Bringing cost and process efficiency to next generation sequencing

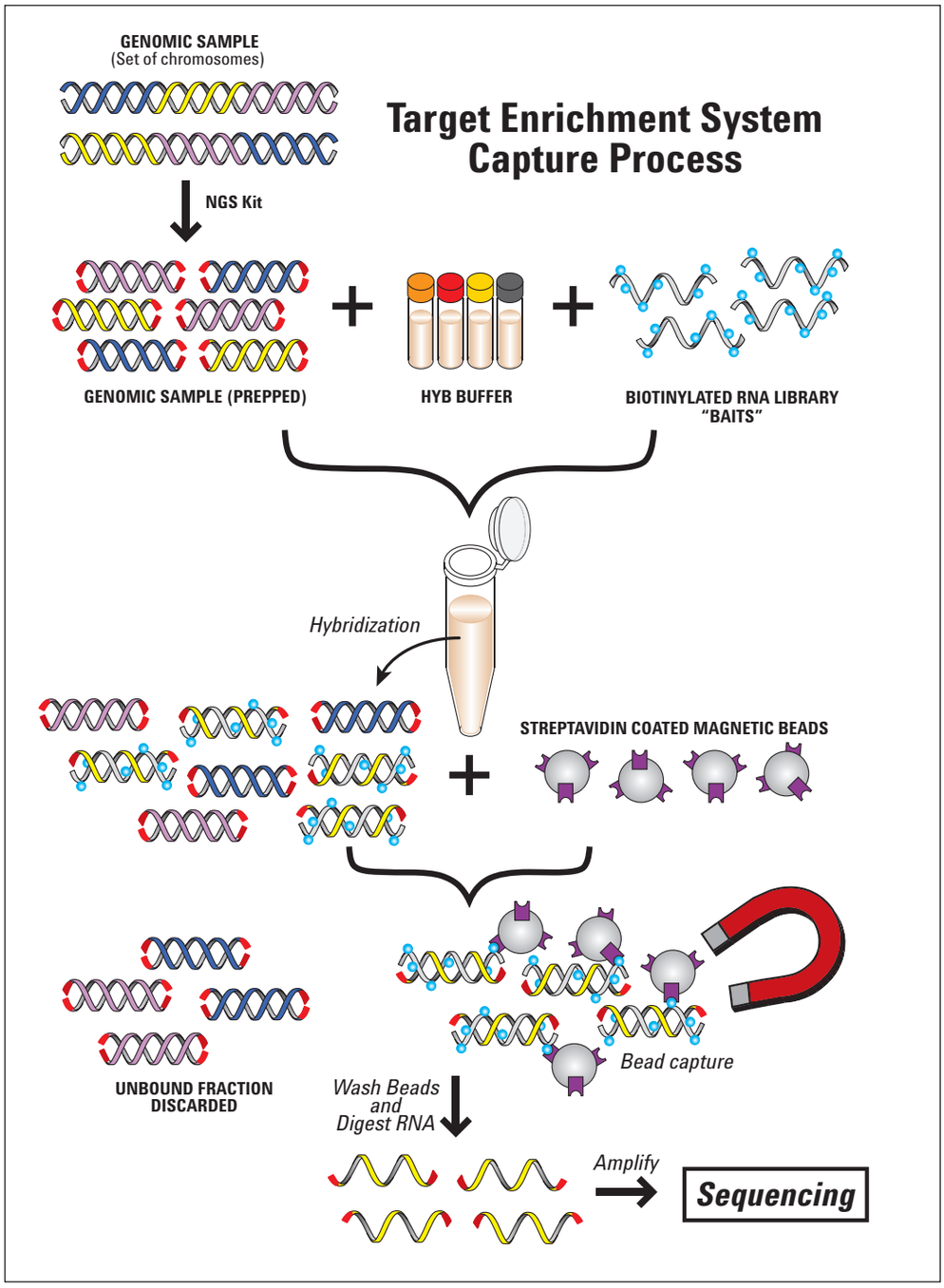
While next-generation sequencing has revolutionised the way genomes are sequenced, this technology possesses a fundamental weakness – the inability to easily target specific regions of a genome. To address this, a method has been developed that uses biotinylated RNA 'baits' to fish targets out of a 'pond' of DNA fragments. The RNA is transcribed from PCR-amplified oligodeoxynucleotides originally synthesised on a microarray, generating sufficient bait for multiple captures at concentrations high enough to drive the hybridisation. This system uses an extremely efficient hybrid selection technique enabling a larger number of samples to be sequenced in a given study. With sample input requirements at or below 3µg of genomic DNA, even the most precious of samples can be utilised for massively-parallel sequencing without risk of depletion. The method can also easily be incorporated into an automated environment, further increasing process efficiencies, while minimising total sample costs.

