**RAD-Seq Genotypes Less, But Offers More**

Floragenex RAD-Seq services leverage the sequencing power of the HiSeq® 2000 to cost-effectively discover SNPs and genotype plants, livestock, and non-model organisms.

### Introduction

Genotyping enables researchers to zero in on what makes an organism unique, with differences in DNA sequence accounting for everything from phenotypic traits, to the ability to combat disease or flourish in harsh environments. For the past decade, researchers have used microarrays to discover the single nucleotide polymorphisms (SNPs), copy number variations (CNVs), insertions, deletions, and duplications that cause these differences. Microarrays are customizable, cost-effective, have simple workflows, and yield high quality data—the perfect SNP discovery and genotyping tool for agrigenomics researchers with limited budgets. With the cost of whole-genome sequencing dropping, agrigenomics meetings have been abuzz with talk that genotyping by sequencing (GBS)—using direct DNA sequence data to determine sample genotypes—was finally within reach. Despite recent advances, directly interrogating every base pair in a genome remains costly and generates huge amounts data, straining the budgets and bioinformatic abilities of most agrigenomics researchers.

Interestingly, a method originally developed for microarrays has provided the foundation for a simpler, more cost-effective sequencing approach for SNP discovery and genotyping. Restriction site-associated DNA (RAD) marker genotyping, developed by Eric Johnson, Ph.D. and a team of researchers at the University of Oregon, provides a genome-wide representation of every site of a particular restriction enzyme. By hybridizing RAD tags to custom microarrays, the team created a tool that performs parallel screening of thousands of polymorphic markers, allowing researchers to map natural variation and induced mutations in diverse organisms.

In 2006, Dr. Johnson and Nathan Lillegard formed Floragenex to commercialize the method and provide RAD genotyping services to budget-constrained researchers studying plants, model, and non-model animals. When the Illumina Genome Analyzer™ System was introduced the next year, the company saw an opportunity to make their RAD fragment libraries compatible with this next-generation sequencing platform. The result was RAD sequencing (RAD-Seq), a method enabling massively parallel and multiplexed sample sequencing of RAD tag libraries that has transformed the company and revolutionized agrigenomics research by facilitating the rapid discovery of thousands of SNPs and the high-throughput genotyping of large populations.

Floragenex now offers RAD-Seq services on the HiSeq 2000 platform to agrigenomics researchers and commercial customers worldwide. iCommunity spoke with Nathan Lillegard (Chief Executive Officer), Rick Nipper, Ph.D. (Senior Vice President of Plant Genomics), and Jason Boone, Ph.D. (Research Director, Animal Genomics) to learn more about the RAD-Seq technology and the benefits it provides for plant and animal research.

**Q. What is the RAD technology and how was it developed?**

Rick Nipper (RN): RAD technology is what is known as a genome “complexity reduction” protocol, designed to reliably interrogate a fraction of a target genome instead of the entire genome sequence. The short fragments (or tags) of genomic DNA that flank the recognition site of particular restriction endonucleases are then screened for the presence of genetic variation. In 2006, Eric and his team at the University of Oregon first hybridized RAD tags to microarrays, where they provided a readout using fluorescent dyes to indicate the presence or absence of a RAD tag sequence. While powerful, the microarray-based RAD technique could only assay a fraction of segregating polymorphisms.

Nathan Lillegard (NL): Eric had the vision to move the technology to Illumina’s first sequencing by synthesis platform, the Genome Analyzer. We made our fragment libraries compatible with the system, developed new RAD tag generation and typing methods, and incorporated nucleotide barcodes for sample tracking that enabled us to sequence many different projects in a single run on the Genome Analyzer.

**RN:** With RAD-Seq we are sequencing anywhere from 0.1% to 15% of the entire genome, rather than directly interrogating every base pair as you would with a whole-genome shotgun (WGS) strategy. The complexity reduction is achieved by using restriction enzymes that nick the genome at specific positions. A series of adapters are then affixed to the fragments and those adapters bind to the Genome Analyzer or HiSeq flow cell. This very powerful technique allows you to directly read the associated genomic fragment, enabling you to perform other genomic studies such as finding and genotyping variants, microindels, and epigenetic changes.

