PROFESSOR LAURA ROSSINI (Orcid ID : 0000-0001-6509-9177)

DR AGOSTINO FRICANO (Orcid ID : 0000-0003-3715-5834)

Article type : Original Article

Segmental duplications are hot spots of copy number variants affecting barley gene content

Gianluca Bretani[1], Laura Rossini[1], Chiara Ferrandi[2], Joanne Russell[3], Robbie Waugh[3], Benjamin Kilian[4†], Paolo Bagnaresi[5], Luigi Cattivelli[5] and Agostino Fricano[5*]

[1] Università degli Studi di Milano – DiSAA, Via Celoria 2, 20133 Milano, Italy

[2] Parco Tecnologico Padano, Via Einstein, Loc. C.na Codazza, 26900 Lodi, Italy

[3] James Hutton Institute, Invergowrie, Dundee, DD2 5DA, UK

[4] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstrasse 3, 06466 Gatersleben, Germany

[5] Council for Agricultural Research and Economics - Research Centre for Genomics & Bioinformatics, Via San Protaso 302, 29017 Fiorenzuola d'Arda (PC), Italy

† Current address, Global Crop Diversity Trust, Platz der Vereinten Nationen 7, 53113 Bonn, Germany

* Correspondence: agostino.fricano@crea.gov.it

Running title

SDs are hot spots of gene CNVs in barley

Keywords

Barley, Copy number variants, Segmental duplications, Exome sequencing

Summary

Copy number variants (CNVs) are pervasive in several animal and plant genomes and contribute to shaping genetic diversity. In barley, there is evidence that changes in gene copy number underlie important agronomic traits. The recently released reference sequence of barley represents a valuable genomic resource for unveiling the incidence of CNVs that affect gene content and identifying sequence features associated with CNV formation.

Using exome sequencing and read count data, we detected 16,605 deletions and duplications that affect barley gene content by surveying a diverse panel of 172 cultivars, 171 landraces, 22 wild relatives and other 32 uncategorized domesticated accessions. The quest for segmental duplications (SDs) in the reference sequence revealed many low-copy repeats, most of which overlap predicted coding sequences. Statistical analyses revealed that the incidence of CNVs increases significantly in SD-rich regions, indicating that these sequence elements act as hot spots for the formation of CNVs.

This study delivers a comprehensive genome-wide study of CNVs affecting barley gene content and implicates SDs in the molecular mechanisms that lead to the formation of this class of CNVs.

Introduction

Copy number variants (CNVs) are a class of unbalanced structural changes within genomes, which represent either a gain of extra sequence copies (duplications or insertions), or a loss of genetic material (deletions) in individuals of the same species (Alkan, Coe and Eichler, 2011). In the human genome, CNVs were generally defined as deletions, insertions and duplications of DNA sequences longer than 1 kb (Feuk, Carson and Scherer, 2006), although small structural changes of 50 bp or larger are now also considered CNVs (Alkan, Coe and Eichler, 2011; Girirajan, Campbell and Eichler, 2011).

While several studies in plants have analysed genomic variability in terms of single nucleotide polymorphisms (SNPs), investigations of the CNV rate, diversity and impact on genomic variation are lagging behind. For example, years of empirical breeding and selection of crops narrowed the number of SNP variants in the cultivated gene pool (Kilian *et al.*, 2007; Fricano *et al.*, 2009), but it is still unclear whether this process might also have eroded CNV diversity. In barley, the contribution of CNVs in shaping genetic diversity is largely unknown: to date systematic analyses for identifying short CNVs have been carried out on a very limited panel of domesticated and wild accessions using a gene-space assembly (Mayer *et al.*, 2012; Muñoz-Amatriaín, Steven R Eichten, *et al.*, 2013).

Genome-wide surveys leading to the discovery of thousands of CNVs revealed a ubiquity of deletions and duplications in maize, tale cress, rice and switchgrass, (Springer *et al.*, 2009; Debolt, 2010; Swanson-wagner *et al.*, 2010; Evans *et al.*, 2015; Bai *et al.*, 2016). Beyond affecting genome structure, CNVs have the potential to modulate or create new gene functions. There is evidence that CNVs along with other structural variants (SVs) play key roles in plant adaptive evolution as well as in human diseases (Freeman, Perry and Feuk, 2006; Kim *et al.*, 2008; Evans *et al.*, 2015; Pinosio *et al.*, 2016; Prunier *et al.*, 2017). In both plant and animal kingdoms, genes exhibiting CNVs are related to defense, biotic and abiotic stress responses (Conrad *et al.*, 2010; Clop, Vidal and Amills, 2012; Pinosio *et al.*, 2016; Prunier *et al.*, 2017). In barley, the genetic dissection of boron-toxicity tolerance demonstrated that duplications of *Bot1* underlie this trait (Sutton *et al.*, 2007), while duplications of *HvFT1* are tied to earlier flowering and have an overriding effect on the vernalization mechanism (Loscos *et al.*, 2014). In wheat, duplications of *Vrn-1A* and *Ppd-1B* were demonstrated to affect vernalization requirement and photoperiod response, respectively (Díaz *et al.*, 2012). Apart from these notable examples, the incidence and the functions of genes exhibiting CNVs are still unknown.

Segmental duplications (SDs) (also termed “low-copy repeats”), are stretches of high complexity DNA sequences longer than 1 kb, which are repeated several times in the genome with nucleotide identity higher than 90% (Eichler, 2001). Genome analyses and the creation of high quality reference sequences of plant and animal species have shown that SDs are common elements of genomes (Pagel *et al.*, 2004; Sharp *et al.*, 2005; Innan and Kondrashov, 2010; Giannuzzi *et al.*, 2011; Zhang *et al.*, 2017). In barley, annotation of the reference sequence revealed that more than 75% of genes belong to families with multiple members, suggesting that duplications of DNA sequences contributed to shaping both gene content and function (Mascher *et al.*, 2017). For instance, the reference sequence of barley cultivar (cv) “Morex” contains five complete genes of *amy1* family, four of which share more than 99.8% nucleotide identity, computed considering intron and exon sequences (Mascher *et al.*, 2017). The abundance of gene families with multiple members hints that low-copy repeats could extend beyond the coding portion of the barley genome and play a fundamental role in shaping CNVs.

Several mammalian genome studies showed that SDs are hotspots of genome instability as they predispose chromosomes to rearrangements, providing templates for non-allelic homologous recombination (NAHR) events (Sharp *et al.*, 2005; Kim *et al.*, 2008; Dittwald *et al.*, 2013; Zhang *et al.*, 2015). Based on the distribution of SDs in the human genome, it was suggested that recent SDs could play a role in the formation of specific classes of CNVs via NAHR (Sharp *et al.*, 2005; Freeman, Perry and Feuk, 2006; PJ Hastings, James R Lupski, 2010). Beyond this mechanism, other types of processes that lead to CNV formation have been proposed, including non-homologous DNA repair (PJ Hastings, James R Lupski, 2010). This class of molecular mechanisms includes non-homologous end-joining (NHEJ), breakage micro-homology-mediated end joining (MMEJ), template switching due to fork stalling or replication slippage and micro-homology-mediated break-induced replication (MMBIR) (PJ Hastings, James R Lupski, 2010). In barley, a portion of short CNVs have a sequence signature of being formed by non-homologous DNA repair (Muñoz-Amatriaín, Steven R Eichten, *et al.*, 2013), although the mechanisms that generate longer CNVs are still unknown.

In this study, we examined the diversity and distribution of CNVs that affect barley gene content. We used exome capture sequencing data from a panel of 397 diverse barley accessions to assess the occurrence and distribution of CNVs across the barley genome. Leveraging the newly created reference sequence of the barley cv “Morex” (Mascher *et al.*, 2017), we show that CNVs occur preferentially in SD-rich regions.

Results

Identification and distribution of CNVs affecting barley gene content

To identify the genome-wide occurrence of gene duplications and deletions we employed a detection strategy based on exome capture sequencing of a panel of 397 (of 403) diverse accessions that have been described previously (Bustos-Korts *et al.*, 2019), which includes 172 cultivars, 171 landraces, 22 wild relatives and other 32 domesticated accessions for which the categorization as cultivar or landrace was questionable. (Supporting Information 1; Table S1). Target regions used to design the exome capture probes were mapped to the reference sequence of barley cv "Morex" (Mascher *et al.*, 2017), which allowed us to establish that the target space covers 170,725 exons or sequence intervals. Overall, the captured sequences encompass 61.3 Mb of non-overlapping genome intervals (Supporting Information 1; Table S4), in accordance to previous estimates computed using the gene-space assembly of barley (Mascher *et al.*, 2013). For computing sequence coverage, only properly mapped paired-ends (PE) reads were considered and on average 24.6 M PE per sample were counted, leading to an average sequencing depth of 40X over the 170,725 captured sequences. Analysis of the average per-target coverage computed across the panel of 397 accessions indicated that 80% of captured sequences show a sequencing depth larger than 5X, which ensured enough coverage for subsequent analyses.

For each sample, properly mapped PE reads were counted within the genome coordinates of the 170,725 capture sequences. The resulting read count data were fitted in a beta-binomial model and used to build optimized reference sets for detecting CNVs using ExomeDepth (Plagnol *et al.*, 2012). As current algorithms for detecting CNVs based on read count data are prone to output results with unsatisfactory levels of type I error (Tan *et al.*, 2014), additional procedures were adopted to increase the confidence of genetic variant calling. First, an average per-target analysis was carried out to remove sites with coverage below 5X, as with this sequencing depth it is challenging to distinguish biases introduced with sequence capture from actual duplications and deletions. The output read count matrix was subsequently used for detecting CNVs. These were categorized based on whether they exhibited a significantly higher or lower number of reads than expected. As our pipeline cannot reliably quantify the number of copies relative to the reference sequence, we collectively refer to these genetic variant groups as duplications or deletions, respectively. Second, duplications and deletions detected in less than three barley accessions were discarded. Overall, this procedure allowed us to call 1,037,381 duplications and deletions over the whole panel of 397 accessions and unveiled that 17.6% of the 170,725 captured sequences exhibit changes in copy number. As captured targets are exons, contiguous duplications or deletions detected in each sample were merged and 197,407 CNV calls were inferred (Supporting Information 1; Table S2). These were then mapped to 16,605 physical positions (CNV sites) across the seven barley chromosomes (Supporting Information 1; Table S2). On average, 497 CNVs per barley sample were detected.

A two-pronged strategy was pursued to assess the reliability of our CNV calling pipeline and estimate the residual type I error. As a first step, a CNV-based phylogeny of the 397 barley accessions was computed using neighbor-joining (NJ) method and Euclidean distance (Figure 1). The resulting phylogeny showed separate clusters of two-row and six-row accessions (Figure 1A) and of wild and domesticated accessions (Figure 1B), reflecting the history of empirical breeding and selection of the genetic material. Similarly, the projection of tree tips onto a world map showed that the barley accessions investigated in this study cluster according to their geographic origin (Figure 1C), demonstrating that our CNV phylogeny was consistent with that obtained using SNPs (Bustos-Korts *et al.*, 2019).

The non-stochastic clustering of barley accessions in the CNV-based phylogeny indicated that CNV detection based on read count data generated reliable calls. In order to further assess the level of type I error, we selected 37 random CNVs, which were subsequently tested by PCR in 150 of the genotypes using primer pairs designed to target detected duplications and deletions (Supporting Information 1; Table S3). For these 37 CNVs, structural changes were correctly identified in 142 out 150 samples (96.6%), demonstrating that CNVs were reliably identified. A very large fraction of the detected CNVs were present in the population at low frequency, although some deletions had a frequency higher than 40 % across the whole panel of accessions (Figure 2).

