Allele Mining in Diverse Accessions of *Urochloa* and *Megathyrsus* spp. Tropical Grasses to Improve Forage Quality and Reduce Environmental Impact

SJ Hanley¹, TK Pellny¹, JJ de Vega², V Castiblanco³, J Arango³, PJ Eastmond¹, JS Heslop-Harrison⁴, RAC Mitchell¹

¹ Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK  
² Earlham Institute, Norwich Research Park, NR4 7UZ, UK  
³ International Center for Tropical Agriculture (CIAT), 6713 Cali, Colombia  
⁴ Department of Genetics & Genome Biology, University of Leicester, Leicester LE1 7RH, UK

Highlight

We found gene variants in a collection of tropical grasses that could help reduce environmental impact of cattle production.

Abstract

The C4 *Urochloa* spp (syn. *Brachiaria*) and *Megathyrsus maximus* (syn. *Panicum maximum*) are used as pasture for cattle across vast areas in tropical agriculture systems in Africa and South America. A key target for variety improvement is forage quality: enhanced digestibility could decrease amount of land required per unit production and enhanced lipid content could decrease methane emissions from cattle. For these traits, loss-of-function (LOF) alleles in known gene targets are predicted to improve them, making a reverse genetics approach of allele mining feasible. We studied allelic diversity of 20 target genes (11 for digestibility, 9 for lipid content) in 104 accessions selected to represent genetic diversity and ploidy levels of *U. brizantha*, *U. decumbens*, *U. humidicola*, *U. ruizienia* and *M. maximum*. We used RNAseq and then bait-capture DNA-seq to improve gene models in a *U. ruizienia* reference genome to assign polymorphisms with high confidence. We found 953 non-synonymous polymorphisms across all genes and accessions; within these, we identified 7 putative LOF alleles with high confidence, including ones in the non-redundant SDP1 and
BAHD01 genes present in diploid and tetraploid accessions. These LOF alleles could respectively confer increased lipid content and digestibility if incorporated into a breeding programme.
Introduction

The environmental impact of cattle production could be decreased by reducing the amount of land required (e.g., land sparing) and amount of methane (CH$_4$) emitted per unit production (i.e., emission intensity). This could be achieved by genetic improvement of pasture grass on which they feed: an increase in digestibility and energy content would allow the same production to be achieved on a smaller land area and an increase in lipid content in vegetative matter would decrease CH$_4$ emitted per unit production, provided that these two traits could be improved without negative side effects, such as reduced growth or susceptibility to biotic or abiotic stresses.

Breeding of commercial tropical forage grass varieties in diploid and polyploid species and interspecific hybrids of *Urochloa* has been achieved by recurrent selection over many years, identifying superior-performing populations for key traits such as biomass production in different environments, resistance to pests and digestibility (Worthington and Miles, 2014). These targets increase efficiency of forage grass, such that less land is required for production. Increasingly, environmental targets such as decreased nitrogen losses (Nuñez et al., 2018; Villegas et al., 2020) and reduced methane emissions from grazing cattle (Gaviria-Uribé et al., 2020) have become public breeding targets for improved pasture grasses. Continued improvement will be accelerated using genetic diversity that is available from accessions of the same genus available in genebank collections; the ploidy and relatedness of 280 of accessions from the International Center for Tropical Agriculture (CIAT) collection of *Urochloa* spp and genome composition of some of the polyploids (P., Tomaszewska, pers. comm.). The relation between these species has previously been studied using microsatellite markers (Triviño et al., 2017).

However, introduction of such material into current breeding programmes is a major undertaking requiring evidence of likely benefit; breeding of *Urochloa* tropical forage grasses is particularly complicated by obligate outcrossing in sexual accessions and occurrence of apomixis in half of the progeny (Worthington and Miles, 2014). For traits where there are known key genes and an understanding of how variants of these might affect the traits, an allele mining approach may be feasible where sequencing of target genes rather than phenotyping can be used to find potentially useful alleles. This reverse genetic approach can find useful loss-of-function variation that would not be found by phenotyping as its effect is hidden due to gene redundancy (Comai, 2005), particularly in polyploid or highly heterozygous material such as *Urochloa*, and provides the basis for perfect markers in crosses for following the alleles. Successful examples of allele mining for natural variation...
in known genes include studies on rice germplasm for starch synthesis genes (Butardo et al., 2017) and on Sorghum germplasm for a gene responsible for Al tolerance (Hufnagel et al., 2018).