Nathan Lillegard is Chief Executive Officer, Rick Nipper, Ph.D. is Senior Vice President of Plant Genomics, and Jason Boone, Ph.D. is Research Director, Animal Genomics at Floragenex, a provider of RAD-Seq services to agrigenomics researchers and commercial customers worldwide.
RAD-Seq GBS is particularly elegant because it whittles down the amount of the genome that you’re querying. It allows you to interrogate a scalable number of loci. If you want to look at 10,000 positions in the genome you can do that. If you want to look at 100,000 unique positions in the genome you can do that too. That kind of dynamic range is really valuable to investigators.

RAD-Seq becomes even more cost-effective as sequencing costs drop. As a result of moving over to Illumina’s HiSeq 2000 platform, we believe over the next year the raw cost to produce RAD-Seq data will begin to approach fixed content arrays in cost and throughput.

“RAD-Seq enables ‘deep sequencing’ of SNPs at 50× coverage, giving researchers confidence that our technology can identify genetic variation linked to a trait or population.”

Q. Who’s using RAD-Seq?

NL: About 35% of our customers are from commercial organizations, primarily crops seed companies, and 65% are academics in universities or research institutes. The projects are 50/50 plant versus animal. While the commercial customers are fewer, those are bigger projects for us. The academic projects come in all shapes and sizes, from work on endangered species to evolutionary studies.

Q. Is RAD-Seq best used for genome-wide genotyping rather than targeted genotyping of organisms?

RN: Yes, RAD-Seq is ideal for generating genome-wide genotype data in situations where there is not much known about the target genome. If you know a priori there is a QTL or a genetic locus that is responsible for a particular trait or phenotype, and you have the physical and genetic positions for that gene, amplicon sequencing or capture sequencing might be better options. RAD-Seq is designed to quickly scan everywhere in the genome to find the target area you’re interested in, then you’d use one of those other tools to dig deeper.

JB: In my opinion a genome-wide scan will be important whether the QTL region is known or not. Currently it is unclear how other regions of a genome might change or be affected during introgression of multiple traits or current breeding strategies. Because of this, many institutions, such as the USDA and FDA, are trying to capture as much information about surrounding genomics regions as possible. For conservation efforts and certain breeding applications, it is important to retain as much genetic diversity in an organism’s background as possible for future breeding applications.

RN: A perfect example is a project in a forage grass (Lolium) that’s grown for grass seed in Oregon’s Willamette Valley. The researchers knew beforehand that there were several traits controlled by different
loci, but the number of markers they had from previous genotyping maps was quite low. We were able to find more markers using RAD-Seq, enabling them to more accurately pin down the QTL important for breeding applications. So even if you have some knowledge of the genome, it can be beneficial to consider a whole-genome scan.

Q. In addition to SNP discovery and genotyping, are there other applications where RAD-Seq is valuable?

RN: You can use RAD-Seq for any application that employs sequencing technology, so there’s a huge range of applications where it can provide value. It can be used to facilitate de novo assembly in genomes or to perform linkage analysis such as a bulk segregant study. This is important in plant genomics, where breeders are interested in finding markers that co-associate with certain desirable phenotypes. They use those markers as a diagnostic to confirm that a particular variety of plant possesses a valuable trait, such as disease resistance.

RAD-Seq can be used to survey the population structure of a particular wild species and perform phylogenetics studies by constructing genomic or linkage maps that provide a portrait of recombination rate across the genome. It also can be used to perform linkage disequilibrium mapping or association mapping studies, or study epigenetics such as changes in methylation status across a genome.

Q. Why is sequence data so much more valuable than array data?

RN: Sequence data is essentially future proof. Once you’ve identified a candidate marker using RAD-Seq, that bit of information is digital and can be integrated with any future genomics resources that are developed. You could look for epigenetic changes in the sequence data, see if it harbors a rare allele or a structural variant, align it to an assembled reference genome, or if you only have a transcript assembly, you can BLAST (Basic local alignment search tool) against that as well. Because the sequence is digital, you have an incredible array of options. That’s the real advantage for any type of sequence-based technology over analog gel-based or even array-based approaches.

