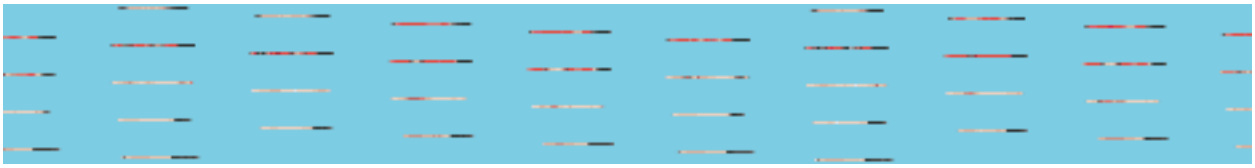
Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture

Nadin Rohland and David Reich

Genome Res. published online January 20, 2012
Access the most recent version at doi:10.1101/gr.128124.111



Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture

Nadin Rohland[1,2,*] and David Reich[1,2]

[1] Department of Genetics, Harvard Medical School, Boston, MA

[2] Broad Institute of Harvard and MIT, Cambridge, MA

* To whom correspondence should be addressed (nrohland@genetics.med.harvard.edu)

Improvements in technology have reduced the cost of DNA sequencing to the point that the limiting factor for many experiments is the time and reagent cost of sample preparation. We present an approach in which 192 sequencing libraries can be produced in a single day of technician time at cost of about $15 per sample. These libraries are effective not only for low-pass whole genome sequencing, but also for simultaneously enriching them in pools of approximately hundred individually barcoded samples for a subset of the genome without substantial loss in efficiency of target capture. We illustrate the power and effectiveness of this approach on about 2,000 samples from a prostate cancer study.

Improvements in technology have reduced the sequencing cost per base by more than a hundred thousand fold in the last decade (Lander 2011). The amount of sequence data that is needed per sample, for example for studying small target regions or low coverage sequencing of whole genomes, is often less than the commercial cost of "library" preparation, so that library preparation is now often the limiting cost for many projects. To reduce library preparation costs, researchers can purchase kits and produce libraries in their own laboratories or use published library preparation protocols (Fisher et al. 2011; Mamanova et al. 2010; Meyer & Kircher 2010). However, this approach has two limitations. First, available kits have limited throughput so that scaling to thousands of samples is difficult without automation. Second, an important application of next generation sequencing technology is to enrich sample libraries for a targeted subsection of the genome (like all the exons) (Albert et al. 2007; Gnirke et al. 2009; Hodges et al. 2007), and then to sequence this enriched pool of DNA, but such experiments are expensive because of the high costs of target capture reagents. One way to save funds is to pool samples prior to target enrichment (after barcoding to allow them to be distinguished after the data are gathered). Although the recently introduced Nextera DNA Sample Prep Kit (Illumina Inc., San Diego, USA) together with 'dual indexing' (12x8

indices and two index reads) allows higher sample throughput for library preparation (Adey et al. 2010) and pooling of up to 96 libraries, the long indexed adapter may interfere during pooled hybrid selection (see below).

We report a method for barcoded library preparation that allows highly multiplexed pooled target selection (hybrid selection or hybrid capture). We demonstrate its usefulness by generating libraries for more than 2,000 samples from a prostate cancer study that we have enriched for a 2.2 Mb subset of the genome of interest for prostate cancer. We also demonstrate the effectiveness of libraries produced with this strategy for whole genome sequencing, both by generating 40 human libraries and sequencing them to 5-fold coverage, and generating 12 microbial libraries and sequencing them to 150-fold coverage. Our method was engineered for high-throughput sample preparation and low cost, and thus we implemented fewer quality control steps and were willing to accept a higher rate of duplicated reads compared with methods that have been optimized to maximize library complexity and quality (Fisher et al. 2011; Meyer & Kircher 2010). Because of this, our method is not ideal for deep sequencing of large genomes (e.g. human genome at 30x), where sequencing costs are high enough that it makes sense to use a library that has as low a duplication rate as possible. However, our method is advantageous for projects where a modest amount of sequencing is needed per sample, so that the savings in sample preparation outweigh costs due to sequencing duplicated molecules or failed libraries. Projects that fall into this category include low pass sequencing of human genomes, microbial sequencing, and target capture of human exomes and smaller genomic targets.

Our method reduces costs and increases throughput by parallelizing the library preparation in 96-well plates, reducing enzyme volumes at a cost-intensive step, using inexpensive paramagnetic beads for size selection and buffer exchange steps (DeAngelis et al. 1995; Lennon et al. 2010; Meyer & Kircher 2010), and automation (Farias-Hesson et al. 2010; Fisher et al. 2011; Lennon et al. 2010; Lundin et al. 2010). To permit highly multiplexed sample pooling prior to target enrichment or sequencing, we attach “internal” barcodes directly to sheared DNA from a sample that is being sequenced, and flank the barcoded DNA fragments by partial sequencing adapters that are short enough that they do not strongly interfere during enrichment (the adapters are then extended after the enrichment step). By combining these individual libraries in pools and enriching them for a subset of the genome, we show that we obtain data that is effective for polymorphism discovery, without substantial loss in capture efficiency.

Outline

Our method is based on a blunt-end-ligation method originally developed for the Roche/454 platform (Stiller et al. 2009), which we have extensively modified for the Illumina/Solexa platform to reduce costs and increase sample processing speed, including by parallelizing the procedure in 96-well format and automating the labor-intensive clean-up steps (Figure 1; Figure 2a; Methods; Supplementary Notes). Some of the modifications adapt ideas from the literature, such as DNA fragmentation on the Covaris E210 instrument in 96-well PCR plates (Lennon et al. 2010) or replacing the gel-based size-selection by a bead-based, automatable, size-selection (Borgstrom et al. 2011; Lennon et al. 2010). Another change is to replace a commonly used commercial kit (AMPure XP kit) for SPRI-clean-up steps with a home-made mix. An important feature of our libraries compared with almost all other Illumina library preparation methods (Cummings et al. 2010; Mamanova et al. 2010; Meyer & Kircher 2010; Teer et al. 2010) is that we add a 6 base pair 'internal' barcoded adapter to each fragment (Craig et al. 2008). These adapters are ligated directly to the DNA fragments, leading to ‘truncated’ libraries with 34 and 33 bp overhanging adapters at the end of each DNA fragment. Adapters at this stage in our library preparation are sufficiently short that they interfere minimally with hybrid capture, compared with what we have found when long adapters are used (64 and 61 bp on either side, including the 6bp internal barcode). The ‘truncated’ adapter sites are then extended to full-length after hybrid capture allowing the libraries to be sequenced (Figure 2a). To assess how this strategy works in different sized pools (between 14 and 95), we applied it to 2.2 Mb of the genome of interest for prostate cancer, where it reduces the capture reagent that is required by two orders of magnitude while still producing highly useful data. Sequencing these libraries shows that we can carry out pooled target capture on at least 95 barcoded samples simultaneously without substantial reduction in capture efficiency.

