 18-28): 3 days

Hybridize probes to genomic DNA: (1.5 days)  
Circularize captured exons: (1 day)  
Exonuclease select for circularized product: (1 hr)

**Amplify and Verify Captured Product (Steps 29-35): 4 hrs**

**Oligonucleotide Sequences:**

Oligonucleotide Name	Sequence (5'→3')
General Format of MIP precursors (100-mers); x's and y's indicate variable targeting arm sequence	AGGACCGGATCAACTxxxxxxxxxxxxxxxxxxxxxCTTCAGCTTCCCGATA TCCGACGGTAGTGTyyyyyyyyyyyyyyyyyyCATTGCGTGAACCGA
Oligo_Fwd_Amp	TGCCTAGGACCGGATCAACT
Oligo_Rev_Amp	GAGCTTCGGTTCACGCAATG
SLXA_Paired_End_CP2_Fwd	AATGATACGGCGACCACCGAGATCTACACGCACGATCCGACGGTA GTGT
SLXA_Paired_End_CP2_Rev	CAAGCAGAAGACGGCATAACGAGATCCGTAATCGGGAAGCTGAAG
PE_Capture_Sequencing Primer	ACACGCACGATCCGACGGTAGTGT
PE_Rev_Capture_Sequencing Primer	CATACGAGATCCGTAATCGGGAAGCTGAAG
Blocking Oligonucleotide	CTTCAGCTTCCCGATATCCGACGGTAGTGT

**Supplementary Protocol 3****Hybrid capture library prep standard operating procedure (array OR solution capture)****Shearing samples to 100-300bp**

For array capture, dilute 20 µg genomic DNA to a total volume of 100 µl with water.

For solution capture, dilute 3 µg genomic DNA to a total volume of 100 µl with water.

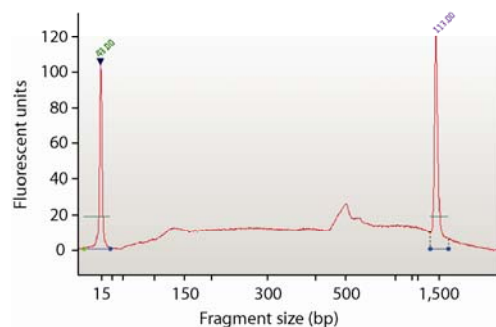
1. Mix and transfer to a 6mmx16mm AFA fibre vial (Covaris cat. no. 520031).
2. Seal the tube using an 8mm metal crimp seal cap (Covaris, cat no. 520028) and crimping tool.
3. Shear with a Covaris, using the settings:

Duty cycle	20 %
Intensity	5
Cycle/burst	200
Time	120 sec

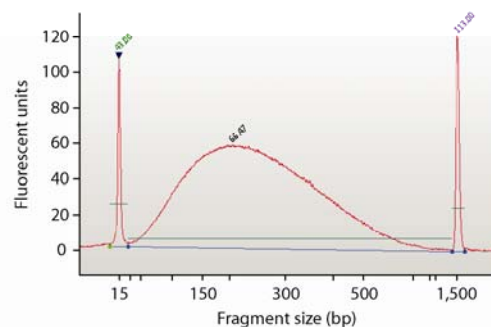
4. Remove the sample from the machine. Open the vial and transfer the sample into a fresh 1.5 ml Eppendorf lo-bind tube. Keep samples on ice.
5. Run 1 µl on an Agilent Bioanalyzer 2100 chip to check the quantity of fragmented DNA and to confirm the success of the fragmentation.

Impure DNA may shear badly. If there is any doubt about the purity of the sample, perform an ethanol precipitation before shearing (Figure 1).

a) without ethanol precipitation



b) with ethanol precipitation



**Figure 1:** Impure DNA sample sheared using identical Covaris settings a) before and b) after ethanol precipitation.

### Purification after Fragmentation

*Use one column for each 10 µg of DNA. For 10-20 µg DNA, divide equally between two columns. Elute with 45 µl EB per 5 µg bound DNA.*

1. Add 5 x volume of Buffer PB to the fragmented DNA (for 100 µl of fragmented DNA add 500 µl buffer). Vortex.
2. Pipette 750 µl of mix per column. Pipette slowly and make sure that all the liquid comes out of the pipette tip. Centrifuge the samples for 1 minute at 13,000 rpm in a benchtop centrifuge.
3. Discard the flow-through from the collection tube. Repeat step 2 if you have any buffer mix left.
4. To wash, add 750 µl of Buffer PE to each of the columns. Centrifuge for 1 minute as above. Discard the flow-through and centrifuge again for a further minute.
5. Leave the tubes in a rack with the lids open to dry for 2 minutes.
6. Transfer the column to a clean labelled 2 ml lo-bind Eppendorf tube. For each 5 µg of bound DNA, add 47 µl of EB buffer to the centre of the column and leave for 1 minute. Centrifuge for 1 minute as above.
7. Combine eluates, if applicable. You should have a volume of 45 µl for every 5 µg in the original sample (approximately 2 µl of buffer is retained by the column).

### End Repair

*This protocol converts the overhangs resulting from fragmentation into blunt ends, using T4 DNA polymerase and E. coli DNA polymerase I Klenow fragment. The 3' to 5' exonuclease activity of these enzymes removes 3' overhangs and the polymerase activity fills in the 5' overhangs.*

1. Prepare a master mix containing the following reaction mix per 5 µg sample, plus a 10 % excess (round the mass of sample UP to the nearest 5 µg):

Water	30 µl
-------	-------

10x T4 DNA ligase buffer with 10mM ATP	10 $\mu$ l
10mM dNTP mix	4 $\mu$ l
3U/ $\mu$ l T4 DNA polymerase	5 $\mu$ l
5U/ $\mu$ l Klenow DNA polymerase	1 $\mu$ l
10U/ $\mu$ l T4 PNK	5 $\mu$ l

- Mix, and aliquot 55  $\mu$ l of master mix into each sample tube containing the 45  $\mu$ l of eluate from the previous step. Mix well and spin down.
- Incubate for 30 minutes at room temperature (20-25  $^{\circ}$ C).
- Clean up using one QIAquick PCR column for up to 10  $\mu$ g of DNA, as described above. Elute in 34  $\mu$ l EB per 5  $\mu$ g DNA. This gives  $\sim$ 32  $\mu$ l of eluate, because approximately 2  $\mu$ l of buffer is retained by the column.

#### Addition of 'A' Bases to the 3' End of the DNA Fragments

*This protocol adds an 'A' base to the 3' end of the blunt phosphorylated DNA fragments, using the polymerase activity of Klenow fragment (3' to 5' exo minus). This prepares the DNA fragments for ligation to the adapters, which have a single 'T' base overhang at their 3' end.*

- Prepare a master mix containing the following reaction mix per 5  $\mu$ g sample, plus a 10 % excess (round the mass of sample UP to the nearest 5  $\mu$ g):

10x Klenow buffer	5 $\mu$ l
1 mM dATP	10 $\mu$ l

- Mix, and aliquot 15  $\mu$ l of master mix into each sample tube containing the 32  $\mu$ l of end-repaired sample. Mix well and spin down.
- Add 3  $\mu$ l 5 U/  $\mu$ l Klenow exo (3' to 5' exo minus). Mix and spin down.
- Incubate for 30 minutes at 37  $^{\circ}$ C in a hot block.
- Clean up using one QIAquick MinElute columns per 5  $\mu$ g of DNA, eluting in 12  $\mu$ l of EB buffer per column, in a 1.5 ml lo-bind Eppendorf tube. This gives  $\sim$ 10  $\mu$ l of eluate, because approximately 2  $\mu$ l of buffer is retained by the column.

**Ligation of Adapters to DNA Fragments**

*This protocol ligates adapters to the ends of the DNA fragments. The procedure uses a 10:1 molar ratio of adapter to DNA insert, based on a starting quantity of 5 µg of DNA before fragmentation. The quantities below are given per 5 µg of DNA. Adjust as appropriate.*

1. Prepare a master mix containing the following reaction mix per 5 µg sample, plus a 10 % excess (round the mass of sample UP to the nearest 5 µg):

2 x DNA ligase buffer	25 µl
Illumina PE Adapter oligo mix	10 µl

2. Mix and aliquot 30 µl master mix into each sample tube containing the 10 µl of A-tailed sample. Mix and spin down.
3. Add 10 µl 2,000 U/ µl T4 DNA ligase. Mix and spin down.
4. Incubate for 15 minutes at room temperature (20-25 °C).
5. Clean ligated samples and C<sub>0</sub>t1 DNA with SPRI beads, eluting in 50 µl water. Use a 5 x excess of C<sub>0</sub>t1 for array capture and a 20 x excess for solution capture.

