Whole exome capture in solution with 3Gbp of data

*Genome Biology* 2010, 11:R62 doi:10.1186/gb-2010-11-6-r62

Matthew N Bainbridge (bainbrid@bcm.edu)
Min Wang (mwang@bcm.edu)
Daniel L Burgess (daniel.burgess@roche.com)
Christie Kovar (kovar@bcm.edu)
Matthew J Rodesch (matthew.rodesch@roche.com)
Mark D’Ascenzo (mark.dascenzo@roche.com)
Jacob Kitzman (jacob.kitzman@roche.com)
Yuan-Qing Wu (qing@bcm.edu)
Irene Newsham (newsham@bcm.edu)
Todd A Richmond (todd.richmond@roche.com)
Jeffrey A Jedeloh (jeffery.jeddeloh@roche.com)
Donna Muzny (donnam@bcm.edu)
Thomas J Albert (thomas.albert@roche.com)
Richard A Gibbs (agibbs@bcm.edu)

ISSN 1465-6906

Article type Method

Submission date 14 April 2010

Acceptance date 17 June 2010

Publication date 17 June 2010

Article URL [http://genomebiology.com/2010/11/6/R62](http://genomebiology.com/2010/11/6/R62)

This peer-reviewed article was published immediately upon acceptance. It can be downloaded, printed and distributed freely for any purposes (see copyright notice below).

Articles in *Genome Biology* are listed in PubMed and archived at PubMed Central.

For information about publishing your research in *Genome Biology* go to [http://genomebiology.com/info/instructions/](http://genomebiology.com/info/instructions/)

© 2010 Bainbridge et al., licensee BioMed Central Ltd. This is an open access article distributed under the terms of the Creative Commons Attribution License ([http://creativecommons.org/licenses/by/2.0](http://creativecommons.org/licenses/by/2.0)). which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Whole exome capture in solution with 3Gbp of data

Matthew N Bainbridge\textsuperscript{1,2}, Min Wang\textsuperscript{1}, Daniel L Burgess\textsuperscript{3}, Christie Kovar\textsuperscript{1}, Matthew J Rodesch\textsuperscript{3}, Mark D'Ascenzo\textsuperscript{3}, Jacob Kitzman\textsuperscript{3}, Yuan-Qing Wu\textsuperscript{1}, Irene Newsham\textsuperscript{1}, Todd A Richmond\textsuperscript{3}, Jeffrey A Jeddeloh\textsuperscript{3}, Donna Muzny\textsuperscript{1}, Thomas J Albert\textsuperscript{3}, Richard A Gibbs\textsuperscript{1} §

\textsuperscript{1}Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA.

\textsuperscript{2}Department of Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA.

\textsuperscript{3}Roche NimbleGen, Inc., 504 S. Rosa Road Madison, WI 53719, USA.

§Corresponding author

Email addresses:

MNB: bainbrid@bcm.edu

MW: mwang@bcm.edu

DLB: daniel.burgess@roche.com

CK: kovar@bcm.edu

MJR: matthew.rodesch@roche.com

MDA: mark.dascenzo@roche.com

JK: jacob.kitzman@roche.com

YQW: qing@bcm.edu
IN: newsham@bcm.edu

TAR: todd.richmond@roche.com

JAJ: jeffery.jeddeloh@roche.com

DM: donnam@bcm.edu

TJA: thomas.albert@roche.com

RAG: agibbs@bcm.edu
Abstract

We have developed a solution-based method for targeted DNA capture-sequencing that is directed to the complete human exome. Using this approach allows the discovery of greater than 95% of all expected heterozygous singe base variants, requires as little as 3Gbp of raw sequence data and constitutes an effective tool for identifying rare coding alleles in large scale genomic studies.

