Heredity (2016) 116, 213–223

www.nature.com/hdy

ORIGINAL ARTICLE

Adaptive selection and coevolution at the proteins of the Polycomb repressive complexes in *Drosophila*

JM Calvo-Martín, P Librado, M Aguadé, M Papaceit and C Segarra

Polycomb group (PcG) proteins are important epigenetic regulatory proteins that modulate the chromatin state through posttranslational histone modifications. These interacting proteins form multimeric complexes that repress gene expression. Thus, PcG proteins are expected to evolve coordinately, which might be reflected in their phylogenetic trees by concordant episodes of positive selection and by a correlation in evolutionary rates. In order to detect these signals of coevolution, the molecular evolution of 17 genes encoding the subunits of five Polycomb repressive complexes has been analyzed in the *Drosophila* genus. The observed distribution of divergence differs substantially among and along proteins. Indeed, CAF1 is uniformly conserved, whereas only the established protein domains are conserved in other proteins, such as PHO, PHOL, PSC, PH-P and ASX. Moreover, regions with a low divergence not yet described as protein domains are present, for instance, in SFMBT and SU(Z)12. Maximum likelihood methods indicate an acceleration in the nonsynonymous substitution rate at the lineage ancestral to the obscura group species in most genes encoding subunits of the Pcl–PRC2 complex and in genes *Sfmbt*, *Psc* and *Kdm2*. These methods also allow inferring the action of positive selection in this lineage at genes *E(z)* and *Sfmbt*. Finally, the protein interaction network predicted from the complete proteomes of 12 *Drosophila* species using a coevolutionary approach shows two tight PcG clusters. These clusters include well-established binary interactions among PcG proteins as well as new putative interactions.
Heredity (2016) **116,** 213–223; doi:10.1038/hdy.2015.91; published online 21 October 2015

INTRODUCTION

Polycomb group (PcG) proteins constitute an epigenetic silencing system with a key role in the stable transcriptional repression of homeotic (*Hox*) genes, X chromosome inactivation, genomic imprinting, prevention of senescence, stem cell regulation, cell fate determination and cancer (as reviewed in Simon and Kingston, 2013). Although first identified in *Drosophila melanogaster*, these proteins are present from fungi to mammals and plants. PcG proteins have been purified as subunits of diverse multimeric complexes that modulate the chromatin state around the target genomic regions called Polycomb response elements (PREs) through posttranslational histone modifications. Genes silenced by the PcG complexes on their binding to PREs are mainly transcription factors and signaling pathway components.

In *Drosophila*, the two main Polycomb repressive complexes (PRCs) are PRC2 and PRC1. The core of PRC2 contains four proteins: enhancer of zeste (E(Z)), extra sex combs (ESC) or alternatively Esc-like (ESCL), suppressor of zeste 12 (SU(Z)12) and chromatin assembly factor 1 subunit (CAF1, also named NURF55). E(Z) has a SET domain and catalyzes the mono-, di- and trimethylation of histone H3 at lysine 27, which is the repressive chromatin hallmark. ESC boosts the enzymatic activity of E(Z), whereas SU(Z)12 and CAF1 are essential for nucleosome binding. A distinct form of the PRC2 complex contains an additional protein: polycomb-like (PCL). This five-protein complex (Pcl–PRC2) is responsible for the high levels of H3K27 trimethylation (H3K27me3) around the PREs that are needed to maintain the repressed state (Nekrasov *et al.*, 2007).

The core of PRC1 also contains four proteins: sex combs extra (SCE, also named dRING), posterior sex combs (PSC), polycomb (PC) and polyhomeotic (PH). SCE has an E3 ubiquitin ligase activity and it monoubiquitylates lysine 118 of histone H2A (H2AK118ub), which is a second chromatin silencing mark. PC has a chromodomain that specifically recognizes the H3K27me3 hallmark deposited by Pcl-PRC2, whereas PSC and PH inhibit chromatin remodeling. PRC1 core complexes can be copurified with the sex comb on midleg (SCM) protein that could contribute to the recruitment of PRC1 at PRE (Wang *et al.*, 2010).

The SCE and PSC proteins of PRC1 are two components of another Polycomb complex named dRAF (dRING-associated factors) that also contains the lysine (K)-specific demethylase 2 (KDM2) protein (Lagarou *et al.*, 2008). This last protein enhances the ubiquitin ligase activity of SCE and has a JmjC demethylase domain, which mediates the demethylation of H3K36 that in its methylated state is a signal of gene activation. Therefore, PcG proteins do not only incorporate silencing hallmarks in the chromatin but they also remove activation marks introduced by proteins of the Trithorax group. Repression of PcG target genes also requires the activity of the PR–DUB complex (Scheuermann *et al.*, 2010) that includes the calypso and additional sex combs (ASX) proteins and has an H2A deubiquitinase activity. Thus, surprisingly both H2A monoubiquinitation and deubiquitination have a role in Polycomb repression.

PhoRC (Pho repressive complex) is an additional Polycomb complex with two components: pleiohomeotic (PHO), or alternatively

Facultat de Biologia, Departament de Genètica, and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain
Correspondence: Dr C Segarra, Facultat de Biologia, Departament de Genètica, Universitat de Barcelona, Diagonal 643, Barcelona 08028, Spain.
E-mail: csegarra@ub.edu
Received 22 January 2015; revised 23 July 2015; accepted 10 August 2015; published online 21 October 2015

Pho-like (PHOL), and the product of the *Scm-related gene containing four MBT domains* gene (SFMBT). PHO and PHOL are the only two described PcG proteins with a DNA-binding affinity and thus PhoRC is crucial for anchoring the other PRCs at PREs. However, the process of recruitment of PcG complexes to chromatin is not fully characterized. According to the classical model, PhoRC would have a pivotal role in the initial recruitment of PRC2 to PREs. Thereafter, the H3K27me3 hallmark deposited by Pcl–PRC2 would trigger the recruitment of PRC1 by means of the chromodomain of PC. Nevertheless, PHO interacts not only with subunits of the PRC2 complex (Wang *et al.*, 2006) but also with subunits of the PRC1 complex (Mohd-Sarip *et al.*, 2002), which could also contribute to anchor PRC1 at PRE. In addition, it has been proposed that the recruitment of PcG complexes to chromatin might require additional not yet identified proteins (Wang *et al.*, 2010). The genome-wide identification of PRE regions by chromatin immunoprecipitation assays in *Drosophila* identified ~200 large Polycomb domains where PcG proteins colocalize (Schuettengruber *et al.*, 2009, and references therein). These elements are included in broader chromatin regions with the H3K27me3 hallmark. The characterized PREs do not share any sequence similarity among them, but contain multiple binding motifs for different DNA-binding proteins (including PHO) that would work combinatorially to recruit PcG complexes (Schuettengruber and Cavalli, 2009). Interestingly, PcG proteins bound to a PRE can silence rather distant promoters, which indicates that chromatin looping is important in the regulation by PcG proteins. In addition, PcG proteins are concentrated in nuclear foci termed PcG bodies, suggesting long-range chromosomal interactions to form distinct repressive compartments in the interphasic nuclei (Pirrotta and Li, 2012).

The multiple interactions among PcG proteins required to form the repressive complexes and to act in concert to repress chromatin might prevent their independent evolution. In fact, interacting proteins in protein–protein networks tend to evolve at similar rates (reviewed by Lovell and Robertson, 2010). Species with a well-characterized interactome, such as yeast, have provided the strongest support for this trend, given that interacting proteins exhibit highly correlated evolutionary rates (Hakes *et al.*, 2007). The correlation of evolutionary rates can be the result of molecular coevolution (or coadaptation according to Juan *et al.*, 2008) between the residues of interacting proteins. Indeed, the substitution of an amino acid in the interaction interface of one protein might be compensated in the binding partner protein by another substitution that maintains the integrity and functionality of the protein complex.

A wide range of bioinformatic methods have been designed to detect coevolution at the molecular level and thus to predict protein interactions (reviewed in Juan *et al.*, 2013). Mirror Tree was the first method developed to analyze the correlation of evolutionary rates between whole proteins. Initially, this method quantified the extent of the interaction between two proteins through the correlation coefficient of the distance matrix of each protein as inferred from the multiple alignments of the orthologous sequences in the same set of species. This method was later modified to correct by the background similarity expected even between non-interacting proteins due to phylogenetic relationships, that is, the common ancestry of the species included in the data set (Fraser *et al.*, 2004; Clark and Aquadro, 2010). An additional modification that also improved Mirror Tree performance is the Context Mirror approach that takes into account the possibility of multiple protein interactions and allows the detection of only those interactions specific to a single protein pair also in an evolutionary context (Juan *et al.*, 2008).

