INTRODUCTION
As part of Life Technologies’ partnership with SG Biofuels, we have undertaken the de novo assembly and annotation of the Jatropha Curas plant genome using Life Tech’s next generation sequencing platforms, SOLID™ and ION PGM™. Jatropha is an important plant for bio-fuel research. Oil from Jatropha seeds can be used to produce biodiesel along with a biomass residue that can be used in electric power plants. Its genome is estimated to contain 380-420 million base pairs (Mb) over 11 diploid chromosomes. The goal of the collaboration is to use sequencing to guide the development of a cultivar that optimizes bio-fuel production.

De novo assembly projects that use NGS platforms typically target small bacterial genomes with less than 50 Mb. Fundamental limits of the assembly algorithms along with practical limits on computational resources make full de novo assembly using only short reads difficult or impossible for medium to large-scale genomes. To scale out to larger genomes, most assembly projects augment NGS data with data acquired from methods that generate longer reads, such as Roche’s 454 or traditional Sanger sequencing.

The size and relative simplicity of the Jatropha genome make it a good target for pushing the limits of de novo assembly using a combination of the SOLID™ and PGM™ platforms. The goal of this work is to use SOLID™ to generate deep coverage of the genome while using the PGM™ to generate longer reads (150 bp) to bridge contigs and fill in gaps.

To our knowledge, this is first genome assembly project to combine SOLID™ and PGM™ reads and the largest genome assembled using SOLID™ reads.

METHODS
We used two sets of sequencing runs to create the reference genome. A long mate-paired SOLID™ library sequenced using SOLID™ 4.0 ECC chemistry generated high-quality short reads at 100x coverage of the genome (464M reads with 3kb inserts, 50 base reads from each end). To validate the run, the SOLID™ reads were mapped to a recently published reference of another jatropha strain [1], a jatropha chloroplast reference, a collection of jatropha ESTs, and a related cassava species.

We evaluated three assemblers, SOAPdenovo [2], ABySS, and Velvet. SOAPdenovo performed the best overall and was used to generate the assembly. All assembly jobs were run on a high-memory (192 GB) compute node.

The initial draft assembly was used to identify putative SNPs between our strain and the Sato reference along with SSRs (simple sequence repeats) to be used as biomarkers.

To generate the longer reads necessary for filling in gaps and bridging the scaffolds, we ran the samples on eight 316 chips using the ION PGM™ and generated approximately 12 million reads spanning 1.8 Gb for roughly 4.5x coverage of the genome.

We used the PGM™ reads to generate two separate assemblies. The first was a simple assembly using only the PGM™ reads and the second combined both the SOLID™ and PGM™ reads.

RESULTS

DE NOVO ASSEMBLY USING SHORT READS
While small genomes are routinely assembled using short reads, large genomes present unique challenges. The number of reads required to attain sufficient depth is a function of the genome size, with approximately 80-100x coverage needed. For a genome the size of Jatropha, this translates to our 800 M 50x50 mate-paired reads for sufficient fragment coverage.

The basic process for assembly is:
1. Split the reads into k-mers, typically 25-29 bp.
2. Remove low frequency k-mers as an error correction step.
3. Build a graph of the k-mers, with each node representing an overlap between two k-mers.
4. Traverse the graph to create contigs.
5. Map reads back to contigs and use mate-pair information to join separate contigs into scaffolds.
6. Repeat with additional sequencing runs to improve the assembly.

ASSEMBLY RESULTS

The SOLID™ assembly, with 279 Mb, is close in size to the published Sato assembly. However, most of the scaffolds consist of contiguous sequences at each end with a large gap in the middle.

The contigs built using only the PGM™ reads were generally in line with the read length, which was expected giving the overall coverage generated by 8 PGM™ runs.

COMPARISON WITH SATO REFERENCE
Comparing the SOLID™ reads against the Sato Jatropha reference genome provided a QC check on the reads and allowed us to identify putative SNPs between the different sub-species used to create the assemblies.

- 273,294,043 bases in the Sato reference are covered by SOLID™ reads (98.7%).
- The average coverage depth is 89 reads per base in the reference.
- F3 Reads
  - with at least one reported alignment: 281,011,417 (60.49%) reads that failed to align: 183,559,376 (39.51%)
- R3 Reads
  - with at least one reported alignment: 310,377,737 (66.80%) reads that failed to align: 154,243,889 (33.20%)

Distribution of SNPs Across the Sato Reference

1,626,639 putative SNPs were detected covering 0.588% of the bases in the reference.

CONCLUSIONS
Over the course of this project, multiple advances were made by the de novo assembly community for generating assemblies from short reads (e.g., [3,4]). The most important was the realization that multiple libraries with insert sizes varying from a few hundred to a few thousand bases are essential for successful assembly of larger genomes. Our SOLID™ run, with 3kb inserts, provides a single set of such reads. The PGM™ runs attempted to cover the lower end, but the PGM™ reads were too few and, likely, too short to effectively cover the range of insert sizes recommended for de novo assembly. However, as PGM™ reads get longer they should be an effective substitute for mate-paired reads with short insert sizes for assembly projects.

Another challenge faced by assembly projects is the sheer scale of the data and the compute resources required by the assemblers. The working data for the SOLID™ assembly was over 2 Tb in size. To complete the assembly in a reasonable amount of time, the assembler required a node with 200 GB of RAM. Large, fast storage systems and high-memory nodes are expensive and rarely found in biology labs. Software improvements will only have a marginal effect on data and resource usage for assembly projects. The size of the target genome dictates the amount of reads (and as a result, compute resources) required for assembly.

De novo assembly using short reads, while still in its infancy, is a promising application for next-generation sequencing platforms. The early results from this project and work done elsewhere suggest that it will be possible to develop a scalable protocol for genome assembly using a combination of SOLID™ and PGM™ runs. For small genomes, the PGM™ is already effective. For larger (50M+) genomes, a combination of SOLID™ long mate-pairs with progressive insert sizes for depth and scaffolding and PGM™ reads for longer fragments along with robust assembly software will make a strong assembly platform from Life Technologies’ products.

REFERENCES

ACKNOWLEDGEMENTS
- Nathan Wood from Life Technologies and Bob Schmidt at SG Biofuels initiated the collaboration.
- Jason Myers and Ilya Zlatkovsky at ION West performed the PGM™ sequencing.
- Dumitrul Birnza from LT provided assembly expertise.
- Antoine Uzzeni and LT IT helped setup the high-memory node used for assembly.

Christopher Mueller1, Eric Mathur2, Joel Brockman3, and Bob Setterquist4 (1) Life Technologies 2130 Woodward St., Austin, TX 78744 (2) SG Biofuels 11260 El Camino Real, San Diego, CA 92130