Targeted enrichment of genomic DNA regions for next-generation sequencing

Florian Mertes, Abdou ElSharawy, Sascha Sauer, Joop M.L.M. van Helvoort, P.J. van der Zaag, Andre Franke, Mats Nilsson, Hans Lehrach and Anthony J. Brookes

Advance Access publication date 26 November 2011

Abstract

In this review, we discuss the latest targeted enrichment methods and aspects of their utilization along with second-generation sequencing for complex genome analysis. In doing so, we provide an overview of issues involved in detecting genetic variation, for which targeted enrichment has become a powerful tool. We explain how targeted enrichment for next-generation sequencing has made great progress in terms of methodology, ease of use and applicability, but emphasize the remaining challenges such as the lack of even coverage across targeted regions. Costs are also considered versus the alternative of whole-genome sequencing which is becoming ever more affordable. We conclude that targeted enrichment is likely to be the most economical option for many years to come in a range of settings.

Keywords: *targeted enrichment; next-generation sequencing; genome partitioning; exome; genetic variation*

INTRODUCTION

Next-generation sequencing (NGS) [1, 2] is now a major driver in genetics research, providing a powerful way to study DNA or RNA samples. New and improved methods and protocols have been developed to support a diverse range of applications, including the analysis of genetic variation. As part of this, methods have been developed that aim to achieve 'targeted enrichment' of genome subregions [3, 4], also sometimes referred to as 'genome partitioning'. Strategies for direct selection of genomic regions were already developed in anticipation of the introduction of NGS [5, 6]. By selective recover and subsequent sequencing of genomic loci of interest, costs and efforts can be reduced significantly compared with whole-genome sequencing.

Targeted enrichment can be useful in a number of situations where particular portions of a

Corresponding author. Florian Mertes, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany. Tel: +49 30 8413 1289; fax +49 30 8413 1128; E-mail: mertes@molgen.mpg.de

Florian Mertes studied biotechnology and earned a Doctorate from the Technical University Berlin. Currently, he is a postdoctoral researcher focusing on applied research to develop test/screening assays based on high-throughput technologies, using both PCR and next-generation sequencing.

Abdou ElSharawy is a postdoctoral researcher (University of Kiel, CAU, Germany), and lecturer of Biochemistry and Cell Molecular Biology (Manusoura University, Egypt). He focuses on disease-associated mutations and miRNAs, allele-dependent RNA splicing, and high-throughput targeted, whole exome, and genome sequencing.

Sascha Sauer is a research group leader at the Max Planck Institute for Molecular Genetics, and coordinates the European Sequencing and Genotyping Infrastructure.

Joop M.L.M. van Helvoort is CSO at FlexGen. He received his PhD at the University of Amsterdam. His expertise is in microarray applications currently focusing on target enrichment.

P.J. van der Zaag is with Philips Research, Eindhoven, The Netherlands. He holds a doctorate in physics from Leiden University. At Philips, he has worked on a number of topics related to microsystems and nanotechnology, lately in the field of nanobiotechnology.

Andre Franke is a biologist by training and currently holds an endowment professorship for Molecular Medicine at the Christian-Albrechts-University of Kiel in Germany and is guest professor in Oslo (Norway).

Mats Nilsson is Professor of Molecular Diagnostics at the Department of Immunology, Genetics, and Pathology, Uppsala University, Sweden. He has pioneered a number of molecular analysis technologies for multiplexed targeted analyses of genes.

Hans Lehrach is Director at the Max Planck Institute for Molecular Genetics. His expertise lies in genetics, genomics, systems biology, and personalized medicine. Highlights include key involvement in several large-scale genome sequencing projects.

Anthony J Brookes is a Professor of Bioinformatics and Genomics at the University of Leicester (UK) where he runs a research team and several international projects in method development and informatics for DNA analysis through to healthcare.

whole genome need to be analyzed [7]. Efficient sequencing of the complete 'exome' (all transcribed sequences) represents a major current application, but researchers are also focusing their experiments on far smaller sets of genes or genomic regions potentially being implicated in complex diseases [e.g. derived from genome-wide association studies (GWAS)], pharmacogenetics, pathway analysis and so on [1, 8, 9]. For identifying monogenetic diseases, exome sequencing can be a powerful tool [10]. Across all these areas of study, a typical objective is the analysis of genetic variation within defined cohorts and populations.

Targeted enrichment techniques can be characterized via a range of technical considerations related to their performance and ease of use, but the practical importance of any one parameter may vary depending on the methodological approach applied and the scientific question being asked. Arguably, the most important features of a method, which in turn reflect the biggest challenges in targeted enrichment, include: enrichment factor, ratio of sequence reads on/off target region (specificity), coverage (read depth), evenness of coverage across the target region, method reproducibility, required amount of input DNA and overall cost per target base of useful sequence data.

