Genomics 94 (2009) 363–368

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Minireview

Enabling technologies of genomic-scale sequence enrichment for targeted high-throughput sequencing

Daniel Summerer *

febit biomed gmbh, Im Neuenheimer Feld 519, 69120 Heidelberg, Germany

ARTICLE INFO

Article history:
Received 13 July 2009
Accepted 22 August 2009
Available online 29 August 2009

Keywords:
Next-generation-sequencing
Sequence capture
Microarrays
Hybridzation probes

ABSTRACT

Next-generation sequencing has still not reached its full potential due to the technical inability of effectively targeting desired genomic regions of interest. Once available, methods adressing this bottleneck will dramatically reduce cost and enable the efficient analysis of complex samples.
Recently, a number of possible approaches for genomic-scale sequence enrichment have been reported using different strategies. All methods basically rely on sequence-specific nucleic acid hybridization, however, they differ in several aspects such as the use of solid phase versus solution phase hybridization, probe design and overall workflows with implications for automation.
Overall, several key challenges of genome-wide sequence enrichment have become clear after these studies that remain to be overcome. We summarize the different technologies and highlight individual characteristics related to general potential and different suitabilities for specific applications.
© 2009 Elsevier Inc. All rights reserved.

Introduction

The vast capacity of next-generation sequencers (NGS) has tremendously increased the scope and comprehensiveness of genomics projects [1-6]. Pro- and eukaryotic genomes are now accessible within days or weeks, fundamentally changing typical project scales in genetics research. Besides the effect of increased throughput, large-scale sequencing studies are now open to many more research laboratories from different disciplines due to associated reduction of cost. This democratizing effect of NGS technologies will accelerate new discoveries and promises to diversify the way in which genetic studies are designed.

However, the enhanced throughput on the sequencing side has not been flanked by the development of suited sample preparation techniques allowing for the focussed analysis of genomic subsets [7]. In fact, until very recently, no efficient methods have been available to enrich DNA sequences out of complex genomic mixtures at a capacity exceeding the low kilobase range of classic PCR. Though PCR as a well-established method of enrichment is basically feasible, its sequence scale and level of multiplexing rather match the throughput of traditional sanger sequencing. The megabase capacity of NGS instruments however may require thousands of PCR reactions for a typical study, including potential optimization of individual reactions, synthesis of primer pairs and normalization. This lack of large scale enrichment methods represents a serious bottleneck for the exploitation of NGS instruments full potential.

The most immediate need for sequence enrichment originates from the current capacities of NGS platforms that do not allow sequencing of whole genomes of complex eukaryotic organisms with reasonable effort [8,9]. This essentially prevents the advantages of NGS for studies involving human and many eukaryotic model organisms. Additionally, large-scale enrichment methods might well play their part even after the next leap in sequencing throughput. Though complete sequencing of a human genome in one instrument run at a cost of about $1000 is certainly a next milestone of DNA sequencing technology [10], a further level is the multiplexing of several genomes within one run [11,12]. Beyond that, applying DNA sequencing to even more complex samples, for example in human population studies, analysis of microbial communities, host–pathogen mixtures or somatic variants might again substantially benefit from sequence enrichment methods of suited efficiency and scale [13]. Since even large regions of interest like a whole exome typically represent only a few percent of a genome for human and many model organisms, efficient targeted sequencing can dramatically reduce cost and effort. This reduction becomes even larger with multi-genome complexities of the analyzed sample. From a practical point of view, sequence enrichment methods also enable the efficient use of the in-built compartments of current NGS platforms for multiplexing of several, separated samples without indexing strategies.

Recently, a number of approaches have been reported that might help to overcome these current bottlenecks [14-22]. All of these rely on complementary hybridization of nucleic acid capture probes to the targeted DNA sequences. However, there are also substantial differences. Some methods use solution phase and others solid phase hybridization and the methods differ in overall workflows and ease of automation. The design of the individual sequence capture

* Fax: +49 6221 6510 390.
E-mail address: daniel.summerer@febit.de.

0888-7543/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2009.08.012

steps additionally influences the accessibility for different types of target regions and overall efficiency. Here, we summarize these recent developments in this highly dynamic and open field and point out similarities and substantial differences between the approaches. We thereby emphasize technological and conceptual differences rather than absolute performance parameters since experiments for direct comparison are not available so far.

Critical parameters of genomic-scale sequence enrichment

It has become clear from recent studies that efficient capture of target sequences on a genomic scale imposes several special requirements on respective enrichment technologies.

For example, many relevant regions in targeted NGS, such as exons, vary in size and sequence properties and are discontiguously distributed in the genome within a context of low complexity sequence. An ideal sequence enrichment method should therefore allow random access to multiple different loci relatively independent of their size, sequence composition and spatial distribution. This must be achieved at a multiplexing grade that matches NGS capacities. Individual loci should not only be generally accessible for capture but should be enriched with equal efficiencies to allow for complete and uniform coverage of the targeted region. This is essential for economic enrichment, since a high uniformity avoids redundant reads from overcaptured regions. In terms of enrichment performance, an ideal enrichment method would thus allow for complete and entirely even coverage of the target region with the minimum depth required for reliable nucleotide calling. This has to be achieved without introduction of bias into allele representations. For economic reasons, enrichment efficiency should finally be such that data output of the NGS intrument is only related to target region with minimal background sequence data.

