Published in final edited form as:

Insect Biochem Mol Biol. 2016 February ; 69: 1–13. doi:10.1016/j.ibmb.2015.07.001.

Transcription factors, chromatin proteins and the diversification of Hemiptera

Newton M. Vidal^{a,b,c}, Ana Laura Grazziotin^{a,b}, Lakshminarayan M. Iyer^a, L. Aravind^{a,*}, and Thiago M. Venancio^{b,c,*}

^a National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

^b Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro, Campos dos Goytacazes, RJ, Brazil

^c Instituto Nacional de Ciência e Tecnologia em Entomologia Molecular (INCT-EM), Rio de Janeiro, RJ

Abstract

Availability of complete genomes provides a means to explore the evolution of enormous developmental, morphological, and behavioral diversity among insects. Hemipterans in particular show great diversity of both morphology and life history within a single order. To better understand the role of transcription regulators in the diversification of hemipterans, using sequence profile searches and hidden Markov models we computationally analyzed transcription factors (TFs) and chromatin proteins (CPs) in the recently available *Rhodnius prolixus* genome along with 13 other insect and 4 non-insect arthropod genomes. We generated a comprehensive collection of TFs and CPs across arthropods including 303 distinct types of domains in TFs and 139 in CPs. This, along with the availability of two hemipteran genomes, *R. prolixus* and *Acyrtosiphon pisum*, helped us identify possible determinants for their dramatic morphological and behavioral divergence. We identified five domain families (i.e. Pipsqueak, SAZ/MADF, THAP, FLYWCH and BED finger) as having undergone differential patterns of lineage-specific expansion in hemipterans or within hemipterans relative to other insects. These expansions appear to be at least in part driven by transposons, with the DNA-binding domains of transposases having provided the raw material for emergence of new TFs. Our analysis suggests that while *R. prolixus* probably retains a state closer to the ancestral hemipteran, *A. pisum* represents a highly derived state, with the emergence of asexual reproduction potentially favoring genome duplication and transposon expansion. Both hemipterans are predicted to possess active DNA methylation

* Corresponding author responsible for the submission: Thiago M. Venancio; thiago.venancio@gmail.com, Av Alberto Lamego 2000 / P5 sala 217, Campos dos Goytacazes / RJ; CEP: 28013-602, Phone number: +55 (22) 27486430. * Corresponding authors aravind@ncbi.nlm.nih.gov (L. Aravind). nwvidal@gmail.com (N.M. Vidal) analauragrazziotin@gmail.com (Ana Laura Grazziotin) lakshmin@ncbi.nlm.nih.gov (L.M. Iyer)

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

systems. However, in course of their divergence aphids seem to have expanded the ancestral hemipteran DNA methylation along with a distinctive linkage to the histone methylation system, as suggested by expansion of SET domain methylases, including those fused to methylated CpG recognition domains. Thus, differential use of DNA methylation and histone methylation might have played a role in emergence of polyphenism and cyclic parthenogenesis from the ancestral hemipteran.

Abstract

Keywords

insect development; lineage-specific expansion; genome evolution; transcription factor; transposable element; *Rhodnius prolixus*

1. Introduction

Insects are among the most species-rich eukaryotic clades on Earth. More than 1,000,000 species of insects have been described, representing 58-67% of all eukaryotic species (Footitt and Adler, 2009). Keeping with this species-richness, insects encompass a striking diversity of ecological, behavioral, morphological and physiological traits. Insects typically undergo ametabolous, hemimetabolous or holometabolous development, corresponding to absence of, partial and complete morphological metamorphosis in their life cycles. The four holometabolous orders, namely Coleoptera, Diptera, Hymenoptera and Lepidoptera comprise the majority of known insect species (81%) (Footitt and Adler, 2009). Moreover, several species from these orders have been extensively used as experimental models, not only to study insect development and biology, but also as more general eukaryotic model organisms (e.g. *Drosophila melanogaster*, *Tribolium castaneum* and *Apis mellifera*). However, outside of these major holometabolous lineages, several aspects of insect biology remain to be elucidated. In this context, the concomitantly published genome sequence of *Rhodnius prolixus* (Mesquita and Vionette-do-Amaral et al., submitted) offers an excellent opportunity to obtain a better understanding of the common features and divergences in the hemimetabolous developmental strategy.

A key question in insect biology is how ancestral developmental and differentiation processes have diverged between and within the clearly monophyletic orders to produce the observed morphological and physiological diversity. In conjunction with the genomic data from a diverse range of insect orders, recent phylogenomic studies have also converged to a large degree on a definitive phylogeny of the insects (Misof et al., 2014), which can now be

used as a framework to objectively interpret the observed patterns of molecular conservation.

Pioneering studies on *D. melanogaster* have shown that transcription factors (TFs) with DNA-binding domains such as homeo (Scott and Weiner, 1984), C2H2 Zn-finger, forkhead, bZIP and bHLH domains play critical roles in body plan determination and cell fate specification (Laurence, 1992; McGinnis and Krumlauf, 1992). These TFs act as regulators of distinct but overlapping developmental processes: 1) in early development to generate basic patterns, such as the morphological differentiated regions along the antero-posterior axis (e.g. the homeo domains of the Hox cluster along with the Homothorax/Extradenticle-like homeo domains) (Mann and Chan, 1996); 2) in broadly specifying organ primordia in development (e.g. the Eyeless/Pax6 proteins in specifying visual organs) (Gehring, 1996) and; 3) in later development to specify cell types corresponding to particular tissues (e.g. Glial-Cell-Missing a WRKY transcription factor that specifies glial cells) (Giangrande, 1996). The unified mechanism of action of these TFs involves binding of specific DNA sequences, and regulating the expression of target genes. TFs frequently interact with a distinct class of proteins (Mishra et al., 2003; Mohd-Sarip et al., 2002; Orsi et al., 2014) that primarily work by altering chromatin structure. In genetic and molecular studies on *D. melanogaster*, these proteins were primarily characterized as the products of the Trithorax and Polycomb genes, which are involved in maintaining open and condensed chromatin states; the former state is typically associated with transcriptional activation and the latter with repression. These chromatin proteins (CPs) can operate upstream and downstream of TFs to facilitate or restrict their access to chromatin or reinforce transcriptional states originally established by the TF by maintaining certain chromatin architectures (Filion et al., 2010; Steffen and Ringrose, 2014).

By analyzing TFs and CPs we sought to understand the genomic basis for the diversity in morphology and differentiation of insects. The availability of complete insect genomes provides an opportunity to comparatively analyze TF and CP families in insects to identify their potential contributions to divergence of different lineages. Several distinct factors contribute to the diversification of TF and CP repertoires over small and large evolutionary distances. Earlier studies showed that lineage-specific expansions (LSE) of gene families, followed by functional divergence of the paralogs, have played a major role in the evolutionary history of TFs and CPs in eukaryotes (Iyer et al., 2008). The LSE of developmental genes is not unprecedented in insects – members of the MADF/SAZ-BESS family of TFs, initially studied in dipteran wing development (Lander et al., 2001; Shukla et al., 2014), and the methuselah/methuselah-like gene family involved in organ morphogenesis (Patel et al., 2012), were found to be expanded in *Drosophila*. While many of these expansions happen via tandem gene duplication, transposable elements (TEs) such as the DINE-1 of *D. melanogaster* may trigger gene amplification and even allow for dispersal of the duplicated genes (Kogan et al., 2012).

Another factor in the evolution of TFs is the “domestication” of TEs (Balaji et al., 2005; Kaessmann, 2010). Systematic analysis of the DNA-binding domains in transcription factors have suggested that transposons directly contribute to TF evolution – the loss of activity of their catalytic core along with the retention of the DNA-binding domain appears to be a

mechanism by which novel TFs have arisen in several lineages (Aravind, 2000; Babu et al., 2006; Iyer and Aravind, 2012). Likewise, the terminal repeats of transposons could contribute to emergence of new regulatory elements that are bound by TFs to regulate genes (Marino-Ramirez et al., 2005).

Currently there are publicly available genome sequences for seven orders of Insecta, including the recently finished *R. prolixus* genome (Mesquita and Vionette-do-Amaral et al., submitted). Here we present a comprehensive analysis of TFs and CPs for 14 insects, 1 crustacean and 3 chelicerates. We identify LSEs of TFs and CPs and present a hypothesis for their possible role in diversification of hemipterans. We also present evidence for the domestication of transposons in the emergence of new lineage-specific TFs in the evolution of insects, hemipterans in particular.

