Pathology (January 2014) **46(1)**, pp. 60–68

MOLECULAR PATHOLOGY

Diagnostic validation of a familial hypercholesterolaemia cohort provides a model for using targeted next generation DNA sequencing in the clinical setting

Marcus Hinchcliffe[1], Huong Le[1], Anthony Fimmel[1], Laura Molloy[1], Lucinda Freeman[1], David Sullivan[2] and Ronald J. Trent[1,3]

[1]Department of Molecular and Clinical Genetics, Royal Prince Alfred Hospital, Camperdown, [2]Department of Clinical Biochemistry, Royal Prince Alfred Hospital, Camperdown, and [3]Sydney Medical School, University of Sydney, Sydney, New South Wales, Australia

Summary

Our aim was to assess the sensitivity and specificity of a next generation DNA sequencing (NGS) platform using a capture based DNA library preparation method. Data and experience gained from this diagnostic validation can be used to progress the applications of NGS in the wider molecular diagnostic setting. A technical cross-validation comparing the current molecular diagnostic gold standard methods of Sanger DNA sequencing and multiplex ligation-dependant probe amplification (MLPA) versus a customised capture based targeted re-sequencing method on a SOLiD 5500 sequencing platform was carried out using a cohort of 96 familial hypercholesterolaemia (FH) samples. We compared a total of 595 DNA variations (488 common single nucleotide polymorphisms, 73 missense mutations, 9 nonsense mutations, 3 splice site point mutations, 13 small indels, 2 multi-exonic duplications and 7 multi-exonic deletions) found previously in the 96 FH samples. DNA variation detection sensitivity and specificity were both 100% for the SOLiD 5500 NGS platform compared with Sanger sequencing and MLPA only when both LifeScope and Integrative Genomics Viewer softwares were utilised. The methods described here offer a high-quality strategy for the detection of a wide range of DNA mutations in diseases with a moderate number of well described causative genes. However, there are important issues related to the bioinformatic algorithms employed to detect small indels.

Key words: Familial hypercholesterolaemia, FH, indel, MLPA, next generation DNA sequencing, NGS, SNP, SNV.

Received 19 March, revised 23 October, accepted 23 October 2013

INTRODUCTION

The decision to introduce next generation DNA sequencing (NGS) technology into routine molecular diagnostics requires careful planning. Considerations include: (1) clinical scenarios for testing; (2) available laboratory expertise; (3) platform(s) to use (including each platform's respective read accuracy and output, complexity of each instrument and associated fixed and variable costs); (4) informatic resources and requirements; and (5) turnaround time. As the raw cost of DNA sequencing continues to plummet it is now tempting to consider exome or even whole genome sequencing for a wide variety of germline related clinical presentations.[1–3] However, many common Mendelian disorders have a small to moderate number of well characterised causative genes. In these clinical situations, a strategy that uses a more targeted NGS approach will present greater localised depth of DNA sequencing coverage and therefore better detection sensitivity and specificity. This approach will also deliver a smaller and better targeted list of DNA variants to sift through for each patient and therefore reduce the burden of DNA variant interpretation for the laboratory. The concern about incidental findings is not relevant with targeted DNA sequencing

There are many NGS platforms available and a number of ways to target genetic loci of interest, including long polymerase chain reaction (PCR),[4] multiplexed PCR[5–7] and hybridisation capture based methods.[8–9] We chose to validate the SOLiD 5500 sequencing platform with the TargetSeq customised capture system for the following reasons.

SOLiD 5500 (Life Technologies, USA): (1) raw read accuracy of the ligation based chemistry is high with a low false positive raw read indel rate; (2) DNA sequencing output, whilst not as large as some other platforms, is adequate for batched targeted resequencing of a moderate set of genes on multiple barcoded samples; (3) output flexibility—one to six lanes can be used per run and each lane can have up to 16 samples at reasonable DNA read depth (coverage).

TargetSeq Custom Enrichment Kit (Life Technologies, USA): (1) no potential primer binding issues unlike PCR based methods; (2) high level of flexibility available for customised probe design with four different sized kits offered (100–500kb, 500kb–2Mb, 2–10Mb, >10Mb); (3) capture probe density has on average three separate probe sequences of length 50–110 bp (DNA based) for each locus, reducing the reliance on any single capture probe sequence; (4) barcoding of four samples per capture offered time and cost efficiencies; (5) with adequate coverage, there was the possibility of using the relative read depth for quantitative assessments as a substitute for multiplex ligation-dependant probe amplification (MLPA) which is needed for detecting multi-exonic deletions or duplications.

The technical cross-validation protocol described here generated data that were analysed with the LifeScope mapping and variant calling software (Life Technologies, USA) combined with the US National Institutes of Health (NIH) Exome Sequencing Project DNA variant data as well as localised relative read depth comparisons for multi-exonic

Print ISSN 0031-3025/Online ISSN 1465-3931
DOI: 10.1097/PAT.0000000000000026

deletion and duplication detection. Mapping alignments using the publically available Integrative Genomics Viewer software (IGV; Broad Institute, USA) were also utilised as a visual tool for a second variant identification method. Data generated were compared to results obtained previously on the same 96 FH samples using Sanger sequencing and MLPA.