The fact that we are using an internal barcoding strategy where barcoded oligonucleotides are ligated directly to fragmented DNA is a non-standard strategy, which deserves further discussion. First, when combined with indexing (introducing a second barcode via PCR after pooling; Figure 2a) (Meyer & Kircher 2010), an almost unlimited number of samples can be pooled and sequenced in one lane. We are currently using this strategy in our prostate cancer study to test library quality and to assess the number of sequence-able molecules per library prior to equimolar pooling for hybrid capture. Second, a potential concern of our strategy of directly ligating barcodes is that differences in ligation efficiency for different barcodes in principle could cause some barcodes to perform less efficiently than others. However, to date we have used each of 138 barcodes at least 15 times, and have not found evidence of particular barcodes performing worse than others as measured by the number of sequenced molecules per library. Third, the blunt-end ligation used in our protocol results in loss of 50% of the input DNA, because two different

adapters have to be attached to either side. This is not a concern for low coverage and small target studies using input DNA amounts of 500ng or higher, but is not the preferred strategy for samples with less input material. Fourth, chimeras of blunted DNA molecules can be created during blunt-end ligation. In our protocol, the formation of chimeras is reduced by using adapter oligonucleotides in such vast excess to the sample DNA that the chance of ligating barcodes to the DNA is much higher than ligating two sample molecules (while the adapters can form dimers, these are removed during bead clean-up). Fifth, when using our internal barcodes, it is important to pool samples in each lane in such a way that the base composition of the barcodes is balanced, as the Illumina base-calling software assumes balanced nucleotide composition especially during the first few cycles. This is of particular importance when only a few barcoded samples are being pooled. To prevent basecalling problems in such pools, a PhiX library can in principle be spiked into the library to increase diversity.

We attempted to carry out a rough calculation breaking cost for our method down into (a) reagents, (b) technician time, and (c) capital equipment (Table 1). The reagents and consumables cost is about $9 per sample without taking into account discounts that would be available for a project that produced large numbers of libraries. The cost for technician time is $3 per sample, assuming an individual makes 480 libraries on 5 plates per week. Capital costs are difficult to compute (because some laboratories may already have the necessary equipment), but if one computes the cost of a Covaris LE220 instrument, a PCR machine, and a Agilent Bravo liquid handling platform, and divides by the cost of 100,000 libraries produced over the 2-3 year lifetime of these instruments, this would add about $3 more to the cost per sample. This accounting does not include administrative overhead, space rental, process management, quality control on the preparation of reagents, bioinformatic support, data analysis, and research and development, all of which could add significantly to cost.

Application 1 - Enrichment of more than 2,000 human samples by solution hybrid capture

For many applications it is of interest to enrich a DNA sample for a subset of the genome; for example in medical genetics, a candidate region for disease risk, or all exons. The target-enriched (captured) sample can then be sequenced. To carry out studies with statistical power to detect subtle genetic effects with genome-wide statistical significance, however, it is often necessary to study thousands of samples (Kryukov et al. 2009; Lango Allen et al. 2010), which can be prohibitively expensive given current sample preparation and target enrichment costs. We designed our protocol with the aim of allowing barcoded and pooled samples to be captured simultaneously. Specifically, our libraries have internal barcodes that are tailored to pooled hybrid capture, whereas most other libraries have external barcodes in the long adaptors. It has been hypothesized that hybridization experiments using libraries that already

have long adapters do not work efficiently in pooled hybridizations because a proportion of library molecules not only hybridize to the 'baits' but also catch unwanted off-target molecules with the long adapter ('daisy-chaining') (Mamanova et al. 2010; Nijman et al. 2010), thus reducing capture efficiency (Figure 2b). In the Supplementary Notes (Influence of Adapter Length in Pooled Hybrid Capture) we present experiments showing that the number of reads mapping to the target region increased from 29% to 73% when we shortened the adapters (Table S1), providing evidence for the hypothesis that interference between barcoded adapters is lowered by short adapters, either by daisy-chaining or another mechanism. In any case, our results show short adapters improve hybridization efficiency.

To investigate the empirical performance of our libraries in the context of target capture, we produce libraries for 189 human samples starting from 0.2-4.8 μg of DNA (98% <1μg for fragmentation), prepared in two 96-well plates as in Figure S1. We combined the samples into differently sized pools of libraries (14, 28, 52 and 95) and then enriched the pooled libraries using a custom Agilent SureSelect™ Target Enrichment Kit in the volume recommended for a single sample (the target was a 2.2 Mb subset of the genome containing loci relevant to prostate cancer). We sequenced the three smaller pools together on one lane of Genome Analyzer II instrument (36 base pair, single reads) and the 95 sample pool on one lane of a HiSeq2000 instrument (50 base pair, paired end reads). We aligned the reads to the human genome using BWA (Li & Durbin 2009), after removing the first six bases of the first read which we used to identify the sample. We removed PCR duplicates using Picard's (http://picard.sourceforge.net) MarkDuplicates and computed hybrid selection statistics with Picard's CalculateHsMetrics. For the 95 sample pool (un-normalized prior to hybrid capture), f_2=93% of samples had a mean target coverage of within a factor of 2 of the median, $f_{1.5}$=67% within a factor of 1.5 of the median, and the coefficient of variation (standard deviation divided by mean coverage) was CV=0.40. For the three smaller pools where normalization was performed, coverage was in general more uniform: for the pool of 14 f_2=93%, $f_{1.5}$=86%, CV=0.66; for the pool of 28 f_2=100%, $f_{1.5}$=96%, CV=0.19; and for the pool of 52 f_2=100%, $f_{1.5}$=94%, CV=0.19 (Table S2). In the 95-sample experiment, the percentage of selected bases, defined as 'on bait' or within 250bp of either side of the baits, was 70-79% across samples (Table 2, Table S2), comparable to the literature for single sample selections (Table S3). Results on the 95-sample pool are as good as the 14-, 28-, and 52-sample pools.

