Published in final edited form as:
Mol Ecol. 2011 December ; 20(23): 4938–4952. doi:10.1111/j.1365-294X.2011.05335.x.

Genomic diversity in switchgrass (*Panicum virgatum)*: from the continental scale to a dune landscape

Geoffrey P. Morris[1,2,3], **Paul Grabowski**[1], and **Justin O. Borevitz**[1]

[1] Department of Ecology and Evolution, University of Chicago, 1101 East 57th St., Chicago, IL 60637.

[2] Argonne National Laboratory, Terrestrial Ecology, 9700 South Cass Avenue, Argonne, Illinois 60439.

Abstract

Connecting broad-scale patterns of genetic variation and population structure to genetic diversity on a landscape is a key step towards understanding historical processes of migration and adaptation. New genomic approaches can be used to increase the resolution of phylogeographic studies while reducing locus sampling effects and circumventing ascertainment bias. Here, we use a novel approach based on high-throughput sequencing to characterize genetic diversity in complete chloroplast genomes and >10,000 nuclear loci in switchgrass, across a continental and landscape scale. Switchgrass is a North American tallgrass species, which is widely used in conservation and perennial biomass production, and shows strong ecotypic adaptation and population structure across the continental range. We sequenced 40.9 billion base pairs from 24 individuals from across the species' range and 20 individuals from the Indiana Dunes. Analysis of plastome sequence revealed 203 variable SNP sites that define eight haplogroups, which are differentiated by 4 to 127 SNPs and confirmed by patterns of indel variation. These include three deeply divergent haplogroups, which correspond to the previously described lowland-upland ecotypic split and a novel upland haplogroup split that dates to the mid-Pleistoscene. Most of the plastome haplogroup diversity present in the northern switchgrass range, including in the Indiana Dunes, originated in the mid- or upper-Pleistocene prior to the most recent postglacial recolonization. Furthermore, a recently colonized landscape feature (~150 ya) in the Indiana Dunes contains several deeply divergent upland haplogroups. Nuclear markers also support a deep lowland-upland split, followed by limited gene flow, and show extensive gene flow in the local population of the Indiana Dunes.

Keywords

Perennial grass; polymorphism; high-throughput sequencing; chloroplast genome; coastal dunes

Introduction

Geographic patterns of genomic diversity are shaped by historical processes of migration, adaptation, and genetic drift, and may reflect changes in landscape and climate over a variety of timescales (Macaulay *et al.* 2005, Baxter *et al.* 2010, Li *et al.* 2010). The spatial and temporal resolution of phylogeographic studies in non-model organisms has been

[3] To whom correspondence should be addressed, gmorris@uchicago.edu. Fax: (773) 702-9740.

Data Accessibility
DNA sequences: NCBI SRA accession: SRP008255
Phylogenetic data: TreeBASE Study accession no. S11893

greatly increased by recent advances in high-throughput *de novo* sequencing. For example, recent work has resolved an early arctic origin for the polar bear, *Ursus maritimus*, (Lindqvist *et al.* 2010) and the successive postglacial spread of the pitcher plant mosquito, *Wyeomyia smithii*, in the northeastern United States (Emerson *et al.* 2010). These methods evaluate many loci, which makes it possible to reconcile incongruent phylogenetic signals between loci due to incomplete lineage sorting (Escobar *et al.* 2011) and locus-specific rate variation (Whittall *et al.* 2010). For example, given high variation of locus-specific *F*st in *Lycaeides* butterflies, the confidence on Fst estimates among populations was greatly improved by genotyping hundreds of loci (Gompert *et al.* 2010). Increasing the number of loci interrogated does not, however, circumvent another limitation of traditional marker systems: the ascertainment bias that arises when polymorphisms discovered in a restricted panel of individuals are used to genotype a wider set of populations (Rosenblum & Novembre 2007). For example, Moragues *et al.* (2010) found that even with a large (1536-plex) single-nucleotide polymorphism (SNP) set ascertained in cultivated barley (*Hordeum vulgare*), genetic diversity and population structure in outgroup barley landraces were underestimated because these SNPs were ascertained from a small set of cultivated varieties. Therefore, when assaying wild genomic diversity across a wide geographic range it is best to obtain non-ascertained SNPs from *de novo* high-throughput sequencing.

A variety of strategies make use of high-throughput sequencing for population genetics and phylogeography of non-model organisms. These strategies include sequencing of amplicons (Bundock *et al.* 2009), expressed sequence tags (ESTs), and restriction-digested reduced-representation libraries (RRLs; Elshire *et al.* 2011). For example, Whittall *et al.* (2010) used amplicon sequencing to generate nearly complete chloroplast genome sequence for Torrey pine *Pinus torreyana* and several related species, and discovered intraspecific variation that was not identified in previous studies using traditional marker systems. Organellar genome sequences can also be obtained by shotgun sequencing due to their small size and high-copy number, as was done for the mitochondrial genome in white-tail deer (Seabury *et al.* 2011). For nuclear polymorphisms, RRL-based studies have generated thousands of markers in non-model organisms such as in swine (Wiedmann *et al.* 2008), cattle (Van Tassell *et al.* 2008), trout (Sanchez *et al.* 2009), and grape (Myles *et al.* 2010). These studies follow a traditional two-step discovery-and-genotyping strategy that is geared towards characterization of mapping lines and breeding populations, but this strategy can introduce ascertainment bias in studies of broader patterns of genetic diversity. A number of recent studies of natural populations have carried out simultaneous polymorphism discovery and genotyping to avoid ascertainment bias while taking advantage of barcode multiplexing to combine many RRLs into a single sequencing run (Emerson *et al.* 2010; Hohenlohe *et al.* 2010; Gompert *et al.* 2010). When possible, genotyping individuals instead of pooled DNA from multiple individuals is preferable as it decreases sampling error in genotyping and enables additional analyses such as evaluating individual-specific admixture and paternity (Gompert *et al.* 2010). In plants, collecting genome-wide data from both nuclear and organellar loci provides additional information because of the differential recombination rates and inheritance of the markers. Mitochondria and chloroplasts generally track the maternal (seed) lineage and are effectively non-recombining markers of seed dispersal; whereas, the nuclear genome, with recombination between markers, is biparentally-inherited and tracks both pollen and seed dispersal (Ennos 1994).

Switchgrass (*Panicum virgatum* L.) is a wind-pollinated perennial tallgrass native to prairies, savannas, and coastal habitats ranging from Florida and New Mexico in the south to Saskatchewan and Nova Scotia in the north. There is a strong signal of climatic and latitudinal adaptation across the species' range, as evidenced by higher survival of local ecotypes in reciprocal transplants and by phenological differences (McMillan 1959; Sanderson *et al.* 1999; Casler *et al.* 2004), such as a 15-week difference in common garden

flowering time between plants from South Dakota and Texas (McMillan 1965). Ecotypic variation correlates with population structure and ploidy levels, with lowland tetraploids predominantly in the south, upland tetraploids in the north, and upland octaploids in the center of the range (Zalapa *et al.* 2011; Zhang *et al.* 2011). The extant patterns of latitudinal and climatic adaptation in the northern range were established following the retreat of Laurentide Ice Sheet around 20,000 B.P., which itself followed several recurrent contractions and expansions of the grassland biome during the Pleistocene (Pielou 1991). The postglacial colonization of the Great Lakes coastal sand dunes during the Holocene resulted in an unusual floral assemblage, with contributions from the western grasslands, the Atlantic coastal plain, the eastern deciduous forest, and the boreal region (Pielou 1991). Interestingly, these regions span several distinct switchgrass gene pools that may have contributed to the Great Lakes switchgrass population (Zhang *et al.* 2011). Unlike most landscapes inhabited by switchgrass, coastal dunes have remained highly dynamic during the Holocene, repeatedly building and destabilizing over timescales from years to millennia and forming chronosequences of recolonized habitats. For this reason, the Indiana Dunes at the southern end of Lake Michigan was the site of foundational studies on ecological succession (Cowles 1899) and its feedback with geomorphology (Cowles 1899; Olson 1958). Dune systems are also sites of endemism and local adaptation due to the harshness of the abiotic environment, where plants must contend with sand burial, poor soil development, and rapidly varying extremes of temperature and moisture (Zhang & Maun 1991; Raduski *et al.* 2010).

Here we applied high-throughput sequencing of multiplexed RRLs to characterize genome-wide nucleotide variation of diverse switchgrass cultivars and wild collected switchgrass from a dune landscape to answer these questions: (i) does a genome-wide, non-ascertained marker set support and/or enhance previously described continental population structure?; (ii) how much genomic variation exists on the local scale relative to larger geographic scales?; (iii) how did ancestral gene pools contribute to postglacial colonization of the Great Lakes dunes and, more recently, newly-formed landscape feature(s) in the Indiana Dunes? We have applied a hybrid approach, combining sequencing of complete plastomes and thousands of nuclear RRL loci, to a sample of 24 switchgrass individuals from across a dune population and 22 plants from the wider continental range. This hierarchically structured sample connects the phylogeographic patterns of genomic diversity at large spatial scales to that of a population on a landscape.

