Title:	**Genome Improvement at JGI-HAGSC**
Project ID:	**0005128**
Prog Mgr:	**Marvin Stodolsky Phone: 301-903-4475 Division: SC-23.2**
PI:	**Richard M. Myers**
Award Register#:	**ER62873**

9/01/09 – 8/31/10

Since the completion of the sequencing of the human genome, the JGI has rapidly expanded its scientific goals in several DOE mission-relevant areas. At the JGI-HAGSC, we have kept pace with this rapid expansion of projects with our focus on assessing, assembling, improving and finishing eukaryotic whole genome shotgun (WGS) projects for which the shotgun sequence is generated at the Production Genomic Facility (JGI-PGF). We follow this by combining the draft WGS with genomic resources generated at JGI-HAGSC or in collaborator laboratories (including BAC end sequences, genetic maps and FLcDNA sequences) to produce an improved draft sequence. For eukaryotic genomes important to the DOE mission, we then add further information from directed experiments to produce reference genomic sequences that are publicly available for any scientific researcher. Also, we have continued our program for producing BAC-based finished sequence, both for adding information to JGI genome projects and for small BAC-based sequencing projects proposed through any of the JGI sequencing programs. We have now built our computational expertise in WGS assembly and analysis and have moved eukaryotic genome assembly from the JGI-PGF to JGI-HAGSC. We have concentrated our assembly development work on large plant genomes and complex fungal and algal genomes.

1. Genome maintenance and follow-up releases

We continue to produce regular improvements to our release genomes. We take feedback from collaborators and make any corrections to the genome sequence to resolve any issues uncovered in the subsequent annotation or in the genome analysis. We also follow-up issues identified by our collaborators and produce parallel data sets for comparison purposes to include in genome publications.

Status of our on-going genome projects

Eukaryotic Plants, Algae and Diatoms:

Sorghum
The second largest potential biofuel crop in the U.S. (after its close relative, maize) and a model for other C4 grasses, 700Mb. Draft published in Science in 2009
Currently working on improving 350Mb of the gene space.

Soybean
A prominent agricultural crop, source of biodiesel and model for nitrogen-fixing legumes, 980Mb
Draft published in Nature in 2010

Populus trichocarpa
black cottonwood tree, 350-550Mb, Wild type poly rate, highly repetitive
Draft Published in Science 2006
Improved assembly V2 2009.

Chlamydomonas reinhardtii
green algae model organism, 112Mb, Difficult sequence and frequent repeats
Draft published in Science 2007
Improved V4 December 2007, Working on V5 assembly.

Physcomitrella patens
A tractable model organism for studies of the cell wall, the principal component of terrestrial biomass and a key target for processing to sugars for bioenergy
Improved V3 released 2009. Working on V4 with BAC ends and genetic map incorporated

Trichoplax adhaerens
simplest known animal, 100Mb, Polymorphic rate 1-4%

Published Nature 2008. Improved assembly released 06-01-10.

Emiliania huxleyi
Algae
Improved release 05-15-10

Bigelowiella natans
Microbial eukaryote 91Mb
Version 1.1 03-15-10

Eukaryotic Fungi:

Laccaria bicolor
Mushroom 61Mb
Finished Version 3 release 06-18-10

Aspergillus niger
black mold, 35Mb, Small number of unclonable region
Finished V1 August 2006, Finished V2 June 2008 Publication in Preparation

Mycosphaerella graminicola
wheat rust, 41Mb, Common repeats
Finished V1 December 2006, Finished V2 May 2008 Publication in Preparation

Mycosphaerella fiijiensis
Black leaf streak disease of bananas
Improved release 12-01-09

Phycomyces blakesleeanus
filamentous fungi, 60Mb, Larg genomic repeats
Improved V2 2009, working on V3 with integrated BAC library and genetic map

Cryphonectria parasitica
chestnut blight, 44Mb
V2 Finished Release 2009 Publication in Preparation

