Title

Diversity and population structure of northern switchgrass as revealed through exome capture sequencing

Journal

The Plant Journal, 84(4)

ISSN

0960-7412

Authors

Evans, Joseph

Crisovan, Emily

Barry, Kerrie

et al.

Publication Date

2015-11-01

DOI

10.1111/tpj.13041

Peer reviewed

the plant journal

SEB Society for Experimental Biology

The Plant Journal (2015) **84,** 800–815 doi: 10.1111/tpj.13041

RESOURCE

Diversity and population structure of northern switchgrass as revealed through exome capture sequencing

Joseph Evans[1,2], Emily Crisovan[1,2], Kerrie Barry[3], Chris Daum[3], Jerry Jenkins[4], Govindarajan Kunde-Ramamoorthy[3], Aruna Nandety[5], Chew Yee Ngan[3], Brieanne Vaillancourt[1,2], Chia-Lin Wei[3], Jeremy Schmutz[3,4], Shawn M. Kaeppler[6,7], Michael D. Casler[6,8] and Carol Robin Buell[1,2,*]

[1]*DOE Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, MI 48824, USA,*
[2]*Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA,*
[3]*Department of Energy, Joint Genome Institute, Walnut Creek, CA 94598, USA,*
[4]*HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806, USA,*
[5]*Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK 73019, USA,*
[6]*DOE Great Lakes Bioenergy Research Center, University of Wisconsin-Madison, 1575 Linden Drive, Madison, WI 53706, USA,*
[7]*Department of Agronomy, University of Wisconsin-Madison, 1575 Linden Drive, Madison, WI 53706, USA, and*
[8]*USDA-ARS, U.S. Dairy Forage Research Center, 1925 Linden Dr., Madison, WI 53706-1108, USA*

Received 19 June 2015; revised 31 August 2015; accepted 3 September 2015; published online 1 October 2015.
*For correspondence (e-mail buell@msu.edu).

SUMMARY

Panicum virgatum **L. (switchgrass) is a polyploid, perennial grass species that is native to North America, and is being developed as a future biofuel feedstock crop. Switchgrass is present primarily in two ecotypes: a northern upland ecotype, composed of tetraploid and octoploid accessions, and a southern lowland ecotype, composed of primarily tetraploid accessions. We employed high-coverage exome capture sequencing (~2.4 Tb) to genotype 537 individuals from 45 upland and 21 lowland populations. From these data, we identified ~27 million single-nucleotide polymorphisms (SNPs), of which 1 590 653 high-confidence SNPs were used in downstream analyses of diversity within and between the populations. From the 66 populations, we identified five primary population groups within the upland and lowland ecotypes, a result that was further supported through genetic distance analysis. We identified conserved, ecotype-restricted, non-synonymous SNPs that are predicted to affect the protein function of *CONSTANS* (*CO*) and *EARLY HEADING DATE 1* (*EHD1*), key genes involved in flowering, which may contribute to the phenotypic differences between the two ecotypes. We also identified, relative to the near-reference Kanlow population, 17 228 genes present in more copies than in the reference genome (up-CNVs), 112 630 genes present in fewer copies than in the reference genome (down-CNVs) and 14 430 presence/absence variants (PAVs), affecting a total of 9979 genes, including two upland-specific CNV clusters. In total, 45 719 genes were affected by an SNP, CNV, or PAV across the panel, providing a firm foundation to identify functional variation associated with phenotypic traits of interest for biofuel feedstock production.**

Keywords: ***Panicum virgatum*****, exome capture, switchgrass, polyploid, genomics, PRJNA280418.**

INTRODUCTION

Panicum virgatum L. (switchgrass) is a perennial C4 grass species native to North America (Vogel, 2004), traditionally grown for conservation and forage, but recently identified as a potential biofuel feedstock. Switchgrass is broadly separated into two ecotypes, upland and lowland, based on a combination of genome ploidy, phenotype, and habitat (Costich *et al.*, 2010). Previous work has demonstrated the presence of population groups within each ecotype associated with geographic habitat and ploidy (Lu *et al.*, 2013). The habitat for upland switchgrass

stretches from central USA to southern Canada, and lowland switchgrass can be found from central USA to northern Mexico. Phenotype and growth habit can vary widely between ecotypes, with southern lowland switchgrass more adapted to longer growing seasons, tending towards larger size and higher biomass, whereas upland switchgrass is frequently smaller yet more capable of overwintering in cold climates. All known switchgrass is polyploid, with lowland switchgrass being tetraploid ($2n = 4\times = 36$) and upland switchgrass being a mix of tetraploid and octoploid accessions ($2n = 8\times = 72$), with aneuploidy observed in both ecotypes (Hopkins *et al.*, 1996; Lu *et al.*, 1998; Costich *et al.*, 2010). The draft genome sequence of the lowland tetraploid individual, AP13, is 1230 Mb with 98 007 annotated genes (version 1.1; http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Pvirgatum).

Whereas considerable interest exists in the development of switchgrass for conservation, forage, and biofuel feedstock, there have been a limited number of reports on genetic diversity of switchgrass populations. For example, conservation efforts with switchgrass often involve the use of germplasm sourced nearby to avoid the importation of alleles that are poorly adapted to local conditions (Clewell and Rieger, 1997; Lesica and Allendorf, 1999). Thus, access to a detailed view of switchgrass population structure would allow the identification of switchgrass populations most suitable for reintroduction to a restored area (Casler *et al.*, 2007a). Until recently, genetic markers for switchgrass were primarily used to identify the geographic origin of individual ecotypes (Hultquist *et al.*, 1996; Missaoui *et al.*, 2006), and were of limited resolution. The development of chloroplast-based simple sequence repeat markers permitted greater differentiation of populations within ecotypes, and the identification of gene flow between upland and lowland populations (Zalapa *et al.*, 2011; Zhang *et al.*, 2011a). More recently, a reduced representation approach that entailed the next-generation sequencing of restriction enzyme sites was combined with *de novo* single-nucleotide polymorphism (SNP) prediction to identify SNPs, leading to the identification of individual populations within the larger ecotype groupings (Lu *et al.*, 2013).

A prime rationale for the use of switchgrass as a biofuel feedstock is its ability to generate substantial biomass under marginal conditions, attributable in part to its large perennial root system (Schmer *et al.*, 2008). Although these natural adaptations make switchgrass valuable as a source of biomass, much opportunity remains for improvement as a biofuel feedstock crop (Vogel, 2004; Sanderson, 2007). Central to this improvement is the development of switchgrass that can perform in variable environments and habitats, for which an understanding of the genetic and phenotypic diversity in native switchgrass populations would lead to improved breeding strategies (Casler *et al.*, 2011; Zalapa *et al.*, 2011; Zhang *et al.*, 2011a,b), as studies have shown a clear genetic and phenotypic division between lowland and upland ecotype switchgrass, even between members of different ecotypes that grow at similar latitudes (Zalapa *et al.*, 2011; Zhang *et al.*, 2011a,b; Lu *et al.*, 2013).

Genome-level analyses of grass genomes is complicated by the large number of repetitive elements that can make up a significant fraction of the total size of the genome (Paterson *et al.*, 2009; Schnable *et al.*, 2009; Zhang *et al.*, 2012). Switchgrass is no exception, with initial analyses indicating that repetitive elements compose approximately 33% of the switchgrass genome (Sharma *et al.*, 2012). Additionally, switchgrass is largely self-incompatible, and thus switchgrass populations have high levels of heterozygosity (Talbert *et al.*, 1983; Martinez-Reyna and Vogel, 2002; Liu and Wu, 2012). Recently, an exome capture oligonucleotide probe set that targets approximately 50 Mb of genic sequence was developed for switchgrass (Evans *et al.*, 2014). When coupled with paired-end sequencing, this probe set permits the capture of an additional 120 Mb of the genome at sufficient depth for high-confidence discrimination of sequence variants between and within switchgrass ecotypes and accessions (Evans *et al.*, 2014).

In this study, we used exome capture sequencing on a panel of 537 individuals of northern switchgrass from 66 populations, representing upland and lowland ecotypes and tetraploid and octoploid ploidy levels. This panel was originally grown from seed at the USDA-ARS Forage Research Center (http://www.ars.usda.gov) in 2007, and was used to demonstrate the UNEAK pipeline by Lu *et al.* (2013) using a genotyping-by-sequencing (GBS) approach, and in assessing the potential for genomic selection in switchgrass (Lipka *et al.*, 2014). The majority of these populations are wild prairie remnants that have been under minimal selection during advancement (Casler *et al.*, 2007a; Lu *et al.*, 2013), representing wide geographical variation from habitats spanning a substantial portion of the USA east of the Rocky Mountains. From ~2444 Gb of exome capture sequence data generated from the panel, we identified 1 590 653 high-confidence SNPs, 129 858 copy-number variants (CNVs) and 14 430 presence/absence variants (PAVs), some of which were ecotype-restricted. Even though the majority of the populations examined were wild populations, we were able to assign nearly all populations into discrete clusters representing differences in ploidy, ecotype, and geographic location. This represents the largest body of identified variants reported to date in switchgrass, and can be coupled with phenotypic variation present within this panel to facilitate genome-wide association studies to identify regions of the genome associated with biofuel feedstock traits.

RESULTS

Exome capture sequencing and detection of single-nucleotide polymorphism

Exome capture sequencing was performed on 537 individuals belonging to 66 switchgrass populations (Table 1) that were initially described in Lu *et al.* (2013), and are referred to in this study as the Northern Switchgrass Panel. A total of 21 of these populations were lowland ecotypes, with 20 tetraploid and one of mixed ploidy. Of the 45 upland populations, one had mixed ploidy, 15 were tetraploid and 29 were octoploid (Table 1; Lu *et al.*, 2013). DNA was extracted from each individual and subjected to exome capture, followed by 150-nucleotide paired end sequencing (2 × 150 nucleotide read pairs). Individuals were pooled and multiplexed (12 per lane) to generate approximately 22 million reads (11 million read pairs) per individual (Appendix S1). In total, approximately 2444 Gb of sequencing data was generated. Filtered reads were aligned with the hardmasked *P. virgatum* 1.1 genome assembly (http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Pvirgatum) using BOWTIE (Langmead *et al.*, 2009).

We achieved an average of at least 1× coverage of 336 985 710 bases of the *P. virgatum* 1.1 reference genome sequence, with 172 478 563 bases having an average depth coverage of five or greater (Table S1). This results in an average of ~13.5× depth of coverage for all bases with any coverage of ~24× depth of coverage if only considering positions with a depth of coverage of 5× or greater (Table S1). An initial set of 27 193 566 positions possessed an SNP in at least one individual of the population, and after filtering to identify loci that are polymorphic in at least two individuals and non-polymorphic in at least two individuals, to reduce the impact of sequencing errors and rare alleles, 12 089 346 loci remained in the sample set. Of these ~12 million positions, 875 267 were predicted to be tri-allelic across the panel and 17 668 were predicted to be tetra-allelic. Additional stringent filtering was performed to exclude loci with missing data or loci where more than 5% of the individuals had low coverage, resulting in a final pool of 1 590 653 bi-allelic polymorphic positions, with an additional 25 074 tri-allelic positions and 259 tetra-allelic positions. This high-confidence, high-coverage bi-allelic SNP set (1 590 653) has been labeled as switchgrass HapMap v1, and is used for all subsequent analyses.

