Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt

Pseudogene Dynamics in Plants

Verena Marina Prade

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende: Prof. Dr. Brigitte Poppenberger-Sieberer

Prüfer der Dissertation:

1. Hon.-Prof. Dr. Klaus F.-X. Mayer
2. apl. Prof. Dr. Ramon A. Torres-Ruiz
3. Prof. Dr. Wilfried Schwab

Die Dissertation wurde am 07. Juni 2018 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 11. September 2018 angenommen.

Abstract

Pseudogenes are gene-like sequences that are often regarded as “evolutionary relics” or “junk DNA”. Gene duplication and the resulting redundancy of information can lead to reduced selection pressure. Redundant copies often pseudogenize and lose their original function. While dysfunctionality used to be one of their defining attributes, recent findings suggest that pseudogenes represent a repertory of potential genes with the capacity to shape an organism during evolution. Additionally, numerous alleged pseudogenes have been shown to adopt regulatory roles and their diagnostic or prognostic potential for human diseases has been recognized.

For many plants, the study of pseudogenes was hampered by their large and complex genomes. Economically important crops like bread wheat (*Triticum aestivum*), barley (*Hordeum vulgare*) or rye (*Secale cereale*) are grasses of the *Poaceae* family. Their large genomes are highly repetitive and often comprised of massive amounts of transposable elements (TEs). Hence, creating full-length chromosome assemblies represented a challenge due to collapsing repetitive regions. Recent methodological and technical advancements paved the way for complete and high-quality genome assemblies. Comprehensive pseudogene annotations became feasible.

In this work, pseudogenes were annotated for three dicotyledonous and 15 monocotyledonous plant species with varying genome sizes, complexities and assembly qualities. Model organisms, but also economically important plants like maize (2.7 Gbp), barley (5.4 Gbp) and bread wheat (16.9 Gbp) were annotated and analyzed. The modular Pseudogene Locus Identification Pipeline (PLIPipeline) manages large datasets by parallelizing calculations. Putative pseudogenes and gene fragments are identified via homology to functional genes. Plant genomes contain massive amounts of TE genes, pseudogenes and smaller gene fragments. Up to eight million putative pseudogenes were identified but most of them are TE-related sequences that were filtered from the final pseudogene set. Most non-TE-related pseudogenes are smaller gene fragments. Per plant, the number of full-length pseudogenes is always below the number of protein-coding genes. Retrotransposition only plays a marginal role in the generation of pseudogenes: While 70% of the human pseudogenes are retroposed, less than 1% of the plant pseudogenes have lost their introns — a feature usually indicative for retrotransposition. However, several other features of processed pseudogenes in plants suggest that they have lost their introns subsequent to duplication. Their non-random distribution resembles the distribution of non-processed pseudogenes. Non-processed pseudogenes are often generated via unequal crossing over or double strand DNA break repair mechanisms. Additionally, many may have been duplicated via incorporation into a TE. Pseudogenes with absent intron sequences exhibit homology beyond their untranslated regions and most are not associated to a poly-A tail. Hence, they are likely not the result of retrotransposition. Larger gene families give rise to disproportionately more pseudogenes and gene family expansion is accompanied by the generation of pseudogenes. In most plants, genes involved in defense-response, translation or photosynthesis frequently give rise to pseudogenes. Transcription evidence was found for 12% of barley's full-length pseudogenes.

This work represents a comprehensive study of plant pseudogenes and suggests an enormous functional potential of these so far under-researched genetic elements.

Zusammenfassung

Pseudogene sind genomische Sequenzen, die funktionellen Genen ähneln. Im Gegensatz zu proteinkodierenden Genen werden sie allerdings häufig als „Relikte der Evolution“ oder „Junk DNA“ (Erbgut-Müll) abgetan. Genduplikation und die daraus resultierende Redundanz der Sequenzinformation kann zu reduziertem Selektionsdruck führen. Die Genkopien können dann pseudogenisieren und ihre ursprüngliche Funktion verlieren. Obwohl Dysfunktionalität definitionsgemäß ein Merkmal von Pseudogenen ist, deuten jüngste Untersuchungen darauf hin, dass Pseudogene ein Repertoir von potentiellen Genen darstellen. Zahlreiche vermeintliche Pseudogene üben zudem regulatorische Funktionen aus und ihr diagnostisches und prognostisches Potential im Bezug auf Krankheiten wie Krebs ist bereits bekannt.

Die großen und komplexen Genome vieler Pflanzen erschweren die Annotation und Erforschung von Pseudogenen. Wirtschaftlich wichtige Kulturpflanzen wie Brotweizen (*Triticum aestivum*), Gerste (*Hordeum vulgare*) oder Roggen (*Secale cereale*) sind Gräser der *Poaceae* Pflanzenfamilie. Ihre großen Genome sind reich an repetitiven Elementen und bestehen zum Großteil aus Transposons. Repetitive Regionen stellen eine besondere Herausforderung für Genomassemblierungen dar, da sie oft nicht vollständig aufgeschlüsselt werden können. Jüngste methodische und technologische Fortschritte bereiteten aber den Weg für vollständige und hochwertige Genomassemblierungen und ermöglichten auch eine umfassende Erforschung von Pseudogenen.

In dieser Arbeit wurden Pseudogene in den Genomen von drei dikotylen und 15 monokotylen Pflanzen annotiert. Sowohl Modelorganismen, als auch wirtschaftlich bedeutende Kuturpflanzen wie Mais (2,7 Gbp), Gerste (5,4 Gbp) oder Brotweizen (16,9 Gbp) wurden untersucht. Die modular aufgebaute „Pseudogene Locus Identification Pipeline“ (PLIPipeline) identifiziert Pseudogene und Genfragmente über Homologie zu proteinkodierenden Genen. Bis zu acht Millionen potentielle Pseudogene wurden annotiert, von denen bis zu 95% als Transposon-Gene klassifiziert und aus dem Pseudogenset gefiltert wurden. Viele der übrigen Pseudogene sind kurze Genfragmente und die Anzahl der volllängen Pseudogene ist stets etwas kleiner als die Anzahl der proteinkodierenden Gene. Retrotransposition spielt nur eine untergeordnete Rolle in der Entstehung von Pseudogenen: Während 70% der Pseudogene im Menschen durch Retrotransposition entstanden sind, können nur 1% der pflanzlichen Pseudogene als prozessiert klassifiziert werden. Einige Eigenschaften von prozessierten Pseudogenen deuten allerdings darauf hin, dass sie ihre Introns erst nach der Duplikation verloren haben. Ihre Verteilung auf dem Genom ist nicht zufällig und ähnelt der Verteilung von nicht-prozessierten Pseudogenen. Diese sind oft das Resultat von ungleichem Crossing-over oder das Nebenprodukt von DNS Reparaturmechanismen. Prozessierte Pseudogene weisen zudem Homologie weit über deren untranslatierten Bereich hinaus auf und haben keinen poly-A-Schwanz. Deshalb sind viele dieser prozessierten Pseudogene höchstwahrscheinlich nicht das Ergebnis von Retrotransposition. Große Genfamilien haben überdurchschnittlich viele Pseudogene und die Expansion von Genfamilien führt zu Pseudogenisierung einiger Genkopien. Gene, die an Abwehrreaktionen, Translation oder Photosynthese beteiligt sind, bringen zudem häufiger Pseudogene hervor. In Gerste sind 12% der volllängen Pseudogene transkribiert.

Diese Arbeit ist eine umfassende Studie zu pflanzlichen Pseudogenen und deutet auf ein enormes funktionelles Potential dieser wenig erforschten genetischen Elemente hin. Sie trägt wesentlich dazu bei, die Genome von Pflanzen besser zu verstehen.

1 Scientific publications

The pseudogenes of barley
Verena M. Prade, Heidrun Gundlach, Sven Twardziok, Brett Chapman, Cong Tan, Peter Langridge, Alan H. Schulman, Nils Stein, Robbie Waugh, Guoping Zhang, Matthias Platzer, Chengdao Li, Manuel Spannagl and Klaus F. X. Mayer
The Plant Journal . 93 : (3) 502-514, 2018.

Shifting the limits in wheat research and breeding using a fully annotated reference genome
The International Wheat Genome Sequencing Consortium
Science . 361 : (6403) , 2018.

Durum wheat genome reveals past domestication signatures and future improvement targets
Marco Maccaferri, Neil S. Harris, Sven O. Twardziok, Heidrun Gundlach, Manuel Spannagl, Danara Ormanbekova, Thomas Lux, **Verena Prade**, Sara Milner, Axel Himmelbach, Martin Mascher, Paolo Bagnaresi, Primetta Faccioli, Paolo Cozzi, Massimiliano Lauria, Barbara Lazzari, Alessandra Stella, Andrea Manconi, Matteo Gnocchi, Raz Avni, Jasline Deek, Sezgi Biyiklioglu, Elisabetta Frascaroli, Simona Corneti, Silvio Salvi, Roberto Tuberosa, Gabriella Sonnante, Francesca Desiderio, Caterina Marè, Cristina Crosatti, Erica Mica, Hakan Ozkan, Pasquale De Vita, Daniela Marone, Reem Joukhadar, Raj K. Pasam, Elisabetta Mazzucotelli, Domenica Nigro, Agata Gadaleta, Shiaoman Chao, Justin Faris, Arthur T. O. Melo, Mike Pumphrey, Nicola Pecchioni, Luciano Milanesi, Krysta Wiebe, Ron P. MacLachlan, John M. Clarke, Andrew G. Sharpe, Kevin Koh, Kevin Y. H. Liang, Gregory J. Taylor, Ron Knox, Hikmet Budak, Anna M. Mastrangelo, Steven S. Xu, Nils Stein, Iago Hale, Assaf Distelfeld, Matthew J. Hayden, Sean Walkowiak, Klaus F. X. Mayer, Aldo Ceriotti, Curtis J. Pozniak and Luigi Cattivelli
Genome Biology (under revision) 2018.

Acknowledgements

This work would not have been possible without the support of my supervisors, colleagues, family and friends. Therefore, I would like to dedicate a few lines to express my sincere gratitude to all of them.

First of all, I would like to thank my advisor Prof. Dr. Klaus F. X. Mayer for supporting my doctoral research study. His continuous support, motivation, patience and guidance helped me tremendously during my research and writing of the thesis. He gave me the opportunity to present my work at an international conference in San Diego and I feel very fortunate for having him as a supervisor of my thesis.

Furthermore, I would like to thank the additional members of my Thesis Committee, Prof. Dr. Ramon Angel Torres-Ruiz and Dr. Heidrun Gundlach, for helpful discussions, insightful comments and encouragement. For temporarily assuming the role of advisor, I thank Prof. Dr. Jörg Durner.

Special thanks go to the entire PGSB group for their unfailing support, fruitful discussions and for their friendship. I enjoyed our daily lunch breaks and our fun activities like barbecuing, canoing, bowling or Bavarian curling. I especially want to thank Heidrun, with whom I shared an office for the entire time of my thesis. Her knowledge and experience helped me a lot. Manuel, especially for taking the lead as deputy PGSB group leader and for always being open for discussions. Jimmy, for his advice with statistical analyses. And special thanks to Michael, for taking care of administrative issues like our computing cluster or the file system. He worked wonders in upgrading our computing power or managing severe technical difficulties. At times, much could have been lost without him. Thank you!

Last but not least, I want to give my warmest thanks to my family and friends. I especially want to thank my parents for their unfailing support. Nicolas, thank you for supporting and encouraging me. Thank you for being there for me.

Contents

2 Introduction

2.1 The green branch of life : plants

The beginning of plant and animal domestication marked a major turning point in the environmental and cultural history of mankind. The birth of agriculture not only introduced the ability to manipulate other organisms, but also led to subsequent technological and cultural changes. Food production via agriculture facilitated a sedentary life style and changed the organization of human communities. While the expansion and advancement of agriculture enabled populations to grow, it also led to massive interference with the environment, for example through the removal of natural vegetation, water redirection or water pollution due to over-fertilization. (Mannion, 1999)

> Germany comprises an area of 35.7 million hectares and almost half is used for farming. Over two thirds of the arable land is used to grow cereals, with a yield of 48.9 million metric tons in 2015. As Europe's leading producer of milk or pork, more than 60 percent of Germany's agricultural farmland is necessary to grow animal fodder. (BMEL, 2016)

With a constantly growing world population, nations worldwide have to invest much effort into the production of food and animal fodder (Figure I1). Since arable land becomes scarce and land clearing is not always an option, yield has to increase. In 1950 Germany, one hectare of farmland produced 2.6 metric tons of wheat. By 2015, yield had increased to 8.1 tonnes per hectare (t/ha) — an increase of 312 percent (BMEL, 2016).

This "green revolution" was made possible by improved high-yielding crop varieties and improved agronomic practices like the application of the Haber-Bosch process for nitrogen fertilizer production. The term "green revolution" was coined in the late 1960s when cereal yields rapidly increased in South Asia (McArthur and McCord, 2017). Research of the "green branch of life" contributes significantly to the green revolution and the fight against undernourishment and food supply shortages by (i) increasing yield and grain quality; (ii) identifying or breeding varieties able to adapt to diverse environments; (iii) shortening growth cycles and (iv) selecting for resistances against biotic or abiotic stresses (Khush, 2001). Ongoing population growth, changing climates and environmental pollution represent current challenges for science.

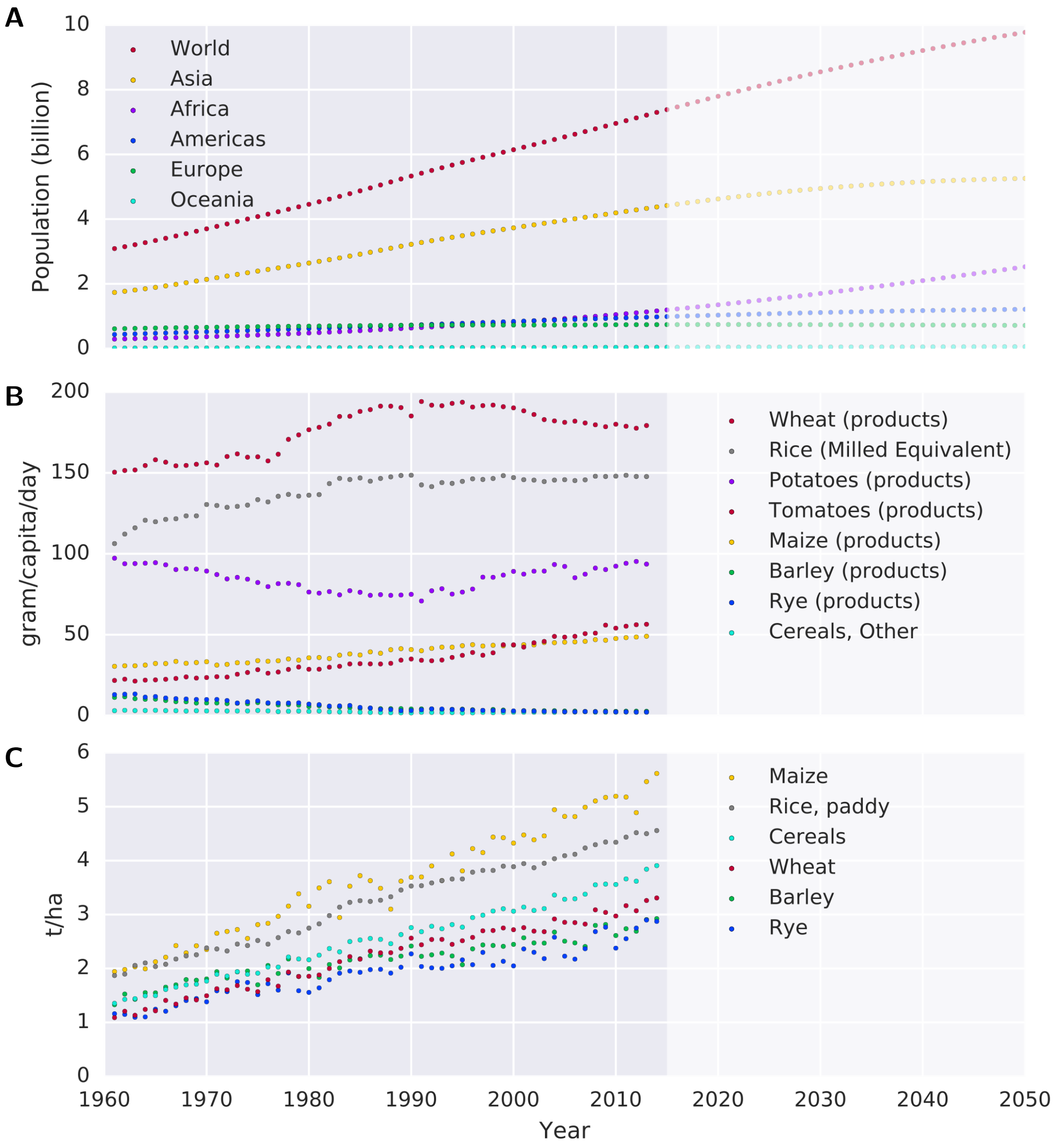


Figure I1: Population growth, per capita consumption and yield increase. A: Worldwide and regional population growth. Data past the year 2014 are estimates. B: Average daily consumption of crop products per capita per day. C: Worldwide yield increase. Potato yields (2014: 20 t/ha) and tomato yields (2015: 35 t/ha) are not shown. The figures were created using data from the Food and Agriculture Organization of the United Nations (2017).

2.2 The race between population growth and yield increase

To meet the demands of a more populated and prosperous world, the average worldwide yield has to be doubled by 2050 (Ray et al., 2013). Today, the yield of the top four crops — maize, rice, wheat and soybean — is increasing by 0.9 to 1.6 percent per year. However, approximately 2.4 percent are necessary to meet the projected demand. While Germany managed to increase wheat yield to over 8 t/ha, the worldwide average is lagging behind with less than 4 t/ha (Figure I1 C).

The Food and Agriculture Organization of the United Nations (FAO) urges all countries and stakeholders to work towards an end of hunger and malnutrition by 2030. In fact, for a decade, the number of undernourished people *was* decreasing. However, this worldwide success story came to an end in 2016, when the number of undernourished people gained 35 million compared to the previous year. In 2016, a total of 815 million people were undernourished, representing eleven percent of the world population or corresponding to almost ten times the number of people currently living in Germany. The situation has particularly worsened in Africa and Asia due to conflicts combined with droughts or floods. Most of the approaches to solve the problems target conflict resolution, but they also address the introduction of resistant crops and livestock. (FAO et al., 2017)

The "food gap" problem needs to be fought on multiple fronts. Plant research has done a great job in the 20th century and accomplished significant breakthroughs. Nevertheless, much more effort is needed to tackle food shortages, fight undernourishment and cover the demands of a growing and prospering world population, while minimizing environmental pollution. New genome research technologies are seen as important contributors for a desperately needed 2nd green revolution.

2.3 Important crop plants from the past to the present

2.3.1 The birth of agriculture 10,000 years ago

Approximately ten thousand years ago during the Neolithic, cereals like wheat or barley were domesticated in the Fertile Crescent, the birth place of western agriculture. The Fertile Crescent ranges from Egypt over the Levant to the northeast — including Israel, Lebanon, Jordan, and Syria — and from there over Mesopotamia to the Persian Golf in the southeast (Figure I2). Considering a more moderate climate at that time, the region was fertile and humid enough to promote the beginning of agriculture, animal husbandry and a sedentary lifestyle. (Haberer et al., 2016; Preece et al., 2017)

Domestication has a strong effect on the genotype and phenotype of plants: Compared to wild progenitors, domesticated grain crops often have a higher yield, greater final plant size, greater seed mass and reduced chaff material (Preece et al.,

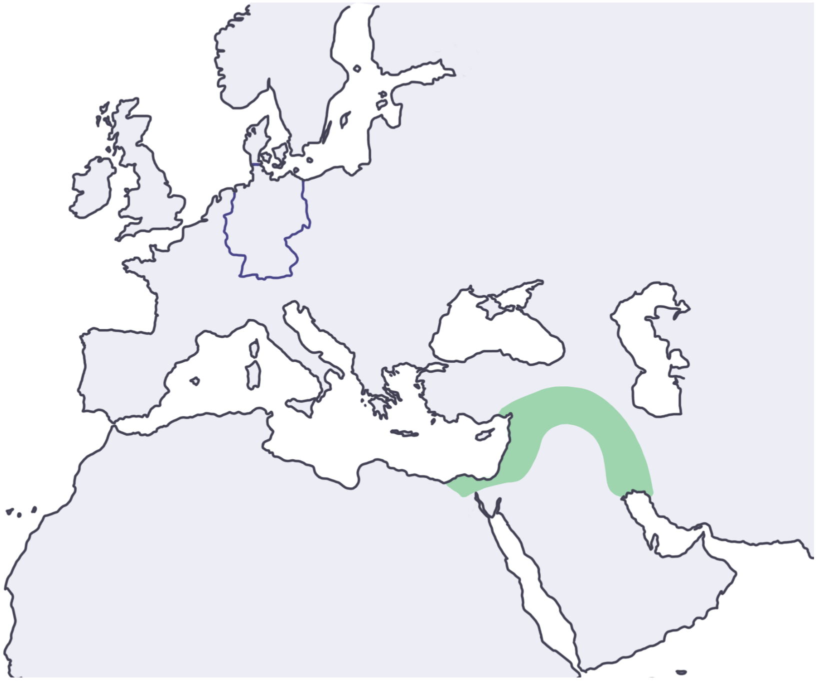


Figure I2: The Fertile Crescent (green).

2017). Nonetheless, one of the first plant phenotypes selected by early farmers was likely a dysfunction in the seed dispersal system (Pourkheirandish, Hensel, et al., 2015). In wild cereals, awns containing ripe grains fall to the ground as the rachis becomes brittle. Collecting grains from the soil is tedious and ineffective. Only two mutations in adjacent, dominant and complementary genes are necessary for the rachis to turn into a non-brittle form (Pourkheirandish, Hensel, et al., 2015). This allowed for effective harvesting by farmers, but also made domesticated plants completely dependent on humans for their dispersal (Preece et al., 2017).

2.3.2 Taxonomy, phylogeny and genomics of important crop plants

Maize, wheat, rice, sorghum and barley are economically important crops providing the major source for caloric intake by humans (Kellogg, 2001). All of them are monocotyledonous species. Monocotyledons (monocots) are a monophyletic clade of flowering plants, mainly characterized by an embryo with a single cotyledon (seed leaf). A large family of monocots are the *Poaceae* (grasses). They comprise many cultivated or edible cereals like wheat, rye or barley and are also an important source for animal fodder, especially in grassland husbandry (BMEL, 2016).

Grasses can have massive differences in genome size with at least 35-fold variation (Davidson et al., 2012). For example, *Brachypodium distachyon* is a wild grass with a smaller genome of 355 megabase pairs (Mbp), while bread wheat has a genome approximately five times larger than the human genome (viz. ∼16.9 gigabase pairs (Gbp)).[1] The main reasons for these massive differences in genome size are polyploidy, transposable element (TE) content and genome repetitivity.

Polyploidy is the result of Whole Genome Duplication (WGD) and describes the presence of more than two homologous sets of chromosomes in a cell. These surplus chromosome sets can originate from the same taxon (autopolyploidy) or from different taxa (allopolyploidy) (Weiss-Schneeweiss et al., 2013).

While polyploidy is common in flowering plants, it is rare in mammals. In human, chromosome number aberrations are often lethal or lead to developmental defects. Thus, true polyploidy is very rare in humans, but can occur in tumor cells. Additionally, specific cell types like liver or bone marrow cells can undergo programmed polyploidization. However, the benefits of this process are still not fully understood. (Van de Peer et al., 2017)

Plants seem to have a higher tolerance for chromosome number changes, because of a fundamentally different mechanism for dosage compensation. Whole-genome duplications, chromosome duplications, segmental duplications and chromosome rearrangements are common in plants. For some plants, the ability for selfing might help them to overcome reproductive isolation and to pass on polyploidy to the next generation. Genomic changes and increased genetic variation can affect the interaction with pollinators or herbivores, which can in the end lead to speciation. It can also help the plant to survive under diverse environmental conditions — a controversial theory, that is supported by the increased frequency of polyploid species living in harsh environments. (Van de Peer et al., 2017)

2.3.3 Tackling large *Triticeae* genomes

The *Triticeae* tribe within the *Poaceae* family comprises agriculturally important genera like wheat, barley and rye. The diploid genome of barley has a size of 5.1 Gbp — which is ∼2 Gbp larger than the human genome (International Barley Sequencing Consortium, 2017). Bread wheat is allohexaploid, which is a specific form of polyploidy, where three homeologous sets of non-recombining chromosomes are present in a cell. The bread wheat genome comprises 42 chromosomes in total, adding up to a genome size of ∼16.9 Gbp. Massive genomes, high repetitivity and high complexity are common in *Triticeae* species.

2.3.3.1 Genome sequencing and assembly Detailed genome analyses can help to gain an understanding of crop varieties that are able to survive under extreme climatic conditions, are resistant to diseases or pests and still produce a

[1] http://data.kew.org/cvalues/CvalServlet?querytype=1, (April 2, 2018, 12:21pm CEST)

high yield. A first step in understanding the genome of a species is DNA sequencing to identify the exact order of nucleotides (nts) in a chromosome or DNA molecule. However, the analysis of many crop plants was hampered by their unusually large and complex genomes, as they are often comprised of massive amounts of repetitive elements. For example, over 80% of most *Triticeae* genomes is TE related (e.g. International Barley Sequencing Consortium, 2017; International Wheat Genome Sequencing Consortium, 2018). Genome assembly algorithms often cannot handle repetitive regions when only supplied with short sequencing reads. Repetitive regions collapse and fragmented assemblies are created. Paired-end or mate-pair sequencing technologies can help bridge smaller regions of repetitivity, but this is still not sufficient to tackle most large and complex plant genomes, because of their larger repetitive regions. Additionally, the repeat landscape of plant genomes differs significantly from that of most vertebrates: The most abundant type of TE in plants are long terminal repeat (LTR)-retrotransposons — compared to non-LTR-retrotransposons in animals. Plant TEs are much larger and younger than animal TEs, further hampering genome assembly attempts (Murat et al., 2012). While TEs and tandemly repeated sequences lead to fragmented and collapsed genome assemblies, the genic regions are generally assembled to a higher quality and synteny information can be used to order contig and scaffold sequences.

The International Wheat Genome Sequencing Consortium (IWGSC) was founded in 2005 by wheat growers, plant scientists and breeders to help develop improved wheat varieties. At the time — due to technological and economic limitations — the sequencing of the hexaploid bread wheat genome was approached using a divide-and-conquer method (International Wheat Genome Sequencing Consortium, 2014): Isolated chromosome arms were sequenced and assembled individually. This approach significantly reduced complexity, as each chromosome arm represents only 1.3 to 3.3 percent of the genome. Chromosome arms were derived from double ditelosomic stocks. Sorting was achieved by staining and suspending them in a fluid flowing as a narrow stream through a fluorescence detector. If the flow karyotype indicates a difference in relative fluorescence intensity, an electrical charge is applied to the broken up liquid stream and deflection plates can be used to redirect chromosome-containing droplets into collection containers (Doležel et al., 2012). Albeit individual chromosome arms have finally been sequenced by a BAC-by-BAC sequencing approach, NRGene — an Israeli company — recently rose to success due to revolutionary achievements in the assembly of large and complex genomes. The IWGSC adopted the NRGene based assembly technology and published a first complete reference sequence of the bread wheat genome in 2018 (International Wheat Genome Sequencing Consortium, 2018). This breakthrough was accomplished by ordering high-quality

NRGene scaffolds using chromosome conformation capture sequencing (Hi-C) technology.

While the NRGene assembly technology uses smaller Next-Generation Sequencing (NGS) reads — albeit from sequencing libraries with varying sequencing length — other technologies can help to overcome the problem of genome repetitivity by producing long reads: Pacific Biosciences (PacBio) and Oxford Nanopore use a single-molecule sequencing approach to produce reads of up to 50 and 100 kilobase pairs (kbp), respectively. One of the biggest challenges in using long read sequencers is their relatively high error rate of 15% (Zimin et al., 2017). However, by combining long reads with smaller NGS reads, high quality genome sequences can be generated, as demonstrated for the *Aegilops tauschii* genome (Zimin et al., 2017).

Also in 2017, a chromosome-scale assembly of the barley genome has been published (International Barley Sequencing Consortium, 2017): Bacterial Artificial Clone (BAC) sequencing using Illumina paired-end and mate-pair technology, physical map information, genetic linkage and a highly contiguous optical map were combined to construct super-scaffolds in a hierarchical approach. After assigning scaffolds to chromosomes, Hi-C was used to obtain three-dimensional proximity information and to order and orient the BAC-based super-scaffolds.

2.3.3.2 Gene annotation When genome sequences are assembled, classical genome analysis can be applied to predict genes and TEs. Protein-coding genes are usually predicted computationally using homology to known genes. Additionally, evidence-based gene prediction utilizes transcriptional evidence like RNAseq data, expressed sequence tags, sequenced proteins or full-length complementary DNAs (Liang et al., 2009). Tools like the TransDecoder[2] can be used on transcript sequences to predict open reading frames and alternate isoforms. Functional annotations (e.g. GO, Pfam, InterPro) can also help to classify and filter TE-related genes. Gene annotations can then be categorized according to several of their attributes: presence of start and stop codons, homology, minimal length of the open reading frame or TE overlap. Finally, the gene annotation is often evaluated with the BUSCO pipeline — a tool to assess genome assembly and annotation completeness via a set of near-universal single-copy orthologs.[3]

2.3.3.3 Transposon annotation and repeat analysis TEs can be detected and classified via homology against transposon libraries like PGSB-REdat (Spannagl et al., 2016) or ClariTeRep[4]. Several software tools are available

[2] `https://transdecoder.github.io`, (May 14, 2018, 10:25am CEST)

[3] `busco.ezlab.org`, (January 15, 2018, 6pm CEST)

[4] CLARITE `https://github.com/jdaron/CLARI-TE/`, (April 2, 2018, 6pm CEST)

to identify and classify specific transposon classes. In plants, the most prevalent group of transposable elements constitute LTR-retrotransposons (class I) (Wicker, Gundlach, et al., 2018). LTRharvest is a tool to identify such LTR-retrotransposons. There is a linear correlation between the number of full-length LTR-retrotransposons and genome size in plants, permitting an assessment of genome assembly quality via LTR-retrotransposons (Wicker, Gundlach, et al., 2018).

2.3.3.4 Further methods in classical genome analysis The standard repertory of classical genome analysis does not only comprise the annotation of protein-coding genes and TEs, but also the analysis of gene families, simple sequence repeats, microRNA (miRNA), transfer RNA (tRNA) or other non-coding RNAs (ncRNAs). Gene families are usually determined via sequence clustering, such that genes with similar protein sequence form a family. The repetitivity of a genome can be assessed via k-mer analyses — the counting of sequence substrings of length k. Such an analysis can also help to compare genomes without computationally expensive genome alignments.

NcRNAs are often identified via homology. For example, miRNAs — RNA molecules involved in post-transcriptional regulation — can be retrieved from miRBase (Kozomara and Griffiths-Jones, 2014). Software like tRNAscan-SE (Lowe and Eddy, 1997) can be used for the prediction and analysis of the secondary structure of tRNA — the molecule that transports amino acids to the messenger RNA (mRNA) during translation.

In recent years, new technologies and methods opened up exciting possibilities. Due to fast and affordable new assembly methods, computational plant genome analysis is no longer limited to genes and sequences of low repetitivity, but can also be used for the study of full-length TEs or the genomic “gray area”. This unannotated gray area contains regulatory elements, TEs degraded to the point where they are no longer discernible, but also other elements previously dismissed as “junk DNA”.

2.4 Pseudogenes

Pseudogenes are often regarded as “evolutionary relics” or “junk DNA”. They are gene-like sequences that have degenerated and lost their original function. Gene duplication and the resulting redundancy of information most often leads to a reduced conservation pressure on one or both of the gene copies (Balakirev and Ayala, 2003). Hence, mutations are accumulated until one copy can no longer be used as a template for a functional protein. Such mutations can for example directly affect the active site of the protein, introduce frameshifts and premature termination codons (PTCs) or affect regulatory regions. Pseudogenes still have characteristics of functional genes, which can make it difficult to distinguish them from genes. It is relatively easy to prove functionality (e.g. with knock-out experiments), but almost impossible to prove the non-functionality of pseudogenes (Zheng and M. B. Gerstein, 2007). Also, the constant flux between gene and pseudogene state can complicate their annotation. These fundamental intricacies lead to numerous and changing definitions and annotations for pseudogenes. While they are non-functional according to their original definition, reports of functional pseudogenes inspired more general interpretations.

2.4.1 The ambiguous definition of pseudogenes

The first pseudogene was identified by Jacq et al. (1977), who reported a tandemly repeated unit of 700 base pairs (bp) in the genome of the frog *Xenopus laevis*. The region contains a 5S-rRNA gene, as well as a pseudogene. The pseudogene “was nearly as long as, and almost an exact repeat of, the gene itself”(Jacq et al., 1977). Since they could not detect RNA products corresponding to the pseudogene, the authors arrived at what became the first definition of a pseudogene:

> [...] it is a relic of evolution. During the evolution [...], a gene duplication occurred producing the pseudogene. Presumably the pseudogene initially functioned [...], but then, by mutation, diverged sufficiently from the gene in its sequence so that it was no longer transcribed into an RNA product.
>
> — Jacq et al. (1977)

This initial discovery was quickly followed by additional findings. In 1980, a processed pseudogene was found in two mice strains (Nishioka et al., 1980; Vanin et al., 1980). The α-3 gene — a deficient homologue of the α-globin gene — comprises not only base changes, but also deletions and insertions causing frameshifts and PTCs. Nishioka et al. (1980) conclude, that the most likely way for the gene to lose all its intervening sequences, would be a recent gene conversion event involving the mediation of mature globin mRNA or its complementary DNA (cDNA) cognate.

Vanin et al. (1980) propose a mechanism, by which the DNA is directly edited "in a fashion comparable to that used at the RNA level". They both ruled out retrotransposition — the reincorporation of reversely transcribed mRNA of the α-globin gene into the genome — since the sequence homology does extend beyond the 5'-cap site of the mRNA.

Intriguingly, Vanin et al. already proposed putative functions of pseudogenes in 1980. The authors speculated that pseudogenes might be "diverting genes", whose transcripts interfere with the transcription of productive genes. Additionally, they offer the possibility, that pseudogenes are "antigenes", and that they are transcribed complementary to the productive gene. The RNA products of gene and pseudogene could then form heteroduplexes. For example, human *PTEN* or *OCT4* genes have been shown to be regulated by anti-sense transcripts of their pseudogenes (Johnsson et al., 2014).

Pseudogenes have been identified in multiple species and various gene families and they are now increasingly being looked at as elements providing a repertory of potential genes, with the capacity to shape an organism during evolution (Brosius and Gould, 1992). The original definition of pseudogenes being "junk DNA" is challenged. The gene-look-alikes particularly attracted attention, since there have been reported cases of "functional pseudogenes" (Balakirev and Ayala, 2003; Poliseno, Salmena, et al., 2010).

> As their dysfunctionality used to be one of their defining properties, the discovery of their potential functionality led to disagreements and inconsistent annotations. The oxymoron "functional pseudogenes" has become a widely used term for transcribed pseudogenes, because transcripts can have regulatory properties, even when a premature termination codon prevents the synthesis of a functioning protein.

In mammals, pseudogenes have been linked to diseases like cancer and they can now be used as diagnostic or prognostic markers or to determine cell identity (Poliseno, Marranci, et al., 2015). Compared to healthy cells, some cancer cells contain pseudogenes that are differentially transcribed or even translated. A proteomic study in humans revealed 140 pseudogenes to be translated into peptides (Kim et al., 2014). They may represent spurious translations without function. However, it is noteworthy that RNA-mediated gene duplicates are "generally considered processed pseudogenes" in contrast to potentially functional retrogenes, because they lack their original promoter (Xu and J. Zhang, 2015). The difficulty in using pseudogenes as markers is their similarity to functional, paralogous genes (parent genes) from which they originally derived. Their sequence similarity can be very high and array-based methods to study the transcription of pseudogenes are not suitable to distinguish them. Instead, RNAseq analysis proves to be the best approach to determine the transcriptional activity of pseudogenes (Poliseno, Marranci, et al., 2015).

The study of pseudogenes in plants fell behind, since the limitations in genome access have been a road blocker. Genomes of many economically important crop plants are large and often highly repetitive. The lack of complete chromosome assemblies led to incomplete gene annotations and limitations in generating whole genome pseudogene annotations. *Triticeae* species like wheat, rye or barley are economically important crops, and due to their genome size and complexity one of the big challenges in plant research. At some point in their evolutionary history, almost all flowering plants experienced a genome duplication event with segmental duplications still traceable in their chromosomes (Tang et al., 2010). Such regions contain duplicated genes, that are especially prone to pseudogenization due to the resulting redundancy of information.

So far, only few plant species were subject to detailed pseudogene studies. The model organism *Arabidopsis thaliana* has been extensively studied. Its sequence and annotation data is manually curated in The Arabidopsis Information Resource (TAIR) (Berardini et al., 2015). In the database versions prior to TAIR8, TE-related genes and pseudogenes had both been categorized as pseudogenes. TAIR10 now contains 3,903 TE genes and 924 pseudogenes.[5] While TAIR10 has been used in this work, the newest data resource for *Arabidopsis thaliana* is Araport11. However, the pseudogene annotation did not change much and only increased to 952 elements (Cheng et al., 2017).

TAIR8 and TAIR10 demonstrate the challenge of comparative studies, since pseudogenes can be a matter of interpretation and detection method. In rice (*Oryza sativa* subsp. *japonica*), 11,956 non TE-related pseudogenes were identified with PseudoPipe (Guo et al., 2009). However, "PseudoPipe is a homology-based computational pipeline that searches a mammalian genome and identifies pseudogene sequences."[6] It can be used on genomes of other organisms as well, but some parameters — like the average and maximum intron length of genes — differ between plants and primates, leading to unexpected and potentially false results.

2.4.2 DNA duplication as a source of pseudogenes

Pseudogenization — gene degeneration and loss of function — is a process mostly affecting non-unique sequences: "The abundance of pseudogenes generally depends on rates of gene duplication and loss" (Tutar, 2012). The degeneration of a unique single-copy gene is believed to be rare, because the loss would likely affect the

[5] The Arabidopsis Information Resource (TAIR), `https://www.arabidopsis.org/portals/genAnnotation/gene_structural_annotation/annotation_data.jsp`, on `www.arabidopsis.org`, (January 3, 2018, 9:10am CET)

[6] `http://pseudogene.org/pseudopipe/`, (January 3, 2018, 10am CET)

phenotype of the organism. Only 76 such unitary pseudogenes have been identified in human by comparison to other primates (Z. D. Zhang et al., 2010).

Thus, the most common hypothesis for the origin of pseudogenes is gene duplication. Duplicated genes encode redundant information and the subsequent loss of one copy does not have an effect compared to the state prior to gene duplication. If the higher amount of transcript does not have a beneficial effect, then the conservation pressure is usually reduced on one of the gene copies.

> Maintaining proper protein and transcriptional balance is vital to sustain normal function. For instance, an imbalance in a highly connected portion of a network likely would result in great negative pleiotropic effects. [...] dosage-sensitive genes must be retained in duplicate following WGD to maintain proper balance of protein and transcriptional networks. However following a smaller scale duplication (e.g. local and tandem duplicates, segmental duplicates, aneuploidy) (i.e. overexpression), duplicates of dosage-sensitive genes will tend to be eliminated to maintain proper balances.
>
> — Edger and Pires (2009)

Plants are more tolerant to duplications due to a genome scale mechanism to harmonize gene expression (Levy and Feldman, 2002). Nevertheless, gradually accumulated mutations, insertions or deletions will in some cases lead to the pseudogenization of one gene copy — an event usually occurring in the first few million years after the loss of selection pressure (Lynch and Conery, 2000). Alternatively, both genes can undergo sub-functionalization or neo-functionalization.

There are various mechanisms leading to small- or large-scale sequence duplication and potentially to the pseudogenization of gene copies. Apart from those duplication mechanisms, the generation of unitary pseudogenes will be explained in the subsequent paragraphs:

- Ectopic recombination / unequal crossing over
- Segmental chromosome duplication
- Chromosome duplication (Aneuploidy)
- Whole genome duplication (Polyploidy)
- DNA double strand break repair mechanisms
- Retrotransposition of host mRNAs
- Transposons carrying host gene fragments
- Unitary pseudogenes

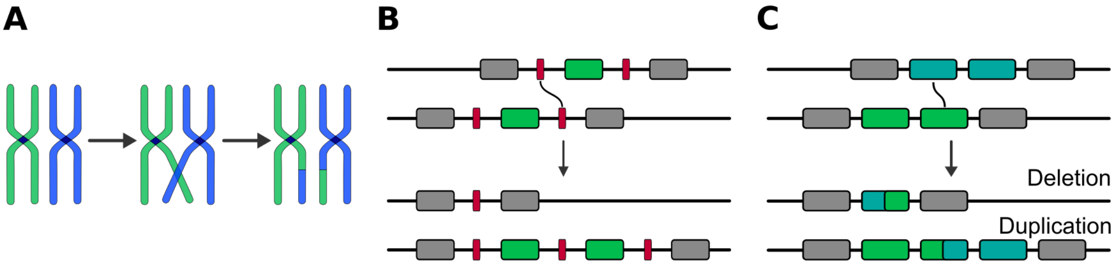


Figure I3: Unequal crossing over of chromosomes and the generation of tandem duplications. **A:** Chromosomal crossing of paired homologous chromosomes during meiosis. **B:** The initial unequal crossing over event requires a region of similar sequence (red). Crossover leads to a deletion in one and a duplication in the other chromosome. **C:** Tandem (gene) clusters are prone to unequal crossing over. Crossover can also occur within a gene, leading to chimeric genes.

2.4.2.1 Ectopic recombination Tandem duplications usually arise through unequal crossing over of chromosomes. During meiosis, homologous chromosomes are paired and cross-overs lead to an exchange of homologous sequences (Figure I3). If the pairing is not precise, part of the sequence is deleted in one chromosome, but added in the other. The latter now contains a tandemly duplicated region, which can contain one or more genes. Analogous, unequal crossing can also occur between sister chromatids during mitosis. (Z. Zhang, Paul M. Harrison, et al., 2003)

2.4.2.2 Segmental chromosome duplication Large scale segmental duplications can be relics of polyploidization, translocations and chromosome rearrangements. While sequence homogenization strongly acts on localized duplicated gene clusters, a duplication and translocation to a physically unlinked chromosome or chromosomal region might allow for the evolution of new functions (Baumgarten et al., 2003).

> [...] mechanisms of tandem and segmental duplication, in combination with recombinational isolation, are supposed to function, in general, in the diversification of gene families.
>
> — Leister (2004)

In Arabidopsis, a series of different gene families have such genomic organisation: LBS-LRR genes (disease resistance), cytochrome *P450* genes, UDPG-glycosyltransferases, receptor-like kinases and mammalian fibroblast growth factor genes. (Leister, 2004)

2.4.2.3 Chromosome duplication Nondisjunction during meiosis leads to chromosome duplications and is a potential source for duplicated pseudogenes (Podlaha and J. Zhang, 2010). Aneuploidy — the gain or loss of entire chromosomes — often has harmful consequences for eukaryotic systems, because the cell-wide ratio of gene products is massively imbalanced (Siegel and Amon, 2012). However, plants are more tolerant to such changes and able to harmonize gene expression "on a genomic scale by global regulation mechanisms leading to silencing, dosage compensation, or, on the other hand, to gene activation." (Levy and Feldman, 2002)

2.4.2.4 Whole Genome Duplication Polyploidy is very common in flowering plants and describes the presence of more than two homologous sets of chromosomes in a cell (Van de Peer et al., 2017; Podlaha and J. Zhang, 2010). Surplus chromosome sets can originate from the same species (autopolyploidy) or from different species (allopolyploidy) (Weiss-Schneeweiss et al., 2013). While aneuploidy (see paragraph above) can lead to an imbalance of gene products in most vertebrates, this is initially not the case in polyploid organisms. The maintained dosage of gene products in young polyploids might explain why redundant chromosomes are not quickly lost after Whole Genome Duplication (WGD): Anything but the simultaneous loss of all surplus chromosomes would result in a massive imbalance of gene products. Nonetheless, at least two complete and mostly redundant gene sets are combined, leading to an increased number of pseudogenization events.

2.4.2.5 Double strand break repair mechanisms DNA break repair mechanisms are important for all organisms, but particularly for plants. Most plants are sessile, autotroph and directly exposed to biotic and abiotic stresses, without the ability to actively avoid them (Schiml et al., 2016). Double strand breaks can either be repaired via homologous recombination or — more common in plants — by illegitimate recombination via non-homologous DNA end joining (NHEJ) (Figure I4 A). Up to 1.2 kbp long "filler DNA" fragments are copied during this double-strand DNA break repair process (Gorbunova and Levy, 1997). The mechanism resulting in filler DNA formation is also frequently called synthesis-dependent strand annealing (SDSA) (Wicker, Buchmann, et al., 2010).

Alternatively, some repair mechanisms can even result in the formation of tandem duplicated DNA sequences (Figure I4 B and C): The repair of two adjacent single strand breaks has been shown to be accompanied with deletions and tandem insertions of up to 100 bp. If adjacent microhomologies are present, protruding 5' strands can hybridize and DNA synthesis at the 3' strands results in tandem duplications. On the other hand, if no microhomologies are present, DNA synthesis at 3' ends results in a regular double strand break, that can be repaired via the

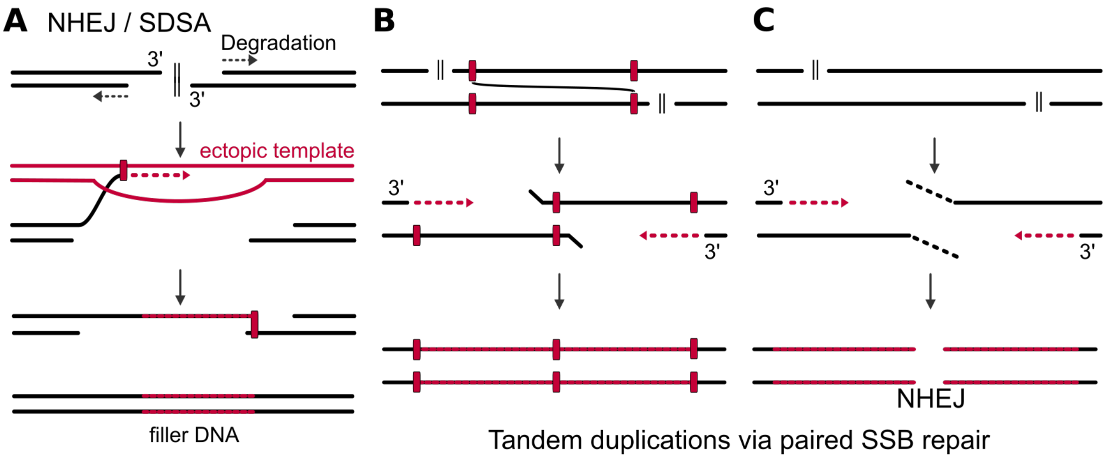


Figure I4: DNA break repair mechanisms leading to DNA duplication. A: Synthesis-dependent strand annealing (SDSA): The 5' ends at the DNA break are degraded. A short region of homology (red) between the 3' end and an ectopic template site is sufficient for invasion and DNA synthesis. The free single strands can re-anneal at a region of microhomology, followed DNA synthesis and nick ligation. Duplicated filler DNA can be incorporated from one or more ectopic template sites. **B and C:** Repair of two neighboring single strand breaks (SSBs) leading to tandem duplications: **B:** Microhomologies (red) between protruding single strands can hybridize. DNA synthesis and nick ligation complete the repair process. **C:** Without present microhomologies, the 3' strands are elongated via DNA synthesis. The double strand break can then be repaired via the NHEJ mechanism. The figure is derived from Gorbunova and Levy, 1997 (A) and Schiml et al., 2016 (B and C).

