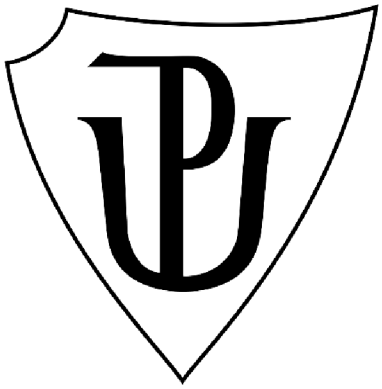
UNIVERZITA PALACKÉHO V OLOMOUCI

Přírodovědecká fakulta

Katedra biochemie



Identification and expression analysis of species-specific genes in plant hybrids

Identifikace a analýza genové exprese druhově specifických genů v rostlinných kříženců

BAKALÁŘSKÁ PRÁCE

Autor:	**Marie Chudecká**
Studijní program:	B1406 Biochemie
Studijní obor:	Bioinformatika
Forma studia:	Prezenční
Vedoucí práce:	**doc. RNDr. David Kopecký, Ph.D.**
Rok:	2023

Prohlašuji, že jsem bakalářskou práci vypracoval/a samostatně s vyznačením všech použitých pramenů a spoluautorství. Souhlasím se zveřejněním bakalářské práce podle zákona č. 111/1998 Sb., o vysokých školách, ve znění pozdějších předpisů. Byl/a jsem seznámen/a s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, ve znění pozdějších předpisů.

V Olomouci dne ....

Podpis studenta

Poděkování

Chtěla bych poděkovat vedoucímu mé práce, doc. RNDr. Davidovi Kopeckému, Ph.D., za věnovaný mi čas, cenné úvahy při návrhu a psaní této práce, a značnou trpělivost, Mgr. Markovi Glombikovi, Ph.D. za obsáhlou konzultaci a vedení v praktické části práce a celé Skupině vzdálené hybridizace za přátelskou atmosféru práce. Dále děkuju také organizaci MetaCentrum za možnost využití výpočetních a úložních serverů.

Bibliografická identifikace

Jméno a příjmení autora	Marie Chudecká
Název práce	Identifikace a analýza exprese druhově specifických genů v rostlinných kříženců
Typ práce	Bakalářská
Pracoviště	Katedra biochemie
Vedoucí práce	doc. RNDr. David Kopecký, Ph.D.
Rok obhajoby práce	2023

Abstrakt

Druhově specifické geny jsou přítomné v genomech drtivé většiny organism, včetně rostlinných a živočišných druhů. Alespoň část z nich má pro své nositele zřejmě význam v adaptaci na okolní prostředí. Regulace jejich exprese je mnohoúrovňový mechanismus, který je značně modifikován po mezidruhové hybridizaci, tj. sloučení genomů dvou různých druhů. Tato práce zabývá se dosud nezkoumaným osudem druhově specifických genů v mezidruhových kříženců. S využitím dostupných sekvenčních dat a *de novo* sestavených transkriptomů byly identifikovány druhově specifické geny pro dva druhy trav, *Festuca pratensis* a *Lolium multiflorum* a studována jejich exprese jak v obou rodičovských druzích, tak v jejich kříženců (*Festulolium*). Byla rovněž provedená jednoduchá funkční anotace s důrazem na genovou ontologii.

Klíčová slova	Druhově specifické geny, mezidruhová hybridizace, genová exprese, genová ontologie
Počet stran	59
Počet příloh	2
Jazyk	Anglický

Bibliographical identification

Author's first name and surname	Marie Chudecká
Title	Identification and expression analysis of species-specific genes in plant hybrids
Type of thesis	Bachelor
Department	Department of biochemistry
Supervisor	doc. RNDr. David Kopecký, Ph.D.
The year of presentation	2023

Abstract

Lineage-specific genes are present in genomes across all kingdoms of life tree and are recognised for their adaptive potential. However, regulation of their expression is a multi-layered mechanism, which undergoes significant changes in a hybrid genome. This thesis examines the previously unexplored interplay of the functionality of these genes with interspecific hybridisation by studying their presence, characteristics and expression patterns in two grass species, *Festuca pratensis* and *Lolium multiflorum*, and their hybrids called *Festulolium*. Species-specific genes are identified for both species using available sequence data and *de novo* transcriptome assembly. Functional annotation, with emphasis on gene ontology, is performed. Finally, existing transcription data are employed in expression analysis.

Keywords	Lineage-specific genes, orphan genes, interspecific hybridisation, gene expression, gene ontology
Number of pages	59
Number of appendices	2
Language	English

TABLE OF CONTENTS

Cíle práce

Teoretická část

- Vypracování literární rešerše na téma identifikace druhově specifických genů a jejich přítomnosti a exprese v rostlinných kříženců

Praktická část

- Sestavení anotovaných transkriptomů druhů *Festuca pratensis* Huds. a *Lolium multiflorum* Lam.
- Identifikace druhově specifických genů, gene ontology (GO) charakterizace
- Porovnání exprese identifikovaných druhově specifických genů v rodičovských rostlinách a kříženců F_1 a F_2 generace

1 INTRODUCTION

The increasing accessibility of whole-genome sequence data has both elevated their relevance in research across many research fields. Although many genomes have already been sequenced, our understanding of their contents, function and evolution is still limited. One of the perspectives on genome structure provided by comparative genomics is the conservation of genes between organisms or lack thereof. The field of interest of the latter approach are lineage-specific genes, which appear only in a specific lineage of a species and species-specific genes are those appearing only in a single species. Species-specific genes are being identified in all domains of life (Khalturin et al., 2009) and their characterisation shows that they can be important for species-specific functions and stress responses. Furthermore, their emergence forces us to reconsider our understanding of gene emergence and extinction.

Emergence of the species-specific genes can be accelerated by the polyploidization. Polyploids, as products of whole genome duplication are common in plant kingdom. Studies of polyploidy and polyploidization have rich traditions and significance in botany and agriculture and thus are far from new in research. However, the advent of genomic era, which revealed prevalence of ancient (Jiao et al., 2011; Ruprecht et al., 2017) and recent polyploidy (Wood et al., 2009; Salman-Minkov et al., 2016), brought possibilities of detailed studies of gene expression and allowed us to observe processes behind stabilisation of polyploid genome. Allopolyploids, as organisms facing challenge of multiplied genetic material from parents of different species, are particularly interesting. The merging of previously independent genomes and their regulatory networks triggers an intensive process of changes and adjustments, especially in gene expression. Many studies have shed light on fate of elements duplicated between the subgenomes, however dynamics of the unique input of parents, specifically their lineage(species)-specific genes remain less clear.

The goal of this thesis is to examine presence of species-specific genes in *Festuca pratensis* (meadow fescue) and *Lolium multiflorum* (Italian ryegrass), two closely related grass species of *Poaceae* family, which are capable of forming viable interspecific hybrids. The species-specific genes will be identified and characterised with emphasis on function prediction. Then the expression of identified genes will be analysed in both parents and their hybrids, *Festulolium*.

2 LITERATURE REVIEW

2.1 Genome and its composition

Genome of an organism represents the complete genetic information, usually defined at the cellular level as all the genetic material found in a single cell. In case of eukaryotes, it includes DNA from nucleus, mitochondria and plastids (if present); however the most common use of the term refers only to nuclear genetic material.

For practical purposes genome refers to a sequence divided into chromosomes. A haploid set of chromosomes is used for genomic characterization of the organisms (species) regardless of their ploidy, although some exceptions apply: for instance human male genome contains both X and Y chromosome from pair 23. The haploid amount of DNA, defined as the amount in a standard (reduced) gamete, is called C-value (1C) and is used interchangeably with the genome size. In allopolyploids genome can be divided into subgenomes by chromosome ancestry, therefore it is important to note which subgenome is referring to which progenitor. For example, bread wheat has three subgenomes: A, B, and D originating from *Triticum urartu*, *Aegilops speltoides* and *Aegilops tauschii*, respectively, given the genome composition of bread wheat BBAADD (recently reviewed by Levy & Feldman (2022)). Thus, the relation between genome size and C-value is more complicated in polyploids and is discussed in detail by Greilhuber et al. (2005).

Chromosome sizes vary across different species, therefore sequence length (in base pairs, bp or their multiples, Mbp or Gbp) or total weight of DNA (in picograms, pg) is used for measuring genome size. In eukaryotes genome size varies greatly, from microsporidian parasite *Encephalitozoon intestinalis* with 1C = 0.0023 pg (Vivares, 1999) to 1C = 152.13 pg in *Paris japonica*, a monocot endemic to Japan (Pellicer et al., 2010), the latter being over 66 000 times larger than the former. Most eukaryotic genomes place themselves in the lower end of the size range (Gregory, 2005) – in plants the mean value is 1C = 5.51 pg (Leitch et al., 2019).

The vast genome size range raises the question of its reason. While polyploidy and increasing evolutional complexity of organisms provide some intuitive explanation, comparison of genome sizes between species revealed that size and complexity of organism actually do not correlate i.e. less complex organism may exhibit larger genome than expected, for instance human genome is around 3.3 pg (Piovesan et al., 2019) compared to 64.62 pg of amphipod *Ampelisca macrocephala* (Traut et al., 2007; Gregory, 2023). This surprising lack of correlation

and general variation in genome size has been historically called the C-value paradox; the analysis of genome contents provided explanation for it and simultaneously highlighted other unanswered questions about genome structure and its evolution, the problem dubbed C-value enigma (Gregory, 2001). From the functional point of view, genome consists of coding and regulatory regions, pseudogenes, gene fragments and different types of repetitive sequences. Content of all genomic components positively correlates with genome size, however to different degree, as contribution of genes decreases and that of repetitive DNA (especially transposable elements) increases with genome size (Elliott & Gregory, 2015).

2.1.1 Genes

In terminology of molecular biology, a gene is a part of genome encoding one or possibly more functional products of one type i.e. either proteins or non-coding RNAs (tRNA, rRNA, miRNA etc.). Several types of sequences participate in synthesis of the product therefore various definitions regarding precise localisation in DNA may be used. Directly contributing exons, together with introns, constitute the open reading frame (ORF), often used synonymously to gene in gene prediction studies. Primary RNA transcript contains also 5' and 3' untranslated regions (UTR); lastly, regulatory sequences (mainly promoters and enhancers) are required for transcription, however they are not present in primary transcript and their location varies greatly: they may appear upstream or downstream, in proximal or distal regions, in other genes or introns. Both types tend to be omitted from gene structure, but are sometimes annotated as gene-associated. Due to alternative splicing, several products may arise from the same DNA sequence, which are matched to genes based on shared exon sequences. Conversely, products may arise from joining transcripts in *trans*-splicing. In result, gene sequence does not need to be continuous, but rather it is a collection of discrete sequences (Gerstein et al., 2007).

2.1.2 Pseudogenes

Pseudogenes are sequences similar to genes, but lack their functionality. They arise in the genome by two main mechanisms: processed pseudogenes emerge as a result of retrotransposition (Fig. 1a), while mutation and resulting degeneration of genes' duplicates gives rise to nonprocessed pseudogenes (Fig. 1b). Though the term might imply that these are regions of no biological meaning, simple by-products of genomic evolution, the definition is broad enough to cover sequences of varying role, with some being translated into proteins or peptides, transcribed into non-coding RNAs, regulating expression of the parent gene and lastly,

the ones that perhaps truly live up to their name and have no discovered function. The many possible functions of pseudogenes (as reviewed by Pink et al. (2011) or Kovalenko & Patrushev (2018)) and cases of genes misidentified as pseudogenes (e.g. Betrán et al. (2002), Prieto-Godino et al. (2016)) highlight the need for careful annotation. Since the identification of pseudogenes is usually performed *in silico* using whole genome sequence data and is predictive in nature, the key role of precise experimental evidence is undeniable (Cheetham et al., 2020).

2.1.3 Repetitive DNA

Genes and pseudogenes are generally regarded as low-copy or single-copy elements; the rest of the genome can be summarised as repetitive DNA. This very broad category can be divided by location into dispersed repeats – scattered around the genome and tandem – forming blocks of neighbouring copies. Dispersed repeats encompass transposons, processed pseudogenes and genes coding tRNA (among others), while prominent examples of tandem repeats include satellite DNA and genes coding rRNA. The number of copies can range from hundreds to hundreds of thousands. In case of multicopy gene families such as genes coding rRNA, tRNA or histone proteins, multiple copies allow the cell to meet high transcriptional demand.

Satellite DNA forms tandem repeats of monomers of variable length – from a few nucleotides to few thousand bp; they are divided accordingly into microsatellites, minisatellites and satellites. Satellite DNA can be primarily found at the centromere, pericentrometic and subtelomeric regions, and in telomeres. Therefore it plays part in heterochromatin structure, chromosome behaviour during cell divisions, additionally its involvement in gene regulation has been proposed (recently reviewed by Garrido-Ramos (2017), Thakur et al. (2021)).

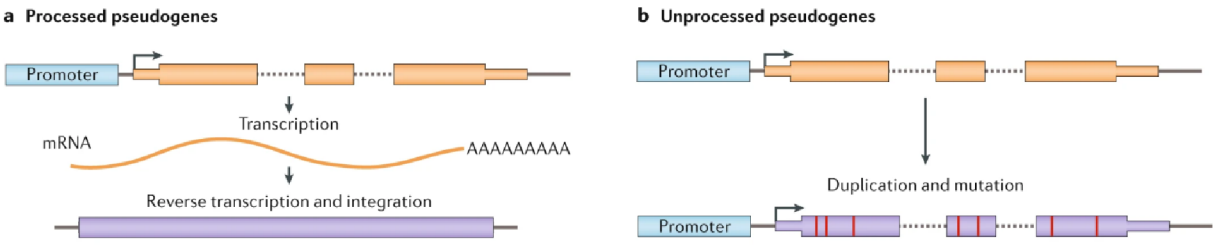


Fig. 1: Two main classes of pseudogenes, defined by their origin: a – processed pseudogenes originate from transcripts of the parent gene than was integrated into genome via retrotransposition; b – unprocessed pseudogenes develop from duplicates of the parent gene via mutation. Adopted from Cheetham et al. (2020)

Another major class are transposable elements (TEs) or simply transposons. Their feature is to replicate and integrate new copies of themselves into the genome, not dissimilar to virus behaviour. Transposons are divided into two main classes: retrotransposons (class I) and DNA transposons (class II). The former use RNA intermediate and reverse transcription to copy their sequence and subsequently integrate into target location; the latter employ DNA intermediate, which is usually excised from its original site or, in case of *Helitron* transposons (Grabundzija et al., 2016), cut from one strand to form a circular intermediate. Autonomous transposons are capable of transporting themselves because they encode all necessary enzymes (usually reverse transcriptase or transposase), while the non-autonomous rely on other elements to propagate.

While transposons *per se* do not perform a function for the genome or organism (and rather resemble parasites in their behaviour), they have significant impact on it. Their mobility is a substantial source of mutation in origin and target site, possibly disrupting gene sequence or regulatory element, causing production of abnormal transcripts or changes in expression level. Additionally, both active and inactive TEs contribute to structural variation in genome (Bourque et al., 2018): homology between their copies in different parts of the genome promotes recombination and rearrangements. They may carry other elements (gene fragments, regulatory sequences or satellite DNA) around genome and lastly, their deletion often encompasses flanking regions. Moreover, they may represent sources of regulatory sequences for nearby genes, be transcribed into non-coding RNA (Piriyapongsa et al., 2007; McCue & Slotkin, 2012) or even adapted into functional genes (e.g. *DAYSLEEPER* (Bundock & Hooykaas, 2005), including *FHY3* and *FAR1* (Lin et al., 2007) and *MUSTANG* family (Joly-Lopez et al., 2012) in *A. thaliana* (see also a review by Jangam et al. (2017) for an extensive list of examples). TEs may represent over half of the genome (e.g. 65-85% in maize (Schnable et al., 2009; Jiao et al., 2017), ~80% in barley (The International Barley Genome Sequencing Consortium, 2012)). Their content has tremendous significance for genome stability and requires careful regulation. Despite the aforementioned potentially positive effects, activity of transposable elements usually has deleterious consequences; therefore they are often epigenetically silenced (but may also be self-regulated e.g. *Ty1* in yeast (Saha et al., 2015)). This in turn may affect neighbouring genes by limiting their expression as well (Hollister & Gaut, 2009).

2.2 Gene expression

A gene is expressed when its product is synthesised; the products of genes include both proteins and non-coding RNA, as discussed earlier. The expression of eukaryotic protein-coding genes consists of several steps, namely transcription, mRNA processing and transport, translation, post-translational modification and transport of protein. Despite identical genetic equipment of all cells in an organism, genes are expressed differently depending on cell or tissue type, developmental stage and external conditions. Thus, gene expression must be precisely regulated throughout all the steps. Regulation (usually at the transcriptional level) differentiates *cis*- and *trans*-acting factors, where the former are regulatory sequences present on the same DNA molecule as the transcribed gene (promoters, enhancers etc.) and the latter encompass molecules interacting externally (e.g. transcription factors).

2.2.1 Transcriptional regulation

Transcription factors (TFs) are proteins containing DNA-binding domains, which recognise regulatory sequences of the gene. TFs are necessary for pre-initiation complex assembly as RNA polymerase is unable to initiate transcription on its own. General transcription factors (GTFs) are essential to connect RNA polymerase and core promoter (Thomas & Chiang, 2008; Luse, 2013), while other TFs are required only for certain genes or altogether optional, modifying the basal transcription level and thus providing expression plasticity. At the sequence level TF binding is determined by regulatory sequences present around or inside a gene; those elements include promoters, enhancers, silencers, insulators and tethering elements.

Promoters are defined as binding sites for TFs obligatory for the transcription of a given gene. The core promoter, typically associated with GTFs, is located near the transcription start site. Enhancers and silencers have complimentary roles: the former increases and the latter decreases transcription activity of a given gene. Their location varies – quite often they are separated by up to hundreds of thousands of nucleotides from the ORF (e.g. (Amano et al., 2009; Shi et al., 2013)). On the other hand, they may appear inside introns or even coding sequences. These elements influence transcription by binding proteins (activators and repressors), which either contribute to or interfere with pre-initiation complex assembly, respectively. The issue of sequential distance is believed to be solved by DNA looping, which positions enhancers and promoters in proximity, allowing for interaction of bound proteins (Levine et al., 2014). The exact mechanism of this interaction is still unknown (Panigrahi & O'Malley, 2021), but both insulators and tethering elements are thought to play a role in it. The

latter help to establish and stabilise enhancer-promoter interaction (Batut et al., 2022). The former hinder those interactions, mainly between topologically associated domains and block spread of chromatin state changes (Brasset & Vaury, 2005; Batut et al., 2022). The distinction between promoters, enhancers and silencers is useful, but rather context-dependent as these elements may perform different functions for different genes and in different conditions (Andersson et al., 2015). Generally, they are rarely specific for a certain gene and conversely, activation of a gene is not specific for a single factor, but rather a result of their specific combination.

In broader perspective, the structure of chromatin (DNA molecules bound to and wrapped around histone proteins) and its changes are crucial for transcription activity. The more densely-packed heterochromatin is overall transcriptionally inactive and requires action of chromatin-remodelling proteins to become euchromatin. This process is mainly linked with histone modifications, mostly methylation and acetylation in specific sites, which constitute so-called histone marking (Bannister & Kouzarides, 2011). Additionally, direct methylation of DNA sequence may be indispensable for recognition by TFs or chromatin-remodelling proteins; conversely it may promote gene silencing, which is widely used in limiting transposon activity (He et al., 2011; Zhang et al., 2018).