To demonstrate that pooled target capture using our libraries is amenable to an experiment on the scale that is relevant to medical genetic association studies, we used the library preparation method to prepare 2,152 DNA samples from one population (African Americans) in the space of two months (1,088 prostate cancer cases and 1,064 controls from the Multiethnic Cohort Study, in collaboration with C. Haiman and B. Henderson). We normalized these samples to the lowest concentrated sample in each pool, combined

them into 15 pools of between 138 and 144 samples, and enriched these 15 pools for the 2.2 Mb target. We sequenced the captured products on a HiSeq 2000 instrument using 75 bp paired end reads to an average coverage of 4.1 in non-duplicated reads (the data will be presented in detail in a manuscript in preparation). The duplication rate of the reads was on average 72%, an elevation above the levels reported in Table 2 and Table S2 that we hypothesize is due to dilution to the lowest complexity library within the pools. We were able to solve this problem by replacing the dilution with a cherry-picking approach that combines samples of similar complexity. We tested this approach by pooling 81 prostate cancer libraries with similar complexity (allowing no more than a 5x difference in molecule count per library), resulting in a duplication rate of 24% on average at 7x coverage (Table S2e).

The experiment was highly sensitive for detecting polymorphisms in the targeted regions. After restricting to sites with at least ¼ of the average coverage, we discovered 35,211 polymorphisms at high confidence (10000:1 probability of being real based on their quality score from BWA). This is more than double the 16,457 sites discovered by the 1000 Genomes Project in 167 African ancestry samples over the same nucleotides (February 2011 data release) (The 1000 Genomes Project Consortium 2010). Exploring this in more detail, we found that we rediscovered 99.7% of sites in the 1000 Genomes Project with minor allele frequency >5% and 83% of 1000 Genomes Project sites with lower frequency in the African samples. As a second measure of the quality of our data, we compared n=1,642 African American samples that had previously been genotyped on an Illumina 1M array at 1,367 SNPs that overlapped between that array and the 2.2Mb target region of the capture experiments. We found that 99.77% of the mapped de-duplicated reads are concordant with the "gold standard" results from genotyping. As a third measure or data quality, we checked for a potential reference bias by counting the reads matching the reference and variant allele at the 1,367 SNPs where we knew the true genotypes. As shown in Figure S2 there is a slight bias (N_{ref}/N_{tot} = 1,289,080/2,537,488 = 50.8%) for the reference allele, which is sufficiently small that we do not expect it to cause a major problem for most applications such as identification of heterozygous sites.

Application 2 – Whole genome sequencing of 40 human libraries to 5x coverage

Whole genome shotgun sequencing (WGS) of mammalian genomes to high coverage (e.g. 30×) is still a process that is dominated by sequencing costs. However, lighter sequencing is of interest for some applications. For example, Genomewide Association Studies (GWAS), which have discovered more than 1,300 associations to human phenotypes (Manolio 2010), cost hundreds of dollars per sample on SNP arrays, which is less than commercial costs of library preparation, and hence sequence-based GWAS are not economical. However, the situation would change if library production costs were lower. If libraries

were inexpensive, sequencing the genome to light coverage followed by imputing missing data using a reference panel of more deeply sequenced or genotyped samples, in theory would allow more cost-effective GWAS (Li et al. 2011). With sufficiently low library production costs, sequencing may begin to compete seriously with SNP array based analysis for medical genetic association studies, as is already occurring in studies of gene expression analysis, where RNA-seq is in the process of replacing array-based methods (Majewski & Pastinen 2010).

To test if our method can produce libraries appropriate for whole genome sequencing, we prepared 40 libraries using an earlier version of our protocol that used microTUBES for shearing instead of plates and a slightly different enrichment PCR procedure (Figure S3). (A more up-to-date protocol, which involves shearing in plates and which we used to produce libraries for the prostate cancer study, further reduces costs by about $5 per sample.) Table 2 and Table S4 show the results of sequencing these libraries to an average of 5.4x coverage using 100bp paired end reads on 58 lanes on Illumina HiSeq 2000 instruments (this is a collaboration with J. Shendure and J. Kitzman, who we thank for allowing us to report these data). A high proportion (95%) of the reads align to the human reference genome (hg19) using BWA (Li & Durbin 2009) and duplicates were removed. We found that 99.86% of the mapped reads are concordant with "gold standard" SNP array data previously collected on these samples (Li et al. 2008) (sequences with quality ≥30 for the 40 libraries were compared at 585,481 SNPs). Thus, we have demonstrated that our protocol can produce libraries that are useful for low-pass whole-genome human sequencing.

Application 3 – Sequencing of 12 *E. coli* strains to 150x coverage

An important application of high-throughput sequencing is the study of microbial genomes, for example in an epidemiological context where it is valuable to study strains from many patients to study the spread of an epidemic, or in the same individual to study the evolution of an infection. Microbial genomes are small so that the required amount of sequencing per sample can be small, and thus the limiting cost is often sample preparation. To explore the utility of our library preparation protocol for microbial sequencing, we produced libraries for 12 *E. coli* strains for a project led by M. Lajoie, F. Isaacs and G. Church (who we thank for allowing us to report the data) (Isaacs et al. 2011). We produced these libraries as a single row on a 96-well plate with an input DNA amount of 1 µg together with human libraries that we were producing for another study following the protocol in Figure S4. Table 2 and Table S5 report the results of the sequencing of these 12 libraries on a single lane of a HiSeq 2000 (50 bp paired end reads). We analyzed the data after separating the libraries by sample using internal barcodes, and mapping to the *E. coli* reference (strain K12 substrain MG1655, Refseq NC_000913) using BWA (Li & Durbin 2009). Overall, 97% of reads mapped, with an average of 147-fold coverage and 1% duplicated reads.

