MSC IN BIOINFORMATICS: MASTER THESIS

EVOLUTIONARY PATTERNS OF piRNA-GENERATING CLUSTERS IN HUMAN GENOME

OLGA DOLGOVA

ACADEMIC TUTOR:
Sònia Casillas Viladerrams

PROJECT SUPERVISER:
Tanya Vavouri

Bellaterra, 2017

UAB Universitat Autònoma de Barcelona

IMPPC Institut de Medicina Predictiva i Personalitzada del Càncer

CONTENTS

SUMMARY

PIWI-interacting RNA (piRNA) is the largest class of small non-coding RNA molecules expressed in animal cells. They are 26-31 nucleotides long and highly expressed in the germline of animals from worms to humans. Their best understood function is genome defense against transposable elements by both epigenetic and post-transcriptional gene silencing of retrotransposons and other genetic elements in germ line cells, particularly those in spermatogenesis (IWASAKI *et al.* 2015). Most piRNA-generating loci are found in a small number of genomic regions referred to as clusters from which long primary transcripts are transcribed and individual piRNAs processed. Because they are among the most recently discovered classes of small non-coding RNAs, many aspects of their biology and evolution remain to be studied.

Our primary objective in this study was to revisit previously published results, extending them to different populations and chromosome datasets from publicly available databases, and elucidate which selective forces are mainly acting on piRNA cluster regions in the human genome at the nucleotide level.

1000 Genomes Project (1000GP) in its GRCh37 version was chosen for this study as a main dataset, from which piRNA clusters and the nearest intergenic regions, as putatively neutral for all genetic comparisons, were extracted following a newly developed pipeline. Polymorphism, mutation rate and divergence from the chimpanzee genome were calculated for both piRNA clusters and intergenic regions in each of five super-populations, and tests for neutrality and selection were conducted using R and the PopGenome package.

A strong difference in the levels of polymorphism was found between piRNA clusters and intergenic regions, the latter being more diverse. Besides, some differences were found in the mutation rates and divergence from chimpanzee, with higher values for intergenic regions. Comparing the observed heterozygosity (Nei's π) with that expected under neutral conditions (Θ of Watterson), we found the latter to be much higher, suggesting the action of purifying selection on weakly deleterious alleles segregating in all populations at low frequencies. This observation was confirmed by Tajima's D and Fu&Li's D tests, which in most of the cases attained negative values due to an excess of rare alleles. The values of alpha in the McDonald and Kreitman test were negative in 53-60% of the cases, being significant in 12-20% of piRNA clusters depending on the super-population. Nevertheless, in 40-47% of significant results, alpha was positive, suggesting recurrent directional selection fixing new advantageous alleles in some clusters.

Altogether, our results are in accordance with the previous findings (LUKIC and CHEN 2011; GOULD *et al.* 2012) that purifying selection is supposed to be the main force driving the evolution of piRNA clusters at the nucleotide level in humans. This suggests that, although piRNA clusters are not well conserved between species, they might be under detectable selective constraint at a shorter time scale. Nevertheless, directional selection also takes place in a substantial part of piRNA clusters, allowing the occurrence of an evolutionary "arm race" between these functional genomic regions and transposable elements.

1 | INTRODUCTION

1.1. PIWI – INTERACTING RNA

Our genome encodes thousands of genes responsible for various cellular functions, and the regulation of the expression levels and patterns of these genes is crucial for development and homeostasis. This regulation is performed by a collection of intramolecular and intermolecular events. The discovery of small non-coding RNAs, including microRNAs (miRNAs) and short interfering RNAs (siRNAs) (HAMILTON and BAULCOMBE 1999; LEE *et al.* 1993; REINHART *et al.* 2000) revolutionised our understanding of how gene expression is regulated. These non-coding RNAs are not translated into proteins but instead act through complementary base pairing with target RNAs. RNA silencing, also referred to as RNA interference (RNAi), has emerged as one of the key gene regulatory pathways in most eukaryotes (SIOMI and SIOMI 2009; GHILDIYAL and ZAMORE 2009). Central to RNA silencing pathways is the generation of small RNAs of 20 to 31 nucleotides (nt) that form an **RNA-induced silencing complex (RISC)** with Argonaute proteins and recognizes their targets via Watson–Crick base pairing. The Slicer endonuclease activity of Argonaute proteins cleaves target transcripts to accomplish gene silencing; posttranscriptional gene silencing, such as translational repression, or transcriptional gene silencing via specific chromatin modifications is performed through recruitment of other proteins (KIM *et al.* 2009).

Argonaute proteins can be phylogenetically separated into two clades based on sequence similarity: the **Ago** clade and the **Piwi** (P-element induced wimpy testis) clade (CARMELL 2002; PETERS and MEISTER 2007). AGO subfamily proteins are ubiquitously expressed and can bind to miRNAs and siRNAs, both of which are processed from double-stranded precursors into mature small RNAs of 20 to 22 nt in length in a Dicer-dependent manner (**Figure 1**; KIM *et al.* 2009). PIWI subfamily proteins (PIWI proteins) are expressed mainly in germline cells and they were first identified in a screen for factors involved in germline stem cell (GSC) maintenance in *Drosophila melanogaster* (CARMELL 2002; LIN and SPRADLING 1997), a finding that was soon expanded to GSCs in other organisms (COX *et al.* 1998). Although hints of the piRNA system were observed in *Drosophila* as early as 2001 (ARAVIN *et al.* 2001), several groups independently reported the identification of a small RNA population called PIWI-interacting RNAs from mouse and rat germ cells by immunoprecipitating PIWI protein in 2006 (ARAVIN *et al.* 2006; GIRARD *et al.* 2006; GRIVNA *et al.* 2006; LAU *et al.* 2006; WATANABE *et al.* 2006). These piRNAs have emerged as an extremely complex population of small RNAs that are highly enriched in the germline tissues of the majority of metazoans analysed to date. PIWI proteins form specific RISCs with piRNAs; these RISCs are termed **piRISCs** (reviewed in SIOMI *et al.* 2011, WEICK and MIKSA 2014; IWASAKI *et al.* 2015).

piRNAs are slightly longer (24–31 nt) than miRNAs and siRNAs. piRNAs have essentially no known defining sequence characteristics beyond a very strong propensity for a 5'-uridine and a weaker bias toward an adenosine at position 10 (**Figure 2**). It also possess 2′-*O*-methyl modification sites at the 3′ terminus, and are processed from single-stranded precursor transcripts expressed from intergenic regions termed **piRNA clusters** via a Dicer-independent mechanism in mammals and *Drosophila* (VAGIN *et al.* 2006; SIOMI *et al.* 2011). piRNAs are in general difficult to predict bioinformatically and must instead be defined biochemically. *Caenorhabditis elegans* piRNAs may be significantly different from mammalian and *Drosophila* piRNAs because they have a different length (21 nt), and there appears to be a conserved promoter motif upstream of many piRNAs (RUBY *et al.* 2006), suggesting that each piRNA is a separate transcription unit, unlike piRNAs in mammals and *Drosophila*, which are typically expressed in long polycistronic transcripts.

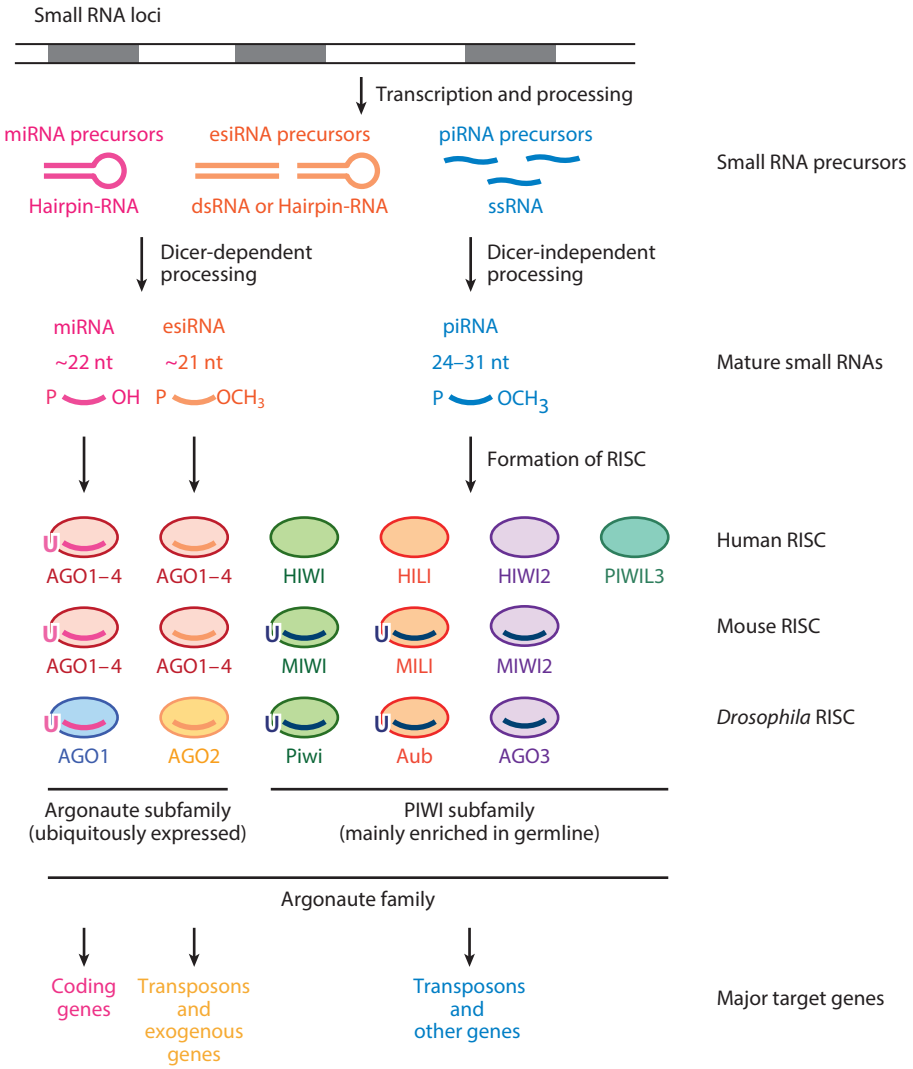


Figure 1: RNA silencing by small RNAs and their partner Argonaute family proteins in human, mouse, and *Drosophila*. Expression of a fourth PIWI protein (PIWIL3) has been detected specifically in humans. Characteristics of piRNA precursors, mature sequences, RISC formation, and target genes are summarized for miRNAs, esiRNAs, and piRNAs. An association between piRNAs and human PIWI proteins has not yet been identified. Abbreviations: dsRNA, double-stranded RNA; esiRNA, endogenous small interfering RNA (siRNA); miRNA, microRNA; nt, nucleotide; piRNA, PIWI-interacting RNA; RISC, RNA-induced silencing complex; ssRNA, single-stranded RNA (from IWASAKI *ET AL.* 2015).

Figure 2: Proposed piRNA structure (Wikipedia: https://en.wikipedia.org/wiki/Piwi-interacting_RNA).

piRNA populations are very complex and piRNAs appear to be produced by quasi-random cleavage of the primary piRNA transcript (BETEL *et al.* 2007). That is, while piRNAs almost always start with a U and there are biases for particular sequences to be cleaved as piRNAs, there is a strong random component that determines which sequences of the primary transcript are processed into piRNAs (hence the term 'quasi- random'). piRNA 3'-end

formation is poorly understood and is an object of active research (KAWAOKA *et al.* 2011). However, piRNA 5'-end formation has been addressed by several key papers (BRENNECKE *et al.* 2007; GUNAWARDANE *et al.* 2007). The authors studied master loci that control transposable element proliferation in *Drosophila* but were molecularly uncharacterized for many years because of the apparent lack of functional sequences at the loci, other than a jumble of transposable element insertions. These master loci were found to produce piRNAs that repress transposable elements in *trans* (BRENNECKE *et al.* 2007; reviewed in COLINE *et al.* 2014). The authors proposed the **"ping-pong" mechanism** (BRENNECKE *ET AL.* 2007; GUNAWARDANE *et al.* 2007; **Figure 3**) in which primary piRNAs cleave sense transposon transcripts and simultaneously produce secondary piRNAs from the sense transposons that then cleave antisense transposon transcripts. Primary piRNAs have a bias toward having uridine (U) at their 5′ nucleic acid (1U bias), whereas secondary piRNAs show 10-nt complementarity with primary piRNAs at their 5′ ends and possess a sense bias with adenosine at the tenth nucleotide (10A bias) (VAGIN *et al.* 2006; ARAVIN *et al.* 2006; ARAVIN *et al.* 2008; GIRARD *et al.* 2006; BRENNECKE *et al.* 2007; GUNAWARDANE *et al.* 2007). This mechanism thus depends on the transcription of both sense and anti-sense transposon transcripts.

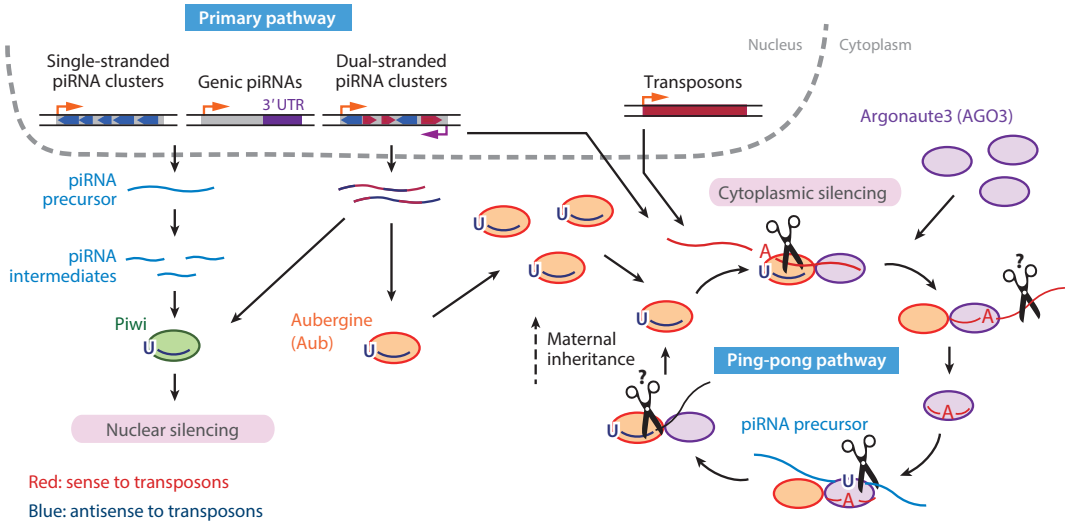


Figure 3: Biogenesis pathway of *Drosophila* piRNAs, consisting of the primary and ping-pong pathways. In the primary pathway, piRNAs are transcribed from genomic regions called piRNA clusters, processed, and loaded onto Piwi or Aub. 3′-UTR sequences of some protein-coding genes can also serve as a source of primary piRNAs. Silencing takes place both in the cytoplasm and nucleus. Piwi performs transcriptional gene silencing in the nucleus. Together with AGO3, the Aub–piRNA complex serves as a trigger to start the ping-pong amplification pathway. The ping-pong pathway silences the target transposon sequence and amplifies the piRNA sequence at the same time. Note that some Aub–piRNA complexes are also maternally inherited. Abbreviations: piRNA, PIWI-interacting RNA; UTR, untranslated region (from IWASAKI *et al.* 2015).

piRNA clusters harbor a large number of and various types of transposons; therefore, piRNAs regulate mainly the activity of transposons, namely their expression and transposition within the genome. Such mobile elements are autonomous pieces of DNA that replicate and insert into the genome and thus have the potential to introduce detrimental DNA damage. Because transposition of transposons has a high risk of damaging the genome intracellularly, the piRNA-mediated regulation of transposons is essential, especially for preserving normal gametogenesis and reproduction. Owing to their processing mechanisms and the variety of source transposons, piRNA sequences are much more diverse than those of any other known class of cellular RNAs and constitute the largest class of noncoding RNAs (SIOMI *et al.* 2011; MOAZED 2009). Transposon regulation by piRNAs is conceptually similar to that in immune

systems, which can achieve "self " and "nonself " recognition. As with our immune systems, piRNAs use a complex mechanism to effectively select and regulate the nonself genes for regulation (MALONE and HANNON 2009). The regulation of mobile sequences by piRNAs canonically involves endonucleolytic cleavage ('slicing') of the target sequence after complementary base-pair recognition through the piRISC. In the germline, this process prevents the accumulation of changes in the genome of the next generation and represents the most thoroughly understood aspect of piRNA biology. The piRNA system is thus a clear example of a Lamarckian mechanism in which environmental factors directly cause heritable genetic changes (KOONIN and WOLF 2009).

Recent studies have started to uncover the hitherto unknown mechanisms of piRNA biogenesis (**Figure 3**). They revealed a large number of cytoplasmic factors that support and maintain piRNA biogenesis, as well as some nuclear factors that either recognize and transcribe piRNA clusters to produce piRNA precursors, or function in piRNA-mediated transcriptional silencing. Additionally, analyses of various eukaryotes have identified piRNAs that target protein-coding genes and piRNAs that are passed through generations to transmit a memory of past transposon activity (ISHIZU *et al.* 2012; ROSS *et al.* 2014; STUWE *et al.* 2014).

Maintaining genome integrity by PIWI proteins and the piRNA system. The prototype of PIWI proteins is encoded by the *Drosophila piwi* (P-element-induced wimpy testes) gene, which was originally identified as an essential gene for germline development (THOMSON and LIN 2009; COX *et al.* 1998). In *Drosophila*, which contains three distinct *PIWI* genes [*ago3*, *aubergine* (*aub*), and *piwi*], *piwi* and *aub* are required for both male and female fertility, whereas *ago3* is essential for female fertility (COX *et al.* 1998; LI *et al.* 2009; SCHMIDT *et al.* 1999; LIN and SPRADLING 1997). Depression of transposons is observed in each of these *PIWI* mutant ovaries, indicating that all three PIWI proteins have non-redundant roles in gonad development and transposon silencing. *Aub* and *ago3* cleave their target transposon transcripts in the cytoplasm, whereas *piwi* can regulate its target transposons at the transcriptional level in the nucleus (LI *et al.* 2009, KALMYKOVA *et al.* 2005; VAGIN *et al.* 2004; SABIN *et al.* 2013; BRENNECKE *et al.* 2007; GUNAWARDANE *et al.* 2007). Interestingly, the *Drosophila* piRNA pathway also regulates transposon activity to maintain the telomeres. Unlike most other eukaryotes, the transposition of a distinct set of transposons to the chromosomal ends maintains the telomeres of *Drosophila* chromosomes, whereas defects of the piRNA pathway in gonads reduce expression of telomere-specific piRNAs and disrupt assembly of the telomere protection complex. Meanwhile, piRNA pathway defects do not affect transposon expression or telomere structure in somatic tissues (SAVITSKY *et al.* 2006; KHURANA *et al.* 2010; PARDUE and DEBARYSHE 2003).

Mice also express three PIWIs (MIWI, MIWI2, and MILI). All three PIWI proteins are expressed at different stages during spermatogenesis (**Figure 4**), but only MILI is expressed, albeit weakly, in female germ cells. Mutations in mouse *PIWI* genes affect the male germline but not the female germline (CARMELL *et al.* 2007; ARAVIN *et al.* 2007b; KURAMOCHI-MIYAGAWA *et al.* 2008; ARAVIN *et al.* 2008). Deficiency in MILI or MIWI2 leads to the activation of long interspersed nuclear element and long terminal repeat (LTR) retrotransposons including L1 and IAP elements, and spermatogenic stem cell arrest is observed. During MIWI depletion, the L1 transposon is also dysregulated, and spermatogenesis is arrested at the early spermatid stage (CARMELL *et al.* 2007; KURAMOCHI-MIYAGAWA *et al.* 2008; ARAVIN *et al.* 2006; GIRARD *et al.* 2006; GRIVNA *et al.* 2006).

Mouse PIWI proteins are bound to piRNAs expressed in two phases: **prepachytene piRNAs** and **pachytene piRNAs**. Prepachytene piRNAs are derived mostly from transposable elements and are associated with MILI and MIWI2 in the gonocyte stage, whereas pachytene piRNAs originate from piRNA clusters located in various regions of the genome and bind to both MILI and MIWI in pachytene spermatocytes to the round spermatid stage.

Although a fraction of pachytene piRNAs originates from transposons, the largest fraction comprises those originating from an unannotated region (ARAVIN *et al.* 2007b; KURAMOCHI-MIYAGAWA *et al.* 2008; ARAVIN *et al.* 2008; GIRARD *et al.* 2006; LAU *et al.* 2006). MIWI and MILI are necessary to maintain the Slicer-dependent silencing of the L1 transposon in the mouse testis after birth, indicating that Slicer activity directly cleaves transposon messenger RNAs (mRNAs) (**Figure 4**; REUTER *et al.* 2011; DE FAZIO *et al.* 2011). Moreover, mouse PIWI proteins function not only in posttranscriptional gene silencing by cleaving transposon transcripts, but also in transcriptional silencing by directing CpG DNA methylation on transposon loci. Indeed, silencing and *de novo* DNA methylation of L1 and IAP elements are decreased in male germlines that are defective for the activity of the *MILI* or *MIWI2* gene (ARAVIN *et al.* 2007b; KURAMOCHI-MIYAGAWA *et al.* 2008; ARAVIN *et al.* 2008). This finding indicates that mouse piRNAs guide specific *de novo* DNA methylation to silence their target transposons. Because mouse PIWI proteins are expressed in a developmental stage–specific manner (ARAVIN *et al.* 2008), the proper removal of PIWI proteins is essential for male germ-cell development.

Although transposons are the major targets of PIWI–piRNA complexes, how regulation of transposons is connected to defects in gametogenesis remains unknown. Activation of transposons in PIWI protein mutants may lead to the generation of double-stranded DNA breaks during abortive or successful transposition that activate a DNA damage checkpoint, resulting in a sterile phenotype (KLATTENHOFF *et al.* 2007). Thus, tissue-specific and developmental timing–specific expression of PIWI proteins and piRNAs may play major roles in maintaining the integrity of the genome and fertility of the organism. However, a mutation in *Drosophila piwi* that leads to transposon activation does not directly affect germline development (KLENOV *et al.* 2011). Also, derepression of L1 transposons by deletion of a part of a pachytene piRNA cluster does not affect spermatogenesis in mouse (XU *et al.* 2008), suggesting that transposon silencing and germline development can be separated.

