CLEMSON
UNIVERSITY

Increasing the Utility of Existing Chestnut DNA and RNA Sequence Data through Bioinformatic Analysis

01/01/2014 – 12/31/2014

$8,580 Requested

Submitted to:
The American Chestnut Foundation

Submitted by:
Clemson University
Sponsored Programs
300 Brackett Hall
Box 345702
Clemson, SC 29634

PI Information:
Dr. Margaret Staton
Staton2@clemson.edu

Authorized Official:
R. Larry Dooley, Ph. D.
Interim Vice President for Research
Phone: (864) 656-2424
Fax: (864) 656-0881
CUOSP@clemson.edu

Signing for

Signature   5/6/15
Date

Title: “Increasing the utility of existing chestnut DNA and RNA sequence data through bioinformatic analysis”

PI:

Margaret Staton, PhD
Research Scientist
Staton2@clemson.edu
864-656-4643
304c BRC/105 Collings St.
Clemson University
Clemson, SC 29634

Co-PI:

John E. Carlson, PhD
Professor of Molecular Genetics
Director of The Schatz Center for Tree Molecular Genetics
jec16@psu.edu
814-863-9164
323 Forest Resources Building
University Park, PA 16802

Narrative:
Prior support from the Forest Health Initiative, the National Science Foundation (NSF) and the American Chestnut Foundation has led to extensive genomic resource development for Chinese and American chestnut species. The DNA and transcriptome sequences have proven valuable for many tasks, however, further bioinformatic analysis is needed to expand their utility and enable a publication on the Chinese chestnut whole genome. Support from the TACF could help accomplish three primary bioinformatic tasks in the next step to integrate genomics resources into TACF activities: identification of SNPs (single nucleotide polymorphisms) between American and Chinese chestnut, structural and functional annotation of the current version of the Chinese chestnut whole genome, and characterization of differences between American and Chinese gene sequences in the blight QTL regions.

I. Species-specific SNP identification
SNPs fixed between American and Chinese chestnut will have an immediate impact on the breeding programs of the TACF by providing a method to differentiate American and Chinese inherited DNA segments in backcross progeny. The RNA sequencing effort originally funded by the NSF led to very extensive set of unigenes (unique expressed gene sequences) for American and Chinese chestnut. Single nucleotide polymorphisms (SNPs) and simple sequence repeats (SSRs) were mined from the unigene sequences to provide polymorphic markers within each species which have been posted online for public use (http://www.fagaceae.org/markers). The identification of SNPs between species will require alterations to the SNP identification pipeline originally created for the NSF project. Tissue from three American genotypes (BA69, AC 4T7, AC 1T6) and two Chinese genotypes (Mahogany and Nanking) were sampled for RNA sequencing [1], increasing the likelihood that SNPs that are truly fixed in the species can be identified *in silico*. By modifying the software code to examine both genotypes and species when calling SNPs, a set of candidate species-specific SNPs can be identified that are more likely to be successful and thereby reduce downstream laboratory costs. To expedite testing, primers will be designed around the SSRs using Primer3 [2].

II. Structural and functional annotation of the draft Chinese chestnut genome
Dr. John Carlson has led two chestnut DNA sequencing efforts in the past 4 years; one to sequence the whole genome of the Chinese chestnut and one to sequence pools of BACs spanning the 3 blight QTL regions [3]. The whole genome sequencing effort has recently achieved a new level of contiguity with 41,270 scaffold sequences with an N50 of 39,580kb and a total length of 724Mb (~90% of estimated genome). This new assembly has yet to be structurally or functionally annotated. A pipeline for this task was developed for smaller QTL sequence assemblies. It relies primarily on the GMOD maker software [4], but also incorporates the repeat finder LTR_Finder [5] and the RNAseq mapping software packages GMAP [6], PASA [7] and GeneWise [8]. While relatively robust, this pipeline is currently too slow to effectively annotate the whole genome while running on a single computer. Part of the hours set forth in this proposal will be used to

