MS. VERENA MARINA PRADE (Orcid ID : 0000-0001-6387-3982)

Article type : Original Article

The Pseudogenes of Barley

Verena M. Prade[1], Heidrun Gundlach[1], Sven Twardziok[1], Brett Chapman[2], Cong Tan[3], Peter Langridge[4], Alan H. Schulman[5], Nils Stein[6,7], Robbie Waugh[8,9], Guoping Zhang[10], Matthias Platzer[11], Chengdao Li[3,12], Manuel Spannagl[1], Klaus F. X. Mayer[1,13,*]

[1]Plant Genome and Systems Biology, Helmholtz Center Munich – German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany

[2]Centre for Comparative Genomics, Murdoch University, 90 South Street, WA6150, Murdoch, Australia

[3]School of Veterinary and Life Sciences, Murdoch University, 90 South Street, WA6150, Murdoch, Australia

[4]School of Agriculture, University of Adelaide, Waite Campus, SA5064, Urrbrae, Australia

[5]Green Technology, Natural Resources Institute (Luke), Viikki Plant Science Centre, and Institute of Biotechnology, University of Helsinki, 00014, Helsinki, Finland

[6]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, 06466 Seeland, Germany

[7]School of Plant Biology, University of Western Australia, Crawley, WA6009, Australia

[8]The James Hutton Institute, Dundee DD2 5DA, UK

[9]School of Life Sciences, University of Dundee, Dundee DD2 5DA, UK

[10]College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, 310058, China

[11]Leibniz Institute on Aging – Fritz Lipmann Institute (FLI), 07745 Jena, Germany

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/tpj.13794

[12]Department of Agriculture and Food, Government of Western Australia, South Perth WA 6151, Australia

[13]TUM School of Life Sciences Weihenstephan, Technical University of Munich, Alte Akademie 8, 85354 Freising, Germany

*Corresponding author:

Klaus F. X. Mayer
PGSB – Plant Genome and Systems Biology
Helmholtz Center Munich – German Research Center for Environmental Health
85764 Neuherberg, Germany
Email: k.mayer@helmholtz-muenchen.de
Phone: +49 89 3187 3584

Email addresses:

Verena M. Prade, verena.prade@helmholtz-muenchen.de

Heidrun Gundlach, h.gundlach@helmholtz-muenchen.de

Sven Twardziok, sven.twardziok@helmholtz-muenchen.de

Brett Chapman, bchapman@ccg.murdoch.edu.au

Cong Tan, cong.tan@murdoch.edu.au

Peter Langridge, peter.langridge@adelaide.edu.au

Matthias Platzer, matthias.platzer@leibniz-fli.de

Alan H. Schulman, alan.schulman@helsinki.fi

Nils Stein, stein@ipk-gatersleben.de

Robbie Waugh, robbie.waugh@hutton.ac.uk

Guoping Zhang, zhanggp@zju.edu.cn

Chengdao Li, c.li@murdoch.edu.au

Manuel Spannagl, manuel.spannagl@helmholtz-muenchen.de

Running head: Pseudogenes of Barley

Keywords: Pseudogenes; Barley; Hordeum vulgare; Plants; Gene fragments; Gene evolution

Abstract

Pseudogenes have a reputation of being 'evolutionary relics' or 'junk DNA'. While they are well characterized in mammals, studies in more complex plant genomes were so far hampered by the absence of reference genome sequences. Barley is one of the economically most important cereals and has a genome size of 5.1 Gb. With the first high-quality genome reference assembly available for a *Triticeae* crop, we conducted a whole genome assessment of pseudogenes on the barley genome. We identified, characterized, and classified 89,440 gene fragments and pseudogenes, scattered along the chromosomes with occasional hotspots and higher densities at the chromosome ends. Full-length pseudogenes (11,015) have preferentially retained their exon-intron structure. Retrotransposition of processed mRNAs only plays a marginal role in their creation. However, the distribution of retroposed pseudogenes reflects the Rabl configuration of barley chromosomes and thus hints towards founding mechanisms. While defense-response related parent genes were found under-represented in cultivated barley, we detected several defense related pseudogenes in wild barley accessions. 7.2% of the pseudogenes are transcriptionally active and may potentially adopt new regulatory roles.The barley genome is rich in pseudogenes and small gene fragments mainly located towards chromosome tips or as tandemly repeated units. Our results indicate non-random duplication and pseudogenization preferences and improve our understanding of gene birth and death dynamics in large plant genomes and the mechanisms that lead to evolutionary innovations.

Background

Pseudogenes are generally referred to as 'evolutionary relics' or 'junk DNA'. They are genomic sequences similar to functional genes, but they contain degenerative features such as mutations like frameshifts or premature stop codons, leading to a loss of their original function. Consequently, pseudogenes have been disregarded in routine plant genome annotations and continuative studies.

Most pseudogenes originate from a duplication event. The functional counterpart is termed 'parent' gene (Tutar 2012). In case the gene copy did not become defective immediately after its duplication, the genetic redundancy will lead to a relaxed selection pressure and the degeneration of one of the copies is tolerated. For scenarios were one copy becomes defective, a gene-pseudogene pair arises (Balakirev & Ayala 2003).

Gene duplication can be triggered by different cellular mechanisms (Podlaha & Zhang 2010). Unequal crossing-over during meiosis can lead to tandem duplications. If sister chromatids are not separated properly during cell division (nondisjunction), chromosome duplications are the result. The duplication of whole genomes, polyploidization, is particularly widespread among plants (Weiss-Schneeweiss et al. 2013). Pseudogenes originating from any of these mechanisms are termed 'duplicated' or 'non-processed' (Podlaha & Zhang 2010). Alternatively, duplication can occur via an mRNA intermediate and re-insertion of reverse transcribed cDNA into the genome. These 'retroposed' or 'processed' pseudogenes are considered as 'dead-on-arrival', because they lose their upstream promoter and regulatory sequences during duplication. Processed pseudogenes are characterized by a loss of intron sequences, poly-A tails near the 3' ends and small flanking direct repeats (Sen & Ghosh 2013). Unitary pseudogenes comprise the third type of pseudogene (Zhang et al. 2010). These are thought to arise rarely and without prior gene duplication. In human, olfactory receptor genes (387) form one of the largest gene families, which has numerous pseudogenes (415). It is hypothesized, that the development of color vision reduced the importance of odor sensing and resulted in the pseudogenization of numerous olfactory receptor genes (Vihinen 2014).

