The road to sorghum domestication: evidence from nucleotide diversity and gene expression patterns

Concetta Burgarella[1,2,$] Angélique Berger[1,2], Sylvain Glémin[3,4], Jacques David[2,5], Nancy Terrier[2,5], Monique Deu[1,2], David Pot[1,2]

[1] CIRAD, UMR AGAP, F- 34398 Montpellier, France
[2] Univ Montpellier, CIRAD, INRAE, Montpellier SupAgro, Montpellier, France
[3] CNRS, Univ. Rennes, ECOBIO - UMR 6553, F-35000 Rennes, France
[4] Department of Ecology and Evolution, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden
[5] INRAE, UMR AGAP, F- 34398 Montpellier, France
[$] present address: Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

Abstract

Native African cereals (sorghum, millets) ensure food security to millions of low-income people from low fertility and drought-prone regions of Africa and Asia. In spite of their agronomic importance, the genetic bases of their phenotype and adaptations are still not well understood. Here we focus on *Sorghum bicolor*, which is the fifth cereal worldwide for grain production and constitutes the staple food for around 500 million people. We leverage transcriptomic resources to address the adaptive consequences of the domestication process. A clear domestication syndrome is visible in cultivated sorghum (plant architecture, shattering, dormancy and seed weight). Gene expression and nucleotide variability were analyzed in 11 crop and 9 wild accessions. We showed a reduction of diversity in the crop pool both in nucleotide polymorphism (30%) and gene expression levels (18%), which supports the occurrence of a genetic bottleneck in the domestication history of sorghum. Nine hundred and forty nine genes were significantly differentially expressed between the two compartments. Among them, genes involved in internode elongation and auxin metabolism were upregulated in the crop, while genes linked to photosynthesis and reduction-oxidation processes were downregulated in the same pool. Coexpression network analyses revealed 21 clusters of genes sharing similar expression patterns. Four clusters (totalizing 2449 genes) were significantly enriched in differentially expressed genes between the wild and crop pools and two were also enriched in domestication and improvement genes previously identified in sorghum. These results reinforce the evidence that domestication and improvement do not only affect the behaviors of a few genes but led to a large rewiring of the transcriptome during the domestication event and the improvement process. Overall, these analyses pave the way towards the identification of key domestication genes valuable for genetic resources characterization and breeding purposes.

Introduction

Domestication is an evolutionary process in which species evolve dramatic functional and phenotypic changes under human selection for characters that fit the agricultural environment and human necessities (such as taste, yield, cultivation, harvesting and storage practices). The resulting domesticated forms are different from unmanaged wild populations and generally unable to reproduce and survive in the wild. The suite of traits that have been modified in the domesticated forms are collectively referred to as “domestication syndrome” (Allaby 2014). In plants they typically include characters related to seed dispersal (e.g. reduction of dehiscence and dormancy), plant architecture (e.g. branching) and properties of the harvested part (such as size, shape, nutritional content of seed, fruit and tubers). Around 250 plants are estimated to have experienced domestication (Dirzo and Raven 2003), with grasses (family Poaceae) having contributed at least 48 species (Glémin and Bataillon 2009).

The extensive phenotypic changes induced in domesticated species are associated with genetic changes, which can be put in evidence by comparing domesticated and wild forms. According to domestication models, an important genome wide signature of the domestication process is a genetic bottleneck experienced by the domesticated forms, which would be due to a small sampling from the wild species associated with the selection for favorable traits (Doebley *et al.* 2006; Meyer and Purugganan 2013). This general reduction of genetic diversity is expected to be stronger for annual than perennial species and in autogamous than allogamous crops (Gaut *et al.* 2015). Experimental evidence supports this hypothesis, as a reduction of genetic diversity has been found in major crop species when compared with their wild progenitors, e.g. maize (Hufford *et al.* 2013), rice (Caicedo *et al.* 2007; Huang *et al.* 2012; Nabholz *et al.* 2014), soybean (Lam *et al.* 2010) and tomato (Koenig *et al.* 2013; Sauvage *et al.* 2017).

The domesticated genome is expected to show also other genome-wide signatures when compared with the wild progenitor, such as increased linkage disequilibrium (e.g. Wright *et al.* 2005; Zhou *et al.* 2015) and increased mutation load, i.e. a genome-wide accumulation of weakly deleterious genetic variants that reduces the fitness of the domesticated forms compared to the wild counterpart. This latter phenomenon has been called ‘domestication cost’ (Lu *et al.* 2006) and has been observed in several domesticated animal and plant species (Moyers *et al.* 2018), including rice (Lu *et al.* 2006; Nabholz *et al.* 2014), maize (Wang *et al.* 2017) and sunflower (Renaut and Rieseberg 2015).

Besides genome-wide effects, the specific genomic regions involved in the phenotypic traits that distinguish domesticated from wild forms are expected to show stronger molecular genetic signatures than the rest of the genome. Although the genetic mechanisms under many complex domestication traits are still unknown, several major domestication genes have been identified along with their causative variation, which includes single nucleotide mutations, copy number polymorphism or in-del changes (Kantar *et al.* 2017; Purugganan 2019). These genes include both coding and regulatory regions, indicating that changes in expression levels may be as important as protein modifications (Meyer and Purugganan 2013, Olsen and Wendel 2015). Supporting this hypothesis, recent surveys of gene expression patterns have put in evidence significant differences between the domesticated and the wild pools (e.g. Bellucci *et al.* 2014; Page *et al.* 2019). In tomato, the rewiring of gene expression levels has been documented for some functional networks (Swanson-Wagner *et al.* 2012; Sauvage *et al.* 2017) up to shutdown complete pathways (Itkin *et al.* 2013), suggesting that the expression changes induced by domestication may have targeted metabolic pathways more than major effect genes (Sauvage *et al.* 2017).

We focus here on sorghum (*Sorghum bicolor* (L.) Moench), the world's 5th cereal for grain production. Being staple food for more than 500 million people, sorghum plays a key role in food security in Africa and Asia. Aside from providing food for human consumption, sorghum is widely cultivated in all continents as a source for feed, fiber and energy (Paterson *et al.* 2009). In the current context of climate change and resource depletion, sorghum is also a cereal with great potential for future agriculture, because it outperforms other crops under low-input and stressful conditions (Hasan *et al.* 2017). Sorghum was first domesticated in North-East Africa, likely around 6000 years ago (Winchell *et al.* 2017, 2018).

The domesticated phenotype is distinguishable from the wild forms in many morphological, physiological and phenological traits, including non-shattering grains, higher seed size, different panicle and plant architecture. Some genes and gene families have been identified as likely targets of the domestication and/or the improvement processes. Notably, the loss of seed shattering has been related to mutations in the *shattering1* (*sh1*) locus, which has likely undergone parallel selection in sorghum, rice and maize (Lin *et al.* 2012). Selection scans and phenotype-genotype association studies have identified other candidate genes. These include known cereal domestication genes Tb1 (*teosinte branched1*) and ba1 (*barren stalk1*), involved in plant architecture (Mace *et al.* 2013; Lai *et al.* 2018), as well as genes associated with grain features (e.g. Psy1, *phytoene synthase1*), photoperiod sensitivity (early maturity Ma genes) and plant height (Dw genes) (Fernandez *et al.* 2008; Mace and Jordan 2010; Mace *et al.* 2013). Genes within key biosynthesis pathways have also been highlighted, tagging in particular the metabolisms of starch (ss1 and sbe3 genes, Campbell *et al.* 2016), grain size and weight (Tao *et al.* 2017, 2018).

In spite of its importance, there is still a limited knowledge about the genetic determinants of sorghum traits and adaptive potential. Comparing the domesticated compartment with its wild progenitor is particularly informative to understand the changes that accompanied the transition from the wild to the crop. By taking the wild compartment as reference, we can quantify the genetic changes we observe in the domesticate both at the genome-wide scale and at the individual gene scale. Despite the simplicity of this approach, just a handful of studies have included wild accessions in analyses of sorghum crop diversity and evolution in a comparative manner. These works have analysed variation at allozyme markers (Aldrich *et al.* 1992), RFLP loci in the nuclear and chloroplast genomes (Aldrich and Doebley 1992), nuclear microsatellite markers (Casa *et al.* 2005; Billot *et al.* 2013) and nucleotide sequence (Hamblin *et al.* 2005; Mace *et al.* 2013). They have all reported a global loss of genetic diversity and for some of them evidence of directional selection in the genome of landraces and improved germplasm.

In this work, we address the genetic bases of the transition from wild to domesticate in sorghum by analysing the transcriptome expressed in leaves, flowers and maturing seeds. Compared to nucleotide diversity screening, transcriptome analysis provides twofold information, about changes in the protein coding sequence and changes in the levels of gene expression. Here we perform complementary analyses with three main objectives: i) to quantify the effect of domestication on nucleotide polymorphism, by looking at gene diversity and differentiation patterns; ii) to unveil the effect of domestication on patterns of gene expression and document changes in single genes and gene networks; iii) to assess the congruence and divergence of polymorphism and expression patterns to identify genes and pathways under selection and responsible for the evolution of the domesticated phenotype.