Background

DNA sequence capture is an effective technique for enrichment of significant subfractions of the genome for targeted analysis. Sequence capture is generally conducted with either a solid-phase substrate, usually a glass microarray, or alternatively, in solution [1-4]. Solution-capture performance, however, has not been thoroughly compared to microarray with respect to uniformity of representation of the targeted DNA bases or evenness of DNA sequencing coverage depth between regions [5, 6]. Further, solution-based sequence capture has not been demonstrated to work effectively at the scale of a human exome (approximately 30Mbp) but typically has been limited to targets <5Mbp in size. Despite this, solution capture has several advantages when compared to microarrays: the reagent cost is lower; less DNA is required, and, because the capture method can be conducted entirely in small laboratory tubes, it is readily scaled and automated. Before solution-capture sequencing can be widely adopted, however, the reproducibility of the method must first be demonstrated, and targets should show similar levels of coverage from capture to capture. Ideally, solution-capture methods should also be able to be coupled to
different sequencing technology platforms, and reliably produce suitable levels of
enrichment that routinely enable the discovery of rare genetic variants.

This report is the first demonstration of whole exome capture in solution (see Table
1). We demonstrate similar levels of specificity to microarray base techniques
without sacrificing reproducibility or specificity of either the capture or variant
discovery while maintaining all the advantages of solution-based techniques over
microarray capture.

To test the reproducibility of our recent innovations of liquid DNA capture, technical
replicate capture experiments were performed and subsequently sequenced on the
SOLiD [7] platform. Capture followed by Illumina [8] sequencing was also
performed with both a fragment (frag) and paired-end (PE) library to test the merits of
employing paired-end data, versus single-ended reads. Finally, we used each of these
data sets to test the ability to discover single nucleotide variants across the exome.

**Results and Discussion**

Here we report the performance of newly developed methods for sequence-capture in
solution. The procedures were tested with respect to reproducibility of capture,
agnosticism to sequencer platform, and the ability to discover genetic variation in
human gene coding regions. In total, six captures of the Consensus Coding Sequence
(CCDS) exons [9, 10] were performed using DNA from one HapMap sample
(NA12812). Four separate, replicate capture libraries were prepared for solid
sequencing and one additional library was prepared for both the Illumina fragment
and PE sequencing. In total, 23Gbp of data was uniquely aligned to the human
reference genome (see Supplementary table S1 in Additional file 1). Variants were called using algorithms tailored to each sequence platform and corresponding data type, and compared to known variants in the HapMap sample (see Methods).

To first test the reproducibility of the entire capture-sequencing process, 4 technical replicate solution captures were performed. Each replicate was sequenced using one quarter of the same SOLiD DNA sequencer ‘slide’ to eliminate run-by-run sequencer variability. The standard procedure for categorizing data from this sequencing platform is to identify individual sequence reads that can be mapped at high stringency to a reference sequence. These ‘mappable’ reads constitute the usually cited yield for each sequence run. Here, the sequences from four technical replicate libraries had an average of 49.6% (S.D. 1.23) of mappable reads derived from the capture target regions with the remainder mapping elsewhere to the genome. This efficiency of properly targeted sequence reads represents a value similar to Ng et al [3], and higher than Choi et al [11]. The final DNA sequence coverage across each target had >98% correlation between all four libraries, (see Figure 1). In 3 of 4 experiments, >65% of the targeted bases were covered 10 or more times, and the observed variation in the coverage levels was primarily accounted for by the total sequence yield of each spot. These results indicate that, for a given amount of sequence data, the average coverage and distribution of coverage is highly predictable and the performance of each individual target region in different experiments is consistent.

One technical artefact of capture-sequencing procedures is the generation of duplicate DNA sequencing reads that represent the repeated sequencing of copies of the same
molecule. These duplicates generally arise when there are too few total molecules present at any stage of the technical manipulations – especially immediately prior to any PCR step. Detection of the duplicate reads by computational analysis is not trivial, and generally relies on observation of the alignment positions. Unfortunately these artefactual duplicates are difficult to distinguish from exactly overlapping reads that naturally occur within deep sequence samples.