On average, using the barley cv 'Morex' reference sequence, the deletions affecting barley gene content were estimated to be 3.81-fold relative to the duplications, spanning from a minimum value of 3.45 of chromosome 1H to a maximum value of 4.20 of chromosome 4H (Table 1).

To assess whether specific barley chromosomes are preferentially enriched in CNVs, the raw number of duplications and deletions detected in each chromosome was normalized relative to the length of per-chromosome captured sequences (Supporting Information 1, Table S4). The density of CNVs, measured as number of deletions or duplications per Mb of captured sequences, was computed to highlight the different incidence of CNV frequency across the coding sequences of barley chromosomes (Table 2). The density of deletions showed large variations as in chromosome 1H 256.04 deletions per Mb of captured sequences were computed, while in chromosome 4H the deletion density was 102.62 (Table 2). A similar trend was observed for duplication densities: in chromosome 1H 74.24 duplications per Mb were computed, while chromosome 4H showed paucity of CNVs with 24.41 duplications per Mb (Table 2).

To test whether the low rate of CNV density observed in chromosome 4H departs significantly from the rates of other chromosomes, CNV densities were modelled as Poisson distributions and tested to assess whether pairs of CNV densities were different. *P* values of the pairwise Poisson's tests revealed that CNV densities were significantly different and that the rate for chromosome 4H was significantly lower than that of the remaining barley chromosomes (Table 3).

The average density of CNVs affecting gene content across all accessions, cultivars and landraces showed that barley wild relatives, and to certain extent landraces, contain a significantly larger fraction of the deletion diversity compared to the cultivars, and this trend was observed in all barley chromosomes (Figure 3). Conversely, the pattern of duplication densities across all barley chromosomes does not show statistically significant differences in landraces and cultivars (Figure 3).

Functional impact of CNVs affecting barley gene content

To obtain insight into the biological and evolutionary implications of CNVs, the whole set of sequences used for designing exome capture probes was annotated using gene ontology (GO) terms. Using a homology-based approach (Conesa and Gotz, 2008), 155,235 out 287,462 sequences (~54 %) used for designing exome capture probes were annotated with GO terms (Mascher *et al.*, 2013). The GO terms of this set of 155,235 sequences were subsequently associated to the barley genes in which captured sequences were unambiguously mapped. With this approach, CNVs were annotated with 4985, 927 and 2679 GO terms of the three domains “biological process”, “cellular component” and “molecular function”, respectively. Categorization of these GO terms using the high-level summary of functions implemented in the GO Slim terms (McCarthy *et al.*, 2006) showed that a large fraction of genes exhibiting changes in copy number are involved in transporter, transferase and hydrolase activities (Figure 4A). Moreover, the examination of GO Slim terms pointed out that genes showing changes in copy number are involved in shaping cellular and membrane components (Figure 4B) (Supporting Information 1; Table S5) and in metabolic and cellular processes (Figure 4C) (Supporting Information 1; Table S5).

To assess the incidence of over-represented GO terms in duplicated and deleted genes, a GO enrichment analysis was carried out considering the whole set of barley genes for which the GO annotation was retrieved. Considering a false discovery rate (FDR) threshold of 0.01, computed using Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995), 193 GO terms were found over-represented in the set of duplicated and deleted genes (Figure 5) (Supporting Information 1; Table S6). GO enrichment analysis showed that genes with kinase, polysaccharide binding and ADP binding functions are more prone to be duplicated or deleted in barley (Figure 5A). Similarly, in duplicated and deleted genes the enrichment analysis uncovered GO terms of the "Cellular Component" domain related to "integral component of membrane" (Figure 5B). Overrepresented GO terms of the "Biological Process" domain and related to functions involved in defense response, DNA integration and protein phosphorylation were also identified in genes showing copy number changes (Figure 5C) (Supporting Information 1; Table S6).

Similarly, a GO enrichment analysis was carried out considering the set of duplicated and deleted genes that were detected exclusively in wild accessions to assess the functional categories of genes exhibiting CNVs that were lost during domestication (Figure 6). This analysis showed that the reduction of CNV diversity during the domestication process leaded to the loss of CNVs affecting genes involved in queuine tRNA-ribosyl-transferase and protein kinase activity (Figure 6A) as well as in cell wall components (Figure 6B). Overrepresented GO terms of the "Biological Process" domain and related to functional categories involved in protein phosphorylation, regulation of stomatal closure and cellular response to nitric oxide were also identified (Figure 6C).

Revisiting of earlier reported CNVs using barley reference sequence

The extent of barley gene CNVs was previously investigated in a limited panel of domesticated and wild accessions using the gene space assembly (Mayer *et al.*, 2012) along with comparative genome hybridization (CGH) technology (Muñoz-Amatriaín, Steven R. Eichten, *et al.*, 2013). These data were revisited in light of the barley reference sequence to lift over the genome coordinates of earlier reported structural variants, which were subsequently compared with the pattern of gene CNVs detected with ES in this study.

As a first step, the whole set of 115,003 contigs used for designing CGH probes (Muñoz-Amatriaín, Steven R. Eichten, *et al.*, 2013) was mapped against the reference sequence (Mascher *et al.*, 2017), and the mapping positions of these contigs were compared along the genome coordinates of ES targeted sequences. Overall, CGH probes target 228,603 non-overlapping chromosome intervals and 46.04 Mb of the barley reference sequence compared to the 170,725 chromosome intervals and 61.3 Mb of ES probes. The CGH and ES targeted regions overlap for 46,814 chromosome intervals, which span 6.33 out 61.3 Mb (10.3 %) of sequences analysed with exome capture technology: although ES and CGH probes were designed using two similar sets of contig sequences, CGH probes cover a small subset of the sequence captured with ES.

As the panel of accessions analysed using ES does not include the whole set of genetic material analysed with CGH (Muñoz-Amatriaín, Steven R. Eichten, *et al.*, 2013), the comparison of CNVs detected with these two technologies was limited to sites in which deletions and duplications were identified. Overall, 8,588 out 33,653 CNV sites identified with CGH and lifted over the barley reference sequence overlap or partially overlap with the 16,605 CNV sites identified with ES (Supplementary Information; Figure 1). The same comparison carried out with the unfiltered dataset of CNV detected with ES revealed that 13,369 overlapping SV sites were identified with both technologies (Supplementary Information; Figure 2). Although the use of different panels of genotypes limits this comparison, the analysis showed that a large fraction of CNV sites detected with ES were previously identified with CGH technology.

Identification and nature of SDs in barley genome

Identification of SDs in the reference sequence of barley cv 'Morex' (Mascher *et al.*, 2017) was pursued adopting a methodology based on sequence similarity search of high complexity regions. After masking interspersed repeats and low complexity regions of the reference sequence using the curated annotation of barley repetitive elements (Wicker *et al.*, 2017), the reference sequence was aligned against itself using chunks of 250 kb as queries to identify high similarity regions. Subsequently, data were parsed to exclude alignment pairs of query sequences matched against themselves and alignments shorter than 1 Kb.

Considering stretches of high complexity repeats with at least 95% identity, 20,853 SDs were identified across the seven barley chromosomes, which encompass circa 40,6 Mb and cover 0.89 % of the genome size. The length distribution (Figure 7A) showed that SDs spanning from 1 kb to 2kb are the most abundant in all chromosomes, while chromosomes 2H and 5H are the most SD-rich (Figure 7A).

Among these SDs, 12,631 and 9,114 have nucleotide identity of 98% and 99%, respectively and represent a subset of SDs that were recently fixed in the barley reference sequence (Table 4).

The density of SDs indicated that the ends of chromosome arms contain more SDs and this trend was observed for all chromosomes (Figure 7B). To unlock the nature of these SDs, their genomic coordinates were compared with the high and low confidence annotations of barley: 5,743 out 20,853 SDs fully or partially overlap high confidence genes, while the remaining SDs are not part of the high confidence annotated gene content. Considering the low confidence annotation (Mascher *et al.*, 2017), 2,714 out 20,853 SDs overlap chromosome intervals in which genes with annotation of unknown function or without functional annotation were detected (Mascher *et al.*, 2017). These findings reflect previous estimates pointing out that a large fraction of barley genes come from duplication events that shaped gene families with multiple members (Mascher *et al.*, 2017).

As the distribution of SDs in barley chromosomes (Figure 7B) shows the same pattern of the predicted coding sequences (Mascher *et al.*, 2017), an association analysis between these genomic regions was carried out based on permutation tests to assess if SDs overlap predicted coding regions more than expected. The average distance of SDs with their closest gene is 47 kb (Figure 8A; green vertical line), while the expected lower bound of the average distance under a random distribution of genomic features is circa 105 kb (Figure 8A; red vertical line), corroborating the finding that SDs and genes are strictly associated in the barley genome. The analysis unveiled that SDs and predicted coding sequences are strictly associated as the 5,743 overlaps between these genomic regions (Figure 8B; green vertical line) are significantly higher than the upper bound of expected overlaps under a random distribution (Figure 8B; red vertical line).

CNVs co-occur with SDs identified in the barley reference sequence

Pioneering studies on structure and function of the human genome pointed out that CNV abundance increases in SD-rich sequence intervals, and SD-mediated NAHR was suggested as a possible mechanism of CNV formation (Freeman, Perry and Feuk, 2006; Goidts *et al.*, 2006; Perry *et al.*, 2006). To assess whether in barley SDs are hot spots for the formation of CNVs, Spearman rank correlation coefficients were computed between the SDs and the CNVs detected in the panel of 397 accessions. SDs were binned into increasing sequence intervals (from 40 kb to 2 Mb) and their associations with the number of CNVs detected in the panel of 397 accessions and mapped within the same bins were examined computing Spearman rank correlation coefficients between these two structural features.

The values of Spearman rank correlation coefficients were finally computed as function of bin sizes (Figure 7C), which show high and statistically significant correlations between SDs and CNVs when bin size equal or larger than 1.5 Mb are used for computation (rank correlation higher than 0.7) (Figure 7C). These high values of rank correlation imply that a monotonic function ties SDs and CNVs and that SD-rich sequence intervals of the reference sequence are those regions that are more prone to gain extra copies or lose DNA sequences. Similarly, an association analysis of the sites where CNVs were detected with SDs was carried out to assess if CNV formation is associated with the closeness of SDs. The results of the association analysis clearly show that CNV sites overlap SDs more than expected under a random distribution (Figure 8C), demonstrating that the presence of CNVs is statistically associated with the closeness of SDs.

Discussion

In this study, we used a sequence-based approach that relies on read count data generated with exome sequencing (ES) to unveil changes in the copy number of barley genes. Considering the large number of accessions and the type of genetic material examined, to date this study delivered the most comprehensive overview of CNVs that affect gene content in cultivars, landraces and wild relatives of barley.

Beyond SNP identification, ES was extensively applied for seeking somatic and germline CNVs in human species. This practice pointed out that methodologies for CNV detection based on read count might output results that are error-prone because of the unsatisfactory FDR (Tan *et al.*, 2014). Currently, several algorithms have been proposed for detecting CNVs using read count data generated with ES to examine genomic aberrations of human individuals, although there is evidence that new statistical paradigms are needed to improve accuracy and sensitivity (Zare *et al.*, 2017). On the other hand, in plants exome capture and sequencing represent groundbreaking technologies for detecting genome-wide DNA variants while maintaining acceptable costs (Warr *et al.*, 2015). In this study, we implemented several strategies to reduce as much as possible the FDR of our CNV detection procedure and we used clustering analyses and targeted amplifications for ascertaining the performance of our procedure. Along with the molecular analyses conducted for validating a subset of duplications and deletions, the CNV-based phylogeny proved that the structural changes identified in this study correctly cluster barley accessions based on their row type (6-row and 2-row) and category (domesticated and wild relatives), corroborating the high quality and performance of our CNV detection strategy.

