GETTING WHAT YOU NEED FROM CHIPS AND BONES – GENOME ENRICHMENT TECHNOLOGIES

Artem Men*, Mark Crowe and Kirby Siemering
Australian Genome Research Facility, University of Queensland, St Lucia QLD 4072
*Corresponding author: artem.men@agrf.org.au

Introduction

All major biological landmarks in the 20th and 21st centuries stand on the discovery of life's essential molecules, such as vitamins, peptides, antibiotics, hormones and, of course, nucleic acids. Both DNA and RNA revealed amazing structural uniformity and simplicity, making them favourite experimental subjects for molecular and biochemical studies for more than half a century. Unraveling DNA's structure in 1953 eventually led to the invention of the Sanger method of sequencing and sequencing of the first genome, the bacteriophage ΦX174, in 1977. Sequencing throughput subsequently grew dramatically on the back of advances in automation allowing deciphering of the genomes of simple phages and microbes through to the completion of human and several other eukaryotic genomes. Those achievements have allowed us to envisage completely new paradigms for the study of biological entities; the complex phenotypic diversity of even single cell organisms no longer prohibits holistic interpretation of their natural history, which can now be deduced from comparison of changes in nucleotide composition.

We are currently at the beginning of the next global revolution in the field of genomics. Coalescence of the disciplines of biotechnology and micro/nanotechnology has resulted in technological innovations that are providing possibly the most significant step-change the field has seen. The technology is driving a transformation of many diverse areas of the life sciences, allowing analyses of genomes, genetic variation, gene expression, DNA modification and other biological processes at unprecedented scale and resolution.

The advent of 'next-generation' sequencing technology in particular means that experiments that were previously not thought to be technically or economically feasible are now being done on a mind-boggling scale. This technological revolution has resulted in a paradigm shift from a model of gene-by-gene analysis to one of genome-by-genome analysis. For example, in moving towards the future of personalised medicine, an international consortium has set out to survey and catalogue human variation by resequencing 1,000 human genomes. Similarly, the International Cancer Genome Consortium is aiming to comprehensively catalogue variation within 50 different tumour types by genome resequencing (1).

Bioinformaticians are also coming to grips with the substantial challenge of *de novo* assembly of complex genomes from next-generation sequence data, with assembly of the genome of the giant panda from exclusively next-generation sequencing data being published earlier this year (2). This is expected to herald an explosion in the sequencing of novel genomes and pave the way for large-scale re-sequencing projects as researchers start to investigate inter-individual variation.

However, despite these tremendous technological advancements, re-sequencing of genomes is still hampered by genome complexity (3) and relatively high costs, with the total cost of re-sequencing and computationally mapping a complex genome still in the tens of thousands of dollars. To address this challenge, a cost-effective solution has been developed recently in the form of genome enrichment technology. This approach provides a powerful means of comparative genomics in population studies by allowing targeted capture of only selected genomic regions of interest, thereby reducing the end sequencing and data analysis costs.

Genome Enrichment Technologies

Three important factors for consideration in sequencing projects are cost per base, time and adequate computing power to analyse the data output. Although whole genome re-sequencing yields a comprehensive catalogue of variation between individuals, in many cases only a partial genome analysis is required, for example, re-sequencing of the protein-coding component of the genome (the exome), or re-sequencing a particular region containing a loss-of-function mutation defined by mapped molecular markers. In these cases, full genome re-sequencing is clearly overkill in terms of both costs and information complexity.

In order to address this issue, new genome enrichment technologies have been developed with the aim of reducing genome complexity to the specific fraction of interest prior to sequencing. Genome enrichment technologies can be broadly divided into two categories, hybridisation-based and PCR-based.

