PROJECT SUMMARY

Structural Polymorphisms as Causes of Heterosis in Populus

Steven H. Strauss, Oregon State University; Stephen DiFazio, West Virginia University; Brian Stanton, Greenwood Resources; Todd Rosenstiel, Portland State University

Collaborators
Gancho Slavov, Aberystwyth University; Matias Kirst, University Florida; Jason Holliday, Virginia Tech; Michele Morgante, University of Udine, Italy; Thomas Byram, Texas A & M University; Andrew Groover, US Forest Service; Matthew Olson, Texas Tech University.

Natural genetic diversity within and among species is the basis of most plant breeding programs. It relies on variation in DNA sequences but also on variability in the presence and absence of genes and regulatory elements, or “structural polymorphisms” (SP). The objectives of this project are to characterize the extent of SP between and within species in cottonwood trees that are used to produce wood and bioenergy, and examine their relationship to growth, stress tolerance, and breeding efficiency. Because interspecies hybrids are the predominant varieties of cottonwoods used in plantations in the Pacific Northwest USA and many other parts of the world, we will focus on the extent to which assay of SPs could improve hybrid breeding compared to alternative approaches, including the use of non-SP DNA sequence variation (i.e., single nucleotide polymorphisms:SNPs) under genomic selection methods. Our preliminary work has already demonstrated that SPs are abundant within and among species, and they are associated with phenotypic variation.

We will use phenotypic and/or genomic information from existing research trials and studies, and also generate substantial new phenotypic and genomic information. We will generate SP genotype data using bait-capture/resequencing of approximately 200,000 loci in 1,150 tree genotypes. The Populus breeding program of GreenWood Resources (GWR), led by coPI Brian Stanton, will provide F1 hybrid genotypes and phenotypic information. We will use phenotypic and genomic resequencing data from the DOE BioEnergy Science Center studies of variation in black cottonwood and advanced generation backcross hybrids. In addition, we will establish a new plantation of F1 hybrids for high precision growth and physiological stress measurements at Oregon State University.

The data we will obtain and the specific hypotheses we will test are:

1) What is the extent of within- and between-species variation in SPs? We will map SPs onto the P. trichocarpa reference genome, to an assembled genome from P. nigra from collaborator Morgante, and to assemblies that we will produce for P. deltoides and P. maximowiczii. We will analyze these polymorphisms in three distinct populations that serve specific roles in the project: a) in a wild P. trichocarpa collection, which has minimal linkage disequilibrium and therefore enables precise isolation of the effects of SPs; b) in an outbred hybrid backcross of (P. trichocarpa x P. deltoides) x P. deltoides, to jointly examine the effects of between- as well as within-species variation in SPs in a large full-sib family; and c) in F1 hybrid families of P. deltoides x P. nigra (DN) and P. deltoides x P. maximowiczii (DM) to evaluate efficiency of detection within the diverse hybrids of an operational breeding program.

2) Are SPs associated with variation in productivity, stability of growth among plantation environments, and temperature stresses? We will examine the degree to which SPs are associated with productivity traits including stem height, diameter, and wood density, and its dependence on plantation environment. We will also use chlorophyll content, fluorescence, carbon assimilation, and respiration under seasonal and imposed heat stress as direct and integrative measures of field-based stress tolerance.

3) Does knowledge of structural variation provide value beyond SNPs as part of genomic selection during hybrid breeding? To assess if the genomic data we produce could be used to accelerate breeding, we will test whether selection using SPs—putatively focusing on selection against homozygosity for gene-containing deletions—can improve selection efficiency and economic gain within families over phenotypic and genomic selection approaches. This work will be based on 622 genotypes chosen from the DN and DM hybrids from the GWR breeding program characterized in hypothesis 1) above.

We have engaged collaborators who are internationally known scholars and practitioners. If successful, this project will lead to methods that can substantially increase the efficiency and rate of genetic improvement, providing economic and environmental benefits for growers and society.
Structural Polymorphisms as Causes of Heterosis in \textit{Populus}

Steven H. Strauss, Department of Forest Ecosystems and Society, Oregon State University, Corvallis, OR 97331-5752 / 541 760 7357 / Steve.Strauss@OregonState.Edu

Co-investigators

Stephen DiFazio, West Virginia University; Brian Stanton, Greenwood Resources; Todd Rosenstiel, Portland State University, Oregon

Collaborators

Gancho Slavov, Aberystwyth University; Michele Morgante, University of Udine, Italy; Matias Kirst, University of Florida; Jason Holliday, Virginia Tech; Thomas D. Byram, Texas A & M University; Andrew Groover, US Forest Service;

Administrative Point of Contact

Ms. Jeffrie Husband, (541) 737-1586, jeffrie.husband@oregonstate.edu

Funding Opportunity Announcement

DE-FOA-0000770, DOE/Office of Science Program Office: Office of Biological & Environmental Research

Program Manager Contact

Dr. Catherine Ronning

PAMS Pre-application tracking number

PRE-0000001577
Introduction and Background

Heterosis refers to the superior productivity and stress tolerance of highly heterozygous individuals. It is most widely known for its application in maize, where hybrids are the basis of nearly all commercial breeding programs worldwide. It is also now widely used in rice cultivation, most notably in China (reviewed in Chen 2010). Many perennial biomass programs also employ heterosis, for example in the form of high levels of fixed heterozygosity imparted by polyploidy in high yielding perennial grasses such as Miscanthus, and in the high productivity of interspecific hybrids in the woody biomass crops Populus, Salix, Acacia, and Eucalyptus. Many of the most advanced and intensive wood, pulp, or bioenergy production programs in the world that employ these species use interspecific hybrids because of their high productivity and stability in the face of environmental stress (White et al. 2007).