2. Materials and methods

2.1. Data sources, sequence and phylogenetic analysis

Predicted proteomes from 18 species were downloaded from different public databases (Table S1). Among the 14 insects, 10 were representatives of the monophyletic holometabolous clade (*Drosophila melanogaster*, *Anopheles gambiae*, *Heliconius melpomene*, *Bombyx mori*, *Tribolium castaneum*, *Dendroctonus ponderosae*, *Harpegnathos saltator*, *Camponotus floridanus*, *Apis mellifera*, *Nasonia vitripennis*); *Pediculus humanus* is a representative of the clade Psocodea; *Rhodnius prolixus* and *Acyrtosiphon pisum* are hemipterans and in turn the higher order clade Condylgnatha; *Zootermopsis nevadensis* is an isopteran and representative of the higher order clade Polyneoptera. Psocodea, Condylgnatha and Polyneoptera form three successive higher-order outgroup clade of Holometabola. Four other non-insect arthropods, namely *Daphnia pulex* a diplostracan crustacean and the scorpion *Mesobuthus martensii*, the spider *Stegodyphus mimosarum*, and the tick *Ixodes scapularis* which are representatives of Arachnida were also used as successive outgroups of the monophyletic Insecta (Table S1).

Hidden Markov Models (HMM) for known TFs and CPs were manually selected from Pfam database version 27.0 (Finn et al., 2014). These were supplemented with additional PSI-BLAST profiles from an in-house database to improve detection of the homologs.

Redundant proteins present in each species proteome were removed using BLASTCLUST (parameters “-S 1.89 -L 0.9 -b F”) (Altschul et al., 1997) in order to avoid alternative transcripts or genome assembly problems. Domain architectures were obtained using the *hmmscan* program of the HMMer package version 3.1 (Eddy, 2009) (E-value 0.01) and PSI-BLAST (3 iterations; inclusion threshold of 0.01) (Altschul et al., 1997). Differences in the number of proteins among species within each TF or CP protein family were evaluated using a one sample t-test with Bonferroni correction for multiple tests in the R package (www.r-project.org).

Phylogenetic analysis of protein domains were performed using sequences covering at least 70% of the analyzed domain model. Sequences were aligned using MAFFT v7.1 (Katoh et al., 2002) or MUSCLE v3.7 (Edgar, 2004). Multiple sequence alignments were manually adjusted based on the secondary structure analysis and profile alignments. Phylogenetic

analyses were carried out using FastTree 2.1 approximately-maximum-likelihood algorithm and JTT model of amino acid evolution (Price et al., 2009).

2.2. Orthology Coefficient (OC)

In order to evaluate the conservation of transcriptional regulators between *R. prolixus* and other arthropod species, all 18 species proteomes were mapped to orthologous groups in the OrthoMCL database version 5 using the website mapping tool (Fischer et al., 2011). Orthology information was used to calculate the OC between species pairs for both transcription factors and chromatin protein categories, and for different protein families. The Orthology Coefficient (Subramanian et al., 2000) was calculated as following: $OC = (No1 + No2)/(N1 + N2)$, where No1 and No2 are the number of members in an orthologous group from the two genomes in question, while N1 and N2 are the number of members of a given protein family or category for the two compared genomes.

3. Results

3.1. Orthology Coefficients reveal conservation differences in transcriptional regulators of hemipterans

Using a list of manually curated Pfam domain profiles and PSI-BLAST profiles of TFs and CPs, we predicted potential transcriptional regulators in the 14 insects and 4 non-insect arthropods. Overall, we identified 303 and 139 types of TF and CP domains, respectively (Table S2). To examine the conservation and divergence patterns of transcriptional regulatory proteins among studied proteomes, we calculated Orthology Coefficients between *R. prolixus* and other insects and arthropods. The OCs between the entire proteomes of *R. prolixus* and other insects ranged from 0.54 to 0.67 (Table S3), while those between *R. prolixus* and non-insect arthropods were slightly lower, from 0.42 up to 0.56; as expected according to phylogenetic relationship among these species (Misof et al., 2014). For the overall category of transcription factors, OC values between *R. prolixus* and *D. melanogaster* (0.57), *A. mellifera* (0.64), *A. pisum* (0.43) and *D. pulex* (0.56) were found to be lower compared to OC values for the overall category of chromatin proteins (0.70, 0.75, 0.66 and 0.62, respectively) (Figure 1). The higher OC values for the CP category as opposed to TF category are consistent with the observation that several of the chromatin modification systems are strongly conserved across eukaryotes whereas TFs are more prone to lineage-specific diversification (Anantharaman et al., 2007; Iyer et al., 2008). Indeed, it has long been recognized that gain and loss of TFs and divergence in the DNA-binding domains of transcription factors are major contributors to phenotypic diversity (Romero et al., 2012).

Strikingly, among all comparisons between *R. prolixus* and other insects, the hemipteran pair *R. prolixus*-*A. pisum* showed the lowest OC value (0.43) for TFs category (Figure 1). Further, inspection revealed that this low OC is not limited to the comparison between the two hemipterans, but also holds for *A. pisum* with respect to all other insects (Table S3). This strongly suggests that whereas the kissing bug has retained a largely conservative TF repertoire the pea aphid has undergone a dramatic transformation of its TF complement relative the ancestral insect. TF domains suggest that this low OC value holds across the

different unrelated families of TFs suggesting that it is related to evolution of regulatory novelty in the aphid. Among these, the SAZ/MADF (Lander et al., 2001) (Pfam: PF10545) and Pipsqueak (Lehmann et al., 1998) (Pfam: PF05225) domains, both versions of the helix-turn-helix domains, are among hemipteran LSEs. However, we found three families of TFs which stood out as exceptions to the low OCs between *R. prolixus* and *A. pisum* in the TF category (Table S3): forkhead (Pfam: PF00250)(Clark et al., 1993; Weigel et al., 1989), homeo (Pfam: PF00046)(Gehring, 1992; Lewis, 1978) and nuclear hormone receptor (Pfam: PF00104)(Tanenbaum et al., 1998) domain families. For these TF families, high OC values were observed across all arthropods investigated here. This suggests that there are strong constraints on their numbers per genome and might reflect their recruitment as regulators of key developmental decisions which are conserved widely across animals (see below for discussion).

Two CP domain families showed a reduced conservation between the two hemipterans. Although POZ/BTB (Pfam: PF00651) (Aravind and Koonin, 1999; Bardwell and Treisman, 1994; Zollman et al., 1994) and SET (Pfam: PF00856)(Tripoulas et al., 1996) domains have similar OC values between other arthropods and *R. prolixus*, the *R. prolixus*-*A. pisum* pair showed lower values for POZ/BTB and SET domains (0.25 and 0.49 respectively). The POZ/BTB domain is present in a large array of animal TFs and mediates homodimerization and recruitment of the Cullin-3 dependent ubiquitin E3 ligase systems (Collins et al., 2001; Perez-Torrado et al., 2006). The SET domain is the protein methyltransferase catalytic domain that methylates particular lysine residues, especially on histones (Aravind et al., 2011; Herz et al., 2013). Closer examination revealed that the low OCs for both domain families were due to unusual LSEs in the pea aphid genome, suggesting the diversification of key chromatin level regulators between the two hemipterans, as discussed in the next section. Overall, the results presented above suggest that diversification of the TFs and, to a lesser extent CPs, might have contributed to the notable morphological and behavioral divergence between the two hemipteran lineages (Romero et al., 2012).

3.2. A phylogenetic gleaning from hemipterans on conservative aspects of TF evolution

Consistent with the OC analysis, phylogenetic analysis revealed orthology patterns approximating a one-to-one correspondence within arthropods for the forkhead, homeo and nuclear hormone receptor families. As an illustration of this, we present the phylogeny of the arthropod representatives of the forkhead family (Figure 2), which has nearly 20 members in most arthropods. While there are few examples of lineage-specific paralog formation via recent duplication, the majority of the tree is dominated by versions, which can be traced back to the common ancestor of the arthropod lineages included in this study. This suggests that, on the whole, the evolution of the forkhead family is dominated by a tendency to preserve ancient lineages rather than being subject to gene gain/loss novel lineage-specific versions.

The homeo domain family, with major developmental roles in Metazoa (Favier and Dolle, 1997; Holland, 2013), displays an average of 83 members in Insecta. This family size is roughly 1/3rd-1/4th the number observed in vertebrates (e.g. 255 in humans and 315 in zebrafish) which have undergone rounds of large-scale duplication of these genes (Holland,

2013; Zhong and Holland, 2011). Notably, the average number of homeo domains in insects is also lower than the observed numbers in crustaceans like *D. pulex* (104) and the scorpion *M. martensii* (132). Given this background, it is important to note that *R. prolixus* with 66 homeo domains has the lowest count for any arthropod in our set. In contrast, *A. pisum* with 90 homeo domains shows a count more typical of insects. We searched for the ten homeo domains present in the ancestral insect Hox complex. Orthologs of labial (lab), proboscipedia (pb), deformed (Dfd), ultrabithorax (Ubx) and abdominal-B (Abd-B) were found in the *R. prolixus* genome. But orthologs of Sex comb reduced (Scr), Antennapedia (Antp) and Abdominal-A (Abd-A) were not found in the current predicted proteome. However, analysis of DNA sequences using TBLASTN revealed fragments of these genes suggesting that they are indeed encoded in the genome but have not been included in the predicted genes due to issues with genome assembly or gene prediction. The Hox genes are found clustered in a collinear chromosomal order in the majority of metazoans. However, in line with above evidence for incomplete assembly of the genomic scaffolds in the current version of the genome we were not able to identify any contiguous genomic region bearing the *R. prolixus* Hox cluster. Orthologs of other Hox-related homeo domains such as bicoid (bcd), zerknullt (zen) and fushi tarazu (ftz), recruited for non-homeotic functions were not detected at all. Thus, the apparently loss of multiple homeo domains in *R. prolixus*, despite the high level of orthology observed with respect to other arthropods, needs closer scrutiny in future studies.