Materials and Methods

Study system and plant materials

In the dune systems of the Great Lakes, switchgrass grows in a variety of contrasting habitats, including open dunes, wooded dunes, and interdunal wetlands (Swink & Wilhelm 1994). Switchgrass, along with little bluestem (*Schizachyrium scoparium*), is a dominant species on the bare sands of interdune troughs or flats, between the marram grass (*Ammophila breviligulata*)-dominated foredune and wooded backdune. Switchgrass populations can persist in interdunes that have undergone succession to savanna or woodland with a canopy of Jack Pine (*Pinus banksiana*) and Black oak (*Quercus velutina*), despite the fact they are considered shade-intolerant (USDA NRCS 2011). Switchgrass also grows alongside obligate wetland species in interdunal swales, a rare globally-imperiled habitat (G2; Kost *et al.* 2007). In Indiana Dunes State Park (SP), interdunal wetlands are found at the base of some blowouts, the U-shaped areas of migrating sand formed after destabilization of backdunes (Swink & Wilhelm 1994). We obtained leaf samples from 21 plants in Indiana Dunes SP from interdune, savanna, and blowout habitat (Table 1). We also obtained seed from three plants in Indiana Dunes SP and one plant from a ruderal site in Chicago that were sown in a greenhouse to generate leaf samples. The coordinates of the

dune plants (Table 1) are means of eight GPS readings taken over the course of several months and are accurate to approximately one meter.

We chose eight switchgrass cultivars from across the native range to represent species-wide genomic diversity (Table 1). Since switchgrass cultivars represent seed increases from source-identified plant materials, with at most one or two generations of selection, they largely represent a sample of the wild genetic diversity at the collection site (Casler *et al.* 2007; Zalapa *et al.* 2011). The best characterized population subdivision in switchgrass is between upland ecotypes, found in the northern Great Plains and Midwest and adapted to short seasons and lower moisture availability, and lowland ecotypes, which are found in the southern Great Plains and Gulf Coast region and are adapted to long growing seasons and moist climates. Six upland cultivars were selected to represent the diversity across the Great Plains and Midwest (Table 1; Dacotah, Sunburst, Forestburg, Cave-In-Rock, Blackwell, and Southlow). Southlow is an ecopool derived from multiple source-identified lines from Southern Michigan, so the coordinates given are approximate. In addition, Kanlow, a lowland cultivar, and High Tide, a mid-Atlantic coastal cultivar were included to represent genetic diversity in the southeastern portion of the switchgrass range (Table 1).

Library preparation and high-throughput sequencing

Leaf tissue was ground in liquid nitrogen using a mortar and pestle, and DNA was extracted using a CTAB-based method adapted from Chen and Ronald (2009). DNA samples with low 260/230 ratios were further purified using a high-salt purification method adapted from Fang *et al.* (1992). DNA quantity and quality was measured on a NanoDrop 1000 (Thermo Fisher Scientific, Waltham, MA) and visualized on a 1% agarose gel. To produce our first reduced-representation libraries (Sequencing run 1, Table 1), 2.0 ug of genomic DNA from 10 samples was digested with ApeKI (NEB, Ipswitch, MA), then purified with a Qiagen PCR purification kit (Qiagen, Valencia CA). The samples were then processed according to the Illumina protocol for preparing libraries for multiplexed paired-end sequencing (Illumina, Inc. San Diego, CA) without the DNA nebulization step. In addition, undigested DNA from two of the samples (Southlow and Blackwell) was processed according to the full protocol, including the nebulization step, to produce random shotgun libraries (RSLs). This multiplexed library therefore contained separate nebulized and ApeKI-digested samples of Blackwell and Southlow. The library was sequenced with a 76 bp single-end run on an Illumina GA-II.

The results from Sequencing Run 1 indicated that digestion with ApeKI (recognition site: G^CWGC) did not reduce representation of the switchgrass genome to the desired level, so subsequent samples were digested with PstI (recognition site: CTGCA^G), which has a longer recognition site. To produce the second set of RRLs (Sequencing run 2, Table 1), genomic DNA (5.0 ug per sample) was digested with PstI-HF (NEB, Ipswich MA), then processed according to the Illumina protocol for preparing libraries for multiplexed paired-end sequencing without the DNA nebulization step. We generated 3 libraries, each containing 12 multiplexed samples. These libraries were sequenced with 100bp paired-end runs on an Illumina GA-II.

Analysis of nuclear loci

The sequencing runs were processed and de-multiplexed using standard methods (GERALD, Illumina). The complete set of reads were assembled using CLC Assembly Cell (clc_novo_assemble; CLC bio, Cambridge, MA) into a *de novo* pseudo-reference, a set of unordered, disjunct contigs for restriction-site associated loci. Sequence reads were then mapped back onto the pseudo-reference with clc_reference_long and the matches to the pseudo-reference were input into a MySQL database for further analysis. Reads that mapped

to identical positions on the pseudo-reference were combined to create “stacks”, alignments anchored at the 5' end by the PstI restriction site. Note, the 3' exonuclease activity of the end repair cocktail in the Illumina protocol leaves only the 3' “G” of the CAGCT^G PstI recognition site, which is insufficient to distinguish reads derived from digested fragments from those derived by shearing. Stacks with unusually low or high total read count, less than 150 or greater than 1500, were dropped from the analysis since they had either insufficient coverage or were derived from clonal amplification or repetitive elements (Emerson *et al.* 2010). Custom R-MySQL scripts were used to identify single nucleotide variants, which are mismatches in the stack alignment that are found in at least two samples and represented in at least four sequence reads. As a heuristic to enrich for true SNPs over substitutions between sub-genome paralogs, we use chi-squared tests (cutoff: $p < 0.01$) to test for even per-sample read count, since true SNPs are unlikely to be heterozygous in all samples (Myles *et al.* 2010). Only candidate SNPs with high base quality (Phred) scores ($Q < 0.001$) for the major and minor alleles were included in the analysis. In cases when several candidate SNPs were in a given stack, only the candidate SNP closest to the PstI site was used to avoid including highly-correlated SNPs in physical linkage. To obtain the nuclear RRL loci, stacks with any reads that mapped to the switchgrass plastome (at least 90% identity over 90% of the read) or *Sorghum bicolor* mitogenome (NC_008360.1; at least 50% identity over 50% of the read) were dropped from the analysis.

Switchgrass is polyploid, so assembly of homeologs can lead to substitutions between subgenomes being erroneously identified as SNPs, and it is difficult to differentiate SNP heterozygosity from homozygosity because the non-varying subgenome masks the SNP genotype. Given that our sample includes individuals with varying ploidy levels (4n and 8n; Costich *et al.* 2011) a dominant marker approach (Elshire *et al.* 2011) is not appropriate. Therefore, we took the approach of sampling alleles randomly from among the observed reads, which is a more robust method for analyzing multiple ploidy levels in STRUCTURE (personal communication, Jonathan Pritchard). One hundred sets of pseudo-haploid genotypes were generated by randomly selecting one sequencing read from each sample at each locus to make allele calls. For ~1500 of the loci in each genotype set the minor alleles were not sampled for any of the individuals, so these loci contained no variation and were not used in the STRUCTURE analysis. Principal coordinates analysis (implemented using cmd_scale function in R; R Development Core Team 2010) and STRUCTURE (Falush *et al.* 2007) were used to characterize population structure in the nuclear data, where each of the 100 pseudo-haploid samples was bootstrapped across loci. STRUCTURE was run on one set of pseudo-haploid genotypes (10,062 nuclear loci) using the admixture model with correlated allele frequencies among populations and ran 10,000 MCMC cycles following 10,000 burnin cycles. Five replications were run for each K ranging from K=1 to K=5 and the best K was determined according to Pritchard *et al.* (2000). The amount of missing data in the pseudo-haploid genotypes inputted in STRUCTURE ranged from 10% to 80% per sample.

Analysis of chloroplast sequence

Complete chloroplast sequences for upland *P. virgatum* cv. Summer (HQ822121) and lowland *P. virgatum* cv. Kanlow (HQ731441) was obtained from Young *et al.* (2011). Reads were mapped to the Summer reference chloroplast (clc_reference_long) with a cutoff of 90% identity over 90% of the read. Plastome sequence differences were identified using find_variants (CLC Bio) with default settings. Complete chloroplast sequences were downloaded from GenBank for three outgroup Panicoideae, maize (*Zea mays*; NC_001666.2), sugarcane (*Saccharum officinarum*; NC_006084.1), and sorghum (*Sorghum bicolor*; NC_008602.1), as well as rice (*Oryza sativa* japonica; NC_001320.1) and wheat (*Triticum aestivum*; NC_002762.1). The relationships of the outgroup species assumed for

phylogenetic analyses were, following Barker *et al.* (2001), as follows: ((rice, wheat), ((maize,(sorghum, sugarcane)), switchgrass)). Multiple alignment of reference sequence was carried out using multiple-LAGAN (Brudno *et al.* 2003) as implemented through mVISTA (Frazer *et al.* 2004). Alignments around indel sites were manually checked for errors or ambiguities in Seaview (Gouy *et al.* 2010), and the indels were partitioned into insertions and deletions based on polarization against the outgroup sequences (see Fig. 2). Phylogenetic analysis was carried using the phangorn R package (Schliep 2011). Molecular dating estimates were made using the chronMPL mean path lengths method of the ape R package (Paradis 2004). For comparison with Zhang *et al.* (2011), we used a divergence time of 65 Mya for the rice-maize split.