Pleurotus ostreatus PC15
oyster mushroom, 35Mb
V2 Finished Release 01-19-10

Mucor circinelloides
fungal plant pathogen, 38Mb
V2 Finished release 10-16-09

Heterobasidion annosum
Fungal Pathogen of Conifers, 33Mb
V2 Finished Release 02-05-10

Cochliobolus heterostrophus
Natural Pathogen of Wheat causing Southern Corn Leaf Blight, 35Mb
V2 Finished Release expected Spring 2010

Thielavia terrestris
Important in the studies of thermophilic enzymes and biomass conversion, 36Mb
In manual finishing

Sporotrichum thermophile
Highly proficient decomposer of cellulose, 38Mb
In manual finishing

Serpula lacrymans
Dry rot fungus, major decomposer, 42.8Mb
In automated finishing

Agarius bisporus
Degrades leaf and needle litter in temperate forests, 30.3Mb
In automated finishing

Schizophyllum commune
White rot model basidiomycete fungus, 38.6Mb
In automated finishing

2. Clone Based Sequencing and Finishing Projects

We have been sequencing and finishing BAC-based projects since the inception of the sequencing group at the SHGC 10 years ago. At the height of the Human Genome Project, we were finishing more than 12 Mb of clone-based projects per month to finish the DOE chromosomes 5, 16 and 19. In recent years, we have continued with BAC and fosmid-based finishing in support of WGS projects and JGI clone-based projects. We have also increased the number of clones we shotgun sequence due to an increased need for WGS quality control, genome improvement and on-going JGI Community Sequencing Projects.

	Mb finished
September 2009	9,055,091
October 2009	8,842,737
November 2009	5,447,520
December 2009	5,602,978
January 2010	4,163,381
February 2010	5,047,143
March 2010	6,827,467
April 2010	4,712,484
May 2010	5,404,409
June 2010	7,233,317
July 2010	7,141,327
August 2010	10,090,839

3. Genomic Resource Development: Maps, BES, ESTs, FLcDNA

The most difficult issues to overcome with a whole genome shotgun sequence are the need to order and orientate the scaffold pieces and infer genome organization from the WGS data set. Most of our projects have no map information, no long-range linking information and no known information about genome organization (although for a few select well-studied genomes, the collaborators do know the number of chromosomes. To overcome these mapping issues, we have been adding new genomic resources when they are needed by a project. These resources include, end-sequencing of BAC libraries, map integration and generation, EST/gene sequence sampling and full-length cDNA sequencing.

BAC End sequencing projects

BAC end sequencing reactions were added in support of the following whole genome sequencing projects:

Switchgrass	202,752
Physcomitrella	201,216
Brassica	73,728

4. Whole Genome Assembly

We began working with whole genome assemblies at JGI-HAGSC when it became apparent we would need to build custom WGS assemblies for genome improvement and finishing. We settled on the Arachne assembler from the Broad

Institute as our assembler of choice and have been working with it now for about four years to assemble whole genomes. We recently expanded our WGS assembly capacity and have taken on assembly and post-assembly integration of mapping resources for the JGI eukaryotic projects with a special emphasis on plant genomes.

The following assemblies were completed (September 2009-August 2010) in support of on-going JGI WGS projects:

09-01-09	*Acremonium alcalophilum*	V1	Fungi	54 Mb
09-15-09	*Setaria italica*	V1	Plant	401 Mb
11-15-09	*Prunus persica*	V1	Plant	225 Mb
01-30-10	*Aquilegia coerulea* Goldsmith	V1	Plant	293 Mb
01-30-10	*Phytophthora capsici* LT1534	V1	Other	56 Mb
03-31-10	*Thellungiella halophila*	V1	Plant	238 Mb
06-30-10	*Eucalyptus grandis*	V1	Plant	643 Mb

Publications September 2009-August 2010

Prochnik SE, Umen J, Nedelcu AM, Hallmann A, Miller SM, Nishii I, Ferris P, Kuo A, Mitros T, Fritz-Laylin LK, Hellsten U, Chapman J, Simakov O, Rensing SA, Terry A, Pangilinan J, Kapitonov V, Jurka J, Salamov A, Shapiro H, Schmutz J, Grimwood J, Lindquist E, Lucas S, Grigoriev IV, Schmitt R, Kirk D, Rokhsar DS. Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri. Science. 2010 Jul 9;329(5988):223-6. PubMed PMID: 20616280; PubMed Central PMCID: PMC2993248.