Despite decades of study, there is still not a convincing mechanistic explanation for why, on a genome scale, heterosis occurs in virtually all plant and animal species (reviews in Hochholdinger and Hoecker 2007, Birchler 2010, Goff 2011, Baranwal et al. 2012, Kaeppler et al. 2012, Schnable and Springer 2013). In fact, there is unlikely to be a single mechanism that explains heterosis for all genes, traits, genotypes, and environments (Kaeppler et al. 2012, Schnable and Springer 2013). However, one of the most powerful and broadly applicable hypotheses for heterosis is the “dominance theory.” It focuses on complementation of deleterious polymorphisms in heterotic individuals. Findings of extensive structural polymorphisms (SPs) within and between plant species have made dominance theories especially compelling (Springer et al. 2009). If these polymorphisms are slightly deleterious and became differentiated in frequency in alternate evolutionary lineages by genetic drift, the resulting genome-wide complementation in F₁ hybrids would cause heterosis (Springer and Stupar 2007). This phenomenon might be particularly powerful for taxa like Populus that have experienced whole genome duplications followed by large-scale gene loss during fractionation (Rodgers-Melnick et al. 2011), and have large repetitive genomes that promote duplications and deletions that contribute to high levels of intraspecific SP. Our preliminary work has indeed demonstrated substantial intraspecific and interspecific variation in gene number and presence in Populus. We hypothesize that these SPs are responsible for much of the extensive variation in heterosis observed among and within families of F₁ hybrids (Stettler et al. 1988), and that knowledge of their phenotypic effects can be used to design crosses that maximize heterosis.

Relevance to Program Goals

The proposed research will test the predictions of a fundamental theory of heterosis in relation to environmental stress, and its application to marker-aided selection to improve the efficiency of Populus breeding for woody biomass production. Thus, our research fits the program goals of: “Research to further understanding of complex interactions between bioenergy feedstock plants and their environment, and how these processes influence plant growth, development, and expression of bioenergy-relevant traits such as biomass yield...” Our focus on biomass growth and photosynthetic performance under heat stress also fits well with the program’s specific area of interest in “adaptation to abiotic stresses.”

Project Objectives

1) What is the extent of within- and between-species variation in SPs? We will discover SPs by performing referenced-based assemblies for P. trichocarpa, P. nigra, P. deltoides and P. maximowiczii. We will analyze these polymorphisms in three distinct populations that serve specific roles in the project (Fig. 1): a) in wild P. trichocarpa populations, which have minimal linkage disequilibrium and therefore enable the most precise isolation of the effects of SPs; b) in a outbred hybrid backcross of (P. trichocarpa x P. deltoides) x P. deltoides, to examine the phenotypic effects of intra- and interspecific SPs in a large full-sib family; and c) in F₁ hybrid families of P. deltoides x P. nigra (DN) and P. deltoides x P. maximowiczii (DM), to evaluate the value of our approach for an operational breeding program.
2) Are SPs associated with variation in productivity, stability of growth among plantation environments, and temperature stresses? We will examine the degree to which SPs are associated with key growth traits well known to impact biomass accumulation, including stem height, diameter, leaf morphology, and wood density—and its dependence on plantation environment. We will also assess the degree to which SPs impact field-based photosynthetic capacity (carbon gain), leaf respiration (carbon loss), and overall stress tolerance through measures of leaf chlorophyll content, leaf specific area, measures of leaf chlorophyll fluorescence, and gas exchange. All of these will be assessed under seasonal and imposed heat stress.

3) Does knowledge of structural variation provide value beyond SNPs as part of genomic selection during hybrid breeding? To assess if the genomic data we produce could be used to accelerate breeding, we will test whether selection using SPs—focusing on selection against homozygosity for gene-containing deletions—can improve selection efficiency and economic gain within families over phenotypic and genomic selection approaches. This work will be based on 622 genotypes chosen from 10 DN and DM hybrid families from the GWR breeding program characterized in objective 1) above.

**Overview of Technical Approach**

We will study an intraspecific association population of 725 wild genotypes where SP effects can be finely dissected due to relatively low LD (Slavov et al. 2012). There is extensive phenotype and genome resequencing data already available for this analysis: all trees were sequenced to 15X depth and have been phenotyped extensively in three common gardens for another study. We will also evaluate the predictive power of SPs in an interspecific backcross family where LD will be higher, but SP effects should be stronger due to differential fixation of slightly deleterious mutations in the different species. The thoroughly characterized *P. trichocarpa* SPs from the first (T) population will also be evaluated for their effects in a hybrid with another well-characterized species, *P. deltoides*.