Comparison of sequence variants generated by exome capture sequencing with genotyping by sequencing

We compared our polymorphism results with previous results using the same diversity panel obtained using a GBS approach coupled with a network analysis approach to identify SNPs (Lu *et al.*, 2013). Approximately 39% of the GBS-derived sequences could be uniquely aligned with the unmasked switchgrass genome, from which only 6460 positions overlapped with polymorphisms in the HapMap v1 panel. The relatively small number of overlapping positions is a result of different approaches to polymorphism detection. Approximately 172 Mb of the genome (~14%) was captured with sufficient coverage to pass our quality filtering using exome capture (which included read depth and coverage throughout the individuals within the panel), whereas the sequencing approach undertaken by Lu *et al.* (2013) sampled *Ape*KI sites across the genome at low coverage, resulting in a high degree of missing data (Appendix S2). Of the 6460 overlapping positions, 5012 (77.5%) had identical alleles in both data sets, with only the alternate (non-reference) allele identified in 1362 (21.1%) positions from the GBS panel, only the reference allele identified in 47 (~0.1%) positions and neither allele identified in 39 (~0.1%) positions, indicating a high level of concordance between the data sets.

Single nucleotide polymorphisms in the switchgrass genome

As expected, the majority of detected polymorphisms in the HapMap v1 data set were located within genes, with nearly 50% of the SNPs located within exons (Table 2), consistent with previous work on exome capture in *Hordeum vulgare* (barley; Mascher *et al.*, 2013), *Oryza sativa* (rice; Henry *et al.*, 2014) and switchgrass (Evans *et al.*, 2014); in total, 36 812 genes (37.6% of annotated genes) contained a HapMap v1 SNP (Appendix S3). A substantially larger number of SNPs were detected in the 3′-untranslated region (3′-UTR) and downstream of genes than were detected in the 5′-UTR or upstream region of genes (Table 2). Analysis of UTR lengths in v1.1 of the annotated switchgrass gene set (http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Pvirgatum) revealed not only more but also longer 3′-UTRs than 5′-UTRs, attributable to biases in annotation, and consistent with our detection of five times more SNPs in the 3′-UTR compared with the annotated 5′-UTR. Substantially more non-synonymous exonic SNPs were detected than synonymous SNPs, with a non-synonymous : synonymous ratio of approximately 1.73 : 1.00 (Table 2). This value is higher than has been observed in rice, where non-synonymous/synonymous SNP ratios of between 1.36 : 1.00 and 1.52 : 1.00 were observed on all 12 chromosomes, excluding sequences not anchored to the 12 pseudomolecules (The 3000 Rice Genomes Project 2014).

Comparing the distribution of SNPs across switchgrass ploidies, ecotypes and population origin types reveals that although the number of variants differ between populations, the distribution of the types of variant remains the same (Figure S1). In all cases, approximately 50% of detected SNPs are located in the exons, 30% are located in the introns and 10% are located in the 3′-UTR, with the remaining 10% of SNPs located in the intergenic space,

Table 1 Population identification, number of individuals, type of population, physical location, ecotype and ploidy of the samples used in this study

Population	Individuals[a]	Population type	Location	Ecotype	Ploidy
Alamo	1	Bred cultivar	TX	Lowland	4
Blackwell	10	Natural track cultivar	OK	Upland	8
Carthage	6+2	Natural track cultivar	NC	Upland	8
Cave in Rock	10	Natural track cultivar	IL	Upland	8
Dacotah	8	Natural track cultivar	ND	Upland	4
ECS-1	5	Natural population	NJ	Lowland	4
ECS-10	6	Natural population	PA	Upland	8
ECS-11	6+	Natural population	PA	Upland	8
ECS-12	7	Natural population	NY	Upland	8
ECS-2	6	Natural population	OH	Upland	8
High Tide	10+1	Natural track cultivar	MD	Lowland	4
Kanlow	10+1	Natural track cultivar	KS	Lowland	4
KY1625	10+1	Natural track cultivar	KY	Upland	8
Pathfinder	10	Bred cultivar	NE	Upland	8
Shelter	9	Natural track cultivar	WV	Upland	8
Sunburst	9	Bred cultivar	SD	Upland	8
SW102	10	Natural population	WI	Upland	4
SW109	9+1	Natural population	WI	Upland	8
SW110	10	Natural population	WI	Upland	8
SW112	10	Natural population	WI	Upland	8
SW114	8	Natural population	WI	Upland	8
SW115	6+1	Natural population	WI	Upland	4
SW116	10	Natural population	WI	Upland	4
SW122	6+1	Natural population	WI	Upland	8
SW123	10+1	Natural population	WI	Upland	8
SW124	10	Natural population	WI	Upland	4
SW127	8+1	Natural population	WI	Upland	8
SW128	9	Natural population	WI	Upland	8
SW129	10	Natural population	WI	Upland	4
SW31	8	Natural population	IN	Upland	4
SW33	6+3	Natural population	IN	Upland	8
SW38	4	Natural population	IN	Upland	8
SW40	7+1	Natural population	IN	Upland	4
SW43	8	Natural population	MI	Upland	4
SW46	9	Natural population	MI	Upland	4
SW49	7	Natural population	MN	Upland	4
SW50	5	Natural population	MN	Upland	8
SW51	7	Natural population	MN	Upland	8
SW58	10	Natural population	MN	Upland	8
SW63	6	Natural population	NY	Upland	4
SW64	10+1	Natural population	OH	Upland	8
SW65	8	Natural population	OH	Upland	8
SW781	6	Natural population	NY	Lowland	4
SW782	7	Natural population	VA	Upland	8
SW786	7	Natural population	MI	Upland	4
SW787	10	Natural population	MI	Upland	4
SW788	10+1	Natural population	NY	Lowland	4
SW789	6+1	Multisite synthetic	MS	Lowland	Mixed
SW790	5	Bred cultivar	MS	Lowland	4
SW793	5	Natural population	NY	Lowland	4
SW795	8+1	Natural population	NY	Lowland	4
SW796	9	Natural population	NY	Lowland	4
SW797	8	Natural population	NY	Lowland	4
SW798	5+1	Natural population	NY	Lowland	4
SW799	7	Natural population	NY	Lowland	4
SW802	5	Natural population	NY	Lowland	4
SW803	9	Natural population	NY	Lowland	4

(continued)

Table 1. (continued)

Population	Individuals[a]	Population type	Location	Ecotype	Ploidy
SW805	9	Natural population	NY	Lowland	4
SW806	9	Natural population	NY	Lowland	4
SW808	10	Natural population	KY	Upland	8
SW809	10	Natural population	KY	Upland	8
SWG32	7	Natural population	IL	Lowland	4
SWG39	7	Natural population	IA	Lowland	4
Timber	7	Multisite synthetic	NC	Lowland	4
WS4U	8	Multisite synthetic	WI	Upland	4
WS98-SB	8	Natural population	WI	Upland	Mixed

All populations used in this study were presented in Lu *et al.* (2013), with the exception of Alamo. Populations were defined as 'Natural' if they were obtained from wild-growing populations, 'Bred cultivars' if they were of mixed descent and had undergone selection by breeders, 'Natural track cultivars' if they were derived from wild germplasm that have been selected for traits by breeders and 'Multisite synthetic' if it was a mixed-descent population grown at multiple locales.
[a]Individuals sequenced more than once are indicated with +, i.e. 10+1 means 10 individuals, with one sequenced twice.

Table 2 Types of genetic variation detected in the 1 590 653 Switchgrass HapMap v1 high-confidence single-nucleotide polymorphism data set

Genomic feature	Number of loci
Exonic	773 368
Synonymous	276 342
Non-synonymous	478 732
Stop gain	17 276
Stop loss	1018
Exonic and splicing	3
Intronic	499 609
Splicing	7146
5′ untranslated region	34 459
3′ untranslated region	171 944
3′ and 5′ untranslated region	230
Upstream	13 115
Downstream	41 837
Upstream and downstream	6067
Intergenic	42 881

5′-UTR and downstream of genes. Additionally, the distribution of the predicted effects of SNPs also remains largely uniform across all sample groupings (Figure S2). When projected onto the reference AP13 annotation, approximately 62% of SNPs within coding regions are non-synonymous, and result in an amino acid change in their corresponding gene, versus approximately 36% of SNPs that are predicted to result in no amino acid change, and the remaining ~2% consisting of mainly predicted stop-gain conditions, with a much smaller number of stop-loss conditions (Figure S2).

When plotted against their physical locations, patterns of SNP distribution become apparent. In general, SNP density peaks in regions of high gene density because of the nature of exome capture corresponding to regions of highest read depth (Figure 1). This uneven read depth is extremely evident at all levels of analysis, including SNP density and CNV density, resulting in a high level of variation in these features (Figure 1). This high level of variation is further complicated by the fragmented nature of the genome assembly, with up to 50% of each scaffold composed of Ns, representing unknown bases (Figure 1). Despite these challenges, several interesting features can be observed. The first is the very low SNP density on Chr08a and Chr08b: these chromosomes are low in gene content (Figure 1c), but even in gene-rich regions the SNP density is lower than in other chromosomes (Figure 1d), and the gene-dense regions also appear to have slightly lower numbers of CNVs (Figure 1e,f). These chromosomes also have a higher than average N content, indicating gaps or missing sequences from the assembly (Figure 1a). Together, these features may indicate that these chromosomes are more difficult to assemble, possibly reflecting high levels of similarity between them, which would result in difficulty detecting SNPs between homeologs and in a lower overall SNP density. This may also be attributable to bias in the capture process or in the SNP calling on these chromosomes. The release of an improved assembly of the genome will resolve this question.

Single-nucleotide polymorphism density is greater in upland than in lowland switchgrass, and octoploid switchgrass accessions have a greater SNP density than both lowland and upland tetraploid switchgrass accessions (Figure 1), consistent with the notion of increased tolerance to polymorphism with increased ploidy. The switchgrass reference genome is based on a lowland tetraploid switchgrass individual genotype (AP13), so a larger number of variants in more diverged (i.e. upland) switchgrass accessions is expected; however, this indicates that upland octoploid switchgrass may have diverged further from the reference than upland tetraploid populations.