Two traits where loss-of-function (LOF) alleles have been identified as beneficial are (1) digestibility where improvements have been gained by knock-out or knock-down of genes involved in cell wall synthesis in grasses and (2) lipid content of vegetative tissue where improvements could be gained by knock-out or knock down of genes involved in lipid metabolism. Increased lipid content of vegetative tissue of forage results in decreased CH₄ emissions from cattle that feed on it, as well as benefiting meat and dairy fatty acid composition, as demonstrated by a transgenic approach in Lolium perenne (Winichayakul et al., 2020).

Here we compile a list of genes identified as targets from work in our labs or elsewhere, and find the orthologues in U. ruziizensis diploid reference species. We then conduct a comprehensive screening of alleles for these genes in 104 diverse accessions of U. brizantha, U. decumbens, U. humidicola, U. ruziizensis and M. maximum using RNAseq and bait capture genomic DNA sequencing.
Materials and Methods

The sections below correspond to the steps shown in red in Figure 1.

Plant materials and RNA sequencing

We collected leaf samples of 104 accessions (Supplemental Table 1) from the field-grown genebank collection at CIAT which were immediately frozen in liquid nitrogen. Samples were ground to a fine powder in liquid nitrogen and subsequently lyophilised. Total RNA was extracted as described in (Pellny et al., 2012) with the difference that prior to DNase treatment the pellets were dried in a rotary evaporator (Eppendorf) and stored/transported at room temperature. Illumina sequencing using standard RNA-seq library preparations with paired reads of length 150 bp was conducted by Novogene, HK. The raw reads were deposited in SRA under Bioproject PRJNA513453. We also collected leaf samples from an overlapping set of accessions for DNA extraction (Supplemental Table 1) and 80 of these were used for DNA sequencing described below.

Orthologue Identification

We searched for U. ruzizienis orthologs of the target genes identified in other species listed in Table 1. Firstly, we identified the putative orthologues of the target genes in Setaria viridis (Setaria) v1.1 genome (Goodstein et al., 2012) as the closest reliably annotated genome using BLASTN with target genes’ CDS as original queries and source genomes (i.e. Arabidopsis, maize, sorghum, Setaria or Brachypodium) of target genes for reciprocal BLASTN of hits. This identified 1-to-1 orthologues for all source genes except for Arabidopsis SDP1 and PXA1 genes where there were two putative paralogues each in Setaria. We repeated this process for the draft U. ruzizienis v1.0 annotated genome (Worthington et al., 2020) and found the same orthologous relationships as for Setaria except for one target gene CGI58 lipase, where an additional parologue was found in U. ruzizienis v1.0. We compared the U. ruzizienis v1.0 gene models with the corresponding Setaria and source genes to judge whether they were correct; for 10 of 22 they were incomplete. We compiled a set of 22 genes using the 12 complete U. ruzizienis v1.0 genes and 10 Setaria genes for the others. We then mapped RNAseq reads from 11 U. ruzizienis accessions to this set. Two U. ruzizienis v1.0 genes with no equivalents in Setaria had almost zero mapped reads, and we removed these as likely pseudogenes, leaving a total of 20 U. ruzizienis target genes. Baits were designed to these 20 U. ruzizienis genomic regions taking account of the mapped RNAseq to customise baits for each species-ploidy group. Bait capture was performed on genomic DNA isolated from 80 accessions Resulting IonTorrent sequencing, RNAseq reads and targeted Sanger sequencing for U. ruzizienis
accessions were together used to check and refine gene models. We annotated the CDS by finding the longest ORF and comparing with that of orthologues. For 19 of these, we found complete coding sequences, but gene Ur.CG158 lacks the first exon. We deposited final annotated sequences for all 20 U. ruziizensis genes in Genbank/EMBL accession numbers (MW323383-MW323402).