Hybridisation-based Technologies

Hybridisation-based methods, often referred to as 'genome capture', were introduced independently by several groups in 2007 (4-7). The idea behind genome capture was that if subregions of well-characterised genome are to be interrogated, then one can design *in silico* a set of overlapping single-stranded probes (oligonucleotides that could be either RNA or DNA) covering (tiling) the region of interest, and then either covalently attach probes to a slide surface, or tag 5'-ends with a biotin tag and use for genome capture in solution. The approach is similar to reverse Southern blotting and is

very straightforward, as is the upfront sample preparation. Genomic DNA of interest is sheared to small fragments, then specific adapters are ligated onto those fragments, and the whole sample undergoes hybridisation to the set of baits (on-array or in-solution). If the array represents a unique set of probes, then only target DNA fragments hybridise to the baits (**Fig. 1**). The non-hybridised sample is washed away, the hybridised fraction eluted, and the eluate amplified to nanogram amounts to enter the next-generation sequencing pipeline. Sequencing reads are then mapped to the region of interest, undergoing another round of filtration against the reference genomic target to discard off-target reads.

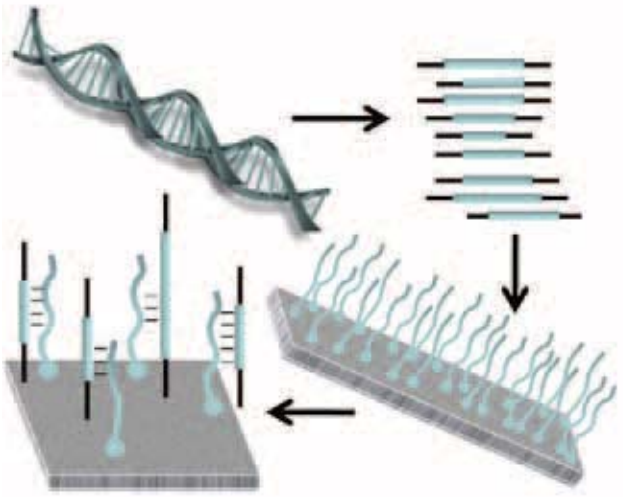

Fig. 1. Schematic representation of 'on-array' genome capture.

Input genomic DNA is sheared to sizes acceptable by the next-gen sequencing platform; pieces are repaired and blunt-ended with DNA polymerase and polynucleotide kinase. Adapters specific for subsequent library production are then ligated onto both ends of polished fragments (at this stage the mixture can be converted into a single-stranded DNA pool). The pool is amplified by PCR, followed by hybridisation onto an array containing capture probes (baits). The slide is subjected to washing off of non-specific fragments, sample elution, secondary PCR and sequencing of captured fragments.

The system is simple and versatile, on the conditions that **a.** a high-quality, or finished, reference genome is readily available for the original design, and **b.** *in silico* repeat masking of the region can effectively remove repeated DNA regions without affecting representation of the target loci in the final sequencing output. If the upfront design is well optimised, the genome can truly be stripped to its essentials. Even the most complex eukaryotic exome usually represents only about 1% of the whole genome, and often targets of interest are spread across only a few megabases of DNA that have to be interrogated in order to find a mutation in the locus of interest.

Three suppliers currently dominate the genome capture market (**Table 1**). Roche-NimbleGen and Agilent were first to release products and currently provide both on-array and in-solution products (**Table 1**). Both companies market human exome kits, targeting approximately 35Mbp of human exons in a single reaction. In addition, researchers can select and design their own specific target regions of interest, comprising up to 30Mbp of contiguous or distributed sequence. The third player who recently entered the market is Febit Biomed, offering an automated HybSelect protocol that combines genome capture with microfluidic Biochips and only takes 30 minutes of hands-on time, according to the company website (**Table 1**). Driven by high demand for re-sequencing of human exomes, major sequencing centres are rapidly picking up on the technology; the Broad Institute, Wellcome Trust and Washington University will all ramp up their exome sequencing to 10,000 human individuals by 2011.

PCR-based Technologies

Although they are a very effective method for genome enrichment, one of the biggest limitations of hybridisation-based methods is that they suffer from sequence-specific biases that are hard to control and optimise. This is related to the very core of the process, such as the kinetics and temperature-dependent thermodynamics of hybridisation (including differences in nucleotide composition of different targets, GC content, etc). As a result, after genome capture, some target areas can end up with insufficient or even no coverage and others with massive oversampling.