Discussion

We have reported a high throughput library preparation method for next generation sequencing, which has been designed to allow an academic laboratory to generate thousands of barcoded libraries at a cost that is 1-2 orders of magnitude less than the commercial cost of library preparation. These libraries are appropriate for whole-genome sequencing of large and small genomes. A particularly important feature of these libraries is that they are effective for pooling approximately a hundred samples together and enriching them for a subset of the genome of interest. We have proven that the method is practical at a scale that is relevant to medical genetics by generating over 2,000 libraries for a prostate cancer study, enriching them for more than 2 Mb of interest, and obtaining sequencing data that are concordant with previously reported genotype calls.

From an engineering point of view, our method was designed with a different set of goals than have driven most previous library preparation methods. In most methods, the emphasis has been on producing libraries with maximal complexity (as measured by the number of unique molecules) and length uniformity (as measured by the tightness of the distribution of insert sizes) given the large amount of sequencing that was planned for each library. Our goal is different: to increase throughput and decrease reagent cost, while building libraries that are appropriate for pooled target capture. In this study, we empirically show that the human libraries produced by our method are complex enough that when shotgun sequenced to a coverage of around 5x, they give duplication rates of 9-20%. This duplication rate is somewhat higher than some published protocols, and the problem of duplication becomes greater as coverage increases, so that for deep sequencing studies (e.g. whole genome sequencing at 30x) where thousands of dollars are invested per sample, it may be more economical to use a more expensive library preparation protocol that minimizes duplication rates. One reason for an increased duplication rate in our libraries is our distribution of fragment insert sizes. As size selection with beads is not as tight as gel-based size selection, fragment insert sizes of the libraries produced with our protocol are less tight. Longer fragments are more prone to duplicated reads (‘optical duplicates’), where the Illumina software identifies one cluster as two adjacent clusters. Another reason for an increased duplication rate is the low input DNA amount per ligation reaction (0.75µg for each of the four ligation reactions per sample), much less than the recommended 3-5µg for standard whole genome sequencing library protocols; we also lose complexity because 50% of molecules are lost during blunt-end ligation due to wrong adapter combinations. Coverages of ~10 fold, a level where our libraries have reasonable duplication rates, have been shown to be highly effective for SNP discovery and genotype imputation (The 1000 Genomes Project Consortium 2010) and thus our libraries are valuable for most medical genetic applications. The high duplication rate for our prostate cancer target capture enrichment study (72% at about 4x coverage)

arose from the normalization strategy of diluting to the lowest complex library within each pool. We were able to lower the duplication rate to 24% at about 7x coverage when we pooled similarly complex libraries and hope to be able to lower this even further in the future.

The method we have presented is tailored to paired-end sequencing using Illumina technology, but is easy to adapt to multiplexing (we recently switched to the Multiplexing-P7 adapter) and to other technologies, for example Roche/454, Solid/ABI, and IonTorrent/LifeTechnologies. While these technologies are different at the detection stage, they are similar in sample preparation, in that technology-specific adapters are attached to DNA fragments, and the fragments are subjected to enrichment PCR to complete the adapter sites, allowing clonal amplification of the libraries and subsequent sequencing-by-synthesis. Thus, a method for one technology can be modified for use with the others. Although we only used the Agilent SureSelect platform for highly multiplexed hybrid selections, similar hybridization based target enrichment systems, such as the Illumina TruSeq Enrichment kits (Clark et al. 2011), the Roche/Nimblegen SeqCap EZ Hybridization kits and array based hybridization (Hodges et al. 2007), are expected to enrich multiplexed samples as efficiently as the Agilent system if the libraries are prepared with short adapters.

There are a number of potential improvements to our method, which should make it possible to produce libraries at even higher throughput, and to further improve library quality. A bottleneck at present is the machine time required for sample shearing. On the Covaris E210 instrument, 21 hours are required to shear to a mean insert size of 200-300 bp for a plate of 96 samples (although this takes negligible technician time), and thus two instruments would be required to produce enough sheared samples for a full time technician. However, this bottleneck could be eliminated by a recently released instrument, the Covaris LE220, which is able to shear eight samples simultaneously. The number of samples that can be pooled per lane is 159 with our 6mer 5'-barcodes, but may not be enough if for example the target size is small and the desired coverage is low. When combining the barcoding strategy with indexing via PCR, a much greater number of samples can be pooled. Another way to increase the number of samples that can be pooled is to either extend the number of barcode nucleotides or to ligate two different adapters on either side of the molecule. Further improvements to the protocol and quality control steps are important directions, which should improve the usefulness of these libraries even further.

Methods

We discuss each of the features of the steps of the protocol in turn, highlighting modifications that achieve substantial savings in terms of reagents or technician time compared with previously published protocols (Figure 1 presents the workflow of the library preparation for the hybrid selection application). A detailed protocol for all three applications (pooled hybrid selection, human shotgun sequencing, microbial sequencing) is given in the Supplementary Notes.

DNA Fragmentation

We used the Covaris E210 AFA instrument (Woburn, MA, USA) (Fisher et al. 2011; Quail et al. 2008; Quail et al. 2009) to shear the DNA into fragments of a desired length. A previous study on automated 454-library preparation (Lennon et al. 2010) showed that it is possible to use 96-well PCR plates on the E210, and we adapted the method to another plate type and a shorter fragment size (Supplementary Notes – DNA Fragmentation). This reduces reagent costs by ~90-fold compared with the microTUBE plates provided by Covaris.

