Project title.
AIM: ancestry informative, transferrable, and affordable DNA markers for chestnut

Summary
We request the second year of funding for development of ancestry informative, transferrable, and affordable DNA markers for chestnut. The markers will be designed for multiple end uses and designed to distinguish seven species of chestnut and every individual tree; detect interspecific ancestry, the species of interspecific ancestry and the degree of same; ascertain recent pedigrees, identify close relative (parents, sibs, grandparents, half-sibs) and locate potential QTL regions.

Principal investigator and institutional affiliation.
Dr. Jeanne Romero-Severson
The University of Notre Dame

Duration of project
12 months

Total amount requested.
$10,000

Short and long-term goals of the project.
The short term goal of this proposal is the development of multi-purpose AIM markers for chestnut. The long term goal is providing a useful and affordable set of AIM markers that will expedite the restoration of the American chestnut.

Narrative.
Pages 1-6. Narrative references p.7. PROJECT UPDATE AND REPORT ON RESULTS TO DATE FOLLOWS TIMELINE.

Timeline
Page 6.

How results will be measured and reported.
The measurement of success for the chloroplast markers will be aligned sequences for at least 90% of the screening samples for the set e of chloroplast sequences which, if examined together, definitively distinguish the chloroplast of American, Chinese, Japanese and European chestnut species, including the American and Chinese chinquapins. These data will be reported as Jalview alignment images, with species specific haplotypes reported in tabular format and visualized in minimum spanning haplotype trees. The measurement of success for the EST-SSR and EST-SNP markers will be aligned sequences for at least 90% of the screening samples for 90% of the markers, which when used as a set, 1) distinguish every individual tested, 2) detect interspecific ancestry with at least 90% confidence and 3) permit finely detailed and affordable admixture analysis for any individual in any chestnut program in which the TACF is involved. Measurement criterion three requires genotyping the all of the baseline chestnut collection and establishment of an affordable method which uses all of the qualified markers.

Breakdown of how and when funds will be spent.
Page 6

Brief curriculum vitae for each principal investigator
Pages 8-9

NARRATIVE

Successful breeding programs have 1) defined proximate and ultimate goals 2) the right parents 3) a quantifiable and reproducible method of evaluating the phenotype(s) at the appropriate time 4)a digital and transparent record keeping system and 5) constancy of purpose. This proposal addresses an aspect of items two and four. The proximate goal of this proposal is the development of multi-purpose ancestry-informative markers (AIM) for chestnut. The ultimate goal is providing a scalable and affordable set of genomic tools that will expedite the restoration of the American chestnut.

The right parents.

Identification. Tree breeding is a long term process. Once the identity of parents and selected progeny is lost, the pedigree and the relationship between genotype and phenotype is also lost. In the long-past days before genetic markers, only meticulous record keeping could prevent this and even then, the published examples of mistaken identity in traditional cultivars of chestnut shows that without genetic markers, these errors are inevitable (McCleary et al. 2013). Given the investment of time and money already made in the American chestnut breeding program, a set of quality-controlled, informative, and platform-independent DNA markers for verifying identity would be a sound investment.

Breeding efficiency. Experience has shown that breeding for resistance to chestnut blight is not a simple or inexpensive progress. The genetic basis of the phenotype is not as simple as once hoped, the influence of local growing conditions and other locations across the range is not well-known and interactions between genetic variance and environmental variance (GxE) are difficult to assess. All this makes identification of specific allelic variants at specific loci difficult. The work of investigators in systems having much higher power and richer genomic resources than any tree breeding program reveal a sobering reality. Even when pedigrees exist across multiple generations, with tested population sizes >30,000 individuals and SNP numbers >30,000, with highly quantifiable phenotypes (e.g. back fat measured by ultrasound), highly partitioned variance components accounting for some of the GxE effects and mixed model analysis using genomic feature best linear unbiased prediction (GFBLUP), detection of causal variants is problematic in traits of low to moderate heritability (Sarup et al. 2016). **Nevertheless, increased efficiency in gain from selection is possible once chromosomal segments are identified that are robustly associated with the phenotype.** A set of quality-controlled, informative, and platform-independent DNA markers for identifying these segments would be a sound investment.