The percentage of duplicate reads, defined as reads with the same start point and direction, was ~22% for 3Gbp of aligned data and was also highly consistent between replicates (see Table 2, libraries 2-4). As expected, when more reads were generated for the SOLiD#1 library the duplicate rate increased to ~33%. Simulations of this data showed an expected ‘natural’ (i.e. not PCR induced) duplication rate of 22 and 14% for 10Gbp and 3Gbp of data, respectively. This means that approximately 33% of our suspected duplicate reads are likely not due to PCR-induced duplication, but are those expected from the random distribution of read ends. As noted above, it is not possible to differentiate true duplicate reads from randomly occurring duplicates, and thus all such reads were removed prior to further analysis.

To test the merits of using paired-end (PE) reads versus fragment (frag) reads we generated both types of library for Illumina sequencing. Target coverage was slightly more uniform for Illumina than SOLiD, however, both sequence types provide more uniform coverage than other solution capture methods (see Supplementary figure S1 in Additional file 1). Although the majority of targets had similar normalized coverage when sequenced by SOLiD or Illumina, approximately equal numbers showed some coverage-bias depending on sequencing platform (Figure 2). Targets
with very different coverage levels typically had low complexity and/or high GC content. PE-sequencing showed both a higher percentage of reads on target and significantly fewer duplicate reads (Table 3). In PE sequencing, reads are generated from both ends of a captured DNA fragment and because the approximate fragment length is known, this information can be used to constrain the alignment of both reads to the human genome. Constraining read alignment can greatly improve accuracy when compared to fragment sequencing and we hypothesized that these inherent advantages of mapping PE vs single end reads resulted in the increased number of reads derived from the target region. We also suspected that the drastic reduction in the duplicate read rate was not because of a difference in library construction but instead the result of improved informatic identification of ‘true’ duplicates. Deep, single-end, fragment sequencing quickly saturates the target-regions, such that any additional reads will likely perfectly overlap an existing read and be identified as a duplicate, even when the reads derive from different DNA molecules. PE sequencing, in contrast, allows us to use information about both the start and the end of the capture-DNA fragment in order to determine whether the data is derived from independent DNA molecules. Thus, the increased information content of PE data allows us to reduce the misidentification of duplicate reads.

To test these theories we also analyzed the PE data as if it were generated from a single-ended fragment library. This caused the on-target alignment rate to drop slightly to 73% and the duplicate rate to nearly quadruple to 27.6%, virtually identical to the Illumina fragment library duplicate rate. The net effect of using PE data instead of fragment data was a significant increase in on-target coverage which resulted in
>90% of the targeted bases covered at 10x-fold or higher using just 2.8Gbp of data, a single 2x75bp lane of Illumina sequencing.

To assess the effect of DNA sequencing coverage depth on our ability to correctly identify variants in the exonic region of NA12812, we conducted variant discovery using both ~3.3Gbp and ~10Gbp of SOLiD capture data and 2.8 and 2.5Gbp of Illumina PE and frag data, respectively. Only a sub-fraction of this data, non-duplicate sequence reads mapped to target regions, was used for variant discovery and it is this data, not the total, that ultimately affects variant discovery quality. Discovered variants were compared to known HapMap SNPs in this sample and dbSNP. Here, the concordance to HapMap is used to measure the false negative discovery rate and the proportion of variants discovered which were also present in dbSNP129, was used to approximate the false positive discovery rate. Others have typically found ~90% of CCDS variants to be present in dbSNP for Europeans [3] and significant deviation below 90% may indicate an increased false positive discovery.

Table 4 shows the total amount of data produced for each sequencing platform, the final amount of data mapped to targets after duplicate removal and the concordance to dbSNP and HapMap. 3Gbp of SOLiD data allowed discovery of approximately 20K variants in the 36Mbp total capture region for the CCDS exome. 6Gbp of additional sequence data facilitated the discovery of approximately 20% more variants. In both cases the proportion of variants present in dbSNP129 was high. Concordance to all known HapMap variants that occur in the target region was 82 and 92% for 3 and 10Gbp of raw data, respectively. Importantly, this statistic includes known SNPs in
regions of little or no coverage where no variant call is possible. When we only considered SNPs which occur in regions with >9x coverage, concordance was ~95%.

