Final Report for “Increasing the utility of existing chestnut DNA and RNA sequence data through bioinformatic analysis”

Objective 1: Species specific SNP identification

We utilized 7 datasets for SNP identification (Table 1). The genome reference sequence was derived from the ‘Vanuxem’ genotype, and John Carlson (Penn State University) provided variant calls against this reference for the Chinese genotype ‘Nanking’, the American genotype ‘Alex R’ from Penn State, and the American genotype ‘Ellis 1’ from SUNY. We also had sequence data from the American genotype ‘GMBig’. Finally, RNASeq reads from a number of cultivars were available.

For the sequence files, the reads were trimmed and mapped against the reference genome, then SNPs were called for each genotype using samtools (further details in Appendices 1 & 2). A custom script was written to process all of the SNP data. Three steps were taken:

1. Compile a list of all SNPs from *C. dentata* samples that were called as homozygous against the reference
2. Remove any SNPs that have been found in any *C. mollissima* sample.
3. Compare SNPs to all individual *C. dentata* reads aligned to that location of the genome to further filter any possible heterozygotes.

The majority of called SNPs are from the ‘GMBig’ resequencing due to the very deep coverage, 46X, of that library. However, the addition of other American chestnut and Chinese chestnut samples allowed us to filter these SNPs significantly, yielding markers that are more likely to be consistently homozygous for opposite alleles between the two species. We further ranked the SNPs by the number of American samples with support for that SNP (Table 2). The full set of 11,194,851 SNPs is provided in the tab-delimited file predicted_diagnostic_SNPs.txt. A smaller subset of 714,039 SNPs supported by sequencing from at least three American genotypes is provided in the Excel spreadsheet predicted_diagnostic_SNPs_HQ.xlsx. These may be provided in additional formats for use in the TACF breeding program as required. The SNPs will also be made available online as part of the chestnut genome browser v1.1 (more information below).

Objective 2: Structural and functional annotation of the draft Chinese chestnut genome

The Chinese chestnut draft reference genome has been submitted to NCBI (accession GCA_000763605.1). The contamination pipeline at NCBI found some areas of suspected non-chestnut DNA, which were removed. The contamination was very low; about 6 in 10,000 bases of the v1.0 genome sequence, and only 10 scaffolds were completely removed. The resulting genome version (v1.1) has the following metrics:

Number of scaffolds	41,260
Total length of scaffolds	724,001,627
Average length of scaffolds	17,547.3
Largest scaffold	429,344
Smallest scaffold	473
N50	39,561
N90	6,866

In lieu of reannotating at this stage, we are instead transferring the original annotations from v1.0 to v1.1. Nathaniel Cannon from John Carlson's laboratory is heading this effort. As soon as the "lifted over" annotations are provided, we will update the website to enable download, browsing, and searching.

This is a major deviation from the proposed work objective. The move of the Staton laboratory from Clemson University to the University of Tennessee in January of 2014 made installation of the genome annotation software Maker on the Clemson high performance computer cluster impossible. However, we now have the maker software installed at the University of Tennessee on our own computational server, and we will be able to reannotate the genome as soon as John Carlson's group releases the next version. We commit to do this when needed without further financial support from TACF.

Objective 3: Manual annotation of genes in QTL regions

The chestnut genome and QTL sequences were originally provided on the hardwood genomics web site (http://www.hardwoodgenomics.org/chinese-chestnut-genome). The assembled sequences are available for download as fasta files or for browsing via Gbrowse. However, Gbrowse does not allow for web users to compare their own sequences to the chestnut sequences or to provide manual annotations for genes. To provide this support, we installed the software JBrowse (http://jbrowse.org/) and the plug-in webApollo (http://apollo.berkeleybop.org/) for the QTL regions. The web link for this functionality:

http://www.hardwoodgenomics.org/a/jbrowse

The JBrowse software allows a user to search for specific QLT scaffolds and view tracks of aligned data, including

- Transcriptome sequence alignments
    - Alignments of Chinese chestnut unigenes by PASA
    - Alignments of Chinese chestnut unigenes by GMAP
    - Alignments of Chinese chestnut unigenes by blastn
    - Alignments of Chinese chestnut unigenes by est2genome
- Ab initio gene predictions
    - Augustus software
    - Glimmer software
- Homologous gene alignments
    - Alignments of *Arabidopsis thaliana* genes by Genewise
    - Alignments of *Medicago truncatula* genes by Genewise
    - Alignments of peach (*Prunus persica*) genes by Genewise
    - Alignments of poplar (*Populus trichocarpa*) genes by Genewise
    - Alignments of grape (*Vitis vinifera*) genes by Genewise
    - Alignment of *Arabidopsis thaliana* and peach (*Prunus persica*) genes by blastx
    - Alignment of *Arabidopsis thaliana* and peach (*Prunus persica*) genes by protein2genome
- Final maker gene predictions
- Repetitive regions identified by RepeatMasker
- Alignments of next generation sequence reads (requires high level of zoom)
    - Alignment of American chestnut "GMBig" whole genome resequence reads
    - Alignment of American chestnut transcriptome reads, multiple genotypes