Read Mapping

We did the mapping and variant calling on the Galaxy platform (Giardine et al., 2005). We first mapped RNAseq of 11 U. ruziizensis accessions to the U. ruziizensis v1.0 reference genome using BWA-MEM (Li and Durbin, 2010). Taking these alignments into Geneious, for each target gene, we combined mapped reads with set of all unmapped reads and conducted a de novo assembly. We compared resulting contigs with U. ruziizensis and Setaria viridis gene models and manually improved U. ruziizensis gene models. We substituted these gene models for the original versions as a first attempt at improving the reference and designed baits based on these. After we completed sequencing of bait capture DNA, we mapped DNA and RNA reads of U. ruziizensis accessions to the modified reference to iteratively improve it until all reads mapped satisfactorily. We substituted the final version of the U. ruziizensis gene models (18 out of 21 were changed from original version) into the U. ruziizensis genome annotation and mapped the RNA and DNA reads of all accessions using HiSAT2 (Kim et al., 2015) and the TMAP mapper within Torrent Suite 5.12.2 software (Thermofisher), respectively. For the latter, only reads greater than 100bp were used. We used the resulting BAM files (104 from RNAseq, 80 from DNAseq) to call variants and to manually inspect alignments on IGV (Robinson et al., 2011) for putative LOF alleles.

Bait Capture

Coding sequences of the 20 genes of interest were targeted using myBaits Custom DNA-seq technology (Arbor Biosciences). A single bait set was designed to capture all genes in any of the species studied. To account for the likely diversity represented within and between species, consensus sequences were derived from the available RNAseq data for all genes in all individuals of each of the species. These were submitted to the design process performed by Arbor Biosciences which resulted in 20,346 baits of 70 nucleotide length with 3x tiling. Genomic DNA was extracted from frozen leaf tissue using a Plant DNeasy Kit (Qiagen) according to the supplied protocol. DNA quality was assessed by agarose gel electrophoresis and quantified using the Qubit dsDNA BR Assay Kit (Thermofisher). Whole genome libraries for use in bait capture were prepared using Ion Plus Fragment Library kits according to the manufacturer’s instructions with a target insert size of 400bp
and unique Ion Xpress barcodes for each sample. Libraries were then amplified using the library kit PCR reagents to generate sufficient DNA for bait hybridisation. All libraries were quantified by qPCR using a Kapa Library Quantification Kit (Roche) and 16 equimolar pools made, each comprising five libraries. For each pool, bait capture was performed according to the manufacturer’s myBaits® Manual v4. Libraries were then quantified by qPCR as before, pooled and sequenced across two runs on an Ion Torrent PGM sequencer, using Ion Hi-Q View OT2 reagents for 400bp templating and the Ion PGM Hi-Q View Sequencing Kit for 400 bp sequencing.

Variant Calling

We used FreeBayes (Galaxy Version 1.0.2.29-3) to call variants, which is a haplotype-based variant calling program capable of dealing with polyploidy (Garrison and Marth, 2012). Both RNAseq and DNA capture BAM files (104 RNA, 80 DNA, overlap of accessions 74) were divided into 11 groups with the same species and ploidy (Table S2) using ploidy information from cytogenetics for the accessions (P. Tomaszewska, pers. comm.). BAM files from each group were submitted together to FreeBayes with the appropriate setting for ploidy, variant calling was limited to the target genes with default parameters for DNA reads; for RNA reads, the minimum fraction of observations supporting an allele (—min-alternate-fraction) was set to 0.05 to allow for low abundance due to nonsense-mediated decay of transcripts (Gutierrez et al., 1999) from LOF alleles. All other FreeBayes parameters were default. We retained variants with quality>=20 using SNPsift v4.0 (Cingolani et al., 2012a). Using custom Perl scripts, we compared polymorphisms from DNA and RNA VCF files produced by FreeBayes for the same group. Polymorphisms observed from the RNAseq were filtered out unless they were also observed in the corresponding DNAseq bait capture sequences for the same accession, or where this was not present, in another accession from the same species.

Variant Effect Prediction

We identified effects on function of the putative polymorphisms with SNPeff v4.0 (Cingolani et al., 2012b), which uses the CDS annotation to predict effects on encoded proteins. We compiled information on all unique variants using custom Perl scripts to process VCF files (summarised in Table 2 and Figs 2, 3).

Classification of missense variants as tolerated or non-tolerated by SIFT

We downloaded orthologs’ protein sequences for the 20 target genes in angiosperms with fully sequenced genomes from Phytozome v12 (Goodstein et al., 2012) and aligned them using Muscle (Edgar, 2004), with default parameters, together with our U. ruiziensis reference protein sequence. We assumed that at least one gene must be functional for each of the 50 species, with the exception
of BAHDO1 and BAHDO4 genes, where we included only the 13 commelinid monocot species as their function is believed to be confined to these species (Mitchell et al., 2007). We removed any paralogues that did not align well. For each gene, we supplied these alignments and the discovered missense variants to the SIFT web server (Sim et al., 2012). We then used the SIFT prediction to classify the missense variants as tolerated (score >0.05) or non-tolerated (score<=0.05); non-tolerated predictions were all flagged as low confidence because of the small number of sequences available for alignment, while tolerated ones were regarded as reliable.
Results

Identification of target genes

We identified genes from the literature, including published work from our own labs, where there was evidence that a loss of function in the gene would confer either increased digestibility (cell wall genes) or increased lipid content in vegetative tissue (lipid genes). This evidence is summarised in Table 1.
Table 1. Evidence from literature for selection of target genes to improve forage quality.