NHEJ/SDSA mechanism. The latter mechanism only results in tandem duplications, if 3' DNA synthesis occurs faster than the degradation of the 5' ends. This DNA repair mechanism is of particular interest, because unequal crossing over cannot account for all tandem gene clusters as it also requires initial microhomologies, that cannot always be found. (Schiml et al., 2016)

2.4.2.6 Retrotransposition of host mRNA The duplication mechanisms described above are sources for so-called duplicated or non-processed pseudogenes, which retain their exon-intron structure. On the other hand, retrotransposition can generate retroposed or processed pseudogenes that have lost their intron sequences due to an intermediate mRNA step (Figure I5). The functional parent gene is transcribed and the mRNA is spliced, resulting in a loss of introns in the mature mRNA. After reverse transcription of the mRNA, the resulting cDNA is reintegrated into the genome at a random location. A prominent feature to

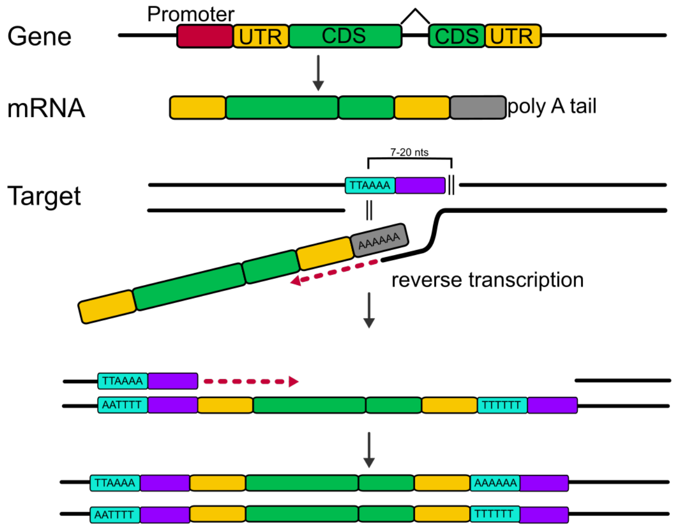


Figure I5: Retrotransposition of genes. A gene is transcribed into mature mRNA resulting in the loss of intron sequences and a poly-A tail at the 3' end. The endonuclease domain of an L1 element creates the first nick at the insertion site (TTAAA sequence), where the mRNA is primed for reverse transcription by the L1 reverse transcriptase domain. A second nick is generated in the other strand and the complementary DNA is synthesized. The flanking region (violet) is duplicated. The figure is derived from Kaessmann et al., 2009.

distinguish retroposed from duplicated pseudogenes is the absence of introns, a poly-adenine (poly-A) tail near the 3' end or direct small repeats. Processed pseudogenes are often called "dead-on-arrival", since they usually do not have an upstream promoter or regulatory sequences associated at the integration site (Sen and Ghosh, 2013). Thus, they are thought to accumulate mutations immediately and diverge faster from their parent genes, than non-processed pseudogenes.

2.4.2.7 Transposons carrying host gene fragments Retroposed pseudogenes are duplicated via reverse transcription and reintegration of host mRNA. They can be identified by an absence of introns. However, another mechanism involving TEs can lead to the duplication of genes without a spliced mRNA intermediate. Such duplicated genes will retain their exon-intron structure.

LTR-retrotransposons are the most abundant type of TE in plants (F. Sabot and Schulman, 2006). Their life cycle comprises the transcription of the entire element including LTRs (Figure I6). The mRNA exits the nucleus, is translated and used for protein synthesis. Within virus-like particles, the mRNA is then

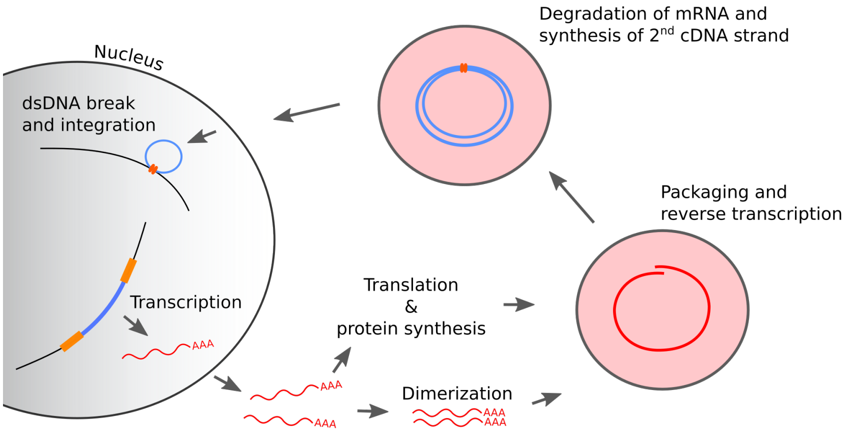


Figure I6: Schematic life cycle of LTR retrotransposons. This figure is derived from F. Sabot and Schulman, 2006.

reverse transcribed before the cDNA reenters the nucleus and is reintegrated into the genome. Sequence insertions within the TE are duplicated during this process as well. Hence, if part of a host gene is inserted into the TE, it is duplicated along with it. Additionally, it does not lose its introns and will not be classified as processed/retroposed, but as duplicated.

Non-autonomous TEs require the activity of another TE to be duplicated, because they do not code for the required proteins themselves. Essentially, only LTR sequences are necessary if the machinery of another TE can be exploited.

Helitrons are DNA transposons that replicate via a rolling-circle mechanism (Kapitonov and Jurka, 2001). They have been shown to frequently capture gene fragments (Barbaglia et al., 2012). At least 2% of the maize genome consists of *Helitrons* — with approximately 2,800 predicted non-autonomous elements of which 94% were found to carry up to nine gene fragments (Du et al., 2009). Additionally, *Helitron* transcription can be accompanied be read-through and conjoined exons from neighboring genes, result in transcripts with potential new functions (Barbaglia et al., 2012).

2.4.2.8 Unitary pseudogenes Unitary pseudogenes have no functional parent gene in the genome and are thus not the result of gene duplication. In human, only 76 unitary pseudogenes have been identified via comparison to the gene sets of other primates. (Z. D. Zhang et al., 2010)

> A functional gene may also become a pseudogene without duplication, if its function no longer confers a fitness advantage to the organism due to a change in the environment or genetic background.
>
> — Xu and J. Zhang (2015)

There are several human unitary pseudogenes that represent a functional loss, because there is no alternative gene with similar sequence: L-gulonolactone oxidase (GULO), major urinary protein (MUP), nephrocan (NEPN), neurotrophin receptor associated death domain (NRADD), threonine aldolase 1 (THA1), and the urate oxidase gene (UOX). Interestingly, most of them share disruptive mutations between human and other primates, indicating that they are at least 30 million years old. (Z. D. Zhang et al., 2010)

2.4.3 Pseudogenes, *poto*genes and proto-genes

Pseudogenes are degenerated genes and usually the result of gene duplication events (Tutar, 2012). They represent a repertory of innovation due to gene-like characteristics and the freedom to accumulate mutations and other sequence changes without immediate consequence (Brosius and Gould, 1992). Their potential of subsequent reactivation and their "capacity to shape an organism during evolution" led to a proposed change in the genomic nomenclature to name them "*poto*genes" (Brosius and Gould, 1992). While this suggested change has been commented on, it did not become an established term (Balakirev and Ayala, 2003; Zheng and M. B. Gerstein, 2007; Sen and Ghosh, 2013).

Another model for gene birth is the emergence of proto-genes from scratch (Carvunis et al., 2012; Schlötterer, 2015). These *de novo* genes are novel protein-coding genes that arise via transcription of non-genic sequences.

Proto-genes are the intermediate step between non-genic sequence and functional genes — in a process that mirrors the pseudogenization mechanism (Figure I7). In *Saccharomyces cerevisiae*, approximately 1,900 such candidate proto-genes have been identified. They also represent a repertory, that "would allow evolutionary innovations to be attempted without affecting existing genes" and they may even be more prevalent than previously assumed. Since the split between baker's yeast and wild yeast, only one to five novel genes were generated via duplication. In contrast, 19 *de novo* open reading frames (ORFs) were found under purifying selection. (Carvunis et al., 2012)

Initially, proto-genes are species-specific and homology-based gene prediction or pseudogene identification tools cannot be used for their annotation. Not only transcriptional but ideally also translational evidence is required for their detection.

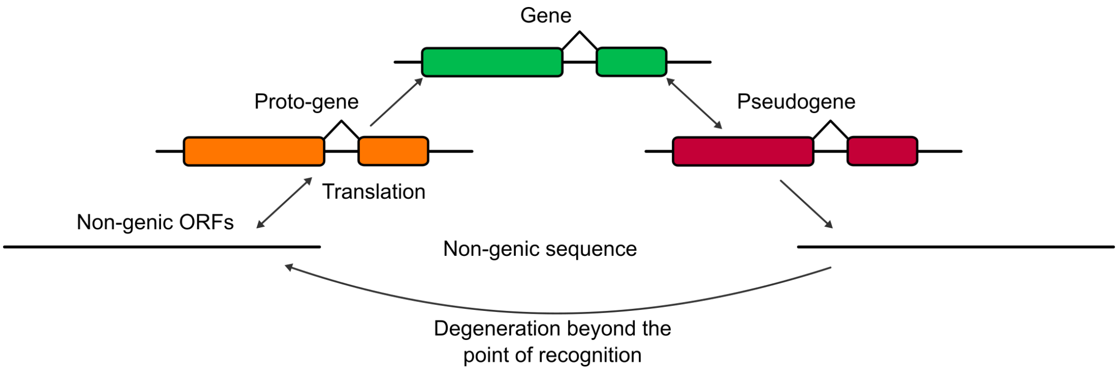


Figure I7: The concept of proto-genes and pseudogenes. The figure is derived from Carvunis et al., 2012 but adapted to contain an arrow connecting the non-genic sequences.

2.4.4 Quantitative analysis of pseudogenes

Most pseudogenes emerge as copies from protein-coding genes (Tutar, 2012). They are either dysfunctional from the moment of their duplication (e.g. due to the loss of promoters) or subsequently degenerate due to the redundancy of information (Podlaha and J. Zhang, 2010). Initially, the nucleotide sequences of parent genes and pseudogenes are highly similar. In time, the defective pseudogene degenerates further and beyond recognition or adopts a new or alternative function. As long as the pseudogene still resembles the parent gene, it can be detected using a homology-based approach. The only data requirements are the genome sequence and gene annotation. An alternative approach would be a *de novo* gene prediction and subsequent identification of pseudogenes in the predicted gene set. This has been done for rice, where the Osa1 Genome Annotation contained 1,439 elements with at least one feature potentially indicative of pseudogenes (Thibaud-Nissen et al., 2009).

There are a few pseudogene prediction tools available (Table I1). However, most are either optimized for mammalian or human pseudogenes, or they focus only on processed pseudogenes. Other pipelines and approaches have been proposed, but are not publicly available. For example, P. M. Harrison (2001) used a homology-based approach to detect pseudogenes in *C. elegans*. The Pseudogene Finder (PSF) was used to identify pseudogenes within 44 selected ENCODE sequences, that were provided as part of the human ENCODE Genome Annotation Assessment Project (EGASP) (Solovyev et al., 2006).

Table I1: Existing pseudogene detection tools

Tool	Target	Description	Source
PseudoPipe	mammals	homology-based pseudogene detection in intergenic regions	Z. Zhang, Carriero, et al., 2006
PPFINDER	mammals, processed	pseudogene detection in (N-SCAN) gene annotations	Baren and Brent, 2006
PseudoGeneQuest	human	online tool to identify pseudogenes given a query sequence	Ortutay and Vihinen, 2008
PseudoDomain	processed	uses conserved protein domain families	Y. Zhang and Sun, 2012

2.4.5 Evolutionary relevance of dysfunctional pseudogenes

Non-transcribed pseudogenes have no apparent function and they have been regarded as evolutionary relics or “junk DNA”. However, they have also been hypothesized to represent a repertory of potential genes with the “capacity to shape an organism during evolution” (Brosius and Gould, 1992). The gene duplicates are free to mutate and can emerge as potentially new genes with new functions. In this mechanism, gene birth is making use of existing gene-like characteristics.

> [...] evolution exploits seemingly dispensable sequences to generate adaptive functional innovation.
>
> — Carvunis et al. (2012)

In rare cases, gene duplication is a mechanism to generate new genes (Kondrashov et al., 2002). Some might stay active and develop new functions, as evolution is a trial-and-error process evaluating variations for their increasing effect on fitness. Similar to most gene mutations being harmful, most pseudogenes will likely remain non-functional and in the end be lost.

Additionally, as it is hard to prove non-functionality, some of the annotated pseudogenes might actually be genes missed in official gene annotations, especially when their function is only observable during a short time period and only in a specific cell type. In human, retroposed genes are generally considered non-functional pseudogenes instead of potentially functional retrogenes (Xu and J. Zhang, 2015).

2.4.6 The significance of “functional pseudogenes”

The term “functional pseudogene” is an oxymoron, since pseudogenes are per definition dysfunctional. However, there is accumulating evidence for pseudogenes,

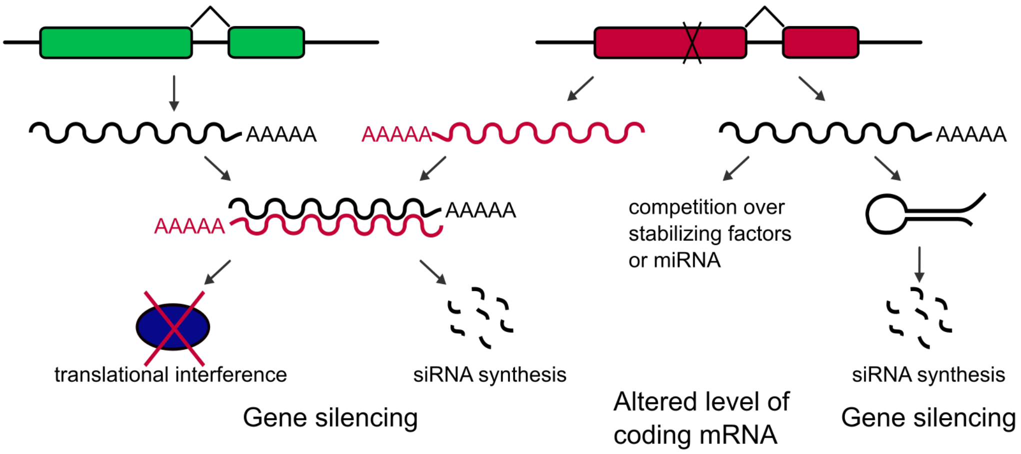

AAAAA
AAAAA
AAAAA
AAAAA
AAAAA
competition over stabilizing factors or miRNA
translational interference
siRNA synthesis
siRNA synthesis
Gene silencing
Altered level of coding mRNA
Gene silencing

Figure I8: Functional potentials of transcribed pseudogenes. The antisense transcript can pair with transcripts of other genes and interfere with the translational machinery. Alternatively, small interfering RNAs (siRNAs) can be generated from the heteroduplex. Both pathways lead to gene silencing. Sense transcripts of pseudogenes can either compete with coding transcripts over stabilizing factors or miRNAs, or they can form hairpin structures, from which siRNAs can be generated. The former leads to an alteration of the coding mRNA level, the latter leads to gene silencing. The figure is derived from Pink et al., 2011.

that are transcribed and take part in regulatory processes. Although these pseudogenes should be regarded and annotated as genes, the term “functional pseudogene” was adopted.

Pink et al. (2011) propose several mechanisms for the functional potential of transcribed pseudogenes (Figure I8): If a pseudogene is transcribed in the same direction as a gene with similar sequence, the transcripts can compete over *trans*-acting stabilizing factors or miRNAs leading to altered expression levels. If transcribed in anti-sense direction, the complementary mRNA strands can form heteroduplexes and interfere with the translational machinery or be used for siRNA synthesis. Both scenarios lead to gene silencing. Finally, the pseudogene transcript itself can form hairpin structures that can be used to generate siRNA, again leading to gene silencing.

In human, approximately 12 percent of the pseudogenes are transcribed and half of them were shown to be conserved and under significant selection pressure (Khachane and Paul M. Harrison, 2009). Approximately 40 percent of long noncoding RNAs (lncRNAs) and pseudogene transcripts are also translated in human (Ji et al., 2015). In 2016, Prieto-Godino et al. reported a pseudo-pseudogene: An olfactory receptor pseudogene in a worm is translated in spite of a premature termination codon due to efficient translational read-through. This natural nonsense suppression can in rare cases increase the error rate of translational termination from less than 0.1% to over 10% (Schueren and Thoms, 2016).

2.4.7 Using pseudogenes to determine neutral mutation rates

Pseudogenes are frequently used to study genome evolution and the rates of neutral substitutions. This, however, has to be pursued cautiously, since non-functionality is hard to determine and there are already numerous examples of pseudogenes being transcribed and conserved. Even for well-studied species, reported pseudogene numbers vary greatly, due to differing identification approaches and underlying pseudogene definitions. At the time of submission in 2006, the pseudogene.org database contained 31,768 pseudogene annotations for the human genome (Karro et al., 2007). Today, the GENCODE annotation (version 21) reports 14,467 human pseudogenes (Harrow et al., 2012). In addition, many pseudogenes are located in tandem situations to functional genes. Gene conversion leads to identical sequences, rendering analyses of neutral mutation rates or pseudogene age unfeasible. Without a reliable pseudogene annotation and certainty of their non-functionality, neutral mutation rates cannot be inferred easily from pseudogene sequences.

2.5 Objectives and Outline

Pseudogenes have long been omitted in genome annotations and analyses, because of their status as "evolutionary relics" or "junk DNA". In mammals — and especially in human — they are now increasingly studied: The discovery of "functional pseudogenes" and their potential role in gene evolution made them an increasingly popular research target. With high-quality reference genome sequences now available for economically important crop plants, their genome-wide assessment and analysis finally become feasible. In this work, pseudogenes are assessed for several plant species, including model organisms but also economically important grasses like bread wheat (*Triticum aestivum*), rye (*Secale cereale*), barley (*Hordeum vulgare*) or rice (*Oryza sativa*) (Figure I9). Pseudogenes are identified computationally by applying a homology-based approach. Requirements for detailed and comparative analyses are a consistent pseudogene definition and a consistent pseudogene detection method. With the generic and modular Pseudogene Locus Identification Pipeline (PLIPipeline) suitable for comparative analyses on large and complex plant genomes, I develop means to analyze pseudogenes in plant genomes. The pipeline is able to handle large and complex datasets — such as the bread wheat genome with 16.9 Gbp — by parallelizing computations. It performs error correction steps to compensate computer malfunctions that can be expected during highly parallelized and time-consuming calculations. After pseudogene candidates are identified, they are classified according to their exon-intron structure and other characteristics like sequence identity, coverage or premature termination codons. An analysis of their parent genes sheds light on preferential duplication and gene birth mechanisms. The functional potential of pseudogenes is determined by their sequence similarity to functional genes and by transcription analysis. Finally, pseudogenes in *Triticeae* are analyzed in more detail. The effect of polyploidization on the pseudogene complement and subgenome specific characteristics are evaluated. This work helps to unlock a significant part of plant genomes and provides the basis for more detailed analyses of individual pseudogenes of interest.

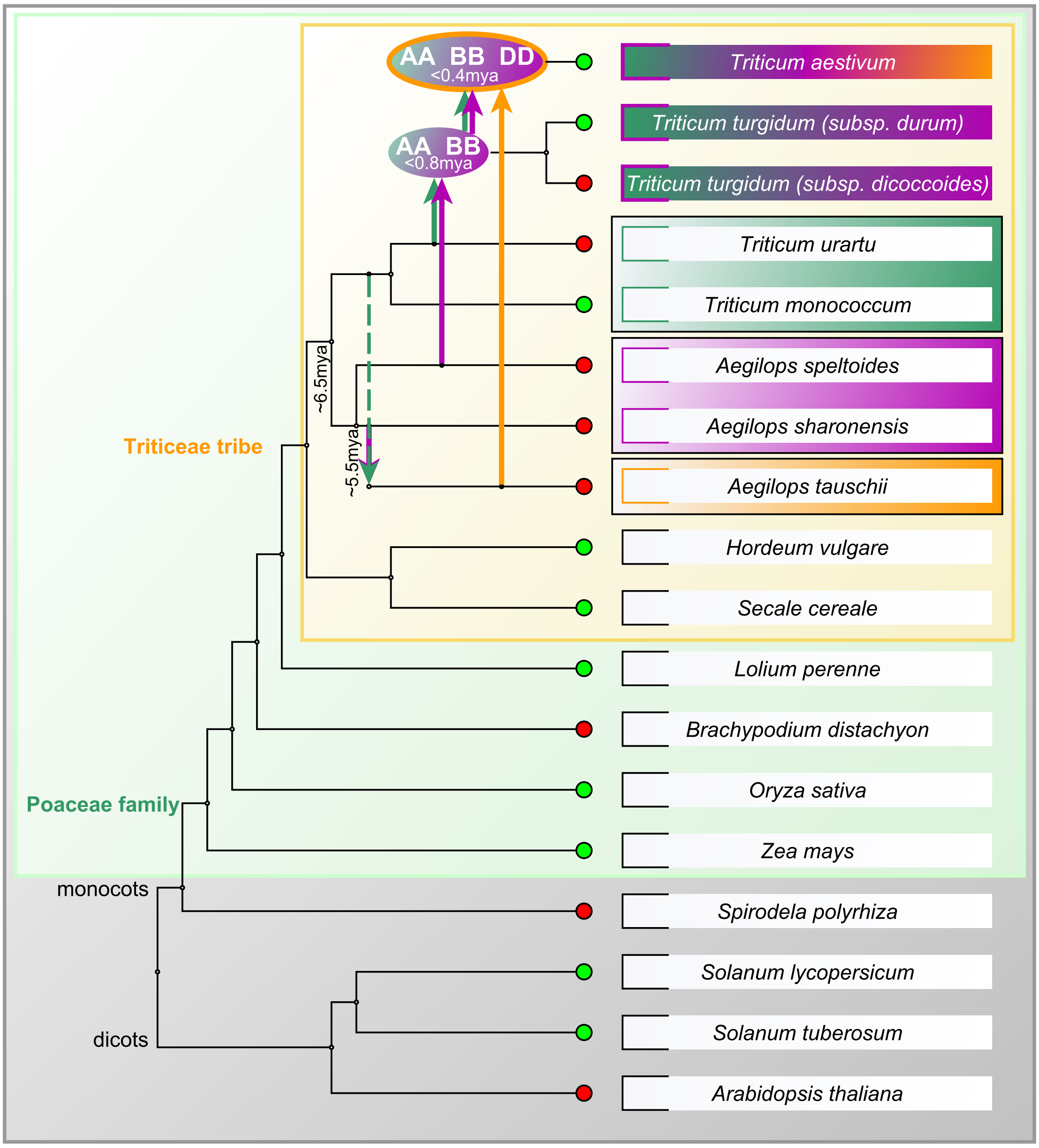


Figure I9: Phylogenetic tree of plant species relevant for this thesis. Tree sections of the *Poaceae* family and the *Triticeae* tribe are highlighted in green and orange, respectively. Cultivated plants are marked by green leaf nodes, while wild plants have red nodes. The closest known subgenome ancestors for polyploid species are framed and arrows indicate polyploidization events. The tree structure was created using the Taxonomy Browser by the National Center for Biotechnology Information (NCBI) (Sayers et al., 2009). Polyploidization data of the *Triticeae* is taken from Marcussen et al. (2014).

3 An introduction to all target species & genome and transcriptome resources

The PLIPipeline was applied on 18 plant species with diverse economic importance and massive differences in genomic complexity (Table I2). Model organisms with smaller and less complex genome structures were investigated, as were crop plants from the *Triticeae* tribe, which stand out due to their massive and highly repetitive genomes.

Table I2: Target species. Binomial nomenclature, English names and additional information for all target plants.

		Abbr.	Latin name	Common name	Cultivar/ Accession
dicots		*A. thaliana*	*Arabidopsis thaliana*	thale cress	Col-0
		S. tuberosum	*Solanum tuberosum*	potato	Phureja DM1-3 516 R44
		S. lycopersicum	*Solanum lycopersicum*	tomato	Heinz 1706
monocots	contigs/scaffolds	*L. perenne*	*Lolium perenne*	perennial ryegrass	P226/135/16
		T. urartu	*Triticum urartu*	red wild einkorn wheat	G1812
		Ae. tauschii	*Aegilops tauschii*	Tausch's goatgrass	AL8/78
		Ae. speltoides	*Aegilops speltoides*	goatgrass	
		T. monococcum	*Triticum monococcum*	einkorn wheat	
		Ae. sharonensis	*Aegilops sharonensis*	Sharon goatgrass	
		S. cereale	*Secale cereale*	rye	Lo7
	pseudomolecules	*S. polyrhiza*	*Spirodela polyrhiza*	common duckweed	
		B. distachyon	*Brachypodium distachyon*	purple false brome	Bd21 (ref. 19)
		O. sativa	*Oryza sativa*	rice	Nipponbare
		Z. mays	*Zea mays*	maize	B73
		H. vulgare	*Hordeum vulgare*	barley	Morex
		T. dicoccoides	*Triticum dicoccoides*	wild emmer	Zavitan
		T. durum	*Triticum durum*	durum wheat (pasta)	Svevo
		T. aestivum	*Triticum aestivum*	bread wheat	Chinese Spring

3.1 *Arabidopsis thaliana*

Arabidopsis thaliana is a flowering plant (angiosperm) and a widely used model organism in plant biology. The dicot is native to Europe, Asia and northern Africa, but it has been introduced or naturalized almost worldwide (Horton et al., 2012). As a weed from the *Brassicaceae* family, its economic and agronomic significance is low. However, its small genome, rapid life cycle and easy cultivation, including a large number of offspring are only some of Arabidopsis' advantages as a research subject (The Arabidopsis Genome Initiative, 2000). The small and diploid genome of Arabidopsis was the first plant genome to be sequenced.

3.2 The *Solanaceae* family

Potato (*Solanum tuberosum*) and tomato (*Solanum lycopersicum*) are dicots of the *Solanaceae* family and valuable crop plants. The potato tuber is a starchy vegetable and a staple food especially in Europe. The second most consumed vegetables are tomato fruits and it is thus a widely used subject for the research of fruit development. (Lall et al., 2013)

Both their genomes have a size of approximately 800 Mbp, with more than 50% being repetitive element related (Mehra et al., 2015). Interestingly, the *Solanum* lineage has undergone one ancient and one recent genome triplication, while only few individual genes remain triplicated in the modern, fully sequenced genome (The Tomato Genome Consortium, 2012).

3.3 *Spirodela polyrhiza*

Spirodela polyrhiza is an aquatic plant growing on freshwater surfaces and found in large parts of the world. Easily mistaken for an algae, it actually belongs to the smallest, fastest growing and morphologically simplest flowering plants, called duckweeds (*Lemnoideae*). Spirodela is of economic interest due to its easy cultivation in aquatic habitats, without competing for arable land. Its rapid growth and starchy composition make it ideal for animal fodder production, or as a source for biofuel. Spirodela has a very small diploid genome of only 147 Mbp showing no sign of recent retrotransposition events, but of two ancient whole-genome duplications, which possibly took place around 95 million years ago (mya). (W. Wang et al., 2014)

3.4 *Zea mays*

Maize (*Zea mays*) is a member of the *Poaceae* family and one of the major sources for the daily calorie uptake by humans. Compared to other grasses, maize has a mid-sized genome with 2.7 Gbp. Although the modern maize genome is diploid, it has undergone a history of duplication and polyploidization events: *Z. mays* is an ancient allotetraploid, which eventually readopted its diploid state. Approximately 70 mya, it had a paleopolyploid ancestor, followed by an additional whole-genome duplication event about 5 to 12 mya. (Schnable et al., 2009)

3.5 *Oryza sativa* subsp. *japonica*

Oryza sativa — commonly known as rice — is one of the main sources for the daily calorie uptake by humans (Haberer et al., 2016). It is a member of the *Poaceae* family and a cereal grain. The rice genome shows evidence for ancient whole-genome duplication, segmental duplications and massive ongoing individual gene duplications (J. Yu et al., 2005). Duplicated chromosome segments are still well discernible and cover over 65% of the genome (J. Yu et al., 2005). These regions also contain duplicated genes, that may very well have pseudogenized or are presently in the process of pseudogenization. Rice is thus a convenient plant to study the impact of segmental duplications on gene death and birth.

3.6 *Brachypodium distachyon*

Brachypodium distachyon is a model organism and a member of the grass family (*Poaceae*) and *Pooideae* subfamily (The International Brachypodium Initiative, 2010). Genome sequencing and assembly efforts of many economically important grasses have been hampered by their large and complex genomes. High sequence repetivity and transposable element content lead to collapsed assemblies and insufficient sequence resolution. Brachypodium is a wild annual grass with a comparatively small genome size of 355 Mbp, making it the perfect model organisms for more complex cereal grasses like barley (5.4 Gbp), rye (8.1 Gbp) or bread wheat (16.9 Gbp).[7]

3.7 *Lolium perenne*

Perennial ryegrass (*Lolium perenne*) is a forage and turf grass species essential for forage-based meat and dairy production in temperate regions worldwide. The genome of *L. perenne* has a size of 2.7 Gbp and is assembled to contigs representing only 54% of the genome. (S. L. Byrne et al., 2015)

3.8 The *Triticeae* tribe

Triticeae include economically important crops like wheat, rye or barley. Ten *Triticeae* species from different genera are analyzed in this work: five species from the *Triticum* genus, three from the *Aegilops* genus, as well as *Hordeum vulgare* and *Secale cereale*.

Barley (*Hordeum vulgare*) is one of the major contributors to the human calorie uptake. Like all cereal crops, it was initially domesticated in the Fertile Crescent. Barley is used for food production, brewing alcoholic beverages or as animal fodder (International Barley Sequencing Consortium, 2017).

Rye (*Secale cereale*) — compared to barley or bread wheat — is more tolerant to abiotic stresses and it is thus used as a model organism for functional analyses.

[7] C-values are taken from `http://data.kew.org/cvalues`, (April 2, 2018, 12:30pm CEST)

In Europe, it is a domesticated cereal of regional importance for food or feed (Martis et al., 2013).

The wild grasses from the *Aegilops* genus are *Ae. sharonensis*, *Ae. speltoides* and *Ae. tauschii*. *Ae. sharonensis* is one of the closest wild known relatives of the B-lineage of bread wheat (International Wheat Genome Sequencing Consortium, 2014). The diploid grass species is known as Sharon goatgrass and especially of interest because of its resistance to diseases like leaf rust and stripe rust (Millet et al., 2014). *Ae. speltoides* and *Ae. tauschii* are of particular interest due to their heat stress resistance (Hairat and P. Khurana, 2015).

Finally, five species from the *Triticum* genus are analyzed in this work: *T. monococcum*, *T. urartu*, *T. dicoccoides*, *T. durum* and *T. aestivum*. Einkorn wheat (*T. monococcum*) is a diploid wheat species and one of the first plants to be domesticated. Even though it is more resistant to disease and drought than common wheat varieties, its yield is considerably lower (Hidalgo and Brandolini, 2014). It is closely related to *T. urartu* and it is the A subgenome donor of the polyploid *T. durum*, *T. dicoccoides* and *T. aestivum*. *T. dicoccoides* and *T. durum* are both subspecies of *T. turgidum*. Hence, their correct binomial nomenclature is *Triticum turgidum* subsp. *dicoccoides* and *Triticum turgidum* subsp. *durum*, respectively. Henceforth, the shortened nomenclature will be used to improve readability. The tetraploid durum wheat (*T. durum*) is grown for pasta production (International Wheat Genome Sequencing Consortium, 2014). The plant with the largest and most complex genome analysed in this work is bread wheat (*T. aestivum*). It is one of the most important crop plants worldwide, providing 18%[8] of the calorie uptake by humans (Food and Agriculture Organization of the United Nations, 2017). It has a highly complex and large genome of ∼16.9 Gbp and comprises 42 chromosomes from three subgenomes (AA BB DD). The closest known relatives of the three subgenomes are *T. urartu* (AA), *Ae. sharonensis* (BB) and *Ae. tauschii* (DD). They diverged approximately 5 mya (International Wheat Genome Sequencing Consortium, 2018). Less than one million years ago, the first polyploidization event combined the A and B subgenomes, a constellation still present in todays durum wheat and wild emmer wheat (International Wheat Genome Sequencing Consortium, 2014). The second polyploidization event happened less than 0.4 mya and formed the now hexaploid genome of bread wheat (Marcussen et al., 2014).

[8] Calculated as 527/2,884 kcal/capita/day; Data retrieved for the year 2013 from `http://www.fao.org/faostat/`, (January 3, 2018, 3:20pm CET)

Table I3: Target species. Source of sequences and annotations.

		Abbr.	Source
dicots		*A. thaliana*	TAIR10, Berardini et al. (2015)
		S. tuberosum	The Potato Genome Sequencing Consortium (2011)
		S. lycopersicum	The Tomato Genome Consortium (2012)
monocots	contigs/scaffolds	*L. perenne*	S. L. Byrne et al. (2015)
		T. urartu	Ling et al. (2013)
		Ae. tauschii	Jia et al. (2013)
		Ae. speltoides	Mario Caccamo, The Genome Analysis Centre TGAC, Norwich, UK (unpublished)
		T. monococcum	Doreen Ware, Cold Spring Harbor Laboratory, USA (unpublished)
		Ae. sharonensis	Burkhard Steuernagel, The Sainsbury Laboratory, Norwich, UK (unpublished)
		S. cereale	Bauer et al. (2017)
	pseudomolecules	*S. polyrhiza*	W. Wang et al. (2014)
		B. distachyon	The International Brachypodium Initiative (2010)
		O. sativa	Kawahara et al. (2013)
		Z. mays	Schnable et al. (2009)
		H. vulgare	International Barley Sequencing Consortium (2017) SRA: SRP076351 (Chengdao Li, wild barley accessions)
		T. dicoccoides	Avni et al. (2017)
		T. durum	Maccaferri et al. (2018)
		T. aestivum	International Wheat Genome Sequencing Consortium (2018)

3.9 Data sources

Sequence and annotation data were taken from both external and in-house sources (Table I3). The transposable element and repeat annotations was provided by Heidrun Gundlach from the Plant Genome and Systems Biology (PGSB) group.

4 Methods

Pseudogenes are defined as gene-like sequences that have degenerated and lost their original function. With gene annotations available, I begin with the assumption that every gene-like sequence not present in the official gene annotation is a pseudogene. Their computational identification was achieved by exploiting their sequence homology to functional genes (Figure M1). Annotated genes were filtered and mapped onto the unmasked genome sequence to identify pseudogene candidates. Those candidates mainly comprise previously uncharted elements. Annotated protein-coding genes with characteristics of pseudogenes — e.g. premature termination codons — are assessed and examined separately. Since pseudogenes were identified via homology to template genes, each one can be assigned to a specific protein-coding parent gene. Sequence coverage, identity and other attributes are determined via comparison to this parent gene. Furthermore, structure comparison was used to determine the mechanisms involved in pseudogene creation.

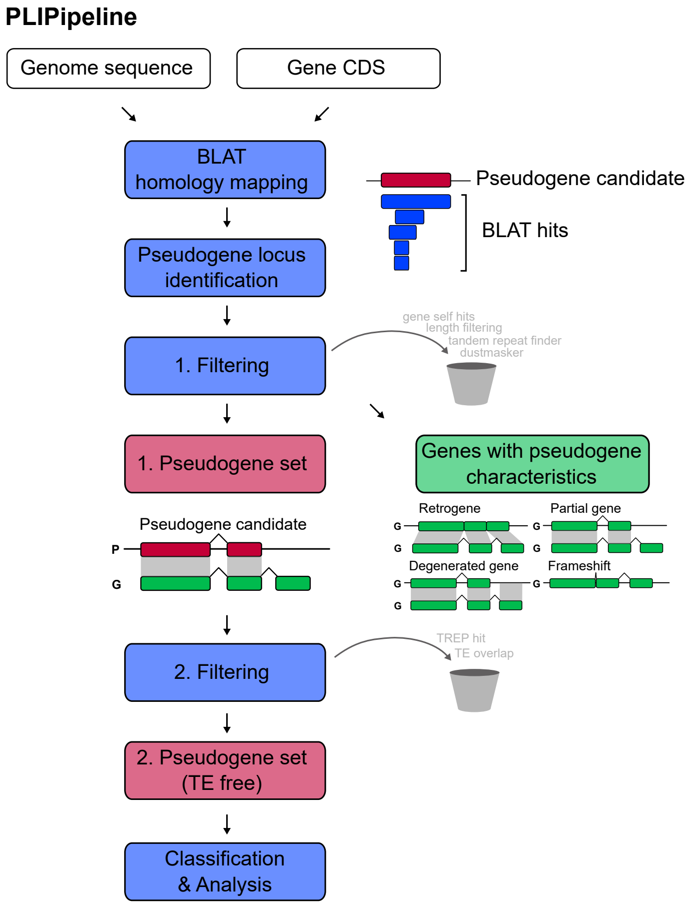


Figure M1: Basic framework of the PLIPipeline.

The PLIPipeline consists of two main parts: First, it identifies pseudogenes via a homology-based approach and stores the data in a database. And second, it combines various functions and tools for the subsequent analysis of pseudogenes — generating simple metrics and figures but also performing more complex tasks like functional analyses.

4.1 Quantitative analysis of pseudogenes: the PLIPipeline

4.1.1 Data preprocessing to facilitate parallel execution

Pseudogenes were assessed via homology to functional protein-coding genes. Thus, two data sets are necessary for their identification: (i) template gene sequences (queries) and (ii) genome sequences (targets). Since a number of the plants have large and complex genomes, extensive preprocessing steps are necessary to reduce both computing time and resource requirements via parallel processing. Thus, genome sequences were split into non-overlapping batches of 1–5 Mbp. Larger and contiguous sequences were split at regions of unknown sequence (Ns).

TE-related genes were filtered from the query gene set via a keyword search in the functional description. In rare cases, the gene annotation contains elements with PTCs or with a coding sequence (CDS) length not divisible by three. Those elements were filtered as well. All isoforms of the remaining query genes were used for the homology mapping on the target genome sequence.

Several additional measures were taken to reduce computing time, but they were later discarded due to quality reduction: (i) The search space for potential pseudogenes was reduced significantly by masking TEs and genes. However, this time-efficient approach resulted in highly fragmented pseudogene candidates and hampered a good structural classification (Figure M2). (ii) The coding sequences of the representative gene isoforms were stringently clustered and only the cluster representative was used for the homology mapping. While this approach reduced the number of query genes, it also introduced a difficulty in determining the best parent gene. Furthermore, retroposed pseudogenes may be copies of smaller splice variants and the representative splice variant does not always contain all exons of the gene locus. If all splice variants are included in the query gene set, the best parent can easily be determined and the results are more accurate.

4.1.2 The homology mapping

The CDS nucleotide sequences of all query gene isoforms were mapped onto the target genome sequences using the BLAST-like alignment tool (BLAT) (Kent, 2002; v36x2) with a minimal identity of 70% and a maximum intron length of 2,500 bp. The average intron length of the query genes lies between 157 and 887 bp, but their

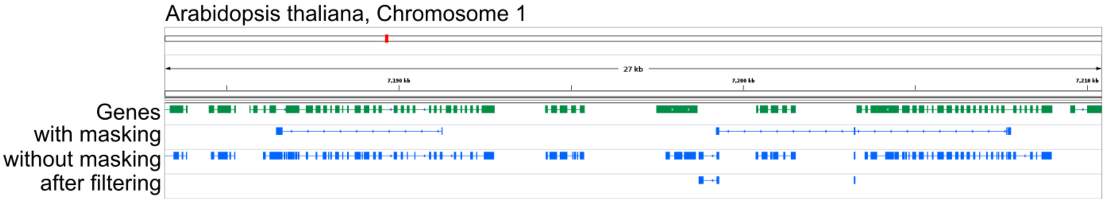


Figure M2: Effect of gene and TE masking on the pseudogene annotation. A region on chromosome 1 of *Arabidopsis thaliana* with genes (green), PLIP pseudogene annotation with prior gene masking (top blue), without masking (middle blue) and without masking, but after filtering pseudogenes overlapping with annotated genes (lower blue).

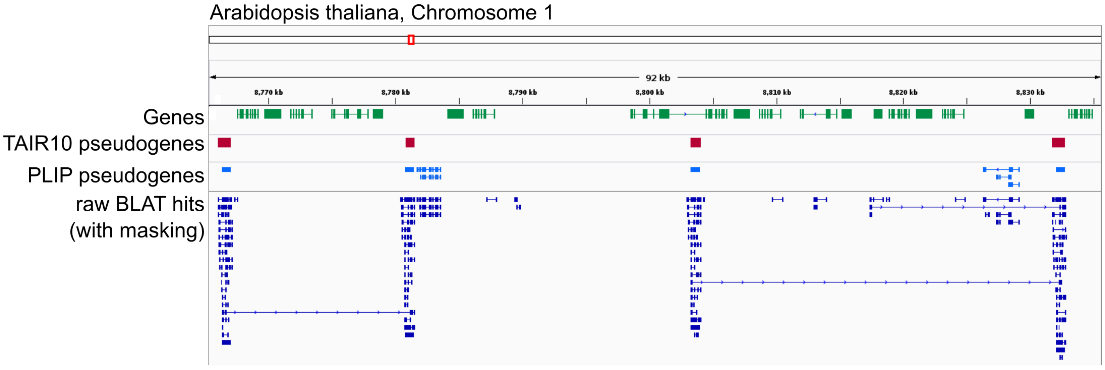


Figure M3: Overlapping BLAT hits with larger intron size. A region on chromosome 1 of *Arabidopsis thaliana* with genes (green), TAIR10 pseudogenes (red), PLIP pseudogene candidates (sky-blue), and raw BLAT hits (dark blue). These are results of an earlier PLIPipeline version where gene sequences had been masked.

maximum often exceeds 10 kbp. Hence, the maximum intron size conditioned for PLIP pseudogenes is much lower than the maximum intron size annotated for genes. This was done to avoid erroneous pseudogene loci (Figure M3). As illustrated, BLAT creates spliced alignments and thus recovers exon-intron structures, allowing for a subsequent intron-based classification of pseudogenes. However, it creates a massive amount of overlapping and highly fragmented hits, that need to be filtered extensively.

4.1.3 Pseudogene post-processing and filtering

BLAT creates spliced alignments an recovers exon-intron structures, but it does not allow insertions. Instead, small gaps are introduced at inserted sequences. To reduce the fragmentation of BLAT pseudogene candidates, gaps smaller than

10 bp were closed. The amount of added sequence was calculated into sequence identity. Second, BLAT hits were filtered to have a length of at least 100 bp and at least one fragment (i.e. exon) with 50 bp. Gene self hits were filtered immediately and hits overlapping with other genes were filtered from the PLIP pseudogene set, but kept for subsequent analyses. Since numerous BLAT hits only covered regions consisting of small tandem repeats, further filtering steps were introduced to reduce spurious hits of low information content. The Tandem Repeats Finder (Benson, 1999) was used to mask repetitive sequences within BLAT hits (version 4.07b, `matching weight=2, mismatching penalty=7, indel penalty=7, match probability=80, indel probability=10, minimum alignment score to report=50, maximum period size to report=10`). Additionally, the Dustmasker from NCBI-BLAST (Gish, 1996–2003; version 2.2.25+, default settings) was used to mask sequences of low information content. Hits masked to more than 65% and those with less than 50 bp left are removed.

Many BLAT hits are overlapping and pseudogene loci were defined with one representative BLAT hit each. Overlapping hits were merged into one locus and the longest hit (excluding introns) was chosen as the representative. Often, the representative hit is of low quality, since it does not span at least 60% of the locus. Figure M3 illustrates this problem: Two pseudogenes are connected by only one hit with a larger intron. The two pseudogenes would collapse in one pseudogene locus, but the representative hit only covers a small portion of the locus. If the "bad" hit with the long intron is removed, the pseudogene locus splits up into two, with two separate representative hits. Bad hits that are removed include the one with the smallest length and all those smaller than half of the current representative hit. Hits are removed and new representative hits are determined in a recursive manner to allow the split of larger pseudogene loci until all representative hits are of good quality. Only the representative hits were then collected in the PLIP pseudogene set.

The pseudogene set comprises many TE-related sequences that were filtered from the final set. First, pseudogene sequences were mapped onto the *Triticeae*-specific TREP database (Wicker, François Sabot, et al., 2007). Sequences with a TREP hit covering at least 75% of the CDS with a minimal sequence identity of 90% were filtered. Additionally, pseudogenes overlapping with the comprehensive TE annotation (≥75% overlap) were filtered as well. All filtering steps combined significantly reduced the number of pseudogene candidates by up to 90%.

4.1.4 Intermediate PLIPipeline versions

Continuous effort was put into the development and improvement of the PLIPipeline. At the same time, is was applied on a growing number of target plant genomes. To finally obtain comparable results for all target plants, the PLIPipeline was re-applied on all target genomes. However, some analyses are too time- and resource-consuming to repeat them.

The comparative analyses of pseudogenes in wild barley accessions and between wild emmer and durum wheat was done using an older PLIPipeline version. Filtering TE-related pseudogenes was done not by mapping their sequence onto the TREP database, but by filtering pseudogenes that originated from parent genes with an unusually high number of pseudogene children (>50).

An analysis of unitary pseudogenes in wild emmer and durum wheat was done by combining the query gene sets of both. Hence, pseudogenes that share sequence similarity to a gene in the other subspecies, but not to a gene within the same species, can be identified.