Alternatively, if doing a pre-hyb PCR (see below), clean up using a QIAquick PCR column as described above, eluting in 50 µl EB.

**SPRI bead cleanup**

Allow SPRI beads to come to room temperature for at least 30 minutes. Reagents need to be mixed well prior to use and should appear homogeneous and consistent in colour.

1. Add 90 µl of SPRI beads per 50 µl of adapter ligated sample in a 1.5 ml Lo-bind Eppendorf tube.
2. Vortex and leave at room temperature for 5 minutes.
3. Place tubes in a magnetic rack.
4. Leave for 5 minutes or until sample is clear.
5. Carefully remove the clear solution from the tubes and discard.

6. Dispense 700  $\mu$ l of 70 % ethanol into each tube while in the magnetic rack taking care not to disturb the magnetic beads. Aspirate and discard ethanol.
7. Repeat the ethanol wash once again (total of two washes).
8. Dry the samples on a heat block (keep the lid of the tube open) at 37 °C for 5 to 10 minutes or until the residual ethanol has evaporated.
9. Add 50  $\mu$ l of molecular biology grade water, vortex and incubate at room temperature for 2 minutes.
10. Place tubes into the magnetic rack and leave for 2-3 minutes or until sample is clear.
11. Carefully remove the water and retain in a new 1.5 ml lo-bind Eppendorf tube.
12. Repeat step 9-12 once more, retaining the water in the same 1.5 ml lo-bind tube. Total volume of elute should be 100  $\mu$ l.
13. Centrifuge the eluate at 13,000 rpm in a bench top centrifuge for 10 minutes
14. Transfer the sample to a new 1.5 ml lo-bind Eppendorf tube leaving behind any precipitated beads.
15. Quantify 1  $\mu$ l of the library using an Agilent DNA 1000 chip on a Bioanalyzer 2100 and proceed to hyb, following the manufacturer's recommended protocols.

**Optional step: pre-hyb PCR**

*Performing a small number of PCR cycles before hybridisation can improve robustness, particularly for clinical samples, and will simplify sample indexing. Amplify each 50  $\mu$ l adapter-ligated library by dividing between 4 PCR reactions.*

1. Prepare a master mix containing the following reaction mix per sample, plus a 10 % excess:

10 $\mu$ M PE2.1	10 $\mu$ l
10 $\mu$ M PEV2.2	10 $\mu$ l
2 x Phusion HF master mix	100 $\mu$ l
water	30 $\mu$ l

- Mix and aliquot 150  $\mu$ l of master mix into each 50  $\mu$ l adapter-ligated library. Mix and spin down.
- Aliquot into four 200  $\mu$ l PCR tubes (50  $\mu$ l each), and perform the following temperature cycling:

98 °C	2 minutes	
98 °C	20 seconds	
65 °C	30 seconds	x 6 cycles
72 °C	30 seconds	
72 °C	5 minutes	
4 °C	indefinitely	

- Combine all 4 reactions and clean up with SPRI beads (see above), adding 360  $\mu$ l beads to the 200  $\mu$ l PCR reactions, and eluting in 50  $\mu$ l water.

Quantify 1  $\mu$ l of library using an Agilent DNA 1000 chip on a Bioanalyzer 2100 and proceed to hybridization (Supplementary Protocol 4 or 5).

## Supplementary Protocol 4

### Array capture standard operating procedure

*This protocol is for hybridization of adapter-ligated or PCR-amplified library DNA, so must be performed after Hybrid Capture Protocol 1.*

#### Sample preparation

Following SPRI bead cleanup or pre-hyb PCR (see Hybrid Capture Protocol 1), lyophilize libraries using a SpeedVac.

1. Add 5.4  $\mu\text{l}$  of molecular biology grade water to each sample to rehydrate.
2. Vortex and centrifuge.
3. Place sample in 70 °C heat block for 10 minutes.
4. Vortex and spin down.
5. To each sample, add:

2 x Hybridization Buffer (Nimblegen)	9.0 $\mu\text{l}$
Hybridization Component A	3.6 $\mu\text{l}$

6. Vortex and spin down.
7. Place each sample in a 95 °C heat block for 10 minutes.
8. Spin down.
9. Store at 42 °C (in hyb station tube slots) until ready for hybridization.

#### Prepare slides and mixers

1. Remove the mixer from the packaging (must be used within 30 minutes of opening). Blow compressed gas across the mixer and slide.

2. Position the precision mixer alignment tool (PMAT) so that the hinge is on the left and then open it.
3. Snap the mixer onto the two alignment pins on the lid of the PMAT, with the tab end of the mixer towards the inside hinge and the mixer's adhesive gasket exposed.
4. While pushing back the plastic spring with a thumb, place the slide in the base of the PMAT so that the barcode is on the right and the corner of the slide sits against the plastic spring. Remove your thumb and make sure the spring is engaging the corner of the slide and the entire slide is registered to the edge of the PMAT closest to you. Gently blow compressed gas across the mixer and slide.
5. Using forceps remove the backing from the adhesive gasket and close the lid of the PMAT so that the gasket makes contact with the slide. Lift the lid while applying pressure through the hole in the lid of the PMAT to free the mixer-slide assembly from the pins.
6. Remove the mixer-slide assembly from the PMAT. Place the mixer-slide assembly on a smooth, dark flat surface.

**NOTE:** You can place the mixer-slide assembly on the back of a 42 °C heating block to facilitate adhesion of the mixer to the slide.

7. Rub the Mixer Brayer over the mixer just hard enough to adhere the adhesive gasket and remove any bubbles. Start in the centre of the array and rub outwards. The adhesive gasket will become clear when fully adhered to both surfaces.

**NOTE:** Mark the sample area on the other side of the working surface on the glass slide!

8. Place the mixer-slide assembly in the slide bay of the Hybridization System. Make sure the assembly is seated completely within the bay.

### Load and hybridize samples

**NOTE:** Before loading samples onto the hyb station ensure that the array area is marked on the back of the slide.

1. Using a Gilson Microman M100 pipette, slowly dispense the sample into the fill port. Dispense slowly to avoid introducing bubbles.
2. Tap the slide to remove any bubbles, dry any sample leaking from the ports with a clean tissue.

3. Adhere a seal tab over each fill and vent port on the mixer and rub the seal gently with the blunt end of the forceps to ensure a tight seal.
4. Close the bay clamp.
5. Turn the MAUI hyb station on/off with the mixing on/off switch. The 2 green mix mode lights on the front panel will flash for 5 seconds. Press the B mix mode switch to select the B mix mode while the light is flashing. The hyb system will recognise the slide in each occupied bay (the red indicator light will change to green if occupied).
6. Hybridize the sample to the array for 72 hours at 42 °C.

## Washing arrays after hybridisation

### Preparing the Wash Solutions

1. Dilute the concentrated wash solutions from the wash buffer kit. **NB:** Shown below are the quantities needed for 1 array. Volumes should be adjusted for the number of arrays to be processed. Wash buffers I, II & III come in 10X concentrated form and the stringent buffer in 2X concentrated form.

Wash Solution	Volume required (of diluted)
Wash buffer I	32 ml
Wash buffer II	164 ml
Wash buffer III	32 ml
Stringent buffer	64 ml

### Preparation of the Elution and Clean-up Solutions

**Method note:** The 125 mM sodium hydroxide (NaOH) and 20 % acetic acid should both be prepared freshly prior to using them in the elution and clean up. Molecular biology grade water (H<sub>2</sub>O) should be used for preparing the solutions.

2. Prepare 125 mM sodium hydroxide as detailed below. **NB:** The amounts below are for processing 1 array. The volumes must be adjusted to reflect the number of arrays being processed.