Aside from the piRNAs that are derived from repetitive elements, there are classes of piRNAs that are not repetitive. For example, the piRNA populations expressed at different stages of mammalian testis development are distinct and those found at the pachytene stage are depleted in repetitive sequences (ARAVIN *et al.* 2007b). In addition, some piRNAs are found in genes (ROBINE *et al.* 2009) and pseudogenes (PANTANO *et al.* 2015) and are assumed to repress their host transcripts. Finally, there is some evidence that piRNAs are functional in the brain in rat (LEE *et al.* 2011) and mouse (NANDI *et al.* 2016). The connection between neural expression of piRNAs and the expression of transposable elements in the mammalian brain (MUOTRI *et al.* 2010; NANDI *et al.* 2016) has been observed and is clearly intriguing, but there is currently no evidence to further connect these two aspects of neuroscience. Recently the evidence was found that PIWI-piRNA complexes contribute to cancer development through aberrant DNA methylation resulting in genomic silencing and promoting a stem-like state of cancer cells and by mutagenic retrotranspositions and genomic instability initiation (SIDDIQI and MATUSHANSKY 2012; SIDDIQI *et al.* 2012; WATANABE and LIN 2014; MOYANO and STEFANI 2015; NG *et al.* 2016, reviewed in LITWIN *et al.* 2017)

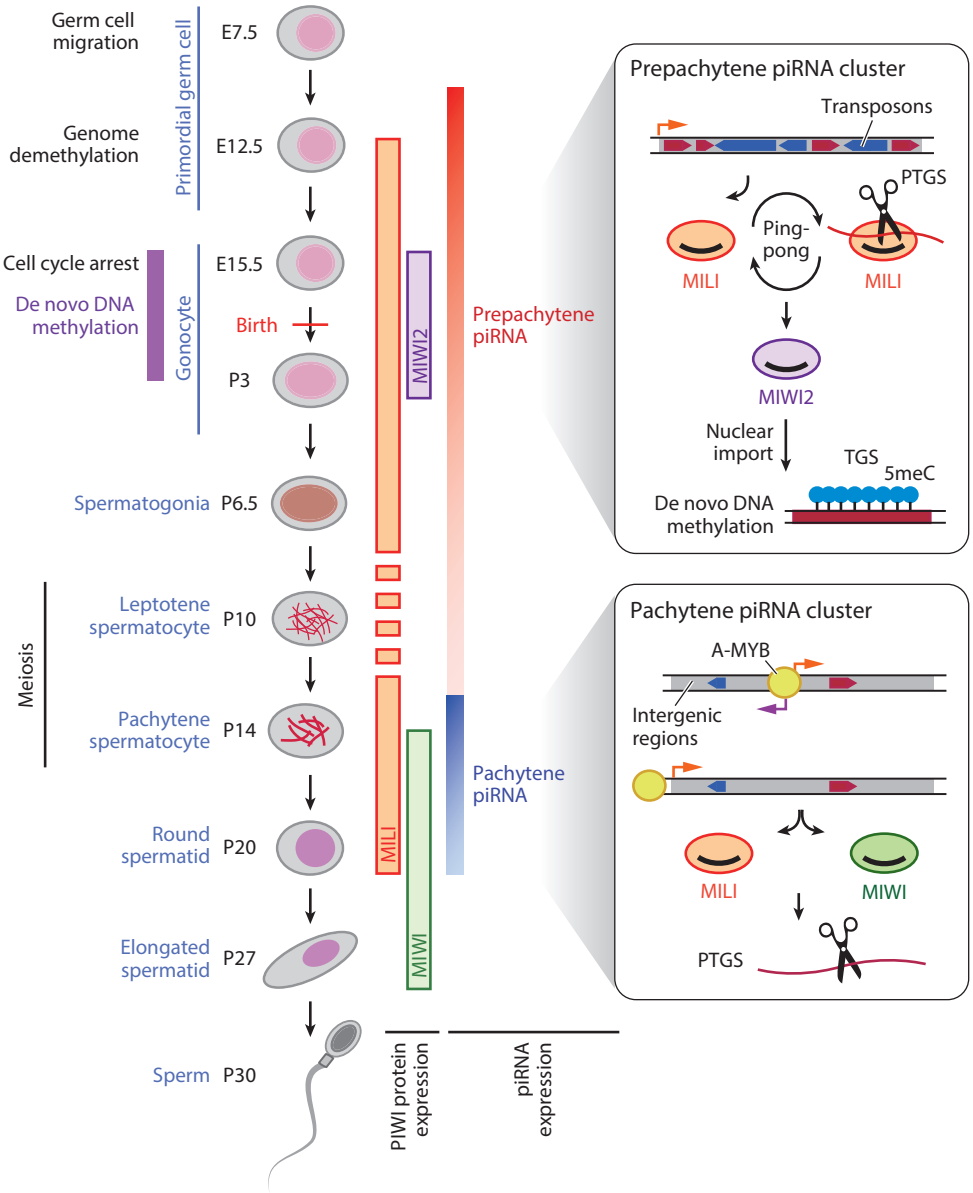


Figure 4: Expression patterns of MILI, MIWI, MIWI2, and mouse piRNAs during spermatogenesis. piRNAs are classified into pachytene piRNAs and prepachytene piRNAs on the basis of the stage in spermatogenesis when they are expressed. In gonocytes, MILI and MIWI2 bind to piRNAs from a prepachytene piRNA cluster, which consists mainly of transposons. MILI performs the homotypic ping-pong cycle to silence targets by PTGS and produce piRNAs that associate with MIWI2. MIWI2 localizes to the nucleus upon piRNA loading to accomplish nuclear silencing by de novo DNA methylation. Beyond the pachytene stage, MIWI and MILI are bound to piRNAs from pachytene piRNA clusters, a large fraction of which consists of intergenic regions. Pachytene piRNAs regulate their target genes by PTGS in the cytoplasm. Abbreviations: piRNA, PIWI-interacting RNA; PTGS, posttranscriptional gene silencing; TGS, transcriptional gene silencing (from IWASAKI *et al.* 2015).

1.2. piRNA CLUSTERS

piRNA clusters have been identified as genomic regions where a large number of piRNA reads are uniquely mapped (**Figure 5**). They often reside within or close to heterochromatin. The length of piRNA clusters ranges from a few kilobases to hundreds of kilobases (BRENNECKE *et al.* 2007; MALONE *et al.* 2009). For example, the *flam* locus, a major somatic primary piRNA cluster in the *Drosophila* X chromosome, is transcribed into a ~180-kb-long single-stranded transcript. Most of the *flam* transcript sequences correspond to transposons, including *gypsy*,

idefix, and *ZAM*, in an antisense orientation (**Figure 5*a***) (ZANNI *et al.* 2013). Thus, piRNAs derived from the cluster can target sense transcripts produced from cognate transposons dispersed throughout the genome. Another type of piRNA cluster, namely the *42AB* cluster in germline cells, is transcribed from both strands as a dual- stranded transcript. piRNAs generated from this type of cluster are integrated into both the primary pathway and the ping-pong cycle (BRENNECKE *et al.* 2007). Unlike the *flam* locus, dual-stranded piRNA clusters contain transposons that are sense and antisense oriented, mostly in a random manner, relative to the polarity of the locus transcription; therefore, it is unclear how the antisense bias observed for the germline primary piRNAs is generated initially. Perhaps Aub may associate with both sense and antisense primary piRNAs. Because the ping-pong cycle requires ongoing expression of the cluster and target transposons, the amplification loop steers piRNA production toward transcriptionally active and highly expressed transposons (**Figure 3**). Therefore, as long as there is an input of active transposon transcripts, antisense piRNAs are preferentially produced for Aub and the bias can be maintained. However, how the antisense bias for PIWI-bound piRNAs is enforced is enigmatic because, as in MIWI2, once loaded with piRNAs, PIWI is imported into the nucleus (SAITO *et al.* 2010; OLIVIERI *et al.* 2010); thus, it is unlikely to participate in the ping-pong cycle with AGO3, which is localized in the cytoplasm.

In mammals, the genomic location or synteny of piRNA clusters is highly conserved, although their primary sequences are not conserved (ARAVIN *et al.* 2007b; ARAVIN *et al.* 2006; GIRARD *et al.* 2006; HIRANO *et al.* 2014). Two types of piRNA clusters exist in mouse: unidirectional piRNA clusters and bidirectional piRNA clusters (**Figure 4**) (ARAVIN *et al.* 2007b; ARAVIN *et al.* 2008; GIRARD *et al.* 2006; LAU *et a*l. 2006). As in *Drosophila*, some mouse piRNA clusters are transcribed as a single strand, frequently spanning a long region of the genome; these are termed **unidirectional piRNA clusters**. **Bidirectional piRNA clusters** are transcribed both from sense and antisense strands, but unlike dual-stranded piRNA clusters in *Drosophila*, they are transcribed from a single central promoter. Transcription then switches to the opposite direction. A MYB-related protein underlies transcription of these piRNA clusters in mouse (LI *et al.* 2013).

Molecular process to define piRNA clusters. It was unknown for a long time how transcripts of the piRNA clusters can produce piRNAs and how they are protected against degradation to be relatively long transcripts. Dual-stranded piRNA cluster transcripts have unique characteristics: they lack a clear promoter, 5′ methyl-guanosine caps, and clear transcription termination (MOHN *et al.* 2014). Also, they seem not to be alternatively spliced, although they harbor intron-like sequences (ZHANG *et al.* 2014). These transcripts somehow escape from transcription termination and RNA decay and are processed into mature piRNAs.

Recent studies have shed some light on how piRNA generation from dual-stranded piRNA clusters in *Drosophila* is initiated (**Figure 5*b***). *Rhi*, an HP1a family gene, plays a major role in the identification of dual-stranded piRNA clusters in germline cells (KLATTENOFF *et al.* 2009). Furthermore, Rhi forms a complex together with Deadlock (Del) and Cutoff (Cuff), and this complex is anchored to H3K9me3-marked chromatin, where it would be defined as a piRNA cluster (MOHN *et al.* 2014). Cuff protects the 5′ ends of piRNA precursor transcripts from the cap-binding complex. This results in inhibition of alternative splicing as well as polyadenylation and transcription termination, leading to continuous transcription of the piRNA cluster transcript. Importantly, depletion of Piwi results in the loss of the Rhi–Del–Cuff complex at a subset of piRNA clusters. Although it seems counterintuitive, the competence of genomic regions to generate piRNAs appears to depend on the presence of a high level of H3K9me3 marks on the cluster. Association of the Rhi–Del–Cuff complex would allow piRNA clusters to be transcribed and would also protect the transcripts from transcription termination. Therefore, Piwi can specify piRNA clusters by guiding H3K9me3 marks to recruit the Rhi–

Del–Cuff complex. Rhi also functions together with Cuff and UAP56 to suppress alternative splicing of piRNA clusters (ZHANG *et al.* 2014). The suppression of alternative splicing would distinguish piRNA clusters from mRNAs, although the underlying mechanism is not yet clear.

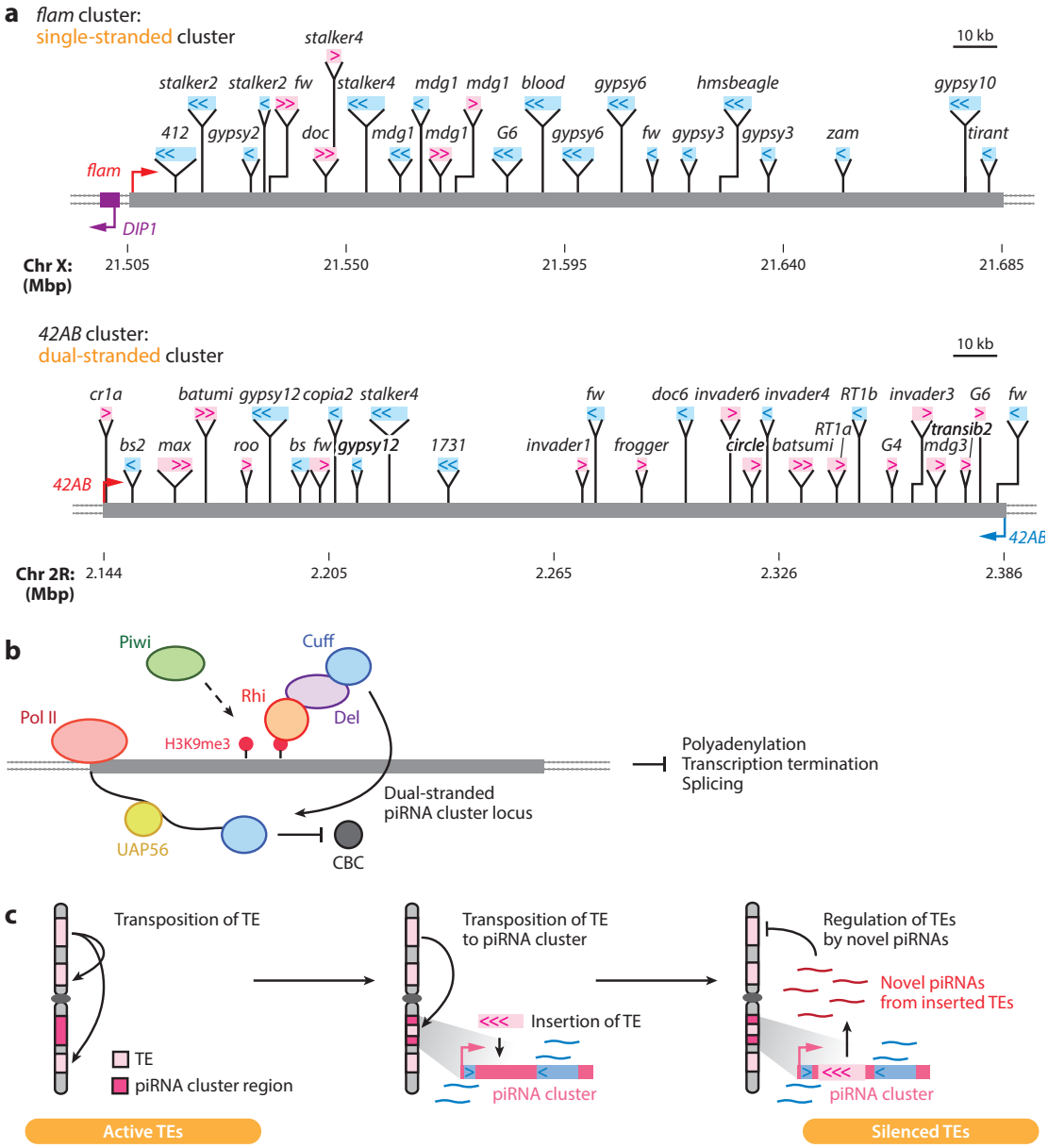


Figure 5: Transposons in *Drosophila* piRNA clusters and generation of novel piRNAs. (*a*) The structure of the *flam* piRNA cluster (single-stranded piRNA cluster) and *42AB* piRNA cluster (dual-stranded piRNA cluster) is illustrated with embedded transposons. In the case of *flam*, the transposons are frequently inserted in the antisense direction, resulting in production of piRNAs complementary to the original transposon. In contrast, no significant bias could be observed for *42AB*. Transposons longer than 2 kb are illustrated with approximate length and genomic position. (*b*) Noncanonical transcription of dual-stranded piRNA cluster transcripts. The dual-stranded piRNA cluster locus is H3K9me3 marked by Piwi, and Rhi is recruited together with Del and Cuff to the region. Cuff then binds to the newly formed 5′ end of a nascent piRNA cluster transcript, preventing polyadenylation and termination of Pol II by competing with the CBC. Cuff also inhibits splicing together with UAP56, leading to export and processing of piRNA cluster transcripts. (*c*) Model showing integration of transposons into piRNA clusters. Transposon integration results in acquisition of novel piRNA sequences, which can regulate the original transposon. Abbreviations: CBC, cap-binding complex; Cuff, Cutoff; Del, Deadlock; piRNA, PIWI-interacting RNA; Pol II, RNA polymerase II; Rhi, Rhino; TE, transposable element. (From IWASAKI *et al.* 2015).

Single-stranded primary piRNA clusters, such as *flam*, seem to be independent of this transcription mechanism, because the loss of Rhi leads to loss of piRNAs only from dual-stranded piRNA clusters. Indeed, *Rhi* is not expressed within ovarian somatic cells, where only single-stranded piRNA clusters are active. The molecular mechanism underlying transcription of single-stranded piRNA clusters remains unknown. Additionally, how piRNA clusters in different species are transcribed and whether they also have specific factors for the identification of piRNA clusters are issues that have only just begun to be addressed (YAMANAKA *et al.* 2014).

1.3. EVOLUTION OF THE piRNA SYSTEM

Nuclear small RNA-mediated silencing has been a topic of great interest in the recent years and many advances have been made in this field over all life kingdoms. In the fungus model *Schizosaccharomyces pombe* co-transcriptional gene silencing by RNAi (TGS) has been particularly well studied (for a review, see CREAMER and PARTRIDGE, 2011). Work on RNA-directed DNA methylation in plants has also yielded insights into small RNA-mediated silencing in higher organisms (reviewed by ZHANG and ZHU, 2011). The plant and algal miRNAs have gene structure, biogenesis and targeting properties distinct from those of animals (JONES-RHOADES *et al.* 2006; MOLNAR *et al.* 2007; ZHAO *et al.* 2007). These differences, considered together with the absence of miRNAs in fungi and all other intervening lineages examined, have led to the conclusion that miRNAs of animals and plants had independent origins (JONES-RHOADES *et al.* 2006). Of the many miRNAs reported in Bilateria (**Figure 6**), ~30 appear to have been present in ancestral bilaterians (PASQUINELLI *et al.* 2000; HERTEL *et al.* 2006; SEMPERE *et al.* 2006; PROCHNIK *et al.* 2007); however, none have been reported in the earliest branching animal lineages, leading to the hypothesis that bilaterian complexity might, in part, be due to miRNA-mediated regulation (SEMPERE *et al.* 2006).

PIWI-proteins and piRNAs are conserved in a wide range of eukaryotes, from sponges to humans (SIOMI *et al.* 2011; GRIMSON *et al.* 2008) and no PIWI homologs have been found outside animals so the piRNA system appears to be an animal-specific innovation. Between closely related species, the genomic locations of many piRNA clusters are conserved, but the sequences of the piRNAs themselves are not conserved between rat and mouse (ASSIS and KONDRASHOV 2009), *C. elegans* and *C. briggsae* (RUBY *et al.* 2006) or *Drosophila melanogaster* and *D. simulans* (MALONE and HANNON 2009). Thus, the overall picture of piRNA evolution at the sequence level is one of very rapid evolution.

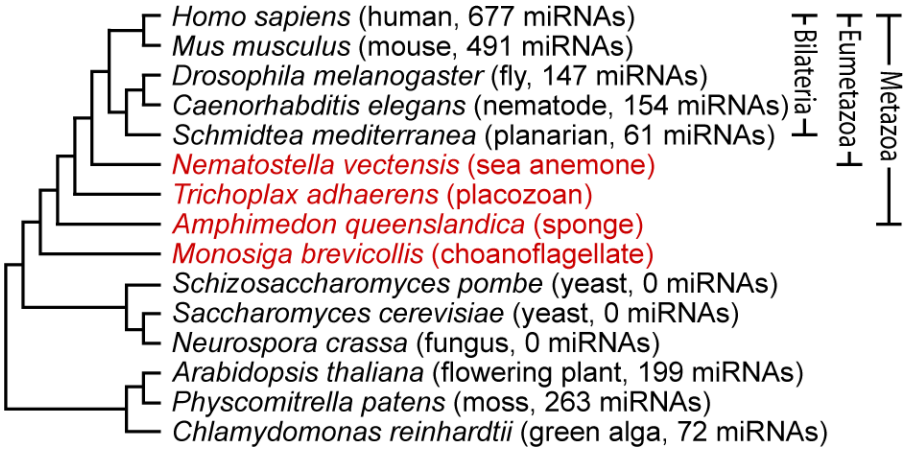


Figure 6: Phylogenetic distribution of annotated miRNAs. Cladogram of selected eukaryotes, with organisms investigated in the study of GRIMSON *et al.* (2008) indicated in red. Branching order of Bilateria is according to BOURLAT *et al.* (2008), and that of basal Metazoa is according to SRIVASTAVA *et al.* (2008). Annotated miRNA tallies are from miRBase (v10.1; GRIFFITHS-JONES *et al.* 2008).

1.3.1. Factors influencing piRNA evolution

Population genetic environment. Three important aspects of the population genetic environment that merit consideration are the effective population size, the generation time of the organisms involved, and the mutation rate.

From population genetics theory, the effective population size for a transposable element family is the effective population size of the host species multiplied by the average number of active copies of the transposable element per haploid genome (DOLGIN *et al.* 2008). The second quantity—the average number of active copies of the transposable element—varies by species and transposable element family. As a concrete example, there are estimated to be 80–100 active LINE-1 elements in the human genome. In the piRNA-transposable element system, the transposable elements are embedded in the host genome and thus are constrained to be replicated in the same generation time as the host genome. The transposable elements also have roughly the same mutation rate as the host genome. Even if the transposable elements are biased to certain parts of the genome, the differences in the local mutation rate are relatively minor. The mutation rate is a significant consideration because there can be mutations in transposable elements that are countered by compensatory mutations in piRNAs (KUMAR and CHEN 2012).

Alternative mechanisms of genomic defense. It is important to place the defense mechanisms described here in the context of other defense mechanisms in the cell. In the case of piRNAs, it was suggested in a comparative study that they might be more effective at repressing transposable elements than CRISPRs are at repressing phages (KUMAR and CHEN 2012). Nonetheless, other molecular mechanisms also play a significant role in the repression of transposable elements in the germline. One important mechanism is DNA methylation that prevents transcription of transposable elements in the germline. While little is understood about the evolutionary properties of DNA methylation or how DNA methyl marks are directed to specific loci in the genome, intriguingly, piRNAs are also implicated in the maintenance of DNA methylation in mammals (WATANABE *et al.* 2011; ARAVIN *et al.* 2008). When this mechanism is more fully worked out at the molecular level, it may be possible to start understanding the interplay between piRNA-mediated regulation of transposable elements at the chromatin level versus the RNA level. On a broader scale, RNAi-related systems in general are known to be involved in genome defense (CERUTTI and CASAS-MOLLANO 2006) and may have even originated for that purpose (KUMAR and CHEN 2012).

The insertion mechanism for new piRNAs. Current evidence suggests that the transposons either jump randomly into piRNA loci or perhaps have a mild preference for inserting into the piRNA loci (KHURANA *et al.* 2011). Selection for relevant piRNA insertions then occurs at the level of individual germ cells, and in this way adaptation to the invasion of the new transposable element can occur over the lifetime of an individual (KHURANA *et al.* 2011).

The piRNA targeting mechanism. Independent of their role as transposon repressors, piRNAs appear to have a role in the control of endogenous gene expression. Such roles include the control of mRNA translation, direction of both euchromatic and heterochromatic histone modifications, and control of higher order chromatin structures (THOMSON and LIN 2009). These nontransposon related roles are expected to apply different evolutionary pressure to some piRNAs, perhaps more similar to the evolutionary properties of miRNAs (CHEN and RAJEWSKY 2007) There is currently no evidence for a seed sequences in piRNAs and current evidence points to a requirement for nearly complete complementarity over the full length of the piRNA for targeting. Another interesting feature of piRNAs is that there are many redundant piRNA

sequences, which would serve to reduce the evolutionary constraint on individual piRNA sequences (KUMAR and CHEN 2012).

Evolution of ping-pong cycle. The piRNA system is known to be ancient as PIWI proteins, and the ping-pong signature is conserved in basal metazoans (GRIMSON *et al.* 2008). Since the ping-pong mechanism is a positive feedback loop, one question is how the ping-pong mechanism is started in the first place. In *Drosophila*, a partial answer is provided by the fact that piRNAs are deposited maternally into the embryo (BRENNECKE *et al.* 2008; BLUMENSTIEL *et al.* 2005). piRNAs can thus be inherited epigenetically across generations. A second answer comes from evidence in *Drosophila*, where primary piRNAs are produced in the somatic follicle cells and delivered to the germline to start the ping-pong cycle (LI *et al.* 2009; MALONE *et al.* 2009). A similar mechanism is found in *Arabidopsis* for a different class of small RNAs (SLOTKIN *et al.* 2009), suggesting that this may be a universal mechanism where transposons are activated outside of the germline to generate small RNAs, thus reducing the chance of deleterious transposon insertions in the germline. A third possibility is suggested by a related system of RNAi and heterochromatin formation in fission yeast, in which degradation products from random abundant transcripts are used to prime Argonaute proteins and start a positive feedback loop (HALIC and MOAZED 2010).

1.3.2. The early origins of small RNAs in animals

piRNAs have not been reported outside of Bilateria, raising the question of whether a rich small-RNA biology is characteristic of more complex animals, or whether these small RNAs might have emerged earlier in metazoan evolution. In the work of GRIMSON *et al.* (2008), small RNAs from animal phyla that diverged before the emergence of the Bilateria were identified. *Eumetazoa* includes the *Bilateria* as well as the *Cnidaria*, which among sequenced genomes is represented by the starlet sea anemone, *Nematostella vectensis*, which possesses an extensive repertoire of miRNA genes, two classes of piRNAs and a complement of proteins specific to small-RNA biology comparable to that of humans. Despite the wholesale shift in their predicted targeting, the *Nematostella* and bilaterian versions of miR-100 had similarity throughout the RNA, suggesting common origins. This result confidently extends the inferred origin of metazoan miRNAs back to at least the last common ancestor of these eumetazoans. Although the short length of miRNAs may cause sequence divergence to obscure common ancestry, it is noteworthy that only one of the 40 *Nematostella* miRNAs appeared homologous to extant bilaterian miRNAs, and even this one seemed to have profoundly different targeting properties.

