Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation

Michael S Werner, Bogdan Sieriebriennikov, Neel Prabh, et al.

Genome Res. published online September 19, 2018
Access the most recent version at doi:10.1101/gr.234872.118

Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation

Michael S. Werner[1], Bogdan Sieriebriennikov[1], Neel Prabh[1], Tobias Loschko[1], Christa Lanz[1] and Ralf J. Sommer[1]

[1]Department of Evolutionary Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany

Corresponding author: ralf.sommer@tuebingen.mpg.de

Keywords: epigenetics, orphan genes, new genes, *Pristionchus pacificus*, *Caenorhabditis elegans*

Abstract

Species-specific, new, or 'orphan' genes account for 10-30% of eukaryotic genomes. Although initially considered to have limited function, an increasing number of orphan genes have been shown to provide important phenotypic innovation. How new genes acquire regulatory sequences for proper temporal and spatial expression is unknown. Orphan gene regulation may rely in part on origination in open chromatin adjacent to pre-existing promoters, although this has not yet been assessed by genome-wide analysis of chromatin states. Here we combine taxon-rich nematode phylogenies with Iso-Seq, RNA-seq, ChIP-seq, and ATAC-seq to identify the gene structure and epigenetic signature of orphan genes in the satellite model nematode *Pristionchus pacificus*. Consistent with previous findings we find young genes are shorter, contain fewer exons, and are on average less strongly expressed than older genes. However, the subset of orphan genes that are expressed exhibit distinct chromatin states from similarly expressed conserved genes. Orphan gene transcription is determined by a lack of repressive histone modifications, confirming long-held hypotheses that open chromatin is important for new gene formation. Yet orphan gene start sites more closely resemble enhancers defined by H3K4me1, H3K27ac and ATAC-seq peaks, in contrast to conserved genes that exhibit traditional promoters defined by H3K4me3 and H3K27ac. While the majority of orphan genes are located on chromosome arms that contain high recombination rates and repressive histone marks, strongly expressed orphan genes are more randomly distributed. Our results support a model of new gene origination by rare integration into open chromatin near enhancers.

Introduction

Gene regulation is a highly orchestrated process that includes transcription factor binding sites (TFBs), non-coding RNAs, histone modifications and chromatin structure (Voss and Hager 2014). The identification and mechanism of these molecular factors have been revealed for several conserved gene networks leading towards a better understanding of development and disease. But how new genes, also referred to as orphan or taxon-restricted, acquire this complex architecture is unknown. For the increasing number of identified new genes that provide important biological function (Cai et al. 2008; Burki and Kaessmann 2004; Rosso et al. 2008; Reinhardt et al. 2013; Chen et al. 2013b; 2010; Mayer et al. 2015; Santos et al. 2017), the evolutionary path from origin to integration into gene networks depends on their precise transcriptional regulation (Carelli et al. 2016). Yet in the majority of cases it is unclear how even the most fundamental *cis*-regulatory elements like promoter and termination sequences are obtained (Tautz and Domazet-Lošo 2011; Long et al. 2013). Orphan genes can originate *de novo* or by duplication, recombination or horizontal gene transfer into pre-existing regulatory architecture (Chen et al. 2012b; Betrán and Long 2003; Kaessmann et al. 2009; Kaessmann 2010; McLysaght and Hurst 2016; Li et al. 2009). But the extent to which this occurs is limited by the potential to disrupt the genes already there (Vinckenbosch et al. 2006). In the few cases where integration has been observed, the presence of nearby regulatory sequences was largely detected by proximity, or sequence homology to known promoters, CpG islands or TFBs (Carvunis et al. 2012; Abrusán 2013; Ruiz-Orera et al. 2015; Li et al. 2016). Given these constraints, the contribution of pre-existing regulatory architecture to new gene transcription is still unknown, and with functional genomic information (e.g. chromatin states and enhancers) could potentially be much larger. Indeed, a recent analysis of mammalian ChIP-seq data sets found 51% of expressed mouse retrogenes (mRNAs that are reverse transcribed and inserted into the genome) exhibit robust H3K4 tri-methylation (Carelli et al. 2016), and transcription of the new gene *QQS* in *Arabadopsis thaliana* is inversely correlated with DNA methylation at 5’

transposable elements (Silveira et al. 2013), suggesting an important role for chromatin regulation in new gene transcription. We sought to employ the rich taxonomic resources of nematodes to first identify young and old genes, and then observe their regulatory architecture by several genome-wide approaches.

The diplogastrid nematode *Pristionchus pacificus* can be found in a necromenic relationship with beetles, but has been developed in the laboratory as a satellite model for comparative studies to *C. elegans* (Sommer and Streit 2011; Sommer and McGaughran 2013)(Fig. 1A-D). More recent genetic analysis of dimorphic mouth-forms (Fig. 1E-G) has led to *P. pacificus* emerging as an important model system for phenotypic plasticity in its own right (Bento et al. 2010; Ragsdale et al. 2013; Serobyan et al. 2016; Kieninger et al. 2016). In addition to the vast taxonomic diversity and corresponding genomes of other nematode species, the recent high quality chromosome-scale genome (Rödelsperger et al. 2017) and reverse genetic tools (Witte et al. 2015) in *P. pacificus* provide a robust framework for studying new genes (Prabh and Rödelsperger 2016; Baskaran et al. 2015). Here we probe the gene structure, expression, and regulatory architecture of *P. pacificus* evolutionary gene classes with long-read PacBio transcript sequencing (Iso-Seq), traditional high depth RNA-sequencing (RNA-seq), chromatin immunoprecipitation (ChIP-seq) of six histone post-translational modifications and assay for transposon-accessible chromatin (ATAC-seq). In addition to our findings, the data sets collected provide the first epigenomic map in *P. pacificus*, which is only the second comprehensive chromatin state annotation in nematodes, creating a resource for future functional and comparative studies.

Results

Partitioning of *P. pacificus* genes into evolutionary classes

The first *P. pacificus* draft genome published in 2008 (Dieterich et al. 2008) had a large number of genes with undetectable homology. Although the confidence in these gene predictions was initially low, every subsequent refinement of both the genome and gene annotation continually detected 20-40% of genes that appear as new, orphan, or taxon-restricted (Sinha et al. 2012; Baskaran et al. 2015; Prabh and Rödelsperger 2016; Baskaran and Rödelsperger 2015). Using our most recent chromosome-scale PacBio genome (Rödelsperger et al. 2017) and 24 other nematode species we re-evaluated the relative abundance of evolutionary gene classes (Fig. 1H). We defined the most highly conserved genes as having 1:1 orthology with *C. elegans* (BLASTP e-value ≤ 0.001), which are estimated to have diverged from *P. pacificus* between 60-90 mya (Cutter 2008; Rota-Stabelli et al. 2013; Hedges et al. 2015). We also defined an intermediate conserved class as 'homologs' if they display homology to at least one gene in the other 24 nematode species (see Methods) – which could represent either relatively young genes, or old genes that have been lost. Finally, we define 'orphan' genes as having no homology to genes in the other 24 queried species. The resulting partition of genes approximates the '30% rule' of new gene composition (Khalturin et al. 2009)(Fig. 1I). We then applied several genomic approaches to molecularly characterize each evolutionary gene class.

Characterization of gene structure by long read RNA sequencing (Iso-Seq)

We sought to improve the overall gene annotation in *P. pacificus*, and then characterize the genetic structures of each evolutionary gene class using PacBio Iso-Seq on mixed-developmental stage RNA (see Supplemental Methods, Supplemental Fig. S1A-C). After alignment we obtained 640,664 reads with a median insert size of 1,363 nucleotides (Supplemental Fig. S1D). Despite low read depth compared to conventional RNA-seq, our Iso-

Seq data covered 17,307 genes (68% of genes in the reference annotation 'El Paco') (Rödelsperger et al. 2017).