Finally, we will evaluate the generality and value of our approach by performing genomic selection (GS) in a commercial breeding program operated by coI Stanton, testing the transferability of markers across families and even in different species backgrounds, and testing whether SPs add significantly to GS based on SNPs.

For each population we will obtain phenotypes over two growing seasons that are together highly related to biomass productivity: stem height, diameter, and density. Because heterosis is often associated with stability and tolerance of diverse stresses (Goff 2011), we will analyze growth in at least two different field sites. This will enable us to assess the statistical interaction of productivity traits with environment. We will also measure leaf morphology as it is highly heritable and thus more precisely reflects genetic differences. It is also highly correlated with hybrid performance even when analyzed over diverse families and environments (Marron et al. 2007). Finally, we will assess the effect of seasonal and imposed temperature variation on leaf chlorophyll content, key chlorophyll...
fluorescence parameters, and leaf gas exchange, including net assimilation and respiration. We have chosen this suite of physiological traits because they are integrative measures of leaf responses to stress, one or more of them have been associated with heterosis in prior studies (Moffat et al. 1990, Peiguo and Mingqi 1996, Chen et al. 2005, Cao 2007), and have potential for use in breeding programs (Baker & Rosenqvist, 2004; Ceulemans & Isebrands, 1996, Condon et al., 2004). We will emphasize heat stress in our assimilation studies because interspecific heterosis has been long known to be highly expressed under high temperature stress (McWilliam et al. 1969), and because of growing concern about high temperature stress during climate change.

SPs will be discovered and genotyped using high-throughput resequencing approaches that take advantage of read depth, read pair discordance, and breakpoint-spanning reads in a population context to derive robust genotype likelihoods (Handsaker et al. 2011). As described in more detail below, we have applied this approach to the *P. trichocarpa* association population and discovered a large number of segregating polymorphisms that affect gene content and show associations with phenotypes. We will apply similar approaches to discover SPs in multiple *Populus* species based on resequencing a discovery population of 24 trees per species. We will then combine this analytical approach with a sequence-capture assay devised to efficiently target the portions of the genome containing the most promising SPs in populations of interspecific hybrids. The sequence capture probes will be designed to flank the insertion break points for the SPs such that a codominant genotype can be determined from sequence data (Fig. 2). This sequence capture assay will also target portions of expressed genes that are conserved across species. This will enable detection of copy number variants segregating in hybrid populations based on sequence depth. Finally, we will capture a large number of non-coding sequences evenly spread across the genome to enable comparison of the performance of random SNP to selected SPs in explaining variance in heterosis in breeding populations. We expect to benefit from collaborations with the Kirst, Holiday, and Olson laboratories in the development of effective, multi-species bait capture designs (letters of collaboration).

**Preliminary Studies**

Below we summarize recent work that helps to demonstrate feasibility of the proposed research.

**Reference Genomes.** Col DiFazio has created a genome assembly for *P. deltoides*, revealing extensive SP polymorphism between species. Approximately 147 million paired, 75 bp Illumina reads (32X genome coverage) were assembled using the ABySS program (Simpson et al. 2009). This yielded 533,296 contigs larger than 100 bp and covered 421 Mb of sequence (N50 of 4 kb). Putative orthologous genes were identified using reciprocal BLAST analyses, homology of non-coding sequences, and synteny between *P. trichocarpa* and *P. deltoides*. Thousands of genes were present in one species and absent in the other (e.g., 1,185 genes were present in *P. deltoides* and absent in the *P. trichocarpa* reference genome). These genes belong to a number of functional categories, including those active in regulatory, metabolic, and signal transduction—supporting the potential importance of SPs for these genes to heterosis.

**Resequencing of *P. trichocarpa.*** Under the auspices of the DOE Bioenergy Science Center, Col DiFazio has established a large association genetics population with 1,100 accessions. These trees have been genotyped for over 34,000 SNP loci using an Illumina BeadArray (Geraldes et al. 2013). As part of this project, all 1,100 trees are
currently being resequenced to a depth of 15X by the DOE Joint Genome Institute. Thus far 917 genomes have been completed; the full set will be complete by the time this project begins. These sequences have been aligned to the *P. trichocarpa* reference genome using BWA (Li and Durbin 2010), and over 47 million robust polymorphisms have been identified using SAMtools and bcftools (Li 2011). These genotype calls are extremely robust: prior to imputation there is a 98% match rate to genotypes assayed independently using the BeadArray described above. Based on extensive analyses, we have identified a core set of 725 genotypes with minimal kinship and population structure (Slavov et al. 2012), thereby maximizing the efficiency and precision of analyses of phenotypic associations. This set will be used in the proposed research.