Within this review, we compare and contrast the most commonly used techniques for targeted enrichment of nucleic acids for NGS analysis. Additionally, we consider issues around the use of such methods for the detection of genetic variation, and some general points regarding the design of the target region, input DNA sample preparation and the output analysis.

ENRICHMENT TECHNIQUES

Current techniques for targeted enrichment can be categorized according to the nature of their core reaction principle (Figure 1):

(i) 'Hybrid capture': wherein nucleic acid strands derived from the input sample are hybridized specifically to preprepared DNA fragments complementary to the targeted regions of interest, either in solution or on a solid support, so that one can physically capture and isolate the sequences of interest;

(ii) 'Selective circularization': also called molecular inversion probes (MIPs), gap-fill padlock probes and selector probes, wherein single-stranded DNA circles that include target region sequences are formed (by gap-filling and ligation chemistries) in a highly specific manner, creating structures with common DNA elements that are then used for selective amplification of the targeted regions of interest;

(iii) PCR amplification: wherein polymerase chain reaction (PCR) is directed toward the targeted regions of interest by conducting multiple long-range PCRs in parallel, a limited number of standard multiplex PCRs or highly multiplexed PCR methods that amplify very large numbers of short fragments.

Given the operational characteristics of these different targeted enrichment methods, they naturally vary in their suitability for different fields of application. For example, where many megabases needs to be analyzed (e.g. the exome), hybrid capture approaches are attractive as they can handle large target regions, even though they achieve suboptimal enrichment over the complete region of interest. In contrast, when small target regions need to be examined, especially in many samples, PCR-based approaches may be preferred as they enable a deep and even coverage over the region of interest, suitable for genetic variance analysis.

An overview of these different approaches is presented in Figure 1, and Table 1 lists the most common methods along with additional information.

Basic considerations for targeted enrichment experiments

The design of a targeted enrichment experiment begins with a general consideration of the target region of interest. In particular, a major obstacle for targeted enrichment is posed by repeating elements, including interspersed and tandem repeats as well as elements such as pseudogenes located within and outside the region of interest. Exclusion of repeat masked elements [11] from the targeted region is a straightforward and efficient way to reduce the recovery of undesirable products due to repeats. Furthermore, at extreme values (<25% or >65%), the guanine-cytosine (GC) content of the target region has a considerable impact on the evenness and efficiency of the enrichment [12]. This can adversely affect the enrichment of the 5′-UTR/promoter region and the first exon of genes, which