- Alignment of Chinese chestnut transcriptome reads, multiple genotypes

A web user may utilize all of these tracks with the webApollo plugin to annotate the genes in this region (Figure 1). The user must first request a username and password from our lab, and after completing annotations, they will be available for other users to view. Eventually, the manual annotations may be curated and merged into a new version of the genome. Equally important, the curator can view and provide comments about the differences between American and Chinese chestnut gene sequences based on the alignments of the American chestnut reads (Figure 2). This is particularly important in the QTL regions as the search for the source of blight resistance continues.

The Staton lab has begun manual annotation of the 44 genes flagged as particularly interesting in the QTL regions based on function (Excel file). We have a conference call scheduled with the webApollo trainer, Monica Munoz-Torres, to learn the best practices of proper gene annotation. The training and subsequent manual annotation of the 44 functionally interesting genes will be completed within 2 months, and a supplemental report will be provided to the TACF with the findings. Additionally, a training manual with instructions for utilizing Jbrowse and webApollo will be created and placed online for users to reference.

Budget

The budget has been fully utilized to provide partial salary support for three individuals:

Mark Cook (undergraduate) – Worked on the webApollo and JBrowse installation

Nathan Henry (research associate) – Led the webApollo and JBrowse installation

Jack Davitt (research associate) – Led the sequence alignments and SNP identification project

All invoices will be submitted by the end of January by Joan P. Webb, UTIA Sponsored Projects Accountant (jwebb44@utk.edu).

Table 1.

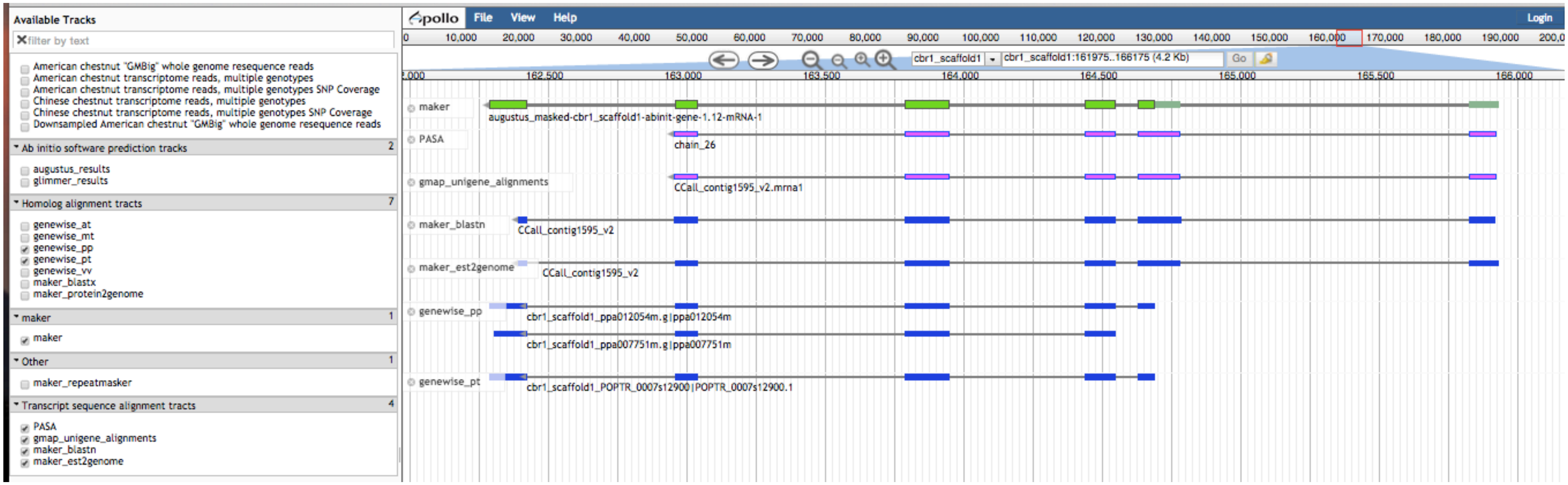
Species	**Genotype**	**Data (File Type)**	**Details**
Castanea dentata	GMBig	Whole genome resequencing (Fastq sequence reads)	2x76 Illumina reads, 490 million reads, 36.6Gbases, 46X coverage. Variant call strategy detailed in Appendix 1. 20,392,649 variants against Vanuxem.
Castanea dentata	Alex R Parent Tree from Penn State	Whole genome resequencing (VCF - variants against Vanuxem)	2,830,062 variants against Vanuxem reference.
Castanea dentata	Ellis 1 from SUNY	Whole genome resequencing (VCF - variants against Vanuxem)	2,503,143 variants against Vanuxem
Castanea dentata	Tree BA69 (RNA from Bill from	454 RNASeq	ACCanker library from canker tissue, 129,508 sequences Variant call strategy detailed in Appendix 2.