Relative to the current reference annotation, Iso-Seq identified a tighter distribution of gene lengths (Fig. 2A, median Iso-Seq=1,452 compared to median reference=1,599, $p < 2.2 \times 10^{-16}$ Wilcoxon rank sum test). This difference appears to be due to a more narrow distribution of exons, with 96.5% of Iso-Seq gene annotations containing between 1 to 20 exons, compared to 85.7% for the reference annotation (Fig. 2B, $p = < 2.2 \times 10^{-16}$ Wilcoxon rank sum test). The tighter distribution is also more consistent with the highly curated gene annotation of *C. elegans* in which 98.0% of genes contain between 1:20 exons (Supplemental Fig. S1E,F)(Michael and Manyuan 1999). This potential improvement in accuracy appears to result from fragmentation of excessively long gene predictions into distinct transcripts (see Supplemental Fig. S1G for example).

Long read Iso-Seq also provides more robust identification of isoforms, which are historically difficult to assemble from standard short-read RNA-sequencing (Conesa et al. 2016). Approximately half (50.6%) of expressed genes in *P. pacificus* exhibit greater than one isoform, and roughly a third (30.9%) exhibit greater than three isoforms (Supplemental Fig. S1H,I). However, some of these transcripts could be artifacts biased by incomplete coverage of 5' ends. Hence, we conservatively defined alternatively spliced isoforms as transcripts with the same start and stop coordinates, but differential exon inclusion or exclusion, intron retention, or differential splice site. Under this classification we observed 3,861 (24%) of expressed genes exhibit alternative splicing in *P. pacificus* (Fig. 2C), similar to the ~25% of genes estimated in *C. elegans* (Ramani et al. 2011). As an example we highlight gene *umm-s259-11.10-mRNA* where the majority of Iso-Seq reads (17/19) cover the entire transcript yielding eight isoforms, in stark contrast to standard short-read sequencing which rarely covers more than 3 exons per read (Fig. 2D). Collectively, a tighter distribution of transcript lengths and exon number, and diversity

of isoforms suggests that Iso-Seq improves the quality and quantity of gene annotation in *P. pacificus*.

Among evolutionary gene classes, most *C. elegans* 1:1 orthologs (88%), and approximately half of homologous and orphan genes (46% and 56%, respectively) exhibit Iso-Seq coverage, demonstrating that our Iso-Seq data is sensitive enough to detect thousands of transcripts from each evolutionary gene class (Fig. 2E). We also performed Iso-Seq on rRNA-depleted 'total RNA' (see Supplemental Methods) to asses whether young genes are un-, or under-polyadenylated, which is typical of non-coding RNAs (Derrien et al. 2012). We found a similar percent coverage from the direct and total RNA methods (Fig. 2E,F), and a consistent polyadenylation read bias for all gene classes (Fig. 2G-I). Hence most young genes, or at least transcribed young genes, are polyadenylated. As polyadenylation is an important component of transcriptional and translational regulation (Proudfoot 2011; Proudfoot et al. 2002), this argues that most young genes have retained, or already acquired, 3' termination and processing sequence architecture.

We then used our Iso-Seq annotation to characterize gene length and exon number between evolutionary gene classes. Consistent with other systems (Stein et al. 2018; Ruiz-Orera et al. 2015), we found a strong bias of *C. elegans* 1:1 orthologs to be longer and contain more exons than homologs, which in turn were longer and contained more exons than orphan genes (Fig. 2J,K, $p = <2.2 \times 10^{-16}$ Wilcoxon rank sum test). The intermediate gene structure of intermediate conserved genes (homologs) is also consistent with a transitional evolutionary path between young and old genes proposed by Carvunis *et al*. in 2012 (Carvunis et al. 2012; Abrusán 2013; Neme and Tautz 2013). In the following sections we sought to characterize and compare the chromatin regulation of young versus old genes.

The *P. pacificus* epigenome

To identify regulatory regions and expression levels of orphan, homolog, and *C. elegans* 1:1 orthologs we performed two to three replicates of ChIP-seq on nine histone modifications and two replicates of RNA-seq in *P. pacificus* adults, and two replicates of ATAC-seq on mixed-stage cultures (Supplemental Fig. S2, Supplemental Table 3, and Supplemental Methods). All data sets showed good correlations between biological replicates (Pearson correlation between 0.70-0.93 for ChIP, 0.88 for ATAC-seq, and 0.98 for gene FPKMs in RNA-seq) (Supplemental Fig. S3). We identified enriched regions (i.e. peaks) for each replicate of ChIP and ATAC-seq using MACS2 (Zhang et al. 2008) (Supplemental Table 1, see Methods). H2bub, H3K9ac and H3K79me2 exhibited less than 50% peak reproducibility and were excluded from further analysis (Supplemental Fig. S4A). The majority of remaining samples exhibited >70% overlap between replicates, except for H3K9me3 (54% reproducibility). However, H3K9me3 is a broadly distributed histone modification that is challenging for peak-finding software (Wang et al. 2013), and although most H3K9me3 antibodies are of low specificity (Nishikori et al. 2012; Hattori et al. 2013), they can nevertheless distinguish constitutive vs. facultative heterochromatin (Trojer and Reinberg 2007).

We also performed ATAC-seq for identifying regions of open chromatin (Buenrostro et al. 2013). Although the standard protocol led to reproducible peaks, initially we could not identify nucleosomal read density, perhaps suggesting a difficulty of obtaining higher resolution fragments from highly differentiated and heterogeneous cell populations. Yet the new *omni* ATAC method (Corces et al. 2017) yielded nucleosomal and sub-nucleosomal read densities (Supplemental Fig. 2E), which we used for subsequent analysis.

We clustered the six high-confidence histone marks and omni ATAC-seq data using a Hidden-Markov Model (ChromHMM) (Ernst and Kellis 2012) into eight chromatin states (Fig. 3A). Each chromatin state is enriched in histone modifications that define specific functional domains, such as actively transcribed regions, heterochromatin, and regulatory loci. We

assigned putative chromatin state annotations based on established classifications (Fig. 3B)(Ernst et al. 2011; Rada-Iglesias et al. 2011)(Supplemental Table 2). We find that, at least at the whole animal level, approximately half (57%) of the genome is repressed, approximately a fifth (16%) represents actively transcribed genes, and more than a quarter (27%) is regulatory (including 6,785 promoters, 13,648 active enhancers, and 3,853 'poised' enhancers) (Fig. 3C).

Next, we verified that histone marks are enriched at the center of promoters and active enhancer annotations, and performed a *de novo* motif search (Heinz et al. 2010) (Fig. 3D-G). Both promoters (30.6%) and enhancers (22.2%) were enriched in a recognition sequence for MBP1, a yeast transcriptional activator that controls cell cycle progression (Koch et al. 1993). There is weak homology (BlastP, e =2 $x10^{-4}$) to MBP1 in *P. pacificus* (UMM-S233-5.4-mRNA-1), and in the future it will be interesting to see if this gene is also involved in cell-cycle control. There were also notable differences between enhancers and promoters, including binding site matches to human homeobox, *Drosophila* GAGA, and eukaryotic GATA transcription factors, demonstrating the precision of promoter and enhancer annotations, and hinting at the existence of deeply conserved regulatory elements.

As expected, promoter annotations were strongly enriched at the 5' end of genes (Fig. 3H). There was also another peak near the 3' end of genes. Enhancers were also enriched at both 5' and 3' ends, although they are more evenly distributed throughout gene bodies. The existence of promoter/enhancer elements at the 3' end of genes has been observed in other species, and while its function is still unclear, there are several reports supporting promoter-3' end chromatin looping to facilitate successive rounds of transcription and enforce directionality (Werner et al. 2017; Lainé et al. 2009; O'Sullivan et al. 2004; Grzechnik et al. 2014).