CNVs contribute to shape barley genome diversity

Along with other structural changes, CNVs were proposed to underlie the speciation of humans from other non-human primates (Perry *et al.*, 2006; Kim *et al.*, 2008; Girirajan, Campbell and Eichler, 2011), which would have led to substantial genome re-arrangements that allowed acquiring new functions, while in plants there is evidence that changes in copy number of genes are pervasive in certain crops and constitute the genetic bases of important agronomic traits (Sutton *et al.*, 2007; Swanson-wagner *et al.*, 2010). In this study, we surveyed genome-wide CNVs affecting gene content in a panel of barley accessions including 172 cultivars, 171 landraces and 22 wild relatives. Previous studies using gene re-sequencing and AFLP technology (Vos *et al.*, 1995) uncovered a loss of diversity in cultivars compared to landraces and wild relatives (Kilian *et al.*, 2006, 2007; Kilian, 2007; Condón *et al.*, 2009; Fricano *et al.*, 2009). Leveraging the

CNVs detected in our study, a reduction of deletions was observed in cultivars and in landraces compared to wild accessions, while the same pattern was not observed for duplications (Figure 3). Similarly, our analysis pointed out a slight reduction of CNV diversity in barley cultivars compared to landraces (Figure 3). While the reduction of deletions can be explained considering that barley domestication and breeding narrowed the genetic diversity in the domesticated accessions (Kilian *et al.*, 2006), the pattern of duplications in cultivars and landraces (Figure 3) hints that newly duplicated sequences would rapidly diverge, accumulating point mutations that mask their formation and our ability to detect these events using exome capture and sequencing.

The results reported in this study limit our conclusions to CNVs that affect gene content and consequently the actual number of deletions and duplications that segregate in our accessions could be underestimated. Moreover, the current availability of a single reference sequence of barley cv "Morex" contributes to shrink our capability to ascertain CNVs of sequences that are not present in this reference.

CNVs are pervasive across barley gene content

Considering the whole panel of 397 diverse accessions of barley, the ES-based pipeline used for detecting CNVs unveiled that 17.6% of the 170,725 captured sequences exhibit changes in copy number. As captured targets represent in most of cases gene exons, contiguous deletions or duplications were merged and 16,605 CNV sites were inferred.

These 16,605 CNV sites represent an estimate of DNA segments that can be duplicated or deleted in barley and their intersection with annotated gene models hints that this genome can bear losses or extra copies of sequences in about 10 % of predicted genes. This figure is comparable to the findings obtained applying comparative genomic hybridization (CGH) technology on a limited set of accessions using the gene space assembly of barley (Muñoz-Amatriaín, Steven R Eichten, *et al.*, 2013). CNV studies carried out in a panel of domesticated maize accessions and teosinte lines showed that more than 10% of the genes annotated in the B73 reference genome exhibit CNVs (Swanson-wagner *et al.*, 2010). Similarly, our findings show evidence that the fraction of genes that exhibit changes in copy number in barley and maize is comparable.

The loss of gene copies found in barley would be explained with the high level of gene families with multiple members annotated in this species (Mascher *et al.*, 2017). It is plausible that genes belonging to the same gene family would have redundant or partially redundant functions, which in turn compensate for possible deleterious effects of losses of gene copies. In barley, there are notorious examples of genes that show CNVs among different accessions. For instance, CNVs of *CBF* genes at *Fr-H2* locus were reported in barley cultivars using a targeted approach based on gene copy quantification (Francia *et al.*, 2016). *CBF* genes underlie frost tolerance trait and their number of copies and paralogs was tied with the level of frost tolerance in barley and other cereals (Francia *et al.*, 2016; Sieber *et al.*, 2016). In this study, CNVs of *CBFs* previously reported were detected in several barley accessions (Francia *et al.*, 2016) along with CNVs of *Vrn-H1*, another important gene that has pleiotropic effects on frost tolerance. Moreover, the detection of duplications affecting gene content hints that these extra copies of DNA would play important roles for barley adaptation to different environmental conditions, as previously reported (Sutton *et al.*, 2007; Francia *et al.*, 2016).

Comparison of the density of deletions or duplications across different chromosomes showed that chromosome 4H contains a significantly lower number of CNVs, confirming the previous report that pointed out the depletion of CNVs in this chromosome using CGH technology (Muñoz-Amatriaín, Steven R Eichten, *et al.*, 2013). Chromosome 4H would undergo a lower rate of events that lead to the formation of deletions and duplications owing to either the lack of regions that promote instability or reduced meiotic recombination as previously suggested (Mayer *et al.*, 2012; Mascher *et al.*, 2017).

Changes in copy number of genes are associated to SD-rich regions

The availability of a high-quality reference sequence allowed us to unlock the extent and occurrence of SDs in the barley genome. A large fraction of newly formed SDs partially or fully overlap predicted genes in both high confidence and low confidence annotations, reflecting the high number of families with duplicated genes that were annotated in the barley genome (Mascher *et al.*, 2017). While predicted genes explain a significant part of SDs identified, the nature of SDs that did not overlap with either annotated mobile elements or coding sequences is still unclear and would be explained postulating the existence of other genes or pseudo-genes that were not considered during the annotation process.

The findings reported in our study demonstrate that CNVs are not randomly distributed across barley coding sequences, but tend to occur in the SD-rich regions identified in the barley reference sequence (Figure 7C). SDs overlap more than expected CNV sites, hinting that they would shape regions of genomic instability, which foster the emergence of new CNVs. Molecular mechanisms that generate CNVs were extensively described in yeast, *Drosophila melanogaster* and primates (Goidts *et al.*, 2006; Kim *et al.*, 2008; Salse *et al.*, 2008; Daines *et al.*, 2009; Conrad *et al.*, 2010; Zecevic *et al.*, 2010; Zhang *et al.*, 2013), but our understanding of their incidence in plant genomes is still limited. An obvious hypothesis is that in barley recent SDs offer adequate nucleotide identity for enabling the formation of new unbalanced structural changes via NAHR. The co-occurrence of CNVs in SD-rich regions is a signature of SD-mediated CNV formation (Figure 7C) that was unveiled in this study and hints that NAHR, similar to mammalian genomes, could shape CNVs that affect barley-coding sequences, although other mechanisms were proposed.

Along with previous findings (Muñoz-Amatriaín, Steven R Eichten, *et al.*, 2013), this study showed that in the barley genome, deletions are about four times more frequent than duplications. Although we cannot exclude that the divergence of newly duplicated sequences masks our ability to detect these events, it is plausible to hypothesize that the formation of duplications and deletions occur at different rates in the barley genome, suggesting that NAHR mediated by SD pairs located in the same chromatids could be more frequent than NAHR mediated by SD pairs located in different chromatids (Chen *et al.*, 2014). Studying the flanking regions of deletions and duplications, sequence signatures of CNV formation based on double-strand break (DSB) repair via single-strand annealing (SSA) were reported on 41.1% of CNVs of barley (Muñoz-Amatriaín, Steven R Eichten, *et al.*, 2013). A possible reason for explaining these seemingly different findings lies in CGH , which was used for detecting CNVs in a small panel of 16 wild and domesticated barley accessions in a previous CNV study conducted in barley (Muñoz-Amatriaín, Steven R Eichten, *et al.*, 2013). As CGH does not allow to examine sequences with high sequence similarity, more probably CNVs in SD-rich regions were not considered in the previous study (Muñoz-Amatriaín, Steven R Eichten, *et al.*, 2013). This study shows evidence of SD-mediated formation of CNVs in barley, a mechanism that in plants has been claimed several times (Muñoz-Amatriaín, Steven R Eichten, *et al.*, 2013; Bai *et al.*, 2016). Further studies on barley CNVs in non-coding sequences are needed to explore the potential role of both NAHR-based and DSB-based mechanisms in the formation of unbalanced structural changes in barley.

Overall, the landscape of CNVs unveiled in this study provides evidence for widespread changes in copy number of genes, which in turn reflect the dynamic nature of the barley genome. Moreover, our findings pave the way to better understand the gene content of core and dispensable genomes of this species for evolutionary studies (Morgante, De Paoli and Radovic, 2007). As already demonstrated for frost and boron-tolerance traits, it is likely that along with SNPs, CNVs significantly contribute to barley phenotypic diversity, although further investigations are necessary to document to which extent these structural variants affect other important traits. The use of CNVs in genome-wide association studies would allow to better understand how these structural variants underlie barley phenotypic variation and enable their exploitation for breeding.

We have demonstrated that changes in copy number of genes are widespread across the barley genome and that these structural variants contribute to shaping the genetic diversity of cultivars, landraces and wild relatives and affect genes with specific functions. Moreover, we reported that SD-rich sequences are regions of the barley genomes in which CNV formation rate is higher than expected and speculated that molecular mechanisms based on similarity of SDs (e.g. NHAR) may be involved in changing copy number of genes. The list of CNVs identified in this study is a new asset for understanding the genome biology and evolution of barley as well as the genetic bases of complex traits.

Methods

Plant materials

The genetic material examined in this study has been extensively described in other reports (Bustos-Korts *et al.*, 2019) and relevant information regarding the classification and the origin, type and of selected accessions is reported in Supporting Information 1, Table S1. In brief, a panel of 397 out 403 barley accessions previously described (Bustos-Korts *et al.*, 2019) was selected for this study, including 172 formally-bred cultivars released in Europe, Asia and Americas, 171 landraces collected in Europe, Asia, Middle East and Africa and 22 wild relatives of barley (*H. spontaneum* subsp. *spontaneum* and *H. spontaneum* subsp. *agriocrithon*) collected in Middle East areas. Other 32 domesticated accessions for which the categorization as cultivar or landrace was questionable were included and examined in this study (Bustos-Korts *et al.*, 2019).

Preparation of Exome Capture library and sequencing

Genomic DNA (gDNA) was extracted from barley leaf material from a single plant for each genotype. DNA samples were checked with a Genomic DNA ScreenTape on Agilent 2200 Tape Station System (Santa Clara, CA, USA) in order to verify gDNA integrity. Samples were quantified by Picogreen assay (Thermo Fisher, CA, USA) and normalised to 20 ng/ul in 10 nM Tris-Hcl (pH 8.0) as suggested in the NimbleGen SeqCap EZ Library SR protocol. The gDNA was fragmented to a size range of 180-200 bp using Covaris microTUBES and a Covaris S220 Instrument (Covaris, MA, USA) and whole genome libraries were prepared according to the Kapa Library Preparation protocol. Libraries were quantified using a Nanodrop (Thermo Fisher, CA, USA) and analysed electrophoretically with an Agilent 2200 Tape Station System using a D1000 ScreenTape. Libraries were pooled in 8-plex and used for the hybridization with the barley SeqCap Ez oligo pool (Design Name: 120426_Barley_BEC_D04) (Mascher *et al.*, 2013) in a thermocycler at 47°C for 48 h. Capture beads were used to pull down the complex of capture oligos and genomic DNA fragments and unbound fragments were removed by washing. Enriched fragments were amplified by PCR and the final library was quantified by qPCR and visualised by Agilent Tape Station. Sequencing libraries were normalised to 2nM, NaOH denatured and used for cluster amplification on the cBot.