4.2 Qualitative analysis of pseudogenes

4.2.1 Basic pseudogene attributes

Pseudogenes have attributes that distinguish them from their parent genes. BLAT returns a sequence identity for each hit, that reflects how many nucleotides are identical between pseudogene and parent gene. Since smaller gaps have been closed subsequently, the added sequences were used to recalculate sequence identities. Sequence coverage reflects how completely the pseudogene represents the parent gene. Thus, complete gene copies have a coverage of 100%, while fragmented pseudogenes have smaller coverages (Figure M4). However, two different approaches have been used to calculate the pseudogene coverage: (i) compared to the complete gene locus and the merged CDSs of all isoforms (ii) compared only to the CDS of the parent gene isoform. If not stated otherwise, coverages are calculated using the first approach.

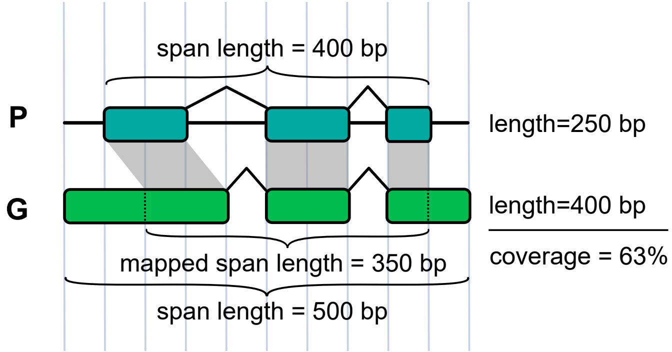


Figure M4: Basic pseudogene attributes.

4.2.2 Mutations affecting the potential functionality of pseudogenes

PTCs were determined independently for each pseudogene fragment (i.e. exon), always starting in the correct frame compared to the parent gene. This was done to compensate potential mapping errors at the fragment borders. Secondly, they were determined on the complete pseudogene sequence with concatenated exons. The second approach is likely to result in more premature termination codons, because frameshifts in one exon carry on to all following exons. PTCs may be located anywhere in the pseudogene sequence, but to significantly affect encoded protein sequences, they have to be located earlier in the sequence. Thus, early PTCs were defined as PTCs located within the first half (5' half) of the CDS compared to the parent gene. Frameshifts are insertions or deletions that affect the reading frame of a pseudogene and often lead to PTCs. Thus insertions or deletions of

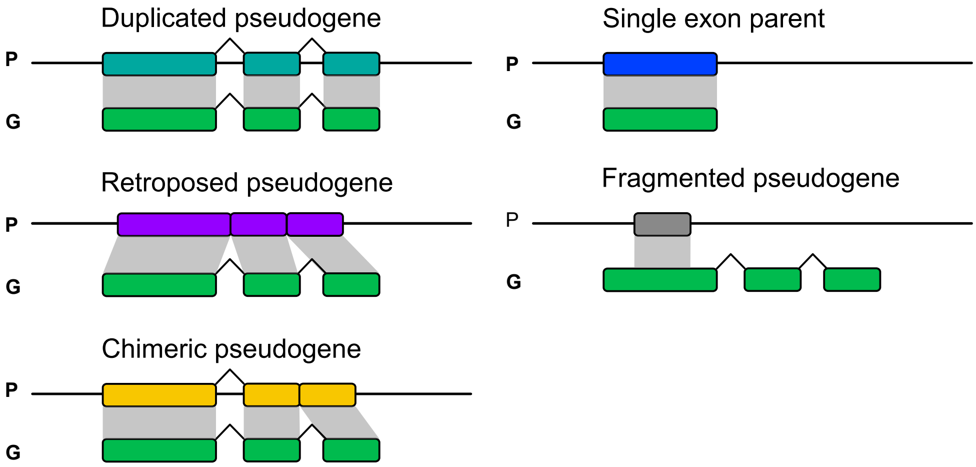


Figure M5: Intron-based classification of pseudogenes.

a length not divisible by three are considered fameshifts. Early frameshifts are defined analogously to early PTCs — within the first half of the CDS.

4.2.3 Pseudogene classification

4.2.3.1 Intron-based classification The exon-intron structure of pseudogenes was used to classify them into duplicated or retroposed pseudogenes. For this intron loss/retention criterion, I defined five pseudogene classes: (i) “duplicated” pseudogenes still contain introns at each covered splice site; (ii) “retroposed” or “processed” pseudogenes have lost all introns; (iii) “chimeric” pseudogenes have both retained and lost introns; (iv) “single-exon” pseudogenes from genes with only one exon; (v) “fragmented” pseudogenes which do not sufficiently cover a splice site.

If the parent gene does not contain any introns of at least 30 bp, the pseudogene is immediately classified as single-exon. Otherwise, all splice sites within the mapped query gene region (+/− 20 bp) are determined. Annotated genic introns with a length smaller than 30 bp are omitted. If the exon-intron structure is preserved between gene and pseudogene, the latter is classified as duplicated. For this, all corresponding pseudogenic introns need to have a length of at least 30 bp. If all or some introns are lost, the pseudogene is classified as retroposed or chimeric, respectively.

Manual scrutiny of pseudogene and parent gene structures showed, that this classification method works very well on duplicated, chimeric and single-exon pseudogenes. However, retroposed pseudogenes often gained insertions at non-intron positions. Those pseudogenes had a particularly fragmented structure and while they indeed have lost all introns, a manual classification would have put them into the “fragmented” or sometimes even “duplicated” group. Thus, processed pseudogenes were reevaluated in an additional step.

To ultimately be classified as retroposed, introns/insertions may not account for more than 20% of the pseudogene. If the span length (exons plus introns) of the mapped parent gene region and the corresponding pseudogene region are highly similar (>90%), the pseudogene is classified as duplicated, even if the splice sites are not completely matching. If more than 20% of the pseudogene represents introns/insertions, it is classified as fragmented.

4.2.3.2 Poly-A regions Retroposed pseudogenes arise via reintegration of reversely transcribed mRNA into the genome. During the maturation of the mRNA, the 3’ end of the transcript is polyadenylated. Since many pseudogenes cannot be classified as duplicated or retroposed because they originate from a single-exon parent gene, the presence of a poly-A tail may help to assess their origin. Poly-A rich regions were determined for the complete genome (min. 16 Adenines in a window of 20 nt; both strands respectively) and their occurrence was counted in the 3’ region close to each pseudogene (<300 bp).

4.2.3.3 UTR mapping The homology of pseudogenes and their parent genes may not extend beyond CDS and untranslated region (UTR) of the parent gene, if they originated in retrotransposition events. Since only the CDSs of genes was used to identify pseudogenes, an additional mapping of the UTR regions is required to test for this homology. The average length of 5’ UTRs and 3’ UTRs in plants is ∼100 and ∼200 bp, respectively (Mignone et al., 2002; Mazumder et al., 2003). Since UTR annotations are not available for all plants, homology was tested for the sequences adjacent to the CDS of the parent gene. For this purpose, the Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990; version 2.3.0+, `e-value=100, max. target sequences=600000, culling limit=5`) was used to remap the complete query regions (+/− 1 kbp) onto pseudogene regions. The extend of the homology beyond the CDS was determined and compared.

4.2.4 Tandem duplications

Gene and pseudogene sequences were first clustered using CD-HIT (Li and Godzik, 2006; Fu et al., 2012; version 4.6.5, `identity=0.8, alignment coverage of the longer sequence=0.3, alignment coverage of the shorter sequence=0.95, global identity=0`) to identify tandem duplications. Elements within one cluster that are separated by less than 50 kbp form a tandem cluster, regardless of whether other genes or pseudogenes are annotated between them.

4.2.5 Gene families and orthologous groups

Gene families were identified using TRIBE-MCL (Dongen, 2000; Enright et al., 2002; version 13–137). First, blastall (Altschul et al., 1990) was used on the representative splice variant CDS of all genes with an E-value threshold of 1×10^-5. Then, mcxdeblast was used to parse the BLAST output files into MCL.

Orthologous groups were defined for barley: High-confidence genes (HC genes) and annotated gene sets of *Sorghum bicolor*, *Brachypodium distachyon*, *Oryza sativa* and *Arabidopsis thaliana* were used to run the OrthoMCL software (version 2.0, default parameters). A total of 170,925 genes were clustered into 24,337 gene families. Sequences from all five genomes were contained in 8,608 clusters. Expanded and contracted gene families were extracted as described in International Barley Sequencing Consortium (2017).

4.2.6 Functional analysis — GO enrichment

The functional analysis was done by assigning Gene Ontology (GO) terms and human-readable description lines from the parent gene set to the pseudogene children. The GO annotation of parent genes was compared to the GO annotation of all genes and enriched or depleted GO terms were identified. This enrichment analysis was done using the GOstats R package (Falcon and Gentleman, 2007) with an adjusted P-value cutoff of 0.05 and only using terms occurring at least 10 times in the "universe" gene set. If required, the resulting GO terms were then grouped using REVIGO (Supek et al., 2011) with a similarity threshold of 0.5 and *A. thaliana* as the GO term database. For visualization purposes, GO terms had to be filtered using more stringent P-value thresholds or GO term level restrictions. If this was necessary, it is mentioned per analysis.

The pairwise comparative analysis between wild emmer and durum wheat was performed differently. The GO enrichment analysis was not performed on the parent genes compared to the complete gene sets, but on the respective pseudogene sets compared to the combined pseudogene set. GO terms were assigned to the pseudogenes based on their parent genes.

4.2.7 Selection pressure on pseudogenes — K_A/K_S analysis

A K_A/K_S analysis was done to study selection pressure on pseudogenes. First, pseudogene and parent gene sequences were aligned and edited. Clustalw2 (Larkin et al., 2007, default parameters) was used for pairwise alignment. The alignments had to be edited, since subsequent tools cannot handle PTCs or unknown sequence. PTCs or aligned codons containing gaps or unknown sequence (Ns) were removed, while protecting the frame of the protein-coding gene. Furthermore, a minimum alignment length of 150 bp is necessary for significant results. Codeml from the PAML package (Yang, 2007) was used to calculate K_A and K_S values (`seqtype=1, CodonFreq=2, runmode=−2, model=0, NSsites=0`). The distribution of the K_A/K_S ratios was tested with the `normaltest` function from the `scipy.stats` package before using `ttest_1samp` from the same package to test whether the distribution is significantly shifted to the left.

4.2.8 Transcribed pseudogenes — RNA-seq analysis

Transcribed pseudogenes may have functional potential. RNA sequencing (RNA-seq) data can be used to assert transcription, but the high sequence similarity between pseudogenes and genes hampers a correct mapping of RNA-seq reads. Hence, only reads mapping uniquely to the pseudogene and not to the parent gene can be used to determine pseudogene expression. Thus, the pseudogene is required to contain mutations that makes it distinguishable from any other sequence. The result is an underrepresentation of transcribed pseudogenes. Furthermore, no conclusions about transcription rates can be drawn.

An RNA-seq analysis was performed for barley. Hisat2 was used to align RNA-seq reads (International Barley Sequencing Consortium, 2017) to the barley genome (`--dta-cufflinks`). Samfiles were then filtered for a minimal mapping quality value of 60. Samtools (version 1.3) was used to convert and sort the files into BAM format. Cufflinks and Cuffcompare (version 2.2.1) were used to assemble the alignment files into a single set of transcripts. Finally, pseudogenes were checked for transcriptional evidence. Pseudogenes often are highly similar to parent genes and mutations may only be present in defined regions. Thus, transcription evidence for 50 bp in either direction of the pseudogene sequence was deemed sufficient to determine transcription.

4.2.9 Identification of syntenic pseudogenes

Syntenic regions were identified via bidirectional BLAST searches. Genes or pseudogenes were blasted against each other with an E-value cutoff of 0.5. Bidirectional hits with an E-value highly similar to the best E-value (>90%) were selected. For visualization purposes, the Dagchainer (Haas et al., 2004; `min. alignment pairs=10/15, gap length=100000/200000, max. distance=1000000/5000000, ignore tandem=False`) was used to identify syntenic blocks.

4.2.10 Visualization tools

Most figures were created with the `Matplotlib` package of `Python`. The Circos software package (version 0.69–6) was used for circular figures of chromosomes.

The Integrative Genomics Viewer (IGV) was used to browse the genomes for regions of interest and to evaluate pseudogene identification and classification (e.g. Figures M2 and M3).

4.2.11 Comparison between Morex barley and wild barley accessions

The assemblies of four wild barley accessions were investigated for differences in gene and pseudogene complements. Contigs and scaffolds were filtered to have a minimum length of 200 bp and a maximum of 35% Ns. The CDSs of representative splice variants of Morex barley genes were then mapped onto the assemblies analogous to the PLIPipeline. The resulting hits were comprised of genes and pseudogenes. Elements were classified as genes, if they do not contain PTCs that shorten the CDS by more than 15 bp. Additionally, their sequence has to be almost identical to the Morex gene (>95%). If it is smaller than 800 bp, the gene has to be covered to at least 98%, otherwise a coverage of 75% is sufficient. This is a very stringent definition of genes and it led to low gene numbers. The remaining hits were divided into pseudogenes that contain PTCs and those that do not. Genes and pseudogenes often are located at the borders of contigs. They are split and have a low coverage. A manual scrutiny of certain pseudogene regions was required to estimate whether an element is truly a pseudogene or a gene. Their sequences was aligned to their respective parent genes using megablast or blastn (Altschul et al., 1990). Syntenic blocks were visualized for contigs containing at least three genes with homologues on a maximum stretch of 1 Mbp of the same Morex barley chromosome. A pairwise comparative visualization of homologous blocks was created between Morex barley and each of the four wild barley accessions, if available. CD-HIT (Li and Godzik, 2006; Fu et al., 2012) was used to cluster the query genes and to determine corresponding pseudogenes (95% sequence identiy, 80% coverage). Any element in a cluster is connected in the visualization of homologous blocks. In total, over 800 homologous regions were visualized and combined if at least one gene is shared between the Morex barley regions. This reduced the number of figures to 203 combined plots that were manually inspected for evaluation purposes and to identify regions of interest.

4.2.12 Identifying genes with pseudogene characteristics

The mapping of query gene sequences onto the genome sequence to identify gene-like elements was done without prior masking of genes or TEs. Gene self-hits were filtered, but genes mapped onto other genes may reveal interesting gene evolution processes. Additionally, a gene clustering was performed using CD-HIT (Li and Godzik, 2006; Fu et al., 2012; `identity=80%, coverage=98%`). Many genes are duplicates that remained functional due to beneficial effects. However, not all of

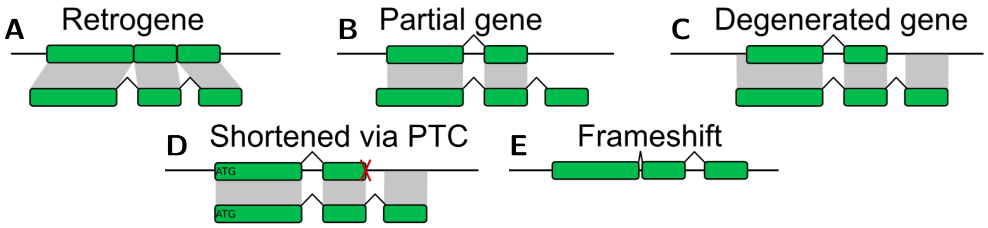


Figure M6: Genes with features typical for pseudogenes.

them still resemble their parent gene perfectly (Figure M6). Retrogenes are genes that arise via a retrotransposition event (Figure M6 A). To remain functional or regain functionality, a promoter is necessary at the insertion site. Retrogenes are identified via sequence clustering. If two genes with similar length (+/− 10 bp) are clustered together and one does not contain introns while the other gene does, it is classified as retrogene. This likely results in an under-estimation of retrogenes, because a high coverage was required in the clustering. However, retrocopies may also be copies of smaller splice variants.

Some genes are partial or degenerated copies of other longer genes (Figure M6 B and C). Partial genes are incomplete copies, while degenerated genes are complete copies that are not annotated over the whole region of homology. This may be due to a PTC leading to a shortened annotation of the CDS (Figure M6 D). Shortened genes have to be covered by the BLAT hit to at least 60%. Additionally, the query gene is required to be at least 100 bp longer than the shortened gene.

There are also genes within the official protein-coding gene sets, that contain PTCs, CDS lengths not divisible by three or putative frame shifts. Introns smaller than 20 bp are likely artifacts. While these micro-introns may in fact be sequencing or assembly errors, they may also be insertions that lead to frameshifts. Annotated genes with such features can be considered pseudogenes.

5 Results

The Pseudogene Locus Identification Pipeline (PLIPipeline) was used to identify pseudogenes in 18 plant species. The target species exhibit different economical importance or genome complexity, but also varying assembly and annotation qualities. In a homology-based approach, coding sequences (CDSs) of annotated genes were used as templates to identify gene-like sequences in the entire genome (Figure M1). Each pseudogene has a parent gene from the homology mapping. Sequence identity, coverage or premature termination codons (PTCs) were determined in respect to that parent gene. The mapping was done without masking genes or transposable elements. Instead, pseudogene candidates were extensively filtered and classified according to their exon-intron structure.

5.1 Evaluation of the PLIPipeline

Whole genome assessments of pseudogenes are available only for a few plant species and numbers can vary significantly. The Arabidopsis Information Resource (TAIR) is a database providing genome annotations for *Arabidopsis thaliana*. The Arabidopsis genome annotation is well curated and in part experimentally verified. The TAIR10 database contains 4,827 pseudogenes or transposable element (TE) genes.[9] Of those, 924 are classified as non-TE-related pseudogenes. The annotation by the PLIPipeline contains 5,157 pseudogenes including TE-related sequences and 2,849 pseudogenes after TE-filtering. 766 (83%) of the TAIR10 pseudogenes overlap with a pseudogene annotated by the PLIPipeline. When filtering TE-related pseudogenes from the PLIPipeline pseudogene set, 669 (72%) overlapping annotations remain. This either indicates that TAIR10 contains pseudogenes overlapping with the in-house PGSB TE annotation, or pseudogenes overlapping with TAIR10 are subsequently filtered by the PLIPipeline due to a match to the (*Triticeae*-specific) TREP database.

Only a small fraction of the TAIR10 pseudogenes are fully covered by a PLIPipeline annotation, as the pipeline often produces fragmented annotations containing gaps. The larger number of pseudogene candidates in the PLIPipeline annotation may be explained by short gene fragments missing from the TAIR10 pseudogene annotation. On the other hand, pseudogenes may be missed by the PLIPipeline due to TE over-masking. Alternatively, some may also represent unitary pseudogenes, that do not have a parent gene in the Arabidopsis gene set. This could be solved by using a combined plant gene set as template genes for the homology mapping, an option which is already implemented in the PLIPipeline and

[9] `https://www.arabidopsis.org/portals/genAnnotation/gene_structural_annotation/annotation_data.jsp`, (December 28, 2017, 1:30pm CEST)

used for comparative analyses of closely related plants. Since the TAIR genome sequence and annotation are well maintained and manually curated, a true positive rate of 83% is more than satisfactory for a purely computational assessment.

5.2 Genomes, genes and repetitive elements — the data base for the pseudogene annotation

The pseudogenes of economically important crops like bread wheat, rice or maize were assessed and examined during this project, as were those of plants with less complex genomes. Most of the 18 target species belong to the *Poaceae* family and to the *Triticeae* tribe, but also three eudicotyledons (dicots) and one additional monocotyledon (monocot) were analyzed (Figure I9). The target species have genome sizes ranging from 156 megabase pairs (Mbp) to 16.9 gigabase pairs (Gbp), reflecting a >100 fold change (Table R1). The two main reasons for this huge difference in genome size are repetitive elements and polyploidy. For example, bread wheat has the largest genome due to its hexaploidy and a repetitivity of over 80%. Differences in the genome assembly quality also affect gene and pseudogene annotation quality. High-quality and continuous genome sequences were available for 11 species. For the others, fragmented contig assemblies were used with N50 values ranging between 683 and 63,687 base pairs (bp). Gene annotations are of varying quality as well, as reflected by differing gene numbers or absent gene isoform annotations (Table R1).

Triticeae genomes are largely composed of TE-related elements, which are distributed over almost the entire length of the chromosomes (Figure R1). In smaller plant genomes, the prevalent class of TEs (long terminal repeat (LTR)-retrotransposons) are mainly located around the centromeric region. The gene space of all investigated plants mirrors the distribution of TEs. This could indicate either a prevalent insertion of TEs into gene-poor regions, a low tolerance of TE insertions into gene space, or a quick removal of TEs from gene-rich regions.

Many genomic regions are not annotated at all. These regions represent a “gray area”, that contains regulatory sequences, highly-degenerated TEs, pseudogenes or other non-classified DNA or “junk-DNA”. One goal of this project is to close this gap a little further to obtain a more complete picture of plant genomes and to estimate the functional potential of pseudogenes.

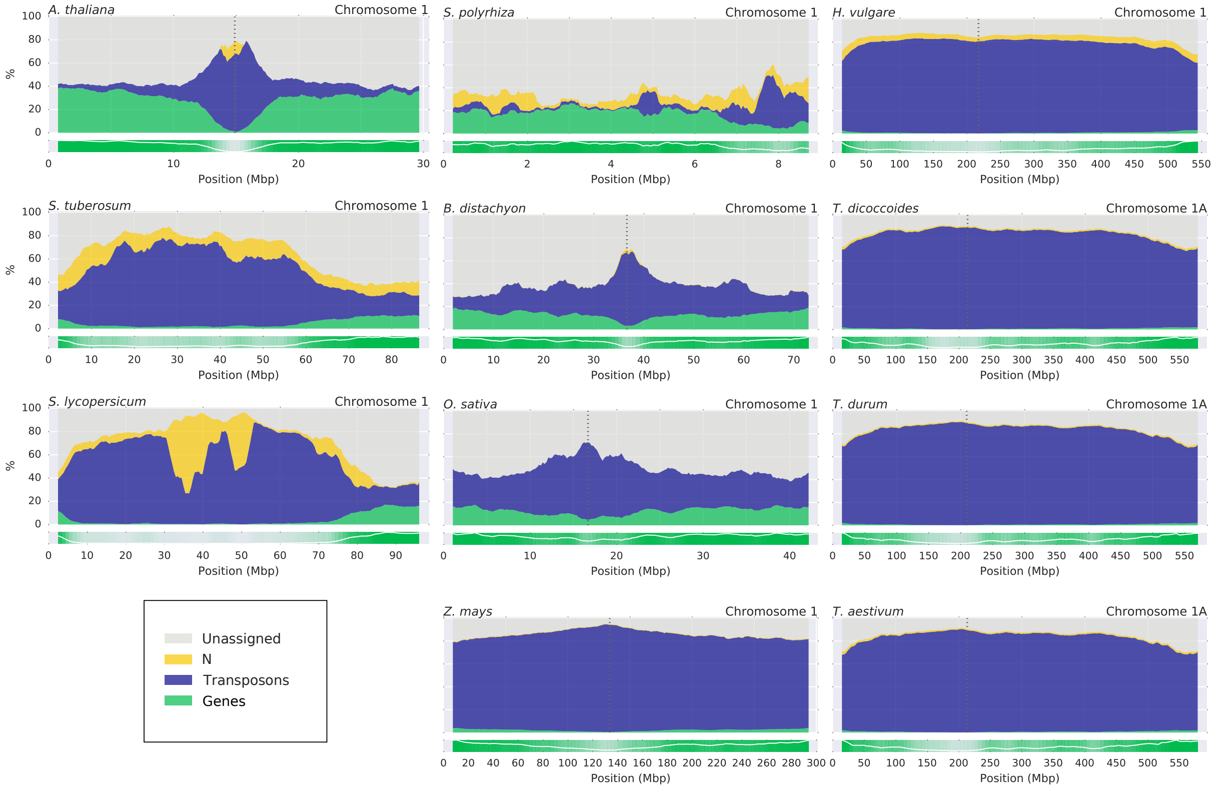


Figure R1: Chromosome composition. Selected chromosomes and their composition of TEs, genes and undefined sequence (N stretches). Centromere positions are indicated with a dotted vertical line if available.

Table R1: Target plant species and genome metrics. The N content represents the amount of unknown sequence.

		Species	Genome size*(Mbp)	Assembly size(Mbp)	Coverage (%)	N content (%)	#Genes	#Isoforms	N50
dicots		*A. thaliana*	156	119	76	0.2	27,416	35,386	–
		S. tuberosum	856	773	90	12.5	39,028	56,215	–
		S. lycopersicum	1,002	823	82	10.5	34,725	–	–
monocots	contigs/scaffolds	*L. perenne*	2,695	1,362	51	34.4	26,710	69,541	15,930
		T. urartu	4,817	4,660	97	15.8	34,879	–	63,687
		Ae. tauschii	4,988	4,147	83	15.2	43,150	–	50,625
		Ae. speltoides	5,037	1,760	35	0.9	43,750	74,607	942
		T. monococcum	6,088	1,208	20	1.0	32,047	61,060	683
		Ae. sharonensis	6,895	1,514	22	1.0	34,406	67,800	1,000
		S. cereale	8,093	1,684	21	1.0	31,869	–	1,708
	pseudomolecules	*S. polyrhiza*	291	145	50	11.7	19,623	–	–
		B. distachyon	355	271	76	0.4	26,552	31,029	–
		O. sativa	489	375	77	0.0	39,049	49,066	–
		Z. mays	2,665	2,066	78	0.6	38,914	63,540	–
		H. vulgare	5,428	4,833	89	8.0	39,734	248,180	–
		T. dicoccoides	12,005	10,509	88	1.8	67,182	205,916	–
		T. durum	12,377	10,463	85	1.5	66,559	196,153	–
		T. aestivum	16,944	14,547	86	1.9	110,790	135,104	–

*C-values are taken from `http://data.kew.org/cvalues`, (April 2, 2018, 12:30pm CEST)

5.3 Pseudogene metrics in plants — a matter of definition, data quality differences or biological differences

Pseudogenes are gene-like sequences that have lost their functionality. This most common definition of pseudogenes allows a wide scope for interpretation. Based on this, the number of annotated pseudogenes can vary tremendously. Published results from different sources are often not comparable. Even within this project, the PLIPipeline evolved and the final version had to be reapplied on most genomes to allow for comparability. In addition to the diversity of interpretation, genome assembly and annotation quality also affect gene and pseudogene annotations.

In this work, pseudogenes were defined as containing a length of at least 100 bp with a sequence corresponding to the CDS of an annotated protein-coding gene. Pseudogenes related to TEs were identified and filtered by mapping their sequence to the *Triticeae*-specific TREP database or by checking for overlaps to available and comprehensive TE annotations. (s. Methods)

For *Triticeae* plants with high-quality reference genome sequences, filtering TE-related sequences reduces the number of pseudogene candidates considerably by 84% to 95%, since their genomes are highly repetitive and TE-rich (Table R2). In contrast, repetitive regions of low-quality assemblies are collapsed, which leads to a reduced search space for the pseudogene detection. This results in a much smaller number of pseudogene candidates for low-quality genome assemblies.

T. urartu and *Ae. tauschii* have an unusual high number of pseudogenes compared to other *Triticeae* with contig/scaffold assemblies. Both those genomes were assembled by the same institution (BGI) and have a significantly higher N50 value compared to the other contig/scaffold assemblies. This indicates a high quality and completeness of the two genome assemblies, even though pseudomolecule sequences are not available.

Potato and tomato are closely related, yet their pseudogene number varies considerably: Potato contains over three times more pseudogenes prior to TE-filtering than tomato. This excess in TE-related pseudogenes cannot be attributed to a more complete genome assembly, since potato has a smaller genome and assembly size, but a higher N-content (unspecific sequence). In tomato, the amount of this unknown sequence peaks at several distinct loci. For plants with smaller genomes, TEs accumulate close to the centromere and the highest density of TE-related pseudogenes is usually found in the same region. If the sequence at this positions is unknown, many pseudogenes cannot be detected. Henceforth, it is possible that the centromeric region in tomato is not well represented in the assembly, leading to a significantly smaller number of TE-related pseudogenes.

Table R2: Pseudogene filtering metrics. Number of pseudogene candidates after various filtering steps. Percentages are calculates using the two preceding columns, respectively.

			Number of pseudogenes						
		Species	**before TE**	**after TE**	**%**	**TE filtered**	**HCov**	**HCov**	**%**
			filtering	**filtering**		**with PTC**		**with PTC**	
dicots		*A. thaliana*	5,167	2,849	55	2,060	373	271	73
		S. tuberosum	287,696	31,727	11	21,840	5,231	3,844	73
		S. lycopersicum	79,493	23,033	29	17,626	6,026	4,756	79
monocots	contigs/scaffolds	*L. perenne*	56,062	41,173	73	21,744	4,200	2,265	54
		T. urartu	1,930,667	299,450	16	140,314	7,185	4,642	65
		Ae. tauschii	1,495,156	270,379	18	142,554	4,497	3,271	73
		Ae. speltoides	100,587	83,463	83	39,062	11,231	6,072	54
		T. monococcum	49,316	44,362	90	22,376	6,191	3,284	53
		Ae. sharonensis	72,893	57,736	79	30,605	7,783	4,611	59
		S. cereale	108,033	80,091	74	45,282	13,259	9,327	70
	pseudomolecules	*S. polyrhiza*	11,454	2,625	23	1,713	277	204	74
		B. distachyon	23,308	11,161	48	6,606	956	687	72
		O. sativa	87,568	20,880	24	11,161	2,693	1,739	65
		Z. mays	210,199	75,954	36	45,105	6,990	4,189	60
		H. vulgare	8,043,394	397,514	5	183,423	22,237	15,236	69
		T. dicoccoides	2,734,893	266,652	10	161,871	25,147	17,199	68
		T. durum	2,074,453	246,211	12	148,054	25,589	17,560	69
		T. aestivum	1,780,537	289,132	16	173,611	48,608	33,611	69

However, this hypothesis is contradicted by the TE landscape of potato (see section 5.12.2): There is no distinct peak of TE-related pseudogenes, but an almost even distribution within gene-poor regions. Henceforth, the most likely explanation might be a contamination of the potato gene annotation with TE-genes.

Compared to other *Triticeae*, barley has a very high pseudogene number prior to TE filtering. Clustering pseudogene sequences indicates that those high numbers are probably the result of very few pseudogenes that exist in high copy numbers (Figure R2). For example, the largest pseudogene cluster in bread wheat contains little over 20,000 sequences, while the largest cluster in barley contains almost 25,000 pseudogenes. Given that bread wheat is hexaploid and barley is only diploid, it would have been expected that bread wheat contains approximately three times the number of pseudogenes compared to barley. In comparison, tetraploid wild emmer and durum wheat only have a maximal cluster size of ~11,000 and ~14,000 pseudogenes. In all target plants, TE filtering preferentially removes pseudogenes from larger clusters. Hence, parent genes from pseudogenes that occur in very high copy number are likely TE genes that have evaded classification. Many of them are of unknown function.

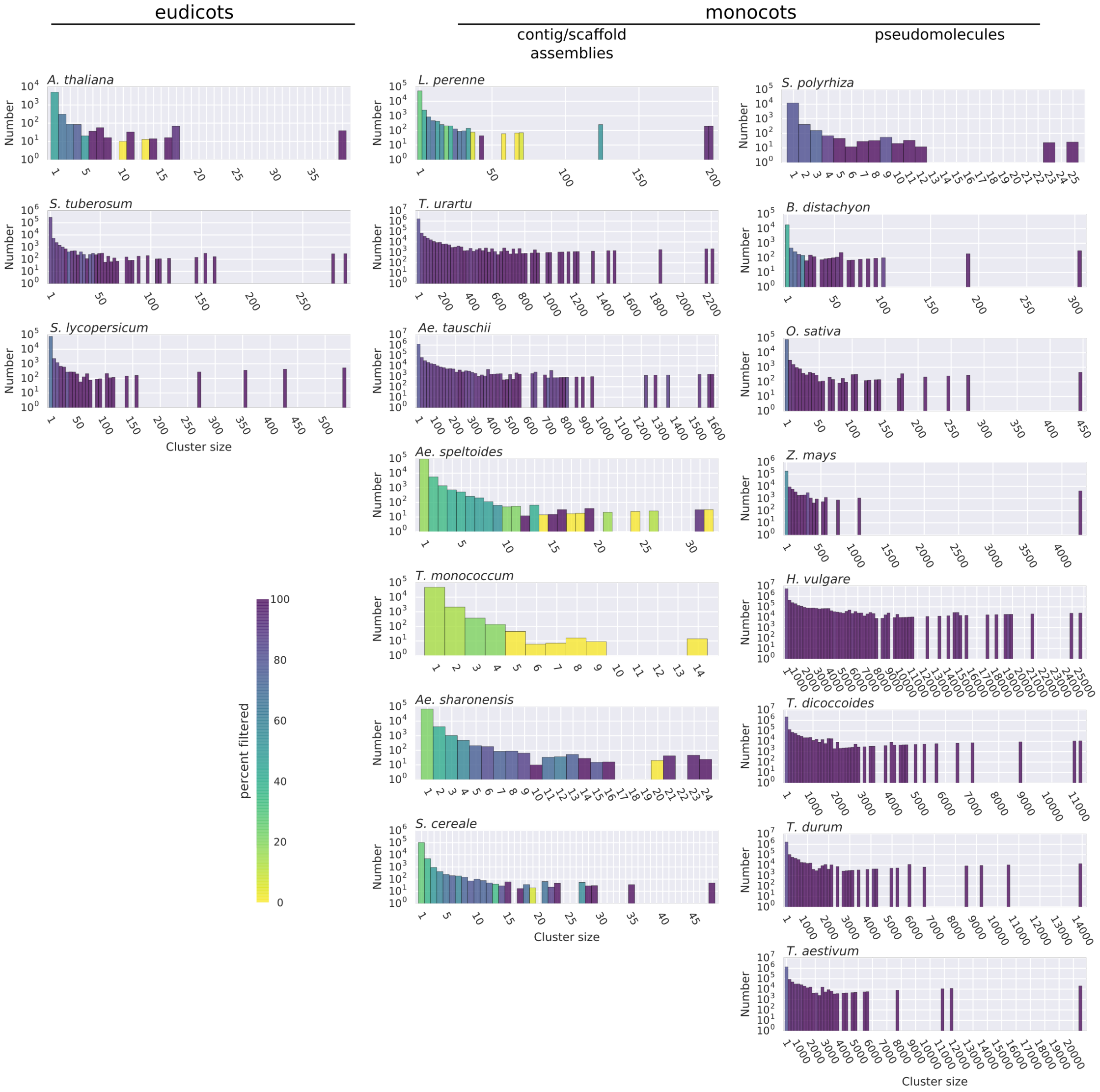


Figure R2: Clustered pseudogene sequences before TE-filtering. Pseudogene sequences were clustered with high stringency (95% identity, 95% coverage). The color gradient shows the percentage of TE filtered pseudogenes per cluster size.

Several reasons are motivation for filtering TE-related sequences from the pseudogene sets: (i) The assembly quality has a pronounced effect on how complete the TE space is represented in the assembled genome sequence. Thus, filtering TE-related pseudogenes reduces quality-dependent differences and renders comparative analyses possible and more significant. (ii) The quality of gene annotations affects pseudogene numbers, because TE genes will be parents to a huge number of pseudogenes. Filtering TE-related pseudogenes helps to correct this difference and contributes to increased significance of comparative analyses. (iii) TE (pseudo)genes are outside of the focus of this work. However, since TE-related pseudogenes and gene fragments occur in such high numbers, their mention is necessary.

Plant genomes contain large amounts of small gene fragments: After TE filtering, 74% to 98% of the elements cover their parent gene's CDS to less than 80% (Table R2). In contrast, the number of high-coverage (HCov) pseudogenes ranges between 277 in Spirodela and 48,608 in bread wheat. It is always below the number of annotated protein coding genes. Smaller gene fragments may represent older pseudogenes with partial sequences degenerated beyond the point of recognition, or pseudogenes that fragmented due to sequence insertions and deletions (e.g. TE insertions). Finally, they can represent partially duplicated genes, for example due to double strand DNA break repair. Their putative origin will be investigated further.

5.4 Pseudogene numbers correlate with genome size

While the number of genes directly affects genome size in bacteria, this is not generally true for eukaryotes (Elliott and Gregory, 2015). For example, rice and barley have almost the same number of annotated protein coding genes, but their genome size differs by a factor of ten (Table R1). Also, since young polyploid genomes can contain several subgenomes, they also tend to harbor a multiple of the diploid gene set and may contain more pseudogenes due to relaxed constraints and degeneration. For the target species in this project, gene and HCov pseudogene numbers do overall correlate with genome size (Figure R3). A reason for the correlation of size and pseudogene number could be a constant duplication and pseudogenization rate affecting the genes. If a higher TE activity led to larger genomes, it could also have led to increased gene duplication and pseudogenization rates. In case of class 2 DNA transposons, "DNA repair following transposon excision is associated with an increased number of mutations in the sequences neighbouring the transposon" (Wicker, Y. Yu, et al., 2018). Alternatively, the duplication of genes into TE space might be less harmful than insertion into the often essential gene space. According to this scenario, pseudogene distribution should mirror gene distribution and resemble the distribution of TEs. Additionally,

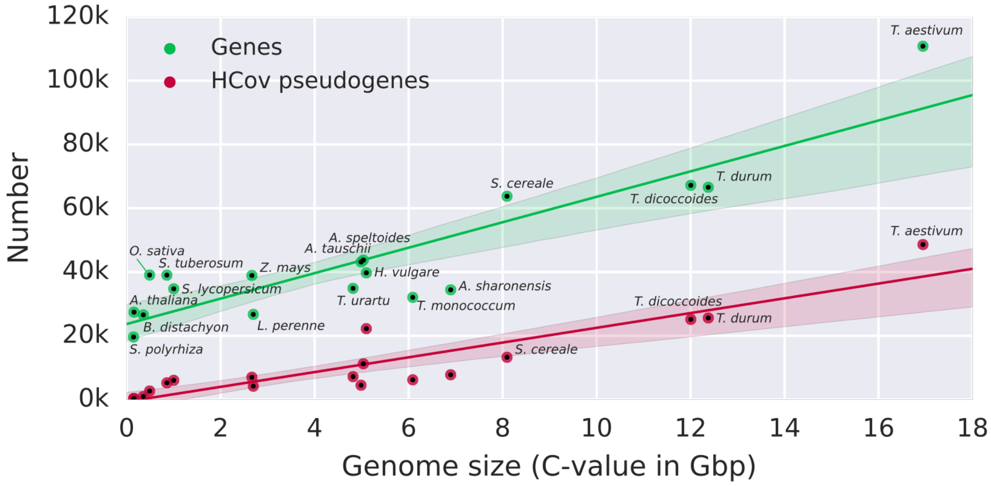


Figure R3: Genome size vs. protein coding gene and HCov pseudogene number. The `seaborn regplot` function of `Python` was used for the linear regression model fit with a confidence interval of 90%.

pseudogenes would be expected to evolve neutrally within TE space and negative selection should quickly degrade and remove pseudogenes from the gene space. All of these hypotheses will be addressed in the following sections.

5.5 Duplication mechanisms and the origin of pseudogenes

Pseudogenes are degenerated copies of functional genes. The two mechanisms commonly described in generating gene duplicates are unequal crossing over during meiosis or retrotransposition, but there are numerous other mechanisms known or suspected to generate gene or gene fragment duplicates (Gorbunova and Levy, 1997; Z. Zhang, Paul M. Harrison, et al., 2003; Wicker, Buchmann, et al., 2010). Structural and positional features of pseudogenes can shed light on prevalent duplication mechanisms and hint towards the origin of pseudogenes.

5.5.1 Structural features as evidence for origin

Pseudogenes were classified according to their exon-intron structure in comparison to the respective parent genes (Table R3). While duplicated pseudogenes are expected to retain intron sequences, retrotransposed pseudogenes are a result of spliced messenger RNA (mRNA) translocation, reverse-transcription and reintegration. Absent intron sequences are the result of the mature mRNA intermediate. Hence, the pseudogene structure is the main feature after which classification is usually achieved, but others are 3' poly-adenine (poly-A) tails or direct small re-

Table R3: HCov pseudogene classification. Pseudogenes were classified according to their exon-intron structure in comparison to the respective parent genes.

		Species	duplicated	%	retroposed	%	chimeric	%	single-exon	%	fragmented	%
dicots		*A. thaliana*	130	35	3	1	2	1	223	60	15	4
		S. tuberosum	1,730	33	108	2	2	0	3,131	60	260	5
		S. lycopersicum	1,771	29	56	1	1	0	3,878	64	320	5
monocots	contigs/scaffolds	*L. perenne*	805	19	27	1	8	0	2,986	71	374	9
		T. urartu	3,323	46	57	1	6	0	3,232	45	567	8
		Ae. tauschii	2,517	56	10	0	0	0	1,596	35	374	8
		Ae. speltoides	2,134	19	29	0	6	0	8,225	73	837	7
		T. monococcum	1,121	18	27	0	5	0	4,338	70	700	11
		Ae. sharonensis	1,593	20	21	0	5	0	5,556	71	608	8
		S. cereale	3,408	26	71	1	12	0	8,708	66	1,060	8
	pseudomolecules	*S. polyrhiza*	149	54	3	1	1	0	96	35	28	10
		B. distachyon	384	40	3	0	1	0	521	54	47	5
		O. sativa	876	33	8	0	3	0	1,631	61	175	6
		Z. mays	1,977	28	56	1	2	0	4,647	66	308	4
		H. vulgare	2,542	11	142	1	3	0	19,185	86	365	2
		T. dicoccoides	11,981	48	280	1	34	0	11,513	46	1,339	5
		T. durum	12,309	48	242	1	32	0	11,603	45	1,403	5
		T. aestivum	21,376	44	346	1	65	0	24,373	50	2,448	5

peats, that are present for retroposed pseudogenes (Rouchka and Cha, 2009; Xiao et al., 2016).

5.5.1.1 Intron-based classification According to the intron-based classification, duplicated HCov pseudogenes outnumber retroposed HCov pseudogenes up to 252 fold (Table R3, Figure R4). On average, 34% of the pseudogenes are duplicated and only 1% are retroposed. The remaining 63% are either single-exon pseudogenes, highly-fragmented, or have a chimeric structure. Of the plants with a complete reference genome sequence, Brachypodium has the highest duplicated to retroposed ratio. That duplicated pseudogenes dominate over retroposed pseudogenes has also been observed in previous studies for Arabidopsis or rice (Thibaud-Nissen et al., 2009; L. Wang et al., 2012). In contrast, most mammals show a different pseudogenization pattern: In human, 70% of the 19,724 pseudogene regions have a retrotranspositional origin (Torrents et al., 2003). As such, it is surprising that plants with highly repetitive and TE-rich genomes do not contain more retroposed pseudogenes. This can be attributed to the different composition of TEs in the genomes of mammals and plants.

5.5.1.2 Transposable elements and pseudogenes More than 75% of the barley genome consists of LTR-retrotransposons. Approximately 25,000 of them are full-length elements and potentially active (International Barley

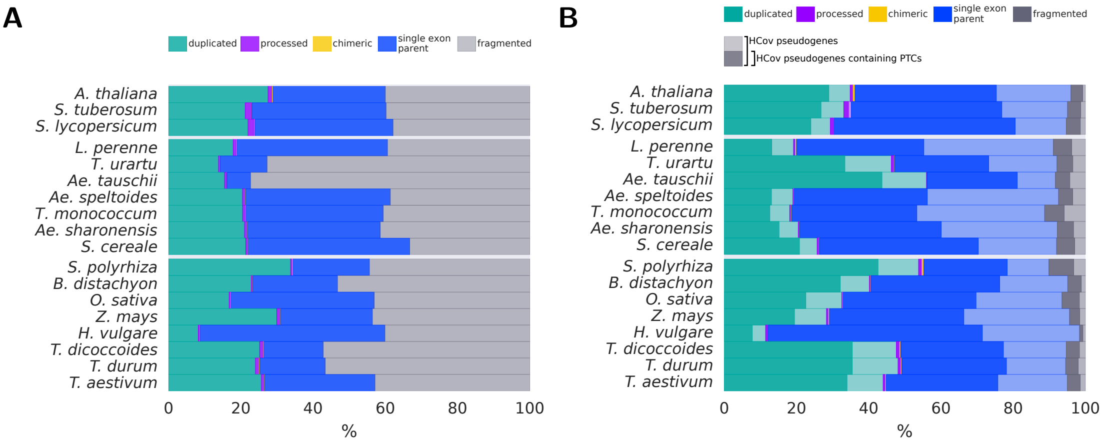


Figure R4: Pseudogene classes. A: Pseudogene classes of the TE filtered set. B: Pseudogene classes of HCov pseudogenes. Color intensity highlights the portion of pseudogenes containing PTCs. Species are separated into dicots (top), monocots with contig/scaffold genome assemblies (middle) and monocots with chromosomes assembled into pseudomolecules (bottom).

Sequencing Consortium, 2017). Only 5.3% of the barley genome is occupied by DNA-transposons and less than one percent by non-LTR-retrotransposons. In contrast, the human TE landscape looks much different: 8% LTR-retrotransposons, 33.7% non-LTR-retrotransposons and ∼3% DNA transposons (Cordaux and Batzer, 2009). Of the human non-LTR-retrotransposons, 16.9% represent long interspersed element 1 (LINE-1)-retrotransposons and 10.6% are Alu sequences (SINEs) — elements which were found in only 0.3% of the barley genome. The machinery of this type of non-LTR-retrotransposon is the source of retroposed pseudogenes in human (Pavlicek et al., 2006) and the small number of LINE-1-retrotransposons in barley and other plants could be a potential explanation for the low number of retroposed pseudogenes in plants.

5.5.1.3 Short gene fragments While the intron-based classification illustrates a remarkable difference between the frequencies of duplicated vs. retroposed pseudogenes, it cannot be applied on all pseudogenes and many remain unclassified. For example, the intron-based classification fails for gene duplicates lacking introns or for highly fragmented pseudogenes, which are too short or do not cover any splice sites. Of all HCov pseudogenes, 59% originate from single-exon genes and 7% are classified as fragmented, because despite their high coverage, splice sites are not sufficiently covered. According to their sequence coverage and identity, pseudogenes can be grouped into two major clusters (Figure R5). Most pseudogenes have a coverage below 20%, but some also represent full-length gene copies. In *Arabidopsis thaliana* and *Oryza sativa* there is even a clearly discernible cluster of pseudogenes with both high sequence coverage and identity. It is possible, that some of those pseudogenes are still functional, but other features like the lack of a promotor might have disabled them despite their CDS similarity to a functional gene.