Solution	Volume H <sub>2</sub> O	Volume 10 M NaOH

125 mM NaOH	987.5 $\mu$ l	12.5 $\mu$ l
-------------	---------------	--------------

3. Prepare 20 % acetic acid as detailed below. **NB:** The amounts below are for processing 1 array. The volumes must be adjusted to reflect the number of arrays being processed.

Solution	Volume H <sub>2</sub> O	Volume glacial acetic acid
20 % acetic acid	400.0 $\mu$ l	100.0 $\mu$ l

4. Open a QIAquickPCR purification kit and take out the buffers to prepare them for use, see below:

Buffer	Prepare by
PE	Adding 24 ml of 100 % molecular biology grade ethanol
PB	Adding 120 $\mu$ l pH indicator I to the 30 ml bottle of PB buffer
EB	Requires no further preparation

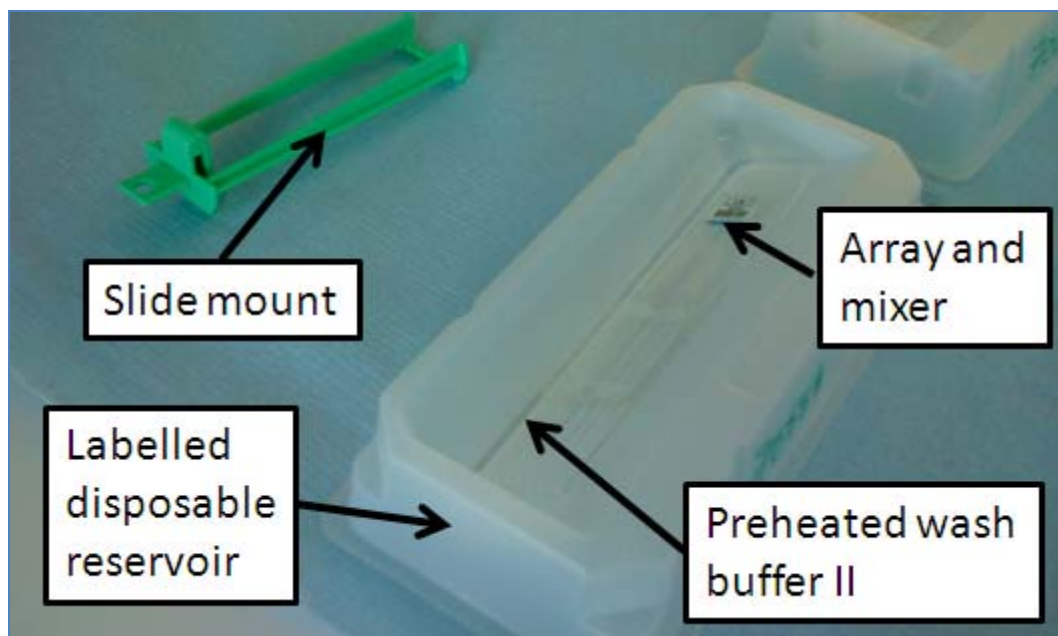
### Washing the Arrays

1. Ensure that there are the wash solutions as detailed below are prepared, aliquoted into falcon tubes where necessary and preheated if needed before starting the washing procedure.

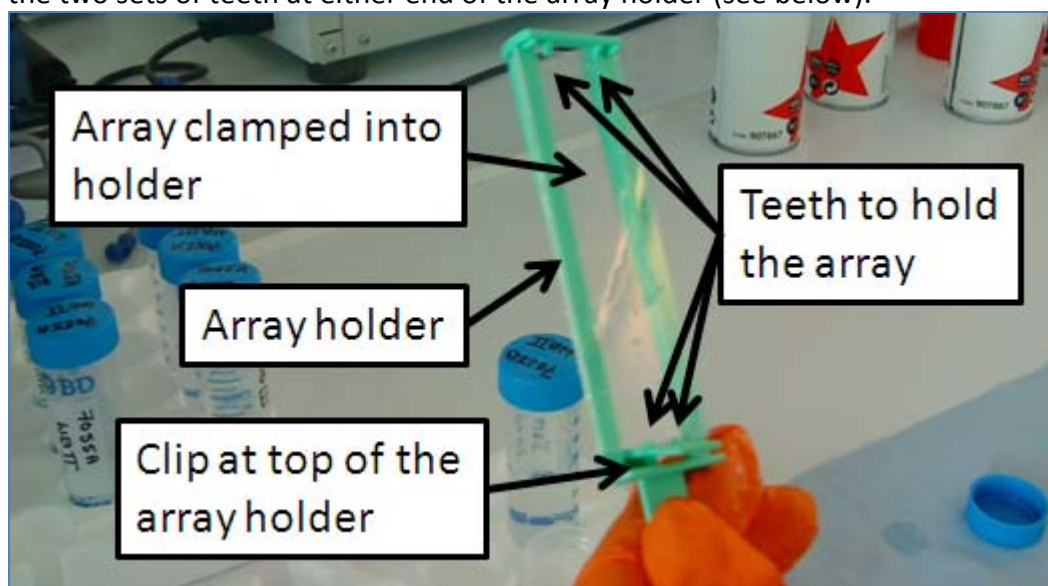
**NB:** These quantities are for one array and must be adjusted to reflect the number of arrays to be processed.

Number of aliquots	Aliquot volume (ml)	Wash solution	Container	Temperature (°C)
1	100	Wash buffer II	Ready for reservoir	42
2	32	Wash buffer II	Labeled 50ml falcon tube	20
2	32	Stringent buffer	Labeled 50ml falcon tube	47.5
1	32	Wash buffer I	Labeled 50ml falcon tube	20
1	32	Wash buffer III	Labeled 50ml falcon tube	20

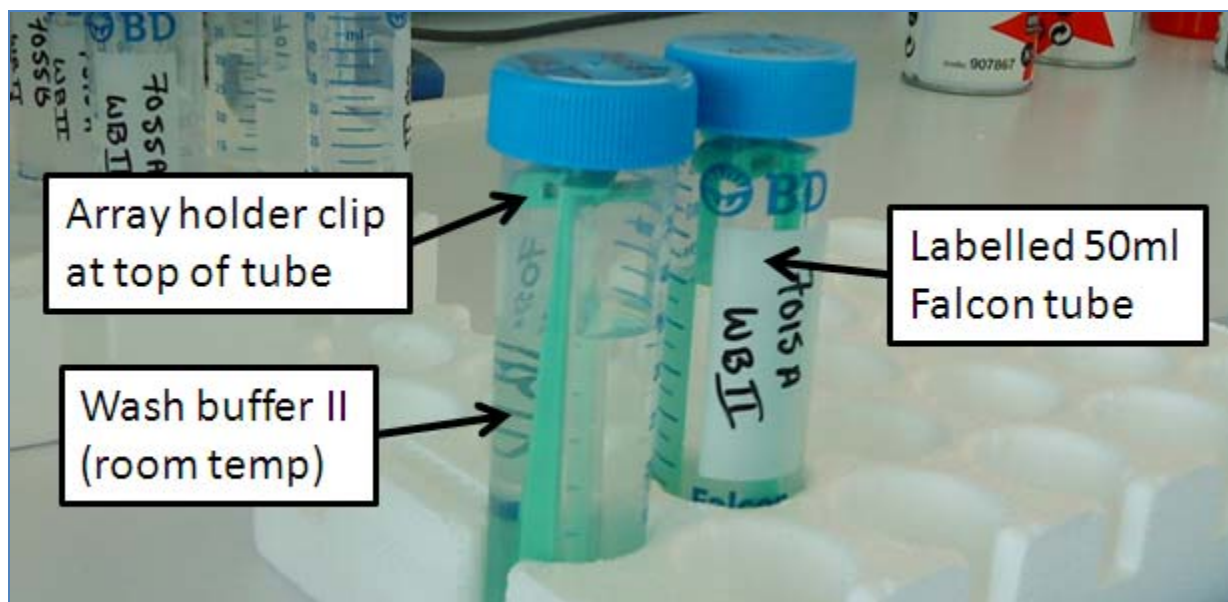
2. Label a new disposable reservoir with the sample number for each of the arrays to be washed. Remove the array from the hybridisation station and place into the reservoir. Cover the array with the 100 ml of preheated wash buffer II (see below).