The poriferan *Amphimedon queenslandica* (sponge), one of the simplest animals and a distant relative of the Bilateria, also possesses miRNAs, both classes of piRNAs and a full complement of the small-RNA machinery. Animal miRNA evolution seems to have been relatively dynamic, with precursor sizes and mature miRNA sequences differing greatly between poriferans, cnidarians and bilaterians. miRNAs appear to have been available to shape gene expression throughout the evolution and radiation of animal phyla. Nonetheless, the numbers identified in simpler animals (8 unique miRNAs in *Amphimedon* and 40 in *Nematostella*) were lower than those reported in more complex animals (**Figure 6**). Although miRNAs expressed only under specific conditions or at restricted developmental stages were possibly missed in these and other animals, the results are consistent with the idea that increased organismal complexity in *Metazoa* correlates with the number of miRNAs and presumably with the number of miRNA- mediated regulatory interactions.

The possibility that piRNAs have early origins was also considered. PIWI proteins, the effectors of bilaterian piRNA pathways, are found in diverse eukaryotic lineages (although not in plants or fungi), implying their presence in early eukaryotes (CERUTTI and CASA-MOLLANO 2006). In cases characterized, however, the small RNAs associated with non-metazoan PIWI proteins resemble siRNAs more than bilaterian piRNAs, raising the question of when piRNAs

of the types found in Bilateria might have emerged. The genomes of both *Amphimedon* and *Nematostella*, but not that of *Trichoplax*, encode PIWI proteins and express many ~27-nucleotide RNAs with a 5′-terminal uridine (5′U)—features reminiscent of piRNAs in vertebrates and flies (Aravin et al. 2007). Moreover, the genomic loci producing a large fraction of the *Nematostella* reads closely resembled the loci producing bilaterian piRNAs, particularly the pachytene piRNAs (ARAVIN *et al.* 2007a). A similar clustering of genomic matches of *Amphimedon* 5′U 24–30-nucleotide RNAs was observed, although the loci were smaller and accounted for fewer reads.

The piRNAs were the type of small RNAs most abundantly sequenced in *Nematostella* and *Amphimedon*. A similar phenomenon is observed in mammalian testes, in which the pachytene piRNAs greatly outnumber the miRNAs and initially obscured detection of a second class of mammalian piRNAs, which resemble the most abundant *Drosophila* piRNAs with respect to both their biogenesis and their apparent role in suppressing transposon activity (ARAVIN *et al.* 2007b). Most of the *Nematostella* and *Amphimedon* genomic loci with clustered piRNA matches resembled the first class of piRNAs, in that they tended to fall outside of annotated genes and spawned piRNAs predominately from only one DNA strand. Therefore, miRNAs and piRNAs, as classes of small riboregulators, have been present since the dawn of animal life, and indeed might have helped to usher in the era of multicellular animal life (GRIMSON *et al.* 2008).

All in all, the two classes of piRNAs found previously in mammals and flies have existed since the origin of metazoans: the class I piRNAs, represented by the mammalian pachytene piRNAs, which have unknown function during germline development; and the class II piRNAs, which use the ping-pong cleavage and amplification cascade to quiet expression of certain genes, particularly those of transposons. Indeed, the sequence-based transposon silencing by piRNAs, which by virtue of the feed-forward amplification process focuses on the most active transposon species, might be one of the main drivers of transposon diversity in animals (GRIMSON *et al.* 2008). piRNA targeting mechanism provides further opportunity for investigation, including evolutionary studies. As they are involved in protection of the genome against retrotransposition, they are expected to evolve very fast between species with low levels of conservation and also to vary within species, co-evolving with their targets and being under detectable selective constraint at a shorter timescale if they are rapidly evolving regions (LUKIC and CHEN 2011). In light of the highly-conserved role for piRNAs in regulating TEs, discoveries from this system have taxonomically broad implications for the evolution of repression (KELLEHER 2016).

1.3.3. Evolution of piRNA system in *Drosophila*

Although the piRNA system is not understood well enough for detailed mathematical modeling, there has been one attempt by LU and CLARK (2010) at modeling piRNA-transposable element co-evolution using computer simulations. From their simulation, they suggested that retrotransposon insertions that are repressed by piRNAs can reach high frequencies or even be fixed in the population because their deleterious effect is attenuated by piRNA repression.

The idea that the piRNA pathway and transposable elements might co-evolve in a Red Queen-like scenario has been explored by a number of authors. In this scenario, alternating rounds of adaptation and counter-adaptation would lead to increased rates of positive selection. In a molecular evolutionary analysis examining species across the *Drosophila* genus, a higher transposable element abundance was found to be positively correlated with greater codon bias in piRNA pathway genes but not with an increased rate of amino acid substitution in these genes (CASTILLO *et al.* 2011). The authors suggested that these observations indicate that positive selection on piRNA pathway genes occurs mainly at the level of translation efficiency mediated by codon usage (although other explanations for codon bias are possible) as opposed

to amino acid substitution (CASTILLO *et al.* 2011). Further, a resequencing study of a number of defense genes in *D. melanogaster* and *D. simulans* concluded that RNAi genes have the highest rate of adaptive evolution over all immune-system genes (OBBARD *et al.* 2009). Subsequent studies also found recurrent adaptation across the twelve sequenced *Drosophila* genomes for a number of piRNA pathway genes, including SPN-E, AUB, KRIMP, SQU, ZUC, as well as Rhino (KOLACZKOWSKI *et al.* 2010; VERMAAK *et al.* 2005; SIMKIN *et a*l. 2013; SONG *et al.* 2014).

The resistance of natural *Drosophila* strains to transposon invasion varies considerably, but the nature of this variability is unknown. A recent study of Russian group (RYAZANSKY *et al.* 2017) discovered that natural variation in the efficiency of primary piRNA production in the germline causes dramatic differences in the susceptibility to expansion of a newly invaded transposon. In one of the most transposon-resistant strains, increased content of primary piRNA is observed in both the germline and ovarian somatic cells, suggesting that polymorphisms in piRNA pathway factors are responsible for increased piRNA production. The authors described a phenotype that might be rapidly evolved under high selective pressure shown affect piRNA pathway proteins. They also demonstrate a likely explanation as to why an overly active piRNA pathway can cause more harm than good in *Drosophila*: Highly efficient piRNA processing leads to elimination of domesticated telomeric retrotransposons essential for telomere elongation, an effect which has been observed in a natural strain that is extremely resistant to transposon invasión (RYAZANSKY *et al.* 2017).

Multiple studies carried out in Drosophila, have demonstrated that the machinery of piRNA biogenesis is often the target of positive selection (reviewed in BLUMENSTIEL *et al.* 2016). Because transposable elements (TEs) are a form of genetic parasite, positive selection in the piRNA machinery is often explained by analogy to the signatures of positive selection commonly observed in genes that play a role in host-parasite dynamics. However, the precise mechanisms that drive positive selection in the piRNA machinery are not known. There is a suggestion that recurrent positive selection in the piRNA machinery can be partly explained by an ongoing tension between selection for sensitivity required by genome defense and selection for specificity to avoid the off-target effects of maladaptive genic silencing by piRNA (BLUMENSTIEL *et al.* 2016).

Overall, these studies proved the elevated rates of evolution on piRNA pathway genes, consistent with its role in genome defense. While the molecular details of the Red Queen scenario for piRNAs and transposable elements are unclear, certain aspects of transposable element evolution, such as a higher global transposition rate, could select for certain features of piRNA-pathway genes, such as stronger binding affinity of the proteins for piRNAs.

piRNA system and canalization. An interesting and somewhat contentious aspect of the role of the piRNA system in evolution is its role in canalization. Canalization, most famously associated with WADDINGTON (1942), refers to the buffering of genetic or environmental insults to ensure developmental robustness. In a seminal paper, RUTHERFORD and LINDQUIST (1998) suggested that Hsp90, a protein chaperone, is a phenotypic capacitor in *Drosophila*, meaning that it buffers genetic variation but when it is compromised, that variation is revealed in multiple mutant phenotypes, at least some of which could be adaptive in certain environments (JAROSZ and LINDQUIST 2010). Similar results were subsequently demonstrated in Arabidopsis (QUEITSCH *et al.* 2002), suggesting that Hsp90 might play an evolutionarily conserved role as a phenotypic capacitor.

The connection between canalization and the piRNA system comes from a recent report that in *Drosophila*, Hsp90 regulates the piRNA pathway, which in turn regulates the insertion of transposons (SPECCHIA *et al.* 2010). It was further suggested that Hsp90 interacts in a protein complex with PIWI protein and mediates canalization by epigenetic silencing of genetic variation and suppressing transposon insertion (GANGARAJU *et al.* 2011). Thus, one potential

mechanism by which the disruption of Hsp90 creates phenotypic variation is not by revealing previously cryptic variation, as suggested by RUTHERFORD and LINDQUIST (1998), but rather through de novo mutations generated by transposon insertions. For this to be true, a strong bias in the preference in genome position for transposition insertion dependent on genetic background is required, and while such a preference is known to exist, it is not clear if it is strong enough to fully explain the results of the RUTHERFORD and LINDQUIST (1998) experiments. Also, the piRNA study (GANGARAJU *et al.* 2011) showed an effect on gene regulation separable from the effect on transposons. Conversely, imprecise transposon deletions could have a mutagenic effect and would necessarily be in the same place in the genome so more work needs to be done to define the exact role of piRNAs in canalization (KUMAR and CHEN 2012).

1.3.4. Human piRNA evolution

The studies concerning evolutionary aspects and selective forces acting on the piRNA system in the human genome are rather scarce and need to be widely completed in different aspects. A recent study of 24,646 human piRNA sequences, clustered in 36 broad clusters, based on derived allele frequency spectrum suggested that there is strong negative selection at the sequence level for human piRNAs but only in the three African populations and not in any of the seven non-African populations studied (LUKIC and CHEN 2011).

It has been reported (LOHMUELLER *et al.* 2008) that Europeans harbor more deleterious polymorphisms than Africans because in general, non-African groups have smaller population sizes than Africans and therefore are more sensitive to the effects of random drift. At first glance, the higher amounts of selective constraint on piRNA sequences that was observed in Africans compared with non-Africans are consistent with these data. However, it could be expected a population-wide effect such as a population size difference to be visible in other classes of functional sites as well, in particular nonsynonymous sites. The fact that in the same study such an effect was not observed suggested that the increased selective constraint in African populations is in fact specific to piRNAs. Because the biological function of piRNAs in humans is still poorly understood, it is difficult at this point to connect the stronger selective constraint in Africans to a particular biological function. However, if a significant fraction of the uniquely mapping piRNAs are involved in transposon defense, then the observed patterns are consistent with data that show a much higher rate of transposon insertions in African compared with non-African populations (EWING and KAZAZIAN 2010). Under this scenario, a higher transposition rate in Africans imposes stronger selective pressure on piRNAs to repress the TEs (LUKIC and CHEN 2011). Even though these authors also demonstrated that piRNAs have evolved quickly between human and chimpanzee. They claimed that the apparent contradiction between their intraspecies analysis and the interspecies analysis can be resolved by one of two nonexclusive interpretations. One explanation is that the strength of selective constraint may simply differ between these two time scales. Such rapid evolution would be expected for genes that mediate defense against parasites such as transposons. The other explanation is that the interspecies substitution rate, but not the derived allele frequency distribution, is affected by mutation rate biases (LUKIC and CHEN 2011). This latter explanation highlights the need of the implementation of neutrality tests based on the comparison of intraspecific polymorphism with interspecific substitution level in future population genetic studies of piRNAs.

A further intriguing observation from the analysis of human piRNAs and transposable elements is the depletion of piRNA matches in the reverse transcriptase region of human LINE-1 elements, though not mouse LINE-1 elements (LUKIC and CHEN 2011). This observation suggests the possibility that at least one reverse transcriptase might be functional for the host and therefore protected from piRNA-mediated repression. The authors measured the sequence conservation of human piRNAs in primates and found that human piRNAs have evolved at a

similar rate to their flanking regions between human and chimpanzee, which was consistent with previous results in rodents, *Drosophila*, and nematodes and has been previously interpreted to mean that the sequences of the piRNAs might not be functionally important (GIRARD *et al.* 2006).

Beyond nucleotide sequence changes, it is also interesting to study the relationship of piRNA clusters and copy number changes, as an increase in copy number could potentially increase the level of gene expression of piRNAs. Assis and Kondrashov studied the evolution of piRNA clusters between mouse and rat and found a very high rate of piRNA cluster duplication, which they suggested is indicative of positive selection for higher expression level of piRNAs (ASSIS and KONDRASHOV 2009). Nevetheless, when GOULD *et al.* (2012) studied the evolution of piRNA-generating loci at the level of copy number variation (i.e. duplications and deletions) using genome-wide copy number variation data from two human populations (European and Yoruban), they showed that at the level of copy number variation there is strong selective constraint and a very high mutation rate in human piRNA-generating loci.

Several of the most basic evolutionary properties that still remain to be elucidated are: (i) the rate of evolution of piRNAs at the sequence level; (ii) the rate of evolution of piRNA-generating loci at the level of copy number variation; and (iii) the true amount of sequence in piRNA that is under selective constraint—particularly the question of whether there is a seed sequence or not. Beyond these basic questions of molecular evolution are broader evolutionary questions such as the interplay of the piRNA system with alternative defense mechanisms against foreign nucleic acids, such as DNA methylation. Once we can compare the different defense mechanisms, we can study the conditions under which the piRNA system might play important role in evolution (GOULD *et al.* 2012).

2 | OBJECTIVES

Nowadays the development of high-throughput DNA sequencing technologies, steadily increasing the number of individual genomes available (especially for humans) in publicly accessible databases, with simultaneous advancement in bioinformatic tools development for population genomic analysis, give us an opportunity to test previously set hypotheses in more precise and deeper manner. Our primary objective in this study was to revisit previously published results, extending them to different populations and chromosome datasets from publicly available databases, and elucidate which selective forces are mainly acting on piRNA cluster regions in the human genome at the nucleotide level. This main objective was divided into several consecutive steps:

1. Data retrieval, filtering and processing.

2. Estimation of genetic diversity in the piRNA clusters within the human species and their divergence from the chimpanzee genome.

3. F_{ST} analysis using polymorphism data to estimate differentiation between several human populations.

4. Estimation of evolutionary rates within different clusters.

5. Application of different neutrality and selection tests, including Tajima's D, Fu and Li's D, and a modified McDonald and Kreitman test, using the chimpanzee genome as outgroup.

3 | MATERIAL AND METHODS

3.1. PRELIMINARY SETTINGS

On the first place, a preliminary assessment was conducted in order to establish the general framework of the project and identify its main necessities. The existing sources of genomic data were reviewed focusing on large international consortia. To date, the 1000 Genomes Project (1000 GP; 1000 GP CONSOTIUM 2015) is the global reference of human variation, containing 2504 individuals from 26 populations all over the world grouped into five super-populations. In our study this data were used to perform genome-wide piRNA cluster variation analysis. For interspecies comparisons, chimpanzee was selected as outgroup based on its evolutionary proximity to humans (divergence time ~ 5-7 Mya) and its extensive utilization in other comparative studies (e.g. VARKI AND ALTHEIDE 2005; BAKEWELL *et al.* 2007; etc.).

As a next step, the genome version on which this study is based was determined among the two most recent versions. In human genomic studies, hg19/GRCh37 is by far the predominant genome assembly: most members of the 1000 GP have released their data mapped to hg19. For human-chimpanzee divergence study, alignment between hg19 and the panTro4 assembly was already available in the Vista Browser (FRAZER *et al.* 2004, http://pipeline.lbl.gov/cgi-bin/gateway2) and it was reasonable to use it in our study. Besides, in this genome assembly the structural variants were already well annotated in phase 3/integrated_sv_map, which allowed us to easity exclude them from the analysis. All in all, the hg19/GRCh37 assembly version is the most suitable to human variation study by this moment.

3.2. DEFINING piRNA CLUSTERS

Previously, several datasets of small RNAs from human testis were analyzed and pachytene piRNA clusters were predicted from them based on the GRCh38 reference genome using the proTRAC software (ROSENKRANZ AND ZISCHLER 2012) by Laura Llobet (Josep Carreras Leukaemia Research Institute). The positions of 254 piRNA clusters, distributed along all 23 chromosomes and ranged between 34 025 and 406 879 bp, were predicted and stored in a BED file, which then were converted into GRCh37 version with *LiftOver* tool from UCSC web page: http://genome.ucsc.edu/cgi-bin/hgLiftOver.

3.3. DATA RETRIEVAL AND PROCESSING

The steps performed for the retrieval, processing and analysis of the data are described below. Following their successful testing in the pilot region (chromosome 22), these steps were integrated in a pipeline for the analysis of whole bunch of piRNA clusters all over the human genome, developed and tested in a Linux environment on the Andromeda UAB Server. The simplified schemes of the pipelines with all implemented tools are represented in **Figure 7.**

The piRNA cluster positions were used to extract the nucleotide information for each cluster from 1000 GP (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/), generated by the 1000 GP Phase 3 and stored in variant call format (VCF) for each chromosome separately. This step was performed with *tabix* package. The VCF is a text file format developed by 1000 GP consortium (DANECEK *et al.* 2011) for efficient storage of DNA polymorphism data, normally in compressed manner. Compared to other formats like FASTA, its main advantage is dramatic reduction in size and, therefore, memory requirements. These files, which contained variants detected across 2 504 individuals, were filtered to exclude 243 individuals found to be inbred by a combination of multi-point approaches, as RELPAIR and Fsuite, exhibiting inbreeding coefficients similar to those for the first-cousin offspring (GAZAL *et*

al. 2015). This and other operations involving VCF files were carried out with *bcftools* (LI 2011), which belongs to the *Samtools* suite and is based on the HTSlib C library (LI *et al.* 2009). Therefore 243 individuals were filtered with *bcftools view*. Next, another filtering step implemented in *substractBed* command was performed to eliminate the positions involved in Structural Variants (SV) listed in one of the papers of the 1000 GP consortium (SUDMANT *et al.* 2015) to avoid possible complications with the diversity analysis of cluster regions. The nucleotides overlapping protein-coding genes and pseudogenes, annotated in GENECODE project (HARROW *et al.* 2012) were also discarded from the dataset using the same *substractBed* tool.

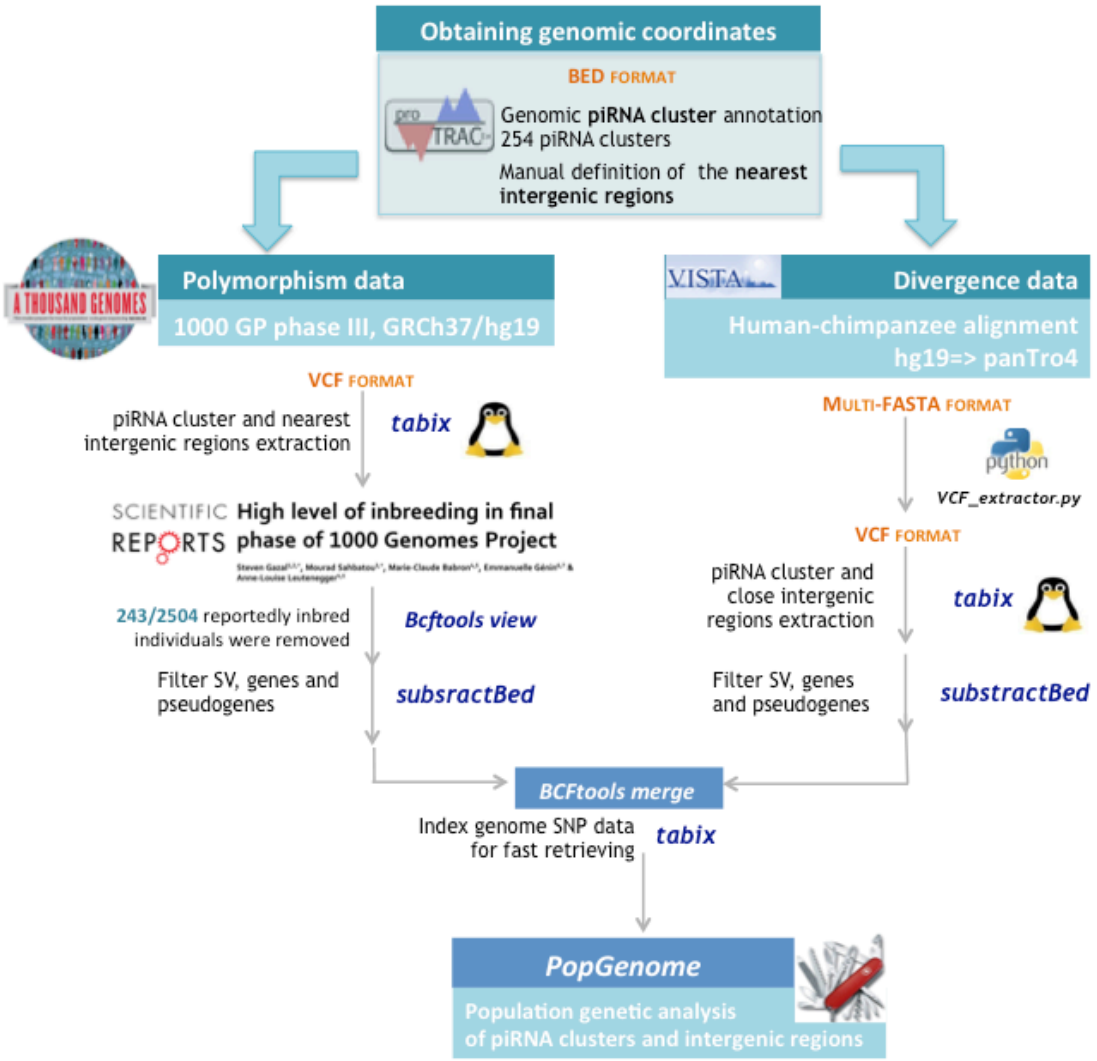


Figure 7. Scheme of the pipeline for the description of variation patterns of piRNA clusters and intergenic regions.

To concatenate all clusters from each chromosome into a single file and to sort them by chromosomal position, two commands, *vcf-concat* and *vcf-sort*, from *VCFtools* were subsequently implemented (DANECEK *et al.* 2011).

For further divergence computation we used the chimpanzee genome assembly as outgroup. Hence, precomputed human-chimpanzee alignments for each chromosome were obtained from the VISTA browser in multi-FASTA format (.mfa) (FRAZER *et al.* 2004). This option was preferred over other sources (such as the UCSC Genome Bioinformatics site, which is regularly updated (SPEIR *et al.* 2016) because it offered a straightforward conversion to VCF, which was convenient for the subsequent integration with the 1000 GP file. To convert multi-

FASTA format files into VCF a custom Python algorithm developed by Roger Multet (laboratory of Bioinforatics of Genome Diversity, UAB) was implemented.

Following the same pipeline for the region extraction, the piRNA clusters were extracted from converted chimpanzee VCF files and filtered for SV with *tabix* and *substractBed* tools subsequently and separately for each chromosome. The obtained files were concatenated and sorted with *VCFtools* as it was described above.

To perform modified McDonald and Kreitman test, described below in the next section, as well as for comparison of variation and evolutionary rates, we selected 254 intergenic regions close to the clusters, which were supposed to share the same evolutionary history as the cluster regions. This procedure was conducted using a number of Linux tools. First of all, the file with GENCODE annotation for GRCh37/hg19 genome reference version was downloaded and the genes were defined, obtaining their coordinates stored in BED files with an *awk* function. Then, *sortBed* and *complementBED* commands, along with *fetchChromSizes.txt* file available in UCSC Browser (http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/fetchChromSizes), were subsequently executed to obtain all intergenic positions across the human genome stored in a BED file. Finally, the intergenic regions, situated as near as possible to the piRNA clusters, were manually chosen, extracted from human and chimpanzee VCF files with subsequent filtering, concatenation and sorting using the same procedures that were already described above. For the rest of this work, when we say "intergenic region" we mean intergenic region that do not contain any piRNA clusters, though the latter constitute only a very small fraction of all intergenic regions.

As a final step of data preparation, human population and chimpanzee *VCF* files were merged together with *bcftools merge*, separately for piRNA clusters and intergenic regions.

