Limited allele-specific gene expression in highly polyploid sugarcane

Gabriel Rodrigues Alves Margarido, Fernando Henrique Correr, Agnelo Furtado, Frederik C Botha and Robert James Henry

Supplemental Methods

Read alignment and SNV calling

The *Sorghum bicolor* genome 454 v3.0.1 (Paterson et al. 2009; McCormick et al. 2018) was used as a reference to identify SNVs. Annotations from Ensembl Plants 49 (Yates et al. 2020) using BioMart v2.44.4 (Durinck et al. 2005, 2009) in R v4.0.5 (R Core Team 2021) were added to the original genome sequence and annotation, downloaded from Phytozome v13 (Goodstein et al. 2012).

Alignment of genomic reads to the sorghum genome was performed with Bowtie 2 v2.4.1 (Langmead and Salzberg 2012). We used end to end (global) alignment with the very sensitive setting, while requiring fragment sizes to be between 50 and 800 bp and further modified the minimum score function (--SCORE-MIN L,0.0,-0.8) and indel penalty (--RDG 4,1 --RFG 4,1) to deal with alignment against sorghum. After read sorting and ordering with SAMtools v1.10 (Li et al. 2009), duplicate reads were marked with Picard tools v2.23.4 (http://broadinstitute.github.io/picard). Reads were filtered to remove secondary and supplementary alignments, as well as PCR and optical duplicates. Only reads with a minimum mapping quality of three were kept, as this allows for a limited number of mismatches between the read and reference sequences.

The GATK v4.1.8.1 pipeline (DePristo et al. 2011) was used to identify SNVs and call genotypes. For doing so we first ran HaplotypeCaller (Poplin et al. 2018) individually for each sample, restricting the genomic intervals to annotated genes and with a padding of 100 bp on both sides. The ploidy was set to 12, with a heterozygosity value of 0.01 and indel-heterozygosity of 0.001. Next we ran GenomicsDBImport and GenotypeGVCFs to identify variants and initially filtered the sites to keep only biallelic SNVs. We loaded the SNVs with VariantAnnotation v1.36.0 (Obenchain et al. 2014) and filtered sites according to the following criteria: total joint depth for all genotypes between 100 and 10,000, quality score >= 50, FISHERSTRAND <= 80 and -8 < READPOSRANKSUM < 8. Monomorphic sites, which were different from the sorghum reference but fixed in our sugarcane samples were removed. We annotated the predicted impact of each SNV based on their relative genomic position with SnpEff v5.0 (Cingolani et al. 2012).

RNA sequencing, data preprocessing and alignment

The raw reads were processed with Trimmomatic v0.39 (Bolger et al. 2014) to remove Illumina adapters, allowing a maximum of two mismatches and using parameters PALINDROMECLIPTHRESHOLD 30 and SIMPLECLIPTHRESHOLD 10. The minimum Phred quality for leading and trailing bases was set to 3, and a sliding window of size four, with an average required quality score of 30, was used to remove low quality bases. The 13 leading bases were removed to account for random hexamer priming bias. After trimming reads shorter than 70 bp were removed. BBDuk v38.79 (sourceforge.net/projects/bbmap) was used to remove contaminating ribosomal RNA reads, based on the SILVA rRNA database (Quast et al. 2013) and a k-mer length of 31.

For aligning the RNA-seq reads we used the same sorghum genome reference as for the WGS data. To account for introns in the genome we used the splice-aware aligner HISAT v2.1.0 (Kim et al. 2015) with the known splice sites from the structural gene annotation, with a maximum intron length of 20 kbp, and again modified the minimum score function (--SCORE-MIN L,0.0,-0.8) and indel penalty (--RDG 4,1 --RFG 4,1).