To verify that our chromatin states correlate with a dynamically regulated gene we looked at *Ppa-pax-3*, which our lab has shown is expressed in early juvenile stages but is repressed during development (Yi and Sommer 2007). Indeed, we found *Ppa-pax-3* is in a large H3K27me3-repressed domain in adults (Fig. 3I). However, we also noticed two putative

enhancers after the first exon and 3' end, perhaps suggesting preparation for activation in developing embryos. Collectively, these data represent the first genome-wide annotation of chromatin regulation in *P. pacificus,* and to our knowledge represents only the second comprehensive data set in nematodes.

Chromatin regulation corresponds to gene expression

We extended the previous single gene example to genome-wide high-depth RNA-seq and binned the adult transcriptome into four expression categories (Fig. 4A), then assessed the chromatin states of each. As predicted, gene bodies (exons and introns) of the highest expressed categories (groups 1 and 2) exhibited enrichment in chromatin states designated as 'transcriptional transition' and 'elongation'. Conversely, repressive chromatin states were virtually absent from genes in the top two categories. In contrast, the lowest two expressed categories (groups 3 and 4) exhibited proportionally greater enrichment in repressive chromatin states and decreased enrichment in transcriptional transition and elongation states (Fig. 4B). While there was a minor enrichment in promoter chromatin states at 5' ends and 5' UTRs between high vs. low expressed categories, there was a larger difference in repressive chromatin states, especially at the 5' ends. There was also an increase in enhancer enrichment at 5' ends and 5' UTRs in the low expressed categories, perhaps reflecting a 'poised' chromatin state that is reactive to environmental influence. While promoters and enhancers exhibit a relatively small portion of the genome (15.6%), they comprise the majority of intergenic regions (Fig. 4B), hinting at a large and mostly unexplored regulatory circuitry in the compact nematode genome.

Chromatin regulation of evolutionary gene classes

Next we assessed the chromatin states of evolutionary gene classes. *C. elegans* 1:1 orthologs resembled the highest expression categories (groups 1 and 2), while conserved and orphan

genes more closely resembled the lower expression categories (groups 3 and 4) (Fig. 4C-E). These histone patterns reflect the higher expression of *C. elegans* 1:1 orthologs compared to less conserved gene classes (Fig. 4F). Nevertheless, we noticed a significant number of orphan and homologous expressed gene outliers (Fig. 4F), and wondered if their chromatin signatures resembled that of expressed *C. elegans* 1:1 orthologs. Here, we found differences. Specifically, strongly expressed (group 1 and 2) orphan and homologous genes, which represent only 9.3% and 12.8% of their respective categories, broadly resembled the general chromatin state pattern of their classes except for having reduced repressive histone marks (Fig. 4G-I). Second, chromatin states 3 and 4 representing transcriptional transition and elongation are more highly represented in *C. elegans* 1:1 orthologs compared to expressed orphan and conserved genes. Third, *C. elegans* 1:1 orthologs exhibit little to no signature of active enhancers (chromatin state 1) at their 5' ends or 5' UTRs, which are instead dominated by the promoter chromatin state consisting of H3K4me3 and H3K27ac. However, expressed homologous and orphan genes exhibited both promoter and enhancer enrichment at their 5' ends and 5' UTRs, and orphan genes in particular exhibited greater enrichment in enhancer vs. promoter chromatin states.

To investigate this difference more closely we examined the distribution of histone ChIP-seq and ATAC-seq around the 5' ends of each evolutionary class. Whereas expressed *C. elegans* 1:1 ortholog TSSs are dominated by H3K4me3 and H3K27ac, expressed orphan and homologous genes exhibit comparatively stronger enrichment of H3K4me1 and ATAC-seq (Fig. 4J). Specifically, *C. elegans* 1:1 orthologs exhibit a 5' H3K4me3/H3K4me1 ratio of 10.1, compared to 2.4 for homologs, and 1.4 for orphan genes. Furthermore, while 54% of expressed *C. elegans* 1:1 ortholog 5' ends are within 1 kb of an annotated promoter, only 27% of expressed homologous gene and 21% of expressed orphan genes are in similar proximity to promoters (p <2.2x10^{-16} Wilcoxon rank sum test)(Fig. 5A). Conversely, 46% of expressed homologous and orphan gene TSSs are in 1 kb of active and poised enhancers, compared to 33% of expressed *C. elegans* 1:1 ortholog TSSs (p <2.2x10^{-16} Wilcoxon rank sum test for both

comparisons)(Fig. 5B). Importantly, the expression of group 1 and 2 orphan and homologous genes are actually higher than *C. elegans* 1:1 orthologs (p <2.2x10^{-16} Wilcoxon rank sum test)(Supplemental Fig. S5), demonstrating that their chromatin architecture is independent of the general correlation with expression. There are two key points from these results. First, the transcription of young genes appears to depend on the absence of repressive heterochromatin, demonstrating a widely held but unconfirmed theory for the requirement of open chromatin in new gene expression. Second, the 5' end of expressed young genes, especially orphan genes, resemble enhancers rather than canonical promoters.

Genomic position affects chromatin regulation of evolutionary gene classes

Finally, we analyzed the general pattern of chromatin marks and evolutionary gene classes at the chromosomal level (Fig. 6). We observed strong patterns of activating marks in the center of autosomes II-V, and the repressive mark H3K27me3 on the chromosome arms. Conversely the X Chromosome was highly enriched in both repressive marks H3K27me3 and H3K9me3. These patterns have been observed in *C. elegans* (Liu et al. 2011), and correspond to general patterns in nematodes of dense clusters of conserved genes and low recombination rates in the center of chromosomes, and species-specific genes and high recombination rates in the chromosome arms (Coghlan 2005). However, Chromosome I in *P. pacificus* is an exception to other autosomes, where we observed two bands of activating marks instead of one. Recent analysis suggests that roughly half of *P. pacficus* Chromosome I is homologous to *C. elegans* Chromosome X, and the other half is homologous to Chromosome V (Rödelsperger et al. 2017). The *P. pacficus* chromosome pattern was viewed as ancestral because this organization is also found in the distantly related nematode *Bursaphelenchelus xylophilus*. However, the bipartite presence of histone modifications and conserved genes hints at an ancient chromosomal fusion from an even earlier origin, or frequent and repeated chromosomal fission and fusion events.