On a molecular level, several of these performance parameters seem to largely depend on two basic process properties. Firstly, the hybridization step itself determines specificity and uniformity of binding capacity for individual target regions and allelic variants and thus influences the enrichment efficiency and consequently the fraction of on-target reads in the NGS instruments output. Good performance and high sequence capacity of the hybridzation step can be adressed by using specific hybridization probe libraries with sizes matching NGS scale. Established design rules for genomic DNA microarray hybridization, e.g., aiming at maximal binding specificity, similar melting temperatures and minimal content of low complexity sequence have thereby been applied.

Secondly, the individual methods have different local coverage distributions at specific target loci as a result of, e.g., capture strategy, probe type and library characteristics. Even for an entirely selective hybridization step, this presents a potential limitation for target accessibility, coverage uniformity and the fraction of target-related data from a sequencing run (Fig. 1, see below).

Another aspect of the methods is their speed and ease of automation to streamline sequencing workflows. This becomes increasingly important for larger facilities like genome centers as the number of instruments grows. Automation can for example be promoted by technical features like a simple overall workflow that avoids extensive manipulations of capture probes or sequencing libraries, use of standard steps that can be integrated into liquid handling systems or the availability of taylored hardware for automated processing. Avoiding manual intervention should also be beneficial for reproducibility, contamination risk and cost. For improvement of process speed, hybridization times seem to be a major concern since these have by far been the most time-consuming steps in previously reported methods.

Solution phase hybridization

Beside PCR and long-range PCR, several methods have been described that make use of solution phase hybridization to target sequences. Two general strategies have been developed so far. One uses different types of circularizing probes, enzymatic manipulation and generic amplification to obtain the enriched sequences [17,21,22]. The other uses enzymatically generated, long RNA probes that are immobilized on beads after hybridization for washing and elution of the desired target fragments [18].

Circularizing probes are known for applications such as FISH or SNP genotyping and have now been developed further for genome-wide sequence capture. In one method, the basic strategy is the use of 70 bp DNA “molecular inversion probes” (MIP) bearing two terminal target recognition sequences that are connected by a common linker [17,21,22]. These are generated by flexible *in situ* DNA microarray synthesis and enzymatic processing (Fig. 2). Both recognition sequences hybridize to the target loci to form a gap of ~60–190 bp that is subsequently filled in by a DNA polymerase. The resulting nick is closed by a DNA ligase and non-circularized probes are removed by an exonuclease digest, resulting in a circular library of target loci copies. An alternative approach using so-called Selector probes [17] relies on a similar probe type of 80 bp that however uses a doublestranded 40 bp common linker to assist ligation of the actual

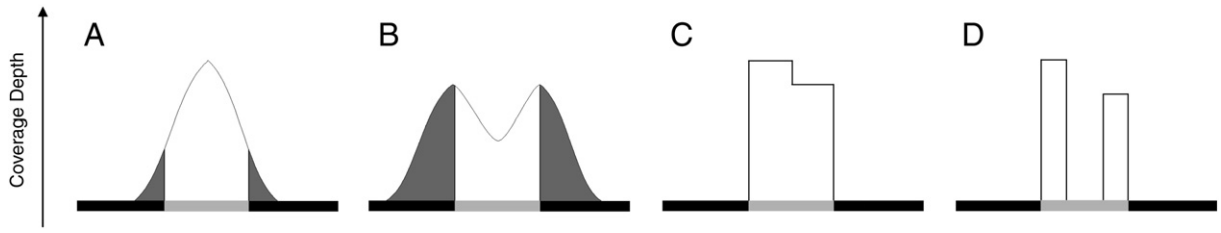

Fig. 1. Schematic view of effects of different sequence capture approaches on relative local coverage depth distributions around target regions. Genomic DNA contig is shown as bold black line with target region in light grey. Coverage depth distributions are shown as continuous black lines with white areas for on-target coverage and dark grey areas as off-target coverage. (A) Scheme of coverage depth distribution for sequencing of shotgun libraries as reported for microarray capture with probes <100 bp in length. Randomly fragmented shotgun NGS library results in binomial-like coverage distribution with maximum in the middle of the target region. A fraction of non-informative off-target reads is generated by fragments overlapping into flanking regions. Absolute proportion of non-informational reads likely depends on size of target region and fragment lengths of the NGS library. (B) Effect of sequence capture of shotgun libraries using long probes (170 bp) in combination with short end sequencing on local coverage depth distribution [18] (see also Fig. 3). Stringent hybridization selects for fragments that contain a substantial proportion of capture probe sequence. This leads to overrepresentation of fragments for which both ends map near or outside the target region boundaries. Fragments generating end sequencing reads near the middle of the target are underrepresented due to low overlap with capture probes during hybridization. This leads to a dip in the middle of the target region and diminishes the fraction of on-target reads. (C and D) Schematic views of coverage depth distributions as reported for sequence capture using molecular inversion probes (MIP, see also Fig. 2). Both ends of target fragments are fixed, resulting in a non-shotgun library. Direct sequencing leads to an even coverage of redundant reads with identical starting points for small regions where read lengths span the whole target (C). In cases where read lengths are not sufficient to span the target region, middle part is not covered (D). Varying coverage depths for both ends of the regions result from the specific setup of paired end adaptor introduction and sequencing.

Download English Version:

https://daneshyari.com/en/article/2821320

Download Persian Version:

https://daneshyari.com/article/2821320

Daneshyari.com