Heredity (2020) 125:50–59
https://doi.org/10.1038/s41437-020-0319-7

the geneticssociety

ARTICLE

Evolution of novel genes in three-spined stickleback populations

Jonathan F. Schmitz[1] · Frédéric J. J. Chain[2] · Erich Bornberg-Bauer[1]

Received: 24 November 2019 / Revised: 27 April 2020 / Accepted: 30 April 2020 / Published online: 4 June 2020

Abstract

Eukaryotic genomes frequently acquire new protein-coding genes which may significantly impact an organism's fitness. Novel genes can be created, for example, by duplication of large genomic regions or de novo, from previously non-coding DNA. Either way, creation of a novel transcript is an essential early step during novel gene emergence. Most studies on the gain-and-loss dynamics of novel genes so far have compared genomes between species, constraining analyses to genes that have remained fixed over long time scales. However, the importance of novel genes for rapid adaptation among populations has recently been shown. Therefore, since little is known about the evolutionary dynamics of transcripts across natural populations, we here study transcriptomes from several tissues and nine geographically distinct populations of an ecological model species, the three-spined stickleback. Our findings suggest that novel genes typically start out as transcripts with low expression and high tissue specificity. Early expression regulation appears to be mediated by gene-body methylation. Although most new and narrowly expressed genes are rapidly lost, those that survive and subsequently spread through populations tend to gain broader and higher expression levels. The properties of the encoded proteins, such as disorder and aggregation propensity, hardly change. Correspondingly, young novel genes are not preferentially under positive selection but older novel genes more often overlap with F_{ST} outlier regions. Taken together, expression of the surviving novel genes is rapidly regulated, probably via epigenetic mechanisms, while structural properties of encoded proteins are non-debilitating and might only change much later.

Introduction

Several studies over the last decades have demonstrated that genomes evolve rapidly, generating abundant genetic diversity at the level of populations (Zhao et al. 2014; Durand et al. 2019; Witt et al. 2019). Gene content and gene order can strongly differ between populations and even between individuals within populations.

For a long time, gene duplication was seen as the only important mode of gene emergence (Long et al. 2003). Gene duplication is an attractive model as it immediately explains the functional potential of the novel sequence. The process often starts with the duplication of a DNA sequence. Such genomic rearrangements first appear as copy number variations (CNVs) at the population level. CNVs are the result of duplications or deletions of genomic regions among individuals that can include the duplication (or multiplication) of genes (Katju and Bergthorsson 2013; Chain et al. 2014). On a short time scale, gene duplications (emerging as CNVs in populations) can lead to expression changes that have an adaptive benefit. Indeed, differences in gene copy numbers between populations can lead to differential gene expression consistent with local adaptation (Huang et al. 2019). Such fitness advantage, however, is not the most prevalent consequence of duplication, but rather expression attenuation of either of the gene copies, or gene silencing and loss (Tautz and Domazet-Lošo 2011; De Smet et al. 2013). Expression changes, foremost attenuation, have been attributed to either *cis*-regulatory changes (Huang et al. 2019) or epigenetic regulation (Keller and Yi 2014; Wang et al. 2017), such as gene-body methylation.

In addition to the duplication of existing genes, novel genes can also emerge from non-coding sequences, i.e., de

Associate Editor: Bastiaan Star

Supplementary information The online version of this article (https://doi.org/10.1038/s41437-020-0319-7) contains supplementary material, which is available to authorised users.

✉ Jonathan F. Schmitz
jonathan.schmitz@uni-muenster.de

[1] Institute for Evolution and Biodiversity, University of Münster, Münster, Germany

[2] Department of Biological Sciences, University of Massachusetts, Lowell, MA, USA

novo. The definition of de novo sensu stricto only includes genisation of intergenic sequences. However, novel genes can also emerge from non-coding genic sequences such as introns (Ruiz-Orera et al. 2015; Prabh and Rödelsperger 2019). The starting point of this process is frequently the spurious expression of large intergenic regions in eukaryotic genomes at low levels (Durand et al. 2019; Witt et al. 2019; Carvunis et al. 2012; Neme and Tautz 2016; Nagalakshmi et al. 2008; Kapranov and Laurent 2012). While most intergenic transcripts do not have any significant function, i.e. one that is selected for or physiologically beneficial, some transcripts have been shown to become fixed and overlap with novel ORFs, which are also randomly acquired (Ruiz-Orera et al. 2014; Schmitz et al. 2018). As a consequence, some transcripts are prone to become either functional RNAs (Heinen et al. 2009; Mercer et al. 2009) or, if translated as well, they become expressed as proteins. Indeed, many transcripts also contain ORFs that are translated into short proteins (Wilson and Masel 2011; Vanderperre et al. 2013), thus exposing the encoded ORFs to selection (Chen et al. 2015; Xie et al. 2012; Zhang et al. 2019). It is to date unclear how often this transition from intergenic to expressed non-coding and further to coding sequences (i.e., de novo gene emergence) happens. In addition, the relative prevalence of transcription first vs. ORF first in this process has not been determined yet. One recent study proposes de novo emergence to be the dominant mechanism of novel gene emergence (Vakirlis et al. 2019). In some cases, de novo genes can emerge and contribute to increased fitness, as they can soon become essential (Zhang et al. 2019), e.g., for reproduction (Gubala et al. 2017). In the case of gene duplication, novel transcripts emerge from duplicated genes that diverge beyond the limits of homology detection. During de novo gene birth, transcripts emerge from non-coding sequences that as such do not have homologous sequences amongst protein-coding genes. In either case, rapid loss of de novo or duplicated "novel" genes is the most likely immediate outcome (Tautz and Domazet-Lošo 2011).

Other proposed mechanisms of how novel, though not strictly de novo, genes are created include the alternative (same of complementary strand) transcription of a CDS (Prabh and Rödelsperger 2019; Van Oss and Carvunis 2019), overprinting (i.e., the reuse of an existing ORF in an alternative reading frame) (Sabath et al. 2012) or the partial extension of reading frames (Bornberg-Bauer et al. 2015; Klasberg et al. 2018; Toll-Riera and Albà 2013).