3. Carefully peel off the mixer from the array taking care not to touch the area of the array under the mixer. The mixer may now be discarded.
4. Holding the array sides only, remove the array from the reservoir and clip it into an array holder by pressing down on the top clip of the holder. The array must go under the two sets of teeth at either end of the array holder (see below).

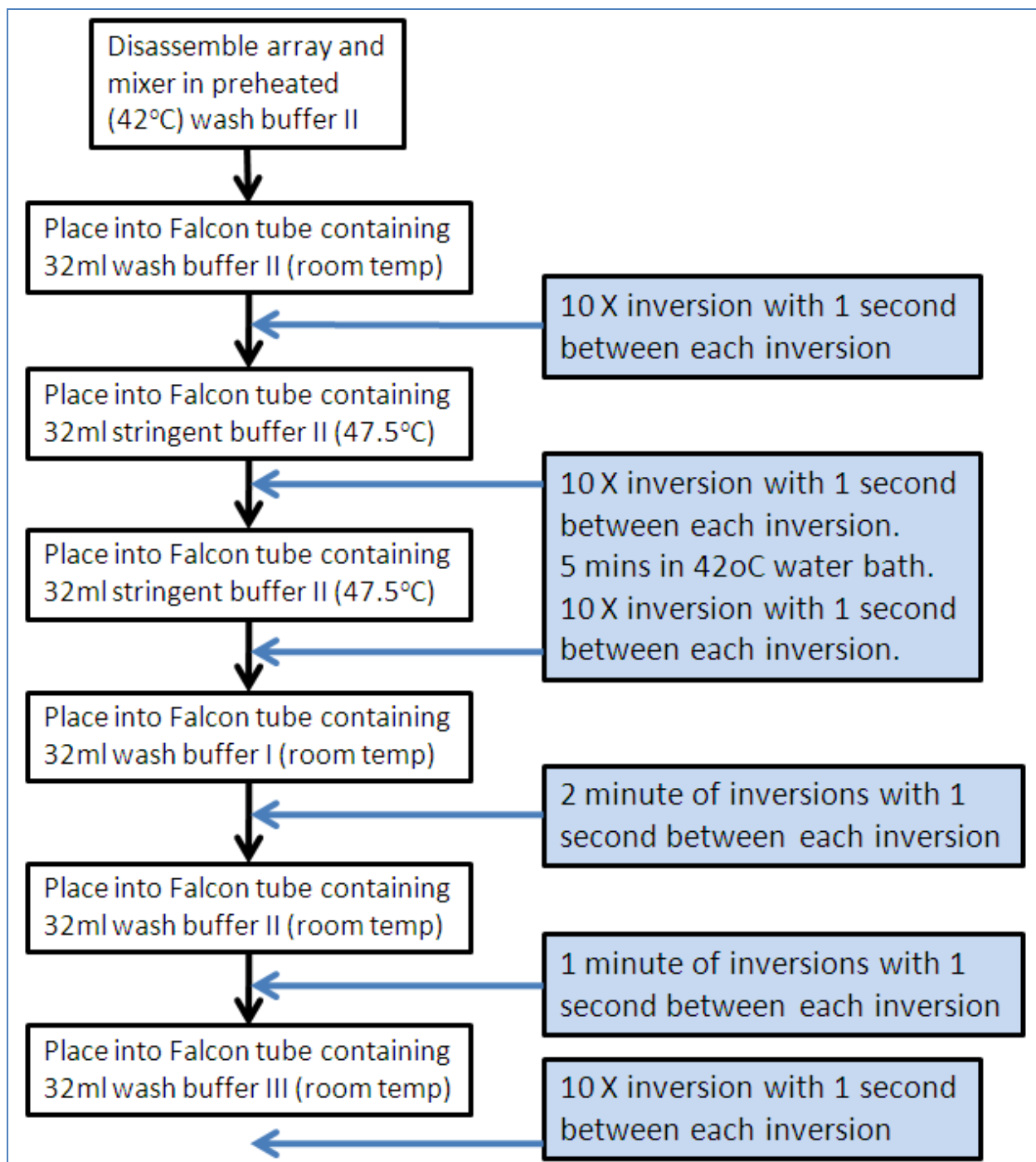


5. Place the array-holder assembly into a labelled 50 ml Falcon tube containing room temperature wash buffer II (32.0 ml). Ensure that the array holder is orientated with the clip at the top of the tube (see below). Replace the lid of the Falcon tube.



6. Invert the Falcon tube containing the array-holder assembly ten times pausing for 1 second after each inversion before inverting again.
7. Remove the array-holder assembly from the Falcon tube and place into a new labelled 50 ml Falcon tube containing preheated stringent buffer (32.0 ml at 47.5 °C). Replace the lid of the Falcon tube.
8. Invert the Falcon tube containing the array-holder assembly 10 times pausing for 1 second after each inversion before inverting again. When 10 inversions have been completed place the falcon tube with the array-holder assembly into a water bath at 42.0 °C for 5 minutes.
9. Following the incubation in the water bath invert the Falcon tube containing the array-holder assembly a further ten times pausing for 1 second after each inversion before inverting again.
10. Remove the array-holder assembly from the Falcon tube and place into a new labelled 50 ml Falcon tube containing a second aliquot of preheated stringent buffer (32.0 ml at 47.5 °C). Replace the lid of the Falcon tube.
11. Invert the Falcon tube containing the array-holder assembly 10 times pausing for 1 second after each inversion before inverting again. When 10 inversions have been completed place the falcon tube with the array-holder assembly into a water bath at 42.0 °C for 5 minutes.

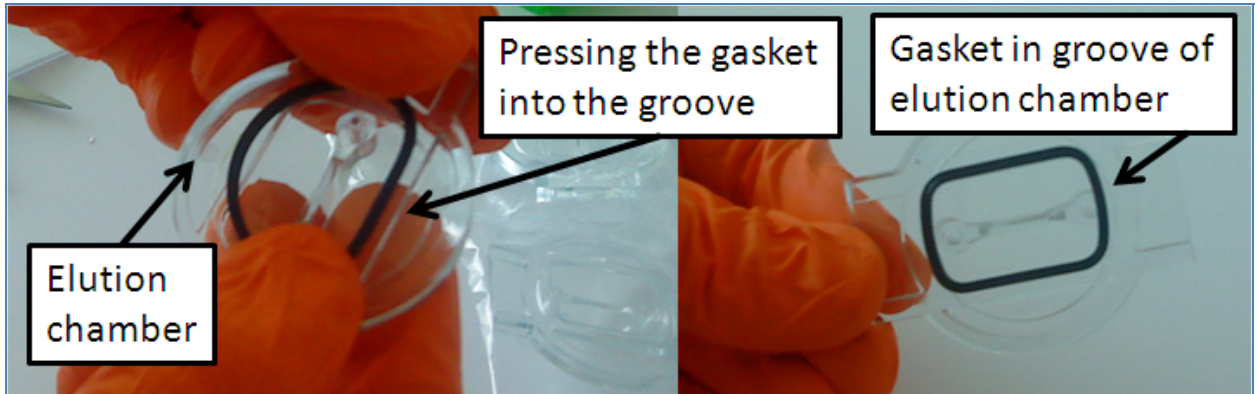
12. Following the incubation in the water bath invert the Falcon tube containing the array-holder assembly a further ten times pausing for 1 second after each inversion before inverting again.
13. Remove the array-holder assembly from the Falcon tube and place into a new labelled 50 ml Falcon tube containing preheated wash buffer I (32.0 ml). Replace the lid of the Falcon tube.
14. Invert the Falcon tube containing the array-holder assembly for 2 minutes pausing for 1 second after each inversion before inverting again.
15. Remove the array-holder assembly from the Falcon tube and place into a new labelled 50 ml Falcon tube containing preheated wash buffer II (32.0 ml). Replace the lid of the Falcon tube.
16. Invert the Falcon tube containing the array-holder assembly for 1 minute pausing for 1 second after each inversion before inverting again.
17. Remove the array-holder assembly from the Falcon tube and place into a new labelled 50 ml Falcon tube containing preheated wash buffer III (32.0 ml). Replace the lid of the Falcon tube.
18. Invert the Falcon tube containing the array-holder assembly ten times pausing for 1 second after each inversion before inverting again (see below for flow diagram of washes).