Genetic diversity. Successful restoration of American chestnut requires trees with resistance to chestnut blight and *Phytophthora*, but these two characteristics are not enough. The enduring resilience of a long-lived, sessile species lies in the standing genetic variation within and among populations. This variation provides a huge array of allelic combinations, some of which will permit at least some individuals to survive most biotic and abiotic stresses. As the nature of these stresses varies across time and place, genomics and mathematical tools cannot untangle the complexities of every interaction in time to rescue what remains of the native chestnut gene pool. What we can do is craft a set of affordable, robust genetic tools that will enable us to generate restoration populations with enough standing genetic variation to lower the risk of regeneration failure across many generations. A set of quality-controlled, informative, and platform-independent DNA markers for identifying genetic diversity would be a sound investment.

Marker qualification.

For the purpose of this proposal, **markers qualified for use in the TACF program or the programs of cooperators must be robust, informative, transferrable, scalable, affordable, and platform-independent.** Robust markers produce a genotype at least 90% of the time on 90% of the individuals tested, under standard laboratory conditions, provided that the DNA sample is of suitable quality. This is a minimum, as many robust markers perform better than this. An informative marker for the purpose of this proposal is one with a PIC value between 0.4 and 0.9 when screened on source populations appropriate for the intended end use. PIC is a measure of allele number weighted by allele frequency. A transferrable marker is one that is both robust and informative across all the species the intended end use is likely to encounter. A scalable set of markers can be used in any combination on any number of trees. An affordable marker set is one that can be deployed (for sample sizes 100-500), for less than the cost of single lane of Illumina sequencing *with the bioinformatics cost included* ($15,000 to $30,000 in materials and staff time, more if done with graduate students). A platform independent marker is a sequenced marker originating from a larger piece of sequenced DNA (e.g. an EST-SSR or a chloroplast intergenic region). With the marker sequence for every individual in hand, an investigator has a choice of genotyping options for the next step of the breeding process, from sequence-capture with bait beads to single marker PCR. Markers that are located on chestnut genetic map, placed on a genome scaffold, or located within a chestnut BAC have the potential for added value.

PLAN OF WORK

The goal in the first 12 months is the development and testing of a set of robust, informative, transferrable, and scalable AIM markers, including chloroplast markers that will permit highly reproducible, scalable, and cost-effective genotyping of breeding. The goal of the second 12 months is assessment of the entire baseline collection and development of an affordable and scalable sequencing protocol for that set of the qualified chloroplast and nuclear markers that has the most resolution. We estimate that this set will contain 4-6 chloroplast sequences and 24-48 nuclear sequences.

Specific aims

1. Detect and test species-specific regions of the chloroplast genome.
2. Develop a set of ancestry-informative markers (AIMs) from mapped EST-SSR sequences.
3. Sequence the baseline collection with both sets and report the results
4. Finish development of the final set of markers designed for maximum informativeness, scalability, and affordability.

The baseline collection. The Romero-Severson program initiated the collection of chestnut germplasm in 2011. Contributors include the Missouri Center for Agroforestry, Michigan State, CAES, the US Forest Service, the Pennsylvania Department of Natural Resources, the Indiana chapter of the American Chestnut Foundation, the Northern Nut Growers Association, private growers, interested private individuals, arboreta, and botanical gardens. The current collection includes 324 putatively different genotypes and seven species, 189 of which are identified as American chestnut, 67 Chinese chestnuts, 12 putative Japanese chestnuts, 12 American chinquapins, six Chinese chinquapins and six European chestnuts with the remainder being putative or strongly suspected hybrids of unknown ancestry (American chestnut x ? or Chinese chestnut x ? or complex multispecies ancestries).