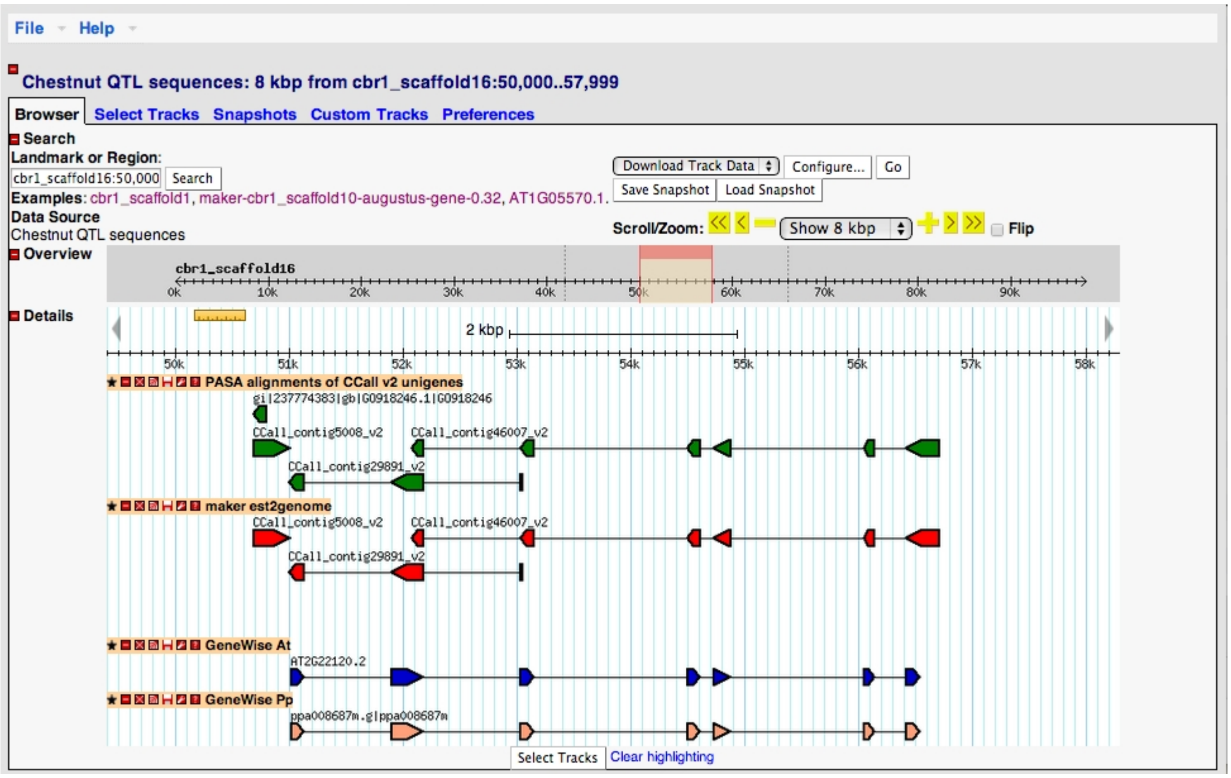
migrate the pipeline to Clemson's palmetto cluster (http://citi.clemson.edu/palmetto/), a grid of thousands of computers that can be used to quickly annotate current and future versions of the whole genome sequence.

The website hosting chestnut genomic resources, Fagaceae Genomics Web (FGW, www.fagaceae.org), is hosted and maintained at Clemson by the project PI. It will be cross-linked and integrated with the Hardwood Genomics Website (HGW, www.hardwoodgenomics.org), also hosted at Clemson by PI Staton. The results of the genome annotation will be uploaded to a Gbrowse instance [9] and made public through FGW and HGW when the genome publication is accepted. This will allow users to find genes, view evidence for those genes, explore repeats, and BLAST sequences of interest to the whole genome.

III. Manual annotation of genes in QTL regions

In order to gain a more comprehensive look at the QTL regions and to identify candidate genes prior to completion of the whole genome reference sequence, targeted sequencing of the blight resistance QTL regions was carried out in 2012. Utilizing the genetic and physical maps, 198 BAC clones inside the QTL regions were selected and sequenced by 454 and Illumina chemistries at Penn State under the direction of PI Carlson. The sequences were assembled and annotated by PI Staton. The sequences and structural annotation are available online for browsing (Figure 1, http://www.hardwoodgenomics.org/cgi-bin/hwg_gb2/gbrowse/hwg_gb2_chestnut_qtls/). As expected, the clone-based sequencing approach yielded larger, more contiguous sequence assemblies than the whole genome "shotgun" approach (Table 1).

Figure 1. Online view of a gene in the QTL region cbr1 using the software Gbrowse. The green and red gene tracks are alignments of the Chinese chestnut unigenes by the software PASA and maker, respectively. The blue and peach-colored alignments are Arabidopsis and peach genes aligned by the software GeneWise, respectively.