The clustered flow cells were sequenced on Illumina HiSeq2000 with an 8-plex strategy (i.e. 8 samples per HiSeq lane) with a 100 bp paired-end run module.

Analysis of whole exome sequencing data

Target regions utilized for designing exome capture probes (http://sequencing.roche.com/content/dam/rochesequence/worldwide/shared-designs/barley_exome.zip) were mapped against the reference sequence of barley cv 'Morex' (Mascher *et al.*, 2017) with bwa-mem 0.7.15 (Li and Durbin, 2009). Mapping positions of captured sequences were extracted from the BAM file of alignments and converted in BED format using bam2bed (Neph *et al.*, 2012). Subsequently overlapping BED records were collapsed using the merge command of bedops 2.4.20 (Neph *et al.*, 2012) to uncover the actual portions of the barley genome that are examined using barley whole exome capture.

Sequence quality control was assessed with FastQC (Andrews, 2010). Raw Illumina reads were then quality trimmed to a base quality of 20 from both ends with Trimmomatic version 0.30 (Bolger, Lohse and Usadel, 2014). Only correctly paired reads longer than 70 bp were used for further processing. Trimmed reads were then mapped to the reference genome with BWA version 0.7.15 using the mem algorithm with default parameters (Li and Durbin, 2009). The resulting BAM files were sorted with Samtools (Li and Durbin, 2009) (http://samtools.sourceforge.net/) and duplicate reads were marked and removed with picard (Broad Institute, 2016) using 'MarkDuplicates' command. Coverage at each captured sequence was computed with samtools depth (Li, 2011) considering only properly mapped paired reads. Captured sequences exhibiting a coverage lower than 5X were removed from all subsequent analyses. The average sequencing coverage across the whole set of captured sequences was computed in R statistical environment using Rsubread package version 1.28 (Liao, Smyth and Shi, 2013; Team, 2015) including the count of PE fragments that overlap contiguous captured sequences. PE fragment counts obtained for each sample, were subsequently merged in R environment for creating a numeric matrix, which was subsequently utilized for detecting copy number variants.

Detection of copy number variants and validation

Read count data were processed in R statistical environment (Team, 2015) with the R package “ExomeDepth” for detecting CNVs (Plagnol *et al.*, 2012) setting the expected exon length at 1,000 bp and the minimum quality mapping score at 30. CNVs detected in less than three barley accessions were discarded and not considered for validation. Contiguous deletions or duplications of captured sequences detected in the same accession were merged and the resulting CNVs were utilized for constructing a phylogeny based on NJ method and Euclidean distance utilizing the R packages “ape” and “phytools” in R statistical environment (Saitou and Nei, 1987; Paradis, Claude and Strimmer, 2004; Revell, 2016).

Identification of segmental duplications in the barley reference sequence

For surveying the occurrence of SDs, all known repetitive elements of the barley reference sequence were masked utilizing the most recent and accurate annotation of transposable elements (Wicker *et al.*, 2017) and subsequently the masked chromosome sequences were split in chunks of 250 kb. These chunks were aligned against the masked reference sequence of barley for identifying homologous sequences using standalone BLAST 2.5.0 (Altschul *et al.*, 1990; Camacho *et al.*, 2009). Alignment records obtained from BLAST analyses were subsequently parsed for identifying homologous sequence pairs sharing a nucleotide identity higher than 95% and larger than 1 KB using python 2.7.9 along with the package Biopython (Cock *et al.*, 2009). Alignment records were transformed in a BED file using custom python scripts and overlapping regions were subsequently collapsed using bedops “merge” command (Neph *et al.*, 2012).

GO ontology and enrichment analysis

For exploring the ontology content of duplicated and deleted genes, the whole set of 283,096 sequences used for designing exome capture probes were annotated with GO terms using Blast2Go (Conesa and Gotz, 2008). Subsequently, GO terms of these sequences were assigned to the genomic coordinates in which captured sequences were unambiguously mapped. The high-level summary of functions implemented in the GO Slim terms (McCarthy *et al.*, 2006) was used for summarizing the ontology content of duplicated and deleted genes.

Enrichment analysis was conducted in R statistical environment using the R package "TopGO" (Alexa, Rahnenführer and Lengauer, 2006; Team, 2015) for identifying GO terms that were over-represented and under-represented in the set of duplicated and deleted genes and functional categories associated to set of duplicated and deleted genes that were lost in the domesticated accessions. For carrying out GO enrichment for the first analysis, the whole set of mapped sequences was utilized as baseline, while the over- and under-represented GO terms were investigated in deleted and duplicated genes, using the "elim" algorithm implemented in TopGO for selecting the most stringent subset of over-represented and under-represented GO terms. For identifying GO terms associated to duplicated and deleted genes that were lost during the domestication process, the whole set of mapped sequences was used as baseline, while the over- and under-represented GO terms were investigated in deleted and duplicated genes that were detected exclusively in wild accessions, using the "elim" algorithm implemented in TopGO.

The false discovery rate threshold was calculated utilizing Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). Bar plots were generated using the package "ggplot2" in R statistical environment (Team, 2015; Wickham, 2016).

Association analysis of SDs with CNV sites and predicted genes

Histograms of SD distribution across barley chromosomes were computed in bins of 50 kb in R statistical environment (Team, 2015) parsing the BED file describing the genome coordinates of SDs having a nucleotide identity higher than 95%.

Association analyses between SDs and CNVs detected in the panel of barley accessions were computed using Spearman rank correlation coefficient, binning barley chromosomes in increasing intervals from 40 kb to 2 Mb. Within each interval Spearman rank correlation coefficient was calculated in R statistical environment (Team, 2015), between the number of SDs unveiled in the reference sequence and the number of CNVs detected in the panel of 397 barley accessions. For assessing the non-random association of SDs with CNV sites or predicted high confidence genes, 1,000 permutation tests were carried out between pairs of features (SD and CNV sites; SD and predicted high confidence genes) randomizing features over the non-masked space of each chromosome for computing the expected number of overlaps under the hypothesis of random distributions of these genomic features. Similarly, the expected average distance of SDs with the closest high confidence gene was computed permuting these genomic features over the non-masked space of each chromosome for 1,000 times. The R package regioneR (Gel *et al.*, 2016) was utilized for these computations and results were plotted utilizing the R package "ggplot2" (Wickham, 2016).

Data availability satement

The raw sequencing data analyzed in this manuscript were deposited in the European Nucleotide Archive under the study number PRJEB33527

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/ 2007-2013) under the grant agreement n°FP7- 613556, WHEALBI project.

Authors' contributions

AF, LR and LC conceived the study. AF led the study and carried out data analysis, AF wrote the paper with significant contributions by RW, BK, LR and LC; BK, RW, JR, LR and LC assembled the panel of barley accessions; LR coordinated exome sequencing of the barley collection; CF carried out library preparation, capture and sequencing; GB carried out validation experiments and PB conducted GO annotations. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

References

Alexa, A., Rahnenführer, J. and Lengauer, T. (2006) 'Improved scoring of functional groups from gene expression data by decorrelating GO graph structure', *Bioinformatics*, 22(13), pp. 1600–1607. doi: 10.1093/bioinformatics/btl140.

Alkan, C., Coe, B. P. and Eichler, E. E. (2011) 'Genome structural variation discovery and genotyping', *Nature Reviews Genetics*, 12(5), pp. 363–376. doi: 10.1038/nrg2958.

Altschul, S. F. *et al.* (1990) 'Basic local alignment search tool', *Journal of Molecular Biology*, 215(3), pp. 403–410. doi: 10.1016/S0022-2836(05)80360-2.

Andrews, S. (2010) *FastQC: A quality control tool for high throughput sequence data.*, *Http://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/*. doi: citeulike-article-id:11583827.

Bai, Z. *et al.* (2016) 'The impact and origin of copy number variations in the Oryza species', *BMC Genomics*. BMC Genomics, 17(1), pp. 1–12. doi: 10.1186/s12864-016-2589-2.

Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society B*, pp. 289–300. doi: 10.2307/2346101.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: A flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp. 2114–2120. doi: 10.1093/bioinformatics/btu170.

Broad Institute (2016) *Picard tools*, *https://broadinstitute.github.io/picard/*. Available at: https://broadinstitute.github.io/picard/%5Cnhttp://broadinstitute.github.io/picard/.

Bustos-Korts, D. *et al.* (2019) 'Exome sequences and multi-environment field trials elucidate the genetic basis of adaptation in barley', *The Plant Journal*. doi: 10.1111/tpj.14414.

Camacho, C. *et al.* (2009) 'BLAST plus: architecture and applications', *BMC Bioinformatics*, 10(421), p. 1. doi: Artn 421\nDoi 10.1186/1471-2105-10-421.

Chen, L. *et al.* (2014) 'Genome architecture and its roles in human copy number variation.', *Genomics & informatics*, 12(4), pp. 136–44. doi: 10.5808/GI.2014.12.4.136.

Clop, A., Vidal, O. and Amills, M. (2012) 'Copy number variation in the genomes of domestic animals', *Animal Genetics*, 43(5), pp. 503–517. doi: 10.1111/j.1365-2052.2012.02317.x.

Cock, P. J. A. *et al.* (2009) 'Biopython: Freely available Python tools for computational molecular biology and bioinformatics', *Bioinformatics*, 25(11), pp. 1422–1423. doi: 10.1093/bioinformatics/btp163.

Condón, F. *et al.* (2009) 'Effect of advanced cycle breeding on genetic gain and phenotypic diversity in barley breeding germplasm', *Crop Science*, 49(5), pp. 1751–1761. doi: 10.2135/cropsci2008.10.0585.

Conesa, A. and Gotz, S. (2008) 'Blast2GO: A comprehensive suite for functional analysis in plant genomics', *International Journal of Plant Genomics*, 2008. doi: 10.1155/2008/619832.

Conrad, D. F. *et al.* (2010) 'Origins and functional impact of copy number variation in the human genome', *Nature*. Nature Publishing Group, 464(7289), pp. 704–712. doi: 10.1038/nature08516.

Daines, B. *et al.* (2009) 'High-throughput multiplex sequencing to discover copy number variants in Drosophila', *Genetics*. doi: 10.1534/genetics.109.103218.

Debolt, S. (2010) 'Copy number variation shapes genome diversity in arabidopsis over immediate family generational scales', *Genome Biology and Evolution*, 2(1), pp. 441–453. doi: 10.1093/gbe/evq033.

Díaz, A. *et al.* (2012) 'Copy number variation affecting the photoperiod-B1 and vernalization-A1 genes is associated with altered flowering time in wheat (Triticum aestivum)', *PLoS ONE*, 7(3). doi: 10.1371/journal.pone.0033234.

Dittwald, P. *et al.* (2013) 'Inverted Low-Copy Repeats and Genome Instability-A Genome-Wide Analysis', *Human Mutation*, 34(1), pp. 210–220. doi: 10.1002/humu.22217.

Eichler, E. E. (2001) 'Recent duplication, domain accretion and the dynamic mutation of the human genome', *Trends in Genetics*, pp. 661–669. doi: 10.1016/S0168-9525(01)02492-1.

Evans, J. *et al.* (2015) 'Diversity and population structure of northern switchgrass as revealed through exome capture sequencing', *Plant Journal*, 84(4), pp. 800–815. doi: 10.1111/tpj.13041.

Feuk, L., Carson, A. R. and Scherer, S. W. (2006) 'Structural variation in the human genome', *Nature Reviews Genetics*, pp. 85–97. doi: 10.1038/nrg1767.