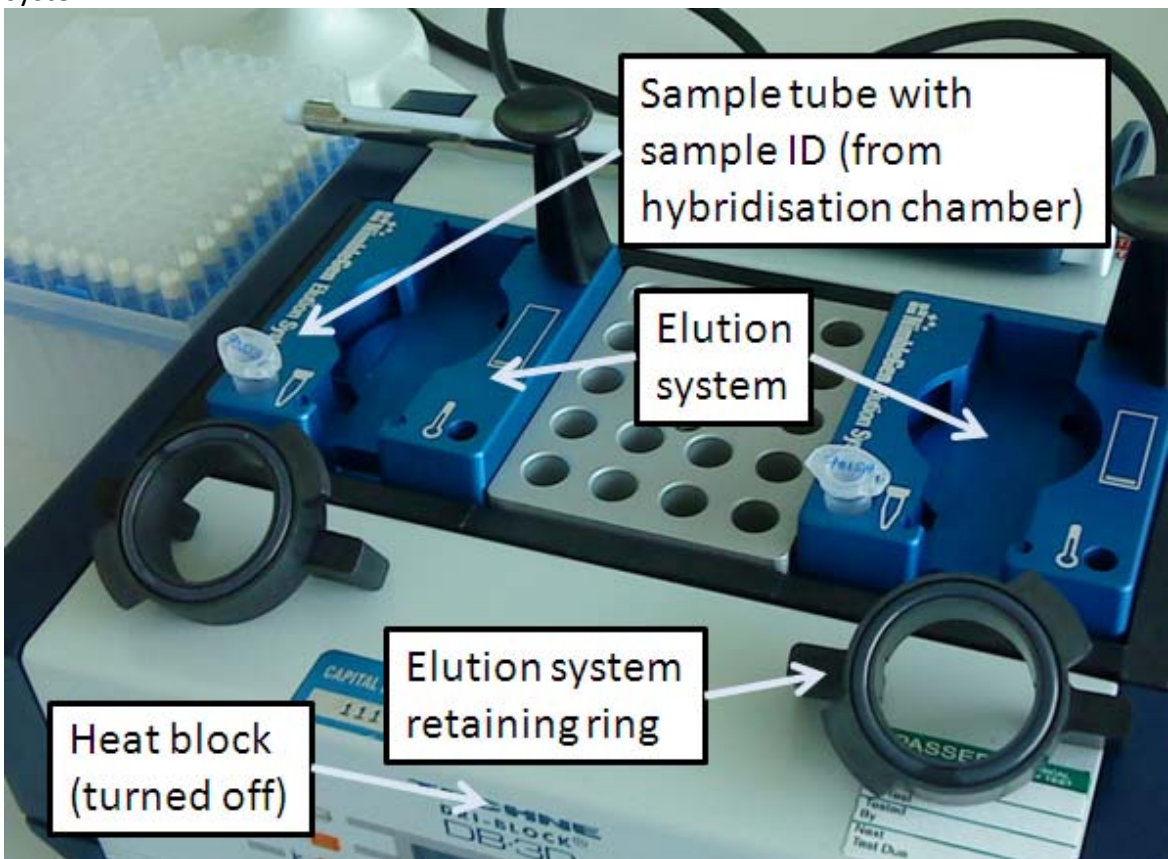
19. The array is now ready for the elution process.

**Eluting the DNA Using Sodium Hydroxide (NaOH)**

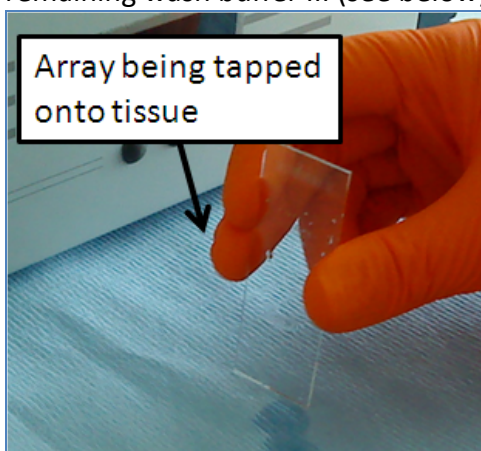
1. Unwrap a new, clean EL1 elution chamber and rubber gasket. Prepare the elution chamber by pressing the gasket into the groove around the edge (see below). Once prepared replace the elution chamber into the plastic packaging until needed.



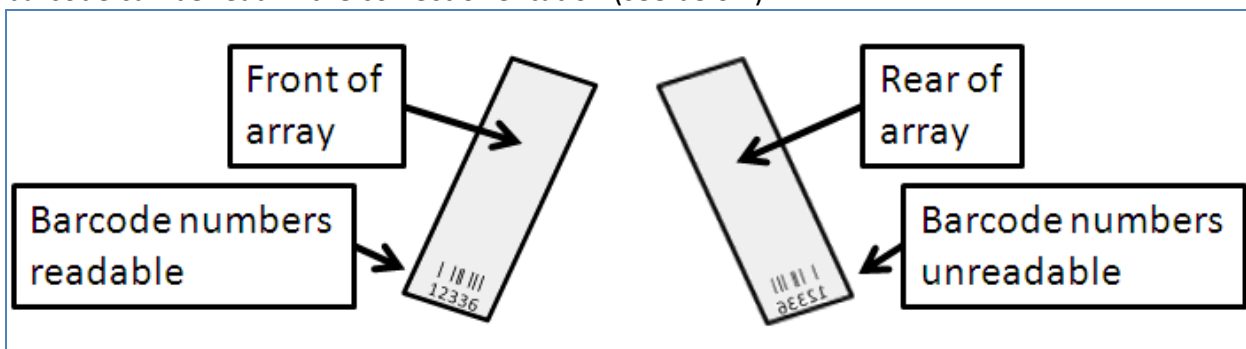
2. Place the elution system blocks into a heat block that is turned off and at room temperature for stability (see below) or place them onto the bench top. Place an elution system retaining ring ready to lock down the elution chamber near to the elution system.



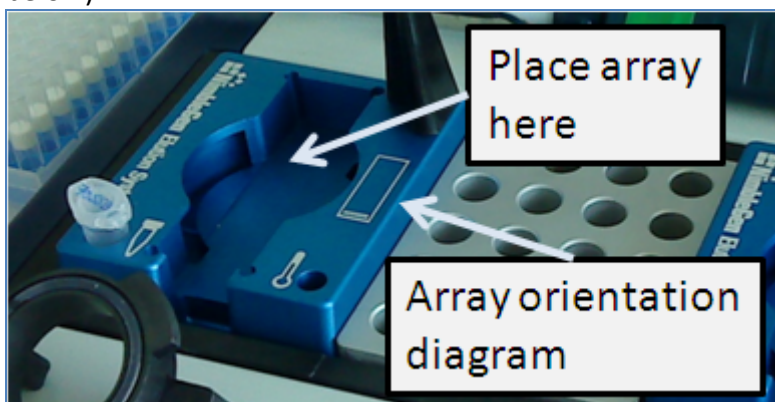
3. Remove the slide from the wash buffer III (final wash above). Unclip the array from the array holder by bending back the clip at the top of the holder. Whilst only holding the sides of the array, gently tap the bottom of the array onto some tissue to remove any remaining wash buffer III (see below)



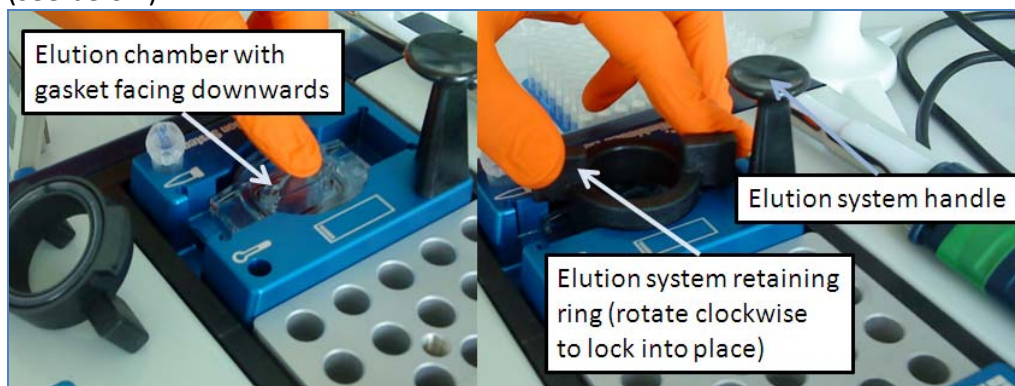
4. Look at the array barcode (which is etched into the glass) to determine the front and back faces of the array. The front face is the side through which the numbers of the barcode can be read in the correct orientation (see below).