Materials and Methods

Plant material

A data-set of 20 sorghum accessions were examined in this study. Accessions were selected to represent the whole diversity of the domesticated (n=11) and the wild relative (n=9) gene pools. Details on the accessions are provided in supplementary table S1, as well as the correspondence of accession names with published works analysing accessions from this data-set (Clément *et al.* 2017; Ranwez *et al.* 2017). All the genotypes were grown in the greenhouse until the mature grain stage. Tissue samples were harvested from three organs: mature leaves (fourth rank below the flag leaf), inflorescence (at the anthesis stage) and maturing grains.

RNA preparation and sequencing

Each tissue was considered independently for RNA extraction, and RNA were then pooled for each accession prior to sequencing. RNA extraction, Illumina library preparation and sequencing conditions are detailed in Sarah *et al*. (2017). A mixture of 65% RNA from the inflorescence, 15% from leaves and 20% from grains for each accession was sequenced using Illumina mRNA-seq, paired-end protocol on a HiSeq2000 sequencer (one run for each compartment). The paired-end reads, in the illumina FASTQ format, were cleaned using cutAdapt (Martin 2011) to trim read ends of poor quality (q score below 20) and to keep only those with an average quality above 30 and a minimum length of 35 base pairs. Those data are freely available on the NCBI Sequence Read Archive database (cultivated: SAMN05277472 to SAMN05277481; wild: SAMN06052464 to SAMN06052472 and SAMN07313361).

Polymorphism analysis

To perform genotype calling, we followed the procedure of Sarah *et al*. (2017) and Nabholz *et al*. 2014. Reads were mapped on sorghum genome version 1.4 with the BWA software (Li and Durbin 2009), allowing at most three mismatches between a given read and the reference. We then excluded reads with more than two insertions/deletions (indels) or with indels larger than 5 bp or mapping on different transcripts. ORFextractor (available at http://fr.softwaresea.com/download-ORF-Extractor-10449769.htm) was used to extract coding sequences for further processing. The read2snp software (Gayral *et al.* 2013) was used to call genotypes and filter out paralogs. Following the approach of Nabholz *et al*. (2014) for non-outcrossing species, we did a first run setting the fixation index $F = 0$, then estimated F from the data and rerun the program once with the new estimated value. We note that the fixation index calculated here accounts for the mating system and potential substructure in each sample. For each individual and each position, we only kept genotypes with a minimum coverage of 10x and with posterior probability higher than 0.95. Otherwise, data was considered missing.

The structure of genetic diversity revealed through the genotype calling steps was first explored using principal component analysis (PCA) on 15,844 SNPs from 10,986 genes common to the wild and domesticated pools (function pca in package LEA v. 3.8, Frichot and François 2015). Diversity statistics per gene and group (SNP number, fixation index F, nucleotide diversity π, Tajima's D) were calculated with dNdSpiNpiS (available at https://kimura.univ-montp2.fr/PopPhyl/) and popGenome (Pfeifer *et al.* 2014). Fst between groups were also calculated with popGenome.

Differential expression analysis

Transcript expression levels have been estimated with the new-Tuxedo pipeline (Pertea *et al.* 2016). Firstly, for each accession, RNA-seq reads were mapped on the sorghum genome assembly Sbicolor_313_v3.1 using Hisat2 (Kim *et al.* 2015). Genes and transcripts were assembled and quantified with stringtie (Pertea *et al.* 2015), using the reference annotation file to guide the assembly process. The output includes expressed reference transcripts as well as any novel transcripts that were assembled. Gffcompare (https://github.com/gpertea/gffcompare) was used to compare transcripts with the reference annotation (gene and transcript predictions) and to identify new genes/transcripts. Assembling and enriched annotation files were used to estimate abundance with stringtie.

The identification of differentially expressed genes between the wild and the crop compartments was done with edgeR (Robinson *et al.* 2010) under the R environment (R Core Team 2018). Only genes assayed in at least 5 accessions and with at least 1 CPM (count per million) in each accession were included in subsequent analyses, considering that the expression level could not be reliably estimated for genes that did not meet these criteria. Since the minimum library size was 12M, one CPM corresponded to 12 reads. To account for gene-specific expression biases at the individual level a normalization step (trimmed mean of M-values (TMM) between each pair of samples) was performed (Robinson and Oshlack 2010). Diversity in expression within each of crop and wild groups was quantified with the coefficient of variation (CV) in read counts. Crop-wild differences in CV were assessed in different gene categories (candidate and non-candidate genes under selection), to assess if the CV loss is due to selection in the crop or drift associated to the domestication/improvement process, following the approach of Bellucci *et al*. (2014). Candidate genes were identified by extreme values of polymorphism and differentiation statistics, non-candidates were all the other genes. The significance of the difference was assessed with 1000 random resamplings. The structure of the sample based on the gene expression levels was explored with a PCA on the normalized counts of 24646 genes with the function dudi.pca of the R package ade4 (Dray and Dufour 2007).

To identify differentially expressed (DE) genes, gene-wise exact tests for differences in the means between the two groups (wild and domesticated pools) were performed. A False Discovery Rate approach was applied to correct for multiple tests. Thresholds of 0.05 and 0.01 FDR were considered to declare DE genes. At 0.05 FDR, we declared a number of DE genes twofold higher than at 0.01 FDR, but the ratio of up- and downregulated genes was qualitatively similar, as well as the main categories of Gene Ontology annotation (see below). We therefore only report the results for the stringent FDR of 0.01.

Polymorphism/expression comparisons

We compared patterns of gene expression and polymorphism to identify genetic signatures potentially linked to selection and the process of domestication. At the genome-wide scale, we expect a stronger purifying selection against deleterious mutations in highly expressed genes, thus we expect that π *and* π_N/π_S would correlate negatively with expression level.

In addition, if significant differential expression is driven by directional selection in the wild or the crop pool, we expect that DE genes would have different signatures of intra-group polymorphism (lower π, π_N/π_S, extreme values of Tajima's D) and/or intergroup differentiation (higher F_{ST}) than non-DE genes. We tested for this hypothesis by checking if DE genes were enriched for values of diversity and differentiation outside the 95% quantile interval for each statistic.

Enrichment of different functional categories

We examined the functional characterization of DE genes and assessed if particular gene categories or functions were overrepresented in DE vs non-DE genes, which could be the result of selection. The categories assessed included the Gene Ontology (GO) categories, Sorghum transcription factors (retrieved from Plant Transcription Factor Database http://planttfdb.cbi.pku.edu.cn/index.php?sp=Sbi), stably expressed genes in sorghum (Shakoor *et al.* 2014), different metabolic pathways (Shakoor *et al.* 2014; Rhodes *et al.* 2014, 2017), domestication and improvement genes from published studies (Mace *et al.* 2013; Lai *et al.* 2018). In general, enrichment in the different functional categories was tested with Fisher's exact test on 2x2 contingency tables. GO annotation per gene was obtained from Sobic 3.1 genome annotation). The correspondence among gene models for the different versions of the Sorghum genome were retrieved with the conversion tool of the Grass Genome Hub (https://grass-genome-hub.southgreen.fr/pseudomolecules-converter). Functional annotation of novel assembled genes was done using Blast2GO (Götz *et al.* 2008). GO enrichment analysis was performed with topGO (Alexa *et al.* 2006) using the "classic" algorithm and Weighted Fisher's test to generate p-values.

Gene expression networks

Gene co-expression networks were built using the WGCNA R package (Langfelder and Horvath 2008, 2012) using the normalized (TMM) and filtered expression data set (as previously described). Networks were built using the "signed" networkType parameter, enabling to capture the direction of the expression variation and grouping genes with the same direction variation in gene expression. This parameter is advised to identify biologically meaningful modules (van Dam *et al.* 2018). According to the mean connectivity and the scale free topology index curves obtained, a power of 20 was used for this analysis. A mergeheight parameter of 0.25 and a minimum module size of 30 were also considered. The modules obtained from this analysis were then tested for their enrichment in genes presenting differential expression patterns between the wild and crop compartments.

Results

Nucleotide diversity patterns

For the analysis of polymorphism patterns the whole transcriptome of the wild and crop pools were analysed separately. We obtained 11 397 and 13 231 genes that passed the quality filters in the wild and the domesticated pool respectively, 10 986 genes being common to both pools. Although a higher number of genes were retrieved for the domesticated pool, a higher total number of polymorphic sites was called in the wild pool (46 147 *versus* 41 179, respectively), suggestive of higher diversity in the wild compartment (see below). We found 15 844 sites from 5240 genes polymorphic in both compartments (Table 1). The estimated fixation index for the domesticated and the wild compartments was F=0.99 and F= 0.42 respectively, which is in agreement with a higher rate of self-fertilization generally observed in the crop (e.g. Djè *et al.* 2004; Muraya *et al.* 2011).