Switchgrass is known to have undergone large changes in habitat and phenotype in the relatively recent past (Clark

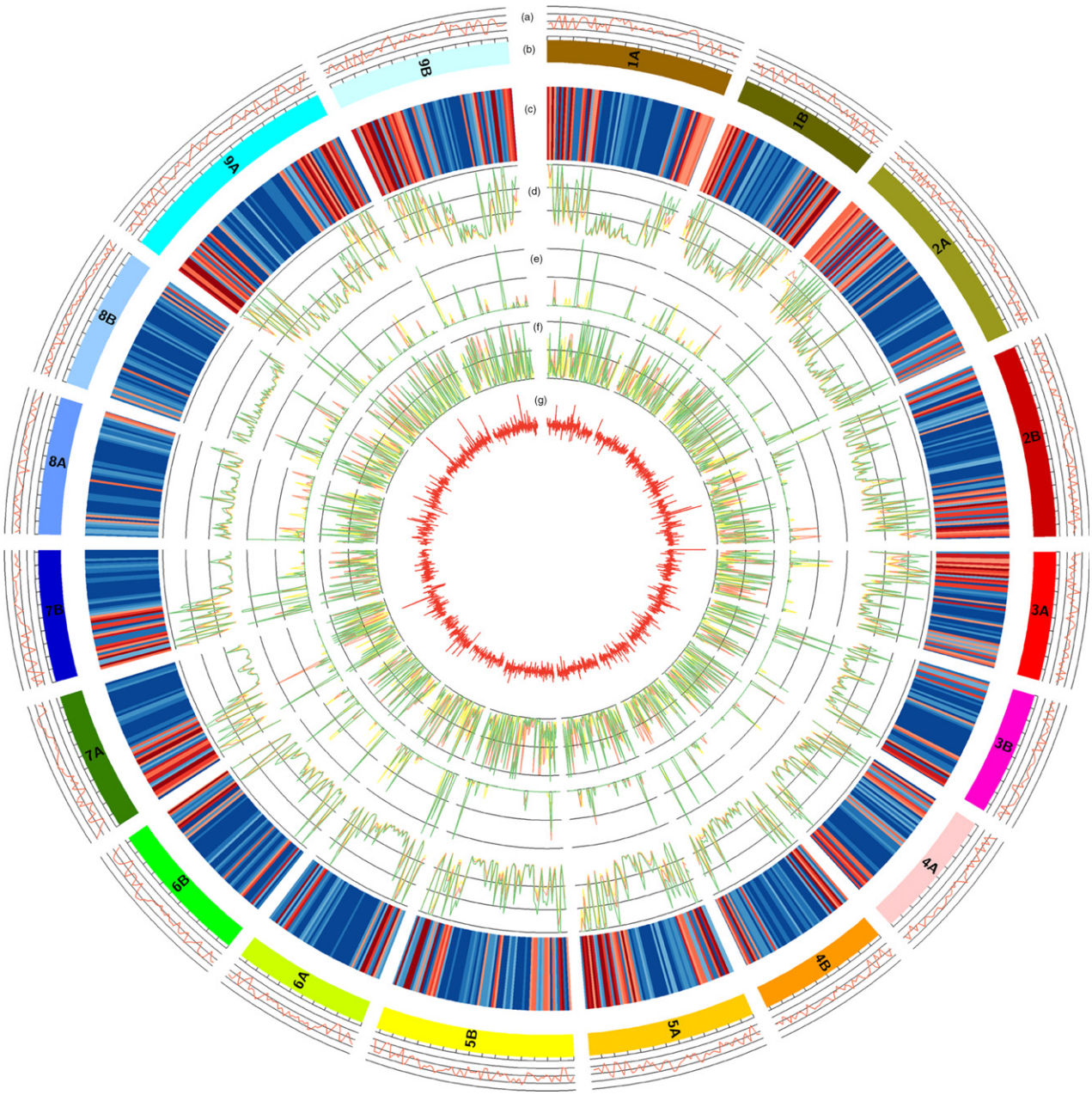

Figure 1. CIRCOS figure showing the genome-level distribution of N content, single-nucleotide polymorphisms (SNPs), copy-number variants (CNVs) and fixation index.(a) Percentage N content in 1-Mb windows, with axis spanning 0–100% N content.(b) Chromosome ideograms with labels.(c) Heat map of gene density in 1-Mb windows. Low-density regions are indicated in blue; high-density regions are indicated in red.(d) SNP frequency in 1-Mb windows. The *y*-axis varies from 500 SNPs per Mb to 30 000 SNPs per Mb. The green line indicates lowland SNP density, the yellow line indicates upland octoploid SNP density and the red line indicates upland tetraploid SNP density.(e) Up-CNV density in 1-Mb windows. The axis ranges from 0 to 24 CNVs per Mb; green, upland octoploid CNVs; red, upland tetraploid CNVs; yellow, lowland CNVs.(f) Down-CNV density in 1-Mb windows. The axis ranges from 0 to 24 CNVs per Mb; green, upland octoploid CNVs; red, upland tetraploid CNVs; yellow, lowland CNVs.(g) Measure of fixation index in 100-kb windows, the axis ranges from 0.0 to 0.4.

et al., 2001; Casler *et al.*, 2004, 2007b; Kelley *et al.*, 2006), and as switchgrass transitioned from warmer southern refuges to northern, cooler habitats, with more variable day lengths, it is possible that certain regions of the genome were under selection to adapt to these new stimuli. In an attempt to identify these regions, we calculated the fixation index of SNPs across the switchgrass genome (Figure 1). The fixation index was averaged in 100-kb windows across the genome and plotted against the 18 assembled switchgrass pseudomolecules (Figure 1). Although small regions of apparently high and low fixation index occur (Figure 1: chromosomes 7a and 9b), these are

present in regions of low gene and SNP density, where individual SNPs can influence the effect on the whole window. In general, we see only minor changes in fixation index, with no obvious regions of abnormally high or low values. This may partly result from the panel consisting largely of northern-adapted switchgrass (with the inclusion of only five southern lowland populations) or from the fragmented nature of the current switchgrass genome assembly (Figure 1). In general, however, the largely uniform level of fixation across all 18 chromosomes indicates that there does not appear to be any large regions of higher or lower fixation, suggesting a lack of selective sweeps. This supports the nature of switchgrass as a largely undomesticated natural species. It is possible that with the addition of more southern populations and cultivated populations, combined with a more complete reference genome sequence, that regions under selection would become identifiable.

Switchgrass population structure

Although switchgrass is largely identified by ecotype, previous research has shown the presence of multiple population groups present in each switchgrass ecotype, consistent with the northern migration of switchgrass following the last glacial period (Clark *et al.*, 2001; Kelley *et al.*, 2006; Zhang *et al.*, 2011b; Lu *et al.*, 2013). The populations in this panel vary in origin from naturally occurring populations (wild accessions with no human-based selection), multisite synthetic populations (amalgams of diverse genotypes from many sites), bred cultivars (generally selected for between one and four generations from wild germplasm) and natural track cultivars (wild populations that have undergone several generations of seed production without selection and have been given a cultivar name). Previous work using 700 236 GBS-derived SNPs identified between three and five upland switchgrass population groups and between two and four lowland population groups within this panel (Zhang *et al.*, 2011a; Lu *et al.*, 2013). Using the HapMapv1 set of ~1.6 million polymorphic loci, we examined population structure and identified geographic locations where populations had intermixed, and to what extent such admixture had taken place. STRUCTURE (Pritchard *et al.*, 2000) was used to calculate population group membership for population group sizes ranging from two to 10 population groups, using a subset of 48 630 randomly selected SNPs. Using the method described in Evanno *et al.* (2005), the correct population distribution was determined as being represented by seven population groups (Figure 2). Two of the groups (Lowland 4× North and Lowland 4× South) represent two lowland population groups: one from southern USA and the other from the north-eastern seaboard area (Figures 2 and 3; Table S2). Two groups (Upland 8× West and Upland 8× East) represent octoploid upland population groups, with one (Upland 8× East) clustering along the Appalachian mountain range, and the other (Upland 8× West) spread throughout central and northern Midwestern USA (Figures 2 and 3; Table S2), and with a final group (Upland 4× North) representing an upland tetraploid population group, spread throughout northern and central Midwestern USA (Figures 2 and 3; Table S2). This supports the results observed previously (Lu *et al.*, 2013), with the exception that we do not observe an Upland 8× South population of individuals with mixed ancestry that cannot be assigned to a single population group (Figures 2 and 4). Close examination indicates that one population group (unknown #2) is present in all individuals at very low levels (Figure 2; Table S2), and may represent miscalled SNPs or loci that

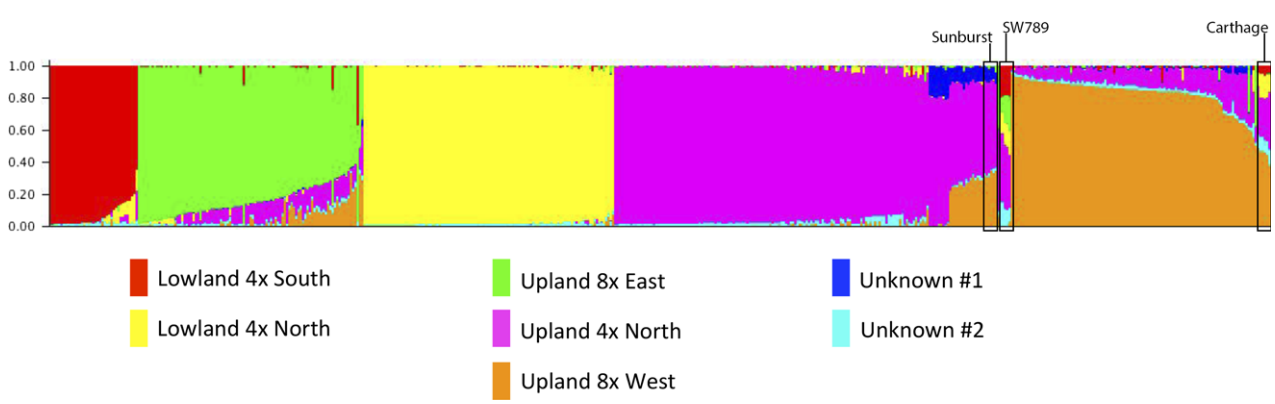

Figure 2. Switchgrass population structure. Using approximately 50 000 single-nucleotide polymorphisms (SNPs), STRUCTURE (Pritchard *et al.*, 2000) identified seven population groups, roughly depicting the geographical distribution of switchgrass ecotypes.Lowland ecotype switchgrass (red and yellow) is clearly distinct from upland switchgrass (magenta, orange and green), and shows very low levels of population group admixture. Upland switchgrass population groups (magenta, orange and green) are much more likely to exhibit population mixing within their ecotype. Three unusual populations have been highlighted: Carthage is a natural track cultivar from North Carolina exhibiting high levels of population group admixture that prevent population group assignment; SW789 is a multisite synthetic population that also exhibits high levels of population admixture, including from lowland populations; Sunburst is a bred cultivar from South Dakota that exhibits membership in an unknown group, potentially a western switchgrass population group that is not well represented in this panel.