<table>
<thead>
<tr>
<th>Target trait</th>
<th>Gene</th>
<th>Species</th>
<th>Suppression mode</th>
<th>Effect on trait</th>
<th>Pleiotropic effects</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digestibility</td>
<td>4CL/Class I</td>
<td><em>Sorghum bicolor</em></td>
<td>missense mutant bmr2</td>
<td>17% increased saccharification</td>
<td>(Saballos et al., 2008; Sattler et al., 2010)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Saccharum officinarum</em></td>
<td>RNAi</td>
<td>52%, 76% improved saccharification (field-grown)</td>
<td>0%, 30% DM yield penalty (field-grown)</td>
<td>(Jung et al., 2016)</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Setaria viridis, Saccharum officinarum</em></td>
<td>RNAi</td>
<td>40-80% increased saccharification</td>
<td>no growth penalty in GH</td>
<td>(de Souza et al., 2018; de Souza et al., 2019)</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Setaria viridis</em></td>
<td>RNAi</td>
<td>10-20% increased saccharification</td>
<td>no growth penalty in GH</td>
<td>(Mota et al., 2020)</td>
</tr>
<tr>
<td>COMT</td>
<td><em>Zea mays</em></td>
<td></td>
<td>bm3 LOF mutant</td>
<td>used in commercial hybrids with improved digestibility for cattle</td>
<td>some reports yield penalty, sometimes no yield penalty</td>
<td>(Sattler et al., 2010; Vignols et al., 1995)</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Sorghum bicolor</em></td>
<td>bmr12 LOF mutant</td>
<td>30% increased tract digestibility</td>
<td>10% DM yield penalty</td>
<td>(Saballos et al., 2008; Sattler et al., 2010)</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Panicum virgatum</em></td>
<td>RNAi</td>
<td>30% increased digestibility</td>
<td>no growth penalty in GH</td>
<td>(Fu et al., 2011a)</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Saccharum officinarum</em></td>
<td>TALEN induced LOF mutations in multiple paralogs</td>
<td>40% improved saccharification (field-grown)</td>
<td>no DM yield penalty (field-grown)</td>
<td>(Kannan et al., 2018)</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Brachypodium distachyon</em></td>
<td>missense mutants</td>
<td>40% increased saccharification</td>
<td>no growth penalty in GH</td>
<td>(Bouvier d’Yvoire et al., 2013)</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Sorghum bicolor</em></td>
<td>missense mutant bmr6-3</td>
<td>20% increased tract digestibility</td>
<td>15% DM yield penalty</td>
<td>(Saballos et al., 2009; Sattler et al., 2010)</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Zea mays</em></td>
<td>bm1</td>
<td></td>
<td>no yield penalty</td>
<td>(Halpin et al., 1998; Sattler et al., 2010)</td>
</tr>
<tr>
<td>Lipid content</td>
<td>Mutant</td>
<td>Gene family</td>
<td>Phenotype</td>
<td>Other phenotype notes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>--------</td>
<td>-------------</td>
<td>-----------</td>
<td>-----------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Panicum virgatum</td>
<td>RNAi</td>
<td>20% increased saccharification</td>
<td>normal GH growth</td>
<td>(Fu et al., 2011b)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zea mays</td>
<td>RNAi</td>
<td>20% increased saccharification</td>
<td>increased growth GH</td>
<td>(Park et al., 2012)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brachypodium distachyon</td>
<td>missense mutant?</td>
<td>10% increased saccharification</td>
<td>no growth penalty in GH</td>
<td>(Whitehead et al., 2018)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zea mays</td>
<td>association with polymorphisms</td>
<td>correlation with fibre digestibility</td>
<td></td>
<td>(Brenner et al., 2010)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>LOF mutant</td>
<td>increase in leaf triacylglycerol</td>
<td>poor seedling establishment in oilseeds</td>
<td>(Kelly et al., 2013)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VIGS</td>
<td>Increase in leaf lipid content</td>
<td>none</td>
<td></td>
<td>(Wijekoon et al.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>LOF mutant</td>
<td>increase in leaf triacylglycerol</td>
<td>none</td>
<td>(James et al., 2010)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>LOF mutant</td>
<td>increase in leaf triacylglycerol</td>
<td>poor seedling establishment in oilseeds, starvation sensitive, jasmonate deficient</td>
<td>(Slcombe et al., 2009)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VIGS</td>
<td>Increase in leaf lipid content</td>
<td>none</td>
<td></td>
<td>(Wijekoon et al.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>leaky mutant</td>
<td>increase in leaf triacylglycerol</td>
<td>embryo defect, growth penalty</td>
<td>(Xu et al., 2005)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>leaky mutant</td>
<td>increase in leaf triacylglycerol</td>
<td>embryo defect, growth penalty</td>
<td>(Awai et al., 2006)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>leaky mutant</td>
<td>increase in leaf triacylglycerol</td>
<td>embryo defect, growth penalty</td>
<td>(Lu et al., 2007)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
We identified orthologues of the genes in Table 1 in *Setaria viridis*, *Setaria italica*, *Sorghum bicolor* (as the most closely related fully sequenced genomes) and in the draft genome of *U. ruziizensis* (Supplemental Table 2). We found additional putative paralogues in *U. ruziizensis* for 4CL, CAD, CCoAOMT and CGI58 and we included these in the analysis, naming them with the suffix “_p1”. We therefore had a final total of 11 cell wall genes and 9 lipid genes as our target set identified in *U. ruziizensis*.