We first looked at the structure of the sample with a principal component analysis. We retrieved two genetic groups corresponding well to the wild and the crop pools (Figure 1). The only exception was the crop accession ec9 (SSM1057) that clustered with the wild sample. This accession belongs to the race *guinea margaritiferum*, which is cultivated but that is genetically close to the wild pool (Sagnard *et al.* 2011; Billot *et al.* 2013; Mace *et al.* 2013).

The crop compartment harbored 20-30% less nucleotide diversity than the wild compartment, in terms of number of polymorphic sites and nucleotide diversity (Kolmogorov - Smirnov test $p < 2.2.e-16$; Table 1). Non-synonymous polymorphism (π_N) was six times lower than synonymous polymorphism (π_S) in both crop and wild pools, in agreement with previous estimates for cultivated sorghum from Hamblin *et al*. (2006). We did not find evidence of accumulation of slightly deleterious mutations in the domesticated pool at the genome scale, since the the ratio of non-synonymous to synonymous polymorphism (π_N/π_S) across genes was similar in the wild and domesticated pools (mean values 0.16 and 0.15, respectively, Table 1).

We also estimated Tajima's *D*, which measures the departure of nucleotide diversity from neutrality due to selective or demographic processes. Crop and wild pools showed comparable close-to-zero mean Tajima's D values across genes (Table 1). Similar values were found genome-wide (Mace *et al.* 2013). The crop pool showed higher variance of Tajima's D gene estimates, suggesting a higher number of genes departing from neutral equilibrium (Table 1).

Assuming the extant wild gene pool as reference, the ratio π_{crop}/π_{wild} may point to genes whose diversity has changed during domestication. Since strong positive selection is expected to reduce diversity in the genomic region linked to the causal allele, $\pi_{crop}/\pi_{wild} < 1$ would characterize genes under positive selection in the crop. In particular, we can consider as strong domestication candidates the genes monomorphic in the crop and polymorphic in the wild. Conversely, $\pi_{crop}/\pi_{wild} > 1$ may indicate a diversity increase in the crop during domestication, either due to other selective dynamics (e.g. balancing selection or diversifying selection across varieties) or to a relaxation of selection in the crop compared to the wild pool. Median π_{crop}/π_{wild} was 0.54 (Table 1) whereas 1734 genes showed $\pi_{crop}/\pi_{wild} = 0$ and 2500 had $\pi_{crop}/\pi_{wild} > 1$.

Finally, we estimated the genetic differentiation between crop and wild by calculating F_{ST} on a gene by gene basis. Overall differentiation was quite low (mean F_{ST}=0.07, Table 1), and did not change when genes with less than 5 polymorphic sites were excluded. Our estimate is lower than previous estimates (F_{ST}=0.13 on SSRs, Casa *et al*. 2005). Two reasons can be proposed here to explain this discrepancy. Firstly, the coding regions analyzed in this study could have experienced a lower neutral drift than the microsatellite markers used by Casa *et al*. (2005). Secondly, it can emerge from the different compositions of the panels used. Indeed, we included a guinea margaritiferum accession in our cultivated pool whereas such type of accession was, to the best of our knowledge, not used in the analysis of Casa *et al*. (2005).

Expression diversity patterns

The analysis of gene expression levels was done on 24 646 genes meeting the criterion of at least 1 read count per million in at least 5 samples. Overall, our sampling strategy (3 different tissues) allowed accessing more than 70 % of the sorghum transcriptome. To compare the diversity in expression between the crop and the wild pools we performed, similarly with the nucleotide diversity information, a principal component analysis and calculated the coefficient of variation (CV) on normalized read counts. PC1 separated the two, cultivated and wild, pools, explaining 25.8% of the read counts diversity (Figure 1). The crop pool harbored 84% of the wild pool expression variability (mean CV_{crop}=0.63 [95% CI 0.13; 1.98], mean CV_{wild}=0.75 [95% CI 0.16; 2.04]), which suggests that domestication in sorghum involved also a loss of expression diversity, besides the loss of polymorphism reported above. We found that the CV reduction is higher for candidate genes harboring signature of selection at the nucleotide level (at least 25%) irrespective of the statistic used to identify them. This signal is stronger for genes with extremely high crop-wild

differentiation. The CV loss was 34%, 32% and 29% in the 99%, 95% and 90% percentile of Fst values respectively. These values are significantly higher (p-value < 0.001) than the CV reduction obtained with 1000 random samples of genes (Figure 2).

We tested for significantly different expression levels between the wild and crop pools, to identify genes whose change in expression may have been driven by selection under domestication. We found 949 differentially expressed (DE) genes at a FDR of 1% (Table S2). The difference in expression levels corresponded to |logFC| > 0.55 (logFC range: -9 to 14). Interestingly, among the DE genes, 82% were over-expressed in the wild compartment, while we did not observe any bias in the direction of expression change in the whole gene set (χ^2 test, p < 2e-16). This result is qualitatively similar for a FDR of 5% (among 2291 DE genes, 72% are overexpressed in the wild pool).

Comparison of expression and nucleotide diversity patterns

We first compared expression and nucleotide patterns across all genes. As expected if higher expressed genes experience stronger selection against new deleterious mutations, gene polymorphism was negatively correlated with expression, either for π_S, π_N or π_N/π_S (Table 2). For π_N/π_S, the correlation was less negative in the crop, which may be the effect of weaker purifying selection and domestication load experienced in the crop pool.

Then we looked at specific features of DE vs non-DE genes to identify signatures associated with the differential expression. DE genes showed lower values of π_N and π_N/π_S than non-DE genes in both crop and wild pools (Table 3), suggesting the action of selection on DE genes, although the distributions were significantly different only for π_N in the crop pool, (Kolmogorov-Smirnov test p-value = 0.008). DE genes were also characterized by significantly lower crop/wild π ratios than non-DE genes (Kolmogorov-Smirnov test, p-value = 0.02 and 0.04 for π_N and π_S, respectively). This difference was largely driven by genes that are monomorphic in the crop pool, which may represent genomic regions targeted by human-driven directional selection. DE genes also showed higher Tajima's D values than non-DE genes, but the difference was significant only in the wild pool (Kolmogorov-Smirnov test, p-value = 0.024). We did not find significant differences between DE and non-DE genes for π_S or crop-wild F_{ST} (Table 3).

Biological functions of differentially expressed genes

We examined the functional characterization of DE genes and assessed if particular gene categories or functions were overrepresented. Gene ontology (GO) analysis of DE genes revealed that terms related to photosynthesis were largely overrepresented among genes with reduced expression in the crop relative to the wild sorghum (n= 51 genes, Table S3). Different metabolic functions were instead enriched in DE genes upregulated in the crop, the more recurrent being related to cytoskeleton organization and intracellular membrane transport (Table S4). Genes upregulated in the crop were enriched also in GO terms associated with Auxin responsive factors (ARF, GO:0005086 and GO:0032012).

We tested DE enrichment in Sorghum transcription factors (Plant Transcription Factor Database http://planttfdb.cbi.pku.edu.cn/index.php?sp=Sbi) and genes from the Sorghum expression Atlas (Shakoor *et al.* 2014). We found that DE genes were significantly depleted of transcription factors (p-value= 0.009) and of genes reported to be stably expressed across tissues (p-value= 3.462e-05) compared to non DE genes.

We also compared our DE gene list with the candidate genes involved in trait domestication and improvement of sorghum according to previous studies. We found a total of 59 genes harboring either a domestication or improvement signature as reported in Mace *et al*. (2013) and Lai *et al*. (2018) (Table S5).

Among the DE genes, we found several genes belonging to known metabolic pathways in sorghum. Among them there were two genes involved in the sucrose metabolism and 4 involved in the phenylpropanoyl monolignol pathway according to Shakoor *et al*. (2014), but these metabolisms were not more represented in DE than in non DE genes. Several genes were previously known to be involved in the biochemical pathways associated with grain development and filling phases (Table S5), such as starch synthesis (Sobic.001G100000, Campbell *et al.* 2016), polyphenols synthesis (4 genes, Rhodes *et al.* 2014) and more generally grain composition (9 genes, Rhodes *et al.* 2017).