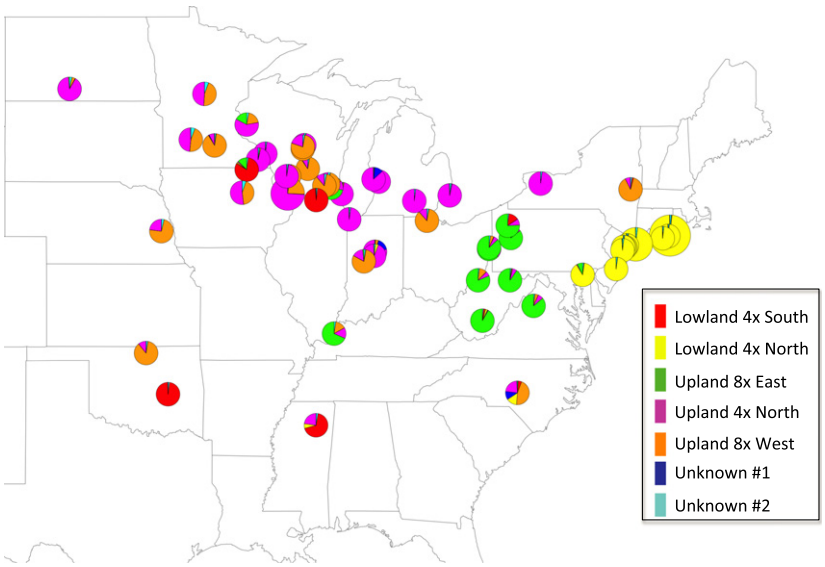

Figure 3. Geographical distribution of switchgrass populations.Pie charts indicate the population group membership of each population, as determined by STRUCTURE (Pritchard *et al.*, 2000), with the membership in each population group indicated by the area of the pie chart that the color occupies. Larger pie charts indicate populations sampled in close geographic proximity that could not be represented separately. Timber, SW789, WS4U and AP13 are not present, as there are no GPS coordinates for the origins of these cultivars.

are polymorphic in nearly all individuals, and does not represent a true population group. Another group (unknown #1) is present in a small number of populations, and is likely to represent an additional population group that was not substantially represented in the Northern Switchgrass Panel (Zhang *et al.*, 2011a). As these two population groups (unknown #1 and #2) do not represent a substantial number of individuals in this panel, all downstream analyses focus on the five major population groups (Upland 4× North, Upland 8× West, Upland 8× East, Lowland 4× North and Lowland 4× South).

Notably, STRUCTURE analysis indicates substantial mixing between the three upland population groups (Figures 2 and 3), despite variations in ploidy, whereas there has been limited mixture between the upland and lowland population groups. Despite the Lowland 4× North population groups being in relatively close proximity to the Upland 8× East population group, limited population mixing has occurred (Figures 2 and 3), indicating the presence of other factors, such as flowering time, that may keep populations reproductively isolated.

To verify the STRUCTURE population results, genetic distance was calculated using PHYLIP (Felsenstein, 1989) with all of the loci in the HapMapv1 set and a consensus neighbor-joining tree was generated. To better visualize the relationship between populations, individuals were colored based on the STRUCTURE-derived population group (Appendix S4; Figure 4). As predicted, the populations clustered into five distinct population groups (Appendix S4; Figure 4). The three upland population groups form a single large cluster, whereas the two lowland population groups form two more readily distinct clusters, consistent with the observation that genetic diversity within species decreases as latitude increases as a result of postglacial founder effects (Hewitt, 1996, 2000; Soltis *et al.*, 1997). This is also reflected in the distribution of SNPs, with Lowland 4× North switchgrass possessing the largest number of private SNPs (SNPs unique to that population group; Figure S3). Additionally, approximately 10% of all SNPs in the panel are shared among all upland population groups, and are absent from all lowland population groups (Figure S3), whereas a slightly smaller number of SNPs are shared among all population groups except the Lowland 4× South population group, which may indicate gene flow from the Lowland 4× North population group into the upland groups, or vice versa. As expected from the population structure, Upland 4× North and Upland 8× West population groups share a substantial pool of variation, whereas the Upland 4× North and Upland 8× East population groups share a much smaller pool (Figure S3).

In nearly all of the populations, all of the individuals within a population were clustered more tightly with each other than with individuals from other populations (Appendix S4; Figure 4). Notably, in addition to the tight population clustering, only three individuals (Carthage-5103, WS98-SB-6207 and WS98-SB-6210) belong to a population group that is discordant with their location on the dendogram (Appendix S4; Figure 4), although several individuals such as SW65-1309 and SW65-1310 have multiple group memberships (Appendix S4; Figures 2 and 4). WS98-SB is a mixed-lineage population that is known to contain individuals from both ecotypes and is extremely diverse genetically, whereas Carthage is a natural track cultivar derived from North Carolina switchgrass with membership in multiple population groups, and was probably created through human-mediated migration of switchgrass from Nebraska to North Carolina in the early 20th century (Appendix S4; Figures 3 and 4; Zhang *et al.*, 2011b).

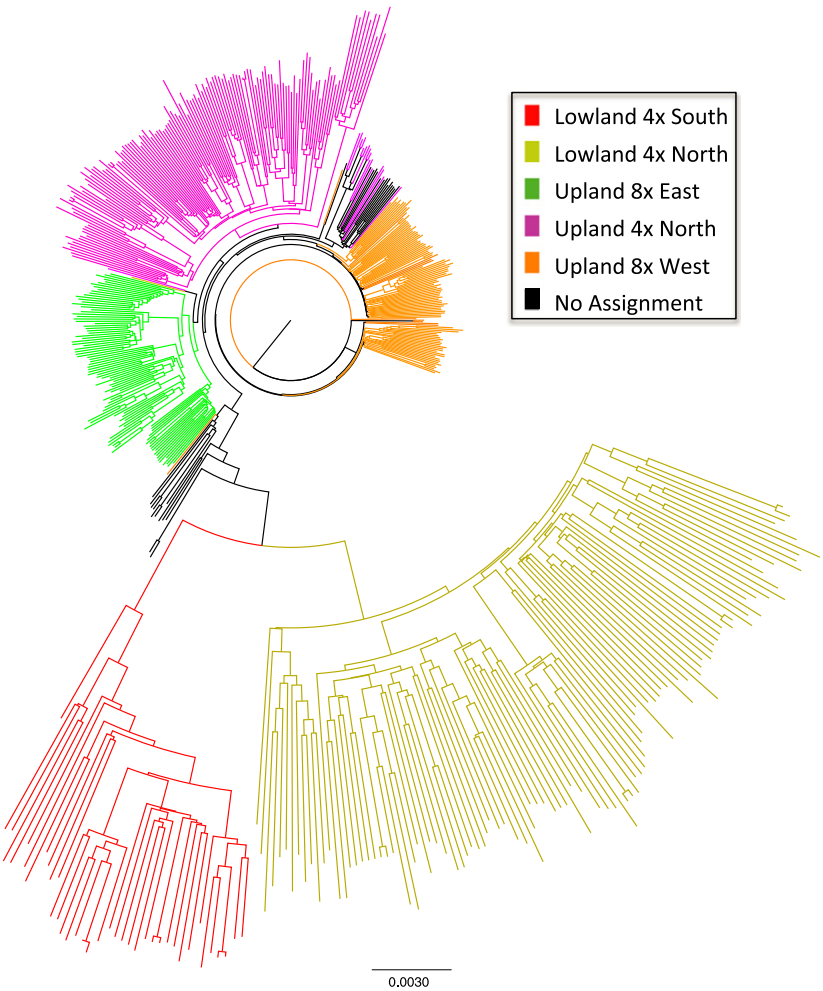

Figure 4. Dendogram representing genetic distance between all 537 individuals in this panel using all ~1.6 million high-confidence single-nucleotide polymorphisms, as calculated by the neighbor-joining method using PHYLIP (Felsenstein, 1989).Individuals have been colored according to population group membership, as shown in Figure 2. In order to assign population group membership, an individual had to have more than 50% membership to that group. Individuals that could not be assigned are colored black. Lowland switchgrass is visibly distinct and distant from upland switchgrass, which are less genetically distant from each other.

Noticeably, the Upland 8× West octoploid population group appears to exhibit much less variation within its populations than the other upland population groups. Although other upland populations are tightly clustered and distinguishable within their population groups, populations occasionally appear to be mixed and indistinguishable within the Upland 8× West population group (Appendix S4; Figure 4). The Upland 8× West population group also possesses the largest number of private SNPs among the upland populations (Figure S3). This suggests that although the Upland 8× West population is genetically distinct from other upland populations groups, there is a limited level of variation between members within the population group, which may indicate the presence of a population bottleneck in the recent past.

Sequence variation in flowering time genes follows ecotype divisions

One of the features that most distinguishes upland and lowland switchgrass is flowering time, which can vary by as much as 6 weeks (Casler *et al.*, 2004, 2007b). Once flowering begins, the accumulation of biomass in the aboveground portion of switchgrass stems declines abruptly (VanEsbroeck *et al.*, 1997), allowing lowland switchgrass, with its later flowering, to accumulate more biomass, even when grown in northern climates (Lemus *et al.*, 2002). Lowland switchgrass often overwinters poorly, however, and suffers stand failure in more northern climates, unlike cold-adapted upland accessions (Casler *et al.*, 2004, 2007b). Thus, altering flowering time in upland switchgrass may provide a path to increase biomass accumulation.

Although the genetic mechanisms controlling flowering time in switchgrass have not yet been elucidated, data exists for these pathways in other grasses (Turck *et al.*, 2008; Colasanti and Coneva, 2009; Jackson, 2009; Valverde, 2011; Xu *et al.*, 2012b; Itoh and Izawa, 2013). We focused on the *Gigantea* (*GI*)–*Constans* (*CO*)–*Flowering locus T* (*FT*) photoperiod-sensitive flowering time pathway, a pathway present in *Arabidopsis thaliana* as well as in rice and *Sorghum* (Valverde, 2011; Andres and Coupland, 2012). In short-day plants such as rice, *Heading Date 1* (*HD1*), an ortholog of *CO*, activates flowering in short days through

the expression of the *FT*-like genes *HD3a* and *RFT1*, whereas in long-day plants such as barley and Arabidopsis, *CO* triggers FT expression during long-day conditions (Valverde, 2011; Andres and Coupland, 2012). *Early heading date 1* (*Ehd1*) acts downstream of *HD1*/*CO*, and also activates the expression of *HD3A* and *RFT1* under short-day conditions in rice, and can act independently of *CO,* but is repressed by *Ghd7* in non-inductive long-day conditions (Doi *et al.*, 2004). In switchgrass, the negative regulation of the expression of *CO* through the transgenic expression of Arabidopsis *LONG VEGETATIVE PHASE 1* (*AtLOV1*) has been shown to delay flowering time (Xu *et al.*, 2012a), indicating the presence of the *Gigantea–CO–FT* floral initiation pathway.