There are several possible scenarios that might account for the massive number of small gene fragments. Gene fragments could have been duplicated via DNA break repair mechanisms. One DNA break repair mechanism that can result in duplicated sequences is the NHEJ/SDSA mechanism. An ectopic template is used to bridge the double-strand break. The filler DNA that is inserted at the double-strand break can have a length of up to 1,200 bp (Gorbunova and Levy, 1997) and can also contain parts of a gene. Alternatively, a repair mechanism of two adjacent single-strand breaks has been proposed to result in tandem duplications (Schiml et al., 2016). Compared to the unequal crossing over mechanism, it does not require microhomologies. However, the resulting tandem duplicate can facilitate further duplications via the unequal crossing over mechanism.

Furthermore, fragmented pseudogenes could represent older and more degenerated elements. The accumulation of mutations may render parts of the sequence unidentifiable. Additionally, DNA insertions or deletions can potentially fragment pseudogenes. The homology mapping via BLAST-like alignment tool (BLAT) allows for introns or insertions within the pseudogene hit. However, the maximum allowed gap length needs to be chosen carefully. The average intron length of genes is between 157 and 889 bp for all 18 analyzed plants, respectively. The maximum intron length often exceeds 10 kilobase pairs (kbp), but only occurs for very few annotated genes. In this work, a maximum gap length of 2.5 kbp was chosen. This guarantees the full-length detection of most gene-like elements, but also prevents questionable pseudogene constructs, that have been observed during initial trial runs. Pseudogenes that originated from genes with larger introns, or that experienced a large-scale sequence insertion, are potentially split up into several annotations — each with a lower sequence coverage.

5.5.1.4 Poly-A tails of retroposed pseudogenes Another feature hinting towards a retrotranspositional origin of a pseudogene is the presence or absence of a poly-A signature. At transcription termination, the 3'-end of mature mRNA is polyadenylated. The spliced mRNA is reversely transcribed into complementary DNA (cDNA) and reintegrated into the genomic DNA at a potentially different chromosomal site. Since annotations for untranslated regions (UTRs) are not available for all target species, potential poly-A tails were identified in the region downstream of their CDS.

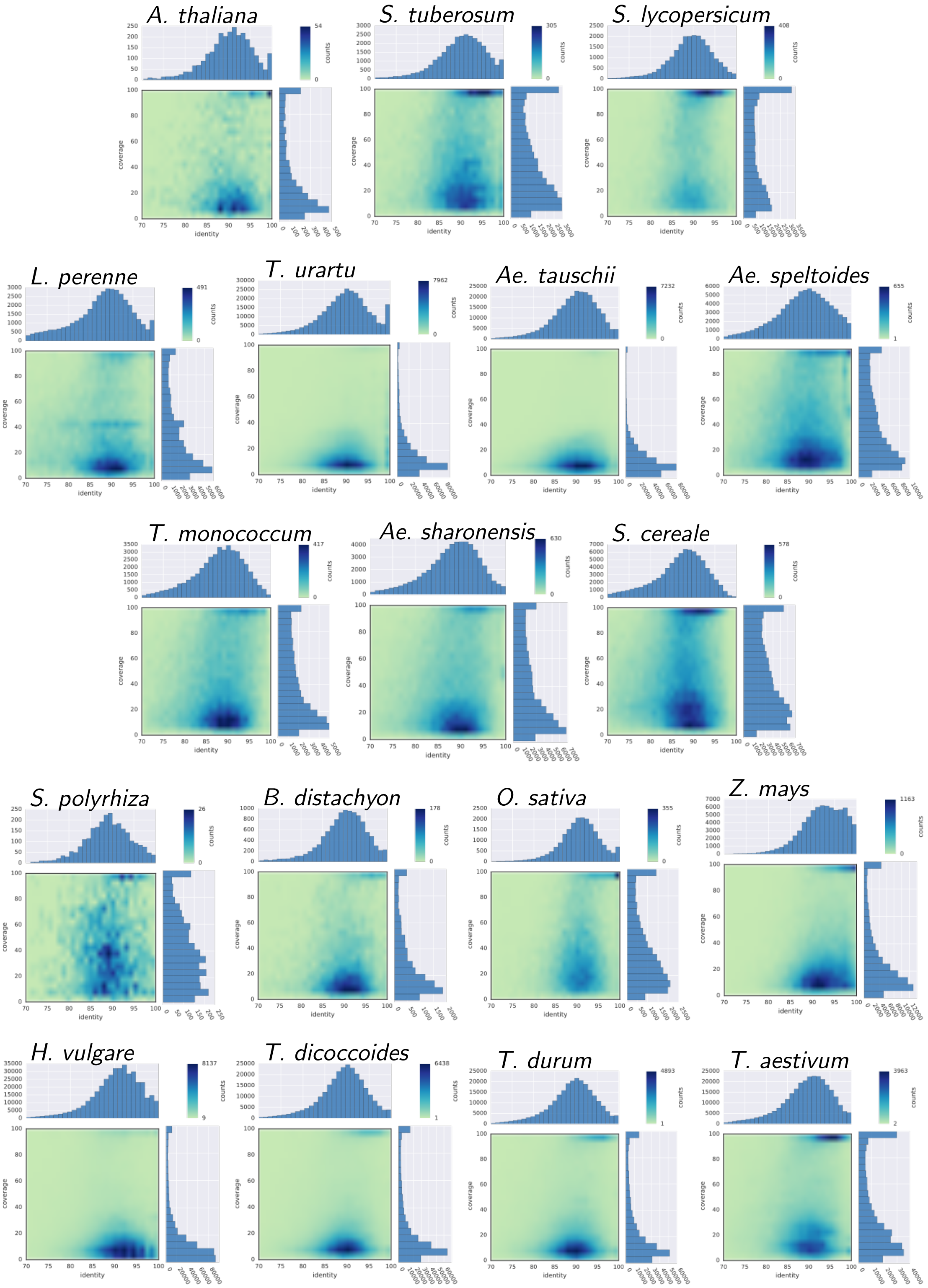


Figure R5: Sequence coverage vs. identity.

In plants, the average 3' UTR length is approximately 200 bp (Mignone et al., 2002; Mazumder et al., 2003). Within 300 bp downstream of HCov pseudogenes, the presence of adenine (A)-rich stretches was determined and their occurrences compared between pseudogene classes. Almost no pseudogene of any class and species features a poly-A region when requiring at least 80% As within a window of 50 bp. When reducing the window size to 10 bp, 60% to 98% of all HCov pseudogenes have an A-rich region downstream of their sequence — percentages that are much too large than would have been expected. Additionally, pseudogenes classified as retroposed do not exhibit poly-A regions more often than those classified as duplicated. Chimeric pseudogenes feature poly-A stretches most frequently in this setting.

Finally, a window size of 20 bp was chosen to determine the presence of putative poly-A tails (Table R4). Such an A-rich region was detected for only 1.8% of the HCov pseudogenes. However, approximately 1.5% of all pseudogenes classified as duplicated and 2.1% of all retroposed pseudogenes exhibit a poly-A stretch. While the larger portion of retroposed pseudogenes with poly-A tail is expected, it is still a very small percentage. This could either indicate, that (i) poly-A tails are not easily sequenced, (ii) most poly-A tails degenerated beyond recognition or (iii) the classification of retroposed pseudogenes generates false positives. Manual scrutiny of a large number of retroposed HCov pseudogenes did not indicate a classification error. Most show clear evidence for retrotransposition due to an absence of introns. However, also other processes may lead to an absence of introns.

Vanin et al. (1980) proposed a model in which mature mRNA or reverse transcribed cDNA participate in gene conversion events. This could lead to a partial or complete loss of introns. Such an intron loss event via reverse transcripts has been demonstrated in *Saccharomyces cerevisiae* (Derr, 1998). Furthermore, recurrent intron loss events have been reported for several grasses including maize, sorghum, rice and Brachypodium (H. Wang et al., 2014).

If most of the pseudogenes classified as retroposed really did lose their introns not via host gene retrotransposition, but subsequent to duplication, then they should not be dubbed “retroposed” but “processed”. While the two terms are often used as synonyms, a discrimination between the two seems advisable.

Alternatively, “processed” pseudogenes might not have lost introns at all. Instead, the parent genes may have experienced intron gain. However, H. Wang et al. (2014) also argue that intron gain is rarer than recurrent intron loss. Overall, the analysis of poly-A tails indicates that retrotransposition does not play a significant role in the generation of pseudogenes in plants.

5.5.1.5 Homology beyond the UTR sequences The identification of pseudogenes was done via homology to the CDS of template genes. The classification

Table R4: Percent HCov pseudogenes with upstream poly-A stretches.

		Species	Total		Duplicated		Retrop.		Chimeric		Single-exon		Fragmented	
			#	%	#	%	#	%	#	%	#	%	#	%
dicots		*A. thaliana*	31	8	9	7	0	0	0	0	22	10	0	0
		S. tuberosum	288	6	97	6	5	5	0	0	169	5	17	7
		S. lycopersicum	482	8	86	5	11	20	0	0	361	9	24	8
monocots	contigs/scaffolds	*L. perenne*	45	1	20	2	1	4	3	38	19	1	2	1
		T. urartu	116	2	73	2	1	2	0	0	38	1	4	1
		Ae. tauschii	103	2	27	1	0	0	0	–	71	4	5	1
		Ae. speltoides	103	1	30	1	0	0	0	0	68	1	5	1
		T. monococcum	37	1	13	1	0	0	0	0	21	0	3	0
		Ae. sharonensis	101	1	22	1	0	0	0	0	71	1	8	1
		S. cereale	118	1	24	1	1	1	1	8	86	1	6	1
	pseudomolecules	*S. polyrhiza*	13	5	9	6	1	33	1	100	2	2	0	0
		B. distachyon	17	2	3	1	0	0	0	0	14	3	0	0
		O. sativa	123	5	35	4	0	0	0	0	84	5	4	2
		Z. mays	144	2	29	1	1	2	0	0	108	2	6	2
		H. vulgare	303	1	39	2	0	0	0	0	261	1	3	1
		T. dicoccoides	387	2	139	1	3	1	0	0	230	2	15	1
		T. durum	397	2	147	1	4	2	1	3	235	2	10	1
		T. aestivum	818	2	270	1	3	1	1	2	479	2	65	3
		Average	3,626	**1.8**	1,072	**1.5**	31	**2.1**	7	**3.7**	2,339	**2.0**	177	**1.6**

into duplicated and processed pseudogenes was achieved by comparing their exon-intron structure to the respective parent gene. While the sequence homology of retroposed pseudogenes is expected to be limited to the CDS and the UTRs, the duplication via unequal crossing over, DNA repair processes, segmental, chromosomal or whole genome duplications can potentially also affect non-genic sequence. Hence, the extend of the homologous region between parent genes and pseudogenes can indicate whether a pseudogene is of duplicated or retroposed origin.

The average length of 5' UTRs and 3' UTRs in plants is ∼100 and ∼200 bp, respectively (Mignone et al., 2002; Mazumder et al., 2003). Homology of the regions adjacent to the CDSs of parent gene and HCov pseudogenes was investigated up to a distance of 1 kbp.

Duplicated and single-exon parent pseudogenes have the same pattern of down- and upstream homology: The homology for most pseudogenes stops either directly at the borders of the CDS, or extends to at least 1 kbp (Figure R6 A, B, E, F). For the latter group, this is a clear indication that they are indeed duplicated and not retroposed. For the former group, however, there are again several possibilities: The duplication could have stopped exactly at this position. This happening by chance for so many HCov pseudogenes is rather unlikely.

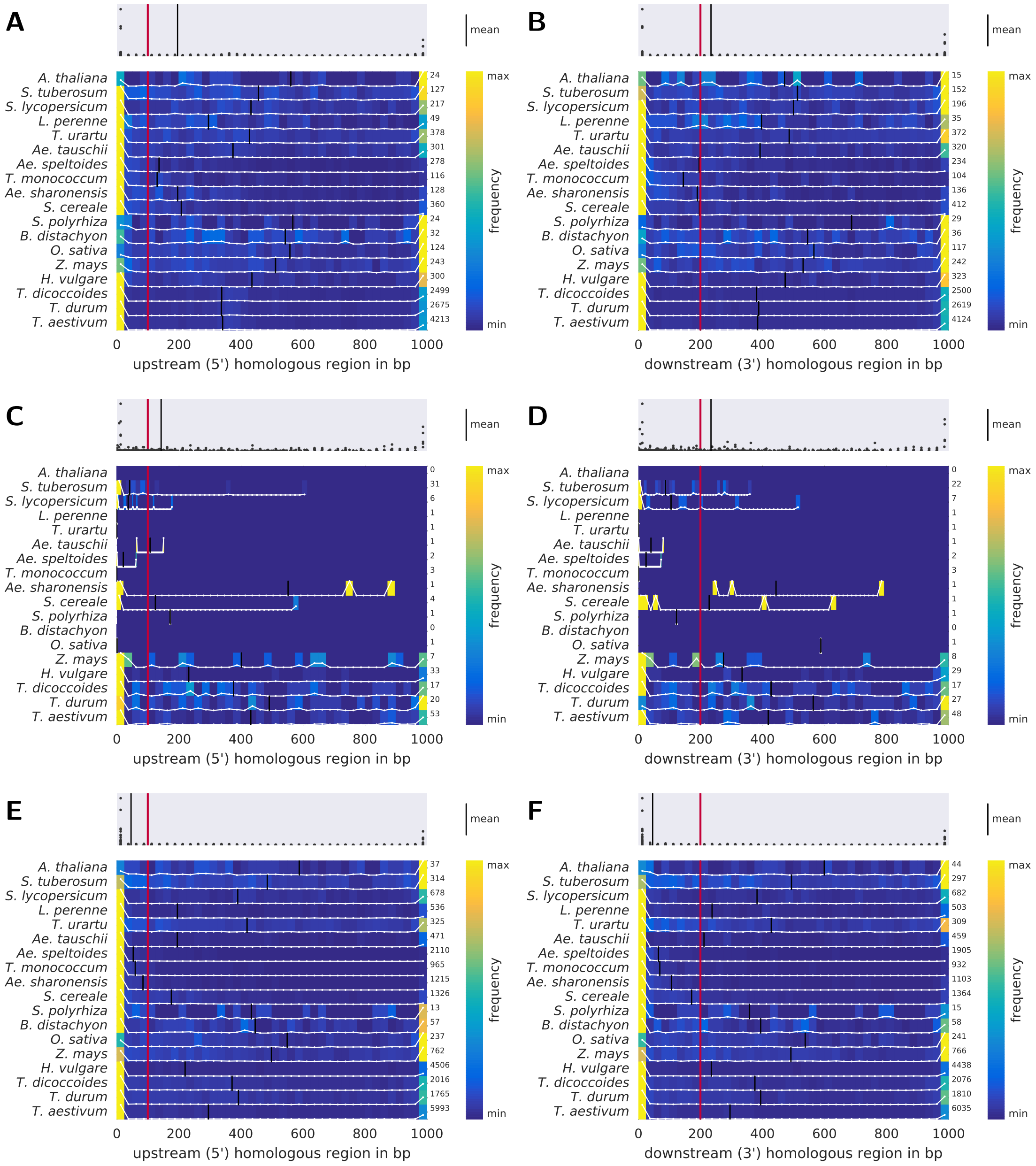


Figure R6: Extent of the homolgous region adjacent to HCov pseudogenes. A: upstream of duplicated pseudogenes; B: downstream of duplicated regions; C: upstream of processed pseudogenes; D: downstream of processed pseudogenes; E: upstream of single-exon parent pseudogenes; F: downstream of single-exon parent pseudogenes. The projected approximate UTR length is marked in red. The mean homologous region length is marked in black.

Alternatively, the gene duplicate could also have been under selection pressure initially, while the UTR diverged. Thus today, there is no significant homology between the UTR sequences. Additionally, if there are multiple genes with highly similar CDS, the wrong one could have been chosen as parent gene. While the CDS sequence might be highly similar, UTR sequences can differ. A comparison to the wrong parent gene UTR yields no homology.

Pseudogenes from single-exon parent genes show the same pattern of homology beyond the UTRs as duplicated pseudogenes, representing additional evidence that they did not originate from retrotransposition.

HCov pseudogenes classified as retroposed according to the intron-based approach express a different pattern of down- and upstream homology (Figure R6 C, D). Similar to the other subclasses, homology often stops at the border of the CDS. In few cases, it extends to 1 kbp. The mean homologous region beyond the CDS is closer to the expected UTR lengths, which is supporting evidence for a potential origin by retrotransposition.

However, maize, barley, wild emmer, durum wheat and bread wheat also harbor a significant amount of processed HCov pseudogenes that show homology up to 1 kbp. Together with the absence of poly-A tails for many pseudogenes classified as retroposed, this can be seen as further evidence, that many pseudogenes which have lost introns are of different origin. Many of them were duplicated via other mechanisms. Either they lost their introns after duplication, or their parent genes gained introns subsequently.

These results are particularly influenced by genome sequence quality, since contig/scaffold borders limit the search space of down- and upstream homologous regions for many genes or pseudogenes. Species with contig/scaffold assemblies thus show fewer pseudogenes with 1 kbp homologous regions.

5.5.1.6 Chimeric pseudogenes Pseudogenes that contain some introns but have lost others are called chimeric pseudogenes. Only ~1% of the annotated HCov pseudogenes are classified as chimeric (Table R3). It is possible that they have lost introns during duplication — for example due to incomplete splicing during retrotransposition — or after duplication. Vanin et al. (1980) proposed a deletion mechanism of introns via a gene conversion event between the pseudogene and the mRNA or cDNA cognate of the parent gene. An interaction with a smaller splice variant might explain a chimeric state. Furthermore, a partial gene conversion event between the duplicated pseudogene and another retroposed pseudogene might result in a chimeric pseudogene.

Chimeric pseudogenes often have qualities typical for retroposed pseudogenes. The probability to find a poly-A region downstream of their CDS is the highest compared to other pseudogene classes (Table R4). Additionally, the pattern of

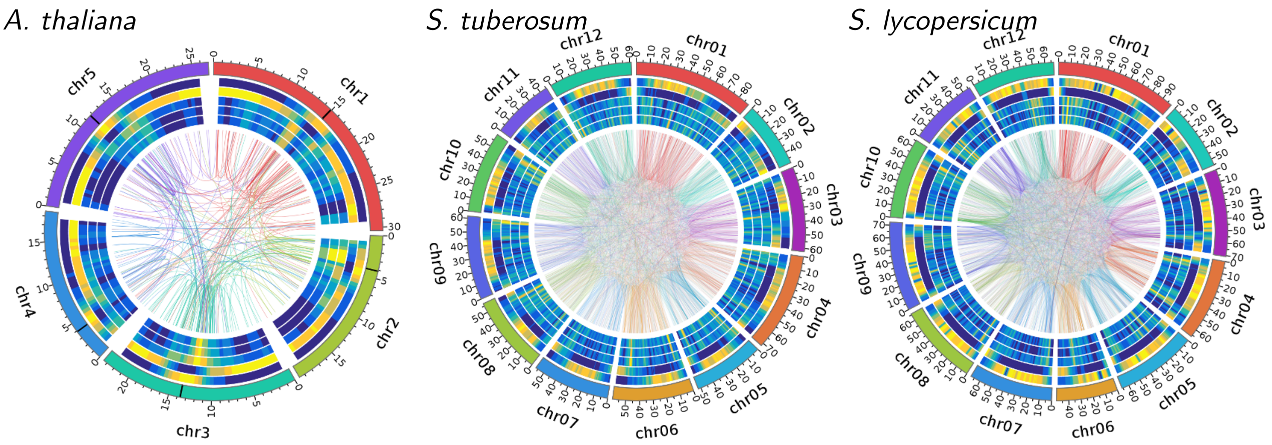


Figure R7: Chromosomal distribution of transposable elements, genes and pseudogenes on three dicots. Tracks from the outside show the (1) chromosomes, the distribution of (2) transposable elements, (3) genes, (4) pseudogenes, (5) HCov pseudogenes, and (6) HCov pseudogenes with PTCs. Links connect parent genes to their pseudogene children and are colored as the chromosome of the parent gene.

homology beyond the CDS is very similar to that of retroposed pseudogenes (Figure not shown). There may be some duplicated pseudogenes among them that have lost introns subsequent to duplication, but most probably originated in a retrotransposition event.

5.5.2 Positional features as evidence for origin

Duplicated pseudogenes arise through unequal crossing over, segmental duplications, chromosome or whole genome duplications or DNA break repair mechanisms. With frequent whole genome duplications in plants, they are prone to contain many gene duplicates and pseudogenes. The distribution and position of pseudogenes compared to their parent genes can shed light on the mechanisms generating gene duplicates.

5.5.2.1 Large-scale distribution of pseudogenes Pseudogenes are more equally distributed than protein-coding genes but they preferentially accumulate in gene-rich regions as opposed to TE-rich regions (Figure R7 and R8). A distance-correlated duplication mechanism like unequal-crossing over or certain DNA break repair mechanisms account for tandem duplications.

However, many pseudogenes are distributed randomly. They may represent retroposed pseudogenes that are randomly reinserted into the genome. However,

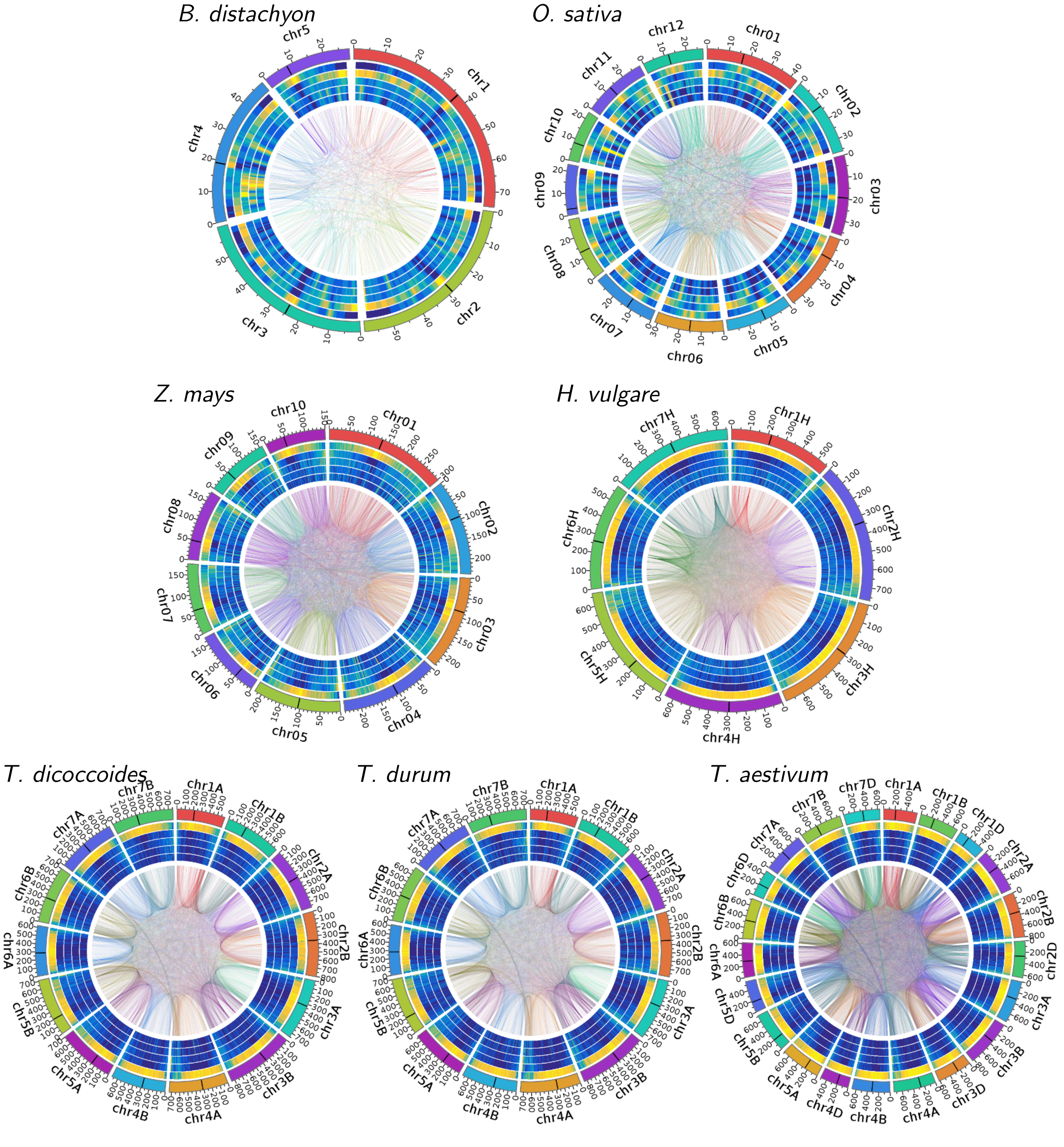


Figure R8: Chromosomal distribution of transposable elements, genes and pseudogenes in monocots. Tracks from the outside show the (1) chromosomes, the distribution of (2) transposable elements, (3) genes, (4) pseudogenes, (5) HCov pseudogenes, and (6) HCov pseudogenes with PTCs. Links connect parent genes to their pseudogene children and are colored as the chromosome of the parent gene.

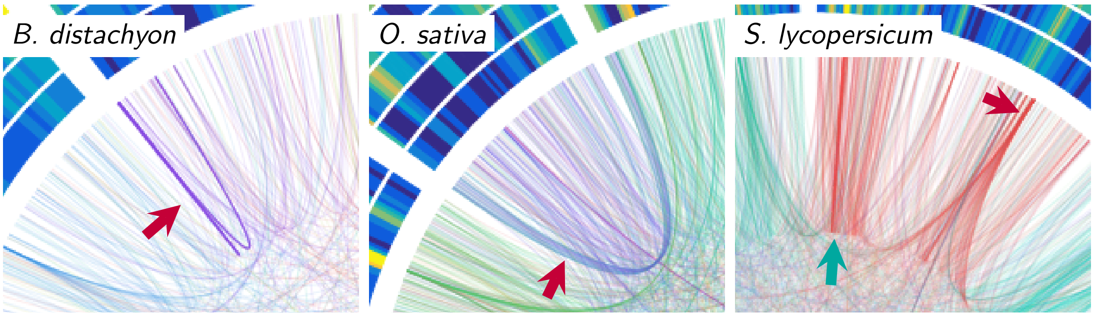


Figure R9: Selected chromosome regions from Figure R8. Chromosome 5 of *B. distachyon* contains two tandem pseudogene cluster with genes and pseudogenes located 4.6 Mbp apart. *O. sativa* contains a segmental duplication on chromosomes 11 and 12. On chromosome 1 of *S. lycopersicum*, tandem duplications are visible (turquoise) as well as a parent gene locus with numerous pseudogene children (red) distributed randomly on other chromosomes.

there is no evidence for a significant role of retrotransposition in the generation of pseudogenes and there is a massive amount of randomly distributed pseudogenes and gene fragments. Alternatively, the NHEJ/SDSA repair mechanism leads to filler DNA originating from loci exhibiting microhomologies. By chance, they may result in the duplication of genes or gene fragments. Lastly, all classes of TEs have been shown to capture and multiply host gene fragments. During the life cycle of LTR-retrotransposons, their transcript exits the nucleus. It is then translated and packaged into virus-like particles in which the reverse transcription is taking place. After reentry into the nucleus, the cDNA is reintegrated into the genome. Insertions into the original LTR-retrotransposon may include parts of a gene and are transcribed and duplicated along with the element itself. This “hitchhiking” mechanism affects both autonomous and non-autonomous TEs. Originally non-TE-related genes can be part of non-autonomous transposons that are copied due to the activity of autonomous TEs of the same family (Slotkin and Martienssen, 2007). One of the simplest non-autonomous TEs consists of a sequence with adjacent terminal inverted repeats (TIRs). If this enclosed sequence contains a gene, its duplicate would retain the original exon-intron structure.

Helitrons have been shown to frequently capture gene fragments (Barbaglia et al., 2012). At least 2% of the genome of maize is made up of *Helitrons* (Du et al., 2009). Approximately 2,800 putative non-autonomous *Helitrons* were predicted and 94% of them were found to carry up to nine gene fragments (Du et al., 2009). Transcription of *Helitron* genes can be accompanied by read-through and conjoined exons from neighboring genes, resulting in transcripts with potential new functions (Barbaglia et al., 2012).

5.5.2.2 Parent genes with numerous pseudogenes While parent genes on average only have 1.3 to 7.9 HCov pseudogene children, almost each of the genomes contains a couple of genes with a particularly high number of pseudogene copies (Table R5). Some of them may be TE genes that were not filtered from the gene nor the pseudogene annotation. Many pseudogene children are distributed randomly. However, a significant amount is located close to the parent gene. For example, chromosome 1 of tomato displays both tandem situations and parent genes with numerous pseudogenes distributed randomly on other chromosomes (Figure R9).

Two particularly prominent pseudogene clusters on chromosome 5 of *Brachypodium distachyon* contain 68 copies of four uncharacterized or unknown proteins with sequence similarities to partial ribosomal RNA genes (Figure R9). Three of the four parent genes are located in the first tandem cluster and the fourth is in a second cluster located 4.6 Mbp distant from the first. Both cluster loci contain pseudogenes from all four parent genes but the second is much smaller containing only six pseudogenes. Pairwise sequence alignments of all four genes unexpectedly revealed them to have a sequence similarity of only 23 to 32%. Thus, since almost all (66/68) of the pseudogenes have a coverage of 100%, the four parent genes cannot be considered “alternative” parents – a group of similar genes where each could be considered the parent of the same pseudogene. Ribosomal RNA (rRNA) genes are transcribed into non-coding RNA (ncRNA) — a component of the ribosome. The universally conserved elements occur in high copy number. In human, it has been shown that rDNA contributes to the production of “junk DNA” and that the resulting pseudogenes might be a by-product of concerted evolution via unequal crossing over (Robicheau et al., 2017). However, since rRNA genes are non-coding, they should not have been in the template gene set of protein-coding genes and pseudogenes of non-coding genes would usually not have been detected with the PLIPipeline.

One of the parent genes with numerous pseudogenes scattered randomly over the genome is located on chromosome 1 of tomato (Figure R9). The AT-rich gene (72%) codes for an unknown protein and has 136 pseudogene descendants. A NCBI BLAST search found hits only on the genome sequences of tomato species without any functional indication. Almost all of the pseudogenes from this putative gene contain PTCs. Most of the genes with numerous pseudogenes distributed in this manner are probably either TE-related or wrong gene calls. However, some might prove to be real and a closer investigation could be insightful.

5.5.2.3 Segmental duplications Due to genomic redundancy, segmental duplications are prone to contain an abundance of pseudogenes. Rice experienced an ancient whole genome duplication between 53 and 94 million years ago (mya)

Table R5: Parent gene metrics

		Species	Query	Parent	%	HCov	%	avg. children	
			genes	genes		parent		all	HCov
dicots		*A. thaliana*	27,179	2,033	7.5	296	1.1	1.4	1.3
		S. tuberosum	38,546	9,389	24.4	2,350	6.1	3.4	2.2
		S. lycopersicum	34,011	6,229	18.3	1,675	4.9	3.7	3.6
monocots	contigs/scaffolds	*L. perenne*	23,583	9,154	38.8	1,417	6.0	4.5	3.0
		T. urartu	33,005	15,172	46.0	2,759	8.4	19.7	2.6
		Ae. tauschii	34,916	14,701	42.1	1,633	4.7	18.4	2.8
		Ae. speltoides	41,695	20,412	49.0	4,912	11.8	4.1	2.3
		T. monococcum	30,463	13,266	43.5	3,070	10.1	3.3	2.0
		Ae. sharonensis	32,453	14,713	45.3	3,532	10.9	3.9	2.2
		S. cereale	29,415	15,171	51.6	3,987	13.6	5.3	3.3
	pseudomolecules	*S. polyrhiza*	19,345	1,164	6.0	147	0.8	2.3	1.9
		B. distachyon	25,994	4,556	17.5	511	2.0	2.4	1.9
		O. sativa	38,692	8,129	21.0	1,501	3.9	2.6	1.8
		Z. mays	38,649	16,138	41.8	2,732	7.1	4.7	2.6
		H. vulgare	39,734	18,299	46.1	2,805	7.1	21.7	7.9
		T. dicoccoides	67,181	28,019	41.7	7,595	11.3	9.5	3.3
		T. durum	66,558	28,275	42.5	7,987	12.0	8.7	3.2
		T. aestivum	110,790	40,483	36.5	12,097	10.9	7.1	4.0

and a very recent segmental duplication between 5 and 21 mya (J. Yu et al., 2005; X. Wang et al., 2005). The distribution of pseudogenes in combination with their parent genes helps to identify segmental duplications and to determine putative unilateral gene loss. Additionally, it can help to date pseudogenes since whole genome or segmental duplication time estimates are often published already. In rice, the pseudogene data does not provide evidence for the ancient whole genome duplication. However, the more recent segmental duplication is clearly visible (Figure R9 and R10 A). Higher pseudogene age and the accompanied gradual accumulation of mutations renders older pseudogenes unidentifiable via homology searches, which might explain why remains of ancient duplication events cannot be detected via pseudogenes.

In potato, several segmental duplications are clearly detectable (Figure R10 B). However, it was not possible to detect a symmetric loss of genes. At first glance, a duplicated segment on chromosome 3 experienced gene loss and contains numerous pseudogenes, while the genes in the corresponding region on chromosome 6 retained their function. A closer look reveals that the duplicated segment on chromosome 6 contains numerous pseudogenes, too. Other clearly discernible segmental duplications in potato are between chromosomes 2 and 3, and between chromosomes 9 and 10. The solanaceous lineage experienced multiple genome duplications and triplications — one whole genome duplication likely preceded the divergence from grape over 100 mya, while other duplication events likely occurred 52 to 91 mya (The Tomato Genome Consortium, 2012). However, it is unlikely that the genes within the duplicated segments pseudogenized over 50 mya ago and remained non-functional while still resembling functional genes.

Several segmental duplications are easily identified and visible in maize (Figure R10 C). The ancient allotetraploid returned to a diploid state, but still contains numerous duplicated regions. Approximately 70 mya, maize had a paleopolyploid ancestor followed by an additional whole genome duplication event 5 to 12 mya. Within the last ∼3 million years, the maize genome expanded due to LTR-retrotransposon activity (Schnable et al., 2009). The pseudogene rich duplicated segments that are still visible today likely originate from the most recent duplication event.

5.5.2.4 Polyploidy In polyploid species containing several subgenomes, almost the complete gene set is duplicated. Gene function can be maintained due to dosage-effects or — especially in plants — their expression can be regulated via dosage-compensation (Edger and Pires, 2009; Heslop-Harrison and Schwarzacher, 2011). However, numerous duplicates will likely degenerate due to the redundancy of information (Prince and Pickett, 2002).

Diploid barley contains ∼40k genes (Table R1). In comparison, tetraploid wild emmer and durum wheat do not contain twice as many genes, but only ∼67k (1.7×) genes. This could already indicate gene loss and potential pseudogenization also caused by functional redundancy. However, barley is no subgenome progenitor of *Triticum turgidum*. It was used for the purpose of comparison, due to its high-quality reference genome sequence and gene annotation. Hexaploid bread wheat contains ∼111k genes, which is remarkably close to the tetraploid gene number of *Triticum turgidum* plus the diploid gene number of barley. This could indicate that there was no accelerated pseudogenization since the last polyploidization event ∼10,000 years ago.

In polyploid *Triticeae*, the subgenomes harbor many pseudogenes in homeologous context (Figure R10 D, E, F). There is no indication for unilateral gene loss — the retention of functional genes in one specific subgenome. The tetraploid *Triticum turgidum* subspecies only have a marginally higher number of HCov pseudogenes compared to barley. However, bread wheat almost harbors twice the number of HCov pseudogenes compared to the *Triticum turgidum* subspecies. This, in turn, indicates that there did happen extensive pseudogenization since the last pseudogenization event. Additionally, plotting parent gene positions against pseudogene positions does not indicate that the D genome contains fewer pseudo-

Table R6: Significant accumulation of pseudogenes on the same chromosome as the parent gene. Checkmarks indicate weather a significant result was obtained in a paired T-test (P-value $\leq 5\%$). FDR-correction was applied. Orange checkmarks indicate a P-value between 2 and 5%.

	Species	all	HCov	duplicated	duplicated	retroposed
					HCov	
dicots	*A. thaliana*	✓	✓	✓	✓	X*
	S. tuberosum	✓	✓	✓	✓	✓
	S. lycopersicum	✓	✓	✓	✓	✓
monocots	*S. polyrhiza*	✓	✓	✓	✓	X*
	B. distachyon	✓	✓	✓	✓	X*
	O. sativa	✓	✓	✓	✓	X*
	Z. mays	✓	✓	✓	✓	✓
	H. vulgare	X	✓	✓	✓	X
	T. dicoccoides	✓	✓	✓	✓	✓
	T. durum	✓	✓	✓	✓	✓
	T. aestivum	✓	✓	✓	✓	✓

*calculation are based on less than one hundred pseudogenes

genes. However, a more detailed analysis of the effects of polyploidization will be provided in the subsequent sections.

5.5.2.5 Tandem gene and pseudogene clusters The distribution of pseudogenes and parent genes can help to determine prevalent gene duplication mechanisms. While unequal crossing over leads to tandem duplications, retrotransposition is expected to result in a random insertion of processed gene duplicates. We know from the intron-based classification that duplicated pseudogenes outnumber retroposed pseudogenes by far. We also saw evidence for duplicated pseudogenes within segmental duplications or within subgenomes. However, how many are the result of tandem duplication events via unequal crossing over or DNA break repair mechanisms?

For most plants — and in-spite of segmental duplications — there is a significantly higher chance for the pseudogenes to be located on the same chromosome as the respective parent gene (Table R6). For the most part, this is due to duplicated pseudogenes within tandem situations that originated mostly via unequal crossing over: A large portion of pseudogenes is located in direct vicinity of the parent gene (Figure R11). As expected, duplicated HCov pseudogenes are significantly accumulated in close vicinity to their parent genes.

Low-coverage pseudogenes and gene fragments in barley are not significantly enriched on the same chromosome as the parent genes (Table R6). This is due to the high contamination of the gene annotation with TE genes. TE genes are

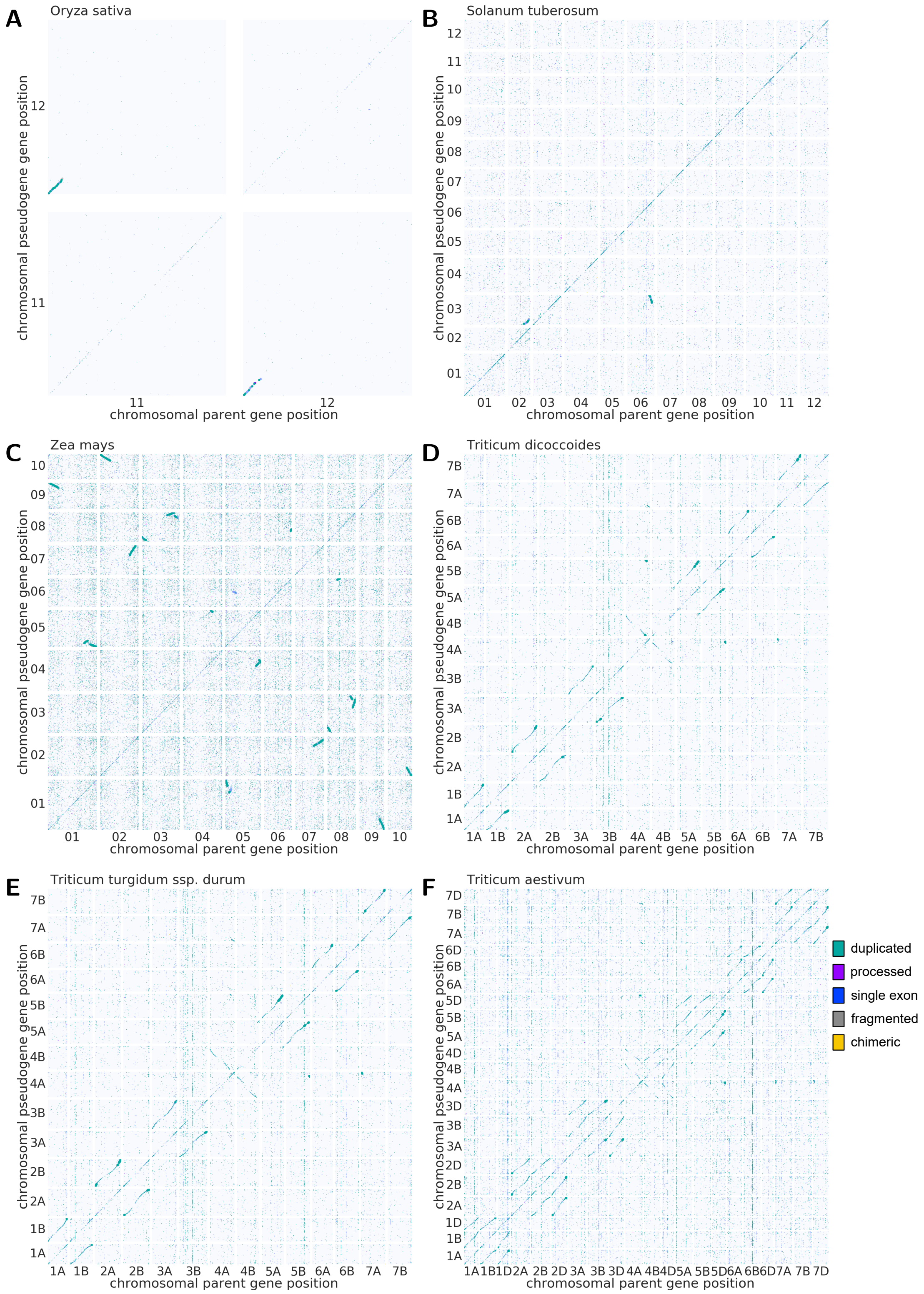


Figure R10: Segmental duplications containing parent and pseudogenes. Duplicated segments were determined using the DAGchainer on parent and pseudogene positions. Segments are accentuated with larger marker sizes. For *O. sativa*, only chromosomes 11 and 12 are shown. For the three *Triticeae* only HCov pseudogenes are plotted.

parents to numerous pseudogene children that are distributed randomly on the genome. In a previous pseudogene annotation, all pseudogenes that occur in very high copy number were filtered based on the assumption that they are TE-related. In this previous annotation, the significant accumulation of barley pseudogenes on the same chromosome as the parent gene was comparable to results in wild emmer, durum wheat and bread wheat. The filtering by pseudogene number per parent gene was replace by alternative filtering procedures involving overlap-checks with existing TE annotations or matches to the TREP database. This was done to avoid the arbitrary choice of a maximum pseudogene number per parent gene for each plant, but it has also led to higher amount of TE-related pseudogenes.

Processed pseudogenes are duplicated via an mRNA intermediate and reverse transcription at the integration site. Previously, they were assumed to be randomly distributed. However, they are often preferentially located close to their parent gene — especially in *Triticeae* genomes (Table R6). This could either be due to misclassification or due the particular Interphase chromosome conformation (Rabl) that has been identified for *Triticeae* (International Barley Sequencing Consortium, 2017). A more detailed analysis of the relationship between processed pseudogenes and Interphase chromosome conformation will be given in section 5.14.

Unequal crossing over occurs at regions featuring microhomologies or larger duplicated sequences (Schiml et al., 2016). The initial homologies may be the result of DNA break repair mechanisms. Once a tandem gene cluster is present, it is an even more likely target for future unequal crossing over events. Thus, tandem gene clusters are prone to gene duplication and pseudogenization.

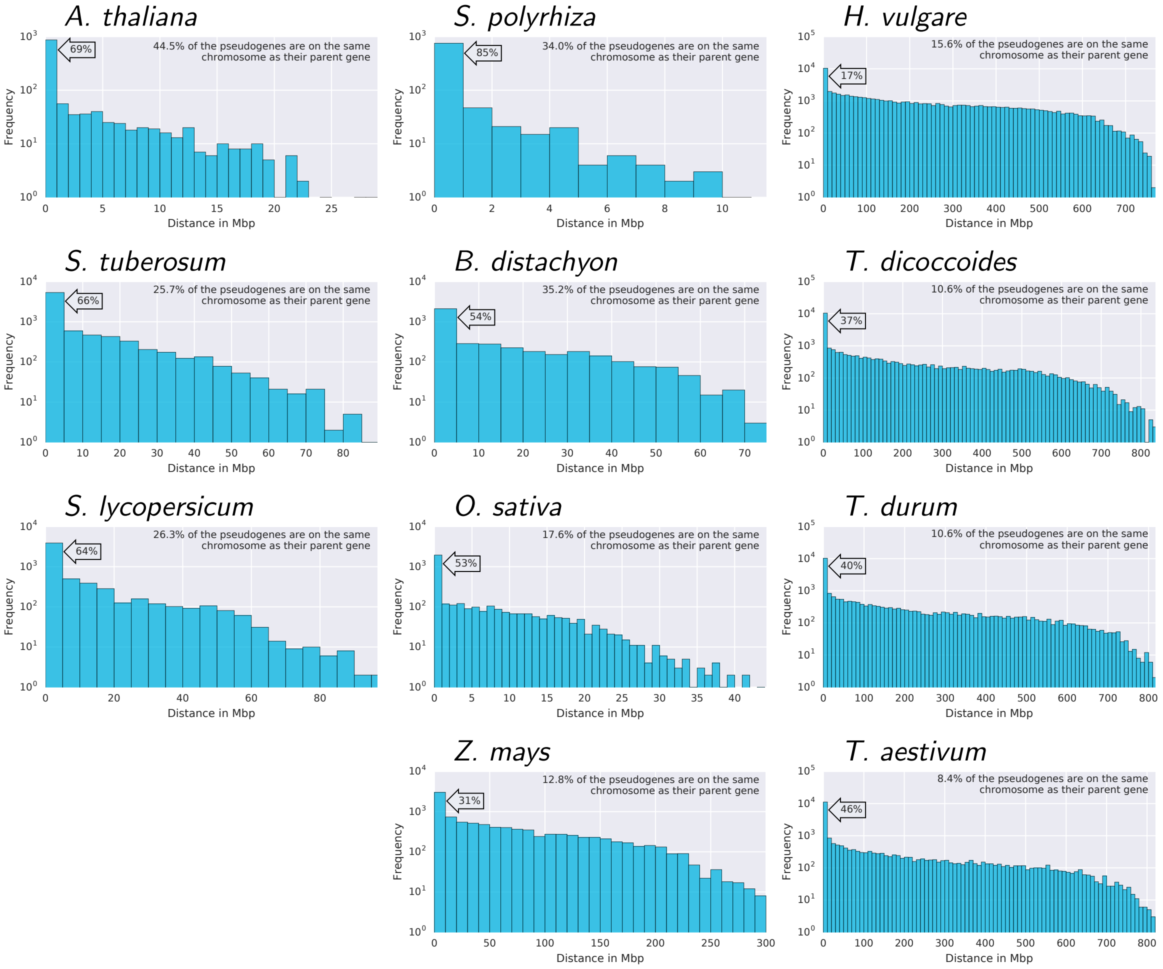


Figure R11: Distribution of parent-pseudogene distances

5.5.2.6 GC content in genes and pseudogenes The GC content of the genome has been shown to be correlated with recombination and gene conversion (Liu et al., 2015). Additionally, in some grasses, genes with higher GC content at the third codon positions (GC3) have been reported to (i) provide more methylation targets, (ii) exhibit more variable expression, (iii) more frequently possess upstream TATAA boxes, (iv) are predominant in certain classes of genes and (v) the GC3 content position increases from 5' to 3' (Tatarinova et al., 2010). The latter statement is true for genes with high GC3 content, but not for those with lower GC3 content. Overall, there is a sharp decrease of the GC3 content gradient from 5' to 3' along most genes in grasses (Glémin et al., 2014).