By Dr Fred P. Ernani, and Dr Emily M. LeProust

Individual human genomes have recently been sequenced on next-generation instruments[1-4], in large part, due to the development and commercialisation of increasingly powerful instrumentation[1,5-7] and significantly reduced cost of sequencing data. Significant economies could be realised in targeting the protein coding fraction, the 'exome', which represents a mere ~1% of the human genome. For many key resequencing targets such as genomic regions implicated by whole genome association scans, as well as exons of sets of protein-coding genes implicated in specific diseases, the savings could be even greater. Highly efficient and cost-effective target of specific fractions of the genome could significantly reduce sequencing costs, thus enabling achievement of research and diagnostic goals.

The development of methods for the massively parallel enrichment of the templates that need to be sequenced will be critical to the sequencing of

Figure 1
Target enrichment system workflow



such targeted regions. Traditional multi- and singlepex methods are not well suited to this purpose, thus several groups have developed 'genome-partitioning' methods for preparing complex mixtures of sequencing templates that are highly enriched for targets of interest[8-15]. Only two of these methods have been tested on target sets complex enough to match the scale of current next-generation sequencing instruments. The first method, employing microarray capture[9,12,13], uses hybridisation to arrays of synthetic oligonucleotides that match the target sequence in order to capture templates that are randomly sheared (adaptor-ligated genomic DNA) and has been applied to more than 200,000 coding exons[12]. Microarray capture-

based methods work best for DNA fragments of more than 500 bases in length, which limits the enrichment and sequencing efficiency to very short dispersed targets, such as human protein-coding exons that have a median size of 12 base pairs (bp)[16]. The second method employs multiplex amplification[14]. It uses oligonucleotides that are synthesised on a microarray, subsequently cleaved off and amplified by PCR, to perform a padlock and molecular inversion reaction[17,18]. This is carried out in solution where the probes are extended and circularised to copy, rather than directly capture, the targets. Such decoupling of synthesis and reaction formats enables reuse, as well as quality control, of single lots of oligonucleotide probes. That said, this method is not well optimised for this type of application, and as recently reported[14], multiplex amplification missed more than 80% of the targeted exons in single reactions. This was, in large part, attributed to poor reproducibility between replicates, uneven recovery of alleles and representation of sequencing targets. Recent studies have suggested ways in which these points could be addressed[15].

A method has been developed that uses biotinylated RNA 'baits' to fish targets out of a 'pond' of DNA fragments[19]. It overcomes some of the weaknesses of previous methods. As described below, it combines the simplicity and robust performance of oligonucleotide hybridisation with the advantages of amplifying array-synthesised oligonucleotides and performing the selection reaction in solution. This approach is more amenable to automation, and it can be scaled to meet the needs of larger sequencing projects, a limitation inherent in other commercially available methods of target enrichment.

RNA-driven DNA capture

The target enrichment method involves oligonucleotides, or 'baits', that are biotinylated for easy capture on to streptavidin-labelled magnetic beads, as well as buffers and blocking agents necessary for performing the capture process (**Figure 1**). To perform the capture, genomic DNA is sheared and assembled into a library format specific to the sequencing instrument utilised downstream. Size selection is performed on the library prior to capture and confirmed by a method such as electrophoresis.