In recent years, the gene-look-alikes attracted particular attention because of reported cases of pseudogene functionality (Pink et al. 2011; Sen & Ghosh 2013). Despite their lost protein coding potential, some are still transcribed and able to play a role in regulatory processes (Balakirev & Ayala 2003; Poliseno et al. 2010). Due to the sequence similarity to bona fide genes, their transcripts can interfere with the translational machinery or be used for gene regulation via siRNA or miRNA

synthesis (Pink et al. 2011). Pseudogenes are now increasingly studied in mammals. For instance, human pseudogenes are of particular interest in the context of diseases (Pink et al. 2011; Sen & Ghosh 2013; Roberts & Morris 2013). Their altered expression has been linked to cancer, where they can now be used as markers for specific cell types (Poliseno et al. 2015). In contrast, for most plant species, there are no genome-wide pseudogene annotations available so far. Until recently, pseudogene studies in more complex plant genomes such as the *Triticeae* (e.g. wheat, barley, rye) were hampered by the absence of high-quality assembled reference genome sequences. Cultivated barley (*Hordeum vulgare L.*) is one of the five economically most important cereal species and part of the *Triticeae* tribe (Spannagl et al. 2013). Its diploid genome has a size of 5.1 gigabases (Gb) – making it 2 Gb larger than the human genome – and comprises 39,734 high-confidence gene loci (Mascher et al. 2017). Sequencing and genome assembly efforts have been hampered by its highly repetitive genome: about 80% consists of transposable elements. With one of the first true reference genome assemblies now available for a *Triticeae* crop and the first BAC-by-BAC assembly of a genome of such size (Mascher et al. 2017), we conducted a genome wide assessment of potential pseudogenes in barley. We exploited the homology of pseudogenes to their parent genes to identify them and then classified them into duplicated or retroposed pseudogenes. We studied their distribution along the chromosomes, their relation to genes and gene families and their functional potential. Then we analyzed syntenic regions between cultivated barley (cv. Morex) and four wild barley accessions (Tan et al. 2017) and investigated pseudogene differences. Our results enable a deeper understanding of pseudogenes in cultivated and wild crops and provide the basis for detailed analyses of potentially functional pseudogenes. Novel insights into the mechanisms underlying pseudogene genesis and thus a major evolutionary force underlying genome evolution are generated. Pseudogenes are a 'playground for innovations', since their usual non-functionality allows them to accumulate mutations without fitness effects. However, their gene-like structure improves their potential for subsequent resurrection and adoption of novel functional roles.

Results

Pseudogenes and gene fragments

The barley genome contains a vast amount of gene fragments and pseudogenes. Using a homology-based approach (Figure S1), we identified 89,440 potential pseudogenes, most of which constitute short gene fragments with an average coding sequence (CDS) length of only 188 base pairs (Table 1, Figure 1). In comparison, protein-coding genes have an average CDS length of roughly one kilobase (Table S1). Similar large quantities of short gene fragments have been found in the genome of hexaploid wheat (Brenchley et al. 2012). In barley, 12.3% (11,015) of the pseudogenes represent full-length copies of genes (Table 1). Those "traditional" pseudogenes cover the CDS of their parent gene by at least 80% and are called high-coverage (HC) pseudogenes hereafter.

The chromosomal and genomic distribution of pseudogenes largely remodels the distribution found for functional genes and gives a mirror image of transposable elements (Figure 2). We observed that some parent genes have a particularly large number of pseudogene homologues. Like in wheat (Brenchley et al. 2012), many of those fragments may actually be common domains found multiplied in the genome. Nevertheless, 1,560 pseudogenes are highly similar to their parent gene, both in length and sequence identity (≥98% similarity). These gene facsimiles are well represented in the duplicated (Figure 3 B), but to a smaller degree in the retroposed pseudogene class (Figure 3 C). This is consistent with the hypothesis, that retroposed pseudogenes accumulate mutations immediately (dead-on-arrival) and thus diverge faster from their parent genes (Thibaud-Nissen et al. 2009).

Retroposed pseudogenes are copies resulting from reinsertion of reverse transcribed mRNA into the genome. In contrast to duplicated pseudogenes, they lose their introns during the maturation of mRNA. Of the full-length HC pseudogenes in barley, 2,151 contain introns at corresponding parent splice sites and can be classified as duplicated pseudogenes (Table 1). In contrast, only 153 HC pseudogenes appear to originate from retrotransposition. The remainders are pseudogenes, that

cannot be classified into duplicated or retroposed based on their exon-intron structure. They are either too short to cover intron junctions (fragmented), chimeric, or their parent gene only comprises a single exon (Figure S1).

Distribution on chromosomes

Duplicated pseudogenes most often arise from unequal crossing-over during meiosis, segmental duplications or chromosome duplications and polyploidization events (Podlaha & Zhang 2010). Most plants have a long evolutionary history of duplications and chromosome rearrangements (Yu et al. 2005; Gaut et al. 2000; Heslop-Harrison & Schwarzacher 2011; Bolot et al. 2009). With a large number of pseudogenes found to be duplicated, we analyzed whether these are located in close vicinity to their parent gene, and thus are likely the result of unequal crossing-over events, or if they are more randomly distributed across the chromosomes, as expected for retroposed pseudogenes and segmental duplications derived by other mechanisms than unequal crossing-over (Figure 4). 3.1% of all HC pseudogenes are located within 50 kb of their respective parent gene. As expected, a significantly larger portion of HC pseudogenes classified as duplicated were found to be located within this close range to their parent gene (4.8%; binomial test, *p*-value 2.2×10^{-5}; Table S1). Also, pseudogenes with a higher sequence similarity to their respective parent genes are likely to have a younger divergence time or are affected by gene conversion. Tandem duplicated pseudogenes are preferentially affected by gene conversion events with their parents that potentially decelerate the sequence divergence between the pair. Indeed, we found pseudogenes in close vicinity to their parent genes to be more similar to them (Figure S2). However, this does not apply exclusively to duplicated pseudogenes, but also to retroposed pseudogenes and is indicative of gene conversions and sequence homogenization events independent of the duplication mechanism.

Moreover, not only duplicated but also retroposed pseudogenes were found to be preferentially located on the same chromosome as their parent gene (20.6%; chi-square test, *p*-value 2.4×10^{-4}). This contradicts the assumed random reinsertion of reverse-transcribed cDNA during

retrotransposition. A preferential reinsertion of the cDNA on the same chromosome, or even in the vicinity of its origin, is unlikely for an LTR-retrotransposon mediated transfer, for which the reverse transcription takes place in the cytosol. However, the presence of retroposed pseudogenes at a significantly higher rate locally or on opposing chromosome arms may be explained by an alternative scenario. As in humans, non-LTR-retrotransposons (LINEs) likely carry out reverse transcription directly at the integration site in the nucleus (Esnault et al. 2000; Kaessmann et al. 2009). In humans, the ORF1p protein of LINE L1 has been shown to bind cellular mRNA, which can serve as templates for reverse transcription (Mandal et al. 2013). The colocalization of transcription and LINE-driven reverse transcription might thus lead to a preferential retrotransposition in physical proximity to the transcribed parent gene. The barley genome contains 7,780 LINE elements within 10 kb of one of the 28,316 high-confidence genes (Wicker et al. 2017), out of the 19,173 LINE elements in the genome in total (Mascher et al. 2017). Compared to the chromosomes of many other eukaryotes, the individual barley chromosomes fold back to juxtapose the long and short arms (Mascher et al. 2017). This so-called Rabl configuration is adopted in interphase nuclei and leads to reduced distances between corresponding chromosome arms (Dong & Jiang 1998). The Rabl configuration thus might increase the probability of retroposed pseudogenes to insert on the same chromosome as the parent gene. Indeed, we found many intrachromosomally retroposed pseudogenes to be located either close to their parent gene or on the opposing chromosome arm, respectively (Figure S3).