Gene co-expression networks

Gene co-expression network analyses revealed 21 clusters of genes sharing similar expression patterns (Table S6). Four of these clusters were enriched in differentially expressed genes between the wild and cultivated compartments. Three were characterized by an over-expression and a higher variability in the wild pool (yellow : 2051 genes, magenta : 173 genes and cyan : 92 genes)) and one was characterized by an over-expression and a higher variability in the crop pool (greenyellow: 133 genes) (Table S7). Among these four clusters, two were also enriched in domestication and improvement candidate genes previously identified through whole genome sequencing of 44 sorghum accessions including 7 wild and weedy genotypes (Mace *et al.* 2013) (Table S8 and S9). More specifically, the yellow and greenyellow modules were enriched in domestication genes whereas the greenyellow module was also enriched in improvement genes. Interestingly, the greenyellow module that harbored an enrichment in improvement genes presented an over-expression in the cultivated pool compared to the wild pool. In addition the black (271 genes) and the lightcyan (45 genes) modules harbored expression patterns revealing the specific properties of the guinea margaritiferum accession (SSM1057). In the black module SSM1057 harbored an over-expression in comparison to all the other accessions analyzed, whereas in the lightcyan module SSM1057 presented an expression pattern more closely related to the wild pool (although it corresponds to the cultivated pool).

Discussion

In this study, we have performed a comparison of nucleotide and expression patterns between wild and cultivated forms of sorghum, to contribute to the understanding of the genetic circumstances of wild-to-crop transition in this major cereal. Much of the genome is not expressed. By targeting the full transcriptome, we focused on genes actually expressed, narrowing the analysis of interest to the part of diversity potentially most relevant. We found different lines of evidence documenting the genetic changes likely associated with the domestication and improvement process in sorghum.

Down-regulation and diversity loss during sorghum domestication

Our genome-wide screen of expressed genes showed a loss of diversity in the cultivated pool up to 30%, both in nucleotide and in expression diversity, supporting the general model of sorghum domestication from a reduced representation of the wild pool (e.g. Hamblin *et al.* 2006) and/or serial founder effects during crop evolution (Allaby *et al. 2019*; Wang *et al.* 2017). Our estimate of

nucleotide diversity loss was higher than previous estimates for domesticated sorghum based on nuclear microsatellite markers (e.g. 14%, Casa *et al.* 2005), but lower than the loss recorded in other main crops with selfing or mixed mating systems. For instance, it was estimated a 62% nucleotide diversity reduction in rice (Caicedo *et al.* 2007), 72% in common beans (Bitocchi *et al.* 2013) and 50% in soybean (Kim *et al.* 2020). The crop showed also lower expression variation than the wild accessions, a finding common to other crops such as maize (Lemmon *et al.* 2014) and common bean (Bellucci *et al.* 2014) but also animal domesticates (Liu *et al.* 2019). We demonstrated that the loss of expression diversity is not merely a consequence of the loss of nucleotide diversity. In fact, the genes with strongest signatures of selection showed greater loss of expression variation, which suggests that selection has driven strongest changes in particular genes, a finding common to other domesticates (Bellucci *et al.* 2014; Liu *et al.* 2019).

However, in contrast with a previous work on sorghum (Smith *et al.* 2018), and evidence in other plant and animal domesticates (Lu *et al.* 2006; Moyers *et al.* 2018), we didn't find clear evidence of accumulation of deleterious load in the crop pool. Interestingly, a recent study in soybean found a genome wide reduction of the deleterious alleles in domesticated soybean relative to wild soybean (Kim *et al.* 2020). These authors suggested that the increase of self crossing associated with domestication may have helped purging the deleterious load, partly by exposing to selection deleterious variants in homozygous state, a process suggested also in cassava (Ramu *et al.* 2017) and grape (Zhou *et al.* 2017). A similar process can be at play in sorghum, where the domesticated pool generally shows a higher selfing rate than the wild pool (as observed in this study with the F index and also in Djè *et al.* 2004; Muraya *et al.* 2011; Sagnard *et al.* 2011). Further, gene flow from wild relatives (and vice versa) is considered frequent across the domesticate distribution (Barnaud *et al.* 2009; Mutegi *et al.* 2011, 2012; Sagnard *et al.* 2011) and has probably strongly contributed to the genetic makeup of the crop pool during the domestication process. Gene flow from the wild compartment has most likely counterbalanced crop drift and divergence genome-wide.

In nature, changes of gene expression are responsible for a great part of phenotypic divergence not explained by amino acid divergence (Carroll 2008). It is also likely that part of the phenotypic evolution associated with the domestication process has involved gene regulation without changes in the coding sequence. This appears to be the case in sorghum, where many DE genes do not show different nucleotide diversity patterns between the crop and the wild pools, suggesting that domestication left a footprint in expression but not in protein composition for these genes.

Our differential expression analysis indicated that domestication largely favored the down-regulation of expression of multiple genes in the crop (82% of DE genes). In agreement with our findings in sorghum, common bean and eggplant landraces showed downregulation of expression when compared with wild relatives, although the analysis was limited to the leaf transcriptome in both studies (Bellucci *et al.* 2014; Page *et al.* 2019). In maize, Lemmon *et al*. (2014), analyzing three different tissues (leaf, ear, stem) found the opposite trend, with differentially expressed genes showing higher expression of the maize allele more often than the teosinte allele. In contrast, Sauvage *et al*. (2017) found a similar number of down- and up-regulated genes in cultivated tomatoes based on the analysis of vegetative and reproductive tissues (leaf, flower and fruits). The lack of consistent patterns across different crops suggest species-specific processes and mechanisms for the evolution of the domesticated phenotype.

Biological significance of Differentially Expressed genes and genes harboring putative signature of selection at the nucleotide level

In this study, we found 949 differentially expressed genes between the wild and cultivated

compartments taking advantage of leaves, flowers and maturing seeds of sorghum. Among the genes significantly down-regulated in the crop compartment, genes with function associated with photosynthesis were disproportionately represented, suggesting that photosynthesis may be a key metabolic pathway contributing to the sorghum domestication syndrome. A number of studies have investigated the effect of domestication on photosynthesis metabolism, suggesting a shift in plant functional strategies over the course of domestication. For instance, higher net photosynthetic rate has been found in modern varieties of wheat compared to ancestral ones (Roucou *et al.* 2018) and in cultivated cassava compared with wild forms (Pujol *et al.* 2008), pointing to selection for increased photosynthetic efficiency during domestication and improvement. Less conclusive results were reported by González *et al*. (1995) in the common bean. Most of the cultivated and wild accessions studied did not show differences in CO2 exchange rate, except for Andean landraces, who showed a decrease in CO2 exchange rate potentially associated with locally specific selective pressures. Another line of evidence suggests that antioxidative responses have been weakened by domestication. For instance, it has been shown that ascorbate, an antioxidant widely associated with photosynthetic functions and stress tolerance, was reduced by domestication in some crop species, likely as a tradeoff of selection for higher fruit size and yield (Gest *et al.* 2013).

Modifications of the photosynthetic metabolism may be also a consequence of changes in the selective pressure experienced by the crop under some stresses. For instance, photosynthetic activity in leaves is overall reduced under drought stress (Osakabe *et al.* 2014). Some authors have suggested that the transition from wild habitats with unpredictable and scarcer water availability to agricultural environments with more regular and abundant water supplies led to a relaxation of selection for water use efficiency while favouring the increase of CO2 uptake and evaporative cooling (Milla *et al.* 2013 and references therein). It can be argued that domestication has led to a relaxation of selection for photosynthesis efficiency under drought stress, because of higher water availability in agricultural ecosystems than in the wild. This scenario would be consistent with positive selection for relatively higher photosynthetic capacity in cultivars adapted to water deficit environments, such as the case of upland rice compared with lowland rice (Zhang *et al.* 2016).

We found that genes upregulated in the crop were enriched in Auxin responsive factors, which are very likely drivers of domestication-related phenotype. Genes involved in the response to auxin stimulus contribute to abiotic stress response in sorghum (Wang *et al.* 2010) and more generally is part of the auxin signaling machinery determining the development of branches and flowers in major crops such as maize (Galli *et al.* 2015). Auxin-responsive genes have been found under selection in sorghum, maize, rice and domesticated emmer (Meyer and Purugganan 2013; Mace *et al.* 2013; Avni *et al.* 2017).

Altogether, our scan for selection signatures and differential expression identified several individual candidate genes particularly interesting for their potential role in the domestication and diversification of cultivated sorghum. Some of these genes are already known to contribute to important phenotypic traits in sorghum or have been previously identified as sorghum domestication genes. For instance, the gene Sobic.006G067700 (dwarf2) has been reported to control the internode length in cultivated sorghum and to have pleiotropic effects on panicle length, seed weight and leaf area (Hilley *et al.* 2017). We found that this gene was upregulated in the crop pool (logFC=-0.95, FDR=4.14E-03), consistently with its involvement in the evolution of the cultivated phenotype. Among genes down-regulated in the crop, we found Sobic.004G272100, which encodes a phosphoribulokinase, an enzyme that catalyzes a key step in carbon fixation as part of the Calvin cycle (dark phase of photosynthesis). This gene has been found to be under

selection in sorghum and maize (Lai *et al.* 2018). We found it was down-regulated in the crop pool (logFC = 1.86; FDR= 0.0096), suggesting that human selection has targeted the reduction in expression level for this gene. In addition to the genes previously detected in the litterature with evidence of signature of selection, our studies also provided a large list of additional genes that contributed or have been impacted by the domestication and improvement process. To further explore the list of genes that potentially contributed to these two key steps of crop evolution we studied their co-expression patterns.