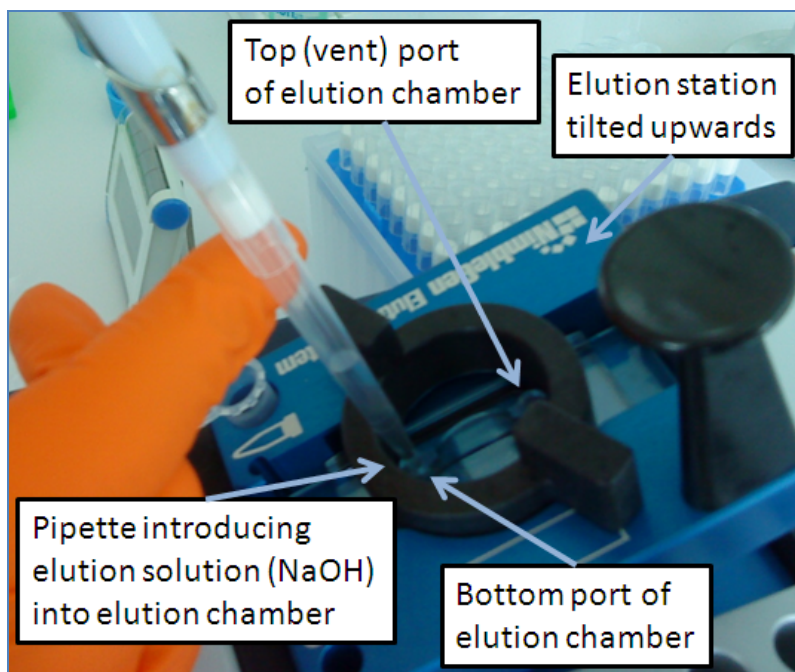
5. Wipe the rear face of the array with a low-lint tissue. Insert the array into the elution system with the barcode orientated as shown on the side of the elution system (see below).



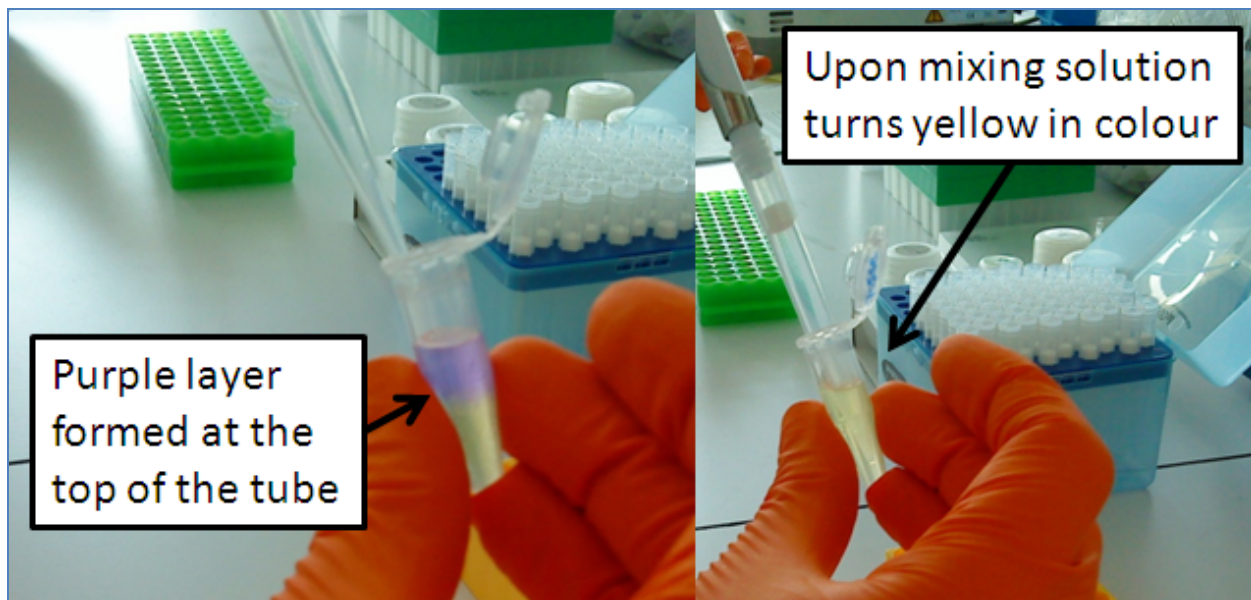
6. Place the elution chamber (prepared as above – Step 1) into the round section of the elution system orientated so that the gasket forms a seal on the array. Place the elution system retaining ring over the elution chamber and twist clockwise to lock into position (see below).



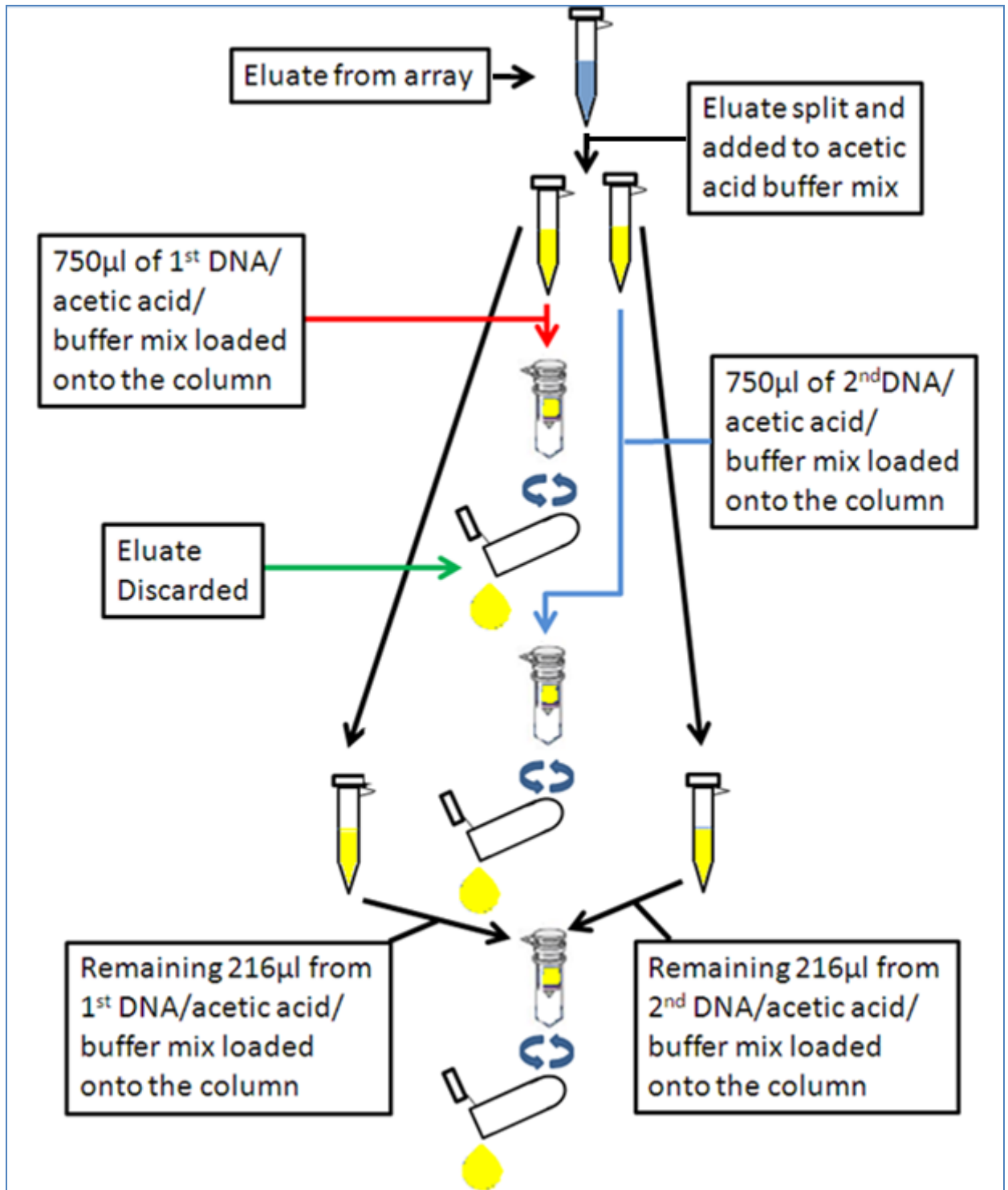
7. Tilt the elution system upwards by pulling on the elution system handle until it is locked in place.
8. Label a clean Eppendorf LoBind 1.5 ml tube with the sample name. Pipette 900  $\mu$ l of the freshly prepared 125 mM NaOH into the tube.
9. Using a pipette introduce the freshly prepared sodium hydroxide (NaOH) solution (section 'Preparation of the Elution and Clean-up Solutions' above) into the elution chamber by placing the pipette tip into the bottom port of the elution chamber (see below). Introduce the liquid slowly and keep adding liquid until liquid can be seen rising up the top (vent) port of the elution chamber. **NB:** For 2.1 M arrays about 900  $\mu$ l of NaOH is needed.



10. Leave the pipette tip in the bottom vent of the elution chamber to prevent any elution liquid from leaking out of the elution chamber. Return the elution system to the horizontal position and remove the pipette tip. Return any unused NaOH to the labelled 1.5 ml Eppendorf LoBind tube that it came from.
11. Incubate the array in the elution system at room temperature for 10 minutes.
12. Following the incubation pipette out eluted DNA and NaOH from the elution chamber and return to the unused, excess NaOH in the labelled Eppendorf LoBind 1.5 ml tube.
13. Twist the elution system retaining ring anti-clockwise to release the elution chamber which can now be discarded. If necessary the elution system can be tilted again and any remaining NaOH be pipetted out and added to the rest of the eluted DNA and NaOH. The array may now be disposed of also. Clean the elution systems and retaining ring with Azo-wipes before storage.
14. Label two 1.5 ml Eppendorf LoBind tubes with the sample details and add 16  $\mu\text{l}$  of 20 % acetic acid and 500  $\mu\text{l}$  of PB buffer to each. Pipette up and down 10 times to mix thoroughly.
15. Split the eluate (~900  $\mu\text{l}$ ) and pipette about 450  $\mu\text{l}$  into each of the labelled 1.5 ml Eppendorf LoBind tubes containing the acetic acid-PB buffer mixes. When first added the eluates sit on the top of the acid/buffer mixes and purple layers form at the top of the tubes (see below). Mix the solutions by pipetting up and down 10 times. The solutions should now be yellow, if it is purple the add 20 % acetic acid, 1  $\mu\text{l}$  at a time mixing after each addition until it turns yellow.

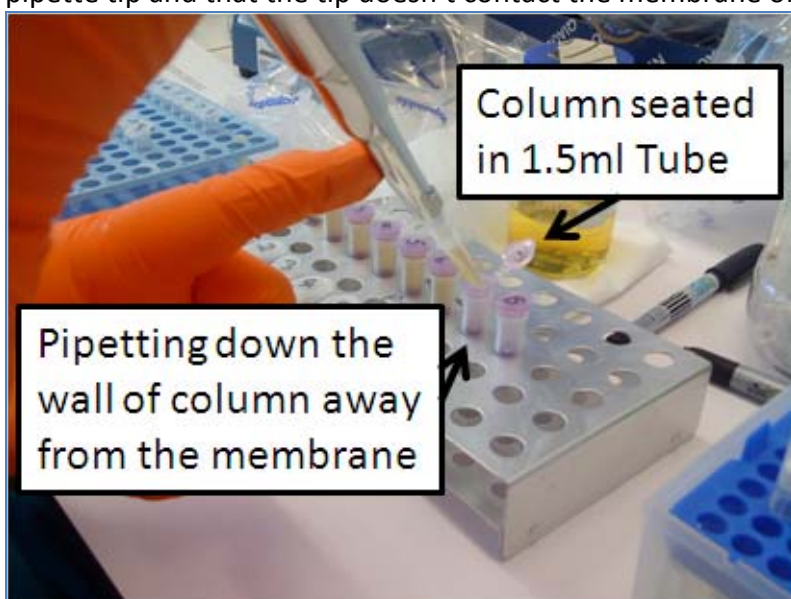


16. Remove a QIAquickMinElute column (already seated in a collection tube) from the kit stored in the fridge and place into a tube rack on the bench. Label the lid of the column with the sample identifying information that appears on the eluted sample tube lids.
17. Carefully pipette 750  $\mu$ l of the first DNA-PB buffer-acetic acid mixes into the appropriately labelled column ensuring that all of the liquid is dispensed from the pipette tip and that the tip does not contact the membrane of the column. **NB:** At this point there will be a column with 750  $\mu$ l of DNA-PB buffer-acetic acid mix loaded onto it and a 1.5 ml Eppendorf containing the remaining 216  $\mu$ l of DNA-PB buffer-acetic acid mix and a second 1.5 ml Eppendorf containing 966  $\mu$ l of DNA-PB buffer-acetic acid mix. Close the lid of the first 1.5 ml Eppendorf LoBind tube and keep on the bench until step 23.
18. Close the lid of the column and centrifuge it at 13,000 rpm for 1 minute.
19. Remove the column from the centrifuge and discard the liquid (eluate) that has been spun into the collection tube in which the column was seated. The column may be re-seated into the same collection tube that it was spun in after discarding the eluate.
20. Carefully pipette 750  $\mu$ l of the second DNA-PB buffer-acetic acid mixes from the second 1.5 ml Eppendorf LoBind tube into the appropriately labelled column. **NB:** At this point there will be a column with a second aliquot (750  $\mu$ l) of DNA-PB buffer-acetic acid mix loaded onto it and two 1.5 ml Eppendorfs containing 216  $\mu$ l of DNA-PB buffer-acetic acid mix each. Close the lid of the second 1.5 ml Eppendorf LoBind tube and keep on the bench until step 23.
21. Close the lid of the column and centrifuge it at 13,000 rpm for 1 minute.
22. Remove the column from the centrifuge and discard the liquid (eluate) that has been spun into the collection tube in which the column was seated. The column may be re-seated into the same collection tube that it was spun in after discarding the eluate.
23. Carefully pipette the two remaining 216  $\mu$ l aliquots of the DNA-PB buffer-acetic acid mixes from both of the 1.5 ml Eppendorf tubes into the appropriately labelled column.



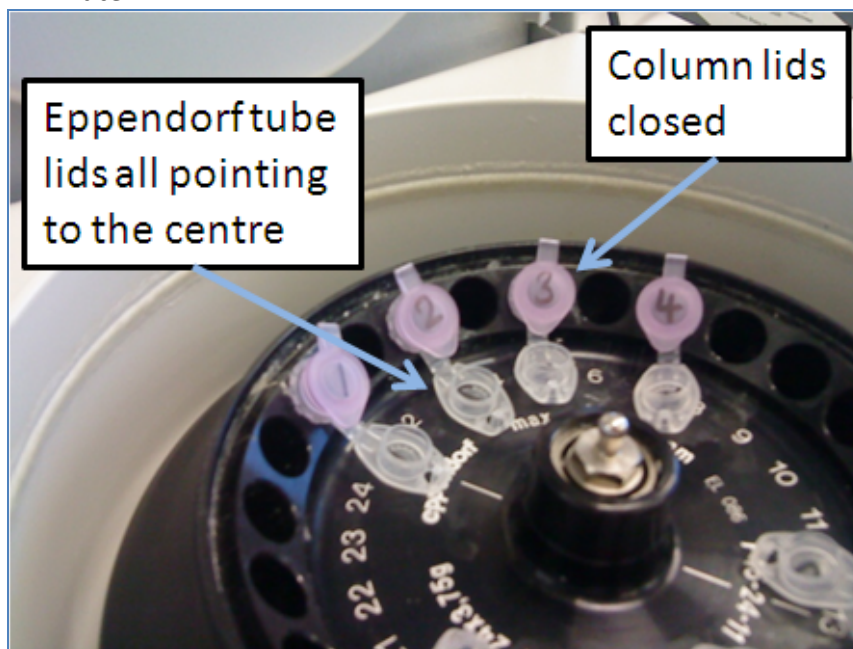
24. Close the lids of the column and centrifuge it at 13,000 rpm for 1 minute.