312 The chromosome-scale distribution of evolutionary gene classes was consistent with
313 histone modification patterns. Specifically, *C. elegans* 1:1 orthologs, which are strongly
314 expressed, are enriched in the active histone-mark chromosome centers. Conversely, the lower
315 expressed homologous and orphan genes are enriched in the chromosome arms, which contain
316 higher recombination rates and a greater density of repressive histone marks. However, these
317 patterns are lost when controlling for expression. Highly transcribed (group 1 and 2) orphan and
318 homologous genes were more randomly distributed throughout chromosomes, if not slightly
319 biased to be closer towards the centers (Fig. 6). This genome-wide perspective also supports
320 the model that location into open chromatin is a critical factor for new gene origination, or at
321 least transcription, of new genes.
322
323 **Discussion**
324 Here we combine the first chromatin state analysis in *P. pacificus* with taxon-rich nematode
325 phylogenies to analyze the transcriptional regulation of young genes. We identified eight
326 chromatin states that partition the genome into varying levels of repression, transcription, or
327 regulatory elements. Expressed young genes were found in open chromatin states, supporting a
328 widely held model of new gene origination. But to our surprise, young gene 5' ends are more
329 similar to enhancers than traditional promoters. We also analyzed young gene transcript
330 structure by long read Iso-Seq, which revealed a unique signature for each evolutionary gene
331 class. Finally, a bipartite pattern of active histone marks in Chromosome I provides molecular
332 evidence of an ancient chromosomal fusion event ~180 mya. The ability to probe over 20,000
333 high confidence promoters and enhancers will be a valuable resource for future mechanistic
334 studies, especially when combined with the powerful array of genetic, phylogenetic, and
335 ecological tools recently available to *P. pacificus.*
336 The origin and subsequent regulation of orphan genes is a widely debated topic that has
337 garnered several theoretical models. Among these are that orphan genes can be transcribed by

integrating into open chromatin, or near gene promoters, effectively hijacking their regulatory sequences and thereby mitigating the need to evolve them *de novo* (Long et al. 2013; Kaessmann 2010; Kaessmann et al. 2009; Tautz and Domazet-Lošo 2011; McLysaght and Hurst 2016; Chen et al. 2012b; 2012a). This model was supported by analyzing the position of transcribed retrogenes in the human genome (Vinckenbosch et al. 2006), however it was also demonstrated that such integration is often deleterious to the host genes. Indeed a recent analysis found only ~14% of mammalian retrogenes utilized pre-existing promoters (Carelli et al. 2016). Nevertheless, to date very few tests of these predictions have been performed beyond retrogenes, and the identify of *cis*-regulatory elements has traditionally been inferred through spatial proximity to genes or known regulatory sequences like TFBs and CpG islands (Betrán and Long 2003; Chen et al. 2012b; Ni et al. 2012; Carvunis et al. 2012; Li et al. 2016; Ruiz-Orera et al. 2015). The phylogenetic diversity of nematode genomes and our recent chromosome-scale genome (Rödelsperger et al. 2017) allowed us to query all orphan genes in *P. pacificus*, including but not limited to retrogenes. By applying ChIP-seq and ATAC-seq we could then interrogate the functional *P. pacificus* genome, including *cis* and *trans* enhancers, and up to eight different chromatin states. The data presented herein support a model of orphan gene integration into open chromatin near enhancers preferentially over promoters.

Enhancers were originally thought of as inactive DNA elements that harbor TFBs (Wasylyk 1988), however over the past decade research from several labs has shown enhancers exhibit bi-directional transcription by RNA polymerase (Andersson et al. 2014; Chen et al. 2013a; Lam et al. 2014). Now there is a growing consensus that enhancers and promoters are similar regulatory elements (Andersson et al. 2015; Kim and Shiekhattar 2015), but promoters have evolved additional sequences to enforce directionality (Grzechnik et al. 2014) or increased expression. Indeed, promoters can even function as enhancers for other genes (Engreitz et al. 2016). Under this paradigm, we propose a model whereby a new gene that originates near an enhancer and is adaptive, will eventually acquire more sophisticated

regulatory architecture – thereby transitioning the enhancer into a promoter (Fig. 5C). This model is complimentary to 'proto-promoters' proposed by the Kaessman lab for 8-9% of rat expressed retrogenes which have H3K4me3 enrichment in rats but H3K4me1 enrichment in syntenic non-expressed regions in mouse (Carelli et al. 2016). However, here we show that proximity of new genes near enhancers correlates with and presumably drives their transcription, and extend this argument beyond retrogenes as a general feature of new genes. If these transcripts prove functional, then selection can convert the enhancer to a promoter. This model potentially solves two problems faced by new genes, (1) expression via origination near enhancers, and (2) introduction near enhancers, especially *trans* enhancers, as opposed to promoters, does not require exchange or competition with the pre-existing gene landscape. Nevertheless, we caution that such interpretation is speculative at this point, and examining and then experimentally manipulating H3K4me1/3 at syntenic loci in closely related strains and species is necessary to test these hypotheses.

In principle, this model could operate regardless of the method of new gene origination (*de novo*, duplication and divergence, or retrotransposition). Transcripts from enhancers or lncRNA promoters generally exhibit less splicing, 3' processing and polyadenylation relative to protein coding genes (Derrien et al. 2012), and are often digested by the nuclear exosome (Schlackow et al. 2017). In *de novo* gene evolution, mutations that recruit sequence specific splicing factors or 3' processing factors such as CPSF73 could stabilize enhancer transcripts allowing for their translation and potential functionalization. In some cases a *de novo* gene that is acting as a functional ncRNA, referred to as 'moonlighting', could lead to greater expression and a greater window of time to accrue such mutations (Jalali et al. 2016). New genes formed by duplication and insertion, or retrotransposition near an enhancer could similarly be transcribed, but without the parental regulatory architecture. In this new genomic context the gene will likely be expressed in different developmental stages or tissues, possibly providing new functions. While misregulation of genes often coincides with deleterious effects and

disease (Lee and Young 2013), in this case the parental gene is still maintained and operating under normal control, while the copied gene is freed, within limits (Geiler-Samerotte et al. 2011), to explore neofunctionalization.

Compared to the current reference annotation ('El Paco'), our Iso-Seq annotation identified shorter genes with fewer exons. The distribution was more similar to *C. elegans* gene structures. Nevertheless, we note that there are still substantially more genes in *P. pacificus* greater than 10 exons compared to *C. elegans* (Supplemental Fig. S1F), arguing that further refinement is still required (although an evolutionary divergence in gene length is formally possible). We also explored the genetic structure of young vs. old genes. Orphan genes displayed the shortest gene lengths and fewest exons, and *C. elegans* 1:1 orthologs were the longest and contained the most exons. The result that homologs appear to be intermediate in length and exon number is consistent with a transitional path between old and young genes (Abrusán 2013; Carvunis et al. 2012). But whether this indicates divergence from old genes, or *de novo* evolution from young genes is unknown, and likely reflects examples of both. Iso-Seq also identified that almost a quarter (24%) of expressed genes in *P. pacificus* have multiple isoforms. Although alternative splicing is often due to stochastic transcriptional errors (Pickrell et al. 2010), there are important examples of alternative isoforms that affect diverse biological processes (Baralle and Giudice 2017). Whether the multiplicity of transcripts observed here are differentially expressed during development or in different environmental conditions, and ultimately if they are functional, will be the focus of future experiments.

Iso-Seq of polyadenylated transcripts and rRNA-depleted total RNA demonstrated that most young genes are polyadenylated. In mammals noncoding RNAs are un-, or underpolyadenylated (Derrien et al. 2012), arguing that most new genes in *P. pacificus* represent coding transcripts. However, retrogene transcripts that contain their polyadenylation 'scar' in the genome may be transcribed directly with a polyA tail, and thus appear as polyadenylated regardless of whether they have been pseudogenized or not. Nevertheless, our

interpretation that they are mostly coding is consistent with a previous investigation of orphan genes in *P. pacificus* that found appreciable peptide coverage from mass spectrometry and evidence of negative selection (Prabh and Rödelsperger 2016). Beyond orphan genes, comparing polyadenylated and total RNA Iso-Seq data sets should also be valuable for investigating gene structures of long non-coding RNAs (lncRNAs), including antisense ncRNA that have been shown to affect phenotypic plasticity in *P. pacificus* (Serobyan et al. 2016).

Genome-wide, we found most young genes are present in chromosome arms where recombination and repressive chromatin in nematodes is the highest (Coghlan 2005; The *C. elegans* Sequencing Consortium 1998). However, the approximately 10% of young genes that are highly transcribed (expression groups 1 and 2) were more randomly distributed. Thus, while recombination in the arms appears to be a furnace for new gene generation, most of these genes are repressed (expression groups 3 and 4), and have a higher barrier for functionalization. This pattern highlights several unresolved questions. In particular, does the presence in open chromatin reflect rare recombination events or *de novo* origination? Further, are these transcribed new genes 'born' into open chromatin and serve as a template for evolution, or have they already acquired nascent function and their presence in open chromatin is a result of translocation to increase their expression? Additional functional genomic comparisons and synteny analysis may shed light on these questions.

