The Forest Health Initiative
Genome Resources and Tools Project
Update
February 1, 2012

The goal for the genome resources and tools project is to provide a high-quality genomic reference sequence for *Castanea mollissima* (Chinese Chestnut) for the identification of genes for resistance to *Cryphonectria parasitica*. In addition this project will demonstrate the power of genomics to address forest health and ecosystem restoration issues in the future.

Scientific Approach

The genome resources and tools project brings together research expertise and cutting-edge facilities in genomics and bioinformatics from Penn State University and the Clemson University Genomics Institute. Our approach to develop a high quality reference genome sequence for Chinese chestnut (*Castanea mollissima*) cv Vanuxem is through deep "next generation" DNA sequencing technologies. The Vanuxem cultivar was chosen for the reference genome due to its key role in The American Chestnut Foundation's breeding program (http://www.acf.org/) and the NSF Fagaceae Tools project. The reference genome will be assembled de novo from 454 sequence data, then the gene sequences corrected and the assembly extended with Illumina sequence data. Pseudo-chromosomes will be built by integration of the genome sequence with the physical and genetic maps for chestnut.

Genome sequencing started in January, 2010, at Penn State. By the end of 2010, we had produced a deep DNA sequence data resource for the Vanuxem genome, totaling 55.4 Gbases from the 454 and Illumina sequencing platforms. This amounted to app. 63-fold depth of sequence. In addition, to guide the correct assembly of the sequence data, we sequenced the ends of 43,143 BAC clones that covered the chestnut genome 1.5 X fold along the "tiling path" of the genome provided by the Chinese chestnut genome physical mapping project.

Progress from July 2011

During the course of sequencing, we produced 12 assemblies of the genome sequence data. Our latest, unguided assembly ('build 12') completed in 2011 placed 925Mb into 1,147,939 contigs, with average size of only 1369bp. Using the BAC end data, the contigs could be merged into 51,766 scaffolds which included 587,208,063 bp from the contig set and an average length of. Three measures of the quality of a genome assembly are how well RNA sequences (gene transcripts) from the species align to the genome, how many genes can be found and the length of those genes. With the build 12 assembly, we found that 96% (46,885) of the Chinese chestnut transcript sequences could be align to the genome. Of the full length gene transcripts from the NSF Fagaceae project, 100% completely mapped to the genome at ≥ 97% sequence identity. Gene-finding algorithms predicted over 66,000 gene models in the assembled scaffolds, with a mean gene length of 2,761 bp. The largest gene known in Arabidopsis, an ATPase over 43,203 bp in length, was found in intact in one scaffold in build 12. The set of predicted genes contains over 550 genes related to disease resistance, including the three major resistance gene categories as well as genes for the transcripts expressed more highly in Chinese chestnut cankers than in American chestnut after inoculations, summarized in table 1.

Table 1. Numbers of genes in major resistance genes categories among predicted genes and stress-response transcripts mapped in the build 12 genome assembly.

Classes of disease resistance genes found	Number
Pathogenesis-response proteins	13
Disease resistance family proteins (TIR-NBS-LRR)	40
Leucine-rich repeat family proteins, total	80
Genes differentially expressed in C. mollissima cankers	433

These statistics indicate that the current assembly is probably a good representation of the genes in Chinese chestnut. However the current assembly is too highly fragmented to serve as a reference for genomes in related species. We spent much of 2011 trying new bioinformatics software and assembly strategies without improvement. Using the genetic and physical maps as guide, we could partially order scaffolds by gene content, as in the following figure -