Tandem gene cluster and larger gene families are birthplaces for pseudogenes

Manual inspection of barley pseudogenes in the Integrative Genomics Viewer (IGV) (Thorvaldsdóttir et al. 2013) hinted towards pseudogene hot-spots at tandem gene clusters. To statistically confirm this, we assessed the proportion of pseudogenes in close vicinity to tandem gene clusters. Considering all HC parent genes, 8.7% of them are located in close vicinity to at least one of their HC pseudogene 'descendants'. However, if we focus only on HC parent genes located within a tandem gene cluster, we find a significantly increased proportion (37.8%; binomial test, *p*-value 2.2×10^{-16}) to

be close to a HC pseudogene 'descendant'. The observed four-fold relative difference supports the hypothesis of an accumulation of pseudogenes in tandem gene clusters.

Additionally, we confirmed a positive correlation between gene family size and HC pseudogene content (Figure 5 A). Not surprisingly, larger gene families are more likely to give rise to pseudogenes (Zou et al. 2009), since expansion gives the opportunity to evolve new functionalities but also to balance eventual pseudogenization of individual gene family members. To study how the pseudogene content changes during gene family expansion or contraction, we compared orthologous groups of barley to *Arabidopsis thaliana*, *Brachypodium distachyon*, rice and Sorghum (Figure 5 B). Barley contains 1,954 expanded and 117 contracted orthologous groups (Mascher et al. 2017). The relative number of pseudogenes per gene family is higher for expanded orthologous groups than for contracted orthologous groups. Consequently, gene duplications leading to an expansion of gene families also go hand in hand with pseudogene creation. Respectively, the contraction of gene families does not lead to a high number of pseudogenes. Either the genes have degenerated beyond recognition or their sequence has been deleted entirely.

Are all pseudogenes non-functional?

Even if degenerated and transcriptionally inactive, pseudogenes may still serve as a repertoire of gene-like sequences with the 'capacity to shape an organism during evolution' (Brosius & Gould 1992). Since it is difficult to prove dysfunctionality – a dogmatic key feature of pseudogenes – there have been several reported cases of pseudogenes which turned out to exert functions (Pink et al. 2011; Sen & Ghosh 2013). To examine the functional potential and background of barley pseudogenes, we first analyzed the functional annotation of the parent gene set and undertook an enrichment analysis (Figure 6). We found that genes involved in transport, pollination or protein processing are over-represented in the parent gene set. In contrast, genes involved in defense response, stress responses, cell wall organization or sexual reproduction give rise to fewer pseudogenes in barley cv. Morex.

Transcribed pseudogenes have the potential to contribute to the regulation of their parent genes (Pink et al. 2011; Sen & Ghosh 2013). However, their sequence similarity hampers transcriptional analysis using RNA-seq data, since reads can map ambiguously to both pseudogenes and functional genes. We therefore used only reads mapping uniquely onto the pseudogene sequences. We found transcription evidence for 6,435 (7.2%) pseudogenes, 1,243 (11.3%) of them from the HC pseudogene set (Table S1). This result is likely an underestimation due to the unique mapping of the RNA-seq data. In comparison, about 20% of the annotated pseudogenes in *Arabidopsis thaliana* and rice are reported to be actively transcribed (Podlaha & Zhang 2010). Many of the transcribed pseudogenes in barley originate from genes involved in glycolysis or glucose metabolic processes (Figure S4). While evidence for transcription does not necessarily imply functionality, it nevertheless can highlight pseudogenes with regulatory potential.

Selective pressure

We applied a Ka/Ks analysis as an indicator for selective pressure on homologous gene pairs. Pseudogenes are usually under neutral evolution (Podlaha & Zhang 2010), and we expected a balanced rate of synonymous and non-synonymous substitutions between pseudogenes and their parent genes. Instead, we obtained a Ka/Ks ratio distribution significantly shifted to the left, usually seen as indicative for conservation pressure (Figure S5 A). Thibaud-Nissen et al. reported similar findings for rice and also gave a convincing explanation (Figure S5 B): The parent gene sequences found do not necessarily reflect the parental genes at the time of duplication. Genes accumulate primarily synonymous substitutions, while pseudogenes are expected to accumulate random mutations. If the present parent gene is compared to the pseudogene sequence, a Ka/Ks ratio below 1 is obtained (Thibaud-Nissen et al. 2009).

Duplicated functional genes with defects

Gene duplication is a genomic process to create new genes and functionalities via neo- and subfunctionalization. In most cases however it leads to pseudogenization (Ho-Huu et al. 2012; Xiao et al. 2016; Kondrashov et al. 2002). We identified functional gene duplicates, which contain a shortened CDS, for example due to premature termination codons. We found 4,100 (10.8%) barley genes with such evidence, 255 of which exhibit premature stop codons, but which are otherwise highly similar to the original version. This result illustrates the other side of the coin: If pseudogenes are interpreted as the byproduct of a mechanism generating new genes, those 4,100 shortened genes can be interpreted as evidence for the generation of new functional genes. Similar to the interpretations by Brosius & Gould (1992) (Brosius & Gould 1992), this also confirms the role of pseudogenes as a repertoire of potential genes. Subfunctionalization, neofunctionalization, and also pseudogene resurrection are possible outcomes of gene duplication events and drivers in the gene evolution of genomes.

Comparing syntenic regions between cultivated and wild barley genotypes

Our pseudogene annotation provides the necessary background for detailed analyses of pseudogene evolution, of their impact on genome structure and dynamics, and of their potential to interfere with gene regulation. To investigate their evolution in barley subspecies and cultivars, we screened syntenic regions between cultivated barley cv. Morex and four wild barley lines for differences in gene and pseudogene content. Two of the additional genome assemblies are from wild barley accessions growing on opposing slopes of the 'Evolution Canyon' I in Israel (Tan et al. 2017). The north- and south-facing slopes (NFS and SFS respectively) of the canyon are only separated by 250 meters but nevertheless exposed to drastically different microclimates. The remaining two genome assemblies are from Tibetan wild barley (Tan et al. 2017). We used high-confidence gene models of barley cv. Morex to find homologous gene-like sequences on all four wild barley genome assemblies. We then scrutinized all identifiable syntenic regions and selected specific loci for in-depth analysis of

very recent pseudogenization events in cultivated and wild barley accessions. While assembly quality and sequencing depth differ and complicate genome wide analysis and statistics, individual gene-pseudogene examples illustrate typical pseudogenization scenarios in closely related subspecies.

