SCIENCE CHINA
Life Sciences

• RESEARCH PAPERS •

October 2011 Vol.54 No.10: 945–952
doi: 10.1007/s11427-011-4232-4

High-performance single-chip exon capture allows accurate whole exome sequencing using the Illumina Genome Analyzer

JIANG Tao[1†], YANG Lei[2,3†], JIANG Hui[1], TIAN Geng[1,2,3] & ZHANG XiuQing[1,2,3*]

[1]*Beijing Genomics Institute, Shenzhen 518000, China;*
[2]*Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 101300, China;*
[3]*Graduate University of Chinese Academy of Sciences, Beijing 100062, China*

Received March 16, 2011; accepted August 16, 2011

Here we present an adaptation of NimbleGen 2.1M-probe array sequence capture for whole exome sequencing using the Illumina Genome Analyzer (GA) platform. The protocol involves two-stage library construction. The specificity of exome enrichment was approximately 80% with 95.6% even coverage of the 34 Mb target region at an average sequencing depth of 33-fold. Comparison of our results with whole genome shot-gun resequencing results showed that the exome SNP calls gave only 0.97% false positive and 6.27% false negative variants. Our protocol is also well suited for use with whole genome amplified DNA. The results presented here indicate that there is a promising future for large-scale population genomics and medical studies using a whole exome sequencing approach.

whole exome sequencing, exon capture, SNP detection, indel detection, high-throughput resequencing

Citation: Jiang T, Yang L, Jiang H, *et al*. High-performance single-chip exon capture allows accurate whole exome sequencing using the Illumina Genome Analyzer. Sci China Life Sci, 2011, 54: 945–952, doi: 10.1007/s11427-011-4232-4

Recently developed massively parallel sequencing technologies offer several orders of magnitude better throughput by producing a large number of short sequencing reads at low cost [1]. In the past two years, several individual human genomes have been resolved on these platforms, providing information on millions of genetic variations [2–4]. However, the cost of resequencing an entire human genome is still too high for many research projects, especially for those that require the batch sequencing of thousands of individuals to determine genetic variations associated with various disease traits. Because most disease causing variants are found in the protein-coding regions of the genome that correspond to less than 2% of the human genome, the development of cost-effective target resequencing approaches for these high-value regions would dramatically reduce the cost and allow sequencing-based research on thousands of samples.

Recently developed exon-capture technologies efficiently enrich targeted regions by hybrid selection. However, there are still severe limitations to their applications, for example, molecular inversion probes miss ~80% of the targeted exons [5]. Biotinylated RNA baits worked well on a set of 1900 human genes, but scalability of this approach to cover all human exons has not yet been demonstrated [6]. Several chip-based exon capture methods have also been tried, but only a small portion of the human exons were probed on a single chip [1,7,8].

We used a single NimbleGen high-density 2.1M-probe array chip to capture the whole human exome and performed accurate sequencing using the Illumina Genome Analyzer. We evaluated our method on an Asian individual whose genome was fully resolved by shot-gun resequencing [3]. The complete and even coverage of the exome and the high accu-

†Contributed equally to this work
*Corresponding author (email: zhangxq@genomics.org.cn)

racy of variation detection that the method achieved demonstrated the usefulness of our protocol and its potential for use in large-scale studies.

1 Materials and methods

1.1 Library construction and sequencing

To adapt whole exome capture to the Illumina sequencing platform, we used a two-step library construction process (Figure 1). In step 1, 20 μg DNA was nebulized into ~500 bp fragments. With linkers for PCR amplification ligated, 5 μg single-stranded fragments were hybridized to a NimbleGen 2.1M-probe exome capture array following the manufacturer's protocol. Non-hybridized fragments were washed out and the probe-hybridized fragments were eluted from the chip. The resulting material was subsequently amplified by ligation-mediated PCR to obtain a primary library with ~500 bp inserts that could be subjected to direct end sequencing on a Roche/454 platform. In step 2, we concatenated the captured fragments by DNA ligase and re-sheared them to short fragments of ~200 bp that were ligated with Illumina sequencing adaptors. Thus, the primary library was transformed to a secondary shot-gun library with the optimal insert size for the Illumina GA-II platform and which could be subjected to the standard Illumina DNA sequencing process. Finally, the exon-enriched shotgun library was sequenced on the Illumina GA-II platform, following the manufacturer's protocols and using the standard sequencing primers. Image analysis and base calling was performed by the Genome Analyser Pipeline version 1.3 with default parameters.