Aside from the availability of binding sites in regulatory sequences, the availability of active TFs is also pivotal in transcriptional regulation. TFs may be present in active or inactive form and require a post-translation modification for activation and/or transport to the nucleus; other TFs are produced "on demand". Interconnection of regulation systems via same TFs or inducing signals results in grouping expression into transcriptional programs (Pope & Medzhitov, 2018). An interesting aspect of TF influence is its ability to travel between cells in plants through plasmodesmata (first discovered in maize by Jackson et al. (1994)), although the factors regulating their mobility and the functional implications are not well understood yet (Kurata et al., 2005; Gundu et al., 2020).

The complete pre-initiation complex contains dozens of proteins: GTFs, other TFs, chromatin remodelling proteins and mediating complexes (e.g. Mediator or Integrator) (Luse, 2013; Sainsbury et al., 2015). Once assembled, it must release RNA polymerase, which starts the transcription properly. The elongation is dependent on availability of nucleotides, activity of elongation factors and regulation of RNA polymerase activity (Chen et al., 2018). Correct termination, inextricably connected to polyadenylation of 3'-end, is needed for formation and

later export of transcript (de Felippes et al., 2023). Alternative polyadenylation may result in different isoforms of varying stability and play a role in protein targeting (Berkovits & Mayr, 2015; Tian & Manley, 2017). Thus, poly-A tail is crucial for mRNA stability and also for translation initiation (Passmore & Coller, 2022).

The primary transcript (pre-mRNA) undergoes the process of splicing (deletion of intron sequences) in spliceosomes located in nucleus. Splicing can be alternative i.e. may result in different transcripts which has direct consequences for amino acid sequence of resulting polypeptide and enables coding of multiple proteins by a single gene. Most common modes are exon skipping and intron retention, the latter being most prevalent in plants (Wang & Brendel, 2006). Apart from determining the sequence of mRNA, splicing facilitates (but is not necessary for) nuclear export as some proteins involved in latter process recognise spliceosome complex; RNA modifications and *cis*-elements also promote nuclear export or retention (Palazzo & Lee, 2018). The transcript is bound by yet another protein complex, which connects to nucleoporins and releases mRNA into cytoplasm.

2.2.2 Post-transcriptional regulation

Abundance of mRNA in cytoplasm is regulated by degradation processes. Poly-A binding proteins protect transcripts from nuclease activity, therefore most general degradation pathways require deadenylation. Subsequently, mRNA can be degraded by decapping enzymes or exosome complex (Garneau et al., 2007). Furthermore, transcript decay may be translation dependent, in nonsense-mediated decay (Brogna & Wen, 2009) or mediated by *cis*-acting AU-rich (Chen & Shyu, 1995) or GU-rich elements (Vlasova-St. Louis & Bohjanen, 2011) in 3' UTR. Transcription and mRNA degradation are mutually regulated forming a "feedback loop" (Hartenian & Glaunsinger, 2019).

Another prominent mechanism of gene silencing and transcript degradation is RNA interference. The effectors of this process are small RNA molecules: siRNA (small interfering), miRNA (micro) and piRNA (piwi-interacting), though the latter class is not present in plants and thus will not be discussed here. In plants, miRNA arises from short dsRNA fragment cut from primary miRNA transcript by Dicer-like 1 enzyme (Kurihara & Watanabe, 2004) and joins Argonaute protein to form RNA-induced silencing complex (RISC), which is later transported out of nucleus (Bologna et al., 2018). siRNA pathways include similar mechanisms, but the primary transcript can arise from various sources, possibly primed by other small RNAs

(secondary silencing) (Carthew & Sontheimer, 2009; Guo et al., 2016; Sanan-Mishra et al., 2021). RISC extracts the guide strand from the dsRNA fragment and binds to transcript with complementary sequence (Meister & Tuschl, 2004; Tomari & Zamore, 2005). siRNA (mainly) recognises completely complementary sequences, while miRNA accepts partial matches; hence miRNA might silence a range of transcripts. In case of a sufficient match, the mRNA is cleaved by Argonaute protein and the fragments are subsequently degraded; the silencing may also be caused by recruitment of decapping and deadenylation enzymes leading to mRNA decay and by translation repression (primarily at the initiation stage) (Carthew & Sontheimer, 2009; Wilczynska & Bushell, 2015; Duchaine & Fabian, 2019). The two classes of small RNAs share some functionality, however miRNA activity is mainly coupled with gene regulation, while siRNA is also involved in response to sequences of exogenous origin (TEs, viral) (Carthew & Sontheimer, 2009). Sequences recognised by small RNAs are often part of 3'-UTR and as such the binding takes place mostly (but not exclusively) post-transcriptionally.

2.2.3 Translational and post-translational regulation

Translation consists of three main stages: initiation, elongation and termination, each enabled by assistance of respective protein factors. The rate-limiting process is thought to be initiation (Palmiter, 1975), perhaps due to multiple factors involved (Merrick & Pavitt, 2018). Initiation depends on effective mRNA concentration (resulting from interplay of degradation, isolation in RNA granules and binding of various proteins), its secondary structure, availability of ribosomes and required initiation factors in active form (Hershey et al., 2019). Elongation factors are less numerable, but their modifications also play a part in regulation of translation (Proud, 2019). RNA structural features and various factors interacting with elongation machinery cause translation recoding (Dever et al., 2018). Availability of particular tRNAs is a prerequisite for translation and is subject to further regulation (Wilusz, 2015).

In order to achieve its functionality, the polypeptide chain is folded and transported to appropriate location; it may be also additionally modified e.g. phosphorylated, glycosylated, acetylated etc. Many of these modifications are reversible and thus may be used for switching protein activity on and off. As protein factors were present throughout all the steps of gene expression, the role of post-translational modifications is evident. In terms of regulation, rate of protein degradation seems to be equally important. A prominent mechanism of degradation management is protein tagging, commonly using ubiquitin, which directs proteins to proteasomes. Ubiquitinylation involves three main proteins (E1, E2 and E3); E3 in particular

ligates ubiquitin to target, with specific enzymes recognising various domains. Plants have well over 1000 variants of E3, indicating its relevance for expression regulation (Chen & Hellmann, 2013). Ubiquitin proteasome systems are incorporated in many processes in plants e.g. in signalling pathways and response to stress (Sadanandom et al., 2012). Protein degradation occurs also non-specifically during general cell processes such as senescence or apoptosis.

2.3 Interspecific hybrids

Interspecific hybrids are formed by crossing organisms from two distinct species. Most hybrids are either homoploid, retaining the ploidy of original species or allopolyploid with increased ploidy, which is usually achieved through whole genome duplication. The first scenario is rarer, one of the possible reasons being the lack of chromosomal homology between parental genomes that hampers the ability to produce viable gametes, due to the inability to properly pair during meiosis. In spite of that, cases of homoploid hybrid species have been reported, with putative examples in butterflies (Salazar et al., 2010; Kunte et al., 2011), fish (Stemshorn et al., 2011) and several plant taxa (as listed by Yakimowski & Rieseberg (2014)) among others. The latter option, allopolyploidy, is fairly common in plants; in fact many well-known and economically important species are allopolyploids e.g. durum wheat, peanut and upland cotton are allotetraploids, whereas bread wheat, tall fescue and oat are allohexaploids. Moreover, polyploidy events occurred multiple times across all lines of plant (angiosperm) evolution (Jiao et al., 2011; Ruprecht et al., 2017).

Whole genome duplication (WGD) can occur by spontaneous DNA doubling in a cell; if this takes place in a hybrid, preferably in zygote or young embryo, proper pairing of homologous chromosomes is restored and viable gametes can be produced. Another mechanism involves a triploid bridge – firstly, an unreduced (2n) and normally reduced (n) gamete form a triploid hybrid, which grows and again may produce an unreduced gamete. Second fusion with normal (reduced) gamete evens out the chromosome number to tetraploidy, restores pairing and fertility (Ramsey & Schemske, 1998; Grotewold et al., 2015).

Spontaneous emergence of interspecific hybrids in nature and traces of hybrid origin in many lineages prompt the question about their significance in adaption and speciation; current opinions on this topic vary (Nieto Feliner et al., 2020). Interspecific hybridisation certainly plays a significant role in cultivation as it can be used to combine desired traits from different species (for instance, *Triticale* has the quality of wheat grain and tolerance to abiotic and biotic

stresses of rye). Furthermore, it may facilitate interspecific gene flow through introgression, which can provide adaptive advantage – for instance introgression from *Arabidopsis lyrata* to *A. arenosa* increased drought resistance in unwelcoming serpentine habitat (Arnold et al., 2016).

Additionally, hybrids often display hybrid vigour (alternatively heterosis): they benefit from greater expression of a certain trait such as biomass production compared to their parents. There are several models for underlying mechanism of allele interaction, particularly the dominance (complementing alleles of independent loci), overdominance (synergy of alleles of one locus) or pseudo-overdominance (complementation of linked loci) theory (Lippman & Zamir, 2007), with indications that many hybrids display their mixture (Chen, 2010). On the other hand, problems with low viability, infertility and even lethality may arise as well. One explanation for lower hybrid fitness is provided by Bateson-Dobzhansky-Muller model of genome incompatibilities, which proposes that divergence between parental lineages in a system of interacting genes can produce a combination of incompatible alleles in hybrid genome. The underlying mechanisms include loss of function in paralogs, for instance *pTAC14* gene in yellow monkeyflower (Zuellig et al., 2018), differences in epigenetic regulation as is the case for *HISNA* and *HISNB* genes in *A. thaliana* (Blevins et al., 2017) and dissimilar small RNA activity (Ha et al., 2009; He et al., 2010). A special case of incompatibility stems from conflict between nuclear and organellar genome (Burton et al., 2013), as the latter is usually inherited only maternally. Other factors include hybridisation load i.e. possible accumulation of deleterious mutations and irregular chromosome pairing.

2.3.1 Genomic changes in allopolyploids

Following hybridisation and WGD, the genome of a newly emerged allopolyploid must tackle several challenges originating from interactions of distinct parental genomes, now enclosed in one nucleus. After initial shock, the hybrid genome stabilises over time and starts to resemble that of diploid – in a process of diploidisation, which, as argued by Dodsworth et al. (2016), contributed greatly to angiosperm evolution. Firstly, hybrid genome's size often differs from the expected sum of parent genomes – this is a result of genome resizing. Downsizing is more common in angiosperms (Leitch & Bennett, 2004) and is thought to occur most drastically directly after polyploidization event and then slow down, with the estimated average of at most several hundred bp loss per generation (Wang et al., 2021), although the exact rates of this process across different species remain to be researched. Molecular disadvantages of larger

genomes (particularly high investment of nitrogen and phosphorus in nuclear acid maintenance and increased cost of metabolism in larger cell) and wide-spread recombination processes leading to DNA loss may be responsible for prevalence of downsizing (Wang et al., 2021); however upsizing is known to occur as well (albeit rarely) e.g. in *Nicotiana* genus (Leitch et al., 2008), presumably due to increased TE activity (Renny-Byfield et al., 2013).

Secondly, the presence of multiple sets of chromosomes leads to changes and abnormalities of meiosis and increased possibilities of recombination. In allopolyploids both homologous and homoeologous pairing is possible, but the latter may be limited due to differences (in structure and sequence) between chromosomes of parental genomes and additionally regulated by molecular regulators – examples of systems including textbook example of *Ph1* in bread wheat controlling pairing can be found among plants (Jenczewski & Alix, 2004). Chromosomal rearrangements are common, especially in recently emerged polyploids (Chester et al., 2012), where redundancy in genetic material may improve tolerance of such changes. They may result in gene loss and variations of chromosome numbers (dysploidy), thus contributing to downsizing (Mandáková & Lysak, 2018). Homoeologous exchanges in particular could influence expression patterns of genes associated with them (Gaeta et al., 2007; Li et al., 2019) and are potentially sources of genetic novelty as they have been found to occur often in coding regions (Zhang et al., 2020). Gene expression is greatly affected in general and is further discussed in next section.

The large-scale changes of downsizing are accompanied with process of genome fractionation at the sequence level. As a result of WGD the genes are duplicated, but the duplicates are gradually lost, presumably due to relaxed selection pressure. This can occur in a number of ways (Innan & Kondrashov, 2010; Freeling et al., 2015): the redundant gene may be degraded to pseudogene or deleted altogether, undergo process of subfunctionalisation and therefore perform the original function only coupled with its partner or gain new function (neofunctionalisation). Gene loss occurs through illegitimate recombination (Devos et al., 2002) and intrachromosomal recombination in tandem with deletion (Woodhouse et al., 2010), while the other fates are driven by accumulating mutations. Given that those changes generally occur in any genome, an interesting take is to focus on genes that are retained duplicated. There is in fact a bias in fractionation concerning different types of genes – those with house-keeping functions, particularly involved in DNA maintenance and chloroplast functioning, are preferentially kept in only one copy, while others (e.g. related to stress response or transcription factors) are more likely to survive in duplicates (De Smet et al., 2013; Li et al., 2016;

Mandáková et al., 2017). Several explanations for this bias had been put forth, most of them as a part of gene balance hypothesis. Genes involved in multi-protein interactions (such as transcription regulation) are more vulnerable to mutations (Freeling & Thomas, 2006); conversely proteins performing solitary functions such as DNA repair may be more tolerant. Significant representation of organelle-related (chloroplast and mitochondria) genes among singletons could result from coordination with organellar genomes (Duarte et al., 2010), which are not included in polyploidy-inducing WGD.

Finally, it is worth mentioning that the processes of genome restructuring described above are rarely symmetrical – usually they are biased in favour of one parental genome. Genome dominance (also subgenome dominance) may evince itself through elimination of chromosomes from the submissive subgenome, replacing them with those of the dominant (e.g. Majka et al. (2023)) or preferential gene loss (Thomas et al., 2006), but mostly through changes in expression (see below).

2.3.2 Dynamics of gene expression in allopolyploids

While structural changes ultimately shape the emergent genome of allopolyploid, changes in expression are usually the front line of genome merger response and have been dubbed 'transcriptomic shock'. Gene dosage of singular homoeolog or a pair of them (or triad etc.) can be affected; this is reflected in two most common types of bias: homoeolog expression bias (HEB) and expression level dominance (ELD). HEB reflects possible expression bias towards homoeologs of one parent, while ELD compares the absolute expression of homoeolog pair with expression level of parents (Grover et al., 2012). Originally these terms refer to single homoeolog system, but can be scaled up to reflect general trends in hybrid genomes. The two types of bias are not necessarily connected, as parental genome dominant in sense of HEB might be submissive in sense of ELD (e.g. as shown by Li et al. (2020) in some tissues of *Brassica napus* or by Glombik et al. (2021) in *Festulolium*).

HEB is linked with genome dominance in sense of fractionation bias discussed above, since more expressed genes tend to be preferentially retained (Schnable et al., 2011). While not all allopolyploid systems exhibit genome dominance (Douglas et al., 2015; Sun et al., 2017), in those that do it has been observed to be hereditary, consistent between natural and resynthesized polyploids and intensifying over generations (Edger et al., 2017), but the direction of bias may differ across organs or tissues e.g. in cotton (Samuel Yang et al., 2006; Flagel et al., 2008) or

blueberry (Colle et al., 2019). Moreover, it is important to underline that HEB at the genome level does not apply universally to every single homoeolog pair – some genes of submissive parental genome may exhibit greater expression than their counterparts from dominant genome (Grover et al., 2012). Given that, the question arises: how many individual biases and in what context are needed to declare one parental genome being dominant? These observations highlight the fact that unambiguous criteria for definition of genome dominance remain to be determined (Alger & Edger, 2020). Several factors deciding which subgenome will become dominant have been proposed, the two leading models indicating TE load and effectiveness of *cis-trans* regulation. Greater amount of methylated TE in close proximity to genes is found in submissive parental genomes (Edger et al., 2017). Differences in *cis*-site affinity and *trans*-factors specificity combined with possibly changed accessibility of binding sites have been theorised to differentiate expression levels between homoeologs (Bottani et al., 2018).

Total expression level of homoeolog system constitutes a scale of possible outcomes. The midway option is additive expression i.e. equal to average of parental expression levels, which can deviate to non-additive expression in ELD (equal to one parent) or transgressive expression (up- or down-regulation beyond both parents). Non-additive expression seems to be more common in older allopolyploids (Boatwright et al., 2021), especially in regard to transgressive expression (Yoo et al., 2013). Similarly to HEB, ELD direction may differ in various tissues (e.g. in cotton (Flagel & Wendel, 2010; Yoo et al., 2013), *B. napus* (Wu et al., 2018; Li et al., 2020) and *Raphanobrassica* (Ye et al., 2016; Zhang et al., 2021)) or conditions (Bardil et al., 2011). Changes in regulation of submissive homoeologs are postulated as the main cause of ELD (Yoo et al., 2013; Cox et al., 2014; Combes et al., 2015; Glombik et al., 2021). The role of *cis-trans* interactions has been highlighted for ELD as well; the strength of *trans*-acting factors has been proposed to determine dominant parental genome (Hu & Wendel, 2019).

Epigenetic regulation seems to be a major player in expression changes. DNA methylation re-patterning and changes in histone modifications have been observed in several allopolyploids (Song & Chen, 2015; Ding & Chen, 2018). DNA methylation and histone deacetylation is responsible for nucleolar dominance i.e. expression of rRNA from only one parental genome in interspecific hybrids (Chen & Pikaard, 1997; Chen et al., 1998). Methylation changes at TEs can lead to their activation, which has plenty of consequences for the whole genome (see below and section 2.1.3). Epigenetic modifications of genes have been shown to have direct impact on phenotypes e.g. flowering time is influenced by histone modifications in *Arabidopsis*

allotetraploid (Wang et al., 2006) and by demethylation in allotetraploid cotton (Song et al., 2017).

Distinct populations of small RNAs, which silence both genes and TEs, naturally affect gene expression in allopolyploids. Non-additive expression of miRNA causes asymmetrical gene silencing (Ha et al., 2009); furthermore, lower levels of siRNA directly after hybridisation may contribute to increased TE activity (Ha et al., 2009; Groszmann et al., 2011). Small RNA preferentially target TE near less expressed homoeologs (Li et al., 2014; Woodhouse et al., 2014), tying into TE-based model of HEB.

The amount of alternative splicing (AS) events, influencing functional effects of transcripts, is also affected by polyploidisation. Decrease in general number of events and differences in AS patterns between homoeologs – loss of AS event in one parental genome or even mutually exclusive distribution of AS events in a manner possibly leading to subfunctionalisation – were reported in *Brassica* (Zhou et al., 2011) and wheat (Yu et al., 2020; Jia et al., 2022). AS events were more common in resynthesized allopolyploids compared to their natural counterparts, suggesting that they are part of initial response and their number diminishes in course of genome stabilisation (Zhou et al., 2011). Their frequency also changed during embryo development in wheat (Jia et al., 2022), implying that patterns may differ depending on development stage.

Lastly, the activation of TEs is worth mentioning here, as it has profound effects on gene expression (and long-term genome structure). Changes in TE repression and relaxed selection could enable TE activation. TE transcription activity was found to increase in polyploids, but not all genomes and TE families respond to new conditions (reviewed extensively by Vicient & Casacuberta (2017)). Moreover, the activation effect seems to be temporary (Kraitshtein et al., 2010). TE insertions may influence expression patterns and gene structure (as discussed in section 2.1.3).