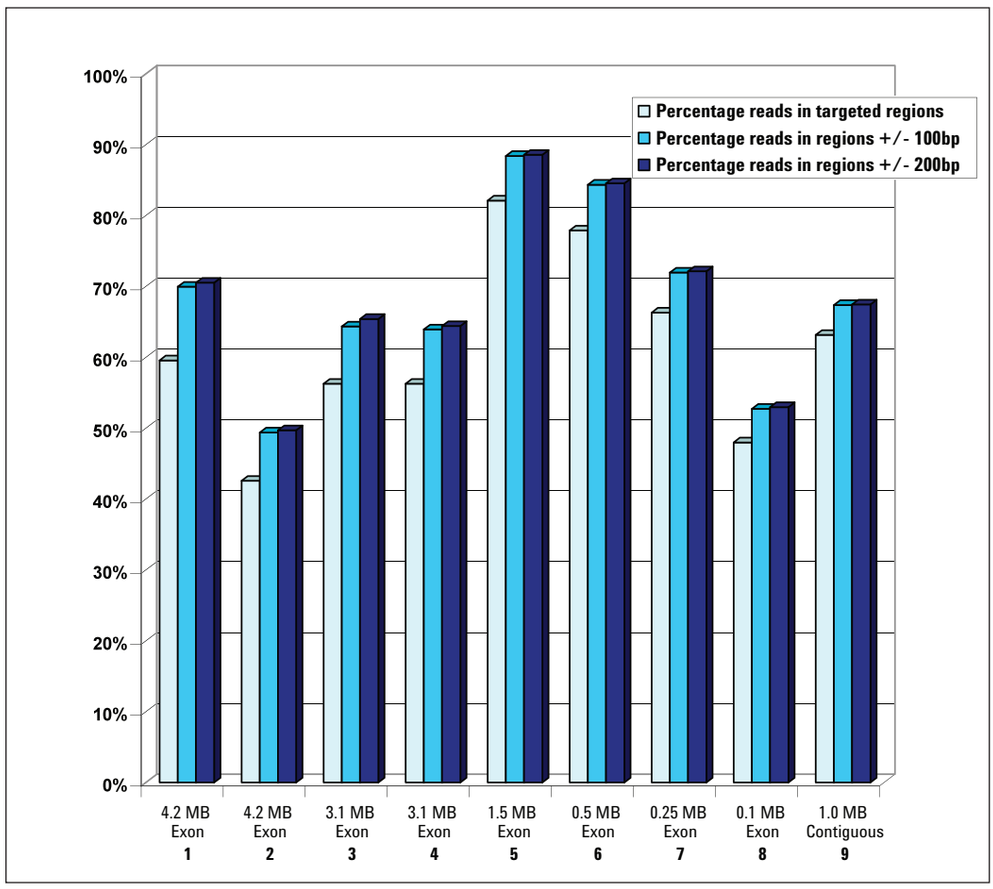


Figure 2
Target enrichment performance over a diverse set of designs. Percentage of sequence reads that map to targeted regions for different library designs. Each set of three values represents the percent of reads that are: on target (light blue), on target plus within 100bp of target (med blue), and on target plus within 200bp of target (dark blue). Samples 1 and 2) 4.2mb Exon Design, different samples; 3 and 4) 3.1mb Exon Design, X Chromosome Demonstration Kit, different samples; 5) 1.5mb Exon Design 6) 0.5mb Exon Design; 7) 0.25mb Exon Design; 8) 0.1mb Exon Design; 9) 1.0mb Contiguous Region Design