Switchgrass homologs of *CO* and *EHD1* were identified in v1.1 of the *P. virgatum* genome assembly (Appendix S5; Figure S4). An examination of SNPs present in Pavir.Da01464 (*CO* homolog) across the panel revealed three polymorphic loci containing SNPs associated with the ecotype (Figure 5a). The first polymorphic allele is predicted to convert the encoded CO protein at amino acid 4 from an asparagine to a lysine (N4K), and was present in 43 upland individuals (32 tetraploid and 11 octoploid) from 14 populations. The second polymorphic allele is predicted to convert a serine to threonine at amino acid position 255 (S255T; Figure 5a), and was present in 63 upland individuals (31 tetraploid and 31 octoploid) from 23 populations. Notably, both of these polymorphisms, found exclusively in upland ecotypes, are predicted by SIFT (Ng and Henikoff, 2001) to be poorly tolerated, although the alignment quality for N4K was below the certainty threshold. The polymorphism restricted to lowland ecotype accessions (19 total) is predicted to convert a methionine to threonine at amino acid 292 (M292T), and is predicted by SIFT to be tolerated (Figure 5a). No similar pattern of polymorphisms was detected in the putative homeolog Pavir.Db01632 (Appendix S3).

A similar pattern of polymorphisms is present in Pavir.Ia04737, the *EHD1* homolog (Figure 5b). One allele is present in 46 upland individuals (32 octoploid and 14 tetraploid) from 17 populations, and causes an isoleucine to threonine substitution at amino acid position 191 (I191T), predicted by SIFT to be poorly tolerated (Figure 5b). The second locus is present in 32 lowland individuals from seven populations, as well as five upland individuals (three octoploid and two tetraploid) from five populations, and results in an isoleucine to threonine change at amino acid position 130 (I130T), predicted by SIFT to be tolerated

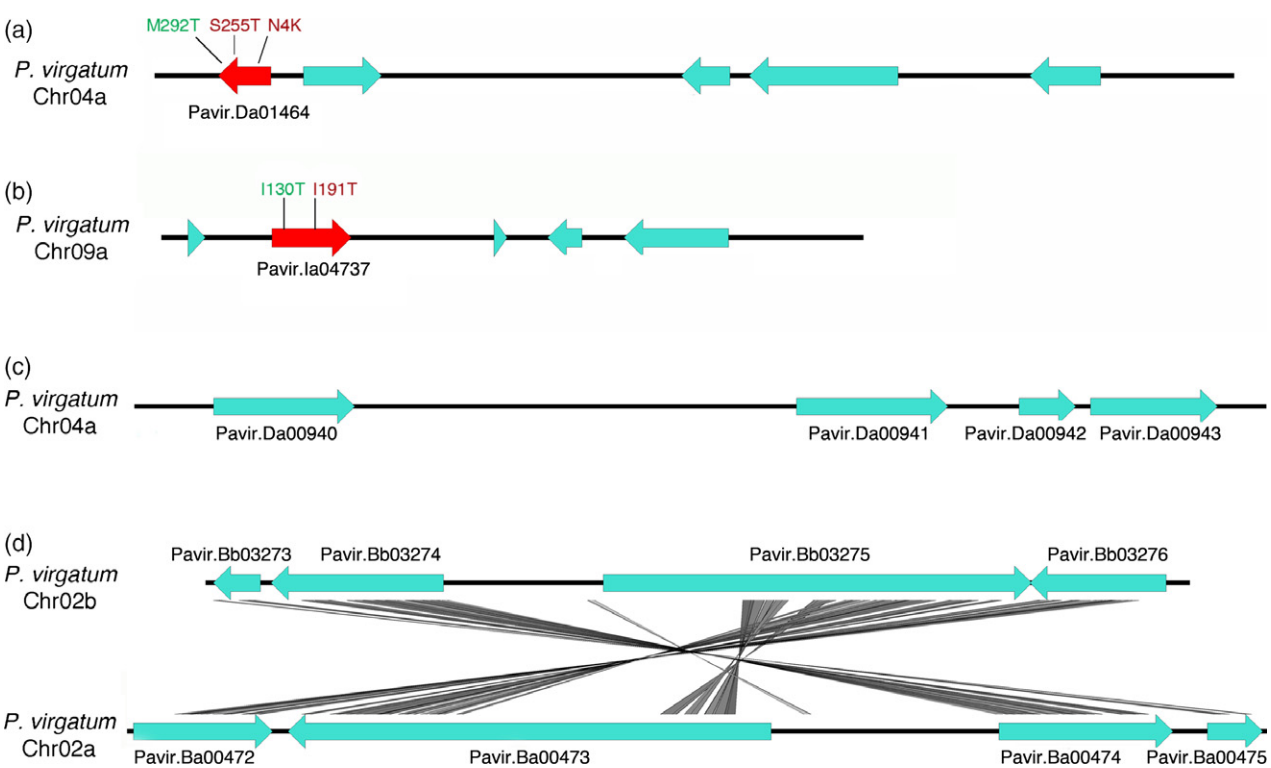

Figure 5. Ecotype-restricted single-nucleotide polymorphisms (SNPs) and copy-number variant (CNV) clusters from the Northern Switchgrass Panel.(a) The switchgrass homolog of *Constans,* Pavir.Da01464, with northern ecotype-restricted non-synonymous SNPs (red text), and lowland ecotype restricted non-synonymous SNPs (green text).(b) The switchgrass homolog of *Early heading date 1,* Pavir.Ia04737, with northern ecotype-restricted non-synonymous SNPs (red text), and lowland ecotype-restricted non-synonymous SNPs (green text).(c) Upland-restricted down-CNV cluster located on chromosome 4a. No homeologous region was detected for this cluster.(d) Upland-restricted down-CNV cluster located on chromosome 2b, and the homeologous region on chromosome 2a.

(Figure 5b). In Pavir.Ib00832, the *EHD1* homeolog, two lowland-specific polymorphic alleles [N138I and I190V; representing 16 individuals (nine populations) and 10 individuals (seven populations), respectively] are predicted by SIFT to be tolerated, although both occur at much lower frequencies (Appendix S3).

These data suggest that variation in genes that govern multiple stages of the flowering time pathway are present in switchgrass, and can be associated, at least in this panel, with the ecotype. Although each polymorphism is not detected in all members of an ecotype or population, it is possible that not all alleles are being captured, especially in the case of the octoploid individuals, where we have enough coverage across the panel to identify the alleles present (Griffin *et al.*, 2011), but not to resolve them with full accuracy in each individual (Uitdewilligen *et al.*, 2013). It has been established that switchgrass flowering time variation occurs on a north–south gradient (Casler *et al.*, 2004, 2007b), so it is possible that different flowering time patterns are a result of different mutations.

Copy number variation

Copy number variation is a form of structural variation in which an individual has an increase or reduction in the number of copies of a gene or locus, relative to another individual or population. Both CNVs and the more extreme form, PAVs have been identified in a wide range of plant species, either through comparative genome hybridizations or through read-depth methods (Swanson-Wagner *et al.*, 2010; Diaz *et al.*, 2012; Winzer *et al.*, 2012; Cook *et al.*, 2014; Evans *et al.*, 2014), with a subset of CNVs and PAVs associated with a phenotype, including biotic defense (Cook *et al.*, 2012), flowering time and vernalization (Diaz *et al.*, 2012), the biosynthesis of pharmacoactive compounds (Winzer *et al.*, 2012), aluminum tolerance in *Zea mays* (maize; Maron *et al.*, 2013), and submergence tolerance in rice (Xu *et al.*, 2006), indicating that CNVs can play a strong role in the phenotype of plants. To detect CNVs and PAVs in the Northern Switchgrass Panel, we calculated normalized read depths in each population relative to the Kanlow population, chosen for being genetically close to the Alamo population from which the AP13 individual used for the creation of the reference genome was selected (Zhang *et al.*, 2011a). After filtering for significant CNVs, we identified 17 228 up-CNVs, 112 630 down-CNVs and 14 430 PAVs (Appendixes S6 and S7; Table S3). To assess the validity of these CNVs, we used the CNVs to calculate genetic distance using PHYLIP (Felsenstein, 1989), treating each CNV as a discrete genetic marker. A genetic distance dendogram was assembled and labeled in the same manner as the SNP dendogram (Figure 6), and was highly concordant with the SNP dendogram, illustrating the high reliability of the two data sets.

In the overall panel, categories of genes enriched in CNVs were similar to those determined previously (Evans *et al.*, 2014). With the availability of 537 individuals from 45 upland and 21 lowland populations, we examined CNVs to identify patterns that may be associated with phenotype variation at the ecotype level. To Identify upland-specific CNVs, we identified genes that had CNVs in at least 40 upland populations and no CNVs in any lowland population, resulting in a set of 62 upland-specific CNVs (Appendix S7). Interestingly, all of these CNVs were down-CNVs, as compared with the reference Kanlow population, potentially indicating an ecotype-specific bias as the probes were designed from primarily AP13 sequences, a lowland tetraploid individual. Predicted gene annotations indicated the presence of several ecotype-specific transposable element-related CNVs, which may indicate amplification of those elements in lowland switchgrass after ecotype divergence. Leucine rich repeat-containing genes and pentatricopeptide repeat-containing genes are also represented, which have been identified to be highly divergent in other plant species (Zmienko *et al.*, 2014). Many of the remaining genes lack annotation, but potential patterns emerge from the annotated genes (Appendix S7). The first is the presence of multiple genes involved in biosynthetic pathways (Appendix S7), including the anthocyanin biosynthesis pathway (Pavir.Da00943), the terpene synthase pathway (Pavir.Ab01716) and the isoquinoline biosynthetic pathway (Pavir.Da00940) (Appendix S7). Anthocyanins are involved in the protection of tissues from photoinhibition during high photon flux: this would be especially necessary for lowland ecotype switchgrass, with habitat ranges through southern USA. The down-CNVs detected in upland switchgrass may be the result of a larger number of copies in the reference population (Kanlow), and reflect a lower photoinhibition protection requirement for upland switchgrass. Products of the terpene and isoquinoline biosynthetic pathway often function as anti-herbivory and anti-microbial compounds (Isman, 2000; Stermitz *et al.*, 2000). The presence of potential members of this biosynthetic pathway among the down-CNVs may indicate that the compounds that the gene products synthesize are in less demand among upland switchgrass, possibly as a result of different biotic threats in their ecosystem.