Francia, E. *et al.* (2016) 'Copy number variation at the HvCBF4–HvCBF2 genomic segment is a major component of frost resistance in barley', *Plant Molecular Biology*, 92(1–2), pp. 161–175. doi: 10.1007/s11103-016-0505-4.

Freeman, J., Perry, G. and Feuk, L. (2006) 'Copy number variation: new insights in genome diversity', *Genome research*, (617), pp. 949–961. doi: 10.1101/gr.3677206.16.

Fricano, A. *et al.* (2009) 'Genetic variants of hvcbf14 are statistically associated with frost tolerance in a european germplasm collection of hordeum vulgare', *Theoretical and Applied Genetics*, 119(7), pp. 1335–1348. doi: 10.1007/s00122-009-1138-7.

Gel, B. *et al.* (2016) 'regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests.', *Bioinformatics (Oxford, England)*, 32(2), pp. 289–91. doi: 10.1093/bioinformatics/btv562.

Giannuzzi, G. *et al.* (2011) 'Analysis of high-identity segmental duplications in the grapevine genome', *BMC Genomics*, 12(1), p. 436. doi: 10.1186/1471-2164-12-436.

Girirajan, S., Campbell, C. D. and Eichler, E. E. (2011) 'Human Copy Number Variation and Complex Genetic Disease', *Annual Review of Genetics*, 45(1), pp. 203–226. doi: 10.1146/annurev-genet-102209-163544.

Goidts, V. *et al.* (2006) 'Identification of large-scale human-specific copy number differences by inter-species array comparative genomic hybridization', *Human Genetics*, 119(1–2), pp. 185–198. doi: 10.1007/s00439-005-0130-9.

Innan, H. and Kondrashov, F. (2010) 'The evolution of gene duplications: classifying and distinguishing between models.', *Nature reviews. Genetics*. Nature Publishing Group, 11(2), pp. 97–108. doi: 10.1038/nrg2689.

Kilian, B. *et al.* (2006) 'Haplotype structure at seven barley genes: Relevance to gene pool bottlenecks, phylogeny of ear type and site of barley domestication', *Molecular Genetics and Genomics*, 276(3), pp. 230–241. doi: 10.1007/s00438-006-0136-6.

Kilian, B. (2007) 'Genetic diversity , evolution and domestication of Triticeae in the Fertile Crescent'.

Kilian, B. *et al.* (2007) ‘Molecular diversity at 18 loci in 321 wild and 92 domesticate lines reveal no reduction of nucleotide diversity during Triticum monococcum (einkorn) domestication: Implications for the origin of agriculture’, *Molecular Biology and Evolution*, 24(12), pp. 2657–2668. doi: 10.1093/molbev/msm192.

Kim, P. M. *et al.* (2008) ‘Analysis of copy number variants and segmental duplication in the human genome: Evidence for a change in the process of formation in recent evolutionary history.’, *Genome Research*, 18, pp. 1865–1874. doi: 10.1101/gr.081422.108.

Li, H. (2011) ‘A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data’, *Bioinformatics*, 27(21), pp. 2987–2993. doi: 10.1093/bioinformatics/btr509.

Li, H. and Durbin, R. (2009) ‘Fast and accurate short read alignment with Burrows-Wheeler transform’, *Bioinformatics*, 25(14), pp. 1754–1760. doi: 10.1093/bioinformatics/btp324.

Liao, Y., Smyth, G. K. and Shi, W. (2013) ‘The Subread aligner: Fast, accurate and scalable read mapping by seed-and-vote’, *Nucleic Acids Research*, 41(10). doi: 10.1093/nar/gkt214.

Loscos, J. *et al.* (2014) ‘HvFT1 polymorphism and effectâ€”survey of barley germplasm and expression analysis’, *Frontiers in Plant Science*, 5. doi: 10.3389/fpls.2014.00251.

Mascher, M. *et al.* (2013) ‘Barley whole exome capture: A tool for genomic research in the genus Hordeum and beyond’, *Plant Journal*, 76(3), pp. 494–505. doi: 10.1111/tpj.12294.

Mascher, M. *et al.* (2017) ‘A chromosome conformation capture ordered sequence of the barley genome’, *Nature*. Nature Publishing Group, 544(7651), pp. 427–433. doi: 10.1038/nature22043.

Mayer, K. F. X. *et al.* (2012) ‘A physical, genetic and functional sequence assembly of the barley genome’, *Nature*. Nature Publishing Group, 491(7426), pp. 711–716. doi: 10.1038/nature11543.

McCarthy, F. M. *et al.* (2006) ‘AgBase: A functional genomics resource for agriculture’, *BMC Genomics*, 7, pp. 1–13. doi: 10.1186/1471-2164-7-229.

Morgante, M., De Paoli, E. and Radovic, S. (2007) ‘Transposable elements and the plant pan-genomes’, *Current Opinion in Plant Biology*, pp. 149–155. doi: 10.1016/j.pbi.2007.02.001.

Muñoz-Amatriaín, M., Eichten, Steven R, *et al.* (2013) ‘Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome’, *Genome Biology*, 14(6), p. R58. doi: 10.1186/gb-2013-14-6-r58.

Muñoz-Amatriaín, M., Eichten, Steven R., *et al.* (2013) ‘Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome’, *Genome Biology*. BioMed Central Ltd, 14(6), p. R58. doi: 10.1186/gb-2013-14-6-r58.

Neph, S. *et al.* (2012) ‘BEDOPS: High-performance genomic feature operations’, *Bioinformatics*, 28(14), pp. 1919–1920. doi: 10.1093/bioinformatics/bts277.

Pagel, J. *et al.* (2004) ‘Segmental duplications within the Glycine max genome revealed by fluorescence in situ hybridization of bacterial artificial chromosomes.’, *Genome / National Research Council Canada = Génome / Conseil national de recherches Canada*, 47(4), pp. 764–8. doi: 10.1139/g04-025.

Paradis, E., Claude, J. and Strimmer, K. (2004) ‘APE: Analyses of phylogenetics and evolution in R language’, *Bioinformatics*, 20(2), pp. 289–290. doi: 10.1093/bioinformatics/btg412.

Perry, G. H. *et al.* (2006) ‘Hotspots for copy number variation in chimpanzees and humans’, *Proceedings of the National Academy of Sciences*, 103(21), pp. 8006–8011. doi: 10.1073/pnas.0602318103.

Pinosio, S. *et al.* (2016) ‘Characterization of the Poplar Pan-Genome by Genome-Wide Identification of Structural Variation’, *Molecular Biology and Evolution*, 33(10), pp. 2706–2719. doi: 10.1093/molbev/msw161.

PJ Hastings, James R Lupski, S. M. R. and G. I. (2010) ‘Mechanisms of change in gene copy number’, *Nat Rev Genet*, 10(8), pp. 551–564. doi: 10.1038/nrg2593.Mechanisms.

Plagnol, V. *et al.* (2012) ‘A robust model for read count data in exome sequencing experiments and implications for copy number variant calling’, *Bioinformatics*, 28(21), pp. 2747–2754. doi: 10.1093/bioinformatics/bts526.

Prunier, J. *et al.* (2017) ‘Gene copy number variations in adaptive evolution: The genomic distribution of gene copy number variations revealed by genetic mapping and their adaptive role in an undomesticated species, white spruce (Picea glauca)’, *Molecular Ecology*, 26(21), pp. 5989–6001. doi: 10.1111/mec.14337.

Revell, L. J. (2016) ‘phytools: Phylogenetic Tools for Comparative Biology (and Other Things)’, *Methods in Ecology and Evolution*. doi: 10.1111/j.2041- 210X.2011.00169.x.

Saitou, N. and Nei, M. (1987) ‘The neighbor-joining method: a new method for reconstructing phylogenetic trees’, *Molecular Biology and Evolution*, 4(4), pp. 406–425. doi: citeulike-article-id:93683.

Salse, J. *et al.* (2008) ‘Identification and Characterization of Shared Duplications between Rice and Wheat Provide New Insight into Grass Genome Evolution’, *the Plant Cell Online*, 20(1), pp. 11–24. doi: 10.1105/tpc.107.056309.

Sharp, A. J. *et al.* (2005) ‘Segmental duplications and copy-number variation in the human genome.’, *American journal of human genetics*, 77(1), pp. 78–88. doi: 10.1086/431652.

Sieber, A. N. *et al.* (2016) ‘Copy number variation of CBF-A14 at the Fr-A2 locus determines frost tolerance in winter durum wheat’, *Theoretical and Applied Genetics*, 129(6), pp. 1087–1097. doi: 10.1007/s00122-016-2685-3.

Springer, N. M. *et al.* (2009) ‘Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content’, *PLoS Genetics*, 5(11). doi: 10.1371/journal.pgen.1000734.

Sutton, T. *et al.* (2007) ‘Boron-toxicity tolerance in barley arising from efflux transporter amplification’, *Science*, 318(5855), pp. 1446–1449. doi: 10.1126/science.1146853.

Swanson-wagner, R. A. *et al.* (2010) ‘Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor', (612), pp. 1689–1699. doi: 10.1101/gr.109165.110.

Tan, R. *et al.* (2014) 'An Evaluation of Copy Number Variation Detection Tools from Whole-Exome Sequencing Data', *Human Mutation*, 35(7), pp. 899–907. doi: 10.1002/humu.22537.

Team, R. D. C. (2015) *R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2014. R Foundation for Statistical Computing*. Vienna.

Vos, P. *et al.* (1995) 'AFLP: A new technique for DNA fingerprinting', *Nucleic Acids Research*, 23(21), pp. 4407–4414. doi: 10.1093/nar/23.21.4407.

Warr, A. *et al.* (2015) 'Exome Sequencing: Current and Future Perspectives', *Genes|Genomes|Genetics*, 5(8), pp. 1543–1550. doi: 10.1534/g3.115.018564.

Wicker, T. *et al.* (2017) 'The repetitive landscape of the 5100 Mbp barley genome', *Mobile DNA*. Mobile DNA, 8(1), pp. 1–16. doi: 10.1186/s13100-017-0102-3.

Wickham, H. (2016) *ggplot 2: Elagant graphics for data analysis*, *Media*. doi: 10.1007/978-0-387-98141-3.

Zare, F. *et al.* (2017) 'An evaluation of copy number variation detection tools for cancer using whole exome sequencing data', *BMC Bioinformatics*. BMC Bioinformatics, 18(1), pp. 1–13. doi: 10.1186/s12859-017-1705-x.

Zecevic, V. *et al.* (2010) 'Genetic and phenotypic variability of yield components in wheat (triticum aestivum l.)', *Bulgarian Journal of Agricultural Science*, 16(4), pp. 422–428.

Zhang, H. *et al.* (2013) 'Gene copy-number variation in haploid and diploid strains of the yeast Saccharomyces cerevisiae', *Genetics*. doi: 10.1534/genetics.112.146522.

Zhang, R. *et al.* (2017) 'Segmental Duplication of Chromosome 11 and its Implications for Cell Division and Genome-wide Expression in Rice', *Scientific Reports*. Springer US, 7(1), p. 2689. doi: 10.1038/s41598-017-02796-9.

Zhang, Z. *et al.* (2015) 'Genome-Wide Mapping of Structural Variations Reveals a Copy Number Variant That Determines Reproductive Morphology in Cucumber', *The Plant Cell*, 27(6), pp. 1595–1604. doi: 10.1105/tpc.114.135848.