We found a duplicated gene triplet in the wild barley accession growing on the SFS of the Evolution Canyon (Figure 7 A). The redundant gene copies contain deletions and insertions, leading to shifts in the reading frame and to premature stop codons. Interestingly, the same triplet is neither duplicated in barley cv. Morex nor the wild barley line growing on the opposing slope of the Evolution Canyon. The gene triplet codes for two leucin-rich repeat receptor-like protein kinases (LRR-RK) and a hexosyltransferase (HT). LRR-RKs comprise a large protein family in plants and regulate developmental and defense-related processes (Torii 2004). The longer one of the two LRR-RK genes in the wild NFS barley line shows a 13 base pair deletion compared to the homologue in barley cv. Morex. This deletion is located in the 5' half of the coding sequence, resulting in a frameshift and premature stop codons, therefore massively disrupting the derived amino acid sequence. Another gene in the same syntenic region, but not part of the duplicated triplet is a polyphenol oxidase (PPO) gene, which became a pseudogene in both wild barley populations from the Evolution Canyon but is found intact in the cultivated barley sequence. Again, a frameshift leads to premature termination codons. Plant PPOs are enzymes responsible for the browning reaction following tissue damage (Tran et al. 2012). They have been suggested to take part in defense response mechanisms. Another syntenic region harboring a potential unitary pseudogene contains a calcium-binding protein (CABP) gene, which is pseudogenized in one wild barley line (Figure 7 B). In humans, CABPs have been shown to be important regulators of key calcium influx channels, which are enriched in neuronal tissue (Haynes et al. 2012). In plants, calcium is an important messenger of external signal transduction cascades and as such plays an essential role in the reaction of plants to external stimuli, such as pathogen attack (Poovaiah et al. 1993). In chloroplasts, calcium is involved in photosynthesis, carbon fixation, CO_2 fixation, protein transport, and protein phosphorylation (Rocha & Vothknecht 2013). The homologues of this CABP gene in both wild barley lines from the Evolution Canyon exhibit

a one base pair deletion at the beginning of the coding sequence. This results in a frameshift and leads to premature stop codons in the NFS accession. However, another one base pair insertion in the SFS accession restores the correct reading frame. The most parsimonious sequence of events is that the CABP gene in the SFS accession first pseudogenized and was subsequently restored by a counteracting mutation. This example illustrates the continuous transitions between genic and pseudogenic states, which sometimes can even lead to small stretches of drastically changed protein sequence by the transitionally out-of-frame sequence. The described events could well represent a common mechanism for novelty introduction and highlight pseudogenes as a 'playground for innovations'.

Transposable elements occupy over 80% of the genomic space in barley and have a strong impact on genome structure. Duplications or rearrangements are often a consequence of transposon mobilization and insertion. We found a greatly expanded genomic region in barley cv. Morex, which experienced repetitive element insertions resulting in rearrangements, duplications and pseudogenization (Figure 7 C). While in all four wild barley accessions the syntenic LRR and NADH kinase (NADK) genes are only separated by ~500 base pairs, the respective functional copies in cultivated barley span over 20,000 base pairs, a 40-fold increase in size. To which extent this massive difference is attributable to underlying assembly problems remains speculative for the time being. However, also the bordering gene containing regions exhibit differences to the barley cv. Morex genome and thus are indicative of pseudogenization. The functional NADK gene is intact over its entire length, but the LRR gene is shortened due to a frameshift and premature stop codon. However, it still can be regarded as a functioning protein coding gene. There is another copy of the NADK gene, which is split into two elements, probably due to repetitive element insertion into the intron of the duplicate. Both fragments are pseudogenized and contain premature termination codons. In conclusion, this region likely experienced massive expansion, rearrangements and duplications leading to pseudogenization in barley cv. Morex, while in all four wild barley genomes, this region is largely similar. Even though assembly differences cannot be excluded, it might also

indicate that transposable element insertions, pseudogene generation and rearrangements in this region occurred during or after domestication less than 10,000 years ago.

Discussion

The barley genome is rich in full-length HC pseudogenes and numerous small gene fragments. While it comprises 39,734 high-confidence gene loci (Mascher et al. 2017), we found more than twice as many pseudogenes and gene fragments (89,440). A major source for pseudogenes seems to be unequal crossing-over leading to tandem gene situations. This is affirmed by their retained exon-intron structure, their gene-like chromosomal distribution and the small distance to their respective parent genes. In mammals, retroposed pseudogenes were found to outnumber duplicated pseudogenes (Sisu et al. 2014; Podlaha & Zhang 2010). In barley, retrotransposition only plays a marginal role in pseudogene creation. This is surprising, because more than 75% of the barley genome is composed of LTR-retrotransposons, including ~25,000 full-length and potentially active elements (Spannagl et al. 2013). In comparison, only 45% of the human genome is derived from transposable elements, including 8% LTR-retrotransposons and large amounts (33.7%) of non-LTR-retrotransposons, specifically 16% LINE-1 (L1) retrotransposons (Cordaux & Batzer 2009). Thus, differences in the prevalence of retroposed pseudogenes cannot be explained by the differential repetitiousness of the genomes, but may be linked to the transposable element composition. The enzymatic machinery of LINEs is responsible for the generation of human processed pseudogenes (Pavlicek et al. 2006). In barley, LINE retrotransposons comprise less than 1% of the genome (Mascher et al. 2017), which could explain the low amounts of retroposed pseudogenes despite the high overall retrotransposon content. The dominance of duplicated pseudogenes over retroposed pseudogenes is not unique for barley. It has also been observed in other plants, such as *Arabidopsis thaliana* and rice (Wang et al. 2012; Thibaud-Nissen et al. 2009).

The non-random chromosomal distribution of retroposed pseudogenes suggests that the reverse transcription of mRNA may not take place exclusively outside the nucleus. The combination of LINE reverse transcription on a chromosomal primer at a nick site (Kaessmann et al. 2009) with the capacity of LINEs to package cellular RNA (Mandal et al. 2013) leads us to propose that LINEs may be responsible for the preferential distribution of retroposed pseudogenes in close spatial proximity to the gene from which they are derived. This LINE-based mechanism also would be consistent with the differential proportions of retroposed pseudogenes in the human and barley genomes. The Rabl conformation of barley chromosomes (Mascher et al. 2017; Dong & Jiang 1998) results in a neighboring arrangement of short and long chromosome arms in the interphase nucleus. Structural constraints imposed by this configuration could support a preferential reinsertion of retroposed pseudogenes on the opposing chromosome arm to the respective parent gene.