At chromosome scale resolution, we observe a double-banding of active histone marks on Chromosome I, in contrast to all other autosomes in both *P. pacificus* and *C. elegans* (Liu et al. 2011). Based on previous phylogeny and synteny analysis (Rödelsperger et al. 2017), we propose this pattern is a remnant from a fusion event that occurred prior to the split between *Diplogasterida* and *Tylenchida*, estimated at ~ 180 mya (Hedges et al. 2015; Cutter 2008). Then more recently, this portion broke off in the *Caenorhabditis* lineage. This interpretation parsimoniously explains the long-standing conundrum that Chromosome V in *C. elegans* has unusual chromosome 'arm-like' characteristics, including relatively high recombination rates and

a low density of conserved genes (Barnes et al. 1995; The *C. elegans* Sequencing Consortium 1998; Parkinson et al. 2004). In essence, it looks like a chromosome arm because it is, or more precisely, was, prior to breaking off. If true, the remarkable stability of histone mark patterns suggests that chromatin organization per se could serve as a molecular fossil of past genomic re-arrangements. Perhaps probing chromatin structure in conjunction with recombination rates could provide a historical record of genome evolution in other nematodes or organisms.

Methods

Evolutionary Gene Classification

Nematode phylogenies were schematically drawn using data downloaded and analyzed from Holterman et al. (Holterman et al. 2017) and Van Megen et al. (van Megen et al. 2009). Evolutionary gene classes were defined in a tiered process. First, we defined conserved genes by blastp and tblastn analysis; we performed all pairwise searches of *P. pacificus* proteins as query against 24 nematode protein sets as target, and the proteins of each of the 24 nematodes as query against *P. pacificus* proteins as target. One blastp hit ($e<10^{-3}$) in any of these 48 comparisons, or one tblastn hit ($e<10^{-5}$) using *P. pacificus* proteins as query against any of the 24 nematode species genomes was enough to classify a gene in *P. pacificus* as conserved. Any gene that did not fit these criteria was defined as a *P. pacificus* 'Orphan gene'. Within the conserved gene class, we then defined 1:1 orthologs as having the best reciprocal blastp hit ($e \leq 10^{-3}$) between *C. elegans* and *P. pacificus* (sorted by e value, and raw scores were used to break ties). Conserved genes that were not in this 1:1 ortholog class, but were previously identified by homology in at least one of the 24 nematodes species, were defined as 'Homologous genes'. We kept the e value cutoff relatively 'high' because of the large phylogenetic distance between *C. elegans* and *P. pacificus*, and hence more conservative with respect to our orphan gene lists.

Nematode synchronization and collection

P. pacificus (PS312) cultures for ChIP-, ATAC-, and RNA-seq were synchronized with bleach and grown on agar to young adults following (Werner et al. 2017, see Supplemental Methods). Worm pellets were flash frozen until processing. For Iso-Seq we used mixed-developmental stage (egg, J2, and J4/young adult) RNA at equimolar ratios.

Native histone ChIP-seq

Native (non-cross-linked) chromatin immunoprecipitation (ChIP) of histone post-translational modifications was performed by combining nematode nuclear isolation (Steiner et al. 2012) with native ChIP (Brand et al. 2008). Co-precipitated DNA was PCR-amplified and converted to Illumina libraries using the TruSeq Nano kit (Illumina) and sequenced on a HiSeq 3000. See Supplemental Methods for detailed protocol.

ATAC-seq

omni ATAC-seq was performed on mixed stage purified nuclei following the Corces et al. 2017 protocol, with a few modifications (see Supplementary Methods) and sequenced on an Illumina HiSeq 3000.

Iso-Seq

RNA was extracted from different developmental time-points separately using TRIzol Reagent (Invitrogen), then after the quality control equal amounts of RNA from different time points were pooled. cDNA synthesis of 'direct' Iso-Seq was performed directly using SMARTer PCR cDNA Synthesis Kit (Clontech Laboratories), while 'total RNA' was first rRNA-depleted with Ribo-Zero rRNA Removal Kit (Human/Mouse/Rat) (Illumina), then *in vitro* polyadenylated with Poly(A) Polymerase (New England Biolabs). Direct and rRNA-depleted cDNA were converted into SMRTbell libraries following the guidelines provided by Pacific Biosciences. SMRT Link software version 4.0.0 (Pacific Biosciences) was used to convert subreads to circular consensus sequences and identify full-length non-chimeric reads, which were mapped to the El Paco genome using GMAP (Wu et al. 2005). See Supplemental Methods for detailed protocol.

RNA-seq

Whole animal young adult (64-68 hour post-bleaching) frozen pellets were freeze-thawed 3x in TRIzol Reagent (Invitrogen) before purification, and converted to sequencing libraries with the NEBNext Ultra Directional RNA-seq for Illumina kit, and sequenced on a HiSeq 3000. See Supplemental Methods for detailed protocol.

Bioinformatics

All sequencing data was mapped to the El Paco genome (Rödelsperger et al. 2017) using GMAP (Wu et al. 2005) for Iso-Seq, Bowtie 2 (Ben Langmead and Salzberg 2012) for ChIP and ATAC-seq, and HISAT2 (Kim et al. 2015) for RNA-seq. Peaks were obtained by MACS2 (Zhang et al. 2008), and only samples containing 50% overlap between replicates were kept. Overlapping peaks were merged using BEDTools (Quinlan and Hall 2010)(Supplemental Table 1). Coverage plots were calculated using BEDTools or HOMER (Heinz et al. 2010) with merged replicate files, and plotted in R (R Core Team 2016). Chromatin states were obtained with ChromHMM (Ernst and Kellis 2012) using merged replicate input files. Distance to nearest promoter or enhancers were performed with BEDTools. See Supplemental Methods for detailed procedure of all Bionformatic steps.

Data Access

Raw and processed data sets from this study have been submitted to the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena) under accession number PRJEB24584.

Acknowledgements

We are thankful to current members of the Sommer laboratory and Dr. Talia Karasov for thoughtful critique of experiments, results, and interpretations. This study was funded by the Max Planck Society.

Author contributions

M.S.W. and R.J.S. conceived and designed all experiments. M.S.W. conducted ChIP- and ATAC-seq with assistance from T.L.; M.S.W, B.S. and C.L. performed Iso-Seq; N.P. conducted phylogenetic analysis and prepared evolutionary gene category data sets; M.S.W. performed all bioinformatic analysis. M.S.W. wrote the manuscript with assistance from R.J.S.

Disclosure Declaration

The authors declare no competing conflicts of interest.

Figure 1. Comparison of *Pristionchus pacificus* and *Caenorhabditis elegans* and phylogenetic relationship. (*A-B*). *P. pacificus* is often found in a necromenic relationship with insect hosts, preferentially scarab beetles, in the dormant dauer state. When the beetle dies, worms exit the dauer stage to feed on bacteria that bloom on the decomposing carcass. (*C-D*) *C. elegans*, the classic nematode model organism is often found in leaf detritus and rotting fruits. Ex. rotting apple photo taken by M.S.W. (*E-G*) *P. pacificus* has become an important model for developmental (phenotypic) plasticity. Adults can adopt (*E*) a narrow mouth form with one tooth (stenostomatous, St) that makes them strict bacterial feeders. However, the 'boom-and-bust' life cycle creates significant competition for resources, and under crowded conditions adults can develop an alternative mouth form (*F*) with a wider buccal cavity and an extra tooth (eurystomatous, Eu) that allows them to prey on other nematodes. (*G*) Shown here is a eurystomatous *P. pacificus* preying on a *C. elegans* larva. (*H*) Schematic phylogeny of nematodes, generated based on the publications of Holterman et. al 2017, and Van Megen et al. 2009. (*I*) Break-down of *P. pacificus* genes by evolutionary category: one to one orthology with *C. elegans* (*C. elegans* 1:1) is the most conserved, followed by genes sharing homology with at least one gene from the 24 other nematodes (homologous), and finally genes that are

only found in *Pristionchus* (orphan). All categories were defined by BLASTP homology (e-value ≤ 0.001, see Methods).