Lineages or species	Algorithms	*E*-value cutoff	Number of *OGs*	Percentage	References
Brassica rapa	BLAST	1E-03	529 real A subgenome-specific *BSGs* (*Brassica*-specific genes) (these 529 *BSGs* also named as *BrOGs*)	1%	Jiang et al., 2018; Jiang M. et al., 2020
Arabidopsis thaliana	BLAST	1E-03	958 lineage-specific genes (*LSGs*)	3%	Donoghue et al., 2011
	BLAST	1E-01	165 *Arabidopsis*-specific genes	1%	Yang et al., 2009
	BLAST	1E-05	1,324 *Arabidopsis* lineage-specific genes (*ALSGs*)	5%	Lin et al., 2010
	BLAST	1E-03	861 species-specific orphan genes (*SSOG*)	3.14%	Cui et al., 2015
Oryza sativa	BLAST	1E-01	638 *Oryza*-specific genes	1%	Yang et al., 2009
	BLAST	1E-04	1,926 *OGs*	3%	Guo et al., 2007
	BLAST and BLAT	1E-02	37 *OGs*	0.0006%	Jin et al., 2019
	BLAST	1E-03	478 *SSOG*	1.18%	Cui et al., 2015
Populus trichocarpa	BLAST	1E-01	109 *Populus*-specific genes	0.2%	Yang et al., 2009
	BLAST	1E-02	40 *Populus trichocarpa*-specific genes (*PtSS*)	0.3%	Lin et al., 2013
Vigna unguiculata	BLAST and Microarray-based genome hybridization	1E-10	578 cowpea *OGs*	2%	Li G. et al., 2019
Poaceae	BLAST	1E-05	861 conserved Poaceae-specific genes (*CPSGs*)	2%	Campbell et al., 2007
Aegiceras corniculatum	BLAST	1E-05	4,823 *Aegiceras*-specific genes (*ASGs*)	12%	Ma et al., 2021
Citrus sinensis	BLAST	1E-05	1,039 *OGs* specific to *Citrus sinensis*	4%	Xu et al., 2015
Citrullus lanatus	BLAST	1E-05	1,652 *OGs*	7.31%	Ma et al., 2022
Lagenaria siceraria	BLAST	1E-05	870 *OGs*	3.87%	Ma et al., 2022
Sechium edule	BLAST	1E-05	627 *OGs*	1.63%	Ma et al., 2022
Cucumis sativus	BLAST	1E-05	2,524 *OGs*	10.38%	Ma et al., 2022
Cucumis melo	BLAST	1E-05	2,287 *OGs*	7.63%	Ma et al., 2022
Cucurbita moschata	BLAST	1E-05	2,498 *OGs*	7.76%	Ma et al., 2022
Trichosanthes anguina	BLAST	1E-05	529 *OGs*	1.65%	Ma et al., 2022
Benincasa hispida	BLAST	1E-05	4,547 *OGs*	16.55%	Ma et al., 2022
Camellia sinensis	BLAST	1E-05	1,701 *Camellia*-specific genes (*CSGs*)	3.37%	Zhao and Ma, 2021
Cajanus cajan	BLAST	1E-02	266 Phaseoleae-restricted ORFans, 169 out of 266 genes are putative pigeonpea-specific ORFan genes.	0.6%	Varshney et al., 2011

Fig. 2: Review of orphan genes identified in various plant species. For species with several identification studies impact of different cut-off values can be observed – the higher the value, the lower the percentage of orphan genes. Adopted from Jiang et al. (2022)

2.4 Lineage-specific genes

Lineage-specific genes (LSGs) are genes present only in given taxon, with no known homologous sequences in others (therefore term 'taxonomically restricted genes' is sometimes used as well). The level of specificity is determined by adding the name of the limiting taxon e.g. *Arabidopsis*-specific, *Brassicaceae*-specific. Usually this name refers to species-specific genes, otherwise called orphan genes or ORFans. In contrast to them, evolutionary conserved genes appear across multiple (not necessarily closely related) species. The existence of orphan genes was first reported when the sequencing of yeast genome was completed (Dujon, 1996). Since then, sets of lineage-specific genes were found in majority of sequenced genomes, although estimates of their content in a species' genome varies greatly, starting with around 1% (Ma et al., 2020) and reaching 70% (Gibson et al., 2013), the typical value being 10-30% of all genes (Khalturin et al., 2009; Tautz & Domazet-Lošo, 2011; Wissler et al., 2013). For plants the general range is 5-15% (Arendsee et al., 2014), however most results place themselves in

the lower end of the range (the typical content is 1-5% as seen in Fig. 2 reviewing the results of recent studies of several plant species (Jiang et al., 2022)). There are several probable reasons for such differences, mainly the dynamic and somewhat unequal character of databases used for finding homologs, which tends to represent some taxa significantly better due to the abundance of the data on closely related species, while having limited information on others. Therefore, with the expansion of available genomic information, homologs of some orphans may be further identified, however results of current research discourage the hypothesis that this should happen to all of them. Moreover, the number of orphan genes found depends on the type of sequence data used or algorithms and criteria deciding on sequence homology. To illustrate, several studies on *Arabidopsis thaliana*, an intensively researched and well annotated species, identified different numbers of LSGs (Lin et al., 2010; Donoghue et al., 2011; Guo, 2013; Arendsee et al., 2014). The impact of different cut-off values is also illustrated in Fig. 2 – the higher the value, the more hits in database are found and the lower the percentage of genes identified as orphans.

2.4.1 Sequence features of lineage-specific genes

Lineage-specific genes have been found to differ from evolutionary conserved genes in more than just the presence/absence of homologs. Their protein length is significantly shorter (Ekman & Elofsson, 2010; Xu et al., 2015; Ma et al., 2020; Ma et al., 2021), however their gene length can be either shorter (Gibson et al., 2013; Yang et al., 2013; Xu et al., 2015) or longer (Ma et al., 2021) than non-orphan genes. They generally have lower number of introns and exons per gene (Lin et al., 2010; Ma et al., 2020). It has been reported that this is to some extent compensated by longer exon or intron sequences (Lin et al., 2010; Xu et al., 2015; Ma et al., 2020), but it does not seem to be a universal trend. Perhaps the most variable characteristic is GC content, which generally distinguishes LSGs from other genes, but has been found to be both lower (Donoghue et al., 2011; Gibson et al., 2013; Ma et al., 2020; Ma et al., 2021) and higher (Lin et al., 2010; Yang et al., 2013; Sun et al., 2015; Xu et al., 2015) in different species, suggesting that this property is species-specific. In conclusion, LSGs prove to be distinct in their sequence features from other genes in given lineage, but the characteristics are not universal across different studies. Considering that the question of homology cannot be yet fully answered, sequence properties provide additional indication for lineage specificity of genes and their evolutionary origin.

2.4.2 Lineage-specific genes' emergence

Evolution of genome is thought to be a steady process of gradually accumulating changes, therefore the existence of species-specific genes (SSGs), completely stranded in world of genes is somewhat perplexing and encourages inquiries about presumably drastic changes that led to their emergence. Two possible explanations are evident: SSGs had homologs in other species, but the similarity has since been lost or they are newly emerged. Indeed, many characteristics of SSGs present them as young genes - particularly their shorter protein length (see above), weaker expression (discussed later) and higher evolutionary rate (Lin et al., 2010; Donoghue et al., 2011). The emergence of novel genes is generally attributed to divergent duplication and *de novo* formation, with several other possible mechanisms occurring less frequently. Contribution of different mechanisms to orphan genes' origin has been suggested to differ for specific species or taxa (Wissler et al., 2013); nevertheless existence of mixed-origin genes and ongoing research into possible means of gene birth complicates its quantification (Prabh & Rödelsperger, 2019).

The first major mechanism is divergent duplication. Genes in general may be duplicated due to transposon activity or recombination (among others). This causes genetic redundancy allowing its significant divergence, particularly if accompanied by considerable modifications of sequence motifs (Tautz & Domazet-Lošo, 2011). Although this model of gene birth is believed to be fairly common, in order to lose recognisable homology and generate a SSG the divergence rate would need to be adequately high; consequently this mechanism is less wide-spread in case of orphan genes' emergence e.g. it is responsible for origin of at most one third of orphan genes in fruit fly, human and yeast (Vakirlis et al., 2020).

The other possibility is *de novo* gene birth from non-coding sequences. Two pathways are proposed: (i) an already transcribed non-genic region may acquire an ORF - "transcription first" scenario (Begun et al., 2006; Levine et al., 2006) or (ii) a region with existing ORF may gain transcriptional activity - "ORF first" scenario (Zhao et al., 2014). The transition from non-coding to coding is explained by continuum model, which presents emergence of novel genes as a spectrum – starting with non-genic sequences that gain ORFs and become transcribed through pathway (i) or (ii); then translational activity raises them to proto-genes, which may either become fixed as fully fledged genes or revert to non-coding status (Carvunis et al., 2012). Substantial evidence shows that many orphan genes might have emerged *de novo* (as reviewed by Long et al. (2013), McLysaght & Guerzoni (2015), Van Oss & Carvunis (2019), Singh & Syrkin Wurtele (2020)). Other plausible and possibly non-exclusive models of *de novo* birth of

orphan genes have been proposed as well (Van Oss & Carvunis, 2019), but distinguishing between them is challenging.

Several alternative mechanisms for gene birth have been suggested. Rise from existing genes (without duplication) through exon rearrangement, gene fusion or fission would, similarly to divergent duplication, require complete loss of homology to parental gene (as reviewed by Long et al. (2013)). A particular scenario is birth of gene from alternative ORF called overprinting (Samandi et al., 2017). Horizontal gene transfer could be considered a special case of gene duplication and likewise is dependent on fast divergence of donor and/or recipient species (Husnik & McCutcheon, 2018) and seems to rarely produce orphan genes (Wissler et al., 2013; Schlötterer, 2015). Similarly, gene transfer between nuclear and non-nuclear (e.g. mitochondrial) genomes may generate species-specific genes as well (O'Conner & Li, 2020). Apart from their possible role in divergent duplication, TEs may be partly (exonisation) or directly (exaptation) adapted into new genes (Toll-Riera et al., 2009; Donoghue et al., 2011; Joly-Lopez & Bureau, 2018; Jin et al., 2021). The process of gene birth in many aspects mirrors pseudogenisation and reuse of pseudogenes has been reported (Brosch et al., 2011), which may be counted as edge case of *de novo* gene emergence.

Lastly, species-specific genes may not be novel, but rather last surviving. Due to approximately constant number of genes in genome (Tautz & Domazet-Lošo, 2011) and at the same time influx of novel genes, balancing gene repertoire by gene loss is a logical explanation. Probability of losing a gene increases with decreasing gene age; consequently lost genes are most likely still young if not novel (Palmieri et al., 2014). In case of novel genes, detection of loss of short-lived homologs may not be possible and orphan genes could be classified as simply *de novo* emerged. The more interesting case is loss of older and well-established genes (Zhao et al., 2015); however this model of orphan gene origin is mainly mentioned as theoretical (Guo, 2013; Zhang et al., 2019).

In summary, orphan genes may originate in several different life stages of sequences (Fig. 3). Plenty of plausible models with some experimental validation were proposed and, as mentioned above, many of them are not mutually exclusive; consequently emergence of an orphan gene is likely to result from several coinciding processes (Prabh & Rödelsperger, 2019).

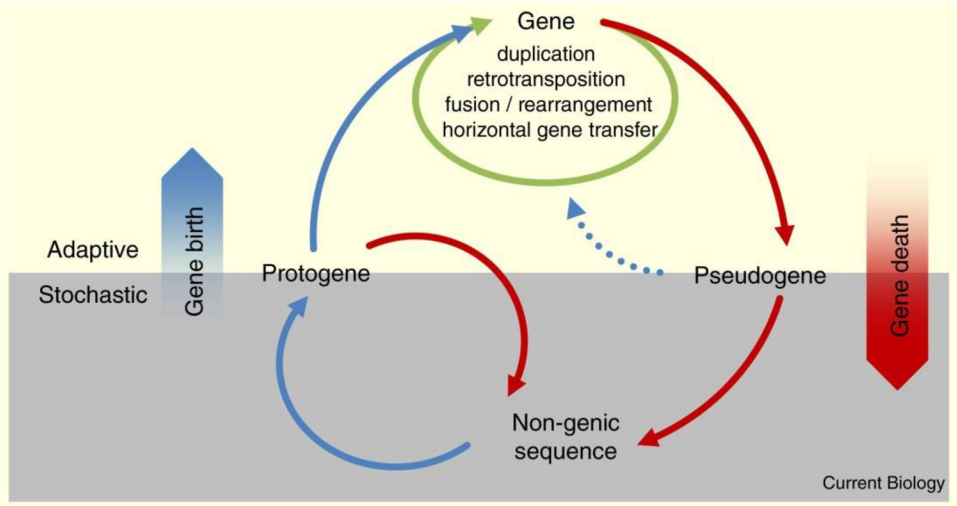


Fig. 3: Life cycle of genes. Blue arrows represent processes leading to birth of new genes from non-coding sequences, green arrow points to the emergence of new ones from previously existing, while red arrows represent the loss of coding potential. Adopted from Neme & Tautz (2014).

2.4.3 Expression of lineage-specific genes

For the justified use of term genes, proof of LSGs' expression is required. However, it is often difficult to confirm it, as the LSGs generally have lower and more tissue-specific expression compared to their conserved counterparts (Donoghue et al., 2011; Xu et al., 2015; Sun et al., 2015; Ma et al., 2020; Ma et al., 2021), with particular emphasis on reproduction-related tissues and organs such as testis in *Drosophila* (Begun et al., 2006; Levine et al., 2006), gonads in zebrafish (Yang et al., 2013) and mature pollen in *Arabidopsis* (Wu et al., 2014). These expression patterns make it easy to miss or dismiss orphans as statistically insignificant. Many are differentially expressed in biotic and especially abiotic (Donoghue et al., 2011; Xu et al., 2015; Cardoso-Silva et al., 2022) stress conditions. The expression study is of utmost importance, as it provides the clues for understanding orphans' functionality.

While many sequences in genome may be expressed at low levels due to pervasive transcription (David et al., 2006; Kapranov et al., 2007) and translation (Ingolia et al., 2014), stabilisation of expression of young orphan genes and their possible fixation in genome is dependent on recruitment of regulatory element(s) as conserved orphans show higher expression (Palmieri et al., 2014). For genes arising through divergent duplication, the necessary elements may have been copied from the original gene as well to ensure gene transcription. Furthermore, genes arising in place of previously existing genes may adapt their regulatory machinery; moreover, novel genes may similarly emerge in locations associated with existing elements (Tautz & Domazet-Lošo, 2011; Abrusán, 2013). Finally, it is possible that

novel regulatory elements form alongside or following the ORF (ORF-first model) (Heinen et al., 2009).

2.4.4 Function of lineage-specific genes

Functional characterisation of LSGs proves difficult. Most computational approaches are based on homology with known protein domains or motifs; however, this method usually fails when it comes to LSGs, given that finding a homologous domain would most likely mean a homology with a certain group of proteins, thus contradicting the very definition of the LSG.

Although many LSGs remain of unknown function, others have been experimentally determined to play a significant role in lineage-specific structures or reactions to stress response to both biotic and abiotic factors, with some being essential (as is the case of a few *Drosophila* genes, whose lack is embryo-lethal (Reinhardt et al., 2013)). *Prod1*, an orphan gene in salamanders with ability to regenerate limbs, is necessary for limb formation (Kumar et al., 2015). In *Hydra* LSGs of *Hym301* family influence formation of tentacles, which is an important adaptive trait given that tentacles are necessary for acquiring food (Khalturin et al., 2008). Surface antigens, vital in parasite-host interactions, are coded by species-specific genes in *Plasmodium* and genus-specific in *Theileria* (Kuo & Kissinger, 2008). *Aegiceras*-specific genes are associated with pathways crucial for adaptation of the mangrove plants to their high saline tidal environment (Ma et al., 2021). Orphans may also play key roles in primary metabolisms, though they are not known to perform catalytic functions (Arendsee et al., 2014). For instance, orphan genes of *Brassica rapa* were found to alter soluble sugar contents (Jiang et al., 2020). Overview of many functions exhibited by orphan genes in plants, with particular emphasis on stress response, is presented in Fig. 4.

Perhaps the best studied example of orphans in plants is QQS (qua-quine starch), a gene of *A. thaliana*. QQS controls carbon and nitrogen distribution to starch, lipid and protein synthesis in leaves and seeds (Li et al., 2009; Li & Wurtele, 2015). Its high expression supports proteosynthesis, while QQS repression shifts the balance of cellular metabolism in favour of starch production. Expression of QQS is influenced by several biotic and abiotic stimuli and thus, it has been suggested that this gene helps regulating primary metabolism in response to environmental and developmental changes (Li et al., 2009; Arendsee et al., 2014). Through interactions with conserved transcription factors NF-YC4, QQS has been shown to function

Gene names	Abbreviations	Gene symbols	GenBank accession numbers	Functions	References
Qua-Quine Starch	*QQS*	*At3g30720*	EU805808	Carbon and nitrogen allocation across species; genetic and environmental perturbations response; pathogens/pests resistance.	Li et al., 2009; Silveira et al., 2013; Arendsee et al., 2014; Li and Wurtele, 2015; Li et al., 2015; Jones et al., 2016; O'Conner et al., 2018; Qi et al., 2019; Tanvir et al., 2022
Brassica rapa Orphan Gene 1	*BrOG1*	*BraA08002322, BraSca000221*		Soluble sugar metabolism regulation.	Jiang M. et al., 2020
Male Sterile 1	*Ms1*		KX447407, KX447408, KX447409	Male fertility and metagenesis regulation, pollen exine development.	Tucker et al., 2017; Wang et al., 2017
Male Sterile 2	*Ms2*		KX533929	Conferment of male sterility.	Ni et al., 2017
Enhancer of Vascular Wilt Resistance 1	*AtEWR1*	*At3g13437*	DQ487672	Drought tolerance; fungal pathogens resistance.	Yadeta et al., 2014
Brassica oleracea Enhancer of Vascular Wilt Resistance 1	*BoEWR1*			Resistance against fungal pathogens.	Yadeta et al., 2014
Big Root Biomass	*BRB*	*SIN_1025576*	MN336257, MN336258, MN336259	Root biomass modulation.	Dossa et al., 2021
Triticum aestivum Septoria-responsive Taxonomically Restricted Gene 6	*TaSRTRG6*	*TraesCS1A01G265600, TraesCS1B01G276500, TraesCS1D01G265800*		Septoria tritici blotch resistance.	Brennan et al., 2020
Triticum aestivum Septoria-responsive Taxonomically Restricted Gene 7	*TaSRTRG7*	*TraesCS3A01G093900, TraesCS3B01G109200, TraesCS3D01G094200*		Septoria tritici blotch resistance.	Brennan et al., 2020
	Xa7		MW467511	Bacterial pathogen resistance.	Wang et al., 2021
Triticum aestivum Fusarium Resistance Orphan Gene	*TaFROG*		KR611570	Resistance to the Fusarium head blight disease.	Perochon et al., 2015; Perochon et al., 2019; Jiang C. et al., 2020
UP12_8740				Drought resistance.	Li G. et al., 2019
Oryza sativa ornithine decarboxylase	*OsODC*	*LOC_Os09g37120*		Biosynthesis of hydroxycinnamoyl putrescine.	Fang et al., 2021
Oryza sativa putrescine hydroxycinnamoyl acyltransferases 3	*OsPHT3*	*LOC_Os09g37180*		Biosynthesis of hydroxycinnamoyl putrescine; immunity and cell death regulation.	Fang et al., 2021
Oryza sativa putrescine hydroxycinnamoyl acyltransferases 4	*OsPHT4*	*LOC_Os09g37200*		Biosynthesis of hydroxycinnamoyl putrescine; immunity and cell death regulation.	Fang et al., 2021
Oryza sativa Pyridoxamine 5'-phosphate oxidase 3	*OsPDX3*	*LOC_Os10g23120*		Phenylpropanoid metabolism; bacterial and fungal pathogen resistance.	Shen et al., 2021
Oryza sativa tyrosine decarboxylase 1	*OsTyDC1*	*LOC_Os10g23900*		Phenylpropanoid metabolism; bacterial and fungal pathogen resistance.	Shen et al., 2021
Oryza sativa tyramine N-Hydroxycinnamoyltransferase 1	*OsTHT1*	*LOC_Os10g23310*		Phenylpropanoid metabolism; bacterial and fungal pathogen resistance.	Shen et al., 2021
Oryza sativa tyramine N-Hydroxycinnamoyltransferase 2	*OsTHT2*	*LOC_Os10g23820*		Phenylpropanoid metabolism; bacterial and fungal pathogen resistance.	Shen et al., 2021
Grain Shape Gene on Chromosome 9	*GS9*	*LOC_Os09g27590*	MF621928	Rice grain shape and appearance quality regulation.	Zhao et al., 2018
GRAINS NUMBER 2	*GN2*			Regulation of grain number, plant height, and heading date.	Chen et al., 2017
Oryza sativa defense-responsive gene 10	*OsDR10*		FJ194952	Negative Regulation of pathogen-induced defense response.	Xiao et al., 2009
Xoo-induced orphan 1	*Xio1*	*Os09g13440*		Bacterial pathogen resistance.	Moon et al., 2022

Fig. 4: Summary of recent findings in orphan genes functional characterisation. Orphan genes are involved in primary metabolism and stress response to biotic and abiotic stimuli in many agriculturally important plants. Adopted from (Jiang et al., 2022).

after introduction into soybean, maize, rice and tobacco and additionally improve pathogen and pest resistance (Li et al., 2015; Qi et al., 2019; Tanvir et al., 2022).