Furthermore, codon bias influences expression patters due to differences in the abundance of transfer RNAs (tRNAs) (Quax et al., 2015). If GC-rich tRNAs are more frequent, the expression of GC-rich genes will be faster than the expression of AT-rich genes. If higher expression levels are beneficial, gene duplication may be as well. Hence, the GC content of genes and pseudogenes was compared to assert a possible duplication and pseudogenization bias due to GC content, which might occur due to the correlation to recombination and gene conversion rates.

First, it was determined whether low-coverage pseudogenes originate preferentially from the 5' end or from the 3' end of the parent gene (Figure R12). If there was a preferential duplication, this could affect the average GC content of pseudogenes compared to parent genes, because there is a slight decrease of GC content from 5' to 3' of the CDS of genes.

Overall, there is no significant preference as to which region of a gene is duplicated in a pseudogene. However, slight differences can be observed and there seems to be a different duplication pattern between plants. In barley, wild emmer and durum wheat, the CDS close to the 5' or 3' end are duplicated slightly more often than CDS from the central part of the genes. This cannot be observed for bread wheat. Additionally, the three subgenomes of bread wheat show a similar pattern (Figures not shown). Hence, it is not the additional D subgenome, that contributes to the different result. However, the bread wheat gene annotation was manually curated compared to gene annotation of wild emmer and durum wheat. Hence, gene annotation differences might influence this analysis.

The GC content distribution along the CDSs of most investigated plants is bimodal between 40% and 75% (Figure R13 A). The mode is at ∼45%, but another smaller peak is often found at ∼70%. The bimodal distribution of the genic GC content is well-known for grasses (Clement et al., 2014). In mammals, a higher GC content has been shown to be positively correlated with expression level (Kudla et al., 2006).

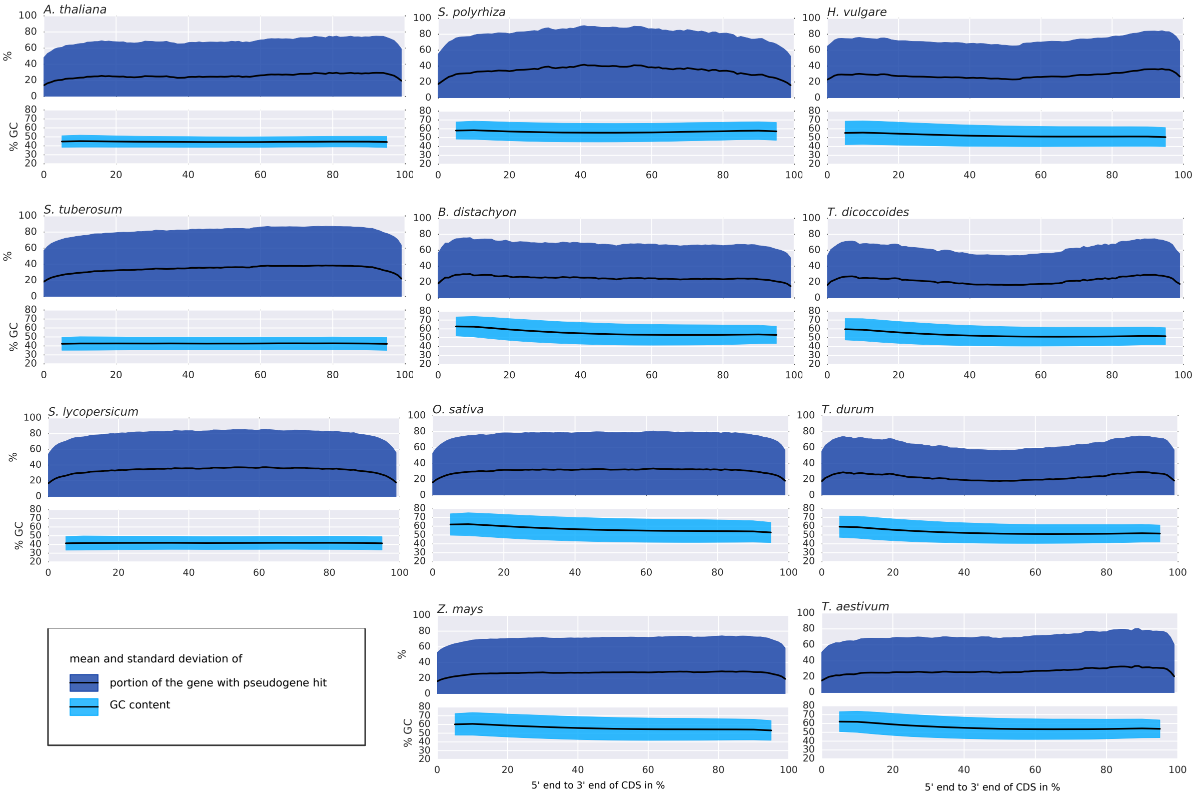


Figure R12: Duplicated regions and GC content along CDS of genes. The upper figure parts (dark blue) show which region of the parent gene's CDS is preferentially duplicated in low-coverage pseudogenes. The lower parts shows the GC content along the CDS of genes. The GC content is calculated using a sliding window (20 windows with a width of 1/10 of the CDS length).

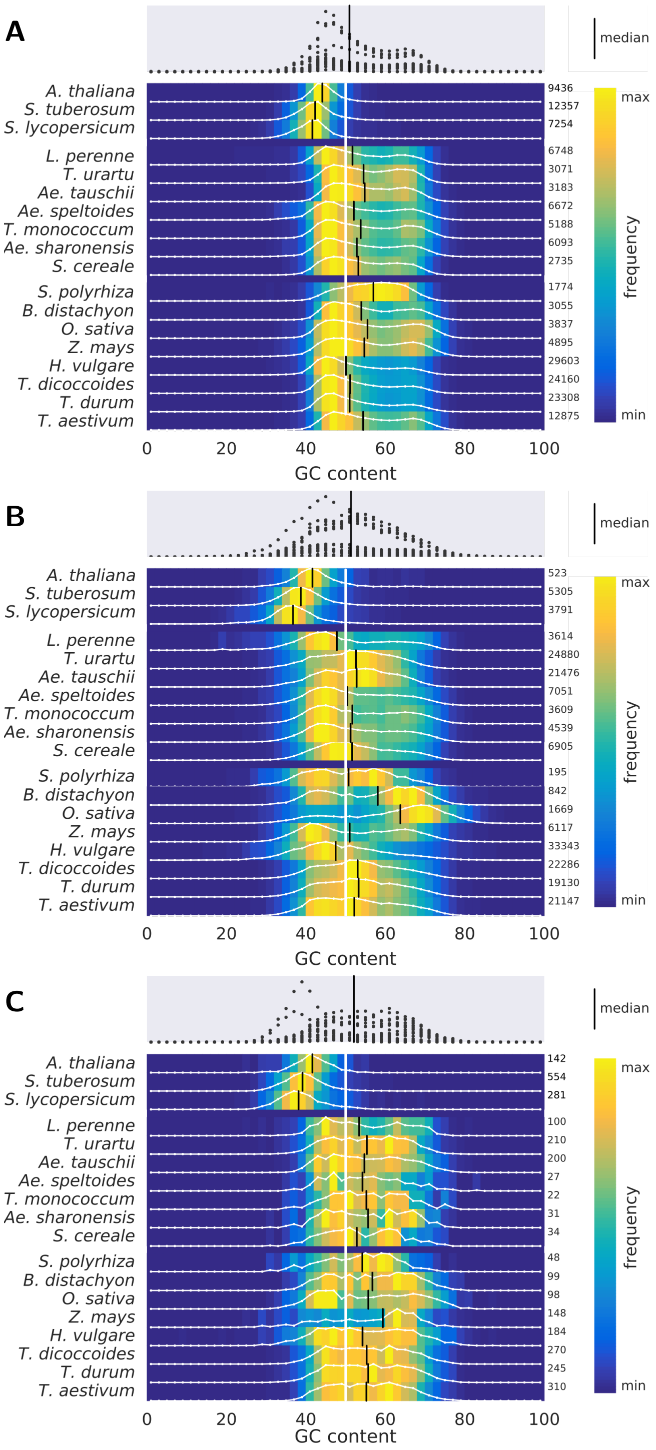


Figure R13: GC content of genes (A), pseudogenes (B) and pseudogenes in tandem situations (C).

Dicots display a completely different GC content distribution with only one peak at ∼40% and without much variance in the distribution. Compared to monocots, the genomes of dicots have been shown to be GC-poor (Z. Zhao et al., 2014).

Width and pattern of the GC content distributions for pseudogenes differ from those for genes (Figure R13 B). The most extreme differences can be observed for *Brachypodium distachyon*, rice, barley and the polyploid *Triticeae* wild emmer and durum wheat. Pseudogenes from these species are clearly GC-richer than genes. For most monocots, pseudogenes that are located close to their parent gene (e.g. in tandem) display an even stronger tendency towards a higher GC content (Figure R13 C).

The correlation between recombination rate and GC content may be a result of "biased repair of heteroduplex mismatches [...] favoring GC residues" (Liu et al., 2015). Since recombination rate and rate of unequal crossing-over can be assumed to covary, tandem duplication may preferentially affect GC-rich regions as well. This would indicate, that from the two potential mechanisms generating tandem duplications — unequal crossing over and paired single-strand break repair — the majority of tandem duplicated pseudogenes originated from unequal crossing over.

5.6 Gene family size correlates with pseudogene number

A correlation between gene family size and pseudogene number has been reported previously (Zou et al., 2009). However, not only does the pseudogene number increase with gene family size, but also the relative number of pseudogenes is larger for gene families with numerous members (Figure R14). Interestingly, most genes of diploid plants are singletons and from them only few pseudogenes originate. Up to a gene family size of 10 members the pseudogene number increases steeply before reaching a plateau. Gene families larger than 20 are more rare and the mean pseudogene content is highly variable.

5.6.1 Gene families in polyploid species

With more than one subgenome and diploid gene set, polyploid species often contain larger gene families. Nevertheless, the most prevalent gene family size usually matches the ploidy level. Just like singletons in diploid species, they give rise to fewer pseudogenes than larger families. However in comparison, real singletons or two-element families in tetraploid or hexaploid plants have multiple pseudogene children. Those singleton genes were originally duplicated or triplicated during polyploidization. Without a dosage effect or compensation taking place, the redundancy of information thus likely led to the pseudogenization of duplicates.

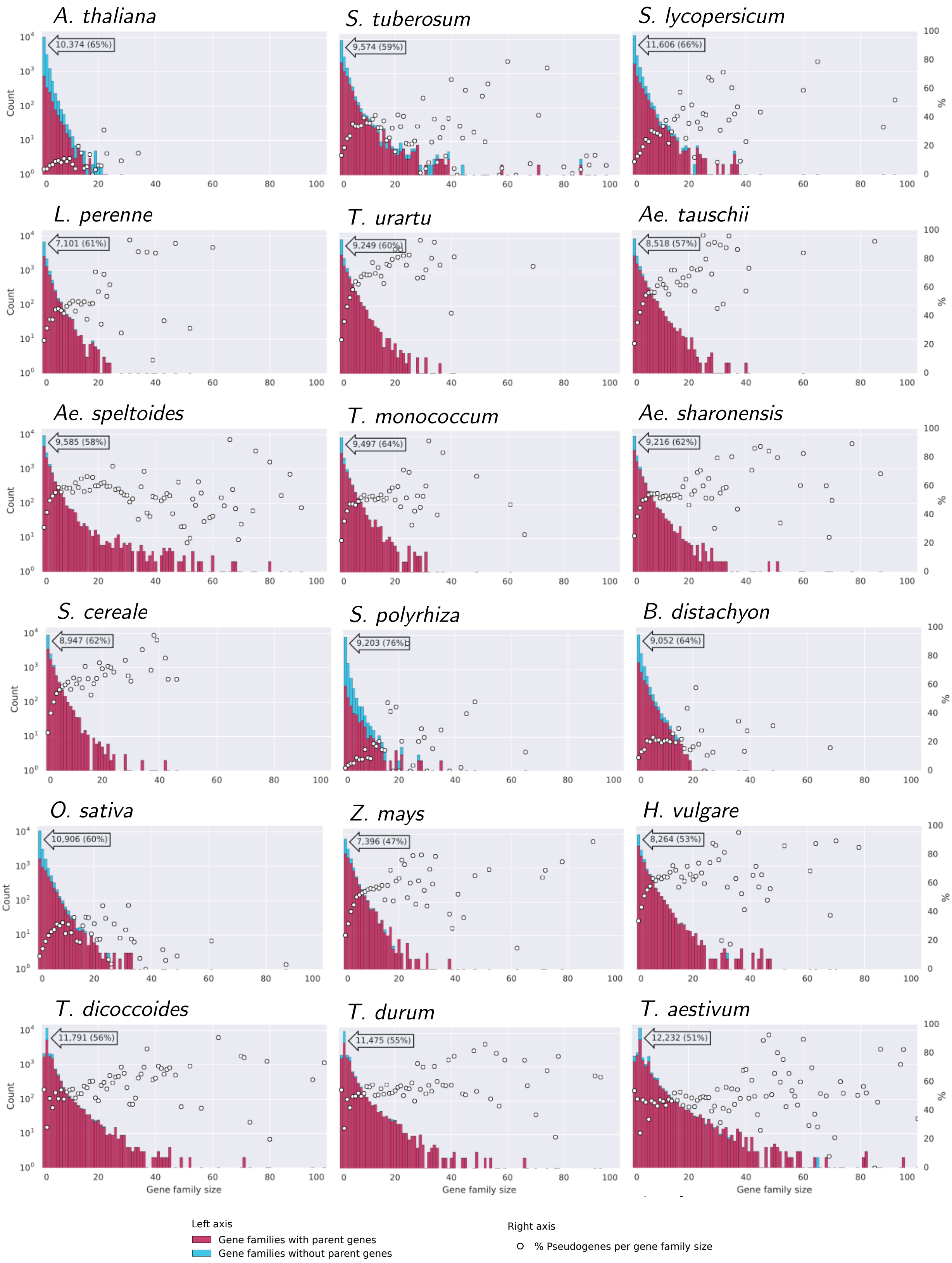


Figure R14: Gene families and pseudogenes. Gene family size vs. HCov pseudogene number. The histogram shows frequencies of gene family sizes with and without parent gene members (left axis). The dot plot shows the HCov pseudogene content in gene families plus their pseudogenes (right axis). The x-axis is limited to a size of 100.

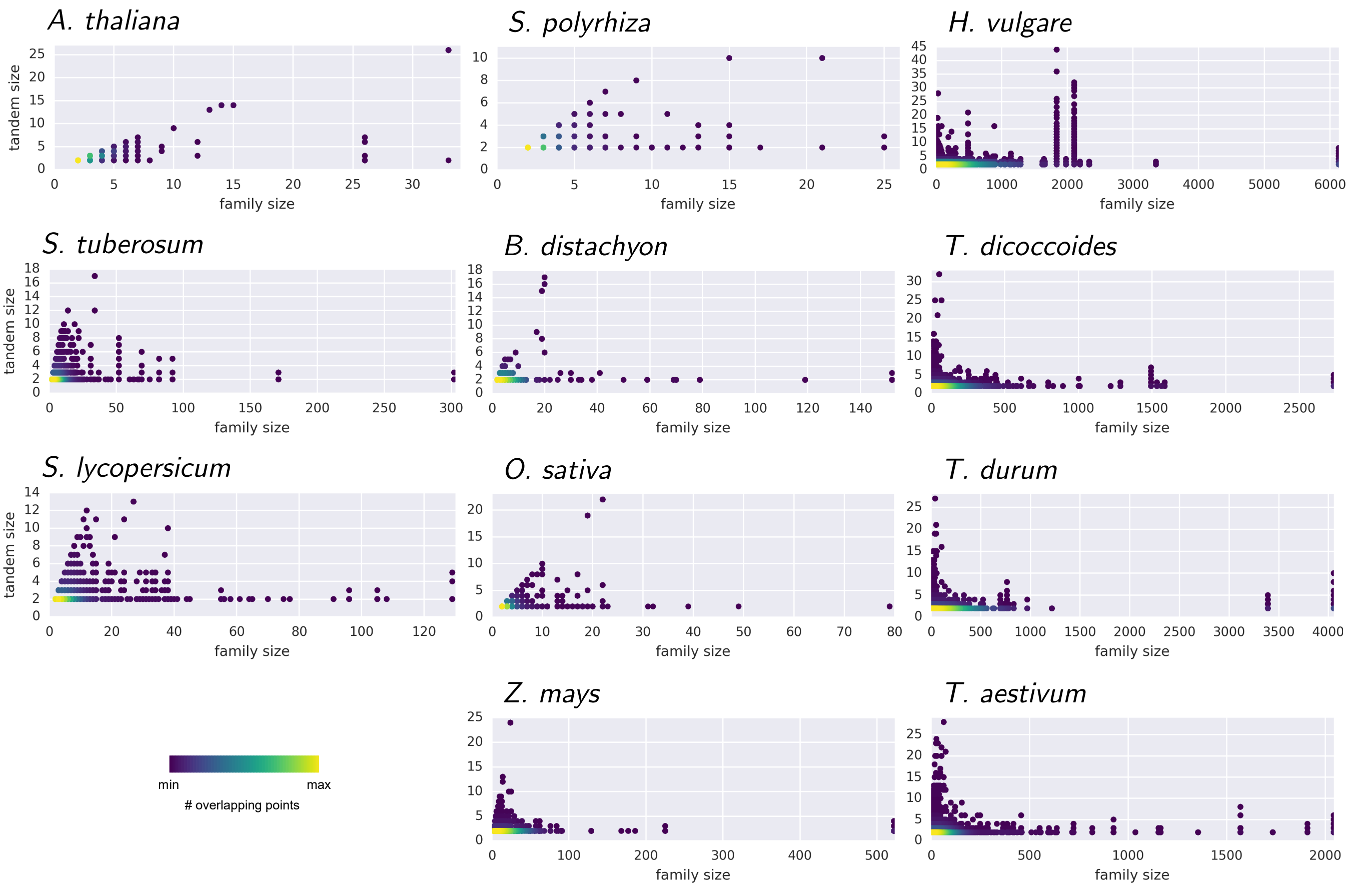


Figure R15: Tandem gene and pseudogene cluster (≥2) vs. family size.

5.6.2 Tandemly duplicated genes and pseudogenes

If larger gene families are predominantly appearing in tandem, unequal crossing over might lead to further expansion and to the symptomatic generation of pseudogenes. Expansion could allow gene families to evolve more rapidly due to increased divergence. In all target plants, larger families often occur in multiple but smaller tandem clusters (Figure R15). Tandem compositions of two genes or pseudogenes are most prevalent. Hence, unequal crossing over or DNA break repair processes do indeed drive pseudogene generation. However, other mechanisms leading to a more random distribution contribute significantly, too. Larger or fast evolving gene families might preferentially occur in tandem, because increased duplication is positively selected for them. Pseudogenes then can evolve and adopt new functions upon reactivation.

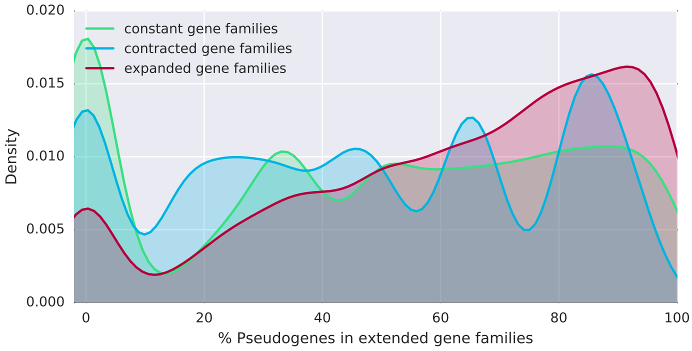


Figure R16: Pseudogene content in expanded, contracted or constant gene families in barley compared with *A. thaliana*, *B. distachyon*, rice and sorghum. Orthologous groups with a minimum size of two were used for this analysis.

On the other hand, it has been shown that tandem gene clusters tend not to diverge quickly due to gene conversion effects (Baumgarten et al., 2003). Sequence homogenization acts less on translocated or physically unlinked chromosomes or chromosomal regions and the evolution of new function might be reserved to pseudogenes duplicated via other mechanisms than unequal crossing over.

5.6.3 Gene family expansion and contraction

Gene family expansion gives the opportunity to evolve new functions from existing gene-like sequences but also has a compensatory effect of sporadic pseudogenization events. If a higher gene copy-number is beneficial, the expansion of gene families is positively selected. A burst of pseudogenes might be a side-effect of gene family expansion. On the other hand, gene family contraction is the result of gene loss due to deletion or degeneration. In this scenario — and if the family contraction happened recently — an increased number of pseudogenes could be expected as well. Pseudogenes of older contracted gene families are expected to exhibit advanced degeneration or deletion.

To study the effect of gene family expansion and contraction on pseudogenes, orthologous groups of barley were compared with *A. thaliana*, *B. distachyon*, rice and *Sorghum*. Barley contains over ten thousand (>25%) more genes than *Arabidopsis thaliana* or *Brachypodium distachyon*. Additionally, it contains 8× more HCov pseudogenes than rice and almost 60× more than *Arabidopsis thaliana*. Hence, it is not surprising that barley contains 1,954 expanded but only 117 contracted orthologous groups (International Barley Sequencing Consortium, 2017). Expanded orthologous groups clearly give rise to numerous pseudogenes (Figure R16). However, contracted orthologous groups or those exhibiting no significant change in numbers show no clear evidence for increased nor decreased pseudogene generation. Gene family contraction may not happen via pseudogenization, but via deletion and possibly via unequal crossing over.

5.7 Functional analysis of pseudogene parents

Pseudogene creation could be a symptom of gene birth and evolution. If fast evolving genes give birth to more pseudogenes than highly conserved genes, then they may share features that promote higher duplication rates. Independent of pseudogene structure or location, the functional annotation of parent genes is of interest in order to determine preferential duplication. Gene Ontology (GO) terms were available for most gene annotations and an enrichment analysis was performed to compare the parent gene set to the complete template gene set of each plant, respectively.

For visualization and interpretation purposes, only GO terms significantly enriched or depleted in at least 9 of 18 plants are depicted. This may introduce a bias towards commonalities. GO terms enriched only in specific plants are of high interest as well and more detailed GO enrichment analyses are performed for pairwise comparisons in subsequent sections.

The pattern of enriched and depleted GO terms is very similar for all target plants (Figure R17). For HCov pseudogenes, parent genes involved in translation, defense response, electron transport and photosynthesis are enriched. Those involved in biological regulation, transport and localization are depleted.

Defense response is a dynamic process that has to adapt quickly to environmental changes and threats. A high gene duplication rate provides the basis for gene evolution and also leads to the generation of pseudogenes. Defense response related genes are enriched in most parent gene sets of HCov pseudogenes. If considering not only HCov pseudogenes, but also smaller gene fragments, defense-related parent genes are only under-represented in barley (Figure R17 A). Additionally, there seems to be no significant difference between domesticated and wild plant species — as was initially suspected due to results from a previous PLIPipeline version.

Overall, the results deviate for *Ae. speltoides* when focusing on under-represented GO terms in the parent gene set off all pseudogenes including smaller gene fragments. However, this might also be an effect of assembly and gene annotation quality. Closely related species like wild emmer and durum wheat show a very similar pattern, even though one is cultivated and one is wild. Differences due to domestication may only appear without filtering for shared GO enrichment results.

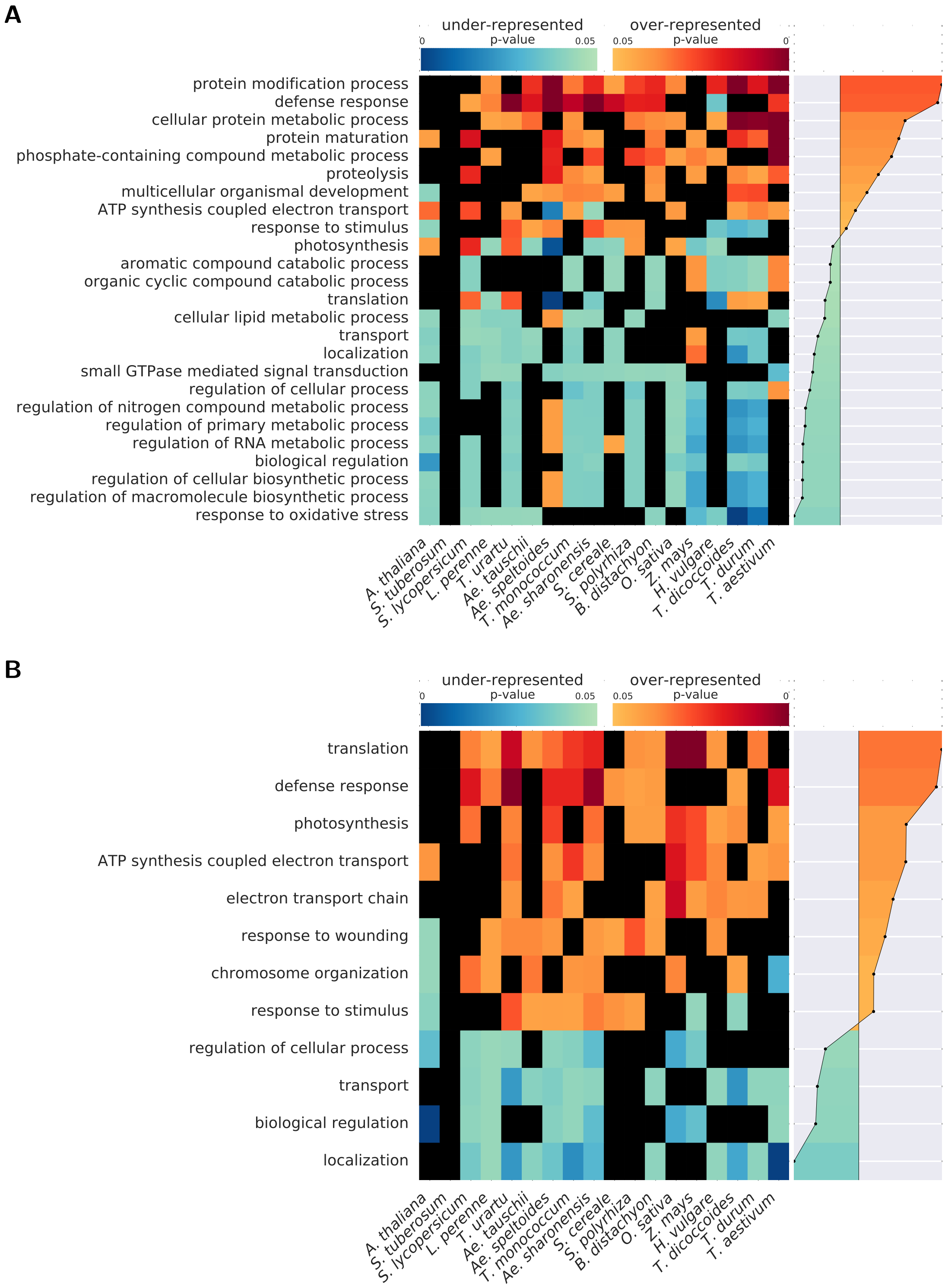


Figure R17: GO enrichment analysis. Over- and under-represented GO terms in the parent gene set of all 18 plants compared to the complete gene sets, respectively. Only terms with a significant P-value in at least nine plants are shown. A: GO enrichment analysis for all pseudogene parents; B: GO enrichment analysis for HCov pseudogene parents. There is no GO annotation available for potato.

5.8 Distinguishing genes from pseudogenes

Distinguishing genes from pseudogenes is a difficult task, since pseudogenes are *gene-like* sequences. Even if a putative pseudogene contains premature termination codons, it can still be transcribed and exert regulatory effects on paralogous genes (Pink et al., 2011). Efficient translational read-through can even lead to pseudogene translation (Schueren and Thoms, 2016). Additionally, some individuals of a species may contain a pseudogenized gene, while others still contain the functional gene. In human, retroposed gene copies are generally considered pseudogenes (until proven otherwise), because they lose their original promoter sequence during duplication and are thus considered dead-on-arrival (E. Khurana et al., 2010; Xu and J. Zhang, 2015).

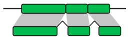
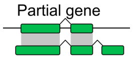
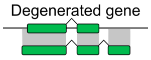
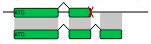
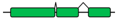
Some pseudogenes are highly similar to their parent genes and could in principle have retained their original function. On the other hand, the official protein-coding gene sets also contain elements, that have arisen via gene duplication and exhibit features that normally would be considered sufficient to classify them as pseudogenes. Elements with features of both functional genes and dysfunctional pseudogenes were identified to estimate the flux between gene and pseudogene state.

5.8.1 Genes containing pseudogene features

Gene duplication is a common mechanism to generate new genes. Hence, many protein-coding genes are found to be related copies of other functional genes. Computational gene prediction often includes homology searches to known genes in related species. Evidence-based gene prediction also utilizes transcriptional evidence like RNAseq data, expressed sequence tags (ESTs), sequenced proteins or full-length complementary DNAs (FLcDNAs) (Liang et al., 2009). In this work, the annotated protein-coding gene sets were used to identify pseudogenes via homology. Gene and TE sequences were not masked on the genome, but subsequently filtered from the pseudogene set. However, mapping one gene onto another can shed light on its origin. Although these genes were annotated as functional and protein-coding using an evidence-based approach, they show pseudogene characteristics and may even degenerate further or adopt modified functions.

The number of potential retrogenes, partial genes and degenerated gene duplicates were determined using the homology mapping of the PLIPipeline. Retrogenes are genes that have been duplicated via retrotransposition but remained functional or were subsequently reactivated. In this work, a stringent definition was used to identify retrogenes: (i) They must not contain any introns and (ii) another homologous gene with very similar CDS length ($\geq$98%) contains at least one intron. Fourteen of the target plants contain such retrogenes (Table R7).

Table R7: Duplicated genes with characteristics of pseudogenes. MOD is the number of genes that have CDS lengths not divisible by three.

		Species	Retrogene 	Partial gene 	Degenerated gene 	Shortened via PTC 	Frameshift 	MOD	PTC
dicots		*A. thaliana*	11	718	268	13	7	36	37
		S. tuberosum	12	5,341	5,231	298	10	0	0
		S. lycopersicum	9	2,660	3,312	177	0	1	0
monocots	contigs/scaffolds	*L. perenne*	1	2,109	1,210	0	0	0	302
		T. urartu	16	3,395	1,086	14	1,484	76	1,427
		Ae. tauschii	3	4,722	2,008	54	22	5,472	7,729
		Ae. speltoides	0	12,207	1,623	0	0	0	793
		T. monococcum	1	5,426	1,023	1	0	0	775
		Ae. sharonensis	0	6,123	1,453	2	0	0	929
		S. cereale	0	3,918	1,064	0	0	0	1,051
	pseudomolecules	*S. polyrhiza*	7	926	529	11	77	122	31
		B. distachyon	0	742	595	40	1	0	3
		O. sativa	2	2,100	1,383	84	3	9	9
		Z. mays	2	3,405	2,359	223	216	6	4
		H. vulgare	18	6,461	5,345	324	19	0	0
		T. dicoccoides	83	6,545	6,781	415	202	0	3
		T. durum	108	6,635	6,955	411	201	0	2
		T. aestivum	375	10,802	9,287	536	1,482	0	0

Most could be identified for the *Triticeae* with high-quality reference genome sequences: barley (18), wild emmer (83), durum wheat (108) and bread wheat (375). As expected, polyploid species contain more putative retrogenes than diploid plant species. This could indicate that they originated before the polyploidization event. However, hexaploid bread wheat does not contain 50% more, but over 300% more retrogenes than the tetraploid wheat species. An effect of *Triticeae* domestication may be possible, also considering that the domesticated durum wheat contains more retrogenes than the wild subspecies wild emmer. Furthermore, the number of retrogenes is likely underestimated: Smaller isoforms can also be used as templates for reverse transcription and the generation of retrogenes. By requiring a CDS length at least 98% similar to the representative (i.e. the longest) splice variant of another gene, retrogenes from smaller splice variants are not taken into consideration.

Three additional types of genes with characteristics of pseudogenes were defined and analyzed: partial, degenerated and genes with putative shifts in the reading frame. Partial genes are incomplete copies of other genes. Homology does not extend beyond their annotated CDS sequence. Among them, some may be retrogenes that have been missed in the previous assessment, as their exon-intron structure is not considered in this assessment. In contrast, degenerated genes are more complete copies of genes, but their annotation does not cover the complete homologous CDS region. A subgroup of degenerated genes are those that are shortened due to a PTC, but are otherwise highly similar. Numerous partial and degenerated genes were identified for all target plants (Table R7). The high number of partial genes in *Ae. speltoides* and other plants with contig assemblies can be explained by poor assembly qualities. However, bread wheat contains almost ten thousand partial and degenerated genes, respectively. Of the degenerated genes, 536 represent elements which usually would be considered classical pseudogenes: a complete duplicate containing a PTC. This does not mean that they are non-functional pseudogenes, since there usually had to be some transcription evidence to support their annotation. Also, the PTC might not shorten the gene sufficiently to affect function.

Sequencing errors and flawed gene structure predictions may contribute to the high number of partial or degenerated genes. In 2006, the results of the human ENCODE Genome Annotation Assessment Project (EGASP) community experiment have been published (Guigó et al., 2006). The experiment was conducted to assess state-of-the-art genome annotation in human. While almost all of the human genes could be identified, the structural prediction was correct for only 50% of the elements. If computational prediction of plant gene structure is similarly flawed, then this would explain many of the “degenerated” genes. However, the

Arabidopsis thaliana gene annotation is of very high quality and it is manually curated. Still, it contains 718 partial and 268 degenerated genes (Table R7).

Most of the plant gene annotations contain gene structures with introns of less than 20 bp. Many even exhibit introns of only one base pair. Such small introns are suspicious, because splicing requires the intron sequence to form a loop structure in order to be removed. If those short introns had not been annotated, most of those genes would have contained frame shifts. Almost 1,500 bread wheat and red wild einkorn wheat genes contain such dubious introns, respectively, making them textbook pseudogene candidates. Nonetheless, such introns may also be the result of sequencing and assembly errors.

Finally, gene annotations were checked for PTCs and CDSs lengths not divisible by three. While the CDS of most genes have a correct length, many genes contain PTCs. The gene set of *Aegilops tauschii* was contaminated with wrong sequences (contigs), that were removed from subsequent annotation versions. Those genes have been filtered from the template gene set that was used to detect pseudogenes. Disregarding this error, some annotations still contain up to 1,427 genes with PTCs. Even the manually curated Arabidopsis genome or the newly generated annotations of wild emmer and durum wheat contain genes with PTCs.

This confirms that annotated protein-coding genes should be investigated for potential pseudogene characteristics. Degenerated genes may represent resurrected genes or be en route to pseudogenization or subfunctionalization. Similarly to pseudogenes, parts of the original CDS can mutate and evolve. If a repairing mutation removes PTCs, a potentially novel function may be adopted.

5.8.2 Functional pseudogenes?

Some of the annotated pseudogenes may in fact be functional genes. In human and other organisms, numerous pseudogenes have been identified, that are transcribed and exert regulatory roles on paralogous genes (Pink et al., 2011). Some might even be translated due to efficient translational read-through (Prieto-Godino et al., 2016). And others may be non-functional only in some individuals of a population.

5.8.2.1 Features affecting functional potential A straightforward way to identify pseudogenes with potential functionality is to examine features that are seen as indicative for disrupted open reading frames (ORFs) or other functional impairment. Sequence coverage and identity compared to the respective parent genes sheds light on putative protein-coding “pseudo”genes (s. Figure R5). Most of the pseudogenes are short gene fragments with a sequence coverage of less than 20%. Of the HCov pseudogenes, most have a sequence identity between 90 and 100%. With this high similarity to the functional parent gene, it is not unlikely that some of them are still functional.

Table R8: HCov pseudogenes with functional impairments. Number and percentage of HCov pseudogenes containing PTCs or frameshifts. *None* contains the metrics for pseudogenes without PTCs, frameshifts, insertions or deletions.

		Species	HCov	PTC	%	Early PTC	%	Frameshift	%	None	%
dicots		*A. thaliana*	373	269	72	188	50	243	65	76	20
		S. tuberosum	5,231	3,827	73	2,726	52	3,588	69	674	13
		S. lycopersicum	6,026	4,704	78	3,433	57	4,531	75	647	11
monocots	contigs/scaffolds	*L. perenne*	4,200	2,249	54	1,380	33	1,961	47	1,127	27
		T. urartu	7,185	4,579	64	3,020	42	4,003	56	1,432	20
		Ae. tauschii	4,497	3,232	72	2,181	48	2,910	65	586	13
		Ae. speltoides	11,231	6,002	53	3,572	32	5,343	48	3,233	29
		Ae. sharonensis	7,783	4,559	59	2,803	36	4,027	52	1,816	23
		S. cereale	13,259	9,230	70	6,084	46	7,937	60	1,992	15
	pseudomolecules	*S. polyrhiza*	277	201	73	132	48	175	63	40	14
		B. distachyon	956	678	71	410	43	688	72	127	13
		O. sativa	2,693	1,680	62	1,038	39	1,860	69	505	19
		Z. mays	6,990	4,125	59	2,530	36	4,251	61	1,762	25
		H. vulgare	6,191	3,246	52	1,890	31	2,766	45	1,765	29
		T. dicoccoides	25,147	16,995	68	11,055	44	15,482	62	3,604	14
		T. durum	25,589	17,352	68	11,296	44	15,741	62	3,721	15
		T. aestivum	48,608	33,290	68	22,126	46	30,903	64	6,384	13

A PTC disrupts the original ORF and shortens the translated sequence. While the pseudogene transcript may still be functional, a complete protein sequence cannot be produced. Approximately 70% of the pseudogenes in plants with high-quality reference genome sequences have PTCs (Table R8). Plants with contig/scaffold genome assemblies contain partial gene and pseudogene annotations. The lower percentage of pseudogenes with PTCs can also be attributed to assembly quality differences. Most of the PTCs are located in the 5' half of the pseudogene "CDS". Thus, a functional protein product is highly unlikely. Frameshifts occur in 45% to 75% of the HCov pseudogenes and often are the cause of PTCs.

While most of the pseudogenes contain features like nonsense-mutations or frameshifts, there are some without any PTCs, frameshifts, insertions or deletions. In barley, wild emmer and durum wheat, there are over 3,000 HCov pseudogenes without such features — in bread wheat, there are 6,384. However, most of them have both a sequence coverage and identity of less than 98%, which could already be sufficient to impair functionality. Additionally, even a perfect copy of CDS of a functional gene can be dysfunctional if the promoter or regulatory regions are defect.

For the newer *Triticeae* assemblies, the gene annotation is divided into a high- and low-confidence set. For example, genes may have been moved into the low-confidence gene sets when there is no annotated start or stop codon. Since the high-confidence gene sets were used as templates to identify pseudogenes via homology, some pseudogenes may overlap with low-confidence genes. A more detailed analysis of overlapping annotations will be provided in section 5.10.

5.8.2.2 Pseudogenes with transcription evidence Transcribed pseudogenes can be involved in regulatory functions affecting protein-coding genes with similar sequence (Pink et al., 2011; Sen and Ghosh, 2013). An RNA sequencing (RNA-seq) analysis can be performed to identify transcribed pseudogenes. However, the high sequence similarity of pseudogenes and parent genes hampers the correct assignment of reads to one locus. Thus, only reads mapping uniquely to one locus can be used to determine whether a pseudogene is transcribed or not. As a consequence, no conclusions about the expression levels of pseudogenes can be inferred and the number of transcribed pseudogenes is likely an underestimation.

In barley, there is transcriptional evidence for 9,312 pseudogenes, 734 are HCov pseudogenes (12% of HCov). In comparison, approximately 20% of the previously annotated pseudogenes in Arabidopsis and rice have been reported to be transcribed (Podlaha and J. Zhang, 2010). The transcribed pseudogenes of barley originate from parent genes functionally related to glycolysis or other metabolic processes involving sugars. Parents from transcribed HCov pseudogenes are primarily involved in carbohydrate metabolic processes and response to endogenous stimuli, organic substances and auxin. Transport and localization related parent genes are under-represented. Transcribed pseudogenes may take part in similar processes as their parent genes. However, transcription evidence does not necessarily imply functionality, but only the potential for functionality. In human, transcription evidence was found for 863 of 11,224 (8%) pseudogenes. (The ENCODE Project Consortium, 2012)

These and other results suggest that some of the pseudogenes are in fact genes that have been missed in the annotation. Some may not be protein-coding, but still adopt regulatory roles on the RNA level. Pseudogenes possess a hidden potential, that needs to be assessed and investigated. An additional way to review the functional properties of pseudogenes is the analysis of selection pressure on pseudogenes.

5.8.2.3 Pseudogenes under selection pressure A K_A/K_S analysis is a common way to determine selective pressure on homologous gene pairs. In essence, the rate of non-synonymous substitutions (K_A) is compared to the rate of synonymous substitutions (K_S). If a sequence is evolving neutrally, a K_A/K_S ratio of one is expected. Under purifying selection, non-synonymous substitutions are preferentially accumulated and the K_A/K_S ratio adopts a value smaller than one. In contrast, conservation pressure prevents non-synonymous but tolerates synonymous substitutions — leading to a K_A/K_S value larger than one. Pseudogenes are expected to be dysfunctional and to evolve neutrally. Comparing their sequence to the functional parent gene should result in a K_A/K_S ratio of one.

A K_A/K_S ratio analysis was performed on HCov pseudogenes that contain PTCs (Figure R18 A). The logarithmic values were plotted to obtain a symmetric value distribution. Neutral evolution would then result in logarithmic value of zero. For all of the 18 target plants, the mean logarithmic K_A/K_S ratio is below zero, which is indicative of conservation pressure. This shift to the left is statistically significant for all plants but *Spirodela polyrhiza*. However, the conclusion, that most of the pseudogenes in plants are conserved — and thus likely functional — is wrong.

Instead, this result can be explained by the concurrent evolution of the parent gene (Figure R18 B). Similar results have been reported for the genome and pseudogene complement of rice (Thibaud-Nissen et al., 2009): The present parent gene is not identical to the originally duplicated gene. Since the duplication event, the pseudogene accumulated random mutations due to neutral evolution, while the functional gene preferentially accumulated synonymous substitutions. Comparing the present gene to its ancestral state would result in a K_A/K_S value below one. In contrast, comparing the present pseudogene to the ancestral gene would result in a K_A/K_S value of one. Today, the ancestral template gene is not known,

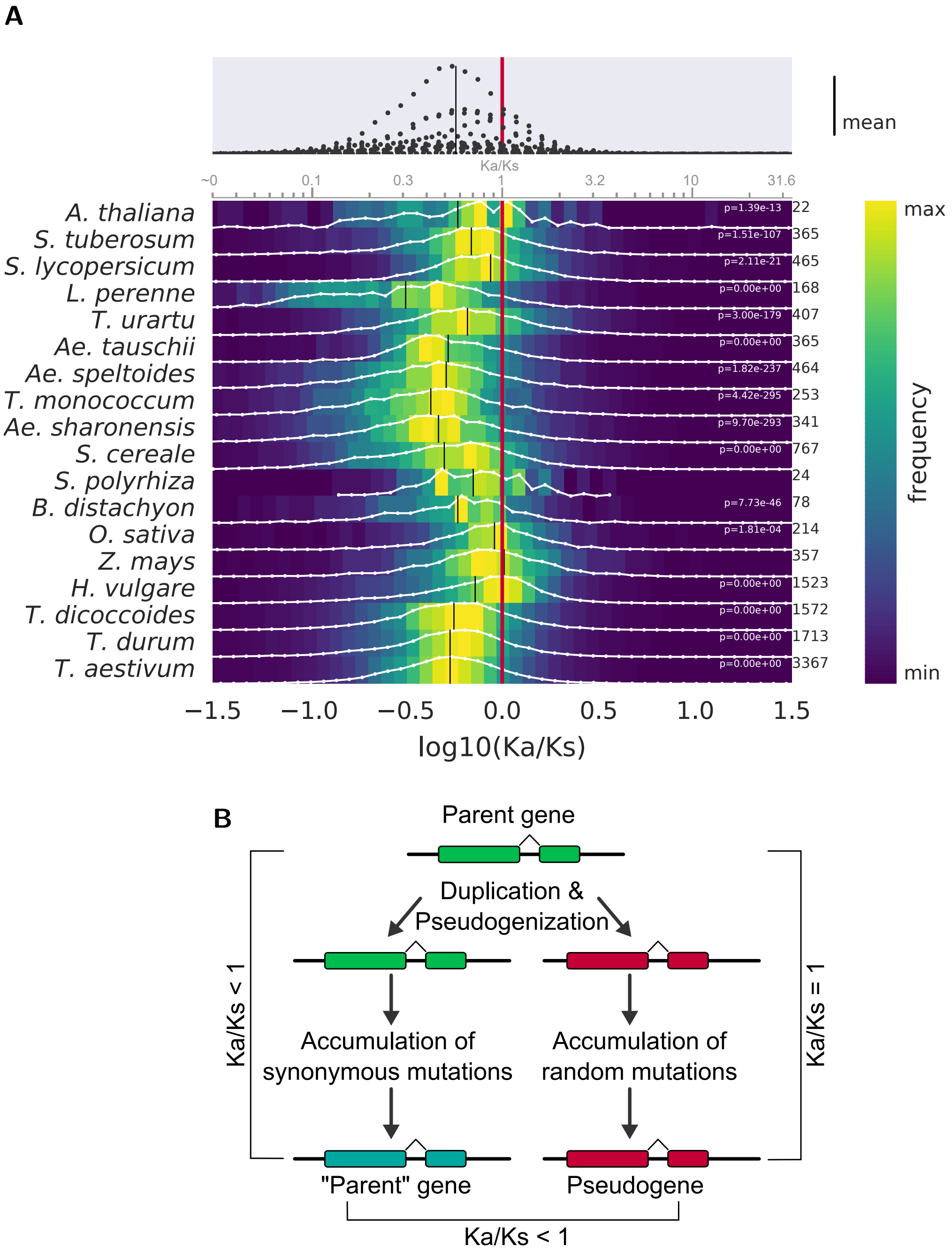


Figure R18: K_A/K_S **ratio analysis.** A: K_A/K_S ratio distribution for HCov pseudogenes containing PTCs compared to their respective parent genes. Mean values are highlighted in black. If the mean is significantly smaller than 0, the respective P-values are given inside the plot area at the right. Maximum frequencies for each species are shown right next to the plot area. B: Schematic representation of the different accumulation of mutations between pseudogenes and their parent genes, that leads to the negative shift depicted in A.