We detected specific differences in the conservation trends of transcriptional regulators – certain families like forkhead, homeo and nuclear hormone receptor showed clear signs of being under greater evolutionary constraints than others, such as POZ/BTB, SET, SAZ/MADF and Pipsqueak. Both forkhead and homeo domains are key developmental regulators, with primary roles in the establishment of body axes during animal embryogenesis (Heffer and Pick, 2013). Additionally, forkhead domains are downstream of different signaling pathways and play key roles in tissue specification and differentiation, with biological consequences including roles in longevity (Myatt and Lam, 2007). Involvement of a large number of their representatives (e.g. all members of Hox and para-Hox clusters in the case of homeo domains) in basic patterning events of the Bauplan, aspects of which are conserved throughout the evolution of bilaterian metazoans, could have placed greater limits on their diversification, especially within crown-group Arthropoda. The nuclear hormone receptors show LSEs in nematodes and vertebrates relative to arthropods. Moreover, in arthropods there are generally fewer nuclear hormone receptors than in either vertebrates or nematodes studies to date (15-25 as opposed to 50 to greater than 100 in vertebrates and rhabditiform nematodes). Thus, our results suggest that within arthropods these are probably utilized only downstream of a small number of ligands, including the molting-related hormones (e.g. ecdysone), which are a common feature throughout Arthropoda.

3.3. Large families of TFs and CPs help understand the trends in expansions and gene loss within Insecta

3.3.1. Transcription Factors—Our results show a list of typically metazoan abundant domains (e.g. C2H2 Zn-finger and bHLH). We took advantage of the 14 insect proteomes

studied here to confirm the insect-specific expansion of the C4DM family and tried to trace its possible evolutionary scenario (Table 1). Certain families containing classical C2H2 Zn-finger domains (represented in Pfam by models zf-C2H2; Pfam: PF00096 and zf-H2C2_2; Pfam: PF13465) were the most abundant TFs not only in hemipterans (Figure 3), but also in other arthropod species (Table 1). In several metazoans the C2H2 Zn-fingers are usually fused to other accessory domains, which mediate specific interactions or recruit other regulatory complexes such as the ubiquitin ligases to chromatin.

One such is the C4DM (also known as ZAD or zf-AD; Pfam: PF07776), a treble-clef fold domain involved in protein-protein interactions (Jauch et al., 2003), which was initially identified in the fruit fly (Lander et al., 2001). We found it in considerably fewer numbers in hemipterans (12 copies in *R. prolixus* and 28 in *A. pisum*) when compared to the 95 members in *D. melanogaster* (Figure 3, Table 1). Among other insects, hymenopterans showed 15-32 C4DM domain proteins, coleopterans 47-60, lepidopterans and dipterans, 87-110 and the termite *Zootermopsis nevadensis* 62 paralogs (Table 1). The domain has been previously shown to be expanded only in *D. melanogaster* (Lespinet et al., 2002). Our analysis showed for the first time that the C4DM is expanded in Insecta compared to other arthropods except for the two paralogs in *D. pulex*. This suggests that the domain probably originated first within the pancrustacean lineage, followed by an initial LSE in Insecta, with further independent LSEs on multiple occasions in different insect lineages. Nevertheless, hemipterans, *R. prolixus* in particular, retain numbers close to the ancestral condition in Insecta. Additional large LSEs of the C4DM domain independently occurred in termites and within holometabolous insects in the common ancestor of beetles, flies and moths. Similar to C4DM, the POZ/BTB is also commonly associated with C2H2 Zn-finger domains, although it might also associate with other DNA-binding domains such as the BEN, AT-hook and WRKY superfamily domains (Aravind and Koonin, 1999; Babu et al., 2006). However, unlike the C4DM domain, the POZ/BTB domain is expanded across all arthropods. Although it is also expanded in vertebrates, it has been previously shown that the two expansions happened independently (Aravind and Koonin, 1999). Within insects, hemipterans display an interesting pattern, with *R. prolixus* displaying a copy number similar to other insects in our set (Table 1), whereas *A. pisum* harbors a massive LSE with 217 POZ/BTB domains, which is greater than any other insect in our set.

The α -helical basic helix-loop-helix (bHLH) (Pfam: PF00010)(Murre et al., 1989) which binds DNA as an obligate dimer, has been widely utilized in several eukaryotic lineages. It is often present in large counts in several distinct eukaryotes, e.g. more than 100 bHLH domain proteins are encoded by the genomes of *Homo sapiens* and *Arabidopsis thaliana* (Lander et al., 2001). TFs with this domain regulate a wide range of developmental processes, including neurogenesis, myogenesis, sex determination, gut development, and hematopoiesis (Jones, 2004). In arthropods we found 40-66 bHLH proteins (Table 1) with an average of 49. *R. prolixus* with just 40 bHLH proteins is significantly below average ($p=7.6 \times 10^{-6}$); here, it deviates from the other hemipteran, *A. pisum* (54 bHLH proteins), which has a more typical count for insects. Thus, in contrast to the C4DM, which retained the ancestral lower counts in hemipterans, the bHLH family appears to be a notable case of gene-loss in *R. prolixus*. The above observations suggest that the differential expansion and

gene loss of TFs in the two hemipteran lineages might have played a major role in their evolution.

3.3.2. Chromatin Proteins—POZ/BTB, PHD finger (Pfam: PF00628)(Schindler et al., 1993) and BACK (Pfam: PF07707)(Stogios and Prive, 2004) domains are the most abundant families of CPs across arthropods (Table 1). Interestingly, these domains show remarkable differences in family size between the two hemipteran species. While *R. prolixus* has respectively 67, 37 and 27 copies of POZ/BTB, PHD finger and BACK proteins, which are in agreement with observed numbers in other arthropods, *A. pisum* has significantly higher protein copy numbers for POZ/BTB (217; $p=4.6 \times 10^{-8}$), PHD finger (111; $p=1.6 \times 10^{-10}$), and BACK (206; $p=8.6 \times 10^{-11}$). In this context it is notable that the only other arthropod in our dataset which show a comparably significant expansion of POZ/BTB domains is *D. pulex* (211; $p=9.5 \times 10^{-8}$). Unlike other arthropods, *A. pisum* and *D. pulex* have large genomes (>30,000 genes), harboring several duplicated genes (Colbourne et al., 2011). POZ/BTB domain proteins are transcriptional regulators that play key roles in the specification of cell fates and differentiations across a wide-range of tissues in arthropods (Aravind et al., 2009). Hence, it is conceivable that the morphological and physiological plasticity observed in these arthropods could be in part associated with the newly emerged POZ/BTB domain proteins adding distinct control steps for or causing altered expression patterns of downstream target genes.

Another notable example of differences between hemipteran gene complements is the SET domain family. SET domains methylate lysine residues on histones H3 and H4. The SET domain co-occurs with distinct CP domains in a variety of domain architectures, which specify distinct histone methylation systems. In contrast to the 20 SET domain proteins in *R. prolixus*, which is within the observed range for insects (20-38 copies), *A. pisum* has an expanded set of 65 SET domain-containing proteins ($p=9.6 \times 10^{-10}$), representing the largest SET group among all sequenced arthropods to date (Table 1). Given the importance of the SET domains in the establishment of chromatin states relevant to gene expression, we thoroughly inspected their structural diversity in order to understand the nature of the LSEs in *A. pisum*. We found the LSEs encompassing three distinct architectures: 1) solo *Suv4-20*-like SET domain (with some versions fused to C2H2 Zn-finger in *A. pisum*), 2) CG8378-like SET domains with an inserted MYND finger domain, and 3) Eggless-like SET domain fused to TUDOR and TAM/MBD domains (Figure 4).

The pea aphid has 27 members of the fruit fly *Suv4-20* SET domain family. Several of these proteins are truncated likely due to incomplete assembly or gene prediction problems, but at least 6 copies are fairly complete, of which five show novel fusions to 1-3 C2H2 Zn-finger domains (Figure 4). In *D. melanogaster*, *Suv4-20* is responsible for 90% and 5% of all dimethylated and trimethylated lysines at histone H4, respectively (Yang et al., 2008). These di- and trimethylation marks facilitate the response to DNA damage and heterochromatin formation. The expansion of this family in *A. pisum*, along with the distinctive fusion to Zn-finger domains, suggests that they could potentially act at localized regions of chromatin to facilitate particular condensed heterochromatin states via H4K20 methylation.