Most of the 89,440 barley pseudogenes are small gene fragments and probably constitute common domains present in high copy numbers. Double-strand DNA break repair mechanisms, so-called non-homologous DNA end joining (NHEJ) or synthesis-dependent strand annealing (SDSA), might be responsible for these short gene fragments, as they are associated with the insertion of filler DNA at the break sites (Wicker et al. 2010; Gorbunova & Levy 1997; Gorbunova & Levy 1999). These processes do not target genes specifically, rendering these short gene fragments symptomatic of the repair mechanism. We found evidence for non-random duplication and pseudogenization preferences especially for genes in tandem clusters, as well as for genes in large or expanded gene families in barley. High duplication rates may be beneficial to rapidly adapt to environmental changes but might also escape dosage compensation mechanisms and thus might be harmful.

We scrutinized syntenic regions between barley cv. Morex and four wild barley accessions for differences in the pseudogene complements and found tandem gene duplications, pseudogenization and sequence rearrangements between the closely related subspecies. However, more detailed comparative analyses were hampered by differences in assembly qualities. While only short contig

assemblies were available for the wild barley accessions, the BAC-by-BAC genome assembly of barley cv. Morex provides more complete chromosome sequences. With improved assemblies available in the near future these limitations will be overcome and more detailed comparative analyses between wild and domesticated species and cultivars will become feasible.

Conclusion

With the availability of an increasing number of genome reference assemblies, comparative analyses for plants with large and complex genome structures become feasible. The barley genome has recently been sequenced and assembled into chromosomal pseudomolecules, enabling us to perform a whole-genome assessment of pseudogenes. We found almost 90.000 pseudogenes and gene fragments, whose analysis sheds light on gene evolution and genome dynamics. There are not only significant differences regarding pseudogenes between mammals and plants, but also between closely related species. The pseudogene complement in domesticated barley and among subspecies growing in different microclimates was found to differ. The main obstacles for comparative analyses remain assembly and annotation qualities. Further studies and conclusions about the effect and the origin of pseudogenes in the evolution and domestication of crop plants will soon be possible and provide an exciting opportunity.

Methods

The identification of pseudogenes was done computationally by exploiting their sequence homology to functional genes. To achieve this, the high-confidence gene set of barley cv. Morex was used as reference to identify gene-like sequences in the genome. Pseudogenes overlapping with high-confidence genes or with transposable element sequences were filtered.

First, the Morex barley pseudomolecules (Mascher et al. 2017) were split into batches to allow for parallel processing. Transposable elements and transposon genes (Mascher et al. 2017) were N-masked to reduce nonspecific hits. The CDS nucleotide sequences of all high-confidence gene isoforms (Mascher et al. 2017) (39,734 loci, 248,180 isoforms) that had no indication of being transposable element related (38,157 loci, 240,113 isoforms) were then mapped onto the genome sequence using BLAT (Kent 2002) (minimal identity 70%, max. intron length 2,500 base pairs), which creates spliced alignments and thus recovers the exon-intron structures. Short BLAT hits with a length smaller than 50 base pairs or containing only fragments (exons) shorter than 25 base pairs were filtered. Gaps (introns) up to a size of 9 base pairs were closed and considered in the calculation of the sequence identity. Premature termination codons were then determined independently for each pseudogene exon, always starting in the correct frame of the parent gene.

Gene self-hits as well as hits overlapping with other high-confidence genes were filtered out completely, but used to determine if a gene is a shortened copy of another gene. Nonspecific hits, as well as hits with low information content, were filtered using the WU-BLAST dust filtering (Gish n.d.) (default settings) and the Tandem Repeats Finder (Benson 1999) (max. 65% masked, ≥50 base pairs remaining). In case BLAT hits were overlapping, the longest hit was chosen as a representative for the locus. If multiple hits with the same maximum length were present at one locus, then the one with the highest sequence identity to its parent was chosen as representative. If the representative covered less than 60% of the locus, then all hits shorter than half of the representative and overlapping with it are removed, as well as the hit with the shortest exon length but also the longest total length. This allowed the hit cluster to split up into multiple loci and newly determined representatives to be of good quality. A final filtering step removed BLAT hits from genes with 50 or more children. Those genes were under strong suspicion to be related to transposable elements.

The low-confidence gene set of barley contains ~41,000 gene-like sequences that do not fulfill the criteria for canonical genes, including potential pseudogenes. 1,863 annotated low-confidence genes (4.6%) overlap with at least 50% of their coding sequence with a pseudogene.

Pseudogene classification

The presence or absence of intron sequences in pseudogenes was used to classify them into duplicated or retroposed pseudogenes. Since not all pseudogenes are complete gene copies, some do not span over splice sites rendering this type of classification impossible for them. For the intron loss/retention criterion, we defined five pseudogene classes (Figure S1): (1) 'duplicated' pseudogenes still containing introns at each covered splice site; (2) 'retroposed' or 'processed' pseudogenes which have lost all introns; (3) 'chimeric' pseudogenes with both retained and lost introns; (4) 'single-exon parent' pseudogenes from isoforms with only one exon; (5) 'fragmented' pseudogenes which do not sufficiently cover a splice site. A splice site is only covered, if at least 10 base pairs of the exons on either side are present in the duplicate. The gap has to be at least 30 base pairs long, to be considered a duplicated intron.

Chromosomal distribution of pseudogenes and other elements

Densities of gene, pseudogenes and transposons along the chromosomes where calculated with a sliding window of 5 Mb and a shift size of 1 Mb as percent sequence coverage. Circular Figures were created using Circos version 0.69-4 (Krzywinski et al. 2009).

Ka/Ks analysis

To determine the selection pressure on pseudogenes, the sequences of pseudogene/parent gene pairs need to be aligned and edited. We used clustalw2 (Larkin et al. 2007) (default) for pairwise alignment and removed codons containing gaps or Ns, as well as premature termination codons. The alignment always was kept in the frame of the gene. In order for the subsequent analysis to work correctly, a minimal alignment length of 150 base pairs was preconditioned. Codeml from the PAML

package (Yang 2007) was used to calculate Ka and Ks values. Highly similar sequences led to extreme log10 Ka/Ks values (e.g. ≥99). For the statistical analysis, we filtered for log-values between minus four and four (32,021 log10 Ka/Ks values remained after all filtering steps). We used the scipy 'normaltest' from python to test for a normal distribution and the scipy one-sample t-test 'ttest_1samp' to test, whether the distribution is significantly shifted from the expected mean of zero.