Transcripts offer a valuable insight into gene emergence as they can be easily verified (in comparison to proteome studies) and do not rely on the difficult processes of gene annotation, which often also relies on homology signals. Consequently, studying novel transcripts allows analysis of the earliest stages of new gene emergence. Only a small fraction of transcripts is expected to survive the purifying effects of selection and eventually become fixed as a new gene. While the term "gene" itself is undergoing frequent semantic changes (Gerstein et al. 2007; Keeling et al. 2019), we here refer to it as a genetic segment transcribing a transcript for which evidence of translation exists, preferably but not necessarily experimentally verified.

New genes have repeatedly been suggested to provide strong adaptive benefits, especially in an ecological context (Chain et al. 2014; Zhang et al. 2019; Khalturin et al. 2009; Kumar et al. 2015). Accordingly, the early stages of gene emergence are of particular interest and need to be further investigated at the level of populations. This is especially important considering that most research on novel genes in general and de novo genes in particular remains controversial. First, signals of the selection of the encoded proteins have been observed in most (Chen et al. 2015; Zhang et al. 2019; Gubala et al. 2017; Palmieri et al. 2014), but not all (Guerzoni and McLysaght 2016) studies. Second, these proteins have been claimed to undergo rapid changes in structural properties which are deemed essential for functioning such as aggregation propensities, size and disorder content in some (Palmieri et al. 2014; Wilson et al. 2017), but not in other studies (Schmitz et al. 2018). In addition, all studies so far have compared genomes of different species and therefore consider evolution over relatively long time scales (i.e., several millions of years and longer).

Another understudied area is the effect and regulation of gene emergence in natural populations. So far, most studies on expression—and all focusing on novel genes—have been conducted in model species (Zhao et al. 2014). Most model species have not been under natural selection for many generations and may be subject to adaptation to laboratory conditions. A well-suited ecological model species is the three-spined stickleback, which is known to undergo rapid adaptation to many environmental conditions across the northern hemisphere (Foster and Bell 1994; McKinnon and Rundle 2002). This coastal fish has extensive genomic (Jones et al. 2012; Roesti et al. 2013; Glazer et al. 2015; Feulner et al. 2013, 2015) and transcriptomic data available (Feulner et al. 2013; Huang et al. 2016; Hanson et al. 2017; Metzger and Schulte 2018) and recent studies have demonstrated patterns of differences in CNVs between populations (Chain et al. 2014; Huang et al. 2019; Hirase et al. 2014), which make the system amenable to study the emergence and disappearance of novel genes.

Here, we take advantage of the genomic and transcriptomic datasets available for three-spined sticklebacks to study the distribution of transcripts between populations. Accordingly, we analysed the emergence and spread of new genes expressed in four tissues across nine populations of the three-spined stickleback. Gene expression levels and tissue specificity were compared across genes of different

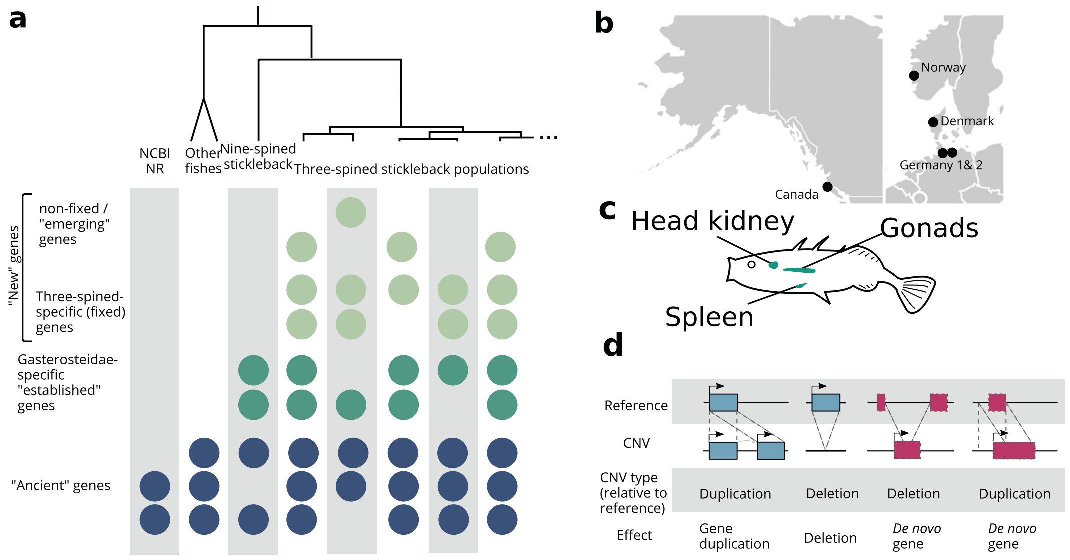

Fig. 1 Illustrations of the gene age and visual presentation of data sources. a Examples for how genes of different ages are distributed across the analysed species. Each circle represents one gene being present in one species. Each gene is displayed on a separate row. Genes in the same row represent orthologs. Colours are used to distinguish age classes. **b** Locations of the population pairs the data was sampled from. **c** Visual representation of the organs sampled. **d** Pictogram depicting how CNV events can lead to gene duplication, deletion or de novo gene emergence.

ages, finding older genes to be expressed more strongly and broadly. We also analysed the overlap of genes with CNVs and found younger new genes to overlap with CNVs more often than older new genes, showing how genomic changes facilitate new gene emergence in populations.

Results and discussion

Transcriptomes were sampled from multiple stickleback individuals from nine different populations taken from lake, river, and marine ecosystems in Europe and North America (as described in Chain et al. 2014 and Feulner et al. 2013, 2015, see also Supplementary Table 1 and Fig. 1). Four tissues were included in the analysis: the head kidney and spleen, as major immunity-associated tissues, and ovaries and testes, as reproductive tissues. Overall, after removing low-quality or outlier transcriptomes, we made use of 93 transcriptomes. The sampling design thus allows to analyse expressed sequences across an array of tissues and along a phylogeographic gradient (see Fig. 1 and Chain et al. 2014 for details). This analysis allows us to closely study the emergence of novel sequences.