Both *A. gambiae* (6 copies) and *A. pisum* (10 copies) showed independent expansions of the *D. melanogaster* CG8378-like Smyd subfamily of SET domains (Figure 4). Smyd-like SET domains contain a MYND finger inserted within the SET domain (Pfam: PF01753) (Gross and McGinnis, 1996). The best characterized Smyd member in insects is the *D. melanogaster smyd4* gene, which represses transcription interacting with histone deacetylases 1 and 3 through protein-protein interaction mediated by its MYND domain (Liu et al., 2007; Rees et al., 2011). The functions of the expanded Smyd4 like families remain poorly understood; however, roles for these proteins in transcriptional repression cannot be ruled out, as suggested for *Drosophila* Smyd4.

The third expanded family of SET domains in *A. pisum* is the ortholog of the *Drosophila* Eggless SET domain, which is fused to Tudor and TAM/MDB domains (Figure 4). The overall similarities of domain architecture and size between the *D. melanogaster* Eggless and the *A. pisum* homologs suggest that these proteins are likely to perform similar functions. Eggless catalyzes the trimethylation of histone H3K9 in somatic and gametic cells during early and late stages of oogenesis (Clough et al., 2007). Mutations in *eggless* (including a complete deletion of the SET domain) result in meiosis arrest and defective oogenesis. In addition, these mutations were also linked to wing and locomotion defects in *D. melanogaster* (Clough et al., 2007; Clough et al., 2014). Given the role of TAM/MDB domains in binding sequences with the CpG dinucleotide with cytosine methylation, the expansion of this family of SET domains in *A. pisum* might correlate with the evolution of a more prominent role for DNA methylation in controlling gene expression and development in the pea aphid. In particular, it would be interesting to investigate their potential role in the aphid-specific morphological polymorphism and the phenomenon of parthenogenesis which is integral to their life history.

In contrast to the above, members of the Trithorax complex (*trithorax* and *ash1*) and *mes-4* that are part of conserved eukaryotic chromatin complexes catalyzing H3K4 and H3K36 methylation (Jiang et al., 2013; Lhoumaud et al., 2014; Ringrose and Paro, 2004), show mostly one-to-one ortholog relationship in arthropods (Figure 4). The high selective pressure against the diversification of these genes likely from their pivotal roles in establishment of chromatin states conserved throughout eukaryotes, which in animals include development maintenance of expression patterns (*trithorax* and *ash1*) (Ringrose and Paro, 2004) and co-regulating the chromatin accessibility to genes flanked by insulators (*mes-4*) (Lhoumaud et al., 2014).

3.4. Role of transposons in provenance and lineage-specific expansions of TF in hemipterans

The above observations suggest that the differential expansion and gene loss of TFs and CPs in the two hemipteran lineages have played a major role in their evolutionary divergence. Hence, we compared TF family sizes in hemipterans with the fruit fly *D. melanogaster* to get a more detailed understanding of the LSEs that could have contributed to the divergence of *R. prolixus* and *A. pisum* along rather distinct lines. Of the several families that stood out in this analysis, Pipsqueak is expanded in both *R. prolixus* and *A. pisum* (Figure 3). While holometabolous insects have from 7 (in *P. humanus*) to 13 (in *N. vitripennis*) pipsqueak

genes, *R. prolixus* and *A. pisum* have 36 ($p=2 \times 10^{-3}$) and 84 ($p=2 \times 10^{-10}$), respectively (Table 1). Phylogenetic reconstructions indicate that these expansions happened independently in each lineage after the split between *R. prolixus* and *A. pisum* (Figure 5, Figure S1). Sequence analysis of these Pipsqueak proteins revealed that in *R. prolixus* three versions are still fused to endonuclease or other domains related to Pogo transposons, whereas in *A. pisum* 71 of the copies show fusions to transposon (transposase or transposon-encoded) domains. Phylogenetic analysis showed two distinct patterns of evolution among these sequences: 1) clades comprising only TFs from multiple species; and 2) clades with interspersed TFs and transposon sequences. An example of the former case is the clade containing the *jim lovell* gene from *D. melanogaster* (Figure 5, Figure S1). One-to-one orthology relationships are found between most insects and *D. pulex*, suggesting that the recruitment of the DNA-binding domain from the transposon to the TF probably happened in the pancrustacean ancestor. On the other hand, the second case comprises clades with transposons and TFs sharing a common DNA-binding domain, suggesting that the DNA-binding domains of several transposons which have lost their transposase domains have been exapted as novel TFs.

In four other cases, there were LSEs only in the *A. pisum* genome but not in *R. prolixus*. One expansion was in the HTH domain SAZ/MADF ($p=9.5 \times 10^{-11}$) while the other three were in families of Zn-binding domains: FLYWCH (a version of the WRKY domains, $p=5.6 \times 10^{-11}$), THAP ($p=4.7 \times 10^{-11}$) and the BED finger ($p=4.5 \times 10^{-11}$). SAZ/MADF domain proteins are present in 11 copies in *N. vitripennis* and *D. ponderosae* (both holometabolous insects), and up to 49 in *D. melanogaster*. About ~30 SAZ/MADF genes are present in *R. prolixus* and *Z. nevadensis* and, remarkably, 178 copies in *A. pisum* (Table 1, Figure S2). Eight out of these 178 copies also show direct fusions to transposon domains of the Harbinger transposon superfamily (Kapitonov and Jurka, 2004).

The FLYWCH domain proteins range in number from 0 to 33 in other arthropods, as opposed to a striking 166 in *A. pisum* (Table 1, Figure S3). Seventy-four members show fusions to endonuclease and other domains of the Mutator superfamily. The THAP domain family ranges from 2-52 members in most arthropods. In remarkable contrast, *A. pisum* displays 430 copies whereas *R. prolixus* has only 3 copies (Table 1, Figure S4). Expansions of THAP domains in other insects have previously been linked to the P-element superfamily of transposons such as Galileo in *D. melanogaster* (Marzo et al., 2013). A similar situation is observed with the BED finger domain (Aravind, 2000), which is present in a single gene in *R. prolixus* in contrast to 103 in the pea aphid. This count is an order of magnitude greater than what has been observed in other arthropods (Table 1, Figure S5). Sixty-seven of the copies in *A. pisum* also harbor fusions to transposon domains of the hAT superfamily of DNA transposons.

LSEs involving TFs harboring DNA-binding domains derived from TEs have long been recognized (De Marco et al., 1982; Hoen et al., 2006; Juretic et al., 2005). The transposition process of DNA transposons in the host genome might contribute to genomic diversity: while still mobile, the transposon may duplicate and increase its copies in distinct *loci*, resulting in genomic plasticity due to recombination. Transposon-derived coding-sequences may serve as raw material for the emergence of stable functional host genes, and in such

case for instance, a protein with co-opted DNA-binding is likely to play a role as a transcription factor (Aravind, 2000; Babu et al., 2006; Iyer and Aravind, 2012; Sinzelle et al., 2009). For example, the Pogo superfamily of DNA transposons and the Pipsqueak transcription factor family share the same HTH pipsqueak DNA-binding domain. Although the Pogo superfamily can be found in several metazoans and fungi species, the domestication of its DNA-binding domain has occurred only in the insect lineage. Particularly within insects, only *R. prolixus* and *A. pisum* show a significantly high number of pipsqueak domains ($p=2 \times 10^{-3}$ and $p=2 \times 10^{-10}$), possibly contributing to functional innovation in both species.

Our findings suggest that the LSEs of DNA-binding proteins (Pipsqueak, SAZ/MADF, FLYWCH, THAP and BED finger) in hemipterans probably emerged from proliferation of TEs bearing the respective DNA-binding domains. It is conceivable that at least a subset of them have been repurposed as new TFs, especially in the versions where the core catalytic domain is disrupted or eroded.

4. Discussion

The availability of diverse arthropod genomes within and outside the insect clade provides an unprecedented view of the molecular basis for their remarkable diversity in morphological form, developmental program, life style, life history and physiology. Our analysis suggests that variations in the transcriptional regulatory potential are likely contributors to this diversity. High OCs of most CPs and one-to-one orthology of certain TFs such as the nuclear hormone receptors, forkhead and homeo domains suggest the early fixation of a conserved regulatory scaffold that formed the backdrop of the transcription program upon which lineage-specific innovations occurred. Indeed, many of these include TFs and CPs that are involved in basic developmental patterning events, tissue specification and differentiation that appear to have been fixed in the stem bilaterian metazoans, or early in arthropod evolution. The occasional gene losses in this ancestral scaffold, though, point to a certain flexibility that in part might be compensated by the emergence of new TFs, or by displacements by alternate TFs (Shigenobu et al., 2010).