	tissue from Fred)		
Castanea dentata	Wisneiwski genotype from CAES	454 RNASeq	ACHS1n library from bark tissue 222,939 sequences ACWP1 library from pooled tissue sample 47,653 sequences Variant call strategy detailed in Appendix 2.
Castanea dentata	Watertown genotype from CAES	454 RNASeq	ACHS2n library from bark tissue 254,810 sequences ACWP2 library from pooled tissue sample 33,288 sequences Variant call strategy detailed in Appendix 2.
Castanea mollissima	Nanking	Whole genome resequencing (VCF - variants against Vanuxem)	1,138,222 variants against Vanuxem
Castanea mollissima	Tree VA37 (Nanking)	454 RNASeq	CCCanker library from canker tissue 235,635 sequences Variant call strategy detailed in Appendix 2.
Castanea mollissima	GR119 (possibly Nanking)	454 RNASeq	CCNHS library from healthy stem tissue 259,859 sequences Variant call strategy detailed in Appendix 2.
Castanea mollissima	Mahogany (probably BX316)	454 RNASeq	CCMHS library from healthy stem tissue 228,594 sequences Variant call strategy detailed in Appendix 2.
Castanea mollissima	Nanking GR119	454 RNASeq	CCWP1 library from pooled tissue samples 60,445 sequences Variant call strategy detailed in Appendix 2.
Castanea mollissima	Mahogany BX316	454 RNASeq	CCWP2 library from pooled tissue samples 53939 sequences Variant call strategy detailed in Appendix 2.

Table 2.

SNPs supported by read evidence in 1 American Chestnut sample	11,194,851
SNPs supported by read evidence in 2 American Chestnut samples	405,348
SNPs supported by read evidence in 3 American Chestnut samples	714,039
SNPs supported by read evidence in 4 American Chestnut samples	12,610
SNPs supported by read evidence in 5 American Chestnut samples	4,212
SNPs supported by read evidence in 6 American Chestnut samples	516

Figure 1. A view of a gene on scaffold 1. The gene (green) has both transcriptome evidence (purple) and alignments to genes from peach and poplar (blue).



Available Tracks

filter by text

□ American chestnut "GMBig" whole genome resequence reads
□ American chestnut transcriptome reads, multiple genotypes
□ American chestnut transcriptome reads, multiple genotypes SNP Coverage
□ Chinese chestnut transcriptome reads, multiple genotypes
□ Chinese chestnut transcriptome reads, multiple genotypes SNP Coverage
□ Downsampled American chestnut "GMBig" whole genome resequence reads

Ab initio software prediction tracks 2

□ augustus_results
□ glimmer_results

Homolog alignment tracts 7

□ genewise_at
□ genewise_mt
☑ genewise_pp
☑ genewise_pt
□ genewise_vv
□ maker_blastx
□ maker_protein2genome