To relate the properties of genes to their age, stickleback genes were grouped into one of three age groups. First, all genes with BLASTP hits with an E-value below 10^{-3} outside of *Gasterosteidae* or Pfam domains were categorised as "ancient" (see also Supplementary Fig. S1). For this homology search we used the NCBI NR database. Second, genes with TBLASTN hits in the nine-spined stickleback's transcriptome were categorised as "established", because they exist in at least two stickleback species and are *Gasterosteidae* specific. For this purpose we used the nine-spined stickleback's transcriptome from the available brain and liver samples (Guo et al. 2013). Genes without hits in any of these homology searches represent recently emerged novel genes and where categorised as "new". If they were present in the genome and all populations (at least once), i.e., they were sub-categorised as "fixed", whereas they were sub-categorised as "emerging" if the transcripts were found in fewer populations. Clearly the line between "fixed" and "emerging" is blurred, because some transcripts might not have been observed in all populations due to low expression under the sampled natural conditions. However, we expect this to be minimal since we include several individual transcriptomes per population and apply a relatively stringent and widely used minimum threshold to consider a transcript as expressed (see "Methods" section). We found 11,413 ancient, 245 established and 991 new genes being expressed across all of our transcriptomes. More than half of the ancient genes are annotated, while only a small fraction (<5%) of the established and new genes is annotated in the three-spined stickleback's Ensembl annotation (Supplementary Fig. S3).

Younger genes are expressed less broadly and at lower levels

To assess the expression patterns of new genes, we analysed the distribution of their expression across all 93

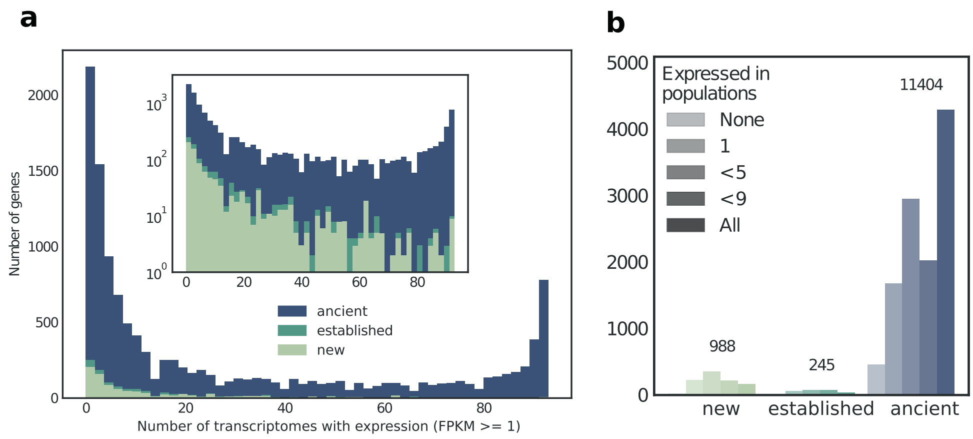

Fig. 2 Distribution of transcript frequencies across transcriptomes and populations. a Number of transcriptomes each gene was found to be transcribed in shown as a stacked bar plot. Inset: same plot with log-scaled *y*-axis. Genes with expression levels >1FPKM were counted as expressed. **b** Numbers of genes found to be expressed in varying numbers of populations.

transcriptomes (Fig. 2a). Out of the 12,637 genes we found expressed in the three-spined stickleback, 2187 (17%) were not found to be transcribed at a level above 1 FPKM (fragments per kilobase of transcript per million mapped reads) in more than one of the transcriptomes. Such low expression levels cannot be properly distinguished from transcriptional noise and were counted as not expressed (Zhao et al. 2014; Chen et al. 2015; Ramsköld et al. 2009). On the other hand, many—especially the "ancient" genes—are expressed above 1 FPKM in more than 90% of the transcriptomes (1708 genes, 14%). In comparison, only 2% of the "new" and "established" genes are found in more than 90% of all the samples, significantly fewer than ancient genes ($p < 10^{-30}$, chi-square).

Among the nine populations, more than one-third (38%) of all "ancient" genes are found to be expressed in all populations, while only 17% of "new" genes are found in all populations ($p < 10^{-37}$, chi-square; Fig. 2b, c). These findings show that younger genes are more often restricted to fewer transcriptomes and are population specific in their expression. This finding also supports previous findings on the unstable transcription of new genes (Li et al. 2019). One caveat here is that we only surveyed expression in four tissues. Consequently, it is possible that new genes are expressed in further populations but were not picked up. This question should be addressed in future, even broader studies.

We also evaluated the expression level of genes according to their age and prevalence among populations to analyse how new emerging genes spread through populations (Fig. 3a). We find that younger genes generally exhibit lower levels of expression than older genes. "New" genes show the lowest average expression levels, followed by "established" genes with intermediate expression levels, while "ancient" genes have the highest expression levels ($p < 10^{-56}$). We also found highly similar results for expression specificity, with older genes being more broadly expressed compared with younger genes (Supplementary Fig. S6).

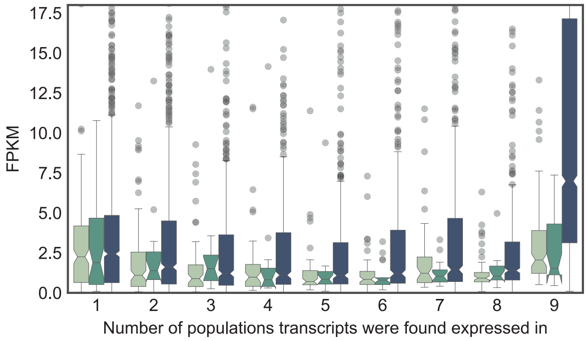

Fig. 3 Box plot of expression strength (in FPKM) of genes by the number of populations the gene was found expressed in. Outliers beyond 17.5 are not shown.

Our findings show that most new genes at first gain low and unstable, i.e. variable between individuals and populations, levels of transcription in single populations (Fig. 3). Such unstable expression is likely to be lost quickly again in a turnover of transcription (Neme and Tautz 2016). Over time, some of the "new" genes appear to spread through populations, i.e., "established" and "ancient" genes are expressed in more populations than "new" genes. This spread occurs in a process that is either biased towards new genes with already high expression levels, or that leads to an increase in expression levels and breadth. These findings are in line with previous studies that have found a high turnover of spurious transcription shortly after de novo gene emergence (Neme and Tautz 2016), and that new genes are spread gradually through populations (Zhao et al. 2014).