Here we analyze the molecular evolution of 17 genes encoding proteins of the Polycomb complexes PhoRC, Pcl–PRC2, PRC1, dRAF and PR–DUB in the 12 *Drosophila* species with whole sequenced genomes (Clark *et al.*, 2007) and in 3 additional species of the obscura group (*Drosophila subobscura*, *Drosophila madeirensis* and *Drosophila guanche*). The main aim of this study is to analyze whether gene divergence in the *Drosophila* genus is affected or not by the physical interactions of the encoded proteins either to form the complexes or to assist in their coordinated function. Indeed, the proteins that form these complexes are good candidates to detect correlated molecular evolution. The results obtained indicate that: (i) the distribution of divergence differs substantially among and across PcG proteins, CAF1 being the most uniformly conserved protein; (ii) highly conserved regions not yet described as protein domains but that might be relevant to protein function are present in SFMBT and SU(Z)12; (iii) an increase in the nonsynonymous divergence in the lineage ancestral to the obscura group species has driven the molecular evolution of most genes coding for subunits of the Pcl–PRC2 complex and also of genes *Sfmbt*, *Psc* and *Kdm2*; (iv) positive selection has acted in this lineage at genes *E(z)* and *Sfmbt*; and (v) there are two tight clusters of coevolving PcG proteins: one with ASX and the subunits of the Pcl–PRC2 complex and another with CALYPSO and subunits of the PhoRC, PRC1 and dRAF complexes.

MATERIALS AND METHODS

Fly samples and DNA sequencing

The *ch cu* strain of *D. subobscura* and highly inbred lines of *D. madeirensis* and *D. guanche* established after 10 generations of sib mating were available in our laboratory. DNA from these lines was purified with the PuregenCore Kit B (Qiagen, Hilden, Germany) according to the manufacturer's instructions. The 17 PcG genes from each species were PCR amplified with primers designed with the OLIGO v4.1 program (Rychlik, 1993) on the *Drosophila pseudoobscura* sequence available at flybase (www.flybase.org). Amplicons were purified with Qiaquick columns (Qiagen) and sequenced using the ABI Prism BigDye Terminators 3.0 Cycle sequencing kit (Applied Biosystems, Waltham, MA, USA). Sequencing reactions were run on an ABI PRISM 3700 automated DNA sequencer. Both strands were completely sequenced by primer walking using internal primers. The sequences of the primers used in the PCR amplification and sequencing, as well as the PCR conditions, are available on request to the authors. The Seqman program of the DNASTAR Lasergene package (Burland, 2000) was used to assemble the partial sequences.

Divergence analysis

The sequences of the 17 Polycomb genes in the 12 *Drosophila* species with sequenced genomes (Clark *et al.*, 2007) were retrieved from flybase (www.flybase.com) after Blast searches. The MUSCLE program (Edgar, 2004) implemented in the software MEGA6 (Tamura *et al.*, 2013) was used to multiple align the orthologous sequences of each gene according to the alignment of the predicted proteins. Alignments were manually checked and corrected whenever necessary. The identification of domains in the *D. melanogaster* protein sequences was performed using the InterProScan v5 program (Jones *et al.*, 2014) available in the EMBL-EBI web site (http://www.ebi.ac.uk/Tools/pfa/iprscan5). This program combines different protein signature databases and recognition methods into one resource. Additional protein motifs relevant for protein function described in the literature but not included in these databases were also identified. The putative presence of intrinsically disordered regions across the *D. melanogaster* proteins was analyzed with the PONDR-FIT predictor (Xue *et al.*, 2010) available at http://www.disprot.org.

The Clustal X v2.1 program (Larkin *et al.*, 2007) was used to infer the scores of amino acid conservation at each site of the multiple alignment of a given protein according to the protein weight matrix BLOSUM62. This analysis was performed with the raw protein alignments including regions with indels and with an uncertain alignment. These regions were manually excluded from the alignment in the subsequent analyses. The accepted phylogenetic tree of the

studied species was analyzed by maximum likelihood with the MEGA6 program (Tamura *et al.*, 2013), to infer and visualize the branch lengths of each protein phylogeny, according to the Jones–Taylor–Thornton model.

Selection analysis

The PAML v4 package (Yang, 2007) was used to compare different evolutionary models with alternative assumptions on the ω value in the accepted phylogeny of the studied species. ω estimates ($\omega = d_N/d_S$, where d_N and d_S correspond to nonsynonymous and synonymous divergence, respectively) were inferred with the CODEML program implemented in PAML. The M0 model that assumes the same ω value in all branches was compared with the free ratio model (FR) that considered a different ω in each branch. The M0 model was also compared with the branch model 2R that assumes that the ω value for particular branches of the phylogeny defined as foreground branches differs from the ω value of the rest of branches (background branches). The branch-site test of positive selection or test 2, as defined by Zhang *et al.* (2005), was also applied to detect the presence of positively selected sites in the foreground branch. In this test, the modified branch-site model A that includes a site class with $\omega_2 > 1$ (that is, under positive selection) in the foreground branch is compared with a null model with a fixed $\omega_2 = 1$ in this branch. For each model, the CODEML program implemented in PAML was run multiple times with different initial values to prevent incorrect parameter estimates caused by local optima. In all cases, a likelihood ratio test (LRT) was used to infer whether the null model could be rejected assuming that twice the log likelihood difference between two nested models differing in n free parameters follows a χ^2-distribution with n degrees of freedom. Sites under positive selection in the foreground branch were identified by the Bayes Empirical Bayes method (Yang *et al.*, 2005). The random effects branch-site model (REL) developed by Kosakovsky Pond *et al.* (2011) and available in the datamonkey web server (www.Datamonkey.org) was also applied to each gene, to detect branch-specific signals of episodic selection in the phylogeny.

Coevolution analysis

The Context Mirror approach (Juan *et al.*, 2008), based on the Mirror Tree method, was used to detect putative coevolution among the studied Polycomb proteins. With this purpose, the complete sets of protein-coding genes for the 12 *Drosophila* species, as well as their pairwise orhologous relationships, were downloaded from FlyBase (*D. melanogaster* release 5.56, *Drosophila simulans* r1.4, *Drosophila sechellia* r1.3, *Drosophila yakuba* r1.3, *Drosophila erecta* r1.3, *Drosophila ananassae* r1.3, *Drosophila persimilis* r1.3, *D. pseudoobscura* r3.1, *Drosophila willistoni* r1.3, *Drosophila grimshawi* r1.3, *Drosophila mojavensis* r1.3, and *Drosophila virilis* r1.2). The retrieved sequences were replaced when required by those reannotated or resequenced in the present study. Pairs of homologous genes were clustered into groups of multiple species with the mcl program available at http://micans.org/mcl. Only orthologous groups with a single gene copy (that is, 1:1) in each *Drosophila* species were retained for further analyses ($n = 6563$). For each 1:1 group, the most likely orthologous combination of protein isoforms was selected using the PALO software (Villanueva-Cañas *et al.*, 2013). These protein isoforms were aligned using the probabilistic framework provided by PRANK v1.4 (Löytynoja and Goldman, 2008). Perl scripts were developed to filter out alignment positions with a posterior probability lower than 99%. The amino acid distances among *Drosophila* species were estimated with the CODEML program of the PAML v4 package (Yang, 2007), separately for each orthologous group. As a proxy for coevolution, these amino acid substitution rates were compared across phylogenetic lineages, using the ContextMirror program with *P*-value cutoffs of 10^{-4} (Juan *et al.*, 2008). CytoScape (Shannon *et al.*, 2003) was used to represent significant instances of coevolution as a network.

RESULTS

Gene identification

Molecular evolution of 17 genes encoding subunits of five PRCs has been analyzed. The studied PcG genes are: *Sfmbt*, *Pho* and its paralog *Phol* (PhoRC complex); *E(z)*, *Esc*, its paralog *Escl*, *Su(z)12*, *Caf1* and *Pcl* (Pcl–PRC2 complex); *Sce*, *Psc*, *Pc*, *Ph-p* and *Scm* (PRC1 complex); *Kdm2* (dRAF complex); and *Calypso* and *Asx* (PR–DUB complex). These genes were identified in *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. pseudoobscura*, *D. persimilis*, *D. willistoni*, *D. mojavensis*, *D. virilis* and *D. grimshawi* after Blast searches against their complete genomes made available by the *Drosophila* 12 Genomes Consortium Initiative (Clark *et al.*, 2007). The first nine species are members of the *Sophophora* subgenus, whereas the last three belong to the *Drosophila* subgenus. The 17 genes were present in the 12 species. The paralog of *Ph-p* (*Ph-d*) was not included in the analysis given its absence in the species of the *Drosophila* subgenus. Several cases of sequencing errors and missannotations were detected (Supplementary Table S1). Most errors are single nucleotide indels that cause frameshift mutations and occasionally affect the predictions of the gene intron–exon boundaries. In addition, there are some cases where the sequence of a gene was partially incomplete in a species. Almost all these problems could be corrected by resequencing the corresponding gene region. The complete coding region of the 17 genes was also sequenced in the three closely related species *D. subobscura*, *D. madeirensis* and *D. guanche* (*subobscura* subgroup) that like *D. pseudoobscura* and *D. persimilis* are included in the *obscura* group of *Drosophila*.