**SP Discovery in *P. trichocarpa***. Using the reference genome alignments described above, we used GenomeSTRiP to discover 7,089 high confidence insertion/deletion polymorphisms ranging in size from 205 bp to a maximum of 678 kb, with a mean and median of 7,287 and 1,465 bp, respectively. Break points for these intervals were estimated with 95% confidence at intervals averaging 13 bp. We validated these genotype calls using a nested PCR assay targeting 15 of the loci in 96 individuals. This revealed a 95% match rate between predicted and observed genotypes. Deleted regions segregated in the population at intermediate frequencies, were significantly enriched for repeat elements, and affected 2,721 unique genes. These genes were significantly enriched with proteins localized to the outer membrane, cellular wall, and periplasmic space. They were also enriched in NBS-LRR, ion channel, and sulfotransferase genes, categories putatively involved in stress resistance. Interestingly, deletions overlapping genes had a significantly lower frequency of the homozygous deletion genotype compared to the total set of deletions (Fig. 3), suggesting that these deletions are indeed deleterious. In a preliminary single locus association analysis using mixed models, we identified 17 large SP loci with suggestive associations with bud flush \((P<0.001)\). Most of these loci were outside of regions identified through association analyses based on 7 million SNPs for the same population and traits, suggesting that the SPs are independently affecting phenotypic variation. We hypothesize that the effects of SPs will be more apparent in interspecific crosses where there are numerous genome-wide differences in their frequency and occurrence.

Finally, as a test of the potential utility of genomic selection in this population, we used randomly-selected SNPs to predict progeny phenotypes using Ridge Regression-BLUP analysis. We found that predictive ability was as high as 0.4 using only 1,000 markers, and that markers selected based on their weightings in the training population provided far-superior predictive ability over randomly-selected markers in the independent test population (G. Slavov, unpubl. data). This suggests that the predictive ability conferred by GS is due to LD between markers and causative polymorphisms, rather than by capturing the effects of population structure and kinship. We hypothesize that determination of SPs will further enhance predictive ability.

**BC Phenotypic Associations and Genomic Selection**. We have extensively characterized the backcross family 52-124 to be used for this study. The family is derived from an F1 of a *P. trichocarpa* x *P. deltoides* cross, which in turn was crossed to an unrelated *P. deltoides* tree (Induri et al. 2012). We have developed an Illumina BeadArray based on whole genome sequence assemblies of the parents to target loci that were expected to segregate 1:1 in the progeny. We have mapped 3,298 of these markers for 710 progeny genotypes. We have also extensively phenotyped this population for productivity traits, crown architecture, rooting, and stress tolerance. Broad sense heritability of these traits ranged from 0.33 to 0.68 in a field study in Morgantown, WV (described below), and we have mapped a total of 115 Quantitative Trait Loci with a LOD > 3.0, individually explaining 3.1 to 13.6% of the phenotypic variance. In addition, as a test of the utility of genomic selection (GS) in this population, we have used the

![Figure 3. Frequency of *P. trichocarpa* genotypes containing insertion (I) or deletion (D) alleles. Note the statistically significant deficiency of homozygous deletions in coding regions.](image-url)
SNP markers to derive Genome Estimated Genetic Values for the progeny. We tested different training population sizes, and compared randomly-selected markers to markers selected based on their estimated effects in analyses with all markers included. The markers selected based on their predictive ability significantly outperformed random markers even with a training population of only 100 progeny (Fig. 4). This demonstrates that GS is an effective approach even with relatively small families. In this project we will evaluate the hypothesis that SPs are especially powerful for predicting heterosis.

**Research Plan**

**Plantations:** For all of the populations we will benefit from plantations and considerable data produced by previous projects (Fig. 1). The three T plantations produced and measured by the Department of Energy (Clatskanie and Corvallis OR, Placerville, CA) were established in 2009 with three clonal replicates in a randomized block design. As described under preliminary results, this population has been extensively characterized genetically and phenotypically. For the proposed work we will select the 725 genotypes that show minimal population structure and relatedness. The BC plantations, also supported by the Department of Energy are in Morgantown, WV (established in 2007 with 680 genotypes and two clonal replicates in a randomized design) and Boardman, OR (established in 2009 with 337 genotypes and 6 clonal replicates in two-tree plots and a randomized block design). We will genotype 480 of these genotypes for this study. Several traits of direct interest to the project will be available to us for analysis, including height and diameter in multiple years. As discussed above, analysis of trait data from these trials has shown strong genetic differences and many significant associations of markers with traits.

There will be two F1 plantations established in Oregon. One will be planted by GWR as part of their breeding program under drip irrigation in eastern Oregon, and the second—referred to as our high precision site—will be established on flat, uniform agricultural land at the OSU Agricultural Experiment Station in Corvallis. The GWR planting is a “Phase II” planting from their breeding program that has undergone no prior genetic selection (Stanton et al. 2010). The GWR F1 planting has 499 progeny from 8 DM families and 2 DN families, and is a subset of the 622 genotypes at the Corvallis high precision site, which includes more progeny per family. The crossing design is an incomplete diallel involving 4 *P. deltoides* females, 5 *P. maximowiczii* males, and 2 *P. nigra* males (see budget justification for full list).