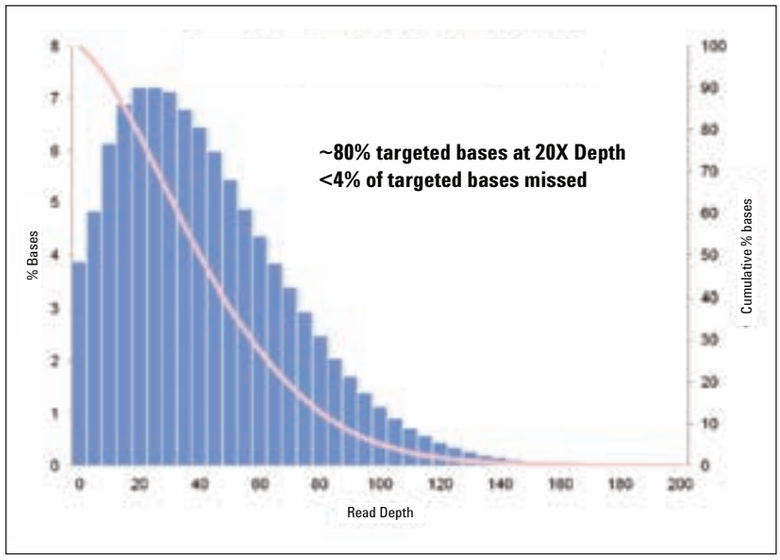


Figure 3: Target enrichment sequence coverage. Sequencing read depth across target intervals for a genomic DNA sample. Sample prepared for the Illumina Genome Analyzer and captured by a target enrichment library covering exonic regions totalling 3.3mb with 50% probe overlap. The columns show the distribution of sequencing read depths per base pair. Read depth is shown on the X-axis, binned percentage of reads at each read depth on the Y-axis (left). The pink line and right Y-axis show the cumulative read depth as a percent of total bases

Size-selected libraries are then incubated with the baits for 24 hours. RNA bait-DNA hybrids are then 'fished' out of the complex mixture by incubation with streptavidin-labelled magnetic beads and captured on a powerful magnet. After the beads have been washed, the RNA bait is then digested so that the only remaining nucleotide is the targeted DNA of interest. A few cycles of DNA amplification are performed at the end of the capture, and the targeted sample is then loaded on to the sequencing instrument.

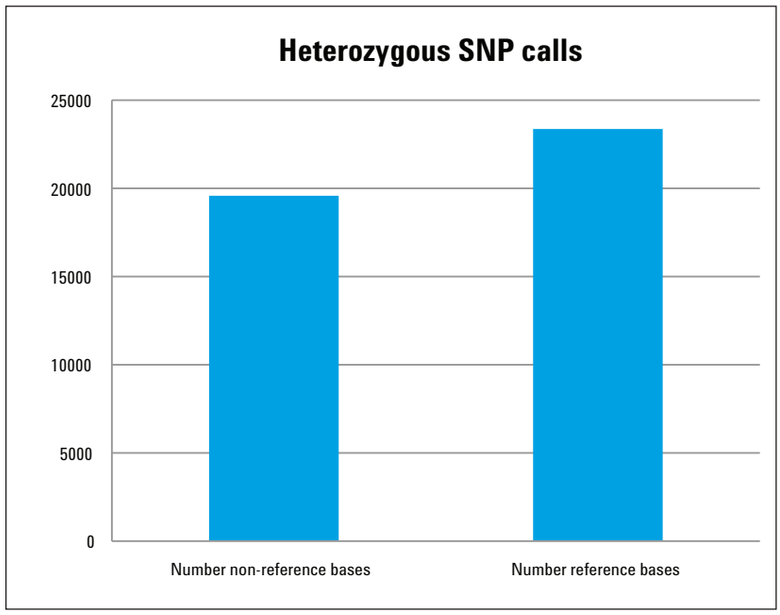


Figure 4: Allele balance. The number of reference bases versus non-reference bases in heterozygous SNP calls