The exon–intron organization of most genes (*Phol*, *Caf1*, *E(z)*, *Su(z) 12*, *Escl*, *Psc*, *Sce*, *Pc*, *Scm* and *Calypso*) as annotated in *D. melanogaster* is conserved in the studied species (Figure 1). In contrast, intron gain or loss events were detected in particular lineages at *Sfmbt*, *Esc*, *Pcl*, *Ph-p*, *Kdm2* and *Asx*. Finally, in *Pho* the exact boundaries of exons 2 and 3 could not be confirmed with certainty in all species given the high divergence of the gene in this region. Each gene is included in scaffolds assigned to the same Muller's element in the 12 species with complete sequenced genomes.

Divergence along the subunits of the PcG complexes

The distribution of the amino acid divergence along the multiple alignment of each protein is shown in Figure 2, where the described protein domains are also indicated. The PhoRC complex is formed by two subunits: PHO (or PHOL) and SFMBT. Divergence at PHO is rather high with some multiple alignment positions being quite uncertain. Only the four C2H2-type zinc fingers and the SP (spacer) domains (Lesley Brown *et al.*, 1998) are highly conserved. SP is a functionally characterized domain that is critical for the interaction of PHO with E(Z) and participates in the interaction with ESC (Wang *et al.*, 2004). PHOL shows a pattern of divergence similar to that of PHO, although it is not so divergent among species. SFMBT has an FCS-type zinc finger domain, four MBT (malignant brain tumor) repeats and a SAM (sterile α-motif) domain that are highly conserved.

The most conserved subunit of the Pcl–PRC2 complex is CAF1 (Figure 2), a member of the WD40 family with a seven-bladed β-propeller structure. Only 7 amino acid replacements (5 of them present in a single species) were detected in the 430 residues long CAF1 protein. The ESC and ESCL proteins that, like CAF1, are members of the WD40 family are also highly conserved not only in the characteristic WD40 region but also in the NTD domain (N-terminal domain), which in ESC binds to histone H3, and thus might contribute to enhance the enzymatic activity of E(Z). The SET catalytic domain of E(Z) characteristic of histone methyltransferases is extremely conserved. Indeed, no replacements were detected in this domain in the 15 species studied. The other E(Z) domains (O'Connell *et al.*, 2001) also show a low divergence (that is, the CXC domain preceding SET, EID (ESC interacting domain), PBD1, PBD2 (PCL binding domains), SANT and domain II that interacts with SU(Z)12). The three domains of SU(Z)12 are also highly conserved: the zinc

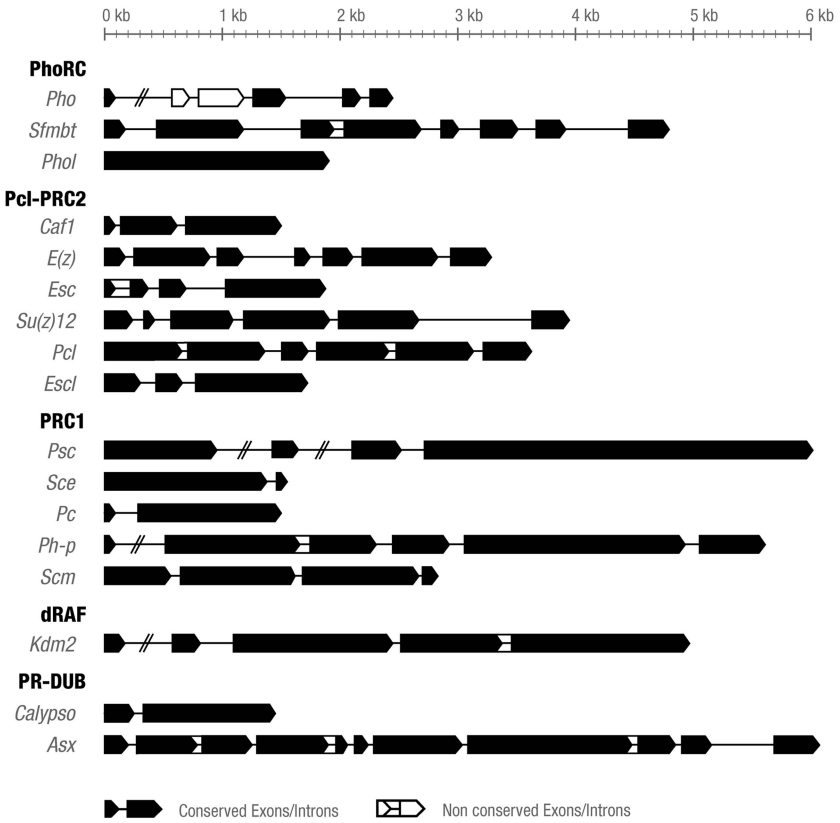

Figure 1 Comparative organization of the coding region in the studied Polycomb genes grouped by Polycomb complex in the 15 *Drosophila* species. Conserved exons are represented by black boxes that indicate the direction of transcription and conserved introns by lines. Empty white boxes indicate nonconserved exons and white boxes with internal lines indicate nonconserved introns. The length of the conserved introns interrupted by dashes is not proportional to the scale.

finger C2H2, VEFS box and the NTD that interacts with CAF1 (Nowak *et al.*, 2011). The presence in SU(Z)12 of two highly conserved regions between NTD and C2H2 not yet characterized as protein domains is noteworthy. PCL is the most divergent subunit of the Pcl–PRC2 complex. Protein conservation is nearly restricted to the described protein domains: Tudor, the two PHD-type zinc fingers, CLD (chromo-like domain; Wang *et al.*, 2004) and EH domain (extended homology domain; Wang *et al.*, 2004).

The core of the PRC1 complex has four subunits: PSC, SCE, PC and PH-P. Divergence along PSC and PH-P is distributed quite heterogeneously and highly divergent regions are interspersed with conserved regions, some of which correspond to the described protein domains: the RING-type zinc finger and the HTH (helix-turn-helix domain; Kyba and Brock, 1998) of PSC, and the FCS zinc finger and SAM domains of PH-P. Conservation is high around the RING-type zinc finger and the two HD domains (Gorfinkiel *et al.*, 2004) of SCE, which, in turn, has a rather high divergence outside the functional domains. A similar pattern is detected in SCM, where a high conservation is detected mainly in the protein domains (that is, the FCS zinc finger, the two MBT repeats, the Scm-like embedded domain (SLED) and the SAM domains). The chromodomain (CD) of PC that recognizes the H3K27me3 hallmark established by the E(Z) subunit of PRC2 is highly conserved, as it is its C-terminal domain (CTD; Franke *et al.*, 1995) and also a central region not previously identified as a protein domain. Divergence along KDM2 (that together with PSC and SCE forms the dRAF complex) shows highly conserved regions that include the JmjC catalytic domain, the CXXC-type zinc finger, the F-box and the Amn1/LRR domains. Finally, the pattern of divergence differs substantially along the two subunits of the PR–DUB complex. CALYPSO with a peptidase C12-ubiquitin carboxyl-terminal hydrolase domain is highly conserved, whereas conservation along ASX is almost only restricted to the ASXH and the PHD-type zinc finger domains.

Selection in the subunits of the PcG complexes

Protein divergence may vary in different lineages due to episodic adaptive selection resulting in an acceleration of the amino acid substitution rate, which can be reflected in the protein phylogenetic tree. In addition, in the case of interacting proteins the same lineages might be affected by positive selection due to protein coevolution. As a first approach to detect adaptive selection, phylogenetic trees were obtained for each PcG protein by maximum likelihood with the MEGA6 program (Tamura *et al.*, 2013) fixing the commonly accepted phylogeny of the 15 studied species, in order to infer branches lengths and therefore the substitution rate for each branch. The most striking result of this analysis is the strong acceleration in the amino acid substitution rate at the branch ancestral to the species of the obscura group (henceforth, *Bobs* branch or lineage) for three proteins of the Pcl–PRC2 complex: the two interacting proteins E(Z) and ESC, and PCL that interacts with E(Z). The same tendency, although not so strong, is detected in SFMBT of the PhoRC complex (Figure 3). Phylogenetic trees based on synonymous divergence in the corresponding genes do not exhibit this acceleration (results not shown), which indicates a decoupling between the synonymous and