maker 1

☑ maker

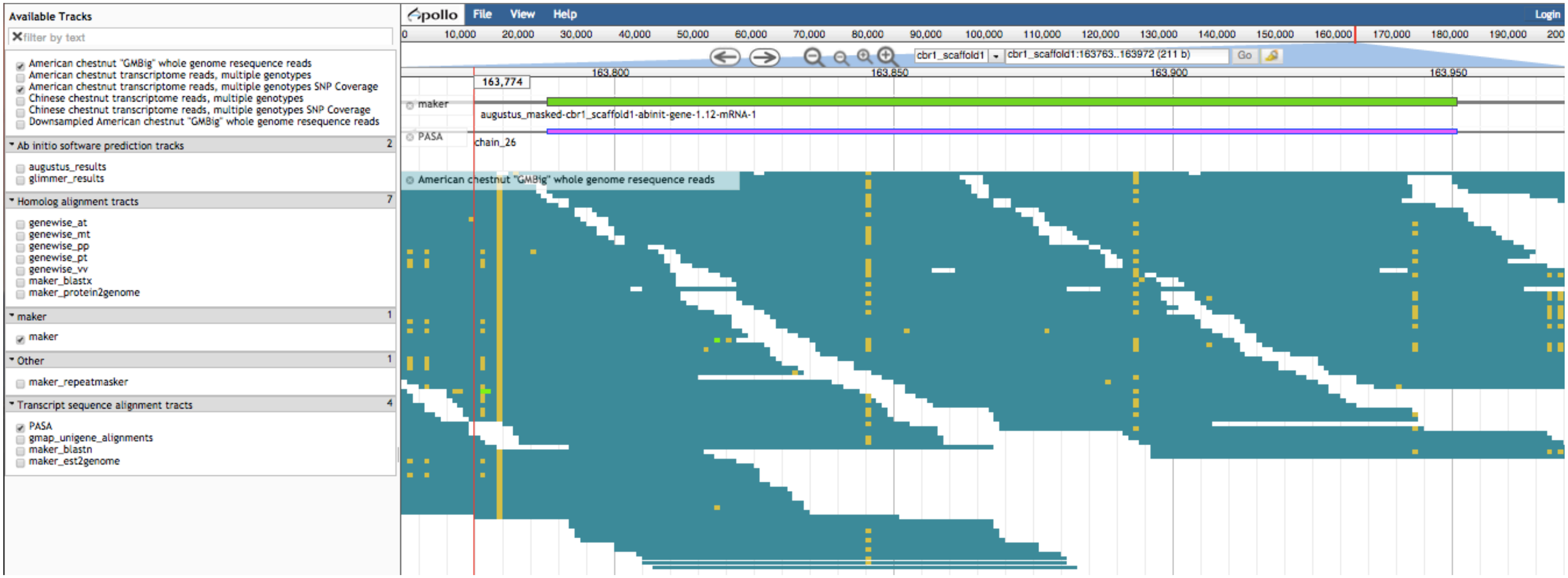
Other 1

□ maker_repeatmasker

Transcript sequence alignment tracts 4

☑ PASA
☑ gmap_unigene_alignments
☑ maker_blastn
☑ maker_est2genome

Figure 2. This is a closer view of a single exon of the gene on scaffold 1. Alignments of American chestnut reads are show in blue. Bases that differ between the Chinese chestnut reference and the American sequence are show in yellow. The yellow bases clearly highlight SNPs between the species within this gene. Homozygous SNPs are a single solid yellow line while heterozygous SNPs are a mixture of yellow and blue bases.



Available Tracks

filter by text

- ☑ American chestnut "GMBig" whole genome resequence reads
- □ American chestnut transcriptome reads, multiple genotypes
- ☑ American chestnut transcriptome reads, multiple genotypes SNP Coverage
- □ Chinese chestnut transcriptome reads, multiple genotypes
- □ Chinese chestnut transcriptome reads, multiple genotypes SNP Coverage
- □ Downsampled American chestnut "GMBig" whole genome resequence reads

Ab initio software prediction tracks 2

- □ augustus_results
- □ glimmer_results

Homolog alignment tracts 7

- □ genewise_at
- □ genewise_mt
- □ genewise_pp
- □ genewise_pt
- □ genewise_vv
- □ maker_blastx
- □ maker_protein2genome

maker 1

- ☑ maker

Other 1

- □ maker_repeatmasker

Transcript sequence alignment tracts 4

- ☑ PASA
- □ gmap_unigene_alignments
- □ maker_blastn
- □ maker_est2genome

Appendix 1. Methods for mapping American chestnut reads to Chinese chestnut reference genome

The trimmed sample was mapped to existing CC reference genome with bwa (Burrows-Wheelers Aligners, 0.7.8) with the “mem” algorithm. The bwa mem aligned file was converted from sam to bam format with samtools (0.1.19) “view”. Flags -b (output in bam format) and -S (sam file input) were used. Bam files were concatenated together and the resulting file was sorted with “sort” option in samtools. Samtools function of “mpileup” was used to generate BCF from bam files. Flags of -I (no indel calling) and -ugf (-u output as uncompressed BCF, -g compute genotype likelihoods and output to BCF format, -f faidx indexed reference files in FASTA format, in this case, CC genome fasta file was previous indexed) were used. The subsequent BCF file was directly piped (|) as the input file for bcftools view -bvcg (-b output in BCF,-v output variant site only,-c enforce -v option,-g call per-sample genotypes at variant sites). The resulting VCF file was used for classifying SNPs within the comparison.

Appendix 2. Methods for mapping 454 RNASeq reads to Chinese chestnut reference genome

The software GSMapper (Roche 454) was used to map the reads to the existing Chinese Chestnut genome in transcriptome mode. The resulting 454Contigs.bam was sorted with samtools sort, followed by BCF file generation as described in Appendix 1.