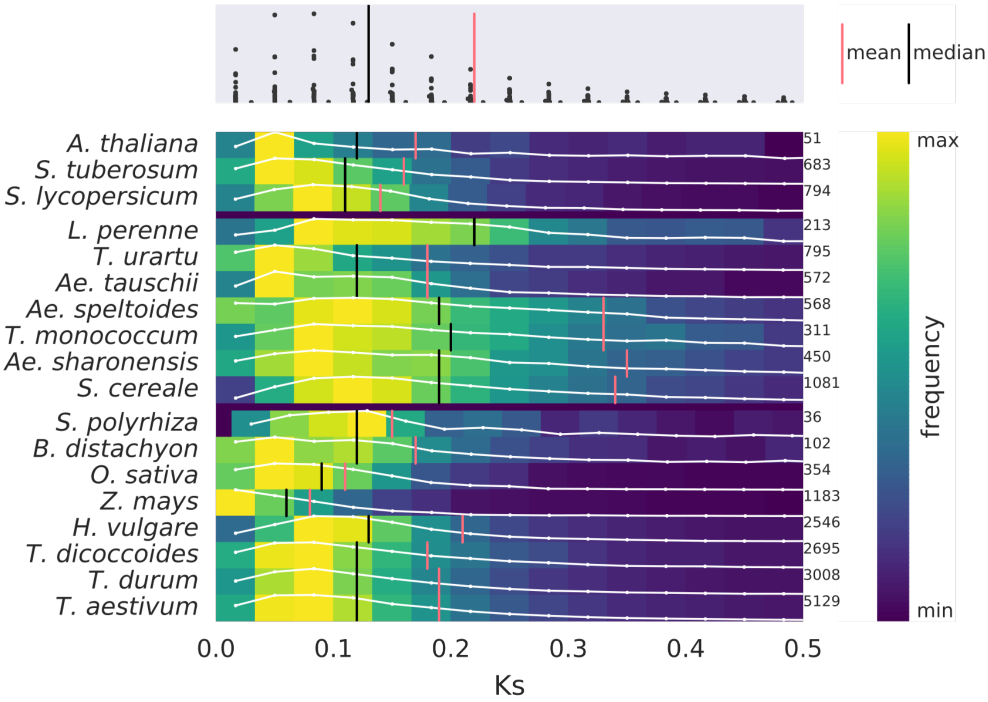


Figure R19: K_S distribution for HCov pseudogenes containing PTCs compared to their respective parent genes.

so only its present state can be compared to the pseudogene and the K_A/K_S ratio distribution is slightly shifted to the left (mean K_A/K_S value between 0.3 and 1). Additionally, many plant pseudogenes may be comparatively young and pronounced K_A/K_S ratios may require more time.

A comparative K_S analysis can help to estimate the relative age of pseudogenes in different plants or subgenomes. When focusing on monocots with high-quality genome assemblies, rice and maize stand out and lower median and mean K_S values are found (Figure R19). Their pseudogenes may be younger and less degenerated owing to a recent segmental duplication (5–21 mya) in rice and a recent Whole Genome Duplication (WGD) in maize (5–12 mya) (J. Yu et al., 2005; Schnable et al., 2009). Furthermore, barley contains slightly more degenerated pseudogenes than wild emmer, durum wheat or bread wheat. While barley has a diploid genome, the other three *Triticeae* have polyploid genomes. Tetraploidization happened 6.5 mya and might have led to progressed pseudogenization of redundant genes. However, hexaploidization in bread wheat happened only ~10,000 years ago and no difference in the Ks value distribution between bread wheat and wild emmer was found.

Table R9: Genes, parent genes and pseudogenes per subgenome.

Species	Subg.	Size %	Genes %	Parents %	Pseudogenes	%	HCov	%
T. dicoccoides	A	48.6	49.8	51.5	122,822	49.4	11,007	46,2
	B	51.4	50.2	48.5	125,780	50.6	12,819	53,8
T. durum	A	48.7	49.6	51.3	109,173	49.0	11,402	47,2
	B	51.3	50.4	48.7	113,740	51.0	12,772	52,8
T. aestivum	A	35.1	33.6	33.2	93,154	34.4	15,248	32,6
	B	36.8	34.0	33.8	101,425	37.4	17,303	37,0
	D	28.1	32.4	32.9	76,451	28.2	14,250	30,4

5.9 Pseudogenes in polyploid plants

Polyploid species contain several subgenomes that originate from the same (autopolyploidy) or from different taxa (allopolyploidy). The presence of multiple and similar gene sets may lead to progressed pseudogenization of redundant genes. Wild emmer and durum wheat both are allotetraploid subspecies of *Triticum turgidum*. The closest relatives of their subgenome progenitors are *Triticum urartu* (AA) and *Aegilops sharonensis*/*Aegilops speltoides* (BB), which diverged approximately 6.5 mya (s. Figure I9). Tetraploidization happened less than ~0.8 mya. Bread wheat has a hexaploid genome that originated from an additional polyploidization event between tetraploid *Triticum turgidum* (AA BB) and *Aegilops tauschii* (DD) less than 0.4 mya (Marcussen et al., 2014).

Gene and pseudogene content were compared among the subgenomes of tetraploid and hexaploid wheat (Table R9). In each of the three *Triticeae*, the B subgenome is the largest subgenome and contains the highest number of genes. In tetraploid wheat, more of the genes on subgenome A are parents to pseudogenes and there are more pseudogenes on subgenome B. This indicates, that genes duplicated due to tetraploization preferentially pseudogenized on the B genome. However, the higher number of functional genes on B opposes this hypothesis.

The D subgenome of bread wheat is smaller and joined the ancestral tetraploid genome much more recently. Nevertheless, gene and parent gene portion per subgenome are very similar. The pseudogene ratio on the three subgenomes correlates with subgenome size. Therefore, compared to sequence amount, there are more pseudogenes on subgenome D than expected for the observations in genomes A and B. A smaller number of pseudogenes was initially anticipated due to the recent addition of the D subgenome to the wheat genome. It was only added less than 0.4 mya. Hence, the small time frame should not have been sufficient for progressed pseudogenization on the subgenome D. Pseudogenization either happened relatively fast or pseudogenes on subgenomes A and B exhibit advanced degeneration affecting their detection. However, if not comparing the number of

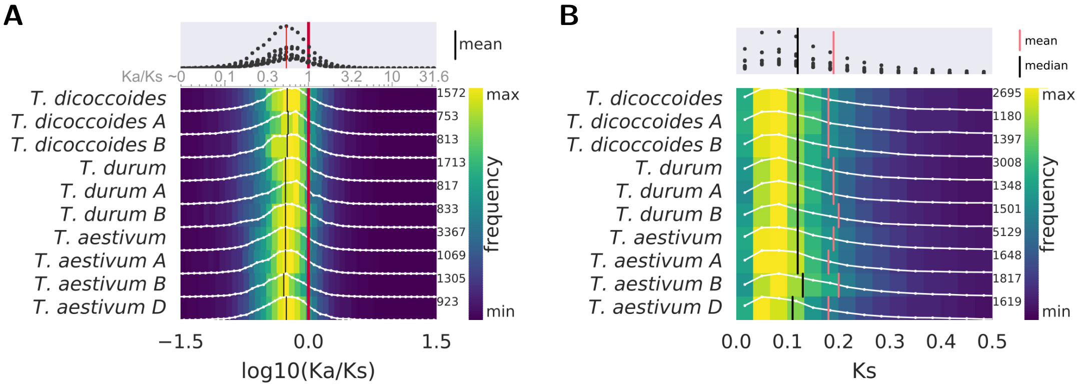


Figure R20: K_A/K_S ratio and K_S value distribution for HCov pseudogenes containing PTCs. A: K_A/K_S ratio distribution. B: K_S value distribution. Mean values are highlighted in black. Maximum frequencies for each species or subgenome are shown right next to the plot area.

pseudogenes to the subgenome size but instead to the gene number, subgenome D does contain significantly fewer pseudogenes (binomial test, P-value<0.05).

If some subgenomes contain younger pseudogenes, this should be reflected in K_A/K_S or K_S values. Overall, the K_A/K_S and K_S distributions are very similar between all polyploid *Triticeae* and their subgenomes (Figure R20). In durum wheat, the K_S distribution of subgenome B is slightly shifted to the right. Larger K_S values indicate more synonymous substitutions and possibly an older age. However, there is no such evidence in wild emmer.

Even though pseudogene numbers in hexaploid bread wheat correlate with subgenome size, the K_S distribution of pseudogenes on the D subgenome is significantly shifted to the left (P-value<0.05), while the K_S values of pseudogenes on B subgenomes are shifted to the right (not significant). Hence, pseudogenes on subgenome D seem to be younger than pseudogenes on A and B.

5.10 High- and low-confidence genes

Gene annotations of recently published *Triticeae* genome assemblies are divided into high-confidence genes (HC genes) and low-confidence genes (LC genes) (International Barley Sequencing Consortium, 2017; Avni et al., 2017; Maccaferri et al., 2018; International Wheat Genome Sequencing Consortium, 2018). Homology to a reference gene, presence of start and termination codons and transcription evidence were used to classify the gene set. The PLIPipeline uses HC genes as query sequences to identify pseudogenes via homology. Hits overlapping with the HC genes are filtered. However, overlaps with LC genes are possible.

Table R10: HC and LC genes and overlaps with pseudogenes. LC genes with TE-evidence have been filtered.

Species	HC genes	LC genes	Overlapping with LC (cov)		
			≥0%	≥90%	≥90% & PTC
H. vulgare	39,734	40,819	6,254	1,512	960
T. dicoccoides	67,182	86,792	20,880	7,549	5,740
T. durum	66,559	98,007	22,333	8,231	6,153
T. aestivum	110,790	105,747	43,578	17,684	12,761

The numbers of non-TE-related LC genes in barley, wild emmer, durum wheat and bread wheat are comparable to the numbers of HC genes (Table R10). Most of the LC genes do not overlap with pseudogenes. This could either mean that they mostly comprise unitary pseudogenes without a functional parent gene in the HC gene. However, it is more likely that most of the LC genes overlap with the TE annotation, which causes the PLIPipeline to filter pseudogenes at such loci.

5.11 Functional background of pseudogenes in hexaploid wheat

Genes on the homeologous chromosomes of bread wheat have been shown to be differentially expressed leading to a genome dominance and bias for certain phenotypes (Pfeifer et al., 2014). For example, the D genome dominates in the wheat hardness locus, which is important for baking quality.

A GO enrichment analysis was performed for pseudogenes of each subgenome compared to the complete bread wheat pseudogene set (Figure R21). For this, GOs from parent genes were transfered to pseudogene children. Enriched GO terms were then semantically clustered and visualized using REVIGO (Supek et al., 2011). Interestingly, the result is different when using all pseudogenes including gene fragments or when focusing only on HCov pseudogenes. Without filtering, most clusters are dominated by subgenome D (Figure R21 A). Clusters constitute functional categories like response to oxidative stress, sexual reproducion, regulation of signaling, photosynthesis or cell cycle. Some of those clusters can also be found when filtering for HCov pseudogenes (Figure R21 B). However, a few clusters are now clearly dominated by subgenome A and B. Defense response pseudogenes are over-represented on subgenome B. Response to oxidative stress is dominated in both A and B. Previous reports show that defense response genes on subgenome D show pronounced response to the fungal plant pathogen *Fusarium graminearum* (Nussbaumer et al., 2015). Hence, a higher pseudogenization rate of defense-related genes on A and B may have contributed to this subgenome dominance.

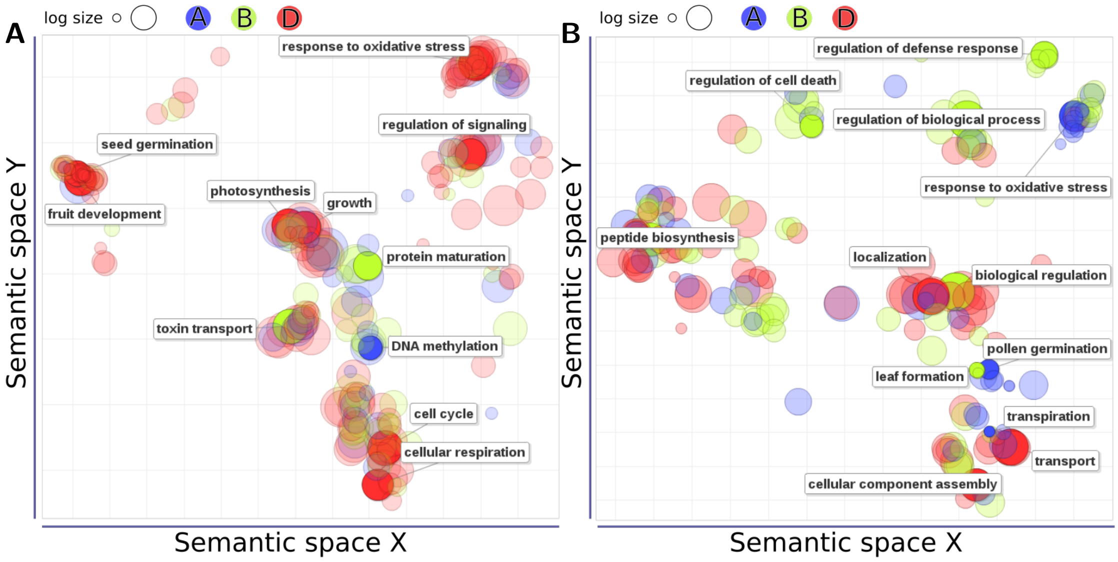


Figure R21: GO enrichment analysis of bread wheat pseudogenes per subgenome. Enriched GO terms were semantically clustered and colored by subgenome (Supek et al., 2011). A: all pseudogenes and gene fragments; B: HCov pseudogenes.

5.12 Comparative analysis of pseudogenes in closely related plants

Pseudogenes are expected to evolve neutrally and to continuously accumulate mutations. Older pseudogenes are likely degenerated beyond recognition, because they do not share sufficient sequence similarity with functional protein-coding genes. When closely related species have pseudogenes in common, they likely have their origin prior to species divergence. Environmental constraints or artificial selection pressure via domestication may affect the gene and pseudogene dynamics. Comparative analyses not only shed light on evolutionary processes, but also helps to determine the age of pseudogenes. Bidirectional Best BLAST hits (BBHs) are a common method to identify orthologous genes and syntenic regions.

5.12.1 *Brachypodium distachyon* and *Oryza sativa*

Brachypodium distachyon and *Oryza sativa* (rice) are both grasses that diverged approximately 46 mya (Bolot et al., 2009). Compared to other plants, they both have smaller genomes with 355 and 489 Mbp, respectively. Rice contains almost 50% more protein-coding genes and even more pseudogenes compared to *Brachypodium distachyon*. While they represent distantly related members of the *Poaceae* family, syntenic regions and segmental duplications are still clearly discernible via

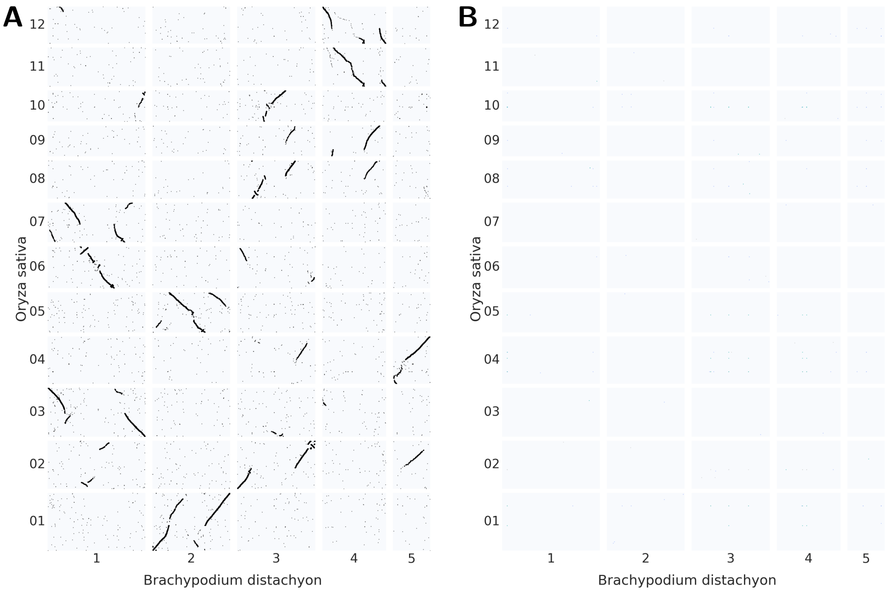


Figure R22: Orthologous sequences and syntenic regions between rice and Brachypodium. A: BBHs of genes. B: BBHs of pseudogenes.

orthologous genes (Figure R22 A). However, there are only very few orthologous pseudogenes and no syntenic blocks can be derived from them (Figure R22 B). Thus, pseudogenes that originated more than 46 mya are highly diverged and degenerated.

In primates, the age of most processed pseudogenes has been dated to ∼40 million years (Ohshima et al., 2003). Thus, their emergence coincides with the the surge of Alu elements. However, there is evidence for even older pseudogenes that originated over 80 mya (Z. Zhang and M. Gerstein, 2003). The majority of identifiable pseudogenes in plants like *Brachypodium distachyon* and rice clearly are not as old as the processed pseudogenes of primates. This is due to a very different rate of molecular evolution, which correlates with generation time. Organisms with shorter generation times are assumed to evolve faster, because frequent genome duplication leads to the accumulation of replication errors (Weller and Wu, 2015). *Brachypodium distachyon* is an annual grass with a life cycle of less than 4 months (Draper et al., 2001). Most rice varieties are annual as well. Under controlled conditions the generation time of the rice cultivar 'Nipponbare' could be reduced to 3 months (Tanaka et al., 2016). Thus, both have much faster generation turnovers than primates accompanied by an increased pseudogene degeneration speed.

5.12.2 *Solanum tuberosum* and *Solanum lycopersicum*

Solanum tuberosum (potato) and *Solanum lycopersicum* (tomato) are species from the same genus that diverged approximately 7.3 mya (The Tomato Genome Consortium, 2012). While potato has a 15% smaller genome size than tomato, it contains 12% more genes. However, previous studies have shown that tomato proteins belong to more gene families and that the tomato genome is more repetitive than the potato genome (Lall et al., 2013). Interestingly, not tomato but potato contains more TE-related pseudogenes and shorter pseudogene fragments, which can be attributed to contaminations of the gene set with TE genes. In contrast, while potato has more TE-related pseudogenes, tomato has 15% more HCov pseudogenes. Apart from assembly and annotation differences, a higher pseudogenization rate in tomato may explain the lower number of genes, as well as the higher number of HCov pseudogenes.

Syntenic regions are clearly detectable when plotting orthologous genes between the two plants (Figure R23 A). Additionally, they can now be identified when visualizing pseudogene BBHs (Figure R23 B). Hence, the origin of these pseudogenes likely predates species divergence ~7.3 mya. Together with previous age estimates derived from the analysis of intraspecies segmental duplications in section 5.5.2.3, this confirms that the oldest plant pseudogenes detectable by the PLIPipeline are ~10 million years old.

5.12.3 *Triticum dicoccoides* and *Triticum durum*

Wild emmer and durum wheat are subspecies of *Triticum turgidum* and as such very closely related. They diverged only ~10,000 years ago approximately at the same time when men started a sedentary lifestyle with the beginning of agriculture (Haberer et al., 2016). Durum wheat is an economically important cultivated cereal that is used for pasta production (Maccaferri et al., 2018). Comparing the genomes of durum wheat and wild emmer may give information about the role of pseudogenes during domestication.

Wild emmer and durum wheat both contain ~12 Gbp genomes that harbor ~67,000 genes and ~25,000 HCov pseudogenes. Their tetraploidy is obvious via orthologous and homeologous genes between durum wheat and wild emmer (Figure R24 A). Interestingly, pseudogene BBHs are mainly detected between homologous and not between homeologous chromosomes (Figure R24 B). Hence, pseudogenes are preferentially found on the respective “founder” chromosomes. To interpret this result, a more detailed explanation of the methodology used to detect “orthologous” elements is necessary: BBHs are identified by mapping genes/pseudogenes from one subspecies onto the genes/pseudogenes of the other subspecies and vice versa. If the best hit of element A is element B and the best

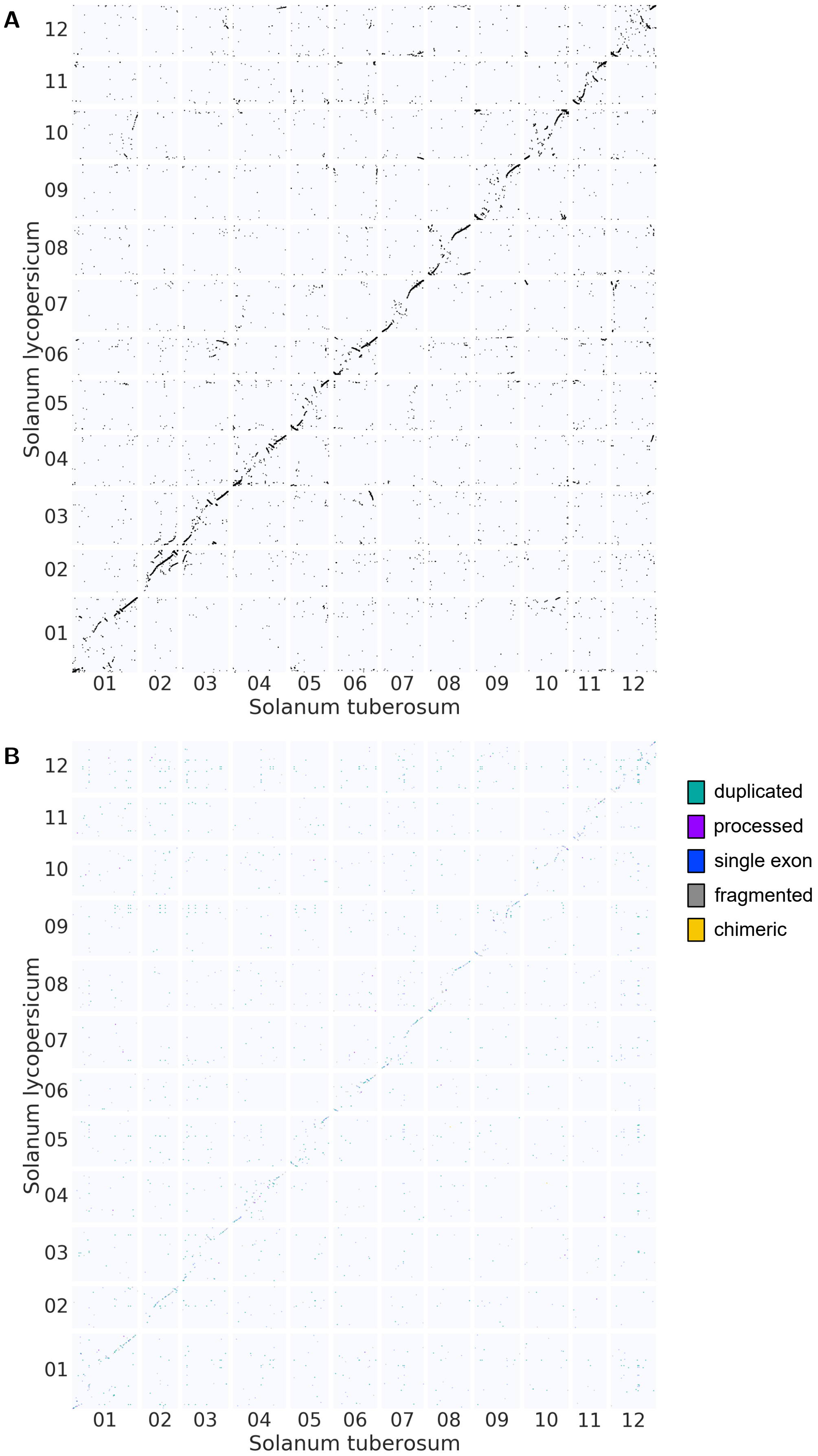


Figure R23: Syntenic regions between potato and tomato. A: BBHs of genes. B: BBHs of pseudogenes. Pseudogene positions are colored according to pseudogene type (Figure M5)

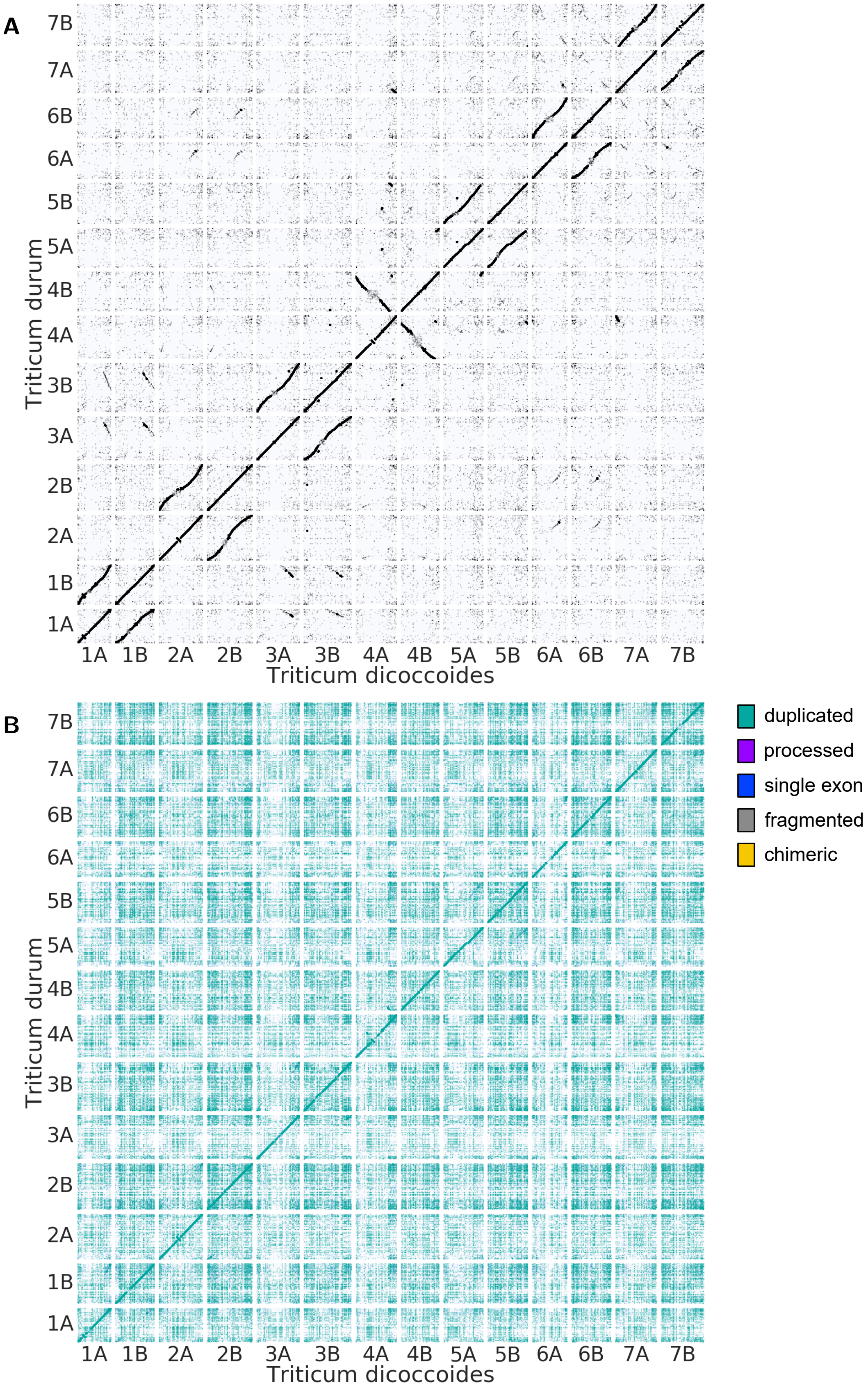


Figure R24: Syntenic regions between wild emmer and durum wheat. A: BBHs of genes. B: BBHs of pseudogenes.

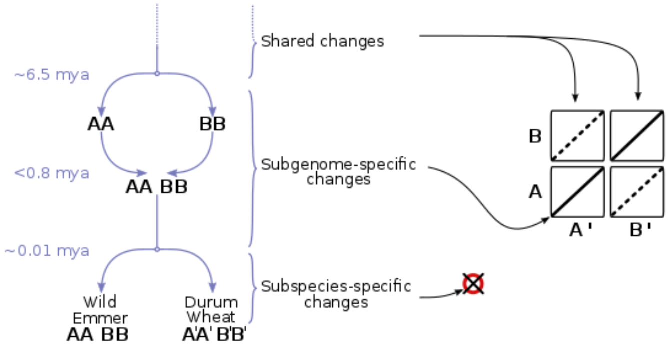


Figure R25: Divergence and polyploidization of *Triticum turgidum* subgenome progenitors.

hit of element B is element A, then the pair is defined a BBH. However, in this analysis, not only the best hit is used, but all with a P-value that is similar (>70%) to the P-value of the actual best hit. As a consequence, one element may appear in multiple BBHs and syntenic regions between all subgenomes become visible. Not only orthologous elements, but also homeologous and some paralogous elements are assessed by this approach.

When plotting pseudogene BBHs, syntenic regions are only visible for homologous chromosomes (Figure R24 B). The diagonal comprises mostly pseudogenes that originated before domestication 10,000 years ago, but after the divergence of the A and B subgenome progenitors ~6.5 mya (Figure R25). Finally, pseudogenes that are older than ~6.5 mya are mostly shared between homologous chromosomes of both subspecies, but also between homoelogous chromosomes within subspecies. They are not as abundant, but still discernible (e.g. chromosome 5). Overall, many pseudogene BBHs are not within a syntenic context, but appear randomly distributed.

The genomes of *Triticum turgidum* subspecies are more than twelve times larger than those of potato or tomato and they contain more than ten times as many pseudogenes. However, this alone does not explain the abundance of BBHs scattered randomly over the chromosomes. They may be the result of relatively recent double-strand break repair processes that lead to insertion of filler DNA.

DNA breaks accumulate during increased TE activity. The activity of TEs has previously been investigated for chromosome 3B of bread wheat: While they have been mostly silent during the last one million years, there was a "global burst" of TE activity 1.2 mya and "smaller bursts" between 1 and 3 mya (Daron et al., 2014). Such TE activity would explain duplicated sequences that are randomly distributed on the chromosomes and share sequence similarity with numerous disjunct loci. The B subgenome of bread wheat is derived from the same ancestor as the B subgenome in *Triticum turgidum*. Tetraploidization happened less than

1 mya. Hence, these "bursts" of TE activity happened prior to polyploidization. Since pseudogenes are estimated to loose their function within the first few million years after the loss of selection pressure (Lynch and Conery, 2000), they are likely not too degenerated and still highly similar to their functional parent genes. Therefore, they may be part of numerous BBH pairs and contribute to the background distribution that is visible in figure R24 B.

Lineage specific pseudogenes A previous version of the PLIPipeline was used to assert more subtle differences between the pseudogene complements of wild emmer and durum wheat. Pseudogenes were detected using the combined gene sets of both subspecies to identify unitary pseudogenes that only have a parent gene in the other subspecies. They likely pseudogenized after subspecies divergence and potentially due to domestication.

Wild emmer and durum wheat contain 9,620 lineage specific genes that were mapped back to the other genome, respectively, to identify pseudogenized or structurally altered elements (Maccaferri et al., 2018): Approximately 26% and 31% of the lineage specific genes are not found in wild emmer and durum wheat, respectively. Hence, they represent deleted or highly degenerated elements. 22% and 26% can be mapped to fragmented or structurally altered HC genes. These genes may still be functional, but they show signs of pseudogenization or functional divergence. Finally, 1,539 (32%) and 1,095 (23%) lineage specific genes were mapped to LC genes or pseudogenes, corresponding to approximately 2.3% and 1.6% of the total wild emmer and durum wheat genes that are degenerated and pseudogenized in the other lineage, respectively. They represent unitary pseudogenes.

This result indicates that large-scale pseudogenization happens within short periods of time and is likely affected by domestication and artificial selection. The functional state of many genes may be in constant flux.

Furthermore, the syntenic context of genes and pseudogenes was investigated. Orthologous and paralogous elements were determined within the identified syntenic regions. Approximately 47,000 syntenic gene pairs were identified in wild emmer and durum wheat. Additionally, ∼92,000 syntenic pseudogene pairs (∼12,000 HCov) were identified between the subspecies.

While the number of syntenic genes and pseudogenes is very similar, there is a significant difference when comparing element pairs that are pseudogenized in one subspecies. Wild emmer harbors 31% more pseudogenes (27% HCov) in syntenic context to a functional gene than durum wheat. This may either indicate (i) a higher pseudogenization rate of protein-coding genes in wild emmer, or (ii) a higher tandem duplication rate accompanied by pseudogenization in wild emmer, or (iii) a more rapid degeneration or deletion of pseudogenes in durum wheat. Wild emmer contains slightly more genes but ~14,000 (9%) more non-TE-related pseudogenes and gene fragments than durum wheat, thus supporting the hypothesis of a higher duplication and pseudogenization rate in wild emmer.

The functional properties of wild emmer and durum wheat pseudogenes were investigated in more detail. GO analyses were performed to identify over- and under-represented GO terms within the parent gene set compared to the complete gene set of each plant, respectively. Here, the combined gene sets of both subspecies were used to identify pseudogenes. Enriched GO terms in the pseudogene set of each subspecies were determined in comparison to the combined pseudogene sets of both. The GO annotation of parent genes was assigned to pseudogene children.

Over three times more GO terms are significantly enriched in durum wheat (Figure R26). They span a broad range of low-level functional categories including metabolic, catabolic or biosynthetic processes (Figure R26 B). More detailed higher-level functional categories are DNA repair, histone modification and processes utilizing glucose or other sugars. Over-represented GO terms in the pseudogene set of wild emmer comprise, amongst others, response to oxidative stress/stimulus or ion transport (Figure R26 A).

The previous enrichment analysis of parent GOs indicated a highly similar pseudogenization pattern between wild emmer and durum wheat (see Figure R17 in section 5.7). When focusing on pseudogenes and on the differences between two subspecies, durum wheat pseudogenes clearly cover a wider range of functional categories than wild emmer pseudogenes.

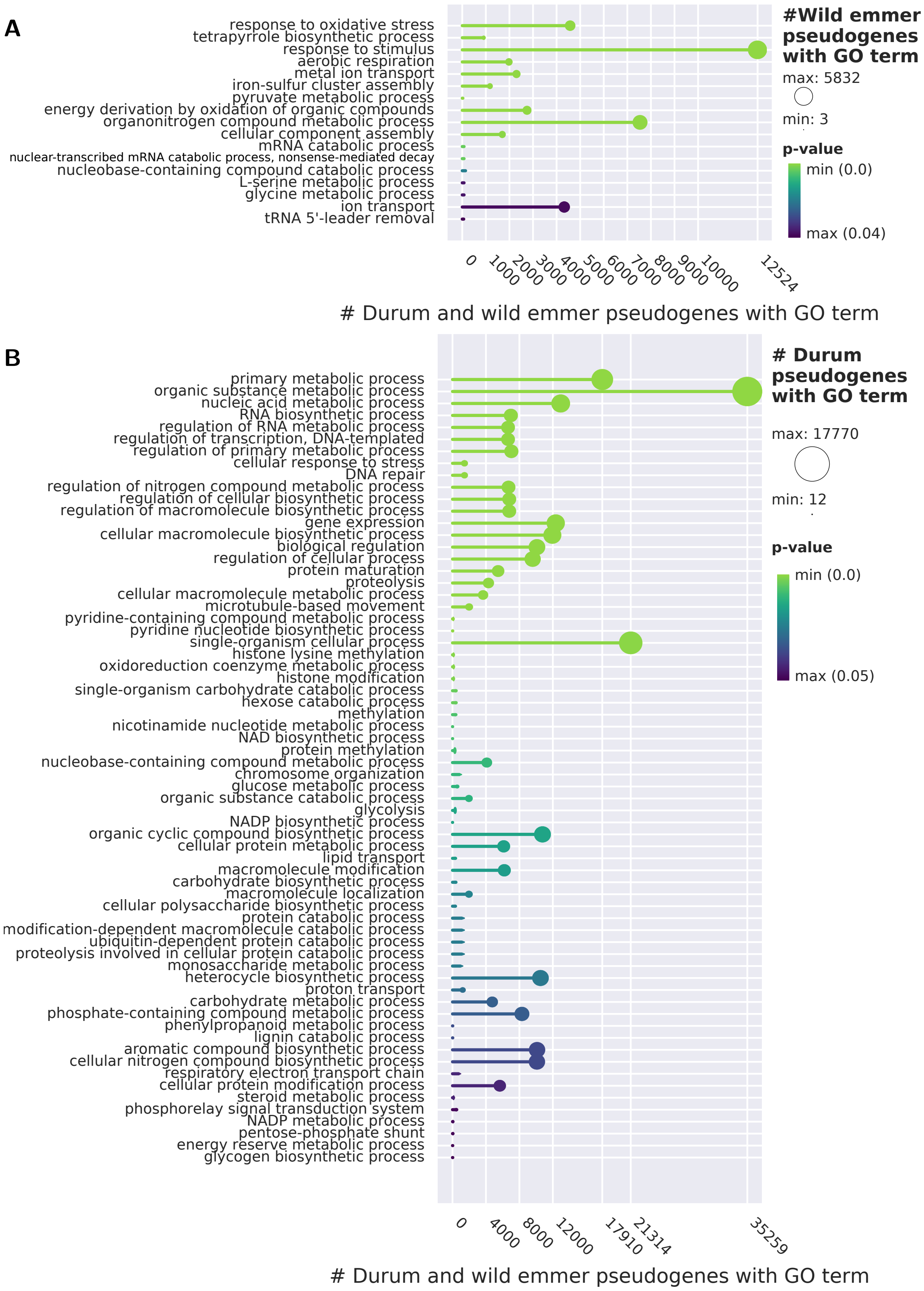


Figure R26: Over-represented GO terms of pseudogenes in wild emmer (A) and durum wheat (B) compared to the combined pseudogene set. GO terms of parent genes are assigned to pseudogenes. Enriched terms on one subspecies are determined by comparison to the combined pseudogene set of both subspecies.

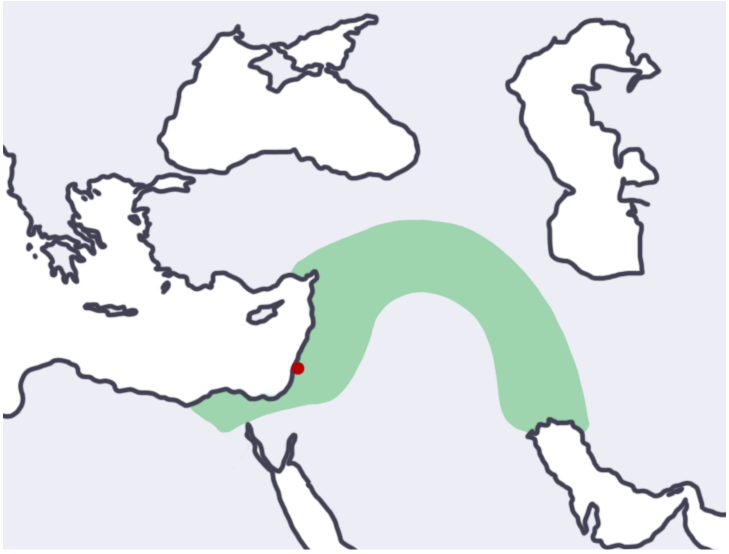


Figure R27: The Evolution Canyon (red) within the Fertile Crescent (green).

5.12.4 Morex barley and two wild barley accessions from the Evolution Canyon I in Israel

Cereals were domesticated approximately 10,000 years ago during the Neolithic (Haberer et al., 2016). The Fertile Crescent is considered the birth place of western agriculture and it is located east of the Mediterranean Sea. One so-called 'Evolution Canyon' is located in Israel (Figure R27).

The contrasting climates on opposing slopes of these canyons make them popular research subjects. The slopes of the 'Evolution Canyon' I in Israel are only separated by ∼250 m but nevertheless exposed to drastically different microclimates (Prade et al., 2018). While the south-facing slope (SFS) of the canyon displays tropical climate — with up to 800% more solar radiation and increased temperature and drought — the north-facing slope (NFS) displays a more temperate climate that is cooler, mesic and forested (Nevo, 2012). Evolution Canyons are frequently used to study climate change and the effect of differing microclimates on the evolution and adaption of various species.

Four wild barley accessions have been used to investigate pseudogene differences between closely related barley subspecies and cultivars. Two of the wild barley accessions originate from the two opposing slopes of the 'Evolution Canyon' I in Israel. The other two are Tibetan wild barley accessions. All four genome assemblies are contig/scaffold assemblies without available pseudomolecule sequences. Before quality filtering, 20–40% of the genome sequence was unknown (Ns). Sequences were filtered for a minimal contig length of 200 bp and a maximum N-content of 35%. The resulting quality filtered contigs and scaffolds have an N50 value of 10–15 kbp. Quality differences between the BAC-by-BAC Morex assembly and the four wild barley assemblies are pronounced.

The HC gene set of Morex barley was used to assess gene-like sequences in the four wild barley accessions. Wild barley genes were assessed via strong homology to Morex genes. Homologous regions were identified, manually scrutinized and regions of interest were selected for in-depth analysis. Regions were selected when featuring pseudogenization or structural differences.

The first region of interest contains pseudogenized genes in wild barley and a triplet of genes that is duplicated in the SFS accession (Figure R28 A). The tandemly duplicated region comprises two leucin-rich repeat protein kinase genes and a Hexosyltransferase. LRR-RKs are a large protein family in plants that regulates developmental and defense-related processes (Torii, 2004). One of the copied *LRR-RK* genes in the SFS accession is fragmented, while the other does not exhibit functional impairments. However, the *LRR-RK* homologue in the NFS accessions is pseudogenized due to a 13 bp deletion within the 5' half of the CDS. Directly adjacent to the duplicated region is a polyphenol oxidase gene (*PPO*), which is pseudogenized in both wild barley accessions from the Evolution Canyon, but found intact in the cultivated barley sequence. *PPO* genes are enzymes responsible for the browning reaction of damaged plant tissues (Tran et al., 2012). They have been suggested to take part in defense-related mechanisms as well.

The second selected region contains a functional calcium-binding protein gene (*CABP*), which is pseudogenized only in the wild barley accession from the NFS (Figure R28 B). In human, such *CABP* genes are enriched in neuronal tissue and act as important regulators of key calcium influx channels (Haynes et al., 2012). In plants, calcium plays an essential role for the reaction to pathogen attacks or other stimuli, as it is an important messenger of external signal transduction cascades (Poovaiah et al., 1993). Additionally, calcium is involved in photosynthesis, carbon fixation, CO_2 fixation, protein transport and protein phosphorylation (Rocha and Vothknecht, 2013). The *CABP* pseudogene in the NFS accession contains a 1-bp deletion right at the beginning of the CDS which is immediately followed by a PTC. Interestingly, this 1-bp deletion is also present in the *CABP* gene of the SFS accession. However, another upstream 1-bp insertion restores the frame without introducing PTCs. According to maximum parsimony, the gene likely pseudogenized in the precursor of both accessions and the counteracting mutation subsequently restored the frame in the SFS accession. Gene and pseudogene states may transition frequently and the introduction of frameshifts that affect smaller stretches of protein-coding sequences may even lead to functional novelty.

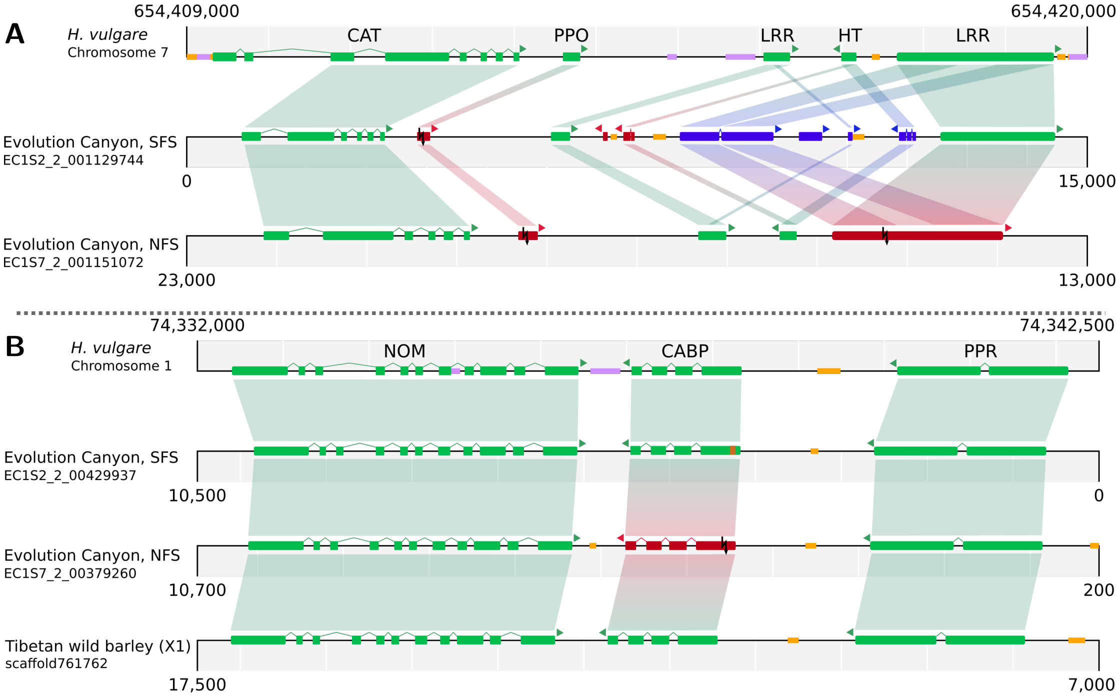


Figure R28: Homologous regions between cultivated Morex and wild barley accessions. Two chromosomal regions are illustrated containing genes (green), pseudogenes with PTCs (red) and potential pseudogenes without PTCs (blue). Regions containing unknown sequence (yellow) and repetitive elements (violet) are highlighted. The locations of the first PTC per pseudogene are marked with a lightning symbol. A: A region on chromosome 7H of Morex barley contains genes that are pseudogenized in both accessions from the Evolution Canyon but only wild barley from the south-facing slope (SFS) contains a tandemly duplicated gene triplet. B: A region on chromosome 1H of Morey barley contains three genes. The *CABP* gene contains a 1 bp deletion in both accessions from the Evolution Canyon. While this deletion leads to a PTC in the accession from the north-facing slope (NFS), the frame is restored in the SFS gene by a subsequent 1 bp insertion. Tibetan wild barley contains an error-free *CABP* gene. Abbreviations: 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).