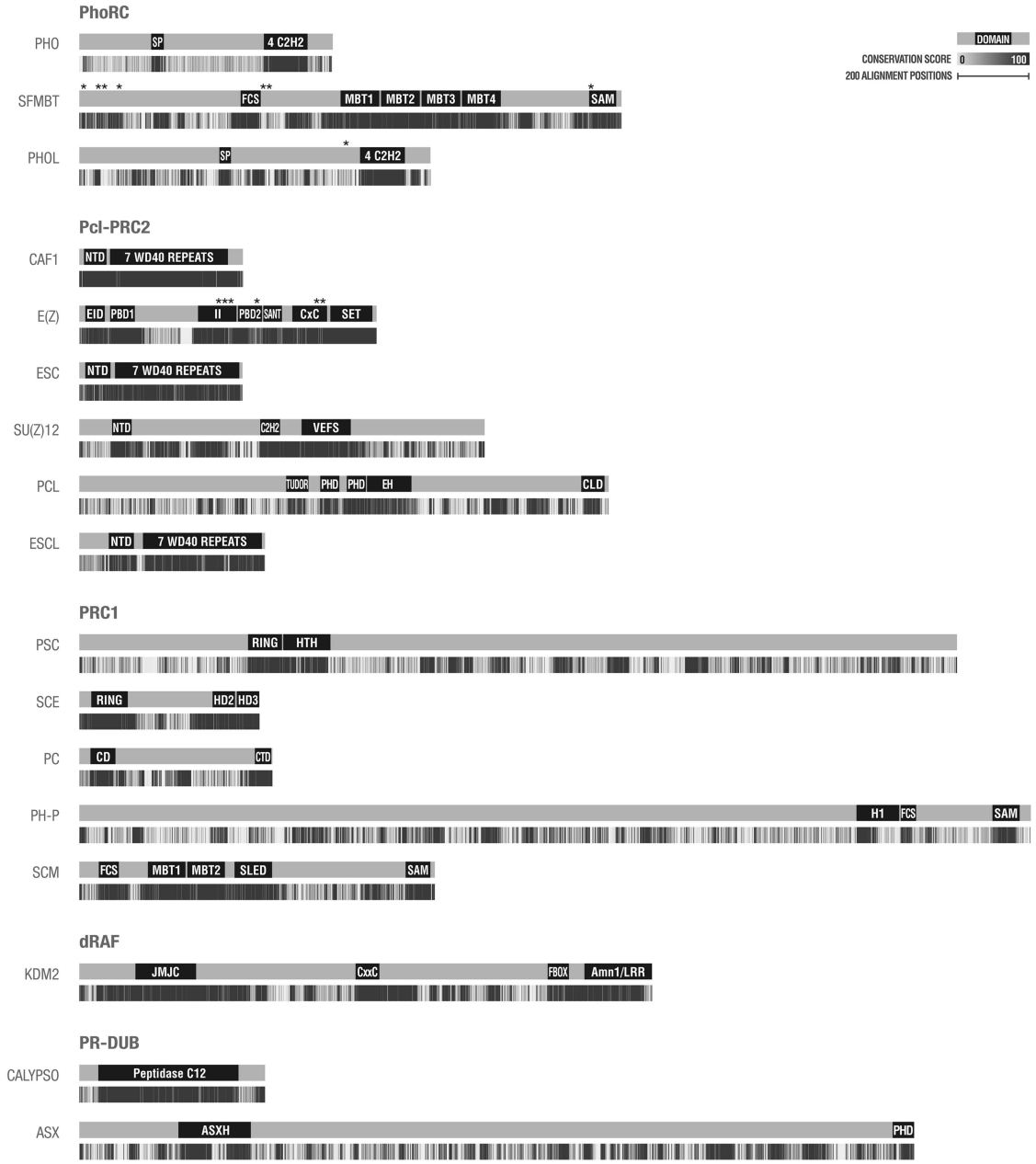

Figure 2 Protein conservation plots inferred with Clustal X v2.1 (Larkin *et al.*, 2007) along the multiple alignment of the studied Polycomb proteins grouped by Polycomb complex. Gray intensity indicates the conservation score of each protein position from 0 (white) to 100 (black). The bar above each conservation plot shows the described protein domains. Asterisks above the bar indicate amino acid positions under positive selection inferred by the Bayes Empirical Bayes method implemented in PAML.

nonsynonymous substitution rate in the *Bobs* branch. This result suggests positive selection acting on these proteins before the split of the *subobscura* and *pseudoobscura* subgroups and it even suggests the coevolution of some PcG proteins.

According to these results, the evolutionary models implemented in PAML (Yang, 2007) were applied to try to detect adaptive selection acting on PcG genes in particular lineages of the phylogeny. As a first approach, the M0 model, which assumes the same ω estimate ($\omega = d_N/d_S$, that is, the ratio between nonsynonymous and synonymous divergence) for all branches was compared with the free ratio (FR) model, which allows a different ω for each branch. Except for genes *Caf1*, *Pc*, *Scm* and *Calypso*, the M0 model could be rejected (Table 1), indicating heterogeneity in the ω estimates across branches in the phylogeny for most studied genes. Next, the M0 model was compared with a branch model ($2R\text{-}f_{Bobs}$) assuming two ω estimates: one for the branch leading to the obscura group species (*Bobs*) that was fixed as foreground branch (ωf_{Bobs}) and one for the rest of branches (ωb). The $2R\text{-}f_{Bobs}$ model is better supported by the data than the M0 model, not only for the four genes with a long *Bobs* branch in Figure 3 (*E(z)*, *Esc*, *Pcl* and *Sfmbt*) but also for *Su(z)12*, *Psc*, *Sce* and *Kdm2*. In *Esc* and *Su(z)12*, ω heterogeneity is exclusively explained by the *Bobs* branch, as the $2R\text{-}f_{Bobs}$ model could not be rejected when compared with the free ratio (FR) model (Table 1). The ML estimates under the best-fit model reveal that ωf_{Bobs} values are

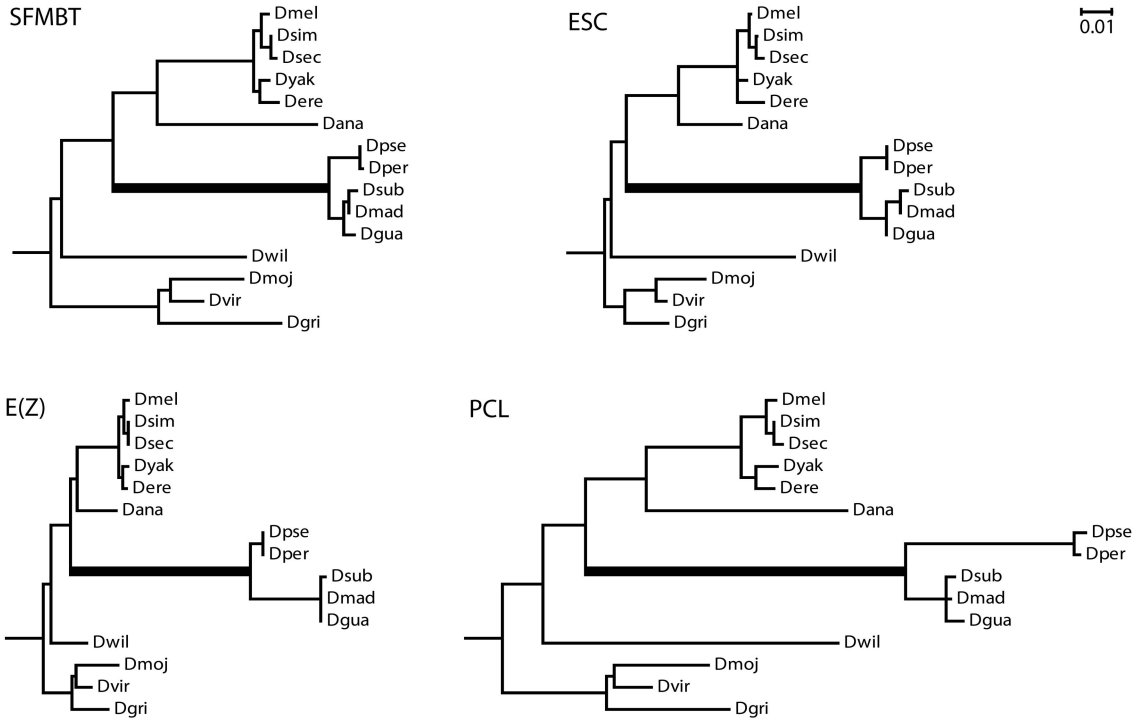

Figure 3 Phylogenetic trees according to protein divergence of the SFMBT subunit of the PhoRC complex and of the E(Z), ESC and PCL subunits of the Pcl–PRC2 complex. Branch lengths were obtained with MEGA6 (Tamura *et al.*, 2013) by maximun likelihood fixing the commonly accepted phylogeny of the studied species. The branch leading to the obscura group species (*Bobs* in the text) is highlighted. Dmel, *D. melanogaster*; Dsim, *D. simulans*; Dsec, *D. sechellia*; Dyak, *D. yakuba*; Dere, *D. erecta*; Dana, *D. ananassae*; Dpse, *D. pseudoobscura*; Dper, *D. persimilis*; Dsub, *D. subobscura*; Dmad, *D. madeirensis*; Dgua, *D. guanche*; Dwil, *D. willistoni*; Dmoj, *D. mojavensis*; Dvir, *D. virilis*; and Dgri, *D. grimshawi*.