The main aim of my BSc. thesis was to identify species-specific genes for *Lolium multiflorum* and *Festuca pratensis*, two related grass species of the Poaceae family. The additional part of the thesis covered analysis of their expression in the particular species and in the interspecific hybrids. Such information is of high value, because to my very best knowledge the analyses on the retention and expression of orphan genes in interspecific hybrids, and their potential contribution to the hybrid genome evolution and function has not been elaborated till now. The only study to investigate the fate of orphan genes in interspecific hybrids was done in *Saccharum* (sugarcane) – the genes were identified based on *Saccharum spontaneum* and expression was studied in *S. spontaneum*, *S. officinarum* and *Saccharum* hybrids (Cardoso-Silva et al., 2022). The authors noted that expression patterns in *S. officinarum* and hybrids shared more similarity than those of *S. spontaneum*, presumably due to greater contribution of *S. officinarum* to hybrid genomes. However, the orphan genes were not specific to one parent; moreover, the hybrid varieties used underwent several backcrosses to *S. officinarum* after initial hybridisation. Therefore, the influence of hybridisation on orphan genes could not be fully elucidated.

3 METHODS AND MATERIALS

3.1 Transcriptome assembly

Transcriptomes of two grass species – *Festuca pratensis* Huds. (meadow fescue) and *Lolium multiflorum* Lam. (Italian ryegrass) were assembled *de novo*. RNA-seq reads were obtained from previous studies done by Czaban et al. (2015) and Stočes et al. (2016), BioProject accessions PRJNA266320 and PRJNA308063, and are hereafter referred to as Czaban and Stočes datasets respectively. Samples in Czaban dataset were collected from mature leafs of *F. pratensis* cv. Laura and *L. multiflorum* cv. Lemtal. In case of Stočes dataset, a total of 10 samples of young leaves from 6 genotypes of each species were used (details in Tab. 1). In both cases, samples were sequenced using Illumina platform.

Quality of the reads was controlled using FastQC software (version 0.11.5, Andrews (2010)). Reads from Stočes dataset were trimmed to remove adapter sequences and low-quality reads were filtered out using Trimmomatic (version 0.39, Bolger et al. (2014)) with minimal required length (MINLEN) of 50 nucleotides; data from Czaban dataset were already trimmed. Following trimming, only read pairs with both reads (forward and reverse) were used in further analysis. Reads from both datasets were assembled using Trinity (version 2.11.0, Grabherr et al. (2011)). In order to lower data redundancy, sequences from transcriptomes were clustered using CD-HIT (version 4.6.1, Li & Godzik (2006), Fu et al. (2012)) with identity threshold of 0.95. Subsequently, TransDecoder (version 3.0.1, Haas) was used to select potential coding regions and predict sequences of their peptide products; the program was trained on annotated barley (*Hordeum vulgare*) data (Mascher, 2021). Assemblies were validated by BUSCO software (version 3.0.2, Simão et al. (2015), Waterhouse et al. (2018)) with Embryophyta lineage dataset (odb9) and wheat as species parameter.

Tab. 1: Sources of samples used for RNA-seq in cited studies. Coloured cultivars are tetraploid, the rest are diploid. Numbers in braces represent number of collected samples. Two genotypes, each with three samples, were used for "Westa" and "Mitos" cultivars.

Cultivars	***F. pratensis***	***L. multiflorum***
Czaban dataset	Laura (1 sample)	Lemtal (1)
Stočes dataset	Fure (1)	Abercomo (1)
	Skawa (1)	Fox (1)
	WSC (1)	Prolog (1)
	Patra (1)	Sikem (1)
	Westa (2 × 3 samples)	Mitos (2 × 3 samples)

In order to annotate the transcriptomes, BLAST searches against Swiss-Prot database (using BLAST+ version 2.12.0, Camacho et al. (2009)) and domain identification using HMMER (HMMER (version 3.3.2), 2020) against Pfam database were performed. The results were integrated using Trinotate (version 3.2.2, Bryant et al. (2017)).

3.2 Identification and characterisation of species-specific genes

ORFs predicted by TransDecoder were used as gene pool in order to identify species-specific genes. Firstly, for both species the ORFs and corresponding peptide sequences were compared to peptide set of the other species using BLASTx and BLASTp, respectively. Only sequences without match in either search were kept. The remaining sequences were then mapped with GMAP (version 2019-05-12, identity threshold 0.7, (Wu & Watanabe, 2005)) to available genome references and filtered so that only genes present in their own genome and missing in the other were retained. For the selected sets, annotations in assembled transcriptomes and possible homologs in general databases (using BLAST against UniprotKB and NCBI non-redundant protein sequences (nr)) were manually checked. Furthermore, in order to validate gene prediction and check for possible undetected genes in broader region, candidate sequences mapped to genome were extracted with margin of 5 kbp on both sides and used to predict genes using Augustus (Stanke & Morgenstern, 2005), run with four pre-trained models – wheat, rice, maize and *A. thaliana*. GC content was calculated using Bedtools (version 2.26.0, (Quinlan & Hall, 2010)) and functional domains were further predicted using web-based InterProScan (Paysan-Lafosse et al., 2023).

Moreover, a set of gene models of *L. multiflorum* was used as alternative set of candidate genes to identify species-specific genes. Similarly to the first pipeline, candidate sequences were BLASTed against *F. pratensis* genome reference and downloaded Swiss-Prot database to eliminate genes with homologs. The remaining species-specific genes constitute Set II for *L. multiflorum* and the previous ones (of both species) are referred to as Set I in the following text. Bedtools and InterProScan (downloaded version 5.55-88.0) were used for characterisation of Set II sequences.

3.3 Expression analysis

Reads from “Mitos” and “Westa” samples (see section 3.1 and Tab. 1) were used to represent expression of parental organisms. *Festulolium* hybrids reads were obtained from study by Glombik et al. (2021) (under accession number PRJNA685345). Hybrids of F_1 generation from

both *F. pratensis* × *L. multiflorum* and *L. multiflorum* × *F. pratensis* crosses were used. Hybrids of F_2 generation were acquired from one self-pollinating F_1 hybrid plant of *F. pratensis* × *L. multiflorum*. Types of collected samples as well as their aliases used throughout this text are listed in Tab. 2. Parental reads were mapped to their respective transcriptomes and hybrid reads were mapped to both transcriptomes; BWA-MEM (Burrows-Wheeler Aligner) was used in all cases. Unmapped reads were further mapped to references used earlier in transcriptome assembly (Trinity assembled and clustered by CD-HIT) to ensure that sources used were truly representative. Mapped reads were counted using Samtools (Danecek et al., 2021) and imported to R Studio (version 4.2.1 (R Core Team, 2022)) for analysis. Using edgeR package (Robinson et al., 2010) genes were filtered by statistical significance (with filterByExpr function) and counts per million (CPM ≥ 1). Subsequently, differential gene expression analysis was performed on paired types of reads (parental and hybrid of given type): negative binomial dispersions were calculated (estimateDisp) and tested for differential expression between the types of samples (exactTest, which implements test proposed by Robinson & Smyth (2007)); p-values were adjusted (p.adjust) using Benjamini & Hochberg (1995) method for controlling false discovery rate. Finally, genes with differential expression between the two sets were identified (p-value < 0.05, log Fold Change < 1 or log Fold Change > -1).

Species-specific genes of *L. multiflorum* Set II were analysed for differential expression between parental and hybrid samples (Lm × Fp F1 and Fp × Lm F1) in similar manner to analysis employed for Set I. Second version of simplified analysis testing for expression in each sample type separately consisted of analogous filtering by statistical significance and counts per million (CPM ≥ 1). Ultimately, overlaps of remaining genes were examined.

Tab. 2: Sources of reads for hybrid plants. Each generation consisted of three plants and nine samples (3 plants × 3 replicates) were collected in each condition. Further details can be found in the original study (Glombik et al., 2021).

Generation	Conditions	Alias
L. multiflorum × *F. pratensis* F_1	normal, 3 weeks	Lm × Fp F1
F. pratensis × *L. multiflorum* F_1	normal, 3 weeks	Fp × Lm F1
	normal, 4 years	F1 aging
	4 years normal + 3 weeks cold treatment	F1 stress
F. pratensis × *L. multiflorum* F_2	normal, 3 weeks	Fp × Lm F2
	normal, 4 years	F2 aging
	4 years normal + 3 weeks cold treatment	F2 stress

4 RESULTS

4.1 Transcriptome assembly

Datasets used for transcriptome assembly contained around 240 million reads for *F. pratensis* and 200 million reads for *L. multiflorum*. Quality check with FastQC was positive and thus all samples were kept for further analyses. As mentioned above, only reads from Stočes dataset required trimming; the percentage of both reads from pair retained was high in all samples, with overall average of 91.49% for *F. pratensis* and 91.18% for *L. multiflorum*. Statistics for each sample are included in Tab. 3. Transcriptomes assembled with Trinity contained over 320 and 410 thousand transcripts for *F. pratensis* and *L. multiflorum* respectively; clustering reduced the numbers approximately by third. Based on those transcripts, 108 and 85 thousand ORFs were predicted; 91.3% and 92.2% complete BUSCOs (of total 1440 BUSCO groups searched) were found and 29% and 21% of transcripts were annotated. Details of these steps can be found in Tab. 4.

4.2 Identification of orphan genes

In course of identifying Set I, after reciprocal BLAST searches 102 and 57 candidate orphan genes were found for *F. pratensis* and *L. multiflorum*, respectively. Mapping to genome references left 23 and 45 sequences when using own genome, 86 and 13 in case of the other. Combining those criteria reduced the numbers to 9 and 6 sequences, though two of *L. multiflorum* genes are isoforms. Furthermore, 5 of *F. pratensis* sequences were excluded after manual inspection and gene prediction validation. As such, final Set I for *F. pratensis* consists of only 4 genes and only 5 genes (one of them having two isoforms) for *L. multiflorum*.

Tab. 3: Number of reads per source and percentage of reads kept after trimming (both reads of pair surviving). See Tab. 1 in section 3.1 for explanation of sources' names. "Laura" and "Lemtal" were already trimmed; "Mitos" and "Westa" statistics refer to all six samples bulked together.

F. pratensis	Input	Paired after filtering (%)	*L. multiflorum*	Input	Paired after filtering (%)
Fure	19385404	87.48	Abercomo	25883513	86.53
Patra	15010568	87.70	Fox	13755323	86.05
Skawa	11764708	88.31	Prolog	8982543	87.93
WSC	11217442	87.18	Sikem	15094707	87.45
Westa	147698823	92.65	Mitos	96942661	94.03
Laura	38927599	-	Lemtal	38363802	-

Tab. 4: Statistics of transcriptome assembly and associated steps.

Step	*F. pratensis*	*L. multiflorum*
Transcripts in assembly	326457	410930
Clustered transcripts	262929	265643
Predicted ORFs	108225	84837
Annotated transcripts	76517	54978

Out of 70886 gene models in diploid assembly of *L. multiflorum* genome, 1616 were left after BLAST search against *F. pratensis* reference and following further refining with search against Swiss-Prot 1572 species-specific genes were identified in Set II.

4.3 Characterisation of orphan genes

Sequence features: GC content, ORF type and protein length for orphan genes of Set I are summarised in Tab. 5. Mean GC content for *F. pratensis* orphan genes is 54.43% in comparison to 53.55% for the rest of genes, while in *L. multiflorum* average of orphan genes is 53.59% and of remaining genes is 56.07%. Average protein length is 136 and 121 amino acids for orphan genes and 959 and 871 for non-orphan genes.

None of sequences in Set I were annotated in assembled transcriptomes; domain predictions yielded limited results. In *F. pratensis*, TRINITY_DN20199_c0_g1 and TRINITY_DN55477_c0_g1 were predicted to contain membrane-embedded regions, while the remaining two to be disordered.

Tab. 5: Sequence features of species-specific genes in Set I.

Sequence name	**GC content**	**ORF type**	**Protein length (aa)**
Festuca pratensis			
TRINITY_DN20199_c0_g1	63.91%	complete	109
TRINITY_DN23520_c0_g1	51.35%	internal	111
TRINITY_DN43871_c0_g4	59.10%	3’ partial	141
TRINITY_DN55477_c0_g1	43.35%	complete	183
Lolium multiflorum			
TRINITY_DN12075_c0_g1	52.55%	3’ partial	170
TRINITY_DN134651_c0_g1	48.11%	3’ partial	106
TRINITY_DN147915_c0_g1	61.64%	internal	106
TRINITY_DN37942_c0_g1	83.33%	internal	142
TRINITY_DN67358_c0_g1*	37.95%	complete	101

* Two isoforms of this gene were identified as orphan.

Additionally, TRINITY_DN20199_c0_g1 had few hits with predicted proteins of unknown function from goatgrass and wheat in NCBI nr and TrEMBL databases; multiple hits with TrEMBL proteins in goatgrass, wild rice and several other plant species were found for TRINITY_DN23520_c0_g1, but most significantly with nucleolin proteins of *A. thaliana* (Q9FVQ1) and rice (Q6Z1C0, Q7XTT4). In *L. multiflorum*, TRINITY_DN134651_c0_g1 and TRINITY_DN37942_c0_g1 were predicted as disordered. Furthermore, the latter had multiple hits in TrEMBL database, most notably with translation initiation factors in several species of bacteria (particularly with reviewed proteins in *Frankia alni* (Q0RDS4) and *Arthrobacter sp.* (A0JUU0)) and artherin in rabbit (Q6SPE9).

Average GC content for Set II is 54.65% and average gene length is 660 nucleotides, whereas corresponding values for the non-orphan genes are 54.33% and 1587 nucleotides. Analysis with InterProScan predicted 1401 (89% of Set II) as proteins with disorder and 71 with coil domain.

Tab. 6: Gene ontology terms found for Set II with their frequency. Ontology types are molecular function (MF), biological process (BP) and cellular component (CC). Note that one sequence could match multiple terms.

Accession	Ontology	Name	Count
GO:0005515	MF	protein binding	15
GO:0004857	MF	enzyme inhibitor activity	3
GO:0005524	MF	ATP binding	2
GO:0008270	MF	zinc ion binding	2
GO:0030598	MF	rRNA N-glycosylase activity	2
GO:0003676	MF	nucleic acid binding	1
GO:0003677	MF	DNA binding	1
GO:0004185	MF	serine-type carboxypeptidase activity	1
GO:0008234	MF	cysteine-type peptidase activity	1
GO:0019829	MF	ATPase-coupled monoatomic cation transmembrane transporter activity	1
GO:0043531	MF	ADP binding	1
GO:0140358	MF	P-type transmembrane transporter activity	1
GO:0006952	BP	defense response	2
GO:0017148	BP	negative regulation of translation	2
GO:0006355	BP	regulation of DNA-templated transcription	1
GO:0006508	BP	proteolysis	1
GO:0006511	BP	ubiquitin-dependent protein catabolic process	1
GO:0006614	BP	SRP-dependent cotranslational protein targeting to membrane	1
GO:0006812	BP	monoatomic cation transport	1
GO:0006886	BP	intracellular protein transport	1
GO:0007166	BP	cell surface receptor signaling pathway	1
GO:0005758	CC	mitochondrial intermembrane space	1
GO:0016021	CC	membrane	1

Assignment of gene ontology terms resulted in mostly singular hits, with notable exceptions of protein binding, enzyme inhibitor activity, ATP binding, zinc ion binding, rRNA N-glycosylase activity, defense response and negative regulation of translation (Tab. 6).

Similarly, mainly singular results were found in domain and protein family predictions – those with at least two matches are presented in Tab. 7. Domain of unknown function and protein of unknown function were most frequent annotations (35 hits across 10 different entries), followed by F-box-like domain (14 hits across 3 entries). 44 entries with singular matches are not presented here. Overall, 163 genes had at least one domain prediction (excluding disorder and coil predictions).

Tab. 7: Matched InterPro entries (domain and protein family predictions) for Set II with at least two matches. Note that one sequence could match multiple entries. Several specific entries for domains and proteins of unknown function, F-box domain and Zinc finger were grouped together and unique gene count across all related entries is presented.

Entry	Name	Count
multiple	Domain of unknown function	34
multiple	F-box-like domain	11
IPR009003	Peptidase S1, PA clan	8
IPR032675	Leucine-rich repeat domain superfamily	7
multiple	Zinc finger (several types)	5
IPR038765	Papain-like cysteine peptidase superfamily	4
IPR004314	Neprosin	3
IPR006501	Pectinesterase inhibitor domain	3
IPR011009	Protein kinase-like domain superfamily	3
IPR016024	Armadillo-type fold	3
IPR035513	Invertase/pectin methylesterase inhibitor domain superfamily	3
IPR011528	Nuclease-related domain, NERD	2
IPR011989	Armadillo-like helical	2
IPR011990	Tetratricopeptide-like helical domain superfamily	2
IPR012340	Nucleic acid-binding, OB-fold	2
IPR015915	Kelch-type beta propeller	2
IPR026961	PGG domain	2
IPR027417	P-loop containing nucleoside triphosphate hydrolase	2
IPR034088	Pla a 1-like	2
IPR036041	Ribosome-inactivating protein superfamily	2
IPR041118	Rx, N-terminal	2
IPR044974	Disease resistance protein, plants	2

4.4 Expression analysis for Set I

Mapping reads to transcriptomes proved them to be representative, with averages of around 80% for parent samples, Lm × Fp F1, Fp × Lm F1 and Fp × Lm F2 and 60-70% for the rest of them. In most cases Set I genes were eliminated as statistically insignificant in first stage of analysis; in fact, only two genes remained to be compared across species. Those were TRINITY_DN23520_c0_g1 in *F. pratensis* and TRINITY_DN12075_c0_g1 in *L. multiflorum*. The former had higher expression in parents when comparing them with Fp × Lm F1 and F1 aging, and comparable across parents coupled with F1 stress, F2 aging or F2 stress. The latter had expression of similar level in Fp × Lm F1 and F1 aging.