Gene-body methylation is a mechanism that could act to rapidly regulate the expression patterns of newly emerged genes (Zemach et al. 2010), similarly to how duplicate genes are regulated (Wang et al. 2015). Gene-body methylation has been shown to be a rapid mechanism of adaptation of expression strength, enabling adaptation to ecological factors (Rando and Verstrepen 2007; Huang et al. 2017). We find the two younger gene classes

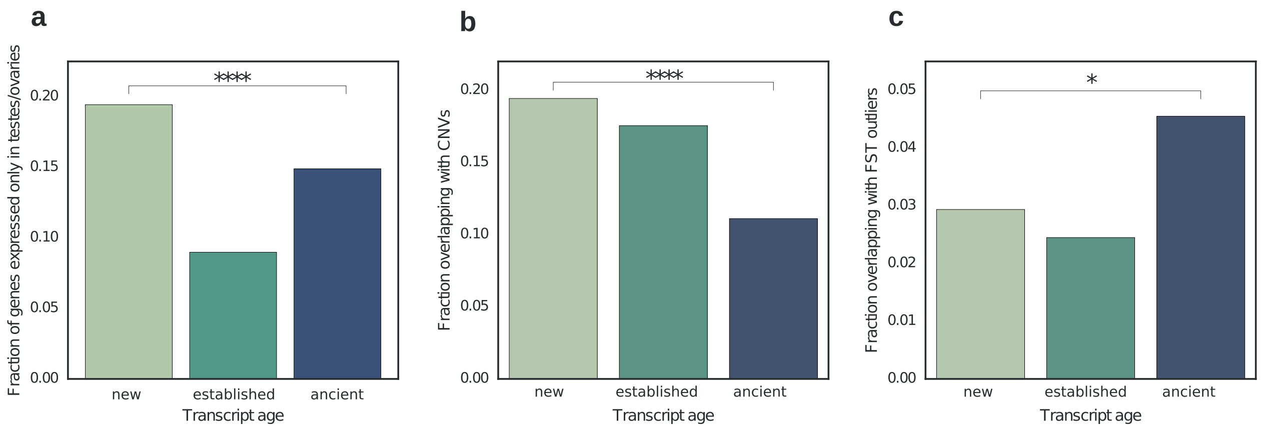

Fig. 4 Differences in gene properties across age classes. a Fraction of genes expressed only in testes or ovaries. The *p* value shown here is the result of a 2 × 2 chi-square test. **b** Fraction of genes overlapping with CNVs. The *p* value shown here is the result of a 2 × 2 chi-square test between the "new" and "ancient" classes. $^{****}p < 0.0001$. **c** Fraction of genes overlapping with regions with F_{ST} indicating positive selection. The *p* value shown here is the result of a 2 × 2 chi-square test between the "new" and "ancient" classes. $^{*}p < 0.05$.

(established and new) to be less CpG depleted compared with older genes (Wilcoxon rank sum tests: $p < 10^{-7}$). Established genes show the strongest CpG depletion. All genes have lower CpG enrichment than intergenic sequences (Supplementary Fig. S4). These findings show that gene-body methylation could play an important role in novel gene fixation, for example, as a first, fast way of regulation after new gene emergence.

Younger genes are more often gonad specific

Our results show that new genes are expressed at lower levels. Previously, new (duplicated and novel) genes have also been suggested to often be gonad biased (Kaessmann 2010; Cui et al. 2014). However, the expression levels of new genes have previously not been tested at a population level. Here, we compared the number of gonad-specific genes between age groups (Fig. 4a) to study the long suspected over-representation of testis bias in new gene emergence. Consistent with previous reports (Cui et al. 2014; Wu et al. 2014; Tobler et al. 2017), we find that "new" genes are 30% more likely to be expressed only in testes or ovaries compared with "ancient" genes ($p < 0.001$, 2 × 2 chi-square test). One could expect this bias to be caused by a lower number of reads required to reach the FPKM threshold to be counted as expressed, however, this is not the case here (see Supplementary Fig. S7).

This gonad specificity of new and established genes aligns with the hypothesis that gonads play an important role in new gene emergence (Kaessmann 2010; Cui et al. 2014). For example, in *Xenopus tropicalis*, young duplicated genes and novel genes are predominantly expressed in one sex and one gonad (Chain 2015). Also, recent findings showing reproductive functions of novel (including some de novo) genes in flies (Gubala et al. 2017; Reinhardt 2013; Kondo et al. 2017) suggest that novel genes often affect reproductive functions. These findings could explain why novel genes are more often expressed specifically in gonads, although it remains unclear whether this trend is causing the out-of-testis effect or caused by it. Furthermore, we also find "new" genes to show a more tissue-specific expression pattern than older genes (Fig. S6).

Younger genes more frequently overlap with structural variations

To study the possible role of CNVs in the emergence of new genes, we compared the fraction of genes originating from genomic regions with duplications and deletions (CNVs) among three-spined stickleback individuals using matched genomes and transcriptomes. Recently, using the same data set, young duplicated genes were shown to evolve rapidly and potentially be involved in local adaptation in sticklebacks (Chain et al. 2014; Huang et al. 2019). Here, we find that "new" and "established" genes more often overlap with CNV regions (17% and 20% of "new" and "established" genes, respectively) compared with older genes (10%; $p < 0.0001$; Fig. 4b).

New genes might have a higher likelihood to overlap with CNVs because most new genes emerge through mutations that are observable as CNVs (see Fig. 1d). Consequently, some of the CNVs we observe might be the very CNVs that have caused the overlapping gene to emerge. De novo genes can emerge from the duplication of an existing, non-coding genomic sequence, if this duplication leads to the formation of a new ORF (Fig. 1d). Deletions can also lead to de novo gene emergence by creating new ORFs or causing existing ORFs to be transcribed (Fig. 1d).