1.2 Public data used

Target DNA fragments from the captured regions of NimbleGen 2.1M-probe Human Exome Array (http://www.nimblegen.com/downloads/annotation/seqcap_exome/index.html) were used. The exon information was downloaded from the Consensus CDS (CCDS) database (version 20080902, ftp://ftp.ncbi.nlm.nih.gov/pub/CCDS/). The consensus YH genome (the first Asian diploid genome) is from http://yh.genomics.org.cn/. The human reference genome sequence (NCBI build 36.1, UCSC hg18) used in this study was downloaded from the UCSC database (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/).

1.3 Read mapping

We used SOAPaligner (v2.01) to align the raw sequencing reads onto the reference genome with at most two mismatches. The parameters were set as: –a –D –o –r 1 –t –c –f 4. Only the uniquely mapped reads were extracted for

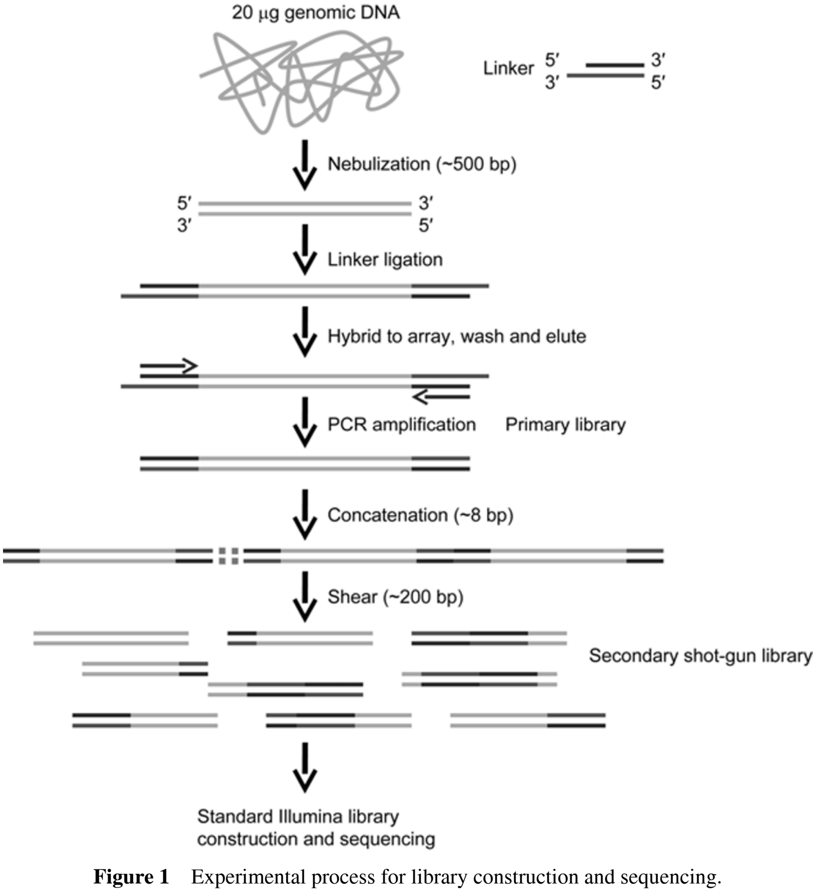

Figure 1 Experimental process for library construction and sequencing.

subsequent analysis and SNP calling.

1.4 SNP calling and accuracy evaluation

Based on the SOAP alignments, SOAPsnp software was used to assemble the consensus sequences and to call genotypes in the target regions. The parameters were set as: –i –d –o –r 0.00005 –e 0.0001 –M –t –u –L –s –2 –T. The YH genome sequence and SNPs were used to evaluate the accuracy of genotype calling based on exome capture sequencing. The raw unaligned Illumina sequences of the YH exome are available at http://yh.genomics.org.cn.