Domestication and improvement steps involve major rewirings of gene expression

Gene co-expression network analyses allowed the identification of 4 clusters of genes (totalizing 2449 genes) significantly enriched in differentially expressed genes between the wild and cultivated pools. In addition, two of these clusters were also enriched in domestication and improvement genes previously identified in sorghum according to their nucleotide diversity patterns. These results reinforce the conclusion that can be drawn from the differential expression analysis indicating that domestication and improvement do not only affect the behaviors of a few "domestication or improvement" genes but led to large rewirings of the transcriptome during the domestication event and the improvement process, as also underlined in tomato (Sauvage *et al.* 2017) and cotton (Gallagher *et al.* 2020). Whereas the "gene based" differential expression analyses can be limited by the sequencing depth achieved for each specific gene to detect significant expression differences, the co-expression gene network approach provides a complementary approach allowing to identify additional potentially relevant genes.

In the evolutionary history of sorghum, the guinea-margaritiferum emergence remains until now quite unclear. Recent works proposed that they can result from a second and more recent domestication in West Africa (Sagnard *et al.* 2011; Mace *et al.* 2013). Our results indicated that at the nucleotide level, as reported by Mace *et al*. (2013) and Morris *et al*. (2013), the guinea-margaritiferum accession presented a high similarity with the wild pool. However, at the expression level, it behaved essentially as a cultivated genotype. Such result, taking into account the relatively recent domestication of the margaritiferum sub-race, would indicate that the domestication process induced a rapid rewiring of the gene co-expression network and a much slower response at the global nucleotide level. It is also interesting to note that among the two clusters of genes for which an atypical behavior of the guinea-margaritferum was observed in comparison to the other cultivated genotypes (the black and lightcyan modules), the gene Sobic.001G341700, which have already been identified in sorghum as contributing to the grain size (Tao *et al.* 2017, 2020) has been identified. This gene is an ortholog of Os DEP1 which has been proved to impact grain yield in rice (Huang *et al.* 2009).

Conclusion

This work highlighted the major impact of domestication not only on the nucleotide diversity but also on the transcriptome landscape. Indeed, it revealed a major rewiring of the gene co-expression networks between the wild and the crop pools. In addition to these global observations, these analyses also identified specific metabolic pathways (photosynthesis, auxin related mechanisms) that are likely to have contributed to the transition from the wild to the crop pools in sorghum. Finally, our results also contributed to a better understanding of the specific attributes of the guinea-margaritiferum genotypes within the sorghum genetic diversity highlighting the specific role of the transcriptome regulation in the emergence of the "cultivated" ideotype.

Fundings

This study was initiated in the framework of ARCAD (http://www.arcad-project.org), a project funded by Agropolis Fondation under the reference ID ARCAD 0900-001. In addition, CB benefited from the support of the Biomass for the Future project (ANR-11-BTBR-0006-BFF), funded by the French National Research Agency (ANR). CB also received support from the Marie Curie grant program.

Acknowledgments

We wish to thank Pierre Roumet (INRAE, Montpellier, France) for useful inputs on the role of photosynthesis in domestication.

Author Contributions Statement

CB: Performed the analyses, interpreted the results, wrote the initial version of the manuscript, aggregated the suggestions of the co-authors and finalized the manuscript
AB: Managed the production of the tissues, took care of the RNA extraction and took part to the gene expression analyses
SG: Wrote and led the ARCAD project together with JD, contributed to the SNP calling section and contributed to the improvement of the manuscript
JD: Wrote and led the ARCAD project together with SG and contributed to the improvement of the manuscript
NT: Contributed to the interpretation of the differential expression results, contributed to the improvement of the manuscript
MD: Contributed to the interpretation of the nucleotide diversity results, and took part to the revision of the manuscript
DP: Contributed to the writing of the sorghum section of the ARCAD project and led the sorghum section of the BFF project. Selected the genotypes that have been used and took part in the analyses and writing of the manuscript.

Conflict of Interest Statement

The authors have declared that no competing interests exist.

References

Aldrich PR, Doebley J. **1992**. Restriction fragment variation in the nuclear and chloroplast genomes of cultivated and wild Sorghum bicolor. *Theoretical and Applied Genetics* **85**: 293–302.
Aldrich PR, Doebley J, Schertz KF, Stec A. **1992**. Patterns of allozyme variation in cultivated and wild Sorghum bicolor. *Theoretical and Applied Genetics* **85**: 451–460.
Alexa A, Rahnenführer J, Lengauer T. **2006**. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**: 1600–1607.
Allaby RG. **2014**. Domestication Syndrome in Plants In: Smith C, ed. *Encyclopedia of Global Archaeology*. New York, NY: Springer, 2182–2184.
Allaby RG, Ware RL, Kistler L. 2019. *A re-evaluation of the domestication bottleneck from archaeogenomic evidence. Evolutionary Applications* **12**: 29–37.

Avni R, Nave M, Barad O, *et al.*2017. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* **357**: 93–97.
Barnaud A, Deu M, Garine E, *et al.*2009. A weed–crop complex in sorghum: The dynamics of genetic diversity in a traditional farming system. *American Journal of Botany* **96**: 1869–1879.
Bellucci E, Bitocchi E, Ferrarini A, *et al.*2014. Decreased Nucleotide and Expression Diversity and Modified Coexpression Patterns Characterize Domestication in the Common Bean. *The Plant Cell*: tpc.114.124040.
Billot C, Ramu P, Bouchet S, *et al.*2013. Massive Sorghum Collection Genotyped with SSR Markers to Enhance Use of Global Genetic Resources. *PLOS ONE* **8**: e59714.
Bitocchi E, Bellucci E, Giardini A, *et al.*2013. Molecular analysis of the parallel domestication of the common bean (Phaseolus vulgaris) in Mesoamerica and the Andes. *New Phytologist* **197**: 300–313.
Caicedo AL, Williamson SH, Hernandez RD, *et al.*2007. Genome-Wide Patterns of Nucleotide Polymorphism in Domesticated Rice. *PLOS Genetics* **3**: e163.
Campbell BC, Gilding EK, Mace ES, *et al.*2016. Domestication and the storage starch biosynthesis pathway: signatures of selection from a whole sorghum genome sequencing strategy. *Plant Biotechnology Journal* **14**: 2240–2253.
Carroll SB. **2008**. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell* **134**: 25–36.
Casa AM, Mitchell SE, Hamblin MT, *et al.*2005. Diversity and selection in sorghum: simultaneous analyses using simple sequence repeats. *Theoretical and Applied Genetics* **111**: 23–30.
Clément Y, Sarah G, Holtz Y, *et al.*2017. Evolutionary forces affecting synonymous variations in plant genomes. *PLOS Genetics* **13**: e1006799.
van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. **2018**. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics* **19**: 575–592.
Dirzo R, Raven PH. **2003**. Global State of Biodiversity and Loss. *Annual Review of Environment and Resources* **28**: 137–167.
Djè Y, Heuertz M, Ater M, Lefèbvre C, Vekemans X. **2004**. In situ estimation of outcrossing rate in sorghum landraces using microsatellite markers. *Euphytica* **138**: 205–212.
Doebley JF, Gaut BS, Smith BD. **2006**. The Molecular Genetics of Crop Domestication. *Cell* **127**: 1309–1321.
Dray S, Dufour A-B. **2007**. The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software* **22**: 1–20.
Fernandez MGS, Hamblin MT, Li L, Rooney WL, Tuinstra MR, Kresovich S. **2008**. Quantitative Trait Loci Analysis of Endosperm Color and Carotenoid Content in Sorghum Grain. *Crop Science* **48**: 1732–1743.
Frichot E, François O. **2015**. LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution* **6**: 925–929.
Gallagher JP, Grover CE, Hu G, Jareczek JJ, Wendel JF. **2020**. Conservation and Divergence in Duplicated Fiber Coexpression Networks Accompanying Domestication of the Polyploid Gossypium hirsutum L. *G3 (Bethesda, Md.)* **10**: 2879–2892.
Galli M, Liu Q, Moss BL, *et al.*2015. Auxin signaling modules regulate maize inflorescence architecture. *Proceedings of the National Academy of Sciences* **112**: 13372–13377.
Gaut BS, Díez CM, Morrell PL. **2015**. Genomics and the Contrasting Dynamics of Annual and Perennial Domestication. *Trends in Genetics* **31**: 709–719.
Gayral P, Melo-Ferreira J, Glémin S, *et al.*2013. Reference-free population genomics from next-generation transcriptome data and the vertebrate–invertebrate gap. *PLoS Genetics* **9**: e1003457.
Gest N, Gautier H, Stevens R. **2013**. Ascorbate as seen through plant evolution: the rise of a successful molecule? *Journal of Experimental Botany* **64**: 33–53.
Glémin S, Bataillon T. **2009**. A comparative view of the evolution of grasses under domestication. *New Phytologist* **183**: 273–290.
González A, Lynch J, Tohme JM, Beebe SE, Macchiavelli RE. **1995**. Characters Related to Leaf Photosynthesis in Wild Populations and Landraces of Common Bean. *Crop Science* **35**: 1468–1476.
Götz S, García-Gómez JM, Terol J, *et al.*2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research* **36**: 3420–3435.