Reaction Clean-Up

Library preparation involves a cascade of enzymatic reactions as well as intermediate clean-up steps for buffer-exchange. For making the clean-up steps efficient, our method heavily uses SPRI-technology, which involves paramagnetic carboxyl coated beads (DeAngelis et al. 1995) as a replacement for column-based clean-up. The beads have three advantages. First, they allow parallelization of the procedure in a way that is impossible using column-based methods (Farias-Hesson et al. 2010; Fisher et al. 2011; Lennon, Lintner, Anderson, Alvarez, Barry, Brockman, Daza, Erlich, Giannoukos, Green et al. 2010; Lundin et al. 2010; Meyer & Kircher 2010). Second, they permit size selection, which is important for removing PCR primer- or adapter-dimers (Quail et al. 2008). Third, they permit a “dual size selection” to reduce not only small DNA molecules but also long molecules (Borgstrom at al. 2011; Lennon et al. 2010) (see below – Fragment Size Selection). The commercially available kits are expensive. Thus, we used a home-made mix by combining Carboxyl-modified Sera-Mag Magnetic Speed-beads (Fisher Scientific, cat.# 65152105050250) in a PEG/NaCl buffer (‘MagNA’, see Supplementary Notes – Reaction Clean-up). We have found empirically that this combination of reagents attains performance that is comparable to the commercial kit with respect to yield and retained fragment sizes for our application (Figure S5). Using commercial bead kits instead of a home-made mix would raise reagent costs for our protocol to about $16.80 (as the cost is about $8 per sample).

Fragment Size Selection

Gel-based fragment size selections are not amenable to high throughput sample preparation even using fully automated systems such as the Pippin-Prep (SageScience, Beverly, MA, USA) or the LabChip XT Chip Prep (Caliper, Hopkinton, MA, USA). Dual SPRI size selection is faster and can be automated more easily (Borgstrom et al. 2011; Lennon et al. 2010). As SPRI-purification is already part of our protocol at all clean-up and concentration steps, we used the beads to perform size selection for whole genome shotgun sequencing applications. We aimed for a mean insert of 300bp, and thus we attempted to remove fragments larger than 400bp and smaller than 200bp (Supplementary Notes – Dual Fragment Size Selection). Size selection can be carried out at any stage of the protocol, although in the examples reported here we carried it out after fragmentation.

Sample Barcoding

We are using a blunt-end ligation procedure to add barcoded, truncated adapter to the fragmented end-repaired DNA (Stiller et al. 2009). Specifically, one of the two truncated partially double stranded adapters includes a 6-mer molecular barcode that is directly ligated to the blunted and 5'-phosphorylated DNA fragments and is therefore detected in the first 6 cycles of the first sequencing read. As the adapters are not 5'-phosphorylated (to prevent adapter dimer formation and to reduce cost) a nick fill-in-step has to be performed before enrichment PCR, which then completes the truncated adapter sites so that the libraries can be sequenced (Figure 2a). This PCR finishes the library preparation for the two WGS applications, but not for hybrid selection. For hybrid selection, we have modified the protocol so that the enrichment PCR (to complete the adapter to full-length) is carried out after hybrid capture, since we have found that the long adapters interfere with hybrid capture (see Supplementary Notes – Influence of Adapter Length in Pooled Hybrid Capture). No indexing read is needed to read out the internal barcode, but as cluster identification is carried out in these cycles in the Illumina/Solexa technology, care has to be taken to equally balance the 4 nucleotides at any of the 6 positions within the barcodes that will be sequenced together. Reaction conditions and overviews about the procedure can be found in Figure 1 and 2a (Supplementary Notes – Sample Barcoding) and Figures S1, S3-S4.

Automation

To achieve high throughput library production, it is crucial to use automated liquid handling robots (although a multi-channel pipettor can be used for 8 or 12 samples at once). The simple liquid handler that we used for the libraries we produced for microbial whole genome shotgun sequencing and hybrid capture has a 96-tip head and is thus appropriate for the clean-up and transfer steps. Over time, we slightly modified the protocol (in particular for elution volumes), so that it is now (as for the microbial

libraries) even more automated. It would be possible to further automate the protocol if the robot were capable of moving plates between positions and had heating and cooling elements. We anticipate that with a robot with all these capabilities, technician time for library production would be about 1.5 hours per plate; in particular, a single dedicated technician could produce 384 libraries in a workday simply by replacing tip boxes and providing plates with master mixes and buffer solutions on two robots in parallel.

Normalization and Pooling

To achieve an even read coverage across samples that are being sequenced simultaneously, it is necessary to measure the number of sequence-able fragments per library before pooling. For the WGS of microbial samples and the smaller sample pools for pooled hybrid selection (14, 28 and 52 samples per pool) a quantitative real-time PCR assay was utilized and for the whole genome shotgun sequencing of human samples one lane of Illumina sequencing was performed to determine the number of sequence-able molecules per library. Libraries were subsequently pooled in equimolar ratios per application and sequenced for the WGS experiments or enriched prior to sequencing for the pooled hybrid capture experiments. No normalization was carried out for the 95-sample pool for hybrid selection, but for the prostate cancer project, we utilized sequencing to determine the copy number per truncated library before pooling for hybrid capture. As we are reusing the barcoded adapters (159 total, Table S7), copy number determination via sequencing for a total of 2,152 libraries was achieved by sequencing these samples all together on just one lane (pooling 138-144 libraries, each, and using indexing PCR to introduce a unique index to one of the adapters to each of these pools via indexing PCR, Figure 2a), a cost of approximately $0.67 per library. Normalization was then performed. Detailed experimental conditions are given in the Supplementary Notes (Copy Number Determination for Equimolar Library Pooling).

Pooled Hybrid Selection

A custom Agilent SureSelect Target Enrichment Kit with a target size of 2.2Mb was developed for a medical genetics study of prostate cancer risk loci. Here, we focus on results for 189 barcoded libraries that we pooled into four pools with 14 to 95 libraries and performed a single hybrid selection per pool. Experimental conditions for the hybridization were as given in the instructions with one modification. Since our libraries only exhibit truncated adapter sites, the blocking oligonucleotides (Block #3) from the kit were replaced by Univ_Block (see Supplementary Notes – Methods and Table S7). We performed 15 cycles of enrichment PCR after hybridization to complete the adapter sites for sequencing.

Data access

Raw sequence data from the human and microbial whole genome shotgun data as well as from the pooled hybrid selection experiment (Influence of Adapter Length) are accessible under the accession number SRA047577 at NCBI's SRA. The dbGaP accession number for the prostate cancer sequence data is phs000306.v3.p1.