25. Remove the column from the centrifuge and discard the eluate that has been spun into the collection tube in which the column was seated. The column may again be re-seated into the same collection tube that it was spun in after discarding the eluate.
26. Using a clean pipette tip, pipette 750  $\mu$ l of PE buffer (from the QIAquickPCR purification kit) into the column ensuring that all of the liquid is dispensed from the pipette tip and that the tip doesn't contact the membrane of the column.
27. Close the lid of the column and centrifuge it at 13,000 rpm for 1 minute.
28. Remove the column from the centrifuge and discard the eluate that has been spun into the collection tube in which the column was seated. This collection tube may now be discarded. Place the column into a clean 1.5 ml Eppendorf LoBind tube and place into a rack on the bench.
29. Leaving the column lid open, incubate the column and tube in the rack on the bench for 2 minutes to dry. While the columns are drying a clean 1.5 ml Eppendorf LoBind tube with the sample identifier information written on the column lid and leave in a rack on the bench.
30. Transfer the column into the appropriately labelled clean 1.5 ml Eppendorf LoBind tube from step 29. Using a clean pipette tip add 34  $\mu$ l of EB buffer (from the QIAquickPCR purification kit) into the column ensuring that all of the liquid is dispensed from the pipette tip and that the tip doesn't contact the membrane of the column (see below).



31. Close the lids of the columns and leave the columns on the bench for 1 minute.

32. Arrange the tube in the centrifuge rotor so that all of the Eppendorf LoBind tube lids all point towards the centre (see below), this will help prevent the Eppendorf LoBind tube lid from breaking off during the centrifugation. Centrifuge the column at 13,000 rpm for 1 minute.



33. Remove the columns and 1.5 ml Eppendorf LoBind tubes from the centrifuge and **DO NOT** discard the eluate that has been spun into the collection tube. The QIAquickMinElute column can now be discarded.

Proceed to the Post-hyb PCR (Supplementary Protocol 6).

## Supplementary Protocol 5

### Solution capture standard operating procedure (Agilent SureSelect protocol)

*This protocol is for hybridization of adapter-ligated or PCR-amplified library DNA, so must be performed after Hybrid Capture Protocol 1.*

#### Sample preparation

Following SPRI bead cleanup (see Hybrid Capture Protocol 1), lyophilize 500 ng library + 7.5 µg C<sub>0</sub>t1 DNA using a SpeedVac..

1. Add 3.4 µl of molecular biology grade water to each sample to rehydrate.
2. Vortex and spin down.
3. To each sample, add:

SureSelect Block #1	2.5 µl
SureSelect Block #2	2.5 µl
SureSelect Block #3	0.6 µl

4. Prepare Hybridization Buffer as follows. Volume for 1 capture:

SureSelect Hyb #1	25 µl
SureSelect Hyb #2	1 µl
SureSelect hyb #3	10 µl
SureSelect Hyb #4	13 µl

Note: Do NOT keep on ice.

5. Incubate 40 µl of Hybridization Buffer and library (from step 3.) at 95 °C for 5 minutes and 65 °C for at least 5 minutes. Keep at 65 °C until RNA baits are prepared (see below).
6. Dilute RNase Block in 1:1 ratio with nuclease-free water and add 1 µl to 5 µl (500 ng) of RNA baits. Incubate for 2 min at 65 °C.

#### Hybridization

1. Mix 13 µl of hybridization buffer with RNA baits.

2. Add the 9  $\mu$ l DNA library to the hyb buffer - RNA bait mixture.
3. Seal plate with the Greiner film and incubate for 24 hours at 65 °C with a heated lid at 105 °C.

## Washing and elution

### Selection with magnetic beads

1. Prepare magnetic beads by washing with SureSelect Binding buffer 3 times and resuspend in 200  $\mu$ l of SureSelect Binding buffer. Incubate the hybrid-capture-bead solution on a Nutator for 30 minutes at room temperature (20 °C).
2. Mix hyb. mixture with beads and separate the beads by removing the supernatant.
3. Resuspend beads in SureSelect Wash Buffer #1.
4. Incubate the samples for 15 minutes at room temperature. Separate the beads and buffer on a Dynal magnetic separator and remove the supernatant.
5. Mix the beads in pre-warmed (65 °C) 500  $\mu$ L SureSelect Wash Buffer #2. Incubate the samples for 10 minutes at 65 °C. Remove wash buffer and repeat these steps for 3 times.
6. Mix the beads in 50  $\mu$ L SureSelect Elution Buffer. Incubate the samples for 10 minutes at room temperature. Separate the beads and buffer on a Dynal magnetic separator. Add 50  $\mu$ L of SureSelect Neutralization Buffer.
7. Desalt the capture solution with a Qiagen MinElute PCR purification column, eluting in 34  $\mu$ L buffer EB.

Proceed to the Post-hyb PCR (Supplementary Protocol 6).

**Supplementary Protocol 6****Hybrid capture eluate PCR standard operating procedure (for array and solution capture eluates)**

*This protocol is for post-elution amplification of captured DNA from an array or solution capture (i.e. following on from Supplementary Protocol 4 or 5).*

PCR primers:

PE.1 = 5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC\*T 3'

PE.2 = 5' CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC\*T 3'

\* indicates phosphorothioate. Both primers are HPLC purified.

1. Prepare the PCR master mix for 1 sample:

10 x PCR Buffer	5 µl
25 mM MgCl <sub>2</sub>	4 µl
2.5 mM dNTPs	4 µl
10 µM PE.1	2.5 µl
10 µM PE.2	2.5 µl
Platinum <sup>®</sup> Pfx DNA Polymerase (Invitrogen)	0.4 µl
Sample (from capture)	31.6 µl

The total volume of the final reaction should be 50 µl.

2. Vortex briefly and spin down.
3. Run on thermocycler with the following program:

94 °C for 5 min

94 °C for 15 sec

58 °C for 30 sec                      x 18 cycles

72 °C for 30 sec

72 °C for 5 min

4 °C indefinitely

4. Transfer PCR product into a 1.5ml Lo-Bind tube. Run the sample on an Agilent DNA 1000 chip on a Bioanalyzer 2100. If PCR was successful, proceed to the SPRI clean up:

**SPRI bead cleanup**

Allow SPRI beads to come to room temperature for at least 30 minutes. Reagents need to be mixed well prior to use and should appear homogeneous and consistent in colour.

1. Take 90  $\mu\text{l}$  of SPRI beads and add them to the 50  $\mu\text{l}$  of PCR sample in a 1.5 ml Lo-Bind tube.
2. Vortex and hold at room temperature for 5 minutes.
3. Place tube in the magnetic rack and leave for 5 minutes or until sample is clear.
4. Carefully remove the clear solution from the tubes and discard.
5. Dispense 500  $\mu\text{l}$  of 70 % ethanol into each tube while in the magnetic rack taking care not to disturb the magnetic beads. Aspirate and discard ethanol.
6. Repeat the ethanol wash once again. Total of two washes.
7. Dry the samples on a heat block (keep the lid of the tube open) at 37 °C for 5 – 10 minutes or until the residual ethanol has evaporated.
8. Add 50  $\mu\text{l}$  of molecular biology grade water, vortex and incubate at room temperature (20 °C) for 2 minutes.
9. Place tubes into the magnetic rack and leave for 2-3 minutes or until sample is clear.
10. Carefully remove the water and retain in a new 1.5 ml Lo-Bind tube.
11. Repeat step 9 -12 once more, retaining the water in the same 1.5 ml Lo-Bind tube. Total volume of elute should be 100  $\mu\text{l}$ .
12. Put the tube into the magnetic tool for 10 min.
13. Transfer the sample to a new 1.5 ml Lo-Bind tube leaving behind any precipitated beads.

After SPRI clean up, quantify by qPCR<sup>1</sup> and sequence.

**References**

1. Quail M.A., Swerdlow H. & Turner D.J. Improved protocols for the Illumina genome analyzer sequencing system. *Curr Protoc Hum Genet* **Chapter 18**, Unit 18 12 (2009).