RNAseq was carried out on RNA collected from leaves of 104 accessions growing in fields at CIAT. These data have been deposited at NCBI under BioProject PRJNA513453. We took reads from *U. ruziizensis* accessions that mapped to the target *U. ruziizensis* genes (Supplemental Table 2) and unmapped reads and re-assembled these target genes. We compared the resulting sequences with target *U. ruziizensis* and *S. viridis* gene models to improve the *U. ruziizensis* gene models and design baits. We carried out bait capture of genomic DNA from a set of 80 accessions, 74 of which were in the RNAseq set. We used *U. ruziizensis* RNAseq, bait capture DNAseq and Sanger sequencing to iteratively improve the *U. ruziizensis* gene models. We submitted the final gene model versions to Genbank/EMBL and substituted them for the original versions into the *U. ruziizensis* reference genome. We then re-mapped all accessions’ RNAseq and bait capture DNAseq to this updated reference. We called variants on the resulting alignments and identified those that were found in RNAseq and confirmed as present in bait capture DNAseq in same accession or, in cases where DNAseq was not available for same accession, from any other accession of same species. We also looked for special case of loss of splice donor or acceptor in DNAseq, which would be expected to change RNAseq read distribution, but did not find any instances of this. We present results below only for variants found in RNAseq and confirmed in bait capture DNAseq since these have good confidence as the two approaches have different sources of error.

The numbers of variants of different types in the target genes are summarised in Table 2 and the complete set is available in Supplemental Table 3.
Table 2. Total numbers of polymorphisms found in RNAseq and confirmed in bait capture DNA of 124 accessions for the 20 target *U. ruziienis* genes. Genes are classified by type or predicted effect on protein. Type ‘snp/mnp’ includes SNPs and a small number of contiguous multiple nucleotide polymorphisms in the same haplotype; ‘complex’ denotes a mixture of SNPs and indels. ‘LOW’ effects are synonymous variants, ‘MODERATE’ are missense, in-frame indels, start or stop lost, and ‘HIGH’ are frameshift or stop gained, predicted to cause loss of function (LOF). From MODERATE missense variants, counts of those predicted as non-tolerated by SIFT web server are shown “(N:)”.