In addition to CNVs, we determined the fraction of genes overlapping with regions exhibiting an increase in sequence differentiation (F_{ST}, Fig. 4c). Here, we looked for instances of increased F_{ST} that have signatures of positive selection (from Feulner et al. 2015). In doing so, we found that older genes more often overlap with regions possessing high F_{ST} with signals of positive selection than younger genes. Consequently, positive selection acting on advantageous variants of older genes might more often be the cause of selective sweeps than selection acting on new genes. However, F_{ST} can only be calculated on regions present in all populations, i.e., regions without CNVs. Because new genes are more likely to overlap with CNVs, we also determined the fraction of genes overlapping F_{ST} outliers after excluding all genes in CNV regions, but the results were consistent whether we included or excluded CNVs.

Protein properties show no difference with gene age

To determine whether the properties of novel proteins encoded by new genes differ from those of older proteins, we analysed the structural and sequence properties of all genes (see Supplementary Fig. S5). We apply IUPred (Dosztanyi et al. 2005) and TANGO (Fernandez-Escamilla et al. 2004) on the primary protein sequence of each gene to predict disorder and aggregation propensity, respectively. These programs are routinely used in comparable studies to allow for the sequence-based analysis of large datasets (Carvunis et al. 2012; Basile et al. 2017; Angyan et al. 2012). Both measures describe important properties of proteins. Protein disorder has significant functional influence because disordered protein regions can, e.g., represent flexible binding regions (Tompa 2011).

The avoidance of aggregation represents a critical force in protein evolution. Aggregating proteins are not functional and pose a fitness burden on the cell because they are toxic (Monsellier and Chiti 2007; Geiler-Samerotte et al. 2011). Consequently, new genes are expected to avoid aggregation and might rely on disordered regions for initial functions. Essentially, genes with high aggregation propensity are expected to lower fitness so much that they will not be observed in adult organism since they, e.g., disrupt developmental processes. Interestingly, the roles of both properties, disorder and aggregation, are controversially discussed in computational (Zhao et al. 2014; Carvunis et al. 2012; Basile et al. 2017) and experimental studies (Tretyachenko et al. 2017).

Here, we do not find structural properties to show measurable differences between any of the three age classes (Supplementary Fig. S2a, b). The properties of the nucleotide sequence, however, are different with the "ancient" genes exhibiting a significantly higher hexamer score and sequence length (Supplementary Fig. S5a, b). Similarly, the amino acid composition of "ancient" genes showed some difference to the two younger categories (Supplementary Fig. S8). These differences were found only between "new" and "ancient" genes. "Established" genes here closely resemble "new" genes, indicating that no adaptation of such sequence properties has taken place since these "established" genes have emerged. This, however, does not preclude further adaptation to take place on longer time scales, as is suggested by the higher chance of overlapping with F_{ST} outliers, which is found more often among "ancient" genes.

Our analyses of sequence properties show that the "new" genes do not differ from "ancient" genes in terms of structural properties. This finding seems to be in contrast to previous studies that have identified differences in structural properties between "new" and "ancient" genes (Zhao et al. 2014; Carvunis et al. 2012; Wilson et al. 2017). However, most of these studies focused on annotated genes. Incidentally, one study also found no difference between protein structural properties between "ancient" and "new" genes when they took unannotated ORFs into account in an additional analysis (Carvunis et al. 2012). Other studies employing similar methods to the ones we applied here came to similar conclusions regarding the properties of novel proteins (Schmitz et al. 2018). However, these studies did not compare new genes between populations and as such could only study later steps of emergence compared with this study. The absence of differences in protein structure properties between "new" and "ancient" genes also suggests a high functional potential for new genes as no structural barriers have to be overcome to reach a structurally functional state. The lack of differences in terms of protein structure properties between "new" and "ancient" genes found here also suggests that selection and adaptation of structural properties do not play important roles during the early stages in new gene evolution.

In contrast to protein structural properties, nucleotide sequence properties differ between "new" and "ancient" genes. Differing nucleotide properties would be expected for new genes that emerged from non-coding sequences whose nucleotide sequence properties differ substantially from coding sequences (Wang et al. 2013). Consequently, some of the new genes we found here might have emerged from non-coding sequences (see next section for further support for this hypothesis). Our findings also suggest that there is a step-wise adjustment of the nucleotide sequence properties of young genes over time as they become older genes. However, it is still unclear what the cause of this adjustment is, be it selection for certain nucleotide sequence properties such as GC content or codon usage bias. Recent studies have identified differing GC (Basile et al. 2017) and amino acid content (Basile et al. 2019) as causes of differences in disorder content of eukaryotic proteins.

Emergence mechanisms of new genes

We searched the sequences of the "new" genes found in our study against the nine-spined stickleback's genome sequence to determine their mechanism of origin (see "Methods" section). We determined new genes mapping against intergenic regions in the nine-spined stickleback to have emerged de novo (McLysaght and Hurst 2016).

We find 800 of 988 new genes (81%) to have a hit with an E-value $\leq 10^{-3}$ in the nine-spined stickleback genome. One hundred and forty-eight (19%) of these genes map against genic regions and 84 (57%) of these gene-mapping genes map against coding sequences. The remaining 652 (of 988 novel) genes map against the nine-spined stickleback's genome in intergenic regions and, therefore, have likely emerged de novo. This fraction of de novo candidates (66% of new genes) seems quite high, but note that we already filtered out all genes showing any homology to ORFs expressed in the nine-spined stickleback or any other species. Therefore, this ratio is not comparable to ratios found in other studies. Further caveats include the possibility that genes might be expressed specifically in tissues we had no access to and that the (relatively recent) annotation of the nine-spined stickleback might lack some genes. Consequently, results here are likely slightly overestimating the frequency of de novo genes in the three-spined stickleback, i.e., new genes vs. established ones.

Interestingly, we found "new" and "established" genes both to be significantly different from "ancient" genes in terms of genomic measures such as overlap with F_{ST} or CNVs (Fig. 4), hexamer score, and sequence length (both Fig. S5). This finding suggests that some genic properties of emerging genes are markedly different from "ancient" genes and only adapt over long time frames (i.e., hundreds of millions of years) as has been proposed before (Schmitz et al. 2018). This is in contrast to protein properties which we did not find to differ between "new" and "ancient" genes.