The switchgrass genome is highly dynamic, having undergone recent whole-genome duplication in addition to the high levels of gene flow caused by being an obligate out-crossing species (Zalapa *et al.*, 2011; Zhang *et al.*, 2011a). To attempt to visualize regions of the genome that have been especially active for structural variation, or that have undergone an ancestral CNV that has been retained, we identified sequential CNVs in all populations under analysis. Sequential CNVs were defined as CNVs located in an unbroken sequence along a chromosome, with the

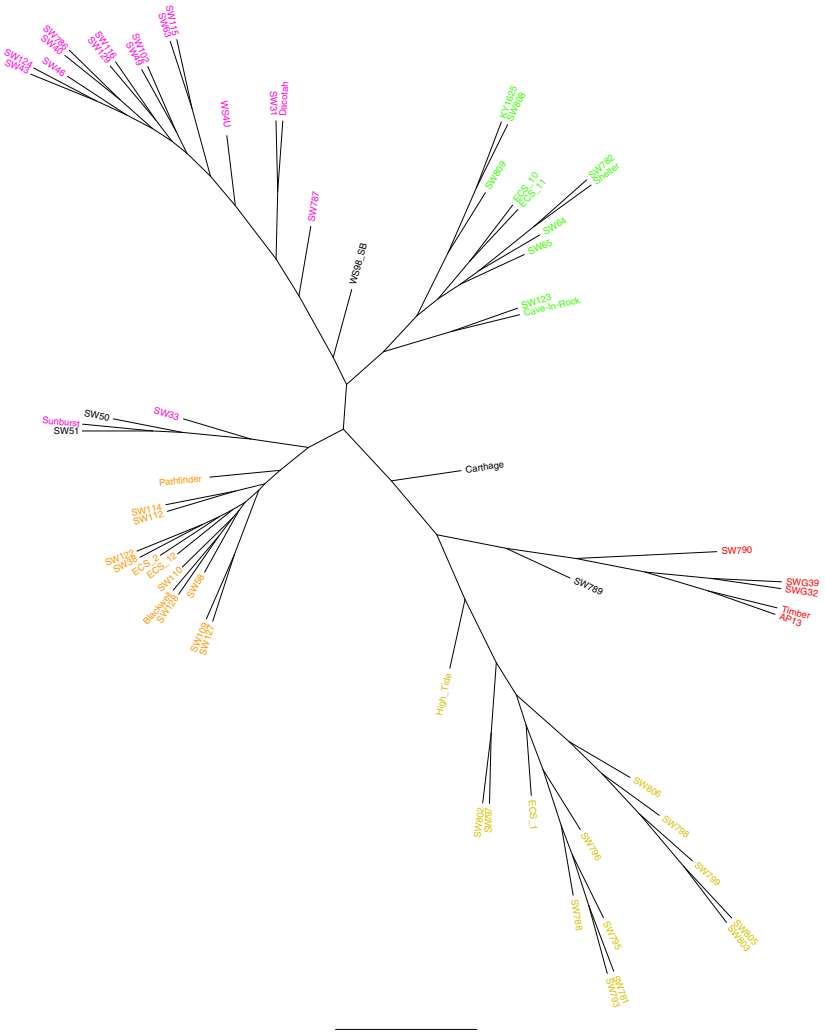

Figure 6. Genetic distance dendogram created using copy-number variants as genetic markers.Populations were assigned a value for each gene – up-copy number variant (up-CNV), down-CNV or neither – and these values were used to calculate genetic distance using the neighbor-joining function in PHYLIP (Felsenstein, 1989). Populations were color-coded based on population group, as shown in Figure 2. With the exception of the Cave in Rock population, all populations group together with other members of their population groups, indicating that the CNVs detected are reliably associated with the population.

caveat that genes for which we had no coverage or were repetitive elements were excluded from this analysis. In total, we identified 9833 CNV clusters (two or more up- or down-CNVs in sequence). It is likely that additional CNV clusters exist, but currently 41 128 of the 98 007 predicted genes reside on unanchored contigs, which makes clustering impossible for those genes. We then evaluated CNV clusters that were present in multiple populations. Two down-CNV clusters were detected on chromosomes 2b and 4a that had substantial representation among the populations (Figures 5c,d and S5; Table S4). The cluster on chromosome 4a (Figure 5c) is of particular note, in that 43 of the 46 upland populations possessed this cluster of down-CNVs, whereas no lowland populations possessed this cluster (Figure S5; Table S4). Genes in this cluster are annotated as two serine–threonine protein kinases (Pavir.Da00941, Pavir.Da00942), a D-lactate dehydrogenase (Pavir.Da00940) and a UDP-glucosyl transferase (Pavir.Da00943) (Appendix S8). As mentioned previously, Pavir.Da00940 and Pavir.Da00943 are potentially involved in photoinhibition protection and defense against biotic stress. As we were unable to detect a co-linear segment of the genome containing these genes, this may provide additional evidence that there is heritable variation in the response to stress across ecotypes.

The down-CNV cluster on chromosome 2b (Figure 5d) is similar though not as pronounced, and is present in 30 of the 46 upland switchgrass populations, but none of the lowland populations (Appendix S8; Figure S5). Interestingly, unlike the cluster on chromosome 4a, all four of these genes appear to have co-linear homeologs on chromosome 2b (Figure 5d), and none of these homeologous genes were detected as CNVs in our analysis (Appendix S7). This may indicate a case where the loss of a section of one subgenome does not result in an extreme phenotype because of the presence of a duplicate copy that can complement the loss; alternatively, the loss of one copy may result in a beneficial phenotype in one ecotype while being detrimental in the other ecotype. This may also represent a duplication of this region in the lowland

population that was used as a reference – additional sequencing will be required to determine which of these hypotheses is correct. There is no apparent functional relationship between the genes within each cluster (Appendix S8), but their close physical proximity and ecotype specificity indicate the cluster was probably lost as a unit. Although additional sequencing will be required to determine the extent of these CNVs, and whether they reflect duplications in the reference ecotype or deletions in the upland ecotype, the presence of large numbers of CNVs in close proximity indicates that CNVs may play an important role in the adaptation of each ecotype to its current environment.

DISCUSSION

We demonstrate here the ability to use exome capture sequencing data on multiple, divergent switchgrass populations, and the ability to distinguish those populations at both the sequence and the gene level. We identify the presence of polymorphic loci that can be evaluated across all individuals in all populations, and these loci can be used to both identify population-associated SNPs, but also to accurately identify population membership of individual samples. Although the coverage of this data set is of sufficient depth and quality to establish consistent population membership and identify ecotype-restricted SNPs, because of the difficulty in resolving low-frequency heterozygous SNPs in polyploid organisms (especially octoploids), it is likely that some variants have been overlooked, especially low-frequency alleles in octoploid individuals. We also demonstrate the ability to identify CNVs at a population level, and demonstrate the accuracy of these CNVs, and that in many cases these CNVs are associated with specific populations and population groups. Several of these CNVs occur in series, and appear to be missing in all upland switchgrass, which may indicate large-scale genomic changes. These results will provide a valuable resource to switchgrass researchers and breeders, and provide a foundation for future evolutionary and population-level analyses. Additionally, this material provides a wide pool of northern switchgrass alleles that can be compared with alleles from other switchgrass habitats, which will allow for insights into the history and adaptation of switchgrass in North America.

EXPERIMENTAL PROCEDURES

Switchgrass population description and DNA isolation

This Northern Switchgrass Panel consists of 66 populations previously described by Lu *et al.* (2013). The panel is composed of 66 populations grown from seed at the USDA-ARS Dairy Forage Research Center glasshouse at the University of Wisconsin, Madison. Ten vegetatively propagated clones were then planted in Ithaca, NY, in 2008 (Table 1). DNA was extracted from freeze-dried leaf tissue using the cetyltrimethylammonium bromide extraction protocol (Saghaimaroof *et al.*, 1984).

Exome capture sequencing, read alignment and analysis

Exome capture sequencing was performed on the individuals listed in Appendix S1, using the established Roche-NimbleGen protocol for SeqCap EZ Developer library preparation and using the Roche-NimbleGen probe set '120911_Switchgrass_GLBRC_R_EZ_HX1' (Evans *et al.*, 2014), as described in Mascher *et al.* (2013), with the exception that Kapa biosystem reagents were used for library preparation (https://www.kapabiosystems.com/, Kapa Biosystems, Wilmington, MA, USA). All capture and sequencing steps were performed by the Joint Genome Institute in Walnut Creek, California, USA. A total of 537 individuals were subjected to capture and sequenced on the Illumina HiSeq 2000 platform, generating 150 nucleotide paired-end reads. Reads underwent initial quality control using FASTQC 0.10.0 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). CUTADAPT v1.1 (http://code.google.com/p/cutadapt) was used to remove PCR primers and adapter sequences, and any base with a quality score less than 20 was trimmed. Reads 35 bases or shorter after trimming were discarded.

After filtering and trimming, reads were aligned with the hard-masked *P. virgatum* v1.1 reference genome (http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Pvirgatum) using BOWTIE 0.12.7 (Langmead *et al.*, 2009). Unique alignments were required, and only a single mismatched nucleotide was allowed in the first 35 bases of the read. Multiplexing of 12 samples per lane of the flow cell was used, generating approximately 3.2 Gb of sequence per sample. In order to improve processing efficiency, unanchored contigs were assigned to numbered scaffolds (ChrUn1–ChrUn15) using custom PERL scripts.

Single-nucleotide polymorphism detection and analysis

Read alignments that met the alignment criteria were processed using the index, sort, merge and mpileup functions of SAMTOOLS 0.1.18 (Li *et al.*, 2009). The –BD and –C 0 flags were used for the mpileup command, and index, sort and merge were all used with the default parameters. Pileup files were filtered with custom PERL scripts to identify polymorphic positions. To term a position polymorphic, at least two individuals must have a read depth of at least five reads, with none of those reads being polymorphic at that position, and at least two other individuals must have a read depth of at least five reads, and at least two reads or 5% of the reads, whichever is greater, must be polymorphic at that position. To generate the SNP set used for phylogenetic and population structure analysis, the initial set was further filtered. Any locus with missing data from any individual was removed from this set, a maximum of 10 individuals were allowed to have insufficient depth of coverage (below five) or a minor allele (number of alternate allele reads <5% of all reads at that position) before the locus was discarded. All positions used for genetic distance calculations and population structure analyses were required to be bi-allelic. SNP annotation was performed using ANNOVAR (Wang *et al.*, 2010) using gene-based annotation. All options were set to their default parameters.

Comparison with genotyping by sequencing-predicted variants

FASTA files containing the sequences used to identify polymorphisms in Lu *et al.* (2013) were aligned with the unmasked switchgrass genome using the same criteria as exome capture reads. SNPs were identified from pileup in the same manner as the exome capture sequence, except that the requirement for coverage depth was removed.

Population structure

Population structure was determined using STRUCTURE (Pritchard *et al.*, 2000). STRUCTURE handles a maximum of 100 million genotypes, so 48 630 SNPs (representing approximately 3% of the total high-confidence SNP set) were randomly selected from the SNP matrix using the PERL rand() function. Based on previous information on switchgrass population structure (Zhang *et al.*, 2011a; Lu *et al.*, 2013), STRUCTURE was run using estimated numbers of subpopulations (K-values) ranging between 2 and 10. Each population size analysis was replicated 10 times, and each analysis involved 20 000 burn-in iterations and 10 000 Monte Carlo iterations. The inferalpha and computeprobs values were set to 1, and all others were left at default values. Analysis was performed using the admixture model, with no prior population knowledge, and to determine the most accurate number of population groups, the method detailed in Evanno *et al.* (2005) was used. Phylogenetic analysis was performed using PHYLIP 3.695 (Felsenstein, 1989). Genetic distances were calculated using the gendist function with default parameters. Bootstrapping was performed using seqboot (100 replicates), and a consensus tree was generated using consense, with default parameters. Dendograms were visualized using FIGTREE 1.40 (http://tree.bio.ed.ac.uk/software/figtree). Colors were applied based on population group, determined in STRUCTURE, with an individual colored according to the majority population (membership $>$ 50%). Individuals with no majority population group were colored black. The fixation index was calculated as described by Holsinger and Weir (2009).