4.5 Expression analysis for Set II

Reads of *Lolium* parent were mapped to genome with 72% success rate, hybrids of F_1 generation with 79% (collectively), of F_2 generation – 87% and the remaining samples with 57-68% rate. Differential analysis was performed for parent and F_1 generation. Again, the majority of candidates were filtered out in the first step; results for those with significant expression are presented in Tab. 8.

12% (188 and 190) of identified genes were expressed in either dataset, of those with difference in expression a few (3 and 6) were expressed more in hybrids, but mostly the parental expression was greater (36 and 33). Genes shared between hybrids from reciprocal crosses included 10 genes with higher expression in parents (compared to hybrids).

Filtering by expression in individual samples revealed a set of 47 genes expressed universally across all samples, whereas 352 genes were detected in at least one sample. Number of *L. multiflorum*-specific genes expressed in each sample type is listed in Tab. 9. F2 aging was most orphan-rich, with 212 (13.5% of Set II) genes with statistically significant expression, followed by 194 (12.3%) of Fp × Lm F1 parent (one of 'Mitos' plants) and 189 (12.0%) of F2 stress. At the other end were Lm × Fp F1 parent (126, 8%) and Fp × Lm F2 (146, 9.3%) sets.

Tab. 8: Differential gene expression analysis of orphan genes for parent (*L. multiflorum*) and F_1 generation (both crosses).

Orphan genes	**Fp × Lm F1**	**Lm × Fp F1**	**Shared**
expressed:	188	190	82
higher expression in hybrids	3	6	0
higher expression in parents	36	33	10

Tab. 9: Number of *L. multiflorum*-specific genes out of Set II expressed in samples.

Sample type	Count	Sample type	Count
Parent of Fp × Lm F1	194	F1 aging	165
Fp × Lm F1	153	F1 stress	177
Parent of Lm × Fp F1	126	F2 aging	212
Lm × Fp F1	160	F2 stress	189
Fp × Lm F2	146		

Number of species-specific genes expressed in sample has partly influenced the fraction of shared genes (Fig. 5) – most shared genes were found for samples with noticeable difference in LSG count (e.g. comparison of parental samples, F2 aging and Fp × Lm F2), although not in all cases (F2 aging and F2 stress both have relatively high counts, parent of Fp × Lm F1 cross has lower percentage of shared genes with the largest set (F2 aging) than with the other parent sample).

The direction of crosses separates samples noticeably – Lm × Fp F1 and its parent generally share least Set II genes with most other samples (originating from the Fp × Lm cross). Expression patterns seem to be similar in given generation (parent, F_1 or F_2) and more related to age than conditions ('aging' and 'stress' samples were taken from 4-year old plants, in contrast to 3-week old crosses Lm × Fp F1, Fp × Lm F1 and F2).

Dataset	Count	Par. of Fp × LmF1	Fp × LmF1	Par. of Lm × Fp F1	Lm × Fp F1	F1 aging	F1 stress	F2 aging	F2 stress	Fp × LmF2
Par. of Fp × Lm F1	194	x	55%	61%	58%	56%	61%	68%	62%	56%
Fp × Lm F1	153	70%	x	54%	58%	80%	79%	81%	75%	65%
Par. of Lm × Fp F1	126	94%	65%	x	76%	66%	74%	74%	70%	61%
Lm × Fp F1	160	70%	55%	60%	x	53%	57%	61%	58%	52%
F1 aging	165	66%	74%	50%	52%	x	83%	82%	78%	66%
F1 stress	177	67%	68%	53%	51%	77%	x	82%	80%	64%
F2 aging	212	62%	58%	44%	46%	64%	68%	x	83%	64%
F2 stress	189	64%	61%	47%	49%	68%	75%	93%	x	68%
Fp × Lm F2	146	74%	68%	53%	57%	75%	77%	93%	88%	x

Fig. 5: Expressed species-specific genes of Set II shared between samples. Percentages were calculated in reference to count at the start of respective row. Par. = Parent.

5 DISCUSSION

Identification of species-specific genes is still somewhat a subjective matter as the choice of filtering criteria influences the results greatly. It is evident that smaller size of Set I can be attributed to definitely more strict criteria and more filtering steps. Consequently, it is quite likely that some of dismissed candidates were false positives and could qualify as species-specific genes if different conditions were used. Therefore, it may be reasonable to treat genes identified in Set I as high-confidence species-specific genes. However, even though the approach was fairly conservative, classification of two genes (namely TRINITY_DN23520_c0_g1 in *F. pratensis* and TRINITY_DN37942_c0_g1 in *L. multiflorum*) as species-specific could be questioned, given that they scored significant BLAST hits, meaning they are probably homologous with other proteins. While in latter case the issue is less relevant, seeing that the gene was not significantly expressed in examined samples, in former it could imply that none 'true' *F. pratensis*-specific genes were expressed. On side note, it is quite surprising that matches with Swiss-Prot proteins were not found in annotated transcriptomes assembled in earlier part of the work; this could potentially be result of differences between command line-based BLAST+ used for transcriptome annotation and BLAST employed on UniProt web server, which was used in manual inspection of Set I. As for the third sequence with potential homology matches, TRINITY_DN20199_c0_g1 in *F. pratensis*, low number and especially automated origin of these hits prompts to disregard them for the time being; though considering changeful nature of used databases, it could be valuable to re-examine them after some time.

In contrast to that, Set II was selected using partly relaxed tests and did not undergo manual inspection; hence it may contain some false positives. Despite that, Set II constitutes only ~2% of all candidate sequences (admittedly the assembly is diploid, signifying that the actual gene number is lower, but then so may be the number of species-specific genes) and falls into lower end of typical LSG content in plants (Arendsee et al., 2014; Jiang et al., 2022); this does not necessarily mean that the value is irregular as several studies have identified similar or even lower numbers of LSG (e.g. (Lin et al., 2014; Zile et al., 2020; Jin et al., 2021)). Refinement of filtering criteria and growing number of known proteins could be responsible for this trend. Finally, it is important to note that while Set I is based on expression data (at the transcript level), Set II was identified from gene models. Of course, expression analysis revealed that both sets contained many genes without statistically significant expression; for Set I this can be at least partially explained by usage of samples not included in analysis for transcriptome

assembly, but no such assumptions can be made for Set II. This further supports possibility of false positives' presence in identified set; additional expression verification would be needed to decipher this issue. Overall, sizes of Set I and Set II determine the probable range for exact number of LSGs in *L. mutliflorum*, which is bound to change over time. Although *F. pratensis* was not analysed in second set, similar size of Set I and close relationship between species could justify similar prediction.

Characterisation of identified genes proved that they exhibit typical properties of LSGs. Shorter protein length is consistent with previous results (as discussed in section 2.4.1) and annotations of several ORFs as partial or internal in Set I. Scarce domain predictions were also expected (see section 2.4.4). Disorder predictions have been reported for orphan and especially *de novo* originated genes in several studies (Mukherjee et al., 2015; Heames et al., 2020; Jin et al., 2021), thus numerous annotations of this type in genes reported here could point to their novel origin. However, it has been also noted that disorder levels are predicted higher for higher GC content of both genes and whole genomes (Basile et al., 2017; Casola, 2018; Vakirlis et al., 2018), which has been detected here; therefore comparison of disorder rate with non-orphan proteins and detailed examination of other origin indicators would be required to evaluate this observation in proper context.

Characterisation of Set I was mostly unsuccessful, probably due to the strict filtering criteria to identify species-specific genes. The two genes annotated by possible homology had gene expression-related functions: nucleolin-like proteins play important part in ribosome assembly (Pontvianne et al., 2007), the role of translation initiation factors is rather self-explanatory and artherin is thought to repress transcription (UniProt entry Q6SPE9 (Bateman et al., 2023)). In Set II, many general structural features were found among domain predictions. Amid them, both protein (e.g. GO:0005515, IPR036047, IPR032675, IPR011990) and nucleic acid binding (e.g. GO:0003676, GO:0003677, IPR012340, IPR027417) were represented. More specific annotations included those related to nuclease and DNA processing activity (IPR011528), peptidase activity (GO:0004185, GO:0008234, IPR009003, IPR038765, IPR004314), ubiquitination (GO:0006511, IPR036047) and plant defense response (GO:0006952, IPR044974, IPR036041, IPR041118). While those inferences seem promising, experimental verification would be needed to determine the accuracy of these predictions.

In both sets most identified genes had statistically insignificant expression and were eliminated at the filtering stage of expression analysis – even in parental samples. This could

be explained in several ways: (i) discarded genes do not exhibit stable expression and cannot be regarded as species-specific, instead their presence might be result of 'pervasive transcription' or annotation artefact; (ii) they exhibit low expression levels typical for young genes and, related to that, (iii) not enough tissues were sampled and thus, given wide-spread pattern of tissue-specific expression reported for orphan genes (see section 2.4.3), their expression might had been missed. The first reason is especially likely for Set II (as mentioned above); nevertheless in both sets different explanations could apply for individual cases and further evidence, preferably at translational level, would be required to discern between them.

Majority of comparisons between parental and hybrid samples revealed comparable levels of expression. All generations were tetraploid, so effects of WGD were not expected to be observed and given that absence of homoeologs was properly ensured, this could indicate no significant interactions between parental genomes in hybrids. On the other hand, cases of higher expression in parental samples may signify that in some instances there could be conflict in regulation or perhaps epistatic interactions with other genes.

As for shared expression patterns examined in Set II, parental lineage was the most dividing factor, while same generation and age increased number of shared expressed genes. In both hybrid generations, number of expressed genes was higher for older plants, perhaps confirming the observations that many orphan genes become expressed after sexual maturation (Domazet-Loso & Tautz, 2003; Guo et al., 2007). The influence of stress conditions was ambiguous judging only by amounts – in F_1 generation more genes were expressed under stress (compared to 'aging' samples), but the opposite was true for F_2 generation. It is possible that changes were more noticeable in expression level, which was not examined here; alternatively biotic stress could be tested, given that defense response annotations reported for few genes are related to that type of stress and lastly, low number of those annotations could mean that this SSGs set is not particularly rich in stress-related genes.

6 CONCLUSION

To conclude, using approach based on the transcriptome data from two parental species, I was able to identify only a few orphan genes (4 and 5) for *F. pratensis* and *L. multiflorum*, respectively. For both species, most of orphan genes were not expressed in examined samples and expression in parental plants and hybrids did not differ. Using alternative approach employing *L. multiflorum* genome data and relaxed criteria, additional 1572 species-specific genes were identified. The expression levels in parents and hybrids were mostly similar, with small fraction of genes down-regulated in hybrids. Indications for function in regulation of gene expression and defense response were found in several cases, but the role of the majority of orphan genes remains unknown. This pioneer study identified the first orphan genes in the two grass species and provided the first insight into their transcription in parental species as well as in their hybrids. Development and annotation of reference genomes and transcriptomes undergoing within the Pangenome consortium will provide an unprecedented opportunity to identify and validate new orphan genes with emphasis on their evolutionary role in hybrid evolution and potential speciation.

7 REFERENCES

Abrusán, G. (2013). Integration of New Genes into Cellular Networks, and Their Structural Maturation. *Genetics*, 195(4), 1407-1417. https://doi.org/10.1534/genetics.113.152256

Alger, E., & Edger, P. (2020). One subgenome to rule them all: underlying mechanisms of subgenome dominance. *Current Opinion in Plant Biology*, 54, 108-113. https://doi.org/10.1016/j.pbi.2020.03.004

Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., & Shiroishi, T. (2009). Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription. *Developmental Cell*, 16(1), 47-57. https://doi.org/10.1016/j.devcel.2008.11.011

Andersson, R., Sandelin, A., & Danko, C. (2015). A unified architecture of transcriptional regulatory elements. *Trends in Genetics*, 31(8), 426-433. https://doi.org/10.1016/j.tig.2015.05.007

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (20.03.2023).

Arendsee, Z., Li, L., & Wurtele, E. (2014). Coming of age: orphan genes in plants. *Trends in Plant Science*, 19(11), 698-708. https://doi.org/10.1016/j.tplants.2014.07.003

Arnold, B., Lahner, B., DaCosta, J., Weisman, C., Hollister, J., Salt, D., Bomblies, K., & Yant, L. (2016). Borrowed alleles and convergence in serpentine adaptation. *Proceedings of the National Academy of Sciences*, 113(29), 8320-8325. https://doi.org/10.1073/pnas.1600405113

Augustus: gene prediction. https://bioinf.uni-greifswald.de/augustus/ (20.03.2023).

Bannister, A., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, 21(3), 381-395. https://doi.org/10.1038/cr.2011.22

Bardil, A., de Almeida, J., Combes, M., Lashermes, P., & Bertrand, B. (2011). Genomic expression dominance in the natural allopolyploid *Coffea arabica* is massively affected by growth temperature. *New Phytologist*, 192(3), 760-774. https://doi.org/10.1111/j.1469-8137.2011.03833.x

Basile, W., Sachenkova, O., Light, S., Elofsson, A., & Dunbrack, R. (2017). High GC content causes orphan proteins to be intrinsically disordered. *PLOS Computational Biology*, 13(3), e1005375. https://doi.org/10.1371/journal.pcbi.1005375

Bateman, A., Martin, M., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E., Britto, R., Bye-A-Jee, H., Cukura, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Garmiri, P., da Costa Gonzales, L., Hatton-Ellis, E., Hussein, A., Ignatchenko, A. et al. (2023). UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research, 51(1), D523-D531. https://doi.org/10.1093/nar/gkac1052

Batut, P., Bing, X., Sisco, Z., Raimundo, J., Levo, M., & Levine, M. (2022). Genome organization controls transcriptional dynamics during development. *Science*, 375(6580), 566-570. https://doi.org/10.1126/Science.abi7178

Begun, D., Lindfors, H., Thompson, M., & Holloway, A. (2006). Recently Evolved Genes Identified From Drosophila yakuba and D. erecta Accessory Gland Expressed Sequence Tags. *Genetics*, 172(3), 1675-1681. https://doi.org/10.1534/genetics.105.050336

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berkovits, B., & Mayr, C. (2015). Alternative 3′ UTRs act as scaffolds to regulate membrane protein localization. *Nature*, 522(7556), 363-367. https://doi.org/10.1038/Nature14321

Betrán, E., Wang, W., Jin, L., & Long, M. (2002). Evolution of the Phosphoglycerate mutase Processed Gene in Human and Chimpanzee Revealing the Origin of a New Primate Gene. *Molecular Biology and Evolution*, 19(5), 654-663. https://doi.org/10.1093/oxfordjournals.molbev.a004124

Blevins, T., Wang, J., Pflieger, D., Pontvianne, F., & Pikaard, C. (2017). Hybrid incompatibility caused by an epiallele. *Proceedings of the National Academy of Sciences*, 114(14), 3702-3707. https://doi.org/10.1073/pnas.1700368114

Boatwright, J., Yeh, C., Hu, H., Susanna, A., Soltis, D., Soltis, P., Schnable, P., & Barbazuk, W. (2021). Trajectories of Homoeolog-Specific Expression in Allotetraploid *Tragopogon castellanus* Populations of Independent Origins. *Frontiers in Plant Science*, 12, 679047. https://doi.org/10.3389/fpls.2021.679047

Bolger, A., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114-2120. https://doi.org/10.1093/bioinformatics/btu170

Bologna, N., Iselin, R., Abriata, L., Sarazin, A., Pumplin, N., Jay, F., Grentzinger, T., Dal Peraro, M., & Voinnet, O. (2018). Nucleo-cytosolic Shuttling of ARGONAUTE1 Prompts a Revised Model of the Plant MicroRNA Pathway. *Molecular Cell*, 69(4), 709-719.e5. https://doi.org/10.1016/j.molcel.2018.01.007

Bottani, S., Zabet, N., Wendel, J., & Veitia, R. (2018). Gene Expression Dominance in Allopolyploids: Hypotheses and Models. *Trends in Plant Science*, 23(5), 393-402. https://doi.org/10.1016/j.tplants.2018.01.002

Bourque, G., Burns, K., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H., Macfarlan, T., Mager, D., & Feschotte, C. (2018). Ten things you should know about transposable elements. *Genome Biology*, 19(1), 199. https://doi.org/10.1186/s13059-018-1577-z

Brasset, E., & Vaury, C. (2005). Insulators are fundamental components of the eukaryotic genomes. *Heredity*, 94(6), 571-576. https://doi.org/10.1038/sj.hdy.6800669

Brogna, S., & Wen, J. (2009). Nonsense-mediated mRNA decay (NMD) mechanisms. *Nature Structural & Molecular Biology*, 16(2), 107-113. https://doi.org/10.1038/nsmb.1550

Brosch, M., Saunders, G., Frankish, A., Collins, M., Yu, L., Wright, J., Verstraten, R., Adams, D., Harrow, J., Choudhary, J., & Hubbard, T. (2011). Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. *Genome Research*, 21(5), 756-767. https://doi.org/10.1101/gr.114272.110

Bryant, D., Johnson, K., DiTommaso, T., Tickle, T., Couger, M., Payzin-Dogru, D., Lee, T., Leigh, N., Kuo, T., Davis, F., Bateman, J., Bryant, S., Guzikowski, A., Tsai, S., Coyne, S., Ye, W., Freeman, R., Peshkin, L., Tabin, C. et al. (2017). A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. *Cell Reports*, 18(3), 762-776. https://doi.org/10.1016/j.celrep.2016.12.063

Bundock, P., & Hooykaas, P. (2005). An *Arabidopsis* hAT-like transposase is essential for plant development. *Nature*, 436(7048), 282-284. https://doi.org/10.1038/Nature03667

Burrows-Wheeler Aligner. https://bio-bwa.sourceforge.net/ (20.03.2023).