Significantly more variants were discovered in the Illumina PE data than were found in the frag data (see Table 4) and consequently there was also higher concordance at HapMap sites. This effect is almost certainly driven by the higher coverage of the target regions achieved by having paired-end reads. As already noted, this occurs because of a slight increase in reads derived from target, but more significantly, a drastic reduction in the number of reads that are incorrectly marked as duplicates (see Figure 3). Interestingly, the proportion of variants which were also in dbSNP was significantly lower in the PE data. This may be due to overall lower read quality at the ends of the PE-reads and improved mapping of reads which would not have been aligned without a mate. Increasing the variant calling stringency (LOD=8, see Methods) on the PE data reduced the number of HapMap concordant variants slightly, but improved the percentage of variant in dbSNP by ~3% (see Table 4). Although the Illumina PE data also had higher HapMap concordance than the SOLiD variant calls, Illumina frag data performed only slightly better than SOLiD, despite having significantly more sequence data on target. When HapMap heterozygous SNP concordance was considered as a function of coverage, SOLiD data out-performed Illumina data at low, <9X, coverage (see Supplementary figure S2 in Additional file 1), however Illumina consistently obtained 2-3% higher concordance at ≥9X coverage. The quality of the variant calls in both data sets was very high with 22,066 (91.6%) of variants shared between SOLiD 10G and the Illumina-PE data sets.
This work demonstrates the practicality of genomic target enrichment using capture-sequencing in solution. For the first time, this technology is used at the scale of the whole exome, comprising over 36Mbp across >170,000K individual targets. Using 4 technical replicate libraries, we show that the average coverage of the targeted regions is highly correlated. Capture performance is also shown to be consistent, with the average coverage of each target having >98% correlation between technical replicates. Thus, it is practical to obtain consistent sequence coverage distributions and reproducible variant discovery for a variety of genomic screening experiments.

This work also shows the feasibility of using either SOLiD or Illumina-based sequencing after capture. PE data was shown to be superior to fragment data, increasing both the on-target number of reads, and greatly improves the correct identification of duplicates. Development of PE sequencing on the SOLiD platform should show a similar effect. Interestingly, Illumina sequencing consistently shows higher levels of enrichment than SOLiD sequencing. This is unexpected because both sequencing platforms yield similar coverage distributions in whole genome sequencing data [12], i.e. without enrichment. Further the capture-sequencing protocols for both methods are almost identical, therefore we suspect that differences in efficiency are due to an increase in initial library complexity arising from better annealing efficiencies of the Illumina adapter. This is probably explained by the fact that Illumina adaptor sequences contain an A/T overhang, whereas the SOLiD adapters rely on less efficient blunt-end ligation. We are currently developing an A/T overhang-type SOLiD sequencing adapter for use in capture to test whether we can improve levels of enrichment for SOLiD-based capture.
Finally, we examine the amount of sequence data required to fully interrogate the single nucleotide variants of the exome in a HapMap sample using either SOLiD or Illumina sequencing. Using ~10Gbp of SOLiD data, ~93% of all HapMap variants are discovered and over 88% of all variants are present in dbSNP. Illumina based sequencing, however, discovers 96% of HapMap variants, with ~85% of variants in dbSNP, using only 3Gbp of sequence data. This result is achieved because our Illumina protocol yields higher overall coverage on the target regions even while producing less raw sequencing data. SOLiD variant calling appears to be more sensitive at <9X coverage, typically obtaining 20% higher concordance. Overall, performance between both platforms is similar, however, and the majority of the observed difference is likely due to differences in the variant discovery pipeline software.

