Robust, Scalable, Low-Coverage Whole Genome Sequencing for High-Throughput Crop Genotyping

Ariele Hanek[1], Jessica Smith[1], Michelle Rahardja[1], Joe Mellor[1], Tobias Schughart[2], Maximilian Weiter[2] and Jesse L. Hoff[3]

[1]seqWell, Inc. Beverly, MA USA, [2]BioEcho Life Sciences GmbH, Cologne, Germany, [3]Gencove, Inc. Los Angeles, CA USA

Introduction

The reduced cost and improved scale of DNA sequencing has allowed a variety of genotyping assays to be approached with sequencing-based alternatives over legacy approaches such as microarrays. One key remaining bottleneck to unlock the efficient use of sequencing instruments for high-throughput genotyping is the creation of scalable workflows and chemistry for sample extraction and library preparation. Here, we demonstrate the robust combination of the EchoLUTION Plant DNA extraction kit (BioEcho Life Sciences) and plexWell™ Low-Pass 384 (LP384) Library Preparation kit (seqWell) across multiple plant species with a range of genome sizes. The combined workflow features normalized multiplexed libraries without the need for time-consuming individual adjustment of input DNA, significantly simplifying the complex task of high-level multiplexing and enabling faster, cost-effective, and reliable results for routine crop genotyping applications such as principal component analysis.

High Throughput Workflow Enables Time and Consumable Savings

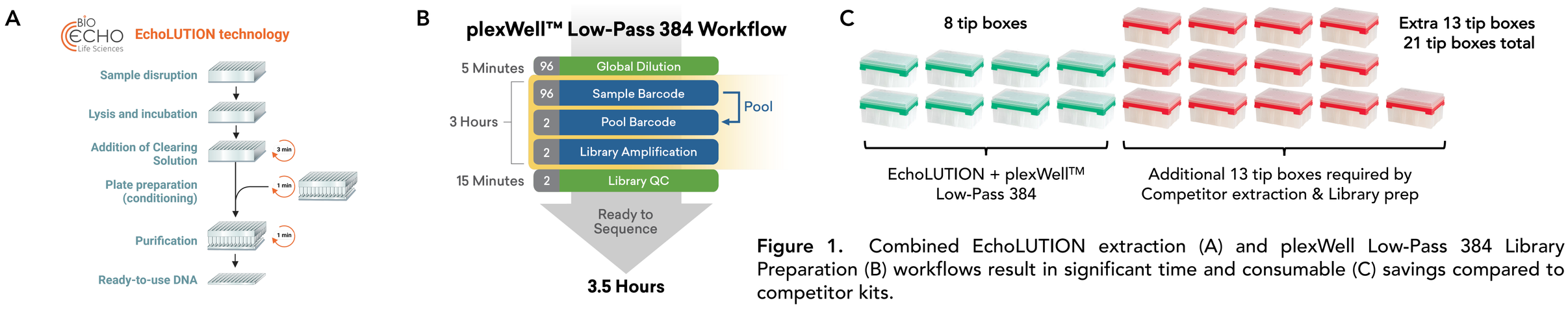

Figure 1. Combined EchoLUTION extraction (A) and plexWell Low-Pass 384 Library Preparation (B) workflows result in significant time and consumable (C) savings compared to competitor kits.

- A collection of plant seeds encompassing 3 varieties each of *Glycine max* (soybean), *Solanum lycopersicum* (tomato), and *Zea mays* (maize) was extracted in triplicate using the EchoLUTION Plant DNA kit (Figure 1A) to generate to 27 purified genomic DNA samples.
- Samples were globally diluted to an average of 10 ng by species. This approach resulted in sample inputs of 5.4-14.5, 8.3-12.2, and 7.0-11.7 ng, respectively, for soybean, tomato, and maize. Following dilutions, samples were processed in duplicate with the plexWell Low-Pass library preparation kit (Figure 1B), resulting in a normalized multiplexed library pool.
- The library was sequenced on the NextSeq2000 (P3 300). Sequencing data was down-sampled to nominal 2X coverage to control for different genome sizes, [1.1 Gb, 950 Mb, and 2.4 Gb for soybean, tomato and maize, respectively] prior to imputation analysis at Gencove and principal component analysis.

Data & Analysis

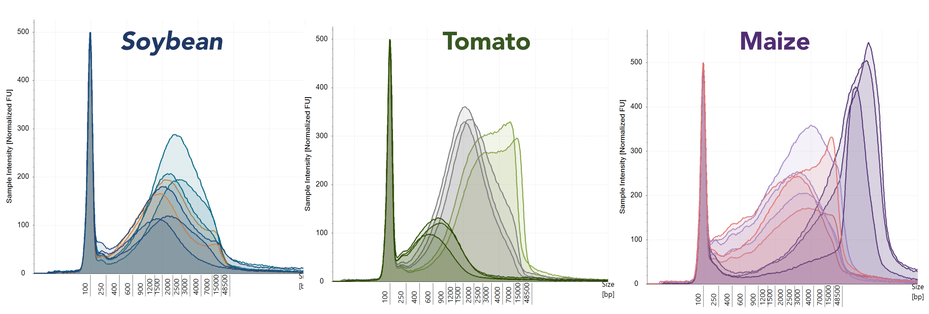

Figure 2. Replicate extractions result in consistent DNA fragment distributions and concentrations.