significantly higher than ωb values in four of the six genes coding for subunits of the Pcl–PRC2 and also in *Sfmbt*, *Psc* and *Kdm2* (Figure 4). The highest difference is detected in *Esc* and *E(z)* with a $\omega f_{\text{Bobs}}/\omega b$ ratio equal to 13.9 and 8.7, respectively. Finally, the branch site test of positive selection (or test 2 in Zhang *et al.*, 2005) was applied, which showed that the differences at *E(z)*, *Sfmbt* and *Phol* in the ωf_{Bobs} and ωb estimates are at least partly explained by positive selection acting in the *Bobs* lineage. In addition, the Bayes empirical Bayes method identified six codons with a high posterior probability of having evolved under positive selection in the *Bobs* lineage at *E(z)*, seven codons at *Sfmbt* and one at *Phol* (Figure 2). All sites under positive selection at *E(z)* are located in gene regions that code for described protein domains. In contrast, only one of the selected codons at *Sfmbt* is in a region coding for a domain, but two of them are close to one such region. The remaining four codons cluster in a region at the beginning of the open reading frame and thus at the conserved N-terminal region of the protein. The single codon under positive selection at *Phol* is neither located in a conserved region nor in a region coding for a protein domain.

The described PAML analysis was performed fixing as foreground the *Bobs* branch suspected to have been under positive selection at least in some genes according to the phylogenetic trees (Figure 3). Kosakovsky Pond *et al.* (2011) developed a random-effects branch site method to detect episodic positive selection in particular lineages of a phylogeny that does not require defining *a priori* foreground and background branches. This likelihood approach (REL) was applied to each gene, to further corroborate the positive selection results. *E(z)*, *Esc*, *Phol*, *Sfmbt* and *Ph-p* showed evidence of episodic selection in the *Bobs* lineage (*P*-value = 0.001, 0.010, 0.042, 0.001 and 0.015, respectively). However, *P*-values remain significant only for *E(z)* and *Sfmbt* when correcting for multiple testing (corrected *P*-value = 0.014 and 0.037, respectively). Therefore, both PAML and REL maximum likelihood approaches infer the action of positive selection in the *Bobs* lineage in *E(z)* and *Sfmbt*.

Coevolution of the PcG proteins

The proteomes of the 12 *Drosophila* species with complete sequenced genomes (Clark *et al.*, 2007) were analyzed for the first time using the Context Mirror approach (Juan *et al.*, 2008), to uncover any evidence of coevolution among the PcG proteins. The interaction network of the proteins with significant partial correlations of evolutionary rates and thus predicted to coevolve is shown in Figure 5. This network includes only 16 of the 17 PcG proteins here studied, as PHOL did not reach the established *P*-value cutoff of significance. PcG proteins form two tight clusters of coevolving proteins (Figure 5). One predicted cluster interconnects linearly four of the five subunits of the Pcl–PRC2 complex: E(Z), ESC, PCL and SU(Z)12. There is, therefore, evidence that these four proteins have coevolved. The other two proteins of this complex (CAF1 and ESCL) are not included in this cluster. Interestingly, the Pcl–PRC2 cluster also includes ASX, one of the subunits of PR–DUB. The second predicted cluster interconnects CALYPSO (the other subunit of PR–DUB) and all the proteins of the PhoRC, PRC1 and dRAF complexes. Thus, PHO and SFMBT (PhoRC complex) are predicted to interact not only between them but also with the five proteins of the PRC1 complex (PC, PSC, SCE, PH-P and SCM), which, in turn, are predicted to interact among them. In addition, CALPYSO and the three proteins of dRAF (KDM2, SCE and PSC) are also interconnected both among them and with the proteins of PRC1. These results clearly indicate the power of the

Table 1 *P*-values of the LRTs to contrast different evolutionary models implemented in PAML (Yang, 2007)

PcG complex gene	*Models* [a]			
	M0 vs FR	*M0 vs 2R-f$_{Bobs}$*	*2R-f$_{Bobs}$ vs FR*	*Test 2*
PhoRC				
Pho	0.016*	0.449	0.013*	0.500
Sfmbt	0.000***	0.000***	0.044*	0.033*
Phol	0.002***	0.068	0.004***	0.017*
Pcl–PRC2				
Caf1	0.201	0.417	0.186	0.500
E(z)	0.000***	0.000***	0.001***	0.031*
Esc	0.000***	0.000***	0.439	0.231
Su(z)12	0.040*	0.004***	0.205	0.409
Pcl	0.000***	0.000***	0.000***	0.500
Escl	0.004***	0.309	0.004***	0.175
PRC1				
Psc	0.000***	0.014*	0.000***	0.500
Sce	0.000***	0.013*	0.001***	0.500
Pc	0.742	0.376	0.735	0.072
Ph-p	0.001***	0.427	0.001***	0.063
Scm	0.056	0.315	0.053	0.500
dRAF [b]				
Kdm2	0.000***	0.006**	0.000***	0.215
PR–DUB				
Calypso	0.780	0.134	0.859	0.500
Asx	0.000***	0.112	0.000***	0.500

Significance: *$0.05 > P > 0.01$; **$0.01 > P > 0.005$; ***$P < 0.005$.
[a]M0, a single ω for all branches; free ratio (FR), a different ω for each branch; *2R-f$_{Bobs}$*, one ω for the branch ancestral to the obscura group species and one ω for the rest of branches. Test 2 of positive selection as described in Zhang *et al.* (2005).
[b]dRAF complex also contains PSC and SCE.

Context Mirror approach to detect interacting proteins from the signals of coevolution reflected in their interespecific divergence.

Despite the good performance of Context Mirror in predicting the interactions between the PcG proteins, the method also displays a certain level of background noise, as it predicts some dubious interactions. Hence, the detected Pcl–PRC2 and the PhoRC/PRC1/dRAF clusters include additional proteins that have, in most cases, an uncertain Polycomb-related function. One exception is SCR (sex combs reduced) that is interconnected with CALYPSO and most proteins of the PhoRC/PRC1/dRAF cluster. It is well established (Gindhart and Kaufman, 1995) that *Scr* is a gene directly regulated by the Polycomb complexes. Our results would suggest a putative interaction of the homeotic protein SCR with Polycomb proteins to regulate some of its target genes.

DISCUSSION

Divergence among the *Drosophila* species differs substantially in the 17 Polycomb proteins studied here. The high conservation of CAF1 stands in contrast with the presence at PHO, PSC, PH-P or ASX of highly divergent regions interspersed with rather short conserved regions. However, the described domains in the different proteins are well conserved in all species without exception (Figure 2). In some cases, conserved regions extend beyond the described domains (for example, the JmjC and the CXXC-type zinc finger domains of KDM2 and the SAM domain of SFMBT). In addition, highly conserved protein regions not yet identified as important motifs for the protein function have been detected, for instance, in the N-terminal region of SFMBT and between the NTD and C2H2 domains of SU(Z)12 (Figure 2). Although evolutionary conservation can not be considered an unequivocal signal of functional relevance, the strong constraints acting on these regions make them good candidates to have an important and conserved role in protein function. Despite the action of strong purifying selection in these conserved regions, four codons with a high posterior probability of having evolved under positive selection were identified by the Bayes Empirical Bayes method implemented in PAML (Yang *et al.*, 2005) in the conserved N-terminal region of SFMBT. Therefore, the detected conserved motifs might be good targets to design experimental studies trying to identify interacting interfaces between proteins or major determinants of the Polycomb complexes function as chromatin modifiers.

However, sequence conservation is not an essential requirement to maintain function, as has been shown for the PcG protein PSC. The C-terminal region of PSC is poorly conserved in sequence across insects but broadly conserved in its ability to inhibit chromatin remodeling. Indeed, it is predicted to be a structurally disordered region, in which the presence of multiple patches of high positive charge and the lack of an extended stretch of contiguous negative charge is important for protein function (Beh *et al.*, 2012). The pattern of divergence of PH-P, PCL, PHO, PHOL and ASX with rather long, poorly conserved regions across the protein is similar to that of PSC. However, when analyzing the predicted structurally disordered regions across these proteins in *D. melanogaster*, only PH-P and ASX present a rather long stretch of ~800 and ~500 residues, respectively, with a high disorder score according to the PONDR-FIT method (Xue *et al.*, 2010). Even though selection on disordered protein might be relaxed, positive selection acting on *Ph-p* noncoding regions has been described after the detection of a selective sweep around the first intron of this gene in a study of nucleotide variation within *D. melanogaster* (Beisswanger and Stephan, 2008).