**Conclusions**

Sequence capture in solution is easier to automate, has higher throughput and is less expensive than microarray based techniques but has not been extensively used because of performance issues. Here we show the high reproducibility and scalability of our capture method and demonstrate that liquid capture can be used in large-scale experiments to yield reliably high levels of coverage that are consistent at a target-by-target level and are similar to microarray based techniques. Further, we establish that the entire CCDS exome can be interrogated with just 2.8Gbp of sequence data, approximately 3% of the required data for whole genome shotgun experiment. At this level of cost and scalability solution capture-sequencing becomes an attractive technique for rare variant discovery in its own right and as follow-up to genome-wide association studies, especially in studies where much of the heritability of the disease
remains unexplained [13] and thus may be due to rare mutation [14]. Capture-sequencing in solution reduces the cost and increases the throughput of rare mutation discovery by focusing on coding regions of the genome and will prove to be a significant addition to the geneticists’ tool chest.

**Materials and methods**

**Sequence Data**

All sequence data is available from the Short Read Archive with the following accession numbers [SRA012614-SRA012615].

**Sequence Alignment & Variant Discovery**

SOLiD sequence data was aligned using ABI's corona_lite package (version: 4.0r2.0) with a maximum allowed mismatch of 6, all other parameters were set at default. Pileup-style files were generated with samtools [15] and were filtered to require a variant score of at least 40, or 30 and the variant to be on both strands, and present in at least 15% of all reads. Illumina data was aligned using BWA (v 0.5.3) [16]. The base quality was recalibrated using GATK [17] (downloaded 2 October 2009). Variants were discovered with a minimum LOD of 5 (unless otherwise stated), and were filtered with the following recommended parameters: -X AlleleBalance:low=0.25,high=0.75 -X ClusteredSnps.

**Library and Capture**

The experimental procedures for preparation of pre- and post-capture libraries have been described in the Supplementary material (Additional file 1) related to this article and are available on-line for the SOLiD [18] and Illumina platforms [19]. Briefly, 5 µg genomic DNA is sheared, end-repaired and ligated with either Illumina
The library is amplified by pre-capture LM-PCR (Linker Mediated-PCR) and hybridized to NimbleGen SeqCap EZ Exome libraries. After washing, amplification by post-capture LM-PCR and a qPCR-based quality check, the successfully captured DNA is ready for sequencing.

**Probe Design**

The CCDS (Consensus Coding Sequence, build 36.2) exome capture oligonucleotide pool was designed by targeting 174,984 exons of 16,008 high-confidence protein-coding genes in CCDS. Chromosomal coordinates were obtained from the UCSC genome browser (human build hg18). Target exons were padded to a minimum length of 80bp, and consolidated to remove redundant overlaps. Coordinates for 528 human miRNA genes were obtained from miRBase (release 10), padded by 25bp on each end, and likewise consolidated. In sum, the coding and miRNA targets comprised 36 Mb of non-redundant sequence, against which 1.9 million probes were selected on the genomic forward strand, with a median probe length of 75bp and median start-to-start spacing of 34 bp. A rebalancing algorithm (described in the Supplementary material in Additional file 1) was used to improve uniformity of coverage across target exons. Probe pools were manufactured by Roche NimbleGen (Madison, WI).

**Abbreviations**

CCDS: Consensus Coding Sequence; Frag: Fragment; Gbp: Giga-base pairs; Mbp: Mega-base pairs; PE: Paired End; SD: Standard deviation; SNP: Single nucleotide polymorphism.
Authors’ contributions

MNB aided in experimental design, sequence alignment and analysis and drafted the manuscript. MW, YQW, IN, DM, CK conducted the capture hybridization and sequencing. DLB and JDJ participated in experimental design and analysis and aided in the manuscript preparation. MJR, MD’A, JK, TAR participated in designing the capture reagent. TJA, RG participated in experimental design and aided in the manuscript preparation.

Acknowledgements

The authors would like to thank Svasti Haricharan for editing the manuscript. This project was supported by Award Number U54HG003273 from the National Human Genome Research Institute.