Methods: Screening Panel. We will select 96 chestnuts from our collection for the screening panel: 30 American chestnuts from different locations including one set of three technical replications (i.e. one sample will be represented four times) for a total of 33 samples, 30 Chinese chestnuts including one set of three technical replications (i.e. one sample will be represented four times) for a total of 33 samples, 12 Japanese chestnuts, 12 American chinquapins, six Chinese chinquapins and six European chestnuts. We would welcome a greater diversity of American chestnuts, as the southern part of the previous native range is under-represented. However, we can do the phase one testing with the collection we have. DNA will be extracted with Qiagen kits and quantified with a nanodrop device.

Methods: Marker selection and genotyping by sequence capture. The approach we propose here, sequence-capture (aka bait-capture) using selected chloroplast regions and microsatellite-containing EST sequences, is scalable, transferrable and platform independent, in that each tree has a set of sequences that determine individual identity and ancestry in the nuclear genome and species identity in the chloroplast genome. Any or all of these sequences may be generated from any chestnut tree using the sequencing technology of choice in future projects.

We propose screening a total of 106 sequences using a capture by hybridization approach (Holliday et al. 2016)with six baits designed for each of 10 chloroplast regions ~400-bp in size and six baits designed for each of 96 EST-SSR sequences ~ 300-400 bp in size. EST-SSR will be chosen from the 121 mapped *C. mollissima* EST-SSR sequences also located to a BAC clone (Kubisiak et al. 2013). We will exclude highly similar sequences and EST-SSR in which the repeat is less than five units or more than 12 units. Although this is certainly not a candidate gene study, we may include sequences from putative candidate genes for blight or phytophthora resistance if other investigators report SNP or EST-SSR polymorphisms in such genes. If we find after exclusion that we have too few EST-SSR candidates for bait capture, we will move to the EST-SNP sequences. In summary, our bait-capture screening project will include 96 Castanea individuals representing seven species (counting the American chinquapins as one) and 444 baits.

The captured pieces of DNA will be sequenced using an Illumina platform, assembled by marker and individual, trimmed, and then assessed for the quality of the result. The criteria are:

1. Robustness i.e. does the same DNA yield the same sequence across all of the markers and does the sequencing work on 90% of the samples at least 90% of time?
2. Informativeness i.e. are the samples polymorphic within and among species?
3. Transferability. Is the level of missing data randomly distributed among species or is there evidence of lack of transferability i.e. some markers work consistently less well in some species?
4. Utility. Are there enough chloroplast sequence to identify candidate chloroplast regions that could be used to reliably determine species identify for chloroplast and is enough nuclear sequence recovered to conclude that the bait-capture technique is appropriate

for screening candidate EST-SSR and EST-SNP sequences for development of the AIM set?

We have generated preliminary Sanger sequence data on a small *Castanea* screening panel for five chloroplast regions known to be informative in other Fagales (hazelnuts, walnuts, hickories, pecans, oaks) (Borkowski et al. 2014). This small panel of 20 trees revealed preliminary indications of high species specificity and detected an individual labeled as Chinese chestnut in which all five chloroplast sequences were identical to the chloroplast sequences of the six Japanese chestnuts, which were identical to each other. Organelle capture is common among sympatric species or species with a recent common ancestry and thus does not prove recent interspecific ancestry. However, it is best to know what the chlorotype of elite breeding stock is, especially in the crosses with American and Chinese chestnut, where chloroplast capture by recent common ancestry is highly unlikely but recent interspecific hybrid ancestry is possible and even certain in chestnut blight resistance breeding programs.

Despite the allopatric speciation of Chinese, and American chestnut, we do not anticipate finding many or any species-specific markers but this is of course possible. In our previous study we found 11 polymorphic chestnut EST-SSR markers that met our criteria in the first set of 20 EST-SSR markers we tried (McCleary et al. 2013), suggesting that tranferability and polymorphism, even in this multi-species situation, will be sufficient to identify species admixture but species specificity is neither expected or required. The use of single markers, whether chloroplast or nuclear, to declare species identity is always unwise, even if the screening panel suggests such specificity may exist.