Hellsten U, Harland RM, Gilchrist MJ, Hendrix D, Jurka J, Kapitonov V, Ovcharenko I, Putnam NH, Shu S, Taher L, Blitz IL, Blumberg B, Dichmann DS, Dubchak I, Amaya E, Detter JC, Fletcher R, Gerhard DS, Goodstein D, Graves T, Grigoriev IV, Grimwood J, Kawashima T, Lindquist E, Lucas SM, Mead PE, Mitros T, Ogino H, Ohta Y, Poliakov AV, Pollet N, Robert J, Salamov A, Sater AK, Schmutz J, Terry A, Vize PD, Warren WC, Wells D, Wills A, Wilson RK, Zimmerman LB, Zorn AM, Grainger R, Grammer T, Khokha MK, Richardson PM, Rokhsar DS. The genome of the Western clawed frog Xenopus tropicalis. Science. 2010 Apr 30;328(5978):633-6. PubMed PMID: 20431018; PubMed Central PMCID: PMC2994648.

Ferris P, Olson BJ, De Hoff PL, Douglass S, Casero D, Prochnik S, Geng S, Rai R, Grimwood J, Schmutz J, Nishii I, Hamaji T, Nozaki H, Pellegrini M, Umen JG. Evolution of an expanded sex-determining locus in Volvox. Science. 2010 Apr 16;328(5976):351-4. Erratum in: Science. 2010 Sep 17;329(5998):1467. PubMed PMID: 20395508; PubMed Central PMCID: PMC2880461.

International Brachypodium Initiative. Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature. 2010 Feb 11;463(7282):763-8. PubMed PMID: 20148030.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang XC, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA. Genome sequence of the palaeopolyploid soybean. Nature. 2010 Jan 14;463(7278):178-83. Erratum in: Nature. 2010 May 6;465(7294):120. PubMed PMID: 20075913.

Schmutz J, Cannon S, Mitros T, Nelson W, Shu S, Goodstein D, Rokhsar D. (2010). "The Draft

Soybean Genome Sequence". Genetics, Genomics, and Breeding of Soybean. 223-44.

Rosa SF, Powell AE, Rosengarten RD, Nicotra ML, Moreno MA, Grimwood J, Lakkis FG, Dellaporta SL, Buss LW. (2010). Hydractinia allodeterminant alr1 resides in an immunoglobulin superfamily-like gene complex. Curr Biol. 20(12):1122-7.

Chain PS, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C, Cole JR, Ding Y, Dugan S, Field D, Garrity GM, Gibbs R, Graves T, Han CS, Harrison SH, Highlander S, Hugenholtz P, Khouri HM, Kodira CD, Kolker E, Kyrpides NC, Lang D, Lapidus A, Malfatti SA, Markowitz V, Metha T, Nelson KE, Parkhill J, Pitluck S, Qin X, Read TD, Schmutz J, Sozhamannan S, Sterk P, Strausberg RL, Sutton G, Thomson NR, Tiedje JM, Weinstock G, Wollam A, Genomic Standards Consortium Human Microbiome Project Jumpstart Consortium, Detter JC. (2009). Genomics: Genome project standards in a new era of sequencing. Science. 326(5950):236-7.

Nicotra ML, Powell AE, Rosengarten RD, Moreno M, Grimwood J, Lakkis FG, Dellaporta SL, Buss LW. (2009). A hypervariable invertebrate allodeterminant. Curr Biol. 19(7):583-9.