Competing interest statement

Harvard University has filed a patent on the techniques discussed in the manuscript. N.R and D.R. are named as co-inventors.

Acknowledgments

We are grateful to Shop Mallick, Heng Li and Andrew Kernytsky for assistance with data analysis. We thank Matthias Meyer, Brendan Blumenstiel and Daniel Herman for technical advice; David Altshuler, George Church, Sheila Fisher, Eric Lander, Erica Mazaika, Steve McCarroll, Matthias Meyer, Bogdan Pasaniuc, Alkes Price, Mark DePristo, Robert Steen and James Wilson for critical comments; Marc Lajoie, Farren Isaacs and George Church for allowing us to report on the bacterial data; Jacob Kitzman and Jay Shendure for allowing us to report on the whole genome sequencing data; and Chris Haiman and Brian Henderson for allowing us to report on the prostate cancer target capture data. We are grateful to the Biopolymers Facility at Harvard Medical School for sequencing services, to Christine and Jonathan Seidman for access to the Covaris E210, and to Stacy Gabriel and Christine Stevens for support. This research was supported by NIH grants CA129435, HL084107, CA63464 and HG004726 and NSF HOMINID grant #1032255.

Authors contribution

NR and DR conceived the experiments and wrote the manuscript, NR performed the experiments.

Figure legends

Figure 1: Experimental workflow of the library preparation protocol for 95 samples for pooled hybrid capture.

Figure 2: (A) Schematic overview of the library preparation procedure using the Illumina PE adapter (internal barcode in red). After a cascade of enzymatic reactions and clean-up steps enrichment PCR can be carried out to complete the adapter sites for Illumina PE sequencing (Rd1 SP, Rd2 SP are PE sequencing primer). Alternatively, libraries can be pooled for hybrid selection (if desired), and then enrichment PCR can be done after hybrid selection. To achieve an even higher magnitude of pooling for sequencing, 'indexing PCR' can be performed instead of 'enrichment PCR', whereby unique indices (in purple) are introduced to the adapter, and a custom index sequencing primer [index-PE-sequencing-Primer] is used to read out the index in a separate read. Finished libraries that have all the adapters necessary to allow sequencing are marked with an X. **(B)** Schematic figure of 'daisy-chaining' during pooled solution hybrid capture, which may explain why a large proportion of molecules are empirically observed to be off-target when using long adapters. Library molecules exhibiting the target sequences hybridize to biotinylated baits, but unwanted library molecules can also hybridize to the universal adapter sites. The adapters of our 'truncated' libraries (including barcode: 34 and 33 bp) are about half the length of regular 'long' adapters (64 and 61 bases), and thus may be less prone to binding DNA fragments that do not belong to the target region.

Table 1: Cost and time assumptions for library preparation

Task	Item	Price per sample	Sample processing time for 192 samples	Technician hands-on time for 192 samples
Library preparation				
Covaris shearing	Plate	$ 0.04	44h	2h
Cleanup	Beads and ethanol	$ 0.54	4h	2h
Blunt end repair	Kit	$ 0.75	1.5h	0.5h
Barcoded adapter ligation	Kit and oligonucleotides	$ 3.30	1.2h	0.5h
Nick fill in reaction	Enzyme and buffer	$ 0.48	1h	0.5h
Amplification	Kit and oligonucleotides	$ 1.58	1-2h	0.5h
Copy number determination	qPCR reagents or sequencing cost*	$ 0.67		
Consumables	Plates and pipette tips	$ 1.40		
	Total for library preparation	*$ 8.76*		*6h*
Technician salary	*Total assuming 480 / week*†	*$ 3.00*		
Capital equipment	*Amortized over 100,000 libraries*††	*$ 3.00*		
	Total for library preparation	*$ 14.76*		

* qPCR for two measurements per sample, or sequencing one lane SR36 and indexing read, divided by 2,152 libraries

† $3/sample for personnel time (assuming salary and benefits of $70,000 per year and processing five 96-well plates/week

†† $3/sample for capital equipment (assuming purchase of a Covaris LE220 instrument, a PCR machine, and a Agilent Bravo liquid handling platform, and dividing over 100,000 libraries).

Table 2: Sequencing results

Application	No. of libraries	Input DNA (µg)	Normalization strategy	PF reads per library	% reads aligning to reference genome	% duplic-ated reads (removed)	Mean target * coverage per library	% selected bases #	% target* with 2x coverage
Human hybrid selection [1]	14	0.6-0.9	dilution	2.8×10^5	73	53.6	0.9	78	23
Human hybrid selection [1]	28	0.2-0.9	dilution	3.3×10^5	72	56.4	1.1	76	31
Human hybrid selection [1]	52	0.3-0.9	dilution	2.7×10^5	74	51.1	1.1	78	29
Human hybrid selection [2]	95	0.2-4.8	unnormalized	1×10^6	89	37.5	7.4	74	79
Human hybrid selection [2]	81	0.6-2.6	cherry picking	5.6×10^5	92	24.4	7.1	92	87
Human whole-genome shotgun[3]	40	0.75		7.1×10^7	95	14.4	5.4	n/a	n/a
Microbial sequencing [2]	12	1		7.2×10^6	97	1	147	n/a	n/a

* Target for the hybrid selection experiment is defined as the regions where baits were designed.
# selected bases is defined as in Picard as 250bp on either side of the bait (target)
[1] 36 cycles of single read sequencing on GAII
[2] 50 cycles of paired end sequencing on HiSeq2000
[3] 100 cycles of paired end sequencing on HiSeq2000; 4 libraries were prepared for each of 10 samples

References

Adey, A., H.G. Morrison, Asan, X. Xun, J.O. Kitzman, E.H. Turner, B. Stackhouse, A.P. MacKenzie, N.C. Caruccio, X. Zhang et al. 2010. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol* **11:** R119.

Albert, T.J., M.N. Molla, D.M. Muzny, L. Nazareth, D. Wheeler, X. Song, T.A. Richmond, C.M. Middle, M.J. Rodesch, C.J. Packard et al. 2007. Direct selection of human genomic loci by microarray hybridization. *Nat Methods* **4:** 903-905.