3.4. ANALYSIS OF PIRNA CLUSTER VARIATION

Considering that data were available in VCF – and that any other formats would have required much more computational power - we reviewed available software able to deal with VCF files. *VCFtools* (DANECEK *et al.* 2011), *P4* (BENAZZO *et al.* 2015) and *DivStat* (SOARES *et al.* 2015) are some of those tools, but in spite of some advantages (4P is extremely fast, for instance), their capabilities were limited for the needs of this project. On the contrary, the R package *PopGenome* (PFEIFER *et al.* 2014) provides an ideal combination of built-in functions and flexibility to integrate user-defined functions. *PopGenome* converts the data into a biallelic matrix constituted by 0 (reference alleles) and 1 (alternative alleles), whose rows correspond to sequences and columns to SNP positions in the alignment. With some basic manipulations, this biallelic matrix can be employed to implement the most common algorithms to measure polymorphism and divergence.

All analyses were performed for each cluster separately as well as grouping the clusters by chromosome and comprising all populations together, to increase computational power. We also analysed five super-populations (from now on will be referred as "populations") defined in 1000 GP (AFR- Africans; AMR- native Americans; EAS- East Asians; EUR- Europeans; SAS- South African) one by one to give us more complete overview of general population genetic patterns. In per-population analysis we used only female individuals to shorten computational time, including African (303 females), American (162 females), European (256 females), East Asian (251 females) and South Asian (182 females).

The difference in variability between piRNA clusters and intergenic regions were examined using the package's default methods: *Pi* (π; NEI and GOJOBORI 1986) and theta of Watterson ($\Theta_{Watterson}$; WATTERSON 1975). The level of genetic differentiation between populations within clusters was estimated with F_{ST} index. The F_{ST} statistics is based on the average number of differences between sequences from two different populations.

Neutrality tests were performed also using *PopGenome* build-in functions. To study possible deviations from neutrality Tajima's *D* (TAJIMA 1989) was calculated, which computes a standardized measure of the total number of segregating sites and the average number of mutations between sequenced pairs. If the mutations presented in the sequences are neutral, D is expected to be equal to zero. A negative Tajima's *D* signifies an excess of low frequency polymorphisms relative to the expectation due to either a selective sweep as a result of directional selection or a recent bottleneck with subsequent population expansion. A positive Tajima's *D* indicates a decrease in population size and/or balancing selection.

Fu and Li's *D* test (FU and LI 1993) assumes that "old" mutations will tend to be found in the older part of the genealogy while "new" mutations will likely be found in the younger part of the genealogy. The older part of the genealogy consists mainly of *internal* branches, while the younger part mainly of *external* branches. A branch is said to be *external* if it directly connects to an external node, otherwise it is said to be *internal*. In the presence of purifying selection there will be an excess of mutations in the external branches because deleterious alleles are present at low frequencies. Also there is likely to be excess of mutations in the external branches if an advantageous allele has recently become fixed in the population, because then the majority of the mutations in the population are expected to be young. On the other hand, if balancing (overdominant) selection is operating at the locus, then some alleles may be old and so there may be deficiency of mutations in the external branches. Therefore, comparing the numbers of mutations in internal and external branches with their expectations under selective neutrality should be a powerful way to detect selection, which is the idea behind this test. We have used *chimpanzee* as outgroup, because without outgroup it is difficult to infer accurately the number of external branches.

For comparison of evolutionary rate between piRNA clusters and intergenic regions, as its proxy, we calculated mutation rate μ, using following formula:

$$\mu = \Theta_{Watterson} \div 4\mathrm{Ne},$$

assuming that effective size of all human populations and chromosomes is equal to 10,000.

Rapid expansion of piRNA clusters during the course of mammalian evolution is most likely driven by positive selection. Because positive selection increases the rate of evolution, but does not induce any long-lasting polymorphisms, the McDonald-Kreitman test would indicate positive selection. If piRNAs are indeed involved in transposon silencing, it is natural to assume that selection for cluster acquisition is caused by an arms race between expanding families of mammalian transposons and piRNA clusters (ASSIS AND KONDRASHOV 2009).

To test this hypothesis we used a modified version of the McDonald-Kreitman test (MKT) (MCDONALD AND KREITMAN 1991; EGEA *et al.* 2008) to study the evolution of human piRNA-generating loci at the SNPs level. Traditionally, the McDonald-Kreitman test contrasts a putatively neutral class of sites (typically synonymous sites) to a putatively selected class of sites (typically nonsynonymous sites). For our study of non-coding piRNA clusters, instead, intergenic regions, tightly linked with clusters, were used as the putatively neutral class of sites and piRNA-generating loci as the putatively selected class of sites. The traditional MKT also contrasts divergence (typically fixed nucleotide substitutions between two species) to polymorphism (typically single nucleotide polymorphisms in one of the species). Under the assumption of neutrality, the ratio of neutral changes between species should be the same as within species.

For our study, the number of biallelic sites were used as polymorphism data and calculated with the standard function of *PopGenome* package. The divergence data in piRNA clusters and intergenic regions for MKT was computed by a custom R script, comparing SNPs of human clusters with orthologous sequences of chimpanzee set as outgroup, counting the number of diverged sites between two species and excluding the polymorphic sites in humans.

This value was divided by the overall number of nucleotides in clusters to obtain the observed divergence *D*, or proportion of sites with divergent nucleotides.

The Neutrality Index (*NI*) was calculated as $NI = \frac{\frac{P_n}{P_s}}{\frac{D_n}{D_s}}$,

where P_n is a number of polymorphic non-neutral sites (polymorphic sites within piRNA clusters), P_s is a number of polymorphic neutral sites (polymorphic sites within intergenic regions from the same chromosome), D_n is the number of divergent non-neutral sites and D_s is the number of divergent neutral sites. Then *a*, which is the proportion of adaptive substitutions (SMITH and EYRE-WALKER 2002) and ranges from -∞ to 1, was calculated as 1-*NI*. If the ratio of fixed differences to polymorphisms is much higher for non-synonymous changes (i.e. $D_n/P_n \gg D_s/P_s$), resulting in a positive value of *a*, this indicates that genetic changes have been subject to positive selection, which promote the fast fixation of advantageous variants. A negative *a* value indicates that there were fewer nonsynonymous substitutions in evolution than expected given the number of non-synonymous polymorphisms. That can be attributed to purifying selection preventing the fixation of harmful mutations (the number of divergent non-neutral changes D_n is lower than expected), but also an excess of non-neutral polymorphisms could be explained by balancing selection. If *a* is approximated to zero the null hypothesis of neutral equilibrium cannot be rejected.

The main goal of this project was to develop a pipeline for the analysis of nucleotide variation in piRNA clusters, and potentially in any particular genomic region. The pipeline was developed with the steps described in the Material and Methods section and was implemented for SNPs of piRNA clusters and intergenic regions for all autosomes from the 1000 Genome Project (1000 GP; 1000 GP CONSOTIUM 2015). We assumed that intergenic regions are evolving neutrally, but if they are in fact evolving under moderate levels of selective constraint, that would only strengthen our results.

4.1. DATA SELECTION

Initially, 5,163,554 nucleotides, non-uniformly distributed along 254 piRNA-generating clusters with the length ranging between 5.5 Kb and 1.2 Mb were considered for this analysis. A total of 162,893 segregating sites were detected in this dataset. For this analysis, we made sure to remove all fragments that overlapped with coding genes and pseudogenes to avoid any spurious signatures of selective constraint. After this filtering step, the remaining dataset consisted of 145 clusters with the overall length of 2,239,144 bp and 48,002 polymorphic sites. To accomplish the comparative analyses of clusters with selectively neutral genomic fragments, 145 intergenic regions were obtained with an overall length of 2,769,292 bp and 134,136 segregating sites in total. The diversity statistics of 2261 human samples were assessed for the entire dataset of each chromosome and separately for five super-populations defined in the 1000 GP (1000 GP CONSOTIUM 2015). To properly conduct all calculations involving divergence we had to exclude chromosomes 13, 17 and 19 from the analysis, as it was not possible to obtain feasible numbers of divergent sites in their intergenic regions probably because of the poor alignment with the chimpanzee genome.

4.2. NUCLEOTIDE DIVERSITY OF PIRNA CLUSTERS AND INTERGENIC REGIONS

First of all, we wanted to know if the difference in the number of segregating sites between piRNA cluster and intergenic regions for the different populations, as well as the whole dataset, was significant using χ^2 statistics. Results are summarized in **Table 1**. The resulting p-values were highly significant in all cases, indicating that the genetic variation at piRNA cluster regions is much smaller than expected under neutrality, which argues previous studies stating that the sequences of the piRNAs might not be functionally important (GIRARD *et al.* 2006).

Table1: Test for the difference in number of segregating sites between piRNA clusters and intergenic regions.

	AFR	AMR	EAS	EUR	SAS	Overall
χ^2	17635.432	8581.359	14907.057	12804.936	13352.603	26376.17
p-value	<E-22	<E-22	<E-22	<E-22	<E-22	<E-22
S_1	34,020	35,766	30,732	28,184	28,196	48,002
S_2	94,295	78,686	83,343	75,210	75,900	135,256

Overall number of studied nucleotides was 2,239,144 and 2,769,292 for piRNA clusters and intergenic regions, respectively. AFR- African population; AMR- American population; EAS- East Asian population; EUR- European population; SAS- South Asian population.
S_1 –number of segregating sites in piRNA clusters.
S_2 – number of segregating sites in intergenic regions.

Nucleotide variation estimates per chromosome for piRNA clusters and corresponding intergenic regions are shown in **Figure 8** and in the box-plots of **Figure 9 (a,b,c)** for the whole dataset, and in **Supplementary Figures F1, F3-F5** for each population apart. Almost the same patterns were observed in the overall dataset, as well as for the five populations separately, with the values of nucleotide diversity and divergence slightly higher for intergenic

regions in all cases. In the **Table 2** the minimum, maximum and average values of all parameters of genetic diversity and divergence for piRNA clusters and intergenic regions in each population are shown (**Supplementary Table S1** contains values for the whole dataset), with overall average divergence (D) equal to 0.013101 and 0.017885; nucleotide diversity (π) equal to 0.001135 and 0.001276; expected heterozygosity (Θ) equal to 0.004994 and 0.007112; for piRNA clusters and intergenic regions respectively. The diversity and divergence in intergenic regions were especially high in chromosome 8 (π=0,00453; D=0,05634; Θ=0,02159), chromosome 3 also showed higher values of nucleotide diversity (π=0,00295; Θ=0,02301). The diversity levels in piRNA clusters visually correlated with that of intergenic regions along the chromosmes with several exceptions in chromosomes 12, 13, 17 and 19 (**Figure 8; Supplemetary Figure F1**), whereas in divergence levels such correlation was not observed.

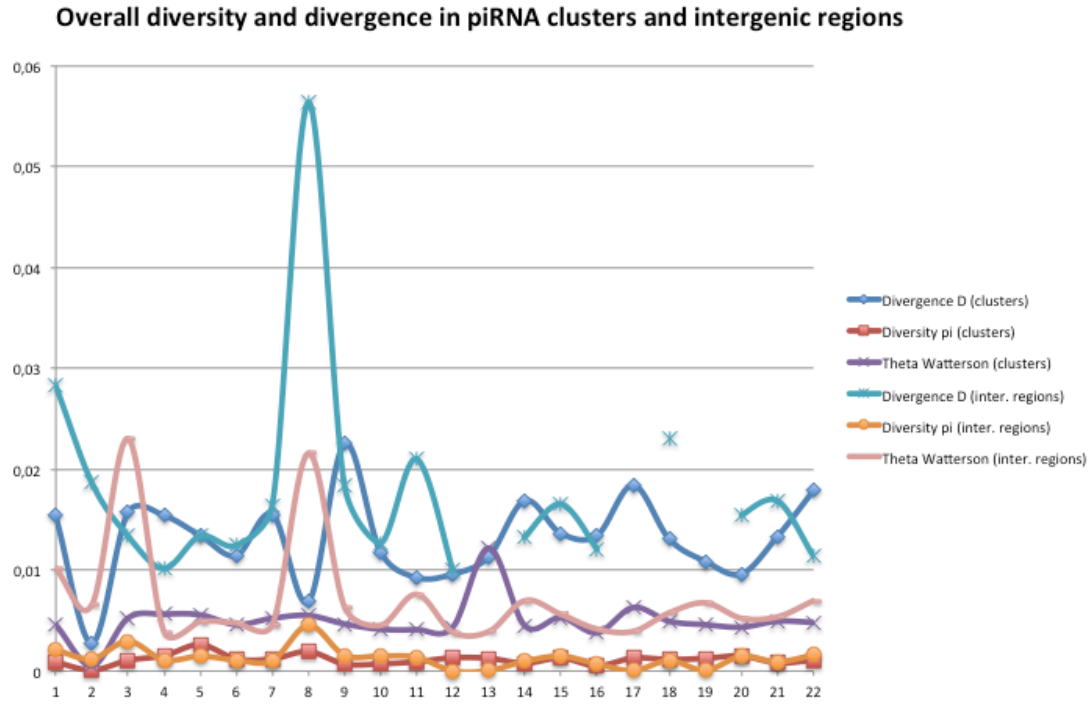


Figure 8: Overall diversity (π and $\Theta_{Watterson}$) and divergence (*D*) in piRNA clusters and intergenic regions per chromosome.

These differences in nucleotide variability and divergence between piRNA clusters and intergenic regions may be indicative that piRNAs are evolving under greater selective constraint compared with intergenic regions in all populations. Below we explain in more detail how these differences may influence the selective patterns.

One way to see the main selective forces acting on the particular genomic regions is to compare its observed heterozygosity (usually denoted as π (*pi*)) with its heterozygosity expected in the condition of neutral variation (usually called Θ (*theta*) of Watterson). When values of observed heterozygosity are greater than those expected, it means that heterozygotes have an advantage in the population and balancing selection is mainly maintaining them. Otherwise, when the values of theta are higher over pi, it points to negative selection acting on deleterious alleles segregating in the population. In our study, comparing *theta* of piRNA regions and intergenic regions with their respective values of *pi*, we found that *theta* in both cases was much higher than *pi* (**Figure 10**). Overall average *theta* was equal to 0.00499 for piRNA clusters and 0.00711 for intergenic regions, and average *pi* was equal to 0.00114 and 0.00128, for the clusters and intergenic regions respectively (**Table 2**). The same trend was found in all populations (**Supplementary Figure F6; Supplementary table S1**) and along nearly all

piRNA clusters (**Figure 11; Supplementary Figure F2**), indicating an excess of slightly deleterious alleles segregating in all populations and weak purifying selection acting on them, which could be considered as background selection in our case.

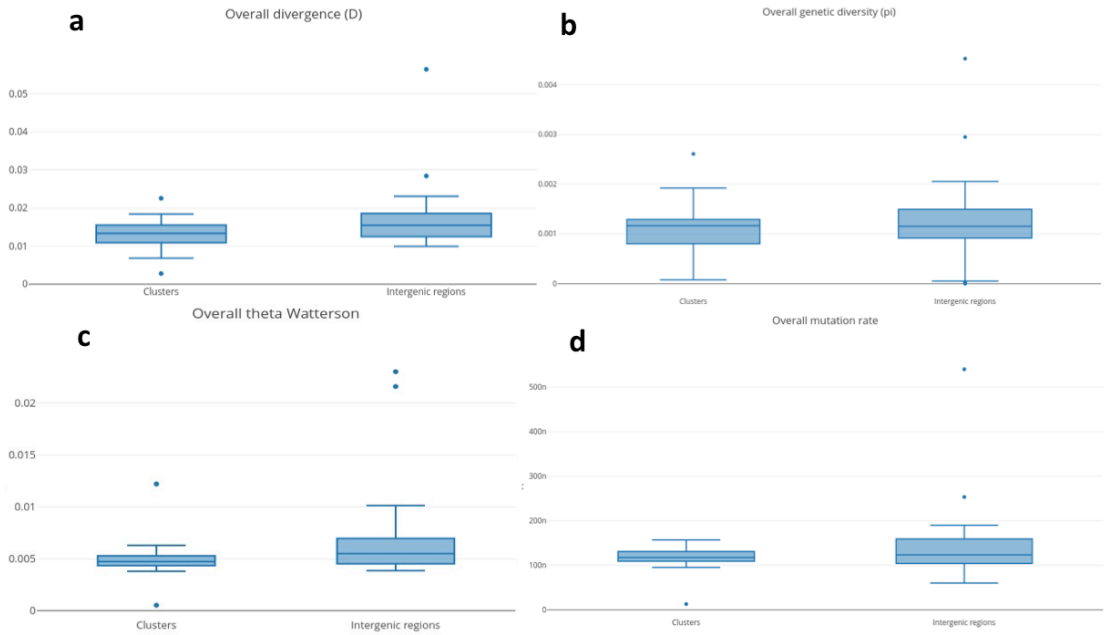


Figure 9: Box-plots of population genetic parameters from the overall dataset of piRNA clusters versus intergenic regions: a) divergence (D); b) nucleotide diversity (π); c) expected heterozygosity ($\Theta_{Watterson}$); d) mutation rate.

Table 2: Minimum, maximum and average values of genetic parameters for piRNA clusters and intergenic regions.

		piRNA clusters			Intergenic regions		
		MIN	MAX	Average	MIN	MAX	Average
Divergence, D	AFR	0,000115	0,022983	0,013522	0,010172	0,058998	0,018684
	AMR	0,002806	0,029758	0,01438	0,010341	0,059906	0,018434
	EAS	0,002797	0,022963	0,013718	0,01031	0,05977	0,018365
	EUR	0,002812	0,023117	0,013891	0,010425	0,060155	0,018587
	SAS	0,002814	0,023201	0,013955	0,010402	0,060541	0,018639
	Overall	0,002778	0,022534	0,013101	0,00996	0,05634	0,017885
Diversity, π	AFR	0,000115	0,002577	0,001357	0,000500	0,005007	0,003783
	AMR	9,577E-05	0,002558	0,001166	0,000221	0,004693	0,001441
	EAS	8,239E-05	0,002596	0,001099	0,000139	0,004112	0,001249
	EUR	8,357E-05	0,002668	0,001067	0,000143	0,003965	0,001269
	SAS	8,454E-05	0,002790	0,00113	0,000176	0,004424	0,001364
	Overall	0,000079	0,002610	0,001135	0	0,004525	0,001276
$\Theta_{Watterson}$	AFR	0,000500	0,005007	0,003783	0,002926	0,0177866	0,005966
	AMR	0,000494	0,007326	0,003618	0,002646	0,016237	0,005503
	EAS	0,000473	0,004788	0,003450	0,002637	0,015828	0,005431
	EUR	0,000456	0,004233	0,003097	0,00225	0,01487	0,005032
	SAS	0,000472	0,00436	0,003223	0,002475	0,01495	0,005169
	Overall	0,000519	0,012198	0,004994	0,00386	0,023012	0,007112
Mutation rate	AFR	2,352E-07	1,252E-07	9,369E-08	7,314E-08	2,445E-06	2,352E-07
	AMR	1,236E-08	1,832E-07	8,963E-08	6,614E-08	2,305E-06	2,194E-07
	EAS	1,236E-08	1,197E-07	8,547E-08	6,594E-08	2,248E-06	2,154E-07
	EUR	1,139E-08	1,058E-07	7,671E-08	5,625E-08	2,305E-06	2,086E-07
	SAS	1,181E-08	1,09E-07	7,982E-08	6,188E-08	2,174E-06	2,067E-07
	Overall	1,296E-08	1,573E-07	1,157E-07	6,009E-08	5,396E-07	1,492E-07

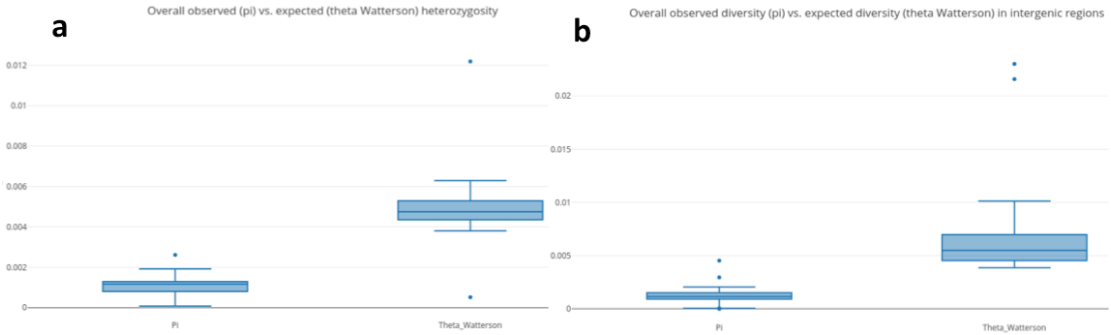


Figure 10: Box-plots of observed (π) versus expected ($\Theta_{Watterson}$) heterozygosity for a) piRNA clusters and b) intergenic regions from the whole dataset.

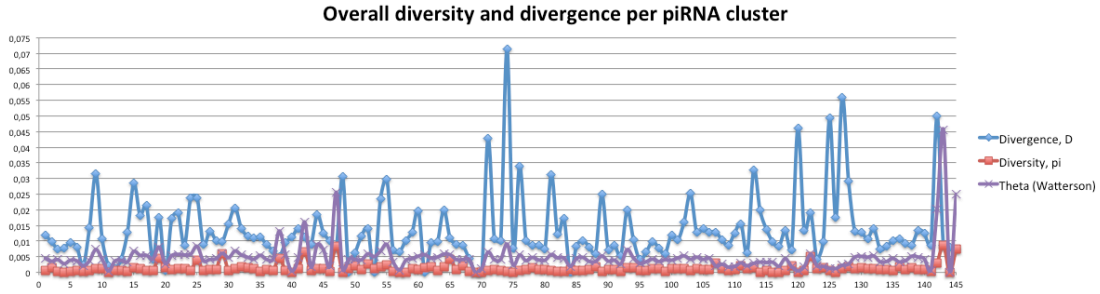


Figure 11: Overall diversity (π and $\Theta_{Watterson}$) and divergence (D) in piRNA clusters.

As for the mutation rate, its average values were higher for intergenic regions in all populations (**Table 2**; **Figure 9d; Supplemetary Figure F8**), especially because of its extraordinarily high level in chromosomes 1, 2, 8, 11 and 19 (**Figure 12, Supplemetary Figures F7; Supplementary Table S1**). Whereas on chromosome 17, the mutation rate was higher in piRNA clusters. The rest of the chromosomes have shown the same mutation rate for clusters and integenic regions.

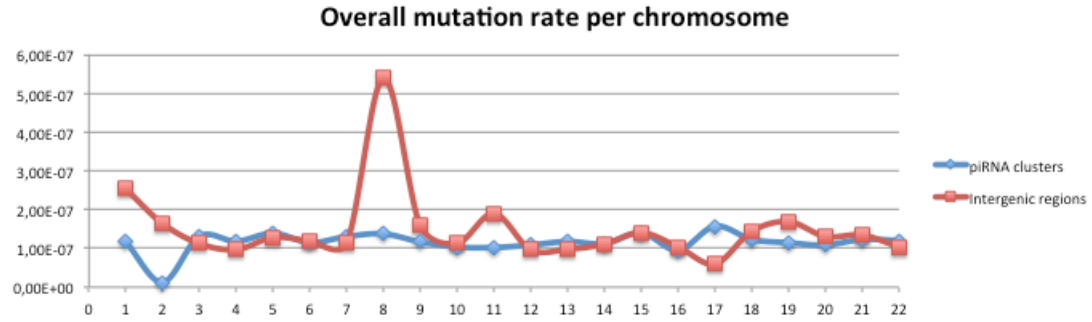


Figure 12: Overall mutation rate of piRNA clusters and intergenic regions per chromosome.

4.3. GENETIC DIFFERENTIATION BETWEEN POPULATIONS WITHIN piRNA CLUSTERS

χ^2 tests were applied to find differences in the number of segregating sites between each pair of populations (**Table 3**), which resulted in very significant p-values in all cases except the comparison European versus South Asian populations, with very similar numbers of polymorphic sites (**Table 3**). The difference in the nucleotide diversity can be seen clearly in the boxplots of **Figure 12b**, where the African population is the most genetically diverse with highest average π equal to 0.00136 versus 0.00107-0.00117 of the remaining populations (**Table 2**). Nevertheless the level of divergence with chimpanzee in piRNA clusters did not differ between populations, attaining similar average values between 0.0135 in African and

0.0144 in American populations (**Figure 12a, Table 2**). At the cluster level, the difference in divergence between populations was not observed, as shown in **Supplementary Figure F9**). The differences between populations at the level of each single cluster are also shown on the same figure.

