BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Contrasting patterns of molecular evolution in metazoan germ line genes

Carrie A. Whittle[1] and Cassandra G. Extavour[1,2*]

Abstract

Background: Germ lines are the cell lineages that give rise to the sperm and eggs in animals. The germ lines first arise from primordial germ cells (PGCs) during embryogenesis: these form from either a presumed derived mode of preformed germ plasm (inheritance) or from an ancestral mechanism of inductive cell-cell signalling (induction). Numerous genes involved in germ line specification and development have been identified and functionally studied. However, little is known about the molecular evolutionary dynamics of germ line genes in metazoan model systems.

Results: Here, we studied the molecular evolution of germ line genes within three metazoan model systems. These include the genus *Drosophila* (*N*=34 genes, inheritance), the fellow insect *Apis* (*N*=30, induction), and their more distant relative *Caenorhabditis* (*N*=23, inheritance). Using multiple species and established phylogenies in each genus, we report that germ line genes exhibited marked variation in the constraint on protein sequence divergence (dN/dS) and codon usage bias (CUB) within each genus. Importantly, we found that *de novo* lineage-specific inheritance (LSI) genes in *Drosophila* (*osk*, *pgc*) and in *Caenorhabditis* (*pie-1*, *pgl-1*), which are essential to germ plasm functions under the derived inheritance mode, displayed rapid protein sequence divergence relative to the other germ line genes within each respective genus. We show this may reflect the evolution of specialized germ plasm functions and/or low pleiotropy of LSI genes, features not shared with other germ line genes. In addition, we observed that the relative ranking of dN/dS and of CUB between genera were each more strongly correlated between *Drosophila* and *Caenorhabditis*, from different phyla, than between *Drosophila* and its insect relative *Apis*, suggesting taxonomic differences in how germ line genes have evolved.

Conclusions: Taken together, the present results advance our understanding of the evolution of animal germ line genes within three well-known metazoan models. Further, the findings provide insights to the molecular evolution of germ line genes with respect to LSI status, pleiotropy, adaptive evolution as well as PGC-specification mode.

Keywords: Germ line genes, Primordial germ cells, Protein divergence, Codon usage, Molecular evolution, Metazoans, Specification mode, *Drosophila*, *Caenorhabditis*, *Apis*

Background

Germ lines are the specialized cell lineage contained in the gonads of sexually reproducing animals that give rise to the sperm and eggs. The germ cell lineages are separate from the soma and are the only source of heritable genetic variation passed between generations, providing them with a crucial role in reproductive success, fitness and evolutionary biology. Extensive experimental and cytological research has focused on the discovery and functionality of germ line genes in animal models, which has led to the identification of a wide range of genes involved in germ line establishment and development [1–5]. At present however, much remains unknown about the molecular evolutionary dynamics of these germ line genes within metazoan model systems.

The germ lines first emerge in early embryogenesis with the formation of primordial germ cells (PGCs). The PGCs in some organisms arise from a presumed evolutionarily derived mode of maternally generated germ line determinants (inheritance), otherwise known as germ

* Correspondence: extavour@oeb.harvard.edu
[1]Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA
[2]Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA

plasm (specialized cytoplasm containing proteins and RNAs), as has been observed in systems including flies (*Drosophila*), wasps (*Nasonia*), nematodes (*Caenorhabditis*), and frogs (*Xenopus*) [2, 4, 6–9]. Alternatively, PGCs may emerge from an ancestral mechanism involving inductive cell-cell signalling pathways (induction) as has been supported by experimental embryological and functional genetic data in taxa such as crickets (*Gryllus*), mice (*Mus*), and salamanders (*Ambystoma*) [4, 5, 8–12]. Gene products involved in germ line specification and in various stages of germ line development have been identified using different model systems (*e.g.,* [1–5, 13–16]), allowing study of how those genes involved in core germ line functions have evolved.

A primary model for the study of germ line genes has been *Drosophila*. In *Drosophila*, PGC-specification depends on maternally provided germ plasm (inheritance), which is asymmetrically located in the posterior region of the oocyte cytoplasm [4, 17]. Genes essential to germ plasm formation and PGC specification in *Drosophila* include two genes believed to have originated only in insects, *oskar* (or *osk*) and *pgc* (*polar granule component*) [2, 4, 18, 19]. *osk* is involved in the recruitment of most other germ plasm components in *D. melanogaster* [2, 4]. *pgc* is involved in transcriptional repression in pole cells, and may be exclusive to *Drosophila*, as this gene has not been reported in other Dipteran insects or other animals to date [19, 20]. Together, these two genes, *osk* and *pgc* appear to be *de novo* genes [21] originating in specific insect lineages that may have facilitated the evolutionary transition to germ plasm and evolved crucial functions in early stages of germ line development, that is PGC specification [2, 4, 5, 17, 19].

Another central model for germ line gene research is the nematode *C. elegans*. In this taxon, similar to *Drosophila*, PGCs are specified under the inheritance mode, via localized molecular entities denoted in that system as P granules [3, 22]. Rather than presence alone, a sufficient concentration and specific conformation of P granules are required for them to act as germ line determinants [23–25]. P granules contain protein products of nematode-specific genes, including the RNA-binding *pgl-1 (P-granule abnormality*) (a partly redundant *pgl-3* also plays a role; [26]), which helps keep P granule components localized and thus functional [4, 26]. *meg-1* (*maternal-effect germ cell defective*), which has been linked to P-granule assembly and fertility [*meg-2* has also been identified as putatively partly redundant; [27, 28]), may be specific to the species *C. elegans*, as it appears to display poor homology to any other animal proteins, including within the same genus [3, 27]. In addition, the *Caenorhabditis* gene *pie-1* (*pharyngeal and intestinal excess*), is involved in transcriptional repression and maintaining germ cell fate [29], and shares functional (but not sequence) similarity to *D. melanogaster pgc* [1, 4, 19]. This gene is essential for PGC establishment and has only been reported in nematodes [4]. In this regard, whilst the inheritance mode (germ plasm) is shared between *Drosophila* and *Caenorhabditis*, the key upstream regulators appear to have arisen independently and in a lineage-restricted manner in these two model organisms [2–4, 19, 30], consistent with the proposed convergent evolution of this mechanism [7, 9].

An additional primary model of study of germ line genes including those involved in PGC-specification, this time using induction mode, or zygotic cell-cell signalling, is mouse [4, 5, 17]. Whilst much of the genetic programming remains to be understood, it has been established that in early embryogenesis ligands including BMPs (Bone Morphogenetic Proteins; particularly BMP2 and BMP4) and WNT3 are crucial for inducing signals that initiate PGC formation [4, 5]. Such signalling activates expression of the transcription factor *Blimp-1* (*B lymphocyte-induced maturation protein-1*), needed for repression of somatic gene expression, epigenetic programming of the genome, and ultimately PGC formation [4, 5, 17, 31]. Recently, similar processes involving *BMP* and *Blimp-1* signalling have been shown to be utilized in the induction of PGCs in a basally branching insect, the cricket *Gryllus bimaculatus* [10, 32] suggesting conservation of this signalling mechanism across these divergent taxa with the induction mode. Although the precise nature of the signalling pathways that may induce PGCs in other insects remains largely unknown, unlike in *Drosophila*, in some insects PGCs are first described as arising from among the mesoderm of the abdominal segments in the vicinity of the future primordial gonad [33]. This is the case, for example, in the honeybee *Apis mellifera*, where cells with the morphology and gene expression typical of animal PGCs are first detected late in embryogenesis in the abdominal segments that will house the primordial gonad [34–38]. Thus, we infer that BMP/Blimp-1 signalling to induce the germ line may likely be shared with other insect models that appear to specify PGCs via inductive-signalling [33] such as *Apis*.

Many germ line genes, including specification genes, appear to play multiple roles in germ line development and/or across animal models [1, 3, 4, 13–16]. The gene *vasa*, an ATP-dependent RNA helicase, is a universal marker of germ lines [1, 39] (although in some animals it plays somatic roles (reviewed by [40]). The products of *vasa* are essential components of *Drosophila* germ plasm [18, 41, 42], localize to *C. elegans* P granules (the *vasa* ortholog *glh-1* is essential for fertility; *glh-2-4* may be nonessential; [43, 44]), and become upregulated in mouse and cricket PGCs following their induction (*vasa* mouse ortholog is *mvh*: 10, 32, [45, 46]), consistent with

a ubiquitous role in germ lines across these divergent systems. In addition, *nanos* (or *nos*) and *pumilio* (*pum*) encode interacting RNA-binding proteins which have each been linked to germ plasm and PGCs in *Drosophila* and been associated with early germ line development across metazoans (reviewed by [1, 4, 47]). In sum, current data suggest that while certain germ line genes are particularly involved in either the inheritance and induction modes [1–5], many germ line genes play various roles across modes and/or throughout germ line development [1, 3, 4, 13–16].

At present, investigations into the molecular evolution of germ line genes remain uncommon, and the few studies available have largely focused on putative genetic regulators in the germ line stem cells of *Drosophila*. For instance, a recent study in *Drosophila* examined 366 genes identified from RNAi screening as involved in adult self-renewing female germ line stem cells [48]. That study suggested those genes were increased targets of short-term selective sweeps, but not typically of recurrent long-term (interspecies) positive selection based on dN/dS [48]. Separate assessments have reported that certain genes expressed in adult germ lines including stem cells (e.g. *pum*, *stone-wall* (*stwl*)) have evolved adaptively and/or in response to Wolbachia infection in *Drosophila* [49–51]. In terms of the investigation of germ line genes involved in PGC specification [19, 52], a study of the *osk* gene in *Drosophila* has shown that dN/dS varies among segments of this protein and shows some signs of positive selection. However, these signs were reported for only one of the 18 species studied (*D. virilis*), and the signal was not consistent across various approaches even in that lineage [52]. Given the limited data on the molecular evolution of germ line genes further study is warranted in metazoan models.

In the present study, we investigate the molecular evolution of genes with experimental or cytological evidence of involvement in germ line specification and/or development, broadly referred to herein as germ line genes, in metazoan models. Specifically, we examine 34 genes in our main target and reference genus *Drosophila* (Phylum Arthropoda, Order Diptera), as well as 23 genes in the divergent genus *Caenorhabditis* (Phylum Nematoda, Order Rhabditida). These genera represent two cases of independently evolved inheritance mode of PGC-specification, and each contains lineage-specific inheritance (LSI) genes, essential to germ plasm functionality (*osk*, *pgc* for *Drosophila* and *pie-1* and *pgl-1* for *Caenorhabditis*). We also study 30 genes in a second insect genus, *Apis* (Phylum Arthropoda, Order Hymenoptera), a taxon likely to use the induction mode. Collectively, the results provide insights into the molecular evolutionary dynamics of germ line genes within each of three distinct metazoan model genera. Most notably, the data show that germ line genes exhibit wide variation in constraint on protein sequence evolution and codon usage within each genus. Further, the LSI genes, which are essential to germ line function and found only in certain lineages, consistently exhibit a striking history of rapid protein sequence evolution relative to other germ line genes in each respective genus. We show this pattern may be explained by adaptive evolution and/or low pleiotropy.

Results and Discussion

We assessed the molecular evolution of 34 germ line genes across six species of the *melanogaster* group in our main target and reference genus *Drosophila* (*D melanogaster*, *D. erecta*, *D. sechellia*, *D. simulans*, *D. yakuba*, and an outgroup species *D. ananassae* (Additional file 1: Table S1 and Figure S1). The gene list included 34 genes with known experimental or cytological evidence of functionality in early germ line development, including PGC specification, and is provided in Table 1. For our study, the genes for *Drosophila* were grouped into four categories as follows: 1) the LSI genes *osk* and *pgc*, which are directly involved in germ plasm function, and found only in certain insects, including *Drosophila* [2, 19]; 2) genes involved in regulating PGC-specification under the inheritance mode ("Inheritance", *N*=13) in *Drosophila* (and other organisms) studied to date; 3) orthologs to genes found to be involved in inductive signalling mode in mice or other models ("Induction", *N*=15); and 4) genes involved in germ line formation regardless of mode ("Inh/Ind", *N*=4) (Table 1). In addition to studying these 34 germ line genes in *Drosophila*, we examined identifiable orthologs (see Methods) to this gene set in *Caenorhabditis* (for categories 2-4; four species studied, Tables S1 and S2), as well as two *Caenorhabditis* LSI genes (*pie-1* and *pgl-1*; *N*=23 genes total, Additional file 1: Table S2, see Methods), and orthologs found in the fellow insect genus *Apis* (N=30 genes; four species studied; Additional file 1: Table S3), which likely exhibits induction.

Molecular evolution in each genus was analyzed fully independently (using within-genus alignments), allowing us to evaluate how these genes evolve in each taxon (cf. [53, 54]). We determined dN/dS, dN and dS for each phylogeny using the free-ratio model (M1) in PAML [55], a model which allows all species branches to have independent values, and using the one-ratio model (M0) which estimates a phylogeny-wide value for each parameter [55–57]. Typically, dN/dS values >1, =1 and <1 indicate positive selection, neutral evolution and purifying selection, respectively [55]. However, even when <1, elevated values of dN/dS are indicative of greater rates of

Table 1 The 34 germ line genes studied in *Drosophila*. For this study, genes have been classified based on known roles in PGC-specification in established models. However, many genes (excluding LSI genes) have been reported to play roles in both germ line specification and at some stage(s) of germ line development or maintenance across model systems and specification modes [1, 3, 4, 13–16].

Gene	Gene Name	FlyBase ID	Length (codons)	Established PGC or Germ Line Role in Animal Models	References
Category 1: Lineage-Specific Inheritance (LSI) (*N*=2)					
osk	*oskar*	FBgn0003015	607	Assembly of germ plasm components	[2, 4, 18]
pgc	*polar granule component*	FBgn0016053	72	Localized to germ plasm; transcriptional repression in pole cells (DM)	[19, 20]
Category 2: Inheritance (*N*=13)					
armi	*armitage*	FBgn0041164	1189	*osk* functionality and mRNA translocation (DM)	[125]
bru-1	*bruno 1*	FBgn0000114	811	*osk* regulation: binding mRNA, translocation (DM)	[126]
capu	*cappuccino*	FBgn0000256	1362	Interacts with *spir*; Localization of *osk* mRNA, VASA and STAUFF to germ plasm, oocyte polarity (DM)	[127, 128]
cup	*cup*	FBgn0000392	1118	*Osk* regulation, with *bruno* (DM)	[129]
cycB	*cyclin B*	FBgn0000405	531	Cell cycle; mRNA localized to germ plasm (DM)	[20, 130]
gcl	*germ cell less*	FBgn0005695	570	Localized to germ plasm; required for PGC specification (DM)	[20]
mago	*mago nashi*	FBgn0002736	148	Required for PGC specification; localization of *osk* mRNA and STAUF to posterior pole of oocyte (DM)	[131]
orb	*oo18 RNA-binding protein*	FBgn0004882	916	Localized to germ plasm; *osk* regulation, oocyte polarity (DM)	[20, 132]
psq	*pipsqueak*	FBgn0263102	1124	Required for germ plasm formation	[133]
spir	*spire*	FBgn0003475	1021	Products are component of germ plasm (DM), interacts with *capu*	[127, 128]
stau	*staufen*	FBgn0003520	1027	Localization to germ plasm, PGC specification	[1, 127, 134]
tud	*tudor*	FBgn0003891	2516	Component of germ plasm, germ cell formation (DM)	[1, 135]
vls	*valois*	FBgn0003978	368	Component of germ plasm, involved in assembly (DM)	[136]
Category 3: Induction (*N*=15)					
Blimp-1	*Blimp-1*	FBgn0035625	1217	BMP signalling pathway for PGC specification, represses somatic expression (MM, GB)	[4, 5, 10, 17, 31, 137, 138]
bnl	*branchless*	FBgn0014135	771	FGF protein; mitogen for PGCs (MM)	[139, 140]
btl	*breathless*	FBgn0005592	1053	FGFR (FGF receptor), linked to PGCs (MM)	[140]
byn	*brachyury*	FBgn0011723	698	Activation of *BLIMP-1;* essential for PGC specification (MM)	[141]
dpp	*decapentaplegic (BMP 2/4-ortholog)*	FBgn0000490	589	BMP2/4 ortholog; involved in BMP-BLIMP1 signalling for PGC specification (MM, GB)	[5, 10, 32, 137]
gbb	*glass bottom boat*	FBgn0024234	456	BMP5/7/8 ortholog; involved in PGC specification (MM, GB)	[5, 32, 142]
mad	*mothers against dpp*	FBgn0011648	526	Ortholog to MM *smad* and GB *mad* genes; PGC specification (MM,GB)	[5, 142, 143]
med	*medea*	FBgn0011655	772	Ortholog to MM *smad4;* PGC induction (MM,GB)	[5, 142]
punt	*punt*	FBgn0003169	521	BMP receptor; putative ortholog in MM Bmpr2 (note; or Acvr2); associated with PGCs in GB	[32, 142, 144]; orthology match in FlyBase.org
sax	*saxophone*	FBgn0003317	583	BMP receptor; putative mammalian ortholog Bmpr1a/b (note: or Acvr) involved PGC specification in birds	[142]; orthology match in FlyBase.org
smox	*smad on X*	FBgn0025800	487	Orthologs are *smad2/3* in MM, which are transcriptional regulators needed for inductive PGC specification	[5]; orthology from FlyBase.org
sog	*short gastrulation*	FBgn0003463	1039	MM ortholog CHRD involved in BMP/CHRD signalling pathway, which relates to PGCs; Regulates DPP (DM)	[145, 146]; Orthology from FlyBase.org
tkv	*thick veins*	FBgn0003716	576	BMP receptor; mammalian ortholog (Bmpr1a/b) involved PGC specification in birds, GB	[142]; orthology match in FlyBase.org
wg	*wingless*	FBgn0284084	469	Ortholog to MM *wnt genes*; essential for PGC specification	[5, 147]

Table 1 The 34 germ line genes studied in *Drosophila*. For this study, genes have been classified based on known roles in PGC-specification in established models. However, many genes (excluding LSI genes) have been reported to play roles in both germ line specification and at some stage(s) of germ line development or maintenance across model systems and specification modes [1, 3, 4, 13–16]. *(Continued)*

Gene	Gene Name	FlyBase ID	Length (codons)	Established PGC or Germ Line Role in Animal Models	References
wit	*wishful thinking*	FBgn0024179	914	BMP receptor; putative ortholog to MM Bmpr2 (similar to punt), GB-punt is linked to PGCs in GB	[32, 144]; orthology match in FlyBase.org
Category 4: Inh/Ind (*N*=4)					
nos	*nanos*	FBgn0002962	402	Germ plasm component (DM), regulates mRNA, associated with PGCs in MM; common germ cell across metazoans	[1, 4, 17, 20, 148]
piwi	*P-element induced wimpy testis*	FBgn0004872	844	Component of germ plasm (DM); germ line essential across metazoans (DM, MM)	[1, 149]
pum	*pumilio*	FBgn0003165	1534	Regulate mRNA in PGCs (DM), conserved germ cell role in humans	[1, 150]
vasa	*vasa*	FBgn0283442	662	mRNA and protein localizes to germ plasm (DM, zebrafish, *C. elegans*), PGCS in Xenopus and MM	[1, 149, 151, 152]

NOTE: Gene identifiers are from the reference species *D. melanogaster*. Genes were placed in one of four categories based on their role in PGC-specification. Experimental or cytological evidence linking genes to lineage-specific inheritance (LSI), to Inheritance or Induction modes, or Inh/Ind are shown. This gene list was used as a reference to identify orthologs in five other *Drosophila* species from the melanogaster group, and in *C. elegans* and *A. mellifera* using reciprocal BLASTX. Length is for the full CDS per gene. Species abbreviations in cited evidence are *Drosophila melanogaster* (DM), *Gryllus bimaculatus* (GB), or *Mus musculus* (MM).

protein sequence evolution and reduced constraint. Thus, dN/dS provides a means to assess the relative constraint among genes involved in specific developmental processes [58]. The distributions of dN/dS, dN and dS for all genes studied in each of the three genera are shown in box plots in Additional file 1: Figure S2 and S3 (values for free-ratio model per branch are shown, see also Tables S4-S6). As indicated therein, dN and dS were unsaturated for all species branches in *Drosophila*, *Caenorhabditis* and *Apis* with values <<1 (Additional file 1: Figure S3A-F). The only species branch among all genera nearing saturation for dS was in the outgroup of *Drosophila*, *D. ananassae*, which had a median value of 1.158, and 25th and 75th percentiles values of 0.989 and 1.705, remaining in a suitable range for analysis [53, 59]. To study dN/dS of each gene per genus, we determined Mean dN/ Mean dS ($\overline{dN}/\overline{dS}$) across all terminal species branches per phylogeny from the free-ratio model. Values obtained using this approach were strongly correlated to the model M0 dN/dS values within each genus (Spearman's ranked correlation R=0.95, 0.99 and 0.99 for genes studied in *Drosophila*, *Caenorhabditis* and in *Apis* respectively, $P<2X10^{-7}$). We present results for $\overline{dN}/\overline{dS}$ throughout, as this provides a phylogeny wide measure of dN/dS, while allowing us to determine branch-specific values of dN and dS (Additional file 1: Figure S2 and S3).