To address this issue, where complete and even coverage of the target is a must and the region is reasonably small, a number of PCR-based genome enrichment methods have been developed as alternatives to genome capture methodology. One of the simplest options is targeted long-range PCR, when the genomic region can be covered by tiling amplicons each about 10kbp in size, followed by shearing of those amplicons and using them as templates for next-generation library production. The difficulty of this method is that laborious PCR optimisation is required for every amplicon. Multiplexing of amplicons is usually impractical due to accumulation of non-specific products when multiple pairs of primers are included in a single tube. Additionally, products have to be normalised before pooling them together for library production. The advantage of long-range PCR is that once the performance of each amplicon is optimised, barcoding of samples can be introduced, and multiple pooled samples can be analysed in a single sequencing run. Several companies are developing targeted PCR approaches that can compete with genome capture. For example, in April 2010, Fluidigm Corporation released the Access Array System that automatically generates 48 uniquely barcoded samples from 48 DNAs and 48 primer sets, ready for sequencing on the 454/Roche FLX platform. This fully automated system can capture up to 24kbp of sequence data per sample, or 1.15Mbp per integrated fluidic circuit device.

Another recent genome enrichment application that allows researchers to perform multiplex PCR amplifications of medium-size genome targets is the RainDance technology. The method takes advantage of emulsion PCR-type microreactors where each emulsion microdroplet contains PCR chemistry, DNA template and a unique pair of primers. The proof of principle,

Table 1. Current capability of genome capture platforms.

Company	Setup	Target Region	Scalability and Automation	Starting Amount of DNA
Agilent	In-solution RNA probes	Up to 6Mbp per reaction (120-mers) or 37Mbp of human exons	Highly scalable	3µg
Roche-NimbleGen	*Array:* DNA probes (micromirror assembly) *In-solution:* SeqCap EZ Exome captures	*Array:* Up to 30Mbp per array (human >60-mer probes) *In-solution:* 39Mbp of human exons (>10 probes/ exon on average)	*Array:* Up to 12 samples on the Hyb System 12 platform *In-solution:* Up to 96 samples at once	1–5µg
Febit Biomed	Array (currently compatible only with SOLiD sequencing)	Up to 2Mbp regions (120kpb probes in eight separate Biochip channels)	Up to 16 pooled barcoded samples on the automated Geniom RT Analyzer	1µg (or less if whole-genome amplification is applied)

where about 4,000 amplicons were sequenced, was first published by Tewhey *et al.* (8), and the most recent publication describes the same approach in which about 2,000 amplicons were designed to cover 86 previously identified XLMR genes (9). The latest offering from RainDance Technologies is a 20,000 primer set for a single tube run covering up to 10Mbp of target. The approach has a great deal of flexibility and may be well suited to many applications where even coverage of the target region is important, such as diagnostic screening.

Finally, another PCR-based method with a twist has recently been developed. This method employs molecular inverted probes (MIPs) and was originally suggested to increase the specificity of target hybridisation and to ramp up multiplexing. MIPs are 70-base oligos that for all probes contain an identical and non-specific linker sequence flanked by two 20-base target-specific 'arms' After hybridisation to the target, the space between two arms is extended by a mixture of DNA polymerase and ligase. Residual linear pieces are then digested with exonuclease before the entire pool is amplified with primers specific to the linker (**Fig. 2**). The design of the probes allows the user to skip the downstream library preparation and the pool can be immediately sequenced on a next-generation sequencing platform. Because, like PCR, the hybridisation event is based on a 'double' annealing event (from both arms), in theory, the specificity of the capture should be higher than for a single annealing event. Unfortunately, this was not borne out by early trials, but more recent publications and protocols are now showing greatly improved specificity (10). Also, according to the publication, MIPs offer more flexibility than conventional genome capture probes; for instance, in order to circumvent hybridisation bias, MIPs can be grouped into sets based on similar capture efficiencies because biases tend to be systematically reproducible. With a current DNA input of only 200ng and with expectations to reach 98% on-target capture (versus 70-80% for other genome capture platforms), MIP technology is probably one to watch in the near future.