We plan to have a promising set of candidate chloroplast and nuclear AIM markers before the end of first 12 month period. We will proceed by testing our candidates on our larger set of Castanea with sequence capture using PCR-generated probes (SCPP) (Peñalba et al. 2014)or molecular inversion probes (MIP) (Niedzicka et al. 2016), both lower cost methods of using a specific set of qualified markers to genotype a moderate number (hundreds) of samples. All results of the project will be reported at the annual TACF meetings and in quarterly reports. If the TACF permits, we will publish the results using those chestnut individuals that are in the public domain.

A digital and transparent record keeping system

All of the sequence data generated for this project will be kept in trimmed FASTA format and identified by a lab index number in a Microsoft Access database in which also exists the original identifiers (as sent by collaborators), the lab index number, the putative species identity, a georeference if available, a cultivar name if available and any other identifiers associated with the sample. Access is not capable of holding huge amounts of data but the data formats and datatypes we generate will be consistent and (in theory) portable to a larger, more complex information storage and retrieval system when available. Sequences for samples in the public domain will be deposited in the appropriate public databases before publication, if publication is permitted.

Project timeline, reporting, budget, and budget justification (next page)

TIMELINE

ACTIVITY	First 12 months				Second 12 months			
Extract and quantify DNA	🟦							
Design baits, perform sequence capture	🟦	🟦						
Chloroplast sequencing	🟦	🟦	🟦					
Sequence captured baits		🟦	🟦	🟦				
Assemble, trim and align EST-SSR and EST-SNP sequences			🟦	🟦				
Analyze all results, chose final AIM set				🟦				
Use AIM set on entire baseline collection				🟦	🟦	🟦	🟦	
Make adjustments to AIM set if necessary						🟦	🟦	
Develop final AIM kit						🟦	🟦	🟦
Report bait design and sequencing		🟦						
Report on progress of sequencing			🟦	🟦				
Report on final results of sequencing				🟦				
Reports on progress of baseline collection sequencing					🟦	🟦		
Analyze results					🟦	🟦	🟦	
Reports on final results							🟦	🟦

PROJECT UPDATE

DNA QC. In the first year of the project, we extracted, quantified and quality checked over 600 chestnut samples from the collection of 1050 samples we started with. All the samples collected before 2014 failed our quality control tests (not much DNA and severely degraded), over 300 samples. The majority of these came from Sandy Anagnostakis, who generously gave of her time to recollect everything in the CAES collection. We now have 330 QC passed samples and intend to extract everything we now have. We have also re-extracted 45 of the samples with the best QC scores for the first sequence capture set. The bait-capture procedure requires a very high quality extract that requires a commercial kit. We save time and money in the long run by doing this “dual” extraction because 1) the standard CTAB extraction (cheap) is good enough for chloroplast sequence and 2) we do not waste expensive kits on bad samples.

Bait design. Our concept requires capture of the EST-SSR flanking sequences and the repeated sequence. This is the only way we can ensure that the captured sequence is the sequence we want. We revised the bait design to use 80bp baits instead of 120 bp baits, using a design and filtering pipeline process/criteria that allows us to target more loci. Bait kits having a total of 878 baits, representing 439 EST-SSR targets, were produced this summer.

We decided not to mix nuclear gene derived baits with chloroplast baits. The classic technique of amplifying selected chloroplast regions and Sanger sequencing is actually more cost effective. One of our undergraduates has chosen this as his multi-year research project.

Bait QC testing. This technology, if used in accordance with published protocols, usually results in sequences containing ~10% target capture and 90% nontarget capture. This means that, of all the sequences captured, only 10% are the desired sequences. While this is considered good enough for most uses, one way to reduce costs and increase quality is to increase the ratio of target to nontarget capture. We have tried a number of different procedural tweaks on a test set of six chestnuts. Our first set of data yield 15% target capture. As our goal is to accurately and precisely identify alleles, we will try more tweaks before we proceed to the first set of 45 chestnuts. This careful testing is necessary because technologies like this are all or nothing technologies. The investigator has one shot at the bait-capture per kit.

