Seeking Out Dyslipidemia Variants with LipidSeq

Robarts Research Institute researchers design a custom targeted resequencing panel to profile lipid-related genetic variants.

Introduction

Cardiovascular disease is one of the leading causes of death in North America.[1] There are numerous genetic disorders and mutations that predispose people toward cardiovascular disease risk factors, such as high cholesterol, high blood pressure, and diabetes. Many of these genetic defects involve lipid metabolism and the regulation of blood lipid levels. Identification of such genetic variants can facilitate patient intervention and increase understanding of the molecular biology of the underlying disease.

Traditionally, genetic variants have been identified with capillary electrophoresis (CE)/Sanger sequencing. However, this iterative process is time-consuming and labor-intensive. New technologies, such as next-generation sequencing (NGS), enable scientists to increase the scope of their studies and shorten the path to discovery.

For over 20 years, Robert Hegele, MD, Director of the Blackburn Cardiovascular Genetics Laboratory at the Robarts Research Institute in London, Ontario, Canada has studied genetic variants associated with dyslipidemias and related metabolic disorders using Sanger sequencing. In 2013, his team transitioned to NGS with the MiSeq® System and Nextera® Rapid Capture Enrichment. iCommunity spoke with Research Manager John Robinson about the development of their LipidSeq targeted resequencing panel and how they are using it to discover new variants.

Q: When did you become involved in lipid research?
John Robinson (JR): I've worked at the Robarts Research Institute for about 22 years and with our Director, Dr. Rob Hegele, for over 12 years. Dr. Hegele is an endocrinologist, specializing in lipid disorders such as dyslipidemias and lipodystrophies in humans.

Q: How did you gather data on potential causative variants of lipid disorders initially?
JR: Dr. Hegele has close to 12,000 distinct human DNA samples in our repository. The typical process workflow begins when Dr. Hegele sees a patient in his clinic with an abnormal blood lipid level. After the patient agrees to become a research subject and signs an informed consent form, we obtain a sample. Based on their clinical diagnosis, lipid levels, and family history, our lab staff chooses candidate genes to investigate.

For example, if there is evidence that the research subject might have familial hypercholesterolemia, we would sequence the low-density lipoprotein receptor gene (*LDLR*). We'd choose an exon that we suspected might house the causative variant, and then move to the other exons in the gene. If we didn't find a mutation in the *LDLR* gene, we'd move to the next gene in the lipid cascade. Each of these analysis workflows was performed exon by exon and 1000 bp at a time. Some of the genes were extremely large, 40–50 exons. It could take up to a month to Sanger sequence each individual DNA sample before we found the causative variant. In addition to research time, there were the costs and the labor to perform the sequencing, sample prep, PCR, etc.

Q: What drove your decision to move to NGS-based targeted resequencing?
JR: About 4 years ago, we became aware that resequencing panel chemistries were available on NGS systems. Our lab staff (which includes PhD students, research associates, technicians, and Dr. Hegele) put together a wish list of genes that we wanted to see represented on a sequencing panel.

We thought of the panel content consisting of several tiers. The first tier was canonical monogenic dyslipidemia genes that have been fairly established. These are genes that if mutated are probably causative for the phenotype. This first tier of genes encompasses elevated low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), and high triglyceride (TG) phenotypes. These are typically the phenotypes that Dr. Hegele sees in the clinic.

John Robinson is the Research Manager in Dr. Robert Hegele's laboratory at the Robarts Research Institute in London, Ontario, Canada.

Dr. Hegele also sees patients that have dyslipidemias due to lipodystrophies, and inherited forms of diabetes, such as maturity onset diabetes of the young. The second tier of the gene panel included genes associated with these secondary causes of dyslipidemia. The third tier was lipid genes for our interest and discovery. These are genes identified in animal models, but where no human gene variants have been described. These 3 gene tiers formed the wish list for the 69 gene panel we now call LipidSeq. We've published papers describing it, including one in the Journal of Lipid Research.[2,3]

> "About 4 years ago, we became aware that resequencing panel chemistries were available on NGS systems. Our lab staff put together a wish list of genes we wanted."