Our results clearly show that LSEs of various TFs are a major evolutionary force shaping the transcriptional regulatory landscape. While this has been noticed as a general trend across eukaryotes (Aravind et al., 2013; Aravind et al., 2009; Iyer et al., 2008; Romero et al., 2012), the availability of diverse representative genomes within arthropods allows us to precisely trace the origin and expansion events of TFs occurring at various stages in the evolution of insects from the ancestral metazoan. These include: 1) expansions of ancient TF families that occurred in the arthropod ancestor (e.g. C2H2 Zn-finger, bHLH, POZ/BTB); 2) derivations within the pancrustacean lineage followed by early or late expansions in different insect lineages (e.g. C4DM) and; 3) independent expansions in terminal insect lineages (e.g. Pipsqueak, SAZ/MADF, FLYWCH, THAP and BED finger). Studies on gene regulation in model metazoans suggest that while several developmental processes are conserved across a wide range of lineages, the TFs that regulate them are often derived in a lineage-specific. For example TFs, such as *ribbon*, *piefke*, *pipsqueak*, and *bric-a-brac1/2*, which contain a POZ/BTB fused to a HTH regulate several key developmental, and tissue

specific differentiation processes in *Drosophila* (Lours et al., 2003; Shim et al., 2001; Siegmund and Lehmann, 2002). Although these developmental processes are conserved across a wider group of metazoans, the fusion and subsequent expansion of POZ/MTB and HTH domains seems to have occurred only in the insect lineage. Similarly, the independent expansions of different TF families in terminal insect lineages as seen in our study (e.g. between the hemipterans) suggest a constant flux of the transcription regulatory regime, even between closely related lineages. Thus, the general trend observed across more divergent eukaryotes that transcriptional regulatory networks undergo massive reorganization and rewiring during the emergence of new lineages, seems to hold even for closely related insect lineages. This phenomenon might also have a significant role in the diversification of morphology and adaptations that are seen in the insects.

Our analysis shows that TEs have been a key driver of LSEs in arthropods, particularly in insects. As has been noticed previously several DNA-binding domains are closely related to versions present in transposases. For example, the THAP, FLYWCH and BED finger domains that are widespread across metazoan can clearly be traced to versions found in TEs (Aravind, 2000; Babu et al., 2006; Hirose et al., 1996; Roussigne et al., 2003; Zhang et al., 2004). That these domains are derived from transposons can be inferred from the several instances where domesticated transposons, which continue to carry the core catalytic domain function as TFs (Babu et al., 2006; Cowan et al., 2005; Lin et al., 2007). More specifically in our analysis, we show that within the hemipterans the expansion of the Pipsqueak, SAZ/MADF, FLYWCH, THAP and BED finger domains in *A. pisum* is apparently related to their linkage to transposase domains. Thus expansions of transposons appear to have provided the raw material in the form of both DNA-binding domains and the specific target DNA sequences for the DNA-binding domains to potentially allow for altered transcription regulation between different insect lineages.

Finally, although CPs are generally highly conserved and show a high OC values across insects, they too can show divergent patterns that might contribute to morphological or developmental variations caused by changes in the epigenetic landscape. DNA cytosine methylation patterns can show great variation in insect lineages, with some such as *Drosophila* only showing methylation in a small amount at particular stages of embryonic development (Boffelli et al., 2014); in contrast to honey bees and termites where differences in DNA methylation appears to be a key determinant in caste differentiation (Elango et al., 2009; Terrapon et al., 2014). Both hemipteran species possess the maintenance DNA cytosine methylase DNMT1, whereas the *de novo* methylase DNMT3 is thus far only detected in *A. pisum*. Both hemipterans also possess a single copy of the TET dioxygenase, which generates oxidized 5-methylcytosine from 5mC in DNA, which could serve both as epigenetic marks and intermediates for demethylation (Pastor et al., 2013). We also identify multiple expansions of predicted histone methylases with the SET domain in *A. pisum*, which is unprecedented in any arthropod studied to date. The expansion of the Eggless family of SET domains fused to the TAM/MBD domain, point to a prominent role for epigenetic regulation by DNA methylation in combination with histone methylation in *A. pisum*. Aphids like *A. pisum* are known to display cyclical parthenogenesis, polyphenism in females with winged forms, and polymorphism in males with winged and wingless forms.

The aphid life-cycle includes asexual and viviparous reproduction in spring and summer, and sexual differentiation with oviparous females in autumn (Duncan et al., 2013). It is possible that, just as in the determination of castes in honeybees and termites, changes in the DNA methylation patterns and consequently the epigenetic landscape might play a role in the expression of multiple phenotypes using the same genetic material in this lineage of hemipterans. In contrast, *R. prolixus* appears to represent the ancestral hemipteran condition with no drastic morphological variation through the life cycle beyond the basic hemimetabolous transformation.

The availability of the *R. prolixus* genome allows us to perform a more general comparison of the transcriptional and chromatin landscape of hemipterans, while providing a careful catalog of all TFs and CPs in this lineage. Investigation of TF and CP through molecular biology techniques (e.g. ChIP-Seq, mutagenesis and tandem-affinity purification followed by mass spectrometry) is likely to provide new vistas in understanding gene regulation not just in *R. prolixus*, but also in other arthropods like *A. pisum*, which might have a bearing on the evolution of gene regulation in insects and provide novel targets that might help control this disease vector.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

NMV is supported by the National Institutes of Health Intramural Research Program, National Library of Medicine, National Institutes of Health; and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). NMV was recipient of a post-doctoral fellowship from Instituto Nacional de Ciência e Tecnologia em Entomologia Molecular (INCT-EM). LMI and LA work is supported by the intramural funds of the National Library of Medicine, National Institutes of Health. TMV is supported by INCT-EM (CNPq) and FAPERJ. TMV is a recipient of an established investigator fellowship award from CNPq.

Abbreviations

bZIP	basic leucine zipper
bHLH	basic helix-loop-helix
TF	transcription factor
CP	chromatin protein
LSE	lineage-specific expansion
HMM	hidden Markov model
OC	orthology coefficient
HTH	helix-turn-helix
DNMT1	DNA cytosine methylase 1
DNMT3	DNA cytosine methylase 3

References

- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped blast and psi-blast: A new generation of protein database search programs. *Nucleic Acids Res.* 1997; 25:3389–3402. [PubMed: 9254694]
- Anantharaman V, Iyer LM, Aravind L. Comparative genomics of protists: New insights into the evolution of eukaryotic signal transduction and gene regulation. *Annu Rev Microbiol.* 2007; 61:453–475. [PubMed: 17506670]
- Aravind L. The bed finger, a novel DNA-binding domain in chromatin-boundary-element-binding proteins and transposases. *Trends Biochem Sci.* 2000; 25:421–423. [PubMed: 10973053]
- Aravind L, Abhiman S, Iyer LM. Natural history of the eukaryotic chromatin protein methylation system. *Prog Mol Biol Transl Sci.* 2011; 101:105–176. [PubMed: 21507350]
- Aravind, L.; Anantharaman, V.; Abhiman, S.; Iyer, LM. Evolution of eukaryotic chromatin proteins and transcription factors, Protein families. John Wiley & Sons, Inc.; 2013. p. 421-502.
- Aravind L, Anantharaman V, Venancio TM. Apprehending multicellularity: Regulatory networks, genomics, and evolution. *Birth Defects Res C Embryo Today.* 2009; 87:143–164. [PubMed: 19530132]
- Aravind L, Koonin EV. Fold prediction and evolutionary analysis of the poz domain: Structural and evolutionary relationship with the potassium channel tetramerization domain. *J Mol Biol.* 1999; 285:1353–1361. [PubMed: 9917379]
- Babu MM, Iyer LM, Balaji S, Aravind L. The natural history of the wrky-gcm1 zinc fingers and the relationship between transcription factors and transposons. *Nucleic Acids Res.* 2006; 34:6505–6520. [PubMed: 17130173]
- Balaji S, Babu MM, Iyer LM, Aravind L. Discovery of the principal specific transcription factors of apicomplexa and their implication for the evolution of the ap2-integrase DNA binding domains. *Nucleic Acids Res.* 2005; 33:3994–4006. [PubMed: 16040597]
- Bardwell VJ, Treisman R. The poz domain: A conserved protein-protein interaction motif. *Genes Dev.* 1994; 8:1664–1677. [PubMed: 7958847]
- Boffelli D, Takayama S, Martin DI. Now you see it: Genome methylation makes a comeback in drosophila. *Bioessays.* 2014
- Clark KL, Halay ED, Lai E, Burley SK. Co-crystal structure of the hnf-3/fork head DNA-recognition motif resembles histone h5. *Nature.* 1993; 364:412–420. [PubMed: 8332212]
- Clough E, Moon W, Wang S, Smith K, Hazelrigg T. Histone methylation is required for oogenesis in drosophila. *Development.* 2007; 134:157–165. [PubMed: 17164421]
- Clough E, Tedeschi T, Hazelrigg T. Epigenetic regulation of oogenesis and germ stem cell maintenance by the drosophila histone methyltransferase eggless/dsetdb1. *Dev Biol.* 2014; 388:181–191. [PubMed: 24485852]
- Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK, Bauer DJ, Caceres CE, Carmel L, Casola C, Choi JH, Detter JC, Dong Q, Dusheyko S, Eads BD, Frohlich T, Geiler-Samerotte KA, Gerlach D, Hatcher P, Jogdeo S, Krijgsvelde J, Kriventseva EV, Kultz D, Laforsch C, Lindquist E, Lopez J, Manak JR, Muller J, Pangilinan J, Patwardhan RP, Pitluck S, Pritham EJ, Rechtsteiner A, Rho M, Rogozin IB, Sakarya O, Salamov A, Schaack S, Shapiro H, Shiga Y, Skalitzy C, Smith Z, Suvorov A, Sung W, Tang Z, Tsuchiya D, Tu H, Vos H, Wang M, Wolf YI, Yamagata H, Yamada T, Ye Y, Shaw JR, Andrews J, Crease TJ, Tang H, Lucas SM, Robertson HM, Bork P, Koonin EV, Zdobnov EM, Grigoriev IV, Lynch M, Boore JL. The ecoresponsive genome of daphnia pulex. *Science.* 2011; 331:555–561. [PubMed: 21292972]
- Collins T, Stone JR, Williams AJ. All in the family: The btb/poz, krab, and scan domains. *Mol Cell Biol.* 2001; 21:3609–3615. [PubMed: 11340155]
- Cowan RK, Hoen DR, Schoen DJ, Bureau TE. Mustang is a novel family of domesticated transposase genes found in diverse angiosperms. *Mol Biol Evol.* 2005; 22:2084–2089. [PubMed: 15987878]
- De Marco A, Hochschwender SM, Laursen RA, Llinas M. Human plasminogen. Proton nmr studies on kringle 1. *J Biol Chem.* 1982; 257:12716–12721. [PubMed: 7130173]