Summary of Molecular Evolution of Germ Line Genes in Each Genus

We first summarize the patterns of $\overline{dN}/\overline{dS}$ and codon usage bias across germ line genes examined within each genus. The $\overline{dN}/\overline{dS}$ for each germ line gene across all six species branches in the *Drosophila* phylogeny and the four species in each of *Caenorhabditis* and *Apis* are reported in Tables 2, 3 and 4 (mean dN and mean dS values are provided in Additional file 1: Tables S4-S6). We found that $\overline{dN}/\overline{dS}$ varied extensively among genes within each genus. As an example, across all 34 germ line genes studied in *Drosophila* the $\overline{dN}/\overline{dS}$ values ranged from a high of 0.1573 in the LSI gene *osk* to a low of 0.0001 (Table 2) in *mago nashi (mago)*. Similar patterns were observed in the genus *Caenorhabditis*, with $\overline{dN}/\overline{dS}$ values ranging from a high of 0.1619 for the LSI gene *pie-1* to a low of 0.0081 for *mag-1*, the ortholog to *Drosophila mago*. For *Apis*, values varied from 0.0001 to 0.1393, for the ortholog to *mago* (tied with *mad* and *nos*) and *piwi* respectively. Together, these patterns show there has been marked variation in selective pressures on germ line genes in each of the three genera, and that these germ line genes are not a homogenous group sharing similar selective profiles.

In addition to $\overline{dN}/\overline{dS}$, we measured codon usage bias (CUB) for each of the germ line genes within each genus. The effective number of codons (ENC) determines the deviation from equal usage of all codons in a gene, wherein values range from 20 to 61 (number of codons in the genetic code), and lower values indicate greater codon usage bias [60]. We used the modified ENC' measure (mENC'), which accounts for abundance of rare amino acids and for nucleotide content (mutational biases) of the genes under study [61, 62]. Preferential usage of codons is thought to usually (but not always) result from weak but persistent selection pressures that promote efficient and accurate transcription

Table 2 $\overline{dN}/\overline{dS}$ and mENC' across the phylogeny of six species of the reference model *Drosophila*

Gene	$\overline{dN}/\overline{dS}$	Mean mENC'	SE	Alignment Length (codons)
Lineage-specific Inheritance *N*=2				
osk	0.1573	52.18	0.29	580
pgc	0.0933	37.25	0.66	71
Inheritance *N*=13				
capu	0.1189	54.26	0.35	1007
orb	0.1102	54.09	0.25	789
stau	0.0885	53.59	0.19	990
cup	0.0878	54.23	0.55	1048
vls	0.0810	51.02	0.44	367
armi	0.0773	54.28	0.44	888
spir	0.0751	54.34	0.17	1001
tud	0.0716	53.77	0.38	1446
cycB	0.0692	50.23	0.41	509
gcl	0.0392	55.13	0.20	568
bru-1	0.0369	51.39	0.19	723
psq	0.0297	49.71	0.27	469
mago	0.0001	42.29	0.72	147
Induction *N*=15				
bnl	0.1303	53.17	0.43	494
btl	0.0935	53.99	0.63	1040
wg	0.0833	47.48	0.95	261
tkv	0.0734	50.42	0.32	561
byn	0.0575	49.71	0.23	651
Blimp-1	0.0572	51.06	0.44	1110
dpp	0.0517	51.64	0.24	430
wit	0.0462	53.52	0.39	897
med	0.0422	52.47	0.57	759
sax	0.0402	50.88	0.80	565
sog	0.0305	49.45	0.48	919
gbb	0.0303	49.58	0.72	451
mad	0.0248	49.75	0.54	510
punt	0.0232	52.42	0.31	433
smox	0.0135	50.54	0.17	468
Inh/Ind *N*=4				
nos	0.1145	50.51	0.38	389
vasa	0.0545	50.59	0.70	572
piwi	0.0446	54.91	0.76	711
pum	0.0320	50.53	0.38	798

NOTE. Results are shown for each of the 34 genes under study a measured using codeml in PAML [55]. Genes are classified based on their role in PGC specification and ranked by $\overline{dN}/\overline{dS}$ within each group. The mean dN and mean dS values are provided in Table S4. SE=standard error for mENC'.

Table 3 $\overline{dN}/\overline{dS}$ and mENC' across the phylogeny of four species of *Caenorhabditis*

CE Gene	DM Gene	$\overline{dN}/\overline{dS}$	Mean mENC'	SE
Lineage-specific Inheritance *N*=2				
pie-1	-	0.1619	46.13	0.42
pgl-1	-	0.1553	50.29	0.45
Inheritance *N*=8				
cyb-2	*cycB*	0.1127	45.01	0.68
cpb-3	*orb*	0.0833	51.56	1.40
gcl-1	*gcl*	0.0719	52.78	0.58
ifet-1	*cup*	0.0570	51.68	0.58
stau-1	*stau*	0.0499	52.63	0.80
par-1	*par-1*	0.0433	51.35	0.46
etr-1	*bru-1*	0.0386	50.54	1.52
mag-1	*mago*	0.0081	38.41	1.79
Induction *N*=10				
dbl-1	*dpp*	0.0963	51.18	0.80
let-756	*bnl*	0.0748	49.36	1.09
sma-6	*sax*	0.0675	50.00	1.30
egl-15	*btl*	0.0639	52.74	0.37
blmp-1	*Blimp-1*	0.0575	52.56	0.41
sma-4	*med*	0.0548	52.34	0.67
tig-2	*gbb*	0.0522	50.24	1.15
crm-1	*sog*	0.0395	47.73	0.45
cwn-1	*wg*	0.0264	49.69	0.42
sma-2	*mad*	0.0128	46.30	0.78
Inh/Ind *N*=3				
glh-1	*vasa*	0.0761	43.98	1.48
puf-8	*pum*	0.0753	50.79	0.92
prg-1	*piwi*	0.0642	46.13	1.81

NOTE. Values are shown for each of the 23 genes under study as measured using codeml in PAML [55]. Genes are ranked from highest to lowest $\overline{dN}/\overline{dS}$ values within each category. *CE Caenorhabditis elegans*, *DM D. melanogaster*. The inheritance gene *par-1* was included for *Caenorhabditis* (see Methods). *SE* standard error for mENC'.

and/or translation, as has been reported in studies of *Drosophila*, *Caenorhabditis* and *Apis* [63–68].

The results showed that of all 34 germ line genes studied in *Drosophila*, the highest and lowest CUB were observed in the two LSI genes, *pgc* and *osk* respectively (mENC'=37.25±0.66 and 52.18±0.29 respectively, values are means and standard errors across species, Table 2). This result reveals marked differences in the degree of codon usage bias of the two LSI genes. In turn, 25 of the remaining 32 studied genes exhibited mENC' values that were >50, implying comparatively low CUB, with respect to *pgc*. For the genus *Caenorhabditis*, among the 23 germ line genes studied in that genus, the CUB was highest for *mag-1* (mENC'=38.41±1.79) and lowest for

Table 4 $\overline{dN}/\overline{dS}$ and mENC' across the phylogeny of four species of *Apis*

Gene	$\overline{dN}/\overline{dS}$	Mean mENC'	SE
Inheritance N=13			
armi	0.1011	54.66	0.13
tud	0.0792	56.67	0.27
cycB	0.0774	53.24	0.07
vls	0.0673	51.78	0.34
spir	0.0564	56.22	0.05
bru-1	0.0560	53.13	0.15
stau	0.0395	51.24	0.10
gcl	0.0326	53.42	0.27
capu	0.0238	53.39	0.53
orb	0.0192	55.04	0.13
psq	0.0173	54.07	0.32
par-1	0.0067	54.84	0.25
mago	0.0001	46.43	0.39
Induction N=13			
dpp	0.1342	52.24	0.14
sax	0.1163	51.80	0.24
wg	0.0854	48.68	0.48
sog	0.0627	55.22	0.21
wit	0.0462	55.48	0.15
Blimp-1	0.0375	49.42	0.72
gbb	0.0363	46.37	0.40
punt	0.0335	52.73	0.39
byn	0.0324	53.20	0.25
med	0.0133	52.72	0.26
tkv	0.0116	53.91	0.09
smox	0.0080	52.44	0.36
mad	0.0001	51.63	0.29
Inh/Ind *N*=4			
piwi	0.1393	54.72	0.39
vasa	0.1155	53.33	0.31
pum	0.0038	53.84	0.29
nos	0.0001	38.89	0.96

NOTE. Values are from codeml in PAML [55]. Genes are listed using the ortholog name from *D. melanogaster*. The inheritance gene *par-1* was included for *Apis* (see Methods). *SE* standard error for mENC

gcl-1 (52.78±0.58), with intermediate values observed for its two LSI genes (Table 3). The broadest range of CUB was observed within the genus *Apis* where mENC' values ranged from 38.89±0.96 to 56.67±0.27 (Table 4), suggesting a propensity for greater variation in CUB of germ line genes in that taxon.

In sum, it is evident that germ line genes exhibit wide variation in selective pressures on protein sequence divergence and in codon usage bias within each of the three genera under study here (see the below section "*Relative Ranking of dN/dS and CUB Between Genera*" for details on how dN/dS and CUB compared between genera). In this regard, the germ line genes are not a homogenous group exhibiting a similar range of $\overline{dN}/\overline{dS}$ values or common CUB profiles. Sex and reproductive gene proteins are thought to often evolve rapidly, particularly those involving gametogenesis and sperm and eggs [69, 70]. However, our results showed that a subset of germ line genes had $\overline{dN}/\overline{dS}$ <0.05, which suggests very high purifying selection (Tables 2, 3 and 4), and thus that pattern may not broadly apply to genes involved in germ line specification or development (see also, [50, 51]). Nonetheless, certain germ line genes studied here, such as the LSI genes, exhibited comparatively rapid evolution, suggesting they may be particularly significant to the evolutionary changes of the molecular mechanisms regulating germ lines in each genus.

Rapid Evolution of LSI Genes in *Drosophila* and in *Caenorhabditis*

Accelerated Divergence of LSI Genes in Drosophila

The LSI germ line genes are of particular interest as they may be crucial to enhancing our understanding the evolution of germ plasm in animals, since they appear to be *de novo* genes that have arisen and developed specialized germ line functions within only certain animal lineages. Our results show that the proteins of LSI genes are among the most rapidly evolving of the 34 germ line genes under study in *Drosophila*. Specifically, the finding that the highest $\overline{dN}/\overline{dS}$ in *Drosophila* was observed for the LSI gene *osk* (0.1573, Table 2) suggests it may have experienced the least constraint among all the studied germ line genes in this genus. The *osk* gene is involved in germ plasm assembly and has only been reported to date in *Drosophila* and certain insects [1, 2, 4]. Osk proteins are essential for recruitment of molecules to germ plasm (e.g. via direct interactions with *Vasa* and *Staufen* proteins) and have RNA-binding functions to *nos* and its own mRNA [18, 42, 71, 72]. The $\overline{dS}$ value for *osk* was intermediate (0.2779), near the median of values across all germ line genes (0.2580; Additional file 1: Table S4). However, its $\overline{dN}$ value (0.0437) was the highest observed across all studied genes (Additional file 1: Table S4), affirming that its elevated $\overline{dN}/\overline{dS}$ value is due to accelerated protein sequence divergence. Such high nonsynonymous changes, if not adaptive, would be apt to often be deleterious in a gene essential for fecundity and fitness [73], and thus would be unlikely to be fixed. In this regard, the high $\overline{dN}/\overline{dS}$ in *osk* appears potentially to be the result of episodic adaptive evolution. That is, changes in amino acids of protein sequences may have been

retained due to positive selection, via a selective advantage of the phenotypes associated with these protein sequence changes [55, 74–76]. Analysis of positive selection using sites analysis in PAML (see Methods, [55]), which we used to test for positive selection at specific codon sites in each gene across all six *Drosophila* species, did not show positive selection within this LSI gene (Table 5). This is similar to results reported in a prior assessment of the gene *osk* within the *melanogaster* group [52] (the previously reported sites under putative positive selection were reported from a branch site test for *D. virilis*, a species outside of the *melanogaster* group and thus not studied herein [52]). However, positive selection analyses (in specific branch-sites or sites) can be highly conservative [56, 76, 77], and often lack sensitivity to detect functional changes [78]. In this regard, positive selection cannot be excluded by the absence of statistically significant sites tests.

Here, based on several lines of evidence we propose that adaptive evolution may have contributed towards the observed rapid protein sequence evolution of *osk*, as compared to all other 33 *Drosophila* germ line genes studied herein (Table 2). First, experimental findings have shown that functionality of *osk* has evolved rapidly, namely based on an inability of the gene in one *Drosophila* species (*D. virilis*) to rescue loss of function mutations in *D. melanogaster* [79], although these two species shared a last common ancestor > 55 My [80]. Second, some functional binding regions within the gene have been shown to have at least two-fold higher dN/dS than other segments [52], which may be deemed consistent with non-random, and thus putatively adaptive, changes. Third, relaxed selection appears unlikely to explain the relatively fast evolution of *osk* given that this lineage-restricted *de novo* gene has evolved crucial functions, involving high protein and RNA interactivity, in the germ plasm and during PGC-specification [2, 4, 18, 42, 71, 72]. These functions may act as constraints that limit relaxed purifying selection. Given that *osk* evolved before the advent of germ plasm in insects [81], such that the ancestral role of *osk* was unlikely to have been be a germ plasm role, we can infer that the evolution of essential roles in germ plasm must have been due to changes that arose following its origin approximately 300 Mya [79, 81], and such episodic changes may have been ongoing in the ~44 Mya history of the melanogaster group of *Drosophila* studied here [80]. Fourth, adaptive changes linked to germ line functions may have been facilitated by the low expression breadth observed for *osk* across development (as shown in the below section, "Expression breadth and pleiotropy in *Drosophila*"). Low pleiotropy appears characteristic of functional *de novo* genes [21, 82, 83], and while this might sometimes cause relaxed selection, it can also allow adaptive protein changes with minimal interference from functions in other tissues [84]. Collectively,

Table 5 Results from sites analysis of positive selection (M7 versus M8) in *Drosophila*, *Caenorhabditis* and *Apis*

Gene	Role	*P*	Sites (BEB Probabilities)
Drosophila			
byn	Induction	**	29 S**
capu	Inheritance	**	252 P**, 253H**, 437 T*
Caenorhabditis			
cpb-3	Inheritance	**	423 N*
let-756	Induction	**	231 S*
par-1	Inheritance	**	468 A**, 625 N*, 647 Q*, 701 G**, 704 T**,705V*
Apis			
bru-1	Inheritance	**	337 T*, 339 A*, 341 A*
cycB	Inheritance	**	4 G**, 5 L**
dpp	Induction	*	80 S*, 81 T*, 86 Q**, 87 L *
med	Induction	**	-
orb	Inheritance	**	-
piwi	Inh/Ind	**	153 H**,318 T** ,320 A **
psq	Inheritance	**	-
pum	Inh/Ind	**	306 Y*
stau	Inheritance	**	-
wit	Induction	**	329 I*

NOTE: Analysis was conducted in PAML [55]. Only those genes at or near statistical significance are shown. ** $P<0.05$ for 2XΔlnL *$P=0.062$. BEB probabilities for specific sites are indicated as ** $P\geq0.95$ (**) and *$0.95\geq P>0.9$

these four lines of evidence suggest that the rapid divergence of *osk* observed here is at least partly the result of a history of episodic adaptive evolution within the *Drosophila* genus.

The other *Drosophila* LSI gene, *pgc*, also evolved relatively rapidly compared with all 34 *Drosophila* genes studied ($\overline{dN}/\overline{dS}$ = 0.0933, ranked 7th of 34 genes). *pgc* encodes a small protein (72 codons in *D. melanogaster*, Table 1) that is involved in transcriptional silencing in PGCs by preventing phosphorylation of RNA PolII and by inhibiting recruitment of the positive transcription elongation factor b (P-TEFb) to transcription sites. The suppression of expression is crucial for preventing germ cells from differentiating into somatic cells, and thus for sustaining the germ line [19]. Our alignment for six species is nearly identical to that produced by Hanyu-Nakamura et al. [19] who studied functionality of this protein in *Drosophila* (including more divergent taxa) and suggested that it is highly conserved. Our findings showing $\overline{dN}/\overline{dS}$ of 0.0933, which is <<1, concurs with conservation, but nonetheless, indicates that with respect to other genes involved in germ line development, this protein evolves notably rapidly. We note that we examined the dN/dS of all orthologous protein-coding genes in the Drosophila clade studied here using the data provided by the flyDIVaS resource [57], and found that *pgc*, like *osk*, was above the average observed across the genome (see Additional file 1: Text file S1). As *pgc* is believed to be restricted to *Drosophila* and is not found in other metazoans, including within fellow Dipteran insects [19], its *de novo* status as a germ plasm regulator, similar to *osk*, appears to be a factor potentially shaping its fast divergence as compared to most other germ line genes.