<table>
<thead>
<tr>
<th>type</th>
<th>effects</th>
<th>GENE</th>
<th>SNP/mnp</th>
<th>indel</th>
<th>complex</th>
<th>LOW</th>
<th>MODERATE</th>
<th>HIGH</th>
</tr>
</thead>
<tbody>
<tr>
<td>cell wall genes</td>
<td></td>
<td>Ur.4CL</td>
<td>200</td>
<td>0</td>
<td>44</td>
<td>186</td>
<td>58 (N: 15)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ur.4CL_p1</td>
<td>195</td>
<td>3</td>
<td>39</td>
<td>169</td>
<td>67 (N: 17)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ur.BAHD01</td>
<td>146</td>
<td>1</td>
<td>33</td>
<td>144</td>
<td>35 (N: 5)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ur.BAHD05</td>
<td>183</td>
<td>0</td>
<td>29</td>
<td>159</td>
<td>53 (N: 19)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ur.CAD</td>
<td>109</td>
<td>0</td>
<td>15</td>
<td>100</td>
<td>24 (N: 6)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ur.CAD_p1</td>
<td>117</td>
<td>1</td>
<td>17</td>
<td>88</td>
<td>46 (N: 11)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ur.CCoAOMT</td>
<td>81</td>
<td>1</td>
<td>6</td>
<td>72</td>
<td>15 (N: 7)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ur.CCoAOMT_p1</td>
<td>49</td>
<td>0</td>
<td>22</td>
<td>64</td>
<td>7 (N: 3)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ur.CCR</td>
<td>76</td>
<td>1</td>
<td>12</td>
<td>74</td>
<td>15 (N: 4)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ur.CG58</td>
<td>54</td>
<td>0</td>
<td>14</td>
<td>39</td>
<td>29 (N: 7)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ur.CG58_p1</td>
<td>68</td>
<td>0</td>
<td>15</td>
<td>47</td>
<td>35 (N: 11)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ur.COMT</td>
<td>104</td>
<td>0</td>
<td>26</td>
<td>108</td>
<td>22 (N: 8)</td>
<td>0</td>
</tr>
<tr>
<td>lipid genes</td>
<td></td>
<td>Ur.GT43A</td>
<td>169</td>
<td>0</td>
<td>30</td>
<td>159</td>
<td>40 (N: 13)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ur.PXA1</td>
<td>284</td>
<td>0</td>
<td>42</td>
<td>228</td>
<td>97 (N: 26)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ur.PXA1-like</td>
<td>337</td>
<td>0</td>
<td>21</td>
<td>228</td>
<td>130 (N: 50)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ur.SDP1</td>
<td>236</td>
<td>2</td>
<td>32</td>
<td>169</td>
<td>99 (N: 33)</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ur.SDP1-like</td>
<td>234</td>
<td>5</td>
<td>28</td>
<td>181</td>
<td>82 (N: 8)</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ur.TGD1</td>
<td>81</td>
<td>0</td>
<td>12</td>
<td>78</td>
<td>15 (N: 3)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ur.TGD2</td>
<td>71</td>
<td>0</td>
<td>7</td>
<td>51</td>
<td>27 (N: 10)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ur.TGD3</td>
<td>118</td>
<td>0</td>
<td>17</td>
<td>90</td>
<td>45 (N: 0)</td>
<td>0</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>2912</td>
<td>14</td>
<td>461</td>
<td>2434</td>
<td>941 (N: 256)</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>
We were most interested in mutations that disrupt function but found only 12 variants predicted to lose function (Table 1); however, among the 941 “moderate” mutations (mostly missense nonsynonymous mutations), the single nucleotide polymorphism results in a different amino acid and it is expected that some of these changes will be disruptive. Using the SIFT web server (Sim et al., 2012), we supplied our protein alignments of orthologs from fully sequenced plant genomes and used resulting SIFT predictions to categorize missense mutations into tolerated and non-tolerated classes. From this analysis, 256 further variants that we discovered may disrupt gene function (Table 2).

We looked at the number of variants found in individual accessions, grouped by species and ploidy (presented as box whisker plots in Figure 2). As expected, we found more variants in accessions from species that are more distantly related to the *U. ruizizensis* reference, i.e. *U. humidicola* and *M. maximum* (Triviño et al., 2017). An outlier accession #26175 for group *U. ruizizensis* with high numbers of variants in these target genes is indicated in Fig. 2; this may be misclassified and is probably not *U. ruizizensis* according to a global analysis of all genes’ SNPs (JJDV, unpublished). We found very similar patterns for low effect and tolerated missense polymorphisms suggesting that these both reflect relatedness to the reference. However, we found a different pattern for non-tolerated polymorphisms predicted to affect protein function by SIFT, which were much more common on groups with high ploidy (Figure 2).