Figure 2. Long read RNA sequencing (Iso-Seq) improves gene annotation, identifies alternative splicing and can distinguish different evolutionary gene classes by gene structure. (*A*) Density distribution of cDNA gene lengths between the 'El Paco' reference (grey) and Iso-Seq annotation (black). The Iso-Seq annotation was derived from guided assembly using Stringtie (Pertea et al. 2016; see Methods), and plots were created using the Density function in R. (*B*) Density distribution of exons per gene between 'El Paco' reference and Iso-Seq annotations. Method and color scheme are similar to (*A*). (*C*) Alternatively spliced isoforms, defined as having multiple detected isoforms with the same start and stop coordinates. White column represents genes containing isoforms that have the same exon-intron structure but different splice-sites, and red columns represent genes containing isoforms with different numbers of exons due to intron retention or exon inclusion/exclusion. (*D*) Example locus of Iso-Seq reads compared to standard short read RNA-seq. Also shown are Iso-Seq assembled isoforms compared to the single reference gene umm-S259-11.10-mRNA-1, made in Integrated Genome Viewer (IGV). (*E,F*) Percent coverage of evolutionary gene classes by Iso-Seq with either the 'direct' method (*E*) or rRNA-depleted total RNA (*F*). (*G-I*) Iso-Seq coverage per gene of each evolutionary class in direct (y-axis) compared to total (x-axis). Coverage was determined by BEDTools, and median ratios of direct/total RNA are presented. Dotted line of slope = 1, y intercept = 0 represents equal coverage between methods. (*J,K*) Similar density distributions of cDNA length and exon number as in (*A,B*) but for the three evolutionary gene classes.

Figure 3. The epigenome of *Pristionchus pacificus*. (*A*) Chromatin states determined through a hidden markov model (ChromHMM) clustered by histone modifications and ATAC-seq, normalized by coverage. Darker blue represents greater enrichment. (*B*) Candidate annotation of each chromatin state according to ENCODE/modENCODE data sets (Ernst et al. 2011; The Roadmap Epigenomics Consortium 2015). Repressive chromatin states are divided into three categories according to standard definitions of constitutive (repressed 3) and facultative (repressed 1 and 2) heterochromatin. Poised enhancers are defined according to previous annotations of loci containing H3K27me3 and DNase sensitivity. (*C*) Genome-wide distribution of chromatin states, and further clustering into three categories: repressive, transcribed, or regulatory. (*D*) Heatmap of indicated histone modifications for promoter chromatin states, where each line represents a single 6 kb locus centered on the promoter. Heatmap matrices were generated in HOMER, clustered from highest to lowest enrichment and plotted in R. (*E*) Position weight matrices of *de novo* sequence motifs in promoters, queried using HOMER. Table also includes percent of promoters containing motif, *p* value and matches to known transcription factors. (*F,G*) Similar to (*D,E*) but for enhancer chromatin states. (*H*) Average density plots of promoter (light blue) and enhancer (dark blue) locations relative to gene bodies, extended 5 kb in each direction from their 5' and 3' ends. Density values measured using HOMER and plotted in excel. (*I*) Epigenomic data of histone modification ChIP-seq, ATAC-seq and RNA-seq surrounding the *Ppa-pax3* gene. Input is included as a reference, and chromatin state annotations are included on the bottom matching the colors in (*C*). ChIP and ATAC-seq coverage is autoscaled per sample, and RNA-seq coverage is in log-scale.

Figure 4. Chromatin states correlate with expression, but expressed young genes exhibit distinct profiles. (*A*) Average expression (fpkm) from two biological replicates of RNA-seq, plotted for each gene from highest to lowest along the x-axis. Expression categories were binned according to approximate inflection points. (*B*) Chromatin state enrichment of each

expression category broken down by genetic element (i.e. TSS, UTRs, exons and introns). (*C-E*) Similar to (*B*) but for each evolutionary gene class. (*F*) Expression of each evolutionary gene class determined from average RNA-seq fpkms. Stars indicate p value < 0.05, Welch's two-tailed test. (*G-I*) Similar to (*B-E*), but only for highly expressed (group 1 and 2) genes belonging to each category. (*J*) Normalized average densities of H3K4me3, H3K4me1, H3K27ac and ATAC-seq over a 7 kb window centered at 5' ends. Densities were measured in HOMER and normalized to the highest and lowest values in each gene class.

Figure 5. Distance of promoters and enhancers to evolutionary gene classes. (A) Distance cumulative frequency distribution of the nearest promoter, or (B) enhancer (active and poised) to transcription start sites (TSSs) from each evolutionary gene category. (C) Model of new gene transcriptional regulation. Enhancers exhibit bi-directional transcription, which can lead to new gene transcription if it originates near an enhancer, either by *de novo* formation or duplication/insertion. If the new gene provides useful function, selection will occur not only on protein function, but also the gene structure leading to more exons, and on regulatory elements to provide more temporal or spatial control, and more or less transcription. Ultimately, evolution on enhancer sequences will convert it to a traditional promoter.

Figure 6: Chromosome-wide distribution of histone modifications reveals distinct patterns for evolutionary gene classes and a double-band pattern on ChrI. Genome-wide patterns of histone modifications from ChIP-seq and ATAC-seq presented as a heatmap with increasing abundance from white to blue, and white to red for RNA-seq (normalized by depth). Also plotted are gene densities of each evolutionary class binned by expressed (group 1 and 2) or transcriptionally repressed (group 3 and 4) for each class.

References

Abrusán G. 2013. Integration of new genes into cellular networks, and their structural maturation. *Genetics* **195**: 1407–1417.

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507**: 455–461.

Andersson R, Sandelin A, Danko CG. 2015. A unified architecture of transcriptional regulatory elements. *Trends in Genetics* **31**: 426–433.

Baralle FE, Giudice J. 2017. Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology* **18**: 437–451.

Barnes TM, Kohara Y, Coulson A, Hekimi S. 1995. Meiotic recombination, noncoding DNA and genomic organization in Caenorhabditis elegans. *Genetics* **141**: 159–179.

Baskaran P, Rödelsperger C. 2015. Microevolution of Duplications and Deletions and Their Impact on Gene Expression in the Nematode Pristionchus pacificus ed. D. Dupuy. *PLoS ONE* **10**: e0131136.

Baskaran P, Rödelsperger C, Prabh N, Serobyan V, Markov GV, Hirsekorn A, Dieterich C. 2015. Ancient gene duplications have shaped developmental stage-specific expression in Pristionchus pacificus. *BMC Evol Biol* **15**: 185.

Ben Langmead, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Bento G, Ogawa A, Sommer RJ. 2010. Co-option of the hormone-signalling module dafachronic acid–DAF-12 in nematode evolution. *Nature* **466**: 494–497.

Betrán E, Long M. 2003. Dntf-2r, a Young Drosophila Retroposed Gene With Specific Male Expression Under Positive Darwinian Selection. *Genetics* **164**: 977–988.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**: 1213–1218.