With respect to codon usage, as described earlier the mENC' value was extremely low for *pgc* in *Drosophila* (37.25±0.66, Table 2), in fact lowest among all 34 genes studied in this insect and was comparatively higher in *osk* (52.18±0.29; Mann Whitney U-test $P<0.002$), suggesting CUB is under greater selective constraint in *pgc*. This is also supported by higher GC3s in *pgc* than *osk* (mean values of 0.712±0.015 and 0.652±0.010 respectively, MWU-test P=0.004, Additional file 1: Table S7), indicating greater use of optimal codons, which have been shown to typically end in G and C in these *Drosophila* taxa [66, 85]. Thus, the greater CUB in the former gene might reflect a crucial role of efficient translation of this gene, potentially due to high translation rates during PGC formation. Alternatively, it might also be connected to the exceptionally short CDS length of *pgc*, which comprises the shortest CDS under study (Table 1), a feature which has been proposed to be linked to elevated selection coefficients on codon usage [86] and to be associated with greater CUB [63, 66].

Accelerated Divergence of LSI genes in Caenorhabditis

With respect to the nematode genus *Caenorhabditis*, we studied two LSI genes, *pie-1* and *pgl-1*, each of which showed markedly elevated $\overline{dN}/\overline{dS}$ among the 23 germ line genes under investigation in this genus (Table 3). For instance, as shown in Table 3, *pie-1* and *pgl-1* in *Caenorhabditis* had the two highest $\overline{dN}/\overline{dS}$ values of all 23 germ line genes under study (0.1619 and 0.1553). This suggests that, similar to the situation in *Drosophila*, genes specifically involved in germ plasm, and that arose independently within a lineage, have experienced accelerated divergence as compared to other germ line genes. *pie-1* and *pgl-1* products are localized to the P granules and essential to fertility, have not been reported in other metazoans and are thought to be specific to *Caenorhabditis* [3, 4, 22, 29, 87, 88]. PIE-1 is crucial to germ line establishment as it represses mRNA transcription in germ line blastomeres and prevents differentiation into somatic cells by inhibiting activity of P-TEFb, which ultimately impedes phosphorylation of RNA PolII and prevents transcription elongation [19, 87, 89]. In this regard, *pie-1* shares functionality (but not sequence homology) with the rapidly evolving *Drosophila* gene *pgc* (Tables 2 and 3) [19]. Remarkably, our data here shows that *pie-1* and *pgc*, which are *de novo* genes with convergent functions that have arisen independently in *Caenorhabditis* and in *Drosophila* respectively [4, 19], each diverge rapidly as compared to other germ line genes in their associated taxonomic group (Tables 2 and 3).

The relatively rapidly diverging *Caenorhabditis* LSI gene *pgl-1* is essential in the P granule assembly pathway [3, 88, 90]. Its protein contains RGG-binding motifs (which are linked to genes involved in transcription, translation, splicing) similar to those found in the protein product of *vasa* (*glh-1* in *Caenorhabditis*) [3, 22]. The P granule pathway presumably involves the localization of *pgl-1* products to the P granules by *glh-1*, as mutants of the latter gene contain PGL proteins dispersed throughout the cytoplasm [3]. We found that *glh-1* had a $\overline{dN}/\overline{dS}$ of 0.0761, which is about half the value of *pgl-1* (0.1553). This suggests that *pgl-1* has evolved at a much faster rate than its localization protein, whilst presumably not interfering negatively with their interaction. While *pgl-1* is part of a gene family with other members *pgl-2* and *pgl-3* that may have arisen from gene duplication, the other paralogs are, unlike *pgl-1*, not essential to P granules or PGCs, as *pgl-2* and *pgl-3* loss of function mutants do not exhibit obvious defects in germ line development on their own [3, 90]. *pgl-3*, however can be partially redundant to *pgl-1* under cold temperatures [3, 90]. Thus, the rapid evolution of

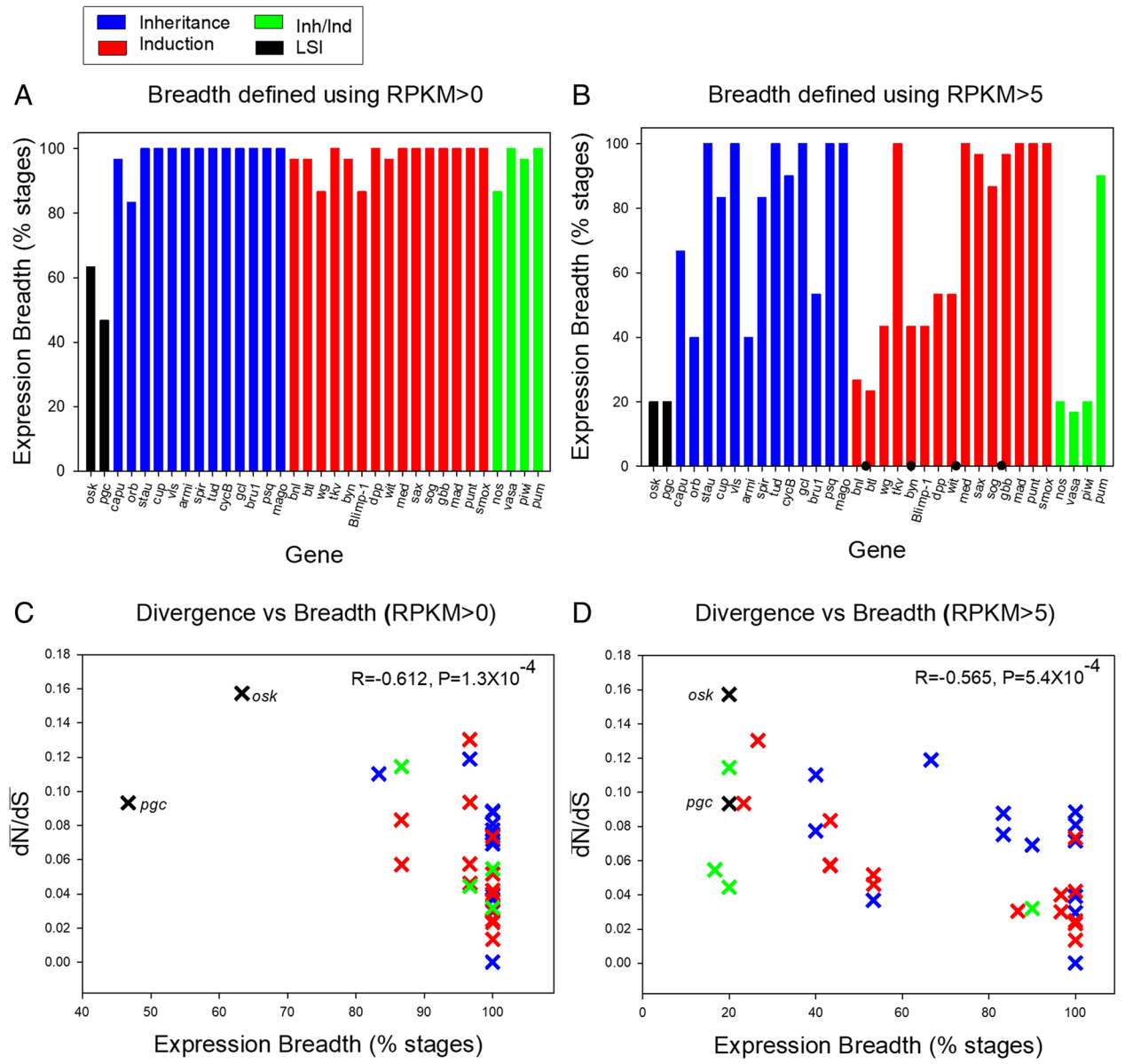


Fig. 1. The expression breadth (percentage of 30 developmental stages expressed) for the 34 *Drosophila* genes under study and its relationship to protein divergence. **a** breadth of expression across developmental stages at a level >0 RPKM; **b** breadth of expression at a level >5 RPKM; **c** $\overline{dN}/\overline{dS}$ versus expression breadth (RPKM>0); **d** $\overline{dN}/\overline{dS}$ versus expression breadth (RPKM>5). The 30 tissues and stages are provided in Table S8. For A-B, genes are listed on the X-axis in the same order as presented in Table 2, from highest to lowest $\overline{dN}/\overline{dS}$ within each of the four categories of genes. Gene names associated with each dN/dS point are provided in Table 2.

pgl-1 might partly result from some redundancy (or relaxed selection) of function under specific conditions. However, the fact that the other LSI genes studied in *Drosophila* and in *Caenorhabditis* do not have apparent paralogs, and each evolve relatively fast among the germ line genes (Tables 1 and 2), suggests that accelerated protein divergence is a common feature of the lineage-restricted germ plasm genes, rather than being an artefact due to the existence of a partially redundant paralog of *pgl-1*.

With respect to CUB, *pie-1* exhibited greater bias (mean mENC' 46.13±0.42) than *pgl-1* (50.29±0.45; MWU test P=0.029), suggesting enhanced selection on CUB in the former gene, perhaps reflecting a higher translation rate. However, similar to *pgc* from *Drosophila*, the short CDS of *pie-1* might contribute to its high CUB. Short CDS not only generally exhibit high CUB, but sometimes also low dN/dS (and/or dN), trends perhaps mediated by protein-protein interactions and/or elevated expression levels [63, 86,

91, 92]. Whilst *pie-1* is markedly shorter than *pgl-1* (note the average lengths per gene across species were 351±8,9 and 770±4.8 codons, a nearly two-fold difference; the conservative alignment lengths had nearly three-fold difference) consistent with elevated CUB, the $\overline{dN}/\overline{dS}$ values were very similar between genes (Table 3), thus suggesting while length might influence the relative CUB, it is unlinked to their amino acid divergence in *Caenorhabditis*. Alternatively, it is possible that adaptive evolution at specific amino acid sites has occurred more frequently for *pgl-1*, leading to selective sweeps at linked sites containing slightly deleterious non-optimal codon mutations [93], which over the long-term can reduce CUB [93, 94]. Although positive selection was not found in the four *Caenorhabditis* species studied here for these two genes using sites analysis (Table 5), further studies including even more species, as data becomes available, will allow greater power of these tests and fuller discernment of the role of positive selection in these genes.

In our assessment, we also wished to examine the *Caenorhabditis* LSI gene *meg-1*, which is a P-granule component required for germ line development [27], but found that identification of *meg-1* orthologs for all four species was ambiguous. The best matches to the *C. elegans meg-1* had an e-value of $5.0X10^{-5}$ for *C. brenneri*, $6.0X10^{-4}$ for *C. briggsae* and 0.053 for *C. remanei*, and were largely unalignable across most of the sequence. We therefore excluded this gene from analysis herein. The lack of clearly identifiable orthologs is suggestive of rapid divergence or potential gene loss in nematodes, or might indicate that this *de novo* gene occurs solely in a single species, *C. elegans* [27]. While *meg-1* is an LSI gene, multiple copies (*meg-1-4*) have been reported in *C. elegans* with partial overlap in function [28], and this might explain potential rapid sequence evolution and/or gene losses in some species of this genus.

Summary of Findings on LSI Genes

Taken together, the collective results from LSI genes from *Drosophila* and *Caenorhabditis* (*osk*, *pgc*, *pie-1* and the *pgl-1*) suggest all these *de novo* genes evolve rapidly as compared to other studied germ line genes. As these four genes appear not to have originated from a gene duplication due to the absence of ancestral orthologs [whilst *pgl* has multiple paralogs in Caenorhabditis, *pgl* genes appear limited to nematode genomes; 3], we speculate that these LSI protein-coding genes might have arisen at least partially from noncoding regions [95–98] or horizontal gene transfer [21, 99]. In fact, recent evidence has supported a putative role of horizontal gene transfer in the origin of *osk* [100]. *De novo* genes have been previously linked to expression in sexual organs, including germ lines [21, 95, 96]. Nonetheless, whilst it appears these LSI genes have not arisen from duplication, we can neither formally exclude nor directly test the hypothesis that they originally arose from a duplication and evolved so rapidly that the orthologs cannot be identified [21].

Regardless of the precise origin, these four *de novo* genes have not degenerated into pseudogenes or been loss from the genome due to lack or loss of function, as frequently occurs for orphan genes [101], but rather play a crucial role in PGC specification and thus fertility. A plausible explanation for their existence and evolution of novel functionalities is their involvement in lineage-specific adaptive processes [102], potentially accompanied by phenotypic novelties, as is thought to occur for *de novo* genes that become functional [98, 102, 103]. For LSI genes, the adaptations would involve their crucial roles in germ plasm, which is believed to be a novelty in the context of nematodes and insects [7]. This notion is further consistent with recent findings in *Drosophila* that surviving (not lost from the genome) *de novo* genes exhibit functionalities specific to a narrow developmental phase or tissue type [82]: germ plasm and the PGCs are limited to the stages involving the egg or early embryo. It can be speculated that adaptive amino acid changes in LSI genes might have been historically mediated by sexual selection, as the pre-formed germ plasm could conceivably indirectly influence sperm-egg fertilization success (and thus, sexual antagonism), from natural selection on germ plasm due to its effect on zygotic or embryonic fitness, and/or from cell-lineage selection among the precursors to PGCs or among PGCs, each of which could accelerate protein sequence divergence [30].

We speculate that a history that includes episodic adaptive evolution in the emergence of functions in the *de novo* LSI genes may have potentially continued after the establishment of their primitive germ plasm roles and extended to within the intra-genus level. This notion appears consistent with multiple lines of evidence for *osk* (Table 2) [52, 79], and could account for elevated $\overline{dN}/\overline{dS}$ observed for all four of the LSI genes studied here (Tables 2 and 3). Nonetheless, we do not exclude a role of neutral functional or non-functional amino acid changes in the rapid divergence of LSI genes relative to other germ line genes, which appears consistent with the absence of detection of positive selection in Table 5; however, such tests can be highly conservative [77] and be prone to substantial inaccuracies [104, 105] (see Additional file 1: Text file S2), and thus cannot reliably exclude positive selection. At present, much remains to be unknown about the evolution of *de novo* functional genes, including about the roles of adaptive and neutral changes in those genes that that form essential roles in genetic networks [21, 97, 98], such as has occurred for the LSI genes involved in germ plasm.

Expression Breadth and Pleiotropy in *Drosophila*

Protein sequence evolution may be influenced by pleiotropy, where the greater a gene's involvement in multiple functions or tissues, the more indispensable it may be to fitness, and thus more likely to exhibit stringent purifying selection [83, 84, 106]. In turn, those genes with reduced pleiotropy may be more dispensable and evolve faster due to relaxed selective pressures (neutral changes), and/or due to adaptive functional changes unimpeded by pleiotropic constraints [84, 107]. Expression breadth across developmental stages and tissues provides a proxy for the pleiotropy of a gene [83, 84, 106]. Further, data from Drosophila suggests that young *de novo* gene products which exhibit functional roles may often have those roles restricted to one or a few developmental stages, and thus may be specialized for specific developmental processes [82]. We thus determined expression breadth across developmental stages/tissues with expression using the large-scale transcriptome data available for our main target and reference taxon *Drosophila*, from its most well studied model species, *D. melanogaster* [48, 108, 109]. For each of the 34 *Drosophila* germ line genes in Table 2, we measured expression breadth across 30 developmental stages (spanning from 0-2 hour embryos, larvae, pupae, to adult males and females, described in Additional file 1: Table S8).

As shown in Fig. 1a, we found that the majority of germ line genes in Table 2 (N=22 of 34) were expressed ubiquitously at a level of >0 RPKM in all the disparate stages (expression breadth=100%; N=28 had values of >95%), indicating high pleiotropy in these germ line genes. The LSI genes *osk* and *pgc* had the lowest values, with a breadth of 63.3 and 46.7% respectively. Using a higher cutoff of >5 RPKM to define specificity (Fig. 1b), we observed even more variation in expression breadth; most genes had values between 40-100%, but remarkably low breadth was observed for Inh/Ind genes such as *vasa* and *piwi* and again for the two LSI genes, which had values of 20% (Fig. 1b). With respect to $\overline{dN}/\overline{dS}$,, we observed a negative correlation between expression breadth and $\overline{dN}/\overline{dS}$ using both criteria of >0 RPKM and >5 RPKM (Spearman R=-0.612, P=1.3X10^{-4} and R=-0.565, P=5.4X10^{-4} respectively, (Fig. 1c and d) consistent with higher protein divergence in narrowly expressed genes. In contrast, no correlation was observed between $\overline{dN}/\overline{dS}$ and the average expression level across developmental stages (P=0.926). mENC' was uncorrelated to expression breadth (P=0.498). Together, these data suggest that most of the germ line genes under study exhibit broad expression, or high pleiotropy. In turn, low expression breadth may substantially contribute towards the accelerated evolution of the LSI genes.

As an example, the extraordinarily low expression breadth for the LSI genes *osk* and *pgc*, using both the criteria of >0 RPKM (Fig. 1a and c) and >5 RPKM (Fig. 1b and d), indicates high specialization and limited pleiotropy of those genes, consistent with their specialized functions in germ plasm. For both genes, the only stages with relatively high expression (for instance, using a cut-off of >15 RPPM) were adult females (42 to 372 RPKM) and 0-4 hour embryos (17 to 270 RPKM), consistent with their PGC specification roles in female sexual cells and young embryos. This observed pattern concurs with the notion that young *de novo* genes exhibit expression limited to one or a few developmental stages as they acquire new functions in an organism [21, 82]. Further, such *de novo* genes may become essential if they integrate into existing networks [21], as appears to be the case for *osk* and *pgc* where these genes have assumed an upstream regulatory role in the pathway of PGC-specification and germ line development [4, 17]. Moreover, their low pleiotropy may facilitate evolution of new functions by permitting adaptive evolution largely unimpeded by co-functionality in other tissues [84], and/or may contribute towards divergence via drift (see above section "Rapid evolution of LSI genes in *Drosophila* and in *Caenorhabditis*"). Further consideration of the putative role of pleiotropy with respect to the $\overline{dN}/\overline{dS}$ of specific germ line genes is provided in Additional file 1: Text File S3.

Relative Ranking of dN/dS and CUB Between Genera

After considering the evolution of LSI genes, we next asked whether the germ line genes studied here shared parallels in their patterns of molecular evolution across genera. For this, we assessed whether the relative ranking of $\overline{dN}/\overline{dS}$ and of mENC' were similar between our reference model *Drosophila* and *Caenorhabditis* and *Apis*. By examining the relative ranking of dN/dS and of CUB, between genera using Spearman's Ranked R (not absolute values), this controls for taxon-specific factors, and allows us to assess if the relative dN/dS and CUB within this group of germ line genes has been retained across genera.

