APPLICATION OF OXFORD NANOPORE MINION PLATFORM IN CANNABIS GENOMICS AND METAGENOMICS: FIRST IMPRESSIONS

John Brunstein, May Cui*, Jerian Reynolds, and Ying Ng
Genotyping and Molecular Lab Services Division, Segra International Corporation, Richmond, B.C. Canada V6W 1M2

Abstract:
The recent emergence of Cannabis cultivation as a legal mainstream activity has opened up opportunities for genomic studies to establish information on plants with other high value industrial plant crops. In particular, next generation sequencing (NGS) technologies show promise in identifying single and multi-trait Mendelian and Quantitative Trait loci as markers for selective breeding programs; in the characterization of structural and copy number variations; and through metagenomics in the discovery, identification, and characterization of associated and possibly pathogenic microorganisms including bacteria, fungi, viruses, and viroids. In contrast to more mainstream NGS platforms with high equipment and per-run costs, the Oxford Nanopore (ONT) MiniION technology has potential to democratize the application of NGS technologies by providing a relatively low entry cost platform with a scale well suited for application to Cannabis. We report here on our initial experiences of application of this platform to Cannabis genomics and metagenomics. These include preliminary development of an inexpensive and technically simple but effective dried plant material DNA isolation protocol enriched for longer fragments well suited to nanopore sequencing; observations on data yield and associated costs and times downstream of the preparation method, with discussion of scalability; a consideration of our experimentally observed read length distributions in the context of potential advantages of long read technologies over short read technologies in establishing genetic scaffolds in the presence of long repeat regions; and a summary of bioinformatics challenges, resources, and possible solutions for Cannabis laboratories considering gathering NGS data. Finally, we present examples of genomic and metagenomic data we have obtained with this method.

Samples:
DNA flower material was obtained from the BC Government online store and accessioned with linked metadata into a custom LIMS for data tracking.

Initial Sample Extraction and QC:
Samples were extracted using Qiagen DNAeasy Plant Mini kits or Thermo Fisher KingFisher Duo Prime MagMAX Plant DNA kits in accordance with manufacturer's instructions. Where large amounts of input DNA were desired, multiple parallel Qiagen extractions were performed with eluates pooled, ethanol precipitated, and resuspended in smaller volume prior to further handling. Eluates were quantified by Qubit and evaluated by A260/A280 for purity. Samples were further evaluated by VNTR genotyping and selected for whole genome sequencing based on phylogenetic considerations.

Sample pre-preparation for Nanopore Sequencing:
We evaluated several pre-process approaches to handling DNA prior to NGS library preparation:
• Direct use of DNA extracts;
• Size fractionated DNA extracts;
• Whole Genome Amplified (WGA) product from DNA extracts; and
• WGA subsequent to size fractionation of DNA extracts. We present data on each of these approaches below.

Whole Genome Amplification (WGA):
Where employed, WGA was performed by Qiagen REPLI-g mini kit per manufacturer's instructions. Post WGA, products were treated with T7 endonuclease digestion to remove hyperbranched products.

Size Fractionation by PEG 8000:
Due to the long read nature of the ONT technology, it is beneficial to preselect DNA library material to minimize short fragments in favour of longer fragments which provide greater data output per successful pore interaction. There exist multiple devices for fragment size preselection for NGS, however most are costly and thus negate one of the appeals of ONT technology (low capital cost). With this in mind, we chose to employ fractional precipitation of eluates by PEG 8000 as a cost effective means to enrich DNA pools for longer strand material.
We evaluated various concentrations of PEG 8000 (4%, 6%, 8%, and 10%). Representative data from a 4% cut and second 10% cut are shown in Table 1 and Figure 1, below.

Table 1:
<table>
<thead>
<tr>
<th>Stage</th>
<th>Conc μg/ml</th>
<th>volume μl</th>
<th>total ng</th>
<th>recovery</th>
</tr>
</thead>
<tbody>
<tr>
<td>raw eluate</td>
<td>35.7</td>
<td>2000</td>
<td>71400</td>
<td>100%</td>
</tr>
<tr>
<td>Pooled EIOH ppt</td>
<td>63.2</td>
<td>1000</td>
<td>63200</td>
<td>88.50%</td>
</tr>
<tr>
<td>4% PEG</td>
<td>14.1</td>
<td>134</td>
<td>1889</td>
<td>2.60%</td>
</tr>
<tr>
<td>10% PEG</td>
<td>570</td>
<td>120</td>
<td>68400</td>
<td>96%</td>
</tr>
<tr>
<td>Supernatant</td>
<td>0.127</td>
<td>1400</td>
<td>178</td>
<td>0.20%</td>
</tr>
</tbody>
</table>