Q: How are you using the panel for variant discovery?
JR: We currently process 24 samples per week on one MiSeq run. The MiSeq System produces 24 pairs of FASTQ files that are then aligned, locally realigned, have PCR duplicates removed, and then have variants called to produce 24 VCF files. The VCF files are annotated so that we can then identify familiar and novel variants that can be attributed to causing the phenotype.

Q: Can you identify polygenic as well as monogenic variants?
JR: Yes, we can identify polygenic variants using the LipidSeq panel. We identified another dimension of genetic susceptibility to a particular trait, called polygenic trait scores. We use a single nucleotide polymorphism (SNP) target identified in a published genome-wide association study (GWAS), such as a report that a GWAS SNP in a research subject had a different allele pattern, with 1 of the alleles causing the trait to be raised in that subject. Previously, we performed those using TaqMan assays, and we would assay a population for a single SNP. For example, in addition to several genes on the panel that are associated with elevated LDL-C, we also have 10 GWAS SNP targets on the panel associated with elevated LDL-C. For each SNP target, a subject that is homozygous for the nontrait raising LDL-C allele would score a 0. A subject would score a 1 for a heterozygous high LDL-C allele, and a 2 for a homozygous high LDL-C allele. For each individual SNP, subjects could score a 0, 1, or 2. If there are 10 SNPs that make up a trait, subjects could have a polygenic trait score between 0 and 20. We then apply a weighting factor.

Q: What challenges did you face in developing the LipidSeq panel?
JR: It seems trivial, but we had to learn how to generate a BED file, which is an Excel file saved as a text file, so that we could design capture probe files for the exon regions of the genes on the LipidSeq panel. We used to map regions of every alternate isoform of a given gene manually by curating tracks of the University of California, Santa Cruz (UCSC) Genome Browser. The issue was how to get a BED file that would take into account the alternate isoforms of some of the transcripts. If we didn't do it correctly, it could compromise the ability of the oligos to capture DNA accurately. We also learned through design iterations that we had to adapt to GC content and repetitive sequence regions to obtain enough depth of coverage to make accurate genotype calls. It was critical that the BED file accurately represented the areas of interest for the probes on the panel. Ultimately, we created scripts that assemble those types of BED files for us. We're using the third version of our design for the LipidSeq panel and are hoping to implement a fourth version in 2017.

Q: What was it like to work with DesignStudio™ Software to develop the LipidSeq panel?
JR: DesignStudio Software was easy to use in designing the LipidSeq panel. After we had BED files designed and the right base pair padding based on our exon boundaries, we realized that our panel comprised of 700 kb of capture oligos for coding exons and SNPs. Then we started to take advantage of the attributes of the MiSeq System. We looked at sample multiplexing and the different sequencing kits to determine how many actual reads we would obtain, and therefore the depth of coverage we were going to achieve.

Q: What is the workflow for the LipidSeq panel?
JR: We process 24 samples a week with about a 2-week turnaround time. Diluting DNA from genomic concentrations down to 5 ng/µl is not trivial, but it's doable, so that takes a couple of days to accomplish. Initially, we ran 12 human DNA samples through the Nextera Rapid Capture Custom Enrichment Kit at the same time. With the advent of the MiSeq v3 reagent kits, we are now able to run 24 samples simultaneously.

Currently, it's a 2-week process. On days 1–3, we run 24 DNA samples using the Nextera Rapid Capture Custom Enrichment Kit, then we sequence them on the MiSeq System at 24-plex. Twenty-four pairs of FASTQ files are delivered to our servers and we use the CLC Genomics Workbench software to batch process those, creating the VCF files that go into our annotation pipeline.

> "It could take up to a month to Sanger sequence each individual DNA sample before we found the causative variant. In addition to research time, there were the costs and labor."

Q: How does the output and turnaround time of the LipidSeq panel on the MiSeq System compare to using Sanger sequencing to identify variants?
JR: The output and turnaround time of LipidSeq panel versus Sanger sequencing isn't even comparable. Sequencing 24 samples with the LipidSeq panel on the MiSeq System delivers 700 kb of sequencing data per person. Within two weeks, we have files that are annotated and categorized for rareness and

pathogenicity, and include monogenic and polygenic variants. Sanger sequencing gives only a 1000 bp per run, which is a best guess for the exon with the causative variant. Plus, you have to use TaqMan assays to obtain polygenic trait scores.