References

1. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, Weinstock GM, Gibbs RA: Direct selection of human genomic loci by microarray hybridization. Nat Methods 2007, 4:903-905.
2. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, McCombie WR: Genome-wide in situ exon capture for selective resequencing. Nat Genet 2007, 39:1522-1527.
3. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhatricharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J: Targeted capture and massively parallel sequencing of 12 human exomes. Nature 2009, 461:272-276.
4. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME: Microarray-based genomic selection for high-throughput resequencing. Nat Methods 2007, 4:907-909.

5. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C: Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 2009, 27:182-189.

6. Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F, Gao Y, Church GM, Shendure J: Multiplex amplification of large sets of human exons. Nat Methods 2007, 4:931-936.

7. Smith DR, Quinlan AR, Peckham HE, Makowsky K, Tao W, Woolf B, Shen L, Donahue WF, Tusneem N, Stromberg MP, Steward DA, Zhang L, Ranade SS, Warner JB, Lee CC, Coleman BE, Zhang Z, McLaughlin SF, Malek JA, Sorenson JM, Blanchard AP, Chapman J, Hillman D, Chen F, Rokhsar DS, McKernan KJ, Jeffries TW, Marth GT, Richardson PM: Rapid whole-genome mutational profiling using next-generation sequencing technologies. Genome Res 2008, 18:1638-1642.

8. Bennett S: Solexa Ltd. Pharmacogenomics 2004, 5:433-438.

9. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Eyre T, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Holland R, Howe KL, Howe K, Johnson N, Jenkinson A, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, et al.: Ensembl 2008. Nucleic Acids Res 2008, 36:D707-714.

10. Consensus Coding Sequence [http://www.ncbi.nlm.nih.gov/CCDS]

11. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakaloglu A, Ozen S, Sanjad S, Nelson-Williams C, Farhi A, Mane S, Lifton RP: Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A 2009, 106:19096-19101.

12. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA, McGuire AL, Zhang F, Stankiewicz P, Halperin JJ, Yang C, Gehman C, Guo D, Irikat RK, Tom W, Fantin NJ, Muzny DM, Gibbs RA: Whole-Genome Sequencing in a Patient with Charcot-Marie-Tooth Neuropathy. N Engl J Med.
13. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 2009, 106:9362-9367.

14. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM: Finding the missing heritability of complex diseases. Nature 2009, 461:747-753.

15. Samtools [http://samtools.sourceforge.net]

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

17. GATK [http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit]

18. SOLiD Protocol [http://www.hgsc.bcm.tmc.edu/documents/Preparation_of_SOLiD_Capture_Libraries.pdf]

19. Illumina Protocol [http://www.nimblegen.com/products/seqcap/ez.html]
Figure legends

Figure 1. Normalized coverage of replicate SOLiD libraries #2-4 vs. normalized coverage of replicate library #1. Average coverage for each target region in library #1 is plotted against library#2 (blue), library#3 (green) and library#4 (red). Coverage for each target is represented as a proportion of the total sequence generated. The line X=Y indicates approximately equal levels of coverage in both libraries.

Figure 2. Average coverage for each target region as sequenced by Illumina 75bp fragment reads is plotted against SOLiD library #1. Coverage for each target is normalized for the total sequence aligned to target. The line X=Y indicates approximately equal levels of coverage in both libraries.

Figure 3 - Coverage distribution across target regions of SOLiD Libraries #1 (10Gbp) SOLiD library #2 (3Gbp) and Illumina PE and Fragment libraries. The number of bases at each level of coverage for each library type is shown for ~10Gbp of SOLiD data (green), ~3Gbp of SOLiD data (red), ~3Gbp of Illumina PE data (yellow) and ~3Gbp of Illumina frag data (blue) after duplicate removal.
# Tables