As shown in Fig. 2, for the *Drosophila* genes having orthologs identified in *Caenorhabditis*, we found $\overline{dN}/\overline{dS}$ was positively correlated between genera (pooled across all categories, excluding the LSI genes, Spearman's rank correlation R=0.44, P=0.05, N=20, Fig. 2a). However, the correlation was not statistically significant for the orthologs between *Drosophila* versus *Apis* (P=0.215, Fig 2b; note, nor for those common to *Caenorhabditis* vs. *Apis* (P=0.112)). Reducing the *Drosophila-Apis* ortholog gene list (which was larger than for the distant nematodes, Fig. 2, Tables 3 and 4) to those also found in *Caenorhabditis*, also did not yield a correlation between *Drosophila* and *Apis* (P=0.181). Thus, this indicates that the germ line gene sets share greater similarity in the relative protein divergence

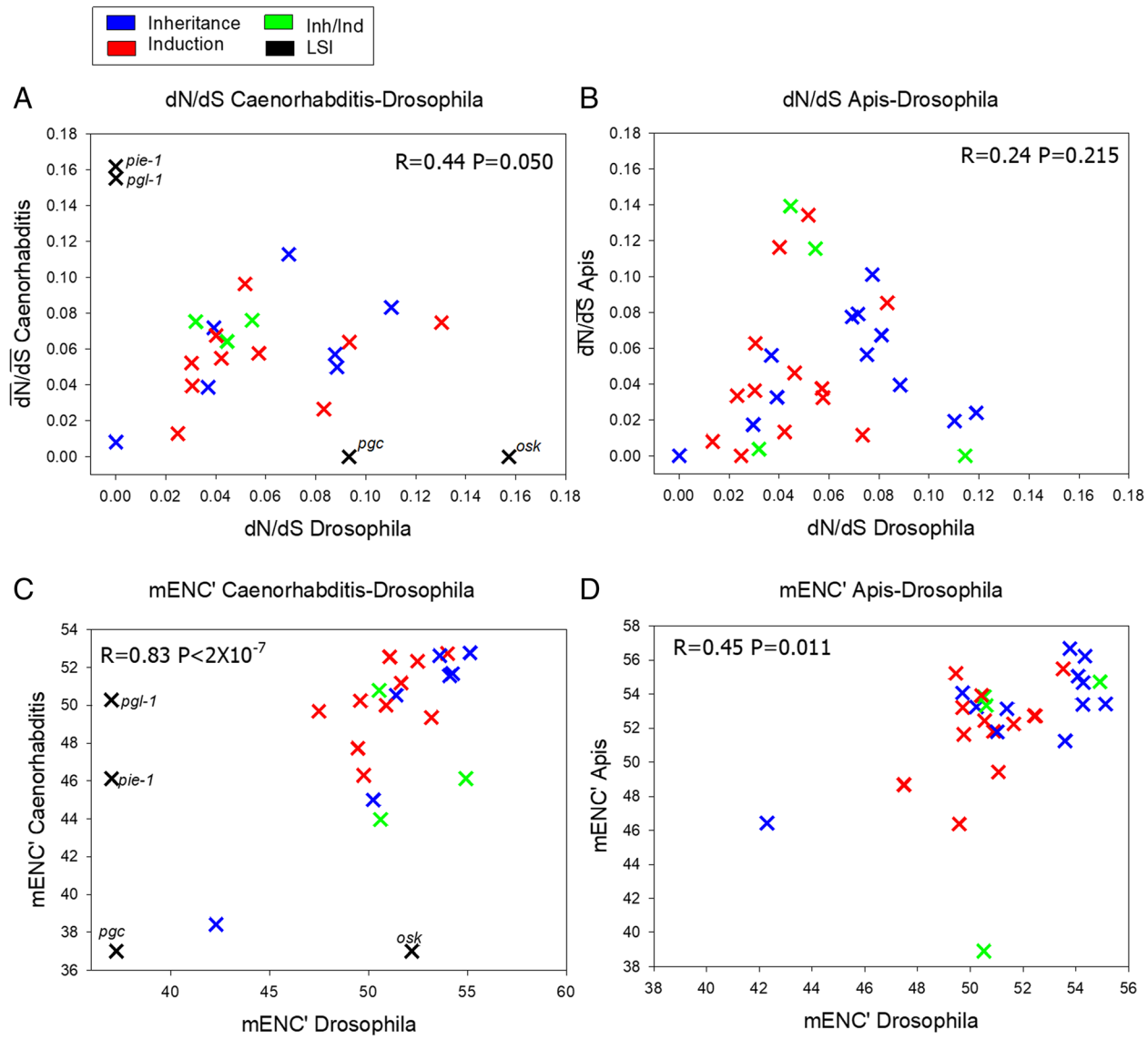


Fig. 2. Correlations in $\overline{dN}/\overline{dS}$ and in mENC' between *Drosophila* and *Caenorhabditis* and *Drosophila* and *Apis*. **a** The $\overline{dN}/\overline{dS}$ values for *Caenorhabditis* versus those from *Drosophila*; **b** the $\overline{dN}/\overline{dS}$ values for *Apis* versus those from *Drosophila*; **c** mENC' for *Caenorhabditis* versus *Drosophila*; **d** mENC' for *Apis* versus *Drosophila*. Spearman's ranked R and P values are shown across all genes for the between genus orthologs (pooled across categories). The lineage-specific inheritance (LSI) genes in *Drosophila* and in *Caenorhabditis* genes are shown on the X and Y axis respectively in A and C and were not included in the calculation of between genus correlations. The gene names associated with each value of $\overline{dN}/\overline{dS}$ and mENC' shown are provided in Tables 2, 3, and 4.

between *Drosophila* and its distant relative *Caenorhabditis* (from different phyla) than between the two insects. The values of $\overline{dN}/\overline{dS}$ in *Drosophila*, *Caenorhabditis* and *Apis* were in a largely similar range across all the three genera examined here (Tables 2, 3 and 4, Fig. 2). Thus in that context the magnitude of $\overline{dN}/\overline{dS}$ did not appear to vary among taxa, but rather the relative dN/dS among genes simply appeared more conserved between *Drosophila* and *Caenorhabditis* than either was to *Apis*. Previous reports have shown that *Drosophila* and *Caenorhabditis*, despite being from different phyla, share striking parallels in gene expression profiles and networks across developmental stages [108]. In this regard, our present data suggest these divergent systems also share similar relative protein divergence patterns of their germ line genes. One obvious similarity between these organisms with respect to germ lines that speculatively could contribute towards this latter pattern is the use of the inheritance mode in both models. However, further study in more genera

would be needed to ascertain whether specification mode plays any role in these shared patterns.

In terms of CUB, we report a strong positive correlation between mean mENC' of genes in *Drosophila* and their orthologs in *Caenorhabditis* (R=0.83, P=2X10^{-7}) and also a significant positive correlation between *Drosophila* and *Apis* (R=0.45, P=0.011) (Fig. 2c and d). This suggests the relative CUB in each germ line gene set has been largely retained across these disparate organisms, particularly between the flies and nematodes. We note that the mENC' levels were fairly high for many germ line genes (values >50 for genes), indicating that these germ line genes as a group do not exhibit exceptionally strong codon bias. Nonetheless, the correlation in CUB values between genera shows that the relative degree of bias tends to be largely conserved across these three divergent animal models. As CUB is believed to often promote translational efficiency of highly translated genes [63], we speculate that germ line genes might have retained their relative translation rates across divergent models.

While the nucleotide composition in germ line genes differed among taxa, including a GC bias in *Drosophila* (GC content across all germ line genes=0.561±0.006) and AT biases in *Caenorhabditis* and *Apis* (GC content=0.450 ±0.020 and 0.405±0.002) (Additional file 1: Table S7), nucleotide content has been accounted for using mENC' [61, 62], and thus Fig. 2 c and d suggests that the relative selective pressure on CUB, despite different types of background nucleotides or optimal codons in these three genera [54, 56, 64, 66, 67, 85, 110] is at least partly retained across orthologous gene sets. The relevancy of correcting for background composition [61] was demonstrated by the fact that traditional ENC showed no correlation between taxa (*Drosophila* and *Caenorhabditis* P=0.392, *Drosophila* and *Apis* P= 0.116, Additional file 1: Figure S4).

Protein Sequence Divergence and the Transition to Inheritance Mode

As a final note, we briefly mention here that a prior hypothesis in the literature had suggested that the transition from induction to inheritance mode results in a release of selective constraint, and accelerated evolution of proteins that is detectable at the genome-wide level [111]. We previously assessed that hypothesis using methods adhering to established principles of molecular evolution, and found some examples disagreeing with its predictions on protein sequence evolution [30, 54]. We had noted that the hypothesis may apply to smaller subsets of genes such as germ line genes, or PGC-specification genes [30, 54]. We thus compared dN/dS of *Drosophila* and *Apis,* which are each from the class Insecta, and have inheritance and induction mode respectively. We did not observe evidence consistent with accelerated evolution (or release of constraint) on proteins of germ line genes of *Drosophila* as compared to *Apis* (Tables 2 and 4, Fig. 2). For instance, dN/dS was not statistically significantly higher for *Drosophila* than for *Apis* using Mann-Whitney (Ranked) U test of all germ line genes with orthologs between these genera (P=0.320), or for the subset of genes with known roles under "Induction" mode (P=0.720). If the PGC-specification hypothesis, as it pertains to protein sequence evolution [111], indeed applied to germ line genes then a tendency towards higher dN/dS would be expected in flies after a transition to inheritance, which is not what we observed. Nonetheless, as this pattern is solely from two genera, we consider it alone anecdotal rather than conclusive or generalizable. Further study across more germ line genes and taxon groups, including even more closely related genera, would be valuable for rigorous testing of any such general relationship.

Conclusions

Our results herein showed that germ line genes exhibit a wide range of $\overline{dN}/\overline{dS}$ values and CUB in each of three genera, *Drosophila*, *Caenorhabditis*, and *Apis*. Relative to other germ line genes, we found evidence that LSI genes in *Drosophila* (*osk*, *pgc*) and in *Caenorhabditis* (*pie-1*, *pgl-1*) have diverged especially rapidly, and we conclude this could be a common property of *de novo* germ plasm genes. Whilst adaptive evolution is a strong candidate to explain this fast divergence of LSI genes, particularly for *osk* which has several lines of evidence consistent with a history of positive selection in *Drosophila*, we do not exclude some role of relaxed purifying selection; both adaptive changes and relaxed selection may be facilitated by the narrow expression breadth and low pleiotropy of LSI genes, as found using data from *Drosophila*. Our findings further show that the relative ranking of $\overline{dN}/\overline{dS}$ and of CUB germ line genes in the reference *Drosophila* were each correlated to their orthologs in *Caenorhabditis* while only CUB was correlated to the orthologs from *Apis*. The molecular evolutionary patterns of germ line genes in the flies and nematodes may be similar to developmental expression profiles [108], wherein striking parallels were observed across these organisms, despite being from different phyla.

Future research should explore how LSI genes evolve within populations, including the study of their amino acid mutational spectra relative to other identified (non-germ line) *de novo* genes [82], and to experimentally assess shifts in in their functionality within or between genera (cf. [79]). Moreover, transcriptome data from the germ lines during early embryogenesis and germ line development in multiple species per genus may provide a means to assess how their gene expression has evolved between species, which may be as

relevant to understanding divergence in their function as protein sequence changes. Furthering our understanding of the molecular evolution of germ line genes will be facilitated by expanding research to species from a wider range of genera, including the induction model systems mice, crickets and salamanders [5, 8, 10, 32, 112] and inheritance species such as wasps and frogs [2, 4, 6–9]. Together, the present findings provide a framework for further study of the molecular evolution of germ line genes in metazoans.

Methods

Identification of Germ Line Genes for Analysis

A set of 34 genes with experimental and/or cytological evidence of involvement in PGC-specification under induction mode (N=13), inheritance mode (N=15), or both modes (N=4) as well as two LSI genes *osk* and *pgc* (Table 1), were selected for study in *Drosophila*. CDS from *D. melanogaster* (longest isoform per gene) were used as the reference CDS set for orthology searching as PGC-specification has been well-studied in that organism, its genome has been well annotated (www.flybase.org), and it is arguably the best annotated species in the genus [48]. For study within the genus, five additional species of *Drosophila* within the *melanogaster* group, *D. erecta*, *D. sechellia*, *D. simulans*, *D. yakuba*, were chosen, as well as a relative outgroup species *D. ananassae* (Additional file 1: Figure S1 and Table S1). All six species are closely related taxa and exhibit a range of dN/dS values (Additional file 1: Figure S2) and largely unsaturated dN and dS (Additional file 1: Figure S3), making them suitable for study of molecular evolution [48, 56]. The procedures used for identification of suitable orthologs for study between *Drosophila* and *Caenorhabditis* and *Apis*, as well for among the various species within each genus, are described in detail in Additional file 1: Text file S4.

Molecular Evolutionary Analysis

The CDS for each gene per genus were aligned by codons using MUSCLE [113] in MEGA [114] set at default parameters with the exception that the gap penalty was set to -1.9, which yielded more effective alignments (than the default of -2.9) across multiple-species. Regions with gaps were removed. It has been proposed that small segments in a gene with poor alignment or greater divergence might influence measures such as dN and dS, and detection of positive selection, and their removal improves such estimates despite loss of some sequence information [115, 116]. Thus, we used a dual approach of filtering using the program GBLOCKS [115] set at default parameters, which accordingly shortened divergent alignments, and inspection of protein alignments by eye, always retaining the start codon [117, 118], to remove residual divergent and putatively misaligned segments. Thus, all alignments and measures of substitution rates herein are considered conservative, and the latter applies specifically to the aligned regions per gene.

Protein sequence divergence per phylogenetic branch was measured using dN and dS under the free ratio model (M1) in codeml of PAML based on an unrooted tree for each genus [55]. Whilst some studies have used the M0 model in PAML to measure dN/dS in taxa including the *melanogaster* group in *Drosophila*, which determines a single dN/dS across all branches in the phylogeny [56], we allowed a separate dN/dS in each branch to include potential species-specific effects on dN/dS (Additional file 1: Figure S2) using the free-ratios model [55]. As noted in the Results and Discussion $\overline{dN}/\overline{dS}$ (using M1 model) and M0 dN/dS were strongly correlated across genes within each genus (Spearman's ranked R>0.95, $P<2X10^{-7}$). The value of $\overline{dN}/\overline{dS}$ was used instead of mean dN/dS across the branches as the latter can be biased towards extremely high values due to rare cases (branches) with extremely low dS and avoids exclusion of a branch (i.e., no dN/dS value) in cases when dN>0 and dS=0 [119]. The phylogeny for *Drosophila* was taken as that provided at FlyBase [109] and was unrooted for PAML. The unrooted *Caenorhabditis* four-species phylogeny was taken from [120] and for *Apis* from [121]. For the latter genus, which is less strongly resolved for (ingroup) positions of *A.dorsata* and *A. cerena*, alternate phylogenies were employed for the ingroup yielding highly similar results. We note that we did not detect orthologs of the germ line gene *par-1* in all six *Drosophila* species, and thus we did not formally include it in the *Drosophila* gene set for study, but a three-species alignment was studied, and *par-1* was examined in *Caenorhabditis* and *Apis*; the results are described in Additional file 1: Text File S5.

Positive selection was assessed using "sites" analysis in PAML across all species per genus [55]. For this we compared M7 versus M8. For those genes exhibiting positive selection using 2XΔlnlikelihood based on the Chi^2 table, we obtained the BEB posterior probabilities identifying the sites with P>0.90.

Codon Usage Bias

The values of mENC', which accounts for abundance of rare amino acids and for nucleotide content of the genes under study [61] was conducted using a program from Satapathy et al. [62]. Standard ENC [60], GC3 content at 3rd synonymous codon positions (GC3s) and GC content per CDS was determined in CodonW [122]. For consistency with $\overline{dN}/\overline{dS}$, all mENC', ENC and GC values were determined using the aligned sequences per gene excluding gaps for each species.

Inter-Genus Contrasts

We compared the $\overline{dN}/\overline{dS}$ values and the mENC' values of orthologs present in each of the two genera per contrast (cf. [123]). As genes varied in sequence between genera and were aligned separately within each genus, under a conservative approach, we compared the relative ranking of germ line genes within each genus (using Spearman Rank Correlations) to assess any relationships between genera.

Pleiotropy

For analysis of pleiotropy in *D. melanogaster*, gene expression breadth was determined using the modENCODE transcriptome database as presented at FlyBase.org [109, 124] across 30 tissues and stages of development. These included twelve stages from embryos, six from larvae, six from pupae, and three for adult males and for adult females (shown in Additional file 1: Table S8). Breadth of expression was quantified as the number of developmental stages in which a gene was expressed [83, 106], and was converted into percentages of the 30 stages studied (Additional file 1: Table S8). Analysis was repeated for those using a cutoff of >5 RPKM.

Additional file

Additional file 1: The file contains the supplementary Tables, Figures and Text which are denoted and Tables S1 to S8, Figures S1, S2,S3 and S4, and Text files S1,S2, S3, S4 and S5. (PDF 509 kb)

Additional file 2: Alignments. (ZIP 213 kb)

Acknowledgements

The authors acknowledge members of the Extavour lab group at Harvard OEB Department for valuable comments on this project. Helpful comments by the two anonymous Reviewers are acknowledged.

Funding

This work was supported by funds from Harvard University and National Institutes of Health grant R01 HD073499-01 to CGE.

Availability of data and materials

All genomic sequence data under study were obtained from public databases and are described in Additional file 1: Table S1. Unique gene identifiers for our reference taxa in each genus *D. melanogaster*, *C. elegans* and *A. mellifera* are provided in Table 1 and Additional file 1: Tables S2 and S3 respectively. Alignments are available in Additional file 2.

Authors' contributions

CAW and CGE designed the study, analyzed data and wrote the manuscript. Both authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare they have no competing interests.

Publisher's Note

Received: 29 June 2018 Accepted: 14 January 2019
Published online: 11 February 2019

References

1. Ewen-Campen B, Schwager EE, Extavour CG. The molecular machinery of germ line specification. Mol Reprod Dev. 2010;77(1):3–18.
2. Lynch JA, Özüak O, Khila A, Abouheif E, Desplan C, Roth S. The Phylogenetic Origin of oskar Coincided with the Origin of Maternally Provisioned Germ Plasm and Pole Cells at the Base of the Holometabola. PLoS Genetics. 2011; 7(4):e1002029.
3. Updike D, Strome S. P granule assembly and function in Caenorhabditis elegans germ cells. J Andro. 2010;31(1):53–60.
4. Strome S, Updike D. Specifying and protecting germ cell fate. Nat Rev Mol Cell Biol. 2015;16(7):406–16.
5. Saitou M, Yamaji M. Primordial germ cells in mice. Cold Spring Harb Perspect Biol. 2012;4(11):a008375.
6. Bull AL. Stages of living embryos in the jewel wasp Mormoniella (Nasonia) vitripennis (Walker) (Hymenoptera: Pteromalidae). Int J Insect Morphol Embryol. 1982;1(11):1–23.
7. Extavour CG, Akam ME. Mechanisms of germ cell specification across the metazoans: epigenesis and preformation. Development. 2003;130(24):5869–84.
8. Johnson AD, Richardson E, Bachvarova RF, Crother BI. Evolution of the germ line-soma relationship in vertebrate embryos. Reproduction. 2011;141(3):291–300.
9. Johnson AD, Crother B, White ME, Patient R, Bachvarova RF, Drum M, Masi T. Regulative germ cell specification in axolotl embryos: a primitive trait conserved in the mammalian lineage. Philos Trans R Soc Lond B Biol Sci. 2003;358(1436):1371–9.
10. Nakamura T, Extavour CG. The transcriptional repressor Blimp-1 acts downstream of BMP signaling to generate primordial germ cells in the cricket Gryllus bimaculatus. Development. 2016;143(2):255–63.
11. Crother BI, White ME, Johnson AD. Inferring developmental constraint and constraint release: primordial germ cell determination mechanisms as examples. J Theor Biol. 2007;248(2):322–30.
12. Johnson AD, Alberio R. Primordial germ cells: the first cell lineage or the last cells standing? Development. 2015;142(16):2730–9.
13. Xie T. Spradling AC: decapentaplegic is essential for the maintenance and division of germline stem cells in the Drosophila ovary. Cell. 1998;94(2):251–60.
14. Zoller R, Schulz C. The Drosophila cyst stem cell lineage: Partners behind the scenes? Spermatogenesis. 2012;2(3):145–57.
15. Song X, Xie T. Wingless signaling regulates the maintenance of ovarian somatic stem cells in Drosophila. Development. 2003;130(14):3259–68.
16. Chuma S, Hosokawa M, Kitamura K, Kasai S, Fujioka M, Hiyoshi M, Takamune K, Noce T, Nakatsuji N. Tdrd1/Mtr-1, a tudor-related gene, is essential for male germ-cell differentiation and nuage/germinal granule formation in mice. Proc Natl Acad Sci U S A. 2006;103(43):15894–9.
17. Marlow F. Primordial Germ Cell Specification and Migration. F1000Res. 2015;4: F1000 Faculty Rev-1462.
18. Ephrussi A, Lehmann R. Induction of germ cell formation by oskar. Nature. 1992;358(6385):387–92.
19. Hanyu-Nakamura K, Sonobe-Nojima H, Tanigawa A, Lasko P, Nakamura A. Drosophila Pgc protein inhibits P-TEFb recruitment to chromatin in primordial germ cells. Nature. 2008;451(7179):730–3.
20. Rangan P, DeGennaro M, Jaime-Bustamante K, Coux R-X, Martinho RG, Lehmann R. Temporal and spatial control of germ plasm RNAs. Current Biology. 2009;19(1):72–7.
21. Schlötterer C. Genes from scratch--the evolutionary fate of de novo genes. Trends Genet. 2015;31(4):215–9.
22. Updike DL, Hachey SJ, Kreher J, Strome S. P granules extend the nuclear pore complex environment in the C. elegans germ line. J Cell Biol. 2011; 192(6):939–48.
23. Gallo CM, Wang JT, Motegi F, Seydoux G. Cytoplasmic partitioning of P granule components is not required to specify the germline in C. elegans. Science. 2010;330(6011):1685–9.
24. Strome S, Martin P, Schierenberg E, Paulsen J. Transformation of the germ line into muscle in mes-1 mutant embryos of C. elegans. Development. 1995;121(9):2961–72.

