sage science

Say Goodbye to Manual Gels: **Automated DNA Size Selection Has Arrived**

Automated size selection is a superior alternative to manual gel extraction, allowing scientists to save time and money, improve the efficiency of sequencing runs and analysis, and perform new applications with next-gen sequencers.

It has been well established in the scientific literature that while next-generation sequencers have used automation to increase throughput, reduce time to results, and lower costs, there continues to be a bottleneck in the sample preparation process due to the reliance on manual gel extraction for DNA size selection (Fisher, Borgström). Accurate, automated size selection is a critical component for cost-effective, high-accuracy DNA sequencing.

The current high-throughput DNA sequencers — including platforms from Illumina, Ion Torrent, and 454 — require tightly sized insert libraries for optimum sequencing performance. Without reliable size selection, valuable sequencing cycles are spent on primer-dimers, adaptor-dimers, and other low molecular weight material that is considered sample contamination. This is a particular challenge for sequencers that require emulsion PCR, which tends to be preferential to smaller fragment amplification. In addition to wasting time and money to sequence this material, this problem takes up resources on the analysis side, where all of these useless sequences must be identified and removed before analysis of the real sequence data can be performed. For paired-end libraries, poor size selection reduces analytical power even more: accurate knowledge of the distance from one read to the next is important for correctly mapping sequence fragments during alignment.

Many scientists use agarose gels as a manual step to size-select their libraries prior to sequencing. This is a laborious and time-intensive activity, fraught with known problems. Manual gel extraction is operator-dependent, resulting in significant variability between the DNA sizing step performed by any two people. Studies have shown that sample-to-sample contamination is a serious concern; scientists must either accept the real risk of contamination in their samples in order to multiplex or always run one sample at a time to prevent contamination, seriously slowing the sample prep process (private communication: The Broad Institute). The other major problem with using manual gels is that it sidelines a highly trained technician or scientist for hours at a time, preventing that person from doing other work or experiments.

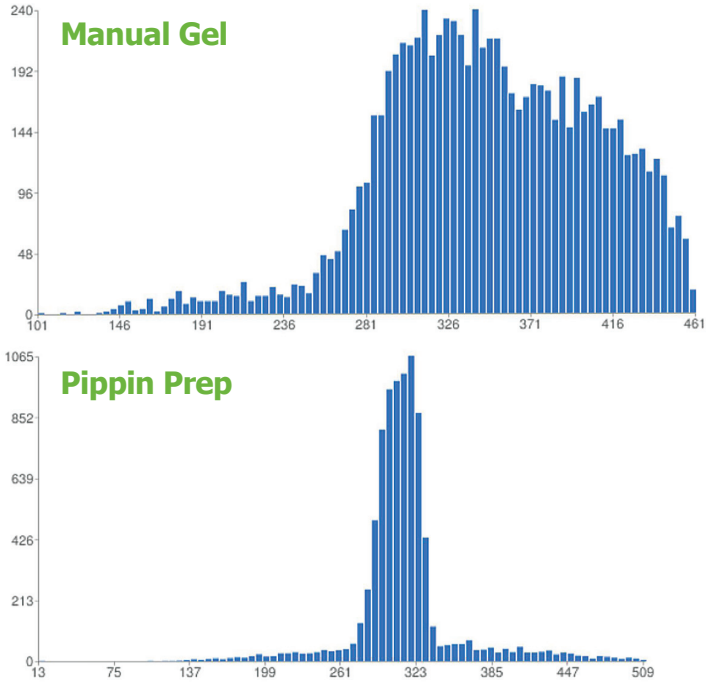

Figure 1. Post-sequencing mapped insert-size distribution graphs for Illumina sequencing libraries prepared from P. falciparum genomic DNA with size-selection using agarose gel electrophoresis and Pippin Prep (Reference: Quail et al.)

Automating Size Selection

As next-gen sequencing has taken hold, a number of vendors have looked to automate the DNA size selection step. In most systems, this takes the form of an instrument loaded with disposable cassettes containing precast agarose gels. The instrument uses electrophoresis along with laser detection or other imaging technology to determine when to start collecting DNA based on size ranges entered by the user. Once the DNA is no longer in the desired size range, collection ceases. The DNA is generally collected in buffer and can be loaded directly onto a sequencer.

In theory, these automated size selection technologies should prevent cross-contamination; boost accuracy and reliability of sizing; free up scientists' time to focus on other tasks; and improve the quality of sequencing and analysis.

Though not all automated systems measure up, the best options do accomplish those goals. Scientists at the genome center at Emory University, for example, say they have been able to shave nearly a full day off the sample prep process for Illumina's mate-pair library prep by using automated DNA size selection. The lab uses the Pippin Prep automated sizing platform from Sage Science to replace the 16-hour runs that had previously been done on gels with a very low agarose concentration. Size selection on the Pippin Prep takes about an hour, according to scientists at the genome center (Sage Science blog, May 2012).