Gene Families and orthologous groups

Gene families were determined by first using BLAST (Altschul et al. 1990) (blastn) on the representative gene splice variants with an e-value threshold of 1×10^{-5}. Then mcxdeblast was used and its output forwarded to mcl (van Dongen 2000; Enright et al. 2002). Orthologous groups were defined from the barley high-confidence class genes and the annotated gene sets of three grasses from diverse grass sub-families (*Sorghum bicolor*, *Brachypodium distachyon,* and *Oryza sativa*) and *Arabidopsis thaliana* using OrthoMCL software version 2.0 (OrthoMCL default parameters). A total of 170,925 coding sequences from these five species were clustered into 24,337 gene families. 8,608 clusters contained sequences from all five genomes. Expanded gene families were extracted as described in Mascher et al. (2017) (Mascher et al. 2017).

GO analysis

To find under- or over-represented Gene Ontology (GO) terms in the parent gene set compared to the complete gene set (subontology: Biological Process), we used the free open-source GOstats R package (Falcon & Gentleman 2007) with a *p*-value cutoff of 0.05. The resulting GO terms were then grouped with REVIGO (Supek et al. 2011) using a similarity threshold of 0.5 and *Arabidopsis thaliana* as GO term database.

RNAseq analysis

Hisat2 was used to align RNA-seq reads (Mascher et al. 2017) to the barley genome (options: --dta-cufflinks). Samfiles were then filtered for a minimal mapping quality value of 60, converted into BAM files and sorted using Samtools (v 1.3). Cufflinks and Cuffcompare (2.2.1) were then used to assemble alignment files to a single set of transcripts. It was then examined whether there is transcriptional evidence for pseudogenes and for HC pseudogenes in particular. A pseudogene was considered to be transcribed if at least 50 base pairs of its sequence overlapped with transcription evidence in either direction.

Shortened genes within the gene set

BLAT hits, which were filtered because they overlapped with annotated genes, were used to determine whether a gene is a shortened copy of another gene. A precondition was that the homology of the shortened gene to the longer gene extended beyond its own CDS. The shorter gene had to be covered by the hit by at least 60%, with either less than 60% of the hit overlapping with the short gene or the hit being at least 100 base pairs longer than the short gene at that position.

Tandem genes and pseudogenes

Coding sequences of genes were clustered using CD-HIT (Fu et al. 2012; Li & Godzik 2006) (80% identity threshold) and tandem gene groups were then defined from the resulting clusters by applying a maximum distance requirement of 50,000 base pairs between any pair of genes. Pseudogene children of tandem genes were considered part of the tandem group, if their distance to any of its gene members did not exceed 50,000 base pairs as well.

Syntenic regions between cultivated and wild barley lines

We used four wild barley genome assemblies (Tan et al. 2017) to investigate differences in gene and pseudogene complements between the two closely related species. A filtering for contigs and scaffolds with a minimal length of 200 base pairs and a maximum of 35% Ns was performed in an

attempt to remove sequences of bad quality. We then used an equivalent of the pseudogene detection pipeline to map the representative isoform of our domesticated barley gene CDS (Mascher et al. 2017) onto the four assemblies. The resulting hits formed a collection of genes and pseudogenes, which all have a parent gene homologue from the Morex barley gene set. Hits were classified as genes if they met all following requirements: (1) Nucleotide differences must not lead to premature termination codons shortening the CDS by more than 15 nucleotides; (2) their sequence identity compared to the Morex homologue is at least 95%; (3) the CDS of the Morex homologue is covered to at least 98% if the hit has a length smaller than 800 base pairs, otherwise it has to be covered to at least 75%. This very stringent definition led to low gene numbers, which is why the remaining hits were divided into pseudogenes with premature stop codons and potential pseudogenes without premature stop codons. Often, potential pseudogenes were located at the borders of a scaffold, resulting in shortened annotations and low coverage. To be able to better estimate whether an element is a gene or a pseudogene, elements of interest were individually examined and aligned to their parent gene using megablast or blastn (Altschul et al. 1990). To investigate syntenic blocks, we focused on and visualized contigs and scaffolds which contain at least three genes with homologues on a stretch of maximal 1 Mb of the same Morex *H. vulgare* chromosome. Since the sequence data for the wild barley populations from the Evolution Canyon was a combination of two assembly versions, possible duplicates of the same locus were removed in the visualizations. The sequence of the higher quality assembly version was kept. A pairwise comparative visualization of syntenic blocks was created between Morex barley and each of the four wild barleys, if available. A CD-HIT clustering (95% identity, 80% coverage in both directions) of the Morex query gene CDS was used to determine the connections of syntenic genes or pseudogenes. Any element pair from the same cluster is connected in the visualization. This resulted in over 800 syntenic block pairs. If they share at least one gene in Morex barley, syntenic block Figures were then combined to allow for the comparison of more than two barley lines. The resulting 203 syntenic shared blocks were manually scrutinized and three loci of interest were selected.

Abbreviations

IBSC: The International Barley Genome Sequencing Consortium; Gb: gigabase; Mb: megabase; CDS: coding sequence; HC: high-coverage (>80%); NFS: north-facing slope; SFS: south-facing slope; LRR-RK: Leucin-rich repeat receptor-like protein kinase; HT: Hexosyltransferase; PPO: Polyphenol oxidase; CABP: Calcium-binding protein; NADK: NADH kinase; LTR: long terminal repeat; NHEJ: non-homologous DNA end joining; SDSA: synthesis-dependent strand annealing; GO: gene ontology.

Acknowledgements

We would like to thank Mats Hansson, Ilka Braumann and Carlsberg Research Laboratory, Copenhagen, Denmark for prepublication access to the Morex barley assembly data (chromosome 6 and 0). Additionally, we thank Jimmy Omony for his help with the statistical analysis.

Declarations

Availability of data and material

The data generated and analyzed during the current study are available from the PGSB ftp site at ftp://plantftp.helmholtz-muenchen.de/barley/. Sequence data for wild barley accessions have been deposited in the sequence read archive (SRA) with the SRA identifier SRP076351.

Competing interests

The authors declare that they have no competing interests.

Funding

This work is funded by the German Ministry of Education and Research (BMBF) grant 0314000 “BARLEX” & 0315954 “TRITEX” to K.F.X.M., M.P. and N.S and 031A536 “de.NBI” to K.F.X.M.

Authors’ contributions

V.M.P. analyzed and interpreted the data. H.G. provided the transposon annotation and interpreted the data. S.T. performed the alignment and assembly of the RNA-seq reads. S.T. provided the gene annotation. N.S., R.W., C.L., G.Z., P.L., A.H.S., R.W. and M.P. provided prepublication access to the assembly and transcription data. C.L., B.C and C.T. contributed to the sequences of the wild barley accessions. V.M.P., K.F.X.M. and H.G. wrote the paper. All authors read, contributed to, and approved the final manuscript.

Short Supporting Information Legends

Figure S1. Computational pseudogene detection and classification pipeline.