25. Hird SN, Paulsen JE, Strome S. Segregation of germ granules in living Caenorhabditis elegans embryos: cell-type-specific mechanisms for cytoplasmic localisation. Development. 1996;122:1303–12.
26. Hanazawa M, Yonetani M, Sugimoto A. PGL proteins self associate and bind RNPs to mediate germ granule assembly in C. elegans. Journal of Cell Biology. 2011;192(6):929–37.
27. Leacock SW, Reinke V. MEG-1 and MEG-2 are embryo-specific P-granule components required for germline development in Caenorhabditis elegans. Genetics. 2008;178(1):295–306.
28. Wang JT, Smith J, Chen BC, Schmidt H, Rasoloson D, Paix A, Lambrus BG, Calidas D, Betzig E, Seydoux G. Regulation of RNA granule dynamics by phosphorylation of serine-rich, intrinsically disordered proteins in C. elegans. eLife. 2014;3:e04591.
29. Mello CC, Schubert C, Draper B, Zhang W, Lobel R, Priess JR. The PIE-1 protein and germline specification in C. elegans embryos. Nature. 1996; 382(6593):710–2.
30. Whittle CA, Extavour CG. Causes and evolutionary consequences of primordial germ cell specification mode in metazoans. Proceedings of the National Academy of Sciences of the United States of America. 2017; 114(23):5784–91.
31. Seki Y, Yamaji M, Yabuta Y, Sano M, Shigeta M, Matsui Y, Saga Y, Tachibana M, Shinkai Y, Saitou M. Cellular dynamics associated with the genome-wide epigenetic reprogramming in migrating primordial germ cells in mice. Development. 2007;134:2627–38.
32. Donoughe S, Nakamura T, Ewen-Campen B. Green Da, Henderson L, Extavour CG: BMP signaling is required for the generation of primordial germ cells in an insect. Proceedings of the National Academy of Sciences of the United States of America. 2014;111(11):4133–8.
33. Ewen-Campen B, Donoughe S, Clarke DN, Extavour CG. Germ cell specification requires zygotic mechanisms rather than germ plasm in a basally branching insect. Current Biology. 2013;23(10):835–42.
34. Bütschli O. Zur Entwicklungsgeschichte der Biene. Zeitschrift für Wissenschaftliche Zoologie. 1870;20:519–64.
35. Fleig R, Sander K. Blastoderm development in honey bee embryogenesis as seen in the scanning electron microscope. International Journal of Invertebrate Reproduction and Development. 1985;8:279–86.
36. Fleig R, Sander K. Embryogenesis of the Honeybee Apis mellifera L (Hymenoptera, Apidae) - an SEM Study. International Journal of Insect Morphology and Embryology. 1986;15(5-6):449–62.
37. Dearden PK. Germ cell development in the Honeybee (Apis mellifera); vasa and nanos expression. BMC Developmental Biology. 2006;6:6.
38. Nelson JA. The embryology of the honey bee. Princeton University Press: Princeton; 1915.
39. Gustafson EA, Wessel GM. Vasa genes: emerging roles in the germ line and in multipotent cells. Bioessays. 2010;32(7):626–37.
40. Yajima M, Wessel GM. The multiple hats of Vasa: its functions in the germline and in cell cycle progression. Molecular Reproduction and Development. 2011;78(10-11):861–7.
41. Mahowald AP. Assembly of the Drosophila germ plasm. International review of cytology. 2001;203:187–213.
42. Breitwieser W, Markussen F-H, Horstmann H, Ephrussi A. Oskar protein interaction with Vasa represents an essential step in polar granule assembly. Genes and Development. 1996;10:2179–88.
43. Kuznicki KA, Smith PA, Leung-Chiu WM, Estevez AO, Scott HC, Bennett KL. Combinatorial RNA interference indicates GLH-4 can compensate for GLH-1; these two P granule components are critical for fertility in C. elegans. Development. 2000;127(13):2907–16.
44. Spike C, Meyer N, Racen E, Orsborn A, Kirchner J, Kuznicki K, Yee C, Bennett K, Strome S. Genetic analysis of the Caenorhabditis elegans GLH family of P-granule proteins. Genetics. 2008;178(4):1973–87.
45. Tanaka M, Kinoshita M, Kobayshi D, Nagahama Y. Establishment of medaka (Oryzias latipes) transgenic lines with the expression of green fluorescent protein fluorescence exclusively in germ cells: A useful model to monitor germ cells in a live vertebrate. PNAS. 2001;5(98): 2544–9.
46. Toyooka Y, Tsunekawa N, Takahashi Y, Matsui Y, Satoh M, Noce T. Expression and intracellular localization of mouse Vasa-homologue protein during germ cell development. Mech Dev. 2000;93(1-2):139–49.
47. Jaruzelska J, Kotecki M, Kusz K, Spik A, Firpo M, Reijo Pera RA. Conservation of a Pumilio-Nanos complex from Drosophila germ plasm to human germ cells. Dev Genes Evol. 2003;213(3):120–6.
48. Choi JY, Aquadro CF. Molecular Evolution of Drosophila Germline Stem Cell and Neural Stem Cell Regulating Genes. Genome Biol Evol. 2015;7(11):3097–114.
49. Choi JY, Aquadro CF. The coevolutionary period of Wolbachia pipientis infecting Drosophila ananassae and its impact on the evolution of the host germline stem cell regulating genes. Mol Biol Evol. 2014;31(9):2457–71.
50. Bauer DuMont VL, Flores HA, Wright MH, Aquadro CF. Recurrent positive selection at bgcn, a key determinant of germ line differentiation, does not appear to be driven by simple coevolution with its partner protein bam. Mol Biol Evol. 2007;24(1):182–91.
51. Flores HA, DuMont VL, Fatoo A, Hubbard D, Hijji M, Barbash DA, Aquadro CF. Adaptive evolution of genes involved in the regulation of germline stem cells in Drosophila melanogaster and D. simulans. G3 (Bethesda). 2015; 5(4):583–92.
52. Ahuja A, Extavour CG. Patterns of molecular evolution of the germ line specification gene oskar suggest that a novel domain may contribute to functional divergence in Drosophila. Development Genes and Evolution. 2014;222(4):65–77.
53. Buschiazzo E, Ritland C, Bohlmann J, Ritland K. Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. BMC Evol Biol. 2012;12:8.
54. Whittle CA, Extavour CG. Refuting the hypothesis that the acquisition of germ plasm accelerates animal evolution. Nature Communications. 2016;7:12637.
55. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution. 2007;24(8):1586–91.
56. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN, et al. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 2007;450(7167):203–18.
57. Stanley CE, Jr., Kulathinal RJ: flyDIVaS: A Comparative Genomics Resource for Drosophila Divergence and Selection. G3 (Bethesda) 2016, 6(8):2355-2363.
58. Kratochwil CF, Geissler L, Irisarri I, Meyer A. Molecular Evolution of the Neural Crest Regulatory Network in Ray-Finned Fish. Genome Biology and Evolution. 2015;7(11):3033–46.
59. Castillo-Davis CI, Bedford TB, Hartl DL. Accelerated rates of intron gain/loss and protein evolution in duplicate genes in human and mouse malaria parasites. Molecular Biology and Evolution. 2004;21(7):1422–7.
60. Wright F. The "effective number of codons" used in a gene. Gene. 1990;87: 23–9.
61. Novembre JA. Accounting for background nucleotide composition when measuring codon usage bias. Molecular Biology and Evolution. 2002;19: 1390–4.
62. Satapathy SS, Sahoo AK, Ray SK, Ghosh TC. Codon degeneracy and amino acid abundance influence the measures of codon usage bias: improved Nc (Nc) and ENCprime (N(')c) measures. Genes Cells. 2017;22(3):277–83.
63. Akashi H. Gene expression and molecular evolution. Current Opinion in Genetics & Development. 2001;11:660–6.
64. Cutter AD, Wasmuth JD, Blaxter ML. The evolution of biased codon and amino acid usage in nematode genomes. Molecular biology and evolution. 2006;23(12):2303–15.
65. Zhou Z, Dang Y, Zhou M, Li L, Yu CH, Fu J, Chen S, Liu Y. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. Proceedings of the National Academy of Sciences of the United States of America. 2016;113(41):E6117–25.
66. Duret L, Mouchiroud D. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America. 1999;96(8):4482–7.
67. Behura SK, Severson DW. Comparative analysis of codon usage bias and codon context patterns between dipteran and hymenopteran sequenced genomes. PLoS One. 2012;7(8):e43111.
68. Behura SK, Severson DW. Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes. Biol Rev Camb Philos Soc. 2013;88(1):49–61.
69. Swanson WJ, Vacquier VD. The rapid evolution of reproductive proteins. Nat Rev Genet. 2002;3(2):137–44.
70. Haerty W, Jagadeeshan S, Kulathinal RJ, Wong A, Ram KR, Sirot LK, Levesque L, Artieri CG, Wolfner MF, Civetta A, et al. Evolution in the fast lane: rapidly evolving sex-related genes in Drosophila. Genetics. 2007;177:1321–35.
71. Yang N, Yu Z, Hu M, Wang M, Lehmann R, Xu RM. Structure of Drosophila Oskar reveals a novel RNA binding protein. Proc Natl Acad Sci U S A. 2015; 112(37):11541–6.

72. Jeske M, Bordi M, Glatt S, Muller S, Rybin V, Muller CW, Ephrussi A. The Crystal Structure of the Drosophila Germline Inducer Oskar Identifies Two Domains with Distinct Vasa Helicase- and RNA-Binding Activities. Cell Rep. 2015;12(4):587–98.
73. Kibota TT, Lynch M. Estimate of the genomic mutation rate deleterious to overall fitness in E. coli. Nature. 1996;381(6584):694–6.
74. Yang W, Bielawski JP, Yang Z. Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. J Mol Evol. 2003;57(2):212–21.
75. Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol. 1998;15(5):568–73.
76. Jeffares DC, Tomiczek B, Sojo V, dos Reis M. A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. Methods Mol Biol. 2015;1201:65–90.
77. Toll-Riera M, Laurie S, Alba MM. Lineage-specific variation in intensity of natural selection in mammals. Mol Biol Evol. 2011;28(1):383–98.
78. Nozawa M, Suzuki Y, Nei M. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. Proc Natl Acad Sci U S A. 2009;106(16):6700–5.
79. Webster PJ, Suen J, Macdonald PM. Drosophila virilis oskar transgenes direct body patterning but not pole cell formation or maintenance of mRNA localization in D. melanogaster. Development. 1994;120(7):2027–37.
80. Tamura K, Subramanian S, Kumar S. Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. Mol Biol Evol. 2004;21(1):36–44.
81. Ewen-Campen B, Srouji JR, Schwager EE. Extavour CG: oskar Predates the Evolution of Germ Plasm in Insects. Current Biology. 2012;22(23):2278–83.
82. Liu HQ, Li Y, Irwin DM, Zhang YP, Wu DD. Integrative analysis of young genes, positively selected genes and lncRNAs in the development of Drosophila melanogaster. BMC Evol Biol. 2014;14:241.
83. Subramanian S, Kumar S. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics. 2004;168(1):373–81.
84. Mank JE, Ellegren H. Are sex-biased genes more dispensable? Biol Lett. 2009;5(3):409–12.
85. Vicario S, Moriyama EN, Powell JR. Codon usage in twelve species of Drosophila. BMC Evolutionary Biology. 2007;7:226.
86. Comeron JM, Kreitman M, Aguade M. Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. Genetics. 1999;151(1):239–49.
87. Lesch BJ, Page DC. Genetics of germ cell development. Nat Rev Genet. 2012;13(11):781–94.
88. Kawasaki I, Shim YH, Kirchner J, Kaminker J, Wood WB, Strome S. PGL-1, a predicted RNA-binding component of germ granules, is essential for fertility in C. elegans. Cell. 1998;94(5):635–45.
89. Seydoux G, Mello CC, Pettitt J, Wood WB, Priess JR, Fire A. Repression of gene expression in the embryonic germ lineage of C. elegans. Nature. 1996;382(6593):713–6.
90. Kawasaki I, Amiri A, Fan Y, Meyer N, Dunkelbarger S, Motohashi T, Karashima T, Bossinger O, Strome S. The PGL family proteins associate with germ granules and function redundantly in Caenorhabditis elegans germline development. Genetics. 2004;167(2):645–61.
91. Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol Biol Evol. 2005;22(5):1345–54.
92. Zhang J. Protein-length distributions for the three domains of life. Trends Genet. 2000;16(3):107–9.
93. Betancourt AJ, Presgraves DC. Linkage limits the power of natural selection in Drosophila. Proceedings Of The National Academy Of Sciences Of The United States Of America. 2002;99(21):13616–20.
94. Kim Y. Effect of strong directional selection on weakly selected mutations at linked sites: implication for synonymous codon usage. Mol Biol Evol. 2004;21(2):286–94.
95. Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. Proc Natl Acad Sci U S A. 2006;103(26):9935–9.
96. Begun DJ, Lindfors HA, Kern AD, Jones CD. Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. Genetics. 2007;176(2):1131–7.
97. Schmitz JF, Bornberg-Bauera E. Fact or fiction: updates on how protein-coding genes might emerge de novo from previously non-coding RNA. F1000Res. 2017;6:57.
98. Tautz D, Domazet-Loso T. The evolutionary origin of orphan genes. Nature reviews Genetics. 2011;12(10):692–702.
99. Wissler L, Gadau J, Simola DF, Helmkampf M, Bornberg-Bauer E. Mechanisms and dynamics of orphan gene emergence in insect genomes. Genome Biol Evol. 2013;5(2):439–55.
100. Blondel L J, TEM, Extavour CG: Bacterial contribution to genesis of the novel germ line determinant oskar. bioRxiv 2018:2018:453514.
101. Palmieri N, Kosiol C, Schlotterer C. The life cycle of Drosophila orphan genes. Elife. 2014;3:e01311.
102. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC. More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. 2009;25(9):404–13.
103. Tautz D, Neme R, Domazet-Loso T. Evolutionary Origin of Orphan Genes. eLS. 2013. https://doi.org/10.1002/9780470015902.a0024601.
104. Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. Genome Biol Evol. 2009;1:114–8.
105. Markova-Raina P, Petrov D. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. Genome Res. 2011;21(6):863–74.
106. Duret L, Mouchiroud D. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol. 2000;17(1):68–74.
107. Park SG, Choi SS. Expression breadth and expression abundance behave differently in correlations with evolutionary rates. BMC Evol Biol. 2010;10:241.
108. Li JJ, Huang H, Bickel PJ, Brenner SE. Comparison of D. melanogaster and C. elegans developmental stages, tissues, and cells by modENCODE RNA-seq data. Genome Res. 2014;24(7):1086–101.
109. Gramates LS, Marygold SJ, Santos GD, Urbano JM, Antonazzo G, Matthews BB, Rey AJ, Tabone CJ, Crosby MA, Emmert DB, et al. FlyBase at 25: looking to the future. Nucleic Acids Res. 2016;45(Database issue):D663–D671
110. Duret L. tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. Trends in Genetics. 2000;16:287-289.
111. Evans T, Wade CM, Chapman FA, Johnson AD, Loose M. Acquisition of germ plasm accelerates vertebrate evolution. Science. 2014;344(6180):200–3.
112. Chatfield J, O'Reilly MA, Bachvarova RF, Ferjentsik Z, Redwood C, Walmsley M, Patient R, Loose M, Johnson AD. Stochastic specification of primordial germ cells from mesoderm precursors in axolotl embryos. Development. 2014;141(12):2429–40.
113. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5:113.
114. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016;33(7):1870–4.
115. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 2007;56(4):564–77.
116. Privman E, Penn O, Pupko T. Improving the performance of positive selection inference by filtering unreliable alignment regions. Mol Biol Evol. 2012;29(1):1–5.
117. Carr M, Richter DJ, Fozouni P, Smith TJ, Jeuck A, Leadbeater BSC, Nitsche F. A six-gene phylogeny provides new insights into choanoflagellate evolution. Mol Phylogenet Evol. 2017;107:166–78.
118. Heidel AJ, Kiefer C, Coupland G, Rose LE. Pinpointing genes underlying annual/perennial transitions with comparative genomics. BMC Genomics. 2016;17(1):921.
119. Wlasiuk G, Nachman MW. Promiscuity and the rate of molecular evolution at primate immunity genes. Evolution. 2010;64(8):2204–20.
120. Felix MA, Braendle C, Cutter AD. A streamlined system for species diagnosis in Caenorhabditis (Nematoda: Rhabditidae) with name designations for 15 distinct biological species. PLoS One. 2014;9(4):e94723.
121. Shullia NI, Raffiudin R, Juliandi B. The Phosphofructokinase and Pyruvate Kinase Genes In Apis andreniformis and Apis cerana indica: Exon Intron Organisation and Evolution. Tropical Life Sciences Research. 2018;29(2). in press.
122. Peden JF. Analysis of Codon Usage. PhD thesis. Nottingham University, Department of Genetics; 1999.