- Duncan EJ, Leask MP, Dearden PK. The pea aphid (*acyrthosiphon pisum*) genome encodes two divergent early developmental programs. *Dev Biol.* 2013; 377:262–274. [PubMed: 23416037]
- Eddy SR. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* 2009; 23:205–211. [PubMed: 20180275]
- Edgar RC. Muscle: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics.* 2004; 5:113. [PubMed: 15318951]
- Elango N, Hunt BG, Goodisman MA, Yi SV. DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, *apis mellifera*. *Proc Natl Acad Sci U S A.* 2009; 106:11206–11211. [PubMed: 19556545]
- Favier B, Dolle P. Developmental functions of mammalian hox genes. *Mol Hum Reprod.* 1997; 3:115–131. [PubMed: 9239717]
- Filion GJ, van Bommel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ, van Steensel B. Systematic protein location mapping reveals five principal chromatin types in *drosophila* cells. *Cell.* 2010; 143:212–224. [PubMed: 20888037]
- Finn RD, Bateman A, Clements J, Coghill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M. Pfam: The protein families database. *Nucleic Acids Res.* 2014; 42:D222–230. [PubMed: 24288371]
- Fischer S, Brunk BP, Chen F, Gao X, Harb OS, Iodice JB, Shanmugam D, Roos DS, Stoeckert CJ Jr. Using orthomcl to assign proteins to orthomcl-db groups or to cluster proteomes into new ortholog groups. *Curr Protoc Bioinformatics.* 2011; 12:11–19. Chapter 6, Unit 6.
- Footitt, RG.; Adler, PH. *Insect biodiversity: Science and society.* First. Wiley-Blackwell; West Sussex, UK: 2009.
- Gehring WJ. The homeobox in perspective. *Trends Biochem Sci.* 1992; 17:277–280. [PubMed: 1357790]
- Gehring WJ. The master control gene for morphogenesis and evolution of the eye. *Genes Cells.* 1996; 1:11–15. [PubMed: 9078363]
- Giangrande A. Development and organization of glial cells in *drosophila melanogaster*. *Int J Dev Biol.* 1996; 40:917–927. [PubMed: 8946240]
- Gross CT, McGinnis W. Deaf-1, a novel protein that binds an essential region in a deformed response element. *EMBO J.* 1996; 15:1961–1970. [PubMed: 8617243]
- Heffer A, Pick L. Conservation and variation in hox genes: How insect models pioneered the evo-devo field. *Annu Rev Entomol.* 2013; 58:161–179. [PubMed: 23317041]
- Herz HM, Garruss A, Shilatifard A. Set for life: Biochemical activities and biological functions of set domain-containing proteins. *Trends Biochem Sci.* 2013; 38:621–639. [PubMed: 24148750]
- Hirose F, Yamaguchi M, Kuroda K, Omori A, Hachiya T, Ikeda M, Nishimoto Y, Matsukage A. Isolation and characterization of cDNA for dref, a promoter-activating factor for *drosophila* DNA replication-related genes. *J Biol Chem.* 1996; 271:3930–3937. [PubMed: 8632015]
- Hoen DR, Park KC, Elrouby N, Yu Z, Mohabir N, Cowan RK, Bureau TE. Transposon-mediated expansion and diversification of a family of ulp-like genes. *Mol Biol Evol.* 2006; 23:1254–1268. [PubMed: 16581939]
- Holland PW. Evolution of homeobox genes. *Wiley Interdiscip Rev Dev Biol.* 2013; 2:31–45. [PubMed: 23799629]
- Iyer LM, Anantharaman V, Wolf MY, Aravind L. Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes. *Int J Parasitol.* 2008; 38:1–31. [PubMed: 17949725]
- Iyer LM, Aravind L. Alog domains: Provenance of plant homeotic and developmental regulators from the DNA-binding domain of a novel class of drosophila-type retrotransposons. *Biol Direct.* 2012; 7:39. [PubMed: 23146749]
- Jauch R, Bourenkov GP, Chung HR, Urlaub H, Reidt U, Jackle H, Wahl MC. The zinc finger-associated domain of the *drosophila* transcription factor grauzone is a novel zinc-coordinating protein-protein interaction module. *Structure.* 2003; 11:1393–1402. [PubMed: 14604529]
- Jiang L, Mu J, Zhang Q, Ni T, Srinivasan P, Rayavara K, Yang W, Turner L, Lavstsen T, Theander TG, Peng W, Wei G, Jing Q, Wakabayashi Y, Bansal A, Luo Y, Ribeiro JM, Scherf A, Aravind L,

- Zhu J, Zhao K, Miller LH. Pfsetvs methylation of histone h3k36 represses virulence genes in plasmodium falciparum. *Nature*. 2013; 499:223–227. [PubMed: 23823717]
- Jones S. An overview of the basic helix-loop-helix proteins. *Genome Biol*. 2004; 5:226. [PubMed: 15186484]
- Juretic N, Hoen DR, Huynh ML, Harrison PM, Bureau TE. The evolutionary fate of mule-mediated duplications of host gene fragments in rice. *Genome Res*. 2005; 15:1292–1297. [PubMed: 16140995]
- Kaessmann H. Origins, evolution, and phenotypic impact of new genes. *Genome Res*. 2010; 20:1313–1326. [PubMed: 20651121]
- Kapitonov VV, Jurka J. Harbinger transposons and an ancient harbi1 gene derived from a transposase. *DNA Cell Biol*. 2004; 23:311–324. [PubMed: 15169610]
- Katoh K, Misawa K, Kuma K, Miyata T. Mafft: A novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res*. 2002; 30:3059–3066. [PubMed: 12136088]
- Kogan GL, Usakin LA, Ryazansky SS, Gvozdev VA. Expansion and evolution of the x-linked testis specific multigene families in the melanogaster species subgroup. *PLoS One*. 2012; 7:e37738. [PubMed: 22649555]
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chisoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, International Human Genome Sequencing, C. Initial sequencing and analysis of the human genome. *Nature*. 2001; 409:860–921. [PubMed: 11237011]
- Laurence, PA. The making of a fly: The genetics of animal design. First. Wiley-Blackwell; 1992.
- Lehmann M, Siegmund T, Lintermann KG, Korge G. The pipsqueak protein of drosophila melanogaster binds to gaga sequences through a novel DNA-binding domain. *J Biol Chem*. 1998; 273:28504–28509. [PubMed: 9774480]
- Lespinet O, Wolf YI, Koonin EV, Aravind L. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res*. 2002; 12:1048–1059. [PubMed: 12097341]
- Lewis EB. A gene complex controlling segmentation in drosophila. *Nature*. 1978; 276:565–570. [PubMed: 103000]
- Lhoumaud P, Hennion M, Gamot A, Cuddapah S, Queille S, Liang J, Micas G, Morillon P, Urbach S, Bouchez O, Severac D, Emberly E, Zhao K, Cuvier O. Insulators recruit histone methyltransferase