Switchgrass flowering time genes

Putative switchgrass homologs of *Hd1* and *CO* were identified by aligning the nucleotide and protein sequence of rice *Hd1* (Yano *et al.*, 2000) and *Sorghum CO* (Yang *et al.*, 2014) with the *P. virgatum* v1.1 genome using BLAST (Altschul *et al.*, 1990) with default parameters and an *E*-value cut-off of −20. The top two matches were selected on the rationale that they had the highest alignment scores and were located on potentially homoeologous regions of chromosomes 4a and 4b. Putative switchgrass homologs of *EHD1* were identified by aligning the nucleotide and protein sequence of rice *EHD1* (Doi *et al.*, 2004) and *Sorghum EHD1* (Yang *et al.*, 2014) with the *P. virgatum* v1.1 genome using BLAST, with default parameters and an *E*-value cut-off of −20. The top two matches were chosen based on score, and were located on potentially homoeologous regions of chromosomes 9a and 9b.

Co-linearity analyses between switchgrass, *Setaria* (Zhang *et al.*, 2012), *Sorghum* (Paterson *et al.*, 2009) and rice (Kawahara *et al.*, 2013) were performed using the tblastx function of BLAST (Altschul *et al.*, 1990), with default parameters. Co-linearity visualization was plotted using EASYFIG 2.1 (Sullivan *et al.*, 2011), with BLAST visualization cut-offs set to a minimum length of 50 and a maximum *E*-value of 10^{-7}. Other options were set to defaults.

Amino acid tolerance was predicted using SIFT (Ng and Henikoff, 2001), with default parameters.

Detection of copy-number variation and dendogram construction

Structural variation was identified by first determining the number of reads mapping to each predicted gene using the htseq-count function of HTSEQ 0.6.1 (Anders *et al.*, 2014). Alignments were processed in BAM format, sorted by position, with stranded set to no. Resulting counts were normalized based on read number using EDGER 3.6.8 (Robinson *et al.*, 2010). Following the methods outlined in Anders *et al.* (2013), all individuals in a population were pooled, and genes with normalized counts per million value of less than one across the panel were removed from the analysis. The ratio for gene coverage depth was determined using the Kanlow population as a comparison, with the rationale that this population was the closest to the reference population (Alamo) and should minimize bias. CNVs were identified by dividing the normalized depth of coverage for each gene by the normalized depth of coverage for that gene in the Kanlow population and taking the $\log_2$ of that value. A gene was identified as an up-CNV if the resulting $\log_2$ value was above the 99th percentile value, had $P < 0.05$, and a false discovery rate (FDR) value of <0.05. A gene was identified as a down-CNV if the $\log_2$ value was lower than the negative of the 99th percentile value of the up-CNV cut-off for that gene. Presence/absence variants were identified as genes that were down-CNVs with at least 5$\times$ depth of coverage in the Kanlow population and zero coverage in all individuals of the comparison population. CNV clusters were identified by examining the physical locations of genes identified as CNVs along the switchgrass chromosomes and identifying sequential genes that were either up- or down-CNVs. Genes with no coverage in any individual were also removed from the analysis. A genetic distance dendogram was created using PHYLIP 3.695 (Felsenstein, 1989), with each gene determined to be a CNV treated as an up-CNV, down-CNV or no-CNV allele. Genetic distance was calculated using the gendist function, bootstrapping was performed using seqboot (1000 replicates) and consensus tree construction was created using consense. Subsequent trees were visualized using FIGTREE 1.40 (http://tree.bio.ed.ac.uk/software/figtree). Colors were applied based on population group determined in STRUCTURE, with a population colored according to the majority population group for that population (membership $>$ 50%). Populations with no majority population group were colored black.

Data access

All exome capture reads are available in the National Center for Biotechnology Information under project accession number PRJNA280418. Because of the large size of the files, the following data are available at the Dryad Digital Repository under doi 10.1111/tpj.13041 (to be released upon publication): unfiltered SNP matrix; filtered SNP matrix; annotation of filtered SNP matrix; annotation of exonic SNPs in filtered SNP matrix; list of up-CNVs; list of down-CNVs; genetic distance dendogram in Newick format; unmasked switchgrass genome sequence (v1.1) with ChrUns; hardmasked switchgrass genome sequence (v1.1) with ChrUns; contig coordinates file for ChrUns; gene annotation list for switchgrass genome (v1.1) with only representative sequences; and ChrUn data.

ACKNOWLEDGEMENTS

This work was funded by the DOE Great Lakes Bioenergy Research Center (DOE BER Office of Science DE-FC02-07ER64494). The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under contract no. DE-AC02-05CH11231.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Stacked bar graphs representing the distribution of single-nucleotide polymorphisms (SNPs) in the high-confidence SNP set across genomic features in various population groupings.

Figure S2. Stacked bar graphs representing the predicted function of single-nucleotide polymorphisms (SNPs) in multiple population groupings.

Figure S3. Venn diagram representing the SNPs from the high-confidence single-nucleotide polymorphism (SNP) set, separated by population group.

Figure S4. Syntenic regions near Pavir.Da01464, the switchgrass homolog of *CO*, and Pavir.Ia04737, the switchgrass homolog of *EHD1*.

Figure S5. Distribution of chromosome 4a and chromosome 2b down-copy number variant (CNV) clusters, as described in Figure 5(c,d).

Table S1. Aggregate read depth coverage information for the panel.

Table S2. Population group membership of switchgrass individuals, as determined by STRUCTURE.

Table S3. Distribution of copy-number variants in switchgrass populations.

Table S4. Populations containing the respective copy-number variant clusters, and their population group membership and ploidy.

Appendix S1. Number of reads and alignment statistics for switchgrass exome capture populations (all samples except AP13 were present in Lu *et al.* 2013).

Appendix S2. Workflow and results of comparison of polymorphic sequences from Lu *et al.* (2013) with polymorphism detected through exome capture sequencing and alignment.

Appendix S3. Matrix containing allele calls generated for all individuals.

Appendix S4. High-resolution searchable version of Figure 4.

Appendix S5. Identification of switchgrass homologs of *Constans* and *Early heading date 1*.

Appendix S6. List of copy-number variants for all populations.

Appendix S7. Predicted annotations for upland-specific copy number variants.

Appendix S8. Predicted annotations for genes in two upland-restricted down-copy number variant clusters.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Anders, S., McCarthy, D.J., Chen, Y.S., Okoniewski, M., Smyth, G.K., Huber, W. and Robinson, M.D. (2013) Count-based differential expression analysis of RNA sequencing data using R and bioconductor. *Nat. Protoc.* **8**, 1765–1786.

Anders, S., Pyl, P.T. and Huber, W. (2014) HTSeq – a python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.

Andres, F. and Coupland, G. (2012) The genetic basis of flowering responses to seasonal cues. *Nat. Rev. Genet.* **13**, 627–639.

Casler, M.D., Vogel, K.P., Taliaferro, C.M. and Wynia, R.L. (2004) Latitudinal adaptation of switchgrass populations. *Crop Sci.* **44**, 293–303.

Casler, M.D., Stendal, C.A., Kapich, L. and Vogel, K.P. (2007a) Genetic diversity, plant adaptation regions, and gene pools for switchgrass. *Crop Sci.* **47**, 2261–2273.

Casler, M.D., Vogel, K.P., Taliaferro, C.M., Ehlke, N.J., Berdahl, J.D., Brummer, E.C., Kallenbach, R.L., West, C.P. and Mitchell, R.B. (2007b) Latitudinal and longitudinal adaptation of switchgrass populations. *Crop Sci.* **47**, 2249–2260.

Casler, M.D., Tobias, C.M., Kaeppler, S.M., Buell, C.R., Wang, Z.Y., Cao, P.J., Schmutz, J. and Ronald, P. (2011) The switchgrass genome: tools and strategies. *Plant Genome*, **4**, 273–282.

Clark, J.S., Grimm, E.C., Lynch, J. and Mueller, P.G. (2001) Effects of holocene climate change on the C-4 grassland/woodland boundary in the Northern Plains, USA. *Ecology*, **82**, 620–636.

Clewell, A. and Rieger, J.P. (1997) What practitioners need from restoration ecologists. *Restor. Ecol.* **5**, 350–354.

Colasanti, J. and Coneva, V. (2009) Mechanisms of floral induction in grasses: something borrowed, something new. *Plant Physiol.* **149**, 56–62.

Cook, D.E., Lee, T.G., Guo, X. *et al.* (2012) Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. *Science*, **338**, 1206–1209.

Cook, D.E., Bayless, A.M., Wang, K., Guo, X.L., Song, Q.J., Jiang, J.M. and Bent, A.F. (2014) Distinct copy number, coding sequence, and locus methylation patterns underlie Rhg1-mediated soybean resistance to soybean cyst nematode. *Plant Physiol.* **165**, 630–647.

Costich, D.E., Friebe, B., Sheehan, M.J., Casler, M.D. and Buckler, E.S. (2010) Genome-size variation in switchgrass (*Panicum virgatum*): flow cytometry and cytology reveal rampant aneuploidy. *Plant Genome*, **3**, 130–141.

Diaz, A., Zikhali, M., Turner, A.S., Isaac, P. and Laurie, D.A. (2012) Copy number variation affecting the photoperiod-B1 and vernalization-A1 genes Is associated with altered flowering time in wheat (*Triticum aestivum*). *PLoS ONE*, **7**, e33234.

Doi, K., Izawa, T., Fuse, T., Yamanouchi, U., Kubo, T., Shimatani, Z., Yano, M. and Yoshimura, A. (2004) Ehd1, a B-type response regulator in rice, confers short-day promotion of flowering and controls FT-like gene expression independently of Hd1 I. *Gene Dev.* **18**, 926–936.

Evanno, G., Regnaut, S. and Goudet, J. (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **14**, 2611–2620.

Evans, J., Kim, J., Childs, K.L. *et al.* (2014) Nucleotide polymorphism and copy number variant detection using exome capture and next-generation sequencing in the polyploid grass *Panicum virgatum*. *Plant J.* **79**, 993–1008.

Felsenstein, J. (1989) PHYLIP – phylogeny inference package (version 3.2). *Cladistics*, **5**, 164–166.

Griffin, P.C., Robin, C. and Hoffmann, A.A. (2011) A next-generation sequencing method for overcoming the multiple gene copy problem in polyploid phylogenetics, applied to Poa grasses. *BMC Biol.* **9**, 19.

Henry, I.M., Nagalakshmi, U., Lieberman, M.C. *et al.* (2014) Efficient genome-wide detection and cataloging of EMS-induced mutations using exome capture and next-generation sequencing. *Plant Cell*, **26**, 1382–1397.