123. Friedman R, Drake JW, Hughes AL. Genome-wide patterns of nucleotide substitution reveal stringent functional constraints on the protein sequences of thermophiles. Genetics. 2004;167(3):1507–12.
124. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, et al. The developmental transcriptome of Drosophila melanogaster. Nature. 2011;471(7339):473–9.
125. Cook HA, Koppetsch BS, Wu J, Theurkauf WE. The Drosophila SDE3 homolog armitage is required for oskar mRNA silencing and embryonic axis specification. Cell. 2004;116(6):817–29.
126. Kim G, Pai CI, Sato K, Person MD, Nakamura A, Macdonald PM. Region-specific activation of oskar mRNA translation by inhibition of Bruno-mediated repression. PLoS Genet. 2015;11(2):e1004992.
127. Liang L, Diehl-Jones W, Lasko P. Localization of vasa protein to the Drosophila pole plasm is independent of its RNA-binding and helicase activities. Development. 1994;120(5):1201–11 (Cambridge, England).
128. Quinlan ME, Hilgert S, Bedrossian A, Mullins RD, Kerkhoff E. Regulatory interactions between two actin nucleators, Spire and Cappuccino. J Cell Biol. 2007;179(1):117–28.
129. Nakamura A, Sato K, Hanyu-Nakamura K. Drosophila cup is an eIF4E binding protein that associates with Bruno and regulates oskar mRNA translation in oogenesis. Dev Cell. 2004;6(1):69–78.
130. Dalby B, Glover DM. 3' non-translated sequences in Drosophila cyclin B transcripts direct posterior pole accumulation late in oogenesis and peri-nuclear association in syncytial embryos. Development. 1992;115(4):989–97.
131. Newmark PA, Boswell RE. The mago nashi locus encodes an essential product required for germ plasm assembly in Drosophila. Development. 1994;120(5):1303–13.
132. Rojas-Rios P, Chartier A, Pierson S, Severac D, Dantec C, Busseau I, Simonelig M. Translational Control of Autophagy by Orb in the Drosophila Germline. Dev Cell. 2015;35(5):622–31.
133. Siegel V, Jongens TA, Jan LY. Jan YN: pipsqueak, an early acting member of the posterior group of genes, affects vasa level and germ cell-somatic cell interaction in the developing egg chamber. Development. 1993;119(4):1187–202.
134. Ephrussi A, Dickinson LK, Lehmann R. Oskar organizes the germ plasm and directs localization of the posterior determinant nanos. Cell. 1991;66(1):37–50.
135. Arkov AL, Wang J-YS, Ramos A, Lehmann R. The role of Tudor domains in germline development and polar granule architecture. Development. 2006;133:4053–62.
136. Anne J, Mechler BM. Valois, a component of the nuage and pole plasm, is involved in assembly of these structures, and binds to Tudor and the methyltransferase Capsuleen. Development. 2005;132(9):2167–77.
137. Vincent SD, Dunn NR, Sciammas R, Shapiro-Shalef M, Davis MM, Calame K, Bikoff EK, Robertson EJ. The zinc finger transcriptional repressor Blimp1/Prdm1 is dispensable for early axis formation but is required for specification of primordial germ cells in the mouse. Development. 2005;132(6):1315–25.
138. Kurimoto K, Yamaji M, Seki Y, Saitou M. Specification of the germ cell lineage in mice: a process orchestrated by the PR-domain proteins, Blimp1 and Prdm14. Cell Cycle. 2008;7(22):3514–8.
139. Matsui Y, Zsebo K, Hogan BL. Derivation of pluripotential embryonic stem cells from murine primordial germ cells in culture. Cell. 1992;70(5):841–7.
140. Takeuchi Y, Molyneaux K, Runyan C, Schaible K, Wylie C. The roles of FGF signaling in germ cell migration in the mouse. Development. 2005;132(24):5399–409.
141. Aramaki S, Hayashi K, Kurimoto K, Ohta H, Yabuta Y, Iwanari H, Mochizuki Y, Hamakubo T, Kato Y, Shirahige K, et al. A mesodermal factor, T, specifies mouse germ cell fate by directly activating germline determinants. Dev Cell. 2013;27(5):516–29.
142. Lochab AK, Extavour CG. Bone Morphogenetic Protein (BMP) signaling in animal reproductive system development and function. Dev Biol. 2017;427:258-269.
143. Itman C, Loveland KL. Smads and cell fate: distinct roles in specification, development, and tumorigenesis in the testis. IUBMB Life. 2013;65(2):85–97.
144. Keshishian H. Is synaptic homeostasis just wishful thinking? Neuron. 2002;33(4):491–2.
145. Yu K, Sturtevant MA, Biehs B, Francois V, Padgett RW, Blackman RK, Bier E. The Drosophila decapentaplegic and short gastrulation genes function antagonistically during adult wing vein development. Development. 1996;122(12):4033–44.
146. Bachiller D, Klingensmith J, Shneyder N, Tran U, Anderson R, Rossant J, De Robertis EM. The role of chordin/Bmp signals in mammalian pharyngeal development and DiGeorge syndrome. Development. 2003;130(15):3567–78.
147. Ohinata Y, Ohta H, Shigeta M, Yamanaka K, Wakayama T, Saitou M. A signaling principle for the specification of the germ cell lineage in mice. Cell. 2009;137(3):571–84.
148. Tsuda M, Sasaoka Y, Kiso M, Abe K, Haraguchi S, Kobayashi S, Saga Y. Conserved role of Nanos proteins in germ cell development. Science. 2003;301(5637):1239–41.
149. Voronina E, Seydoux G, Sassone-Corsi P, Nagamori I: RNA granules in germ cells. Cold Spring Harb Perspect Biol 2011, 3(12).
150. Moore FL, Jaruzelska J, Fox MS, Urano J, Firpo MT, Turek PJ, Dorfman DM, Pera RA. Human Pumilio-2 is expressed in embryonic stem cells and germ cells and interacts with DAZ (Deleted in AZoospermia) and DAZ-like proteins. Proc Natl Acad Sci U S A. 2003;100(2):538–43.
151. Ikenishi K, Tanaka TS. Involvement of the protein of Xenopus vasa homolog (Xenopus vasa-like gene 1, XVLG1) in the differentiation of primordial germ cells. Dev Growth Differ. 1997;39(5):625–33.
152. Knaut H, Pelegri F, Bohmann K, Schwarz H, Nusslein-Volhard C. Zebrafish vasa RNA but not its protein is a component of the germ plasm and segregates asymmetrically before germline specification. J Cell Biol. 2000;149(4):875–88.

Additional file 1 containing supplementary materials for

Contrasting patterns of molecular evolution in metazoan germ line genes

Carrie A. Whittle and Cassandra G. Extavour

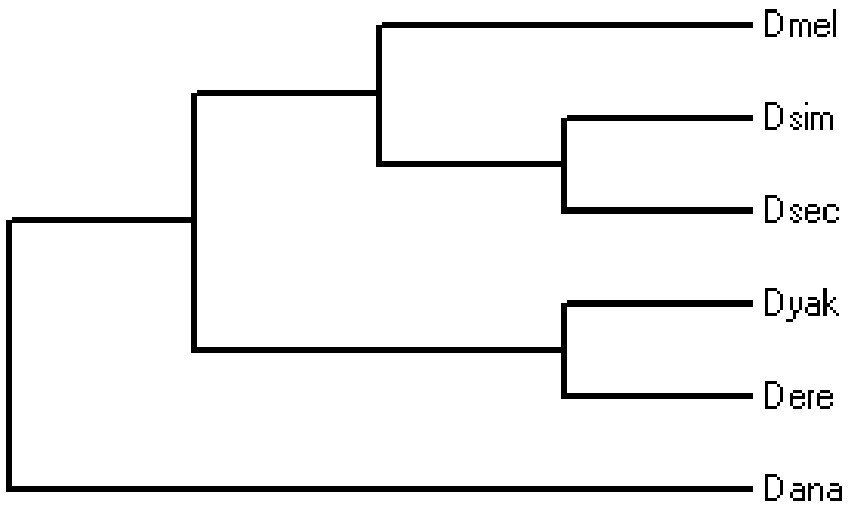


Figure S1. The *Drosophila* phylogeny for six species under study herein. The tree was unrooted for PAML analysis [1] . Each taxon name is abbreviated using the first three letters of the species name. Full names are provided in Table S1. Phylogeny is as provided by FlyBase [2].

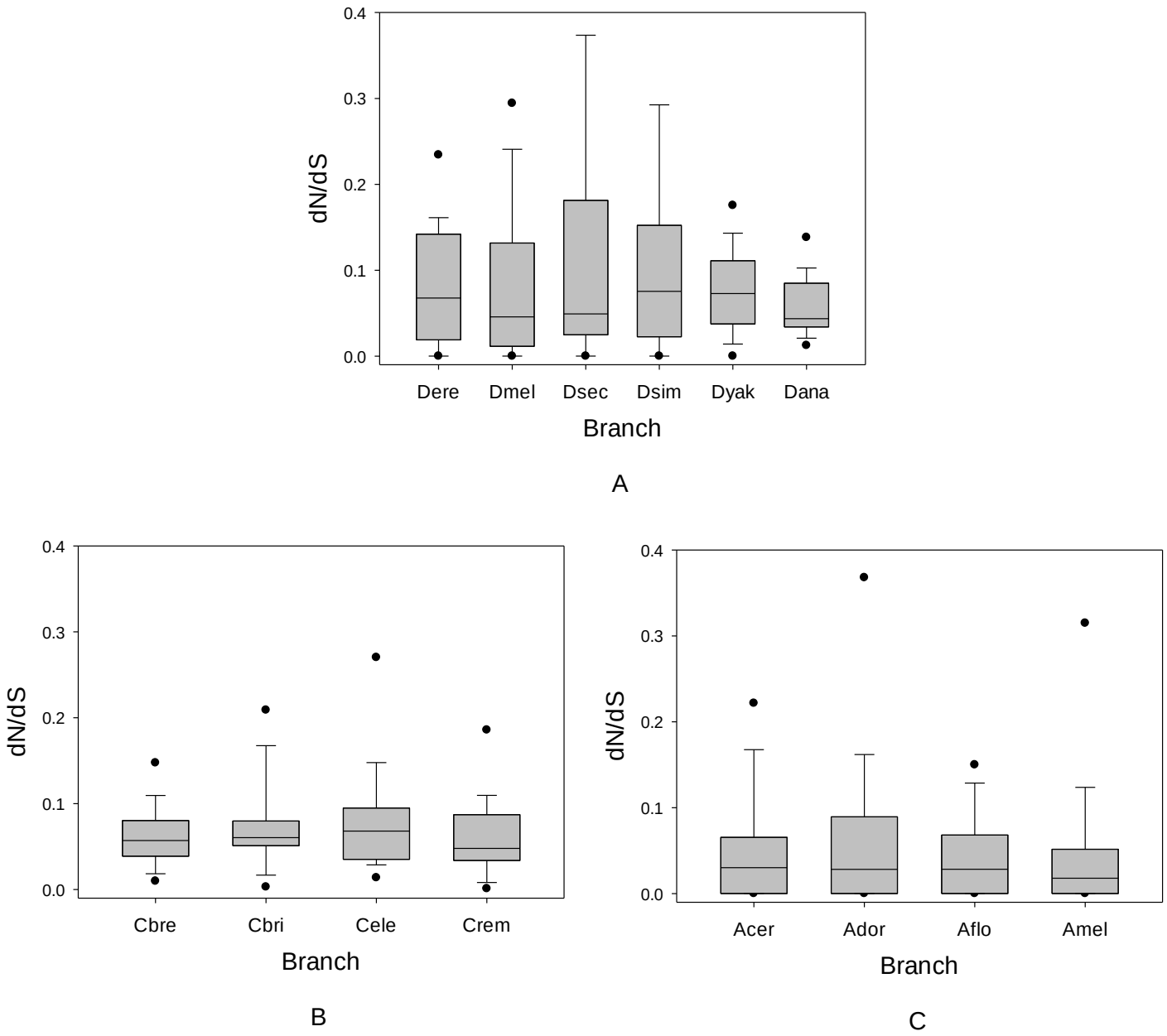


Figure S2. Box plots for dN/dS for each terminal branch per genus. (A) All six *Drosophila* species under study; (B) *Caenorhabditis* species under study; and (C) *Apis* species under study. Each taxon name is abbreviated using the first three letters of the species name.

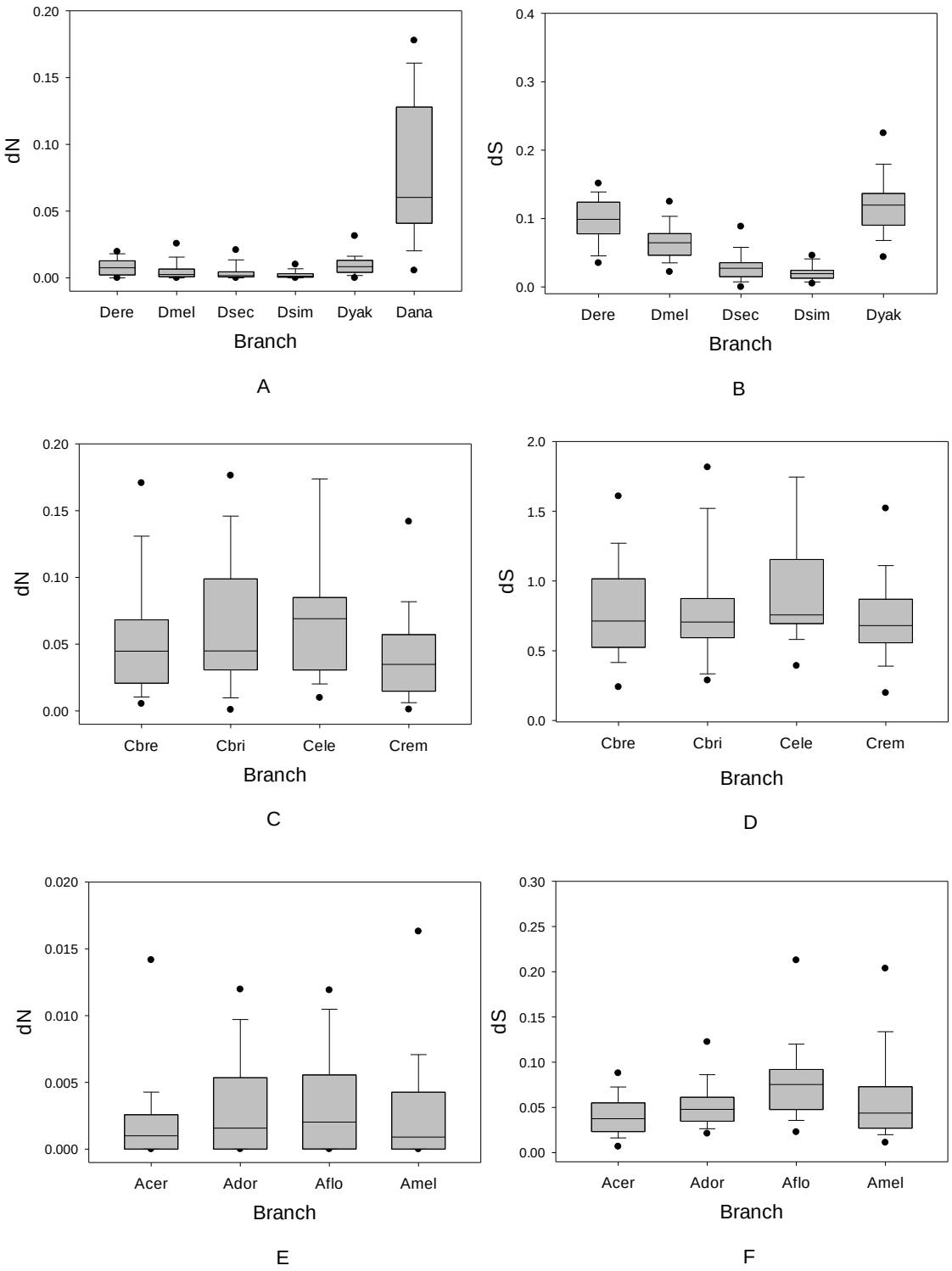
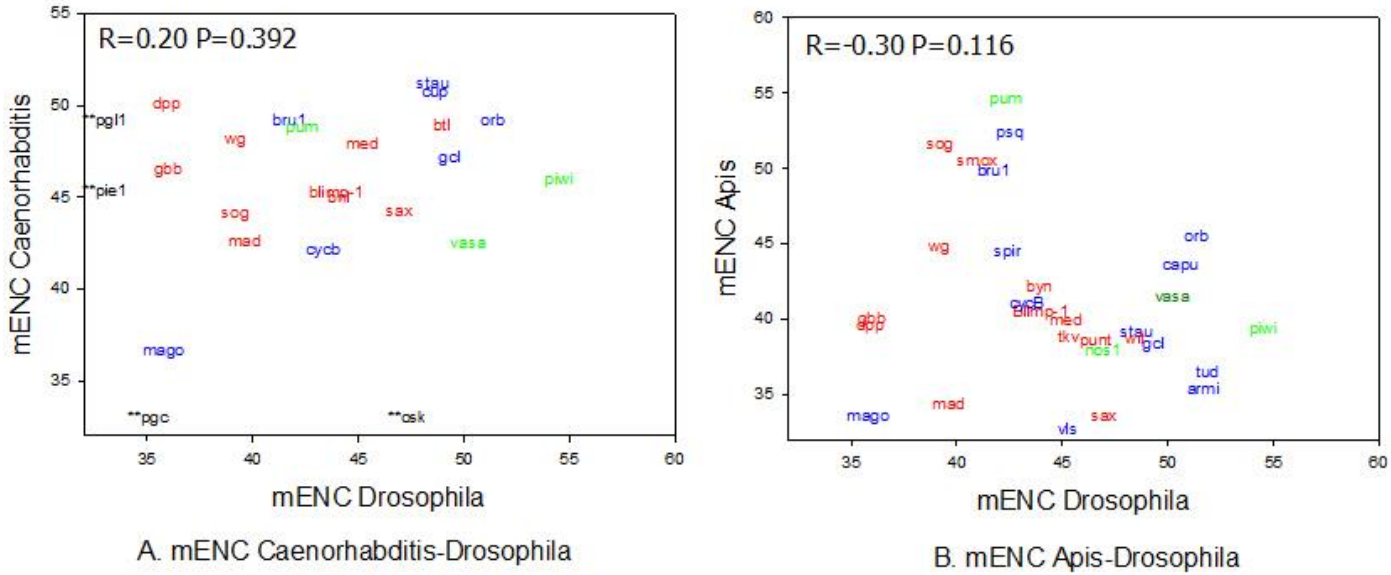


Figure S3. Box plots for dN and dS for all studied genes for each lineage per genus. (A) dN in each *Drosophila* branch; (B) dS in each *Drosophila* branch (C) dN in each *Caenorhabditis* branch; (D) dS in each *Caenorhabditis* branch; (E) dN in each *Apis* branch; (F) dS in each *Apis* branch. Note that for panel B, dS was omitted for *D. ananassae* for visualization purposes, as its values were higher than those of the other species. The median dS for *D. ananassae* was 1.158, and 25th and 75th percentile values were 0.989 and 1.705 respectively, whilst only two values were >2. Each taxon name is abbreviated using the first three letters of the species name.