Table 3: Test for the difference in number of segregating sites of piRNA clusters in each pair of population

	AFR/ AMR	AFR/ EAS	AFR/ EUR	AFR/ SAS	AMR/ EAS	AMR/ EUR	AMR/ SAS	EAS/ EUR	EAS/ SAS	EUR/ SAS
χ^2	44,375	169,408	555,248	552,862	724,657	911,955	1277,518	111,665	110,593	0,003
p-value	2,7E-11	10E-39	9,1E-123	3E-122	1,3E-159	2,5E-200	8,7E-280	4,2E-26	7,3E-26	0,959
S_1	34020	34020	34020	34020	35766	35766	35766	30732	30732	28184
S_2	35766	30732	28184	28196	30732	28184	28196	28184	28196	28196

Overall number of studied nucleotides was 2239144 for piRNA clusters in all populations.
S_1 –number of segregating sites in the population before the slash.
S_2 – number of segregating sites in the population after the slash.

As it is shown with the box-plots on the **Figure 12d**, the highest amutation rate was attained by African population (9.369E-08 on average), followed by East-Asian and American populations, the smallest mutation rate (average=7.671E-08) was observed in European population. The observed patterns are consistent with data that show a much higher rate of transposon insertions in African compared with non-African populations (EWING and KAZAZIAN 2010). At the cluster level, the highest mutation rate was observed in the ninth cluster of chromosome 9 (coordinates: 84523917-84548943) attaining the value 1.754E-06 in the European population and 1.874E-06 in American population (**Supplementary Figure F10; Supplementary Table S2**). Also an extraordinary high mutation rate was found in the first piRNA cluster on chromosome 5 (coordinates: 42989118-42997891) in the African population (3,95543E-07; **Supplementary Figure F10**; **Supplementary Table S2**).

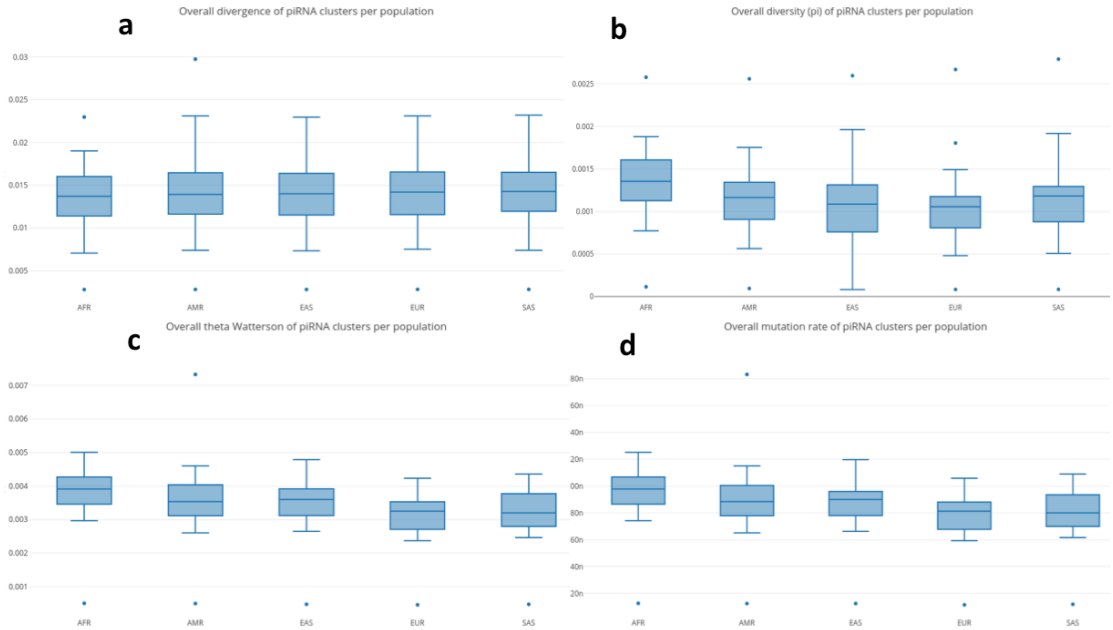


Figure 12: Box-plots of population genetic parameters of piRNA clusters for five populations: a) divergence (D); b) nucleotide diversity (π); c) expected heterozygosity ($\Theta_{Watterson}$); d) mutation rate.

The genetic differentiation of piRNA clusters between populations was also estimated with the fixation index F_{ST}, which is the average number of differences in the sequences of a population in relation to the average number of differences between two populations (HUDSON *et al.* 1992). At the chromosomal level, F_{ST} values were positive in all cases, suggesting that the

genetic difference in clusters per chromosome between each pair of populations is larger than the difference between individuals within the same population (**Figure 13c**; **Supplementary Table S4; Supplementary Figure 10**). At the level of separate clusters, F_{ST} values very close to zero or even negative were observed in several cases (**Figure 13d**; **Supplementary Table S3**). The most differentiated population was the African one in the pairwise comparisons with the other populations (average F_{ST} ranged between 0.1084 and 0.1374; **Table 4**; **Figure 13 a, b**), as well as versus all populations joined together (F_{ST} = 0.123 on average). The East Asian population also showed a high differentiation especially with the African (average F_{ST}= 0.1374) and European populations (average F_{ST} = 0.0829). The clusters of American and European populations were found to be the least differentiated (average F_{ST}= 0.0227; **Table 4**; **Figure 13 a, b**). Several peaks of differentiation were also observed at the cluster level in clusters 23 (chromosome 4, coordinates: 41981218-41989614), 78 (chromosome 9, coordinates: 93340401-93351022) and 108 (chromosome 15, coordinates: 24269210-24279095), mostly between the African population and the rest, ranging between 0.3151 and 0.6606 (**Figure 13d**; **Supplemetary Figure F10** and **Supplemetary Table S3**). The highest F_{ST} values were found in cluster 108 for differentiation between East Asian/European (F_{ST}=0.3785) and East Asian/South Asian (F_{ST}=0.3335) populations (**Supplemetary Figure 10**; **Supplementary Table S3**). So, piRNA generating clusters from the African population show the greatest differentiation compared to the rest of the populations, which in accordance with their highest nucleotide variability.

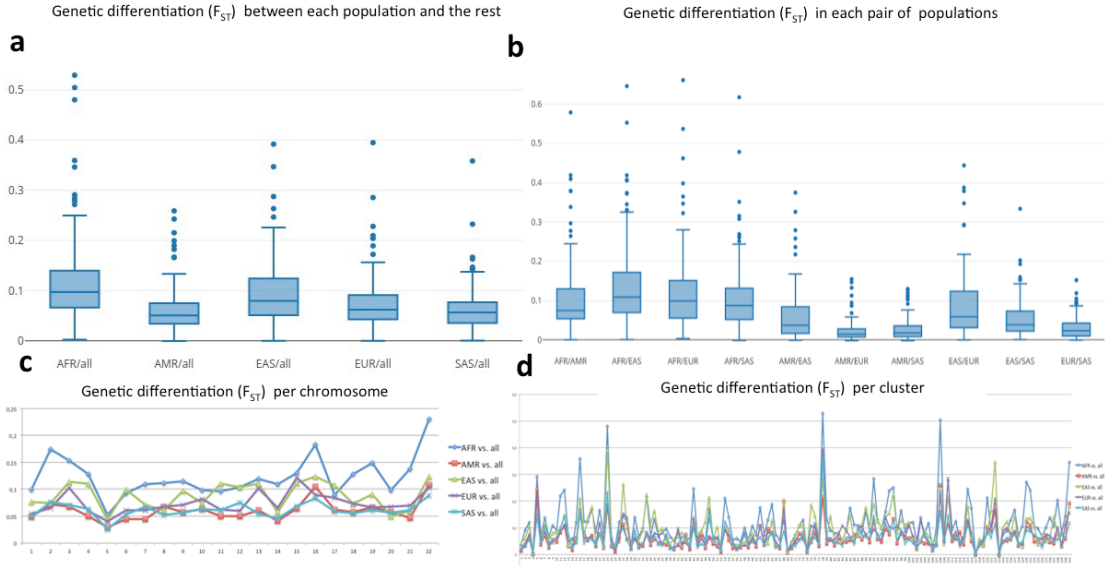


Figure 13: Genetic differentiation between populations: a) box-plots for each population vs. the rest of them; b) box-plots for each pair of populations; c) F_{ST} per chromosome for each population versus the remaining four populations; d) F_{ST} per cluster for each population versus the remaining four populations.

Table 4: Minimum, maximum and average value of differentiation index FST per cluster in each pair of populations and each single population versus the rest of them.

	AFR/ AMR	**AFR/ EAS**	**AFR/ EUR**	**AFR/ SAS**	**AMR/ EAS**	**AMR/ EUR**	**AMR/ SAS**	**EAS/ EUR**	**EAS/ SAS**	**EUR/ SAS**
MIN	0,0275	0,0621	0,0648	0,0448	0,0125	0,0061	0,0049	0,016	0,0144	0,0035
MAX	0,2581	0,2675	0,2376	0,1903	0,1011	0,0603	0,0644	0,1835	0,1182	0,0785
AV.	0,1128	0,1374	0,1330	0,1084	0,0537	0,0227	0,0286	0,0829	0,058	0,0323
	AFR/ all	**AMR/ all**	**EAS/ all**	**EUR/ all**	**SAS/ all**					
MIN	0,0522	0,0289	0,0435	0,0404	0,0261					
MAX	0,2296	0,1080	0,1235	0,1217	0,0887					
AV.	0,1231	0,0599	0,0883	0,0741	0,0612					

4.4. TESTS FOR NEUTRALITY AND PATTERNS OF SELECTION

Several tests were conducted to measure deviation from neutrality in the piRNA clusters. The values of Tajima's *D* (TAJIMA 1989) were negative at the chromosomal level in all populations and in almost all clusters, suggesting an excess of low frequency polymorphism, which may be indicative of purifying selection maintaining deleterious alleles at low frequencies. However, the same signature could be due to a recent population expansion after the Out of Africa bottleneck (STRINGER 2003), as supposedly neutral intergenic regions also showed negative D values in all populations and chromosomes (**Figures 14, 15**; **Supplementary Figure 11**; **Supplementary Tables S1, S2**). Despite this general trend toward an excess of rare alleles, several clusters have shown positive values of D in the five populations: the third cluster of chromosome 3 (coordinates: 124462898-124469870) with extremely high D values in all populations (Tajima's D ranging from 10.13 in Africa to 16.79 in America), second cluster of chromosome 5 (coordinates: 178250015-178259870; Tajima's D: from 0.42 in Africa to 1.7 in Europe); forth cluster on chromosome 9 (coordinates: 44017295-44031054; Tajima's D: from 0.5 in America to 5.13 in East Asia). In these cases, balancing selection with heterozygote advantage might be maintaining nucleotide variability.

.

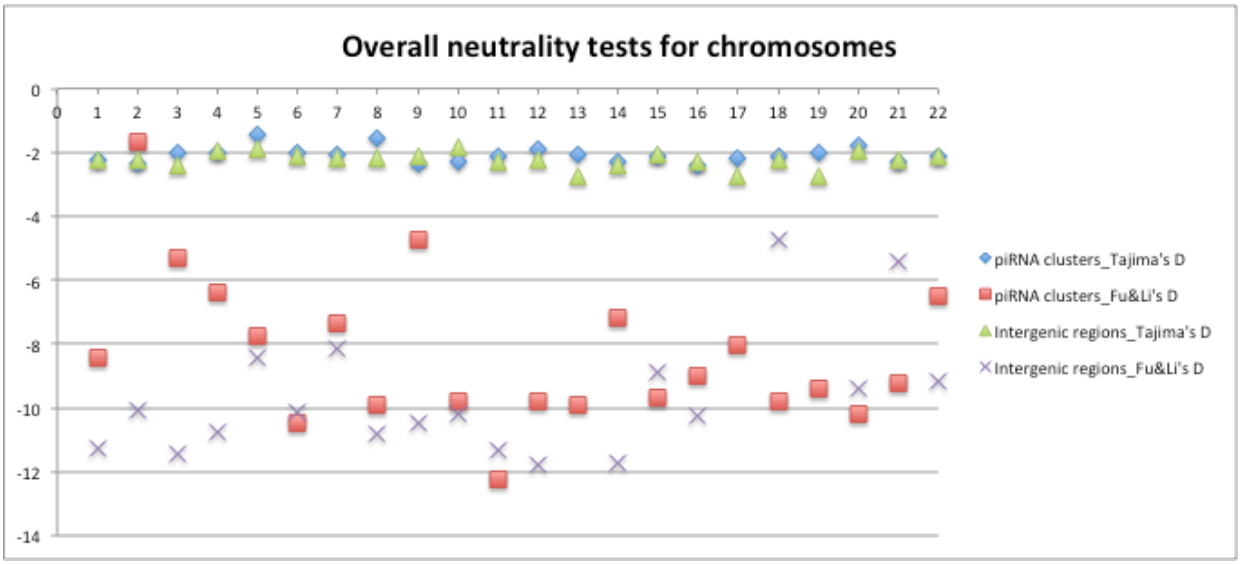


Figure 14: Tajima's D and Fu&Li's D tests for neutrality in piRNA clusters and intergenic regions, per chromosome.

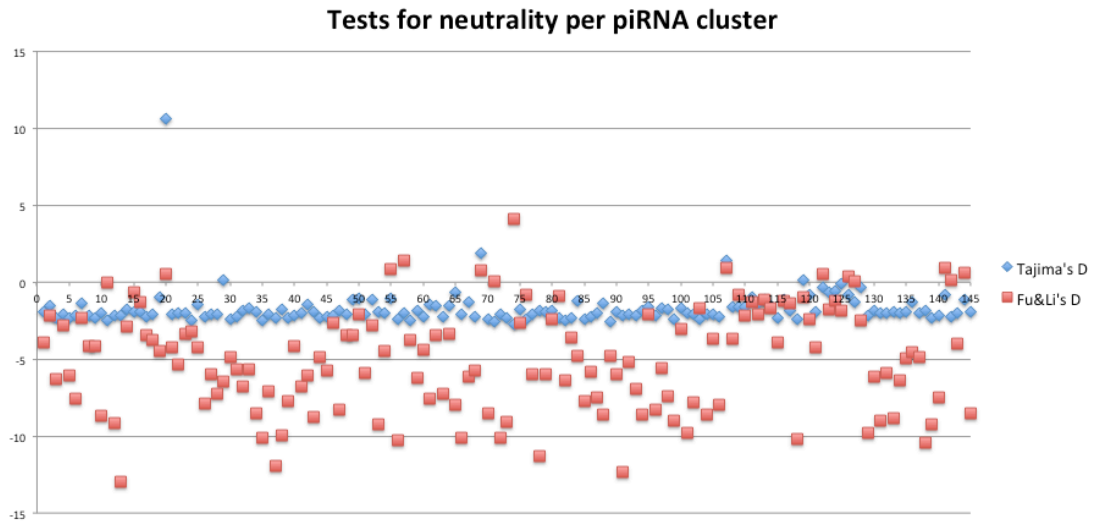


Figure 15: Tajima's D and Fu&Li's D tests for neutrality in each piRNA cluster.

As Tajima's D, Fu&Li' *D* (FU AND LI 1993) is also a test of selection whose values might also reflect demographic changes in the population. Both tests are based on the fact that, under the neutral model, estimates of the number of segregating sites and of the average number of nucleotide differences are correlated. If the value of D is too large or too small, the neutral null hypothesis is rejected. Unlike Tajima's test, Fu&Li's D is directly based on the coalescent, requiring data both from intraspecific polymorphism and from an outgroup species (chimpanzee in our case). The results of both tests, based on the allelic variation, may not clearly distinguish between selection and demographic events (bottleneck, population subdivision, migration), but this problem only applies to the analysis of a single locus: demographic changes affect all loci whereas selection is expected to be locus-specific, showing distinguishable footprints when multiple loci are analyzed.

Fu&Li's test also showed the prevalence of negative values at the chromosomal as well as the cluster levels (**Figures 14**, **15**; **Supplementary Figure F11**; **Supplementary Tables S1** and **S2**). Nevertheless, some chromosomes showed positive values of this statistics in piRNA clusters and intergenic regions, although some of them were near to zero. Chromosomes 2, 3, 4, 5 and 9 attained positive D values in all populations (**Figure 14; Supplementary Figure 11; Supplementary Table S1**). At the cluster level, a considerable fraction of the clusters had positive but very low values of Fu&Li's D. The percentages of positive D values slightly differed among populations, ranging between 32.17 (Africa) and 41.38 (America). This differential pattern reject the hypothesis of demographic changes as a major reason of an excess of rare alleles and confirms that weak purifying selection is acting on piRNA clusters, while 30-40% of them are rather under balancing selection, although the majority of them have D values of Fu&Li test not very different from zero, so the "null" hypothesis of neutrality can not be rejected for them.

The MKT is less sensitive to demographic effects than other similar selection tests (NIELSEN 2001). Under the neutral model, any demographic factor affecting variability would be expected to act equally on neutral and non-neutral differences. However, when some of the nonsynonymous mutations are slightly deleterious, then P_n/P_s itself is sensitive to demographic effects. For example, in the case of an increase in the population size during the expansion process, slightly deleterious mutations could become fixed by genetic drift and P_n/P_s could become smaller than D_n/D_s (EYRE-WALKER 2002). When interpreting results and distinguishing between different types of selection, it is important to note that the hypothesis of positive selection predicts more fixed replacement differences than predicted by the neutral mutation random drift hypothesis, since these are the changes that may have fitness effects. The

hypothesis of balancing selection makes a different prediction: a preponderance of replacement polymorphisms maintained by selection. Negative selection is much more frequent than positive selection, therefore, in many cases positive selection is masked and difficult to detect. Another assumption of the test that must be taken into account is that it considers that selective constraints are constant over time. If there has been a recent change in selective constraints, either by a relaxation of selection or increased selective constraint, the proportions D_n/D_s and P_n/P_s would not have been expected to be equal. In addition, it is also assumed that the sites being compared are closely linked, and the null hypothesis can be wrongly rejected in cases of intermediate levels of recombination (ANDOLFATTO 2008).

To properly conduct the MKT, we had to exclude chromosomes 13, 17 and 19 from the analysis, as it was not possible to obtain faithful numbers of divergent sites in their intergenic regions probably because of the poor alignment with the chimpanzee genome. Main results of the MKT at the level of single clusters per population and for the overall dataset are summarized in **Table 5**. The proportion of significant values of *alpha* differed depending on the populations, attaining values around 11.9% in the European population as a minimum, and a maximal percentage of 19.7% in the African population. Joining all populations together, the number of significant *alpha* values increased until 30.7% (**Table 5, Figure 16**). Most of these values were negative, due to an excess of non-neutral polymorphic sites, prevented from fixation by purifying or balancing selection in the majority of the clusters. Nevertheless, a substantial fraction of the significant *alpha* values were positive, with a maximum of 46.7% in the European population and a minimum of 40% in the African population, which is in accordance with the result of Lukic and Chen (2011). They previously found the African population as the most selectively constrained at the level of piRNA sequences (**Table 5**; **Supplementary Figure F12**, **Supplementary Table S6**). When the whole chromosomes were considered in the overall dataset, only five of them (chromosomes 3, 5, 6, 10 and 21) showed non-significant values of *alpha*. In the other 15 chromosomes, half of the significant values were positive (**Supplementary Table S5**).

Table 5: Proportion of significant and positive alpha along all cluster data set per population.

	% of significant α	% of positive α
AFR	19,7	40
AMR	15	42,1
EAS	15,7	45
EUR	11,9	46,7
SAS	13,4	41,2
Overall	30,7	46,15

The percentages of positive alpha was calculated in relation to only significant alpha ($p<0.05$).

To compare proportions of divergent and polymorphic sites (Div/Pol) in piRNA clusters versus intergenic regions, we summarized the calculations for each population and joint dataset in **Table 6**. In all cases the proportion Div/Pol attained higher values in the piRNA clusters compared to the intergenic regions due to a relative excess of divergent sites in the clusters, indicating higher fixation rate of advantageous alleles, although this results were not significant when we made per chromosome comparison for each population.

This result makes an interesting supplement to the previous findings of LUKIC and CHEN (2011) in respect to conservation and substitution rates in piRNAs, based on the number of substitutions between humans and chimpanzees. They found slightly higher rates of conservation in piRNAs compared to flanking regions in primates that were not significant. Furthermore, they observed no significant difference in the substitution rates between humans and chimpanzees, which was consistent with the previous results in rodents, *Drosophila* and nematodes.

Despite of the prevalence of purifying selective forces, directional selection also takes place in a substantial fraction of piRNA clusters, allowing the occurrence of an evolutionary "arm race" between these functional genomic regions and transposable elements. Negative selection is supposed to be the main force driving the evolution of piRNA clusters at the nucleotide level in all human populations. This result is consistent with the previous study at the level of copy number variants by GOULD *et al.* (2012) and confirmed the notion that, although piRNAs are not well conserved between species, they might be under detectable selective constraint at a shorter time scale if they are rapidly evolving regions.

Table 6: The proportion of divergent to polymorphic sites (Div/Pol) in piRNA clusters and intergenic regions per population.

	piRNA clusters	Intergenic regions
AFR	0,54363	0,46573
AMR	0,64047	0,58757
EAS	0,6059	0,55177
EUR	0,66348	0,59851
SAS	0,66794	0,61385
Overall	0,38066	0,32189

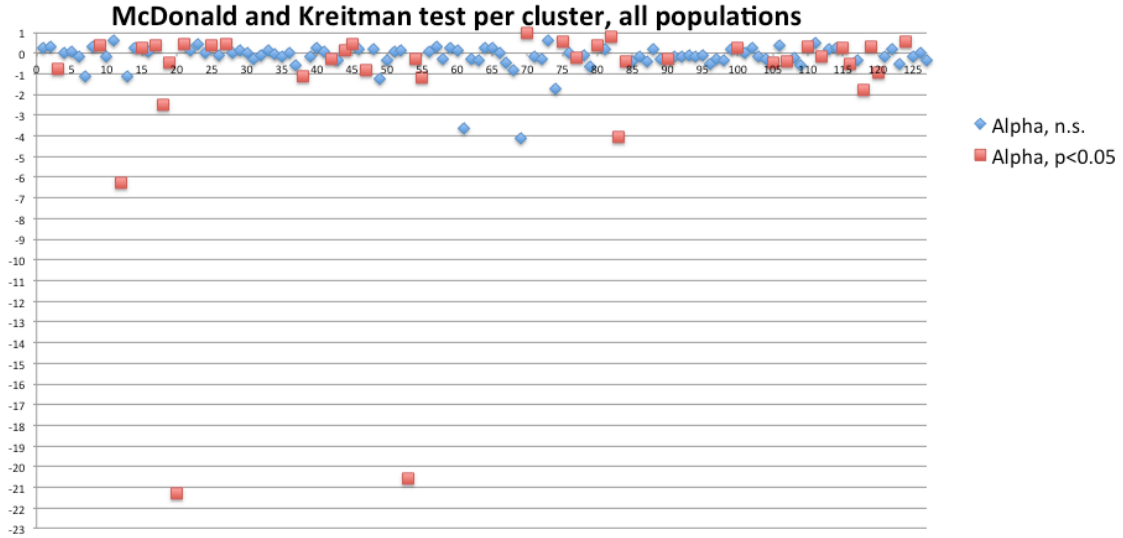


Figure 17: Alpha of MKT for each piRNA cluster in joint dataset of all populations.

Our results represent an important supplement to the results of LUKIC and CHEN (2011), whose study of selective regimes acting on piRNA sequences was based on the derived allele frequency distribution. Their statistical tests showed that piRNAs are indeed evolving under significantly greater selective constraint compared to intergenic regions, but only in the African populations. In seven non-African populations they observed a trend for piRNA to be under greater selective constraint than intergenic regions, but the results were not statistically significant, making them to conclude that there is strong statistical support for selective constraint on piRNAs in the African populations and only weak evidence for selective constraint in non-African populations (LUKIC and CHEN 2011). In our study we used the tests for neutrality and selection, such as the MKT and Fu&Li, which, unlike the tests based only on intraspecies polymorphism, such SFS used in the previous study, allowed us to compare the intraspecific variation with the interspecific divergence and observe selective constraint in all studied populations.

