DOE Joint Genome Institute FY15 Q1 Progress Report
Computational developments for complex genome assembly and genetic mapping at the DOE JGI

Q1 Metric: Report on new computational developments for improving microbial, metagenomic, or plant genome assemblies

Background

A critical problem for computational genomics is the problem of "genome assembly": the development of robust scalable methods for transforming short randomly sampled "shotgun" sequences into the contiguous and accurate reconstruction of genomes. Currently, advanced methods exist for satisfactorily assembling the small haploid genomes of prokaryotes from both short and long read sequencing technologies. While also comprised of haploid cells, assembly of sequenced bacterial communities (metagenomes) and single cells are complicated by the non-uniform representation of genomic regions in the data and for that reason they represent special cases that, while the focus of intensive efforts at JGI, will not be further discussed in this report. The genomes of eukaryotes are more complex than bacteria & archaea. In particular the large, polyploid, repetitive, and outbred genomes of plants, including those of important biofuel feedstock candidates such as switchgrass, miscanthus, and sugarcane, are especially challenging, combining complexity with scale. This report is focused on recent JGI advances in assembling these genomes.

One challenge is that plant genomes can be large; for example, the genome of wheat is 17 gigabases (Gbp), nearly six times larger than human. The expanded genome size of plants is due to several factors. Replication and transposition of repetitive elements can inflate a genome. Repetitive elements within and between genes make the determination of local and global linkage along a chromosome difficult to decipher. The detailed positioning of these elements can affect function by altering gene regulation, so they cannot be dismissed as simply "repetitive" elements to be ignored relative to gene space.

A second factor common to many plants is polyploidy, the presence of two or more closely related ("homeologous") sub-genomes in each cell. Polyploidy not only multiplies the size of a genome, but also provides an additional challenge for genome assembly, since each genomic segment then has one or more closely related "homeologous" segments that must be separated computationally. Many DOE JGI flagship plants, including the biofuel feedstocks switchgrass, miscanthus, and sugarcane, are polyploid.

Finally, while some plants can be inbred or rendered into doubled haploids, many (especially the biofuel feedstocks noted above) are outbreeding species with levels of heterozygosity (allele-allele variation) that can be an order of magnitude larger than found in human genomes. Assembly of heterozygous genomes must differentiate allelic variation from differences between other highly similar sequences, including repetitive elements and homeologous genes. To produce a single composite reference sequence, these allelic variants must be identified and a single variant chosen, but a more general solution would be to "phase" the variants to recover the multiple haplotypes found in the genome. One general method for phasing is to use the sequence of a defined cross for linkage mapping.

The value of robust genome assembly is clear. The starting point for characterizing the genes of an organism is a genome assembly, and for comparative analysis across species. The assembly forms the basis for assessing genetic variation, including single nucleotide polymorphisms, insertions and deletions, and larger-scale structural variants. Finally, genome assembly provides a reference against which new sequenced based methods for assessing

gene regulation and dynamics can be aligned. For all these reasons, this has been an important focus of the DOE JGI computational genomics group.

Progress

In the past year substantial progress has been made through the development of novel algorithms for whole genome shotgun assembly and large-scale "shotgun" genetic mapping. Many of these efforts are part of collaboration between DOE JGI computational scientists and researchers at the Computational Research Division (CRD) of LBNL, and other scientific collaborators at the DOE JGI.

1) **Development of meraculous2.** Meraculous2 is a state-of-the-art de novo assembler for short reads developed at the DOE JGI. Meraculous is a hybrid assembler that combines aspects of several earlier assemblers, incorporates base quality scores, and shorter sequence reads to yield nearly error-free contiguous sequences. Relative to the original Meraculous, the new version deals more effectively with allelic variation, has improved gap closing, and an improved scaffolding approach that produces more complete assemblies . The speed and bandwidth efficiency of the new parallel implementation have been substantially improved, allowing the assembly of a human genome to be accomplished in 24 hours on the DOE JGI/NERSC Genepool system. A previous version of Meraculous was at or near the top in scaffolding completeness and accuracy, and sequence accuracy, in the Assemblathon II comparison with other state-of-the-art assemblers (Bradnam et al 2013). A manuscript currently in review (Chapman et al., 2015b) highlights the features of Meraculous2 by presenting the assembly of the diploid human genome NA12878, and comparing it with previously published assemblies of the same data using other algorithms. The Meraculous2 assemblies are shown to have better completeness, contiguity, and accuracy than other published assemblies for these data. Practical considerations including pre-assembly analyses of polymorphism and repetitiveness are described.