A. mENC Caenorhabditis-Drosophila

B. mENC Apis-Drosophila

Figure S4. The (A) traditional ENC for *Caenorhabditis* versus *Drosophila* and (B) traditional ENC for *Apis* versus *Drosophila***.** Data points are listed using the gene identifiers. ******Genes that are specific to *Caenorhabditis* or *Drosophila* are shown on the axes. Gene names are those from *Drosophila*.

Table S1. The protein coding sequences (CDS) used in present study to identify germ line genes under study in *Drosophila*, *Caenorhabditis* and *Apis***.** The number of CDS includes all known isoforms.

Taxon	Version	Database[a]	No. CDS
Drosophila melanogaster	release 6.13	Flybase	30,482
D. ananassae	r1.05	Flybase	21,191
D. erecta	r1.05	Flybase	19,592
D. sechellia	r1.3	Flybase	16,471
D. simulans	r2.02	Flybase	24,119
D. yakuba	r1.3	Flybase	16,082
Caenorhabditis elegans	PRJNA13758.WS237	Wormbase	26,189
C. brenneri	PRJNA20035.WS257	Wormbase	30,672
C. briggsae	PRJNA10731.WS257	Wormbase	25,332
C. remanei	PRJNA53967.WS257	Wormbase	31,450
Apis mellifera	Refseq r. 79	Genbank-NCBI	22,460
A. cerena	Refseq r. 79	Genbank-NCBI	19,247
A. dorsata	Refseq r. 79	Genbank-NCBI	18,146
A. florea	Refseq r. 79	Genbank-NCBI	17,664

[a] CDS data are available from FlyBase at http://flybase.org/, for WormBase at http://www.wormbase.org/ and for Genbank-NCBI at https://www.ncbi.nlm.nih.gov/genbank/. Complete and annotated CDS files are available at FlyBase and WormBase for each species under their release number. For *Apis* CDS were downloaded from the Genbank Refseq database using the search criteria: "Species name[organism]". This was followed by up selecting the "Refseq" option for output, and were downloaded using the send to option "coding sequences". The Refseq database contains the non-redundant sequences available for each species, providing a stable reference genome and mRNA. *Apis* sequences were downloaded in December 2016.

Table S2. The 23 *Caenorhabditis* germ line genes under study. Putative *C.elegans* orthologs to the *D. melanogaster* gene list in Table 1 were identified by orthology searches in FlyBase (with *C. elegans* as the taxon of interest). Two germ plasm genes are *Caenorhabditis*-specific (LSI). Only those with orthologs also identified in all three fellow *Caenorhabditis* species (Table S1) using reciprocal BLASTX (with *C. elegans* as the reference) are listed below and were used for analyses.

D. melanogaster	***C. elegans* ortholog**[a]	***C. elegans* Gene ID**
Blimp-1	*blmp-1*	WBGene00003847
bnl	*let-756*	WBGene00002881
bru1	*etr-1*	WBGene00001340
btl	*egl-15*	WBGene00001184
cup	*ifet-1*	WBGene00004132
cycB	*cyb2.1*	WBGene00000866
dpp	*dbl-1*	WBGene00000936
gbb	*tig-2*	WBGene00006570
gcl	*gcl-1*	WBGene00013382
mad	*sma-2*	WBGene00004856
mago	*mag-1*	WBGene00003123
med	*sma-4*	WBGene00004858
NA	*pgl-1*	WBGene00003992
NA	*pie-1*	WBGene00004027
orb	*cpb-3*	WBGene00000772
par-1	*par-1*	WBGene00003916
piwi	*prg-1*	WBGene00004178
punt	*daf-4*	WBGene00000900
sax	*sma-6*	WBGene00004860
sog	*crm-1*	WBGene00007103
stau	*stau-1*	WBGene00018857
vasa	*glh-1*	WBGene00001598
wg	*cwn-1*	WBGene00000857

[a] The *C. elegans* CDS were used as a reference to identify orthologs in the sister species of *Caenorhabditis*. *D. melanogaster* genes (34 in Table 2) that are not listed in the table fell into one of three categories: they did not have high confidence matches in *C. elegans*, the C. *elegans* ortholog also matched another second *D. melanogaster* gene, and/or ortholog matches were not found or poorly aligned across all four *Caenorhabditis* species. *par-1* was added for *Caenorhabditis*, but not studied in all six species in *Drosophila*.

Table S3. The 30 *Apis* genes under study. Putative *A. mellifera* orthologs to the *D. melanogaster* gene list in Table 1 were identified by reciprocal BLASTX. The e-value represents the BLASTX result with *D. melanogaster* CDS as the query against the *A. mellifera* CDS list. As *Apis* has the least well annotated genomes studied, the list of orthologs are putative and comprise the best reciprocal hits to the reference *A. mellifera* and *D. melanogaster* gene list. The *A. mellifera* CDS were used as a reference to identify orthologs in the remaining three sister species of *Apis* using reciprocal BLASTX. Only those genes with putative orthologs in all four *Apis* species are provided below and used for analyses. Functions are those described for the *Apis* CDS from NCBI.

***D. melan.* Gene**	**Gene ID of *Apis mellifera* orthologs in GenBank**	**Putative Function**	**BLASTX e-value**
armi	XM_006571686.2	probable RNA helicase armi	1.00E-130
Blimp-1	XM_006566008.2	probable serine/threonine-protein kinase DDB_G0282963 isoform X1	1.00E-100
bru-1	XM_016911026.1	CUGBP Elav-like family member 2	3E-74
byn	XM_006565966.2	brachyury protein isoform X2	9E-89
capu	XM_016911606.1	protein cappuccino	7E-88
cycB	XM_624245.5	uncharacterized protein LOC551860	5E-65
dpp	XM_006569786.2	protein decapentaplegic	7E-81
gbb	XM_394252.4	protein 60A	1.00E-101
gcl	XM_624700.5	protein germ cell-less isoform X1	1.00E-100
mad	XM_006567242.2	protein mothers against dpp	0
mago	XM_016911522.1	protein mago nashi homolog	4E-76
med	XM_392838.6	mothers against decapentaplegic homolog 4 isoform X1	e-106
nos	XM_016913300.1	protein nanos isoform X1	5E-17
orb	XM_395376.5	cytoplasmic polyadenylation element-binding protein 1 isoform X1	e-120
par-1	XM_006570585.2	serine/threonine-protein kinase MARK2 isoform X2	0
piwi	NM_001165906.1	Aubergine	0
psq	NM_001011602.1	Pipsqueak	2E-95
pum	XM_006565992.2	pumilio homolog 2 isoform X2	0
punt	XM_395928.6	activin receptor type-2A isoform X2	1.00E-150
sax	XM_016917500.1	activin receptor type-1 isoform X1	1.00E-148
smox	XM_006571160.2	mothers against decapentaplegic homolog 3 isoform X1	1.00E-132
sog	XM_393520.6	dorsal-ventral patterning protein Sog isoform X3	0
spir	XM_006570180.2	protein spire isoform X3	8E-91

stau	XM_006564450.2	double-stranded RNA-binding protein Staufen homolog 2 isoform X1	2E-69
tkv	XM_391989.6	bone morphogenetic protein receptor type-1B isoform X1	1.00E-161
tud	XM_016916602.1	LOW QUALITY PROTEIN: maternal protein tudor	8E-53
vasa	XM_006571702.2	ATP-dependent RNA helicase vasa isoform X1	1.00E-149
vls	XM_006565093.2	methylosome protein 50 isoform X1	4E-14
wg	XM_006571239.2	protein Wnt-1	1.00E-169
wit	XM_397334.6	bone morphogenetic protein receptor type-2	1.00E-124

Table S4. $\overline{dN}/\overline{dS}$, mean dN and mean dS and across the phylogeny of six species of *Drosophila* herein for each of the 34 genes under study. Values were measured using codeml in PAML [1]. SE=standard error. Note that additional decimal places of dN and dS than shown were used for calculation of $\overline{dN}/\overline{dS}$.

Gene	$\overline{dN}/\overline{dS}$	Mean dN	SE	Mean dS	SE
Lineage-specific Inheritance					
osk	0.1573	0.0437	0.0350	0.2779	0.2290
pgc	0.0933	0.0290	0.0267	0.3108	0.2808
Inheritance					
capu	0.1189	0.0301	0.0189	0.2527	0.1908
orb	0.1102	0.0239	0.0162	0.2172	0.1695
stau	0.0885	0.0249	0.0216	0.2818	0.2417
cup	0.0878	0.0305	0.0245	0.3472	0.2967
vls	0.0810	0.0279	0.0203	0.3440	0.3032
armi	0.0773	0.0355	0.0294	0.4596	0.4307
spir	0.0751	0.0133	0.0110	0.1770	0.1309
tud	0.0716	0.0304	0.0242	0.4246	0.3949
cycB	0.0692	0.0158	0.0113	0.2284	0.1911
gcl	0.0392	0.0125	0.0092	0.3189	0.2747
bru1	0.0369	0.0083	0.0072	0.2250	0.1794
psq	0.0297	0.0023	0.0015	0.0789	0.0579
mago	0.0001	0.0000	0.0000	0.2459	0.2003
Induction					
bnl	0.1303	0.0328	0.0245	0.2519	0.1829
btl	0.0935	0.0342	0.0287	0.3663	0.3207
wg	0.0833	0.0183	0.0139	0.2199	0.1597
tkv	0.0734	0.0141	0.0091	0.1924	0.1316
byn	0.0575	0.0128	0.0094	0.2230	0.1689
Blimp-1	0.0572	0.0118	0.0070	0.2068	0.1384
dpp	0.0517	0.0123	0.0085	0.2384	0.2028
wit	0.0462	0.0128	0.0099	0.2763	0.2267
med	0.0422	0.0094	0.0085	0.2233	0.2003
sax	0.0402	0.0109	0.0084	0.2712	0.2294
sog	0.0305	0.0080	0.0074	0.2634	0.1913
gbb	0.0303	0.0147	0.0126	0.4865	0.4122
mad	0.0248	0.0053	0.0044	0.2120	0.1752
punt	0.0232	0.0091	0.0070	0.3923	0.3543
smox	0.0135	0.0026	0.0028	0.1933	0.1631
Inh/Ind					
nos	0.1145	0.0342	0.0270	0.2984	0.2706
vasa	0.0545	0.0263	0.0139	0.4822	0.4201
piwi	0.0446	0.0214	0.0186	0.4794	0.4394
pum	0.0320	0.0055	0.0045	0.1727	0.1356

Table S5. $\overline{dN}/\overline{dS}$, mean dN and mean dS across the phylogeny of four species of *Caenorhabditis* herein for each of the 23 genes under study. Values were measured using codeml in PAML [1]. SE=standard error. Note that additional decimal places of dN and dS than shown were used for calculation of $\overline{dN}/\overline{dS}$.

***C. elegans* Gene Name**	**DM Name**	**Mean dN/dS**	**Mean dN**	**SE**	**Mean dS**	**SE**
Lineage-specific Inheritance						
pie-1	-	0.1619	0.1243	0.0376	0.7681	0.1056
pgl-1	-	0.1553	0.1580	0.0053	1.0170	0.1834
Inheritance						
cyb-2	*cyb2*	0.1127	0.0826	0.0498	0.7324	0.2730
cpb-3	*orb*	0.0833	0.0714	0.0127	0.8576	0.1878
gcl-1	*gcl*	0.0719	0.1137	0.0257	1.5809	0.1900
ifet-1	*cup*	0.0570	0.0410	0.0032	0.7192	0.0571
stau-1	*stau*	0.0499	0.0957	0.0083	1.9185	0.4262
par-1	*par-1*	0.0433	0.0209	0.0041	0.4818	0.0576
etr-1	*bru1*	0.0386	0.0334	0.0044	0.8651	0.1563
mag-1	*mago*	0.0081	0.0054	0.0008	0.6591	0.1921
Induction						
dbl-1	*dpp*	0.0963	0.0247	0.0092	0.2568	0.0537
let-756	*bnl*	0.0748	0.0732	0.0113	0.9795	0.1237
sma-6	*sax*	0.0675	0.0537	0.0100	0.7961	0.0917
egl-15	*btl*	0.0639	0.0648	0.0189	1.0131	0.1962
blmp-1	*Blimp-1*	0.0575	0.0385	0.0133	0.6698	0.0429
sma-4	*med*	0.0548	0.0383	0.0101	0.6985	0.0689
tig-2	*gbb*	0.0522	0.0349	0.0198	0.6683	0.1161
crm-1	*sog*	0.0395	0.0238	0.0025	0.6027	0.0749
cwn-1	*wg*	0.0264	0.0256	0.0094	0.9688	0.1297
sma-2	*mad*	0.0128	0.0073	0.0058	0.5707	0.0986
Inh/Ind						
glh-1	*vasa*	0.0761	0.0419	0.0071	0.5506	0.0809
puf-8	*pum*	0.0753	0.0669	0.0088	0.8878	0.1105
prg-1	*piwi*	0.0642	0.0544	0.0214	0.8474	0.2268

Table S6. $\overline{dN}/\overline{dS}$, mean dN and mean dS across the phylogeny of four species of *Apis* herein for each of the germ line genes under study. Values were measured using codeml in PAML [1]. SE=standard error. Note that additional decimal places of dN and dS than shown were used for calculation of $\overline{dN}/\overline{dS}$.

Gene	$\overline{dN}/\overline{dS}$	Mean dN	SE	Mean dS	SE
Inheritance					
armi	0.1011	0.0068	0.0018	0.0673	0.0138
tud	0.0792	0.0054	0.0007	0.0682	0.0102
cycB	0.0774	0.0014	0.0008	0.0176	0.0024
vls	0.0673	0.0038	0.0004	0.0566	0.0170
spir	0.0564	0.0031	0.0012	0.0555	0.0096
bru-1	0.0560	0.0024	0.0028	0.0437	0.0155
stau	0.0395	0.0014	0.0010	0.0347	0.0137
gcl	0.0326	0.0011	0.0008	0.0336	0.0107
capu	0.0238	0.0041	0.0017	0.1699	0.0495
orb	0.0192	0.0006	0.0003	0.0332	0.0065
psq	0.0173	0.0018	0.0013	0.1060	0.0677
par-1	0.0067	0.0002	0.0002	0.0245	0.0041
mago	0.0001	0.0000	0.0000	0.0624	0.0335
Induction					
dpp	0.1342	0.0058	0.0044	0.0431	0.0146
sax	0.1163	0.0083	0.0033	0.0714	0.0092
wg	0.0854	0.0042	0.0039	0.0490	0.0082
sog	0.0627	0.0045	0.0024	0.0712	0.0080
wit	0.0462	0.0025	0.0013	0.0542	0.0126
Blimp-1	0.0375	0.0043	0.0017	0.1139	0.0253
gbb	0.0363	0.0015	0.0010	0.0411	0.0133
punt	0.0335	0.0016	0.0006	0.0473	0.0070
byn	0.0324	0.0020	0.0014	0.0624	0.0136
med	0.0133	0.0005	0.0003	0.0352	0.0049
tkv	0.0116	0.0008	0.0006	0.0656	0.0084
smox	0.0080	0.0003	0.0003	0.0314	0.0058
mad	0.0001	0.0000	0.0000	0.0443	0.0136
Inh/Ind					
piwi	0.1393	0.0105	0.0030	0.0755	0.0170
vasa	0.1155	0.0093	0.0018	0.0806	0.0144
pum	0.0038	0.0003	0.0003	0.0671	0.0147
nos	<0.0001	0.0000	0.0000	0.0460	0.0087

Table S7. The mean GC content and GC content at 3rd synonymous positions of codons (GC3s) for all genes under study in *Drosophila*, *Caenorhabditis* and *Apis*. SE=standard error.

	Drosophila					***Caenorhabditis***					***Apis***			
Gene	**GC**	**SE**	**GC3s**	**SE**	**Gene**	**GC**	**SE**	**GC3s**	**SE**	**Gene**	**GC**	**SE**	**GC3s**	**SE**
armi	0.500	0.011	0.588	0.032	*blmp-1*	0.429	0.020	0.279	0.024	*armi*	0.277	0.001	0.092	0.002
Blimp-1	0.591	0.006	0.708	0.017	*cpb-3*	0.454	0.020	0.373	0.025	*Blimp-1*	0.525	0.005	0.724	0.012
bnl	0.564	0.004	0.699	0.011	*crm-1*	0.449	0.020	0.339	0.038	*bru-1*	0.512	0.003	0.616	0.005
bru-1	0.590	0.004	0.737	0.012	*cwn-1*	0.443	0.021	0.425	0.046	*byn*	0.415	0.002	0.218	0.004
btl	0.541	0.010	0.639	0.028	*cyb-2*	0.487	0.021	0.577	0.034	*capu*	0.528	0.001	0.719	0.007
byn	0.605	0.003	0.687	0.009	*dbl-1*	0.450	0.021	0.395	0.024	*cycB*	0.332	0.002	0.214	0.004
capu	0.564	0.006	0.626	0.017	*egl-15*	0.411	0.022	0.359	0.026	*dpp*	0.368	0.002	0.160	0.006
cup	0.543	0.003	0.663	0.004	*etr-1*	0.474	0.022	0.425	0.042	*gbb*	0.578	0.002	0.734	0.005
cycB	0.561	0.003	0.711	0.008	*gcl-1*	0.384	0.022	0.318	0.012	*gcl*	0.326	0.001	0.174	0.004
dpp	0.623	0.008	0.835	0.021	*glh-1*	0.470	0.020	0.318	0.007	*mad*	0.361	0.001	0.093	0.003
gbb	0.595	0.008	0.822	0.024	*ifet-1*	0.482	0.020	0.388	0.002	*mago*	0.309	0.002	0.126	0.007
gcl	0.534	0.005	0.662	0.008	*let-756*	0.445	0.021	0.338	0.054	*med*	0.415	0.001	0.158	0.003
mad	0.574	0.003	0.772	0.007	*mag-1*	0.502	0.021	0.676	0.032	*nos*	0.429	0.003	0.343	0.004
mago	0.545	0.008	0.789	0.022	*par-1*	0.477	0.020	0.419	0.039	*orb*	0.396	0.001	0.244	0.003
med	0.628	0.002	0.674	0.006	*pgl-1*	0.474	0.020	0.446	0.008	*par-1*	0.455	0.002	0.362	0.006
nos	0.559	0.006	0.656	0.020	*pie-1*	0.480	0.020	0.423	0.026	*piwi*	0.361	0.001	0.158	0.004
orb	0.533	0.005	0.604	0.013	*prg-1*	0.403	0.020	0.513	0.009	*psq*	0.477	0.014	0.464	0.041
osk	0.515	0.004	0.652	0.010	*puf-8*	0.471	0.019	0.387	0.017	*pum*	0.508	0.004	0.410	0.011
pgc	0.540	0.007	0.712	0.015	*sma-2*	0.424	0.019	0.423	0.031	*punt*	0.360	0.001	0.166	0.003
piwi	0.468	0.009	0.513	0.026	*sma-4*	0.446	0.020	0.318	0.065	*sax*	0.321	0.001	0.112	0.003
psq	0.605	.002	0.711	0.006	*sma-6*	0.399	0.020	0.300	0.025	*smox*	0.459	0.001	0.380	0.003
pum	0.618	0.003	0.725	0.010	*stau-1*	0.449	0.015	0.394	0.032	*sog*	0.461	0.005	0.340	0.013
punt	0.541	0.003	0.682	0.010	*tig-2*	0.440	0.012	0.358	0.045	*spir*	0.383	0.001	0.218	0.003
sax	0.531	0.009	0.633	0.024						*stau*	0.384	0.000	0.163	0.001
smox	0.584	0.004	0.751	0.010						*tkv*	0.361	0.001	0.161	0.003
sog	0.596	0.005	0.761	0.013						*tud*	0.293	0.001	0.104	0.003
spir	0.601	0.004	0.765	0.009						*vasa*	0.374	0.003	0.232	0.009
stau	0.560	0.005	0.630	0.012						*vls*	0.303	0.001	0.088	0.005
tkv	0.551	0.006	0.663	0.016						*wg*	0.569	0.001	0.693	0.003
tud	0.486	0.013	0.569	0.038						*wit*	0.322	0.000	0.142	0.003

vasa	0.506	0.013	0.506	0.038									
vls	0.583	0.009	0.687	0.020									
wg	0.577	0.008	0.739	0.019									
wit	0.561	0.004	0.645	0.013									
Mean	0.561	0.006	0.683	0.016	0.450	0.020	0.400	0.029	0.405	0.002	0.293	0.006

Table S8. The stages of development wherein expression was quantified in *D. melanogaster* using transcriptome data from www.flybase.org.