1.5 Indels detection

To call indels, we first used the GATK pipeline. The sequence reads were re-aligned to the hg18 reference genome with BWA [9], and GATK base quality score recalibration, indel realignment, duplicate removal, and indels discovery and genotyping for YH exome were simultaneously applied using the standard hard filtering parameters or the variant quality score recalibration. We then used the Dindel pipeline. Candidate indels were first extracted from the BWA alignment and realignment windows were defined. For every realignment window, candidate haplotypes that represent an alternative sequence to the reference were generated. The reads were then realigned to the candidate haplotypes and finally the posterior probability of candidate indels was estimated using a Bayesian method. For both the indel calling methods, a high-confident indel was called if it had at least three reads to support it.

2 Results

2.1 Whole exome sequence capture for Illumina sequencing

We used a NimbleGen 2.1M-probe array capturing chip with custom-synthesized 60–90-mers as capture probes to enrich exome DNA from the YH individual. The designed 2.1M hybridization probes covered 34 Mb target regions, including 180640 coding exons (Table 1) of 18654 well-annotated CCDS genes (http://www.ncbi.nlm.nih.gov/CCDS) and their 200 bp flanking sequences, as well as 698 microRNA genes. Although some genes that are generally repetitive and hard to hybridize were not covered, this high-density array chip tiled 92.8% of the well-annotated genes, providing a near-complete whole exome.

We performed the steps (Figure 1) of the exome capturing protocol on the YH genomic DNA [3]. We obtained 67.95 million reads (5.10 Giga bases of sequence) with an average read size of 75 bp. As hybridized linkers and sequencing adaptors were introduced during library construction and were also sequenced, we first searched for possible linker and adaptor sequences using a Smith-Waterman algorithm (Figure 2). Substrings of reads that could be matched to linker or adaptor sequences with 12 bp long alignments and allowing up to two mismatches were identified as linker/adaptor contamination. The reads were then truncated and used in further analysis. This filtering step greatly improved the data quality and the effective sequencing depth for the target region (Table 2). 3.38 Gb reads out of 4.82 Gb linker/adaptor masked sequences were uniquely aligned to the complete reference genome using SOAP software [10]. Among the uniquely aligned data (Table 3), 1.14 Gb (34%) unambiguously mapped to the targeted exonic regions and 1.57 Gb (46%) mapped to the extended 500 bp flanking regions. Of the 180640 targeted exons, 175181 (97%) were represented by at least one uniquely mapped read. The average sequencing depth of the target regions and the extended flanking regions were 33

Table 1 Composition of target regions designed for the NimbleGen 2.1M-probe exome capturing array[a)]

Category	# of elements	Length (bp)	Percent (%)
Promoter region	3062	50063	0.15
5′-UTR	14604	436388	1.28
Coding exons	180640	23939542	70.19
Introns	149008	8771941	25.72
3′-UTR	14505	477440	1.40
MicroRNA	545	46776	0.14
CCDS genes	18654	33720506	98.86
Intergenic regions	–	388304	1.14

a) The total length of the target region was 34108810 bp which is smaller than the sum of the categories because some elements may overlap.

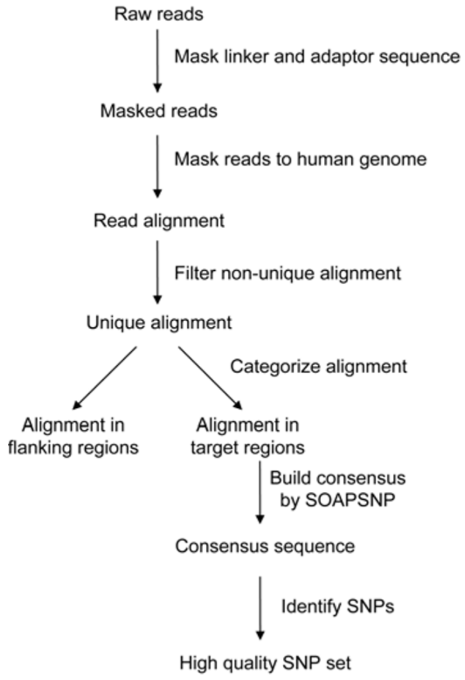
Raw reads
Mask linker and adaptor sequence
Masked reads
Mask reads to human genome
Read alignment
Filter non-unique alignment
Unique alignment
Categorize alignment
Alignment in flanking regions
Alignment in target regions
Build consensus by SOAPSNP
Consensus sequence
Identify SNPs
High quality SNP set

Figure 2 Bioinformatics analysis pipeline.