Hewitt, G.M. (1996) Some genetic consequences of ice ages, and their role in divergence and speciation. *Biol. J. Linn. Soc.* **58**, 247–276.

Hewitt, G. (2000) The genetic legacy of the Quaternary ice ages. *Nature*, **405**, 907–913.

Holsinger, K.E. and Weir, B.S. (2009) Fundamental concepts in genetics genetics in geographically structured populations: defining, estimating and interpreting F-St. *Nat. Rev. Genet.* **10**, 639–650.

Hopkins, A.A., Taliaferro, C.M., Murphy, C.D. and Christian, D. (1996) Chromosome number and nuclear DNA content of several switchgrass populations. *Crop Sci.* **36**, 1192–1195.

Hultquist, S.J., Vogel, K.P., Lee, D.J., Arumuganathan, K. and Kaeppler, S. (1996) Chloroplast DNA and nuclear DNA content variations among cultivars of switchgrass, *Panicum virgatum* L. *Crop Sci.* **36**, 1049–1052.

Isman, M.B. (2000) Plant essential oils for pest and disease management. *Crop Prot.* **19**, 603–608.

Itoh, H. and Izawa, T. (2013) The coincidence of critical day length recognition for florigen gene expression and floral transition under long-day conditions in rice. *Mol. Plant.* **6**, 635–649.

Jackson, S.D. (2009) Plant responses to photoperiod. *New Phytol.* **181**, 517–531.

Kawahara, Y., de la Bastide, M., Hamilton, J.P. *et al.* (2013) Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice*, **6**, 4.

Kelley, D.W., Brachfeld, S.A., Nater, E.A. and Wright, H.E. (2006) Sources of sediment in Lake Pepin on the upper Mississippi River in response to Holocene climatic changes. *J. Paleolimnol.* **35**, 193–206.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, r25.

Lemus, R., Brummer, E.C., Moore, K.J., Molstad, N.E., Burras, C.L. and Barker, M.F. (2002) Biomass yield and quality of 20 switchgrass populations in southern Iowa, USA. *Biomass Bioenergy*, **23**, 433–442.

Lesica, P. and Allendorf, F.W. (1999) Ecological genetics and the restoration of plant communities: mix or match? *Restor. Ecol.* **7**, 42–50.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Proc, G.P.D. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lipka, A.E., Lu, F., Cherney, J.H., Buckler, E.S., Casler, M.D. and Costich, D.E. (2014) Accelerating the switchgrass (*Panicum virgatum* L.) breeding cycle using genomic selection approaches. *PLoS ONE*, **9**, e112227.

Liu, L.L. and Wu, Y.Q. (2012) Identification of a selfing compatible genotype and mode of inheritance in switchgrass. *Bioenerg. Res.* **5**, 662–668.

Lu, K., Kaeppler, S.M., Vogel, K., Arumuganathan, K. and Lee, D. (1998) Nuclear DNA content and chromosome numbers in switchgrass. *Great Plains Res.* **8**, 269–280.

Lu, F., Lipka, A.E., Glaubitz, J., Elshire, R., Cherney, J.H., Casler, M.D., Buckler, E.S. and Costich, D.E. (2013) Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genet.* **9**, e1003215.

Maron, L.G., Guimaraes, C.T., Kirst, M. *et al.* (2013) Aluminum tolerance in maize is associated with higher MATE1 gene copy number. *Proc. Natl Acad. Sci. USA*, **110**, 5241–5246.

Martinez-Reyna, J.M. and Vogel, K.P. (2002) Incompatibility systems in switchgrass. *Crop Sci.* **42**, 1800–1805.

Mascher, M., Richmond, T.A., Gerhardt, D.J. *et al.* (2013) Barley whole exome capture: a tool for genomic research in the genus *Hordeum* and beyond. *Plant J.* **76**, 494–505.

Missaoui, A.M., Paterson, A.H. and Bouton, J.H. (2006) Molecular markers for the classification of switchgrass (*Panicum virgatum* L.) germplasm and to assess genetic diversity in three synthetic switchgrass populations. *Genet. Resour. Crop Evol.* **53**, 1291–1302.

Murphy, R.L., Klein, R.R., Morishige, D.T., Brady, J.A., Rooney, W.L., Miller, F.R., Dugas, D.V., Klein, P.E. and Mullet, J.E. (2011) Coincident light and clock regulation of pseudoresponse regulator protein 37 (PRR37) controls photoperiodic flowering in sorghum. *Proc. Natl Acad. Sci. USA*, **108**, 16469–16474.

Ng, P.C. and Henikoff, S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874.

Paterson, A.H., Bowers, J.E., Bruggmann, R. *et al.* (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, **457**, 551–556.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Saghaimaroof, M.A., Soliman, K.M., Jorgensen, R.A. and Allard, R.W. (1984) Ribosomal DNA spacer-length polymorphisms in barley – mendelian inheritance, chromosomal location, and population-dynamics. *Proc. Natl Acad. Sci. USA*, **81**, 8014–8018.

Sanderson, M. (2007) The US experience developing switchgrass as a bioenergy crop – applications to Canada. *Can. J. Plant Sci.* **87**, 527–527.

Schmer, M.R., Vogel, K.P., Mitchell, R.B. and Perrin, R.K. (2008) Net energy of cellulosic ethanol from switchgrass. *Proc. Natl Acad. Sci. USA*, **105**, 464–469.

Schnable, P.S., Ware, D., Fulton, R.S. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.

Sharma, M.K., Sharma, R., Cao, P.J., Jenkins, J., Bartley, L.E., Qualls, M., Grimwood, J., Schmutz, J., Rokhsar, D. and Ronald, P.C. (2012) A genome-wide survey of switchgrass genome structure and organization. *PLoS ONE*, **7**, e33892.

Soltis, D.E., Gitzendanner, M.A., Strenge, D.D. and Soltis, P.S. (1997) Chloroplast DNA intraspecific phylogeography of plants from the Pacific Northwest of North America. *Plant Syst. Evol.* **206**, 353–373.

Stermitz, F.R., Lorenz, P., Tawara, J.N., Zenewicz, L.A. and Lewis, K. (2000) Synergy in a medicinal plant: antimicrobial action of berberine potentiated by 5′-methoxyhydnocarpin, a multidrug pump inhibitor. *Proc. Natl Acad. Sci. USA*, **97**, 1433–1437.

Sullivan, M.J., Petty, N.K. and Beatson, S.A. (2011) Easyfig: a genome comparison visualizer. *Bioinformatics*, **27**, 1009–1010.

Swanson-Wagner, R.A., Eichten, S.R., Kumari, S., Tiffin, P., Stein, J.C., Ware, D. and Springer, N.M. (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res.*, **20**, 1689–1699.

Talbert, L.E., Timothy, D.H., Burns, J.C., Rawlings, J.O. and Moll, R.H. (1983) Estimates of genetic-parameters in switchgrass. *Crop Sci.* **23**, 725–728.

The 3000 Rice Genomes Project (2014) The 3,000 rice genomes project. *GigaScience*, **3**, 7.

Turck, F., Fornara, F. and Coupland, G. (2008) Regulation and identity of florigen: flowering locus T moves center stage. *Annu. Rev. Plant Biol.* **59**, 573–594.

Uitdewilligen, J.G., Wolters, A.M., D'Hoop B, B., Borm, T.J., Visser, R.G. and van Eck, H.J. (2013) A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PLoS ONE*, **8**, e62355.

Valverde, F. (2011) CONSTANS and the evolutionary origin of photoperiodic timing of flowering. *J. Exp. Bot.* **62**, 2453–2463.

VanEsbroeck, G.A., Hussey, M.A. and Sanderson, M.A. (1997) Leaf appearance rate and final leaf number of switchgrass cultivars. *Crop Sci.* **37**, 864–870.

Vogel, K. (2004) Switchgrass. In *Warm-season (C4) Grasses* (Moser, L.E., ed.). Madison, WI: ASA, CSSA, SSSA, pp. 561–588.

Wang, K., Li, M.Y. and Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164.

Winzer, T., Gazda, V., He, Z. *et al.* (2012) A *Papaver somniferum* 10-gene cluster for synthesis of the anticancer alkaloid noscapine. *Science*, **336**, 1704–1708.

Xu, K., Xu, X., Fukao, T., Canlas, P., Maghirang-Rodriguez, R., Heuer, S., Ismail, A.M., Bailey-Serres, J., Ronald, P.C. and Mackill, D.J. (2006) Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. *Nature*, **442**, 705–708.

Xu, B., Sathitsuksanoh, N., Tang, Y.H., Udvardi, M.K., Zhang, J.Y., Shen, Z.X., Balota, M., Harich, K., Zhang, P.Y.H. and Zhao, B.Y. (2012a) Overexpression of AtLOV1 in switchgrass alters plant architecture, lignin content, and flowering time. *PLoS ONE*, **7**, e47399.

Xu, J., Liu, Y.X., Liu, J., Cao, M.J., Wang, J., Lan, H., Xu, Y.B., Lu, Y.L., Pan, G.T. and Rong, T.Z. (2012b) The genetic architecture of flowering time and photoperiod sensitivity in maize as revealed by QTL review and meta analysis. *J. Integr. Plant Biol.* **54**, 358–373.

Yang, S.S., Weers, B.D., Morishige, D.T. and Mullet, J.E. (2014) CONSTANS is a photoperiod regulated activator of flowering in sorghum. *BMC Plant Biol.* **14**, 148.

Yano, M., Katayose, Y., Ashikari, M. *et al.* (2000) Hd1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the Arabidopsis flowering time gene constans. *Plant Cell*, **12**, 2473–2484.

Zalapa, J.E., Price, D.L., Kaeppler, S.M., Tobias, C.M., Okada, M. and Casler, M.D. (2011) Hierarchical classification of switchgrass genotypes using SSR and chloroplast sequences: ecotypes, ploidies, gene pools, and cultivars. *Theor. Appl. Genet.* **122**, 805–817.

Zhang, Y.W., Zalapa, J., Jakubowski, A.R., Price, D.L., Acharya, A., Wei, Y.L., Brummer, E.C., Kaeppler, S.M. and Casler, M.D. (2011a) Natural hybrids and gene flow between upland and lowland switchgrass. *Crop Sci.* **51**, 2626–2641.

Zhang, Y.W., Zalapa, J.E., Jakubowski, A.R., Price, D.L., Acharya, A., Wei, Y.L., Brummer, E.C., Kaeppler, S.M. and Casler, M.D. (2011b) Post-glacial evolution of *Panicum virgatum*: centers of diversity and gene pools revealed by SSR markers and cpDNA sequences. *Genetica*, **139**, 933–948.

Zhang, G.Y., Liu, X., Quan, Z.W. *et al.* (2012) Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nat. Biotechnol.* **30**, 549–554.

Zmienko, A., Samelak, A., Kozlowski, P. and Figlerowicz, M. (2014) Copy number polymorphism in plant genomes. *Theor. Appl. Genet.* **127**, 1–18.