Hamblin MT, Casa AM, Sun H, *et al.*2006. Challenges of Detecting Directional Selection After a Bottleneck: Lessons From Sorghum bicolor. *Genetics* **173**: 953–964.
Hamblin MT, Salas Fernandez MG, Casa AM, Mitchell SE, Paterson AH, Kresovich S. **2005**. Equilibrium processes cannot explain high levels of short- and medium-range linkage disequilibrium in the domesticated grass Sorghum bicolor. *Genetics* **171**: 1247–1256.
Hasan SA, Rabei SH, Nada RM, Abogadallah GM. **2017**. Water use efficiency in the drought-stressed sorghum and maize in relation to expression of aquaporin genes. *Biologia Plantarum* **61**: 127–137.
Hilley JL, Weers BD, Truong SK, *et al.*2017. Sorghum Dw2 Encodes a Protein Kinase Regulator of Stem Internode Length. *Scientific Reports* **7**: 4616.
Huang X, Kurata N, Wei X, *et al.*2012. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**: 497–501.
Huang X, Qian Q, Liu Z, *et al.*2009. Natural variation at the DEP1 locus enhances grain yield in rice. *Nature Genetics* **41**: 494–497.
Hufford MB, Lubinksy P, Pyhäjärvi T, Devengenzo MT, Ellstrand NC, Ross-Ibarra J. **2013**. The Genomic Signature of Crop-Wild Introgression in Maize. *PLoS Genet* **9**: e1003477.
Itkin M, Heinig U, Tzfadia O, *et al.*2013. Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science (New York, N.Y.)* **341**: 175–179.
Kantar MB, Nashoba AR, Anderson JE, Blackman BK, Rieseberg LH. **2017**. The Genetics and Genomics of Plant Domestication. *BioScience* **67**: 971–982.
Kim D, Langmead B, Salzberg SL. **2015**. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**: 357–360.
Kim M-S, Lozano R, Kim JH, *et al.*2020. Purging of deleterious mutations during domestication in the predominant selfing crop soybean. *bioRxiv*: 2020.03.12.989830.
Koenig D, Jiménez-Gómez JM, Kimura S, *et al.*2013. Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. *Proceedings of the National Academy of Sciences* **110**: E2655–E2662.
Lai X, Yan L, Lu Y, Schnable JC. **2018**. Largely unlinked gene sets targeted by selection for domestication syndrome phenotypes in maize and sorghum. *The Plant Journal* **93**: 843–855.
Lam H-M, Xu X, Liu X, *et al.*2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat Genet* **42**: 1053–1059.
Langfelder P, Horvath S. **2008**. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**: 559.
Langfelder P, Horvath S. **2012**. Fast R Functions for Robust Correlations and Hierarchical Clustering. *Journal of statistical software* **46**.
Lemmon ZH, Bukowski R, Sun Q, Doebley JF. **2014**. The Role of cis Regulatory Evolution in Maize Domestication. *PLOS Genetics* **10**: e1004745.
Li H, Durbin R. **2009**. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**: 1754–1760.
Lin Z, Li X, Shannon LM, *et al.*2012. Parallel domestication of the Shattering1 genes in cereals. *Nature Genetics* **44**: 720–724.
Liu W, Chen L, Zhang S, *et al.*2019. Decrease of gene expression diversity during domestication of animals and plants. *BMC Evolutionary Biology* **19**: 19.
Lu J, Tang T, Tang H, Huang J, Shi S, Wu C-I. **2006**. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends in Genetics* **22**: 126–131.
Mace ES, Jordan DR. **2010**. Location of major effect genes in sorghum (Sorghum bicolor (L.) Moench). *Theoretical and Applied Genetics* **121**: 1339–1356.
Mace ES, Tai S, Gilding EK, *et al.*2013. Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. *Nature Communications* **4**: ncomms3320.
Martin M. **2011**. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10–12.
Meyer RS, Purugganan MD. **2013**. Evolution of crop species: genetics of domestication and diversification. *Nature Reviews Genetics* **14**: 840–852.
Milla R, de Diego-Vico N, Martín-Robles N. **2013**. Shifts in stomatal traits following the domestication of plant species. *Journal of Experimental Botany* **64**: 3137–3146.
Morris GP, Ramu P, Deshpande SP, *et al.*2013. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proceedings of the National Academy of*

Sciences **110**: 453–458.
Moyers BT, Morrell PL, McKay JK. **2018**. Genetic Costs of Domestication and Improvement. *Journal of Heredity* **109**: 103–116.
Muraya MM, de Villiers S, Parzies HK, *et al.*2011. Genetic structure and diversity of wild sorghum populations (Sorghum spp.) from different eco-geographical regions of Kenya. *Theoretical and Applied Genetics* **123**: 571–583.
Mutegi E, Sagnard F, Labuschagne M, *et al.*2012. Local scale patterns of gene flow and genetic diversity in a crop–wild–weedy complex of sorghum (Sorghum bicolor (L.) Moench) under traditional agricultural field conditions in Kenya. *Conservation Genetics* **13**: 1059–1071.
Mutegi E, Sagnard F, Semagn K, *et al.*2011. Genetic structure and relationships within and between cultivated and wild sorghum (Sorghum bicolor (L.) Moench) in Kenya as revealed by microsatellite markers. *Theoretical and Applied Genetics* **122**: 989–1004.
Nabholz B, Sarah G, Sabot F, *et al.*2014. Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice (Oryza glaberrima). *Molecular Ecology* **23**: 2210–2227.
Osakabe Y, Osakabe K, Shinozaki K, Tran L-SP. **2014**. Response of plants to water stress. *Frontiers in Plant Science* **5**: 86.
Page A, Gibson J, Meyer RS, Chapman MA. **2019**. Eggplant Domestication: Pervasive Gene Flow, Feralization, and Transcriptomic Divergence. *Molecular Biology and Evolution* **36**: 1359–1372.
Paterson AH, Bowers JE, Bruggmann R, *et al.*2009. The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**: 551–556.
Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. **2016**. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols* **11**: 1650–1667.
Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. **2015**. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33**: 290–295.
Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, Lercher MJ. **2014**. PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. *Molecular Biology and Evolution* **31**: 1929–1936.
Pujol B, Salager J-L, Beltran M, Bousquet S, McKey D. **2008**. Photosynthesis and Leaf Structure in Domesticated Cassava (Euphorbiaceae) and a Close Wild Relative: Have Leaf Photosynthetic Parameters Evolved Under Domestication? *Biotropica* **40**: 305–312.
Purugganan MD. **2019**. Evolutionary Insights into the Nature of Plant Domestication. *Current Biology* **29**: R705–R714.
R Core Team. **2018**. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Ramu P, Esuma W, Kawuki R, *et al.*2017. Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nature Genetics* **49**: 959–963.
Ranwez V, Serra A, Pot D, Chantret N. **2017**. Domestication reduces alternative splicing expression variations in sorghum. *PLOS ONE* **12**: e0183454.
Renaut S, Rieseberg LH. **2015**. The Accumulation of Deleterious Mutations as a Consequence of Domestication and Improvement in Sunflowers and Other Compositae Crops. *Molecular Biology and Evolution* **32**: 2273–2283.
Rhodes DH, Hoffmann L, Rooney WL, *et al.*2017. Genetic architecture of kernel composition in global sorghum germplasm. *BMC Genomics* **18**: 15.
Rhodes DH, Hoffmann L, Rooney WL, Ramu P, Morris GP, Kresovich S. **2014**. Genome-wide association study of grain polyphenol concentrations in global sorghum [Sorghum bicolor (L.) Moench] germplasm. *Journal of Agricultural and Food Chemistry* **62**: 10916–10927.
Robinson MD, McCarthy DJ, Smyth GK. **2010**. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140.
Robinson MD, Oshlack A. **2010**. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11**: R25.
Roucou A, Violle C, Fort F, Roumet P, Ecarnot M, Vile D. **2018**. Shifts in plant functional strategies over the course of wheat domestication. *Journal of Applied Ecology* **55**: 25–37.
Sagnard F, Deu M, Dembélé D, *et al.*2011. Genetic diversity, structure, gene flow and evolutionary relationships within the Sorghum bicolor wild–weedy–crop complex in a western