The high conservation of the four zinc finger domains responsible for the recognition and binding of PHO at the specific target sequences present at PREs reflects the selective pressures exerted on this protein domain. It is noteworthy in this context that PREs, which are complex *cis*-regulatory DNA elements containing multiple combinations of binding sites exhibit a high sequence plasticity. Indeed, interspecific comparisons show a high variation in the number, genomic location and motif design in *Drosophila* species (Hauenschild *et al.*, 2008). The rapid evolution of PREs might be related with the high divergence of PHO in regions outside the established functional domains.

Polycomb proteins interact among them and form functional complexes that repress gene expression. Numerous biochemical studies have focused on analyzing these interactions in *D. melanogaster*, yielding a rather well-characterized Polycomb network in this species. In addition, the regions directly involved in binary interactions have been identified for different PcG protein pairs as summarized in Figure 6. Therefore, Polycomb proteins are good candidates to contrast whether interacting proteins evolve coordinately, which might be reflected by concordant episodes of positive adaptive selection in their phylogenetic trees and by signs of coevolution. Most genes (*E(z)*, *Esc*, *Su(z)12* and *Pcl*) coding for subunits of the Pcl–PRC2 complex share a pattern of molecular evolution characterized by a ω value significantly higher in the branch leading to the obscura group species (*Bobs* branch) than in the background branches. In addition, *Sfmbt*,

Psc and *Kdm2* of the PhoRC, PRC1 and dRAF complexes, respectively, show the same pattern (Figure 4). Therefore, an acceleration of the nonsynonymous divergence in the *Bobs* branch seems to have driven the evolution of several Polycomb genes, which, in turn, suggests some kind of coevolution among the encoded proteins. However, the evolutionary models implemented in PAML and REL inferred the action of positive selection only in *E(z)* and *Sfmbt* that encode subunits of different PcG complexes between which no binary direct interactions have been detected (Figure 6).

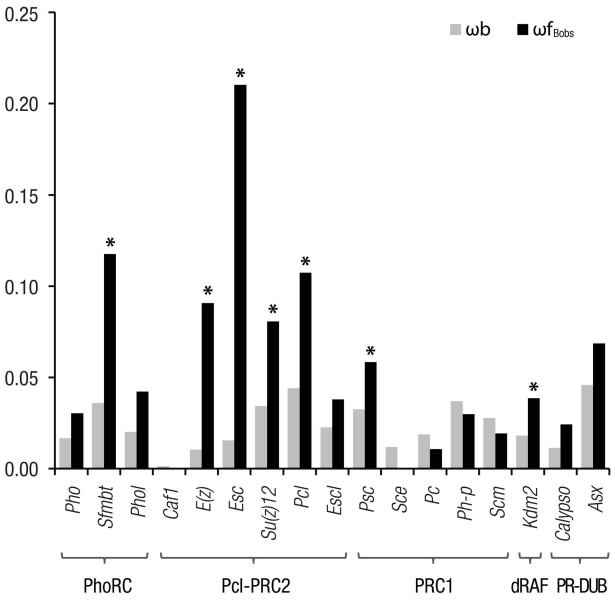

Figure 4 Maximum likelihood ω estimates inferred for each Polycomb gene by the 2*R*-f_{Bobs} branch model implemented in PAML v4 package (Yang, 2007) assuming two ω estimates: one fixing as foreground the branch ancestral to the obscura group species (ωf_{Bobs}) and one for the rest or background branches (ωb). Genes in which the 2*R*-f_{Bobs} model fits better the data than the M0 model (a single ω estimate for all branches) being $\omega f_{Bobs} > \omega b$ are indicated by an asterisk. Genes are grouped according to Polycomb complex.

The maximum-likelihood PAML and REL approaches analyze the phylogenetic tree of each gene independently and, although concordant results in different genes might indicate a coordinated evolution, the detection of coevolution requires more powerful bioinformatic methods, such as Mirror Tree and Context Mirror. These methods directly compare the phylogenetic trees of different proteins to detect a correlated evolution due to the presence of compensatory amino acid changes required to maintain the integrity and functionality of the protein complex. It has been questioned (Hakes *et al.*, 2007) whether the molecular coevolution of physically interacting amino acids might be reflected in the evolution of the whole protein given that the fraction of aminoacidic residues directly involved in the interaction between proteins is often small. However, several factors apart from selection on protein structure and function affect the rate of protein evolution. Some of these factors can cause similar constraints on the evolutionary rate of interacting proteins and therefore a correlated evolution. Likely, the most important of these external or indirect factors would be the similar expression level of the genes encoding interacting proteins, which ensures proper stoichiometry between the interacting components of the protein complex (Fraser *et al.*, 2004). These factors causing similar constraints between proteins and molecular coevolution are not mutually exclusive and both can act, to some extent, to cause a correlated evolution of interacting proteins (Juan *et al.*, 2008).

The proteomes of the 12 *Drosophila* species with complete genome sequences have been analyzed from a coevolutionary perspective focusing on the detection of the footprint of coevolution in the Polycomb proteins. The network of interacting proteins shows two tight clusters of coevolving proteins, one of them including four subunits of the Pcl–PRC2 complex (Figure 5). Among the predicted interactions in this cluster, only the binary interaction between E(Z) and ESC (E(Z)/ESC) has been well characterized experimentally (Jones *et al.*, 1998; Tie *et al.*, 1998). Context Mirror does not predict the proved interactions PCL/E(Z) (O'Connell *et al.*, 2001), E(Z)/SU(Z)12

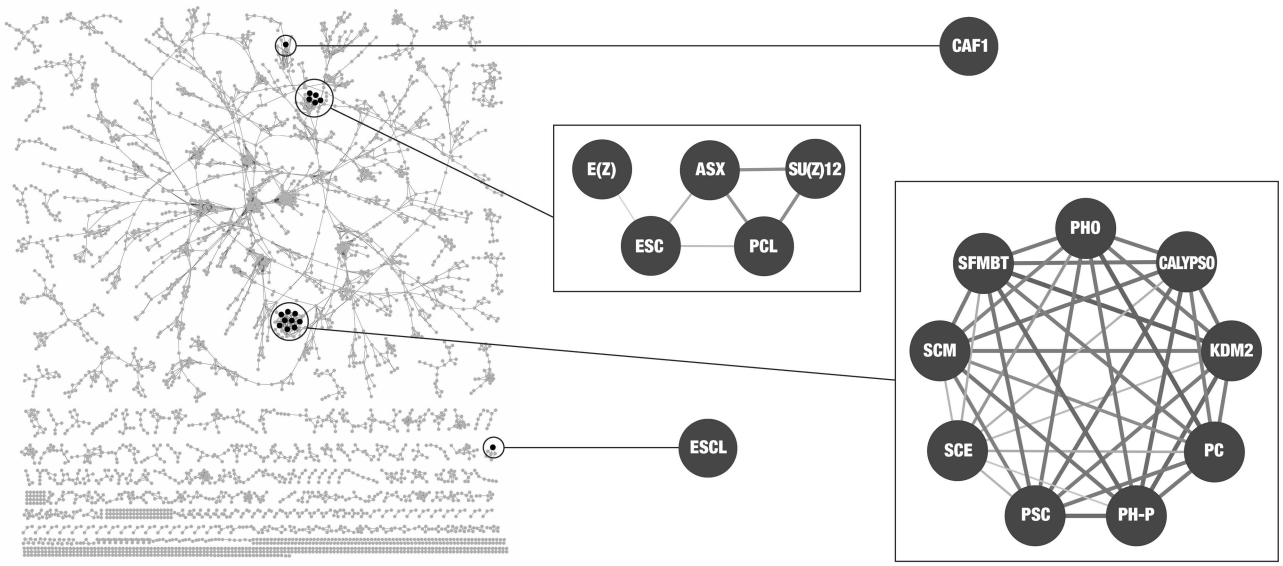

Figure 5 Left: interaction network predicted by the Context Mirror approach (Juan *et al.*, 2008) for proteins with significant partial correlations in evolutionary rates from the 12 *Drosophila* species proteomes (Clark *et al.*, 2007). Proteins in the network are represented by gray dots, except PcG proteins that shown by black dots. Right: zoom-in image of the predicted interactions between PcG proteins. The width and gray intensity of the connecting lines reflect the degree of the predicted interaction, which is stronger as the line gets wider and darker.