Results

Sequencing and mapping

For this study we obtained 211 million reads from 4 lanes of Illumina sequencing, totalling 40.9×10^9 bp of sequence in two sequencing runs (Table 1). We obtained substantial sequencing output for most of the samples: the ten ApeKI RRLs generated an average (+/- S.D.) of 2.27 (0.41) millions reads, the two random shotgun libraries (RSLs) generated an average of 3.39 (0.28) million reads, and the 34 PstI RRLs generated an average of 5.07 (2.21) million reads. One PstI RRL (BS57) failed and another (Kanlow 2) generated a normal amount of sequencing output but from only a small amount of clonally-derived sequences, so both samples were dropped from further analysis. In our pilot run (Table 1, Sequencing Run 1), we generated several thousand low-coverage ApeKI RRL loci and observed a large fraction of reads generated from random shearing of DNA (lacking the signature of the ApeKI overhang) that contained organellar sequence. Therefore, for the second sequencing run, we used PstI for greater complexity reduction and generated RRLs that also contained large amounts of shotgun organellar sequence. The portion of reads mapping to the *Sorghum bicolor* mitochondrial genome was low (0.6% for ApeKI RRLs, 0.4% for RSLs, and 0.8% for PstI RRLs) and yielded coverage that was too sparse to warrant phylogenetic analysis. In contrast, the higher proportion of chloroplast reads (1.2% for ApeKI RRLs, 0.8% for RSLs, and 2.7% for PstI RRLs), along with the smaller size of the plastome relative to the mitogenome (~140kb vs. ~470kb), lead to sufficient plastome coverage for phylogenetic analysis.

Plastome variation

We obtained complete or partial plastome sequence for 44 individual switchgrass plants ranging from 1x to 786x median coverage. Thirty-two PstI RRLs had high median coverage (44x to 786x), ten ApeKI RRLs and two PstI RRLs had low coverage (1x and 10x), and the RSLs had moderate (15x and 17x) coverage (Fig. 1, Table 1). We identified 203 chloroplast candidate SNPs, 11 chloroplast small (5 to 6 bp) segmental duplications, and 11 mono- or di-nucleotide indels (not including singletons) at poly-A tracts. We identified three major haplogroups, two upland-ecotype haplogroups (A and B; Fig. 2) and one lowland-ecotype haplogroup, which are well-supported by the high-coverage (>10X) samples (Fig. 2a). Based of the SNP differences among these haplogroups, we estimate the age of the upland/lowland ecotype divergence to c. 0.7–1.0 Mya (lower-Pleistocene) and the divergence of upland haplotypes A and B to *c.* 0.3–0.5 Mya (mid-Pleistocene). Within the upland haplogroups, five sub-haplogroups are resolvable within Group A, and 2 sub-haplogroups are resolvable within Group B (Fig. 2a). Each of the sub-haplogroups are separated from other haplogroups by at least four shared haplogroup-specific SNPs. Samples with 10x or less median plastome coverage could be differentiated into lowland/upland haplogroups and upland-A/upland-B haplogroups using the identified candidate SNPs, but could not be further resolved due to low coverage and sequencing errors (Fig. 2, unfilled lines). Two

samples, Sunburst-1 and DT29 were assigned to the upland haplogroup but could not be further resolved and were removed from subsequent analysis. The presence/absence of indels (15 segmental duplications and deletions at tandem repeats) is entirely congruent with the SNP-based phylogeny, including the lower-coverage samples (Fig. 2b). Of the seven sub-haplogroups, five are present in the Indiana Dunes (A1, A2, A4, A5, and B1; Fig. 2a) with two found only in the Indiana Dunes samples (A5 and B1), while four of the sub-haplogroups contain samples from two or more disparate populations (A1, A2, A3, and A4; Fig. 2a).

To test for purifying selection on the plastome SNPs and provide further validation, we classified SNPs in 48 tRNA (2.9 kb) and synonymous and non-synonymous SNPs in the 77 coding genes (~60 kb) based on the reference annotation (Young *et al.* 2011). If the observed SNPs resulted from sequencing error they would be equally distributed among genic and intergenic regions; in fact, there was a significantly lower fraction of SNPs in the coding sequence (Chi-squared test: $p < 10^{-10}$) and the tRNA sequences ($p = 0.046$) than in intergenic regions. As further evidence for purifying selection on plastome SNPs, we also observe that the ratio of non-synonymous to synonymous SNPs is greater ($p < 0.05$) for the younger upland-specific SNPs (16:3) than the older SNPs which predate the upland-lowland split (23:16).

Nuclear variation

To generate a set of genome-wide nuclear RRL loci, we built a pseudo-reference, created stacks of reads, and filtered out stacks containing organellar sequence, as described above. The pseudo-reference created by *de novo* assembly of the PstI RRLs consisted of 2.8×10^6 contigs. Given that there are on the order of 10^5 PstI sites per 700 Mb (Young *et al.* 2010) single-copy switchgrass genome (the frequency of the 6bp PstI recognition site is 1/4096 bp) and the vast majority of the contigs have low coverage, most contigs likely represent random regions generated by shearing of DNA during RRL preparation. A total of 19,067 PstI RRL loci (stacks) passed our cutoffs (mean per-sample coverage = 17.4X +/- 59.9X), of which 11,647 contained at least one putative SNP and were used for further analysis.

For the results presented here, we generated sets of pseudo-haploid genotypes by randomly sampling reads at each of the RRL loci for each sample and identifed population structure with Principle Coordinate Analysis (PCoA) and the Bayesian clustering algorithm STRUCTURE. Strong differentiation between individuals with the lowland plastome and the rest of the samples is shown by both PCo-1 and STRUCTURE posterior probabilities (Fig. 3). Furthermore, the upland cultivar and IDSP individuals can be separated along PCo-1, and, as expected, the magnitude of the genetic differentiation among these upland individuals is less than the differentiation between the lowland and upland individuals (Fig. 3a). One individual from Southlow is indistinguishable from the Indiana Dunes SP plants by both PCoA and STRUCTURE, but this is not surprising given the Southlow ecopool includes samples in Michigan nearby to Indiana Dunes SP (Fig. 3, Fig. 4). There is also one outlier plant from Indiana Dunes SP, DT17, which is more similar to several of the cultivars than to any of the other plants from Indiana Dunes SP. Principal coordinate analysis (Fig. 3a) shows a significant lane effect for PCo 2 (ANOVA: $p < 10^{-4}$), which has been removed in Fig. 3, but correcting for lane effect or not does not alter our interpretation based on PCo1.

Discussion

Assaying genomic diversity

In this study, we generated complete plastome sequence and thousands of nuclear markers from *de novo* sequencing of RRLs. Our approach is complementary to a number of methods that have been developed to take advantage of multiplexed RRL libraries (Baird *et al.* 2008; Andolfatto *et al.* 2011; Elshire *et al.* 2011). Multiplex Shotgun Genotyping (MSG; Andolfatto *et al.* 2011) involves light coverage sequencing followed by parental genome imputation and is appropriate in organisms with reference genomes, but when no reference sequence is available higher coverage is necessary for *de novo* assembly. Our method is similar to the RAD approach of Baird *et al.* (2008) except we do not shear the restriction fragments. Therefore, the size selection step acts as an additional complexity reduction step, similar to the GBS approach of Elshire *et al.* (2011). However, due the high specificity of the sticky-end adapters employed in RAD and GBS, these methods do not provide whole plastome sequence, as in our hybrid RRL approach. If plastome sequence is desired, low-cost multiplexing strategies like RAD and GBS could be adapted to incorporate end-polishing and blunt-end ligation similar to the Illumina multiplexing protocol used here. In fact, this modification would allow our strategy to be scaled up to hundreds of individuals by eliminating the costly separate Illumina library preparation for each individual.

Despite recent technological advances, SNP genotyping in polyploids is inherently problematic because homology between subgenomes makes it difficult to differentiate segregating SNPs from substitutions between homeologs. Methods for developing SNPs in polyploids include deep-sequencing amplicons to identify SNPs in a dose-dependent fashion (Bundock *et al.* 2009), sequencing amplicons from multiple related individuals to differentiate segregating SNPs and homeologous substitutions (Durstewitz *et al.* 2010), and using special lines, like double haploids, to identify homeologous substitutions (Trick *et al.* 2009). Morever, since gene conversion between homeologs may be rampant in polyploid plants (Salmon *et al.* 2010), individual nucleotide variants may be present both as segregating SNPs and homeologous substitutions. Likelihood approaches can be used to estimate allele frequency from high-coverage sequencing data (Lynch 2009; Hohenlohe *et al.* 2010), but these methods are not designed to deal with multiple ploidy levels and our sequencing depth was not sufficient to obtain reliable estimates. STRUCTURE can be used to evaluate population structure in polyploids (Falush et al. 2007). However, differential uncertainty in the heterozygous genotype calls across ploidy levels could lead to spurious clustering of individuals with the same ploidy level (personal communication, Jonathan Pritchard). Given the same sequencing effort, each subgenome in an octaploid will have half the coverage of each subgenome in a tetraploid. The reduced power to call genotypes in higher ploidy individuals, particularly given asymmetric heterozygosity, could lead to an underestimate of heterozygosity in octaploids. Therefore, we took a conservative approach of randomly selecting one allele at each RRL locus for each sample to generate comparable pseudo-haploid genotypes across ploidy levels. While this reduces power to identify genetic differentiation at each given RRL locus it allows us to assess genetic differentiation using many thousands of loci where alleles were not exhaustively sampled for all ploidy levels. A recent simulation study showed that a mean coverage of 20X per allele is a good target in diploid systems (Catchen *et al.* 2011), so exhaustive sampling of alleles in polyploids is likely to require much higher coverage.

A potential drawback to RRL methods is the large number of sequencing reads originating from random shearing rather than the restriction digest; however, we show that, given a reference sequence, these reads can be used to produce fully-resequenced organellar genomes. High-throughput sequencing methods may also have high error rates, which are especially unreliable at low sequencing depth (Harismendy *et al.* 2009; Hubisz *et al.* 2011).