For the high precision F1 planting, we will obtain dormant sticks (stems) from GWR in January and contract with a commercial propagator such as Broadacres Nursery or Mt. Jefferson Farms—both of whom we have used for this purpose with success in the past—to produce rooted, hardened trees suitable for planting under irrigation in May 2014. PI Strauss has conducted and published numerous field studies with hybrid poplars over the last 15 years and thus has defined cultural conditions to provide rapid and uniform growth in this region (e.g., Busov et al. 2011, Han et al. 2011, Voelker et al. 2011). His recent study of semi-dwarf trees (Elias et al. 2012) was conducted on the same field site as proposed for this study. Trees will be planted in a randomized block design with one genotype in each of three blocks, all surrounded by a border row. The blocks are included so that we can sample blocks for physiological data collection consecutively (by dates and/or times). This is done because those traits are likely to be temporally and diurnally variable; the timing/block effects can then be removed in ANOVA. Irrigation will not be used during the second growing season because this will allow the trees to undergo moderate drought and heat stress during the summer drought and high temperatures of August-September that are common in Corvallis. The trees will be

![Figure 4. Predictive ability of randomly-selected SNPs compared to markers selected based on weightings from Genomic Selection analyses in BC hybrids. Analysis was based on date of bud flush, examined at different family sizes in the BC population to be used in this study. Selected markers, which we expect will include SPs, substantially improved phenotype prediction capability.](image-url)
“singled” (i.e., multiple shoots removed) as needed within the first month after planting so that all the trees have a single dominant stem, facilitating estimation of stem biomass.

Phenotypic assessment. Each of the plantations will be intensively phenotyped for this project as follows. Productivity of the trees will be estimated as a “volume index” the product of tree height x basal diameter squared, which will then be weighted by stem density determined after harvest to produce a stem “biomass index” for each tree. cols DiFazio and Stanton also have total shoot biomass, and both height and diameter, from a number of 
P. trichocarpa and hybrid field studies of similar age and environment; we will derive allometric relationships of size to biomass for the most similar datasets (genotype, environment, age) and examine its correlation with our biomass index, then adjust the index statistically if there is significant non-linearity. Height and diameter have been measured, or will be measured, on all trees at the conclusion of each of two growing seasons. We will collect two fully expanded leaves per tree from two T and one BC population in Oregon (the West Virginia planting is no longer adequately maintained), and from both F₁ populations, and determine leaf area using a Li-COR LI-3100 scanning area meter present in the Strauss laboratory. The leaves will then be oven dried and dry weight determined to enable estimation of specific leaf weight (weight per unit area). Just prior to removal of the field site in the second year, a disc from the main stem of each tree will be taken to estimate wood density. The bark will be removed, the disc oven dried, and its weight and volume (water displacement) determined.

We will determine leaf chlorophyll content and fluorescence parameters (Fv/Fm, PSII efficiency, and steady-state fluorescence) using portable Chlorophyll Meters (SPAD 502) and hand-held fluorometers (OptiScience OS30P, OSSP). These instruments are non-intrusive and ideal for analyzing the large numbers of trees in our field studies (Baker and Rosenqvist, 2004). Similar to the leaf traits, analyses will occur at each Oregon field site (two F₁, two T and the BC). Measurements will occur once in the first growing season after trees are beyond planting shock, and twice in the second year. We will aim for the physiological measurements to occur during high temperature conditions in year one (e.g., August), and during both low temperature (e.g., May) and high temperature conditions in year two, as each may yield distinct fluorescence responses among genotypes. After rapid analysis of chlorophyll data from the first growing season we will pick a subset of contrasting genotypes for comprehensive analyses of simultaneous leaf gas exchange and chlorophyll fluorescence in relation to high temperature stress. Stress will be imposed by exposing leaves to a series of increasing leaf temperatures using a portable photosynthesis system with a temperature controlled leaf cuvette (LiCor Li6400XT); net assimilation, stomatal conductance, electron transport, and dark respiration will be determined at each temperature. The precise temperatures will be determined from preliminary studies, and chosen to induce high-temperature photosynthetic stress. col Rosenstiel has extensive experience in measures of gas exchange and chlorophyll fluorescence in poplar (Rosenstiel et al., 2003 & 2004; Monson et al., 2007), and has used such measures to evaluate differential patterns of physiological stress tolerance in a wide variety of plants (e.g., Adams et al 2001, Cheeke et al 2012). As genomic data becomes available, we will also use it to help identify additional genotypes most likely to differ due to heterotic causes (i.e., based on their SP genotypes), and include them in gas exchange studies during the second growing season.