Table 2 Comparison between sequence alignment results with or without masking linker and adaptor sequences[a)]

	Without linker/ adaptor masked	With linker/adaptor masked
Raw data amount (Mb)	5096	4818
Aligned reads (Mb)	3178	3707
Uniquely aligned (Mb)	2897	3379
Uniquely aligned on TR (Mb)	1001	1136
Average depth of TR	29.3×	33.3×

a) About 14% more data were aligned to the genome after masking linkers and adaptors even though there was less raw data input.

and 13 folds, respectively (Table 3). SNP identification was performed on the aligned target regions using SOAPsnp [11] (section 1.4).

2.2 Evaluation of whole exome capturing performance

We evaluated our whole exome capturing experiment based on three parameters: specificity, uniformity, and reproducibility.

Specificity refers to the fraction of sequences that were derived from the targeted regions by specific selection rather than through systematic errors or stochastic effects. Here, we considered only the percentage of uniquely aligned reads that specifically mapped on or near selected exons as the indicator of capture-specificity. As shown in Table 3, about 2.70 Gb (80%) of the 3.38 Gbp uniquely aligned read data were on targeted exons or on their flanking regions. Based on this result, we estimated the specificity of our protocol to be 80%. Normalized base depth profiling showed a dramatic breakdown of depth when it was extended beyond the target regions' boundaries, further indicating the high specificity of the capture (Figure 3). Our results suggest that the specificity of exon-capture for the whole exome was similar to that reported earlier in an experiment on a small subset of selected human exons [6].

Uniformity refers to the randomness of sampling by the whole exome capturing and sequencing process. This is critical to the unbiased discovery of variations. Our procedure achieved as high as 95.6% unique coverage in target regions and 86.2% in extended regions (Table 3). By bootstrapping subsets of the data aligned to target regions, the unique coverage already exceeded 90% at a sequencing depth of only 6-fold. Only a marginal improvement of coverage was achieved at higher sequencing depths (Figure 4). These results suggest that the exome was randomly sampled in general and that our sequencing almost reached the practical upper limit of coverage of the whole exome.

Table 3 Summary of the GA-II sequence alignment[a)]

	Target region	Flanking region	Genome background	Total
Genomics length (Mb)	34	117	2707	2858
Uniquely aligned (Mb)	1136	1566	677	3379
Average depth	33.3	13.4	0.25	–
Coverage (%)	95.6	86.2	9.5	–

a) The average depth of the target regions was 33 folds, 2.5 times greater than that of the flanking regions. This is consistent with the principle that fragments with more bases hybridized to oligonucleotide probes are preferentially selected relative to fragments with fewer bases hybridized. The 0.25× genome background DNA, which was not a result of specific hybridization, was also sequenced.

Of the 4.5% target regions that were not covered, 2% could be covered by reads with multiple placements (repeat hits). The remaining 2.5% were probably inaccessible regions that are hard to capture by hybridization. Subsequent analysis showed that these regions generally had an extreme G+C content, and mostly contained sequences with a G+C content higher than 60% (Figure 5). We also checked the per-base depth distribution of the covered region to examine the evenness of coverage (Figure 6). The distribution approximately followed the expected theoretical model (Poisson distribution when sampling is completely random) with five-times larger standard deviation. Our analysis showed that the variation in coverage was mainly associated with GC content (Figure 7). Target regions with moderate GC

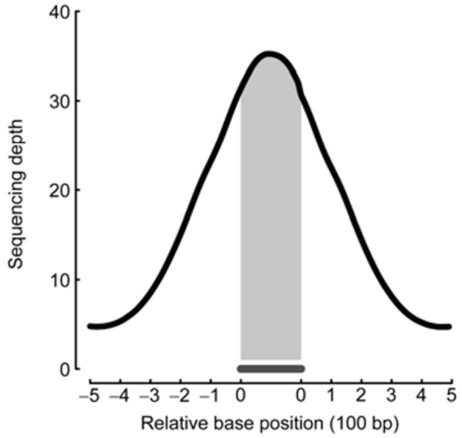

Figure 3 Distribution of aligned data relative to the target region. The target region (the gray area with an underline) had the highest depth (~33×). The 500 bp flanking regions were also enriched with aligned data, and the depth decayed when extended into the non-target regions.

