An Automated Workflow for Multiplexed NGS Library Construction in Low-Coverage Whole-Genome Sequencing of *Cannabis sativa*

Michelle Rahardja[1], Ariele Hanek[1], Marcel Brun[2], Jessica Smith[1], Yanyan Liu[1], Joseph Mellor[1], Matt Stull[2], Emma Schulze[2], Alex Watros[2], Joshua Hill[2], Mark Scheel[1], Charles D. Johnson[2]

1. seqWell, Inc. Beverly, MA USA 2. Texas A&M AgriLife Genomics and Bioinformatics Service, College Station, TX USA

Introduction

- The flowering plant, *Cannabis sativa*, has been grown for centuries for a variety of purposes. In addition to producing a huge variety of secondary metabolites, it also exhibits significant phenotypic diversity. The current system of classification of *Cannabis* and its legal regulation is based on the chemistry characteristics, branching patterns, leaf shape and other phenotypic characteristics. However, these traits are not correlated when lineages are crossed, which is a common breeding practice.
- Low-coverage whole genome sequencing (lcWGS) is a cost-effective approach to genotyping. This "low pass" genome sequencing is a useful tool in agricultural population studies, as it allows for the generation of genomic information on a large scale, which can be used for classification purposes.
- Here, we demonstrate the ability to generate highly multiplexed pools of lcWGS libraries from *Cannabis sativa* using an automation workflow on a Sciclone G3 NGSx Workstation liquid handling workstation from PerkinElmer, Inc. using seqWell's purePlex™ library preparation technology for sequencing on an Illumina NovaSeq instrument.

Automated purePlex Library Prep Enables Scalable Low-Coverage Whole-Genome Sequencing

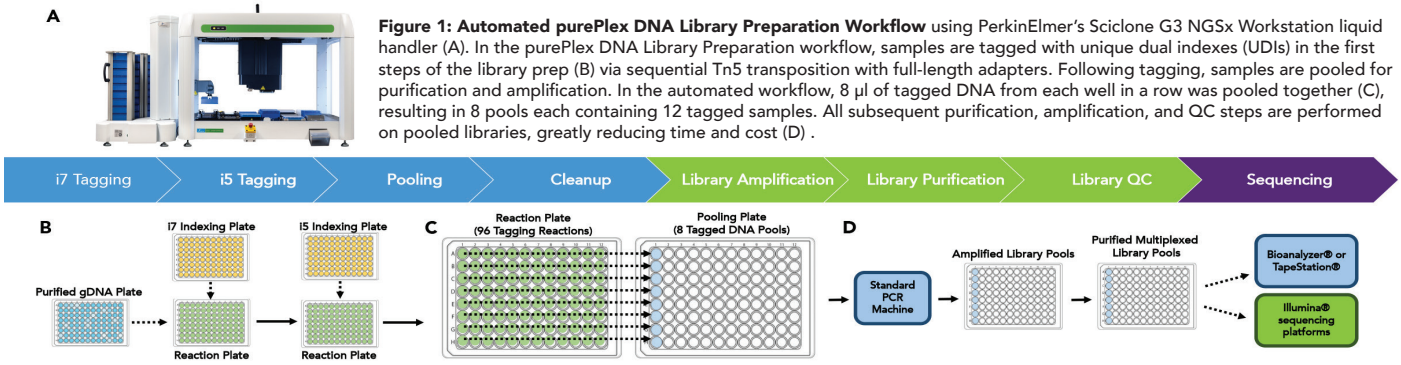

Figure 1: Automated purePlex DNA Library Preparation Workflow using PerkinElmer's Sciclone G3 NGSx Workstation liquid handler (A). In the purePlex DNA Library Preparation workflow, samples are tagged with unique dual indexes (UDIs) in the first steps of the library prep (B) via sequential Tn5 transposition with full-length adapters. Following tagging, samples are pooled for purification and amplification. In the automated workflow, 8 µl of tagged DNA from each well in a row was pooled together (C), resulting in 8 pools each containing 12 tagged samples. All subsequent purification, amplification, and QC steps are performed on pooled libraries, greatly reducing time and cost (D) .