Many of the polymorphisms were shared between multiple accessions and we present a summary of this in Figure 3. We found that polymorphisms predicted to disrupt (non-tolerated missense) or eliminate (LOF) function were shared between fewer accessions than other polymorphisms. Polymorphisms characteristic of subgenomes (homeologues) would be expected to be present in all accessions with these subgenomes (typically >20 in the set used here), whereas allelic variants would be present in fewer accessions. From our analysis, it appears that non-tolerated missense and LOF mutations are much more likely to be allelic than homeologous compared with other mutations (Fig. 3).

We manually examined the alignments for the 12 putative LOF alleles we found initially by our automated pipeline (Table 2) and found that 3 were frameshifts in stretches of homopolymer or low complexity, with low coverage in some cases. It is likely that these are real since they were found in the same accessions in RNAseq and gDNA sequencing, but it is also possible they are artefacts due to systematic errors common to both DNA and RNAseq approaches. We found two others were predicted to truncate protein close to C-terminus so were less certain to knock-out function. We therefore designate these 5 as “low confidence” and the remaining 7 as “high confidence”. We show
the alignments of these 7 LOF alleles, 3 in cell wall genes (Figure 4), 4 in lipid genes (Figure 5). From a breeding perspective, it is more difficult to transfer an allele from an accession with higher ploidy to a line with lower ploidy; since commercial varieties of these species are tetraploid, this may make the alleles of PXA1, SDP1-like, 4CL_p1 found in accessions with ploidy >4 of less immediate value. This leaves the putative LOF alleles in BAHDO1 in tetraploid *U. brizantha*, in CAD_p1 in diploid *U. ruziziensis* and SDP1 in diploid *U. decumbens* as of most potential interest. All these alleles appear to be present in heterozygous form so further breeding would be required even in diploid accessions to achieve complete loss of function.
Discussion

We developed a new methodology for allele discovery of candidate genes in a collection of diploid and polyploid lines with only a draft genome sequence for one diploid species as reference. Our approach of combining RNA-seq and bait capture (Fig. 1) provides a means of avoiding pseudogenes and resolving complexities. As part of the process, we improved gene models for 18 key genes and confirmed 2 more as accurate in the U. ruizienis v1 genome. Our approach could be adapted for allele discovery in other plant collections or populations.

Our motivation in this work was the hope that breeding tropical forage grass with increased digestibility and lipid content could reduce environmental impact of cattle production by respectively decreasing land requirement and CH4 emissions. We selected target genes from our work or the literature where reduction in function improves either digestibility or lipid content of vegetative tissue (Table 1). In the case of digestibility, the evidence comes directly from grass species, whereas the target genes for lipid content have so far only been tested in dicots. These genes differ substantially in the effects of knock-downs or knock-outs and in the evidence for any adverse pleiotropic effects. For some, it is thought that a complete knock-out of function is required for the beneficial effect and this causes little or no side-effects (e.g. COMT in maize; (Vignols et al., 1995)). For the BAHDO1 and BAHDO5 genes putatively involved in addition of hydroxycinnamic acids to arabinoxylan, no complete knock-outs have been reported but knock-downs can have substantial effects (de Souza et al., 2018; de Souza et al., 2019). In general, LOF alleles have less effect the greater the redundancy from other genes, so are recessive, although dosage effects can occur. In polyploid species like wheat, it can be necessary to stack homozygous LOF alleles in all homeologs to achieve a phenotype (Borrill et al., 2019).

We found 941 non-synonymous variants for our 20 target genes within 104 CIAT Genebank accessions confirmed in RNAseq and DNA (Table 2). Most of these likely have little or no effect on function but to gauge which ones are more likely to be detrimental we used the SIFT webserver (Sim et al., 2012) to identify a subset of 256 non-tolerated missense variants. Since these are predicted by SIFT based on alignments of all orthologs, they do not reflect relatedness to the U. ruizienis reference and their frequency in accessions was principally dependent on ploidy (Fig. 2). This is most simply explained by the increasing copy number of the genes. An additional effect might be expected where detrimental mutations accumulate in lines with higher ploidy as purifying selection will act less on highly redundant genes, but we could not judge this from our data. In fact, these non-tolerated missense variants were only predicted to be detrimental with low confidence by SIFT due
to insufficient diversity of orthologues from fully sequenced plant genomes. These predictions could
be improved in future as more genomes are sequenced and knowledge of the proteins improves.

The most secure predictions for disrupted function are the LOF variants with premature stop codons
or frameshifts. We found that both non-tolerated missense and LOF variants tended to be shared
between fewer accessions compared to other variants (Fig. 3), indicating they were more likely to be
allelic than homeologous. Homeologous variants are more ancient than allelic variants so it may be
that purifying selection has tended to remove detrimental variants over longer timescales.