Figure S2. Sequence identity vs. distance of pseudogenes compared to their parent genes.

Figure S3. Distribution of transposable elements, genes and intrachromosomally retroposed pseudogenes on the seven chromosomes of barley.

Figure S4. Over- and under-represented Gene Ontology (GO) terms for the parent gene set of transcribed pseudogenes compared to the complete gene set of the barley genome.

Figure S5. Analysis of the relative rates of synonymous and non-synonymous substitutions between pseudogenes and their parent genes.

Table S1. Metrics for pseudogenes and high-coverage (HC) pseudogenes, as well as template and parent genes in the barley genome.

References

Altschul, S.F. et al., 1990. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215, pp.403–410.

Balakirev, E.S. & Ayala, F.J., 2003. Pseudogenes: Are They "Junk" or Functional DNA? *Annual review of genetics*, 37, pp.123–151.

Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, 27(2), pp.573–580.

Bolot, S. et al., 2009. The "inner circle" of the cereal genomes. *Current Opinion in Plant Biology*, 12(2), pp.119–125.

Brenchley, R. et al., 2012. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, 491(7426), pp.705–710.

Brosius, J. & Gould, S.J., 1992. On "genomenclature": A comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA." *Proceedings of the National Academy of Sciences of the United States of America*, 89(22), pp.10706–10.

Cordaux, R. & Batzer, M.A., 2009. The impact of retrotransposons on human genome evolution. *Nature reviews. Genetics*, 10(10), pp.691–703.

Dong, F. & Jiang, J., 1998. Non-Rabl Patterns of Centromere and Telomere Distribution in the Interphase Nuclei of Plant Cells. *Chromosome Research*, 6(7), pp.551–558.

van Dongen, S., 2000. *Graph Clustering by Flow Simulation.* University of Utrecht.

Enright, A.J., Van Dongen, S. & Ouzounis, C.A., 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7), pp.1575–1584.

Esnault, C., Maestre, J. & Heidmann, T., 2000. Human LINE retrotransposons generate processed pseudogenes. *Nature genetics*, 24(4), pp.363–7.

Falcon, S. & Gentleman, R., 2007. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2), pp.257–258.

Fu, L. et al., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23), pp.3150–3152.

Gaut, B.S. et al., 2000. Maize as a model for the evolution of plant nuclear genomes. *PNAS*, 97(13), pp.7008–7015.

Gish, W., WU BLAST.

Gorbunova, V. & Levy, A.A., 1999. How plants make ends meet: DNA double-strand break repair. *Trends in Plant Science*, 4(7), pp.263–269.

Gorbunova, V. & Levy, A.A., 1997. Non-homologous DNA end joining in plant cells is associated with deletions and filler DNA insertions. *Nucleic Acids Research*, 25(22), pp.4650–4657.

Haynes, L.P., McCue, H. V & Burgoyne, R.D., 2012. Evolution and functional diversity of the Calcium Binding Proteins (CaBPs). *Frontiers in molecular neuroscience*, 5(February), p.9.

Heslop-Harrison, J.S.P. & Schwarzacher, T., 2011. Organisation of the plant genome in chromosomes. *The Plant Journal*, 66(1), pp.18–33.

Ho-Huu, J. et al., 2012. Contrasted patterns of selective pressure in three recent paralogous gene pairs in the Medicago genus (L.). *BMC evolutionary biology*, 12, p.195.

Kaessmann, H., Vinckenbosch, N. & Long, M., 2009. RNA-based gene duplication: mechanistic and

evolutionary insights. *Nature reviews. Genetics*, 10(1), pp.19–31.

Kent, W.J., 2002. BLAT — The BLAST-Like Alignment Tool. *Genome research*, 12, pp.656–664.

Kondrashov, F.A. et al., 2002. Selection in the evolution of gene duplications. *Genome biology*, 3(2), p.research0008.1-0008.9.

Krzywinski, M.I. et al., 2009. Circos: an information aesthetic for comparative genomics. *Genome Research*, 19(9), pp.1639–1645.

Larkin, M.A. et al., 2007. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21), pp.2947–2948.

Li, W. & Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13), pp.1658–1659.

Mandal, P.K. et al., 2013. Enrichment of processed pseudogene transcripts in L1-ribonucleoprotein particles. *Human Molecular Genetics*, 22(18), pp.3730–37.

Mascher, M. et al., 2017. A chromosome conformation capture ordered sequence of the barley genome. *Nature*, 544, pp.427–433.

Pavlicek, A. et al., 2006. Retroposition of processed pseudogenes: The impact of RNA stability and translational control. *Trends in Genetics*, 22(2), pp.69–73.

Pink, R.C. et al., 2011. Pseudogenes: Pseudo-functional or key regulators in health and disease? *RNA*, 17, pp.792–798.

Podlaha, O. & Zhang, J., 2010. Pseudogenes and Their Evolution. *Encyclopedia of Life Sciences (ELS)*, pp.1–8.

Poliseno, L. et al., 2010. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, 465(7301), pp.1033–1038.

Poliseno, L., Marranci, A. & Pandolfi, P.P., 2015. Pseudogenes in human cancer. *Frontiers in Medicine*, 2, p.68.

Poovaiah, B., Reddy, A. & Feldman, L., 1993. Calcium and Signal Transduction in Plants. *Critical Reviews in Plant Sciences*, 12(3), pp.185–211.

Roberts, T.C. & Morris, K. V, 2013. Not so pseudo anymore: pseudogenes as therapeutic targets. *Pharmacogenomics*, 14(16), pp.2023–34.

Rocha, A. & Vothknecht, U., 2013. Identification of CP12 as a Novel Calcium-Binding Protein in Chloroplasts. *Plants*, 2(3), pp.530–540.

Sen, K. & Ghosh, T.C., 2013. Pseudogenes and their composers: delving in the “debris” of human genome. *Briefings in Functional Genomics*, 12(6), pp.536–547.

Sisu, C. et al., 2014. Comparative analysis of pseudogenes across three phyla. *Proceedings of the National Academy of Sciences of the United States of America*, 111(37), pp.13361–13366.

Spannagl, M. et al., 2013. Analysing complex Triticeae genomes - concepts and strategies. *Plant Methods*, 9, p.35.

Supek, F. et al., 2011. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PloS one*, 6(7), p.e21800.

Tan, C. et al., 2017. Sympatric ecological speciation of wild barley driven by microclimate at Evolution Canyon, Israel. *submitted*.

Thibaud-Nissen, F., Ouyang, S. & Buell, C.R., 2009. Identification and characterization of pseudogenes in the rice gene complement. *BMC genomics*, 10, p.317.

Thorvaldsdóttir, H., Robinson, J.T. & Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): high-

performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2), pp.178–192.