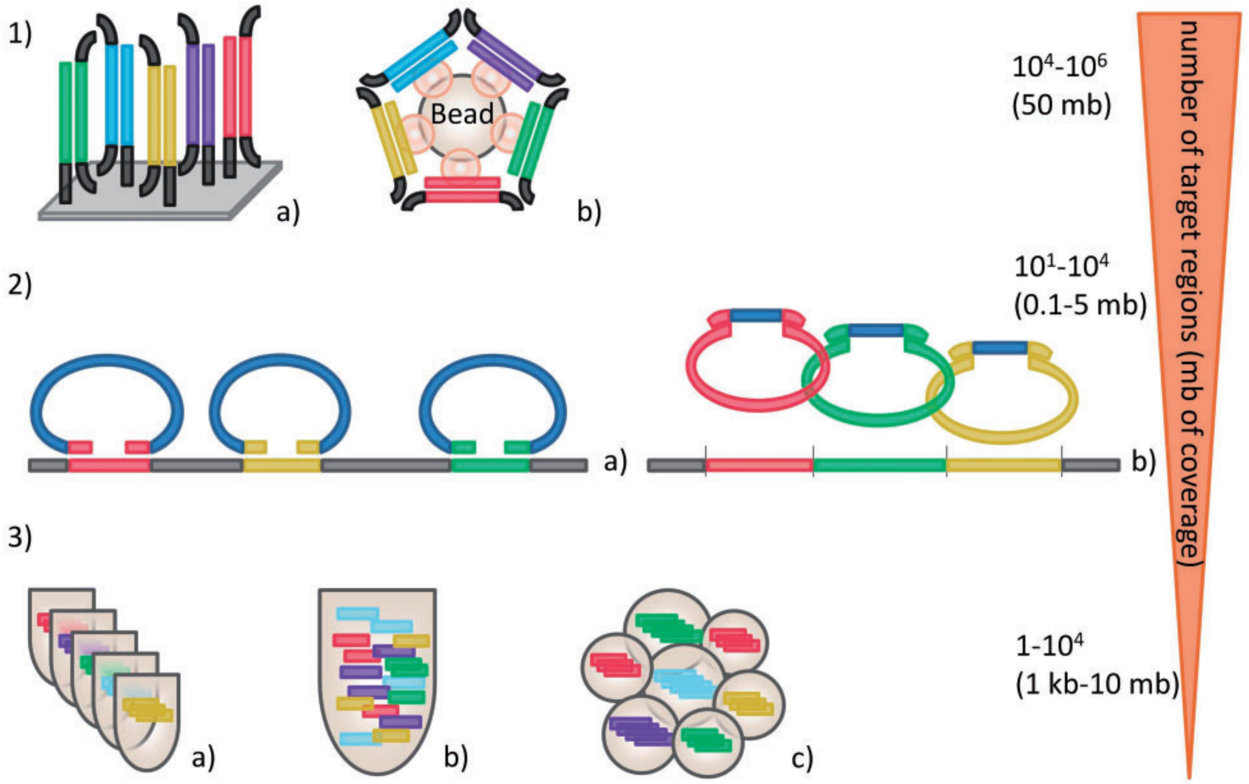

Figure 1: Commonly used targeted enrichment techniques. (1) Hybrid capture targeted enrichment either on solid support-like microarrays (a) or in solution (b). A shot-gun fragment library is prepared and hybridized against a library containing the target sequence. After hybridization (and bead coupling) nontarget sequences are washed away, the enriched sample can be eluted and further processed for sequencing. (2) Enrichment by MIPs which are composed of a universal sequence (blue) flanked by target-specific sequences. MIPs are hybridized to the region of interest, followed by a gap filling reaction and ligation to produce closed circles. The classical MIPs are hybridized to mechanically sheared DNA (a), the Selector Probe technique uses a restriction enzyme cocktail to fragment the DNA and the probes are adapted to the restriction pattern (b). (3) Targeted enrichment by differing PCR approaches. Typical PCR with single-tube per fragment assay (a), multiplex PCR assay with up to 50 fragments (b) and RainDance micro droplet PCR with up to 20 000 unique primer pairs (c) utilized for targeted enrichment.

are often GC rich [13]. Therefore, expectations regarding the outcome of the experiment require careful evaluation in terms of the precise target region in conjunction with the appropriate enrichment method.

The performance of a targeted enrichment experiment will also depend upon the mode and quality of processing of the input DNA sample. Having sufficient high-quality DNA is key for any further downstream handling. When limited genomic DNA is available, whole-genome amplification (WGA) is usually applied. Since WGA produces only a representation and not a replica of the genome, a bias is assumed to be introduced though the impact of this on the final results can be compensated for, to a degree by identically manipulating control samples [14].

All three major targeted enrichment techniques (hybrid capture, circularization and PCR) differ in terms of sample library preparation workflow enabling sequencing on any of the current NGS instruments (e.g. Illumina, Roche 454 and SOLiD). Enrichment by hybrid selection relies on short fragment library preparations (typically range from 100 to 250 bp) which are generated before hybridization to the synthetic library comprising the target region. In contrast, enrichment by PCR is performed directly on genomic DNA and thereafter are the library primers for sequencing added. Enrichment by circularization offers the easiest library preparation for NGS because the sequencing primers can be added to the circularization probe, thus eliminating the need for any further library preparation steps.

Table 1: Currently employed targeted enrichment techniques

Enrichment technique	Vendor	Features	Pros	Cons	Number of loci (target size)	Library prep for NGS
Hybrid capture						
Solid support	Agilent SureSelect, Roche NimbleGen SeqCap EZ	Medium to large target regions, custom and preconfigured target regions (i.e. whole exome), Multiplexing possible, ready to use kits	Ease of production, large target sets	Large amount of input DNA, high-tech equipment (3–10 μg)	10^4–10^6 (1–50 Mb)	Before enrichment
In solution	Agilent SureSelect, FlexGen FleXelect, MYcroarray MYselect, Roche NimbleGen SeqCap EZ		Ease of use, small amount of input DNA (<1–3 μg)			
Circularization						
Molecular inversion probes			No dedicated instruments, high specificity, input DNA (<1 μg)		10^2–10^4 (0.1–5 Mb)	During enrichment (incorporated into hybridization probes
Selector probes	HaloGenomics	Custom kits and clinically relevant panel kits		exome kit not available yet	10–200 (0.1–1.5 Mb)	
PCR						
Long range	Invitrogen SequalPrep, Qiagen SeqTarget system	Smaller target regions, coverage by tiling	Relatively easy to set up and automatable, even coverage	PCR conditions largely influence effectiveness, >10 μg DNA for Large sets	10^2–10^4 (0.1–5 Mb)	After enrichment
Multiplex	Multiplicom, Fluidigm	Smaller target regions, coverage by tiling, multiplex PCR of 150–200 amplicons (150–450 bp)	Easy to perform, reasonably economical in terms of, even coverage			
Micro droplet	RainDance	Smaller target regions, coverage by tiling, micro droplet PCR of up to 20 000 amplicons (150–1500 bp)	Even coverage, low input DNA	Relatively expensive, specialist equipment	10^3–10^4 (up to 10 Mb)	

All major targeted enrichment techniques show relative pros and cons.