Reference genomes: Most studies of copy number variation on a genome-wide scale have been carried out using microarray-based comparative genomic hybridization (e.g., Flibotte et al. 2009, Fu et al. 2010, Haun et al. 2011). However, as sequencing costs continue to decline it is becoming clear that resequencing offers a more powerful, precise, and information-rich representation of SPs (Alkan et al. 2011). Sequence-based methods are most effective when a quality reference genome is available, as is currently true for 
P. trichocarpa, but not for other 
Populus species. We will therefore produce improved reference genomes for 
P. deltoides and 
P. maximowiczii. We already have a preliminary reference assembly for 
P. deltoides based on a 35X Illumina GAII dataset (paired 75 bp reads, as described in preliminary studies), and we will obtain a draft reference genome for 
P. nigra from collaborator Morgante (see letter). We will produce a similar dataset for 
P. maximowiczii by performing deep (>35X) sequencing using 250 bp inserts and paired overlapping 150 bp reads on the Illumina HiSeq 2000. To enhance contiguity of these assemblies, we will create two mate pair libraries for each species using the Illumina Mate Pair library kit, v2.0 with target insert sizes of 2 kb and 4 kb. These will be sequenced to a depth of 15X with paired 250 bp reads on the Illumina MiSeq. Assemblies will be performed using ALLPATHS-LG (Butler et al. 2008). We will assess assembly quality based on a variety of advanced metrics (Earl et al. 2011), and through comparison to EST datasets and BAC
assemblies produced by Sanger sequencing for *P. deltoides* (Bresson et al. 2011). Although we recognize that this assembly will contain many gaps, it should be adequate for discovery of SPs, which will only require assembly around the junctions of insertion/deletion sites.

**SP discovery:** We plan to perform SP discovery in all of the parental genotypes and additional species used in the hybrid crosses. We have successfully used our collection of 15X *P. trichocarpa* genome sequences to discover SPs using GenomeSTRiP (Handsaker et al. 2011), as described in more detail in preliminary results. This requires alignment of multiple resequenced genomes to a reference genome. Based on our experience with *P. trichocarpa*, a 10X depth is adequate for high confidence genotyping of SPs. We will perform discovery with 23 individuals (including parents of the crosses) for *P. deltoides* and *P. maximowiczii*, plus the two *P. nigra* parents, for a total of 48 libraries. These libraries will be multiplexed in sets of 6 per lane and sequenced as paired 100 bp reads on a HiSeq2000 to generate 10X coverage. Sequences will be assembled to the appropriate reference genomes using BWA (Li and Durbin 2010). We will use GenomeSTRiP to identify large SPs and mpileup and bcftools to identify SNPs and small insertion/deletion polymorphisms, and calculate genotype likelihoods (Li and Durbin 2010; Li 2011). We will then use BEAGLE (Browning and Browning 2007) to perform genotype imputation jointly for SNPs, small indels, and large SPs. Given the level of linkage disequilibrium in *Populus* (Slavov et al. 2012), this depth should be adequate to impute most moderate-frequency polymorphisms in this population (Marchini and Howie 2010).

**Genotyping:** Our primary method for genotyping hybrid populations will use a sequence-capture assay devised to efficiently target the portions of the genome containing the most polymorphic gene containing SPs within and among species. The sequence capture probes will be designed to flank the insertion break points for the SPs such that a codominant genotype can be determined from sequence data (Fig. 2). We will prioritize SPs that: 1) involve coding regions; 2) show phenotypic associations in *P. trichocarpa*; 3) are closely linked to known yield and stress tolerance QTL in the BC family; 4) consist of non-repetitive flanking sequence (to maximize sequence yield across all probes); and 5) yield high confidence genotypes in the SP discovery assay. This sequence capture assay will also target portions of expressed genes based on currently-available expression data (approximately 32,000; Rodgers-Melnick et al. 2012). Four probes will be designed per gene, targeting regions that are conserved across species to maximize the general value of the methods for poplar and willow breeding. This will enable detection of copy number variants segregating in hybrid populations based on sequence depth that are not captured in the discovery phase. Finally, we will design 40,000 probes to capture non-coding sequences evenly spread across the genome to enable comparison of the performance of random SNPs to selected SPs in explaining variance in heterosis in breeding populations. We expect to benefit from collaborations with the Kirst, Holiday, and Olson laboratories in the development of effective, multi-species bait capture designs (letters of collaboration).

Illumina-compatible bar-coded libraries with 400 to 500 bp inserts will be prepared for each template using custom oligonucleotides and a sonicator (Covaris, Inc.). This work will take place using robots at the OSU Center for Genomics and Biocomputing (robot purchase currently planned, C. Rosato, pers. comm.), and at the WVU Genomics Core Facility (robot purchase planned, col DiFazio is Facility Director). We may also use robotics at cooperating laboratories (e.g., Stoilov laboratory at WVU, letter of collaboration). Libraries will be multiplexed in sets of 6 for hybridization to the capture probes, which will consist of 200,000 120mer RNA baits designed in conjunction with MycroArray, Inc (Roffard letter of collaboration). Captured targets will be further multiplexed in sets of 144 (each with a unique bar code) and sequenced with paired 150 bp reads on an Illumina HiSeq. This will provide approximately 10X depth per target. Sequences will be assembled with Bowtie2 (Langmead and Salzberg 2012) to a custom reference library consisting of predicted sequences from the appropriate species with and without insertions. Genotypes will be called using a metric like RPKM as calculated by Tophat and Cufflinks (Trapnell et al. 2009, 2010). Programs under development by collaborators Kirst and Groover (letters of collaboration) may also facilitate calling SPs based on sequence depth.