2) **Licensing of Meraculous through LBNL Technology Transfer Office.** Meraculous2 software is publicly available in a Perl & C++ implementation and supports "local" and "cluster" modes of execution where the latter can be configured to run on any Grid Engine-line cluster with relatively little effort. The software, along with installer, user manual, and a test dataset, is freely available at
http://sourceforge.net/projects/meraculous20/
For more information, also visit http://DOE JGI.doe.gov/data-and-tools/meraculous/
A handbook of written protocols for using meraculous is available at
http://1ofdmq2n8tc36m6i46scovo2e.wpengine.netdna-cdn.com/wp-content/uploads/2014/12/Manual.pdf

3) **Parallelization of meraculous for use in high performance computing (HPC) settings.** In collaboration with members of the LBNL Computational Research Division and the UC Berkeley Computer Science Department, two of the most computationally intensive steps of the Meraculous algorithm have been adapted for use with Unified Parallel C (UPC) (Georganas *et al*. 2014). These new parallel algorithms enable assembly calculations to be performed rapidly, with linear scaling over thousands of nodes, such as are available at the DOE National Energy Research Scientific

Computing Center (NERSC) at LBNL. The most time-consuming phases of Meraculous have been parallelized in this way. The memory requirements for this stage have been reduced by a factor of nearly 7-fold using Bloom filters and probabilistic counting techniques, and remarkable scaling up to 15K cores for this I/O- and communication-intensive computation was achieved (Georganas *et al.* 2014).

To overcome a previous memory bottleneck, we devised a novel algorithm that takes advantage of novel communication optimizations and the design of a lightweight synchronization scheme. Overall results show unprecedented performance and efficient scaling on up to 15,360 cores of a Cray XC30, on human genome as well as the challenging wheat genome, with performance improvement from days to seconds.

This work shows that efficient utilization of distributed memory architectures is possible for this irregular problem. Hence, it removes the requirement of large memory machines by exploiting Unified Parallel C's PGAS (Partitioned Global Address Space) capabilities. Our implementation is portable and executable on both shared- and distributed-memory architectures without modifications.

4) **Development of efficient methods for shotgun genetic mapping.** Long-range linkages in assembly are typically provided by mate-pairs at modest separations up to tens of kilobases. Longer-range linkage data is difficult to generate, but organisms naturally maintain coherence of entire chromosomes through meiosis. High-density meiotic (linkage) maps can therefore be used to achieve chromosome-scale assembly if methods can be developed for producing linkage maps with millions of sequence-based markers. The current generation of genetic mapping algorithms was designed for the small data setting. These methods are limited by the prohibitively slow clustering algorithms they employ in the genetic marker clustering stage for millions of markers. We have developed a new approach to genetic mapping, BubbleCluster, based on a fast clustering algorithm that exploits the geometry of the data (Strnadova, *et al.* 2014). Our theoretical and empirical analysis shows that the algorithm can correctly recover linkage groups. Using synthetic and real-world data, including the grand-challenge wheat genome (JGI is part of an international consortium working on the wheat genome), we demonstrate that our approach can quickly process orders of magnitude more genetic markers than existing tools while retaining — and in some cases even improving — the quality of genetic marker clusters.

While current approaches to genetic mapping typically use algorithms that scale quadratically in the number of markers, our approach exploits the underlying linear structure of chromosomes to avoid expensive comparisons between (quadratically) many pairs of markers. The resulting linkage groups (i.e., marker clusters) are highly concordant with computationally expensive quadratic calculations, but our improved scaling allows far denser maps to be constructed with minimal computation. After the formation of linkage groups, the next step in constructing a high quality genetic map is inferring the detailed ordering of markers along chromosomes. Since our method takes into account the linear structure of chromosomes from the start, the result is an approximate marker ordering

that is an excellent starting point for detailed marker order by simulated annealing or other methods that explore short-range perturbations of our approximate ordering.