Chestnut QTL sequences: 8 kbp from cbr1_scaffold16:50,000..57,999

Landmark or Region: cbr1_scaffold16:50,000

Examples: cbr1_scaffold1, maker-cbr1_scaffold10-augustus-gene-0.32, AT1G05570.1.

Data Source: Chestnut QTL sequences

Scroll/Zoom: Show 8 kbp

Of particular interest in utilizing these results is how American chestnut differs from Chinese chestnut sequence in these regions. Seven lanes of American chestnut paired end DNA sequence was produced by the Carlson laboratory. These reads can be aligned to the Chinese sequence assembly with software such as bowtie2 [10] to discover mutations between the species, particularly possible loss of function mutations in genes related to biotic stress. While some of the sequence analysis can be automated, the final cataloguing and prioritization of these mutations will have to be addressed manually. Support from this proposal will fund the hours of work needed to accomplish this important task, providing gene targets for blight resistance.

Table 1. Assembly results from the sequencing of BACs spanning the three blight resistance QTL regions in the Chinese chestnut genome

QTL	cbr1	cbr2	cbr3
Genetic Map Location	LGB (40.9-50.4 cM)	LGF (38.1-46.8 cM)	LGG (35.7-39.5 cM)
# of BACs in Pool	99	51	40
Sequence scaffolds	214	128	53
Avg length	31,657	32,151	56,410
N50*	75,056	72,331	158,218
Total length	6,774,520	4,115,273	2,989,748

Conclusion

The extensive genomic resource development in chestnut, including RNA sequence data, gene sequences, a dense genetic map, a physical map, a draft genome and a public website for access to these data sets, has created a fertile environment for chestnut research. However, leveraging these resources for specific research goals often requires bioinformatic expertise to integrate these diverse datasets and produce useful results for further field/laboratory utilization. Here we propose to increase the value of the existing chestnut genomic data in three key areas: identification of fixed SNPs to differentiate between Chinese and American chestnut genomes, robust structural annotation of the Chinese chestnut draft genome, and manual annotation and prioritization of candidate disease resistance genes across the blight resistance QTLs. These tasks are all key next steps to furthering the American Chestnut Foundation goal of the restoration of healthy American chestnut trees to North American forests.

Timeline: To be completed within calendar year 2014

Budget: Salary support for bioinformatic analysis: $8,580. This is 9.4% (app. 5 weeks) of PI Staton's annual salary including fringe. (File storage systems and computation equipment are already available to accomplish the outlined tasks. This is not intended as a cost share, but rather to indicate the facilities and equipment are in place.)

Citations

1. Barakat A, Staton M, Cheng CH, Park J, Yassin NB, Ficklin S, Yeh CC, Hebard F, Baier K, Powell W *et al*: **Chestnut resistance to the blight disease: insights from transcriptome analysis**. *BMC Plant Biol* 2012, **12**:38.
2. Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers**. *Methods Mol Biol* 2000, **132**:365-386.
3. Kubisiak TL, Nelson CD, Staton ME, Zhebentyayeva T, Smith C, Olukolu BA, Fang GC, Hebard FV, Anagnostakis S, Wheeler N *et al*: **A transcriptome-based genetic map of Chinese chestnut (Castanea mollissima) and identification of regions of segmental homology with peach (Prunus persica)**. *Tree Genetics & Genomes* 2013, **9**(2):557-571.
4. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M: **MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes**. *Genome Res* 2008, **18**(1):188-196.
5. Xu Z, Wang H: **LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W265-268.
6. Wu TD, Watanabe CK: **GMAP: a genomic mapping and alignment program for mRNA and EST sequences**. *Bioinformatics* 2005, **21**(9):1859-1875.
7. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Jr., Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD *et al*: **Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies**. *Nucleic Acids Res* 2003, **31**(19):5654-5666.
8. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise**. *Genome Res* 2004, **14**(5):988-995.
9. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A *et al*: **The generic genome browser: a building block for a model organism system database**. *Genome Res* 2002, **12**(10):1599-1610.
10. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**(4):357-359.

Staton Budget Justification:

Senior Personnel Salary: $6,500 is requested for 1 the PI's salary for the year of the project. Fringe benefits of $2,080 have been calculated at 32.0% rate.