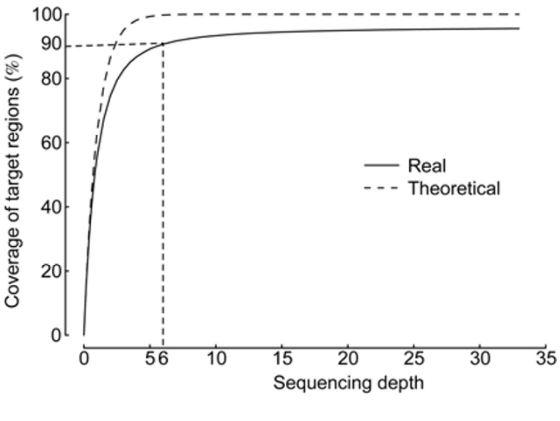

Figure 4 Correlation between sequencing depth and coverage of the target region obtained by exome capture.

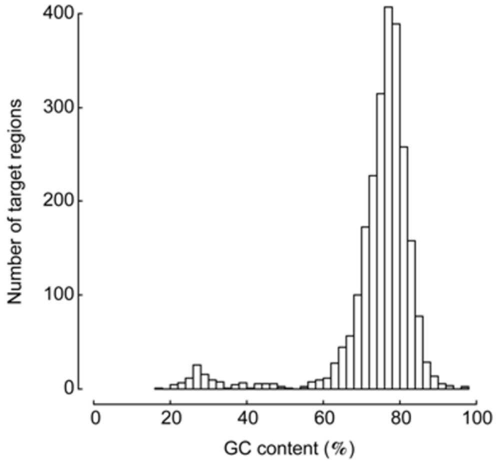

Figure 5 Distribution of GC content of the target regions that failed in the capturing process.

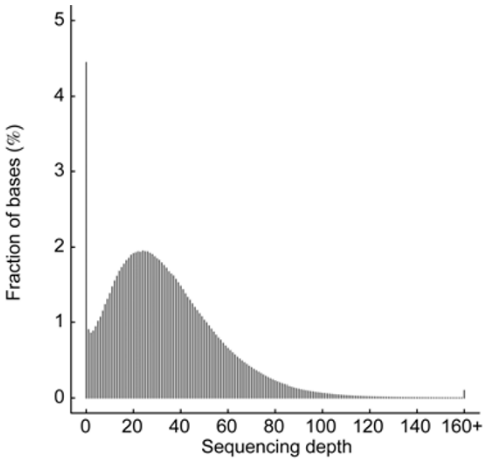

Figure 6 The per-base depth distribution of the target region. Despite the failing part, the distribution roughly followed a single-peak distribution in agreement with the theoretical Poisson model.

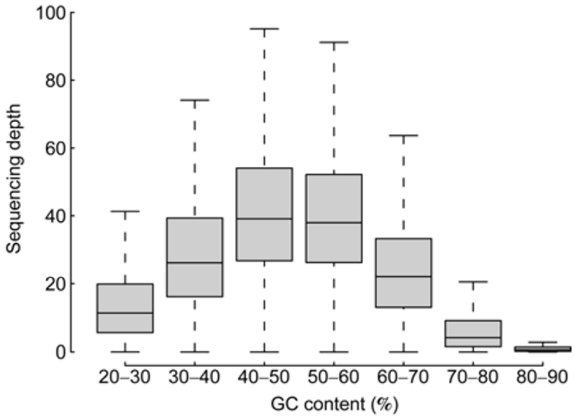

Figure 7 Average sequencing depth of target regions for different levels of GC content. The box denotes the range between the first and third quartile. The target regions with GC content in the 40%–60% range had the most depth; both low-GC and high-GC target regions had less sequencing depth.

content (40%–60%) had higher coverage rates and regions with high GC or low GC content were poorly covered. This result might be explained by the fact that the hybridization and PCR processes in sequence capturing and sequencing were optimized for modest GC content, leading to the difficulty of retrieving sequences with high or low GC content. However, the chosen experimental conditions were a reasonable trade-off that successfully covered most parts of the exome. Thus, 76.8% of bases in target regions had a sequencing depth of >16-fold (half of the average depth); a better uniformity than that obtained in a previous study on a limited number of captured exons [6].

Reproducibility refers to whether or not the protocol can be replicated and comparably applied to other samples, a crucial requirement in large-scale applications with multiple samples. To assess reproducibility, we carried out another independent whole-exome capture and generated approximately the same amount of data for another sample genome (named EMC). For both the YH and EMC samples, the depths of the target regions were normalized to the average depths of the entire set of target regions. The correlation coefficient of the normalized depth between the two samples was 0.7 and the log ratio of the normalized depth was distributed around 0, which demonstrated that our whole exome capturing method was well replicated in the different sample (Figure 8).

2.3 Variation detection by exome sequencing

In our pioneering research on the resequencing of whole human genomes, we confirmed that SNP identification using the Illumina Genome Analyzer has high accuracy and consistency rates, and showed that its concordance rate with alleles found by array-based genotyping was higher than 99.9% [3]. These results provided an unprecedented opportunity to assess the accuracy of SNP detection by exome-capture at the whole genome level. To this end, we built the consensus sequence of the YH exome from the exome-capture sequencing data by SOAPsnp [11] using the

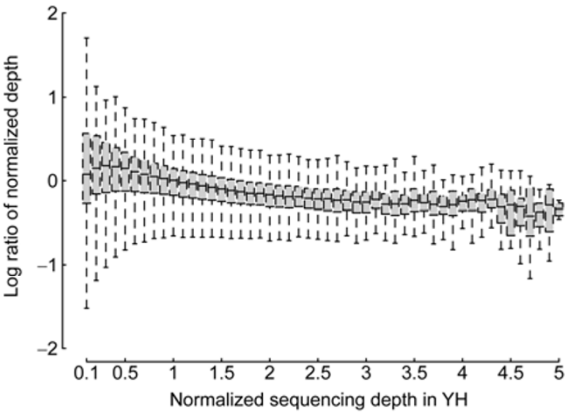

Figure 8 Reproducibility of normalized depth distribution between YH and EMC, an independent exome capture sample.

Table 4 Comparison of the YH exome with the YH genome[a)]

	Allele type	YH Exome			Total	Rate of difference
		HOM ref.	HOM mut.	HET		
YH genome	HOM ref.	32329370	6	189	32329565	6.03×10^{-6}
	HOM mut.	29	8413	11	8453	4.70×10^{-3}
	HET	1294	39	11314	12647	1.05×10^{-1}

a) HOM ref. indicates genotypes that were the same as in the hg18 human reference genome; HOM mut. indicates a homozygous difference from hg18; HET is for the heterozygote.

same parameters that were used in the YH genome resequencing project. A total of 32387638 (95%) of the exome consensus genotype calls passed the Q20 threshold, 32350665 of them also overlapped with the Q20 part of YH genome consensus sequence. Only 1568 (4.8×10^{-5}) discrepant base pairs were detected between the consensus calls of the YH genome and the YH exome (Table 4). The high consistency between the two sets of consensus calls indicated accurate genotype calling in the YH exome analysis and confirmed that our method only introduced minor errors.

A total of 19972 potential high-quality SNPs were detected by comparing the YH exome consensus with the hg18 human reference. Taking the YH genome consensus calls as the gold standard, 195 of the YH exome SNPs were overcalls and 1323 of the YH genome SNPs were undercalls, indicating that the overall SNP error rate was 0.97% false positives and 6.27% false negatives in the Q20 consensus. Thus, very few false positives were introduced by this method. However, the false negative rate for the exon-capture method was a little higher. Further optimization of the method should help improve this. The problem might simply be that because the hybridization probes were designed based on the reference sequence they preferentially capture the reference alleles. In some regions that contained a high number of SNPs compared to the reference sequence hybridization by the capture probes may have failed and, thus, were missing in the sequencing.

To determine the sequencing depth effect on SNPs coverage and accuracy, we randomly extracted subsets of reads at different depths, and detected SNPs using the same methods. The increase of sequence depth should enhance the power of SNP detection. Indeed, a steady increase of SNP-calling accuracy was observed as sequence depth increased. At depths of about 11×, we found 65% of the high quality YH SNPs with an accuracy rate of about 91%. The overall sensitivity of SNP detection presented here (65%) is a little higher than that reported by Choi *et al.* (50%) [12] where the lower accuracy was largely because of a high false negative rate. However, at SNP positions shared by the YH exome and the YH genome, the concordance rate was as high as 99.65%, indicating that the sensitivity and accuracy of SNP detection were limited only by sequencing depth.

We then called indels for the YH exome datasets using both the GATK [13] and Dindel [14] packages. In summary, 1128 indels including 98 coding indels were detected by GATK, while 1130 indels including 167 coding indels were identified using Dindel (Table 5). The difference in the number of the indels detected can be attributed to the different algorithms used in the two softwares.

2.4 Evaluation of exome capturing on whole genome amplified samples

The amount of input DNA required for whole exome capturing and sequencing (20 μg) may restrict the application of our method on medical samples with small amounts of available DNA. Recent studies have successfully tested sequence capturing on whole genome amplified DNA samples [7]. We also applied whole genome amplification (WGA) to YH DNA and tested our protocol on the WGA DNA. The WGA-captured DNA was sequenced to an average depth of approximately 12-fold in the target regions. The coverage was 86.4%, which is lower than it was without WGA at the same sequencing depth (93.8%).

The per-base depth distribution also showed less uniformity than for the sample without WGA because the distribution shows poor fit to the theoretical model curve (Figure 9). The SNP error rate of WGA consensus calls is 1.73% false positives and 9.90% false negatives at the Q20 level, indicating a larger bias in sampling from the WGA exome. Nevertheless, this result is comparable with a previous study [7] and this method could still provide meaningful results for medical studies when only small amounts of genomic DNA are available.

Table 5 Indel identification in the YH exome

Method	GATK	Dindel
Total number of indels	1128	1130
Insertion-coding	29	46
Deletion-coding	69	121
Splice site	42	43
Intron	916	848
5'-UTRs	35	24
3'-UTRs	35	45
Intergenic	2	3
Total insertion	438	433
Total deletion	690	697
Heterozygous indels	669	716
Homozygous indels	459	414

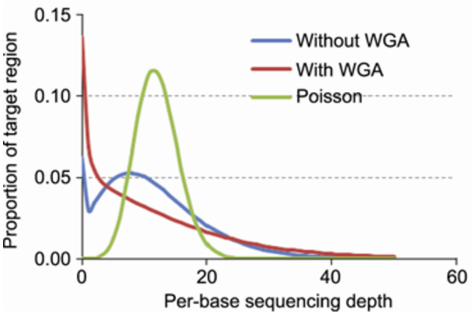

Figure 9 Distribution of per-base sequencing depth in exome capturing and sequencing of the WGA sample. At an average depth of 12-fold, 13.4% of the target regions were still not covered by any reads. In the non-WGA sample this fraction was only 6.2%.

3 Discussion

To significantly reduce the cost of using new generation sequencing technology for personal genomics and large sample genome wide association (GWA) studies, targeted resequencing of the whole exome enriched by probe hybridization is one of the methods that have been used. In this study, we have performed deep-sequencing and extensive analysis of captured whole-exome regions from an individual whose genome has already been very well resequenced. Thus, we could systemically evaluate the efficiency of exon-capture and the potential sequencing errors that might be induced during the experimental and sequencing processes. We showed that the captured sequences were specifically enriched in exon regions, and produced only a few unexpected sequences at the flanking regions. Although high-GC content regions might be missed during capture, the sequence reads were produced at high sequence depth for most of the targeted regions and were evenly distributed on the reference genome, thereby creating high-accuracy at the single base level. Moreover, this method reproduced the results at the whole genome level for a different sample, illustrating the possibility of using this protocol universally.

Most previous GWA studies were carried out using microarrays which were designed based on the current knowledge from SNP databases such as dbSNP [15]. Our high resolution resequencing of the YH genome has identified thousands of novel SNPs that were absent in dbSNP [3]. This result suggests that many undiscovered SNPs may still exist in different populations. One of the significant limitations of using microarrays in GWA studies is that rare SNPs, which may play important functional roles, are ignored in the analyses. Our whole genome survey of the SNP calling accuracy of the present method supported the possibility of using exon-capture in target resequencing of the whole exon regions in GWA studies. Although a weak GC bias introduced by the hybridization process should be remedied in future applications, the high coverage of target regions and the high accuracy of the detected SNPs suggested that this bias will not significantly affect the further analysis. An important issue in GWA studies is the limitations of the amount of sample that can be collected. We showed that whole genome amplification could help by drastically reducing the need for large amounts of sample DNA. However, because of the amplification bias in whole genome amplification and the error introduced by the amplification enzyme, whole genome amplification introduces a high rate of false positives and false negatives.