An important application of our method is in the efficient construction of ultra-dense genetic maps for large and complex genomes that are filled with repetitive sequences that frustrate genome assembly but do not limit the number of genetic markers.

The most economically important of these genomes are various grasses, including crops grown for food (e.g., barley and wheat, whose genome sizes are two- to seven-fold larger than the human genome) or as biofuel feedstocks (*e.g*, switchgrass and miscanthus, polyploids that contain multiple, subtly different copies of a basic genome). We contributed a genetic map for hexaploid wheat produced using our method to the recent chromosome-scale shotgun assembly publication by the the International Wheat Genome Sequencing Consortium (IWGSC 2014).

5) **Whole genome shotgun assembly of hexaploid wheat.** Polyploid species have long been thought to be recalcitrant to whole-genome assembly. By combining high-throughput sequencing, recent developments in parallel computing, and genetic mapping, we have derived, *de novo*, a sequence assembly representing 9.1 Gbp of the highly repetitive 16 Gbp genome of hexaploid wheat, *Triticum aestivum*, and assigned 7.1 Gb of this assembly to chromosomal locations (Chapman et al. 2015, in press). The genome representation and accuracy of our assembly is comparable or even exceeds that of a chromosome-by-chromosome shotgun assembly (IWGSC 2014). Our assembly and mapping strategy uses only short read sequencing technology and is applicable to any species where it is possible to construct a mapping population. The Chapman et al. 2015 manuscript also describes a new class of genetic markers that are easily genotyped by shotgun sequencing without requiring alignment to a reference.

6) **K-mer-based alignment.** Aligning a set of query sequences to a set of target sequences is an important task in bioinformatics, and is especially time consuming for genome assembly, where billions of reads must be aligned to a *de novo* assembly (or intermediate). Since this alignment is done once, it benefits from indexing methods that are designed to be used to align multiple datasets to the same assembly. We have developed merAligner, a highly parallel sequence aligner that implements a seed–and– extend algorithm and employs parallelism in all of its components (Georganas et al. 2015). MerAligner relies on a high performance distributed hash table (seed index) and uses one-sided communication capabilities of the Unified Parallel C to facilitate a fine-grained parallelism. We leverage communication optimizations at the construction of the distributed hash table and software caching schemes to reduce communication during the aligning phase. Additionally, merAligner preprocesses the target sequences to extract properties enabling exact sequence matching with minimal communication. Finally, we efficiently parallelize the I/O intensive phases and implement an effective load-balancing scheme. Results show that merAligner exhibits efficient scaling up to thousands of cores on a Cray XC30 supercomputer using real human and wheat genome data while significantly outperforming existing parallel alignment tools.

Impact:

- *Assembly*. As noted above, three of the most computationally intensive stages of the Meraculous assembly pipeline have been adapted for use with Unified Parallel C, enabling these stages to be run at NERSC in minutes across thousands of cores. These developments have dramatically improved the speed of genome assembly. Large-scale assemblies like the wheat genome previously required prohibitively large shared memory nodes, but are now accessible through the parallel implementation that removes this memory bottleneck and allows nearly linear speedup with the number of cores. When the entire Meraculous pipeline is adapted for Unified Parallel C, rapid de novo assemblies will be accessible through NERSC's high performance computing systems. This in turn will allow multiple assemblies to be run for each dataset, enabling optimizations that were previously inaccessible. Rapid de novo assemblies will also allow the de novo detection of genetic variation and structural variants, complementing alignment-based variant detection.
- *Genetic mapping*. The development of shotgun genetic maps for barley and wheat are already complete, but ongoing efforts for miscanthus and switchgrass will be instrumental in enabling chromosome-scale sequences for these biofuel feedstocks and DOE JGI flagships. These shotgun mapping methods are being integrated into other eukaryotic genome projects at the DOE JGI.