Q. How do you handle organisms with reference genomes versus those without reference genomes?

RN: One of the benefits of RAD-Seq is that it can be used for organisms where there is a lot of genomic information and for those that have never been sequenced. If there is a reference genome available, such as for rice, we can use our technology to sequence 1% of the genome, identify genetic variants, position them on the reference genome, and provide the marker and genetic information to our customer.

For an organism that does not have a reference genome, such as sunflower, we would generate the exact same content; there just wouldn’t be a reference genome to determine the position of those markers. The nice thing about RAD-Seq is that the data is forward compatible. With the revolution in genomics that’s going on right now, there will be reference genomes available in an increasing number of species. Once these are available, one will be able to pull out the RAD-Seq data and determine the physical marker positions.

Q. Why would a customer choose to have Floragenex perform their RAD-Seq analyses?

NL: We’ve worked with a variety of commercial customers as well as academics that often have experience with next-generation sequencing. Even so, they’re looking for a turnkey solution and want the
expertise we bring to RAD-Seq library preparation, sequencing, and bioinformatics. Rather than having to learn the RAD-Seq protocols and bioinformatics, our customers get to focus on the actual biology of their work and leave the technical challenges to us.

RN: Our customers value our ability to quickly build next-generation RAD-Seq libraries. We have several years of experience using Illumina sequencing technology, first on the Genome Analyzer and now on the HiSeq 2000. Unlike most of the academic labs that are performing RAD-Seq, we’re in a production environment. We’re not just making one or two RAD libraries a week, we’re making tens if not hundreds over the course of a month, so we need to be able to look very quickly at sequence data and make intelligent decisions about what the sequence data means, where do we go from here, and what needs to be done to complete the project.

The most common questions we get are about how deeply we query each locus in the genome. For various reasons, if you don’t interrogate a specific nucleotide to about 15× coverage, there can be ambiguities in what the actual sequence genotype for that locus is. Someone performing RAD-Seq themselves may not realize until the end of the analysis that the sequencing depth isn’t enough. We can’t afford to have that happen, so we developed customized quality control and quality assurance reports that provide us with immediate feedback on the number of reads per sample, the sequencing depth, etc.

“One of the benefits of RAD-Seq is that it can be used for organisms where there is a lot of genomic information and for those that have never been sequenced.”

On the backend, we have a dedicated bioinformatics pipeline that’s focused on analysis of RAD data. We’ve developed tools that help us hammer out the analysis and produce a report that provides all the information a customer will need about their samples.

Q. What types of projects have you worked on so far?

JB: One of our more interesting projects was working with Scott Baker, Ph.D., an international forensic expert on cetaceans who did the forensic work for the movie “The Cove.” Dr. Baker and his team wanted to look deeper into the possibility of hybridization between Blue and Fin whales in the Northern Pacific Ocean. Not only an interesting phenomenon, but possibly useful for conservation genetic efforts. Together with Dr. Baker’s team, Floragenex used the RAD-Seq local de novo assembly technique for each Blue and Fin whale. These results allowed the team to rapidly identify common and divergent SNPs between the two species, as well as generate the most sequence data for both species to date.

RN: We’re conducting genome development work in over 35 distinct angiosperm species, most of the major row crops, and some of the minor vegetable crops. We’ve constructed a genetic map in barley⁴, identified SNP and SSR markers in eggplant⁵, conducted polyploidy SNP discovery in canola¹, performed SNP discovery in sorghum², and just completed work for scientists in Brazil conducting eucalyptus research.

Q. The eucalyptus sequence was just published. Why was RAD-Seq performed and what did the data reveal?

RN: Eucalyptus is a fast growing tree and is of interest because of its genetic diversity and potential as a biofuel. The U.S. Department of Energy’s Joint Genome Institute (JGI) released a ~700 Mb eucalyptus genome⁶ in May. We were contracted by customers in Brazil to perform paired-end RAD sequencing in the same accession that was sequenced in the JGI study—BRASZU1—to evaluate the feasibility of the technique for de novo assembly and SNP development.