Based on an increase in the copy number variation across the evolution of piRNA generating loci, ASSIS AND KONDRASHOV (2009) proposed a hypothesis of strong positive selection for these copies in the mammalian genomes. Interestingly, the recent study of selection

fingerprint on copy number variation in piRNA generating loci in two human populations (Yoruban and European), designed differently from ours, came to a similar conclusion about mainly negative selection acting on these loci (GOULD *et al.* 2012), arguing the hypothesis of ASSIS and KONDRASHOV (2009). Contrarily to our work, they studied copy number variation in the II highly repetitive type of piRNAs, mainly found in the euchromatin (while we studied the action of selection in the pachytene piRNA clusters, which are not represented by repetitive sequences, being mainly found in the heterochromatin).

To find the fingerprint of selection, they also used a modified version of the MKT, by comparing piRNA-generating loci with intergenic regions in the human and chimpanzee genomes. They found that the divergence of piRNA-generating loci (7.6%) was higher than that for intergenic regions (6%), consistent with a higher CNV mutation rate in piRNA-generating loci. Thus, despite the strong enrichment of CNVs in piRNA-generating loci, there was a large excess of polymorphism in piRNA-generating loci compared to divergence, assuming that intergenic regions as a whole are evolving neutrally with respect to CNVs. This pattern was interpreted as an indication of negative selection on piRNA-generating loci at the CNV level.

Altogether, these results point out that global negative selection is the primary selective force acting on piRNA-generating loci across populations, not balancing selection caused by local adaptation. Negative selection would cause an excess of mildly deleterious polymorphisms and a depletion of divergence, which is what was observed in all these piRNA data.

There is still a long way to go in understanding the basic molecular biology of piRNAs and detailed quantitative models of their evolution are not easy to formulate at the present time. Nonetheless, we hope that by elucidating the main selective forces acting on piRNA clusters in human genome we can help to further elucidate the evolutionary mechanisms of the piRNA system.

5 | CONCLUSIONS

1. A pipeline for the data retrieval from the 1000 Genomes Project, filtering, processing and analysis of piRNA clusters and intergenic regions was developed.

2. A strong difference in the levels of polymorphism was found between piRNA clusters and intergenic regions, the latter being more diverse. Some differences were also found in the mutation rates and divergence from chimpanzee, with higher values for intergenic regions.

3. Heterozygosity expected under neutrality (Θ of Watterson) was much higher than observed heterozygosity (Nei's π), suggesting the action of purifying selection on weakly deleterious alleles segregating in all populations at low frequencies. This observation was confirmed by Tajima's D and Fu&Li's D tests, which in most of the cases attained negative values due to an excess of rare alleles.

4. The values of alpha in the McDonald and Kreitman test were negative in most of the cases, being significant in 12-20% of piRNA clusters depending on the super-population. Nevertheless, in 40-47% of significant resuls a positive alpha was detected, suggesting recurrent directional selection fixing new advantageous alleles.

5. The African super-population was the most differentiated one, attaining highest values of fixation index F_{ST} in all pairwise comparisons, as well as versus all remaining populations taken together. It showed higher selective constraint (60% of significant values have negative alphas), which is in accordance with the previous findings.

BIBLIOGRAPHY

ARAVIN A, GAIDATZIS D, PFEFFER S, LAGOS-QUINTANA M, LANDGRAF P, IOVINO N, MORRIS P, BROWNSTEIN MJ, KURAMOCHI-MIYAGWA S, NAKANO T, CHIEN M, RUSSO JJ, JU J, SHERIDAN R, SANDER C, ZAVOLAN M, TUSCHL T. 2006. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* 442:203–7.

ARAVIN AA, NAUMOVA NM, TULIN AV, VAGIN VV, ROZOVSKY YM, GVOZDEV VA. 2001. Double-stranded RNA–mediated silencing of genomic tandem repeats and transposable elements in the *D. melanogaster* germline. *Curr. Biol.* 11:1017–27.

a) ARAVIN AA, HANNON GJ, BRENNECKE J. 2007. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* 318:761–764.

b) ARAVIN A, SACHIDANANDAM R, GIRARD A, FEJES-TOTH K, HANNON GJ. 2007. Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* 316:744–7.

ARAVIN A, SACHIDANANDAM R, BOURC'HIS D, SCHAEFER C, PEZIC D, TOTH KF, BESTOR T, HANNON GJ. 2008. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Mol Cell* 31:785–99.

ASSIS R, KONDRASHOV AS. 2009. Rapid repetitive element-mediated expansion of piRNA clusters in mammalian evolution. *PNAS USA* 106(17): 7079–82.

BAKEWELL MA, SHI P, ZHANG J. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. *PNAS USA* 104: 7489-7494.

BENAZZO A, PANZIERA A, BERTORELLE G. 2015. 4P: fast compution of population genetics statistics from large DNA polymorphism panels. *Ecol Evol.* 5(1): 172-5.

BETEL D, SHERIDAN R, MARKS D, SANDER C. 2007. Computational analysis of mouse piRNA sequence and biogenesis. *PLoS Comput Biol* 3:e222.

BLUMENSTIEL JP, ERWIN AA, HEMMER LW. 2016. What Drives Positive Selection in the *Drosophila* piRNA Machinery? The Genomic Autoimmunity Hypothesis. *The Yale Journal of Biology and Medicine*. 89(4):499-512.

BLUMENSTIEL J, HARTL D. 2005. Evidence for maternally transmitted small interfering RNA in the repression of transposition in *Drosophila virilis*. *Proc Natl Acad Sci* 102:15965–70.

BOURLAT SJ, NIELSEN C, ECONOMOU AD, TELFORD MJ. 2008. Testing the new animal phylogeny: a phylum level molecular analysis of the animal kingdom. *Mol Phylogenet Evol.* 49(1):23-31.

BRENNECKE J, ARAVIN AA, STARK A, DUS M, KELLIS M, SACHIDANANDAM R, HANNON GJ. 2007. Discrete small RNA–generating loci as master regulators of transposon activity in *Drosophila*. *Cell* 128:1089–103.

BRENNECKE J, MALONE C, ARAVIN A, STARK A, SACHIDANANDAM R, HANNON GJ. 2008. An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science* **322**:1387–92.

CARMELL MA. 2002. The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. *Genes Dev.* **16**: 2733-42.

CARMELL MA, GIRARD A, VAN DE KANT HJ, BOURC'HIS D, BESTOR TH, BESTOR TH, DE ROOIJ DG, HANNON GJ. 2007. MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev. Cell* **12**:503–14.

CASTILLO D, MELL J, BOX K, BLUMENSTIEL JP. 2011. Molecular evolution under increasing transposable element burden in *Drosophila*: a speed limit on the evolutionary arms race. *BMC Evol Biol* **11**:258.

CERUTTI H, CASAS-MOLLANO J. 2006. On the origin and functions of RNA-mediated silencing: from protists to man. *Curr Genet* **50**:81–99.

GHILDIYAL M, ZAMORE PD. 2009. Small silencing RNAs: an expanding universe. *Nat. Rev. Genet.* **10**:94– 108

CHEN K, RAJEWSKY N. 2007. The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet* **8**: 93–103.

COX DN, CHAO A, BAKER J, CHANG L, QIAO D, LIN H. 1998. A novel class of evolutionarily conserved genes defined by *piwi* are essential for stem cell self-renewal. *Genes Dev.* **12**:3715–27.

CREAMER KM, PARTRIDGE JF. 2011. RITS-connecting transcription, RNA interference, and heterochromatin assembly in fission yeast. *Wiley Interdiscip. Rev. RNA* **2**:632-646.

DANECEK, P., AUTON, A., ABECASIS, G., ALBERS, C.A., BANKS, E., DEPRISTO, M.A., HANDSAKER , R.E., LUNTER, G., MARTH, G.T., SHERRY, S.T., MCVEAN G., DURBIN, R., 1000 GENOMES PROJECT ANALYSIS GROUP. 2011. The variant call format and VCFtools. *Bioinformatics* **27**(15): 2156–8.

DOLGIN E, CHARLESWORTH B, CUTTER A. 2008. Population frequencies of transposable elements in selfing and outcrossing *Caenorhabditis* nematodes. *Genet Res*, Camb **90**:317–29.

EGEA, R., CASILLAS, S., BARBADILLA, A. 2008. Standard & Generalized McDonald and Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Res* **36**: W157-W162.

EWING A, KAZAZIAN H. 2010. Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. *Genome Res* **21**:985–90.

DE FAZIO S, BARTONICEK N, DI GIACOMO M, ABREU-GOODGER C, SANKAR A, FUNAYA C, ANTONY C, MOREIRA PN, ENRIGHT AJ, O'CARROLL D. 2011. The endonuclease activity of MILI fuels piRNA amplification that silences LINE1 elements. *Nature* **480**:259–63.

FRAZER KA, PACHTER L, POLIAKOV A, RUBIN EM, DUBCHAK I. 2004. VISTA: computational tools for comparative genomics. *Nucleic Acids Res* **32** (Web Server issue):W273-9.

FU Y-X, LI W-H. 1993. Statistical tests of neutrality of mutations. *Genetics* **133**: 693-709.

GANGARAJU V, YIN H, WEINER M, WANG J, HUANG XA, LIN H. 2011 *Drosophila* PIWI functions in Hsp90-mediated suppression of phenotypic variation. *Nat Genet* **2**:153–8.

GAZAL, S., SAHBATOU, M., BABRON, M.C., GÉNIN, E., LEUTENEGGER, A.L. 2015. High level of inbreeding in final phase of 1000 Genomes Project. *Sci Rep.* **5**: 17453.

GHILDIYAL M, ZAMORE PD. 2009. Small silencing RNAs: an expanding universe. *Nat. Rev. Genet.* **10**:94–108.

GIRARD A, SACHIDANANDAM R, HANNON GJ, CARMELL MA. 2006. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442**:199–202.

COLINE G, THÉRON E, BRASSET E, VAURY C. 2014. History of the discovery of a master locus producing piRNAs: the *flamenco/COM* locus in *Drosophila melanogaster*. *Frontiers in Genetics* **5**:257.

GOULD DW, LUKIC S, CHEN KC. 2012. Selective constraint on copy number variation in human piwi-interacting RNA Loci. *PloS One* **7**(10): e46611.

GRIFFITHS-JONES S, SAINI HK, VAN DONGEN S, ENRIGHT AJ. 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36**:D154–D158.

GRIMSON A, SRIVASTAVA M, FAHEY B, WOODCROFT BJ, CHIANG HR, KING N, DEGNAN BM, ROKHSAR DS, BARTEL DP. 2008. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* **455**:1193–97.

GRIVNA ST, BEYRET E, WANG Z, LIN H. 2006. A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev.* **20**:1709–14.

GUNAWARDANE LS, SAITO K, NISHIDA KM, MIYOSHI K, KAWAMURA Y, NAGAMI T, SIOMI H, SIOMI MC. 2007. A Slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in *Drosophila*. *Science* **315**:1587–90.

HALIC M, MOAZED D. 2010. Dicer-independent primal RNAs trigger RNAi and heterochromatin formation. *Cell* **140**: 504–16.

HAMILTON AJ, BAULCOMBE DC. 1999. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* **286**: 950-952.

HARROW J, FRANKISH A, GONZALEZ JM, ET AL. 2012. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research* **22**(9):1760-1774.

HERTEL J, LINDEMEYER M, MISSAL K, FRIED C, TANZER A, FLAMM C, HOFACKER IL, STADLER PF. 2006. The expansion of the metazoan microRNA repertoire. *BMC Genomics* **7**:25.

HIRANO T, IWASAKI YW, LIN ZY, IMAMURA M, SEKI NM, SASAKI K, OKANO H, SIOMI MC, SIOMI H. 2014. Small RNA profiling and characterization of piRNA clusters in the adult testes of

the common marmoset, a model primate. *RNA* **20**:1223–37.

HUBISZ M. 2008. Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**:994–997.

ISHIZU H, SIOMI H, SIOMI MC. 2012. Biology of PIWI-interacting RNAs: new insights into biogenesis and function inside and outside of germlines. *Genes Dev.* **26**:2361–73.

IWASAKI YW, SIOMI MC, SIOMI H. 2015. PIWI-Interacting RNA: Its Biogenesis and Functions. *Annu Rev Biochem.* **84**:405-33.

JAROSZ D, LINDQUIST S. 2010. Hsp90 and environmental stress trans- form the adaptive value of natural genetic variation. *Science* **330**:1820–4.

JONES-RHOADES MW, BARTEL DP, BARTEL B. MicroRNAS and their regulatory roles in plants. *Annu Rev Plant Biol.* 2006 **57**:19–53.

KALMYKOVA AI, KLENOV MS, GVOZDEV VA. 2005. Argonaute protein PIWI controls mobilization of retro transposons in the *Drosophila* male germline. *Nucleic Acids Res.* **33**:2052–59.

KAWAOKA S, IZUMI N, KATSUMA S, TOMARI Y. 2011. 3′ end formation of PIWI-interacting RNAs in vitro. *Mol. Cell* **43**:1015–22.

KELLEHER ES. Reexamining the *P*-Element Invasion of *Drosophila melanogaster* Through the Lens of piRNA Silencing. 2016. *Genetics.* **203**(4):1513-31.

KHURANA JS, XU J, WENG Z, THEURKAUF WE. 2010. Distinct functions for the *Drosophila* piRNA pathway in genome maintenance and telomere protection. *PLOS Genet.* **6**:e1001246.

KHURANA JS, WANG J, XU J, KOPPETSCH BS, THOMSON TC, NOWOSIELSKA A, LI C, ZAMORE PD, WENG Z, THEURKAUF WE. 2011. Adaptation to P element transposon invasion in *Drosophila melanogaster*. *Cell* **147**:1551–63.

KIM VN, HAN J, SIOMI MC. 2009. Biogenesis of small RNAs in animals. *Nat. Rev. Mol. CellBiol.* **10**:126–39.

KLATTENHOFF C, BRATU DP, MCGINNIS-SCHULTZ N, KOPPETSCH BS, COOK HA, THEURKAUF WE. 2007. *Drosophila* rasiRNA pathway mutations disrupt embryonic axis specification through activation of an ATR/Chk2 DNA damage response. *Dev. Cell* **12**:45–55.

KLENOV MS, SOKOLOVA OA, YAKUSHEV EY, STOLYARENKO AD, MIKHALEVA EA, LAVROV SA, GVOZDEV VA. 2011. Separation of stem cell maintenance and transposon silencing functions of Piwi protein. *PNAS* **108**:18760–65.

KOLACZKOWSKI B, HUPALO D, KERN A. 2010. Recurrent adaptation in RNA interference genes across the Drosophila phylogeny. *Mol Biol Evol* **28**:1033–42.

KOONIN E, WOLF Y. 2009. Is evolution Darwinian or/and Lamarckian? *Biol Direct* **4**:42.

KUMAR MS, CHEN KC. 2012. Evolution of animal Piwi-interacting RNAs and prokaryotic

CRISPRs. *Briefings in Functional Genomics*. 11(4):277-288.

KURAMOCHI-MIYAGAWA S, WATANABE T, GOTOH K, TOTOKI Y, TOYODA A, IKAWA M, ASADA N, KOJIMA K, YAMAGUCHI Y, IJIRI TW, HATA K, LI E, MATSUDA Y, KIMURA T, OKABE M, SAKAKI Y, SASAKI H, NAKANO T. 2008. DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. *Genes Dev.* 22:908–17.

LAU NC, SETO AG, KIM J, KURAMOCHI-MIYAGAWA S, NAKANO T, BARTEL DR, KINGSTON RE. 2006. Characterization of the piRNA complex from rat testes. *Science* 313:363–67.

LEE RC, FEINBAUM RL, AMBROS V. 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 75:843-854.

LEE E, BANERJEE S, ZHOU J, JAMMALAMADAKA A, ARCILA M, MANJUNATH BS, KOSIK KS. 2011. Identification of piRNAs in the central nervous system. *RNA* 17:1090–9.

LI H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27(21): 2987-93.

LI H, HANDSAKER B, WYSOKER A, FENNELL T. RUAN J, HOMER N, MARTH G, ABECASIS G, DURBIN R, 1000 GENOME PROJECT DATA PROCESSING SUBGROUP. 2009. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 25(16): 2078-9.
LI XZ, ROY CK, DONG X, BOLCUN-FILAS E, WANG J, HAN BW, XU J, MOORE MJ, SCHIMENTI JC, WENG Z, ZAMORE PD. 2013. An ancient transcription factor initiates the burst of piRNA production during early meiosis in mouse testes. *Mol. Cell* 50:67-81.

LIN H, SPRADLING AC. 1997. A novel group of *pumilio* mutations affects the asymmetric division of germline stem cells in the *Drosophila* ovary. *Development* 124:2463–76.

LOHMUELLER K, INDAP A, SCHMIDT S, BOYKO A, HERNANDEZ R, LU J, CLARK A. 2010. Population dynamics of PIWI-interacting RNAs (piRNAs) and their targets in *Drosophila*. *Genome Res* 20:212–27.

LUKIC S, CHEN K. 2011. Human piRNAs are under selection in Africans and repress transposable elements. *Mol Biol Evol* 28:3061–7.

MALONE CD, HANNON GJ. 2009. Small RNAs as guardians of the genome. *Cell* 136:656–68.

MALONE C, BRENNECKE J, DUS M, STARK A, MCCOMBIE WR, SACHIDANANDAM R, HANNON GJ. 2009. Specialized piRNA pathways act in germline and somatic tissues of the drosophila ovary. *Cell* 137:522–35.

MCDONALD JH, KREITMAN M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351(6328):652-4.

MOAZED D. 2009. Small RNAs in transcriptional gene silencing and genome defence. *Nature* 457:413–20.

MOHN F, SIENSKI G, HANDLER D, BRENNECKE J. 2014. The Rhino–Deadlock–Cutoff complex

licenses noncanonical transcription of dual-strand piRNA clusters in *Drosophila*. *Cell* **157**:1364–79.

MOLNAR A, SCHWACH F, STUDHOLME DJ, THUENEMANN EC, BAULCOMBE DC. 2007. miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii. *Nature* **447**:1126–1129.

MOYANO M, STEFANI G. 2015. piRNA involvement in genome stability and human cancer. *J Hematol Oncol* **8**:38.

MUOTRI AR, MARCHETTO MC, COUFAL NG, OEFNER R, YEO G, NAKASHIMA K, GAGE FH. 2010. L1 retrotransposition in neurons is modulated by MeCP2. *Nature* **468**: 443–6.

NANDI S, CHANDRAMOHAN D, FIORITI L, MELNICK AM, HÉBERT JM, MASON CE, RAJASETHUPATHY P, KANDEL ER. 2016. Roles for small noncoding RNAs in silencing of retrotransposons in the mammalian brain. *Proceedings of the National Academy of Sciences of the United States of America* **113**(45):12697-702.

NEI M, GOJOBORI T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**:418-26.

NG KW, ANDERSON C, MARSHALL EA, MINATEL BC, ENFIELD KS, SAPRUNOFF HL, LAM WL AND MARTINEZ VD. 2016. Piwi-interacting RNAs in cancer: Emerging functions and clinical utility. *Mol Cancer* **15**: 5.

OBBARD D, WELCH J, KIM K, JIGGINS FM. 2009. Quantifying adaptive evolution in the *Drosophila* immune system. *PLoS Genet* **5**:e1000698.

OLIVIERI D, SYKORA MM, SACHIDANANDAM R, MECHTLER K, BRENNECKE J. 2010. An *in vivo* RNAi assay identifies major genetic and cellular requirements for primary piRNA biogenesis in *Drosophila*. *EMBO J.* **29**:3301–17

PANTANO L, JODAR M, BAK M, BALLESCÀ JL, TOMMERUP N, OLIVA R, VAVOURI T. 2015. The small RNA content of human sperm reveals pseudogene-derived piRNAs complementary to protein-coding genes. *RNA* **21**(6):1085-95.

PARDUE ML, DEBARYSHE PG. 2003. Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. *Annu. Rev. Genet.* **37**:485–511

PASQUINELLI AE, REINHART BJ, SLACK F, MARTINDALE MQ *et al.* 2000. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* **408**:86–89.

PETERS L, MEISTER G. 2007. Argonaute proteins: mediators of RNA silencing. *Mol. Cell* **26**:611–23.

PFEIFER B, WITTELSBÜRGER U, RAMOS-ONSINS SE, LERCHER MJ. 2014. PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. *Mol Biol Evol* **31**(7): 1929–36.

PROCHNIK SE, ROKHSAR DS, ABOOBAKER AA. 2007. Evidence for a microRNA expansion in

the bilaterian ancestor. *Dev Genes Evol.* **217**:73–77.

QUEITSCH C, SANGSTER T, LINDQUIST S. 2002. Hsp90 as a capacitor of phenotypic variation. *Nature* **417**:618–24.

REINHART BJ, SLACK FJ, BASSON M, PASQUINELLI AE, BETTINGER JC, ROUGVIE AE, HORVITZ HR, RUVKUN G. 2000. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* **403**:901-6.

REUTER M, BERNINGER P, CHUMA S, SHAH H, HOSOKAWA M, FUNAYA C, ANTONY C, SACHIDANANDAM R, PILLAI RS. 2011. Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing. *Nature* **480**:264–67.

ROBINE N, LAU NC, BALLA S, JIN Z, OKAMURA K, KURAMOCHI-MIYAGAWA S, BLOWER MD, LAI EC. 2009. A broadly conserved pathway generates 3′UTR-directed primary piRNAs. *Curr. Biol.* **19**:2066–76.

ROSENKRANZ, D., ZISCHLER, H. 2012. ProTRAC - a software for probabilistic piRNA cluster detection, visualization and analysis. *BMC Bioinformatics* **13**(1): 5.

ROSS RJ, WEINER MM, LIN H. 2014. PIWI proteins and PIWI-interacting RNAs in the soma. *Nature* **505**:353–59.

RUBY JG, JAN C, PLAYER C, AXTELL MJ, LEE W, NUSBAUM C, GE H, BARTEL DP. 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* **127**:1193–207.

RUTHERFORD S, LINDQUIST S. 1998. Hsp90 as a capacitor for mor- phological evolution. *Nature* **396**:336–42.

RYAZANSKY S, RADION E, MIRONOVA A, AKULENKO N, ABRAMOV Y, MORGUNOVE V, KORDYUKOVE MY, OLOVNIKOV I, KALMYKOVA A. 2017. Natural variation of piRNA expression affects immunity to transposable elements. Feschotte C, ed. *PLoS Genetics*. **13**(4):e1006731.

SABIN LR, DELAS MJ, HANNON GJ. 2013. Dogma derailed: the many influences of RNA on the genome. *Mol. Cell* **49**:783–94.

SAITO K, ISHIZU H, KOMAI M, KOTANI H, KAWAMURA Y, NISHIDA KM, SIOMI MC. 2010. Roles for the Yb body components Armitage and Yb in primary piRNA biogenesis in *Drosophila*. *Genes Dev.* **24**:2493–98.

SAVITSKY M, KWON D, GEORGIEV P, KALMYKOVA A, GVOZDEV V. 2006. Telomere elongation is under the control of the RNAi-based mechanism in the *Drosophila* germline. *Genes Dev.* **20**:345–54.

SCHMIDT A, PALUMBO G, BOZZETTI MP, TRITTO P, PIMPINELLI S, SCHÄFER U. 1999. Genetic and molecular characterization of sting, a gene involved in crystal formation and meiotic drive in the male germ line of *Drosophila melanogaster*. *Genetics* **151**:749–60.

SEMPERE LF, COLE CN, MCPEEK MA, PETERSON KJ. 2006. The phylogenetic distribution of

metazoan microRNAs: insights into evolutionary complexity and constraint. *J Exp Zool.* 306:575– 588.

SIDDIQI S, MATUSHANSKY I. 2012. Piwis and piwi-interacting RNAs in the epigenetics of cancer. *J Cell Biochem* 113:373-80.

SIDDIGI S, TERRY M, MATUSHANSKY I. 2012. Hiwi mediated tumorigenesis is associated with DNA hypermethylation. *PLoS One* 7: e33711.