How this fits in with the overall goal. The purpose of using the bait-capture approach is to screen through a lot of markers looking for the best combinations for our purpose. Once we have the right combination (done by genotyping the first set for all of the sequences and analyzing the result in detail), then we are free to choose a different technology to genotype a much larger set of *Castanea* samples. In the first year of this project, we spent money primarily on technology. In the second year, we will spend more money on human time. This project will continue into a third year, due to a grant from the Chestnut Growers of America and the Northern Nut Growers. The growers are primarily interested in *C. mollissima, C. sativa, C. crenata* and their various hybrids. However, all the work done in these first two years benefits the grower community and all the work done in the third year will benefit the TACF, as we refine our ability to detect ancestral interspecific hybrids.

BUDGET

Item.	**Year 1(spent)**	**Year 2**
Bait design and bait-bead kit	4027.08	
Bait sequencing	2500	2500
Norgen kits	277.76	600
Second bait-bead kit	0	2500
Genomic Core technician time	3605.01	1500
Graduate student summer stipend (partial)	0	2700
Lab consumables	11.20	200
TOTAL	10421.05	10000

Budget justification

The bait design and bait-bead kits were purchased from Mycroarray, a bead kit provider based in Ann Arbor, MI. This vendor has reasonably priced kits and was willing to design the first of baits. Bait design should be done experienced people. Bait sequencing will be done by the Notre Dame Genomics Core Facility on an Illumina sequencer, although we may use the Cornell facility for the next round of sequencing (lower costs, reliable service). A Core facility technician designed the library construction approach and executed the bait-bead capture the first time, with graduate student assistance. The student will do the procedure the second time, with Core Facility supervision. We used Norgen rather Qiagen kits because the Norgen kit protocol is faster than Qiagen, while the price is comparable. The actual sum spent on this project YTD is $10421.05. The partial summer stipend in 2018 is for the assembly, alignment, and trimming of the second set of baits. Lab consumables cover part of the cost of tubes, tips, gloves, ultrapure water and other consumables.

REFERENCES

Holliday JA, Zhou L, Bawa R, Zhang M, Oubida RW (2016) Evidence for extensive parallelism but divergent genomic architecture of adaptation along altitudinal and latitudinal gradients in Populus trichocarpa New Phytologist 209:1240-1251 doi:10.1111/nph.13643

Kubisiak TL et al. (2013) A transcriptome-based genetic map of Chinese chestnut (*Castanea mollissima*) and identification of regions of segmental homology with peach (*Prunus persica*) Tree Genetics & Genomes 9:557-571 doi:10.1007/s11295-012-0579-3

McCleary T, McAllister M, Coggeshall M, Romero-Severson J (2013) EST-SSR markers reveal synonymies, homonymies and relationships inconsistent with putative pedigrees in chestnut cultivars Genet Resour Crop Evol 60:1209-1222 doi:10.1007/s10722-012-9912-9

Niedzicka M, Fijarczyk A, Dudek K, Stuglik M, Babik W (2016) Molecular Inversion Probes for targeted resequencing in non-model organisms Scientific Reports 6:24051 doi:10.1038/srep24051

http://www.nature.com/articles/srep24051#supplementary-information

Peñalba JV et al. (2014) Sequence capture using PCR-generated probes: a cost-effective method of targeted high-throughput sequencing for nonmodel organisms Molecular Ecology Resources 14:1000-1010 doi:10.1111/1755-0998.12249

Sarup P, Jensen J, Ostersen T, Henryon M, Sørensen P (2016) Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred Danish Duroc pigs BMC Genetics 17:1-16 doi:10.1186/s12863-015-0322-9