Borgstrom, E., S. Lundin, and J. Lundeberg. 2011. Large scale library generation for high throughput sequencing. *PLoS One* **6:** e19119.

Clark, M.J., R. Chen, H.Y. Lam, K.J. Karczewski, G. Euskirchen, A.J. Butte, and M. Snyder. Performance comparison of exome DNA sequencing technologies. *Nat Biotechnol*. Advanced online publication

Craig, D.W., J.V. Pearson, S. Szelinger, A. Sekar, M. Redman, J.J. Corneveaux, T.L. Pawlowski, T. Laub, G. Nunn, D.A. Stephan et al. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. *Nat Methods* **5:** 887-893.

Cummings, N., R. King, A. Rickers, A. Kaspi, S. Lunke, I. Haviv, and J.B. Jowett. 2010. Combining target enrichment with barcode multiplexing for high throughput SNP discovery. *BMC Genomics* **11:** 641.

DeAngelis, M.M., D.G. Wang, and T.L. Hawkins. 1995. Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res* **23:** 4742-4743.

Farias-Hesson, E., J. Erikson, A. Atkins, P. Shen, R.W. Davis, C. Scharfe, and N. Pourmand. 2010. Semi-automated library preparation for high-throughput DNA sequencing platforms. *J Biomed Biotechnol* **2010:** 617469.

Fisher, S., A. Barry, J. Abreu, B. Minie, J. Nolan, T.M. Delorey, G. Young, T.J. Fennell, A. Allen, L. Ambrogio et al. 2011. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol* **12:** R1.

Gnirke, A., A. Melnikov, J. Maguire, P. Rogov, E.M. LeProust, W. Brockman, T. Fennell, G. Giannoukos, S. Fisher, C. Russ et al. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* **27:** 182-189.

Hodges, E., Z. Xuan, V. Balija, M. Kramer, M.N. Molla, S.W. Smith, C.M. Middle, M.J. Rodesch, T.J. Albert, G.J. Hannon et al. 2007. Genome-wide in situ exon capture for selective resequencing. *Nat Genet* **39:** 1522-1527.

http://picard.sourceforge.net.

Isaacs, F.J., Carr, P.A., Wang, H.H., Lajoie, M.J., Sterling, B., Kraal, L., Tolonen, A.C., Gianoulis, T.A., Goodman, D.B., Reppas, N.B. et al. 2011. Precise manipulation of chromosomes in vivo enables genome-wide codon replacement. *Science* **333:** 348-353

Kryukov, G.V., A. Shpunt, J.A. Stamatoyannopoulos, and S.R. Sunyaev. 2009. Power of deep, all-exon resequencing for discovery of human trait genes. *Proc Natl Acad Sci U S A* **106:** 3871-3876.

Lander, E.S. 2011. Initial impact of the sequencing of the human genome. *Nature* **470:** 187-197.

Lango Allen, H. K., Estrada, G., Lettre, S.I., Berndt, M.N., Weedon, F., Rivadeneira, C.J., Willer, A.U., Jackson, S., Vedantam, S., Raychaudhuri et al. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467:** 832-838.

Lennon, N.J., R.E. Lintner, S. Anderson, P. Alvarez, A. Barry, W. Brockman, R. Daza, R.L. Erlich, G. Giannoukos, L. Green et al. 2010. A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454. *Genome Biol* **11:** R15.

Li, H. and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25:** 1754-1760.

Li, J.Z., D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319:** 1100-1104.

Li, Y., C. Sidore, H.M. Kang, M. Boehnke, and G. Abecasis. 2011. Low coverage sequencing: Implications for the design of complex trait association studies. *Genome Res*. **21:** 940-951

Lundin, S., H. Stranneheim, E. Pettersson, D. Klevebring, and J. Lundeberg. 2010. Increased throughput by parallelization of library preparation for massive sequencing. *PLoS One* **5:** e10029.

Majewski, J. and T. Pastinen. 2010. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet* **27:** 72-79.

Mamanova, L., A.J. Coffey, C.E. Scott, I. Kozarewa, E.H. Turner, A. Kumar, E. Howard, J. Shendure, and D.J. Turner. 2010. Target-enrichment strategies for next-generation sequencing. *Nat Methods* **7:** 111-118.

Manolio, T.A. 2010. Genomewide association studies and assessment of the risk of disease. *N Engl J Med* **363:** 166-176.

Meyer, M. and M. Kircher. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc* **2010:** pdb prot5448.

Nijman, I.J., M. Mokry, R. van Boxtel, P. Toonen, E. de Bruijn, and E. Cuppen. 2010. Mutation discovery by targeted genomic enrichment of multiplexed barcoded samples. *Nat Methods* **7:** 913-915.

Quail, M.A., I. Kozarewa, F. Smith, A. Scally, P.J. Stephens, R. Durbin, H. Swerdlow, and D.J. Turner. 2008. A large genome center's improvements to the Illumina sequencing system. *Nat Methods* **5:** 1005-1010.