Figure 1: CNV-based phylogeny of the 397 barley accessions. (**A**) In this phylogeny, two-row and six-row barley accessions are depicted in red and green, respectively. Accessions exhibiting mutant phenotypes for spikelet formation (*Hordeum vulgare* L. convar. *deficiens, Hordeum vulgare* L. convar. *intermedium* and *Hordeum vulgare* L. convar. *labile*) were depicted in yellow, white and brown, respectively. (**B**) In this phylogeny, domesticated barley accessions (*Hordeum vulgare* subsp. *vulgare*) and wild relatives (*Hordeum vulgare* subsp. *spontaneum* and feral *Hordeum vulgare* subsp. *agriocrithon*) are depicted in brown, orange and red, respectively. (**C**) Projection of the CNV-based phylogeny onto a world map according to the geographic origin of barley accessions.

Figure 2. Distribution and frequency of CNVs detected across the seven barley chromosomes. Plots show the genome coordinates of CNVs along the seven barley chromosomes (x-axis), while the frequency (in %) of each CNV in the panel of 397 accessions is reported in the y-axis. Red and blue points of the plots indicate deletions and duplications, respectively.

Figure 3. Average of per chromosome CNV density computed in different categories of barley accessions. Bars report the average density of deletions (left bar plot) and duplications (right bar plot) detected in wild relatives (violet bars), landraces (light blue bars), cultivars (green bars) and in the whole panel of accessions (red bars).

Figure 4. Overview of the ontology content of duplicated and deleted genes. Bars show the description of GO Slim Term (y-axis) of duplicated and deleted genes, while the count of each GO Slim term is reported in the x-axis. **(A)** In this bar plot, the count of high-level GO terms of “Molecular Function” domain are reported, while in **(B)** and **(C)** the count of high-level GO terms of “Cellular Component” and “Biological Process” domains are reported, respectively.

Figure 5. GO enrichment in duplicated and deleted genes. The 193 GO terms (y-axis) (FDR threshold ≤ 0.01) overrepresented in duplicated and deleted genes are plotted along the corresponding negative logarithm of their Fisher's *P* value (x axis). **(A)** Overrepresented GO terms of the "Molecular Function", **(B)** "Cellular Component", and **(C)** "Biological Process" domains are reported, respectively.

Figure 6. GO enrichment of duplicated and deleted genes differentially detected in wild and domesticated accessions. The 39 GO terms (y-axis) (FDR threshold ≤ 0.01) overrepresented in duplicated and deleted genes of wild accessions are plotted along the corresponding negative logarithm of their Fisher's *P* value (x axis). **(A)** Overrepresented GO terms of the "Molecular Function", **(B)** "Cellular Component", and **(C)** "Biological Process" domains are reported, respectively.

Figure 7. Frequency and length spectra of SDs and correlation with CNVs. (A) Length spectrum of SDs detected in barley cv "Morex"; **(B)** Histograms of SD distribution across the seven barley chromosomes; **(C)** For each of the seven plots, in the y-axes the values of Spearman rank correlation coefficient between SDs and CNVs were plotted, while in the x-axes the values of bin size utilized for computing the Spearman rank correlation coefficient were reported. Only statistically significant values of Spearman rank correlation coefficient with *P* values lower that 0.001 were plotted.

Figure 8. Association analysis of SDs based on permutation tests. In all plots, the measured value (green line) and the expected value (black line) obtained after the randomization of sequence intervals are reported. **(A)** In this plot, the average distance of SDs (x-axis) with their closest genes was compared with the lower bound of the expected average distance (red vertical line); **(B)** In this plot the number of overlaps (x-axis) between SDs and annotated genes was compared with the upper bound (red line) of the expected number of overlaps in case of random distribution. **(C)** In this plot the number of overlaps (x-axis) between SDs and CNV sites was compared with the upper bound (red line) of the expected number of overlaps.

Number of CNVs and deletion/duplication ratios across barley chromosomes

Chromosome	Total number of CNV	Number of deletions	Number of duplications	Deletion/duplication ratio
Chromosome 1H	2,558	1,983	575	3.45
Chromosome 2H	2,941	2,355	586	4.02
Chromosome 3H	2,496	2,001	495	4.04
Chromosome 4H	968	782	186	4.20
Chromosome 5H	2,498	1,973	525	3.76
Chromosome 6H	2,104	1,663	441	3.77
Chromosome 7H	3,040	2,393	647	3.70
All chromosomes	16,605	13,150	3,455	3.81

Table 1: Distribution of CNVs across the seven barley chromosomes

Density of deletions and duplications in barley coding sequences

Chromosome	Density of deletions [a]	Density of duplications [b]
Chromosome 1H	256.04	74.24

Chromosome 2H	238.03	59.23
Chromosome 3H	204.94	50.70
Chromosome 4H	102.62	24.41
Chromosome 5H	200.75	53.42
Chromosome 6H	229.19	60.78
Chromosome 7H	260.11	70.33
All chromosomes	213.10	56.16

Table 2: Distribution of CNVs affecting coding sequences across the seven barley chromosomes

[a] Number of deletions per Mb of per-chromosome captured targets.

[b] Number of duplications per Mb of per-chromosome captured targets.

Pairwise Poisson's test *P* values for comparing CNV densities of barley chromosomes

	Chromosome 1H	Chromosome 2H	Chromosome 3H	Chromosome 4H	Chromosome 5H	Chromosome 6H
Chromosome 2H	0.36	-				
Chromosome 3H	**5.05*x10^{-11*}**	**1.76x10^{-05*}**	-			
Chromosome 4H	**1.33x10^{-113*}**	**1.97x10^{-102*}**	**3.36x10^{-64*}**	-		
Chromosome 5H	**4.89x10^{-13*}**	**5.00x10^{-07*}**	1	**5.35x10^{-60*}**	-	
Chromosome 6H	**0.02***	1	**1.64x10^{-2*}**	**7.77x10^{-81*}**	0.15*	-
Chromosome 7H	1	**0.04***	**6.51x10^{-14*}**	**5.09x10^{-127*}**	**2.90x10^{-16*}**	0.15

Table 3 *P* values of pairwise Poisson's tests for comparing the rates of CNV densities in barley chromosomes.

Survey of old and recent segmental duplications in barley cv ‘Morex’

Number of SDs	**Identity (%)**	**Length (bp)**
20,853	>95	>1,000
18,873	>96	>1,000
16,107	>97	>1,000
12,631	>98	>1,000
9,114	>99	>1,000

Table 4 Number of segmental duplications (SDs) identified in the reference sequence of barley cv “Morex” using different identity thresholds.

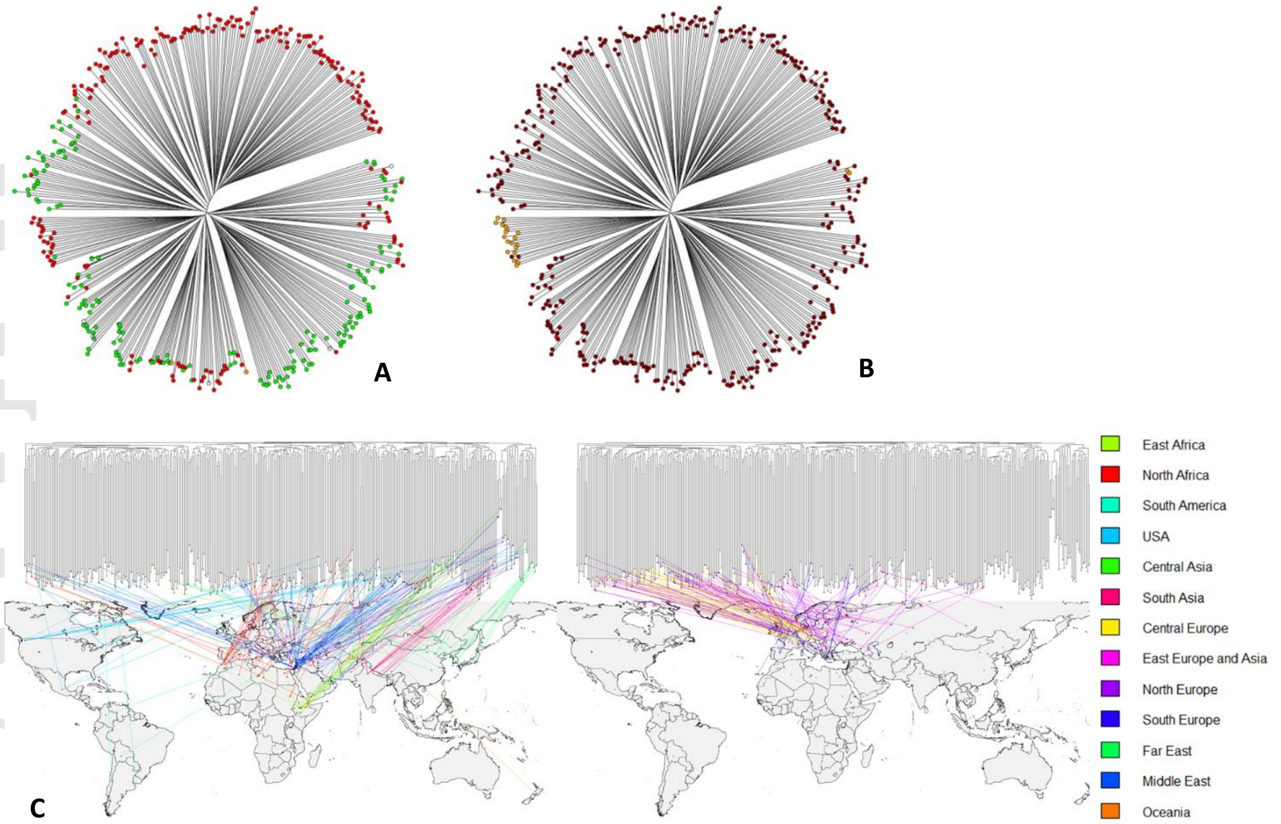

Figure 1: CNV-based phylogeny of the 397 barley accessions. (**A**) In this phylogeny, two-row and six-row barley accessions are depicted in red and green, respectively. Accessions exhibiting mutant phenotypes for spikelet formation (*Hordeum vulgare* L. convar. *deficiens, Hordeum vulgare* L. convar. *intermedium* and *Hordeum vulgare* L. convar. *labile*) were depicted in yellow, white and brown, respectively. (**B)** In this phylogeny, domesticated barley accessions (*Hordeum vulgare* subsp. *vulgare*) and wild relatives (*Hordeum vulgare* subsp. *spontaneum* and feral *Hordeum vulgare* subsp. *agriocrithon*) are depicted in brown, orange and red, respectively. (C) Projection of the CNV-based phylogeny onto a world map according to the geographic origin of barley accessions.

