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[1].

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[2].



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.

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.

Jason Boone (JB): 2011 is turning into a breakout year in terms of the exposure of the technology to the broader community. We estimated that there would be about 20 RAD-Seq papers in 2011, but in just the first 8 months there have been 11 publications featuring RAD-Seq for everything from *de novo* assembly to linkage mapping. We've also assisted in 10 to 15 grant applications, so the outlook for 2012 looks promising.

Q. How is RAD-Seq better than other GBS methods?

RN: WGS and transcriptome sequencing enable you to find millions upon millions of genetic variations between any two individuals. Yet, all that information is generally unnecessary for scientists interested in standard molecular breeding techniques, performing limited linkage analysis, or even just looking for several thousand SNP variants. Researchers don't need the massive output of WGS sequencing, but desire higher multiplexing ability than can be obtained from exome sequencing. RAD-Seq sits in between those two methods, providing a lot of information at a relatively low cost point, without the complexity of large genome sequencing and assembly efforts.

NL: RAD-Seq enables us to quickly provide researchers with useful, high-quality genotyping information. RAD enables "deep sequencing" of SNPs at 50× or greater coverage, giving researchers confidence that our technology can identify genetic variation linked to a trait or population. Unlike WGS sequencing data which can be difficult to assemble and hard to compare, RAD-Seq data is easier to handle and allows you to easily compare multiple individuals or populations from different samples. We have worked with sample sizes ranging from 2 to over 1,000.

Q. How does RAD-Seq compare from a library preparation and bioinformatics standpoint?

JB: Library preparation for RAD-Seq is simpler and faster than other applications. For WGS sequencing, each sample is carried all the way through the process individually. In RAD-Seq, after the first couple of steps you can pool all 96 samples or even 384 samples, and process it as one sample to the end of the protocol.

NL: RAD libraries and data are easier to analyze and smaller to process because there is less overall data produced than in WGS analysis. RAD multiplexing enables us to break down a lane of 100 million reads into 25 different files. Because we know that those 25 files have certain characteristics, it's easier to manage the data, giving us a huge advantage in bioinformatics over WGS or expressed sequence tag (EST) sequencing.

Q. What is driving people to be interested in GBS and in RAD-Seq?

RN: Frequently it's cost. Sequencing costs have come down from a billion dollars to one million dollars to $5,000 for generating 30× coverage of a whole human genome. Thanks to the efforts of companies like Illumina, it will inevitably drop to the magic $1,000 per genome price point, which will make GBS extremely cost efficient in the future.

Certainly, I would say researchers recognize there's value in performing whole-genome genotyping. If you were to fragment the entire genome and sequence every nucleotide—you'd get a portrait of genotype information for every position in the genome. That is particularly useful in cancer genetics and other high-density genomic analysis. However, in some instances researchers don't need that level of information.

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[3]. 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 genetic 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.

> "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."

Q. Can you reconstruct a genome using RAD-Seq?

RN: Yes, we can use RAD sequencing to perform what is called "local" *de novo* assembly. The goal is to construct a small sequence contig of 300–500 base pairs around a nuclease digestion site. This local assembly is possible because of the unique architecture of a RAD fragment, which consists of a 5' end anchored to the restriction enzyme site and a randomly sheared 3' end. Both the 5' and 3' ends have adapters. When you do a RAD preparation, you perform a size selection of the fragments that are produced and create a series of ostensibly overlapping fragments. By using paired-end Illumina sequencing, you can sequence 50–100 base pairs from the restriction enzyme cut site and 50–100 bp in the randomly sheared genomic region. When you get the RAD-Seq data back you can then assemble or "stitch" it together into a contiguous DNA sequence. From there, you can align the contig back to a reference genome from a related species or EST database. Eric's laboratory at the University of Oregon has also pioneered major improvements in this strategy that enable longer local assemblies, up to several kilobases in length[4].

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 select the correct enzyme?

RN: We consult with the customer, evaluate how many markers they require, and assess the genetic diversity of the species they are studying. We look at the available sequence data for the species and run some in silico analyses on the sequence information to try to predict how many sites we might uncover. Then we select the particular restriction enzyme that is going to generate the results they need. We have worked with over 60 species, so often we have empirical knowledge of a particular genome, enabling us to accurately guide customers to the right enzyme and advise them about what to expect from the output that will be generated.

JB: Sometimes it takes a little bit of detective work to figure out the optimal solution. Certain enzymes are overkill for what a customer needs and you have to consider the cost benefit. If you choose a restriction enzyme that produces more fragments, it will cost more to perform the analysis. A customer's budget is an important parameter in selecting the right enzyme.

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 informatics, 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've 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 barley[5], identified SNP and SSR markers in eggplant[6], conducted polyploidy SNP discovery in canola[7], performed SNP discovery in sorghum[8], 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 genome[9] 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 × 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 complied 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.

"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: 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

1. Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA. (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res 17:240–248.
2. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, et al. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3:e3376.
3. Chutimanitsakun Y, Nipper RW, Cuesta-Marcos A, Cistue L, Corey A, et al. (2011) Construction and application for QTL analysis of a Restriction Site Associated DNA (RAD) linkage map in barley. BMC Genomics 12:4.
4. Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA (2011) Local de novo assembly of RAD paired-end contigs using short sequencing reads. PLoS ONE 6: e18561.
5. Pfender WF, Saha MC, Johnson EA, Slabaugh MB (2011) Mapping with RAD (restriction-site associated DNA) markers to rapidly identify QTL for stem rust resistance in Lolium perenne. Theoretical and Applied Genetics 122(8):1467–1480.
6. Barchi L, Lanteri S, Portis E, Acquadro A, Vale G, et al. (2011) Identification of SNP and SSR markers in eggplant using RAD tag sequencing, BMC Genomics 12:304.
7. Tang S, Wiggins M, Nipper R, Gribbin J, Johnson E, et al. SNP discovery using restriction-site associated DNA (RAD) long-read sequencing in Brassica napus, a polyploid species. Plant and Animal Genome 2010
8. Nelson JC, Wang S, Wu Y, Li X, Antony G, et al. (2011) Single-nucleotide polymorphism discovery by high-throughput sequencing in sorghum. BMC Genomics 12:352.
9. www.phytozome.net/eucalyptus.php
10. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, et al. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet. 12:499-510.

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

FOR RESEARCH USE ONLY

Pub. No. 070-2011-016 Current as of 06 November 2014

illumina®