SIMKIN A, WONG A, POH Y-P, THEURKAUF WE, JENSEN JD. 2013. Recurrent and recent selective sweeps in the piRNA pathway. *Evolution; international journal of organic evolution* 67(4):1081-1090.

SIOMI, M.C., SATO, K., PEZIC, D., ARAVIN, A.A. 2011. PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol* 12: 246-58.

SIOMI H, SIOMI MC. 2009. On the road to reading the RNA-interference code. *Nature* 457:396–404.

SLOTKIN R, VAUGHN M, BORGES F, TANURDZI`C M, BECKER JD, FEIJÓ JA, MARTIENSSEN RA. 2009. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell* 136:461–72.

SMITH, N.G., EYRE-WALKER, A. 2002. Adaptive protein evolution in *Drosophila*. *Nature* 415: 1022–4.

SOARES I., MOLEIRINHO A., OLIVERA GN, AMORIM A. 2015. DivStat: a user-friendly tool for single nucleotide polymorphism analysis of genomic diversity. *PLoS One* 10(3): e0119851.

SONG J, LIU J, SCHNAKENBERG SL, HA H, XING J, CHEN KC. 2014. Variation in piRNA and Transposable Element Content in Strains of *Drosophila melanogaster*. *Genome Biology and Evolution*. 6(10):2786-2798.

SPECCHIA V, PIACENTINI L, TRITTO P, TRITTO P, FANTI L, D'ALESSANDRO R, PALUMBO G, PIMPINELLI S, BOZZETTI MP. 2010. Hsp90 prevents phenotypic variation by suppressing the mutagenic activity of transposons. *Nature* 463:662–5.

SPEIR, M.L., ZWEIG, A.S., ROSENBLOOM, K.R., RANEY, B.J., PATEN, B., NEJAD, P., LEE, B.T., LEARNED, K., KAROLCHIK, D., HINRICHS, A.S., HEITNER, S., HARTE, R.A., HAEUSSLER, M., GURUVADOO, L., FUJITA, P.A., EISENHART, C., DIEKHANS, M., CLAWSON, H., CASPER, J., BARBER, G.P., HAUSSLER, D., KUHN, R.M., KENT, W.J. 2016. The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res* 44: 717–25.

SRIVASTAVA M, BEGOVIC E, CHAPMA J, PUTNAM NH *et al.* 2008. The Trichoplax genome and the nature of placozoans. *Nature* 454:955– 960.

STRINGER CH. 2003. Human evolution: Out of Ethiopia. *Nature* 423(6941): 692–3, 695.

STUWE E, TOTH KF, ARAVIN AA. 2014. Small but sturdy: small RNAs in cellular memory and epigenetics. *Genes Dev.* 28:423–31.

SUDMANT PH, *et al.* 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**:75–81.

TAJIMA F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–95.

THE 1000 PROJECT CONSORTIUM. 2015. A global reference for human genetic variation. *Nature* **526**:68-74.

THOMSON T, LIN H. 2009. The biogenesis and function of PIWI proteins and piRNAs: progress and prospect. *Annu. Rev. Cell Dev. Biol.* **25**:355–76.

VAGIN VV, SIGOVA A, LI C, SEITZ H, GVOZDEV V, ZAMORE PD. 2006. A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313**:320–24.

VARKI A, ALTHEIDE TK. 2005. Comparing the human and chimpanzee genomes: Searching for needles in a haystack. *Genome Res***15**:1746-58.

VERMAAK D, HENIKOFF S, MALIK H. 2005. Positive selection drives the evolution of rhino, a member of the heterochro- matin protein 1 family in Drosophila. *PLoS Genet* **1**: 96–108.

WADDINGTON C. 1942. Canalization of development and the inheritance of acquired characters. *Nature* **150**:563–5.

WATANABE T, LIN H. 2014. Posttranscriptional regulation of gene expression by Piwi proteins and piRNAs. *Mol Cell* **56**:18-27.

WATANABE T, TAKEDA A, TSUKIYAMA T, MISE K, OKUNO T, SASAKI H, MINAMI N, IMAI H. 2006. Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. *Genes Dev.* **20**(13):1732-43.

WATANABE T, TOMIZAWA S, MITSUYA K, TOTOKI Y, YAMAMOTO Y, ET AL. 2011. Role for piRNAs and noncoding RNA in de novo DNA methylation of the imprinted mouse Rasgrf1 locus. *Science* **332**:848–52.

WATTERSON GA. 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**(2): 256–276.

WEICK E-M, MISKA EA. 2014. piRNAs: from biogenesis to function. *Development* **141**(18):3458–71.

XU M, YOU Y, HUNSICKER P, HORI T, SMALL C, GRISWOLD MD, HECHT NB. 2008. Mice deficient for a small cluster of Piwi- interacting RNAs implicate Piwi-interacting RNAs in transposon control. *Biol. Reproduct.* **79**:51–57.

YAMANAKA S, SIOMI MC, SIOMI H. 2014. piRNA clusters and open chromatin structure. *Mobile DNA* **5**:22.

ZANNI V, EYMERY A, COIFFET M, ZYTNICKI M, LUYTEN I, QUESNEVILLE H, VAURY C, JENSEN

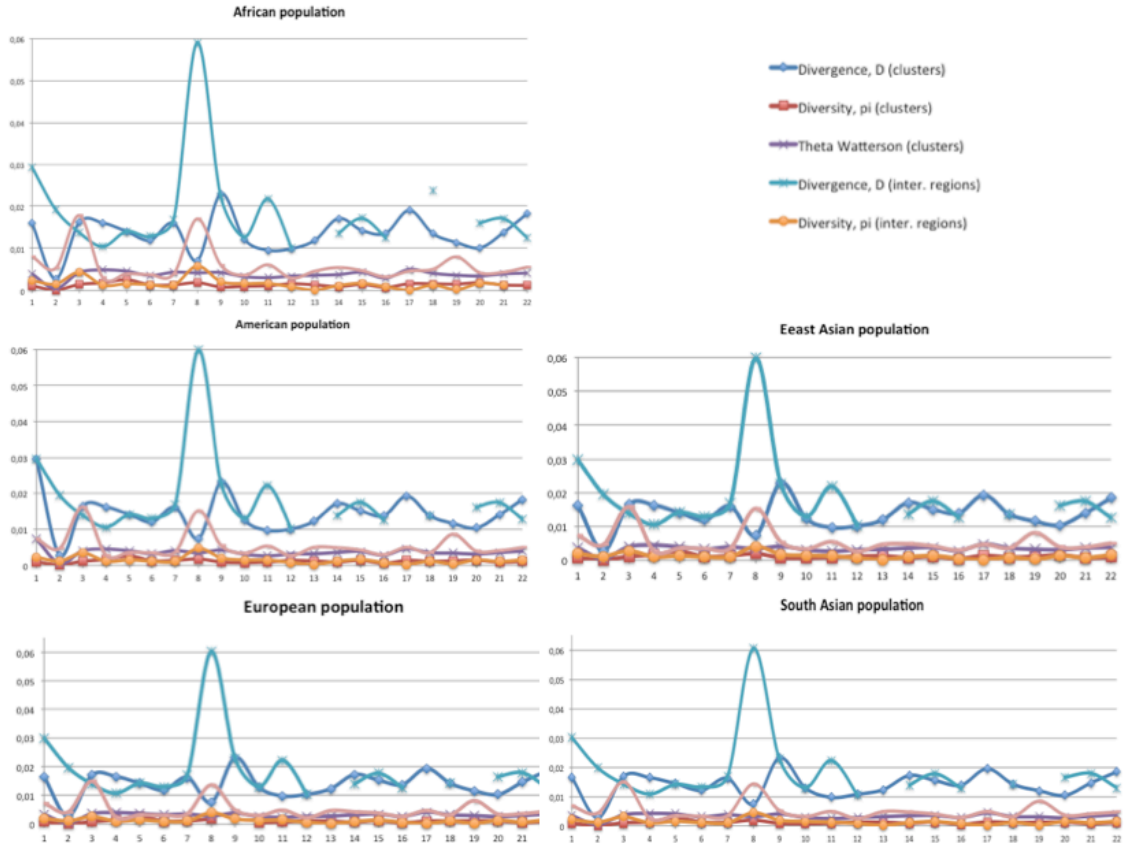
S. 2013. Distribution, evolution, and diversity of retrotransposons at the *flamenco* locus reflect the regulatory properties of piRNA clusters. *PNAS* **110**:19842–47.

ZHANG Z, WANG J, SCHULTZ N, ZHANG F, PARHAD SS, TU S, VREVEN T, ZAMORE PD, WENG Z, THEURKAUF WE. 2014. The HP1 homolog Rhino anchors a nuclear complex that suppresses piRNA precursor splicing. *Cell* **157**:1353–63.

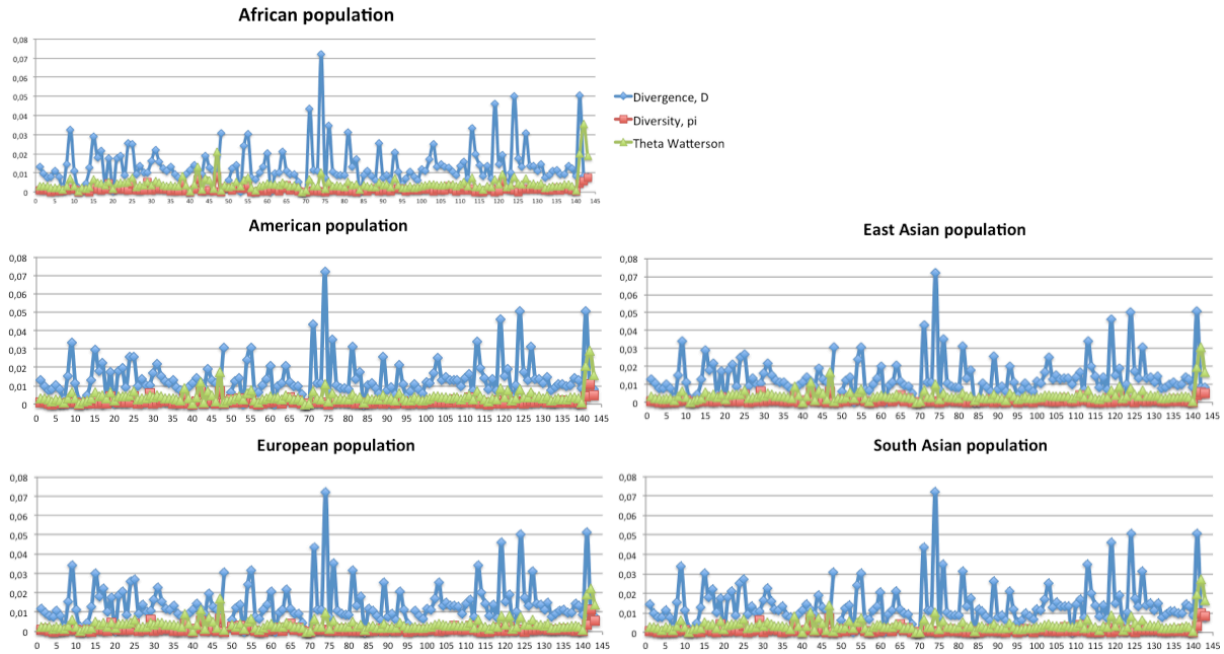
ZHANG H, ZHU J-K. 2011. RNA-directed DNA methylation. *Curr. Opin. Plant Biol.* **14**:142-147.

ZHAO T, LI G, MI S, LI S, HANNON GJ, WANG XJ, QI Y. 2007. A complex system of small RNAs in the unicellular green alga *Chlamydomonas reinhardtii*. *Genes Dev.***21**:1190–203.

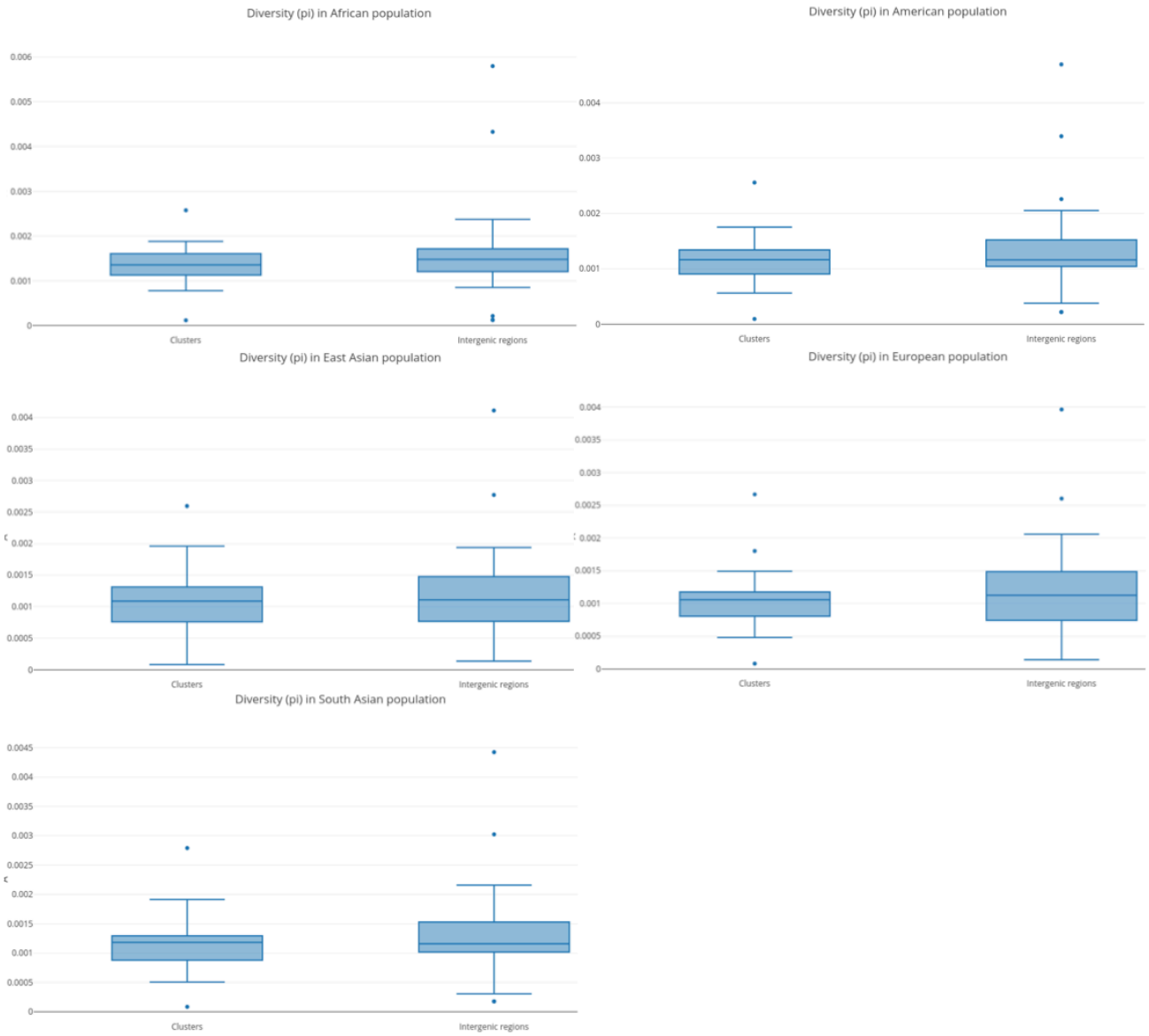
SUPPLEMENTARY FIGURES



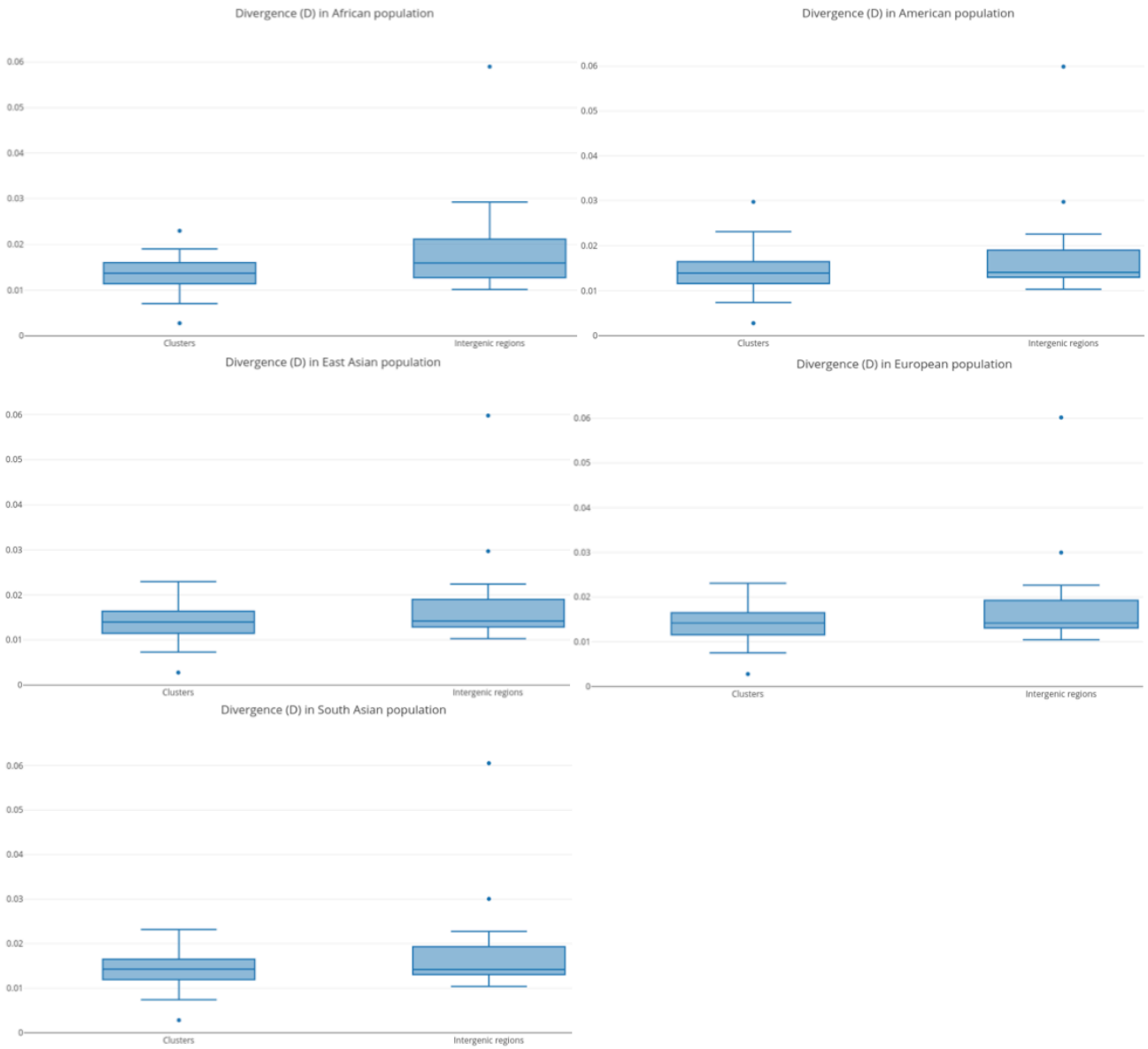
Supplementary Figure F1: Overall diversity and divergence per chromosome and population in piRNA clusters and intergenic regions.



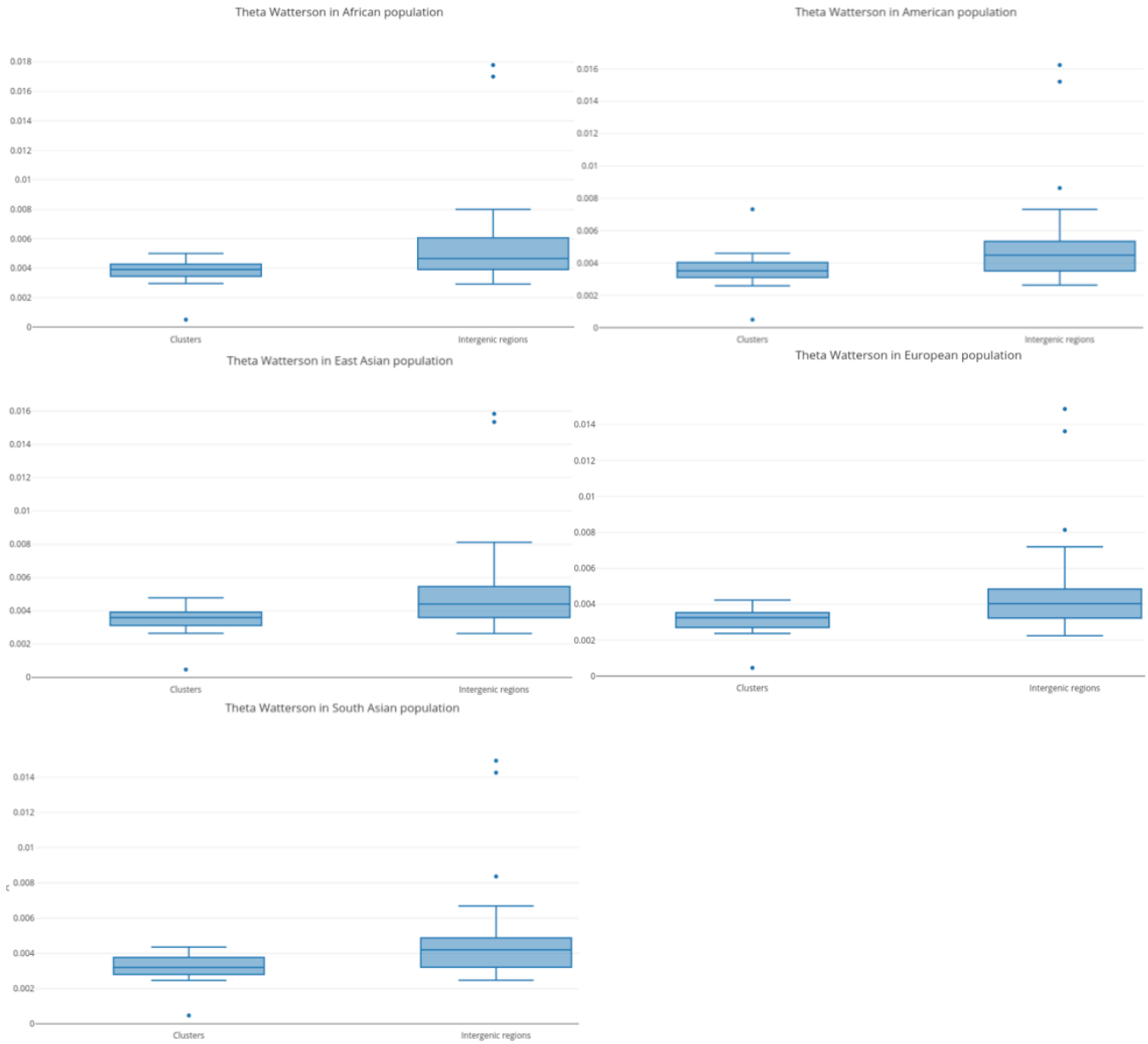
Supplementary Figure F2: Diversity and divergence per piRNA cluster and population.