Conclusions

We used multiple transcriptomes from natural populations along a geographic gradient to analyse the expression patterns and properties of new genes and encoded proteins in the three-spined stickleback. The chosen datasets allowed the investigation of dynamics and mode of new gene emergence at the level of populations and therefore at very short evolutionary time scales. By focusing on sequences that neither showed sequence similarity to proteins outside of the analysed species nor sequences being expressed in a sister species, the nine-spined stickleback, we analysed different stages of gene evolution and genetic novelty.

In conclusion, we find 988 new (i.e., three-spined stickleback specific) and 245 established (also present in nine-spined stickleback) genes that have emerged over the course of stickleback evolution (see Fig. 2b).

Most of these younger genes exhibit lower expression levels and were also less broadly expressed across tissues. Younger genes were not as universally expressed across populations and also overlapped with CNVs more often. These findings suggest that younger genes less often encode for essential functions compared with older genes, which preferentially overlap with population-differentiated regions under positive selection. In addition, the findings further stress the importance of analysing weakly expressed transcripts when looking for new genes.

In addition, we find new gene expression to emerge preferentially in gonads and start with relatively low expression levels, gaining higher expression over evolutionary time. We did not find the properties of novel proteins to differ from older proteins. This finding suggests that general order or disorder properties of the encoded proteins do not play a decisive role for many new genes to be retained over longer time scales.

Future studies should investigate how the patterns of new gene emergence found here, i.e. gain of transcription of (presumably intergenic) sequences, relate to their emergence mechanisms, e.g., in the evolution of expression patterns. This information could help infer how frequent the different emergence mechanisms are.

Methods

Transcriptome assembly

In this analysis, we used transcriptomes from four tissues: head kidney, spleen, ovaries and testes. Transcriptomes from two immune-related tissues were acquired from a previous study (Huang et al. 2016; accession PRJEB8677 in the European Nucleotide Archive), and from gonads (accession PRJEB26492 in the European Nucleotide Archive). These tissues were derived from the same set of populations and individuals (Supplementary Table 1) whose genomes were sequenced for genome-wide surveys of population differentiation (Chain et al. 2014; Feulner et al. 2015) (accession PRJEB26492 in the European Nucleotide Archive). Raw reads were trimmed using Trimmomatic (Bolger et al. 2014). Trimmed reads were then aligned to the stickleback reference genome (BROADS1 assembly) using HISAT2 (Kim et al. 2015) using default parameters. Transcriptomes for each sample were assembled with StringTie using default arguments (Pertea et al. 2015). The resulting transcriptomes were then merged using the StringTie merge functionality, using the reference annotation (Ensembl68 annotation) as a guide GTF. StringTie was used again on each sample with the merged transcriptome as reference to calculate the per-sample expression strengths. Protein-coding ORFs were predicted using the "transcripts_to_best_scoring_ORFs" script from the PASA suite that employs the TransDecoder algorithm (Haas et al. 2003). Here, we required a minimum protein length of 75 amino acids.

The nine-spined stickleback transcriptome (Guo et al. 2013) was assembled by first trimming the raw reads using Trimmomatic (parameters: ILLUMINACLIP:TruSeq3-SE.fa:3:30:10 LEADING:10 TRAILING:10 SLIDINGWINDOW:4:10 MINLEN:80). Trimmed reads were then assembled using Trinity de novo assembly employing default options (Grabherr et al. 2011).

New gene detection

Protein age was detected by first blasting all proteins against two outgroup proteomes (mouse, GRCm38; zebra fish, GRCz10). All proteins with hits with an E-value $\leq 10^{-3}$ were considered "ancient". In addition, all proteins without hits were searched against the NCBI NR database and all proteins with hits outside of stickleback species were again classified as "ancient". All proteins were also searched against the Pfam protein domain database (v30.0) and all proteins with hits were classified as "ancient". The remaining "not ancient" proteins were then searched against the previously assembled nine-spined stickleback's transcriptome using TBLASTN. Proteins with hits with an E-value $\leq 10^{-3}$ were then categorised as "established". All genes that did not have hits in any of these searches were considered "new" (see Fig. 1 and Supplementary Fig. S1).

Mapping against nine-spined stickleback genome

Protein sequences of new (three-spined stickleback specific) genes were searched against the nine-spined stickleback genome (Varadharajan et al. 2019) using TBLASTN. An E-value of $\leq 10^{-3}$ was used as a cut-off for valid hits. For each of the new proteins, the genome annotation at the position overlapping the best hit was analysed. "Gene" or "CDS" annotations were counted as genes or CDS, respectively.

Analysis of sequence properties

Sequence properties of gene sequence and the encoded proteins were analysed using a number of programs. Disorder was determined using IUPred (Dosztanyi et al. 2005). IUpred was used in the "short" mode and the fraction of amino acids with a disorder prediction above 0.5 was calculated. TANGO (Fernandez-Escamilla et al. 2004) was used to analyse the aggregation propensity of protein sequences. To do this, the fraction of the sequence with an aggregation propensity higher than 5% was determined. CPAT (Wang et al. 2013) was used to calculate hexamer score. The *Danio rerio* model files were used in CPAT.

Analysis of expression

Expression patterns were analysed based on FPKM. The cut-off for counting absence/presence of expression was set at 1 FPKM. Genes with lower expression levels were still included in other analyses.

Expression specificity was calculated as τ. τ can vary between 0 and 1, with 0 representing ubiquitously expressed genes and 1 specifically expressed genes (Yanai et al. 2004). τ is calculated as follows:

$$\tau = \frac{\sum_{i=1}^{N} 1 - x_i}{N - 1},$$

where N represents the number of tissues and x_i a normalised expression value in a tissue.

Analysis of F_{ST} and CNVs

F_{ST} measurements were acquired from a genome-wide divergence analysis between the same population pairs of sticklebacks as used in this study (Feulner et al. 2015). CNV regions were established in a previous study that evaluated read depth across the genome of these stickleback individuals (Chain et al. 2014). For the analysis of CNVs, only CNVs present in at least five individuals, but <50% of the individuals were considered. This was done because CNVs present in most of the analysed transcriptomes might represent a mutation in the reference genome. In addition, at least 25% of a gene's sequence was required to overlap with a CNV to be considered as overlapping.