Burton, R., Pereira, R., & Barreto, F. (2013). Cytonuclear Genomic Interactions and Hybrid Breakdown. *Annual Review of Ecology, Evolution, and Systematics*, 44(1), 281-302. https://doi.org/10.1146/annurev-ecolsys-110512-135758

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1), 421. https://doi.org/10.1186/1471-2105-10-421

Cardoso-Silva, C., Aono, A., Mancini, M., Sforça, D., da Silva, C., Pinto, L., Adams, K., & de Souza, A. (2022). Taxonomically Restricted Genes Are Associated With Responses to Biotic and Abiotic Stresses in Sugarcane (*Saccharum* spp.). *Frontiers in Plant Science*, 13, 923069. https://doi.org/10.3389/fpls.2022.923069

Carthew, R., & Sontheimer, E. (2009). Origins and Mechanisms of miRNAs and siRNAs. *Cell*, 136(4), 642-655. https://doi.org/10.1016/j.cell.2009.01.035

Carvunis, A., Rolland, T., Wapinski, I., Calderwood, M., Yildirim, M., Simonis, N., Charloteaux, B., Hidalgo, C., Barbette, J., Santhanam, B., Brar, G., Weissman, J., Regev, A., Thierry-Mieg, N., Cusick, M., & Vidal, M. (2012). Proto-genes and de novo gene birth. *Nature*, 487(7407), 370-374. https://doi.org/10.1038/Nature11184

Casola, C. (2018). From de novo to 'de nono': The majority of novel protein coding genes identified with phylostratigraphy are old genes or recent duplicates. *Genome Biology and Evolution*, 10(11), 2906-2918. https://doi.org/10.1093/gbe/evy231

Colle, M., Leisner, C., Wai, C., Ou, S., Bird, K., Wang, J., Wisecaver, J., Yocca, A., Alger, E., Tang, H., Xiong, Z., Callow, P., Ben-Zvi, G., Brodt, A., Baruch, K., Swale, T., Shiue, L., Song, G., Childs, K. et al. (2019). Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry. *GigaScience*, 8(3), giz012. https://doi.org/10.1093/gigaScience/giz012

Combes, M., Hueber, Y., Dereeper, A., Rialle, S., Herrera, J., & Lashermes, P. (2015). Regulatory Divergence between Parental Alleles Determines Gene Expression Patterns in Hybrids. *Genome Biology and Evolution*, 7(4), 1110-1121. https://doi.org/10.1093/gbe/evv057

Cox, M., Dong, T., Shen, G., Dalvi, Y., Scott, D., Ganley, A., & Stajich, J. (2014). An Interspecific Fungal Hybrid Reveals Cross-Kingdom Rules for Allopolyploid Gene Expression Patterns. *PLoS Genetics*, 10(3), e1004180. https://doi.org/10.1371/journal.pgen.1004180

Czaban, A., Sharma, S., Byrne, S., Spannagl, M., Mayer, K., & Asp, T. (2015). Comparative transcriptome analysis within the *Lolium/Festuca* species complex reveals high sequence conservation. *BMC Genomics*, 16(1), 249. https://doi.org/10.1186/s12864-015-1447-y

Danecek, P., Bonfield, J., Liddle, J., Marshall, J., Ohan, V., Pollard, M., Whitwham, A., Keane, T., McCarthy, S., Davies, R., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2), giab008. https://doi.org/10.1093/gigaScience/giab008

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C., Bofkin, L., Jones, T., Davis, R., & Steinmetz, L. (2006). A high-resolution map of transcription in the yeast genome. *Proceedings of the National Academy of Sciences*, 103(14), 5320-5325. https://doi.org/10.1073/pnas.0601091103

de Felippes, F., Waterhouse, P., & Marquardt, S. (2023). Plant terminators: the unsung heroes of gene expression. *Journal of Experimental Botany*, 74(7), 2239-2250. https://doi.org/10.1093/jxb/erac467

De Smet, R., Adams, K., Vandepoele, K., Van Montagu, M., Maere, S., & Van de Peer, Y. (2013). Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences*, 110(8), 2898-2903. https://doi.org/10.1073/pnas.1300127110

Dever, T., Dinman, J., & Green, R. (2018). Translation Elongation and Recoding in Eukaryotes. *Cold Spring Harbor Perspectives in Biology*, 10(8), a032649. https://doi.org/10.1101/cshperspect.a032649

Devos, K., Brown, J., & Bennetzen, J. (2002). Genome Size Reduction through Illegitimate Recombination Counteracts Genome Expansion in *Arabidopsis*. *Genome Research*, 12(7), 1075-1079. https://doi.org/10.1101/gr.132102

Ding, M., & Chen, Z. (2018). Epigenetic perspectives on the evolution and domestication of polyploid plant and crops. *Current Opinion in Plant Biology*, 42, 37-48. https://doi.org/10.1016/j.pbi.2018.02.003

Dodsworth, S., Chase, M., & Leitch, A. (2016). Is post-polyploidization diploidization the key to the evolutionary success of angiosperms?. *Botanical Journal of the Linnean Society*, 180(1), 1-5. https://doi.org/10.1111/boj.12357

Domazet-Loso, T., & Tautz, D. (2003). An Evolutionary Analysis of Orphan Genes in Drosophila. *Genome Research*, 13(10), 2213-2219. https://doi.org/10.1101/gr.1311003

Donoghue, M., Keshavaiah, C., Swamidatta, S., & Spillane, C. (2011). Evolutionary origins of *Brassicaceae* specific genes in *Arabidopsis thaliana*. *BMC Evolutionary Biology*, 11(1). https://doi.org/10.1186/1471-2148-11-47

Douglas, G., Gos, G., Steige, K., Salcedo, A., Holm, K., Josephs, E., Arunkumar, R., Ågren, J., Hazzouri, K., Wang, W., Platts, A., Williamson, R., Neuffer, B., Lascoux, M., Slotte, T., & Wright, S. (2015). Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid *Capsella bursa-pastoris*. *Proceedings of the National Academy of Sciences*, 112(9), 2806-2811. https://doi.org/10.1073/pnas.1412277112

Duarte, J., Wall, P., Edger, P., Landherr, L., Ma, H., Pires, P., Leebens-Mack, J., & dePamphilis, C. (2010). Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evolutionary Biology*, 10(1). https://doi.org/10.1186/1471-2148-10-61

Duchaine, T., & Fabian, M. (2019). Mechanistic Insights into MicroRNA-Mediated Gene Silencing. *Cold Spring Harbor Perspectives in Biology*, 11(3), a032771. https://doi.org/10.1101/cshperspect.a032771

Dujon, B. (1996). The yeast genome project: what did we learn?. *Trends in Genetics*, 12(7), 263-270. https://doi.org/10.1016/0168-9525(96)10027-5

Edger, P., Smith, R., McKain, M., Cooley, A., Vallejo-Marin, M., Yuan, Y., Bewick, A., Ji, L., Platts, A., Bowman, M., Childs, K., Washburn, J., Schmitz, R., Smith, G., Pires, J., & Puzey, J. (2017). Subgenome Dominance in an Interspecific Hybrid, Synthetic Allopolyploid, and a 140-Year-Old Naturally Established Neo-Allopolyploid Monkeyflower. *The Plant Cell*, 29(9), 2150-2167. https://doi.org/10.1105/tpc.17.00010

Ekman, D., & Elofsson, A. (2010). Identifying and Quantifying Orphan Protein Sequences in Fungi. *Journal of Molecular Biology*, 396(2), 396-405. https://doi.org/10.1016/j.jmb.2009.11.053

Elliott, T., & Gregory, T. (2015). What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678). https://doi.org/10.1098/rstb.2014.0331

Flagel, L., & Wendel, J. (2010). Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. *New Phytologist*, 186(1), 184-193. https://doi.org/10.1111/j.1469-8137.2009.03107.x

Flagel, L., Udall, J., Nettleton, D., & Wendel, J. (2008). Duplicate gene expression in allopolyploid *Gossypium* reveals two temporally distinct phases of expression evolution. *BMC Biology*, 6(1), 16. https://doi.org/10.1186/1741-7007-6-16

Freeling, M., Scanlon, M., & Fowler, J. (2015). Fractionation and subfunctionalization following genome duplications: mechanisms that drive gene content and their consequences. *Current Opinion in Genetics & Development*, 35, 110-118. https://doi.org/10.1016/j.gde.2015.11.002

Freeling, M., & Thomas, B. (2006). Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Research,* 16(7), 805-814. https://doi.org/10.1101/gr.3681406

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23), 3150-3152. https://doi.org/10.1093/bioinformatics/bts565

Gaeta, R., Pires, J., Iniguez-Luy, F., Leon, E., & Osborn, T. (2007). Genomic Changes in Resynthesized *Brassica napus* and Their Effect on Gene Expression and Phenotype. *The Plant Cell*, 19(11), 3403-3417. https://doi.org/10.1105/tpc.107.054346

Garneau, N., Wilusz, J., & Wilusz, C. (2007). The highways and byways of mRNA decay. *Nature Reviews Molecular Cell Biology*, 8(2), 113-126. https://doi.org/10.1038/nrm2104

Garrido-Ramos, M. (2017). Satellite DNA: An Evolving Topic. *Genes*, 8(9), 230. https://doi.org/10.3390/genes8090230

Gerstein, M., Bruce, C., Rozowsky, J., Zheng, D., Du, J., Korbel, J., Emanuelsson, O., Zhang, Z., Weissman, S., & Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Research*, 17(6), 669-681. https://doi.org/10.1101/gr.6339607

Gibson, A., Smith, Z., Fuqua, C., Clay, K., & Colbourne, J. (2013). Why so many unknown genes? Partitioning orphans from a representative transcriptome of the lone star tick *Amblyomma americanum*. *BMC Genomics*, 14(1). https://doi.org/10.1186/1471-2164-14-135

Glombik, M., Copetti, D., Bartos, J., Stoces, S., Zwierzykowski, Z., Ruttink, T., Wendel, J., Duchoslav, M., Dolezel, J., Studer, B., & Kopecky, D. (2021). Reciprocal allopolyploid grasses (*Festuca* × *Lolium*) display stable patterns of genome dominance. *The Plant Journal*, 107(4), 1166-1182. https://doi.org/10.1111/tpj.15375

Grabherr, M., Haas, B., Yassour, M., Levin, J., Thompson, D., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B., Nusbaum, C., Lindblad-Toh, K. et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7), 644-652. https://doi.org/10.1038/nbt.1883

Grabundzija, I., Messing, S., Thomas, J., Cosby, R., Bilic, I., Miskey, C., Gogol-Döring, A., Kapitonov, V., Diem, T., Dalda, A., Jurka, J., Pritham, E., Dyda, F., Izsvák, Z., & Ivics, Z. (2016). A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. *Nature Communications*, 7(1), 10716. https://doi.org/10.1038/ncomms10716

Gregory, T. (2005). The C-value Enigma in Plants and Animals: A Review of Parallels and an Appeal for Partnership. *Annals of Botany*, 95(1), 133-146. https://doi.org/10.1093/aob/mci009

Gregory, T. (2001). Coincidence, coevolution, or causation? DNA content, cellsize, and the C-value enigma. *Biological Reviews*, 76(1), 65-101. https://doi.org/10.1111/j.1469-185X.2000.tb00059.x

Gregory, T. (2023). Animal Genome Size Database. http://www.genomesize.com (02.02.2023).

Greilhuber, J., Doležel, J., Lysák, M., & Bennett, M. (2005). The Origin, Evolution and Proposed Stabilization of the Terms 'Genome Size' and 'C-Value' to Describe Nuclear DNA Contents. *Annals of Botany*, 95(1), 255-260. https://doi.org/10.1093/aob/mci019

Groszmann, M., Greaves, I., Albertyn, Z., Scofield, G., Peacock, W., & Dennis, E. (2011). Changes in 24-nt siRNA levels in *Arabidopsis* hybrids suggest an epigenetic contribution to hybrid vigor.

Proceedings of the National Academy of Sciences, 108(6), 2617-2622. https://doi.org/10.1073/pnas.1019217108
Grotewold, E., Chappell, J., & Kellogg, E. (2015). Plant genes, genomes and genetics. Wiley Blackwell.
Grover, C., Gallagher, J., Szadkowski, E., Yoo, M., Flagel, L., & Wendel, J. (2012). Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytologist*, 196(4), 966-971. https://doi.org/10.1111/j.1469-8137.2012.04365.x
Gundu, S., Tabassum, N., & Blilou, I. (2020). Moving with purpose and direction: transcription factor movement and cell fate determination revisited. *Current Opinion in Plant Biology*, 57, 124-132. https://doi.org/10.1016/j.pbi.2020.08.003
Guo, Q., Liu, Q., A. Smith, N., Liang, G., & Wang, M. (2016). RNA Silencing in Plants: Mechanisms, Technologies and Applications in Horticultural Crops. *Current Genomics*, 17(6), 476-489. https://doi.org/10.2174/1389202917666160520103117
Guo, W., Li, P., Ling, J., & Ye, S. (2007). Significant Comparative Characteristics between Orphan and Nonorphan Genes in the Rice (*Oryza sativa* L.) Genome. *Comparative and Functional Genomics*, 2007, 1-7. https://doi.org/10.1155/2007/21676
Guo, Y. (2013). Gene family evolution in green plants with emphasis on the origination and evolution of *Arabidopsis thaliana* genes. *The Plant Journal*, 73(6), 941-951. https://doi.org/10.1111/tpj.12089
Haas, B. TransDecoder (version 3.0.1). https://github.com/TransDecoder/TransDecoder (20.03.2023).
Ha, M., Lu, J., Tian, L., Ramachandran, V., Kasschau, K., Chapman, E., Carrington, J., Chen, X., Wang, X., & Chen, Z. (2009). Small RNAs serve as a genetic buffer against genomic shock in *Arabidopsis* interspecific hybrids and allopolyploids. *Proceedings of the National Academy of Sciences*, 106(42), 17835-17840. https://doi.org/10.1073/pnas.0907003106
Hartenian, E., & Glaunsinger, B. (2019). Feedback to the central dogma: cytoplasmic mRNA decay and transcription are interdependent processes. *Critical Reviews in Biochemistry and Molecular Biology*, 54(4), 385-398. https://doi.org/10.1080/10409238.2019.1679083
Heames, B., Schmitz, J., & Bornberg-Bauer, E. (2020). A Continuum of Evolving De Novo Genes Drives Protein-Coding Novelty in *Drosophila*. *Journal of Molecular Evolution*, 88(4), 382-398. https://doi.org/10.1007/s00239-020-09939-z
He, G., Zhu, X., Elling, A., Chen, L., Wang, X., Guo, L., Liang, M., He, H., Zhang, H., Chen, F., Qi, Y., Chen, R., & Deng, X. (2010). Global Epigenetic and Transcriptional Trends among Two Rice Subspecies and Their Reciprocal Hybrids. *The Plant Cell*, 22(1), 17-33. https://doi.org/10.1105/tpc.109.072041
Heinen, T., Staubach, F., Häming, D., & Tautz, D. (2009). Emergence of a New Gene from an Intergenic Region. *Current Biology*, 19(18), 1527-1531. https://doi.org/10.1016/j.cub.2009.07.049
Hershey, J., Sonenberg, N., & Mathews, M. (2019). Principles of Translational Control. *Cold Spring Harbor Perspectives in Biology*, 11(9), a032607. https://doi.org/10.1101/cshperspect.a032607
He, X., Chen, T., & Zhu, J. (2011). Regulation and function of DNA methylation in plants and animals. *Cell Research*, 21(3), 442-465. https://doi.org/10.1038/cr.2011.23
HMMER (version 3.3.2): biosequence analysis using profile hidden Markov models. (2020). http://hmmer.org/ (20.03.2020).
Hollister, J., & Gaut, B. (2009). Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Research*, 19(8), 1419-1428. https://doi.org/10.1101/gr.091678.109
Hu, G., & Wendel, J. (2019). Cis – trans controls and regulatory novelty accompanying allopolyploidization. *New Phytologist*, 221(4), 1691-1700. https://doi.org/10.1111/nph.15515
Husnik, F., & McCutcheon, J. (2018). Functional horizontal gene transfer from bacteria to eukaryotes. *Nature Reviews Microbiology*, 16(2), 67-79. https://doi.org/10.1038/nrmicro.2017.137
Cheetham, S., Faulkner, G., & Dinger, M. (2020). Overcoming challenges and dogmas to understand the functions of pseudogenes. *Nature Reviews Genetics*, 21(3), 191-201. https://doi.org/10.1038/s41576-019-0196-1
Chen, F., Smith, E., & Shilatifard, A. (2018). Born to run: control of transcription elongation by RNA polymerase II. *Nature Reviews Molecular Cell Biology*, 19(7), 464-478. https://doi.org/10.1038/s41580-018-0010-5
Chen, C., & Shyu, A. (1995). AU-rich elements: characterization and importance in mRNA degradation. *Trends in Biochemical Science*s, 20(11), 465-470. https://doi.org/10.1016/S0968-0004(00)89102-1

Chen, L., & Hellmann, H. (2013). Plant E3 Ligases: Flexible Enzymes in a Sessile World. *Molecular Plant*, 6(5), 1388-1404. https://doi.org/10.1093/mp/sst005

Chen, Z. (2010). Molecular mechanisms of polyploidy and hybrid vigor. *Trends in Plant Science*, 15(2), 57-71. https://doi.org/10.1016/j.tplants.2009.12.003

Chen, Z., Comai, L., & Pikaard, C. (1998). Gene dosage and stochastic effects determine the severity and direction of uniparental ribosomal RNA gene silencing (nucleolar dominance) in *Arabidopsis* allopolyploids. *Proceedings of the National Academy of Sciences*, 95(25), 14891-14896. https://doi.org/10.1073/pnas.95.25.14891

Chen, Z., & Pikaard, C. (1997). Epigenetic silencing of RNA polymerase I transcription: a role for DNA methylation and histone modification in nucleolar dominance. *Genes & Development*, 11(16), 2124-2136. https://doi.org/10.1101/gad.11.16.2124

Chester, M., Gallagher, J., Symonds, V., Cruz da Silva, A., Mavrodiev, E., Leitch, A., Soltis, P., & Soltis, D. (2012). Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (*Asteraceae*). *Proceedings of the National Academy of Sciences*, 109(4), 1176-1181. https://doi.org/10.1073/pnas.1112041109

Ingolia, N., Brar, G., Stern-Ginossar, N., Harris, M., Talhouarne, G., Jackson, S., Wills, M., & Weissman, J. (2014). Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. *Cell Reports*, 8(5), 1365-1379. https://doi.org/10.1016/j.celrep.2014.07.045

Innan, H., & Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. Nature Reviews Genetics, 11(2), 97-108. https://doi.org/10.1038/nrg2689

InterPro. https://www.ebi.ac.uk/InterPro/ (20.03.2023).