**Data analysis:** For the *P. trichocarpa* population, we have already determined that population structure and relatedness are relatively weak in the subset of the population selected for this project (Slavov et al. 2012; unpublished data). We will therefore use the SP markers in phenotype-genotype association analyses in the *P. trichocarpa* population, taking advantage of the relatively low levels of LD observed across most of the genome...
We will utilize the EMMAX mixed-model approach (Kang et al. 2010) for analysis; the model will include an identity by state kinship matrix to account for residual relatedness within our samples (estimated in EMMAX), as well as principal component axes (estimated in EIGENSOFT) to correct for population stratification if needed (Kang et al. 2010, Price et al. 2010). We will control for multiple testing and Type 1 errors in two ways. First, we will use the QVALUE approach of Storey (2002) to control the false discovery rate. Second, loci significantly associated with traits should be surrounded by increasing P-value support for associations at nearby variants, observable in Manhattan plots. We will search for these characteristic patterns, combined with our assessment of significance via QVALUE.

Because the effective population size of the BC and F1 populations is relatively low, the strength of inference and the interpretation of single marker association analyses is limited. Instead, using marker data we will apply a genomic selection approach to estimate genotypic values for growth, leaf and physiological measures (Heslot et al. 2012). Within each family we will use the Ridge Regression-BLUP R package (Endelman 2011) to estimate marker effects using 10-fold cross-validation (Heslot et al. 2012).

Because our sequence capture genotyping assay will yield a large number of densely-spaced markers throughout the genome, we can determine the relative influence of SNPs and SPs in phenotype prediction by directly comparing the effects estimated for each type of marker in the GS model. Although LD is likely to be high in the hybrid populations, there should be sufficient recombination to differentiate the predictive abilities of the markers in these relatively large families. To evaluate the relative predictive ability conferred by randomly-selected markers versus targeted SPs, we will examine the estimated effects of the individual markers (BLUPs from the GS analysis) and rank each marker type. We will also compare the overall predictive ability (the correlation between the genomic estimated genetic values and observations) of models using subsets of the markers, including: 1) SPs only; 2) SNPs within and near genes; 3) evenly-distributed SNPs and 4) polymorphisms of all classes with the highest estimated effects in the full model. We will test different numbers of markers in each category by resampling to determine the relative efficiency of each marker type. Finally, we will evaluate the robustness and generality of our approach by using models trained based on phenotypes measured in one environment to predict phenotypes in a different environment (e.g., Corvallis versus Boardman). We will also test models across families with varying degrees of relatedness, from half-sib to families of the same cross type to families of different cross types (e.g., using models trained in a DM family to predict genetic values in DN families). We hypothesize that SP predictive abilities will be more repeatable across environments and families than SNPs because they are more often the primary causative polymorphisms underlying heterosis.

Economics of MAS: We will collaborate with Tom Byram of Texas A & M (see letter of collaboration) to analyze the expected economic benefits for within family selection from MAS under genomic selection with and without SPs. col Stanton will provide economic data and relevant timelines from the GWR operational breeding programs, and PI Strauss will obtain genotyping cost estimates from commercial genomics companies at the time of analysis. In brief, Byram has developed a Simitar® (Excel add-on) module that accounts for breeding cycle length, breeding costs, and end product values adjusted for financial discount rates. The result is a stochastic estimation that enables the probability of different outcomes to be compared.

Outreach to communicate results: If the results of this project are promising with respect to application in breeding programs, we will actively communicate our work to bioenergy industries through newsletters, conferences, email, social media, and our project web sites. col Stanton is intimately involved with poplar breeding programs worldwide and has in place a network to make this rapid and effective.

Personnel Qualifications and Responsibilities

1. PI Strauss has worked in poplar genetic & genomics for two decades, and has conducted numerous field trials. He will lead on phenotyping as well as coordinate the overall study.
2. Col DiFazio has worked on poplar genomics for >10 years. He will lead on bioinformatics, including genome assembly, identification of SPs, and GS with SPs.
3. Col Stanton has bred hybrid poplars for ~30 years. He will provide plant materials and co-lead on translation, including to work with collaborator Byram on economics of MAS.

4. Col Rosenstiel has worked extensively in poplar ecophysiology; he will actively advise, provide and train in the use of physiological equipment, and help analyze and interpret data.

5. Postdoctoral Associate Kelly Vining has worked in poplar genomics and Illumina sequencing for four years. She has constructed numerous sequencing libraries and done extensive bioinformatic analysis on Illumina data. She will produce libraries, conduct bait enrichment, oversee all phenotyping, and take part in bioinformatic analysis.

6. Collaborator Slavov has worked extensively in poplar genomics, including on genome resequencing, genome-wide linkage disequilibrium, and GS. He will work closely with us on GS.

7. Collaborator Morgante works on poplar genomics and will share *P. nigra* genome and assembly data.

8. Collaborators Kirst and Holliday have developed sequence-capture technologies for poplar. They will share the sequences and associated analytical methods for data analysis.

9. Collaborator Groover works on genomics of poplar heterosis, and will share results and coordinate research with their project on poplar aneuploidy and genomics of heterosis.

10. Collaborator Byram has worked extensively on economics of marker aided selection, and will help to implement his software for analysis of potential financial benefits from GS with SPs.