We have high (15x to 786X) median plastome coverage in 34 samples, which greatly increases the confidence in sequencing calls. We can estimate the error rate of our consensus SNP calls for the high-coverage samples using the A1 haplogroup, which contains six individuals that are identical to the reference sequence (Summer), except for (i) a SNP found in all of our samples and the Kanlow reference (likely a derived SNP in the reference Summer individual) and (ii) a single difference found in BS66. If we assume, conservatively, that the difference in BS66 is an error, then the error rate for the high-coverage samples is about 1 error per Mbp. There also may be systematic biases in high-throughput sequencing errors (Harismendy *et al.* 2009), and in fact, we see a lane effect on the observed genetic differentiation (Fig 3a, PCo2). Lane effects may be a result of samples from the same lane sharing the same sequencing errors, which are then interpreted as shared polymorphisms. In contrast, our chloroplast results do not show lane effects (i.e. seven of eight haplogroups were found on multiple lanes or sequencing runs), probably due to the high sequencing depth of the plastome. Additional technical variation is seen in the number of the reads from each sample at each RRL locus, with mean per-sample coverage of 17.4X but ~200X more variance than expected from a Poisson distribution. This overdispersion may be due to moderate clonal amplification during the library preparation.

Population structure in switchgrass

Genetic diversity and population structure in switchgrass has been the subject of many studies due to interest in identifying regional gene pools for conservation use, locally-adapted varieties for perennial forage, and most recently, to develop switchgrass as a sustainable bioenergy crop (Casler *et al.* 2007). Early studies using chloroplast sequences (Hultquist *et al.* 1996; Missaoui *et al.* 2006), restriction-fragment length polymorphisms (Missaoui *et al.* 2006), and randomly-amplified polymorphic DNA (Gunter *et al.* 1996; Casler *et al.* 2007) were able to differentiate upland and lowland ecotypes but were unable to resolve population structure within ecotypes. By characterizing the complete chloroplast sequence, we confirmed the deep upland-lowland split and discovered enough variation to detect substantial haplotype diversity within the upland ecotype, including two deep haplogroups (A and B) and seven sub-haplogroups (A1-A5, B1-B2). Both the upland and lowland haplogroups, which originated in the lower Pleistocene *c.* 0.8 - 1.0 million years ago, and the upland-A and upland-B haplogroups, which originated in the mid-Pleistocene *c.* 0.3 - 0.5 million years ago, have persisted through several range expansion/contractions during glacial cycles. Moreover, all the sub-haplogroups we identify likely originated prior to the most recent (holocene) range expansion (c. 2 - 20 kya), and do not represent *in situ* diversification in the northern range. Interestingly, the level of divergence between upland and lowland plastomes (127 SNPs) is considerably more than the subspecies divergence between *Oryza sativa* ssp. *indica* and ssp. *japonica* (72 SNPs; Tang *et al.* 2004), and even the upland A and B divergence (50 SNPs) is similar in magnitude to the rice subspecies divergence. In comparison, the variation we see in the chloroplasts of switchgrass is much greater than is seen in North American pines, where nearly complete plastome sequences show just 0-17 nucleotide differences in pairwise comparisons among species (Whitall *et al.* 2010). The northern and southern populations of the wide-ranging Western White Pine, *Pinus monticola*, had about 20X less nucleotide divergence than northern and southern switchgrass populations (i.e. upland vs. lowland), which may be due to the slower molecular clock in pines versus grasses (Willyard *et al.* 2007).

Several recent studies have investigated nuclear polymorphism using simple sequence repeats derived from expressed sequence tags (EST-SSRs; Narasimhamoorthy *et al.* 2008; Cortese *et al.* 2010; Zalapa *et al.* 2011; Zhang *et al.* 2011) and were able to identify several lowland and upland sub-populations. However, all these EST-SSR studies used markers ascertained in four Kanlow individuals (Tobias *et al.* 2005) or single individuals of Kanlow

and Summer (Tobias *et al.* 2008). These markers were used to genotype accessions that span spatially-structured populations where ascertainment bias can lead to an underestimate of diversity (Rosenblum & Novembre 2007). Here we confirm previously observed patterns using a genome-wide set of non-ascertained markers. The population structure we observed using our nuclear loci is consistent with the findings of Cortese *et al.* (2010) using 16 EST-SSR markers, with Sunburst grouping with Shawnee (a selection of Cave-In-Rock) and High Tide grouping with Kanlow. Also, we confirm previous findings that morphological upland type individuals may genetically be more similar to lowland individuals. For instance, morphologically-upland High Tide, was found as more closely related to lowland cultivars than to upland cultivars from the Great Plains (Cortese *et al.* 2010) and Zhang *et al.* (2011) found that several other mid-Atlantic coast accessions have lowland genotypes despite their upland phenotype. All of the lowland varieties that have been investigated are tetraploid, while many upland varieties are octaploid, including several that are sympatric with lowland varieties in the southern Great Plains (Costich *et al.* 2010; Zhang *et al.* 2011). Consistent with Narasimhamoorthy *et al.* (2008), we find that Blackwell is a mix of genetically-upland and -lowland (similar to Kanlow) individuals, despite being classified as upland. Note, since the plastome and nuclear signals are congruent for these Blackwell individuals, this is evidence of mixing of ploidy levels in the seed source and not evidence of gene flow between ploidy levels. Indeed, the overall congruence of plastome and nuclear data differentiating lowland and upland genotypes (Fig. 4) implies that substantial barriers to gene flow have prevented major introgression since the split of the ecotypes (Zalapa *et al.* 2011; Zhang *et al.* 2011). In contrast, we do not observe nuclear/organellar congruence among the upland-A and upland-B groups, which indicates nuclear gene flow within ecotypes (compare Fig. 2 and Fig. 3).

Dispersal and gene flow on a dune landscape

While previous studies have identified the major continental gene pools for switchgrass (Zalapa *et al.* 2011; Zhang *et al.* 2011), little is known about the spatial scale of gene flow and dispersal within a region or across a landscape. We would expect gene flow by pollen dispersal to be rampant at a regional scale since switchgrass is wind-pollinated and is a near-obligate out-crosser (Martinez-Reyna & Vogel 2002). Conversely, we would expect gene flow by seed dispersal to be relatively slow since the seeds of switchgrass have no adaptations for wind dispersal, in marked contrast to other dominant tallgrass species like big bluestem (*Andropogon gerardii*), little bluestem (*Schizachyrium scoparium*), and Indiangrass (*Sorghastrum nutans*). In keeping with the view that switchgrass has relatively limited seed dispersal, we foundindividuals occurring only sporadically (>100 m apart) along much of the interdune (e.g. there is no switchgrass for ~1 km between IS00 and the patch that includes IS13 and IS36) and no switchgrass in several other blowout swales in Indiana Dunes SP, despite abundant bare ground in these open dune habitats and the high potential rate of seed germination and seedling emergence (~90%; Zhang & Maun 1990). Given this putatively low seed dispersal rate, as well as the fact that rhizomatous grasses may form large clonal stands (e.g. *A. gerardii*, Keeler *et al.* 2002; *Ammophila breviligulata*, Fant *et al.* 2008) and that the Big Blowout formed recently (~150 years ago; Noel Pavlovic, personal communication), we hypothesized there may be little or no haplotype diversity among the switchgrass in the Big Blowout swale. However, in sampling just six individuals we found representatives from both of the deeply diverging upland haplogroups (A and B) and two sub-haplogroups within A (A1 and A5; Fig. 3). Further sampling will be required to determine whether the high levels of deep haplotype diversity in the Indiana Dunes are due to the unusual biogeography and ecology of the Great Lakes dunes or if it is a consistent feature of switchgrass populations, including those in prairies. The lack of differentiation in nuclear markers among the Indiana Dunes individuals is consistent with extensive gene flow at the local level. For example, two individuals from the Big Blowout swale, BS36 and

BS40, are nearly identical based on the nuclear RRL loci, but have deeply divergent chloroplast haplotypes (BS36 = A1, BS40 = B1; Fig. 4). The Indiana Dunes, like all habitats in the tallgrass region, have been subject to habitat loss, fragmentation, and disturbance that may have affected geographic patterns of genomic diversity (Gustafson *et al.* 2004; Casler *et al.* 2007). For instance, Indiana Dunes SP is adjacent to transportation and utility corridors that may have been routes for the introduction of adventive genotypes (e.g. one individual, DT17, was an outlier for both the plastome and nuclear signal), but more comprehensive spatial sampling would be needed to address possible contribution to the local gene pool.

Since switchgrass is found in a variety of habitats with starkly different abiotic and biotic environments in Great Lakes coastal dunes there is a possibility of habitat-specific ecotypes, as is the case in bitter panicgrass (*P. amarum)*, a sister species (or conspecific subspecies) of switchgrass found in Atlantic coastal dunes (Barkworth *et al.* 2007). In *P. amarum* there is sympatric divergence between subspecies adapted to wet and dry dune habitats, with *P. amarum* ssp. *amarum* found predominantly on dry foredunes and *P. amarum* ssp. *amaralum* found in swales. Therefore, we were interested to see if there was any preliminary evidence for genetic differentiation between the individuals in the interdune and the swale. We observed no clear evidence for overall genomic differentiation between plants from the blowout swale as compared with those from the interdune in our nuclear data (Fig. 3). Furthermore, the chloroplast haplotypes of the swale individuals (Fig. 2; A1, A5, and B1) are also found among the interdune individuals.