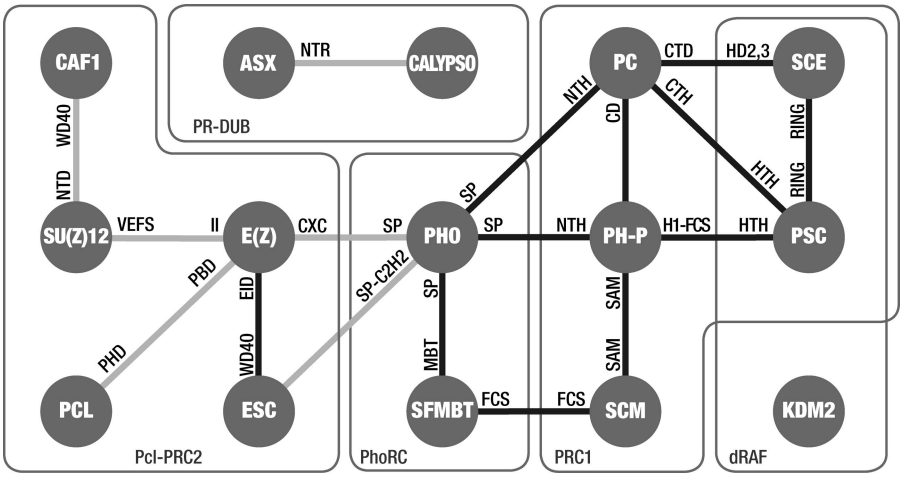

Figure 6 Binary interactions well characterized experimentally between PcG proteins of the Pcl–PRC2, PhoRC, PRC1 and dRAF complexes. Interactions predicted by Context Mirror (Figure 5) are shown by black lines, in contrast to those interactions not predicted by this bioinformatic method that are shown by gray lines. The protein domains included in the regions involved in the binary interactions are shown on each line close to the corresponding protein. Protein domains are indicated as in Figure 2, except NTH (N-terminal half), CTH (C-terminal half) and NTR (N-terminal region). See text for references.

(Ketel *et al.*, 2005; Joshi *et al.*, 2008) and SU(Z)12/CAF1 (Nowak *et al.*, 2011) (Figure 6). Indeed, CAF1 is the only subunit of the Pcl–PRC2 complex that is not predicted to interact with any of the other subunits of the complex in the network, although such interactions have been experimentally inferred (Nowak *et al.*, 2011). However, CAF1 is a component of different complexes that regulate chromatin metabolism (Song *et al.*, 2008) and it is under strong functional constraints, as reflected by its high conservation. Like CAF1, ESCL is not interconnected with any other Polycomb proteins in the network. Although ESCL can be assembled *in vitro* in functional PRC2 complexes (Wang *et al.*, 2006), this result would support that *in vivo* ESCL is not a usual subunit of PRC2 when ESC is present (Kurzhals *et al.*, 2008).

The most striking result of the coevolution analysis is that no interactions are predicted between the subunits of the PhoRC complex (PHO and SFMBT) and those of the Pcl–PRC2 complex, despite the experimental identification of the regions involved in the PHO/E(Z) and PHO/ESC interactions (Wang *et al.*, 2004). In contrast, Context Mirror predicts the PHO/SFMBT interaction (Alfieri *et al.*, 2013) and the well-characterized interactions between these two subunits of the PhoRC complex and those of PRC1: PHO/PC (Mohd-Sarip *et al.*, 2002), PHO/PH-P (Mohd-Sarip *et al.*, 2002) and SFMBT/SCM (Grimm *et al.*, 2009). The bioinformatic method also predicts the well-established interactions between subunits of the PRC1 complex, that is, PH-P/PC (Strutt and Paro, 1997), PH-P/PSC (Kyba and Brock, 1998), PH-P/SCM (Peterson *et al.*, 1997), PC/PSC (Kyba and Brock, 1998), PC/SCE and PSC/SCE (Gorfinkel *et al.*, 2004). In contrast, no interaction is predicted between the two subunits of the PR–DUB complex, although the N-terminal region of ASX is directly involved in its interaction with CALYPSO (Scheuermann *et al.*, 2010).

Therefore, our analysis corroborates 11 of the 17 well-established binary interactions between PcG proteins and fails to predict the remaining 6 interactions (Figures 5 and 6). In contrast, it predicts new interactions between the subunits that form the Pcl–PRC2, PRC1 and dRAF complexes, and also between subunits of PhoRC and PRC1. In addition, ASX is predicted to interact with subunits of Pcl–PRC2 complex and CALYPSO with subunits of the PhoRC, PRC1 and dRAF complexes. Further biochemical studies would be required to confirm these predicted interactions. In fact, the three subunits of dRAF (KDM2, PSC and SCE) coimmunoprecipitate (Lagarou *et al.*, 2008) and thus they have to interact to some extent, although the directly interacting subunits and the regions involved in these interactions have not been characterized yet.

It is also noteworthy that strong interactions between subunits of the PhoRC and PRC1 complexes are predicted in the coevolution analysis, which might be relevant to understand the recruitment of PRC1 at PRE. In fact, the strong evidence of coevolution detected between PHO and both PC and PH-P would indicate that these interactions have a pivotal role in the recruitment of PRC1 at PRE, in contrast to the classical view, suggesting that histone trimethylation by the E(Z) subunit of PRC2 is necessary for this recruitment. In addition, it has been proposed that SCM as a subunit of PRC1 would also be important for the recruitment of this complex at PRE (Wang *et al.*, 2010), which is consistent with the results of the coevolution analysis. The importance of PRC1 in the cooperative interactions to recruit PcG complexes at PRE has received further support in more recent studies (Kahn *et al.*, 2014). On the other hand, it should be noted that Context Mirror predicts the interactions SCM/PC, SCM/PSC and SCM/SCE, although SCM is present in substoichiometric quantities relative to the other subunits of the PRC1 complex (Wang *et al.*, 2010).

Finally, the acceleration in the nonsynonymous divergence detected in the *Bobs* branch in some genes by maximum likelihood (Figure 4) is neither necessary nor sufficient to explain the detected signals of coevolution (Figure 5). Although the higher ω in this branch (ωf_{Bobs}) than in the background branches (ωb) might have contributed to the predicted coevolution between subunits of the Pcl–PRC2 complex and also between SFMBT and KDM2 or PSC, other factors acting on the molecular evolution of the genes encoding these proteins in the 12 *Drosophila* species have to be likewise relevant. In fact, the partial correlations between protein evolutionary rates do not predict coevolution between SFMBT and the subunits of the Pcl–PRC2 complex despite the similar pattern of divergence in the *Bobs* branch in the genes coding these proteins and even despite the detected action of positive selection in this branch at *Sfmbt* and *E(z)*.

In summary, divergence in the *Drosophila* genus of the subunits that form the PhoRC, Pcl–PRC2, PRC1, dRAF and PR–DUB Polycomb complexes has allowed identification of putative new protein domains and the detection of episodic selection in the lineage ancestral to the

obscura group species in genes *E(z)* and *Sfmbt*, with six and seven codons, respectively, with a high posterior probability of having evolved under positive selection. In addition, the analysis of the proteomes of the species sequenced by the *Drosophila* 12 Genomes Consortium in a coevolutionary context clearly detects the footprint of coevolution not only between subunits of four Polycomb complexes, but also between subunits of different complexes. Indeed, our analysis confirms some of the well-characterized binary interactions between Polycomb proteins and predicts new interactions that deserve to be further investigated in future biochemical studies.

DATA ARCHIVING

The newly reported sequences are deposited in the EMBL/GenBank Data Libraries under accession numbers LN864764 to LN864814.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank G Mas de Xaxars, I Salvador and NL Clark for their input in an early phase of the project, and E Puerma for her contribution in the final stages.
We also thank Serveis Científico-Tècnics, Universitat de Barcelona, for automated DNA sequencing facilities. This work was supported by a predoctoral fellowship from Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), Generalitat de Catalunya, Catalonia, Spain, to JMC-M; and grants BFU2007–63229 and BFU2012–35168 from Ministerio de Economía y Competitividad, Spain, and 2009SGR-1287 and 2014SGR10555 from Comissió Interdepartamental de Recerca i Innovació Tecnològica, Generalitat de Catalunya, Catalonia, Spain, to MA.

Alfieri C, Gambetta MC, Matos R, Glatt S, Sehr P, Fraterman S *et al.* (2013). Structural basis for targeting the chromatin repressor Sfmbt to Polycomb response elements. *Genes Dev* **27**: 2367–2379.

Beh LY, Colwell LJ, Francis NJ (2012). A core subunit of Polycomb repressive complex 1 is broadly conserved in function but not primary sequence. *Proc Natl Acad Sci USA* **109**: E1063–E1071.

Beisswanger S, Stephan W (2008). Evidence that strong positive selection drives neofunctionalization in the tandemly duplicated *polyhomeotic* genes in *Drosophila*. *Proc Natl Acad Sci USA* **105**: 5447–5452.

Burland TG (2000). DNASTAR's Lasergene sequence analysis software. *Methods Mol Biol* **132**: 71–91.

Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA *et al.* (2007). Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**: 203–218.

Clark NL, Aquadro CF (2010). A novel method to detect proteins evolving at correlated rates: identifying new functional relationships between coevolving proteins. *Mol Biol Evol* **27**: 1152–1161.

Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797.