### Timetable

<table>
<thead>
<tr>
<th>DNA SEQUENCING, PCR</th>
<th>2013</th>
<th>2014</th>
<th>2015</th>
<th>2016</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Year 1</td>
<td>Year 2</td>
<td>Year 3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fall</td>
<td>Winter</td>
<td>Spring</td>
<td>Summer</td>
</tr>
<tr>
<td>1. Label, collect shoots in field</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2. DNA extractions</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3. Sequence capture</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>4. Sequencing library construction</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>5. Reference genome sequencing</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>6. Sequencing for SP discovery</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>7. Multiplex sequencing</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>8. PCR/sequence verification-SPs</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PHENOTYPING</th>
<th>2013</th>
<th>2014</th>
<th>2015</th>
<th>2016</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Year 1</td>
<td>Year 2</td>
<td>Year 3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fall</td>
<td>Winter</td>
<td>Spring</td>
<td>Summer</td>
</tr>
<tr>
<td>8. Collect dormant cuttings</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>9. Multiply, propagate &amp; root</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>10. Field: Plant, irrigate, manage</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>11. Measure height, diam, lf morph</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>12. Field physiology/stress analysis</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>13. Harvest experiment</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>14. Measure stem density</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Year 1</td>
<td>Year 2</td>
<td>Year 3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fall</td>
<td>Winter</td>
<td>Spring</td>
<td>Summer</td>
</tr>
<tr>
<td>15. Genome assembly/alignment</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>16. SP/SNP discovery</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>17. Bait design for sequence capture</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>18. Adapt and test SP/GS algorithms</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>19. Analysis genotype/phenotype assoc</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>20. Economics of marker aided-selection</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
Data Sharing Plan

We will make all data public by placing it in public databases and websites. We will release data to NCBI as appropriate: For example, sequencing data will be released to trace archives as it is generated; SNP data to dbSNP, structural data to dbVar, and phenotype-genotype data to dbGaP. The _P. deltoides_ and _P. maximowiczii_ assemblies will be released to GenBank/WGS. All of the other data will also be uploaded to our public poplar gbrowse site and our main project web site as they are produced and checked for accuracy. Publication-sensitive data will be made available to collaborators, where appropriate, upon request. To help promote community, summaries and links will be provided to specialty databases [e.g., PopGenIE (http://www.popgenie.org/) and Phytozome (http://www.phytozome.net/poplar)] for poplar genomics. We also have an active working relationship with the DOE Kbase project, and will integrate our genotype and phenotype data into their platforms, and work collaboratively to develop a GS workflow.

Rationale for Scope of Study

This proposal has been extensively revised in light of reviewer comments on related versions that had been submitted to this program in two prior years. Mainly, in the current proposal we have chosen to focus on a single genomic mechanism, SPs, to enable a rigorous analysis of its importance. In addition, we think that SPs will be most important for heterosis and easiest to implement in a breeding program.

After careful consideration, we have decided to conduct all of our phenotypic measurements in the field, rather than in greenhouse or laboratory environments. This is because of the distinct physiology of trees in field environments; the nearly exclusive reliance of breeding programs on field data; and the widely known importance of stressful environments to promote expression of heterosis (Goff 2011). In addition, we have used existing plantations where possible to leverage the considerable efforts already put into making crosses and phenotyping; this project would otherwise be impossible to accomplish in a three year period.

After careful consideration we decided not to include stable isotope measurements in the study plan. Though they are integrative measures, they appear to be rarely useful in breeding programs as WUE measured by stable isotopes is often negatively correlated with yield potential (Condon et al. 2004, Blum 2005, Araus et al. 2010). Moreover, our planned gas exchange measures will measure instantaneous assimilation/transpiration ratios (WUE) and thus will give us an indication of how WUE is related to the productivity traits and SPs. Additionally, col Stanton is part of an ongoing project that is examining carbon isotope-derived water use efficiency as a breeding tool. The project is characterizing isotope-based WUE, biomass production, and osmotic potential in 152 hybrid cottonwood genotypes under varying soil moisture and drought stress. If those results are promising we may revise our plan and measure C-isotopes for at least a subset of genotypes (e.g., for those where we are measuring gas exchange).

We have had multiple contacts with eXtension personnel about developing resources from this project (A. Stone and J. McQueen, pers. comm.). We have chosen not to pursue such work at this time because it is impossible to judge if a large investment in this direction is warranted until the results of this project are known. However, if project results warrant it we will renew our contacts with eXtension and consider making the results available in their standardized format.

Finally, we note that the proposed work is distinct from a DOE Feedstock Genomics-funded project on poplar heterosis directed by Drs. Comai and Groover of the University of California at Daivs (letter of collaboration). Their funded project is complementary to ours in that it focuses on a different hypothesized mechanism of heterosis (i.e., theirs is focused on the effects of gene dosage in relation to aneuploidy, while this proposal is focused on complementation). Other specific areas where this proposal is different from theirs include the relationship of heterosis to stress; measurement of physiological mechanisms underpinning the interaction of heterosis and environmental stress; fine scale rather than karyotype scale structural variation; and analysis of translatability of results for breeding.