Jackson, D., Veit, B., & Hake, S. (1994). Expression of maize KNOTTED1 related homeobox genes in the shoot apical meristem predicts patterns of morphogenesis in the vegetative shoot. *Development*, 120(2), 405-413. https://doi.org/10.1242/dev.120.2.405

Jangam, D., Feschotte, C., & Betrán, E. (2017). Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. *Trends in Genetics*, 33(11), 817-831. https://doi.org/10.1016/j.tig.2017.07.011

Jenczewski, E., & Alix, K. (2004). From Diploids to Allopolyploids: The Emergence of Efficient Pairing Control Genes in Plants. *Critical Reviews in Plant Sciences,* 23(1), 21-45. https://doi.org/10.1080/07352680490273239

Jiang, M., Li, X., Dong, X., Zu, Y., Zhan, Z., Piao, Z., & Lang, H. (2022). Research Advances and Prospects of Orphan Genes in Plants. *Frontiers in Plant Science*, 13, 947129. https://doi.org/10.3389/fpls.2022.947129

Jiang, M., Zhan, Z., Li, H., Dong, X., Cheng, F., & Piao, Z. (2020). *Brassica rapa* orphan genes largely affect soluble sugar metabolism. *Horticulture Research*, 7(1). https://doi.org/10.1038/s41438-020-00403-z

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M., Wang, B., Campbell, M., Stein, J., Wei, X., Chin, C., Guill, K., Regulski, M., Kumari, S., Olson, A., Gent, J., Schneider, K., Wolfgruber, T., May, M., Springer, N. et al. (2017). Improved maize reference genome with single-molecule technologies. *Nature*, 546(7659), 524-527. https://doi.org/10.1038/Nature22971

Jiao, Y., Wickett, N., Ayyampalayam, S., Chanderbali, A., Landherr, L., Ralph, P., Tomsho, L., Hu, Y., Liang, H., Soltis, P., Soltis, D., Clifton, S., Schlarbaum, S., Schuster, S., Ma, H., Leebens-Mack, J., & dePamphilis, C. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature*, 473(7345), 97-100. https://doi.org/10.1038/Nature09916

Jia, Z., Gao, P., Yin, F., Quilichini, T., Sheng, H., Song, J., Yang, H., Gao, J., Chen, T., Yang, B., Kochian, L., Zou, J., Patterson, N., Yang, Q., Gillmor, C., Datla, R., Li, Q., & Xiang, D. (2022). Asymmetric gene expression in grain development of reciprocal crosses between tetraploid and hexaploid wheats. *Communications Biology*, 5(1), 1412. https://doi.org/10.1038/s42003-022-04374-w

Jin, G., Zhou, Y., Yang, H., Hu, Y., Shi, Y., Li, L., Siddique, A., Liu, C., Zhu, A., Zhang, C., & Li, D. (2021). Genetic innovations: Transposable element recruitment and de novo formation lead to the birth of orphan genes in the rice genome. *Journal of Systematics and Evolution*, 59(2), 341-351. https://doi.org/10.1111/jse.12548

Joly-Lopez, Z., & Bureau, T. (2018). Exaptation of transposable element coding sequences. *Current Opinion in Genetics & Development*, 49, 34-42. https://doi.org/10.1016/j.gde.2018.02.011

Joly-Lopez, Z., Forczek, E., Hoen, D., Juretic, N., Bureau, T., & Bennetzen, J. (2012). A Gene Family Derived from Transposable Elements during Early Angiosperm Evolution Has Reproductive Fitness Benefits in *Arabidopsis thaliana*. *PLoS Genetics*, 8(9), e1002931. https://doi.org/10.1371/journal.pgen.1002931

Kapranov, P., Cheng, J., Dike, S., Nix, D., Duttagupta, R., Willingham, A., Stadler, P., Hertel, J., Hackermüller, J., Hofacker, I., Bell, I., Cheung, E., Drenkow, J., Dumais, E., Patel, S., Helt, G., Ganesh, M., Ghosh, S., Piccolboni, A. et al. (2007). RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription. *Science*, 316(5830), 1484-1488. https://doi.org/10.1126/Science.1138341

Khalturin, K., Anton-Erxleben, F., Sassmann, S., Wittlieb, J., Hemmrich, G., Bosch, T., & Patel, N. (2008). A Novel Gene Family Controls Species-Specific Morphological Traits in Hydra. *PLoS*, 6(11), e278. https://doi.org/10.1371/journal.pbio.0060278

Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R., & Bosch, T. (2009). More than just orphans: are taxonomically-restricted genes important in evolution?. *Trends in Genetics*, 25(9), 404-413. https://doi.org/10.1016/j.tig.2009.07.006

Kovalenko, T., & Patrushev, L. (2018). Pseudogenes as Functionally Significant Elements of the Genome. *Biochemistry (Moscow)*, 83(11), 1332-1349. https://doi.org/10.1134/S0006297918110044

Kraitshtein, Z., Yaakov, B., Khasdan, V., & Kashkush, K. (2010). Genetic and Epigenetic Dynamics of a Retrotransposon After Allopolyploidization of Wheat. *Genetics*, 186(3), 801-812. https://doi.org/10.1534/genetics.110.120790

Kumar, A., Gates, P., Czarkwiani, A., & Brockes, J. (2015). An orphan gene is necessary for preaxial digit formation during salamander limb development. *Nature Communications*, 6(1), 8684. https://doi.org/10.1038/ncomms9684

Kunte, K., Shea, C., Aardema, M., Scriber, J., Juenger, T., Gilbert, L., Kronforst, M., & Moran, N. (2011). Sex Chromosome Mosaicism and Hybrid Speciation among Tiger Swallowtail Butterflies. *PLoS Genetics*, 7(9). https://doi.org/10.1371/journal.pgen.1002274

Kuo, C., & Kissinger, J. (2008). Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites *Plasmodium* and *Theileria*. *BMC Evolutionary Biology*, 8, 108. https://doi.org/10.1186/1471-2148-8-108

Kurata, T., Okada, K., & Wada, T. (2005). Intercellular movement of transcription factors. *Current Opinion in Plant Biology*, 8(6), 600-605. https://doi.org/10.1016/j.pbi.2005.09.005

Kurihara, Y., & Watanabe, Y. (2004). *Arabidopsis* micro-RNA biogenesis through Dicer-like 1 protein functions. *Proceedings of the National Academy of Sciences*, 101(34), 12753-12758. https://doi.org/10.1073/pnas.0403115101

Leitch, I., & Bennett, M. (2004). Genome downsizing in polyploid plants. *Biological Journal of the Linnean Society*, 82(4), 651-663. https://doi.org/10.1111/j.1095-8312.2004.00349.x

Leitch, I., Hanson, L., Lim, K., Kovarik, A., Chase, M., Clarkson, J., & Leitch, A. (2008). The Ups and Downs of Genome Size Evolution in Polyploid Species of *Nicotiana* (*Solanaceae*). *Annals of Botany*, 101(6), 805-814. https://doi.org/10.1093/aob/mcm326

Leitch, I., Johnston, E., Pellicer, J., Hidalgo, O., & Bennett, M. (2019). Plant DNA C-values Database (Release 7.1). https://cvalues.science.kew.org/ (02.02.2023).

Levine, M., Jones, C., Kern, A., Lindfors, H., & Begun, D. (2006). Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proceedings of the National Academy of Sciences*, 103(26), 9935-9939. https://doi.org/10.1073/pnas.0509809103

Levine, M., Cattoglio, C., & Tjian, R. (2014). Looping Back to Leap Forward: Transcription Enters a New Era. *Cell*, 157(1), 13-25. https://doi.org/10.1016/j.cell.2014.02.009

Levy, A., & Feldman, M. (2022). Evolution and origin of bread wheat. *The Plant Cell*, 34(7), 2549-2567. https://doi.org/10.1093/plcell/koac130

Li, A., Liu, D., Wu, J., Zhao, X., Hao, M., Geng, S., Yan, J., Jiang, X., Zhang, L., Wu, J., Yin, L., Zhang, R., Wu, L., Zheng, Y., & Mao, L. (2014). mRNA and Small RNA Transcriptomes Reveal Insights into Dynamic Homoeolog Regulation of Allopolyploid Heterosis in Nascent Hexaploid Wheat. *The Plant Cell*, 26(5), 1878-1900. https://doi.org/10.1105/tpc.114.124388

Li, L., Foster, C., Gan, Q., Nettleton, D., James, M., Myers, A., & Wurtele, E. (2009). Identification of the novel protein QQS as a component of the starch metabolic network in *Arabidopsis* leaves. *The Plant Journal*, 58(3), 485-498. https://doi.org/10.1111/j.1365-313X.2009.03793.x

Li, L., & Wurtele, E. (2015). The QQS orphan gene of *Arabidopsis* modulates carbon and nitrogen allocation in soybean. *Plant Biotechnology Journal*, 13(2), 177-187. https://doi.org/10.1111/pbi.12238

Li, L., Zheng, W., Zhu, Y., Ye, H., Tang, B., Arendsee, Z., Jones, D., Li, R., Ortiz, D., Zhao, X., Du, C., Nettleton, D., Scott, M., Salas-Fernandez, M., Yin, Y., & Wurtele, E. (2015). QQS orphan gene regulates carbon and nitrogen partitioning across species via NF-YC interactions. *Proceedings of the National Academy of Sciences*, 112(47), 14734-14739. https://doi.org/10.1073/pnas.1514670112

Li, M., Wang, R., Wu, X., & Wang, J. (2020). Homoeolog expression bias and expression level dominance (ELD) in four tissues of natural allotetraploid *Brassica napus*. *BMC Genomics*, 21(1), 330. https://doi.org/10.1186/s12864-020-6747-1

Lin, H., Moghe, G., Ouyang, S., Iezzoni, A., Shiu, S., Gu, X., & Buell, C. (2010). Comparative analyses reveal distinct sets of lineage-specific genes within *Arabidopsis thaliana*. *BMC Evolutionary* Biology, 10(1). https://doi.org/10.1186/1471-2148-10-41

Li, N., Xu, C., Zhang, A., Lv, R., Meng, X., Lin, X., Gong, L., Wendel, J., & Liu, B. (2019). DNA methylation repatterning accompanying hybridization, whole genome doubling and homoeolog exchange in nascent segmental rice allotetraploids. *New Phytologist*, 223(2), 979-992. https://doi.org/10.1111/nph.15820

Lin, R., Ding, L., Casola, C., Ripoll, D., Feschotte, C., & Wang, H. (2007). Transposase-Derived Transcription Factors Regulate Light Signaling in *Arabidopsis*. *Science*, 318(5854), 1302-1305. https://doi.org/10.1126/Science.1146281

Lin, W., Cai, B., & Cheng, Z. (2014). Identification and characterization of lineage-specific genes in *Populus trichocarpa*. *Plant Cell, Tissue and Organ Culture (PCTOC)*, 116(2), 217-225. https://doi.org/10.1007/s11240-013-0397-9

Lippman, Z., & Zamir, D. (2007). Heterosis: revisiting the magic. *Trends in Genetics*, 23(2), 60-66. https://doi.org/10.1016/j.tig.2006.12.006

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13), 1658-1659. https://doi.org/10.1093/bioinformatics/btl158

Li, Z., Defoort, J., Tasdighian, S., Maere, S., Van de Peer, Y., & De Smet, R. (2016). Gene Duplicability of Core Genes Is Highly Consistent across All Angiosperms. *The Plant Cell*, 28(2), 326-344. https://doi.org/10.1105/tpc.15.00877

Long, M., VanKuren, N., Chen, S., & Vibranovski, M. (2013). New Gene Evolution: Little Did We Know. *Annual Review of Genetics*, 47(1), 307-333. https://doi.org/10.1146/annurev-genet-111212-133301

Luse, D. (2013). The RNA polymerase II preinitiation complex. *Transcription*, 5(1), e27050. https://doi.org/10.4161/trns.27050

Ma, D., Ding, Q., Guo, Z., Zhao, Z., Wei, L., Li, Y., Song, S., & Zheng, H. (2021). Identification, characterization and expression analysis of lineage-specific genes within mangrove species *Aegiceras corniculatum*. *Molecular Genetics and Genomics*, 296(6), 1235-1247. https://doi.org/10.1007/s00438-021-01810-0

Majka, J., Glombik, M., Doležalová, A., Kneřová, J., Ferreira, M., Zwierzykowski, Z., Duchoslav, M., Studer, B., Doležel, J., Bartoš, J., & Kopecký, D. (2023). Both male and female meiosis contribute to non-Mendelian inheritance of parental chromosomes in interspecific plant hybrids (*Lolium* × *Festuca*). *New Phytologist*, 238(2), 624-636. https://doi.org/10.1111/nph.18753

Mandáková, T., Li, Z., Barker, M., & Lysak, M. (2017). Diverse genome organization following 13 independent mesopolyploid events in *Brassicaceae* contrasts with convergent patterns of gene retention. *The Plant Journal*, 91(1), 3-21. https://doi.org/10.1111/tpj.13553

Mandáková, T., & Lysak, M. (2018). Post-polyploid diploidization and diversification through dysploid changes. *Current Opinion in Plant Biology*, 42, 55-65. https://doi.org/10.1016/j.pbi.2018.03.001

Ma, S., Yuan, Y., Tao, Y., Jia, H., & Ma, Z. (2020). Identification, characterization and expression analysis of lineage-specific genes within *Triticeae*. *Genomics*, 112(2), 1343-1350. https://doi.org/10.1016/j.ygeno.2019.08.003

Mascher, M. (2021). Pseudomolecules and annotation of the third version of the reference genome sequence assembly of barley cv. Morex [Morex V3]. e!DAL - Plant Genomics and Phenomics Research Data Repository (PGP), IPK Gatersleben, Seeland OT Gatersleben, Corrensstraße 3, 06466, Germany. https://doi.org/10.5447/IPK/2021/3

McCue, A., & Slotkin, R. (2012). Transposable element small RNAs as regulators of gene expression. *Trends in Genetics*, 28(12), 616-623. https://doi.org/10.1016/j.tig.2012.09.001

McLysaght, A., & Guerzoni, D. (2015). New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678), 20140332. https://doi.org/10.1098/rstb.2014.0332

Meister, G., & Tuschl, T. (2004). Mechanisms of gene silencing by double-stranded RNA. *Nature*, 431(7006), 343-349. https://doi.org/10.1038/Nature02873

Merrick, W., & Pavitt, G. (2018). Protein Synthesis Initiation in Eukaryotic Cells. *Cold Spring Harbor Perspectives in Biology*, 10(12), a033092. https://doi.org/10.1101/cshperspect.a033092

Mukherjee, S., Panda, A., & Ghosh, T. (2015). Elucidating evolutionary features and functional implications of orphan genes in *Leishmania major*. *Infection, Genetics and Evolution*, 32, 330-337. https://doi.org/10.1016/j.meegid.2015.03.031

Neme, R., & Tautz, D. (2014). Evolution: Dynamics of De Novo Gene Emergence. *Current Biology*, 24(6), R238-R240. https://doi.org/10.1016/j.cub.2014.02.016

Nieto Feliner, G., Casacuberta, J., & Wendel, J. (2020). Genomics of Evolutionary Novelty in Hybrids and Polyploids. *Frontiers in Genetics*, 11. https://doi.org/10.3389/fgene.2020.00792

O'Conner, S., & Li, L. (2020). Mitochondrial Fostering: The Mitochondrial Genome May Play a Role in Plant Orphan Gene Evolution. *Frontiers in Plant Science*, 11, 600117. https://doi.org/10.3389/fpls.2020.600117

Palazzo, A., & Lee, E. (2018). Sequence Determinants for Nuclear Retention and Cytoplasmic Export of mRNAs and lncRNAs. *Frontiers in Genetics*, 9, 440. https://doi.org/10.3389/fgene.2018.00440

Palmieri, N., Kosiol, C., & Schlötterer, C. (2014). The life cycle of *Drosophila* orphan genes. *eLife*, 3, e01311. https://doi.org/10.7554/eLife.01311

Palmiter, R. (1975). Quantitation of parameters that determine the rate of ovalbumin synthesis. *Cell*, 4(3), 189-197. https://doi.org/10.1016/0092-8674(75)90167-1

Panigrahi, A., & O'Malley, B. (2021). Mechanisms of enhancer action: the known and the unknown. *Genome Biology*, 22(1), 108. https://doi.org/10.1186/s13059-021-02322-1

Passmore, L., & Coller, J. (2022). Roles of mRNA poly(A) tails in regulation of eukaryotic gene expression. *Nature Reviews Molecular Cell Biology*, 23(2), 93-106. https://doi.org/10.1038/s41580-021-00417-y

Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B., Salazar, G., Bileschi, M., Bork, P., Bridge, A., Colwell, L., Gough, J., Haft, D., Letunić, I., Marchler-Bauer, A., Mi, H., Natale, D., Orengo, C., Pandurangan, A., Rivoire, C. et al. (2023). InterPro in 2022. *Nucleic Acids Research,* 51(1), D418-D427. https://doi.org/10.1093/nar/gkac993

Pellicer, J., Fay, M., & Leitch, I. (2010). The largest eukaryotic genome of them all?. *Botanical Journal of the Linnean Society*, 164(1), 10-15. https://doi.org/10.1111/j.1095-8339.2010.01072.x

Pink, R., Wicks, K., Caley, D., Punch, E., Jacobs, L., & Francisco Carter, D. (2011). Pseudogenes: Pseudo-functional or key regulators in health and disease?. *RNA*, 17(5), 792-798. https://doi.org/10.1261/rna.2658311

Piovesan, A., Pelleri, M., Antonaros, F., Strippoli, P., Caracausi, M., & Vitale, L. (2019). On the length, weight and GC content of the human genome. *BMC Research Notes,* 12(1), 106. https://doi.org/10.1186/s13104-019-4137-z

Piriyapongsa, J., Mariño-Ramírez, L., & Jordan, I. (2007). Origin and Evolution of Human microRNAs From Transposable Elements. *Genetics*, 176(2), 1323-1337. https://doi.org/10.1534/genetics.107.072553

Pontvianne, F., Matía, I., Douet, J., Tourmente, S., Medina, F., Echeverria, M., Sáez-Vásquez, J., & Bickmore, W. (2007). Characterization of AtNUC - L1 Reveals a Central Role of Nucleolin in Nucleolus Organization and Silencing of AtNUC - L2 Gene in *Arabidopsis*. *Molecular Biology of the Cell*, 18(2), 369-379. https://doi.org/10.1091/mbc.e06-08-0751

Pope, S., & Medzhitov, R. (2018). Emerging Principles of Gene Expression Programs and Their Regulation. *Molecular Cell*, 71(3), 389-397. https://doi.org/10.1016/j.molcel.2018.07.017

Prabh, N., & Rödelsperger, C. (2019). De Novo , Divergence, and Mixed Origin Contribute to the Emergence of Orphan Genes in *Pristionchus* Nematodes. *G3 Genes|Genomes|Genetics*, 9(7), 2277-2286. https://doi.org/10.1534/g3.119.400326

Prieto-Godino, L., Rytz, R., Bargeton, B., Abuin, L., Arguello, J., Peraro, M., & Benton, R. (2016). Olfactory receptor pseudo-pseudogenes. *Nature*, 539(7627), 93-97. https://doi.org/10.1038/Nature19824

Proud, C. (2019). Phosphorylation and Signal Transduction Pathways in Translational Control. *Cold Spring Harbor Perspectives in Biology*, 11(7), a033050. https://doi.org/10.1101/cshperspect.a033050

Qi, M., Zheng, W., Zhao, X., Hohenstein, J., Kandel, Y., O'Conner, S., Wang, Y., Du, C., Nettleton, D., MacIntosh, G., Tylka, G., Wurtele, E., Whitham, S., & Li, L. (2019). QQS orphan gene and its interactor NF - YC 4 reduce susceptibility to pathogens and pests. *Plant Biotechnology Journal*, 17(1), 252-263. https://doi.org/10.1111/pbi.12961

Quinlan, A., & Hall, I. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841-842. https://doi.org/10.1093/bioinformatics/btq033

R Core Team. (2022). R: A language and environment for statistical computing. https://www.R-project.org/ (20.03.2023).