Scientists at the genome center at Emory University, for example, say they have been able to shave nearly a full day off the sample prep process for Illumina's mate-pair library prep by using automated DNA size selection.

The Emory team had previously tried a different automated size selection platform, but it would not work for fragments larger than 1 kb. Once the lab started running large mate-pair libraries that could range from 3 kb to 20 kb, the scientists assessed alternatives and chose the Pippin instrument.

Scientists at other labs report that automated size selection has freed up resources: technicians or researchers who would previously have spent hours preparing and slicing gels can instead load those samples, often several at once, on a size selection platform that requires minimal hands-on time. While those samples are running, the same person is able to perform other critical steps in the experimental workflow. In an era of uncertain funding and layoffs at genome centers and academic labs, this technology allows scientists to increase efficiency in their labs.

Similarly, the ability of accurate size selection to improve the efficiency of sequencing runs also fills an important need. With tighter size selection, scientists might fit an experiment that would have taken two Illumina lanes onto a single lane, saving money and time. The technology also boosts the accuracy of sequence analysis, making for more reliable results with less time spent manually reviewing alignments or assemblies.

Case Study: Better Genome Assemblies

At the DNA Technologies Laboratory at the National Research Council of Canada, scientists are using automated size selection to improve the quality of their genome assemblies. Assembly projects tend to focus on large plant and fungal genomes, for which the scientific team relies on Illumina and 454 sequencing, often combining them to make a hybrid genome assembly that takes advantage of both platforms (Sage Science blog, July 2012).

Most of the libraries run in the lab are relatively short, standard paired-end libraries ranging from 200 bases to 400 bases, according to Andrew Sharpe, Research Officer and Group Leader of the Saskatoon-based laboratory, which also serves as a core facility for NRC and other Canadian government agencies. The team also runs longer mate libraries, usually in the range of 3 kb to 10 kb. Automated size selection has proven beneficial for both types of libraries.

Sharpe and his team shifted away from manual gel extractions and now use the Pippin Prep and the BluePippin from Sage Science to perform more efficient, automated size selection. In addition to saving time, Sharpe says that using the Pippin platform enables his lab to create multiple paired-end libraries of different insert sizes for the same sample. Those libraries, which could for example be set up as 200-base, 300-base, and 400-base inserts, are then sequenced and assembled together using SOAPdenovo or other tools.

"If you assemble one of the libraries, then you'll end up with an assembly. But if you assemble all three together using three different lengths, you get quite a bit better product," Sharpe says. "The nice thing with the Pippin Prep is being able to easily get those discrete size ranges."

Less DNA, More Applications

The Broad Institute has used automated size selection to increase throughput and reduce variability in its Genome Sequencing Platform. Sheila Fisher, director of operations and development for the sequencing platform, worked with Sage Science to implement the Pippin platform for DNA sizing. She found that one of the benefits of using this solution was boosting the sample recovery. With manual gel extraction, a lot of the DNA is lost, but with the Pippin automated size selection, sample yield is significantly higher (Bridger).

In cases where Fisher's team would have needed 3 micrograms or more of DNA to start, they can now begin with just 100 nanograms — a 40-fold improvement. Because of this advance, the Genome Sequencing Platform can now accept samples that were previously considered unsuitable for sequencing.

"If you assemble one of the libraries, then you'll end up with an assembly. But if you assemble all three together using three different lengths, you get quite a bit better product," Sharpe says. "The nice thing with the Pippin Prep is being able to easily get those discrete size ranges."

Beyond precious samples with limited DNA, automated size selection has enabled a number of applications that were difficult or even impossible with a manual gel extraction approach, including ChIP-seq and a new method for massive-scale genotyping for hundreds of markers across hundreds or thousands of samples.

Case Study: Size Selection for ChIP-Seq

The large-scale study of chromatin immunoprecipitation with a next-gen sequencer (ChIP-seq) has gained traction quickly in the genomics field, but still has technical limitations due to the small amounts of sample typically available for such studies.

Thomas Westerling, a scientist at the Dana-Farber Cancer Institute's Center for Functional Cancer Epigenetics, evaluated the Pippin Prep platform for size selection, since that preparative step often contributes to sample loss or cross-contamination. He ran several ChIP studies on the Pippin, as well as on E-gel® SizeSelect™ gels for comparison purposes (Westerling).

Across the experiments, Westerling determined that Pippin Prep was well suited to the particular sample prep needs of ChIP-seq studies. Despite nanogram-scale starting material, Pippin returned good enrichment levels with strong agreement to the original sample.

In the study, Westerling noted that the Pippin platform produced higher-accuracy sizing with far less manual work than the E-gel alternative. "BioAnalyzer tracings confirm that the Pippin Prep size selection selected more accurate broad size ranges with good yield across the full size range selected," Westerling wrote in an application note detailing the work. "In contrast, the BioAnalyzer tracing for the E-gel run confirmed a lower than expected size range, with a marked absence of above 300bp fragments in the ChIP sample. In addition, the analysis revealed that smaller primer-dimers and adaptor-dimers were not adequately separated from the amplified ChIP DNA."