Torii, K.U., 2004. Leucine-Rich Repeat Receptor Kinases in Plants: Structure, Function, and Signal Transduction Pathways. *International Review of Cytology*, 234, pp.1–46.

Tran, L.T., Taylor, J.S. & Constabel, C.P., 2012. The polyphenol oxidase gene family in land plants: Lineage-specific duplication and expansion. *BMC Genomics*, 13(1), p.395.

Tutar, Y., 2012. Pseudogenes. *Comparative and functional genomics*, 2012, p.424526.

Vihinen, M., 2014. Contribution of Pseudogenes to Sequence Diversity. In L. Poliseno, ed. *Pseudogenes: Functions and Protocols*. pp. 15–24.

Wang, L. et al., 2012. Genome-Wide Survey of Pseudogenes in 80 Fully Re-sequenced Arabidopsis thaliana Accessions. *PloS one*, 7(12), p.e51769.

Weiss-Schneeweiss, H. et al., 2013. Evolutionary Consequences, Constraints and Potential of Polyploidy in Plants. *Cytogenetic and genome research*, 140(2–4), pp.137–150.

Wicker, T. et al., 2017. Transposable elements are a major evolutionary force shaping the 5,700 Mbp barley genome. *submitted*.

Wicker, T., Buchmann, J.P. & Keller, B., 2010. Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome research*, 20(9), pp.1229–1237.

Xiao, J. et al., 2016. Pseudogenes and Their Genome-Wide Prediction in Plants. *International Journal of Molecular Sciences*, 17(12), p.1991.

Yang, Z., 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8), pp.1586–1591.

Yu, J. et al., 2005. The Genomes of Oryza sativa: A History of Duplications. *PLoS biology*, 3(2), p.e38.

Zhang, Z.D. et al., 2010. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. *Genome Biology*, 11(3), p.R26.

Zou, C. et al., 2009. Evolutionary and Expression Signatures of Pseudogenes in Arabidopsis and Rice. *Plant Physiology*, 151(1), pp.3–15.

Table 1. Basic metrics for all pseudogenes and high-coverage (HC) pseudogenes found in the barley genome.

	pseudogene class	number	%	mean length (bp)	mean coverage (%)	mean identity (%)
all	all pseudogenes	89,440		188	33.5	91.4
	duplicated	12,556	14.0	329	40.7	93.8
	processed	1,834	2.1	238	29.3	91.4
	chimeric	571	0.6	423	35.5	93.4
	single exon parent	38,424	43.0	190	46.4	90.3
	fragmented	36,055	40.3	130	17.4	91.7

	pseudogene class	number	%	mean length (bp)	mean coverage (%)	mean identity (%)
HC	all pseudogenes	11,015		376	94.6	93.0
	duplicated	2,151	19.5	540	95.1	95.1
	processed	153	1.4	509	93.6	90.1
	chimeric	41	0.4	713	90.7	94.2
	single exon parent	8,224	74.7	329	94.8	92.5
	fragmented	446	4.1	378	89.8	93.3

Figure legends

Figure 1. Gene and pseudogene metrics. Number of genes and parent genes, as well as pseudogene classes for all and HC pseudogenes, respectively.

Figure 2. Distribution of transposable elements, genes and pseudogenes on the seven chromosomes of barley. The first (outer) track shows the seven barley chromosomes with positions in Mb and highlighted centromeres. The second to fourth track show densities of transposable element sequences (min. 47% to max. 85% sequence coverage), genes (min. 0% to max 5% sequence coverage) and pseudogenes (min. 0% to max. 2% sequence coverage), respectively. Densities have been calculated using a sliding window of 5 Mb shifted by 1 Mb. Links in the center connect parent genes with their pseudogene 'descendants' and are colored in the chromosome of the respective parent gene. Tandem duplicates can be easily recognized as straight lines, in particular at the chromosome ends.

Figure 3. Sequence coverage vs. identity of barley pseudogenes and their subclasses compared to their respective parent genes. **A** all pseudogenes; **B** duplicated pseudogenes; **C** processed pseudogenes; **D** pseudogenes from single-exon parent genes; **E** chimeric pseudogenes; **F** fragmented pseudogenes.

Figure 4. Distance distribution of pseudogenes to their respective parent genes.

Figure 5. Gene families and pseudogenes. **A** Relationship of gene family size to HC pseudogene number. The histogram depicts frequencies of gene family sizes with and without parent gene members (left axis). The dot plot shows the HC pseudogene content in 'extended' families, which are gene families combined with their HC pseudogenes (right axis). **B** Pseudogene content in 'extended' gene families, that are expanded, contracted or constant in barley compared to rice, sorghum, *Brachypodium distachyon* or *Arabidopsis thaliana*. Only orthologous groups with a minimal size of five were used for this analysis.

Figure 6. Over- and under-represented Gene Ontology (GO) terms for the parent gene set compared to the complete high-confidence gene set of the barley genome. The sub-ontology 'Biological Process' was used for this analysis.

Figure 7. Three syntenic regions containing pseudogenes in barley cv. Morex and four wild barley accessions. Chromosomal regions are displayed with gene coding sequences (green), pseudogenes with premature stop codon (red), and potential pseudogenes without premature stop codon (blue). Syntenic elements are connected. Stretches of Ns in the sequence are highlighted in orange; the annotation of repetitive elements on the barley cv. Morex chromosomes is highlighted in violet. **A** A tandem duplication in one of the EC accessions. Wild barley from the south-facing slope (SFS) experienced a tandem gene duplication event with a subsequent pseudogenization of redundant copies. A respective LRR gene in wild barley from the north-facing slope (NFS) contains a 13 base pair deletion in the first half of the coding sequence resulting in a frameshift and premature stop codons (lightning symbol). Both wild barley populations share another pseudogene with premature stop codons resulting from a frameshift. **B** The shifted reading frame of a CABP gene is restored in the SFS accession, but not the NFS accession. A 1 base pair deletion is present in both EC accessions, but only the frame of the gene from the SFS accession is restored due to another 1 base pair insertion. The shifted region is marked in orange and does not contain premature termination codons. **C** Transposable elements result in rearrangements and pseudogene creation in barley cv. Morex. The region in barley cv. Morex is greatly expanded (x12 scale difference) due to repetitive element insertion resulting in duplications and rearrangements. Copied gene fragments are degenerated. A copy of a LRR gene is shorted due to a frameshift. The pseudogene and gene connected with a dashed line have no sequence similarity, but were both detected through their homology with different isoforms of the same gene. Gene names are abbreviated: catalase (CAT), polyphenol oxidase, chloroplastic (PPO), leucin-rich repeat protein kinase (LRR), hexosyltransferase (HT), nucleolar MIF4G domain-containing protein (NOM), calcium-binding protein (CABP), pentatricopeptide repeat-containing protein (PPR), NAD(H) kinase (NADK).