African region. *Theoretical and Applied Genetics* **123**: 1231–1246.
Sarah G, Homa F, Pointet S, *et al.*2017. A large set of 26 new reference transcriptomes dedicated to comparative population genomics in crops and wild relatives. *Molecular Ecology Resources* **17**: 565–580.
Sauvage C, Rau A, Aichholz C, *et al.*2017. Domestication rewired gene expression and nucleotide diversity patterns in tomato. *The Plant Journal* **91**: 631–645.
Shakoor N, Nair R, Crasta O, Morris G, Feltus A, Kresovich S. **2014**. A Sorghum bicolor expression atlas reveals dynamic genotype-specific expression profiles for vegetative tissues of grain, sweet and bioenergy sorghums. *BMC Plant Biology* **14**: 35.
Smith O, Nicholson WV, Kistler L, *et al.*2018. A domestication history of dynamic adaptation and genomic deterioration in sorghum. *bioRxiv*: 336503.
Swanson-Wagner R, Briskine R, Schaefer R, *et al.*2012. Reshaping of the maize transcriptome by domestication. *Proceedings of the National Academy of Sciences* **109**: 11878–11883.
Tao Y, Mace E, George□Jaeggli B, *et al.*2018. Novel Grain Weight Loci Revealed in a Cross between Cultivated and Wild Sorghum. *The Plant Genome* **11**: 170089.
Tao Y, Mace ES, Tai S, *et al.*2017. Whole-Genome Analysis of Candidate genes Associated with Seed Size and Weight in Sorghum bicolor Reveals Signatures of Artificial Selection and Insights into Parallel Domestication in Cereal Crops. *Frontiers in Plant Science* **8**: 1237.
Tao Y, Zhao X, Wang X, *et al.*2020. Large-scale GWAS in sorghum reveals common genetic control of grain size among cereals. *Plant Biotechnology Journal* **18**: 1093–1105.
Wang S, Bai Y, Shen C, *et al.*2010. Auxin-related gene families in abiotic stress response in <Emphasis Type="Italic">Sorghum bicolor</Emphasis>. *Functional & Integrative Genomics* **10**: 533–546.
Wang L, Beissinger TM, Lorant A, Ross-Ibarra C, Ross-Ibarra J, Hufford MB. **2017**. The interplay of demography and selection during maize domestication and expansion. *Genome Biology* **18**: 215.
Winchell F, Brass M, Manzo A, *et al.*2018. On the Origins and Dissemination of Domesticated Sorghum and Pearl Millet across Africa and into India: a View from the Butana Group of the Far Eastern Sahel. *African Archaeological Review* **35**: 483–505.
Winchell F, Stevens CJ, Murphy C, Champion L, Fuller DorianQ. **2017**. Evidence for Sorghum Domestication in Fourth Millennium BC Eastern Sudan: Spikelet Morphology from Ceramic Impressions of the Butana Group. *Current Anthropology* **58**: 673–683.
Wright SI, Bi IV, Schroeder SG, *et al.*2005. The Effects of Artificial Selection on the Maize Genome. *Science* **308**: 1310–1314.
Zhang Z-F, Li Y-Y, Xiao B-Z. **2016**. Comparative transcriptome analysis highlights the crucial roles of photosynthetic system in drought stress adaptation in upland rice. *Scientific Reports* **6**: 19349.
Zhou Z, Jiang Y, Wang Z, *et al.*2015. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature Biotechnology* **33**: 408–414.
Zhou Y, Massonnet M, Sanjak JS, Cantu D, Gaut BS. **2017**. Evolutionary genomics of grape (Vitis vinifera ssp. vinifera) domestication. *Proceedings of the National Academy of Sciences* **114**: 11715–11720.

List of Tables

Table 1. Characteristics of the data sets, coding sequence polymorphism and differentiation in wild and cultivated pools of *Sorghum bicolor*.
Table 2. Correlations of polymorphism estimates and expression levels
Table 3. Comparison of polymorphism patterns between differentially expressed (DE) and non differentially expressed (non-DE) genes between wild and cultivated accessions of *Sorghum bicolor*

List of Figures

Figure 1. Principal Component analysis of polymorphism (top) and expression data (bottom) for wild (green) and crop (red) accessions of *Sorghum bicolor*. es: wild; ec: crop

Figure 2. Observed (red triangles) and expected (black boxplots) wild to crop reduction of the coefficient of variation in expression (CV). Observed values were calculated on genes with extreme values of crop-wild differentiation (Fst outliers) at different Fst thresholds: 99%, 95% and 90% upper percentiles. Expected values were calculated on 1000 random samples of genes with sample size equal to the number of ouliers at each Fst threshold. Sample sizes: 92, 457, 913 respectively. CV reduction was calculated as 1-(mean CV_{CROP} / mean CV_{WILD})

Supplementary material

Table S1. Origin of Sorghum accessions used in this study
Table S2. Annotation, expression and nucleotide statistics for genes differentially expresses between crop and wild pools at 0.01 FDR
Table S3. Gene Ontology term enrichment in genes that were significantly over-expressed in the wild pool at FDR < 0.01
Table S4. Gene Ontology term enrichment in genes over-expressed in the crop pool at FDR < 0.01
Table S5. Differentially expressed genes found in common with other studies
Table S6: Membership of the 24646 genes to the 21 modules identified through the WGCNA analysis
Table S7: Enrichment of the modules detected through the WGCNA analysis in genes differentially expressed between the wild and cultivated pools
Table S8: Enrichment of the modules detected through the WGCNA analysis in genes detected by Mace et al., 2013 as harbouring signature of selection compatible with a domestication event
Table S9: Enrichment of the modules detected through the WGCNA analysis in genes detected by Mace et al., 2013 as harbouring signature of selection compatible with an improvement event

Table 1. Characteristics of the data sets, coding sequence polymorphism and differentiation in wild and cultivated pools of *Sorghum bicolor*.

	Wild	**Crop**	**Wild and crop**
No. of accessions	9	11	
No. of raw contigs	26274	26552	
No. of retained contigs	11397	13231	10986
Length of retained coding sequences Mean [95% quantiles]	1610 [326; 3882]	1596 [312; 3830]	
No. of polymorphic sites [a]	46147	41179	15844 [c]
F [95%CI] [a]	0.42 [0.408; 0.424]	0.99 [0.987; 0.989]	
mean π [b] **[95%CI]**	1.6 [0; 6.7]	1.1*** [0; 5.4]	
mean π_S [a] **[95%CI]**	4.1 [3.97; 4.17]	3.3*** [3.15; 3.35]	
mean π_N [a] **[95%CI]**	0.7 [0.64; 0.67]	0.5*** [0.48; 0.51]	
mean π_N/π_S [a] **[95%CI]**	0.16 [0.156; 0.166]	0.15 [0.147; 0.158]	
Tajima's D [b] **Median [95% quantiles]**	-0.4 [-1.6; 1.6]	-0.39*** [-1.8; 1.8]	
π_{CROP}/π_{WILD} Median [95% quantiles]			0.54 [0; Inf]
Mean F_{ST} [b] **[95% quantiles]**			0.07 [0; 0.42]

F: fixation index (Weir & Cockerham 1984).
π: nucleotide diversity (x 10^{-3}). All polymorphic positions (π), synonymous sites (π_S), non-synonymous sites (π_N)
F_{ST}: pairwise differentiation (Hudson, R. R., M. Slatkin, and W.P. Maddison (1992))
[a] calculated with dNdSpiNpiS, after UTR sequence removal and reads2snp aligned site cleaning
[b] calculated with popGenome, after UTR sequence removal and reads2snp aligned site cleaning
[c] SNPs polymorphic in the two pools are located in 5240 contigs
*** Kolmogorov-Smirnov test for difference between crop and wild distributions, two-side p-value < 2.2. e-16

Table 2. Correlations of polymorphism estimates and expression levels

		Slope	**adjR^2**	**p-value**
π_S	Crop	-0.39	0.13	<2e-16
	Wild	-0.26	0.06	<2e-16
π_N	Crop	-0.54	0.24	<2e-16
	Wild	-0.44	0.18	<2e-16
π_N/π_S	Crop	-0.14	0.014	<2e-12
	Wild	-0.19	0.03	<2e-16

π: nucleotide diversity (x 10^{-3}). synonymous sites (π_S), non-synonymous sites (π_N)

Table 3. Comparison of polymorphism patterns between differentially expressed (DE) and non differentially expressed (non-DE) genes between wild and cultivated accessions of *Sorghum bicolor*