Ramsey, J., & Schemske, D. (1998). PATHWAYS, MECHANISMS, AND RATES OF POLYPLOID FORMATION IN FLOWERING PLANTS. *Annual Review of Ecology and Systematics*, 29(1), 467-501. https://doi.org/10.1146/annurev.ecolsys.29.1.467

Reinhardt, J., Wanjiru, B., Brant, A., Saelao, P., Begun, D., Jones, C., & Betran, E. (2013). De Novo ORFs in Drosophila Are Important to Organismal Fitness and Evolved Rapidly from Previously Non-coding Sequences. *PLoS Genetics*, 9(10), e1003860. https://doi.org/10.1371/journal.pgen.1003860

Renny-Byfield, S., Kovarik, A., Kelly, L., Macas, J., Novak, P., Chase, M., Nichols, R., Pancholi, M., Grandbastien, M., & Leitch, A. (2013). Diploidization and genome size change in allopolyploids is associated with differential dynamics of low- and high-copy sequences. *The Plant Journal*, 74(5), 829-839. https://doi.org/10.1111/tpj.12168

Robinson, M., & Smyth, G. (2007). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2), 321-332. https://doi.org/10.1093/biostatistics/kxm030

Robinson, M., McCarthy, D., & Smyth, G. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139-140. https://doi.org/10.1093/bioinformatics/btp616

Ruprecht, C., Lohaus, R., Vanneste, K., Mutwil, M., Nikoloski, Z., Van de Peer, Y., & Persson, S. (2017). Revisiting ancestral polyploidy in plants. *Science Advances*, 3(7), e1603195. https://doi.org/10.1126/sciadv.1603195

Sadanandom, A., Bailey, M., Ewan, R., Lee, J., & Nelis, S. (2012). The ubiquitin–proteasome system: central modifier of plant signalling. *New Phytologist*, 196(1), 13-28. https://doi.org/10.1111/j.1469-8137.2012.04266.x

Saha, A., Mitchell, J., Nishida, Y., Hildreth, J., Ariberre, J., Gilbert, W., Garfinkel, D., & Sundquist, W. (2015). A trans -Dominant Form of Gag Restricts Ty1 Retrotransposition and Mediates Copy Number Control. *Journal of Virology*, 89(7), 3922-3938. https://doi.org/10.1128/JVI.03060-14

Sainsbury, S., Bernecky, C., & Cramer, P. (2015). Structural basis of transcription initiation by RNA polymerase II. *Nature Reviews Molecular Cell Biology*, 16(3), 129-143. https://doi.org/10.1038/nrm3952

Salazar, C., Baxter, S., Pardo-Diaz, C., Wu, G., Surridge, A., Linares, M., Bermingham, E., Jiggins, C., & Walsh, B. (2010). Genetic Evidence for Hybrid Trait Speciation in *Heliconius* Butterflies. *PLoS Genetics*, 6(4). https://doi.org/10.1371/journal.pgen.1000930

Salman-Minkov, A., Sabath, N., & Mayrose, I. (2016). Whole-genome duplication as a key factor in crop domestication. *Nature Plants*, 2(8), 16115. https://doi.org/10.1038/nplants.2016.115

Samandi, S., Roy, A., Delcourt, V., Lucier, J., Gagnon, J., Beaudoin, M., Vanderperre, B., Breton, M., Motard, J., Jacques, J., Brunelle, M., Gagnon-Arsenault, I., Fournier, I., Ouangraoua, A., Hunting, D., Cohen, A., Landry, C., Scott, M., & Roucou, X. (2017). Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *eLife*, 6, e27860. https://doi.org/10.7554/eLife.27860

Samuel Yang, S., Cheung, F., Lee, J., Ha, M., Wei, N., Sze, S., Stelly, D., Thaxton, P., Triplett, B., Town, C., & Jeffrey Chen, Z. (2006). Accumulation of genome-specific transcripts, transcription factors and phytohormonal regulators during early stages of fiber cell development in allotetraploid cotton. *The Plant Journal*, 47(5), 761-775. https://doi.org/10.1111/j.1365-313X.2006.02829.x

Sanan-Mishra, N., Abdul Kader Jailani, A., Mandal, B., & Mukherjee, S. (2021). Secondary siRNAs in Plants: Biosynthesis, Various Functions, and Applications in Virology. *Frontiers in Plant Science*, 12, 610283. https://doi.org/10.3389/fpls.2021.610283

Shi, J., Whyte, W., Zepeda-Mendoza, C., Milazzo, J., Shen, C., Roe, J., Minder, J., Mercan, F., Wang, E., Eckersley-Maslin, M., Campbell, A., Kawaoka, S., Shareef, S., Zhu, Z., Kendall, J., Muhar, M., Haslinger, C., Yu, M., Roeder, R. et al. (2013). Role of SWI/SNF in acute leukemia maintenance and enhancer-mediated Myc regulation. *Genes & Development*, 27(24), 2648-2662. https://doi.org/10.1101/gad.232710.113

Schlötterer, C. (2015). Genes from scratch – the evolutionary fate of de novo genes. *Trends in Genetics*, 31(4), 215-219. https://doi.org/10.1016/j.tig.2015.02.007

Schnable, J., Springer, N., & Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proceedings of the National Academy of Sciences*, 108(10), 4069-4074. https://doi.org/10.1073/pnas.1101368108

Schnable, P., Ware, D., Fulton, R., Stein, J., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T., Minx, P., Reily, A., Courtney, L., Kruchowski, S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B. et al. (2009). The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science*, 326(5956), 1112-1115. https://doi.org/10.1126/Science.1178534

Simão, F., Waterhouse, R., Ioannidis, P., Kriventseva, E., & Zdobnov, E. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), 3210-3212. https://doi.org/10.1093/bioinformatics/btv351

Singh, U., & Syrkin Wurtele, E. (2020). How new genes are born. *eLife*, 9, e55136. https://doi.org/10.7554/eLife.55136

Song, Q., & Chen, Z. (2015). Epigenetic and developmental regulation in plant polyploids. *Current Opinion in Plant Biology*, 24, 101-109. https://doi.org/10.1016/j.pbi.2015.02.007

Song, Q., Zhang, T., Stelly, D., & Chen, Z. (2017). Epigenomic and functional analyses reveal roles of epialleles in the loss of photoperiod sensitivity during domestication of allotetraploid cottons. *Genome Biology*, 18(1), 99. https://doi.org/10.1186/s13059-017-1229-8

Stanke, M., & Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research*, 33(), W465-W467. https://doi.org/10.1093/nar/gki458

Stemshorn, K., Reed, F., Nolte, A., & Tautz, D. (2011). Rapid formation of distinct hybrid lineages after secondary contact of two fish species (*Cottus* sp.). *Molecular Ecology*, 20(7), 1475-1491. https://doi.org/10.1111/j.1365-294X.2010.04997.x

Stočes, Š., Ruttink, T., Bartoš, J., Studer, B., Yates, S., Zwierzykowski, Z., Abrouk, M., Roldán-Ruiz, I., Książczyk, T., Rey, E., Doležel, J., & Kopecký, D. (2016). Orthology Guided Transcriptome Assembly of Italian Ryegrass and Meadow Fescue for Single-Nucleotide Polymorphism Discovery. *The Plant Genome*, 9(3). https://doi.org/10.3835/plantgenome2016.02.0017

Sun, H., Wu, S., Zhang, G., Jiao, C., Guo, S., Ren, Y., Zhang, J., Zhang, H., Gong, G., Jia, Z., Zhang, F., Tian, J., Lucas, W., Doyle, J., Li, H., Fei, Z., & Xu, Y. (2017). Karyotype Stability and Unbiased Fractionation in the Paleo-Allotetraploid Cucurbita Genomes. *Molecular Plant,* 10(10), 1293-1306. https://doi.org/10.1016/j.molp.2017.09.003

Sun, W., Zhao, X., & Zhang, Z. (2015). Identification and evolution of the orphan genes in the domestic silkworm, *Bombyx mori*. *FEBS Letters*, 589(19), 2731-2738. https://doi.org/10.1016/j.febslet.2015.08.008

Tanvir, R., Ping, W., Sun, J., Cain, M., Li, X., & Li, L. (2022). AtQQS orphan gene and NtNF-YC4 boost protein accumulation and pest resistance in tobacco (*Nicotiana tabacum*). *Plant Science*, 317, 111198. https://doi.org/10.1016/j.plantsci.2022.111198

Tautz, D., & Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews Genetics,* 12(10), 692-702. https://doi.org/10.1038/nrg3053

Thakur, J., Packiaraj, J., & Henikoff, S. (2021). Sequence, Chromatin and Evolution of Satellite DNA. *International Journal of Molecular Sciences*, 22(9), 4309. https://doi.org/10.3390/ijms22094309

The International Barley Genome Sequencing Consortium. (2012). A physical, genetic and functional sequence assembly of the barley genome. *Nature*, 491(7426), 711-716. https://doi.org/10.1038/Nature11543

Thomas, B., Pedersen, B., & Freeling, M. (2006). Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Research*, 16(7), 934-946. https://doi.org/10.1101/gr.4708406

Thomas, M., & Chiang, C. (2008). The General Transcription Machinery and General Cofactors. *Critical Reviews in Biochemistry and Molecular Biology*, 41(3), 105-178. https://doi.org/10.1080/10409230600648736

Tian, B., & Manley, J. (2017). Alternative polyadenylation of mRNA precursors. *Nature Reviews Molecular Cell Biology*, 18(1), 18-30. https://doi.org/10.1038/nrm.2016.116

Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X., & Mar Alba, M. (2009). Origin of Primate Orphan Genes: A Comparative Genomics Approach. *Molecular Biology and Evolution*, 26(3), 603-612. https://doi.org/10.1093/molbev/msn281

Tomari, Y., & Zamore, P. (2005). Perspective: machines for RNAi. *Genes & Development*, 19(5), 517-529. https://doi.org/10.1101/gad.1284105

Traut, W., Rees, D., Dufresne, F., Glémet, H., & Belzile, C. (2007). Amphipod genome sizes: first estimates for Arctic species reveal genomic giants. *Genome*, 50(2), 151-158. https://doi.org/10.1139/G06-155

Trinotate: Transcriptome Functional Annotation and Analysis. https://github.com/Trinotate/Trinotate/wiki (20.03.2023).

Vakirlis, N., Carvunis, A., & McLysaght, A. (2020). Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife*, 9. https://doi.org/10.7554/eLife.53500

Vakirlis, N., Hebert, A., Opulente, D., Achaz, G., Hittinger, C., Fischer, G., Coon, J., & Lafontaine, I. (2018). A Molecular Portrait of De Novo Genes in Yeasts. *Molecular Biology and Evolution*, 35(3), 631-645. https://doi.org/10.1093/molbev/msx315

Van Oss, S., & Carvunis, A. (2019). De novo gene birth. *PLOS Genetics*, 15(5), e1008160. https://doi.org/10.1371/journal.pgen.1008160

Vicient, C., & Casacuberta, J. (2017). Impact of transposable elements on polyploid plant genomes. *Annals of Botany*, 120(2), 195-207. https://doi.org/10.1093/aob/mcx078

Vivares, C. (1999). On the genome of Microsporidia. *Journal of Eukaryotic Microbiology*, 46(1), 16A.

Vlasova-St. Louis, I., & Bohjanen, P. (2011). Coordinate regulation of mRNA decay networks by GU-rich elements and CELF1. *Current Opinion in Genetics & Development*, 21(4), 444-451. https://doi.org/10.1016/j.gde.2011.03.002

Wang, B., & Brendel, V. (2006). Genomewide comparative analysis of alternative splicing in plants. *Proceedings of the National Academy of Sciences*, 103(18), 7175-7180. https://doi.org/10.1073/pnas.0602039103

Wang, J., Tian, L., Lee, H., Wei, N., Jiang, H., Watson, B., Madlung, A., Osborn, T., Doerge, R., Comai, L., & Chen, Z. (2006). Genomewide Nonadditive Gene Regulation in *Arabidopsis* Allotetraploids. *Genetics*, 172(1), 507-517. https://doi.org/10.1534/genetics.105.047894

Wang, X., Morton, J., Pellicer, J., Leitch, I., & Leitch, A. (2021). Genome downsizing after polyploidy: mechanisms, rates and selection pressures. *The Plant Journal*, 107(4), 1003-1015. https://doi.org/10.1111/tpj.15363

Waterhouse, R., Seppey, M., Simão, F., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E., & Zdobnov, E. (2018). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution*, 35(3), 543-548. https://doi.org/10.1093/molbev/msx319

Wilczynska, A., & Bushell, M. (2015). The complexity of miRNA-mediated repression. *Cell Death & Differentiation*, 22(1), 22-33. https://doi.org/10.1038/cdd.2014.112

Wilusz, J. (2015). Controlling translation via modulation of tRNA levels. *Wiley Interdisciplinary Reviews: RNA*, 6(4), 453-470. https://doi.org/10.1002/wrna.1287

Wissler, L., Gadau, J., Simola, D., Helmkampf, M., & Bornberg-Bauer, E. (2013). Mechanisms and Dynamics of Orphan Gene Emergence in Insect Genomes. *Genome Biology and Evolution*, 5(2), 439-455. https://doi.org/10.1093/gbe/evt009

Woodhouse, M., Cheng, F., Pires, J., Lisch, D., Freeling, M., & Wang, X. (2014). Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. *Proceedings of the National Academy of Sciences*, 111(14), 5283-5288. https://doi.org/10.1073/pnas.1402475111

Woodhouse, M., Schnable, J., Pedersen, B., Lyons, E., Lisch, D., Subramaniam, S., Freeling, M., & Wolfe, K. (2010). Following Tetraploidy in Maize, a Short Deletion Mechanism Removed Genes Preferentially from One of the Two Homeologs. *PLoS Biology*, 8(6), e1000409. https://doi.org/10.1371/journal.pbio.1000409

Wood, T., Takebayashi, N., Barker, M., Mayrose, I., Greenspoon, P., & Rieseberg, L. (2009). The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences*, 106(33), 13875-13879. https://doi.org/10.1073/pnas.0811575106

Wu, D., Wang, X., Li, Y., Zeng, L., Irwin, D., & Zhang, Y. (2014). "Out of Pollen" Hypothesis for Origin of New Genes in Flowering Plants: Study from *Arabidopsis thaliana*. *Genome Biology and Evolution*, 6(10), 2822-2829. https://doi.org/10.1093/gbe/evu206

Wu, J., Lin, L., Xu, M., Chen, P., Liu, D., Sun, Q., Ran, L., & Wang, Y. (2018). Homoeolog expression bias and expression level dominance in resynthesized allopolyploid *Brassica napus*. *BMC Genomics*, 19(1), 586. https://doi.org/10.1186/s12864-018-4966-5

Wu, T., & Watanabe, C. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9), 1859-1875. https://doi.org/10.1093/bioinformatics/bti310

Xu, Y., Wu, G., Hao, B., Chen, L., Deng, X., & Xu, Q. (2015). Identification, characterization and expression analysis of lineage-specific genes within sweet orange (*Citrus sinensis*). *BMC Genomics*, 16(1). https://doi.org/10.1186/s12864-015-2211-z

Yakimowski, S., & Rieseberg, L. (2014). The role of homoploid hybridization in evolution: A century of studies synthesizing genetics and ecology. *American Journal of Botany*, 101(8), 1247-1258. https://doi.org/10.3732/ajb.1400201

Yang, L., Zou, M., Fu, B., & He, S. (2013). Genome-wide identification, characterization, and expression analysis of lineage-specific genes within zebrafish. *BMC Genomics*, 14(1). https://doi.org/10.1186/1471-2164-14-65

Ye, B., Wang, R., & Wang, J. (2016). Correlation analysis of the mRNA and miRNA expression profiles in the nascent synthetic allotetraploid *Raphanobrassica*. *Scientific Reports*, 6(1), 37416. https://doi.org/10.1038/srep37416

Yoo, M., Szadkowski, E., & Wendel, J. (2013). Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity*, 110(2), 171-180. https://doi.org/10.1038/hdy.2012.94

Yu, K., Feng, M., Yang, G., Sun, L., Qin, Z., Cao, J., Wen, J., Li, H., Zhou, Y., Chen, X., Peng, H., Yao, Y., Hu, Z., Guo, W., Sun, Q., Ni, Z., Adams, K., & Xin, M. (2020). Changes in Alternative Splicing in Response to Domestication and Polyploidization in Wheat. *Plant Physiology*, 184(4), 1955-1968. https://doi.org/10.1104/pp.20.00773

Zhang, H., Lang, Z., & Zhu, J. (2018). Dynamics and function of DNA methylation in plants. *Nature Reviews Molecular Cell Biology*, 19(8), 489-506. https://doi.org/10.1038/s41580-018-0016-z

Zhang, L., He, J., He, H., Wu, J., & Li, M. (2021). Genome-wide unbalanced expression bias and expression level dominance toward *Brassica oleracea* in artificially synthesized intergeneric hybrids of *Raphanobrassica*. *Horticulture Research*, 8(1), 246. https://doi.org/10.1038/s41438-021-00672-2

Zhang, L., Ren, Y., Yang, T., Li, G., Chen, J., Gschwend, A., Yu, Y., Hou, G., Zi, J., Zhou, R., Wen, B., Zhang, J., Chougule, K., Wang, M., Copetti, D., Peng, Z., Zhang, C., Zhang, Y., Ouyang, Y. et al. (2019). Rapid evolution of protein diversity by de novo origination in *Oryza*. *Nature Ecology & Evolution*, 3(4), 679-690. https://doi.org/10.1038/s41559-019-0822-5

Zhang, Z., Gou, X., Xun, H., Bian, Y., Ma, X., Li, J., Li, N., Gong, L., Feldman, M., Liu, B., & Levy, A. (2020). Homoeologous exchanges occur through intragenic recombination generating novel transcripts and proteins in wheat and other polyploids. *Proceedings of the National Academy of Sciences*, 117(25), 14561-14571. https://doi.org/10.1073/pnas.2003505117

Zhao, L., Saelao, P., Jones, C., & Begun, D. (2014). Origin and Spread of de Novo Genes in *Drosophila melanogaster* Populations. *Science*, 343(6172), 769. https://doi.org/10.1126/Science.1248286

Zhao, Y., Tang, L., Li, Z., Jin, J., Luo, J., & Gao, G. (2015). Identification and analysis of unitary loss of long-established protein-coding genes in *Poaceae* shows evidences for biased gene loss and putatively functional transcription of relics. *BMC Evolutionary Biology*, 15(1), 66. https://doi.org/10.1186/s12862-015-0345-x

Zhou, R., Moshgabadi, N., & Adams, K. (2011). Extensive changes to alternative splicing patterns following allopolyploidy in natural and resynthesized polyploids. *Proceedings of the National Academy of Sciences*, 108(38), 16122-16127. https://doi.org/10.1073/pnas.1109551108

Zile, K., Dessimoz, C., Wurm, Y., Masel, J., & Alba, M. (2020). Only a Single Taxonomically Restricted Gene Family in the *Drosophila melanogaster* Subgroup Can Be Identified with High Confidence. *Genome Biology and Evolution*, 12(8), 1355-1366. https://doi.org/10.1093/gbe/evaa127

Zuellig, M., Sweigart, A., & Malik, H. (2018). Gene duplicates cause hybrid lethality between sympatric species of *Mimulus*. *PLOS Genetics*, 14(4), e1007130. https://doi.org/10.1371/journal.pgen.1007130

8 LIST OF ABBREVIATIONS

AS	Alternative splicing
ELD	Expression level dominance
GTF	General transcription factor
HEB	Homoeolog expression bias
LSG	Lineage-specific gene
RISC	RNA-induced silencing complex
SSG	Species-specific gene
TE	Transposable element
TF	Transcription factor
UTR	Untranslated region
WGD	Whole genome duplication

9 APPENDICES

Electronic:

- *Chudecka_appendix_1.zip* – zipped folder with files *Festuca_pratensis_transcriptome.xlsx* and *Lolium_multiflorum_transcriptome.xlsx* containing annotated transcriptomes for respective species (reports generated with Trinotate).
- *Chudecka_appendix_2.zip* – zipped folder with file *Set_II_Interpro_entries.xlsx* containing full list of matched Interpro entries for Set II with counts.