- dmes4 to regulate chromatin of flanking genes. *EMBO J.* 2014; 33:1599–1613. [PubMed: 24916307]
- Lin R, Ding L, Casola C, Ripoll DR, Feschotte C, Wang H. Transposase-derived transcription factors regulate light signaling in arabidopsis. *Science.* 2007; 318:1302–1305. [PubMed: 18033885]
- Liu Y, Chen W, Gaudet J, Cheney MD, Roudaia L, Cierpicki T, Klet RC, Hartman K, Laue TM, Speck NA, Bushweller JH. Structural basis for recognition of smrt/n-cor by the mynd domain and its contribution to aml1/eto's activity. *Cancer Cell.* 2007; 11:483–497. [PubMed: 17560331]
- Lours C, Bardot O, Godt D, Laski FA, Couderc JL. The drosophila melanogaster btb proteins bric a brac bind DNA through a composite DNA binding domain containing a pipsqueak and an at-hook motif. *Nucleic Acids Res.* 2003; 31:5389–5398. [PubMed: 12954775]
- Mann RS, Chan SK. Extra specificity from extradenticle: The partnership between hox and pbx/exd homeodomain proteins. *Trends Genet.* 1996; 12:258–262. [PubMed: 8763497]
- Marino-Ramirez L, Lewis KC, Landsman D, Jordan IK. Transposable elements donate lineage-specific regulatory sequences to host genomes. *Cytogenet Genome Res.* 2005; 110:333–341. [PubMed: 16093685]
- Marzo M, Liu D, Ruiz A, Chalmers R. Identification of multiple binding sites for the thap domain of the galileo transposase in the long terminal inverted-repeats. *Gene.* 2013; 525:84–91. [PubMed: 23648487]
- McGinnis W, Krumlauf R. Homeobox genes and axial patterning. *Cell.* 1992; 68:283–302. [PubMed: 1346368]
- Mishra K, Chopra VS, Srinivasan A, Mishra RK. Trl-gaga directly interacts with lola like and both are part of the repressive complex of polycomb group of genes. *Mech Dev.* 2003; 120:681–689. [PubMed: 12834867]
- Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, Niehuis O, Petersen M, Izquierdo-Carrasco F, Wappler T, Rust J, Aberer AJ, Aspöck U, Aspöck H, Bartel D, Blanke A, Berger S, Böhm A, Buckley TR, Calcott B, Chen J, Friedrich F, Fukui M, Fujita M, Greve C, Grobe P, Gu S, Huang Y, Jermiin LS, Kawahara AY, Krogmann L, Kubiak M, Lanfear R, Letsch H, Li Y, Li Z, Li J, Lu H, Machida R, Mashimo Y, Kapli P, McKenna DD, Meng G, Nakagaki Y, Navarrete-Heredia JL, Ott M, Ou Y, Pass G, Podsiadlowski L, Pohl H, von Reumont BM, Schütte K, Sekiya K, Shimizu S, Slipinski A, Stamatakis A, Song W, Su X, Szucsich NU, Tan M, Tan X, Tang M, Tang J, Timelthaler G, Tomizuka S, Trautwein M, Tong X, Uchifune T, Walz MG, Wiegmann BM, Wilbrandt J, Wipfler B, Wong TK, Wu Q, Wu G, Xie Y, Yang S, Yang Q, Yeates DK, Yoshizawa K, Zhang Q, Zhang R, Zhang W, Zhang Y, Zhao J, Zhou C, Zhou L, Ziesmann T, Zou S, Li Y, Xu X, Zhang Y, Yang H, Wang J, Wang J, Kjer KM, Zhou X. Phylogenomics resolves the timing and pattern of insect evolution. *Science.* 2014; 346:763–767. [PubMed: 25378627]
- Mohd-Sarip A, Venturini F, Chalkley GE, Vrijzener CP. Pleiohomeotic can link polycomb to DNA and mediate transcriptional repression. *Mol Cell Biol.* 2002; 22:7473–7483. [PubMed: 12370294]
- Murre C, McCaw PS, Baltimore D. A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, myod, and myc proteins. *Cell.* 1989; 56:777–783. [PubMed: 2493990]
- Myatt SS, Lam EW. The emerging roles of forkhead box (fox) proteins in cancer. *Nat Rev Cancer.* 2007; 7:847–859. [PubMed: 17943136]
- Orsi GA, Kasinathan S, Hughes KT, Saminadin-Peter S, Henikoff S, Ahmad K. High-resolution mapping defines the cooperative architecture of polycomb response elements. *Genome Res.* 2014; 24:809–820. [PubMed: 24668908]
- Pastor WA, Aravind L, Rao A. Tetonic shift: Biological roles of tet proteins in DNA demethylation and transcription. *Nat Rev Mol Cell Biol.* 2013; 14:341–356. [PubMed: 23698584]
- Patel MV, Hallal DA, Jones JW, Bronner DN, Zein R, Caravas J, Husain Z, Friedrich M, Vanberkum MF. Dramatic expansion and developmental expression diversification of the methuselah gene family during recent drosophila evolution. *J Exp Zool B Mol Dev Evol.* 2012; 318:368–387. [PubMed: 22711569]
- Perez-Torrado R, Yamada D, Defosse PA. Born to bind: The btb protein-protein interaction domain. *Bioessays.* 2006; 28:1194–1202. [PubMed: 17120193]

- Price MN, Dehal PS, Arkin AP. Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol.* 2009; 26:1641–1650. [PubMed: 19377059]
- Rees JS, Lowe N, Armean IM, Roote J, Johnson G, Drummond E, Spriggs H, Ryder E, Russell S, St Johnston D, Lilley KS. In vivo analysis of proteomes and interactomes using parallel affinity capture (ipac) coupled to mass spectrometry. *Mol Cell Proteomics.* 2011; 10 M110 002386.
- Ringrose L, Paro R. Epigenetic regulation of cellular memory by the polycomb and trithorax group proteins. *Annu Rev Genet.* 2004; 38:413–443. [PubMed: 15568982]
- Romero IG, Ruvinsky I, Gilad Y. Comparative studies of gene expression and the evolution of gene regulation. *Nat Rev Genet.* 2012; 13:505–516. [PubMed: 22705669]
- Roussigne M, Kossida S, Lavigne AC, Clouaire T, Ecochard V, Glories A, Amalric F, Girard JP. The thap domain: A novel protein motif with similarity to the DNA-binding domain of p element transposase. *Trends Biochem Sci.* 2003; 28:66–69. [PubMed: 12575992]
- Schindler U, Beckmann H, Cashmore AR. Hat3.1, a novel arabidopsis homeodomain protein containing a conserved cysteine-rich region. *Plant J.* 1993; 4:137–150. [PubMed: 8106082]
- Scott MP, Weiner AJ. Structural relationships among genes that control development: Sequence homology between the antennapedia, ultrabithorax, and fushi tarazu loci of drosophila. *Proc Natl Acad Sci U S A.* 1984; 81:4115–4119. [PubMed: 6330741]
- Shigenobu S, Bickel RD, Brisson JA, Butts T, Chang CC, Christiaens O, Davis GK, Duncan EJ, Ferrier DE, Iga M, Janssen R, Lin GW, Lu HL, McGregor AP, Miura T, Smaghe G, Smith JM, van der Zee M, Velarde RA, Wilson MJ, Dearden PK, Stern DL. Comprehensive survey of developmental genes in the pea aphid, *acyrthosiphon pisum*: Frequent lineage-specific duplications and losses of developmental genes. *Insect Mol Biol.* 2010; 19(Suppl 2):47–62. [PubMed: 20482639]
- Shim K, Blake KJ, Jack J, Krasnow MA. The drosophila ribbon gene encodes a nuclear btb domain protein that promotes epithelial migration and morphogenesis. *Development.* 2001; 128:4923–4933. [PubMed: 11731471]
- Shukla V, Habib F, Kulkarni A, Ratnaparkhi GS. Gene duplication, lineage-specific expansion, and subfunctionalization in the madf-bess family patterns the drosophila wing hinge. *Genetics.* 2014; 196:481–496. [PubMed: 24336749]
- Siegmund T, Lehmann M. The drosophila pipsqueak protein defines a new family of helix-turn-helix DNA-binding proteins. *Dev Genes Evol.* 2002; 212:152–157. [PubMed: 11976954]
- Sinzelle L, Izsvak Z, Ivics Z. Molecular domestication of transposable elements: From detrimental parasites to useful host genes. *Cell Mol Life Sci.* 2009; 66:1073–1093. [PubMed: 19132291]
- Steffen PA, Ringrose L. What are memories made of? How polycomb and trithorax proteins mediate epigenetic memory. *Nat Rev Mol Cell Biol.* 2014; 15:340–356. [PubMed: 24755934]
- Stogios PJ, Prive GG. The back domain in btb-kelch proteins. *Trends Biochem Sci.* 2004; 29:634–637. [PubMed: 15544948]
- Subramanian G, Koonin EV, Aravind L. Comparative genome analysis of the pathogenic spirochetes *borrelia burgdorferi* and *treponema pallidum*. *Infect Immun.* 2000; 68:1633–1648. [PubMed: 10678983]
- Tanenbaum DM, Wang Y, Williams SP, Sigler PB. Crystallographic comparison of the estrogen and progesterone receptor's ligand binding domains. *Proc Natl Acad Sci U S A.* 1998; 95:5998–6003. [PubMed: 9600906]
- Terrapon N, Li C, Robertson HM, Ji L, Meng X, Booth W, Chen Z, Childers CP, Glastad KM, Gokhale K, Gowin J, Gronenberg W, Hermansen RA, Hu H, Hunt BG, Huylmans AK, Khalil SM, Mitchell RD, Munoz-Torres MC, Mustard JA, Pan H, Reese JT, Scharf ME, Sun F, Vogel H, Xiao J, Yang W, Yang Z, Yang Z, Zhou J, Zhu J, Brent CS, Elsik CG, Goodisman MA, Liberles DA, Roe RM, Vargo EL, Vilcinskas A, Wang J, Bornberg-Bauer E, Korb J, Zhang G, Liebig J. Molecular traces of alternative social organization in a termite genome. *Nat Commun.* 2014; 5:3636. [PubMed: 24845553]
- Tripoulas N, LaJeunesse D, Gildea J, Shearn A. The drosophila ash1 gene product, which is localized at specific sites on polytene chromosomes, contains a set domain and a phd finger. *Genetics.* 1996; 143:913–928. [PubMed: 8725238]