## Table 1. A comparison of different capture methodologies

| Study          | Capture Type | Reactors    | Capture Size | Sequencer Type |
|----------------|--------------|-------------|--------------|----------------|
| Ng *et al* [3] | Array        | Multiple Array | >30Mbp       | Illumina       |
| Choi *et al* [11] | Array       | Single Array | >30Mbp       | Illumina       |
| Gnirke *et al* [5] | Solution  | Single Tube  | <5Mbp        | Illumina       |
| This study     | Solution     | Single Tube  | >30Mbp       | SOLiD/ Illumina |

## Table 2. Alignment statistics for SOLiD-fragment sequencing libraries

|                       | SOLiD#1        | SOLiD#2        | SOLiD#3        | SOLiD#4        |
|-----------------------|----------------|----------------|----------------|----------------|
| Total Reads Aligned   | 199,704,874    | 53,558,797     | 47,640,266     | 46,687,848     |
| Total Data Aligned (Gbp) | 9.99          | 2.68           | 2.38           | 2.33           |
| Reads on Target (%)   | 50.1           | 50.9           | 49.5           | 48.0           |
| Duplicate Reads (%)   | 32.8           | 22.1           | 22.5           | 20.9           |
| Mean Coverage (x)*    | 43             | 22             | 20             | 19             |
| Median Coverage(x)*   | 42             | 19             | 16             | 15             |
| Targets Hit (%)       | 99.3           | 98.44          | 98.4           | 98.1           |
| Bases ≥1X Coverage (%)| 97.6           | 94.31          | 93.5           | 92.5           |
| Bases ≥10X Coverage (%)| 89.4           | 70.8           | 65.9           | 64.1           |
| Bases ≥20X Coverage (%)| 78.9           | 48.2           | 42.0           | 40.5           |

* calculated after duplicate read removal.

## Table 3. Alignment statistics for Illumina-PE and fragment sequencing libraries

|                       | IlluminaFrag   | IlluminaPE     |
|-----------------------|----------------|----------------|
| Total Reads Aligned   | 33,524,973     | 37,832,835     |
| Total Data Aligned (Gbp) | 2.51          | 2.84           |
| Reads on Target (%)   | 67.62          | 78.0           |
| Duplicate Reads (%)   | 30.97          | 8.3            |
| Mean Coverage (x)*    | 24             | 52             |
| Median Coverage(x)*   | 20             | 40             |
| Targets Hit (%)       | 99.36          | 99.6           |
| Bases ≥1X Coverage (%) | 96.39 | 98.9 |
|------------------------|-------|------|
| Bases ≥10X Coverage (%)* | 71.33 | 90.8 |
| Bases ≥20X Coverage (%)* | 51.23 | 76.9 |

* calculated after duplicate read removal.

Table 4. Variant discovery and HapMap concordance for different sequencing types and varying amounts of sequence data

|                        | Illumina Frag | Illumina PE | Illumina PE (High Strigency) | SOLiD#1 | SOLiD#1 |
|------------------------|---------------|-------------|-----------------------------|---------|---------|
| Bases produced (Gbp)   | 2.51          | 2.84        | 3.4                         | 9.99    |         |
| Bases on target after duplicate removal (Gbp) | 1.04 | 2.01 | 0.59 | 1.72 |
| Total SNPs             | 21,239        | 27,953      | 26,489                      | 19,790  | 24,077  |
| dbSNP SNPs             | 19,525        | 23,745      | 23,133                      | 18,016  | 21,350  |
| dbSNP (%)              | 91.9          | 84.95       | 87.3                        | 91.04   | 88.67   |
| HapMap Variant Concordance (%) | 83.0 | 96.0 | 95.8 | 81.6 | 92.9 |
| Variant Concordance (>9x coverage) (%) | 95.5 | 98.5 | 98.2 | 94.5 | 97.2 |

Additional files

Additional file 1. Supplementary table, figures and materials and methods.

Supplementary table S1 and figures S1 & S2 as well as detailed materials and methods.
Figure 1
Figure 3
Additional files provided with this submission:

Additional file 1: SUPP1.doc, 236K
http://genomebiology.com/imedia/1589081824081059/supp1.doc