Stage of Development
embryo 00-02hr
embryo 02-04hr
embryo 04-06hr
embryo 06-08hr
embryo 08-10hr
embryo 10-12hr
embryo 12-14hr
embryo 14-16hr
embryo 16-18hr
embryo 18-20hr
embryo 20-22hr
embryo 22-24hr
larva L1
larva L2
larva L3 12hr old
larva L3 puffstage 1-2
larva L3 puffstage 3-6
larva L3 puffstage 7-9
white prepupae new
white prepupae 12hr
white prepupae 24hr
pupae 2d postWPP
pupae 3d postWPP
pupae 4d postWPP
adult male 01day
adult male 05day
adult male 30day
adult female 01day
adult female 05day
adult female 30day

Text File S1: Comparison of Drosophila dN/dS to genome-wide values

For our reference and main target taxon *Drosophila*, we assessed the rate of divergence in the PGC-specification gene set to genome-wide values. For the six *Drosophila* species from the *melanogaster* group studied here, a database of dN/dS using the M0 model in PAML [1] has been generated for 8,510 genes with 1:1 single-copy orthologs in all taxa and is available at FlyBase [2, 3]. Using those datasets [3], we determined the average value of dN/dS was 0.0876 (standard error $\pm 9.0X10^{-4}$) and $\overline{dN}/\overline{dS}$ was 0.0837. Using the more recent and updated *Drosophila* database for the *melanogaster* group at flyDIVaS (downloadable version 1, also generated using M0 in PAML) [4], we obtained highly similar results with a genome-wide average value for dN/dS of 0.0864 ± 9.210^{-6} (N=8,656). Based on these values, the *Drosophila* $\overline{dN}/\overline{dS}$ results herein for *osk*, *bnl*, *capu*, *nos*, *orb*, *btl* and *pgc* (Table 2: 0.0933 to 0.1573) indicate these genes have diverged at a rate above the genome-wide average, consistent with particularly rapid evolution. Further, the germ line genes when taken as a collective group appear to span values observed in the genome as a whole, and thus do not exhibit strongly conserved or rapidly evolving as compared to the rest of the genome (as a group).

Text File S2:

Positive selection in Drosophila

A prior genome-wide study of the six species from the *Drosophila melanogaster* group has indicated that 878 of 8,150, or 10.7%, of genes examined have been subjected to positive selection [3] based on sites analysis in PAML [1]. This implies that the 34 germ line genes studied herein for *Drosophila* exhibit a lower percentage of genes with positive selection than observed in the genome as a whole; as we found relatively few genes (only 2 of 34 studied, or 6%) exhibited positive selection at codon sites (Table 5). However, this may be partly due to our conservative method of generating alignments: we aimed to improve divergence estimates by using GBLOCKS on our MEGA alignments [5] to remove any divergent segments that may have contained misaligned regions [6], and further aligned sequences by eye to remove residual ambiguous and divergent segments [7, 8]. We chose this conservative approach as prior data has demonstrated that ambiguous alignments can result in inflated estimates of positive selection at specific sites (e.g., ≥48% false positives), including those reported for the *Drosophila* genome-wide analyses [3, 9, 10]. Excluding the estimated 48% of false positives suggested to have occurred in the whole-genome *Drosophila* dataset [3] we are left with 52% (estimated true positives) of the 10.7% of genes in the genome that show signs of positive selection. This, level of positive selection (5.6%) is highly similar to that found in our gene line gene set in this taxon (6%), suggesting that the germ line genes studied here exhibit levels of positive selection that are typical for genes in the *Drosophila* genome as a whole. Nonetheless, whist our approach was highly conservative, we do not fully exclude the possibility that some genes in Table 5 might still have occasionally exhibited positive selection due to alignment ambiguity at specific sites [10].

Positive selection in Apis

It is worthwhile mentioning that among the three genera under study herein, sites analysis in PAML suggested positive selection was uncommon for *Drosophila* and *Caenorhabditis* germ line genes (found for two and three genes respectively, Table 5) and was most common in *Apis* (nine of the 30 genes; M7 versus M8 $2X\Delta lnL > 5.99$, $P<0.05$, Table 5). *dpp* was near statistical significance ($2X\Delta lnLiklihood=5.56$, $P=0.062$) in *Apis*, with four codon sites showing signs of positive selection by BEB posterior possibilities ($P>0.90$)). The genes exhibiting positive selection in *Apis* span all categories of germ line specification studied (categories 2-4 as there

were no lineage-specific genes in *Apis*, Table 5). Nonetheless, while positive selection appears most common in this genus, under a conservative interpretation, we note that *Apis* contains the least well-annotated genomes of the three genera studied here. Thus, we cannot fully exclude that the issues inherent to positive selection tests, which have been reported to be very site-sensitive and to have been consistently overestimated by up to an order of magnitude in the literature (as an example, reportedly overestimated in chimpanzees vs. humans) when based on genomes containing sequencing errors, incomplete annotation, imprecise ortholog matching, and ambiguous alignment segments [9], might partly contribute towards these findings. As an example, the alignment of *dpp* using BLASTp [11] yields an alignment nearly identical to that obtained using our alignment approach (see Methods), but that differs slightly (albeit is equally divergent) in positions 80-87, the precise segment of the protein in which we noted positive selection (Table 5). Thus even these conservative alignments are not perfectly unambiguous. Nevertheless, the elevated propensity for positive selection in the germ line genes in *Apis* suggests a striking difference from *Drosophila* and *Caenorhabditis*, and thus warrants further study as more data become available.

Text File S3: Expression breadth and gene evolution

It is notable that *nos*, the fastest evolving Inh/Ind gene, and *bnl* and *btl*, the fastest evolving Induction genes in *Drosophila* ($\overline{dN}/\overline{dS}$ = 0.1145, 0.1303 and 0.0935 respectively), each had low expression breadth using the >5 RPKM criterion in all 30 developmental stages (20 to 26.7%, Fig. 2BD). In contrast, genes that evolved exceptionally slowly, with $\overline{dN}/\overline{dS} \leq 0.0232$ ((e.g. *mago*, *punt*, and *smox*) were expressed in all 30 tissues using the criterion of >0 and >5RPKM (Fig. 2A-D). These trends in germ line specification genes are consistent with greater conservation of widely expressed genes, as has been often reported in animals [12-14]. Nonetheless, this pattern was not observed universally, as *stau* had 100% expression breadth but relatively elevated $\overline{dN}/\overline{dS}$ (0.0885) within the Induction group (ranked 3rd), while the Inh/Ind genes *vasa* and *piwi* displayed relatively low $\overline{dN}/\overline{dS}$ despite having low expression breadth (≤20%) using the >5RPKM criterion (Fig. 2A-D). Collectively, these results suggest that the level of pleiotropy likely significantly contributes to molecular evolution of most germ line genes. Further, low pleiotropy may contribute to the fast evolution of LSI genes (Table 2, Fig. 2B), where it may promote reduced interference from functions in other non-PGC tissues and thus facilitate positive selection [14] in these genes.

While *nos* was the fastest evolving Inh/Ind gene in *Drosophila* ($\overline{dN}/\overline{dS}$ =0.1145, mENC'=50.5 (±0.40)), and had comparatively low pleiotropy (Fig. 1B), the putative orthologous gene identified in *Apis* was surprisingly extremely conserved ($\overline{dN}/\overline{dS}$ =0.0001, mENC'= 38.9 (±0.1) (Fig. 2B, Table 4). We note also that the inter-genus orthologs were very divergent in protein sequence for *nos* between these genera, as observed previously for this gene across animal models [15]. As further developmental expression data becomes available for *Apis*, it will be worthwhile to assess whether high expression breadth is observed for the *nos* gene in that taxon (including embryonic expression [16]), putatively contributing to the much different evolution of this gene in these two genera. In addition, it is worth consideration of whether another putative *nos* ortholog can be identified in *Apis*.

Text File S4: Additional methods

To identify orthologs in all six *Drosophila* species, we compared the 34 CDS of *D. melanogaster* (using longest isoform per gene in the reference species, with rare exception the second longest) to the complete CDS list of each of the five remaining species (including all known isoforms, Table S1) using legacy BLASTX [11] as is employed in FlyBase [2], and for consistency was also used in standalone BLASTX (v. 2.2.18) searches; all cited e-values were obtained from that approach. The modified version BLAST+, which contains the same core programs (e.g., BLASTX), yielded the same interspecies matches per gene (with typically slightly lower e-values, data not shown). The best hit with the lowest e-value and $e<10^{-6}$ was taken as the match. Reciprocal searches were conducted using the best hit for each CDS back to the complete *D. melanogaster* CDS list. Those CDS that were best hits in reciprocal BLASTX searches were used for analysis. All gene identifiers of orthologs across species concurred with those predicted in the www.flybase.org orthology search tool. In some cases, a well-established germ line gene was part of a gene family in *D. melanogaster*: in those cases, we studied the single gene (or copy) most strongly linked to PGC-specification based on the literature, and its best match ortholog in compared taxa (Table 1). Thus, our analysis is of one-to-one best match orthologs of germ line genes [17], and excludes paralogs with presumably relatively weak or absent germ line roles, as such paralogous genes have often undergone changes of substrate or ligand specificity [17, 18].

For *Caenorhabditis* and *Apis*, we examined four model species wherein the genomes were sufficiently complete to allow identification of orthologs (containing the full CDS) for the genes under study (Table S1). The CDS list for the well-studied system *C. elegans* was used as the reference species for *Caenorhabditis*. In the nematodes, we identified orthologs to the gene list from *Drosophila* (Table 1) using the Orthology Search tool available at FlyBase (www.flybase.org) with the *D. melanogaster* FlyBase gene identifier as input, and *C. elegans* as the target taxon of interest. Using the resultant gene list in *C. elegans*, plus three additional nematode-specific genes involved in germ plasm (*pgl-1*, *pie1*, and *meg-1*) we searched for orthologs in the remaining three *Caenorhabditis* species under study, *C. brenneri*, *C. briggsae*, and *C. remanei*, using reciprocal BLASTX searches as described above. For *Apis*, where orthology to *D. melanogaster* genes has been less well studied and thus unavailable in orthology search tools, we conducted reciprocal BLASTX of the 34 *D. melanogaster* CDS (plus *par-1*) to

the complete CDS list for the most well studied model *A. mellifera*. Using the gene list from *A. mellifera*, we then searched for orthologs in the three additional *Apis* species, *A. cerena*, *A. dorsata*, and *A. florea*, using reciprocal BLASTX (S1 Table). The gene *par-1* was included for *Caenorhabditis* and *Apis*, but was not included the formal gene list for *Drosophila* as only three alignable orthologs (to the longest *D. melanogaster* isoform) were found among the six fly species. Results for *par-1* are provided in Supporting Text File 5. The unique gene identifiers for all genes studied per genus have been provided in Table 1 and Tables S2-S3. The final gene list for *Drosophila*, *Caenorhabditis* and *Apis* contained 34, 23 and 30 genes respectively.

For all genera, ortholog matches were further confirmed using alignments (see below). Only genes that had quality matches from ortholog searches (criteria described above) in all species under study per genus and could be confidently aligned were further examined. When two or more CDS, including isoforms for a single gene, had an identical e-value in a BLASTX search, we chose the one with the highest bit score as the best match. Using different isoforms may alter dN/dS or CUB marginally for some genes, but for consistency we always chose the isoform with the best reciprocal BLAST hit to the reference species. Occasional CDS with ambiguous sites were processed in ORF-predictor [19], which sometimes yielded a considerably shortened but highly reliable ORF (e.g., *tud* in *Drosophila*) and alignment for a gene.

Finally, we note that the *D. melanogaster* genes (34 in Table 2) that are not listed as having orthologs in *Caenorhabditis* (Table 3, Table S2) or in *Apis* (Table 4, Table S3) fell into one of three categories: 1) they did not have high confidence matches in *C. elegans* or *A. mellifera* using criteria defined above; 2) the C. *elegans* or *A. mellifera* ortholog also matched another second *D. melanogaster* gene; and/or 3) the ortholog matches were not found or poorly aligned across all four *Caenorhabditis* species or among all four *Apis* species. We chose to employ a conservative approach to ortholog identification in the study and thus any gene that fell into any one of these categories was excluded from analysis in the respective genus. Thus, we do not exclude that some of these genes excluded have orthologs, but did fall not within these criteria.

Text File S5: Results for par-1

It is noteworthy that for the gene *par-1*, which is involved in cell polarity and in stabilization of Osk in *D. melanogaster* and PIE-1 in *Caenorhabditis* [20-23] we could identify unambiguous orthologs for the longest isoform in only three of the *Drosophila* species studied here (*D melanogaster*, *D. erecta*, *D. simulans*); we thus did not include it in Table 2. Nonetheless using CDS for those three species, we found a $\overline{dN}/\overline{dS}$ of 0.0672, which would place it below the 10th place of the 13 Inheritance genes in Table 2. For *Caenorhabditis*, where we found *par-1* orthologs in all four species studied, we observed *par-1* had also had a relatively low $\overline{dN}/\overline{dS}$ within that genus (0.0433), well below the median for Inheritance genes (0.0751). However, it did exhibit signs of positive selection (2XΔLnLikelihood=22.5, P<0.05), with three sites identified by BEB analysis (P>0.95, and six sites with P>0.90, Table 5), suggesting adaptive evolution of this gene in nematodes, in *Apis*, the *par-1* gene evolved remarkably slowly with a value of 0.0067 (2nd slowest behind *mago*). Speculatively, the *par-1* protein might have been subjected to adaptive evolution in *Drosophila*, possibly in response to rapid changes observed in its phosphorylation target protein Osk [22, 24, 25], which had the highest $\overline{dN}/\overline{dS}$ (Table 2).

References

1. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood**. *Molecular Biology and Evolution* 2007, **24**(8):1586-1591.

2. Gramates LS, Marygold SJ, Santos GD, Urbano JM, Antonazzo G, Matthews BB, Rey AJ, Tabone CJ, Crosby MA, Emmert DB *et al*: **FlyBase at 25: looking to the future**. *Nucleic Acids Res* 2016.

3. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN *et al*: **Evolution of genes and genomes on the *Drosophila* phylogeny**. *Nature* 2007, **450**(7167):203-218.

4. Stanley CE, Jr., Kulathinal RJ: **flyDIVaS: A Comparative Genomics Resource for Drosophila Divergence and Selection**. *G3 (Bethesda)* 2016, **6**(8):2355-2363.

5. Kumar S, Stecher G, Tamura K: **MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets**. *Mol Biol Evol* 2016, **33**(7):1870-1874.

6. Talavera G, Castresana J: **Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments.** *Systematic Biology* 2007, **56**(4):564-577.

7. Carr M, Richter DJ, Fozouni P, Smith TJ, Jeuck A, Leadbeater BSC, Nitsche F: **A six-gene phylogeny provides new insights into choanoflagellate evolution**. *Mol Phylogenet Evol* 2017, **107**:166-178.

8. Heidel AJ, Kiefer C, Coupland G, Rose LE: **Pinpointing genes underlying annual/perennial transitions with comparative genomics**. *BMC Genomics* 2016, **17**(1):921.

9. Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D: **Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment**. *Genome biology and evolution* 2009, **1**:114-118.

10. Markova-Raina P, Petrov D: **High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes**. *Genome Res* 2011, **21**(6):863-874.

11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *Journal Of Molecular Biology* 1990, **215**(3):403-410.

12. Duret L, Mouchiroud D: **Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate**. *Mol Biol Evol* 2000, **17**(1):68-74.

13. Subramanian S, Kumar S: **Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome**. *Genetics* 2004, **168**(1):373-381.

14. Mank JE, Ellegren H: **Are sex-biased genes more dispensable?** *Biol Lett* 2009, **5**(3):409-412.

15. Subramaniam K, Seydoux G: ***nos-1* and *nos-2*, two genes related to *Drosophila nanos*, regulate primordial germ cell development and survival in *Caenorgabditis elegans***. *Development* 1999(126):4861-1871.

16. Dearden PK: **Germ cell development in the Honeybee (*Apis mellifera*); *vasa* and *nanos* expression**. *BMC Developmental Biology* 2006, **6**:6.

17. Hulsen T, Huynen MA, de Vlieg J, Groenen PM: **Benchmarking ortholog identification methods using functional genomics data**. *Genome Biol* 2006, **7**(4):R31.

18. Li WH, Yang J, Gu X: **Expression divergence between duplicate genes**. *Trends Genet* 2005, **21**(11):602-607.

19. Min XJ, Butler G, Storms R, Tsang A: **OrfPredictor: predicting protein-coding regions in EST-derived sequences**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W677-680.

20. Shulman JM, Benton R, St. Johnston D: **The Drosophila Homolog of *C. elegans* PAR-1 Organizes the Oocyte Cytoskeleton and Directs *oskar* mRNA Localization to the Posterior Pole**. *Cell* 2000(101):377-388.

21. Reese KJ, Dunn MA, Waddle JA, Seydoux G: **Asymmetric segregation of PIE-1 in C. elegans is mediated by two complementary mechanisms that act through separate PIE-1 protein domains**. *Mol Cell* 2000, **6**(2):445-455.

22. Riechmann V, Gutierrez GJ, Filardo P, Nebreda AR, Ephrussi A: **Par-1 regulates stability of the posterior determinant Oskar by phosphorylation**. *Nature cell biology* 2002, **4**(5):337-342.

23. Morais-de-Sa E, Vega-Rioja A, Trovisco V, St Johnston D: **Oskar is targeted for degradation by the sequential action of Par-1, GSK-3, and the SCF(-)Slimb ubiquitin ligase**. *Dev Cell* 2013, **26**(3):303-314.

24. Benton R, Palacios IM, St Johnston D: **Drosophila 14-3-3/PAR-5 is an essential mediator of PAR-1 function in axis formation**. *Dev Cell* 2002, **3**(5):659-671.

25. Ephrussi A, Lehmann R: **Induction of germ cell formation by *oskar***. *Nature* 1992, **358**(6385):387-392.