Local genetic diversity is also of interest for determining the appropriate range of seed sourcing for habitat restoration projects, which will preserve local genetic diversity without unduly restricting restoration projects for lack of locally-sourced seed (Jones 2003; Gustafson *et al.* 2004; Casler *et al.* 2007). Our observations that (i) switchgrass plants from the Lake Michigan area (Indiana Dunes and Southlow) show genetic differentiation from the plants from rest of the upland switchgrass (Fig. 4) and (ii) that two chloroplast haplogroups (A5 and B1) were only found in the Indiana Dunes support the view that use of seed from nearby remnants will help capture and propagate local genotypes (Gustafson *et al.* 2004). At the same time, the evidence for extensive historical gene flow across the continental range lends support to the view that using switchgrass germplasm from a wider ecoregional gene pool, instead of narrowly defined local seed source, will not “contaminate” local gene pools (Casler *et al.* 2007).

Acknowledgments

We thank Indiana Dunes State Park and Indiana Department of Natural Resources for help with sampling, Christian Tobias and Hugh Young for access to reference switchgrass chloroplast sequences prior to publication, Jonathan Pritchard for advice on use of STRUCTURE, Michael Casler for early access to unpublished data, Nina Noah for help with plant growth and manuscript editing, and two anonymous reviewers for helpful suggestion. G.M. and J.B. were supported by University of Chicago Energy Initiative and the Argonne University of Chicago Strategic Initiative. P.G. was partially supported by National Institutes of Health Training Grant T32 GM007197.

References

Andolfatto P, Davison D, Erezyilmaz D, Hu TT, Mast J, Sunayama-Morita T, Stern DL. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. Genome Research. 2011; 21:610–617. [PubMed: 21233398]

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE. 2008; 3:e3376. [PubMed: 18852878]

Barker NP, Clark LG, Davis J, Duvall MR, Guala GF, Hsiao C, Kellogg EA, Linder HP. Phylogeny and subfamilial classification of the grasses (Poaceae). Annals of the Missouri Botanical Gardens. 2001; 88:373–457.

Barkworth, ME.; Anderton, LK.; Capels, KM.; Long, S.; Piep, MB. Manual of Grasses for North America. Utah State University Press; Logan, UT, USA: 2007.

Baxter I, Brazelton JN, Yu D, Huang YS, Lahner B, Yakubova E, Li Y, Bergelson J, Borevitz JO, Nordborg M, Vitek O, Salt DE. A coastal cline in sodium accumulation in Arabidopsis thaliana is driven by natural variation of the sodium transporter AtHKT1;1. PLoS Genetics. 2010; 6:e1001193. [PubMed: 21085628]

Britton T, Oxelman B, Vinnersten A, Bremer K. Phylogenetic dating with confidence intervals using mean path lengths. Molecular Phylogenetics and Evolution. 2002; 24:58–65. [PubMed: 12128028]

Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S. LAGAN and Multi-LAGAN: Efficient Tools for Large-Scale Multiple Alignment of Genomic DNA. Genome Research. 2003; 13:721–731. [PubMed: 12654723]

Bundock PC, Eliott FG, Ablett G, Benson AD, Casu RE, Aitken KS, Henry RJ. Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. Plant Biotechnology Journal. 2009; 7:347–354. [PubMed: 19386042]

Casler MD, Vogel KP, Taliaferro CM, Wynia RL. Latitudinal Adaptation of Switchgrass Populations. Crop Science. 2004; 44:293–303.

Casler MD, Stendal CA, Kapich L, Vogel KP. Genetic diversity, plant adaptation regions, and gene pools for switchgrass. Crop Science. 2007; 47:2261–2273.

Catchen JM, Amores A, Hohenlohe PA, Cresko WA, Postlethwait JH. Stacks: building and genotyping loci de novo from short-read sequences. G3: Genes, Genomes, Genetics. 2011; 1:171–182.

Chen DH, Ronald PC. A rapid DNA minipreparation method suitable for AFLP and other PCR applications. Plant Molecular Biology Reporter. 1999; 17:53–57.

Cortese LM, Honig J, Miller C, Bonos SA. Genetic diversity of twelve switchgrass populations using molecular and morphological markers. Bioenergy Research. 2010 doi: 10.1007/s12155-010-9078-2.

Costich DE, Friebe B, Sheehan MJ, Casler MD, Buckler ES. Genome-size variation in switchgrass (Panicum virgatum): Flow cytometry and cytology reveal rampant aneuploidy. Plant Genome. 2010; 3:130–141.

Cowles, HC. The Ecological Relations of the Vegetation on the Sand Dunes of Lake Michigan. University of Chicago Press; Chicago, IL, USA: 1899.

Durstewitz G, Polley A, Plieske J, Luerssen H, Graner EM, Wieseke R, Ganal MW. SNP discovery by amplicon sequencing and multiplex SNP genotyping in the allopolyploid species Brassica napus. Genome. 2010; 53:948–956. [PubMed: 21076510]

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE. 2011; 6:e19379. [PubMed: 21573248]

Emerson KJ, Merz CR, Catchen JM, Hohenlohe PA, Cresko WA, Bradshaw WE, Holzapfel CM. Resolving postglacial phylogeography using high-throughput sequencing. Proceedings of the National Academy of Sciences of the United States of America. 2010; 107:16196–16200. [PubMed: 20798348]

Ennos RA. Estimating the relative rates of pollen and seed migration among plant populations. Heredity. 1994; 72:250–259.

Escobar JS, Scornavacca C, Cenci A, Guilhaumon C, Santoni S, Douzery EJ, Ranwez V, Glemin S, David J. Multigenic phylogeny and analysis of tree incongruences in Triticeae (Poaceae). BMC Evol Biol. 2011; 11:181. [PubMed: 21702931]

Fang G, Hammar S, Grumet R. A quick and inexpensive method for removing polysaccharides from plant genomic DNA. Biotechniques. 1992; 13:52–54. 56. [PubMed: 1503775]

Fant JB, Holmstrom RM, Sirkin E, Etterson JR, Masi S. Genetic Structure of Threatened Native Populations and Propagules Used for Restoration in a Clonal Species, American Beachgrass (*Ammophila breviligulata* Fern.). Restoration Ecology. 2008; 16:594–603.

Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: dominant markers and null alleles. Molecular Ecology Notes. 2007; 7:574–578. [PubMed: 18784791]

Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Research. 2004; 32(Web Server issue):W273–9. [PubMed: 15215394]

Gunter LE, Tuskan GA, Wullschleger SD. Diversity among populations of switchgrass on randomly amplified polymorphic DNA (RAPD) markers. Crop Science. 1996; 36:1017–1022.

Gompert Z, Forister ML, Fordyce JA, Nice CC, Williamson RJ, Buerkle CA. Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of Lycaeides butterflies. Molecular Ecology. 2010; 19:2455–2473. [PubMed: 20497324]

Gouy M, Guindon S, Gascuel O. SeaView version 4 : a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular Biology and Evolution. 2010; 27:221–224. [PubMed: 19854763]

Gustafson DJ, Gibson DJ, Nickrent DL. Using local seeds in prairie restoration: Data support the paradigm. Native Plants Journal. 2004; 6:25–28.

Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, Frazer KA. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biology. 2009; 10:R32. [PubMed: 19327155]

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genetics. 2010; 6:e1000862. [PubMed: 20195501]

Huang S, Su X, Haselkorn R, Gornicki P. Evolution of switchgrass (Panicum virgatum L.) based on sequences of the nuclear gene encoding plastid acetyl-CoA carboxylase. Plant Science. 2003; 164:43–49.

Hubisz MJ, Lin MF, Kellis M, Siepel A. Error and error mitigation in low-coverage genome assemblies. PLoS ONE. 2011; 6:e17034. [PubMed: 21340033]

Hultquist SJ, Vogel KP, Lee DJ, Arumuganathan K, Kaeppler S. Chloroplast DNA and nuclear DNA content variations among cultivars of switchgrass, Panicum virgatum L. Crop Science. 1996; 36:1049–1052.

Jones TA. The restoration gene pool concept: Beyond the native versus non-native debate. Restoration Ecology. 2003; 11:281–290.

Keeler KH, Williams CF, Vescio LS. Clone Size of Andropogon Gerardii Vitman (Big Bluestem) at Konza Prairie, Kansas. American Midland Naturalist. 2002; 147:295–304.

Kost, MA.; Albert, DA.; Cohen, JG.; Slaughter, BS.; Schillo, RK.; Weber, CR.; Chapman, KA. Natural Communities of Michigan: Classification and Description. Report 2007-21. Michigan Natural Features Inventory; Lansing, Michigan: 2007.

Lindqvist C, Schuster SC, Sun Y, Talbot SL, Qi J, Ratan A, Tomsho LP, Kasson L, Zeyl E, Aars J, Miller W, Ingólfsson O, Bachmann L, Wiig O. Complete mitochondrial genome of a Pleistocene jawbone unveils the origin of polar bear. Proceedings of the National Academy of Sciences of the United States of America. 2010; 107:5053–5057. [PubMed: 20194737]

Lynch M. Estimation of allele frequencies from high-coverage genome-sequencing projects. Genetics. 2009; 182:295–301. [PubMed: 19293142]

Martinez-Reyna JM, Vogel KP. Incompatibility systems in switchgrass. Crop Science. 2002; 42:1800–1805.

Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, Blackburn J, Semino O, Scozzari R, Cruciani F, Taha A, Shaari NK, Raja JM, Ismail P, Zainuddin Z, Goodwin W, Bulbeck D, Bandelt HJ, Oppenheimer S, Torroni A, Richards M. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science. 2005; 308:1034–1036. [PubMed: 15890885]

McMillan C. The role of ecotypic variation in the distribution of the central grassland of North America. Ecological Monographs. 1959; 29:285–308.