Sequencing can be performed either as single read or paired-end reads of the fragment library. In general, mate–pair libraries are not used for hybridization-based targeted enrichments due to the extra complications this implies in terms of target region design.

In general, a single NGS run produces enough reads to sequence several samples enriched by one of the mentioned methods. Therefore, pooling strategies and indexing approaches are a practical way to reduce the per sample cost. Depending on the method used for targeted enrichment, different multiplexing strategies can be envisaged that enable multiplexing in different stages of the enrichment process: before, during and after the enrichment. For targeted enrichment by hybrid capture, indexing of the sample is usually performed after the enrichment but to reduce the number of enrichment reactions, the sample libraries can alternatively be indexed during the library preparations and then pooled for enrichment [15]. Enrichment by PCR and circularization offers indexing during the enrichment by using bar-coded primers in the product amplification steps [16]. Furthermore, two multiplexing strategies can be combined in a single experiment. First, multiple samples can be enriched as a pool, with each harboring a unique pre-added bar-code. Then second, another bar-coding procedure can be applied postenrichment, to each of these pools, giving rise to a highly multiplexed final pool. If such extensive multiplexing is used, great care must be taken to normalize the amount of each sample within the pool to achieve sufficiently even representation over all samples in the final set of sequence reads. In addition, highly complex pooling strategies also imply far greater challenges when it comes to deconvoluting the final sequence data back into the original samples.

The task of designing the target region is relatively straightforward, and this can be managed with web-based tools offered by UCSC, Ensembl/BioMart, etc. and spreadsheet calculations (e.g. Excel) on a personal computer. Web-based tools like MOPeD offer a more user-friendly approach for oligoncleotide probe design [17]. Far more difficult, however, is the final sequence output analysis, which needs dedicated computer hardware and software. Fortunately, great progress has recently been made in read mapping and parameter selection for this process, leading to more consistent and higher quality final results [18]. Reads generated by hybrid selection will always tend to extend into sequences beyond the target region and the longer the fragment library is, the more of these 'near target' sequences will be recovered. Therefore, read mapping must start with a basic decision regarding the precise definition of the on/off target boundaries, as this parameter is used for counting on/off target reads and so influences the number of sequence reads considered as on target. This problem is not so critical for enrichments based on PCR and circularization as these methods do not suffer from 'near target' products. Another major consideration in data analysis is the coverage needed to reliably identify sequence variants, e.g. single nucleotide polymorphisms (SNP). This depends on multiple factors such as the nature of the region of interest in question, the method used for targeted enrichment. In different reports, it has ranged from 8x coverage [19], which was the minimum coverage for reliable SNP calling and up to 200x coverage [20], in this case the total average coverage for the targeted region.

Enrichment by hybrid capture

Enrichment by hybrid capture (Figure 1.1a and b) builds on know-how developed over the decade or more of microarray research that preceded the NGS age [21, 22]. The hybrid capture principle is based upon the hybridization of a selection 'library' of very many fragments of DNA or RNA representing the target region against a shotgun library of DNA fragments from the genome sample to be enriched. Two alternative strategies are used to perform the hybrid capture: (i) reactions in solution [4] and (ii) reactions on a solid support [3]. Each of these two approaches brings different advantages, as listed in Table 1.

Selection libraries for hybrid capture are typically produced by oligonucleotide synthesis upon microarrays, with lengths ranging from ~60 to ~180 bases. These microarrays can be used directly to perform the hybrid capture reaction (i.e. surface phase methods), or the oligonucleotide pool can be harvested from the array and used for an in-solution targeted enrichment (i.e. solution phase methods). The detached oligonucleotide pool enables versatile downstream processing: if universal 5′- and 3′-end sequences are included in the design of the oligonucleotides, the pool can be reamplified by PCR and used to process many genomic samples. Furthermore, it is possible to introduce T7/SP6 transcription start sites via these PCRs [23], so that the pool can be transcribed into RNA before being used in an enrichment experiment.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.

WHAT WILL YOU BUILD? | sales@docketalarm.com | 1-866-77-FASTCASE

fastcase®
Smarter legal research.