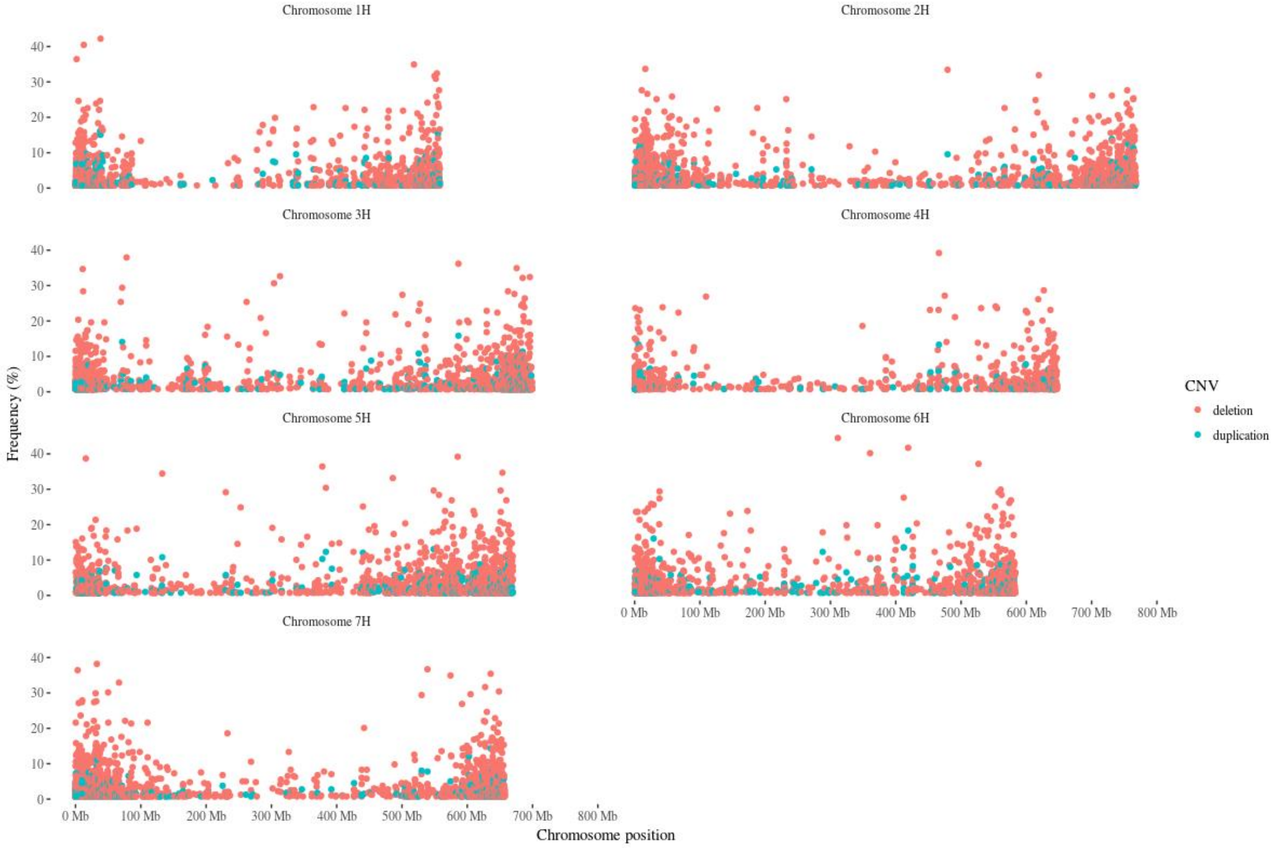

Figure 2. Distribution and frequency of CNVs detected across the seven barley chromosomes. Plots show the genome coordinates of CNVs along the seven barley chromosomes (x-axis), while the frequency (in %) of each CNV in the panel of 397 accessions is reported in the y-axis. Red and blue points of the plots indicate deletions and duplications, respectively.

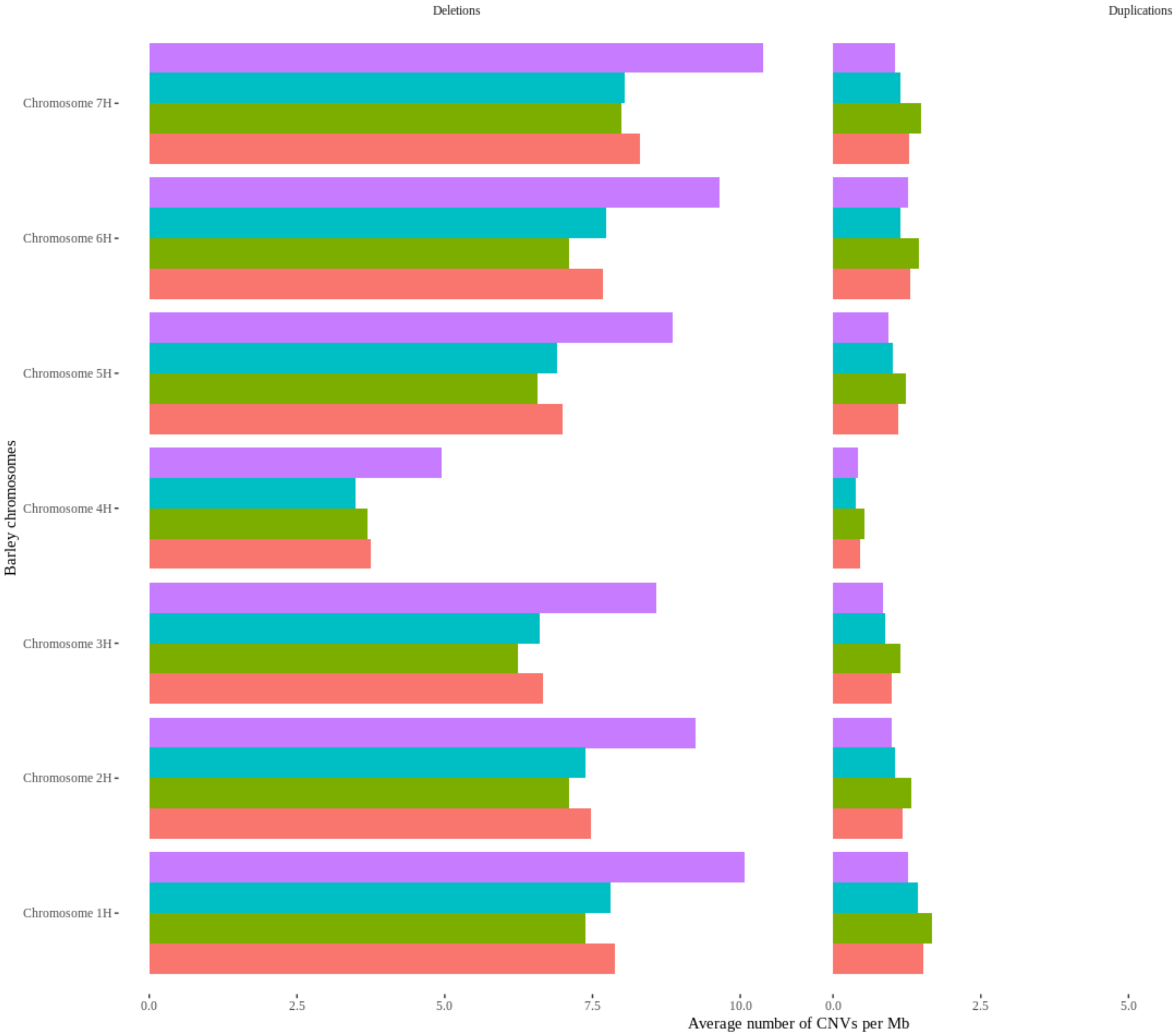

Figure 3. Average of per chromosome CNV density computed in different categories of barley accessions. Bars report the average density of deletions (left bar plot) and duplications (right bar plot) detected in wild relatives (violet bars), landraces (light blue bars), cultivars (green bars) and in the whole panel of accessions (red bars).

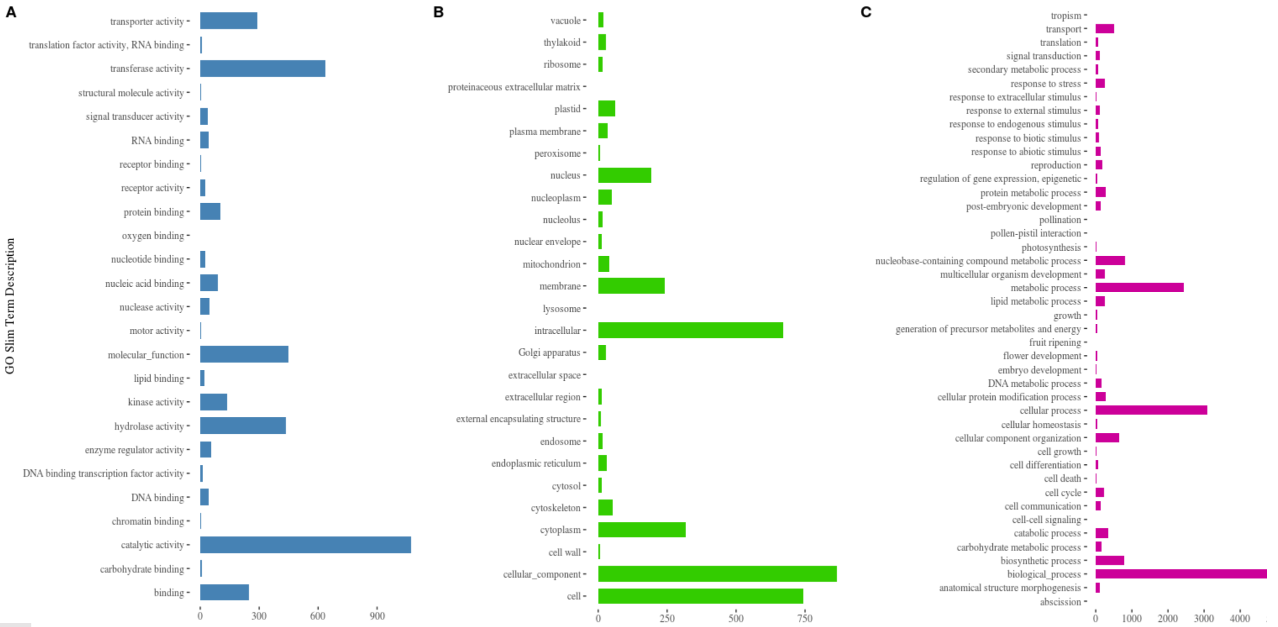

Figure 4. Overview of the ontology content of duplicated and deleted genes. Bars show the description of GO Slim Term (y-axis) of duplicated and deleted genes, while the count of each GO Slim term is reported in the x-axis. **(A)** In this bar plot, the count of high-level GO terms of "Molecular Function" domain are reported, while in **(B)** and **(C)** the count of high-level GO terms of "Cellular Component" and "Biological Process" domains are reported, respectively.