Case Study: Massively Multiplexed Genotyping

A paper published in PLoS One in May of 2012 reports a novel method enabled by automated size selection to perform low-cost, massively parallel genotyping that does not require prior knowledge of an organism's genome sequence (Case Study). The publication, entitled "Double Digest RADseq: An Inexpensive Method for *De Novo* SNP Discovery and Genotyping in Model and Non-Model Species," came from the Harvard University lab of Hopi Hoekstra, a professor in the departments of Organismic & Evolutionary Biology and Molecular & Cellular Biology (Peterson).

The method reported builds on reduced-representation genome sequencing, commonly called RADseq, improving the approach by lowering costs and increasing accuracy.

> The Hoekstra team found that Pippin automated size selection was more than twice as precise as even the best-case manual gel practitioner, with better yield as well.

RADseq deploys restriction enzymes across the genomes of many individuals, cutting at certain predefined sequences to generate a slew of fragments to interrogate. This 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. For applications ranging from evolutionary development to population studies to QTL mapping, having even a small fraction of the genome can be very informative.

Reducing the fraction of genome to sequence allows scientists to study far more genomes at a reasonable cost than would otherwise be possible. However, a major challenge for the utility of the RADseq approach has been the accuracy and reproducibility of size selection. For RADseq to work, every genome that has been randomly reduced by the restriction enzymes must produce the same selection of fragments for scientists to be able to learn anything by comparing them. Successful RADseq, therefore, relies on near-perfect size selection; manual gel extractions, on the other hand, are widely known to be subject to both operator-to-operator variability as well as each person's own variation in slicing.

Brant Peterson, PhD, a postdoctoral fellow in the Hoekstra lab and lead author on the paper, says that the need for such precise size selection has been a limiting factor for RADseq. If an experiment sampling many regions across many individuals winds up generating fragments that are not comparable, "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."

The Hoekstra lab acquired the Pippin Prep from Sage Science to determine whether automated size selection offered a precision that could make RADseq significantly more useful for population genetics or quantitative genetics studies, where scientists might need to examine hundreds of variants across hundreds or thousands of samples.

In this experiment, the Hoekstra team found that Pippin automated size selection was more than twice as precise as even the best-case manual gel practitioner, with better yield as well. For instance, 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.

With 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."

Conclusion

Just as DNA sequencing itself has seen major advances in automation, so too have the techniques for sample preparation. Manual gel preparation, which is one of the most time-consuming, laborious, and variable steps in the preparative process, is slowly but surely being replaced by automated solutions. The best offerings in automatic size selection improve coverage and yield while introducing reproducibility, efficiency, and the ability to multiplex. This tighter size distribution contributes to lower sequencing costs and higher-quality analysis. While the technology underlying automated size selection will be expanded to cover more areas, today it is already a superior alternative to manual gels for paired-end and mate-pair sequencing, bead template generation, ChIP-seq, and microRNA library isolation.

With 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."

References

Borgström E, Lundin S, Lundeberg J (2011) Large Scale Library Generation for High Throughput Sequencing. PLoS ONE 6(4): e19119. doi:10.1371/journal.pone.0019119

Bridger, Haley. (April 2012) A Sage partnership. http://www.broadinstitute.org/blog/sage-partnership

Case Study: New Take on RADseq Enables High-Throughput Variant Discovery. (2012) http://www.sagescience.com/wp-content/uploads/2012/08/sage_casestudy_Hoekstra_4_0712.pdf

Fisher, Sheila, et al. (2011) A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. Genome Biology 12:R1 doi:10.1186/gb-2011-12-1-r1

Peterson BK, et al. (2012) Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. PLoS ONE 7(5): e37135. doi:10.1371/journal.pone.0037135

Quail, M. A. et al. (2012) Evaluation and optimisation of preparative semi-automated electrophoresis systems for Illumina library preparation. Electrophoresis. Accepted article online. doi: 10.1002/elps.201200128

Sage Science blog. (May 17, 2012) At Emory, Size Selection Saves Time for NGS Prep. http://www.sagescience.com/blog/uncategorized/at-emory-size-selection-saves-time-for-ngs-prep/

Sage Science blog. (July 24, 2012) For NRC Team, Pippin Platform Leads to Higher-Quality Assemblies. http://www.sagescience.com/blog/uncategorized/for-nrc-team-pippin-platform-leads-to-higher-quality-assemblies/

Westerling, Thomas. (2011) Application Note: ChIP-Seq Library Prep. PDF downloaded from http://www.sagescience.com/wp-content/uploads/2012/08/Dana_Farber_Pipp_ChIP_Westerling-App_Note.pdf

For additional information, contact us at info@sagescience.com or 978-922-1932, or visit our website at www.sagescience.com.

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