Functional enrichment tests

To search for enrichment of gene groups among those with ASE, OrthoFinder v.2.3.12 (Emms and Kelly 2015, 2019) was used to find clusters of genes conserved in the Liliopsida. The proteomes of *Aegilops tauschii* v4.0, *Ananas comosus* v3, *Brachypodium distachyon* v3.1, *Dioscorea rotundata* v1.0, *Eragrostis curvula* v1.0, *Hordeum vulgare* v2, *Leersia perrieri* v1.4, *Musa acuminata* v1, *Miscanthus sinensis* v7.1, *Oropetium thomaeum* v1.0, *Oryza sativa* v7.0, *Panicum virgatum* v5.1, *Saccharum spontaneum*, *Setaria italica* v2.2, *Sorghum bicolor* v3.1.1, *Triticum aestivum* v2.2 and *Zea mays* v4 from Phytozome v13 (Goodstein et al. 2012) and Ensembl Plants 49 (Yates et al. 2020) were downloaded. OrthoFinder was used with default parameters and identified clusters of genes with at least one representative in each of these species. Based on this analysis we also identified groups of paralogs that are exclusive to sorghum, the reference genome, but absent from all other species. In parallel, OrthoDB v10.1 (Kriventseva et al. 2019) was used to find genes that are present in single copy in sorghum, rice and Brachypodium.

For enrichment tests SNVs with at least 50 genomic reads and 10 or more RNA-seq reads were selected, and genes were considered to show ASE when they included at least two SNVs with significant ASE. For each genotype and internode combination we tested for enrichment using Fisher's exact test, and obtained FDR-corrected p-values according to Benjamini and Hochberg (1995). In all cases enrichment was tested among Liliopsida-conserved, single copy and sorghum-exclusive paralogous genes. The depth threshold was used to remove noisy, less informative loci, but the statistical model accounted for variation in depth, and the majority of sites had higher coverage (median depths of roughly 1,000 and 200 for higher and lower coverage genotypes, respectively) (Supplemental Fig S3).

We also searched for enriched Gene Ontology terms among genes with ASE (The Gene Ontology Consortium et al. 2000). GO terms were assigned to SNV present within each annotated gene. SNVs with a minimum of 50 genomic reads and 10 RNA-seq reads were kept and genes with less than 10 SNVs removed. For each gene we calculated the median absolute deviation between the expressed allele ratio and the genomic allele ratio of its corresponding filtered SNVs. Genes were ranked according to this median deviation and GSEAPreranked v4.0.3 (Subramanian et al. 2005; Mootha et al. 2003) was used to test for enrichment of functional terms. We tested GO terms with five or more genes, with 10,000 permutations and no collapsing of gene identifiers.

References

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* **57**: 289–300.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Ruden DM, Lu X. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**: 80–92.

DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491–8.

Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W. 2005. BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**: 3439–3440.

Durinck S, Spellman PT, Birney E, Huber W. 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* **4**: 1184–1191.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**: 238.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* **16**: 157.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. 2012. Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res* **40**: D1178–D1186.

Kim D, Langmead B, Salzberg SL. 2015. HISAT: A fast spliced aligner with low memory requirements. *Nat Methods* **12**: 357–360.

Kriventseva E V, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, Zdobnov EM. 2019. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res* **47**: D807–D811.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–9.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

McCormick RF, Truong SK, Sreedasyam A, Jenkins J, Shu S, Sims D, Kennedy M, Amirebrahimi M, Weers BD, McKinley B, et al. 2018. The Sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *Plant J* **93**: 338–354.

Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, et al. 2003. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34**: 267–273.

Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, Morgan M. 2014. VariantAnnotation: A Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics* **30**: 2076–2078.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, et al. 2009. The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**: 551–556.

Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D, et al. 2018. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**: D590–D596.

R Core Team. 2021. R: A Language and Environment for Statistical Computing. https://www.r-project.org/.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**: 15545–15550.

The Gene Ontology Consortium, Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, et al. 2000. Gene Ontology: Tool for the unification of biology. *Nat Genet* **25**: 25–29.

Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, et al. 2020. Ensembl 2020. *Nucleic Acids Res* **48**: D682–D688.