sage science

Case Study :: **Hoekstra Lab, Harvard University**

New Take on RADseq Enables High-Throughput Variant Discovery

A new, massively parallel genotyping technique from a research team at Harvard relies on automated size selection and next-gen sequencing to boost accuracy and lower costs.

While the much-hyped $1,000 genome will make a big difference for some projects, scientists who study large populations or need to look across hundreds or thousands of samples require a far more cost-effective approach.

And it's those scientists who will benefit most from a significant advance out of Hopi Hoekstra's lab at Harvard University. Hoekstra, a professor in the departments of Organismic & Evolutionary Biology and Molecular & Cellular Biology, focuses on population genetics, development, speciation, and behavioral genetics. In a paper published in *PLoS One* in May of 2012, she and her team present a method for low-cost, massively parallel genotyping that does not require prior knowledge of an organism's genome sequence. ["Double Digest RADseq: An Inexpensive Method for *De Novo* SNP Discovery and Genotyping in Model and Non-Model Species."]

"Anyone who does genotype work has this sense that the right way to be doing this is by sequencing," says Brant Peterson, PhD, a postdoctoral fellow in the Hoekstra lab and lead author on the paper. Array-based approaches tend to look for known variants, preventing the discovery of new or rare variants. Assembling a panel of common polymorphic sites to build an array takes a lot of time and work — and, as Peterson notes, "if you work on non-model species, you're probably the only person who will ever use that array."

"The two things that make the biggest difference for us in this approach were the ability to do precise sizing and the ability to do it reproducibly."

Brant Peterson, Ph.D, Post Doctoral Researcher at the Department of Organismic and Evolutionary Biology, Harvard University

For the past two years, Peterson and other members of the Hoekstra lab have been trying to come up with something better. Their method of choice builds on reduced-representation genome sequencing, commonly called RADseq, improving the approach by lowering costs and increasing accuracy. "The two things that make the biggest difference for us in this approach were the ability to do precise sizing and the ability to do it reproducibly," Peterson says.

In essence, RADseq allows scientists to sample large numbers of individuals at once, looking for hundreds or thousands of variants in each — but doing so in just, say, half a percent of that organism's genome. That keeps sequencing costs manageable while still allowing scientists to get a good survey of genomic variation. "If you could sequence every genome of everything all the time, maybe that would be better," Peterson says. "But in the real world, this means that we can do experiments for the right scale at a cost we can get to."

RADseq

RADseq, or restriction site associated DNA sequencing, draws from the idea that scanning even a small fraction of a genome can provide a lot of information. The method, which was pioneered by the Cresko lab at the University of Oregon, involves deploying restriction enzymes across the genomes of many individuals; the enzymes cut at a certain predefined sequence, generating a slew of sequence fragments to interrogate.

Peterson and his colleagues began using RADseq because their organism of interest, the deer mouse *(Peromyscus)*, is only distantly related to the more heavily studied lab mouse and "there are very few formal resources in terms of genome sequence," he says. "The easiest thing we could think of to do is to sequence the same bit of genome of many different individuals and look for variable sites in those intervals."

Where RADseq's advantage is no requirement for prior knowledge of the organism's genome sequence, its tradeoff is that the genome fragments are somewhat randomly selected. Unlike exome sequencing — another approach to reduced-representation genome sequencing — there's no guarantee that the genomic fragments will be from coding regions, for example.

But for certain types of studies, that's not a limitation. From evolutionary development to population studies to QTL mapping, there are several applications where having the same genomic fraction from many organisms is just as informative. "Let's say you're curious about the history of a group of organisms in the context of their evolutionary relatedness — how long ago did they last share an ancestor? Are individuals from two different populations exchanging genetic material, and how frequently do they come into contact?" Peterson says. "The key to both of those is you don't actually care specifically which bit of genome you sample; you just want a good survey." Other areas of interest include genome-wide association studies and QTL mapping. For those, "you want to sample some fraction of DNA that's variable and close by the gene or enhancer that's doing the heavy lifting. That'll report on what's going on with the functional site," Peterson adds.

> "This means that we can do experiments for the right scale at a cost we can get to."

The restriction digest approach, then, was a strong candidate for the type of studies performed in the Hoekstra lab. "Restriction enzymes should cut in the same place in every individual genome, so you should get the same sized fragments from each region of the genome from many individuals subject to the same restriction digest," Peterson explains. That was the theory, anyway. "The only thing that was missing was the ability to actually do that," he says.