McMillan C. Ecotypic differentiation within four North American prairie grasses: II. Behavioral variation within transplanted community fractions. American Journal of Botany. 1965; 52:55–65.

Missaoui AM, Paterson AH, Bouton JH. Molecular markers for the classification of switchgrass (Panicum virgatum L.) germplasm and to assess genetic diversity in three synthetic switchgrass populations. Genetic Resources and Crop Evolution. 2006; 53:1291–1302.

Moragues M, Comadran J, Waugh R, Milne I, Flavell AJ, Russell JR. Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. Theoretical and Applied Genetics. 2010; 120:1525–1534. [PubMed: 20157694]

Myles S, Chia J-M, Hurwitz B, Simon C, Zhong GY, et al. Rapid Genomic Characterization of the Genus *Vitis*. PLoS ONE. 2010; 5:e8219. doi: 10.1371/journal.pone.0008219. [PubMed: 20084295]

Narasimhamoorthy B, Saha MC, Swaller T, Bouton JH. Genetic diversity in switchgrass collections assessed by EST-SSR markers. Bioenergy Research. 2008; 1:136–146.

Olson JS. Lake Michigan dune development. II. Plants as agents and tools in geomorphology. Journal of Geology. 1958; 66:345–351.

Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004; 20:289–290. [PubMed: 14734327]

Pielou, EC. After the Ice Age: the return of life to glaciated North America. University of Chicago Press; Chicago, IL: 1991.

Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155:945–959. [PubMed: 10835412]

Raduski AR, Rieseberg LH, Strasburg JL. Effective population size, gene flow, and species status in a narrow endemic sunflower, *Helianthus neglectus*, compared to its widespread sister species, *H. petiolaris*. International Journal of Molecular Sciences. 2010; 11:492–506. [PubMed: 20386650]

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2010. ISBN 3-900051-07-0, URL http://www.R-project.org

Rosenblum EB, Novembre J. Ascertainment bias in spatially structured populations: a case study in the eastern fence lizard. Journal of Heredity. 2007; 98:331–336. [PubMed: 17611259]

Salmon A, Flagel L, Ying B, Udall JA, Wendel JF. Homoeologous nonreciprocal recombination in polyploid cotton. New Phytologist. 2010; 186:123–34. [PubMed: 19925554]

Sanderson MA, Reed RL, Ocumpaugh WR, Hussey MA, Van Esbroeck G, Read JC, Tischler CR, Hons FM. Switchgrass cultivars and germplasm for biomass feedstock production in Texas. Bioresource Technology. 1999; 67:209–219.

Sanchez C, Smith T, Wiedmann R, et al. Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. BMC Genomics. 2009; 10:559. [PubMed: 19939274]

Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011; 27:592–593. [PubMed: 21169378]

Seabury CM, Bhattarai EK, Taylor JF, Viswanathan GG, Cooper SM, Davis DS, Dowd SE, Lockwood ML, Seabury PM. Genome-Wide Polymorphism and Comparative Analyses in the White-Tailed Deer (Odocoileus virginianus): A Model for Conservation Genomics. PLoS ONE. 2011; 6:e15811. [PubMed: 21283515]

Swink, F.; Wilhelm, G. Plants of the Chicago Region. Indiana Academy of Science; Indianapolis, IN, USA: 1994.

Tang J, Xia H, Cao M, Zhang X, Zeng W, Hu S, Tong W, Wang J, Wang J, Yu J, Yang H, Zhu L. A comparison of rice chloroplast genomes. Plant Physiology. 2004; 135:412–20. [PubMed: 15122023]

Tobias CM, Sarath G, Twigg P, Lindquist E, Pangilinan J, Penning B, Barry K, Carpita N, Lazo GR. Comparative genomics in switchgrass using 61,585 high-quality EST. Plant Genome. 2008; 1:111–124.

Trick M, Long Y, Meng J, Bancroft I. Single nucleotide polymorphism (SNP) discovery in the polyploid Brassica napus using Solexa transcriptome sequencing. Plant Biotechnology Journal. 2009; 7:334–346. [PubMed: 19207216]

USDA, NRCS. The PLANTS Database. National Plant Data Team; Greensboro: 2011. (http://plants.usda.gov, 17 June 2011)NC 27401-4901 USA

Van Tassell CP, Smith TPL, Matukumalli LK, et al. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature Methods. 2008; 5:247–252. [PubMed: 18297082]

Wang YW, Samuels TD, Wu YQ. Development of 1,030 genomic SSR markers in switchgrass. Theoretical and Applied Genetics. 2010; 122:677–686. [PubMed: 20978736]

Weller SG, Keeler KH, Thomson BA. Clonal growth of Lithospermum caroliniense (Boraginaceae) in contrasting sand dune habitats. American Journal of Botany. 2000; 87:237–242. [PubMed: 10675311]

Whittall JB, Syring J, Parks M, Buenrostro J, Dick C, Liston A, Cronn R. Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. Molecular Ecology. 2010; 19:100–114. [PubMed: 20331774]

Wiedmann R, Smith T, Nonneman D. SNP discovery in swine by reduced representation and high throughput pyrosequencing. BMC Genetics. 2008; 9:81. [PubMed: 19055830]

Willyard A, Syring J, Gernandt DS, Liston A, Cronn R. Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for pinus. Molecular Biology and Evolution. 2007; 24:90–101. [PubMed: 16997907]

Young HA, Hernlem BJ, Anderton AL, Lanzatella CL, Tobias CM. Dihaploid Stocks of Switchgrass Isolated by a Screening Approach. BioEnergy Research. 2010; 3:305–313.

Young H, Lanzatella CL, Sarath G, Tobias C. Chloroplast genome variation in upland and lowland switchgrass. PLoS ONE. 2011; 6:e23980. doi:10.1371/journal.pone.0023980. [PubMed: 21887356]

Zalapa JE, Price DL, Kaeppler SM, Tobias CM, Okada M, Casler MD. Hierarchical classification of switchgrass genotypes using SSR and chloroplast sequences: ecotypes, ploidies, gene pools, and cultivars. Theoretical and Applied Genetics. 2011; 122:805–817. [PubMed: 21104398]

Zhang J, Maun MA. Sand Burial Effects on Seed Germination, Seedling Emergence and Establishment of Panicum virgatum. Holarctic Ecology. 1990; 13:56–61.

Zhang Y, Zalapa J, Casler M. Post-glacial evolution of Panicum virgatum: centers of diversity and gene pools revealed by SSR markers and cpDNA sequences. Genetica. 2011; 139:933–948. [PubMed: 21786028]

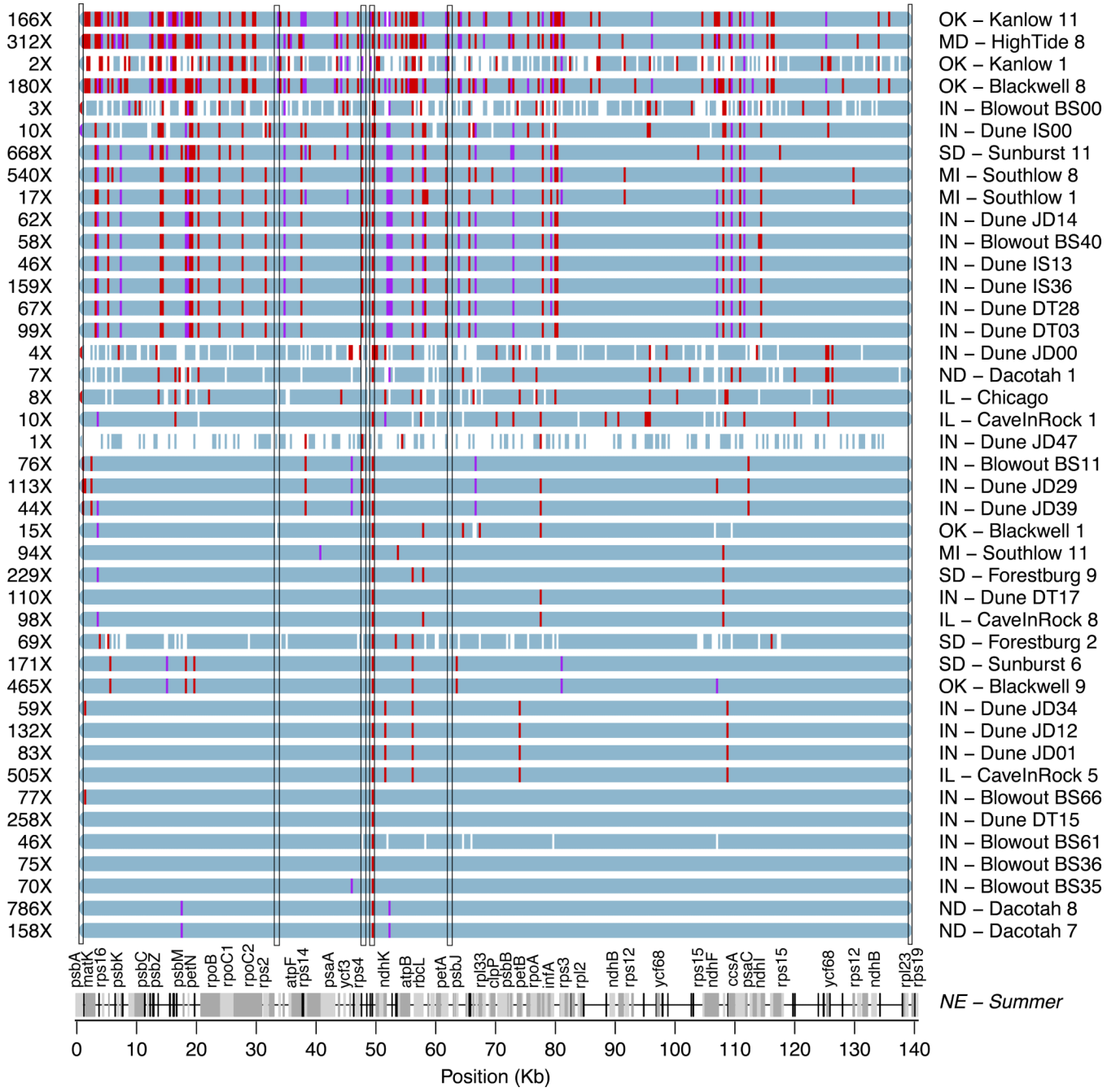

Figure 1.