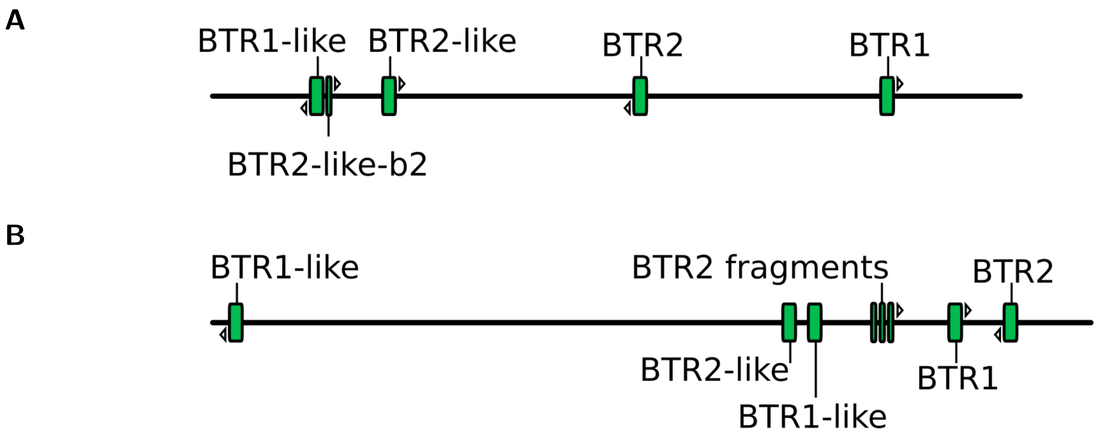


Figure R29: BTR locus on chromosome 3H of barley. A: Schematic representation of the published *BTR* locus (NCBI, GenBank: KR813336.1). B: Schematic representation of the *BTR* locus found in the *H. vulgare* cv. Morex genome assembly.

5.13 The BTR locus in barley and how sequencing and assembly strategies affect data quality

The Brittle Rachis (BTR) locus is of particular interest in wild and domesticated cereals, as the genes within this locus are important for the grain dispersal system. In wild cereals, awns containing ripe grains fall to the ground as the rachis becomes brittle. Early farmers selected plants with a dysfunction in this seed dispersal system making harvesting seeds much easier. Only two mutations in adjacent, dominant and complementary genes are necessary for the rachis to turn into a non-brittle form. (Pourkheirandish, Hensel, et al., 2015)

The locus comprises *BTR1*, *BTR2* and several *BTR1/2*-like elements (Figure R29 A). A homology search of the published nucleotide sequences was performed to identify the genes in the Morex barley genome assembly. As expected, *BTR1* and *BTR2* are not within the HC gene set. However, a pseudogene is annotated at the *BTR1* gene position. This pseudogene was detected via homology to the annotated *BTR1*-like parent gene. *BTR2*-like elements are not in the HC gene annotation. Thus, without a suitable parent gene present, the *BTR2* pseudogene could not be identified with the PLIPipeline.

Results are often influenced by assembly quality differences, that render comparative analyses challenging. Repetitive and TE-rich regions are only assembled for plants with smaller genomes or for those with high-quality BAC-by-BAC or NRGene assemblies. High-quality *Triticeae* genome sequences assembled into com-

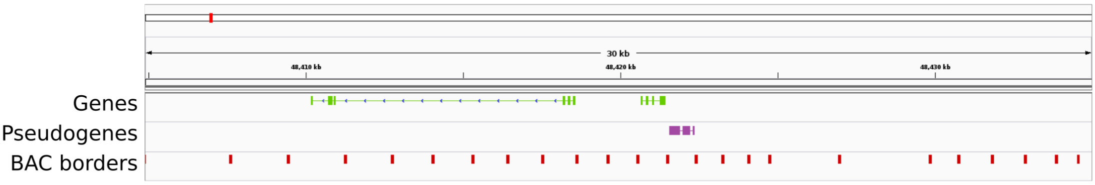


Figure R30: Region on Chromosome 3H of the Morex barley assembly. Tracks include genes (green), pseudogenes (violet) and BAC borders (red).

plete pseudomolecules were available only for Morex barley, wild emmer, durum wheat and bread wheat. The Morex barley assembly was done with a BAC-by-BAC approach, while the other three are NRGene assemblies (International Barley Sequencing Consortium, 2017; Avni et al., 2017; Maccaferri et al., 2018; International Wheat Genome Sequencing Consortium, 2018). The large-scale structure of the barley chromosomes is of very high quality. However, the small-scale structures show irregularities. For example, elements from the *BTR* locus are not ordered or spaced correctly in the Morex barley genome assembly (Figure R29 B).

This small-scale structural problem occurs due to inaccurate BAC contig ordering. Frequently, larger BAC contigs are ordered correctly, while smaller ones are sorted by size and inserted at the approximate position of the BAC (Figure R30). As a consequence, genes or pseudogenes may be unordered, fragmented or contain unusually large introns that span several BAC contigs.

When comparing Morex barley to the four wild barley accessions, several syntenic regions attracted attention due to apparent rearrangements and TE insertion (Figure R31). Although some of these structural peculiarities might be genuine and potentially the consequence of domestication, they are more likely ascribable to BAC-by-BAC assembly problems.

5.14 Chromosome conformation during Interphase may affect pseudogene distribution in *Triticeae*

Retroposed pseudogenes are the result of retrotransposition. Duplication occurs via an mRNA intermediate, which is reverse-transcribed into cDNA and reintegrated into the genome sequence. The reintegration step was assumed to happen at a random location (Z. Zhang, Paul M. Harrison, et al., 2003; Wen Wang et al., 2006; Williams et al., 2009). The distribution of retroposed pseudogenes suggests a different scenario: Retroposed pseudogenes in *Solanaceae* and *Triticeae* are prevalently located on the same chromosome as their parent gene (Table R6). Since polyploid species contain multiple diploid gene sets, it is possible that par-

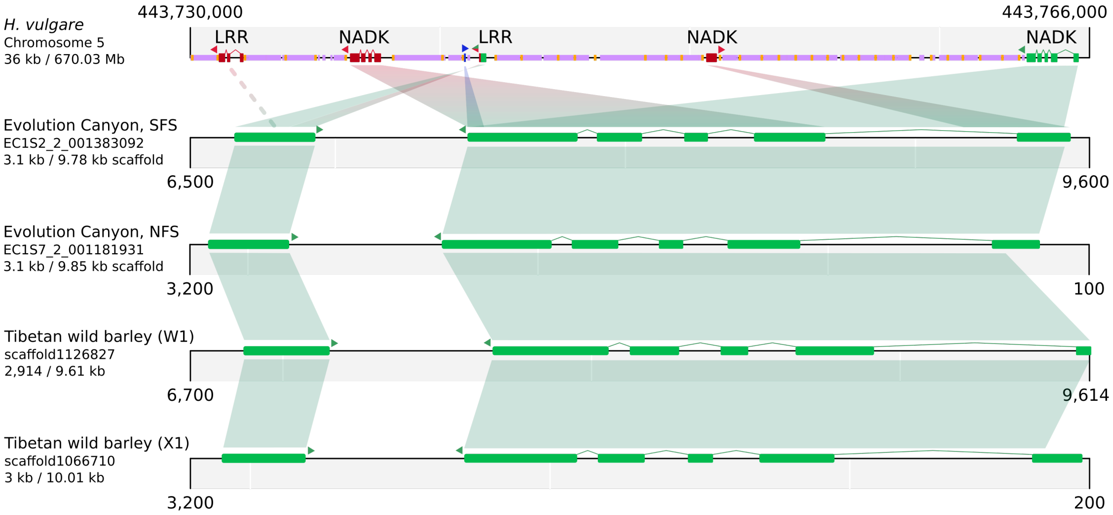


Figure R31: Syntenic region between cultivated Morex and wild barley accessions. The chromosomal regions contain genes (green), pseudogenes with PTCs (red) and potential pseudogenes without PTCs (blue). Regions containing unknown sequence (yellow) and repetitive elements (violet) are highlighted. Abbreviations: Leucin-rich repeat protein kinase (LRR), NAD kinase (NADK).

Table R11: Significant accumulation of pseudogenes on the same homeologous chromosomes as the parent gene. Checkmarks indicate weather a significant result was obtained in a paired T-test (P-value $\leq$5%). FDR-correction was applied.

Species	all	HCov	duplicated	duplicated HCov	retroposed
T. dicoccoides	✓	✓	✓	✓	✓
T. durum	✓	✓	✓	✓	✓
T. aestivum	✓	✓	✓	✓	✓

ent genes from the "wrong" homeologous chromosomes were chosen. However, duplicated and retroposed pseudogenes are also prevalently located on the same homeologous chromosomes (Table R11).

Retroposed pseudogenes are not randomly distributed. Is there a mechanism that could introduce a preferential insertion into specific chromosome regions? Retrotransposition via the LINE-1 mechanism takes place inside the nucleus. The mRNA is primed for reverse transcription at a nick directly at the insertion site (Kaessmann et al., 2009). The preferential distribution of retroposed pseudogenes in close spatial proximity suggests a distance depended insertion. Interphase chromosomes of barley and other plants with large genomes have been shown to adopt a Rabl conformation (Dong and Jiang, 1998; International Barley Sequencing Consortium, 2017). The neighboring arrangement of short and long chromosome arms may impose structural constraints supporting preferential reinsertion nearby or on opposing chromosome arms (Figure R32).

However, during the analysis of pseudogenes that have been classified as retroposed due to the absence of intron sequences, evidence accumulated that some of them may have lost their introns subsequent to duplication. Alternatively, their parent genes may have gained introns. Hence, such alleged "retroposed" pseudogenes should rather be dubbed "processed" pseudogenes. Those pseudogene likely originated from unequal crossing over or DNA repair mechanisms, but they appear retroposed due to the absence of introns. One mechanism that could account for intron loss is the participation of mRNA or cDNA in gene conversion events (H. Wang et al., 2014). This would explain why processed pseudogenes often exhibit features of duplicated pseudogenes (e.g. homology beyond the UTR). Again, the Rabl conformation of *Triticeae* chromosomes could impose structural constraints on gene conversion events between parent gene transcripts and pseudogene children.

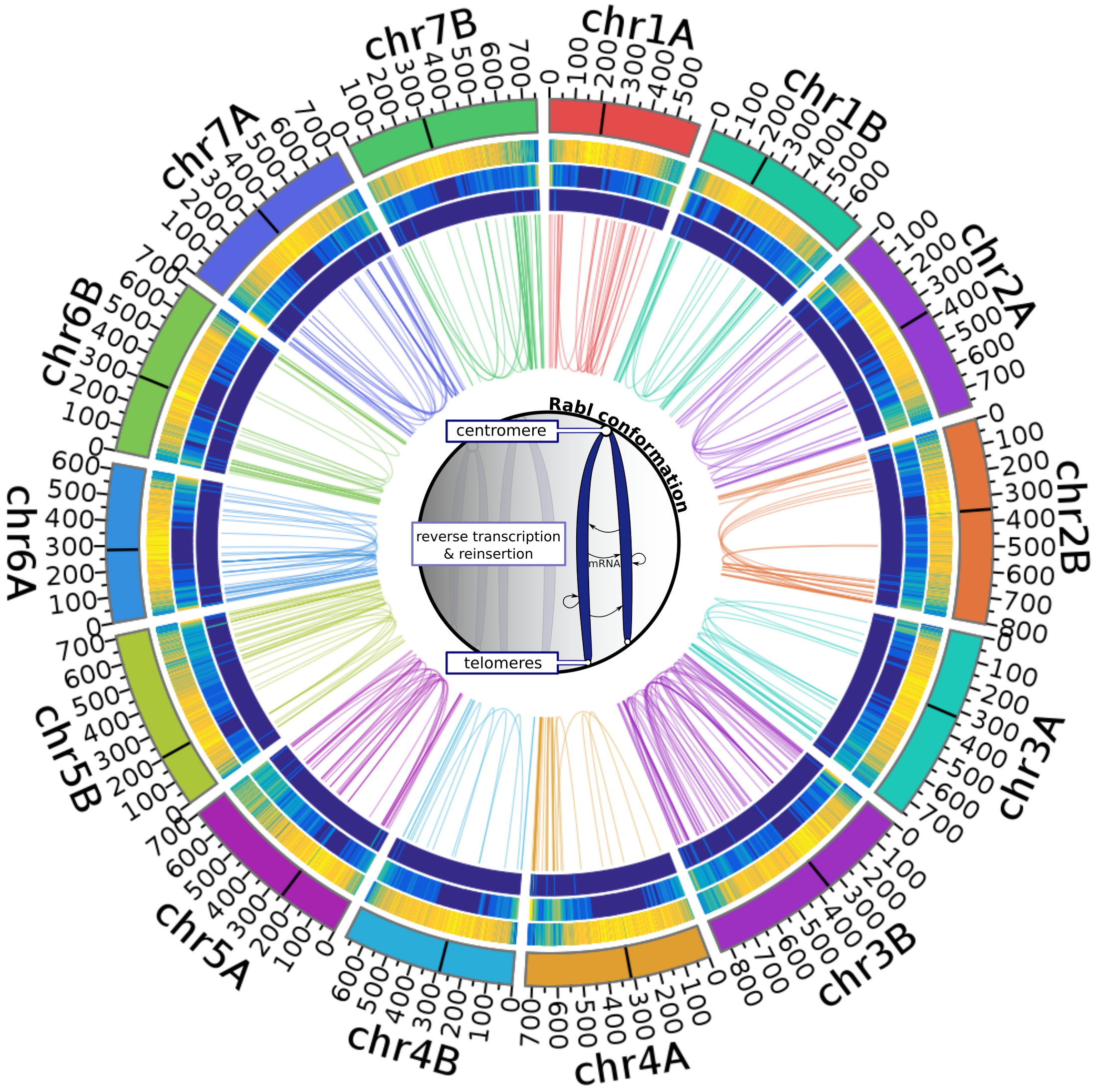


Figure R32: Retroposed pseudogenes located on the same chromosome as their parent gene in wild emmer. Heatmap tracks show the distribution of TEs, genes and pseudogenes. Links connect pseudogenes with parent genes. A schematic representation of the Rabl conformation and its putative effect on the distribution of retroposed pseudogenes is shown in the center.

5.15 Closing the gap — reducing the amount of "gray area"

Larger plant genomes are highly repetitive and TE-rich (Figure R33). More than 80% of the bread wheat genome is occupied by TEs. In plants with smaller genomes, TEs accumulate close to the centromere and their distribution generally mirrors the chromosomal gene space. While TEs and genes occupy much space in plant genomes, there is still a lot of empty space or "gray area" for which no annotation is available. Those regions harbor highly degenerated TEs, ncRNAs, regulatory elements or "junk DNA".

The annotation of pseudogenes represents one further step to close this annotation gap and to unlock the full potential of genome analysis. Although pseudogenes do not rival genes nor TEs when it comes to the amount of annotated sequence, their extremely high numbers, functional potential or role during evolution make them an important research subject.

The quality of the genome sequence and annotation is of utmost importance when attempting comparative analyses. Filtering of TE-genes from the query gene set strongly affects pseudogene numbers. For example, barley contains more than 8 million pseudogenes prior to TE filtering, while wild emmer, durum wheat or bread wheat only contain 1.8–2.7 million pseudogenes. Most of them overlap with TE annotations and are subsequently filtered. They were detected, because the barley gene set contains TE genes, for which thousands of copies could be identified (s. Figure R2).

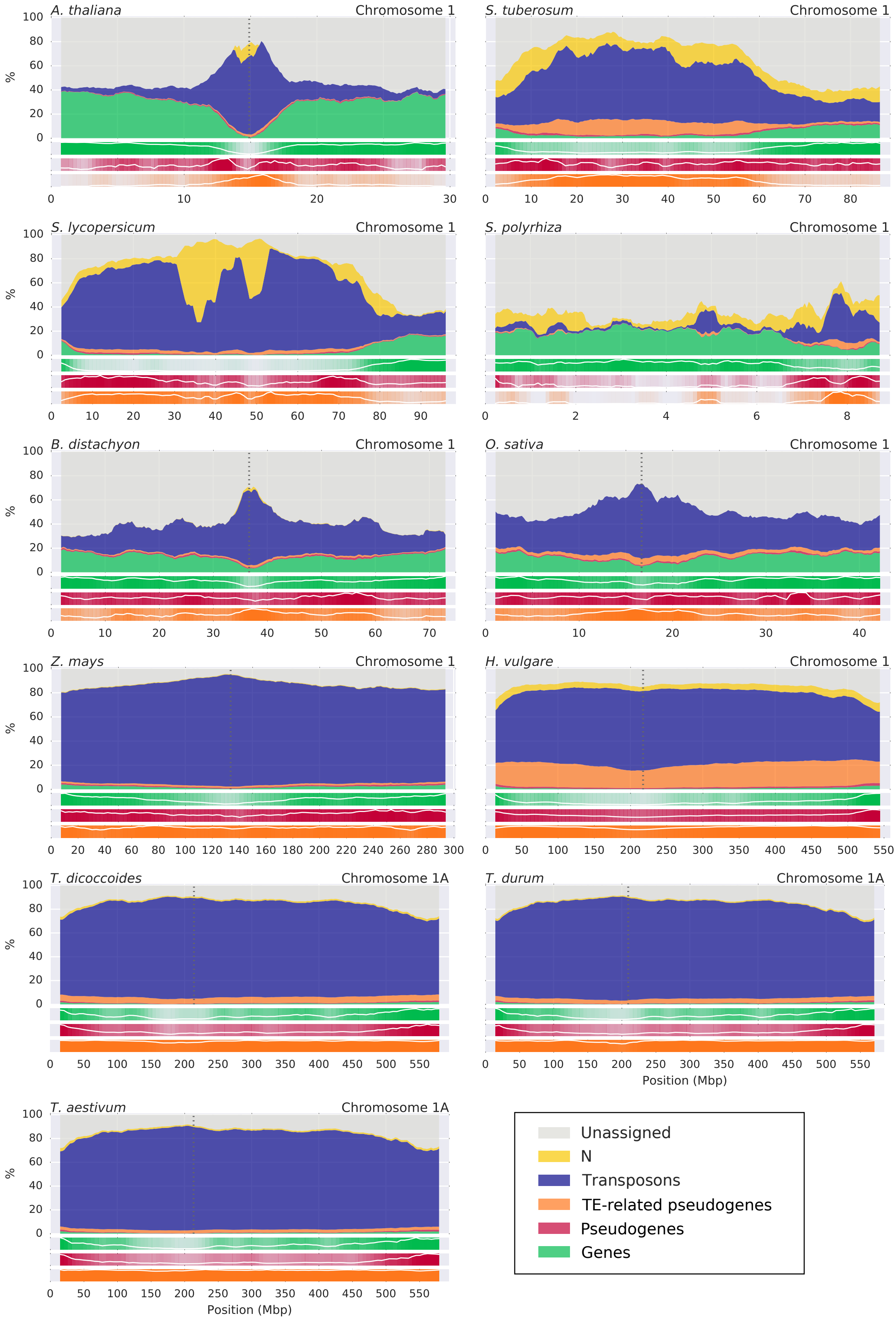


Figure R33: Chromosome composition. Selected chromosomes and their composition of TEs, genes, pseudogenes, TE-related pseudogenes and undefined sequence (N stretches). Centromere positions are indicated with a dotted vertical line if available.

6 Discussion

Pseudogenes are often referred to as "evolutionary relics" or "junk DNA". With mounting evidence for the functional potential of these gene-like sequences, they are now increasingly studied in mammals (Xiao-Jie et al., 2015). The oxymoron "functional pseudogene" is a widely used term for transcribed or translated pseudogenes. While critics are justified to argue that functional pseudogenes should be reclassified and added to official gene annotations, a discrimination between gene and pseudogene is not trivial. For conceptual reasons, non-functionality is difficult to prove, for example, since non-tested functions might be present. Transcription may happen rarely or only in very specific cell types. Sequencing or assembly errors may introduce features normally indicative for functional impairment. And although not being expressed at present, a pseudogene may be reactivated, adopt new functions or participate in gene conversion events. Thus, pseudogenes possess a hidden potential to "shape an organism during evolution" (Brosius and Gould, 1992).

For many plants, comprehensive annotations and analyses of pseudogenes were hampered due to their large and complex genomes. Also, until very recently the classical main focus of any plant genome analysis was the detection and description of classical *bona fide* genes. Recent technological advancements — especially in sequencing and genome assembly — opened up exiting new possibilities. With full-length pseudomolecules now available for many economically important crop plants like barley (5.4 Gbp), durum wheat (12.4 Gbp) or bread wheat (16.9 Gbp), comprehensive pseudogene annotations are now feasible.[10] In this work, pseudogenes of 18 plant genomes featuring different genome sizes, complexities and economical significance were annotated and analyzed (Table R1). Most of the plants are monocots from the *Poaceae* family and the *Triticeae* tribe. Additionally, three eudicotyledons (dicots) — *Arabidopsis thaliana*, *Solanum tuberosum* (potato) and *Solanum lycopersicum* (tomato) — are investigated.

The Pseudogene Locus Identification Pipeline (PLIPipeline) was developed to annotate pseudogenes using a homology-based approach: The coding sequence (CDS) of functional genes was mapped back to the genome sequence to identify gene-like elements (Figure M1). Hence, each pseudogene can be associated with a functional parent gene. Sequence coverage, identity or features like premature termination codons (PTCs) are determined in comparison to the parent gene (Figure M4). The pipeline is able to identify 83% of the TAIR10 pseudogenes plus many more smaller gene fragments. Pseudogenes missed by the PLIPipeline include filtered transposable element (TE)-related pseudogenes and unitary pseudogenes without parent gene in the protein-coding gene set. Given that the manually

[10] C-values are taken from `http://data.kew.org/cvalues`, (April 2, 2018, 12:30pm CEST)

curated *Arabidopsis thaliana* genome assembly and annotation are of particularly high quality, a true positive rate of 83% is respectable.

6.1 Data quality affecting pseudogene annotation

Filtering TE-related sequences reduces the number of potential pseudogenes by up to 95% (Table R2). Hence, the majority of the gene-like sequences overlap with TE annotations or have a match to the TREP repeat database. Many TE-related pseudogenes originate from only few TE genes. Their presence in the gene data sets is frequently observed and leads to these classifications. For example, before TE filtering, over eight million potential pseudogenes were found in the barley genome. The most frequent element has a copy number of over 25,000 (Figure R2). Filtering TE-related sequences reduces the number to under 400,000 pseudogenes.

Filtering TE-related sequences improves comparability and compensates for contaminations in the gene set or other data quality differences. Like in barley, the potato gene annotation contains TE genes. While potato and tomato are both from the *Solanaceae* family and thus closely related, their pseudogene number differs massively. Before TE filtering, more than three times as many pseudogene candidates were found in potato (Table R2). After filtering, the pseudogene numbers of both species are in a similar order of magnitude. Thus, high-quality TE annotations are also a fundamental requirement for a robust pseudogene annotation.

In addition to quality differences due to contaminations in the gene annotations, genome assembly coverages affect pseudogene metrics. The PLIPipeline was applied on 18 plants (Table R1). Three of them are dicots. Eight of the 15 monocotyledonous plant genomes are assembled into pseudomolecules. The remaining seven genome assemblies are contig/scaffold collections with N50 values as low as 683 base pairs. Repetitive regions are collapsed, leading to low genome coverages and reduced search space for the pseudogene detection, but also to a reduced number of TE-related pseudogenes. Additionally, gene and pseudogene annotations contain fragmented and incomplete elements, because they are only partially located on contig sequences. Fragmented annotations also have an effect on pseudogene metrics like sequence length or coverage.

Finally, even high-quality pseudomolecule assemblies are often of varying quality. The genome assembly of barley was made using a BAC-by-BAC approach (International Barley Sequencing Consortium, 2017). Chromosome conformation capture sequencing (Hi-C) was used to obtain three-dimensional proximity information and to order and orient BAC-based super-scaffolds. While the large-scale chromosome structures are of high quality, small-scale ordering of contigs can still be inaccurate. For example, the Brittle Rachis (BTR) locus on chromosome 3H was found to be rearranged and fragmented due to inaccurate contig ordering (Figure R29). Additionally, some barley genes exhibit particularly large introns due to the erroneous placement of contigs within the gene structure (Figure R30). Erroneous contig ordering can also lead to fragmented annotations and may contribute to the high number of pseudogenes observed in barley.

TE-related pseudogenes were filtered and exempted from most analyses to reduce the influence of data quality differences. Significant comparative analyses require complete and high-quality genome sequences and annotations. With high-quality genome assemblies available for several economically important crops, comparative analyses now become feasible.

6.2 Numerous pseudogenes and small gene fragments

Plant genomes contain massive amounts of TE genes, pseudogenes and smaller gene fragments. Filtering TE-related sequences reduces the number of potential pseudogenes up to 95%. After filtering, bread wheat contains 289,132 non-TE-related pseudogenes (Table R2). Only 48,608 of them are high-coverage (HCov) pseudogenes — complete copies (>80%) of functional genes. Most pseudogenes are smaller gene fragments, that correspond to less than 20% of a parent gene's CDS (Figure R5). While there is a significant amount of pseudogenes located in close vicinity to their respective parent gene, many pseudogenes and gene fragments are also randomly distributed on the chromosomes. Various duplication mechanisms can explain a random distribution of duplicated sequences:

Retrotransposition of host mRNA Retroposed pseudogenes are duplicated via retrotransposition of host messenger RNA (mRNA) (Figure I5). They were originally believed to reinsert at a random location and lose their function immediately (dead-on-arrival) due to absent nearby promoter sequences (Thibaud-Nissen et al., 2009). Reverse transcription of a mature mRNA happens directly at the integration site via products of long interspersed element 1 (LINE-1)-retrotransposons. However, classification of pseudogenes via their exon-intron structure indicates that pseudogenes with lost intervening sequences are not randomly distributed, but preferentially located close to their parent genes. Furthermore, only 1% of the pseudogenes exhibit an absence of intron sequences. Thus, retrotransposition is not the prevalent mechanism in the generation of pseudogenes. The implications of the non-random distribution of pseudogenes with lost intron sequences will be discussed further in section 6.4.

Double strand break repair mechanisms Double strand DNA break repair via non-homologous DNA end joining (NHEJ) or synthesis-dependent strand annealing (SDSA) can lead to filler DNA (Gorbunova and Levy, 1997; Wicker, Buchmann, et al., 2010). Such filler DNA can originate from an ectopic template sequence that is used to bridge the break (Figure I4). DNA repair mechanisms are particularly important for plants, because most plants are sessile, autotroph and directly exposed to biotic and abiotic stresses (Schiml et al., 2016). Furthermore, TE activity increases the frequency of double strand breaks. Class 1 elements (retrotransposons) are duplicated via an RNA intermediate, reverse transcription and reintegration into the target site (Figure I6). In contrast, class 2 elements (DNA transposons) jump to a new location via a cut-and-paste mechanism. In many organisms, double strand breaks introduced due to TE excision are repaired via gene conversion, but a deficiency in this process may cause a switch to other pathways (Mehta and Haber, 2014). For example, sequence motifs found at the borders between duplicated fragments and TEs in *Brachypodium distachyon*, rice and sorghum indicate that they are the result of double strand break repair via the SDSA mechanism (Wicker, Buchmann, et al., 2010).

Transposable elements capturing host genes Some types of TEs have been shown to capture host gene fragments. Approximately 2% of the maize genome is composed of *Helitrons* (Barbaglia et al., 2012). Almost 2,800 non-autonomous elements were computationally predicted and 94% of them were found to carry fragments of transcribed host genes (Du et al., 2009). The most abundant type of TE in plants are long terminal repeat (LTR)-retrotransposons (F. Sabot and Schulman, 2006). Insertions into their sequence will be amplified along with the element itself. Essentially, only LTRs or terminal inverted repeats (TIRs) are required for non-autonomous elements to exploit the machinery of autonomous TEs. Any sequence in between the repeat pair is duplicated.

All three of the described mechanisms — retrotransposition of host mRNA, double strand break repair and transposable elements capturing host genes — may contribute to the huge number of fragmented and randomly distributed pseudogenes. However, retrotransposition of host mRNA is not a very frequent duplication mechanism in plants. Hence, the most likely origin of most pseudogene fragments is either DNA repair or duplication via TE hitchhiking.

6.3 Pinpointing pseudogenes

Pseudogenes are usually considered dysfunctional gene-like sequences. However, it is conceptually impossible to prove non-functionality. For example, PTCs disrupt the open reading frame (ORF) of genes leading to functional impairment of the encoded proteins. Nevertheless, pseudogenes with PTCs can still be transcribed and exert regulatory effects on paralogous genes (Pink et al., 2011). Additionally, efficient translational read-through has been observed in mammals and insects (Schueren and Thoms, 2016; Prieto-Godino et al., 2016).

> In many cases, pseudogenes confer no observable selective advantage to the host organism and may be on a path towards removal from the genome. However, pseudogenes can also serve as raw material for the exaptation of novel functions, particularly in relation to the regulation of gene expression. Many pseudogenes are resurrected as noncoding RNA genes, which function in RNA-based gene regulatory circuits. As such, functional pseudogenes might simply be considered as 'genes'.
>
> — Roberts and Morris (2013)

The presence of PTCs is often used to distinguish pseudogenes and genes. For the 18 different plants, between 52 and 78% of the HCov pseudogenes contain PTCs and a similar portion contains frameshifts (Table R8). However, 11 to 29% do not exhibit any PTCs, frameshifts, insertions or deletions within their CDS. While there are other defects that can impair function, the annotated pseudogene set thus might also contain some functional elements or genes missed in the official gene annotation.

Comparing pseudogene sequences to their respective parent genes did reveal a higher rate of synonymous nucleotide changes, which is usually interpreted as indicative for selection pressure (Figure R18 A). However, the concurrent evolution of pseudogene and parent gene would also result in a higher rate of synonymous changes between the two elements (Figure R18 B). Furthermore, there is transcription evidence for at least 12% of barley's HCov pseudogenes. In Arabidopsis and rice, approximately 20% of the annotated pseudogenes have previously been shown to be transcribed (Podlaha and J. Zhang, 2010). This can be seen as further indication that some of the pseudogenes have functional potential. Alternatively, however, these cases might be examples for genes undergoing non-functionalization.

Gene annotations might also be contaminated with pseudogenes. Approximately 20,000 bread wheat genes contain features that identify them as degenerated duplicates (Table R7). Some are retrogenes, some exhibit a fragmented structure and some contain PTCs that are disrupting their original ORFs. However, they may still be transcribed and translated into functional products. For example, while retroposed pseudogenes are considered dead-on-arrival due to the loss of promoter, the insertion site may provide an alternative promoter. Gene duplicates shortened due to PTCs may still be transcribed and translated. Apart from pseudogenization, neo-functionalization or sub-functionalization are alternative pathways following gene duplication.

Thus, in consequence, the detection of pseudogenes is a non-trivial analytical task and numerous transitory states need to be considered. To a certain degree, the functional potential of pseudogenes can be determined computationally. However, a higher degree of certainty can only be achieved via an in-depth analysis of individual pseudogenes of interest. Olfactory receptor pseudogenes in the fly *Drosophila sechellia* illustrate the challenges: The *Ir75a* (pseudo)gene contains a PTC that is fixed in the population, but it encodes a functional receptor (Prieto-Godino et al., 2016). For this specific pseudogene, efficient translational read-through seems to depend on tissue type and on the sequence downstream of the PTC. The authors also present other PTC-containing loci in different species and suggest that such "pseudo-pseudogenes" may represent a widespread phenomenon.

6.4 The origin of pseudogenes

Most pseudogenes have their origin in a gene duplication event (Tutar, 2012). The resulting redundancy of information allows one copy to degenerate or adopt new functions without consequence for the original gene (Balakirev and Ayala, 2003). Genes rarely pseudogenize without prior duplication and in human only 76 unitary pseudogenes could be identified since the human-mouse divergence 75 million years ago (mya) (Z. D. Zhang et al., 2010). Most of these pseudogenes are from gene families with several members. Hence, there are still functional genes present that have a similar coding sequence. In this work, unitary pseudogenes were defined as elements with no sequence similarity to a functional protein-coding gene. Since a homology-based approach was chosen, such unitary pseudogenes cannot be detected when only using template/query gene sequences from the same species. However, by using homology for their identification, putative pseudogenes can always be associated to paralogous counterpart (parent) genes within the protein-coding gene set.

Various DNA duplication mechanisms can generate pseudogenes. A classification into duplicated and retroposed pseudogenes was attempted via structural comparisons: Duplicated pseudogenes retain their exon-intron structure, while retroposed pseudogenes lose intron sequences during retrotransposition. Only ∼1% of the HCov pseudogenes exhibit an absence of intron sequences (Table R3). In contrast, ∼34% have retained their exon-intron structure. Most of the remaining pseudogenes are copies of single-exon genes. Hence, while plant genomes are rich in TEs, they do not contain large amounts of retroposed pseudogenes. Previous studies on the origin of pseudogenes in Arabidopsis and rice provide similar results (Thibaud-Nissen et al., 2009).

There are other means to estimate the significance of retrotransposition for the origin of pseudogenes — especially since the structure-based classification is not applicable for the large amount of single-exon pseudogenes. During the duplication of retroposed pseudogenes, the reverse transcription of mature mRNA from the parent gene happens directly at the integration site via products of LINE-1-retrotransposons. Since transcription includes polyadenylation of the 3' end, retroposed pseudogenes should feature poly-adenine (poly-A) tails downstream of their CDS. Additionally, homology should be restricted to CDSs and untranslated regions (UTRs).

Only a small percentage (1.8%) of HCov pseudogenes feature a poly-A tail and most of these are not classified as processed (Table R4). Even 1.5% of HCov pseudogenes with intact exon-intron structure have a putative poly-A tail downstream of their CDS. On the one hand, this structure-independent analysis confirms that only few plant pseudogenes originate from retrotransposition. However, depending on the choice of parameters, almost all or none of the pseudogenes can be associated with a putative poly-A tail. Since the presence of poly-A tails does not significantly correlate with the structure-based classification, most of the detected poly-A regions likely are random occurrences. There are three possible explanations for this result: (i) Many processed pseudogenes did not originate from retrotransposition, (ii) poly-A tails degenerate more quickly or (iii) sequencing errors particularly affect poly-A regions.

In addition to the presence of poly-A tails, homology beyond the UTRs can be used to draw conclusions about the origin of pseudogenes. For retroposed pseudogenes, homology between pseudogene and parent gene should not extend beyond the UTRs. However, even pseudogenes without intron sequences feature homology far beyond the estimated UTR sequences (Figure R6). Hence, many of the processed pseudogenes may not have lost their introns via retrotransposition, but subsequent to duplication by other mechanisms. Intron loss may happen if mature mRNA or reverse transcribed complementary DNA (cDNA) participates in gene conversion (Vanin et al., 1980). Intron loss via cDNA has been demonstrated in yeast (Derr, 1998). Additionally, recurrent intron loss has been reported for several grasses including maize, sorghum, rice and Brachypodium (H. Wang et al., 2014). H. Wang et al. also demonstrate that intron loss events significantly outnumber intron gain events. This suggests that many processed pseudogenes did not originate from retrotransposition. Instead, the introns of duplicated pseudogenes were removed subsequently.

Results from the structure-based classification, poly-A tail assessment and homology beyond the UTR sequences all indicate a low frequency of retroposed pseudogenes in plants. Some pseudogenes that exhibit intron loss likely lost them subsequent to duplication. Instead of characterizing them as 'retroposed' pseudogenes, it would be more accurate to call them 'processed' pseudogenes. Until now, the two terms are frequently used as synonyms.

Many pseudogenes are located in tandem to their parent genes (Figure R11). Larger gene families and tandem gene clusters are prone to duplication via unequal crossing over. One consequence is the generation of pseudogenes (Figure R14). Additionally, recombination rates correlate with GC content, possibly due to GC biased gene conversion (Liu et al., 2015). The pseudogenes of most monocots exhibit a higher average GC content than genes (Figure R13). This may be because recombination rate and unequal crossing over can also be assumed to covary and numerous tandemly duplicated pseudogenes originate from unequal crossing over. The arrangement of genes in tandem clusters may promote unequal crossing over, resulting in higher duplication rates and potentially in the evolution of new functions. However, sequence homogenization strongly acts on tandem gene clusters, thereby impeding the divergence of genes within physically linked chromosomal regions (Baumgarten et al., 2003). Segmental duplications and homeologous chromosomes also contain numerous duplicated genes and pseudogenes. Compared to tandem gene clusters, their physically unlinked state may permit divergence and functional reorientation of genes.

The high abundance of duplicated vs. retroposed pseudogenes matches results obtained for Arabidopsis or rice (Thibaud-Nissen et al., 2009). However, it contradicts recent findings for the genome of *Aegilops tauschii* (G. Zhao et al., 2017): The authors were able to identify 25,893 full-length pseudogenes featuring PTCs or frameshift mutations. When including smaller gene fragments, the number of pseudogenes amounts to 267,546 elements. These metrics are in a similar range as the results obtained for diploid *Triticeae* via the PLIPipeline. However, G. Zhao et al. report that 29% of the full-length pseudogenes that have multi-exon ancestor genes had lost their introns. They thus conclude that they originated from retrotransposition. Since no detailed methodological explanation was provided on how they define intron-loss, it is difficult to make conclusions about the cause for this discrepancy. Additionally, they do not state the exact number of pseudogenes that have multi-exon parent genes. Hence, I shall offer potential explanations based on my experiences during the detection and analysis process: First, when comparing the exon-intron structure of parent and pseudogene, corresponding introns are often not located at the exact same position. To counteract mapping errors at splice sites, I consider all introns in close proximity to the corresponding splice site (+/−20 base pairs). An experiment evaluating state-of-the-art genome annotations in human demonstrated that structural prediction was only correct for 50% of the elements (Guigó et al., 2006). Hence, a less stringent identification of corresponding splice sites is necessary. Secondly, I only use introns that have a length of at least 30 base pairs. The published gene annotations of several *Triticeae* contain many genes with suspiciously small introns. For example, 1,482 bread wheat genes contain introns smaller than 20 base pairs. Since the splicing of introns requires them to form a loop structure, a minimal length requirement was used for pseudogene classification. Lastly, I performed extensive manual validation via visual inspection and comparison of exon-intron structures. This guarantees a high true positive rate for the classification via exon-intron structure.

6.5 Estimating the age of pseudogenes

The comparison of closely related species, but also the analysis of segmental duplications allowed for an estimate of pseudogene age. For example, Brachypodium and rice diverged approximately 46 mya (Bolot et al., 2009). While most human pseudogenes have been estimated to be ~40 million years old (Ohshima et al., 2003), almost no orthologous pseudogenes can be identified between Brachypodium and rice (Figure R22). Hence, after 46 million years, plant pseudogenes disappeared or are too degenerated to be identified via the PLIPipeline.

Potato and tomato diverged approximately 7.3 mya and orthologous pseudogenes are detectable (Figure R23). Similarly, most but not all of the orthologous or homeologous pseudogenes in wild emmer and durum wheat originated after the divergence of the A and B subgenome progenitors 6.5 mya (Figure R24). Also the most recent segmental duplication in rice (5 to 21 million years old) contains numerous gene-pseudogene pairs (Figure R10 A). Approximately 70 mya, maize had a paleopolyploid ancestor and experienced an additional whole genome duplication 5 to 12 mya (Schnable et al., 2009). Segmental duplications can be detected when plotting parent gene vs. pseudogene positions (Figure R10 C). Additionally, many pseudogenes in maize are distributed randomly. This may be a consequence of increased LTR-retrotransposon activity within the last ~3 million years (Schnable et al., 2009).

Recapulating these results for various plants, most pseudogenes that are identifiable via the PLIPipeline seem to be younger than 21 mya. Given that only few orthologous pseudogenes can be identified between potato and tomato, most are likely younger than ~10 mya.

Most primate pseudogenes originated ~40 mya and some have even been found to be over 80 million years old (Z. Zhang and M. Gerstein, 2003). Thus, their emergence coincides with the the surge of Alu elements (Ohshima et al., 2003). Plant pseudogenes are much younger, implying a faster degeneration of pseudogenes. The rate of molecular evolution correlates with generation time. Hence, organisms with shorter generation times are assumed to evolve faster, because of the accumulation of replication errors (Weller and Wu, 2015). Most of the plants that are analyzed in this work have generation times of less than one year. For example, *Brachypodium distachyon* is an annual grass with a life cycle of less than 4 months (Draper et al., 2001). All analyzed plants have much faster generation turnovers than primates accompanied by an increased pseudogene degeneration speed.

6.6 Functional background of pseudogenes

Functional Gene Ontology (GO) annotations of pseudogene parents were investigated to analyse for duplication and pseudogenization preferences. GO enrichment analyses were performed on each parent gene set compared to the complete gene set, respectively. Functional categories that are enriched in most plants comprise translation, photosynthesis or defense response (Figure R17). Localization, biological regulation or transport are under-represented GO terms. Amongst others, the GO term ‘translation’ is assigned to ribosomal protein genes. In human, ribosomal protein-like elements constitute more than 2,000 pseudogenes (Z. Zhang, P. Harrison, et al., 2002). Apart from their role in ribosome assembly and translation, ribosomal proteins have been suggested to contribute to the activation of a tumor suppressor pathway or in various physiological and pathological processes (Zhou et al., 2015). Only few ribosomal protein pseudogenes have been reported in plants (M. E. Byrne, 2009). However, the latter statement is largely based on results for *Arabidopsis thaliana*, one of the few plants where no enrichment could be determined for this GO term.

Other functional categories that are over-represented for parent genes are defense response, response to wounding and response to stimulus. Defense processes profit from high gene duplication rates to quickly adapt to changing threats. Hence, the large number of defense-related pseudogenes could be a symptom of high duplication rates caused, for example, by structural characteristics. Larger gene families tend to have more pseudogenes. Additionally, they are often clustered in several tandem situations (Figure R15). The largest class of resistance genes in flowering plants are *NBS-LRR* genes (nucleotide-binding site and C-terminal leucine-rich repeat containing), which have been shown to be distributed and separated via tandem and segmental duplications (Leister, 2004). In general, fast-evolving genes that exhibit a high duplication rate may generate more pseudogenes than highly conserved genes.

6.7 Chromosome conformation and pseudogene structure

All pseudogenes with absent intron sequences were originally classified as retroposed pseudogenes. However, several findings suggest that many may have lost their introns subsequent to duplication (section 6.4). Alternatively, their parent genes may have gained introns. Processed pseudogenes were believed to be distributed randomly. Instead — in *Triticeae* and *Solanaceae* — they are preferentially located on the same chromosome as the parent gene (Table R6). Either the reversely transcribed mRNA is not reintegrated at a random position, or these pseudogenes originated from unequal crossing over or DNA repair mechanisms and lost their introns subsequently.

The chromosomes of barley have been shown to adopt a Rabl conformation during Interphase (Dong and Jiang, 1998; International Barley Sequencing Consortium, 2017). The neighboring arrangement of short and long chromosome arms may impose structural constraints and support preferential reinsertion of retroposed pseudogenes near or on the opposing chromosome arm to the respective parent gene (Prade et al., 2018). For example, degradation of mRNA may reduce the probability of retrotransposition with increasing distance.

However, many processed pseudogenes exhibit features that suggest an origin other than retrotransposition. To account for the non-random distribution of processed pseudogenes, the mechanism leading to the deletion of intron sequences has to be affected by structural constraints as well. One mechanism that could account for intron loss is the participation of mRNA or cDNA in gene conversion events (Derr, 1998; H. Wang et al., 2014). Again, the Rabl conformation of *Triticeae* chromosomes could impose structural constraints on gene conversion events between parent gene transcripts and pseudogene children. This would explain why processed pseudogenes often exhibit features of duplicated pseudogenes (e.g. homology beyond the UTR).

6.8 Genome size, repetitivity and polyploidy

In plants, the number of genes and pseudogenes are positively correlated with genome size (Figure R3). Genome size mainly depends on the amount of repetitive sequence and TEs. For example, *Triticeae* like barley (5.4 Gbp), durum wheat (12.4 Gbp) and bread wheat (16.9 Gbp) have huge genomes, but over 80% consists of transposable elements (International Barley Sequencing Consortium, 2017; International Wheat Genome Sequencing Consortium, 2018). Even after filtering TE-related pseudogenes, the number of non-TE-related pseudogenes increases with genome size.

Polyploidy — the presence of multiple subgenomes — significantly affects the number of both genes and pseudogenes. Duplicated sequences represent a redundancy of information, leading to reduced selection pressure. While gene function can be maintained via dosage effect, especially in plants, their expression can also be regulated via dosage compensation (Edger and Pires, 2009; Heslop-Harrison and Schwarzacher, 2011). The subgenomes of polyploid *Triticeae* harbor many pseudogenes in homeologous context and there is no indication for subgenome-dependent unilateral gene loss (Figure R10). However, in wild emmer, durum wheat and bread wheat, subgenome B contains significantly more pseudogenes (Table R9). Since subgenome B also contains more genes, this would point to an increased duplication rate that is accompanied by increased pseudogenization. Subgenome D of bread wheat is the smallest subgenome and was only added approximately 10,000 years ago. Hence, the short time frame since polyploidization can be expected to result in a smaller number of pseudogenes on subgenome D. While this could not be confirmed, a K_A/K_S ratio analysis revealed that pseudogenes on subgenome D are younger, than pseudogenes on subgenome A or B (Figure R20). In the long run, polyploidization leads to pseudogenization. Additionally, subgenome-dependent duplication and pseudogenization rates are observed in polyploid plants.

The three subgenomes of bread wheat show preferential pseudogenization patterns. For example, compared to subgenome D, subgenomes A and B contain many pseudogenes with parent genes involved in defense response or response to oxidative stress (Figure R21). Subgenome D has previously been shown to exhibit pronounced response to the fungal plant pathogen *Fusarium graminearum* and also contributed significantly to overall defense response (Nussbaumer et al., 2015). Hence, this subgenome dominance correlates with increased pseudogenization in the other subgenomes.