Q: How is the data quality that the MiSeq System produces and how is the data analyzed?
JR: The MiSeq System has worked well, and it produces good data sets. Our bioinformatics software can handle batches of samples simultaneously and produce good quality VCF and annotation files. After we have those files, we ask our research associates to curate the individual data. Ultimately 700,000 nucleotides of capture are reduced to about 800 nucleotides or scaffold positions on a VCF file. Those 800 nucleotides get reduced to about 20 candidates for deleteriousness and rareness for causality for various dyslipidemia syndromes. At that point, our staff is curating a small subset that then forms a report that Dr. Hegele can communicate back to his research subjects.

> "The output and turnaround time of the LipidSeq panel versus Sanger sequencing isn't even comparable. Sequencing 24 samples with the LipidSeq panel on the MiSeq System delivers 700 kb of sequencing data per person…Sanger sequencing gives only 1000 bp per run."

Q: How do you add new loci to the LipidSeq panel?
JR: We create a standard BED file in the DesignStudio Software import format that has all the isoforms and duplicate probe designs built into the spreadsheet. DesignStudio Software takes over from there and gives me the design. If there's a low coverage or GC content warning, we often have already replicated the probes for that. We can always link out to the UCSC Genome Browser and see what the probes are going to look like.

Q: Have there been any surprises as you've used the LipidSeq panel in your research?
JR: There have been different types of variation patterns in some of the candidate genes and different polygenic trait score profiles, depending on which phenotype we're looking at or the extremity of the phenotype. They are the focus of scientific manuscripts that we're preparing.

The LipidSeq panel enables us to obtain complete data sets on all 2000 research subjects that we've sequenced so far, regardless of phenotype. The VCF files are reporting back 800 to 1000 variants per person. That provides us with the opportunity to move from assessing individual cases (what variation accounts for a subject's phenotype) to association testing using tools like sequence kernel association tests. So we can move from individuals to cohorts and populations, using NGS to perform experiments that other researchers are conducting with microarrays. In our case, we're using the VCF files as the variant call files for the subjects.

Q: Are you using the MiSeq System for any other studies in your lab?
JR: We keep the MiSeq System busy running samples from our lab. We're a customer of the London Regional Genomics Center (LRGC), which is a core lab Dr. Hegele directs at the Robarts Research Institute. Our lab is also acting as a core lab for the Ontario Neurodegeneration Research Initiative (ONDRI), part of the Ontario Brain Institute of Ontario. The MiSeq System is processing samples with their ONDRISeq resequencing panel, which is designed for identifying variants associated with neurodegenerative diseases. The LRGC has another customer who has developed a pharmacogenomics-based panel. The LRGC also sequences many bacterial genomes and transcriptomes, and performs microbiome analysis on their MiSeq System.

Q: What do researchers think of the MiSeq System data?
JR: They like the fact that they can provide us with DNA and RNA samples, and receive good quality data in return. For example, we work with the staff of the LRGC to set up experiments that meet their customers' needs. They like that fact that the MiSeq System can sequence small bacterial genomes and transcriptomes, that they can multiplex samples on one run cartridge to save them money, and ultimately provide them with high-quality data sets. The CLC Genomics Workbench software enables them to perform DNA and RNA analysis of different formats and experimental designs. These data sets produce expression differentials between different conditions in the bacterial transcriptomes and microbiome profiles of complex environments.

> "The MiSeq System became the platform of choice. Customers need high-quality data and lots of it. Other sequencing technologies may have faster turnarounds, but they can't compete with the high-quality data that comes off the MiSeq System."

Q: What made you choose the MiSeq System?
JR: We were one of the first labs in Canada to have an Ion Torrent Personal Genome Machine (PGM). It became obvious to us that the PGM output, I think one chip was 500 Mb and the large chip was 1 Gb in theoretical output, didn't match the 14 Gb theoretical output of the MiSeq System. The MiSeq System became the platform of choice. Customers need high-quality data and lots of it. Other sequencing technologies may have faster turnarounds,

but they can't compete with the high-quality data that comes off the MiSeq System.