Coverage and sequence variation for switchgrass plastomes. Vertical bars depict the presence of a single-nucleotide variants (red) or indels (purple) with respect to the published upland switchgrass reference sequence from *P. virgatum* cv. Summer (HQ822121) in 350 bp sliding windows. Regions with less than 2X median coverage are unshaded. Samples are sorted by similarity to the upland reference Summer. Median coverage is given on the left. Regions of the chloroplast that have been used for previous studies are designated with open black rectangles. Positions of tRNA (black bars) and coding sequences (light and dark gray segments) are from the Summer reference sequence (Young *et al.* 2011), with select coding sequences noted.

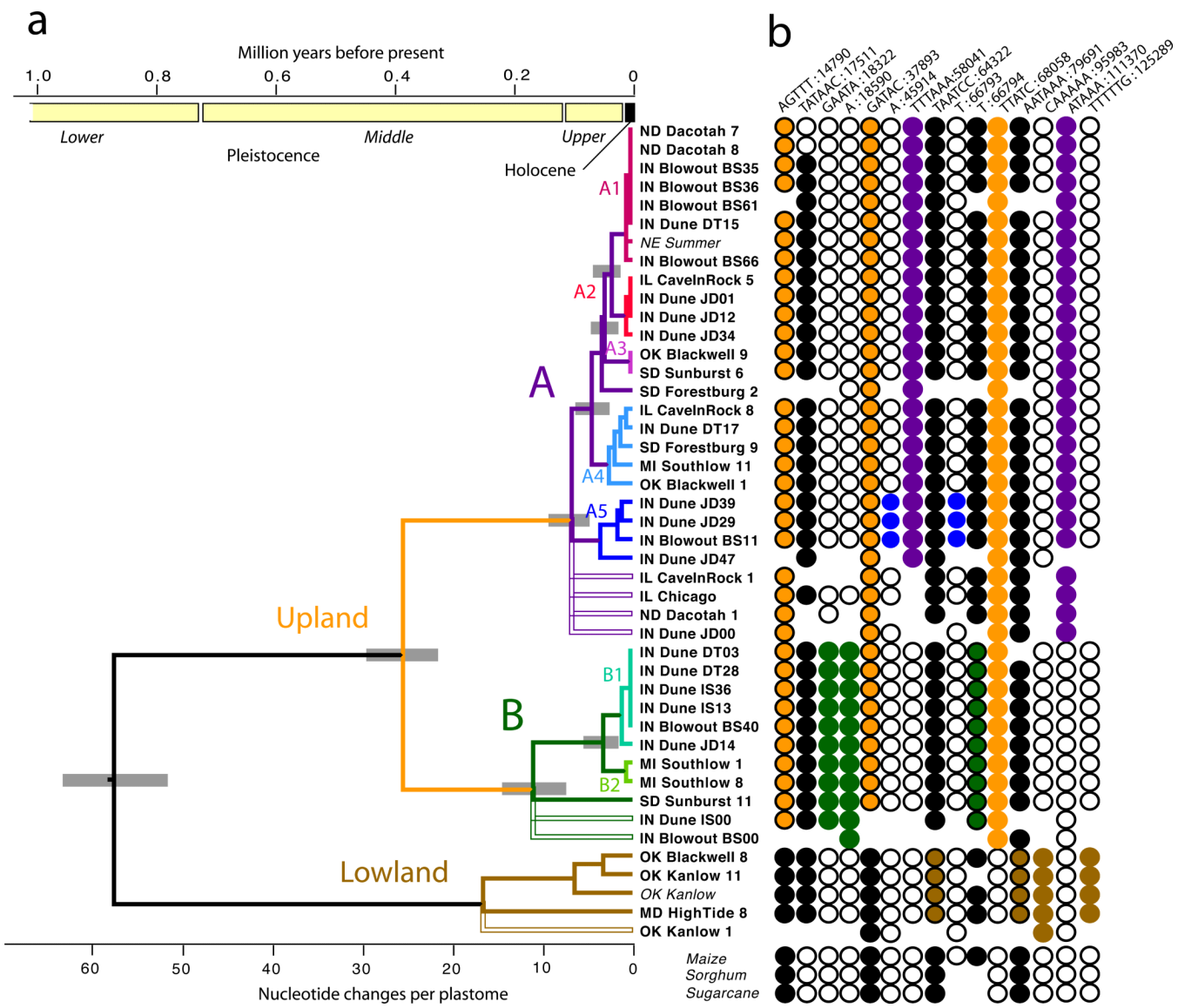

Figure 2.

Plastome haplotypes and phylogenetic relationships inferred by (a) SNPs and (b) indels. (a) Phylogeny is based on pairwise SNP differences using the UPGMA method. All named haplogroups have >50% bootstrap support using the neighbor joining method. The switchgrass (Kanlow and Summer) and Andropogoneae (maize, sorghum, and sugarcane) reference sequences used for the analysis are denoted in italics. Lower-coverage samples with partial haplogroup resolution are designated by unfilled lines. (b) Indel sequences, followed by the position of the indel with respect to the Summer reference sequence, are noted above the character matrix. Solid circles indicate the presence of the given sequence while outlined circles indicate its absence. Color coding is as in (a) and indicates the derived state, as inferred using the Andropogoneae outgroup sequences given at the bottom. Absence of a circle indicates missing data for the given sample.

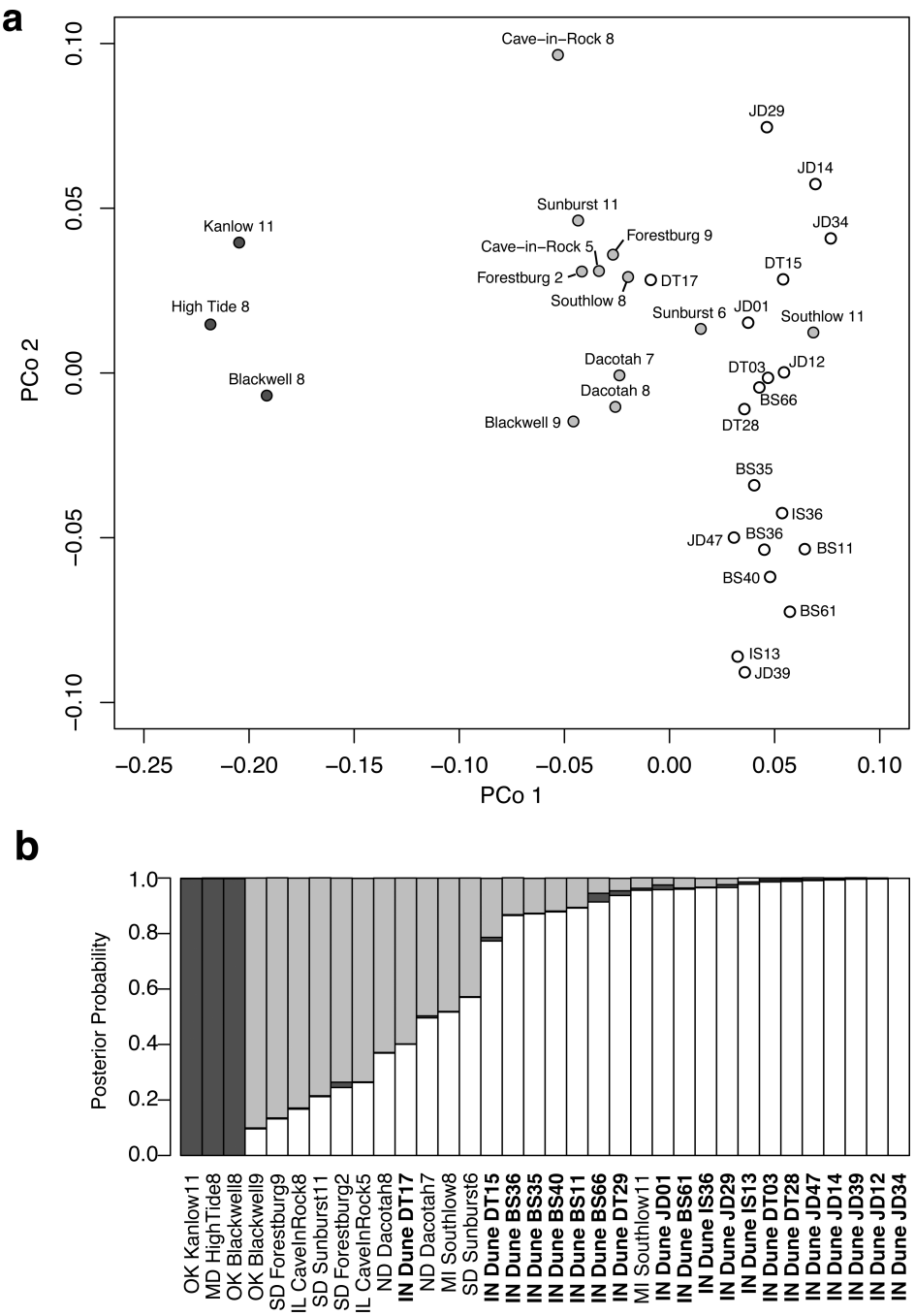

Figure 3.