Limitations of Genome Enrichment Technologies

Theoretically, genome enrichment methodology is only limited by the size and complexity of the genomic target. In practice, both hybridisation-based and PCR-based techniques have a number of limitations. As outlined above, hybridisation-based methods suffer from sequence-specific hybridisation biases that mean that some targets may be inefficiently captured or not at all. The consequence of this is that higher levels of overall sequence coverage may often be required to obtain sufficient representation of poorly captured targets. Furthermore, all design has to go through careful repeat masking; therefore, potential targets located in centromeric or other repeated regions might be rejected by the software design. Hybridisation-based protocols also usually involve PCR amplification of the final capture eluate, potentially introducing further bias due to preferential amplification of certain regions depending on nucleotide composition of amplimers. Hybridisation-based methods may also cause allelic imbalances or drop-out. Because of the effects of mismatches on hybridisation efficiency, one allele of a polymorphic target may be preferentially captured over the other, resulting in false homozygote calls. Finally, sequencing libraries made from hybridisation-based capture methods usually result in a significant proportion of 'off-target' reads. These are thought to result from illegitimate hybridisation events or 'daisy chaining' of DNA fragments onto the ends of captured fragments.

Although PCR-based methods overcome many of the challenges associated with hybridisation-based methods, they are fundamentally limited in the target space they can cover by the requirement to synthesise large numbers of target-specific primer pairs and carry out PCR in a highly parallel fashion. Technology platforms such as Raindance that utilise microfluidic circuits offer the greatest scope for improvement in this regard.

Finally, the one fundamental drawback of all genome enrichment technologies is that because they only allow interrogation of about 1% of genome at one time, there is a risk that biologically important genome polymorphisms

may escape discovery due to inefficient capture or if they do not lie in the target region. Also, these technologies are not well suited to detecting polymorphisms such as copy number variations and other complicated genome structural rearrangements.

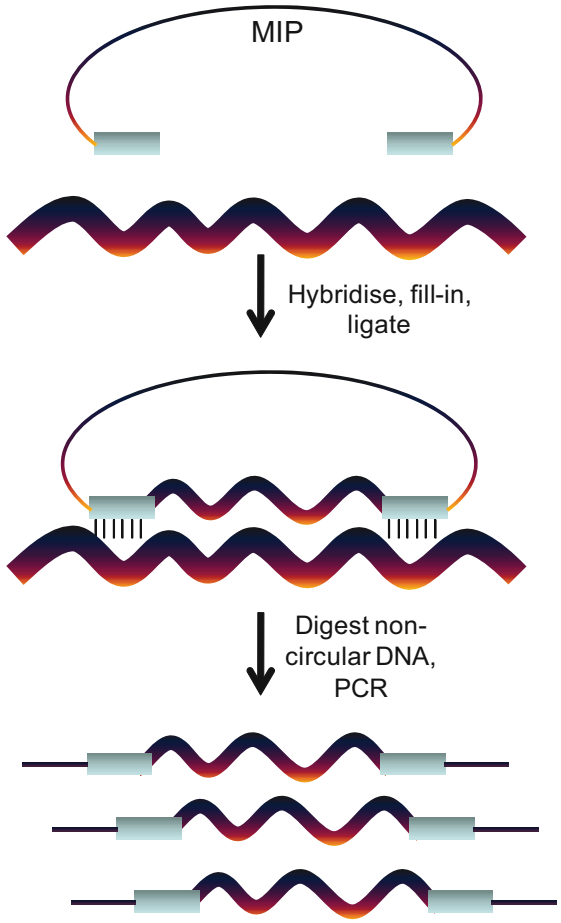

Fig. 2. Molecular Inversion Probe (MIP) technology.
MIP is a single-stranded DNA probe consisting of two target-specific fragments (light blue) that flank a 'spacer' that is common for all MIPs and contains primer-annealing sites for the subsequent PCR of captured material. Specific fragments anneal at both sides of the genomic target, and the MIP is filled in and circularised by DNA polymerase and ligase. Non-circular fragments are digested away, and the target DNA can be amplified and sequenced without the need for library preparation. MIP technology fits nicely into the 'few targets, many samples' niche, as it is conducted in-solution and can be formatted for microtitre plates and standard laboratory robotics.