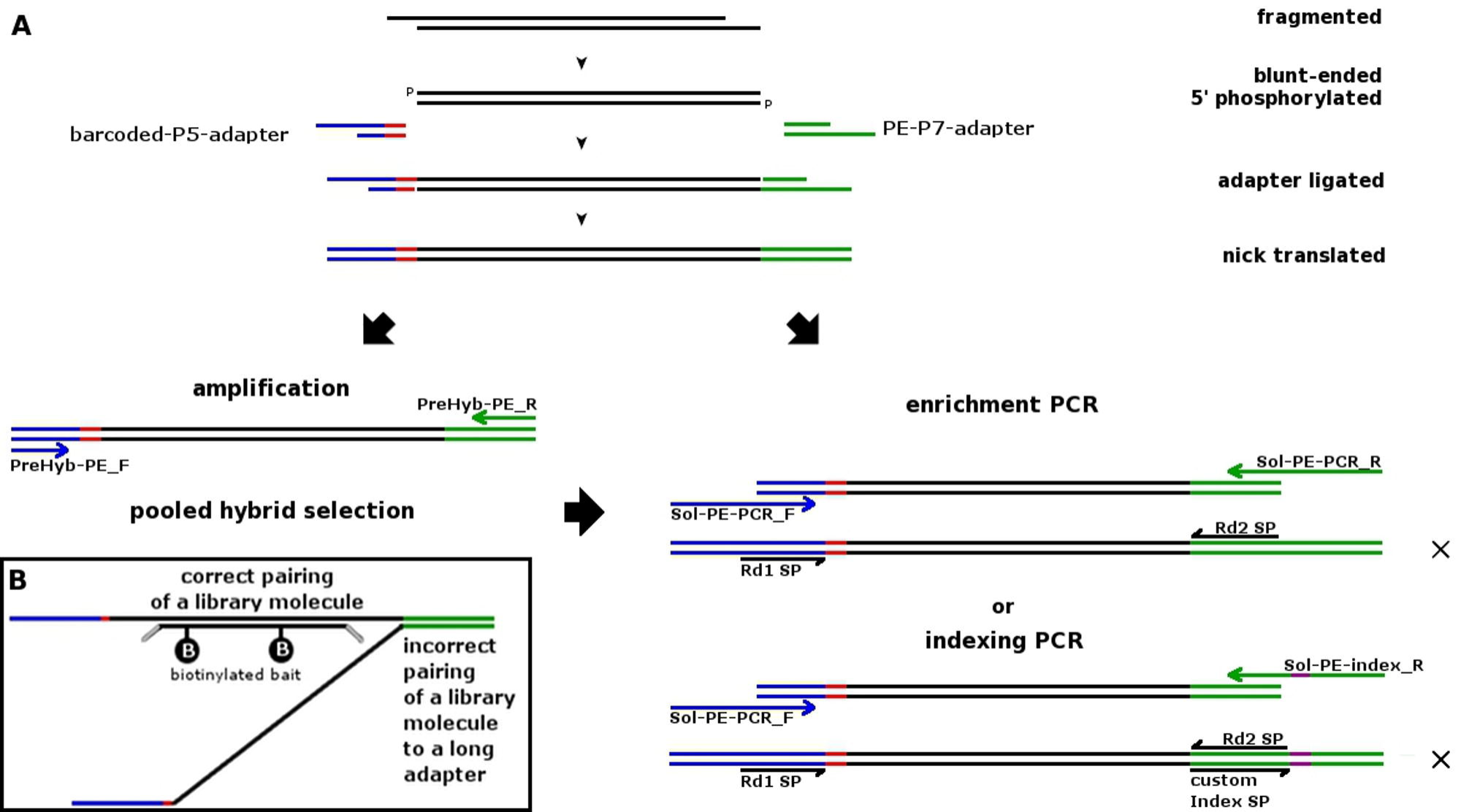
Quail, M.A., H. Swerdlow, and D.J. Turner. 2009. Improved protocols for the illumina genome analyzer sequencing system. *Curr Protoc Hum Genet* **Chapter 18:** Unit 18 12.

Stiller, M., M. Knapp, U. Stenzel, M. Hofreiter, and M. Meyer. 2009. Direct multiplex sequencing (DMPS)--a novel method for targeted high-throughput sequencing of ancient and highly degraded DNA. *Genome Res* **19:** 1843-1848.

Teer, J.K., L.L. Bonnycastle, P.S. Chines, N.F. Hansen, N. Aoyama, A.J. Swift, H.O. Abaan, T.J. Albert, E.H. Margulies, E.D. Green et al. 2010. Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. *Genome Res* **20:** 1420-1431.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467:** 1061-1073.

		95 samples
Covaris shearing	≤1µg genomic DNA, final volume 200µl in ABI MicroAmp Fast 96-well reaction plate E210 settings: Duty cycle: 10%, Intensity: 5, Cycles/burst: 200, 780sec	volume 200µl
Concentration of DNA	2x MagNA (400µl)	elution volume 30µl 19µl 11µl backup
Blunt-end repair	Quick Blunting Kit (NEB: E1201L): 1x Blunting Buffer, 100µM dNTP Mix, 0.25µl Enzyme Mix 20' @ 12ºC, 15' @ 37ºC	reaction volume 25µl
Reaction clean-up	2x MagNA (50µl)	elution volume 28.8µl 28.8µl
Barcoded adapter ligation	Quick Ligation Kit (NEB: M2200L): 1x Quick Ligation Reaction Buffer,1.2µl Quick T4 DNA Ligase, 6.7µM barcoded-P5-adapter, 6.7µM PE-P7-adapter 25' @ RT	reaction volume 60µl
Reaction clean-up	1.6x MagNA (96µl)	elution volume 40µl 40µl
Nick fill-in reaction	*Bst* DNA Polymerase, Large Fragment (NEB: M0275L): 1x ThermoPol Reaction Buffer, 250µM dNTP Mix, 16U *Bst* DNA Polymerase 15' @ 37ºC	reaction volume 50µl
Reaction clean-up	1.6x MagNA (80µl)	elution volume 35µl 35µl
Amplification	AccuPrime *Taq* DNA Polymerase High Fidelity (Invitrogen: 12346-086): 1x AccuPrime PCR Buffer II, 1U AccuPrime *Taq* DNA Polymerase High Fidelity, 200nM PreHyb-PE_F, 200nM PreHyb-PE_R, 200µM dNTP Mix 3' @ 98ºC, 6x (80" @ 98ºC, 45" @ 55ºC, 60" @ 68ºC), 10' @ 72ºC)	reaction volume 50µl
Reaction clean-up	1.6x MagNA (80µl)	finished libraries: elution volume 50µl


A
fragmented
blunt-ended
5' phosphorylated
barcoded-P5-adapter
PE-P7-adapter
adapter ligated
nick translated
amplification
PreHyb-PE_R
PreHyb-PE_F
pooled hybrid selection
enrichment PCR
Sol-PE-PCR_R
Sol-PE-PCR_F
Rd2 SP
Rd1 SP
or
indexing PCR
Sol-PE-index_R
Sol-PE-PCR_F
Rd2 SP
Rd1 SP
custom
Index SP
B
correct pairing
of a library molecule
biotinylated bait
incorrect
pairing
of a library
molecule
to a long
adapter