Genetic similarity and population structure based nuclear loci. (a) Principal coordinates analysis for 100 random allele samples with a bootstrap sampling of loci. Indiana Dunes samples are noted in white, while other samples are indicated in light gray for genetically upland samples and dark gray for genetically lowland samples (b) STRUCTURE analysis using a random allele sample from 10,062 RRL loci with 10,000 burn-in and 10,000 runs.

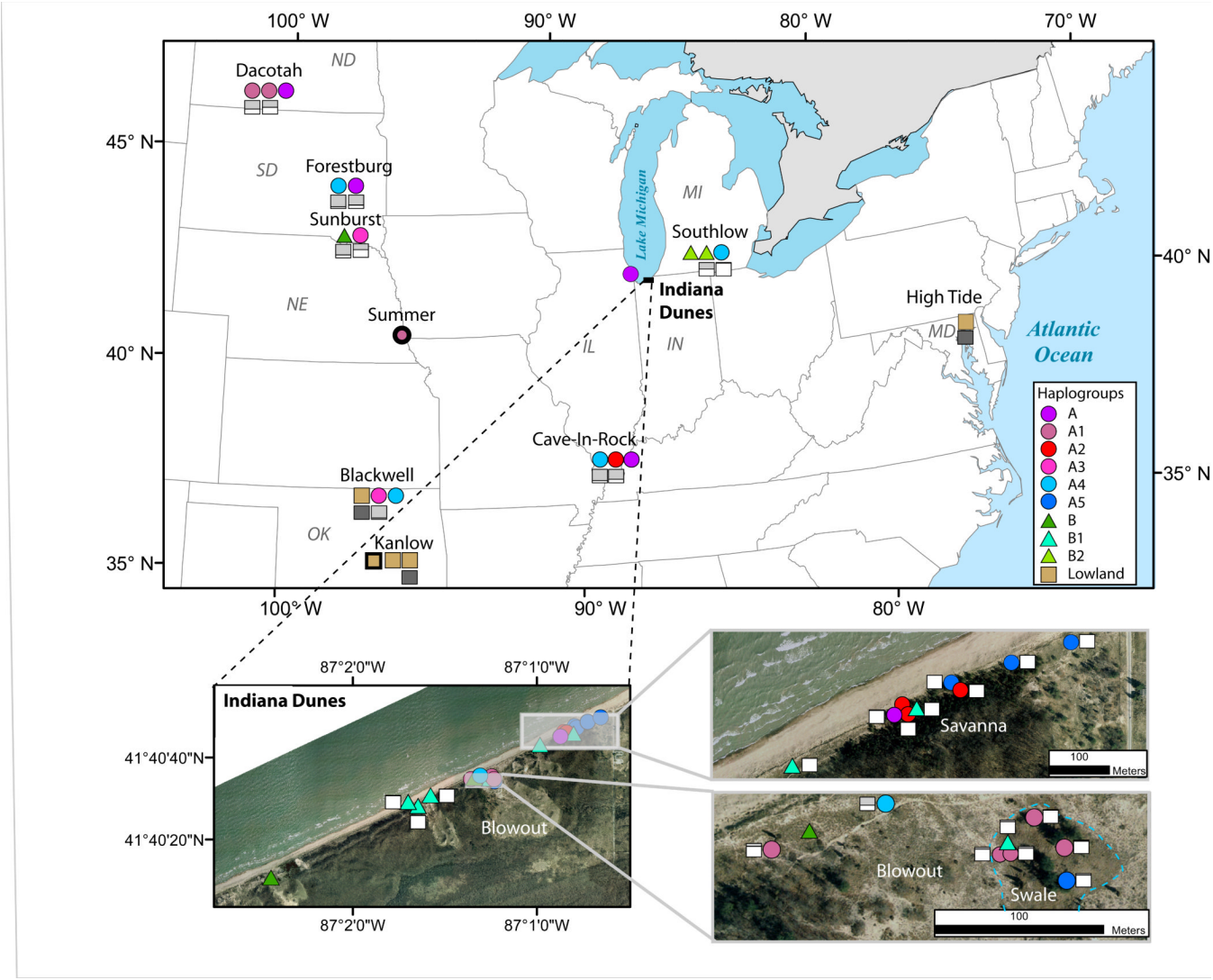

Figure 4.

Map of central United States and Indiana Dunes State Park with origins of sampled switchgrass individuals. Colored points indicate plastome haplogroup as in Fig. 2, while the adjacent gray-scale barplots indicates nuclear genome ancestry as in Fig. 3b. Kanlow and Summer reference haplotypes (Young *et al.* 2011) are noted with the outlined symbols.

Mol Ecol. Author manuscript; available in PMC 2012 December 01.

Table 1

Plant materials, library preparation, and sequencing output

Sample name	Locality information	Latitude	Longitude	Source	Sequencing Run (Lane)	Fragmentation	Sequencing reads
Blackwell 1	Blackwell, OK	36.82	-97.28	Ernst Seed	1	nebulization	3 593 118
Blackwell 8	”	36.82	-97.28	”	2 (1)	PstI	5 394 166
Blackwell 9	”	36.82	-97.28	”	2 (3)	PstI	7 632 272
Dacotah 1	Breien, ND	46.38	-100.94	Ernst Seed	1	ApeKI	2 349 374
Dacotah 7	”	46.38	-100.94	”	2 (3)	PstI	4 976 378
Dacotah 8	”	46.38	-100.94	”	2 (1)	PstI	8 777 446
Cave-in-Rock	Cave-in-Rock, IL	37.47	-88.16	Elsberry PMC	1	ApeKI	2 905 210
Cave-in-Rock 5	”	37.47	-88.16	”	2 (1)	PstI	5 268 646
Cave-in-Rock 8	”	37.47	-88.16	”	2 (3)	PstI	5 092 772
Chicago	Chicago, IL	41.8606	-87.6301	This study	1	ApeKI	2 441 452
Forestburg 2	Forestburg, SD	44.02	-98.11	Ernst Seed	2 (1)	PstI	2 731 532
Forestburg 9	”	44.02	-98.11	”	2 (3)	PstI	7 127 594
BS00	Indiana Dunes State Park, IN	41.6765	-87.0223	This study	1	ApeKI	1 491 598
BS11	”	41.6761	-87.0205	”	2 (1)	PstI	6 107 842
BS35	”	41.6763	-87.0210	”	2 (1)	PstI	6 069 018
BS36	”	41.6763	-87.0209	”	2 (1)	PstI	4 792 042
BS40	”	41.6764	-87.0210	”	2 (1)	PstI	3 122 278
BS57	”	41.6766	-87.0207	”	2 (2)	PstI	8 686
BS61	”	41.6765	-87.0208	”	2 (2)	PstI	2851078
BS66	”	41.6763	-87.0205	”	2 (2)	PstI	7 360 878
DT03	”	41.6753	-87.0263	”	2 (2)	PstI	5 805 340
DT15	”	41.6764	-87.0226	”	2 (3)	PstI	10 136 892
DT17	”	41.6766	-87.0218	”	2 (3)	PstI	4 653 522
DT28	”	41.6787	-87.0164	”	2 (2)	PstI	7 639 154
DT29	”	41.6788	-87.0161	”	2 (2)	PstI	5 152 670
IS00	”	41.6697	-87.0406	”	1	ApeKI	2 140 140
IS13	”	41.6746	-87.0276	”	2 (3)	PstI	2 942 958
IS36	”	41.6747	-87.0276	”	2 (2)	PstI	7 009 108
JD00	”	41.6793	-87.0145	”	1	ApeKI	2 350 982

Sample name	Locality information	Latitude	Longitude	Source	Sequencing Run (Lane)	Fragmentation	Sequencing reads
JD01	”	41.6795	-87.0142	”	2 (2)	PstI	6 853 616
JD12	”	41.6795	-87.0141	”	2 (2)	PstI	5 889 736
JD14	”	41.6795	-87.0139	”	2 (2)	PstI	5 096 484
JD29	”	41.6799	-87.0132	”	2 (2)	PstI	5 159 228
JD34	”	41.6798	-87.0131	”	2 (2)	PstI	4 170 848
JD39	”	41.6802	-87.0120	”	2 (3)	PstI	1 163 094
JD47	”	41.6805	-87.0109	”	2 (3)	PstI	60 308
High Tide 8	Perryville, MD	39.56	-76.07	Big Flats PMC	2 (1)	PstI	6 859 480
Southlow	Southern Michigan ecopool	42	-84	Rose Lake PMC	1	ApeKI	2 055 008
Southlow 1	”	42	-84	”	1	nebulization	3 191 758
Southlow 11	”	42	-84	”	2 (3)	PstI	6 509 868
Southlow 8	”	42	-84	”	2 (1)	PstI	3 876 554
Kanlow 1	Wetumka, OK	35.24	-96.24	Big Flats PMC	1	ApeKI	1 815 296
Kanlow 11	”	35.24	-96.24	”	2 (1)	PstI	3 121 720
Kanlow 2	”	35.24	-96.24	”	2 (3)	PstI	3 909 786
Sunburst 1	Yankton, SD	42.88	-97.39	Ernst Seed	1	ApeKI	2 646 684
Sunburst 11	”	42.88	-97.39	”	2 (1)	PstI	4 225 690
Sunburst 6	”	42.88	-97.39	”	2 (3)	PstI	4 889 424