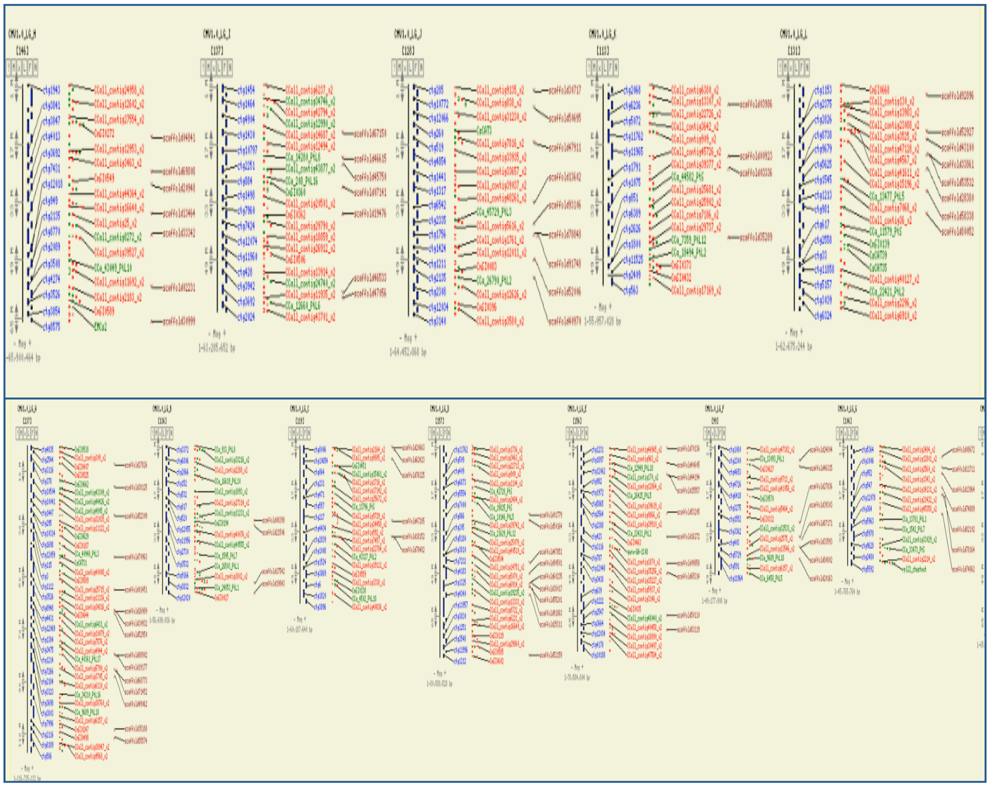

Fig 1. Partial Alignment of Genome Scaffolds to the Genetic and Physical Maps for Chinese Chestnut. Blue font, physical map contig names; red and green font genetic map marker names; purple font, genome scaffold names; numbers to the left of the vertical lines are genetic map distances in centiMorgans.

It was obvious from this exercise that the chestnut genome is highly complex, and that orienting all 51,766 current scaffolds would not be possible given the much smaller number of markers (1156) on the genetic map. Thus to attempt to achieve a better reference genome assembly, we are currently sequencing a new preparation of Vanuxem genomic DNA. We are optimistic that improvements to 454 sequencing technology, and a better DNA preparation will improve results.

While the new round of general genome sequencing is underway, we have also undertaken a highly focused approach to specifically obtain the Chinese chestnut blight resistance genes by separately sequencing each blight resistance QTL. This is possible because the 3 major blight resistance QTL are completely spanned in the recently completed physical map for Chinese chestnut by 4 large contigs of BAC clones. The contig covering the resistance QTL on linkage group LGG contains 40 BACs. The contig covering the resistance QTL on linkage group LGF contains 51 BACs. And the resistance QTL on linkage group LGB is composed of two contigs containing a total of 97 BAC clones. We sequenced each of the QTL as pools of BAC clones, which assembled into 2.94 Mb with 410 genes for QTL LGG, 3.99Mb with 548 genes for QTL LGF, and 6.60 Mb with 994 genes for QTL LGB.

The 3 QTL appear to be major disease resistance regions in the chestnut genome. Examples of genes in the chestnut QTL which appear to be good candidates for blight resistance include:

Disease resistance family proteins in CC-NBS-LRR class / family
Disease resistance proteins in TIR-NBS-LRR class / family
Disease resistance proteins in the LRR family, for response to fungi
Leucine-rich repeat kinase (LRR-RLK), receptor for bacterial associated EF-Tu
NB-ARC domain-containing disease resistance proteins
Homologs of barley mildew resistance locus O (MLO) protein
Callose synthase, required for callose formation in response to fungal pathogens
Plasma-membrane receptor-like kinase, involved in powdery mildew infection
Sugar isomerase (SIS) family protein involved in defense response to fungi
Callose synthase required for wound formation in response to fungal pathogens
Isoflavone reductase, involved in response to oxidative stress
TAO1 protein for resistance to Pseudomonas syringae AvrB
Pathogenesis-related thaumatin superfamily protein
Wound-responsive family protein

The blight-resistance candidate genes discovered are being provided to the FHI transformation project for functional studies. A publically accessible web portal has been created at the Fagaceae Genomics website. The public web portal (http://www.fagaceae.org/FHI) will provide searchable databases for access to the genome sequences, along with an interactive browser for viewing the assembled genome and assembled resistance QTL sequences, as soon as the assemblies are finalized.