On manual inspection of LOF variants, 5 were such that we were not completely confident they were
real or were likely to knock-out function. Of the other 7 (Figs. 4, 5), three were of particular interest
since they occur in tetraploid or diploid accessions that are more easily incorporated into breeding
programmes; these occurred in CAD_p1, BAHDO1 and SDP1 genes. However, CAD_p1 is putatively
redundant with CAD and we have not found a report of effect of repressing CAD_p1 without also
repressing CAD. No complete knock out of BAHDO1 has been reported but partial suppression had a
substantial effect on digestibility in Setaria (de Souza et al., 2018), so it is possible that dosage
effects of this allele we found in CIAT accession #16141 might be observed even in tetraploid lines
retaining some functional BAHDO1 alleles. The SDP1 LOF allele occurs in a diploid U. decumbens
accession (CIAT #26308) in heterozygous form. This accession is sexual so could be crossed to
compatible diploid lines the descendants of which could be crossed to produce a homozygous
diploid line. Knock-out of SDP1 alone increases storage lipid content by many fold in vegetative
tissue of Arabidopsis, (i.e., it is not redundant with SDP1-like) (Kelly et al., 2013), so such a line could
be used to test for this effect in Urochloa genus. If successful, the line could be crossed into
tetraploid commercial breeding populations, e.g. using a chromosome doubling step.

In future, the allele mining approach we describe here could be applied to other genes and need not
be confined to alleles detrimental to molecular function. For example, candidate genes underlying
apomixis (Worthington et al., 2016) and spittlebug resistance (Ferreira et al., 2019) traits have
recently been identified in Urochloa; with improving ability to predict consequences of variants in
these, an allele mining approach could be of value. Also, for gene targets such as these where
dominant alleles may affect phenotype, candidate gene association genetics could be a useful
approach, as successfully applied for the FT gene and flowering time in Lolium perenne (Skøt et al.,
2011). As knowledge of genes improves, allele mining of diverse germplasm will become an
increasingly powerful tool to identify lines that could be beneficially brought into many crop
breeding programmes.
**Author Contributions**

RACM, JJDV and JSHH conceived the project. TKP collected samples with assistance from VC and JA and carried out RNA isolation. SJH carried out DNA sequencing and bait capture. SJH and RACM carried out bioinformatic steps. RACM wrote the manuscript with contributions from SJH, TKP, JJDV, JA, VC, PJE and JSHH.

**Acknowledgements**

This work was supported under the RCUK-CIAT Newton-Caldas Initiative “Exploiting biodiversity in Brachiaria and Panicum tropical forage grasses using genetics to improve livelihoods and sustainability”, with funding from UK’s Official Development Assistance Newton Fund awarded by UK Biotechnology and Biological Sciences Research Council (BB/R02828/1). Additional funding for this study was received from the CGIAR Research Programs on Livestock; and Climate Change, Agriculture and Food Security (CCAFS).
Figure 1. Workflow for analyses. Steps in red text are described in Methods section.
Figure 2 Number of variants in individual accessions grouped by species and ploidy. Only the 66 accessions with both RNA and DNA sequencing were used for this analysis. Ur-2 *U. ruizienis* 2x, Ud-2 and Ud-4 *U. decumbens* 2x and 4x, Ub-5 and Ub-5 *U. brizantha* 4x and 5x, Uh-6 and Uh-7 *U. humidicola* 6x and 7x, Mm-4 *M. maximus* 4x.
Figure 3: Numbers of accessions that share polymorphisms, grouped by predicted effect.
Figure 4 Three cell wall genes with LOF alleles. Read coverage for both RNAseq and DNAseq is shown for accessions indicated which carry LOF allele. Scale of coverage in number of reads and LOF polymorphism are indicated.
Figure 5. Four lipid genes with LOF alleles. Read coverage for both RNAseq and DNAseq is shown for accessions indicated which carry LOF allele. Scale of coverage in number of reads and LOF polymorphism are indicated.
References


Wijekoon C, Singer SD, Weselake RJ, Petrie JR, Chen G, Singh S, Eastmond PJ, Acharya SN. Down-regulation of key genes involved in carbon metabolism in Medicago truncatula results in increased lipid accumulation in vegetative tissue. *Crop Science* n/a.