Enrichment design algorithms enable target region specificity

Enrichment design algorithms have been developed that enable unique probe sets to be optimised for a target enrichment method. These have been proven to work over a diverse set of genomic locations, such as small and large exons, short and long contiguous genomic targets, genome targets within repeat areas and non-coding DNA. To assess the performance of these algorithms for target enrichment, a diverse set of enrichment designs were created and tested in parallel against multiple samples (**Figure 2**). To measure performance, one of the most direct measures of enrichment efficiency was used – analysing the sequencing reads for percent in the targeted regions. As shown in **Figure 2**, the percent on target for seven designs is within approximately 40-80%, demonstrating efficiency in focusing on only regions of interest across several diverse probe configurations for target enrichment. This translates into enrichments of 300-7,400-fold; fold enrichment is largely dependent upon the size of the targeted region as well as the number of reads per sequencing run. Small capture designs typically yield a higher level of enrichment due to the kit's efficiency in focusing sequencing reads on a smaller subset of the targeted genome. Because the shearing of DNA prior to creation of prepared libraries is random, no matter how specific the capture methodology, both the targeted region and near DNA targets will be captured. These design algorithms and the protocol to prepare libraries prior to capture can be optimised in such a way as to limit the near target sequences captured. These near-target sequences are of little to no use to researchers and may also entrain repeat regions close to targeted regions, in particular, exons. To this end, the specificity of the target enrichment method was measured by analysing sequencing reads exactly on target, within 100bp of the target, and within 200bp of the target. **Figure 2** shows the off-target capture rate is not substantially increased by including sequence reads within up to 200bp of the target.

Read distribution and sequence coverage

Another metric of great importance to DNA sequencing research is read distribution, as it affects the ability to adequately cover genomes to

References

1 Bentley, DR et al. Accurate whole genome sequencing using reversible terminator chemistry. Nature 456, 53-59 (2008).
2 Ley, TJ et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome, Nature 456. 66-72 (2008).
3 Wang, J et al. The diploid genome sequence of an Asian individual. Nature 456, 60-66 (2008).
4 Wheeler, DA et al. The complete genome of an individual by massively parallel DNA sequencing. Nature 452, 872-876 (2008).
5 Margulies, M et al. Genome sequencing in microfabricated high density picolitre reactors. Nature 437, 376-380 (2005).
6 Shandure, J et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309, 1128-1732 (2005).
7 Smith, DR et al. Rapid whole-genome mutational profiling using next-generation sequencing technologies. Genome Res. 18, 1638-1642 (2008).
8 Dahl, F, Gullberg, M, Stenberg, J, Landegren, U and Nilsson, M. Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. Nucleic Acids Res. 33. e71 (2005).
9 Albert, TJ et al. Direct selection of human genomic loci by microarray hybridization. Nat. Methods 4, 903-905 (2007).
10 Dahl, F et al. Multigene amplification and massively parallel Sequencing for cancer mutation discovery. Proc. Natl. Acad. Sci. USA 104, 9387-9392 (2007).
11 L Fredriksson, S et al. Multiplex amplification of all coding sequences within 10 cancer genes by Gene-Collector Nucleic Acids Res. 35. e47 (2007).
12 Hodges E, et al. Genome-wide in situ exon capture for selective resequencing, Nat, Genet. 39, 1522-1527 (2007).