Ongoing developments:

- *Development of a complete Meraculous-UPC implementation***.** A goal for the remainder of FY2015 is the complete parallelization of the Meraculous2 pipeline. We estimate that genomes like human will be capable of assembly in minutes at NERSC's Edison system.
- *Optimization of BubbleCluster algorithm to allow for missing data*. Light shotgun sequencing provides a low-cost way to genotype individuals at a dense set of genetic markers. To keep sequencing costs down, however, we must target low depth (e.g., ~1x), which generates many missing data points. In our experiments with simulated data, we found a high concordance between representative point order and the simulated map order, with a high ρ in most cases and a perfect order in many examples with 35% missing data, but these methods are still under development.
- *Adaptation to metagenomes*. The core algorithms of meraculous have been designed for and applied to whole genome random shotgun datasets where a single genome is uniformly covered at a single depth. There are, however, natural applications of these methods to metagenomes, which consist of a mixture of different microbial species, each with their own abundance and intra-specific variation. We anticipate that the core algorithms of Meraculous can also enable high-throughput metagenome assembly, pending characterization of the impact of intra-specific variation and sequencing error in this context.
- *K-mer based alignment with other alphabets*. The core methods developed for merAligner (Georganas et al. 2015) are, in principle, applicable to alignments with

other alphabets besides the four letter DNA code. In particular, they may be suitable for some protein sequence alignment problems. This will require additional testing and development.

DOE JGI team: Jarrod Chapman, Eugene Goltzman, Isaac Ho, Daniel Rokhsar (DOE Joint Genome Institute);
Jeremy Schmutz and Jerry Jenkins (DOE Joint Genome Institute and HudsonAlpha Biotechnology Institute)

External collaborators:

- **UPC Meraculous and dense mapping algorithms:** Evangelos Georganas, Aydin Buluc, Leonid Oliker, Kathy Yelick (LBNL Computational Research Division); Veronika Strnadova, and John Gilbert (UC Santa Barbara)
- **POPSEQ and wheat genome**: Martin Mascher and Nils Stein (Leibniz Institute of Plant Genetics and Crop Plant Research); Gary Muehlbauer (University of Minnesota); Jesse Poland (Kansas State University); Robbie Waugh (The James Hutton Institute, Dundee, UK)

Related publications

1. K.R. Bradnam, J.N. Fass, A. Alexandrov, *et al.* (2013). "Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species." *Gigascience*. 22;2(1):10.
2. M. Mascher, G.J. Muehlbauer, D.S. Rokhsar, J. Chapman, J. Schmutz, K. Barry, M. Muñoz-Amatriaín, T.J. Close, R.P. Wise, A.H. Schulman, A. Himmelbach, K.F. Mayer, U. Scholz, J.A. Poland, N. Stein, R. Waugh (2013). "Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ)." Plant J. 76(4):718-27. (November 2013).
3. International Wheat Genome Sequencing Consortium (IWGSC) (2014) "A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome." *Science*. 345(6194):1251788.
4. E. Georganas, A. Buluc, J. Chapman, L. Oliker, D. Rokhsar, & K. Yelick (2014). "Parallel de bruijn graph construction and traversal for de novo genome assembly." Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC'14).
5. V. Strnadova, A. Buluc, J. Gonzalez, S. Jegelka, J Chapman, J. Gilbert, D. Rokhsar, and L. Oliker (2014). "Efficient and accurate clustering for large-scale genetic mapping." The IEEE International Conference on Bioinformatics and Biomedicine (BIBM'14), Belfast, UK, November 2014.
6. E. Georganas, A. Buluc, J. Chapman, L. Oliker, D. Rokhsar, & K. Yelick (2015). "merAligner: A Fully Parallel Sequence Aligner." Proceedings of the 29th IEEE International Parallel & Distributed Processing Symposium (accepted).
7. J. Chapman, M. Mascher, et al. (2015a). "A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome." *Genome Biology* (in press)
8. 8. J. Chapman, I.Y. Ho, E. Goltsman, D.S. Rokhsar. (2015b) "Meraculous2: fast accurate short-read assembly of large polymorphic genomes." in preparation.