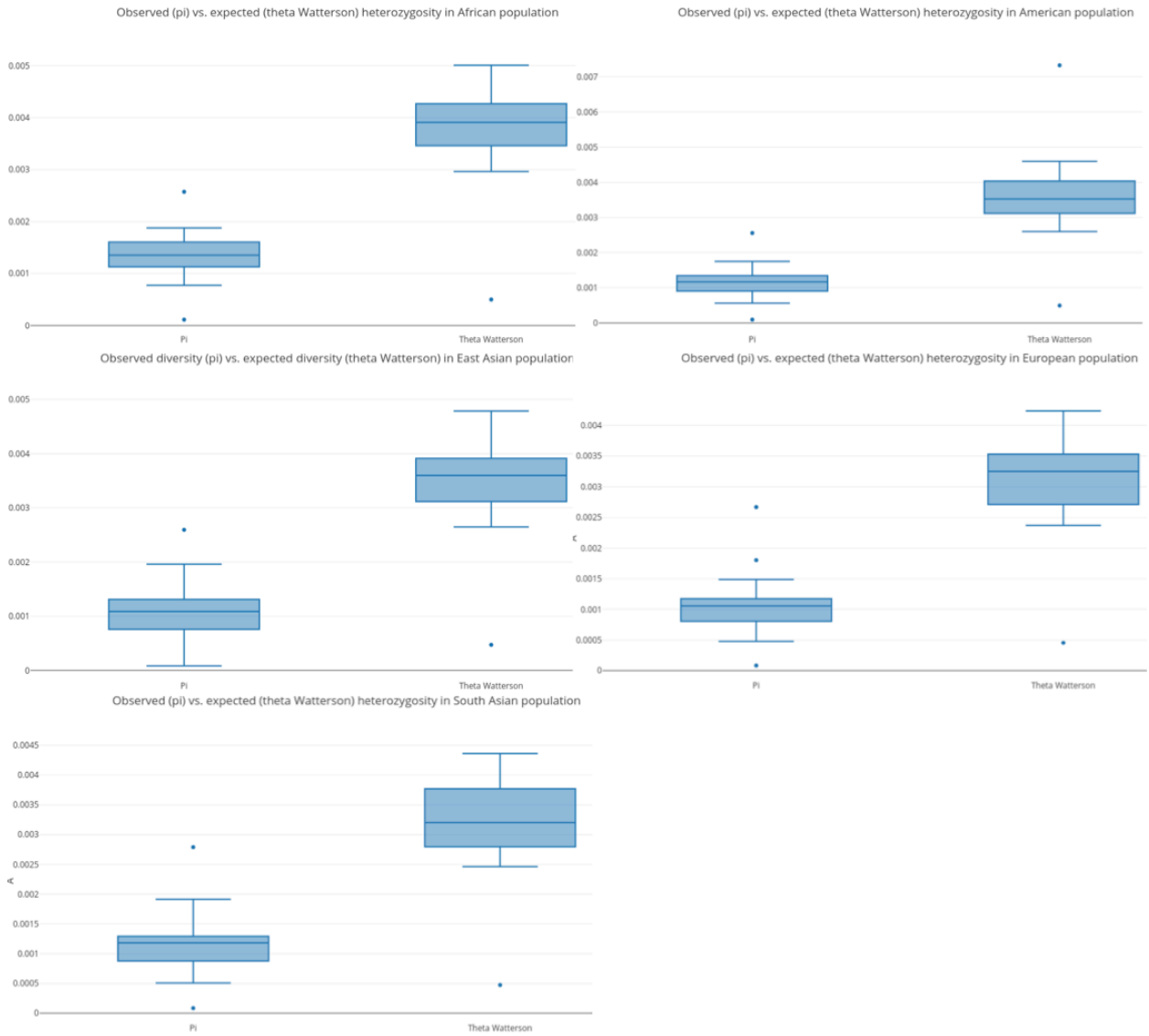
Supplementary Figure F3: Box-plots of genetic diversity π in piRNA clusters and intergenic regions in five populations.



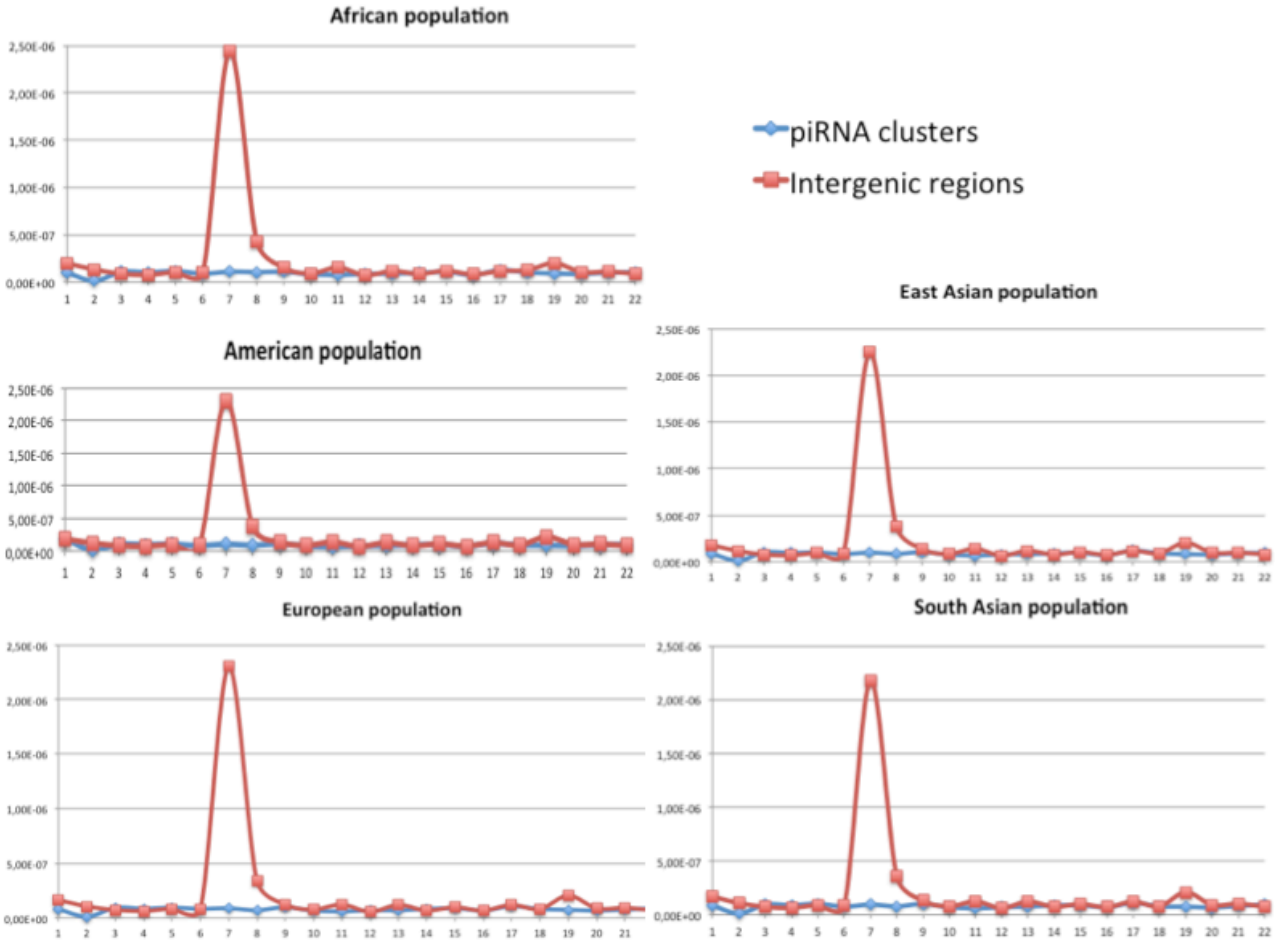
Supplementary Figure F4: Box-plots of genetic divergence *D* in piRNA clusters and intergenic regions in five populations.



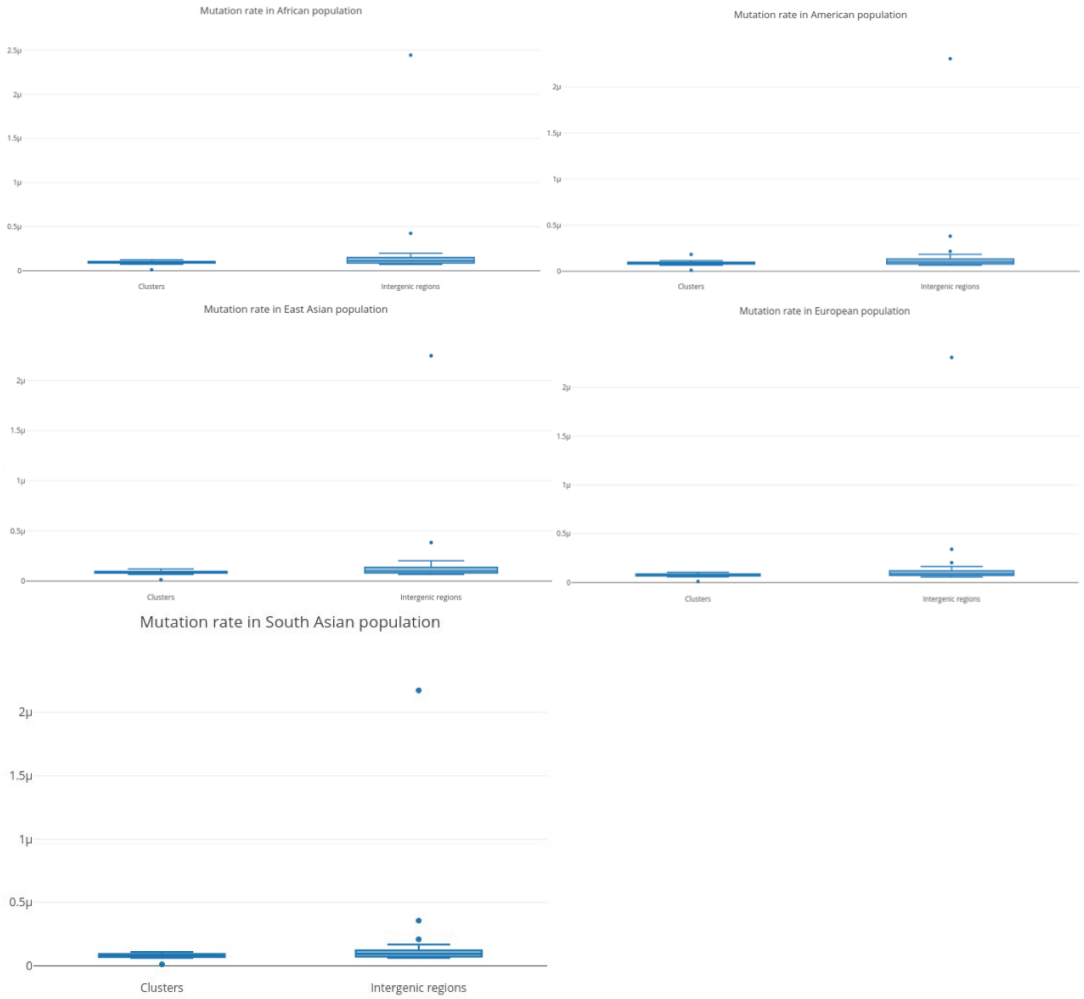
Supplementary Figure F5: Box-plots of expected heterozygosity $\Theta_{Watterson}$ in piRNA clusters and intergenic regions in five populations.



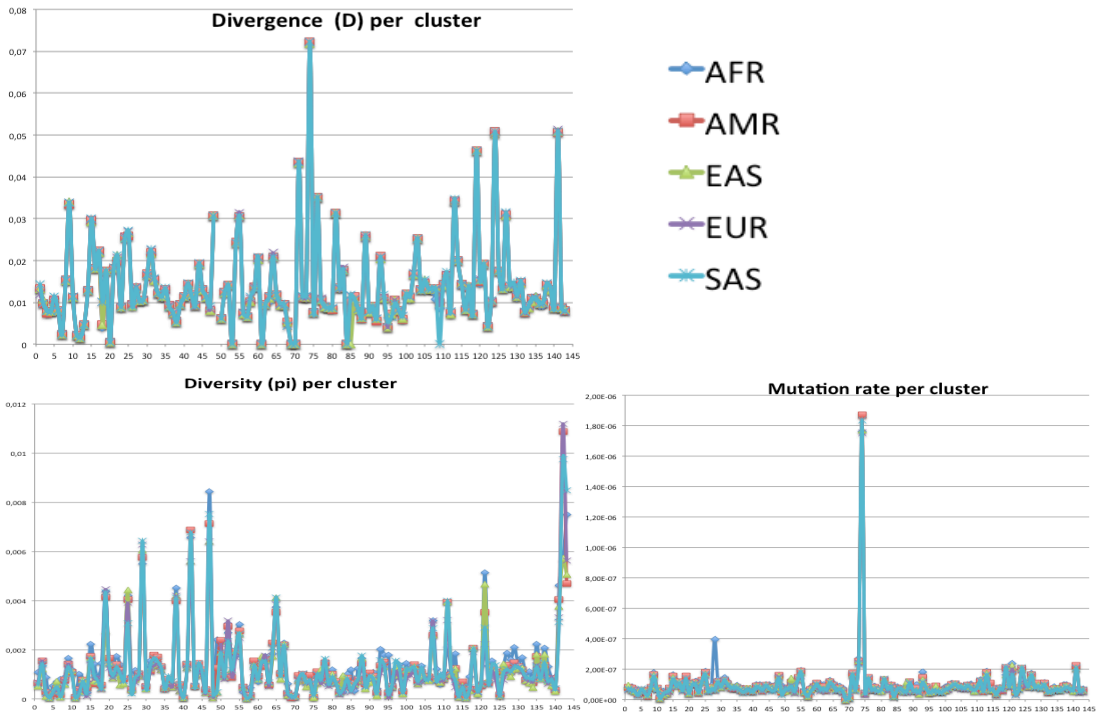
Supplementary Figure F6: Box-plots of observed π vs. expected heterozygosity $\Theta_{\text{Watterson}}$ in piRNA clusters and intergenic regions in five populations.



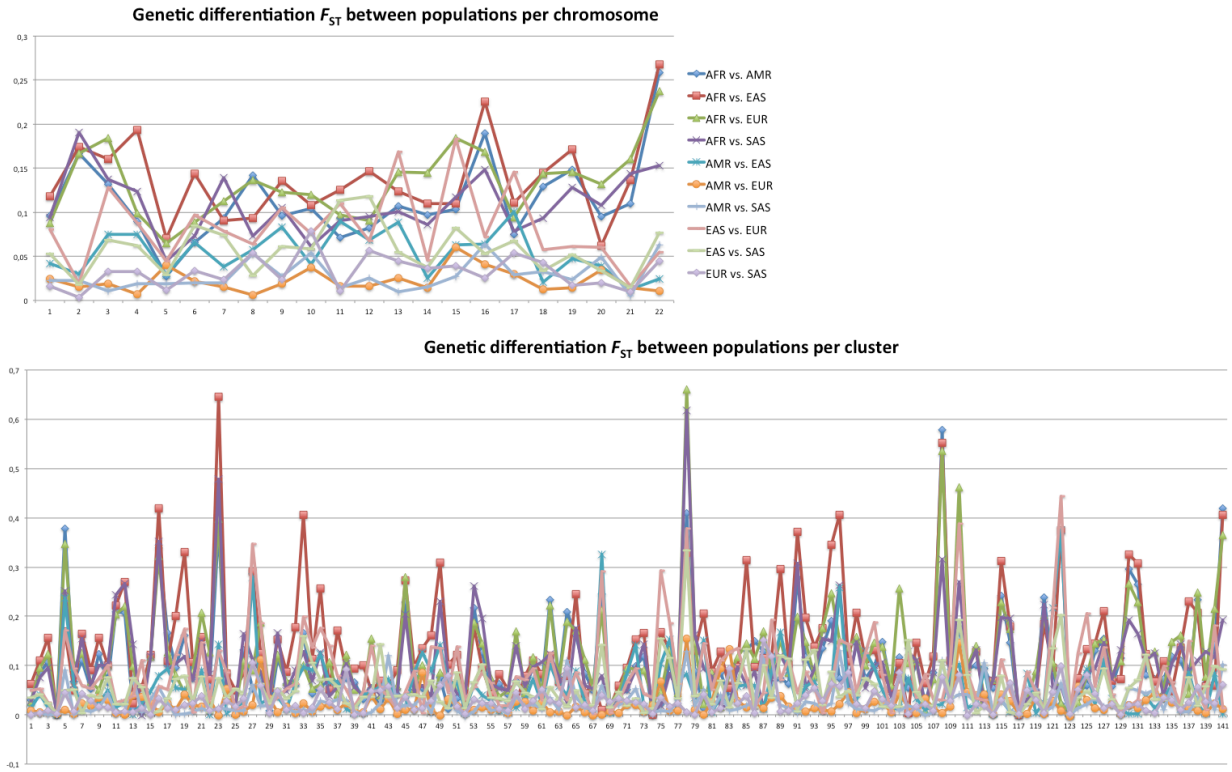
Supplementary Figure F7: Overall mutation rate per chromosome and population in piRNA clusters and intergenic regions.



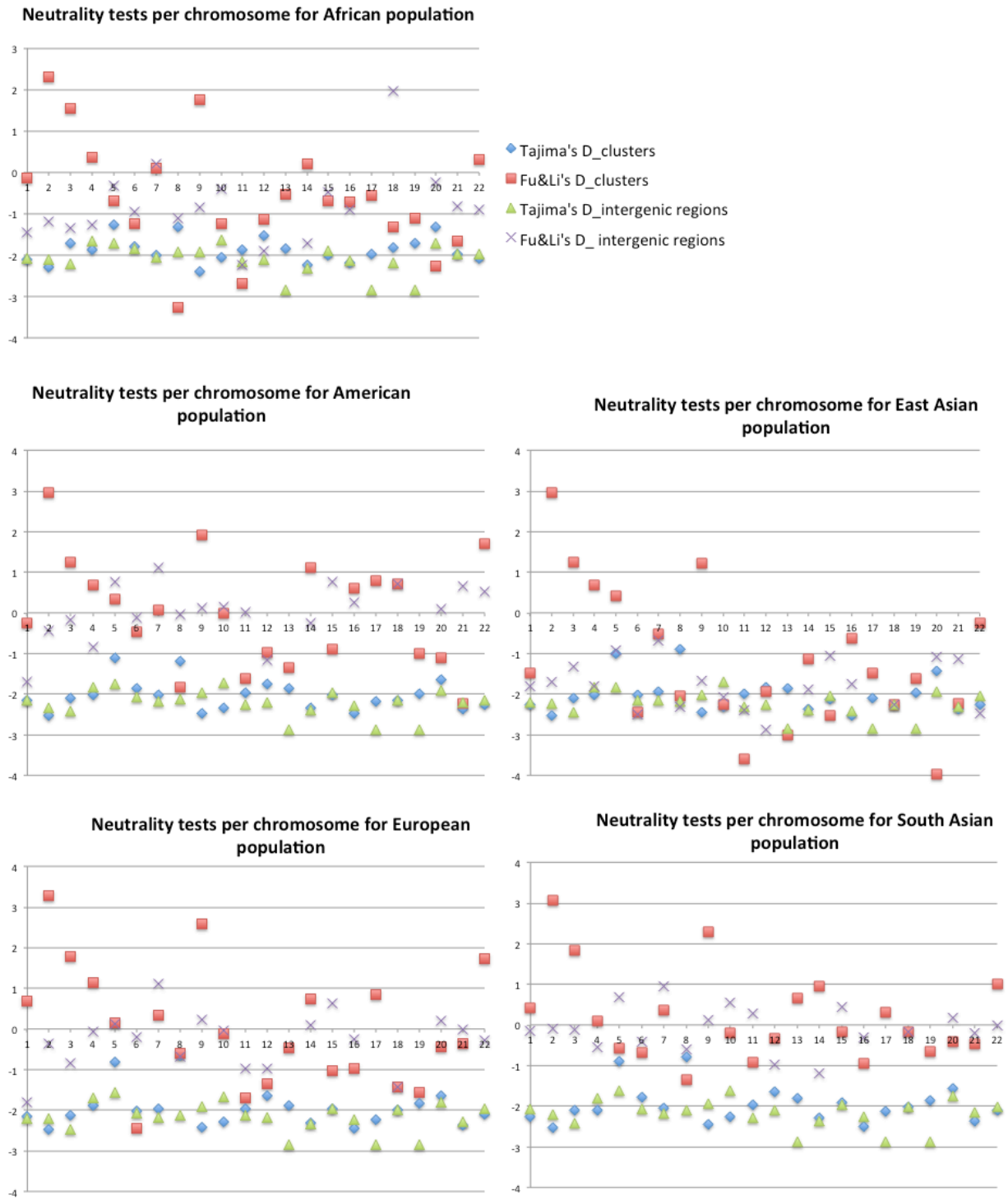
Supplementary Figure F8: Box-plots of mutation rate μ in piRNA clusters and intergenic regions in five populations.



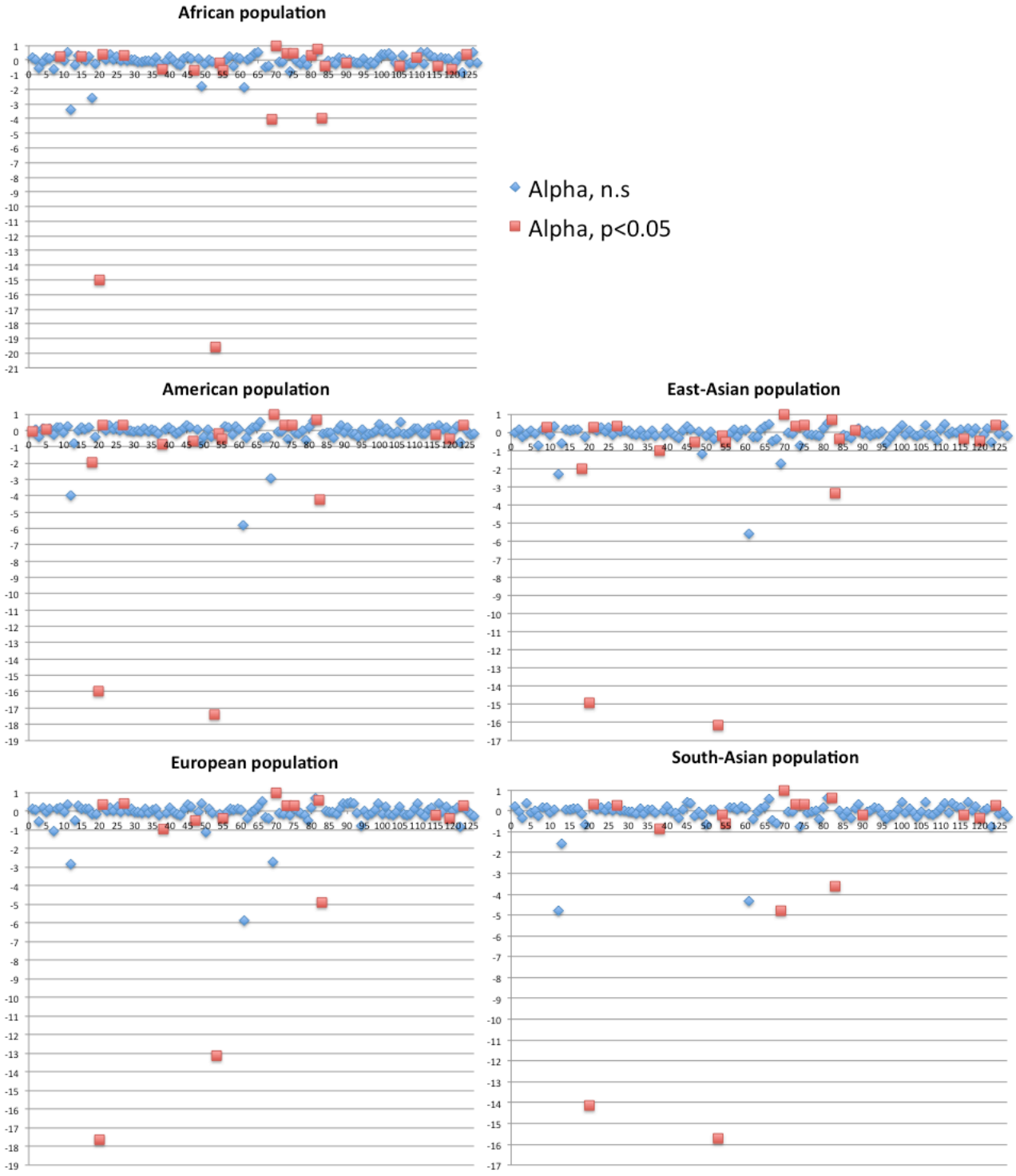
Supplementary Figure F9: Population genetic parameters in piRNA clusters of five populations.



Supplementary Figure F10: Genetic differentiation in piRNA clusters between populations per cluster and chromosome.



Supplementary Figure F11: Neutrality tests per chromosome and population in piRNA clusters and intergenic regions.



Supplementary Figure F12: McDonald and Kreitman test per piRNA cluster for five populations.