Clinical samples

FH is an autosomal dominant genetic disorder resulting from mutations in one of several genes that affect the function of low-density lipoprotein (LDL) receptors that remove LDL from the blood. Patients with heterozygous FH have LDL cholesterol levels that are approximately twice as high as unaffected individuals. As a result, cardiovascular disease (CVD) is accelerated by two to four decades, resulting in premature coronary heart disease, stroke and peripheral vascular disease. Clinical diagnosis is based on a personal and family history of hypercholesterolaemia, with or without physical signs, such as tendon xanthomas and a history of premature CVD events. The sensitivity of physical signs is limited, particularly in childhood, whilst LDL cholesterol levels are non-specific, especially in later adult life. Diagnostic certainty is required as a basis for family screening and lifelong therapy, so molecular genetic tests offer an appealing means for improved diagnostic capability.[10] This is important as it has been shown that early (pre-adult) intervention is valuable.[11] Approximately 1 in 500 people across all racial groups have FH.[12] The Australian population has over 45,000 affected individuals, some of whom will have been treated from a clinical diagnosis alone. However, the great majority of patients with FH have not had their respective mutation(s) identified and many remain undiagnosed. Indeed, given that early intervention can have a significant impact upon life expectancy, any scheme that appreciably increases the identification and early treatment of such a prevalent and serious condition is important. Hence, a further aim of the method described here was to increase the number of cases that can be processed in a laboratory by utilising individual sample nucleotide based barcoding and NGS. A valid and efficient new methodology will also have wider relevance for molecular diagnostics beyond FH.

MATERIALS AND METHODS

FH cohort

The study cohort comprised patients attending a lipid clinic that specialises in the detection and management of FH. Samples were collected from patients who gave informed consent for genetic testing. Approval was also gained from the Hospital Ethics Clinical Practice Subcommittee to introduce NGS for targeted molecular diagnostics. Our DNA samples came from patients who were regarded as 'probable' or 'definite' FH, as determined by the Dutch Lipid Clinic's score >5.[13] Sanger DNA sequencing was obtained for all exons and the exon-intron boundaries in the *Low-Density Lipoprotein Receptor* (*LDLR*) gene and part of exon 26 of the *APOB* gene as well as MLPA testing of the *LDLR* gene if no mutation was found on initial sequencing. Previous experience with DNA sequencing has suggested that the rate of mutation detection in this group exceeds 70%, with mutations in the *LDLR* gene accounting for over 90% of these. Of the laboratory's large FH patient cohort, 96 were chosen for this cross validation specifically to include a wide range of mutation types (Table 1).

Eighty patients had a single heterozygous mutation. Twelve patients were double heterozygotes (the second identified variations in at least seven of these were likely benign), and four had no identified pathological mutation after Sanger sequencing and MLPA. The majority of the variations included in this technical cross-validation were common single nucleotide polymorphisms (SNPs). These common SNPs were equally valid and useful for determining this NGS method's technical sensitivity and specificity. All specimens were numbered and processed through the laboratory's NGS pipeline in a blinded sample procedure.

Table 1 List of DNA variation types in validation

Variation type/subcategory	No.
Point mutations	
Common SNPs (coding and non-coding)	488
Missense mutations	73
Nonsense mutations	9
Splice-site mutations	3
Small indels	
Frameshift	8
In-frame	5
Large duplications (multi-exonic)	2
Large deletions (multi-exonic)	7
Total	595

SNPs, single nucleotide polymorphisms.

Laboratory methods

All 96 specimens initially had their *LDLR* gene exons and exon-intron boundaries and part of exon 26 of the *APOB* gene progressively Sanger sequenced beginning with the 'hotspot' exons. If a mutation was found early in that process that matched the phenotype, then sequencing was stopped. As a consequence, not all 96 specimens had their entire *LDLR* gene sequenced. Only the regions that were originally Sanger sequenced were then analysed in the subsequent NGS comparison. Samples where a pathogenic mutation was not identified after completion of the Sanger sequencing had an MLPA assay targeting the *LDLR* gene to detect large deletions and duplications.

Re-extraction of DNA from the original blood samples into a low TE buffer was carried out for the NGS cross-validation process. Samples were sequentially numbered and fully processed through to the end of the bioinformatic stage before result cross comparisons were completed.

The 96 NGS DNA samples in 2 μg aliquots were fragmented on a S220 sonicator (Covaris, USA) and then barcoded adapters were ligated using a standard automated Library Builder assay (Life Technologies). Libraries were nick-translated and amplified with seven amplification cycles and then excess adapter and primers were removed using Agencourt AMPure XP beads (Beckman Coulter, USA). Samples were then quantified on a high sensitivity dsDNA chip using a 2100 Bioanalyzer (Agilent Technologies, USA) in addition to a SOLiD Library TaqMan quantitation assay (PN A12120; Life Technologies).

A TargetSeq Custom Enrichment Kit (PN A14227; Life Technologies) was specifically designed to have a greater fold coverage over the exonic regions of the *LDLR* and *APOB* genes versus the intronic and intergenic regions in a ratio of eight (exons) to three (intronic, upstream and downstream regions). This enabled greater certainty in capturing the coding regions whilst allowing the possibility of using the intronic and intergenic sequencing read depth to detect a relative coverage alteration that would indicate the presence of a multi-exonic deletion or duplication. The customised kit was rationalised for the absence of genomic homologs and repeats.

Samples were pooled in 24 groups of four in an equimolar amount (125 ng per sample, total of 500 ng). Capture hybridisation was carried out using the TargetSeq Custom Enrichment Kit. The 96 barcoded samples were captured in 24 RNase free PCR tubes at 47°C over 72 h then washed, amplified by 10 cycles and cleaned using Agencourt AMPure XP beads. The 24 pooled post-capture samples were then quantified on a high sensitivity dsDNA chip using the Agilent 2100 Bioanalyser and a SOLiD Library TaqMan quantitation kit. A series of SYBR green qPCR assays were run at this stage on four *LDLR* targets and compared the delta C_T results with matching TargetSeq Exome capture kit SYBR green results to gauge the relative effectiveness of using this narrowly targeted capture method compared with a more diffuse exome capture process.

The 24 pooled sample libraries were then processed through a series of six standard E20 emulsion and enrichments (Life Technologies) such that there