- Weigel D, Jurgens G, Kuttner F, Seifert E, Jackle H. The homeotic gene fork head encodes a nuclear protein and is expressed in the terminal regions of the drosophila embryo. *Cell*. 1989; 57:645–658. [PubMed: 2566386]
- Yang H, Pesavento JJ, Starnes TW, Cryderman DE, Wallrath LL, Kelleher NL, Mizzen CA. Preferential dimethylation of histone h4 lysine 20 by suv4-20. *J Biol Chem*. 2008; 283:12085–12092. [PubMed: 18296440]
- Zhang X, Jiang N, Feschotte C, Wessler SR. Pif- and pong-like transposable elements: Distribution, evolution and relationship with tourist-like miniature inverted-repeat transposable elements. *Genetics*. 2004; 166:971–986. [PubMed: 15020481]
- Zhong YF, Holland PW. Homeodb2: Functional expansion of a comparative homeobox gene database for evolutionary developmental biology. *Evol Dev*. 2011; 13:567–568. [PubMed: 23016940]
- Zollman S, Godt D, Prive GG, Couderc JL, Laski FA. The btb domain, found primarily in zinc finger proteins, defines an evolutionarily conserved family that includes several developmentally regulated genes in drosophila. *Proc Natl Acad Sci U S A*. 1994; 91:10717–10721. [PubMed: 7938017]

- Distinct families of transcription factors and chromatin proteins have undergone proliferation in the pea aphid *Acyrtosiphon pisum* and kissing bug *Rhodnius prolixus*.
- Of these lineage-specific expansions (LSEs) of transcription factors (TFs) SAZ/MADF, THAP, FLYWCH and BED finger-domain and multiple families of histone methylating SET domain proteins have occurred in *A. pisum* but not in *R. prolixus*.
- Pipsqueak domain TFs were independently expanded in both the hemipterans, *R. prolixus* and *A. pisum*.
- The five LSE families have genes coding for DNA-binding, dimerization and/or transposase domains, indicating that their expansion in hemipterans was associated with mobility of transposable elements.

Figure 1.

Orthology Coefficients of transcriptional regulators of *R. prolixus* and other arthropods. A) *R. prolixus* and *D. melanogaster*, B) *R. prolixus* and *A. mellifera*, C) *R. prolixus* and *A. pisum*, D) *R. prolixus* and *D. pulex*.

Legend: Proteome: non-redundant predicted proteomes; Transcription factors: all proteins with at least one of the transcription factor domains from the Pfam curated list (Table S2, “TF Pfam” tab); Chromatin proteins: all proteins with at least one of the chromatin protein domains from the Pfam curated list (Table S2, “CP Pfam” tab). POZ/BTB, SET, nuclear hormone receptor, forkhead, homeo, pipsqueak, SAZ/MADF: all proteins containing domain (Table S2, Pfam list).

Figure 2.

Phylogenetic reconstruction of forkhead domain proteins across 18 arthropod species.

Legend: Species names are abbreviated according to Table 1 and the color scheme is as following: Diptera (Dmel, Agam) in purple, Lepidoptera (Bmor, Hmel) in blue, Coleoptera (Tcas, Dpon) in gray, Hymenoptera (Amel, Cflo, Hsal, Nvit) in orange, Rpro in red, Apis in green, Phum in yellow, Znev in pink, and crustaceans (Dpul) and chelicerates (Mmar, Smim, Isca) in brown.

Figure 3. Comparison of transcription factor domains (A) and chromatin protein domains (B) between *R. prolixus*/*A. pisum* and *D. melanogaster*. Most prevalent domains are labeled.

Figure 4. Phylogenetic reconstruction of SET domain proteins across 18 arthropod species. Legend: Species names are abbreviated according to Table 1 and the color scheme is as following: Diptera (Dmel, Agam) in purple, Lepidoptera (Bmor, Hmel) in blue, Coleoptera (Tcas, Dpon) in gray, Hymenoptera (Amel, Cflo, Hsal, Nvit) in orange, Rpro in red, Apis in green, Phum in yellow, Znev in pink, and crustacens (Dpul) and chelicerates (Mmar, Smim, Isca) in brown.

Figure 5.

Phylogenetic reconstruction of pipsqueak domain proteins across 18 arthropod species.

Legend: Species names are abbreviated according to Table 1 and the color scheme is as following: Diptera (Dmel, Agam) in purple, Lepidoptera (Bmor, Hmel) in blue, Coleoptera (Tcas, Dpon) in gray, Hymenoptera (Amel, Cflo, Hsal, Nvit) in orange, Rpro in red, Apis in green, Phum in yellow, Znev in pink, and crustacean (Dpul) and chelicerates (Mmar, Smim, Isca) in brown. *D. melanogaster* transcription factors have the gene name indicated in the sequence identifier. Sequences containing endonuclease and/or other transposon domains are labeled in black and with '_Tnp' indicated in the sequence identifier.

Table 1

Number of genes in each transcription factor and chromatin protein domain families identified in 18 arthropods using 901 Pfam HMM-profiles.

Phylogenetic group		Diptera		Lepidoptera		Coleoptera		Hymenoptera				Hemiptera		Psocodoea	Isoptera	Crustacea		Chelicerata		
Category	Domain \ Species	Dmel	Agam	Bmor	Hmel	Tcas	Dpon	Amel	Cflo	Hsal	Nvit	Rpro	Apis	Phum	Znev	Dpul	Mmar	Smim	Isca	
Transcription Factors	zf-H2C2_2	204	215	257	187	180	181	142	103	29	106	142	333	137	298	113	141	597	221	
	zf-C2H2	190	217	242	164	161	161	179	130	68	115	146	338	135	288	102	132	479	153	
	C4DM/ZAD	95	110	105	87	60	47	32	28	15	24	12	28	25	62	2	0	0	0	
	bHLH	58	48	48	40	47	44	53	47	51	41	40	54	50	50	50	66	50	45	
	Nuclear Hormone Receptor	20	21	18	15	19	19	20	19	18	17	19	16	18	17	22	25	19	19	
	Forkhead	19	21	17	18	17	18	16	16	17	16	16	16	18	17	14	17	29	17	
	Homeo	96	86	90	81	86	83	91	77	76	80	66	90	89	70	104	138	75	91	
	Pipsqueak	9	10	11	10	8	7	8	9	9	11	13	36	84	7	11	6	29	3	
	SAZ/MADF	49	19	40	30	13	11	15	26	21	11	33	178	0	26	8	43	20	14	
	THAP	10	16	24	13	28	13	9	35	34	26	3	430	2	13	48	34	52	49	
Chromatin Proteins	FLYWCH	12	21	10	13	13	5	2	1	1	1	5	166	10	3	33	0	6	2	
	BED finger	11	2	8	7	5	2	5	9	4	3	1	103	11	2	13	3	5	2	
	POZ/BTB	75	103	66	52	56	62	62	74	52	109	67	217	64	114	211	60	113	44	
	PHD	37	43	37	30	39	29	36	32	28	34	37	111	32	39	61	28	46	34	
	BACK	27	24	19	19	22	21	21	22	17	33	27	206	22	61	28	57	56	30	
	SET	28	36	23	21	20	28	28	24	20	38	20	65	25	27	32	11	20	15	