6.9 Barley cv. Morex and wild barley accessions

Human sedentism and domestication of plants and animals first occurred approximately 10,000 years ago during the Neolithic (Haberer et al., 2016). The birth place of western agriculture is located east of the Mediterranean Sea in the Fertile Crescent (Figure R27). So-called Evolution Canyons are used to study microclimates, global warming, biodiversity divergence, adaptation and sympatric speciation (Nevo, 2012). Evolution Canyons have opposing slopes, that exhibit drastically different microclimates. For example, the south-facing slope (SFS) of the Evolution Canyon I in Israel exhibits up to 800% more solar radiation and increased temperature and drought, while the north-facing slope (NFS) displays a more temperate climate (Nevo, 2012).

Also, the Tibetan Plateau was suggested to be one of the centers of domestication of cultivated barley (Dai et al., 2012): Tibetan wild barley and wild barley from the Near East diverged approximately 2.8 mya, but some modern barley cultivars exhibit a closer relationship to Tibetan wild barley than to wild barley from the Near East. Hence, some barley cultivars (e.g. Chinese hulless barley) seem to descend from Tibetan wild barley. However, a recent analysis of domestication genes of agriocrithon barley disagrees with this hypothesis (Pourkheirandish, Kanamori, et al., 2018): *Hordeum vulgare* subsp. *vulgare* f. *agriocrithon* is a six-rowed spike barley with brittle rachis, whose origin is disputed. In the study, all of the collected material from Tibet indicate that its origin is a hybridization of two six-rowed barley landraces, followed by recombination and thereby restoration of the brittle rachis.

Domestication and adaptation to different microclimates influence gene evolution. The pseudogene complements of the domesticated barley cultivar Morex and two wild barley accessions from the opposing slopes of the Evolution Canyon I in Israel were compared (Prade et al., 2018). Additionally, two Tibetan wild barley accessions were analyzed. The comparison of Morex barley and the four wild barley accessions, as well as the analysis of the BTR locus are hampered by drastically different assembly qualities. The Morex barley pseudomolecules cover 4,833 Gbp (89%) of the genome. In contrast, the wild barley assemblies have N50 values between 10 and 15 kilobase pairs (kbp). Even after quality filtering, the sequences still contain 10 to 20% of unknown sequence. After mapping the high-confidence genes (HC genes) of barley cv. Morex onto the genome assemblies of the four wild barley accessions, only 10,832 to 20,859 high-quality genes could be identified. Hence, whole-genome comparative analyses of pseudogene complements became non-feasible due to the strong influence of data quality. Instead, ∼200 syntenic regions were identified and manually scrutinized for pseudogenization differences. Indeed, a number of functional Morex genes were found pseudogenized in wild barley accessions. A particularly interesting region contains pseudogenenized Polyphenol oxidase genes in wild barley from the Evolution Canyon (Figure R28 A). Additionally, a tandem duplication of three consecutive genes is only present on the accession from the SFS of the Evolution Canyon. Another region of interest reveals a mutation in a Calcium-binding protein gene on the two wild barley accessions from the Evolution Canyon, but not in Morex barley or Tibetan wild barley (Figure R28 B). However, while the mutation introduces a frameshift, another nearby mutation in the SFS gene restores its reading frame and prevents the disruption of the ORF by a PTC. While these small-scale examples illustrate possible pseudogenization differences in separated barley lineages, no comprehensive conclusions about whole-genome metrics can be drawn. Additionally, Morex barley pseudogenes can only be identified if another gene of similar sequence is present in the Morex gene annotation. This introduces a bias and most of the scrutinized syntenic regions only exhibit pseudogenes in wild barley accessions. Potential pseudogenes in Morex barley, that are homologous to a gene in wild barley, may have escaped detection.

6.10 Wild emmer and durum wheat

Wild emmer and durum wheat are both subspecies of *Triticum turgidum*. They are closely related and only diverged approximately 10,000 years ago (Avni et al., 2017). Durum wheat is an economically important cultivated cereal that is used for pasta production (International Wheat Genome Sequencing Consortium, 2014). The genomes of both plants are tetraploid and have a size of approximately 12 Gbp harboring ~67,000 genes and ~25,000 HCov pseudogenes (Tables R1 and R2).

The two subspecies are of particular interest since insights into gene-pseudogene complements can shed light on gene evolution and pseudogenization during domestication. Compared to barley cv. Morex and the four wild barley accessions, the genome sequences of wild emmer and durum wheat are of very high quality. Both the assemblies and gene annotations were created analogously — increasing the informative value of comparative analyses.

Most of the syntenic pseudogenes in wild emmer and durum wheat originated after the divergence of the A and B subgenome progenitors ~6.5 mya, but before the divergence of wild emmer and durum wheat (Figure R24 and R25). Additionally, wild emmer and durum wheat contain 9,620 lineage-specific genes. The analysis of these lineage-specific genes is published in Maccaferri et al. (2018):

Approximately a quarter of the 9,620 lineage-specific genes were found to be fragmented or structurally altered HC genes that may still be functional, but that show signs of pseudogenization or functional divergence. Another quarter represents either deleted or highly degenerated elements. Finally, 1,539 (32%) and 1,095 (23%) lineage specific genes were mapped to low-confidence genes (LC genes) or pseudogenes, corresponding to approximately 2.3% and 1.6% of the total wild emmer and durum wheat genes that are degenerated and pseudogenized in the other lineage, respectively. They represent unitary pseudogenes and may be targets relevant for domestication. The relatively high number of lineage-specific (pseudo)genes, suggests a quick process that may be amplified by domestication and breeding.

The functional background of pseudogenes in wild emmer and durum wheat differ. In wild emmer, only 17 functional categories are over-represented, including, but not limited to: response to oxidative stress, response to stimulus, nitrogen compound metabolism and ion transport (Figure R26). In contrast, over 60 functional descriptions are over-represented in the pseudogene complement of durum wheat. Many are general, low-level categories like metabolic, catabolic or biosynthetic processes. Other categories include DNA repair, histone modification or a number of sugar-related processes. Domestication might affect a range of diverse genes, either due to increased duplication rates, or due to progressed gene loss. While breeding often targets specific phenotypic traits like yield or nutritional value, various other traits are potentially not selected for and their loss is thus accepted.

6.11 Reactivating pseudogenes

Artificial selection may have led to genetic hitchhiking and the loss of some functions, not because the loss of a gene is beneficial, but because it is genetically linked to a locus associated with favorable traits. For example, breeding resulted in high-yielding wheat varieties, but reduced genetic diversity (Khush, 2001). Domestication influences physical/morphological traits and nutritional content, but often also leads to impaired plant defenses (Chen et al., 2015). Optimizing yield has increased the reliance on artificial crop protection (Mitchell et al., 2016). Gene duplicates involved in defense response are over-represented in the pseudogene sets of almost all plants. Hence, their reactivation may establish new or restore lost defenses. Reactivating pseudogenes may only require small changes — for example via the CRISPR/Cas9 system. In Germany — since point mutations can be the result of natural processes — organisms edited via the high-precision technique CRISPR/Cas9 are not considered genetically modified organisms.[11] Hence, reactivating pseudogenes may be a way to (re)introduce beneficial functions without applying large-scale genome editing.

In addition to reactivating pseudogenes by restoring their previous state, their diverged sequence may be artificially expressed. In 2014, Shidhi et al. computationally estimated the consequences of artificially expressing pseudogenes into novel proteins. For this purpose, they assessed 16 full-length *Saccharomyces cerevisiae* pseudogenes that are not interrupted by PTCs. Most of the pseudogenes were predicted to encode enzymes with stable tertiary structures. However, further experimental validation is necessary to validate functional prediction of translated pseudogenes. Still, utilizing the gene-like characteristics of pseudogenes through reactivation may present new opportunities for breeding.

6.12 Pseudogene origin in plants and mammals

The study of mammalian pseudogenes has attracted more attention than the study of pseudogenes in plants. Searching the PubMed database[12] returns 7,809 results for “pseudogene”, 944 results for “pseudogene plant” and 4,797 results for “pseudogene human”. Human pseudogenes are actively explored in the context of diseases (Pink et al., 2011; Roberts and Morris, 2013). The human genome comprises

[11] `https://www.bvl.bund.de/SharedDocs/Downloads/06_Gentechnik/Stellungnahme_rechtliche_Einordnung_neue_Zuechtungstechniken.pdf?__blob=publicationFile&v=11`, (April 9, 2018, 3:20pm CEST)

[12] `https://www.ncbi.nlm.nih.gov/pubmed`, (April 12, 2018, 10:25am CEST)

~3.2 Gbp of sequence, ~20,000 to 25,000 protein-coding genes and ~15,000 to 20,000 pseudogenes (Torrents et al., 2003; The International Human Genome Sequencing Consortium, 2004; assembly GRCh38.p12 data[13]). Pseudogenes can now be used as diagnostic or prognostic markers or to determine cell identity (Poliseno, Marranci, et al., 2015). Cancer cells were shown to contain differentially expressed pseudogenes. Additionally, approximately 12% of the human pseudogenes are transcribed and half of them are conserved and under significant selection pressure (Khachane and Paul M. Harrison, 2009). Some pseudogenes have even been shown to be translated into peptides (Kim et al., 2014).

For many plants, the study of pseudogenes was hampered due to their large and complex genomes. Hence, for a long time, their functional potential remained unexplored. In this work, genome-wide assessments of pseudogenes were performed for 18 different plants. Numerous pseudogenes and gene fragments were identified (Table R2). The vast majority of pseudogenes that originated from multi-exon ancestor genes have an intact exon-intron structure (Figure R4). Hence, most pseudogenes are duplicated and non-processed. Only 1% of the pseudogenes have lost their introns. While this is usually an indicator that they originated from retrotransposition, other features like the lack of poly-A tails, homology beyond the UTRs or their distribution on the genome indicate that they are not retroposed. Instead, they may have lost their intron sequences or their parent genes have gained intron sequences subsequent to duplication. Instead of calling them “retroposed”, it seems more accurate to call them “processed” pseudogenes.

In human, 70% of the pseudogenes are of retrotranspositional origin (Torrents et al., 2003). However, retroposed genes are generally considered non-functional pseudogenes instead of potentially functional retrogenes (Xu and J. Zhang, 2015). The high percentage of retroposed pseudogenes in mammals may be the result of a completely different composition of TEs. For example, the most abundant type of TE in *Triticeae* are LTR-retrotransposons. More than 75% of the barley genome consists of LTR-retrotransposons, compared to less than 1% of non-LTR-retrotransposons. In contrast, the most abundant type of TE in human are non-LTR-retrotansposons (33.7%) (Cordaux and Batzer, 2009). Most of the non-LTR-retrotransposons are LINE-1-retrotransposons, which have been shown to provide the machinery for the retrotransposition of host mRNA (Pavlicek et al., 2006). Hence, the small number of LINE-1-retrotransposons in plants may be the reason for the small number of retroposed pseudogenes. However, since plants contain large numbers of duplicated pseudogenes — which are more rare in mammals — other duplication mechanisms seem to be more prevalent. Mammals might tolerate retroposed pseudogenes better than duplicated pseudogenes, because there is a

[13] `https://www.ensembl.org/Homo_sapiens/Info/Annotation`, (April 10, 2018, 10am CEST)

higher chance for them to be dead-on-arrival. On the other hand, plants might tolerate duplications better due to their fundamentally different mechanism for dosage compensation.

7 Outlook

Pseudogenes have long been considered "evolutionary relics" or "junk DNA". However, their potential role in gene evolution makes them an increasingly popular research target. In mammals, numerous alleged pseudogenes have been shown to be transcribed or to exhibit regulatory functions. The annotation and analysis of pseudogenes in plants was hampered by incomplete genome assemblies. Many plants — including economically important crops like wheat, rye or barley — have massive genomes with high sequence complexity and repetitivity. Recent technological and methodological advancements finally paved the way for complete and high-quality genome assemblies. With reference sequences available for many crop plants, pseudogenes could now be annotated and analyzed comprehensively and on the whole-genome level. In this work, the Pseudogene Locus Identification Pipeline (PLIPipeline) was created to annotate pseudogenes in 18 plant genomes. The pipeline identifies pseudogenes using homology to functional protein-coding genes. Pseudogenes are then classified according to their structure, distribution and features affecting functionality. While an in-depth analysis of pseudogenes is provided, continuative research is certainly of interest.

Pangenomics With more and more high-quality genome assemblies on the horizon, plant researchers can direct their attention to Pangenomics. A pangenome describes the complete genome of a clade — including all variations between the sequenced individuals. Pangenomics was born with the understanding that "the genome of a single individual is insufficient to represent the gene diversity within a whole species" (Golicz et al., 2016). With more than one reference genome sequence, structural variations like copy number or presence/absence variants can be determined comprehensively and in-depth. A study of 80 fully resequenced *Arabidopsis thaliana* accessions suggests that disruptive mutations occur in at least ~28% of the protein-coding genes (L. Wang et al., 2012). Hence, sporadic pseudogenization may affect a large portion of genes.

Comparative analyses With the availability of high-quality genome assemblies, comparative analyses gain significance. In the past, correcting quality differences was time-consuming and often affected results. For example, in this work, *Triticeae* genomes assembled into contigs/scaffolds consistently stand out compared to *Triticeae* with high-quality assemblies and annotations. Additionally, the comparative analysis of wild barley and barley cv. Morex was hampered by differing assembly qualities. In contrast, high-quality assemblies were available for wild emmer, durum wheat or bread wheat. This significantly improved comparative analyses of pseudogene complements.

Follow-up analyses Not all results from this work can be explained conclusively and some may require additional analyses. For example, in barley, wild emmer and durum wheat, fragmented pseudogenes are preferentially duplicates of the 5' or 3' end of the parent gene's coding sequence (CDS). The central regions were found less frequently or may have degenerated more quickly. The increased frequencies of peripheral regions may be a result of initial functionality up to a premature termination codon or beginning from an alternative start codon. However, this pattern was not found for all plants. Even the closely related bread wheat exhibits a different pattern which cannot be attributed to subgenome D. A follow-up analysis may provide an explanation for this finding.

Reactivating pseudogenes The artificial reactivation of pseudogenes may be of interest for breeding purposes. Especially defense-related pseudogenes may represent targets of interest, since domestication often led to impaired plant defenses (Chen et al., 2015). A recent study demonstrated that chemical defenses in maize are better at targeting specialized insects compared to generalists (Gaillard et al., 2018). This was explained by artificial selection against "specialized herbivores that have coexisted with the crops throughout their domestication". However, maize defense was also shown to be consistently inferior to defense mechanisms in teosinte (wild maize). Reactivating defense-related pseudogenes that were lost due to yield-focused breeding may reduce the need for artificial crop protection and contribute to environment protection.

The PLIPipeline assesses pseudogenes in plants and can manage large and complex data sets. Annotations and detailed analyses were performed for 18 plants. Furthermore, this work also provides an essential basis for continuative studies.

List of Figures

List of Tables

Glossary

A adenine.
Araport The Arabidopsis Information Portal.

BAC Bacterial Artificial Clone.
BBH Bidirectional Best BLAST hit.
BLAST Basic Local Alignment Search Tool.
BLAT BLAST-like alignment tool.
bp base pair.
BTR Brittle Rachis.

cDNA complementary DNA.
CDS coding sequence.

dicot eudicotyledon.
DNA deoxyribonucleic acid.

FAO Food and Agriculture Organization of the United Nations.

Gbp gigabase pairs.
GO Gene Ontology.

HC gene high-confidence gene.
HCov high-coverage.
Hi-C chromosome conformation capture sequencing.

IGV Integrative Genomics Viewer.
IWGSC International Wheat Genome Sequencing Consortium.

kbp kilobase pairs.

LC gene low-confidence gene.
LINE-1 long interspersed element 1.
lncRNA long non-coding RNA.
LTR long terminal repeat.

Mbp megabase pairs.
miRNA microRNA.
monocot monocotyledon.
mRNA messenger RNA.
mya million years ago.

NCBI National Center for Biotechnology Information.

ncRNA non-coding RNA.
NFS north-facing slope.
NGS Next-Generation Sequencing.
NHEJ non-homologous DNA end joining.
nt nucleotide.

ORF open reading frame.

PacBio Pacific Biosciences.
PGSB Plant Genome and Systems Biology.
PLIPipeline Pseudogene Locus Identification Pipeline.
poly-A poly-adenine.
PSF Pseudogene Finder.
PTC premature termination codon.

RNA ribonucleic acid.
RNA-seq RNA sequencing.
rRNA ribosomal RNA.

SDSA synthesis-dependent strand annealing.
SFS south-facing slope.
siRNA small interfering RNA.

t/ha tonnes per hectare.
TAIR The Arabidopsis Information Resource.
TE transposable element.
TIR terminal inverted repeat.
tRNA transfer RNA.

UTR untranslated region.

WGD Whole Genome Duplication.

References

Altschul, Stephen F. et al. (1990). “Basic Local Alignment Search Tool”. In: *Journal of Molecular Biology* 215, pp. 403–10.

Avni, Raz et al. (2017). “Wild emmer genome architecture and diversity elucidate wheat evolution and domestication”. In: *Science* 357 (6346), pp. 93–7.

Balakirev, Evgeniy S. and Ayala, Francisco J. (2003). “Pseudogenes: Are They “Junk” or Functional DNA?” In: *Annual review of genetics* 37, pp. 123–51.

Barbaglia, Allison M. et al. (2012). “Gene capture by Helitron transposons reshuffles the transcriptome of maize”. In: *Genetics* 190.3, pp. 965–75.

Baren, Marijke J. van and Brent, Michael R. (2006). “Iterative gene prediction and pseudogene removal improves genome annotation.” In: *Genome research* 16, pp. 678–85.

Bauer, Eva et al. (2017). “Towards a whole-genome sequence for rye (Secale cereale L.)” In: *Plant Journal* 89.5, pp. 853–69.

Baumgarten, Andrew et al. (2003). “Genome-level evolution of resistance genes in Arabidopsis thaliana”. In: *Genetics* 165.1, pp. 309–19.

Benson, Gary (1999). “Tandem repeats finder: a program to analyze DNA sequences.” In: *Nucleic Acids Research* 27.2, pp. 573–80.

Berardini, Tanya Z. et al. (2015). “The Arabidopsis Information Resource: Making and Mining the ”Gold Standard” Annotated Reference Plant Genome”. In: *Genesis* 53.8, pp. 474–85.

Bolot, Stéphanie et al. (2009). “The ‘inner circle’ of the cereal genomes”. In: *Current Opinion in Plant Biology* 12.2, pp. 119–25.

Brosius, J. and Gould, S. J. (1992). “On “genomenclature”: A comprehensive (and respectful) taxonomy for pseudogenes and other “junk DNA””. In: *Proceedings of the National Academy of Sciences* 89.22, pp. 10706–10.

Bundesministerium für Ernährung und Landwirtschaft (2016). *Understanding Farming - Facts and figures about German farming.* Tech. rep., p. 32.

Byrne, Mary E. (2009). “A role for the ribosome in development”. In: *Trends in Plant Science* 14.9, pp. 512–19.

Byrne, Stephen L. et al. (2015). “A synteny-based draft genome sequence of the forage grass Lolium perenne”. In: *Plant Journal* 84.4, pp. 816–26.

Carvunis, Anne-Ruxandra et al. (2012). “Proto-genes and de novo gene birth”. In: *Nature* 487.7407, pp. 370–4.

Chen, Yolanda H., Gols, Rieta, and Benrey, Betty (2015). “Crop Domestication and Its Impact on Naturally Selected Trophic Interactions”. In: *Annual Review of Entomology* 60.1, pp. 35–58.

Cheng, Chia Yi et al. (2017). “Araport11: a complete reannotation of the Arabidopsis thaliana reference genome”. In: *Plant Journal* 89.4, pp. 789–804.

Clement, Yves et al. (2014). “The bimodal distribution of genic GC content is ancestral to monocot species”. In: *Genome Biology and Evolution* 7.1, pp. 336–48.

Cordaux, Richard and Batzer, Mark A. (2009). “The impact of retrotransposons on human genome evolution.” In: *Nature reviews. Genetics* 10.10, pp. 691–703.

Dai, F. et al. (2012). “Tibet is one of the centers of domestication of cultivated barley”. In: *Proceedings of the National Academy of Sciences* 109.42, pp. 16969–73.

Daron, Josquin et al. (2014). “Organization and evolution of transposable elements along the bread wheat chromosome 3B”. In: *Genome Biology* 15.12, p. 546.

Davidson, Rebecca M. et al. (2012). “Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution”. In: *Plant Journal* 71, pp. 492–502.

Derr, L. K. (1998). “The involvement of cellular recombination and repair genes in RNA-mediated recombination in Saccharomyces cerevisiae.” In: *Genetics* 148, pp. 937–45.

Doležel, Jaroslav et al. (2012). “Chromosomes in the flow to simplify genome analysis”. In: *Functional and Integrative Genomics* 12.3, pp. 397–416.

Dong, Fenggao and Jiang, Jiming (1998). “Non-Rabl Patterns of Centromere and Telomere Distribution in the Interphase Nuclei of Plant Cells”. In: *Chromosome Research* 6.7, pp. 551–8.

Dongen, Stijn van (2000). “Graph Clustering by Flow Simulation”. PhD thesis. University of Utrecht.

Draper, John et al. (2001). “Brachypodium distachyon. A New Model System for Functional Genomics in Grasses”. In: *Plant physiology* 127.4, pp. 1539–55.

Du, C. et al. (2009). “The polychromatic Helitron landscape of the maize genome”. In: *Proceedings of the National Academy of Sciences* 106.47, pp. 19916–21.

Edger, Patrick P. and Pires, J. Chris (2009). “Gene and genome duplications: The impact of dosage-sensitivity on the fate of nuclear genes”. In: *Chromosome Research* 17.5, pp. 699–717.

Elliott, Tyler A. and Gregory, T. Ryan (2015). “What’s in a genome? The C-value enigma and the evolution of eukaryotic genome content”. In: *Philosophical Transactions of the Royal Society B* 370.1678, p. 20140331.

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). “An efficient algorithm for large-scale detection of protein families.” In: *Nucleic Acids Research* 30.7, pp. 1575–1584.

Falcon, S. and Gentleman, R. (2007). “Using GOstats to test gene lists for GO term association.” In: *Bioinformatics* 23.2, pp. 257–8.

FAO et al. (2017). *The State of Food Security and Nutrition in the World 2017.* Tech. rep. Rome: Food and Agriculture Organization of the United Nations.

Food and Agriculture Organization of the United Nations (2017). *FAOSTAT Statistics Database*. Rome. URL: `http://www.fao.org/faostat/`.

Fu, Limin et al. (2012). "CD-HIT: accelerated for clustering the next-generation sequencing data". In: *Bioinformatics* 28.23, pp. 3150–3152.

Gaillard, Mickaël D. P. et al. (2018). "Fine-tuning the 'plant domestication-reduced defense' hypothesis: specialist vs generalist herbivores". In: *New Phytologist* 217.1, pp. 355–66.

Gish, Warren (1996–2003). *WU BLAST*. URL: `http://blast.wustl.edu`.

Glémin, Sylvain et al. (2014). "GC content evolution in coding regions of angiosperm genomes: A unifying hypothesis". In: *Trends in Genetics* 30.7, pp. 263–70.

Golicz, Agnieszka A., Batley, Jacqueline, and Edwards, David (2016). "Towards plant pangenomics". In: *Plant Biotechnology Journal* 14.4, pp. 1099–105.

Gorbunova, Vera and Levy, Avraham A. (1997). "Non-homologous DNA end joining in plant cells is associated with deletions and filler DNA insertions". In: *Nucleic Acids Research* 25.22, pp. 4650–4657.

Guigó, Roderic et al. (2006). "EGASP: the human ENCODE Genome Annotation Assessment Project." In: *Genome biology* 7.Suppl 1, S2.

Guo, Xingyi et al. (2009). "Small RNAs originated from pseudogenes: cis- or trans-acting?" In: *PLoS computational biology* 5.7, e1000449.

Haas, Brian J. et al. (2004). "DAGchainer: A tool for mining segmental genome duplications and synteny". In: *Bioinformatics* 20.18, pp. 3643–6.

Haberer, Georg, Mayer, Klaus F. X., and Spannagl, Manuel (2016). "The big five of the monocot genomes". In: *Current Opinion in Plant Biology* 30, pp. 33–40.

Hairat, Suboot and Khurana, Paramjit (2015). "Evaluation of Aegilops tauschii and Aegilops speltoides for acquired thermotolerance: Implications in wheat breeding programmes". In: *Plant Physiology and Biochemistry* 95.Supplement C, pp. 65–74.

Harrison, P. M. (2001). "Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome". In: *Nucleic Acids Research* 29.3, pp. 818–30.

Harrow, Jennifer et al. (2012). "GENCODE: the reference human genome annotation for The ENCODE Project." In: *Genome research* 22.9, pp. 1760–74.

Haynes, Lee P., McCue, Hannah V., and Burgoyne, Robert D. (2012). "Evolution and functional diversity of the Calcium Binding Proteins (CaBPs)." In: *Frontiers in molecular neuroscience* 5.February, p. 9.

Heslop-Harrison, John Seymour Pat and Schwarzacher, Trude (2011). "Organisation of the plant genome in chromosomes". In: *The Plant Journal* 66.1, pp. 18–33.

Hidalgo, Alyssa and Brandolini, Andrea (2014). "Nutritional properties of einkorn wheat (*Triticum monococcum* L.)" In: *Journal of the Science of Food and Agriculture* 94.4, pp. 601–12.

Horton, Matthew W. et al. (2012). "Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel". In: *Nature Genetics* 44.2, pp. 212–6.

International Barley Sequencing Consortium (2017). "A chromosome conformation capture ordered sequence of the barley genome". In: *Nature* 544, pp. 427–33.

International Wheat Genome Sequencing Consortium (2014). "A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome". In: *Science* 345.6194, p. 1251788.

International Wheat Genome Sequencing Consortium (2018). "Shifting the limits in wheat research and breeding using a fully annotated reference genome". In: *Science* 361.6403.

Jacq, Claude, Miller, J. R., and Brownlee, G. G. (1977). "A Pseudogene Structure in 5S DNA of Xenopus laevis." In: *Cell* 12.1, pp. 109–20.

Ji, Zhe et al. (2015). "Many lncRNAs, 5′UTRs, and pseudogenes are translated and some are likely to express functional proteins". In: *eLife* 4, e08890.

Jia, Jizeng et al. (2013). "Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation." In: *Nature* 496.7443, pp. 91–5.

Johnsson, P., Morris, K. V., and Grandér, D. (2014). "Pseudogenes: A Novel Source of trans-Acting Antisense RNAs". In: *Pseudogenes: Functions and Protocols.* Ed. by Laura Poliseno. Vol. 1167.

Kaessmann, Henrik, Vinckenbosch, Nicolas, and Long, Manyuan (2009). "RNA-based gene duplication: mechanistic and evolutionary insights." In: *Nature reviews. Genetics* 10.1, pp. 19–31.

Kapitonov, V. V. and Jurka, J. (2001). "Rolling-circle transposons in eukaryotes". In: *Proceedings of the National Academy of Sciences* 98.15, pp. 8714–9.

Karro, John E. et al. (2007). "Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation." In: *Nucleic acids research* 35.Database issue, pp. D55–60.

Kawahara, Yoshihiro et al. (2013). "Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data." In: *Rice* 6.1, p. 4.

Kellogg, Elizabeth A. (2001). "Evolutionary History of the Grasses". In: *Plant physiology* 125, pp. 1198–205.

Kent, W. James (2002). "BLAT — The BLAST-Like Alignment Tool". In: *Genome research* 12, pp. 656–64.

Khachane, Amit N. and Harrison, Paul M. (2009). “Assessing the genomic evidence for conserved transcribed pseudogenes under selection.” In: *BMC genomics* 10, p. 435.

Khurana, Ekta et al. (2010). “Segmental duplications in the human genome reveal details of pseudogene formation”. In: *Nucleic Acids Research* 38.20, pp. 6997–7007.

Khush, G. S. (2001). “Green revolution: the way forward”. In: *Nature Reviews Genetics* 2.10, pp. 815–22.

Kim, Gunjune et al. (2014). “Genomic-scale exchange of mRNA between a parasitic plant and its hosts”. In: *Science* 345.6198, pp. 808–11.

Kondrashov, Fyodor A. et al. (2002). “Selection in the evolution of gene duplications.” In: *Genome biology* 3.2, research0008.1–9.

Kozomara, Ana and Griffiths-Jones, Sam (2014). “miRBase : annotating high confidence microRNAs using deep sequencing data”. In: *Nucleic Acids Research* 42.Database, pp. 68–73.

Kudla, Grzegorz et al. (2006). “High guanine and cytosine content increases mRNA levels in mammalian cells”. In: *PLoS Biology* 4.6, pp. 933–42.

Lall, Rohit et al. (2013). “Comparative genome analysis of Solanum lycopersicum and Solanum tuberosum.” In: *Bioinformation* 9.18, pp. 923–8.

Larkin, M. A. et al. (2007). “Clustal W and Clustal X version 2.0.” In: *Bioinformatics* 23.21, pp. 2947–8.

Leister, Dario (2004). “Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes”. In: *Trends in genetics : TIG* 20.3, pp. 116–22.

Levy, Avraham A. and Feldman, Moshe (2002). “The impact of polyploidy on grass genome evolution.” In: *Plant physiology* 130.4, pp. 1587–93.

Li, Weizhong and Godzik, Adam (2006). “Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences”. In: *Bioinformatics* 22.13, pp. 1658–9.

Liang, Chengzhi et al. (2009). “Evidence-based gene predictions in plant genomes”. In: *Genome Research* 19.10, pp. 1912–23.

Ling, Hong-Qing et al. (2013). “Draft genome of the wheat A-genome progenitor Triticum urartu.” In: *Nature* 496.7443, pp. 87–90.

Liu, Haoxuan et al. (2015). “Causes and consequences of crossing-over evidenced via a high-resolution recombinational landscape of the honey bee”. In: *Genome Biology* 16.1.

Lowe, Todd M. and Eddy, Sean R. (1997). “tRNAscan-SE : a program for improved detection of transfer RNA genes in genomic sequence”. In: *Nucleic Acids Research* 25.5, pp. 955–964.

Lynch, M. and Conery, John S. (2000). “The Evolutionary Fate and Consequences of Duplicate Genes”. In: *Science* 290.5494, pp. 1151–5.

Maccaferri, Marco et al. (2018). “Durum wheat genome reveals past domestication signatures and future improvement targets”. In: *Genome Biology (under revision)*.

Mannion, A. M. (1999). “Progress in Physical Geography Domestication and the origins of agriculture: an appraisal”. In: *Progress in Physical Geography* 23.1, pp. 37–56.

Marcussen, Thomas et al. (2014). “Ancient hybridizations among the ancestral genomes of bread wheat”. In: *Science* 345.6194, p. 1250092.

Martis, Mihaela M. et al. (2013). “Reticulate Evolution of the Rye Genome.” In: *The Plant cell* 5, pp. 1–15.

Mazumder, Barsanjit, Seshadri, Vasudevan, and Fox, Paul L. (2003). “Translational control by the 3′-UTR: The ends specify the means”. In: *Trends in Biochemical Sciences* 28.2, pp. 91–98.

McArthur, John W. and McCord, Gordon C. (2017). “Fertilizing growth: Agricultural inputs and their effects in economic development”. In: *Journal of Development Economics* 127, pp. 133–52.

Mehra, Mrigaya, Gangwar, Indu, and Shankar, Ravi (2015). “A deluge of complex repeats: The Solanum genome”. In: *PLoS ONE* 10.8, pp. 1–38.

Mehta, Anuja and Haber, James E. (2014). “Sources of DNA Double-Strand Breaks and Models of Rec”. In: *Cold Spring Harbor Perspectives in Biology* 6, pp. 1–19.

Mignone, Flavio et al. (2002). “Untranslated regions of mRNAs.” In: *Genome biology* 3.3, reviews0004.1–10.

Millet, E. et al. (2014). “Introgression of leaf rust and stripe rust resistance from Sharon goatgrass (Aegilops sharonensis Eig) into bread wheat (Triticum aestivum L.)” In: *NRC Researcj Press* 57.August, pp. 309–16.

Mitchell, Carolyn et al. (2016). “Plant Defense against Herbivorous Pests: Exploiting Resistance and Tolerance Traits for Sustainable Crop Protection”. In: *Frontiers in Plant Science* 7.1132, pp. 1–8.

Murat, Florent, Van De Peer, Yves, and Salse, Jérôme (2012). “Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes”. In: *Genome Biology and Evolution* 4.9, pp. 917–28.

Nevo, E. (2012). ““Evolution Canyon”, a potential microscale monitor of global warming across life”. In: *Proceedings of the National Academy of Sciences* 109.8, pp. 2960–5.

Nishioka, Y., Leder, A., and Leder, P. (1980). “Unusual alpha-globin-like gene that has cleanly lost both globin intervening sequences.” In: *Proceedings of the National Academy of Sciences* 77.5, pp. 2806–9.

Nussbaumer, Thomas et al. (2015). “Joint Transcriptomic and Metabolomic Analyses Reveal Changes in the Primary Metabolism and Imbalances in the Subgenome Orchestration in the Bread Wheat Molecular Response to *Fusarium graminearum*”. In: *Genes|Genomes|Genetics* 5.12, pp. 2579–92.

Ohshima, Kazuhiko et al. (2003). “Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates”. In: *Genome Biology* 4.11, pp. 1–14.

Ortutay, Csaba and Vihinen, Mauno (2008). “PseudoGeneQuest – Service for identification of different pseudogene types in the human genome”. In: *BMC Bioinformatics* 9.1, p. 299.

Pavlicek, Adam et al. (2006). “Retroposition of processed pseudogenes: The impact of RNA stability and translational control”. In: *Trends in Genetics* 22.2, pp. 69–73.

Pfeifer, M. et al. (2014). “Genome interplay in the grain transcriptome of hexaploid bread wheat”. In: *Science* 345.6194, p. 1250091.

Pink, Ryan Charles et al. (2011). “Pseudogenes: Pseudo-functional or key regulators in health and disease?” In: *RNA* 17, pp. 792–798.

Podlaha, Ondrej and Zhang, Jianzhi (2010). “Pseudogenes and Their Evolution”. In: *Encyclopedia of Life Sciences (ELS)*, pp. 1–8.

Poliseno, Laura, Marranci, Andrea, and Pandolfi, Pier Paolo (2015). “Pseudogenes in human cancer”. In: *Frontiers in Medicine* 2, p. 68.

Poliseno, Laura, Salmena, Leonardo, et al. (2010). “A coding-independent function of gene and pseudogene mRNAs regulates tumour biology.” In: *Nature* 465.7301, pp. 1033–1038.

Poovaiah, B. W., Reddy, A. S. N., and Feldman, L. (1993). “Calcium and Signal Transduction in Plants”. In: *Critical Reviews in Plant Sciences* 12.3, pp. 185–211.

Pourkheirandish, Mohammad, Hensel, Goetz, et al. (2015). “Evolution of the Grain Dispersal System in Barley”. In: *Cell* 162.3, pp. 527–539.

Pourkheirandish, Mohammad, Kanamori, Hiroyuki, et al. (2018). “Elucidation of the origin of ‘*agriocrithon*’ based on domestication genes questions the hypothesis that Tibet is one of the centers of barley domestication”. In: *The Plant Journal* 94 (3), pp. 525–34.

Prade, Verena M. et al. (2018). “The pseudogenes of barley”. In: *The Plant Journal* 93, pp. 502–14.

Preece, Catherine et al. (2017). “How did the domestication of Fertile Crescent grain crops increase their yields?” In: *Functional Ecology* 31.2, pp. 387–97.

Prieto-Godino, Lucia L. et al. (2016). “Olfactory receptor pseudo-pseudogenes”. In: *Nature* 539.7627, pp. 93–97.

Prince, Victoria E. and Pickett, F. Bryan (2002). “Splitting pairs: the diverging fates of duplicated genes.” In: *Nature reviews. Genetics* 3.11, pp. 827–37.

Quax, Tessa E. F. et al. (2015). “Codon Bias as a Means to Fine-Tune Gene Expression”. In: *Molecular Cell* 59.2, pp. 149–61.

Ray, Deepak K. et al. (2013). “Yield Trends Are Insufficient to Double Global Crop Production by 2050”. In: *PLoS ONE* 8.6.

Roberts, Thomas C. and Morris, Kevin V. (2013). “Not so pseudo anymore: pseudogenes as therapeutic targets.” In: *Pharmacogenomics* 14.16, pp. 2023–34.

Robicheau, Brent M. et al. (2017). “Ribosomal RNA genes contribute to the formation of pseudogenes and junk DNA in the human genome”. In: *Genome Biology and Evolution* 9.2, pp. 380–97.

Rocha, Agostinho and Vothknecht, Ute (2013). “Identification of CP12 as a Novel Calcium-Binding Protein in Chloroplasts”. In: *Plants* 2.3, pp. 530–540.

Rouchka, Eric Christian and Cha, Elizabeth (2009). “Current Trends in Pseudogene Detection and Characterization”. In: *Current Bioinformatics* 4, pp. 112–9.

Sabot, F. and Schulman, A. H. (2006). “Parasitism and the retrotransposon life cycle in plants: A hitchhiker’s guide to the genome”. In: *Heredity* 97.6, pp. 381–8.

Sayers, Eric W. et al. (2009). “Database resources of the National Center for Biotechnology Information”. In: *Nucleic Acids Research* 37.Database issue, pp. 5–15.

Schiml, Simon, Fauser, Friedrich, and Puchta, Holger (2016). “Repair of adjacent single-strand breaks is often accompanied by the formation of tandem sequence duplications in plant genomes”. In: *PNAS* 113.26, pp. 7266–71.

Schlötterer, Christian (2015). “Genes from scratch – the evolutionary fate of de novo genes”. In: *Trends in Genetics* 31.4, pp. 215–219.

Schnable, Patrick S. et al. (2009). “The B73 maize genome: complexity, diversity, and dynamics.” In: *Science* 326.5956, pp. 1112–5.

Schueren, Fabian and Thoms, Sven (2016). “Functional Translational Readthrough: A Systems Biology Perspective”. In: *PLoS Genetics* 12.8, pp. 1–12.

Sen, Kamalika and Ghosh, Tapash Chandra (2013). “Pseudogenes and their composers: delving in the ‘debris’ of human genome”. In: *Briefings in Functional Genomics* 12.6, pp. 536–547.

Shidhi, P. R. et al. (2014). “Making novel proteins from pseudogenes”. In: *Bioinformatics* 31.1, pp. 33–39.

Siegel, Jake J. and Amon, Angelika (2012). “New Insights into the Troubles of Aneuploidy”. In: *Annual Review of Cell and Developmental Biology* 28.1, pp. 189–214.

Slotkin, R. Keith and Martienssen, Robert (2007). "Transposable elements and the epigenetic regulation of the genome". In: *Nature Reviews Genetics* 8.4, pp. 272–85.

Solovyev, Victor et al. (2006). "Automatic annotation of eukaryotic genes, pseudogenes and promoters." In: *Genome biology* 7.Suppl 1, S10.1–12.

Spannagl, Manuel et al. (2016). "PGSB PlantsDB : updates to the database framework for comparative plant genome research". In: *Nucleic Acids Research* 44, pp. 1141–7.

Supek, Fran et al. (2011). "REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms." In: *PloS one* 6.7, e21800.

Tanaka, Junichi, Hayashi, Takeshi, and Iwata, Hiroyoshi (2016). "A practical, rapid generation-advancement system for rice breeding using simplified biotron breeding system". In: *Breeding Science* 66.4, pp. 542–551.

Tang, H. et al. (2010). "Angiosperm genome comparisons reveal early polyploidy in the monocot lineage". In: *Proceedings of the National Academy of Sciences* 107.1, pp. 472–77.

Tatarinova, Tatiana V. et al. (2010). "GC3 biology in corn, rice, sorghum and other grasses". In: *BMC Genomics* 11.1.

The Arabidopsis Genome Initiative (2000). "Analysis of the genome sequence of the flowering plant Arabidopsis thaliana." In: *Nature* 408.6814, pp. 796–815.

The ENCODE Project Consortium (2012). "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414, pp. 57–74.

The International Brachypodium Initiative (2010). "Genome sequencing and analysis of the model grass Brachypodium distachyon." In: *Nature* 463.7282, pp. 763–8.

The International Human Genome Sequencing Consortium (2004). "Finishing the euchromatic sequence of the human genome". In: *Nature* 431.7011, pp. 931–45.

The Potato Genome Sequencing Consortium (2011). "Genome sequence and analysis of the tuber crop potato". In: *Nature* 475.7355, pp. 189–95.

The Tomato Genome Consortium (2012). "The tomato genome sequence provides insights into fleshy fruit evolution". In: *Nature* 485, pp. 635–41.

Thibaud-Nissen, Françoise, Ouyang, Shu, and Buell, C. Robin (2009). "Identification and characterization of pseudogenes in the rice gene complement." In: *BMC genomics* 10, p. 317.

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

Torrents, David et al. (2003). "A genome-wide survey of human pseudogenes." In: *Genome research* 13.12, pp. 2559–67.

Tran, Lan T., Taylor, John S., and Constabel, C. Peter (2012). “The polyphenol oxidase gene family in land plants: Lineage-specific duplication and expansion.” In: *BMC Genomics* 13.1, p. 395.

Tutar, Yusuf (2012). “Pseudogenes.” In: *Comparative and functional genomics* 2012, p. 424526.

Van de Peer, Yves, Mizrachi, Eshchar, and Marchal, Kathleen (2017). “The evolutionary significance of polyploidy”. In: *Nature Reviews Genetics* 18.7, pp. 411–424.

Vanin, E. F. et al. (1980). “A mouse alpha-globin-related pseudogene lacking intervening sequences”. In: *Nature* 286.5770, pp. 222–226.

Wang, Hao, Devos, Katrien M., and Bennetzen, Jeffrey L. (2014). “Recurrent Loss of Specific Introns during Angiosperm Evolution”. In: *PLoS Genetics* 10.12.

Wang, Long et al. (2012). “Genome-Wide Survey of Pseudogenes in 80 Fully Resequenced Arabidopsis thaliana Accessions.” In: *PloS one* 7.12, e51769.

Wang, Wen et al. (2006). “High Rate of Chimeric Gene Origination by Retroposition in Plant Genomes”. In: *Plant Cell* 18.8, pp. 1791–1802.

Wang, W. et al. (2014). “The Spirodela polyrhiza genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle.” In: *Nature communications* 5, p. 3311.

Wang, Xiyin et al. (2005). “Duplication and DNA segmental loss in the rice genome: Implications for diploidization”. In: *New Phytologist* 165.3, pp. 937–46.

Weiss-Schneeweiss, H. et al. (2013). “Evolutionary Consequences, Constraints and Potential of Polyploidy in Plants.” In: *Cytogenetic and genome research* 140.2-4, pp. 137–150.

Weller, C. and Wu, M. (2015). “A generation-time effect on the rate of molecular evolution in bacteria”. In: *Evolution* 3.69, pp. 2541–58.

Wicker, Thomas, Buchmann, Jan P., and Keller, Beat (2010). “Patching gaps in plant genomes results in gene movement and erosion of colinearity.” In: *Genome research* 20.9, pp. 1229–1237.

Wicker, Thomas, Gundlach, Heidrun, et al. (2018). “Impact of transposable elements on genome structure and evolution in bread wheat”. In: *Genome Biology (under revision).*

Wicker, Thomas, Sabot, François, et al. (2007). “A unified classification system for eukaryotic transposable elements”. In: *Nature Reviews Genetics* 8.12, pp. 973–82.

Wicker, Thomas, Yu, Yeisoo, et al. (2018). “DNA transposon activity is associated with increased mutation rates in genes of rice and other grasses”. In: *Nature Communications* 7, p. 12790.

Williams, Diana L. et al. (2009). “Implications of high level pseudogene transcription in Mycobacterium leprae.” In: *BMC genomics* 10, p. 397.

Xiao, Jin et al. (2016). “Pseudogenes and Their Genome-Wide Prediction in Plants”. In: *International Journal of Molecular Sciences* 17.12, p. 1991.

Xiao-Jie, Lu et al. (2015). “Pseudogene in cancer: real functions and promising signature.” In: *Journal of medical genetics* 52.1, pp. 17–24.

Xu, Jinrui and Zhang, Jianzhi (2015). “Are Human Translated Pseudogenes Functional ?” In: *Molecular Biology and Evolution* 33.3, pp. 755–60.

Yang, Ziheng (2007). “PAML 4: Phylogenetic Analysis by Maximum Likelihood.” In: *Molecular Biology and Evolution* 24.8, pp. 1586–1591.

Yu, Jun et al. (2005). “The Genomes of Oryza sativa: A History of Duplications.” In: *PLoS biology* 3.2, e38.

Zhang, Yuan and Sun, Yanni (2012). “PseudoDomain: identification of processed pseudogenes based on protein domain classification”. In: *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pp. 178–185.

Zhang, Zhaolei, Carriero, Nicholas, et al. (2006). “PseudoPipe: an automated pseudogene identification pipeline.” In: *Bioinformatics* 22.12, pp. 1437–9.

Zhang, Zhaolei and Gerstein, Mark (2003). “The human genome has 49 cytochrome c pseudogenes, including a relic of a primordial gene that still functions in mouse”. In: *Gene* 312.1-2, pp. 61–72.

Zhang, Zhaolei, Harrison, Paul M., et al. (2003). “Millions of years of evolution preserved: A comprehensive catalog of the processed pseudogenes in the human genome”. In: *Genome Research* 13.12, pp. 2541–58.

Zhang, Zhaolei, Harrison, Paul, et al. (2002). “Identification and Analysis of Over 2000 Ribosomal Protein Pseudogenes in the Human Genome Identification and Analysis of Over 2000 Ribosomal Protein Pseudogenes in the Human Genome”. In: *Genome Research* 12 (10), pp. 1466–82.

Zhang, Zhengdong D. et al. (2010). “Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates.” In: *Genome Biology* 11.3, R26.

Zhao, Guangyao et al. (2017). “The Aegilops tauschii genome reveals multiple impacts of transposons”. In: *Nature Plants* 3.12, pp. 946–955.

Zhao, Zhixin et al. (2014). “Genome-Wide Analysis of Tandem Repeats in Plants and Green Algae”. In: *Genes|Genomes|Genetics* 4.1, pp. 67–78.

Zheng, Deyou and Gerstein, Mark B. (2007). “The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they?” In: *Trends in genetics : TIG* 23.5, pp. 219–24.

Zhou, Xiang et al. (2015). “Ribosomal proteins: Functions beyond the ribosome”. In: *Journal of Molecular Cell Biology* 7.2, pp. 92–104.

Zimin, Aleksey V. et al. (2017). “Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm”. In: *Genome Research* 27.5, pp. 787–792.

Zou, Cheng et al. (2009). “Evolutionary and Expression Signatures of Pseudogenes in Arabidopsis and Rice.” In: *Plant Physiology* 151.1, pp. 3–15.