Q: How do you choose when to use Nextera Rapid Capture Custom Enrichment Kit or TruSeq® Custom Amplicon?
JR: We were aware of the Nextera Rapid Capture Custom Enrichment Kit and TruSeq Custom Amplicon as we started the process. I found in several proof-of-principle designs that TruSeq Custom Amplicon is based on producing PCR amplicon designs, which are susceptible to repetitive elements and GC content components of the target area in question. If you input 100% of your BED file coordinates when you submit a design for TruSeq Custom Amplicon, you may get a 90–95% success rate on the design. Conversely, if you submit 100% of your BED file to the Nextera Rapid Capture Custom Enrichment, you will obtain probes designed for 100% of that design. Then you have to figure out if you have poor capture in some areas, and if you do, determine how to bump that up. With TruSeq Custom Amplicon, you'll never have that opportunity because it will not design probes for those areas.

We've found that using Nextera Rapid Capture Custom Enrichment has been beneficial in our LipidSeq studies. We've learned to play with the base-pair padding to make sure that we obtain adequate coverage over our target regions. We typically achieve 200–400× coverage over most of our coding regions on our Nextera Rapid Capture Custom Enrichment targets.

> "The LipidSeq panel enables us to obtain complete data sets on all 2000 research subjects that we've sequenced so far…So we can move from individuals to cohorts and populations, using NGS to perform experiments that other researchers are conducting with microarrays."

Certainly, TruSeq Custom Amplicon has its place. For example, we became aware of a new gene and wanted to sequence that gene in all 2000 of our subjects. I performed a design for a TruSeq Custom Amplicon–based process for that single gene. I realized that I could perform the assay, sequence it on the MiSeq System at 96-plex on one run, and it would be cheaper than Sanger sequencing the backlog of samples.

Q: What is the perception in your lab of the transition from Sanger sequencing to NGS?
JR: Some of our research associates and senior staff have been part of the transition and have seen how successful it has been in terms of time and costs. However, our fourth-year and new grad students have come to expect that they're going to get 700 kb of data in a 2-week turnaround.

Q: How did the culture of the Robarts Research Institute support the creativity needed to design the LipidSeq panel?
JR: The culture of the Robarts Research Institute provides a solid infrastructure to support its scientists. Dr. Hegele is certainly a world leader in his research, creating an excellent work environment and scientific program that attracts creative people. It makes us think outside the Sanger box, and embrace new technologies such as NGS. A perfect example is turning the concept of performing exon resequencing and polygenic trait score SNP scoring on the same resequencing assay and turning that into reality on the LipidSeq panel. That was the result of creativity and good insight, that combined Sanger and TaqMan assay workflows and transformed them into a high-throughput process using NGS.

Learn more about the Illumina systems mentioned in this article:

Nextera Rapid Capture Custom Enrichment Kit, www.illumina.com/products/by-type/sequencing-kits/library-prep-kits/nextera-rapid-capture-custom-enrichment.html

MiSeq System, www.illumina.com/systems/miseq.html

DesignStudio Custom Assay Design Tool, www.illumina.com/informatics/research/experimental-design/designstudio.html

TruSeq Custom Amplicon Panel, www.illumina.com/products/truseq_custom_amplicon.html

Targeted Resequencing, www.illumina.com/techniques/sequencing/dna-sequencing/targeted-resequencing.html

References

1. American Heart Association. Heart disease and stroke statistics – 201y Update: A Report From the American Heart Association. *Circulation*. 2017;135:e146–e603.
2. Johansen CT, Dube JB, Loyzer MN, et al. LipidSeq: A next-generation clinical resequencing panel for monogenic dyslipidemias. *J of Lipid Res*. 2014;55(4):765–772.
3. Hegele RA, Ban MR, Cao H, McIntyre AD, Robinson JF, Wang J. Targeted next-generation sequencing in monogenic dyslipidemias. *Curr Opin Lipidol*. 2015;26(2):103–113.

Illumina, Inc. • 1.800.809.4566 toll-free (US) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

illumina®