Burki F, Kaessmann H. 2004. Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. *Nat Genet* **36**: 1061–1063.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**: 2012–2018.

Cai J, Zhao R, Jiang H, Wang W. 2008. De Novo Origination of a New Protein-Coding Gene in Saccharomyces cerevisiae. *Genetics* **179**: 487–496.

Carelli FN, Hayakawa T, Go Y, Imai H, Warnefors M, Kaessmann H. 2016. The life history of

659 retrocopies illuminates the evolution of new mammalian genes. *Genome Res* **26**: 301–314.

660 Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B,
661 Hidalgo CA, Barbette J, Santhanam B, et al. 2012. Proto-genes and *de novo* gene birth.
662 *Nature* **487**: 370.

663 Chen RAJ, Down TA, Stempor P, Chen QB, Egelhofer TA, Hillier LW, Jeffers TE, Ahringer J.
664 2013a. The landscape of RNA polymerase II transcription initiation in C. elegans reveals
665 promoter and enhancer architectures. *Genome Res* **23**: 1339–1347.

666 Chen S, Krinsky BH, Long M. 2013b. New genes as drivers of phenotypic evolution. *Nature*
667 *Reviews Genetics* **14**: 645–660.

668 Chen S, Ni X, Krinsky BH, Zhang YE, Vibranovski MD, White KP, Long M. 2012a. Reshaping of
669 global gene expression networks and sex□biased gene expression by integration of a
670 young gene. *EMBO J* **31**: 2798–2809.

671 Chen S, Spletter M, Ni X, White KP, Luo L, Long M. 2012b. Frequent Recent Origination of
672 Brain Genes Shaped the Evolution of Foraging Behavior in Drosophila. *Cell Rep* **1**: 118–
673 132.

674 Chen S, Zhang YE, Long M. 2010. New Genes in Drosophila Quickly Become Essential.
675 *Science* **330**: 1682–1685.

676 Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak
677 MW, Gaffney DJ, Elo LL, Zhang X, et al. 2016. A survey of best practices for RNA-seq data
678 analysis. *Genome Biol* **17**: 13.

679 Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A,
680 Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. 2015. Integrative analysis of
681 111 reference human epigenomes. *Nature* **518**: 317–330.

682 Coghlan, A. 2005. Nematode genome evolution. WormBook, ed. The C. elegans Research
683 Community, WormBook, doi/10.1895/wormbook.1.15.1, http://www.wormbook.org.

684 Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S,
685 Satpathy AT, Rubin AJ, Montine KS, Wu B, et al. 2017. An improved ATAC-seq protocol
686 reduces background and enables interrogation of frozen tissues. *Nat Methods* **14**: 959–962.

687 Cutter AD. 2008. Divergence times in Caenorhabditis and Drosophila inferred from direct
688 estimates of the neutral mutation rate. *Mol Biol Evol* **25**: 778–786.

689 Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel
690 A, Knowles DG, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs:
691 analysis of their gene structure, evolution, and expression. *Genome Res* **22**: 1775–1789.

692 Dieterich C, Clifton SW, Schuster LN, Chinwalla A, Delehaunty K, Dinkelacker I, Fulton L, Fulton
693 R, Godfrey J, Minx P, et al. 2008. The *Pristionchus pacificus* genome provides a unique
694 perspective on nematode lifestyle and parasitism. *Nat Genet* **40**: 1193–1198.

695 Engreitz JM, Haines JE, Perez EM, Munson G, Chen J, Kane M, McDonel PE, Guttman M,

Lander ES. 2016. Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539**: 452–455.

Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**: 215–216.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**: 43–49.

Geiler-Samerotte KA, Dion MF, Budnik BA, Wang SM, Hartl DL, Drummond DA. 2011. Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proc Natl Acad Sci USA* **108**: 680–685.

Grzechnik P, Tan-Wong SM, Proudfoot NJ. 2014. Terminate and make a loop: regulation of transcriptional directionality. *Trends Biochem Sci* **39**: 319–327.

Hattori T, Taft JM, Swist KM, Luo H, Witt H, Slattery M, Koide A, Ruthenburg AJ, Krajewski K, Strahl BD, et al. 2013. Recombinant antibodies to histone post-translational modifications. *Nat Methods* **10**: 992–995.

Hedges SB, Marin J, Suleski M, Paymer M, Kumar S. 2015. Tree of life reveals clock-like speciation and diversification. *Mol Biol Evol* **32**: 835–845.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**: 576–589.

Holterman M, Karegar A, Mooijman P, van Megen H, van den Elsen S, Vervoort MTW, Quist CW, Karssen G, Decraemer W, Opperman CH, et al. 2017. Disparate gain and loss of parasitic abilities among nematode lineages ed. W.O. Wong. *PLoS ONE* **12**: e0185445.

Jalali S, Gandhi S, Scaria V. 2016. Navigating the dynamic landscape of long noncoding RNA and protein-coding gene annotations in GENCODE. *Hum Genomics* **10**: 35.

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

Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. *Nature Reviews Genetics* **10**: 19–31.

Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends in Genetics* **25**: 404–413.

Kieninger MR, Ivers NA, Rödelsperger C, Markov GV, Sommer RJ, Ragsdale EJ. 2016. The Nuclear Hormone Receptor NHR-40 Acts Downstream of the Sulfatase EUD-1 as Part of a Developmental Plasticity Switch in Pristionchus. *Curr Biol*.

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

Kim T-K, Shiekhattar R. 2015. Architectural and Functional Commonalities between Enhancers and Promoters. *Cell* **162**: 948–959.

Koch C, Moll T, Neuberg M, Ahorn H, Nasmyth K. 1993. A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. *Science* **261**: 1551–1557.

Lainé J-P, Singh BN, Krishnamurthy S, Hampsey M. 2009. A physiological role for gene loops in yeast. *Genes & Development* **23**: 2604–2609.

Lam MTY, Li W, Rosenfeld MG, Glass CK. 2014. Enhancer RNAs and regulated transcriptional programs. *Trends Biochem Sci* **39**: 170–182.

Lee TI, Young RA. 2013. Transcriptional Regulation and Its Misregulation in Disease. *Cell* **152**: 1237–1251.

Li L, Foster CM, Gan Q, Nettleton D, James MG, Myers AM, Wurtele ES. 2009. Identification of the novel protein QQS as a component of the starch metabolic network in Arabidopsis leaves. *The Plant Journal* **58**: 485–498.

Li Z-W, Chen X, Wu Q, Hagmann J, Han T-S, Zou Y-P, Ge S, Guo Y-L. 2016. On the Origin of De Novo Genes in Arabidopsis thaliana Populations. *Genome Biol Evol* **8**: 2190–2202.

Liu T, Rechtsteiner A, Egelhofer TA, Vielle A, Latorre I, Cheung M-S, Ercan S, Ikegami K, Jensen M, Kolasinska-Zwierz P, et al. 2011. Broad chromosomal domains of histone modification patterns in C. elegans. *Genome Res* **21**: 227–236.

Long M, VanKuren NW, Chen S, Vibranovski MD. 2013. New Gene Evolution: Little Did We Know. *Annu Rev Genet* **47**: 307–333.

Mayer MG, Rödelsperger C, Witte H, Riebesell M, Sommer RJ. 2015. The Orphan Gene dauerless Regulates Dauer Development and Intraspecific Competition in Nematodes by Copy Number Variation ed. S.K. Kim. *PLoS Genet* **11**: e1005146.

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

Michael D, Manyuan L. 1999. Intron—exon structures of eukaryotic model organisms. *Nucleic Acids Res* **27**: 3219–3228.

Neme R, Tautz D. 2013. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics 2012 13:1* **14**: 117.