Franke A, Messmer S, Paro R (1995). Mapping functional domains of the Polycomb protein of *Drosophila melanogaster*. *Chromosome Res* **3**: 351–360.

Fraser HB, Hirsh AE, Wall DP, Eisen MB (2004). Coevolution of gene expression among interacting proteins. *Proc Natl Acad Sci USA* **101**: 9033–9038.

Gindhart Jr JG, Kaufman TC (1995). Identification of *Polycomb* and *trithorax* group responsive elements in the regulatory region of the *Drosophila* homeotic gene *Sex combs reduced*. *Genetics* **139**: 797–814.

Gorfinkiel N, Fanti L, Melgar T, García E, Pimpinelli S, Guerrero I *et al.* (2004). The *Drosophila* Polycomb group gene *Sex combs extra* encodes the ortholog of mammalian Ring1 proteins. *Mech Dev* **121**: 449–462.

Grimm C, Matos R, Ly-Hartig N, Steuerwald U, Lindner D, Rybin V *et al.* (2009). Molecular recognition of histone lysine methylation by the Polycomb group repressor dSfmbt. *EMBO J* **28**: 1965–1977.

Hakes L, Lovell SC, Oliver SG, Robertson DL (2007). Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proc Natl Acad Sci USA* **104**: 7999–8004.

Hauenschild A, Ringrose L, Altmutter C, Paro R, Rehmsmeier M (2008). Evolutionary plasticity of Polycomb/Trithorax response elements in *Drosophila* species. *PLoS Biol* **6**: e261.

Jones CA, Ng J, Peterson AJ, Morgan K, Simon J, Jones RS (1998). The *Drosophila* esc and E(z) proteins are direct partners in Polycomb group-mediated repression. *Mol Cell Biol* **18**: 2825–2834.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C *et al.* (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**: 1236–1240.

Joshi P, Carrington EA, Wang L, Ketel CS, Miller EL, Jones RS *et al.* (2008). Dominant alleles identify SET domain residues required for histone methyltransferase of Polycomb repressive complex 2. *J Biol Chem* **283**: 27757–27766.

Juan D, Pazos F, Valencia A (2008). High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc Natl Acad Sci USA* **105**: 934–939.

Juan D, Pazos F, Valencia A (2013). Emerging methods in protein co-evolution. *Nat Rev Genet* **14**: 249–261.

Kahn TG, Stenberg P, Pirrotta V, Schwartz YB (2014). Combinatorial interactions are required for the efficient recruitment of Pho repressive complex (PhoRC) to Polycomb response elements. *PLoS Genet* **10**: e1004495.

Ketel CS, Andersen EF, Vargas ML, Suh J, Strome S, Simon JA (2005). Subunit contributions to histone methyltransferase activities of fly and worm Polycomb group complexes. *Mol Cell Biol* **25**: 6857–6868.

Kosakovsky Pond SLK, Murrell B, Fourment M, Frost SDW, Delport W, Scheffler K (2011). A random effects branch-site model for detecting episodic diversifying selection. *Mol Biol Evol* **28**: 3033–3043.

Kurzhals RL, Tie F, Stratton CA, Harte PJ (2008). *Drosophila* ESC-like can substitute for ESC and becomes required for Polycomb silencing if ESC is absent. *Dev Biol* **313**: 293–306.

Kyba M, Brock HW (1998). The *Drosophila* Polycomb group protein Psc contacts ph and Pc through specific conserved domains. *Mol Cell Biol* **18**: 2712–2720.

Lagarou A, Mohd-Sarip A, Moshkin YM, Chalkley GE, Bezstarosti K, Demmers JAA *et al.* (2008). dKDM2 couples histone H2A ubiquitylation to histone H3 demethylation during Polycomb group silencing. *Genes Dev* **22**: 2799–2810.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H *et al.* (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–2948.

Lesley Brown J, Mucci D, Whiteley M, Dirksen M-L, Kassis JA (1998). The *Drosophila* Polycomb group gene *pleiohomeotic* encodes a DNA binding protein with homology to the transcription factor YY1. *Mol Cell* **1**: 1057–1064.

Lovell SC, Robertson DL (2010). An integrated view of molecular coevolution in protein-protein interactions. *Mol Biol Evol* **27**: 2567–2575.

Löytynoja A, Goldman N (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**: 1632–1635.

Mohd-Sarip A, Venturini F, Chalkley GE, Verrijzer CP (2002). Pleiohomeotic can link Polycomb to DNA and mediate transcriptional repression. *Mol Cell Biol* **22**: 7473–7483.

Nekrasov M, Klymenko T, Fraterman S, Papp B, Oktaba K, Köcher T *et al.* (2007). Pcl-PRC2 is needed to generate high levels of H3-K27 trimethylation at Polycomb target genes. *EMBO J* **26**: 4078–4088.

Nowak AJ, Alfieri C, Stirnimann CU, Rybin V, Baudin F, Ly-Hartig N *et al.* (2011). Chromatin-modifying complex component NURF55/P55 associates with histones H3, H4 and Polycomb repressive complex 2 subunit SU(Z)12 through partially overlapping binding sites. *J Biol Chem* **286**: 23388–23396.

O'Connell S, Wang L, Robert S, Jones CA, Saint R, Jones RS (2001). Polycomblike PHD fingers mediate conserved interaction with enhancer of zeste protein. *J Biol Chem* **276**: 43065–43073.

Peterson AJ, Kyba M, Bornemann D, Morgan K, Brock HW, Simon J (1997). A domain shared by the Polycomb group proteins Scm and ph mediates heterotypic and homotypic interactions. *Mol Cell Biol* **17**: 6683–6692.

Pirrotta V, Li H-B (2012). A view of nuclear Polycomb bodies. *Curr Opin Genet Dev* **22**: 101–109.

Rychlik W (1993). Selection of primers for polymerase chain reaction. In: White BA(ed) *PCR Protocols: Current Methods and Applications. Methods in Molecular Biology*. Humana Press Inc.: Totowa, NJ, vol 15, pp 31–40.

Scheuermann JC, Gaytán de Ayala Alonso A, Oktaba K, Ly-Hartig N, McGinty RK, Fraterman S *et al.* (2010). Histone H2A deubiquitinase activity of the Polycomb repressive complex PR-DUB. *Nature* **465**: 243–247.

Schuettengruber B, Cavalli G (2009). Recruitment of Polycomb group complexes and their role in the dynamic regulation of cell fate choice. *Development* **136**: 3531–3542.

Schuettengruber B, Ganapathi M, Leblanc B, Portoso M, Jaschek R, Tolhuis B *et al.* (2009). Functional anatomy of Polycomb and Trithorax chromatin landscapes in *Drosophila* embryos. *PLoS Biol* **7**: e1000013.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D *et al.* (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498–2504.

Simon JA, Kingston RE (2013). Occupying chromatin: Polycomb mechanisms for getting to genomic targets, stopping transcriptional traffic, and staying put. *Mol Cell* **49**: 808–824.

Song J-J, Garlick JD, Kingston RE (2008). Structural basis of histone H4 recognition by p55. *Genes Dev* **22**: 1313–1318.

Strutt H, Paro R (1997). The Polycomb group protein complex of *Drosophila melanogaster* has different compositions at different target genes. *Mol Cell Biol* **17**: 6773–6783.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* **30**: 2725–2729.

Tie F, Furuyama T, Harte PJ (1998). The *Drosophila* Polycomb group proteins ESC and E(Z) bind directly to each other and co-localize at multiple chromosomal sites. *Development* **125**: 3483–3496.

Villanueva-Cañas JL, Laurie S, Albà MM (2013). Improving genome-wide scans of positive selection by using protein isoforms of similar length. *Genome Biol Evol* **5**: 457–467.

Wang L, Brown JL, Cao R, Zhang Y, Kassis JA, Jones RS (2004). Hierarchical recruitment of Polycomb group silencing complexes. *Mol Cell* **14**: 637–646.

Wang L, Jahren N, Vargas ML, Andersen EF, Benes J, Zhang J *et al.* (2006). Alternative ESC and ESC-like subunits of a Polycomb group histone methyltransferase complex are differentially deployed during *Drosophila* development. *Mol Cell Biol* **26**: 2637–2647.

Wang L, Jahren N, Miller EL, Ketel CS, Mallin DR, Simon JA (2010). Comparative analysis of chromatin binding by sex comb on midleg (SCM) and other Polycomb group repressors at a *Drosophila Hox* gene. *Mol Cell Biol* **30**: 2584–2593.

Yang Z, Wong WSW, Nielsen R (2005). Bayes Empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol* **22**: 1107–1118.

Yang Z (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–1591.

Xue B, Dunbrack RL, Williams RW, Dunker AK, Uversky VN (2010). PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim Biophys Acta* **1804**: 996–1010.

Zhang J, Nielsen R, Yang Z (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* **22**: 2472–2479.

Supplementary Information accompanies this paper on Heredity website (http://www.nature.com/hdy)