The Size Selection Challenge

While reduced-representation approaches like RADseq have been around since 2000, the stumbling block for boosting accuracy has been size selection. If you randomly reduce a genome to fragments, and then compare those to the randomly generated fragments of another genome, you have to be comparing the same random fragments from each genome for your results to make any sense.

"The idea is that if you could get the same sizes of DNA from each individual from the same restriction digests, you'd get the same regions — that's assuming you're perfect at your size selection," Peterson says. What became clear in the Hoekstra lab as they studied this problem was that size selection on manual gels was subject to both operator-to-operator variability as well as each person's own variation in slicing.

In the end, those sources of variation can derail the best-planned experiment. "You've sampled many regions from many individuals, but when you go to stack them all up, no one has everything and no spot is sampled in everyone," Peterson says. "The devil in the detail is that your probability of getting it right has to be really, really high each time for each region in each individual — or else you end up not being able to do your analysis."

That's where the Pippin Prep from Sage Science came in. Launched the same year Peterson was starting this project, the Hoekstra lab acquired the instrument to see how automatic size selection compared to their standard manual gel methods. As it turned out, the Pippin's precise sizing allowed the team to make progress in a way manual gels could not. In the absolute best-case scenario, Peterson and his colleagues estimate that a manual gel practitioner can achieve up to 50 percent of the "precision and repeatability of automated DNA size selection," they note in the *PLoS One* paper. For example, if running the RADseq experiment through Pippin Prep would have generated 20,000 shared regions across 100 individuals, Peterson says, "you might get 4,000 or 5,000 regions in the same 100 individuals running it on a gel." That loss compounds as the number of individuals and number of markers increase. "As the scale of the project gets bigger, the ability to repeat the same operation becomes more crucial," he adds.

Thanks to the Pippin platform, size selection "is no longer dependent on one operator," Peterson says. "There's very little difference from one sizing reaction to the next, which is the key to this approach working." He notes that sample recovery is better on the Pippin as well, and the automation means that people who would be toiling over gels can be working on more interesting things while the instrument is performing size selection.

New Possibilities

Beyond implementing automated size selection, the Hoekstra lab made other improvements to the RADseq approach, including the addition of a second restriction enzyme and eliminating the preparative shearing step. In combination with automated sizing, these changes allowed the team to increase efficiency, drastically reduce costs, and maximize the number of samples they could sequence in a single Illumina lane. Ultimately, the technique costs "fractions of a penny per individual per site … and requires little starting material (i.e., 100 ng of DNA)," the authors write.

The Hoekstra team expects this genome-reduction technique to enable a range of projects that weren't possible before, particularly for scientists who study organisms that aren't as comprehensively analyzed as human or lab mouse. "We can't make meaningful inferences about the history of populations or the influence of natural selection without hundreds of samples, and it's currently completely impractical for us to imagine doing that on a whole-genome scale," Peterson says. "In population genetics and quantitative genetics, everyone's experiment needs to sample hundreds or thousands of individuals."

In the Hoekstra lab, the double-digest RADseq approach is already being used for new studies. One group is using it to study relatedness in thousands of lizards, while other projects involve "looking at phenotype associations across populations where we only need to sample a couple of hundred individuals, but we need to sample them at half a million sites genome-wide in order to capture the phenomenon that we're looking for," Peterson says.

"When you reduce the fraction of the genome you're looking at," he adds, "it becomes possible to do certain kinds of analyses that wouldn't be possible on the entire genome."

The Pippin Prep system is an automated gel electrophoresis platform designed to save scientists time and money in DNA size selection. The platform uses optical fluorescence detection of DNA separations to automatically collect size-selected fragments from pre-cast agarose gel cassettes. DNA is electro-eluted from agarose according to user-input settings, and up to five samples may be independently size selected per cassette. Samples are collected in buffer and removed by standard pipettes. Compared to manual gel purification, DNA fragments are collected with much higher accuracy and reproducibility — and with no contamination. For additional information, contact us at info@sagescience.com or 978-922-1932, or visit our website at www.sagescience.com.

sage science

Suite 3150
500 Cummings Center
Beverly, MA 01915
978.922.1832
www.sagescience.com

© 2012 Sage Science, Inc. All rights reserved. Pippin Prep is a trademark of Sage Science.