The exon-capture method for whole exome resequencing has been successfully applied to identify causative genes for rare Mendelian disorders [16–19] and to make a genetic diagnosis [12]. We earlier reported whole exome resequencing using the capture method here of 50 Tibetan individuals that revealed adaptation to high altitude [20]. We have also completely resequenced 200 human exomes and identified an excess of low frequency non-synonymous coding variants, a large proportion of which were predicted to be rare deleterious mutations [21]. We now plan to use the same strategy, a combination of exon-capture and new generation sequencing technology, to elucidate the genetic risk of metabolic diseases. In the present study, we have provided clear evidence that this strategy is applicable to large scale whole exome resequencing projects. In particular, we have confirmed that relevant SNP sets with high accuracy can be obtained using this protocol. This study should motivate the further application of this method to disease association studies.

This work was supported by the Chinese Academy of Sciences (Grant Nos. GJHZ0701-6 and KSCX-YWN-023), the National Natural Science Foundation of China (Grant Nos. 30725008, 90403130, 90608010, 30221004, 90612019 and 30392130), the National Basic Research Program of China (Grant Nos. 2007CB815701, 2007CB815703 and 2007CB815705), the Danish Platform for Integrative Biology, and the Lundbeck Foundation Centre of Applied Medical Genomics for Personalized Disease Prediction, Prevention and Care (LUCAMP). Lars Bolund, Li QiBin, and Hu YuJie edited the manuscript.

1 Albert T J, Molla M N, Muzny D M, *et al*. Direct selection of human genomic loci by microarray hybridization. Nat Methods, 2007, 4: 903–905

2 Levy S, Sutton G, Ng P C, *et al*. The diploid genome sequence of an individual human. PLoS Biol, 2007, 5: e254

3 Wang J, Wang W, Li R, *et al*. The diploid genome sequence of an Asian individual. Nature, 2008, 456: 60–65

4 Wheeler D A, Srinivasan M, Egholm M, *et al*. The complete genome of an individual by massively parallel DNA sequencing. Nature, 2008, 452: 872–876

5 Porreca G J, Zhang K, Li J B, *et al*. Multiplex amplification of large sets of human exons. Nat Methods, 2007, 4: 931–936

6 Gnirke A, Melnikov A, Maguire J, *et al*. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol, 2009, 27: 182–189

7 Hodges E, Xuan Z, Balija V, *et al*. Genome-wide *in situ* exon capture for selective resequencing. Nat Genet, 2007, 39: 1522–1527

8 Okou D T, Steinberg K M, Middle C, *et al*. Microarray-based genomic selection for high-throughput resequencing. Nat Methods, 2007, 4: 907–909

9 Li H, Durbin R. Fast and accurate short read alignment with Bur-

rows-Wheeler transform. Bioinformatics, 2009, 25: 1754–1760

10 Li R, Yu C, Li Y, *et al*. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics, 2009, 25: 1966–1967

11 Li R, Li Y, Fang X, *et al*. SNP detection for massively parallel whole-genome resequencing. Genome Res, 2009, 19: 1124–1132

12 Choi M, Scholl U I, Ji W, *et al*. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci USA, 2009, 106: 19096–19101

13 McKenna A, Hanna M, Banks E, *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res, 2010, 20: 1297–1303

14 Albers C A, Lunter G, Macarthur D G, *et al*. Dindel: Accurate indel calls from short-read data. Genome Res, 2010, 21: 961–973

15 Frazer K A, Murray S S, Schork N J, *et al*. Human genetic variation and its contribution to complex traits. Nat Rev, 2009, 10: 241–251

16 Ng S B, Turner E H, Robertson P D, *et al*. Targeted capture and massively parallel sequencing of 12 human exomes. Nature, 2009, 461: 272–276

17 Ng S B, Bigham A W, Buckingham K J, *et al*. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet, 2010, 42: 790–793

18 Ng S B, Buckingham K J, Lee C, *et al*. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet, 2010, 42: 30–35

19 Bilguvar K, Ozturk A K, Louvi A, *et al*. Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. Nature, 2010, 467: 207–210

20 Yi X, Liang Y, Huerta-Sanchez E, *et al*. Sequencing of 50 human exomes reveals adaptation to high altitude. Science, 2010, 329: 75–78

21 Li Y, Vinckenbosch N, Tian G, *et al*. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. Nat Genetics, 2010, 42: 969–972