Results: Metagenomics Applications:
One of the easiest applications of resulting NGS data is in the analysis of metagenomics – that is, what else other than cannabis DNA was present in each sample. Beyond pure curiosity, such data may prove useful in identifying novel or adventitious pest organisms (bacterial, fungal, viral, or viroid in nature). The WIMP (What's In My Pot) application on the Epi2Me bioinformatics portal available to ONT users provides a rapid analysis of this, and one which can be done meaningfully even on relatively low coverage data sets.

Figure 2 represents one such result, from only approximately 1 Gbase (~1x genome coverage).

Figure 1: Fragment size distribution of 4% PEG 8000 cut

Library Preparation and Flow Cells:
Following pre-processes as described above, all templates were prepared for sequencing by the ONT SQK-LSK109 protocol (1D Genomic DNA by Ligator) [1] with all ancillary reagents as recommended. We adopted the extended incubation period modifications (15 minutes for end preparation, 30 minutes for adapter ligation) suggested by Nippon Genetics [2]. Prepared libraries were loaded on FLO-MIN106 flow cells and runs controlled by MinIT dedicated computer device.

Summary of Observations:
1. Not unexpectedly, direct sequencing of low amounts of poor quality DNA gave poor results ~ 2 Mbase total per flow cell in 24 hours.
2. 6% PEG 8000 provided a good balance between size fractionation and mass recovery.
3. Simple pre-process methods greatly improved results.
   a) 1.5 μg pooled, EIOH precipitated Qiagen kit extract yielded ~1 Gbase of sequence in 48 hours
   b) 1.5 μg post-WGA material derived from 20ng KingFisher extract yielded ~8 Gbase of sequence in 48 hours, with average read lengths about 1 kb
c) 1.5 μg post-WGA material derived from 20ng 6% PEG preselected Qiagen extract yielded ~7 Gbase of sequence in 48 hours, with average read lengths about 3 kb (see Figure 4).
4. Removal of PEG 8000 from library templates is essential prior to loading on flow cells.

Data Handling and Bioinformatics:
In general, the increasing technical simplicity and dropping cost barriers to NGS have outstripped bioinformatics; obtaining the data is simpler than meaningfully interpreting it in many cases. Multiple options are none the less available for the novice through more advanced bioinformatics to process the ONT data. Some of these include:
• Epi2Me. Limited but simple functionality; good for metagenomics.
• Cyverse. Free and hosts GUI driven forms of multiple useful tools, but primarily a training and education environment not suited for large data sets.
• BaseSpace (illumina). A GUI driven bioinformatic solution on the AWS platform, BaseSpace provides scalable computational power and many useful NGS applications pre-packaged for novice to intermediate users at reasonable per-compute-load cost. Unfortunately, at present this platform neither provides a simple way to accept long read raw data, nor many of the applications optimized for this data type.
• CLC Genomics Workbench (Qiagen). A Windows-based local compute solution, this is capable of conducting reference assemblies from 8-10 Gbase ONT data sets vs. published cannabis chromosome sequences in reasonable time frames (1 – 2 hours depending on hardware, whole genomes within a day). Figure 3 is an example of the coverage map for a Chromosome 1 reference alignment derived from an 8 Gbase (~9x coverage) run.

Figure 3: Reference Alignment Coverage vs Purple Kush Chromosome 1

Although de novo assembly routes are available in this package, we found these unable to run effectively on 10x coverage data sets. The software can also treat one or more ONT data sets as databases against which to conduct BLAST searches (particularly useful for identifying multiple gene copies in a specimen).

• Cloud based Linux clusters, while requiring the most innate bioinformatics skill, remain the best alternative for most meaningful bioinformatics processing of data sets of the size obtained here (7-10x genome coverage).

Conclusions:
We have found the ONT MiniION system to provide a simple, cost effective platform for on-demand sequencing of Cannabis whole genomes. When linked to WGA, only small amounts of input material (~20ng) are required and a few hours total work in sample and library preparation routinely yields meaningful (~10x) genome coverage.

Figure 4: Read Length Distribution on Library post size fractionation and WGA

References: