Only a single taxonomically restricted gene family in the *Drosophila melanogaster* subgroup can be identified with high confidence

Karina Zile[1,*], Christophe Dessimoz[2,3,4], Yannick Wurm[5,6], Joanna Masel[7]

[1] University College London, Division of Biosciences, Gower Street, London, UK WC1E 6BT
[2] Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
[3] University of Lausanne, Department of Computational Biology and Center for Integrative Genomics, 1015 Lausanne, Switzerland
[4] University College London, Department of Genetics, Evolution & Environment and Department of Computer Science, London WC1E 6BT, United Kingdom
[5] Queen Mary University of London, School of Biological and Chemical Sciences, Mile End Road, London E1 4NS, United Kingdom
[6] Alan Turing Institute, London NW1 2DB, United Kingdom
[7] University of Arizona, Department of Ecology and Evolutionary Biology, Tucson, AZ, 85721, USA
[*] Corresponding author, email: karina.zile@gmail.com

Abstract

Taxonomically restricted genes (TRGs) are genes that are present only in one clade. Protein-coding TRGs may evolve *de novo* from previously non-coding sequences: functional ncRNA, introns or alternative reading frames of older protein-coding genes, or intergenic sequences. A major challenge in studying *de novo* genes is the need to avoid both false positives (non-functional open reading frames and/or functional genes that did not arise *de novo*) and false negatives. Here we search conservatively for high confidence TRGs as the most promising candidates for experimental studies, ensuring functionality through conservation across at least two species, and ensuring *de novo* status through examination of homologous non-coding sequences. Our pipeline also avoids ascertainment biases associated with preconceptions of how *de novo* genes are born. We identify one TRG family that evolved *de novo* in the *Drosophila melanogaster* subgroup. This TRG family contains single copy genes in *D. simulans* and *D. sechellia*. It originated in an intron of a well-established gene, sharing that intron with another well-established gene upstream. These TRGs contain an intron that pre-dates their ORF. These genes have not been previously reported as *de novo* originated, and to our knowledge they are the best *Drosophila* candidates identified so far for experimental studies aimed at elucidating the properties of *de novo* genes.

Introduction

Some genes are present only in one clade, and are therefore called taxonomically restricted genes (TRGs). They are also referred to as *orphans* or simply novel genes. Some of these may have originated *de novo*. We use the origin version of the selected effect definition of function (Linquist et al., 2020) to determine when a sequence becomes a protein-coding gene. This means that de novo birth occurs at the moment beyond which a mutation leading to loss of the protein product would have a negative effect on fitness. For the birth to occur, a *de novo* TRG needs not only the amino acid sequence itself, but also the right environment and expression regulation pattern, to confer an advantage to the organism. Protein-coding genes may evolve *de novo* from non-coding regions (Vakirlis et al., 2017; McLysaght and Guerzoni, 2015), in alternative frames of established genes (Willis and Masel, 2018; Guan et al., 2018), or as a result of genome rearrangement (Chen et al., 2015; Stewart and Rogers, 2019).

The research enterprise is biased toward studying ancient gene families with homologs across multiple model organisms, and so the properties and evolutionary dynamics of young TRGs are not well understood. TRGs are likely to include proteins with as yet undocumented functions and, especially in the case of *de novo* genes, new protein domains or other structural forms that are yet to be discovered (Bungard et al., 2017). Mounting evidence suggests that TRGs can acquire important functions. For example, a TRG in the tardigrade *Ramazzottius varieornatus* produces a protein that protects DNA and improves radio-tolerance (Hashimoto et al., 2016). TRGs in *Hymenoptera* are implicated in the speciation of parasitoid wasps and in the production of diverse venoms characteristic of this clade (Werren et al., 2010). Albertin et al. (2015) identified numerous cephalopod-specific genes and were able to find hints about their diverse functions based on their tissue-specific expression profiles. These examples remain anecdotal since functional characteristics of TRGs cannot be inferred computationally due to the lack of homologs outside a specific clade.

Most previous studies aimed at elucidating properties and rates of emergence of *de novo* TRGs have used an approach known as "phylostratigraphy" that focuses on protein-coding genes with protein homologs within a particular clade and no detectable homology outside that specific clade. This approach is incapable of discriminating between *de novo* genes and highly diverged copies of well-established genes (Weisman et al., 2020). Hence, the properties of "young genes" reported in these studies are averages computed across the two groups, and risk attributing to TRGs properties that instead reflect the disappearance of the ability to detect homology. For example, most of the studies reported that new genes tend to be shorter (Wissler et al., 2013; Zhao et al., 2014; Ruiz-Orera et al., 2015; Sun et al., 2015) and evolve faster than well established genes (Domazet-Loso, 2003; Toll-Riera et al., 2008; Donoghue et al., 2011). It is *a priori* plausible that TRGs have these properties, but phylostratigraphy does not provide clear evidence to support this claim. It is harder to detect homology for shorter and/or faster evolving genes, and this is sufficient to explain at least the qualitative direction of the observed trend. Including synteny information in the phylostratigraphy analysis changes the inferred gene ages (Arendsee et al., 2019), demonstrating that by itself phylostratigraphy approach is not

sufficient.

Related to this are substantive disputes about the frequency of *de novo* gene birth (Casola, 2018), even though the existence of extremely well-documented case studies (Cai et al., 2008; Baalsrud et al., 2017) has made indisputable the qualitative claim that *de novo* gene birth is ongoing. Vakirlis et al. (2020) used synteny conservation to show that genes originate *de novo* from ancestral non-coding sequences as well as via divergence from ancestral genes. More quantitatively, synteny-based methods suggest that sequence divergence is not the main source of orphan genes (Vakirlis et al., 2020).

There are also compelling arguments for plausibility. Purifying selection is expected to screen occasionally translated open reading frames (ORFs) in a way that makes them more viable as raw material (Wilson and Masel, 2011). The physicochemical properties and secondary structures of evolved and random sequences are very similar, and randomly created sequences can be tolerated *in vivo* by *Escherichia coli* (Tretyachenko et al., 2017). Indeed, Neme et al. (2017) showed that at least two non-coding and one protein-coding gene could be selected from around a million randomly generated sequences (mimicking *de novo* evolution) in lab conditions. While the beneficial nature of these genes is disputed (Weisman and Eddy, 2017; Knopp and Andersson, 2018), Knopp et al. (2019) similarly selected three random peptides conferring antibiotic resistance. At minimum, substantial tolerance clearly exists.

While these arguments apply to *de novo* gene birth overall, the only way to be confident that a particular putative TRG is not merely a rapidly evolving gene duplicate is to find evidence of how it emerged. If we can identify homologous DNA region(s) in the species outside the clade from which a gene has emerged (i.e. the outgroup species), if these DNA regions are non-coding, and if we can rule out pseudogenization in this outgroup via synteny-based evidence of absence in more distant outgroups, then we have the evidence that the gene is specific to this particular clade, as well as information about the nature of the origination process. When a putative TRG has simply diverged beyond detection of its protein-coding homologs, no homology in non-coding sequence will be detectable either (although a syntenic homologous coding sequence may be found upon close scrutiny), and so a false positive *de novo* gene identification will be avoided.

A false positive could, however, arise from a horizontal gene transfer followed by pseudogenization in one lineage. Fortunately, such cases can often be excluded when homology to the donor clade is detectable. Both lack of donor sequence and pseudogenization in a member of the focal clade are required to generate such a false positive, a scenario that in combination should be reasonably rare.

One important scenario to consider is when, following a gene duplication, the ortholog in the outgroup is lost or diverges beyond detectable homology. It is therefore important to consider all likely homologous DNA regions in outgroup species, not only the single most likely region. One way to do this is to check whether the identified region in the outgroup species is homologous to any other regions in that genome. This is made relatively easy when the duplicated DNA region contains flanking, better-conserved genes, such that lo-

cal synteny information can be exploited.

Even with synteny, detecting homologous non-coding sequences can be difficult. Non-coding regions of the genome are either under little evolutionary constraint, or under constraint very different from that of protein-coding regions, depending on their function or lack thereof. What constraint they have might apply to very general properties rather than to specific nucleotides at specific positions, and hence might not be enough to prevent rapid degradation of sequence similarity (Frigola et al., 2017). This means that it is necessary to confine analysis to closely related genomes in order to identify evolutionary origins of TRGs. A measure of "evolutionary traceability" of a protein family can quantify the evolutionary distance beyond which homologous proteins can no longer be identified (Jain et al., 2019). No similar metric exists for homologous non-coding DNA regions, but it is prudent to stick to closely related species.

Some analyses restrict their search for putative TRGs to the set of already-annotated protein-coding genes. Gene annotations are based largely on ORF length, transcription, and homology to known genes. Hence, a short TRG that has no previously known homologs is likely to be missed by an annotation algorithm, despite the fact that TRGs are expected *a priori* to be short. An alternative approach is to start with all ORFs present in the genome and exclude the ones that have no evidence for being functional. Previous studies used different types of evidence of functionality: Blevins et al. (2017) analysed deep RNA sequencing and ribosome profiling data, Ruiz-Orera et al. (2018) combined that with proteomics data and single nucleotide polymorphism analysis, while Vakirlis et al. (2017) developed a logistic regression classifier trained on coding and non-coding sequences using such properties as codon frequency, hydrophobicity and aromaticity scores and structural predictions (secondary structures, transmembrane and disordered regions). However, TRGs are expected to have a narrow expression profile (Wu and Knudson, 2018) and they may have sequence properties distinct from well-studied protein families. There is thus a trade-off between false positives (non-functional ORFs) and false negatives (true TRGs excluded from the analysis). Beginning with annotated protein-coding genes tilts the balance toward false negatives, while beginning with all ORFs tilts it toward false positives. Regardless of how stringent or relaxed the requirements for evidence of functionality are, the resulting set of putative TRGs is unlikely to be both high confidence and exhaustive, limiting the potential for novel biological insights.

To advance our knowledge about *de novo* TRGs, resource-intensive experimental investigations of the most promising candidates are required, including knockout studies and structural biology experiments. Candidates need to be chosen from studies that prioritize avoiding false positives over avoiding false negatives. For example, BSC4, which is found only in *Saccharomyces cerevisiae*, has synthetic lethal knockouts (Cai et al., 2008). This strong functional evidence made it a good candidate for structural biology experiments, which showed that it folds to a partially specific three-dimensional structure (Bungard et al., 2017). Absent such direct experimental data as synthetic lethal screens, the best indication of functionality is sequence conservation between several species (Graur et al., 2013), which is by definition unavailable for single-species TRGs, even when they are functional.

Several studies have focused on identifying the evolutionary origins of putative TRGs in primates, insects and rosids, as a way of confirming their *de novo* nature (Toll-Riera et al., 2008; Zhou et al., 2008; Wissler et al., 2013; Sun et al., 2015; Donoghue et al., 2011). Unfortunately, these studies extensively ruled out TRG candidates based on thinly justified *a priori* assumptions about TRGs, in some cases discarding up to 61% of candidate genes (Vakirlis et al., 2017). For example, one study excluded genes with more than one coding exon because "it is difficult to distinguish the absence of coding potential due to frame-shifts and stop codons from the alternative explanation of evolutionary change of intron-exon boundaries" (Guerzoni and McLysaght, 2016), perhaps also believing that the evolution of both a long ORF and an intron splicing signal is highly improbable (Knowles and McLysaght, 2009). Interestingly, other studies excluded single coding-exon genes, either to avoid promoter- or enhancer-associated transcripts (PROMPTS and eRNAs) (Ruiz-Orera et al., 2015), or to avoid possible contamination of TEs incorrectly annotated as genes (Toll-Riera et al., 2008). Similarly, many studies excluded genes whose length is below a certain threshold (Yang and Huang, 2011), genes with compositions too far from an average established protein-coding gene, and genes that are evolving too fast (Vakirlis et al., 2017). In perhaps the most extreme case, Casola (2018) excluded TRG candidates which are present in several copies in a genome due to a belief that young genes could not have had the time to duplicate.

Once they have identified TRGs, a second major limitation of studies focussed on establishing the mechanism of origination is testing hypothesised mechanisms sequentially instead of looking holistically at the evidence available for each of the genes to establish their evolutionary origin. *De novo* protein-coding genes might be born within functional ncRNA, within introns or alternative frames of older protein-coding genes, or from intergenic sequences. Despite our desire to classify new genes into discrete categories, the evolutionary journey from an ancestral sequence to a new protein-coding gene might involve multiple steps, or vary along the gene's length. For example, TRGs might contain both previously non-coding sequences and fragments of well-established genes. McLysaght and Hurst (2016) proposed the classification of TRGs into several groups based on the proportion of the sequence that has previously been under natural selection for protein-coding properties. However, the distinction can blur, e.g. if previously protein-coding genes are pseudogenized or rearranged into non-coding sequence (see review by Balakirev and Ayala (2003)), and are then resurrected as part of a TRG. While pre-existing transcription may obviously be an advantage, most of the genome is likely to be transcribed across relatively short evolutionary time in at least one cell type (Neme and Tautz, 2016). Non-functional transcripts have been hypothesized to be a reservoir of genomic raw material that can increase organisms' ability to adapt (Brosius, 2005). On the other hand, their GC content makes ORFs from them more ordered and hence less suitable as raw material than for example the alternative reading frames of existing genes (Ángyán et al., 2012; Wilson et al., 2017; Casola, 2018).

Here we aim to identify high-confidence protein-coding genes that emerged *de novo*, hoping to provide a good starting point for experimental investigation. We focus on the *Drosophila melanogaster* subgroup, which is not only experimentally tractable, but also

has compact genomes of ∼140Mb, and genome assemblies of five closely related species that range in quality from good to excellent. We look for taxonomically restricted gene families (TRGFs) that emerged after the split of the *simulans-sechellia-melanogaster* clade from the *yakuba-erecta* clade and before the speciation of *D. simulans* and *D. sechellia* (Figure 1). We use conservative but strongly justified criteria to identify putative *de novo* genes among annotated protein-coding genes that have homologs in at least two of the three species in the *simulans-sechellia-melanogaster* clade. By focussing on TRGFs instead of singleton TRGs, we hope to avoid genome sequencing and assembly artefacts. We used ORF conservation across two to three species as a proxy for functionality under the selected-effect definition (Graur et al., 2013), as the half-life of a non-functional ORF is small given the probability of acquiring a stop codon by chance. A dN/dS signal of selection would be still stronger evidence for functionality, but short sequences in three closely related species do not contain enough information to reliably distinguish deviations from dN/dS = 1. By identifying the evolutionary origins of TRGs that have passed our conserved-ORF criterion for functionality, which we do by examining the homologous DNA region in the most closely related species that lack(s) the ORF, we aim to both validate their *de novo* origin (providing vetted experimental candidates) and improve our understanding of how *de novo* genes emerge.

Results

The five species we study in the *Drosophila melanogaster* subgroup (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, and *D. erecta*) had a common ancestor ∼3.3 Mya (Obbard et al., 2012) (Figure 1). Each has a genome of ∼140Mb containing ∼14,000 protein-coding genes. There is no evidence of major segmental genome duplications in this clade, reducing complications in identifying homologous non-coding sequences. The genome assembly for *D. sechellia* is highly fragmented, as confirmed by N50 metric and a BUSCO (Waterhouse et al., 2017) estimate that ∼8% of the genes likely present in the genome are missing from the assembly (Table 1). The quality of the *D. sechellia* genome assembly leads to a different distribution of annotated protein lengths compared to other species in this clade (Figure 2). For this reason, we should be especially cautious of inferring anything based on absence from *D. sechellia*.

Based on the OMA homology inference algorithm (Altenhoff et al., 2017), these five *Drosophila melanogaster* subgroup species contain 14,149 gene families. Amongst the inferred gene families there were 205 families with genes in at least 2 of the species in the *simulans-sechellia-melanogaster* clade and no genes from species outside the clade. Protein sequence similarity searches against the RefSeq database revealed diverged homologs outside the clade for 170 of these families. We used sequence similarity searches in nucleotide space to identify homologous DNA regions corresponding to the 35 putative TRGFs in all five genomes. Out of these 35 families, 18 contained conserved but unannotated ORF(s) covering $\geq 50\%$ of the putative TRGF ORF in at least one of the *yakuba-erecta* clade species, indicative of an earlier origin of these TRGs. A conserved ORF in an outgroup was considered strong evidence that the gene family originated before the speciation of the clade. We were unable to obtain a continuous alignment of

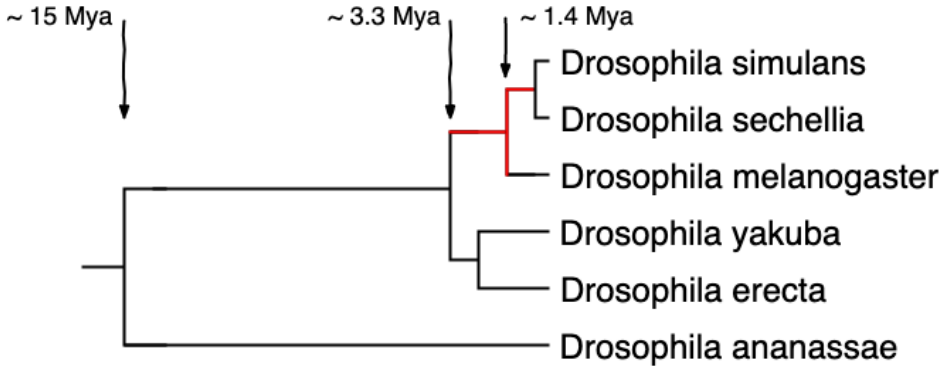

Figure 1: Species tree of the *Drosophila melanogaster* subgroup. Branch lengths correspond to divergence time estimates by Obbard et al. (2012). We looked for TRGFs that emerged during the evolutionary time marked in red, i.e. between ∼0.5 and ∼3.3 Mya. We ultimately confirm one TRGF shared only by *D. simulans* and *D. sechellia*, i.e. that originated between ∼0.5 and ∼1.4 Mya.

inferred homologous DNA regions in the *yakuba-erecta* clade for five putative TRGFs. It is unknown whether this is due to genome rearrangements and the lack of sequence conservation or simply because the true homologous DNA regions are missing from the genome assemblies. Only the 12 putative TRGFs for which we were able to obtain a continuous alignment of homologous DNA regions in all five species and show that the ORFs were only present in the *simulans-sechellia-melanogaster* clade were considered in further analyses.

Manual examination of genome annotations revealed problems and inconsistencies with ten putative TRGF annotations, such that we were uncertain about the nature or location of the ORF. For example, some of the gene families were missing a start codon, had annotated exons that overlapped in alternative frames, exons misaligned with splicing signals, or inconsistent start/stop codons and/or splicing signals across species. These putative TRGF were removed from further analysis as they did not satisfy our requirement for a conserved ORF in more than one independently annotated species.

To infer the evolutionary origins of the two putative TRGFs that remained following these filters to remove potential false positives (summarised in Figure 3), we looked at the homologous non-coding sequences whose common ancestry with the TRGF preceded the origin of the TRGF. In the process, we were able to confirm the recent *de novo* status of the first, and refute the apparent taxonomic restriction of the second.

The first TRGF evolved *de novo* in the *simulans-sechellia* clade on chromosome 3R, giving rise to *Dsim_GD19764* and *Dsec_GM10790*. These are annotated uncharacterised protein-coding genes with two CDSs and a conserved canonical GU—AG splicing signal. The protein is 129 amino acids long in *D. simulans* and 113 in *D. sechellia*. The conserved intron is 52 nucleotides long (not a multiple of 3), hence it is likely to pre-date the ORF (otherwise, later intronisation would have resulted in a frame-shift; see Yang and Huang (2011) for a detailed explanation). BUSCA predicts that this TRGF contains a transmembrane alpha helix and hence localises to the endomembrane system. We checked

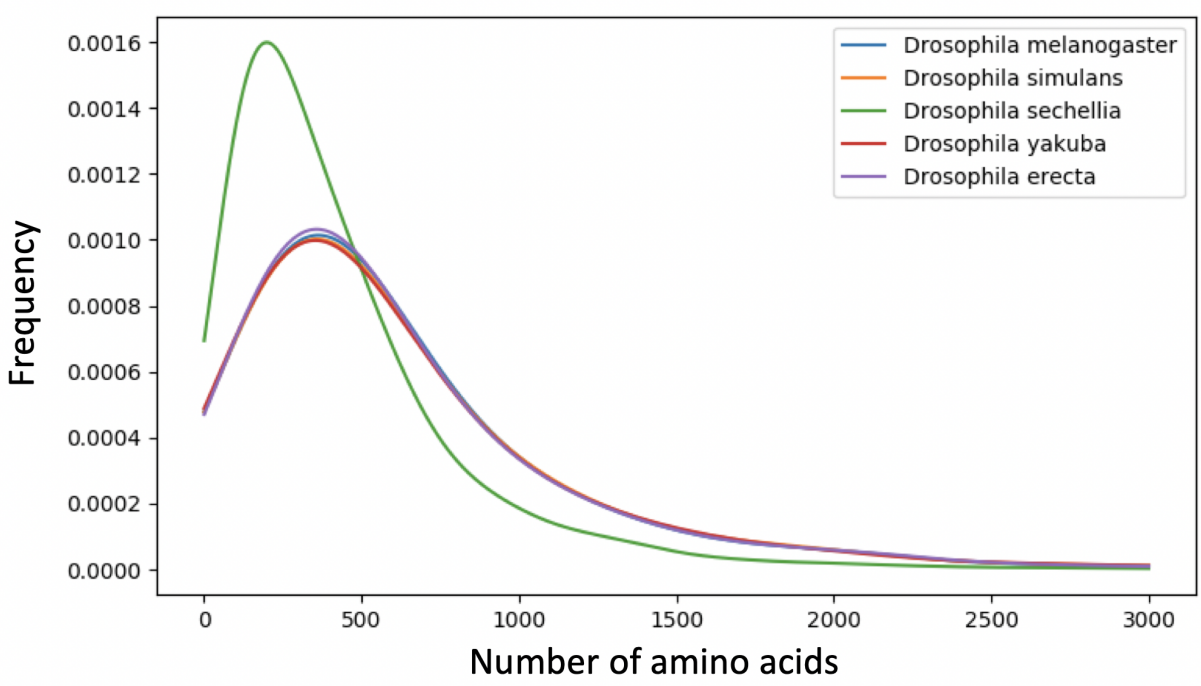

Figure 2: Protein length distributions in five *Drosophila melanogaster* subgroup species.

Dsim_GD19764 and *Dsec_GM10790* for presence of known protein domains, but found no hits. TANGO predicts that *Dsim_GD19764* and *Dsec_GM10790* have no regions prone to aggregation. There is transcriptomic evidence that *Dsim_GD19764* is expressed in the male reproductive system (Drosophila 12 Genomes Consortium, 2007), which is in line with previous results showing that TRGs are predominantly expressed in testes (Levine et al., 2006).

Dsim_GD19764 is located in an intron of a conserved protein-coding gene *Dsim_GD19765*, downstream of conserved protein-coding gene *Dsim_GD19763* located inside the same intronic region (Figure 4). In *D. sechellia*, the *Dsec_GM10791* gene harbouring two genes inside its intron appears to have lost the first two exons, and thus *Dsec_GM10790* is located in a similar genomic context but not inside an intron. The DNA regions that we presume to be homologous to TRGs in *D. melanogaster*, *D. yakuba* and *D. erecta* are located between the genes homologous to the ones neighbouring TRGs in the *simulans-sechellia* clade. There is too little nucleotide conservation for a good alignment to this region in *D. melanogaster*, which contains no ORF. Alignment can be achieved with the *yakuba-erecta* clade, where the ORF is disrupted by an early stop codon in *D. yakuba* and several indels including an early frameshift plus loss of splicing signal in *D. erecta*.

Note that Hild et al. (2003) previously inferred a protein-coding gene in *D. melanogaster* located in this region on the opposite strand, but this gene is no longer part of the official genome annotations. Because of this "homologous" hit, Heames et al. (2020) classified this TRGF as originating through rapid divergence rather than *de novo*. However, even if this no longer annotated sequence did encode a functional protein, the fact that it is on the opposite strand means that it should not be classified as a diverged homolog. *De*

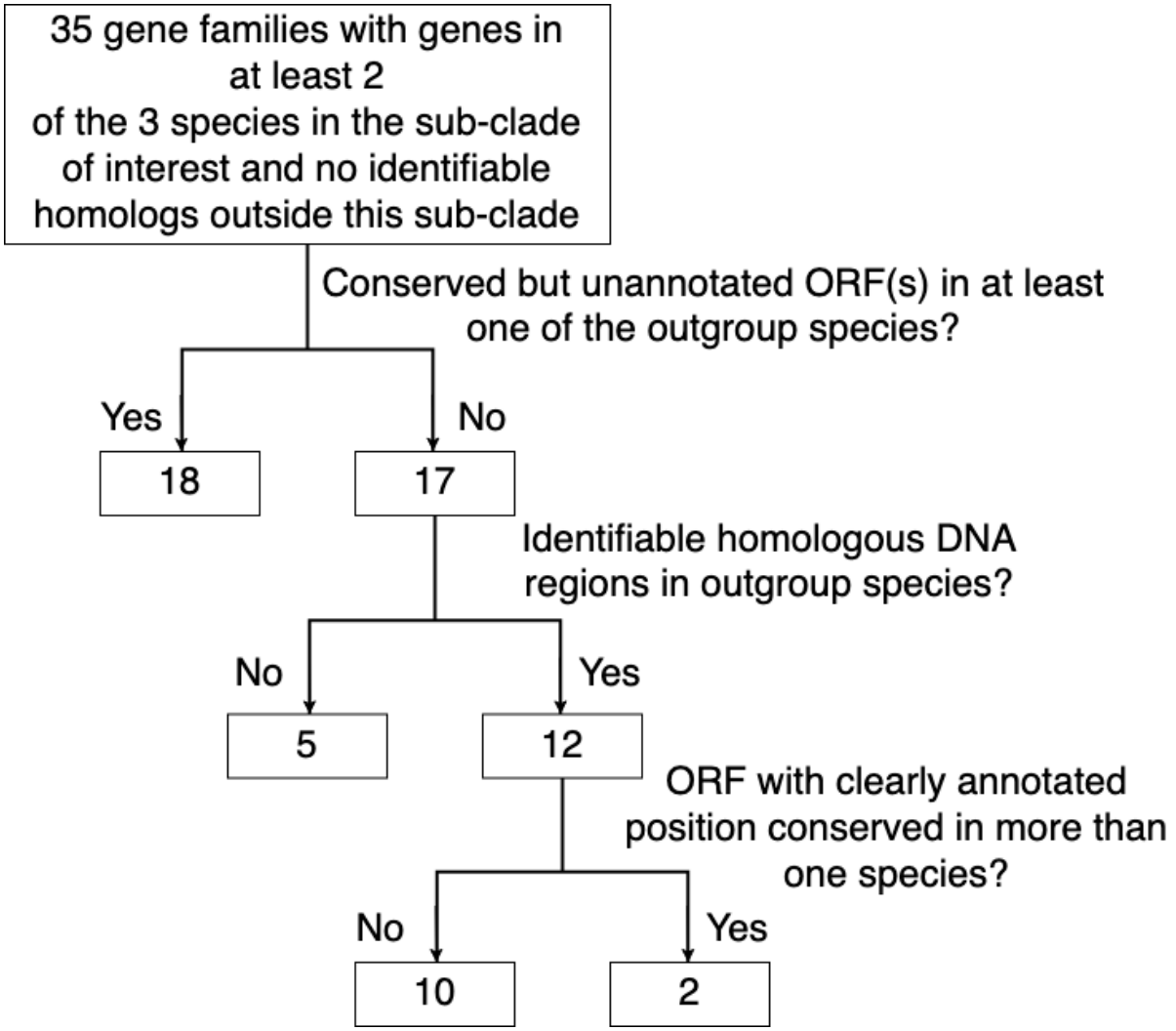

Figure 3: The elimination of TRGFs with either evidence of being false positive, or with insufficient evidence available.

novo origination that occurs in alternative reading frames is still *de novo* origination.

We identified homologous DNA regions in four additional outgroup species (*D. ananassae*, *D. suzukii*, *D. pseudoobscura* and *D. miranda*), and while the sequence conservation level was insufficient to provide precise information about the most likely ancestral state, no start codon was present in these homologous DNA regions. We can thus rule out the possibility that two independent pseudogenization events, one in *D. melanogaster* and one in the basal lineage of the *D. yakuba-erecta* clade, created the illusion of a TRGF as a false positive. The homologous regions in *D. ananassae* and *D. pseudoobscura* contain three (orange, yellow and blue in Figure 4) and two (yellow and blue in Figure 4) syntenic homologs respectively, while the putative homologous regions in *D. suzukii* and *D. miranda* (identified via BLASTn alone) contain none. Using protein sequences of the TRGF to perform tBLASTn search resulted in partial hits (covering 45-58% of the sequence) with 47-53% sequence similarity in *D. eugracilis, D. ficusphila, D. rhopaloa, D. elegans* and *D. biarmipes*. Due to the lack of syntenic evidence we were unable to confirm whether these

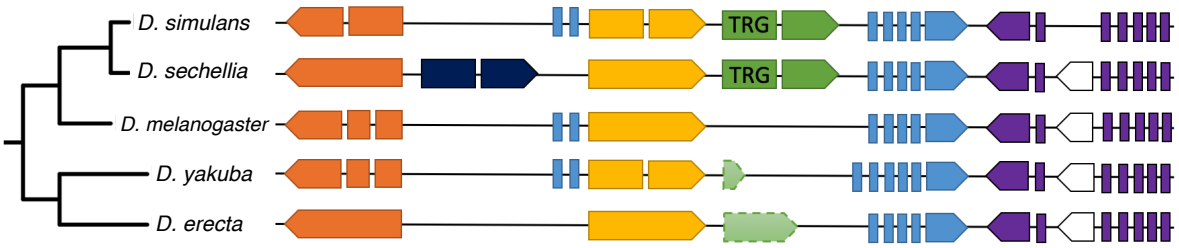

Figure 4: DNA regions homologous to the TRGF containing *Dsim_GD19764* and *Dsec_GM10790*. Homologous protein-coding genes are the same color (each element corresponds to an exon), small nuclear RNA (snRNA) genes are white. The direction of the arrow shows which strand the gene is located on. Features with dashed outlines are not annotated. The diagram is not to scale. In the order from top to bottom, the orange genes are *Dsim_GD29138*, *Dsec_GM10660*, *Dmel_CG12589*, *Dyak_GE25310*, *Dere_GG11200* (with a syntenic homolog *Dana_GF16073* in *D. ananassae*); the yellow genes are *Dsim_GD19763*, *Dsec_GM10789*, *Dmel_CG12590*, *Dyak_GE25451*, *Dere_GG12627* (with syntenic homologs *Dana_GF18925* and *Dpse_GA11706* in *D. ananassae* and *D. pseudoobscura* respectively); the blue genes are *Dsim_GD19765*, *Dsec_GM10791*, *Dmel_CG12591*, *Dyak_GE25452*, *Dere_GG12638* (with syntenic homologs *Dana_GF18926* and *Dpse_GA11707* in *D. ananassae* and *D. pseudoobscura* respectively); the purple genes are *Dsim_GD19639*, *Dsec_GM10658*, *Dmel_CG12161*, *Dyak_GE25306*, *Dere_GG11178*.

are truly homologous regions, but absence of hits from the syntenic region in closer relatives makes this unlikely. It was not possible to obtain an informative multiple sequence alignment of these highly diverged sequences, and hence no additional information was acquired from these hits.

The second gene family, that our pipeline mistakenly identified as a TRGF, contains uncharacterised protein-coding genes *Dsim_GD20667* and *Dsec_GM19408*, and an unannotated homologous ORF in *D. melanogaster*. These annotated genes are located on the 3R chromosome and contain a single CDS of length 155 in *D. simulans* and *D. melanogaster*. In *D. sechellia*, a frameshift close to the end of the CDS results in a conserved stop codon becoming in-frame and thus shortening the CDS to 139 amino acids. BUSCA predicts that the proteins localise in the nucleus.

These putative TRGs are located amongst protein-coding gene families syntenically conserved in all five subgroup species, ∼70Kb downstream from a conserved pair of overlapping genes and ∼25Kb upstream from a conserved seven exon gene. The region between these two gene families is shown in Figure 5.

A number of protein-coding genes are annotated in *D. sechellia* but have no detectable homologs in other species in the subgroup. *D. melanogaster* has a number of annotated ncRNAs, one of which overlaps with parts of *D. sechellia*-specific genes. Since these protein-coding genes are present in only one species, we did not include them in our analysis, because in the absence of conservation, we lack sufficient evidence that they are

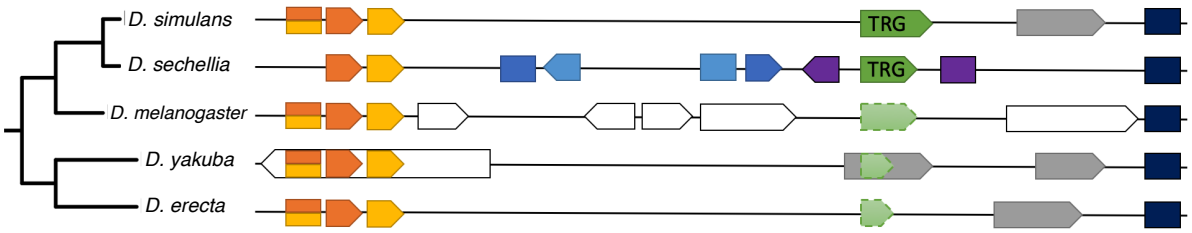

Figure 5: DNA regions homologous to the gene family containing *Dsim_GD20667* and *Dsec_GM19408*. Protein-coding genes are shown in color, pseudogenes in grey and ncRNA genes in white. Homologous protein-coding genes are marked by the same color, each element corresponds to an exon. Only the first of the seven exons of the dark blue gene is shown. Genes shown directly above/below each other share sequence similarity, but we did not infer homology of all pseudogenes and ncRNA in a rigorous way. The direction of the arrow shows which strand the gene is located on. Features with dashed outlines are not annotated. The diagram is not to scale.

functional. The region containing the putative TRGs is annotated as an intron of one of these *D. sechellia* protein-coding genes. The downstream region annotated as a pseudogene in *D. simulans*, *D. yakuba* and *D. erecta*, and as a ncRNA in *D. melanogaster*, is well-conserved in all species. The annotation boundaries vary among species.

The region containing the TRGF is extremely well-conserved in all five species and is annotated as a pseudogene in *D. yakuba*. Using BLASTn for similarity searches to identify the parent gene of this putative pseudogene we were only able to find a self-hit and numerous matches covering <10% of the sequence in all species with an exception of *D. simulans* where we identified a 219 nucleotide long unannotated contig with 97.7% sequence identity. We were unable to find any other evidence about the parent gene of this putative pseudogene, casting doubt on its pseudogenic nature.

The start codon of the putative TRGF is in a different frame than that of the *D. yakuba* putative pseudogene, suggesting that it evolved *de novo* in an alternative frame, but upon closer scrutiny we realised that this is not the case. The start codon of the putative TRGF is flanked by two indels, which brings the frame of the annotated *D. yakuba* pseudogene in frame with the putative TRGF following its annotated start codon (see Figure 6). A *TG*-dinucleotide repeat region in the middle of the putative TRGF ORF appears to be poorly conserved; this could be either because of a genuinely higher mutation/indel rate, or merely because of a poor quality of reads/assembly in this region. The uncertainty created by this region and the fact that the length of the pseudogene is not a multiple of three makes it difficult to infer whether the putative TRGF shares the frame with the pseudogene throughout the whole sequence.

More telling information comes from six stop codons conserved across the five species and located between the repeat region and the stop codon shared by both the pseudogene and the putative TRGF. Four stop codons are in +2 frame of the putative TRGF and 2 stop codons are in the +3 frame, leaving +1 as the only frame of the putative TRGF free from stop codons conserved across the five species. If we assume that the pseudogene was free

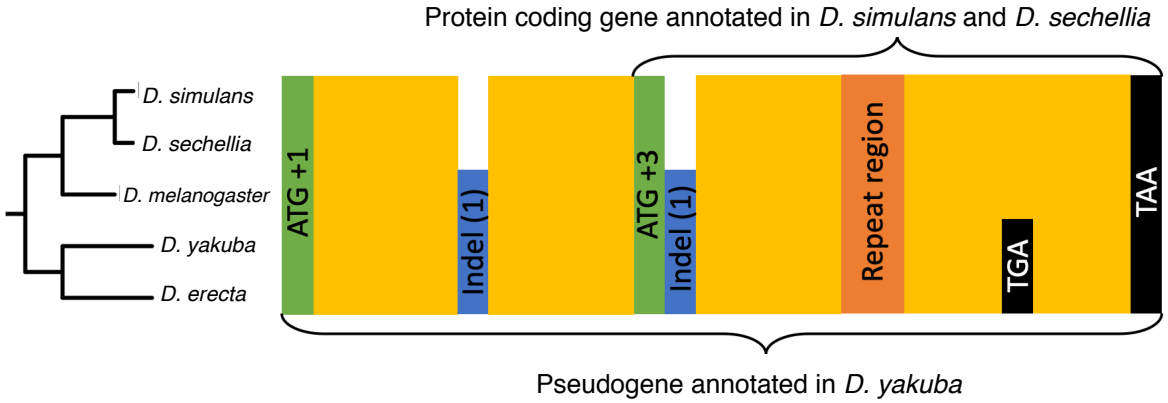

Figure 6: Sequence features of the ancestral ORF, which is annotated as a pseudogene in *D. yakuba*. Start codons of the annotated pseudogene and of the shorter putative TRG in the *simulans-sechellia-melanogaster* clade are in green, well conserved regions in yellow, frame-shift causing indels in blue, repetitive DNA in orange, and stop codons in black. We use the following frame numbering convention: the start codon of the putative pseudogene is denoted the +1 frame, the other two frames on the same strand are denoted +2 and +3 frames. The numbers in parentheses indicate how many more nucleotides (modulo 3) the species it is marked in has. The frames of the stop codons are not marked due to uncertainty about frame created by the repeat region. The two stop codons shown are located in the same frame.

from stop codons when it originated, then this implies that the putative TRGF sequence following the repeat region is in the same frame as the original pseudogene sequence.

The fact that we were unable to identify the ancestral gene associated with this pseudogene, the high level of sequence conservation (95% sequence identity between *D. yakuba* and *D. simulans*, excluding 25 out of 587 nucleotides corresponding to indels), and the more than 110 amino acid long ORF still present in *D. yakuba* all add up to substantial evidence that this *D. yakuba* ORF annotated as a pseudogene might in fact be a mis-annotated functional gene. Regardless of whether this pseudogene is a true remnant of a previously functional gene or a mis-annotated gene that is still functional today, we conclude that the putative TRGF in *simulans-sechellia* did not evolve *de novo*. Instead it evolved from an ancestral protein encompassing all yellow regions shown in Figure 6, via truncation of the N-terminal, whose homologous sequence does not appear to be in the same frame as it is in the rest of the clade.

Discussion

The aim of this study was to identify high confidence TRGFs as most promising candidates for experimental studies of protein-coding genes that emerged in the past 3.3 million years, while avoiding ascertainment biases associated with preconceptions of how *de novo* genes are born. We will only learn about how *de novo* genes are different from well-established genes if we look for them with an open mind and without assumptions that

their sequences must be similar to well-established genes in order to be functional. Unlike other studies, we did not filter out genes with composition distinct from average composition of sequences in protein databases (Vakirlis et al., 2017) nor assume that TRGFs cannot contain splicing signals (Knowles and McLysaght, 2009). To avoid including candidates that are not functional protein-coding genes, without making such assumptions, we used ORF conservation as a proxy for selection and hence evidence for functionality, in addition to using NCBI genome annotations as the most comprehensive synthesis of evidence for transcription and/or translation. To avoid including candidates that were not born *de novo*, we conducted extensive investigation of homologous non-coding sequences in sister species.

We identified a single TRGF with annotated single copy genes in *D. simulans* and *D. sechellia*. This TRGF is located in a syntenic context conserved across all examined species in *Drosophila melanogaster* subgroup. It contains an intron that pre-dates birth as an ORF. Due to the lack of sequence conservation outside the *Drosophila melanogaster* subgroup we were unable to establish whether enabling mutations (indels and substitutions) occurred after divergence of the *simulans-sechellia-melanogaster* clade, or earlier followed by loss in the common ancestor of *D. yakuba* and *D. erecta*. Our results highlight that *de novo* gene studies should under no circumstances exclude candidate TRGs just because they have introns.

The number of *de novo* genes reported in any study depends on the balance of false positives and false negatives that has been achieved by the authors. This is shaped by decisions as to what counts as evidence for functionality and what properties of the candidate genes signify that they are not true *de novo* genes. When we began this study, our requirement that a *de novo* gene must have homologous non-coding DNA sequence(s) in outgroup species as evidence for the time of emergence was stricter than most. Since then two papers have been published that described (Vakirlis and McLysaght, 2018) and applied (Zhang et al., 2019) a similar requirement for homologous non-coding DNA sequences in outgroup species.

Zhang et al. (2019) examined *de novo* genes in the *Oryza* clade and concluded that about 51.5 *de novo* genes per million years are generated and retained in this clade. While care was taken to show *de novo* status, this number is nevertheless likely inflated by lenient criteria for functionality. Intact gene structure and some transcription and translation were considered sufficient, with no requirement for functional evidence or evolutionary conservation. The estimated rate of *de novo* gene birth is also potentially deflated (but not by as much) by the assumption that recent *de novo* genes cannot be present in more than one copy. Another limitation of the study is that only the single best hit to a genome was considered. Since hits were accepted if they covered $\geq$20% of an ORF, this could lead to selecting a short highly similar region (for example, to a low complexity region) and ignoring a longer truly homologous region with a slightly lower match score. Accepting matches that cover as little as 20% of an ORF is contradictory to the idea presented in the paper that indels and substitutions are the main ORF triggers, and may have deflated the estimate.

In contrast, our study, which was designed to identify high-confidence experimental candidates, is likely an underestimate, in part because homologous sequence in orthologs might be missing or unrecognizable, but mostly because it cannot find a TRGF unless it is already present in the NCBI gene annotations of two species. The incomplete nature of genome annotation is more of a problem when a gene must be annotated in two species than when it must merely be annotated in one. Abascal et al. (2018) shows that about 12% of human genes have different annotations across the three most popular databases (RefSeq, Ensembl/GENCODE and UniProtKB), and that some genes that are listed as non-coding actually have more experimental evidence for producing a protein than some genes listed as protein-coding. Even in relatively simple species like *Escherichia coli*, about 35% of the annotated genes lack experimental evidence of function (Ghatak et al., 2019). The annotation quality for the *Drosophila melanogaster* subgroup is unlikely to be better than for the human genome. Nevertheless, we believe that synthesis of evidence from all data sets submitted to NCBI is by far better than the evidence that we could have gathered and synthesised ourselves without performing experimental work.

The availability of evidence for functionality is the limiting factor in identifying very young genes. Without it, short young proteins are often left out of genome annotations, and hence alternative approaches like screening all ORFs present in a genome (Ruiz-Orera et al., 2018) are required to identify them. Given the frequency of premature stop codon mutations, conservation of an ORF across several species can be used as a proxy for functionality, as we do here. However, sufficiently short ORFs can still be conserved by chance sometimes across several species.

One reason we find a lower rate of de novo gene birth might be that false positive evidence of functionality inflates single-species estimates in other studies, whereas false negative failure to reproduce such evidence in two species deflates it in our study. However, it is also possible that both estimates are approximately correct, with the discrepency arising from the fact that rapid emergence of functional ORFs is counter-balanced by rapid loss, as discussed by Schlötterer (2015). Since newborn proteins are not yet integrated in the protein interaction network, they might be relatively dispensable; even if adaptive at first, they might not remain adaptive as the environmental and genetic context changes. In this case our approach, in using evolutionary conservation to exclude non-functional polypeptides, also excludes functional proteins whose functionality is short-lived.

There have been several previous papers aimed at identifying TRGs in *Drosophila melanogaster* subgroup: the pioneering work of Levine et al. (2006) focused on *de novo* genes, followed by a survey of all TRGs (Zhou et al., 2008), a study about essentiality of TRGs (Chen et al., 2010), an in-depth analysis of the evolution and function of six candidate *de novo* genes (Reinhardt et al., 2013), and a study of very young *de novo* genes in *D. melanogaster* that are still segregating in the population (Zhao et al., 2014). These studies collectively reported 16 *de novo* protein-coding genes and two *de novo* ncRNAs (Dme_CR32582, Dmel_CR32690) that are fixed in *D. melanogaster* and not present outside of the *Drosophila melanogaster* subgroup. Three of these protein-coding genes (Dmel_CG33235, Dmel_CG33666, Dmel_CG34434) are present only in *D. melanogaster* and hence were not included in our analysis, and another seven of them (Dmel_CG2042,

Dmel_CG32582, Dmel_CG32690, Dmel_CG32824, Dmel_CG40384, Dmel_CG9284, Dmel_CG32582) have been removed from the genome annotations since the time of publication. For the remaining six previously reported *de novo* protein-coding genes, we were either able to identify homologous genes outside the *Drosophila melanogaster* subgroup (Dmel_CG31882, Dmel_CG30395, Dmel_CG31406, Dmel_CG32712), or we were unable to identify homologous DNA regions in any of the outgroup species (Dmel_CG15323, Dmel_CG31909). Note that these last two could still be *de novo* genes. Here we have identified a TRGF containing *Dsim_GD19764* and *Dsec_GM10790* in *D. simulans* and *D. sechellia* respectively that evolved *de novo*. This TRGF is not present in *D. melanogaster* and hence was not part of these previous studies. We did not identify any TRGFs in this clade that evolved *de novo* and contain an annotated *D. melanogaster* gene.

The most recent study by Heames et al. (2020) identified 32 putative *de novo* TRGFs in *simulans-sechellia-melanogaster* clade. None of these 32 were supported by our analysis. For 25 of them, we identified BLASTp hits outside the *Drosophila melanogaster* subgroup, and for 3 of them we identified conserved but unannotated ORFs in outgroup species. This indicates an earlier origin of these TRGs, as well as emphasizing the importance of these two quality control steps. For 2 of them (one consisting of FBgn0269617 in *D. simulans* and FBgn0169891 in *D. sechellia*, the other consisting of FBgn0268387 in *D. simulans* and FBgn0168374 in *D. sechellia*), the candidate gene in *D. sechellia* was no longer part of the official genome annotations (meaning that we failed to get our minimum of two annotated homologs). While these two might still be genuine TRGFs, we note that poorly assembled genomes contain more spurious genes, and that this is reflected in the relative numbers of singleton TRGs reported by Heames et al. (2020), with 41 in *D. melanogaster*, 251 in *D. simulans*, and 958 in *D. sechellia*. For the remaining two putative *de novo* TRGFs, the two homologs did not meet our length tolerance ratio of 61% of aligned homology : length of shorter protein (see Methods), so that our pipeline did not infer them to be homologous. The gene pair of FBgn0268561 in *D. simulans* and FBgn0266534 in *D. melanogaster* had a ratio of 43.06%, while the gene pair of FBgn0269153 in *D. simulans* and FBgn0267104 in *D. melanogaster* had a ratio of 54.17%. As discussed earlier, the one TRGF that we did identify with high confidence was not found by Heames et al. (2020) because a homologous nucleotide sequence used to have a protein-coding gene annotated on the opposite strand, and this was taken by Heames et al. (2020) to be evidence of origination by divergence instead of *de novo*.

Our results show that while *de novo* genes that are conserved across several species undoubtedly do exist, their number is probably on the lower side of the spectrum of estimates reported in previous studies. We have identified only a single TRGF in the *Drosophila melanogaster* subgroup, which does not allow us to identify a common pattern of emergence of *de novo* genes. High confidence in its annotation as *de novo* and as conserved may make this *de novo* gene the best candidates in *Drosophila melanogaster* subgroup identified so far for the experimental studies needed to drive the field forward.

Materials and Methods

Data

The genome assemblies for *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and *D. erecta* were downloaded from RefSeq (Haft et al., 2017) along with the genome annotations (O'Leary et al., 2015). The completeness of the protein sets was assessed using BUSCO (Waterhouse et al., 2017), using 2799 Hidden Markov Models (HMMs) of single-copy orthologs found in >90% of species in the order Diptera. Table 1 summarizes genome statistics for each species.

Species (RefSeq assembly accession)	Assembly size (Mbp)	Molecule count	N50 (Mbp)	Proteins	BUSCO (%)
D. simulans (GCF_000754195.2)	124.96	7	0.45	14179	98.6%
D. sechellia (GCF_000005215.3)	166.59	1	0.042	16467	92.2%
D. melanogaster (GCF_000001215.4)	116.52	8	19.48	13916	99.3%
D. yakuba (GCF_000005975.2)	165.71	8	0.12	14824	98.4%
D. erecta (GCF_000005135.1)	152.71	0	0.45	13605	99%

Table 1: Genome assemblies of *Drosophila melanogaster* subgroup species used in this study.

Homology predictions

We used OMA v2.2.0 implementation of the OMA algorithm (Altenhoff et al., 2017) with default parameters to infer groups of homologous genes across six genomes: five *Drosophila melanogaster* subgroup species and *D. ananassae* as an outgroup. The length tolerance ratio was set to the default value of 61%, meaning that if the length of the alignment between a putative pair of homologous proteins is less than 61% of the length of the shortest of the two proteins, then no homology was inferred. All the genes annotated as protein-coding in the assemblies described above were used, regardless of their length. We selected orthologous families with genes in at least 2 of the species in the *simulans-sechellia-melanogaster* clade and no genes outside this clade as putative TRGFs for further analysis.

Validation of putative TRGFs

Putative TRGFs were first validated with sequence similarity searches in amino acid space against all non-redundant proteins in the RefSeq database, using BLASTp v2.7.1+ (Camacho et al., 2009) with default parameters. All hits with e-value $\leq 1e-03$ and covering $\geq 50\%$ of the query were considered. If every gene in a putative TRGF had at least one hit to the species outside the clade, the family was removed from further validation. We did not try to identify highly diverged homologs that are beyond detectability with BLASTp using more advanced methods like PSI-BLAST (Schaffer, 2001), HHMER (Eddy, 2011) or HHblits (Remmert et al., 2011) that rely on building a sequence profile. There were two reasons for this. First, given that the protein is only present in two species the resulting sequence profile would not contain much more information than a single sequence and

hence it would be unlikely to yield useful results. Second, we relied on our assumption that if a homologous gene is present in an outgroup genome it would be included in the BLASTn hits against that genome. This assumption doesn't necessarily hold at large evolutionary distances, but for closely related species it would be extremely unlikely to identify a good DNA sequence match covering all of the gene and at the same time to miss a homologous gene that diverged beyond detectable similarity in nucleotide sequence space.

Remaining putative TRGFs were validated with sequence similarity searches in nucleotide space against the five genomes in *Drosophila melanogaster* subgroup, using BLASTn v2.7.1+ (Camacho et al., 2009) with default parameters. We did not use tools like FASTA3 (Pearson, 2000) that take into account synonymous codons or amino acid similarity because the homologous DNA sequences are protein-coding in some species but not the others. BLASTn makes no additional assumptions about the evolutionary constraints specific to the query sequence, and hence is most suitable tool for this problem. For each gene we used both the whole gene sequence and the set of coding sequences (CDSs) as a query. This approach ensures that hits to even very short CDSs are retained, while also using the information in the non-coding parts of the gene when the information contained in a short CDS is insufficient. All hits with e-value $\leq 1e-03$ and covering $\geq 50\%$ of the query (a whole gene or a CDS) were considered, and overlapping hits were amalgamated. In cases where the total number of hits exceeded 1000, we ordered the hits by e-value and selected the five best hits per species. Hits (including self-hits to the genes) were aligned with MAFFT v7.407 (Katoh and Standley, 2013) using E-INS-i algorithm that makes minimal assumptions about the nature of the resulting alignment. We used the "--adjustdirectionaccurately" option to align hits located on different strands and the "--addfragments" option to subsequently add CDSs to the alignment of hits. Alignments were examined manually to remove the hits that were only covering parts of introns or untranslated regions (UTRs) and to extend promising hits that ended in the middle of the gene. After these amendments the remaining/extended hits were realigned and the resulting alignments were examined for presence of homologous ORFs in the *yakuba-erecta* clade. If an ORF was identified in at least one of the two outgroup species, it was considered as evidence that the putative TRGF originated prior to the speciation of the *Drosophila melanogaster* subgroup and the family was removed from further validation. Putative TRGFs that passed sequence similarity validations were manually examined for quality and consistency of annotations.

Inferring the origin of TRGFs

To infer the origins of TRGFs, we extracted genome annotations corresponding to the identified homologous DNA regions in all five species. Synteny conservation in these DNA regions was used as evidence of homology for the less conserved sequences. We also identified homologous DNA regions in four additional species - two in the *melanogaster group* (*D. ananassae* and *D. suzukii*) and two in its sister clade *obscura group* (*D. pseudoobscura* and *D. miranda*). We checked for presence of known protein domains with HMMER v3.1b2 (Eddy, 2011) using Pfam v31 database (Finn et al., 2015). We used the BUSCA web server to predict protein sub-cellular localization (Savojardo et al., 2018),

TANGO to predict protein aggregation (Fernandez-Escamilla et al., 2004), and Wasabi for visualising multiple sequence alignments (Veidenberg et al., 2015). All analysis was performed in Python v3.7.0, using packages biopython v1.73 (Cock et al., 2009) and gffutils v0.9. The code is available at `https://github.com/KarinaZile/TRGs_in_Drosophila_melanogaster_subgroup`.

Author Contributions

KZ and JM designed the project. KZ developed the methods, performed the analysis and wrote the manuscript. JM provided feedback and support throughout all of the stages of this work, and edited the manuscript. CD and YW supervised earlier unpublished work upon which this project was built. All the authors reviewed and approved the final manuscript.

Acknowledgements

KZ was supported by the Biotechnology and Biological Sciences Research Council grant BB/M009513/1. JM was supported by the National Institutes of Health grant GM104040. CD was supported by the Swiss National Science Foundation grant 183723. YW was supported by the Natural Environment Research Council grant NE/L00626X/1. The computations were performed on the Department of Computer Science cluster at University College London.

References

Abascal, F., D. Juan, I. Jungreis, L. Martinez, M. Rigau, J. M. Rodriguez, J. Vazquez, and M. L. Tress
2018. Loose ends: almost one in five human genes still have unresolved coding status. *Nucleic Acids Research*, 46(14):7070–7084.

Albertin, C. B., O. Simakov, T. Mitros, Z. Y. Wang, J. R. Pungor, E. Edsinger-Gonzales, S. Brenner, C. W. Ragsdale, and D. S. Rokhsar
2015. The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature*, 524(7564):220–224.

Altenhoff, A. M., N. M. Glover, C.-M. Train, K. Kaleb, A. W. Vesztrocy, D. Dylus, T. M. de Farias, K. Zile, C. Stevenson, J. Long, H. Redestig, G. H. Gonnet, and C. Dessimoz
2017. The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Research*, 46(D1):D477–D485.

Ángyán, A. F., A. Perczel, and Z. Gáspári
2012. Estimating intrinsic structural preferences of de novo emerging random-sequence proteins: Is aggregation the main bottleneck? *FEBS Letters*, 586(16):2468–2472.

Arendsee, Z., J. Li, U. Singh, P. Bhandary, A. Seetharam, and E. S. Wurtele
2019. fagin: synteny-based phylostratigraphy and finer classification of young genes. *BMC Bioinformatics*, 20(1):440.

Baalsrud, H. T., O. K. Tørresen, M. H. Solbakken, W. Salzburger, R. Hanel, K. S. Jakobsen, and S. Jentoft
2017. De novo gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data. *Molecular Biology and Evolution*, 35(3):593–606.

Balakirev, E. S. and F. J. Ayala
2003. Pseudogenes: Are they "junk" or functional DNA? *Annual Review of Genetics*, 37(1):123–151.

Blevins, W., M. Albà, and L. Carey
2017. Comparative transcriptomics and ribo-seq: Looking at de novo gene emergence in saccharomycotina. *PeerJ*, P. 3030.

Brosius, J.
2005. Waste not, want not – transcript excess in multicellular eukaryotes. *Trends in Genetics*, 21(5):287–288.

Bungard, D., J. S. Copple, J. Yan, J. J. Chhun, V. K. Kumirov, S. G. Foy, J. Masel, V. H. Wysocki, and M. H. Cordes
2017. Foldability of a natural de novo evolved protein. *Structure*, 25(11):1687–1696.e4.

Cai, J., R. Zhao, H. Jiang, and W. Wang
2008. De novo origination of a new protein-coding gene in saccharomyces cerevisiae. *Genetics*, 179(1):487–496.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden
2009. BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421.

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.

Chen, S., Y. E. Zhang, and M. Long
2010. New genes in drosophila quickly become essential. *Science*, 330(6011):1682–1685.

Chen, X., S. Jung, L. Y. Beh, S. R. Eddy, and L. F. Landweber
2015. Combinatorial DNA rearrangement facilitates the origin of new genes in ciliates. *Genome Biology and Evolution*, P. evv172.

Cock, P. J. A., T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon
2009. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.

Domazet-Loso, T.
2003. An evolutionary analysis of orphan genes in drosophila. *Genome Research*, 13(10):2213–2219.

Donoghue, M. T., C. Keshavaiah, S. H. Swamidatta, and C. Spillane
2011. Evolutionary origins of brassicaceae specific genes in arabidopsis thaliana. *BMC Evolutionary Biology*, 11(1).

Drosophila 12 Genomes Consortium
2007. Evolution of genes and genomes on the drosophila phylogeny. *Nature*, 450(7167):203–218.

Eddy, S. R.
2011. Accelerated profile HMM searches. *PLoS Computational Biology*, 7(10):e1002195.

Fernandez-Escamilla, A.-M., F. Rousseau, J. Schymkowitz, and L. Serrano
2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature Biotechnology*, 22(10):1302–1306.

Finn, R. D., P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, and A. Bateman
2015. The pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44(D1):D279–D285.

Frigola, J., R. Sabarinathan, L. Mularoni, F. Muiños, A. Gonzalez-Perez, and N. López-Bigas
2017. Reduced mutation rate in exons due to differential mismatch repair. *Nature Genetics*, 49(12):1684–1692.

Ghatak, S., Z. A. King, A. Sastry, and B. O. Palsson
2019. The y-ome defines the 35% of escherichia coli genes that lack experimental evidence of function. *Nucleic Acids Research*, 47(5):2446–2454.

Graur, D., Y. Zheng, N. Price, R. B. R. Azevedo, R. A. Zufall, and E. Elhaik
2013. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biology and Evolution*, 5(3):578–590.

Guan, Y., L. Liu, Q. Wang, J. Zhao, P. Li, J. Hu, Z. Yang, M. P. Running, H. Sun, and J. Huang
2018. Gene refashioning through innovative shifting of reading frames in mosses. *Nature Communications*, 9(1).

Guerzoni, D. and A. McLysaght
2016. De novo genes arise at a slow but steady rate along the primate lineage and have been subject to incomplete lineage sorting. *Genome Biology and Evolution*, 8(4):1222–1232.

Haft, D. H., M. DiCuccio, A. Badretdin, V. Brover, V. Chetvernin, K. O'Neill, W. Li, F. Chitsaz, M. K. Derbyshire, N. R. Gonzales, M. Gwadz, F. Lu, G. H. Marchler, J. S. Song, N. Thanki, R. A. Yamashita, C. Zheng, F. Thibaud-Nissen, L. Y. Geer, A. Marchler-Bauer, and K. D. Pruitt
2017. RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Research*, 46(D1):D851–D860.

Hashimoto, T., D. D. Horikawa, Y. Saito, H. Kuwahara, H. Kozuka-Hata, T. Shin-I, Y. Minakuchi, K. Ohishi, A. Motoyama, T. Aizu, A. Enomoto, K. Kondo, S. Tanaka, Y. Hara, S. Koshikawa, H. Sagara, T. Miura, S. ichi Yokobori, K. Miyagawa, Y. Suzuki, T. Kubo, M. Oyama, Y. Kohara, A. Fujiyama, K. Arakawa, T. Katayama, A. Toyoda, and T. Kunieda
2016. Extremotolerant tardigrade genome and improved radiotolerance of human cultured cells by tardigrade-unique protein. *Nature Communications*, 7(1):12808.

Heames, B., J. Schmitz, and E. Bornberg-Bauer
2020. A continuum of evolving de novo genes drives protein-coding novelty in drosophila. *Journal of Molecular Evolution*, 88(4):382–398.

Hild, M., B. Beckmann, S. Haas, B. Koch, V. Solovyev, C. Busold, K. Fellenberg, M. Boutros, M. Vingron, F. Sauer, J. Hoheisel, and R. Paro
2003. An integrated gene annotation and transcriptional profiling approach towards the full gene content of the drosophila genome. *Genome Biology*, 5(1):R3.

Jain, A., D. Perisa, F. Fliedner, A. von Haeseler, and I. Ebersberger
2019. The evolutionary traceability of a protein. *Genome Biology and Evolution*, 11(2):531–545.

Katoh, K. and D. M. Standley
2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780.

Knopp, M. and D. I. Andersson
2018. No beneficial fitness effects of random peptides. *Nature Ecology & Evolution*, 2(7):1046–1047.

Knopp, M., J. S. Gudmundsdottir, T. Nilsson, F. König, O. Warsi, F. Rajer, P. Ädelroth, and D. I. Andersson
2019. De novo emergence of peptides that confer antibiotic resistance. *mBio*, 10(3).

Knowles, D. G. and A. McLysaght
2009. Recent de novo origin of human protein-coding genes. *Genome Research*, 19(10):1752–1759.

Levine, M. T., C. D. Jones, A. D. Kern, H. A. Lindfors, and D. J. Begun
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.

Linquist, S., W. F. Doolittle, and A. F. Palazzo
2020. Getting clear about the f-word in genomics. *PLOS Genetics*, 16(4):e1008702.

McLysaght, A. and D. Guerzoni
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.

McLysaght, A. and L. D. Hurst
2016. Open questions in the study of de novo genes: what, how and why. *Nature Reviews Genetics*, 17(9):567–578.

Neme, R., C. Amador, B. Yildirim, E. McConnell, and D. Tautz
2017. Random sequences are an abundant source of bioactive RNAs or peptides. *Nature Ecology & Evolution*, 1(6):0127.

Neme, R. and D. Tautz
2016. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *eLife*, 5:09977.

Obbard, D. J., J. Maclennan, K.-W. Kim, A. Rambaut, P. M. O'Grady, and F. M. Jiggins
2012. Estimating divergence dates and substitution rates in the drosophila phylogeny. *Molecular Biology and Evolution*, 29(11):3459–3473.

O'Leary, N. A., M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and

K. D. Pruitt
2015. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745.

Pearson, W. R.
2000. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.*, 132:185–219.

Reinhardt, J. A., B. M. Wanjiru, A. T. Brant, P. Saelao, D. J. Begun, and C. D. Jones
2013. De novo ORFs in drosophila are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLoS Genetics*, 9(10):e1003860.

Remmert, M., A. Biegert, A. Hauser, and J. Söding
2011. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, 9(2):173–175.

Ruiz-Orera, J., J. Hernandez-Rodriguez, C. Chiva, E. Sabidó, I. Kondova, R. Bontrop, T. Marqués-Bonet, and M. Albà
2015. Origins of de novo genes in human and chimpanzee. *PLOS Genetics*, 11(12):e1005721.

Ruiz-Orera, J., P. Verdaguer-Grau, J. L. Villanueva-Cañas, X. Messeguer, and M. M. Albà
2018. Translation of neutrally evolving peptides provides a basis for de novo gene evolution. *Nature Ecology & Evolution*, 2(5):890–896.

Savojardo, C., P. L. Martelli, P. Fariselli, G. Profiti, and R. Casadio
2018. BUSCA: an integrative web server to predict subcellular localization of proteins. *Nucleic Acids Research*, 46(W1):W459–W466.

Schaffer, A. A.
2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research*, 29(14):2994–3005.

Schlötterer, C.
2015. Genes from scratch – the evolutionary fate of de novo genes. *Trends in Genetics*, 31(4):215–219.

Stewart, N. B. and R. L. Rogers
2019. Chromosomal rearrangements as a source of new gene formation in drosophila yakuba. *PLOS Genetics*, 15(9):e1008314.

Sun, W., X.-W. Zhao, and Z. Zhang
2015. Identification and evolution of the orphan genes in the domestic silkworm, bombyx mori. *FEBS Letters*, 589(19PartB):2731–2738.

Toll-Riera, M., N. Bosch, N. Bellora, R. Castelo, L. Armengol, X. Estivill, and M. M. Alba
2008. Origin of primate orphan genes: A comparative genomics approach. *Molecular Biology and Evolution*, 26(3):603–612.

Tretyachenko, V., J. Vymětal, L. Bednárová, V. Kopecký, K. Hofbauerová, H. Jindrová, M. Hubálek, R. Souček, J. Konvalinka, J. Vondrášek, and K. Hlouchová
2017. Random protein sequences can form defined secondary structures and are well-tolerated in vivo. *Scientific Reports*, 7(1):15449.

Vakirlis, N., A.-R. Carvunis, and A. McLysaght
2020. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife*, 9.

Vakirlis, N., A. S. Hebert, D. A. Opulente, G. Achaz, C. T. Hittinger, G. Fischer, J. J. Coon, and I. Lafontaine
2017. A molecular portrait of de novo genes in yeasts. *Molecular Biology and Evolution*, 35(3):631–645.

Vakirlis, N. and A. McLysaght
2018. Computational prediction of de novo emerged protein-coding genes. In *Methods in Molecular Biology*, Pp. 63–81. Springer New York.

Veidenberg, A., A. Medlar, and A. Löytynoja
2015. Wasabi: An integrated platform for evolutionary sequence analysis and data visualization. *Molecular Biology and Evolution*, 33(4):1126–1130.

Waterhouse, R. M., M. Seppey, F. A. Simão, M. Manni, P. Ioannidis, G. Klioutchnikov, E. V. Kriventseva, and E. M. Zdobnov
2017. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, 35(3):543–548.

Weisman, C. M. and S. R. Eddy
2017. Gene evolution: Getting something from nothing. *Current Biology*, 27(13):R661–R663.

Weisman, C. M., A. W. Murray, and S. R. Eddy
2020. Many but not all lineage-specific genes can be explained by homology detection failure. *bioRxiv*.

Werren, J. H., S. Richards, C. A. Desjardins, O. Niehuis, J. Gadau, J. K. Colbourne, L. W. Beukeboom, C. Desplan, C. G. Elsik, C. J. P. Grimmelikhuijzen, P. Kitts, J. A. Lynch, T. Murphy, D. C. S. G. Oliveira, C. D. Smith, L. v. d. Zande, K. C. Worley, E. M. Zdobnov, M. Aerts, S. Albert, V. H. Anaya, J. M. Anzola, A. R. Barchuk, S. K. Behura, A. N. Bera, M. R. Berenbaum, R. C. Bertossa, M. M. G. Bitondi, S. R. Bordenstein, P. Bork, E. Bornberg-Bauer, M. Brunain, G. Cazzamali, L. Chaboub, J. Chacko, D. Chavez, C. P. Childers, J.-H. Choi, M. E. Clark, C. Claudianos, R. A. Clinton, A. G. Cree, A. S. Cristino, P. M. Dang, A. C. Darby, D. C. de Graaf, B. Devreese, H. H. Dinh, R. Edwards, N. Elango, E. Elhaik, O. Ermolaeva, J. D. Evans, S. Foret, G. R. Fowler, D. Gerlach, J. D. Gibson, D. G. Gilbert, D. Graur, S. Grunder, D. E. Hagen, Y. Han, F. Hauser, D. Hultmark, H. C. Hunter, G. D. D. Hurst, S. N. Jhangian, H. Jiang, R. M. Johnson, A. K. Jones, T. Junier, T. Kadowaki, A. Kamping, Y. Kapustin, B. Kechavarzi, J. Kim, J. Kim, B. Kiryutin, T. Koevoets, C. L. Kovar, E. V. Kriventseva, R. Kucharski, H. Lee, S. L. Lee, K. Lees, L. R. Lewis, D. W. Loehlin,

J. M. Logsdon, J. A. Lopez, R. J. Lozado, D. Maglott, R. Maleszka, A. Mayampurath, D. J. Mazur, M. A. McClure, A. D. Moore, M. B. Morgan, J. Muller, M. C. Munoz-Torres, D. M. Muzny, L. V. Nazareth, S. Neupert, N. B. Nguyen, F. M. F. Nunes, J. G. Oakeshott, G. O. Okwuonu, B. A. Pannebakker, V. R. Pejaver, Z. Peng, S. C. Pratt, R. Predel, L.-L. Pu, H. Ranson, R. Raychoudhury, A. Rechtsteiner, J. G. Reid, M. Riddle, J. Romero-Severson, M. Rosenberg, T. B. Sackton, D. B. Sattelle, H. Schluns, T. Schmitt, M. Schneider, A. Schuler, A. M. Schurko, D. M. Shuker, Z. L. P. Simoes, S. Sinha, Z. Smith, A. Souvorov, A. Springauf, E. Stafflinger, D. E. Stage, M. Stanke, Y. Tanaka, A. Telschow, C. Trent, S. Vattathil, L. Viljakainen, K. W. Wanner, R. M. Waterhouse, J. B. Whitfield, T. E. Wilkes, M. Williamson, J. H. Willis, F. Wolschin, S. Wyder, T. Yamada, S. V. Yi, C. N. Zecher, L. Zhang, and R. A. G. and
2010. Functional and evolutionary insights from the genomes of three parasitoid nasonia species. *Science*, 327(5963):343–348.

Willis, S. and J. Masel
2018. Gene birth contributes to structural disorder encoded by overlapping genes. *Genetics*, 210(1):303–313.

Wilson, B. A., S. G. Foy, R. Neme, and J. Masel
2017. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nature Ecology & Evolution*, 1(6):0146.

Wilson, B. A. and J. Masel
2011. Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biology and Evolution*, 3:1245–1252.

Wissler, L., J. Gadau, D. F. Simola, M. Helmkampf, and E. Bornberg-Bauer
2013. Mechanisms and dynamics of orphan gene emergence in insect genomes. *Genome Biology and Evolution*, 5(2):439–455.

Wu, B. and A. Knudson
2018. Tracing the de novo origin of protein-coding genes in yeast. *mBio*, 9(4):0102418.

Yang, Z. and J. Huang
2011. De novo origin of new genes with introns inPlasmodium vivax. *FEBS Letters*, 585(4):641–644.

Zhang, L., Y. Ren, T. Yang, G. Li, J. Chen, A. R. Gschwend, Y. Yu, G. Hou, J. Zi, R. Zhou, B. Wen, J. Zhang, K. Chougule, M. Wang, D. Copetti, Z. Peng, C. Zhang, Y. Zhang, Y. Ouyang, R. A. Wing, S. Liu, and M. Long
2019. Rapid evolution of protein diversity by de novo origination in oryza. *Nature Ecology & Evolution*, 3(4):679–690.

Zhao, L., P. Saelao, C. D. Jones, and D. J. Begun
2014. Origin and spread of de novo genes in drosophila melanogaster populations. *Science*, 343(6172):769–772.

Zhou, Q., G. Zhang, Y. Zhang, S. Xu, R. Zhao, Z. Zhan, X. Li, Y. Ding, S. Yang, and W. Wang
2008. On the origin of new genes in drosophila. *Genome Research*, 18(9):1446–1455.