Data and Analysis

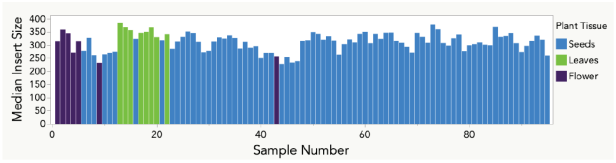

Figure 2: Consistent insert size for 95 *Cannabis sativa* samples.

Table 1: Sequence alignment and insert size by sample type

	Seeds	Leaves	Flowers
# of Samples	80	8	7
% Reads Aligned	97.6%	97.5%	98.4%
Average Median Insert	309	355	299

95 of 96 samples yielded sufficient sequencing data for further analysis. Poor DNA quality accounted for the failed sample. All sample types demonstrated high alignment rates and median inserts well suited to 2 X 150 Illumina sequencing (Table 1), and consistent insert size (Figure 2), characteristic of purePlex library prep. 12 samples were selected for principal component analysis (PCA), which distinguishes between varieties. The 12 samples (Table 2) comprise a parent-offspring trio as well as seeds and flowers from the same plant. Figure 3 demonstrates the successful application of low-coverage whole genome sequencing and PCA to distinguish sample varieties, those with high sequence similarity will co-locate on the PCA plot.

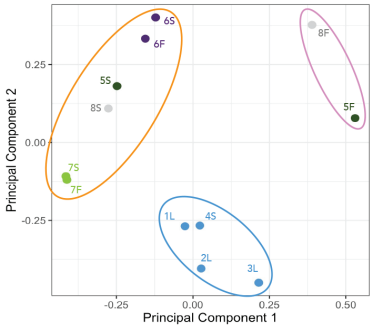

Table 2: Summary of samples analyzed for principal component analysis

Plant ID	Sample Type
Plant 1	Leaf – Parent
Plant 2	Leaf – Parent
Plant 3	Leaf – Offspring
Plant 4	Seed
Plant 5	Seed & Flower
Plant 6	Seed & Flower
Plant 7	Seed & Flower
Plant 8	Seed & Flower

Figure 3: Principal component analysis (PCA) for 12 cannabis samples. Samples that cluster together are likely to have high sequence similarity. Use of lcWGS data here shows 3 distinct clusters of samples. Further demonstrating the validity of this approach, the offspring (3L) co-locates near the parents (1L and 2L). Seed and flower samples from Plant 6 and 7 co-locates. Seed and flower don't always co-locate due to contaminating DNA on the flower samples.

Methods

- DNA was extracted from an assortment of *Cannabis sativa* seeds, leaves, and flowers using the EchoLUTION Plant DNA kit (BioEcho Life Sciences).
- All samples were diluted to 5.0 ng/µl prior to library preparation.
- Multiplexed library pools were prepared using the PerkinElmer Sciclone G3 NGSx Workstation liquid handler with purePlex DNA library preparation kit.
- The resulting pools were sequenced on an Illumina NovaSeq6000 SP 300 cycle kit.

Acknowledgment

- Thank you to BioEcho Life Sciences for the loan of bead beating equipment and donation of extraction kits.
- Thank you to David R. Gang (Washington State University) and Jessica Tonani (Verda Bio) for the *Cannabis* samples.

Summary

- The purePlex DNA Library Prep Kit supports truly multiplexed and highly scalable construction of library pools for low-pass WGS by generating reproducible library characteristics despite sample plant tissue origins (Figure 2, Table 1). A 300–400 bp insert size is optimal for generating unique data from a 2 X 150 run and increasing the usable data per dollar of sequencing.
- An automated workflow of the purePlex DNA Library Prep accelerates high-throughput processing of low-coverage whole genome sequencing, which can be utilized in agricultural population studies for classification purposes (Figure 3).