Ni X, Zhang YE, Nègre N, Chen S, Long M, White KP. 2012. Adaptive Evolution and the Birth of CTCF Binding Sites in the Drosophila Genome ed. H.S. Malik. *PLOS Biol* **10**: e1001420.

Nishikori S, Hattori T, Fuchs SM, Yasui N, Wojcik J, Koide A, Strahl BD, Koide S. 2012. Broad Ranges of Affinity and Specificity of Anti-Histone Antibodies Revealed by a Quantitative Peptide Immunoprecipitation Assay. *Journal of Molecular Biology* **424**: 391–399.

O'Sullivan JM, Tan-Wong SM, Morillon A, Lee B, Coles J, Mellor J, Proudfoot NJ. 2004. Gene loops juxtapose promoters and terminators in yeast. *Nat Genet* **36**: 1014–1018.

768 Parkinson J, Mitreva M, Whitton C, Thomson M, Daub J, Martin J, Schmid R, Hall N, Barrell B,
769 Waterston RH, et al. 2004. A transcriptomic analysis of the phylum Nematoda. *Nat Genet*
770 **36**: 1259–1267.

771 Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis
772 of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc 11: 1650–1667.

773 Pickrell JK, Pai AA, Gilad Y, Pritchard JK. 2010. Noisy Splicing Drives mRNA Isoform Diversity
774 in Human Cells ed. E.T. Dermitzakis. *PLoS Genet* **6**: e1001236.

775 Prabh N, Rödelsperger C. 2016. Are orphan genes protein-coding, prediction artifacts, or non-
776 coding RNAs? *BMC Bioinformatics 2016 17:1* **17**: 226.

777 Proudfoot NJ. 2011. Ending the message: poly(A) signals then and now. *Genes & Development*
778 **25**: 1770–1782.

779 Proudfoot NJ, Furger A, Dye MJ. 2002. Integrating mRNA Processing with Transcription. *Cell*
780 **108**: 501–512.

781 Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. 2011. A unique
782 chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**: 279–
783 283.

784 Ragsdale EJ, Müller MR, Rödelsperger C, Sommer RJ. 2013. A Developmental Switch Coupled
785 to the Evolution of Plasticity Acts through a Sulfatase. **155**: 922–933.

786 Ramani AK, Calarco JA, Pan Q, Mavandadi S, Wang Y, Nelson AC, Lee LJ, Morris Q,
787 Blencowe BJ, Zhen M, et al. 2011. Genome-wide analysis of alternative splicing in
788 Caenorhabditis elegans. *Genome Res* **21**: 342–348.

789 Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, Jones CD. 2013. De Novo ORFs in
790 Drosophila Are Important to Organismal Fitness and Evolved Rapidly from Previously Non-
791 coding Sequences ed. E. Betrán. *PLoS Genet* **9**: e1003860.

792 Rosso L, Marques AC, Weier M, Lambert N, Lambot M-A, Vanderhaeghen P, Kaessmann H.
793 2008. Birth and Rapid Subcellular Adaptation of a Hominoid-Specific CDC14 Protein ed. M.
794 Long. *PLOS Biol* **6**: e140.

795 Rota-Stabelli O, Daley AC, Pisani D. 2013. Molecular timetrees reveal a Cambrian colonization
796 of land and a new scenario for ecdysozoan evolution. *Curr Biol* **23**: 392–398.

797 Rödelsperger C, Meyer JM, Prabh N, Lanz C, Bemm F, Sommer RJ. 2017. Single-Molecule
798 Sequencing Reveals the Chromosome-Scale Genomic Architecture of the Nematode Model
799 Organism Pristionchus pacificus. *Cell Rep* **21**: 834–844.

800 Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabidó E, Kondova I, Bontrop R, Marques-
801 Bonet T, Albà MM. 2015. Origins of De Novo Genes in Human and Chimpanzee ed. J.
802 Noonan. *PLoS Genet* **11**: e1005721.

803 Santos ME, Le Bouquin A, Crumière AJJ, Khila A. 2017. Taxon-restricted genes at the origin of
804 a novel trait allowing access to a new environment. *Science* **358**: 386–390.

Schlackow M, Nojima T, Gomes T, Dhir A, Carmo-Fonseca M, Proudfoot NJ. 2017. Distinctive Patterns of Transcription and RNA Processing for Human lincRNAs. *Mol Cell* **65**: 25–38.

Serobyan V, Xiao H, Namdeo S, Rödelsperger C, Sieriebriennikov B, Witte H, Röseler W, Sommer RJ. 2016. Chromatin remodelling and antisense-mediated up-regulation of the developmental switch gene eud-1 control predatory feeding plasticity. *Nat Commun* **7**: 12337.

Silveira AB, Trontin C, Cortijo S, Barau J, Del Bem LEV, Loudet O, Colot V, Vincentz M. 2013. Extensive Natural Epigenetic Variation at a De Novo Originated Gene ed. M.D. Purugganan. *PLoS Genet* **9**: e1003437.

Sinha A, Sommer RJ, Dieterich C. 2012. Divergent gene expression in the conserved dauer stage of the nematodes Pristionchus pacificus and Caenorhabditis elegans. *BMC Genomics 2012 13:1* **13**: 254.

Sommer RJ, McGaughran A. 2013. The nematode Pristionchus pacificus as a model system for integrative studies in evolutionary biology. *Molecular Ecology* **22**: 2380–2393.

Sommer RJ, Streit A. 2011. Comparative Genetics and Genomics of Nematodes: Genome Structure, Development, and Lifestyle. *http://dxdoiorg/101146/annurev-genet-110410-132417*.

Stein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, Zhang C, Chougule K, Gao D, Iwata A, Goicoechea JL, et al. 2018. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. *Nat Genet* **50**: 285–296.

Steiner FA, Talbert PB, Kasinathan S, Deal RB, Henikoff S. 2012. Cell-type-specific nuclei purification from whole animals for genome-wide expression and chromatin profiling. *Genome Res* **22**: 766–777.

Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. *Nature Reviews Genetics* **12**: 692–702.

Trojer P, Reinberg D. 2007. Facultative Heterochromatin: Is There a Distinctive Molecular Signature? *Mol Cell* **28**: 1–13.

van Megen H, van den Elsen S, Holterman M, Karssen G, Mooyman P, Bongers T, Holovachov O, Bakker J, Helder J. 2009. A phylogenetic tree of nematodes based on about 1200 full-length small subunit ribosomal DNA sequences. *Nematology* **11**: 927–950.

Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci USA* **103**: 3220–3225.

Voss TC, Hager GL. 2014. Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nature Reviews Genetics* **15**: 69–81.

Wang J, Lunyak VV, Jordan IK. 2013. BroadPeak: a novel algorithm for identifying broad peaks in diffuse ChIP-seq datasets. *Bioinformatics* **29**: 492–493.

Wasylyk B. 1988. Enhancers and transcription factors in the control of gene expression. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* **951**: 17–35.

Werner MS, Sullivan MA, Shah RN, Nadadur RD, Grzybowski AT, Galat V, Moskowitz IP, Ruthenburg AJ. 2017. Chromatin-enriched lncRNAs can act as cell-type specific activators of proximal gene transcription. *Nat Struct Mol Biol* **24**: 596–603.

Witte H, Moreno E, Rödelsperger C, Kim J, Kim J-S, Streit A, Sommer RJ. 2015. Gene inactivation using the CRISPR/Cas9 system in the nematode Pristionchus pacificus. *Dev Genes Evol* **225**: 55–62.

Yi B, Sommer RJ. 2007. The pax-3 gene is involved in vulva formation in Pristionchus pacificus and is a target of the Hox gene lin-39. *Development* **134**: 3111–3119.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.