Success Stories

There are already several publications that illustrate the utility of genome enrichment technology. Here, we would like to present three interesting examples of how genome capture technology has been used in different ways. The first example deals with maize, an important agricultural crop but also a 'genomic nightmare' as 85% of its genome comprises repetitive sequence. Instead of going into the cumbersome preparation of maize Cot-1 DNA to filter out repeats, Fu and co-authors (11) applied a double subtraction protocol, where a Roche-NimbleGen array covered with 720,000 oligos representing repeats was hybridised first, followed by the genome capture protocol on 42,000 probes derived from 70 tiled bacterial artificial chromosomes covering the region. As a result, in combining Roche-NimbleGen genome capture with Roche FLX sequencing, a 2.2Mbp chromosomal interval containing 43 genes was sequenced and annotated. Undoubtedly, these results give a lot of hope to those working with other agricultural crops with challenging genomes, such as wheat and barley.

Another paper appearing in 2009 in *Nature* described a re-sequencing experiment of a 27Mbp exome target in 12 individuals, four of whom were affected by Freeman–Sheldon syndrome (FSS). Through multiple rounds of filtering of single nucleotide polymorphisms (SNPs) found in the study, several coding SNPs were found in the MYH3 gene sequence of all affected individuals, thus confirming it as a molecular cause for FSS (12). This finding is particularly remarkable because **a.** FSS is a dominantly inherited disorder (only one mutant allele is sufficient to cause the disease), and **b.** there was no degree of kinship between the individuals in the cohort. The authors speculate that gene identification was facilitated by the availability of eight reference exomes in the HapMap database that enabled them to carefully curate all found SNPs. Studies like this show that it is possible to clone human disease genes for which no prior linkage data or candidate gene exists.

The last interesting example is the recent report on genome capture of a partial Neanderthal exome. Burbano *et al.* (13) designed a one million feature Agilent array and recovered more than a megabase of target regions from Neanderthal bone that was 99.8% contaminated with microbial DNA. They were able to identify 88 amino acid substitutions between modern humans and the Neanderthals, which are being investigated further. Again, this example shows the power of genome capture not only to capture essential information in a cost and time efficient manner, but also to filter out the high levels of contamination that make it impractical to sequence original sample by standard shotgun methods.

Conclusion

Genome enrichment technologies have proven themselves a valuable addition to the arsenal of next-generation sequencing related technologies. Within their inherent limitations, genome enrichment technologies provide a powerful means of comparative genomics in population studies by allowing targeted capture of only selected genomic regions of interest, thereby reducing the end sequencing costs. Despite this, as sequencing and data analysis costs continue to drop, at some point it will become more effective to simply sequence whole genomes. Until this time comes, however, genome enrichment technologies continue to play an important role in the field of next-generation sequencing.

References continued on page 21

References continued from page 31

1. Nicholls, H. (2008) *Nat. Biotechnol.* **26**, 722
2. Li, R.Q., Fan, W., Tian. G., *et al.* (2010) *Nature* **463**, 311-317
3. Dolgin, E. (2009) *Nature* 462, 843-845
4. Albert, T.J., Molla, M.N., Muzny, D.M. et al. (2007) *Nat. Methods* 4, 903-905
5. Hodges ,E., Xuan ,Z., Balija, V., *et al.* (2007) *Nat. Genet.* **39**, 1522–1527
6. Okou, D.T., Steinberg, K.M., Middle, C., *et al.* (2007) *Nat. Methods* **4**, 907-909
7. Porreca, G.J., Zhang, K., Li, J.B., *et al.* (2007) *Nat. Methods* **4**, 931-936
8. Tewhey, R., Warner, J.B., Nakano, M., *et al.* (2009) *Nat. Biotechnol.* **27**, 998-999
9. Hu, H., Wrogemann, K., Kalscheuer, V., *et al.* (2009) *HUGO J.* **3**, 41-49
10. Mamanova, L., Coffey, A.J., Scott, C.E., *et al.* (2010) *Nat. Methods* **7**, 111-118
11. Yan Fu, Y., Springer, N.M., Gerhardt, D.J., *et al.* (2010) *Plant J.* **62**, 898-909
12. Ng, S.B., Turner, E.H., Robertson, P.D., *et al.* (2009) *Nature* **461**, 272-278
13. Burbano, H.A., Hodges, E., Green, R.E., *et al.* (2010) *Science* **328**, 723-725