Using the RAD-Seq local de novo assembly protocol, we obtained approximately 9 million 2 x 80 reads on the Genome Analyzer. At the end of the assembly, we had about 71,000 contigs, with an N50 size of 310 base pairs, which represents about 22 Mb of the Eucalyptus genome. That’s about 3% of the haploid genome length and probably about 30–50% of the assembly that you might expect to obtain from an EST or transcriptome effort in the average plant genome.

The overall sequence coverage in the RAD-Seq assembly was just under 10×, with very high consensus quality. Over 99% of the nucleotides had greater than Q30 quality scores, with over 75% of the assembly having Q40 or above scores, representing 99% and 99.99% accuracy rates, respectively. We aligned the RAD assembly with the reference genome from JGI and over 59,000 of the contigs assembled were exact matches to the JGI assembly.

Q. How have Illumina systems enhanced Floragenex RAD-Seq services?

NL: Illumina sequencing systems work perfectly with RAD-Seq and, because your company keeps innovating and increasing capacity of its sequencing systems, we’ve been able to keep our costs nearly flat for the amount of data that we can give our customers. We realized in 2007 that we could sequence four RAD libraries in one Illumina lane on the Genome Analyzer. With the HiSeq 2000, we’re regularly sequencing 96 samples in a lane. That kind of multiplexing is what our customers need, because their budgets only have room for a few lanes of sequencing.

RN: There’s so much enabled by the Illumina technology that it really keeps us on our toes. I’ve spent quite a bit of time re-engineering our bioinformatics pipeline to handle the 5× increase in data, soon to be a 20× increase in data, offered by the HiSeq. We’re taking advantage of parallel processing, distributing analysis over multiple computers, and also building more programs in compiled computer languages to speed data processing. We’ve developed robust computational solutions to handle 100 million reads per lane, all the way up to about 300 million reads per lane that the new Illumina reagents support.

JB: The HiSeq 2000 system is so powerful we can batch projects now, putting them all on one to three runs. With just three runs of a HiSeq, we’re sequencing as much as we did almost all of last year.

NL: It’s also great to work with a company that’s supporting agri-genomics research, particularly in plants. Illumina is breaking down the barriers, reducing the cost of sequencing so that more researchers can use it to advance their research.
Q. What do you see as the growth areas for RAD-Seq and your service offering?

JB: As people become more familiar with the value RAD-Seq can bring to their research, we see ourselves growing in three ways. The first is in the plant sciences, particularly the agricultural seed market where we are becoming a specialty expert in plant genomics. The second is in terms of non-plant organisms, where we’re happy to work on projects for non-model organisms, livestock animals, and even the occasional exotic animal. Finally, we’re focused on increasing the RAD technology user base through licensing agreements, and the development of reagents and bioinformatic tools so people can do it themselves.

NL: We see a lot of potential growth in agrigenomics companies, especially seed companies. The larger players have in-house genomics expertise, but there is still a big portion of that industry that has yet to join the genomics world.

RN: From a bioinformatics perspective, we’ll continue to accommodate the increase in Illumina sequencing throughput by creating more and better analysis tools, especially ones that take advantage of cloud computing infrastructures.

References
9. www.phytozome.net/eucalyptus.php

“The HiSeq 2000 system is so powerful we can batch projects now, putting them all on one to three runs. With just three runs of a HiSeq, we’re sequencing as much as we did almost all of last year.”

Illumina • 1.800.809.4566 toll-free • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

FOR RESEARCH USE ONLY
© 2011-2012, 2014 Illumina, Inc. All rights reserved.
Illumina, IlluminaDx, BaseSpace, BeadArray, BeadXpress, cBot, ChIPseq, ChIP, Infinium, iSelect, MiSeq, Nextera, Sentrix, SeqMonitor, Solexa, TruSeq, VeraCode, the pumpkin orange color, and the Genetic Energy streaming bases design are trademarks or registered trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners.
Pub. No. 070-2011-016 Current as of 06 November 2014