Analysis of CpG depletion

CpG depletion is a measure for gene-body methylation since an absence of CpG dinucleotides hints at the presence of previous methylation. This is because methylated CpG dinucleotides mutate with a higher likelihood than non-methylated CpG dinucleotides. Here, CpG depletion was determined by calculating the ratio of GC dinucleotides over the product of the frequencies of G and C nucleotides:

$$D = \frac{f_{\mathrm{GC}}}{f_{\mathrm{G}} f_{\mathrm{C}}},$$

here f_{GC} is the observed fraction of CpG dinucleotides, while $f_{\mathrm{G}} f_{\mathrm{C}}$ is the expected fraction.

For each gene, the CpG depletion for all combined exons was calculated. In addition, the median CpG depletion of all intergenic sequences was calculated.

Data availability

The transcriptomic data analysed in this study can be found under the accessions PRJEB8677 and PRJEB26492 in the European Nucleotide Archive (see also Supplementary Table S1). Transcript sequences, expression values and calculated transcript ages as well as scripts are available under https://doi.org/10.6084/m9.figshare.11889771.

Acknowledgements JFS was supported by an HFSP grant to EBB. Open access funding provided by Projekt DEAL.

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Ángyán AF, Perczel A, Gáspári Z (2012) Estimating intrinsic structural preferences of de novo emerging random-sequence proteins: is aggregation the main bottleneck? FEBS Lett. 586:2468–2472

Basile W, Salvatore M, Bassot C, Elofsson A (2019) Why do eukaryotic proteins contain more intrinsically disordered regions? PLoS Comput. Biol. 15:e1007186

Basile W, Sachenkova O, Light S, Elofsson A, High GC (2017) Content causes orphan proteins to be intrinsically disordered. PLoS Comput. Biol. 13:e1005375

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics 30:2114–2120

Bornberg-Bauer E, Schmitz J, Heberlein M (2015) Emergence of de novo proteins from 'dark genomic matter' by 'grow slow and moult'. Biochem. Soc. Trans. 43:867–873

Carvunis A-R et al. (2012) Proto-genes and de novo gene birth. Nature 487:370–374

Chain FJJ et al. (2014) Extensive copy-number variation of young genes across stickleback populations. PLoS Genet. 10:e1004830

Chain FJJ (2015) Sex-biased expression of young genes in Silurana (Xenopus) tropicalis. Cytogenetic Genome Res. 145:265–277

Chen J-Y et al. (2015) Emergence, retention and selection: a trilogy of origination for functional de novo proteins from ancestral LncRNAs in primates. PLoS Genet. 11:e1005391

Cui X et al. (2014) Young genes out of the male: an insight from evolutionary age analysis of the pollen transcriptome. Mol. Plant 8:935–945

Dosztányi Z, Csizmok V, Tompa P, Simon I (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21:3433–3434

Durand É et al. (2019) Turnover of ribosome-associated transcripts from de novo ORFs produces gene-like characteristics available for de novo gene emergence in wild yeast populations. Genome Res. 29:932–943

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

Feulner PGD et al. (2013) Genome-wide patterns of standing genetic variation in a marine population of three-spined sticklebacks. Mol. Ecol. 22:635–649

Feulner PGD et al. (2015) Genomics of divergence along a continuum of parapatric population differentiation. PLoS Genet. 11:e1004966

Foster SA, Bell M (1994) The evolutionary biology of the threespine stickleback. Oxford University Press, Oxford

Geiler-Samerotte KA et al. (2011) Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. Proc. Natl Acad. Sci. 108:680–685

Gerstein MB et al. (2007) What is a gene, post-encode? History and updated definition. Genome Res. 17:669–681

Glazer AM, Killingbeck EE, Mitros T, Rokhsar DS, Miller CT (2015) Genome assembly improvement and mapping convergently evolved skeletal traits in sticklebacks with genotyping-by-sequencing. G3: Genes Genom. Genet. 5:1463–1472

Grabherr MG et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644–652

Gubala AM et al. (2017) The goddard and saturn genes are essential for Drosophila male fertility and may have arisen de novo. Mol. Biol. Evol. 34:1066–1082

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

Guo B, Chain FJ, Bornberg-Bauer E, Leder EH, Merilä J (2013) Genomic divergence between nine- and three-spined sticklebacks. BMC Genomics 14:756

Haas BJ et al. (2003) Improving the arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 31:5654–5666

Hanson D, Hu J, Hendry A, Barrett R (2017) Heritable gene expression differences between lake and stream stickleback include both parallel and antiparallel components. Heredity 119:339

Heinen TJAJ, Staubach F, Häming D, Tautz D (2009) Emergence of a new gene from an intergenic region. Curr. Biol. 19:1527–1531

Hirase S, Ozaki H, Iwasaki W (2014) Parallel selection on gene copy number variations through evolution of three-spined stickleback genomes. BMC Genomics 15:735

Huang X et al. (2017) Rapid response to changing environments during biological invasions: DNA methylation perspectives. Mol. Ecol. 26:6621–6633

Huang Y et al. (2016) Transcriptome profiling of immune tissues reveals habitat-specific gene expression between lake and river sticklebacks. Mol. Ecol. 25:943–958

Huang Y et al. (2019) Genome-wide genotype-expression relationships reveal both copy number and single nucleotide differentiation contribute to differential gene expression between stickleback ecotypes. Genome Biol. Evol. 11:2344–2359

Jones FC et al. (2012) The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484:55

Kaessmann H (2010) Origins, evolution, and phenotypic impact of new genes. Genome Res. 20:1313–1326

Kapranov P, St. Laurent G (2012) Dark matter RNA: existence, function, and controversy. Front. Genet. 3:60

Katju V, Bergthorsson U (2013) Copy-number changes in evolution: rates, fitness effects and adaptive significance. Front. Genet. 4:273

Keeling DM, Garza P, Nartey CM, Carvunis A-R (2018) The meanings of 'function' in biology and the problematic case of de novo gene emergence. Elife 8:e47014

Keller TE, Yi SV (2014) Dna methylation and evolution of duplicate genes. Proc. Natl Acad. Sci. 111:5932–5937

Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC (2009) More than just orphans: are taxonomically-restricted genes important in evolution. Trends Genet. 25:404–413

Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat. Methods 12:357–360

Klasberg S, Bitard-Feildel T, Callebaut I, Bornberg-Bauer E (2018) Origins and structural properties of novel and de novo protein domains during insect evolution. FEBS J. 285:2605–2625

Kondo S et al. (2017) New genes often acquire male-specific functions but rarely become essential in Drosophila. Genes Dev. 31:1841–1846

Kumar A, Gates PB, Czarkwiani A, Brockes JP (2015) An orphan gene is necessary for preaxial digit formation during salamander limb development. Nat. Commun. 6:8684

Li J, Arendsee Z, Singh U, Wurtele ES (2019) Recycling rna-seq data to identify candidate orphan genes for experimental analysis. bioRxiv. https://doi.org/10.1101/671263

Long M, Betrán E, Thornton K, Wang W (2003) The origin of new genes: glimpses from the young and old. Nat. Rev. Genet. 4:865–875

McKinnon JS, Rundle HD (2002) Speciation in nature: the threespine stickleback model systems. Trends Ecol. Evol. 17:480–488

McLysaght A, Hurst LD (2016) Open questions in the study of de novo genes: what, how and why. Nat. Rev. Genet. 17:567–578

Mercer TR, Dinger ME, Mattick JS (2009) Long non-coding rnas: insights into functions. Nat. Rev. Genet. 10:155

Metzger DC, Schulte PM (2018) Similarities in temperature-dependent gene expression plasticity across timescales in threespine stickleback (Gasterosteus aculeatus). Mol. Ecol. 27:2381–2396

Monsellier E, Chiti F (2007) Prevention of amyloid-like aggregation as a driving force of protein evolution. EMBO Rep. 8:737–742

Nagalakshmi U et al. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320:1344–1349

Neme R, Tautz D (2016) Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. Elife 5:e09977

Van Oss SB, Carvunis A-R (2019) De novo gene birth. PLoS Genet. 15:e1008160

Palmieri N, Kosiol C, Schlötterer C (2014) The life cycle of Drosophila orphan genes. Elife 3:e01311

Pertea M et al. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33:290–295

Prabh N, Rödelsperger C (2019) De novo, divergence, and mixed origin contribute to the emergence of orphan genes in pristionchus nematodes. G3: Genes Genom. Genet. 9:2277–2286

Ramsköld D, Wang ET, Burge CB, Sandberg R (2009) An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput. Biol. 5:e1000598

Rando OJ, Verstrepen KJ (2007) Timescales of genetic and epigenetic inheritance. Cell 128:655–668

Reinhardt JA et al. (2013) De novo ORFs in Drosophila are important to organismal fitness and evolved rapidly from previously non-coding sequences. PLoS Genet. 9:e1003860

Roesti M, Moser D, Berner D (2013) Recombination in the threespine stickleback genome–patterns and consequences. Mol. Ecol. 22:3014–3027

Ruiz-Orera J et al. (2015) Origins of de novo genes in human and chimpanzee. PLoS Genet. 11:e1005721

Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM (2014) Long non-coding RNAs as a source of new peptides. Elife 3:e03523

Sabath N, Wagner A, Karlin D (2012) Evolution of viral proteins originated de novo by overprinting. Mol. Biol. Evol. 29:3767–3780

Schmitz JF, Ullrich KK, Bornberg-Bauer E (2018) Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. Nat. Ecol. Evol. 2:1626

De Smet R et al. (2013) Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proc. Natl Acad. Sci. 110:2898–2903

Tautz D, Domazet-Lošo T (2011) The evolutionary origin of orphan genes. Nat. Rev. Genet. 12:692–702

Tobler R, Nolte V, Schlötterer C (2017) High rate of translocation-based gene birth on the Drosophila Y chromosome. Proc. Natl Acad. Sci. 114:11721–11726

Toll-Riera M, Albà MM (2013) Emergence of novel domains in proteins. BMC Evol. Biol. 13:47

Tompa P (2011) Unstructural biology coming of age. Curr. Opin. Struct. Biol. 21:419–425

Tretyachenko V et al. (2017) Random protein sequences can form defined secondary structures and are well-tolerated in vivo. Sci. Rep. 7:15449

Vakirlis N, Carvunis A-R, McLysaght A (2019) Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. Elife 9:e53500

Vanderperre B et al. (2013) Direct detection of alternative open reading frames translation products in human significantly expands the proteome. PLoS ONE 8:e70698

Varadharajan S et al. (2019) A high-quality assembly of the nine-spined stickleback (pungitius pungitius) genome. Genome Biol. Evol. 11:3291–3308

Wang H et al. (2015) CG gene body DNA methylation changes and evolution of duplicated genes in cassava. Proc. Natl Acad Sci. 112:13729–13734

Wang L et al. (2013) CPAT: coding-potential assessment tool using an alignment-free logistic regression model. Nucleic Acids Res. 41:e74

Wang X et al. (2017) Gene-body CG methylation and divergent expression of duplicate genes in rice. Sci. Rep. 7:2675

Wilson BA, Masel J (2011) Putatively noncoding transcripts show extensive association with ribosomes. Genome Biol. Evol. 3:1245–1252

Wilson BA, Foy SG, Neme R, Masel J (2017) Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. Nat. Ecol. Evol. 1:0146

Witt E, Benjamin S, Svetec N, Zhao L (2019) Testis single-cell rna-seq reveals the dynamics of de novo gene transcription and germline mutational bias in drosophila. Elife 8:e47138

Wu D-D et al. (2014) "Out of pollen" hypothesis for origin of new genes in flowering plants: study from *Arabidopsis thaliana*. Genome Biol. Evol. 6:2822–2829

Xie C et al. (2012) Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. PLoS Genet. 8:e1002942

Yanai I et al. (2004) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21:650–659

Zemach A, McDaniel IE, Silva P, Zilberman D (2010) Genome-wide evolutionary analysis of eukaryotic dna methylation. Science 328:916–919

Zhang L et al. (2019) Rapid evolution of protein diversity by de novo origination in oryza. Nat. Ecol. Evol. 3:679–690

Zhao L, Saelao P, Jones CD, Begun DJ (2014) Origin and spread of de novo genes in Drosophila melanogaster populations. Science 343:769–772