Continued on page 81

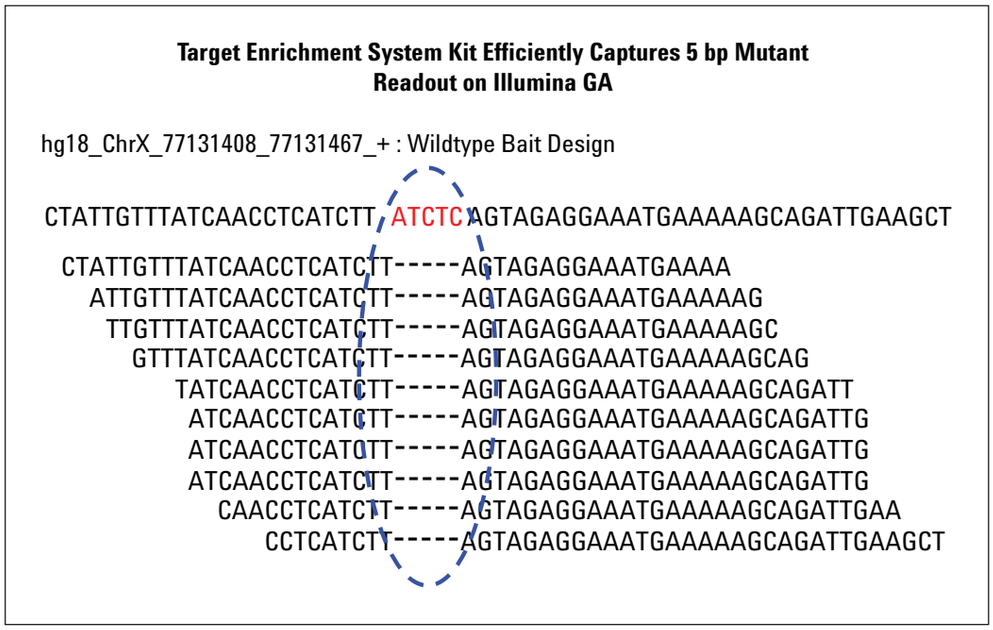


Figure 5: Target enrichment captures region with large deletion. Sequencing results from captures of a 5BP deletion on the X-Chromosome: Menke's Syndrome

identify sequence variation. However, none of the commercially available next-generation sequencers can output a perfectly even read distribution across a genome of interest due to the nature of massively parallel sequencing. Statistically there will always be some stretches of DNA that are read out more than others. The challenge for any sample preparation method aimed at specifically selecting targets for sequencing is to not introduce any further sequence bias. There are two ways of looking at this metric: 1) plot the distribution of the actual percent of bases with a certain number of reads; and 2) plot the cumulative number of bases with at least a certain depth. **Figure 3** shows a plot of a representative sample run on an Illumina Genome Analyzer. This plot shows that roughly 80% of all bases have at least a 20x read depth, indicating the suitability of this method for identifying SNPs with targeted resequencing. Plots like this will vary considerably based upon the capture design and run parameters set on the sequencer, but despite this variability, the sequence coverage seen with the target enrichment method is very even, showing negligible bias. For example, if a capture design is made to target 3.3mb of genomic locations, and 10 million reads are obtained from the sequencing run, the distribution of read depth is typically centred around 30x with 70% of the reads within 1.7 logs. Additionally, in this type of scenario one could expect approximately 95%+ of the targeted bases to have at least one read, 80%+ to have at least five reads, and 50%+ to have at least 20 reads, making this target enrichment method ideally suited for interrogating genomes for mutations.

Robust and reproducible results

To demonstrate that the target enrichment process is both robust and reproducible, the same sample was captured with the same kit twice and run on two separate lanes of an Illumina Genome Analyzer. Sequence coverage results were then compared between replicates. The read depth for corresponding genomic locations between experiments shows strong correlation, indicating experiment-to-experiment reproducibility. Additionally, the technical replicates gave results that were very consistent for the following metrics: actual bases represented, percentage on target bases, and percentage of reads with 30x read depth. Note that the differences in read depth from one replicate to the other is driven by the difference in total reads from the sequencer – the sample represented on the Y axis generated more reads than the sample represented on the X axis. However, the R2 value of 0.9616 shows strong correlation between sample runs, indicating high reproducibility of the target enrichment.

Little or no allele bias

Any target enrichment process performed prior to next-generation sequencing has the potential to bias allele representation, complicating the identification of sequence variation. To demonstrate that bias is limited in the target enrichment process, sequence data were analysed for allele balance. **Figure 4** shows data derived from a control sample of known heterozygosity for >20,000 SNP reads. This information was used to deconvolute the allelic origin of each associated read. The results show little to no bias as the coverage across both alleles for several SNPs is quite similar. Thus, the Agilent target enrichment approach is efficient at capturing DNA regardless of whether the DNA strand targeted is 'wild type' or contains mutations, an indicator of the utility of ultra-long oligonucleotides for capture.