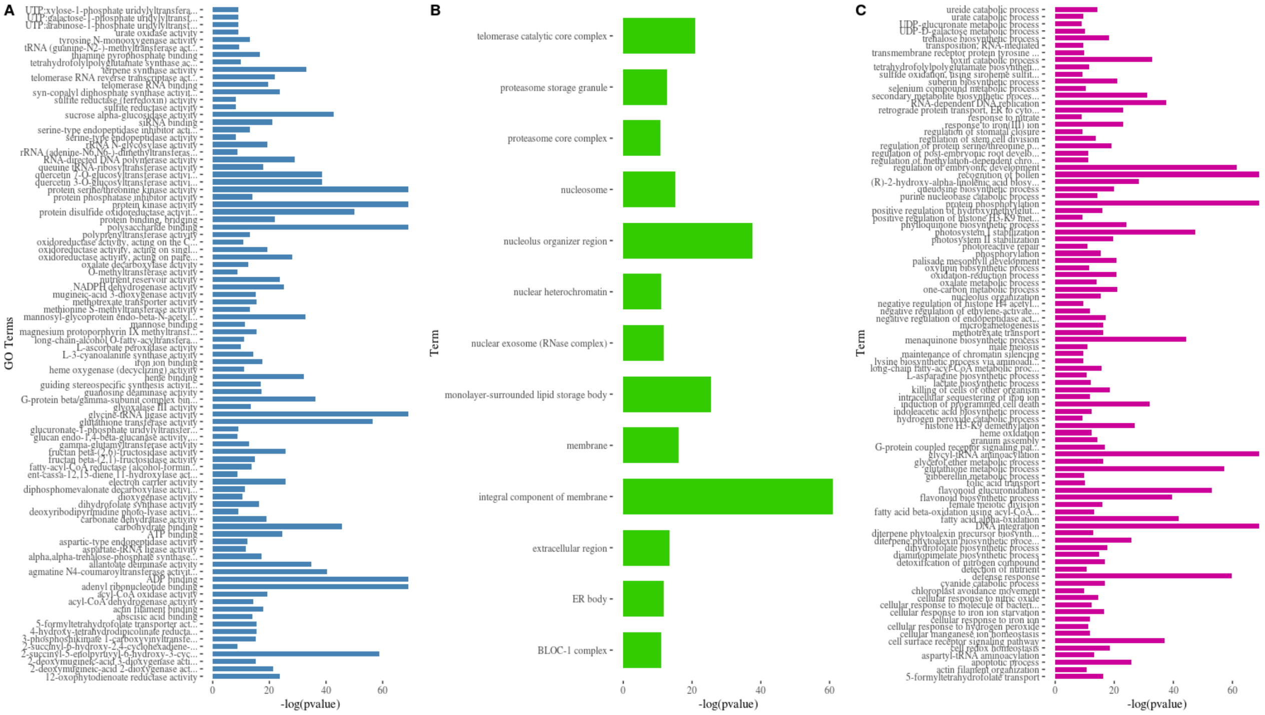

Figure 5. GO enrichment in duplicated and deleted genes. The 193 GO terms (y-axis) (FDR threshold ≤ 0.01) overrepresented in duplicated and deleted genes are plotted along the corresponding negative logarithm of their Fisher’s *P* value (x axis). **(A)** Overrepresented GO terms of the “Molecular Function”, **(B)** “Cellular Component”, and **(C)** “Biological Process” domains are reported, respectively.

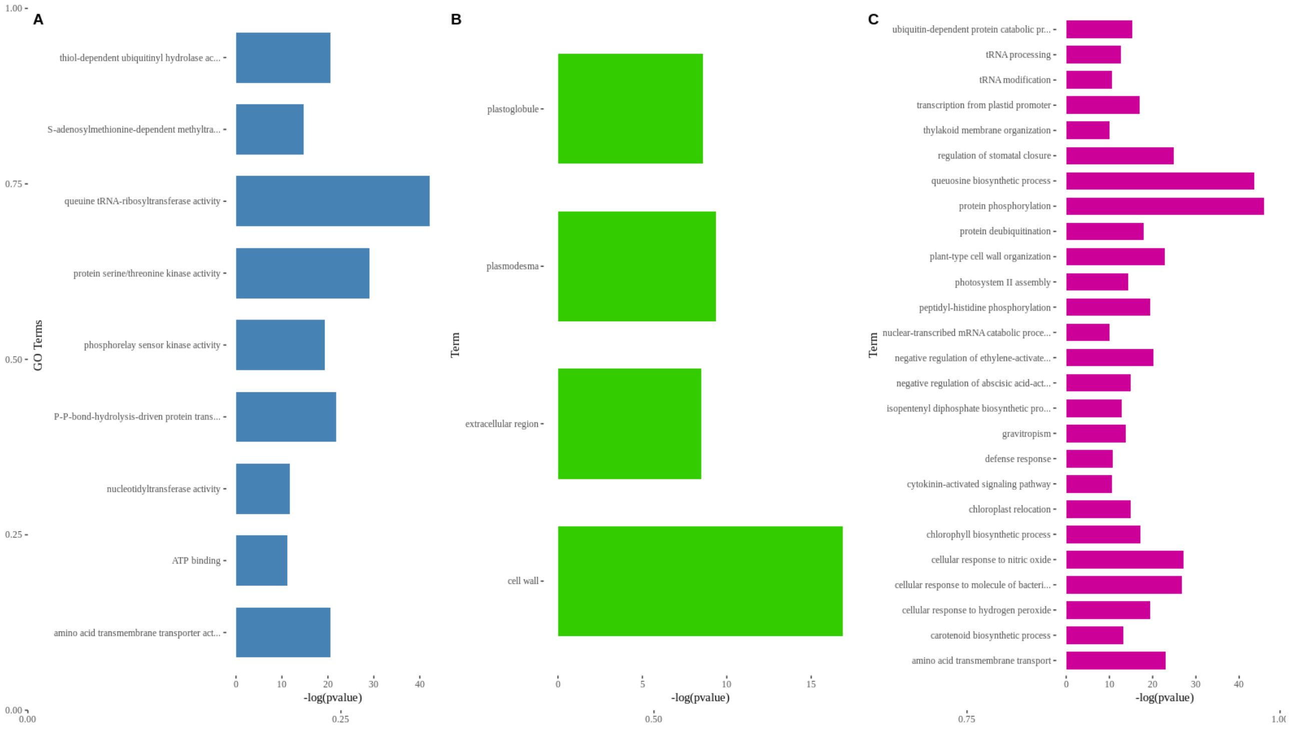

Figure 6. GO enrichment of duplicated and deleted genes differentially detected in wild and domesticated accessions. The 39 GO terms (y-axis) (FDR threshold ≤ 0.01) overrepresented in duplicated and deleted genes of wild accessions are plotted along the corresponding negative logarithm of their Fisher's *P* value (x axis). **(A)** Overrepresented GO terms of the "Molecular Function", **(B)** "Cellular Component", and **(C)** "Biological Process" domains are reported, respectively.

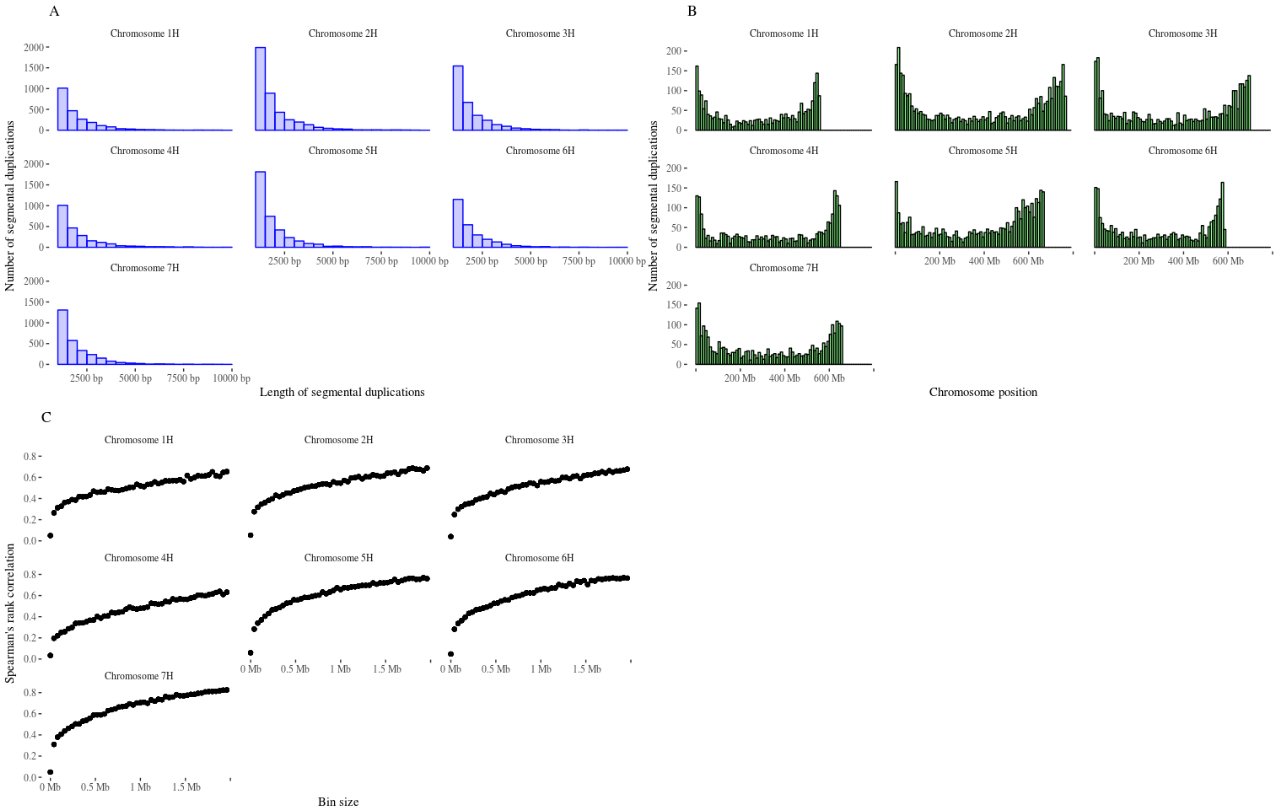

Figure 7. Frequency and length spectra of SDs and correlation with CNVs. (A) Length spectrum of SDs detected in barley cv "Morex"; **(B)** Histograms of SD distribution across the seven barley chromosomes; **(C)** For each of the seven plots, in the y-axes the values of Spearman rank correlation coefficient between SDs and CNVs were plotted, while in the x-axes the values of bin size utilized for computing the Spearman rank correlation coefficient were reported. Only statistically significant values of Spearman rank correlation coefficient with *P* values lower that 0.001 were plotted.

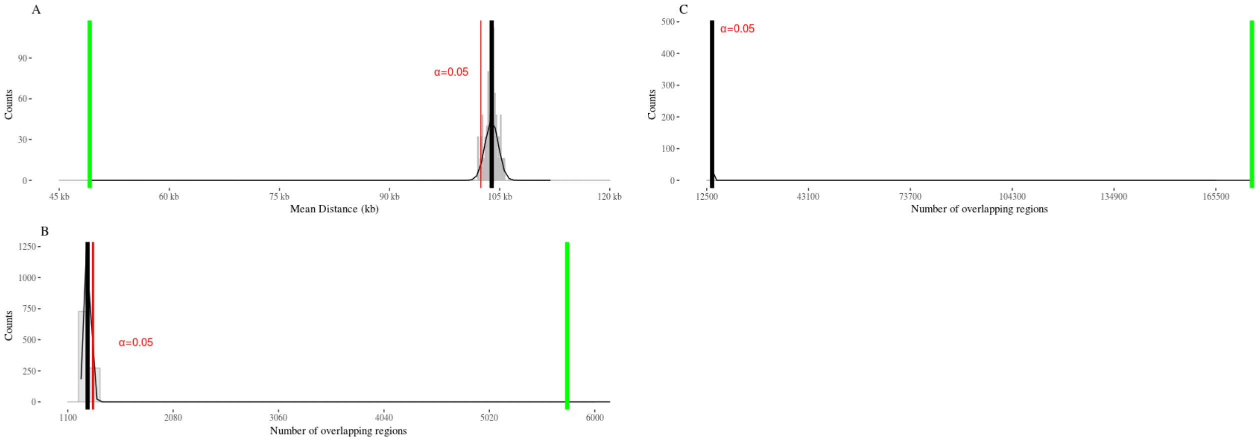

Figure 8. Association analysis of SDs based on permutation tests. In all plots, the measured value (green line) and the expected value (black line) obtained after the randomization of sequence intervals are reported. **(A)** In this plot, the average distance of SDs (x-axis) with their closest genes was compared with the lower bound of the expected average distance (red vertical line); **(B)** In this plot the number of overlaps (x-axis) between SDs and annotated genes was compared with the upper bound (red line) of the expected number of overlaps in case of random distribution. **(C)** In this plot the number of overlaps (x-axis) between SDs and CNV sites was compared with the upper bound (red line) of the expected number of overlaps.