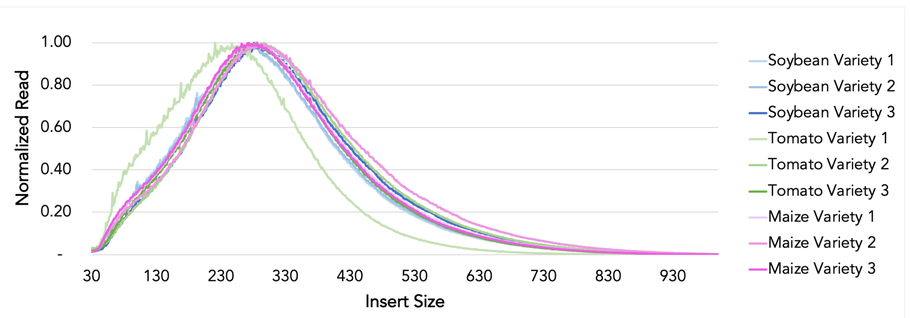

Figure 3. Library preparation produces consistent insert size.

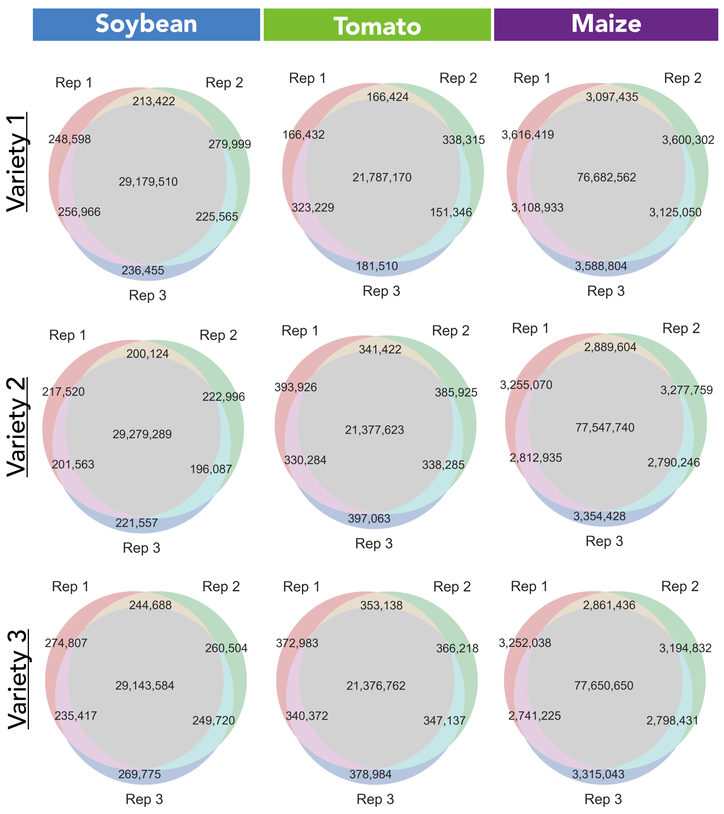

Figure 4. High SNP concordance amongst replicates of soybean (95%), tomato (92%), and maize (80%)

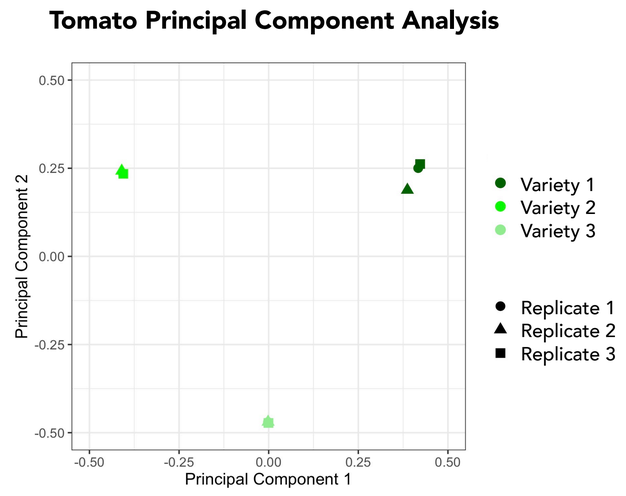

Figure 5. Principal component analysis for three tomato strains distinguishes between varieties. Principal component 1 (x-axis) accounts for 64% of variation while principal component 2 (y-axis) accounts for 18%. Samples clustering together are likely to have high sequence similarity. Replicates from each variety are clearly distinguishable clusters, indicating genetic differences between tomato strains can be resolved with low-pass sequencing.

Summary

- High-throughput processing is accelerated by consistent genomic DNA yield and fragment distribution during extraction (Figure 2), obviating the need to individually normalize samples prior to library preparation. Instead, a single global dilution can be applied to bring the DNA within the 5-25 ng auto-normalizing input range of the plexWell LP384 library preparation kit. Early pooling within the library preparation workflow further reduces the library QC and pooling burden prior to sequencing.
- In addition to read count, plexWell LP384 library prep normalizes insert size distribution (Figure 3) with median inserts of ~300 or greater, enabling full advantage of 2x150 sequencing while minimizing redundant data due to read overlap. Removal of additional small fragments is easily implemented by adjustment of the final pool purification, if desired.
- >98% of sequencing reads align to the expected reference and generate robust imputation results with >98% (soybean), 97% (tomato) and 91% (maize) of SNPs called with high confidence. Figure 4 demonstrates the high concordance amongst the called SNPs which can be utilized for downstream analysis such as principal component analysis (Figure 5). The robust combination of EchoLUTION and plexWell Low-Pass multiplexed NGS technologies provides a streamlined, scalable, and reproducible solution for routine crop genotyping applications.