Robust and reliable SNP identification and confirmation

One of the main applications of target enrichment prior to next-generation sequencing is the follow-up to whole genome association studies to: 1) confirm SNPs; 2) define new SNPs; and 3) correlate these mutations to disease states. We wanted to verify that the method could reliably capture known and unknown SNPs and that newly identified SNPs are genuine. Using the SureSelect method, we selected both known SNPs and a novel SNP identified after an enrichment experiment (confirmed by PCR and Sanger sequencing).

Capture method for genomes with large deletions

One of the biggest challenges for target enrichment sequence capture methods is the ability to capture regions of genomes that possess large deletions. Methods that utilise shorter oligonucleotides capture these regions less efficiently than systems utilising complex mixtures of longer (>100bp) oligonucleotides, simply due to the less favourable hybridisation kinetics. To improve capture performance across regions with insertions and deletions, the target enrichment method utilises complex mixtures of RNA-based oligonucleotides of 120bp. Capture efficiency around stretches of deletions is improved by both the longer oligonucleotides and the chemical nature of the capture oligonucleotide. RNA's stronger affinity for DNA improves the SureSelect Target Enrichment System's efficiency in binding targeted regions that possess mutations such as insertions and deletions. The long capture probes are also more tolerant of mutations, as shown in **Figure 5** where a deletion of five nucleotides on human chromosome X from a patient with Menkes Disease did not prevent this region from being represented with 10x depth on an Illumina Genome Analyzer.

Summary

The data presented here demonstrate the robustness and reliability of the described target enrichment method. By requiring 10-fold less input DNA than other commercial systems and by focusing DNA sequencing on genomic areas of interest, this method enables studies that were previously unfeasible due to the rarity of DNA sample and/or the overall cost of a sequencing study. In effect, it will enable studies that were previously unfeasible due to the rarity of DNA sample and/or the overall cost of a sequencing study, by allowing researchers to process anywhere from a handful to thousands of samples, using the same approach.

Acknowledgements

We thank the Broad Institute Genome Sequencing Platform and Genetic Analysis Platform for generating sequencing and genotyping data. DDW

Dr Fred Ernani *is currently Senior Product Manager, Emerging Genomic Applications at Agilent Technologies. Previously he was at Microchip Biotechnologies as well as Applied Biosystems and GE Healthcare. Prior to the commercial world, Dr Ernani was Senior Research Associate at the University of Texas, Health Science Centre.*

Dr Emily LeProust *joined the Genomics division of Agilent Technologies in 2000 and has held several technical and management positions in R&D and manufacturing focusing on the development and deployment of chemical processes for the industrial scale synthesis of Microarrays and Oligo Libraries. Most recently, Dr LeProust has been directing the Applications and Chemistry R&D team developing quantitative and structural Genomic applications powered by Microarray and Next Generation Sequencing technologies. Dr LeProust holds a MS degree in Industrial Chemistry from the Ecole Supérieure de Chimie Industrielle de Lyon (Lyon School of Industrial Chemistry, France) and a PhD in Organic Chemistry from the University of Houston. Dr LeProust has 14 granted patents and 17 publications.*

Continued from page 80

13 Okouj DT et al. Microarray-based genomic selection for high-throughput resequencing. Nat. Methods 4, 907-909 (2007).
14 Porreca, GJ et al. Multiplex amplification of large sets of human exons. Nat. Methods 4,931-936 (2007).
15 Krishnakumar, S at al. A comprehensive assay for targeted multiplex amplification of human DNA sequences. Proc, Natl. Acad. Sci USA 105, 9296-9301 (2008).
16 Clamp, M et al. Distinguishing protein-coding and noncoding genes in the human genome. Proc. Natl. Acad. Sci. USA 104,19428-19433 (2007).
17 Nilsson, M et al. Padlock probes, circularizing oligonucleotides for localized DNA detection. Science 255, 2085-2088 (1994).
18 Hardenbol, P et al. Multiplexed genotyping with sequence-tagged molecular inversion probes. Nat. Biotechnol. 21, 673-678 (2003).
19 Gnirke, A et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 27, 182-9 (2009).