	DE genes (n=949)		Non-DE genes (n=23698)	
	Wild	**Crop**	**Wild**	**Crop**
No. upregulated genes	773c	176	11287	12410
mean π_S [a] [95%CI]	4.67 [0.03; 20.5]	3.73 [0.03; 24.9]	4.09 [0.03; 22.7]	3.28 [0.03; 21.5]
mean π_N [a] [95%CI]	0.61 [0.03; 4.1]	0.48** [0.03; 3.8]	0.65 [0.03; 4.5]	0.5 [0.03; 3.4]
mean π_N/π_S [a] [95%CI]	0.13 [0; 1.1]	0.13 [0; 1]	0.16 [0; 1.1]	0.15 [0; 0.97]
Tajima's D [b] Median [95% quantiles]	-0.25* [-1.5; 1.5]	-0.19 [-1.6; 1.7]	-0.31 [-1.5; 1.4]	-0.26 [-1.7; 1.6]
$\pi_{N\text{-}CROP}/\pi_{N\text{-}WILD}$ Median [95% quantiles]	0.799		0.801	
$\pi_{S\text{-}CROP}/\pi_{S\text{-}WILD}$ Median [95% quantiles]	0.767		0.769	
Mean F_{ST} [b] [95% quantiles]	0.09 [0; 0.4]		0.07 [0; 0.3]	

[a] calculated on 469 genes with polymorphism estimates

[b] calculated on 9867 genes with polymorphism estimates

[c] The number of genes overexpressed in the wild pool is significantly higher among DE than non-DE genes (χ 2 test p-value < 2.2e-16)

* distribution significantly different in DE vs non-DE genes (Kolmogorov-Smirnov test *p-value < 0.05, ** p-value < 0.01)

π: nucleotide diversity (x 10^{-3}). synonymous sites (π_S), non-synonymous sites (π_N)

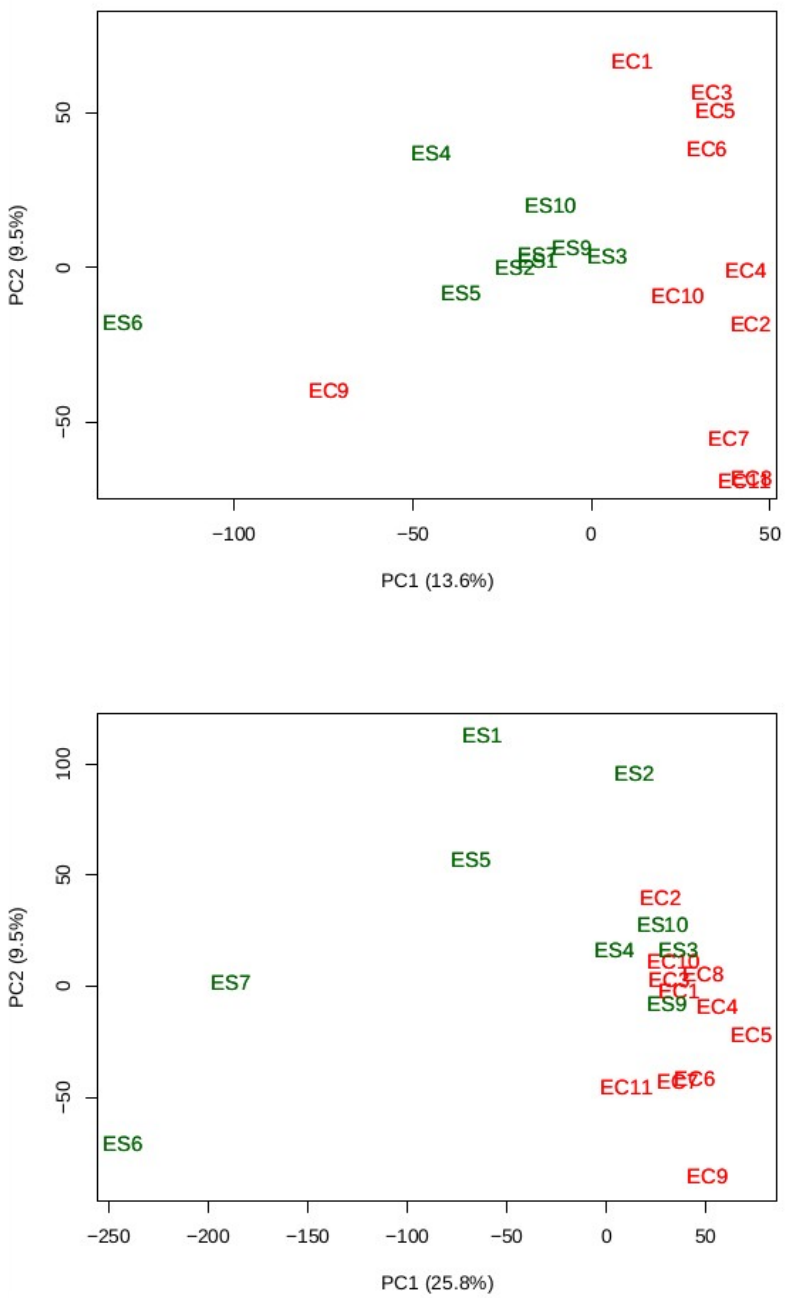

Figure 1. Principal Component analysis of polymorphism (top) and expression data (bottom) for wild (green) and crop (red) accessions of *Sorghum bicolor*. ES: wild; EC: crop

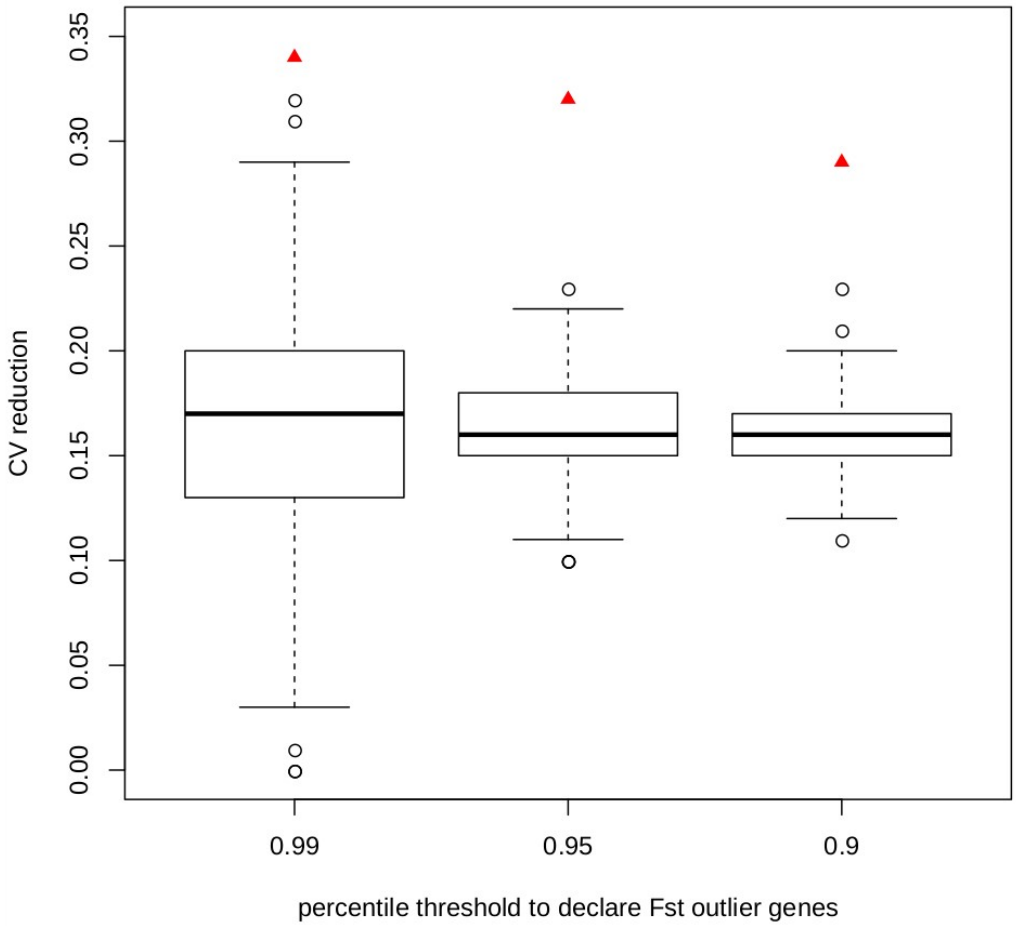

Figure 2. Observed (red triangles) and expected (black boxplots) wild to crop reduction of the coefficient of variation in expression (CV). Observed values were calculated on genes with extreme values of crop-wild differentiation (Fst outliers) at different Fst thresholds: 99%, 95% and 90% upper percentiles. Expected values were calculated on 1000 random samples of genes with sample size equal to the number of ouliers at each Fst threshold. Sample sizes: 92, 457, 913 respectively. CV reduction was calculated as 1-(mean CV_{CROP} / mean CV_{WILD})