High Level of Structural Polymorphism Driven by Mobile Elements in the *Hox* Genomic Region of the Chaetognath *Spadella cephaloptera*

Ferdinand Marlétaz*[,1], Gabor Gyapay[2], and Yannick Le Parco*[,1]

[1]Centre d'Océanologie de Marseille, CNRS UMR 6540 DIMAR, Université de la Méditerranée (Aix-Marseille II), Station Marine d'Endoume, Marseille, France

[2]Genoscope (CEA), CNRS UMR 8030, Université d'Evry, Evry, France

*Corresponding author: E-mail: ferdinand.marletaz@me.com; yannick.leparco@univmed.fr.

Insert sequences have been deposited in GenBank under the following accessions numbers: FP700129 (clone 41B16), FP700130 (clone 46L09), FP700127 (clone 76O01), FP700128 (clone 102L13).
Sequences were deposited in GenBank: Anchor locus (GU257444–GU257468), L36a (GU257420–GU257443), L40a (GU257399–GU257419), COI (GU257469–GU257487) and BAC inserts including *Hox1* gene (FP700127–FP700130).

Accepted: 27 July 2010

Abstract

Little is known about the relationships between genome polymorphism, mobile element dynamics, and population size among animal populations. The chaetognath species *Spadella cephaloptera* offers a unique perspective to examine this issue because they display a high level of genetic polymorphism at the population level. Here, we have investigated in detail the extent of nucleotide and structural polymorphism in a region harboring *Hox1* and several coding genes and presumptive functional elements. Sequencing of several bacterial artificial chromosome inserts representative of this nuclear region uncovered a high level of structural heterogeneity, which is mainly caused by the polymorphic insertion of a diversity of genetic mobile elements. By anchoring this variation through individual genotyping, we demonstrated that sequence diversity could be attributed to the allelic pool of a single population, which was confirmed by detection of extensive recombination within the genomic region studied. The high average level of nucleotide heterozygosity provides clues of selection in both coding and noncoding domains. This pattern stresses how selective processes remarkably cope with intense sequence turnover due to substitutions, mobile element insertions, and recombination to preserve the integrity of functional landscape. These findings suggest that genome polymorphism could provide pivotal information for future functional annotation of genomes.

Key words: structural polymorphism, mobile elements, *Hox* genes, population genomics, Chaetognatha, metazoan evolution.

Introduction

Population genetics studies have only recently paid some attention to noncoding single-copy nuclear regions (Lynch 2007). Although not encoding any protein product, these regions include functional elements such as transcription factor binding sites and regulatory RNAs that control gene expression (Carthew and Sontheimer 2009; Ponting et al. 2009) and are therefore pivotal to morphological evolution (Prud'homme et al. 2007; Wray 2007). Evidence from reassociation kinetics originally suggested that single-copy nuclear regions display high levels of intraspecific variation (Britten et al. 1978), and the accumulation of whole genome sequences has recently allowed the investigation of polymorphism at a broader genome scale (Luikart et al. 2003; Hahn 2008). Notably, significant progress has been made in determining the extent of structural variation in the human genome, that is, polymorphism that is not related to nucleotide substitutions (single nucleotide polymorphisms [SNPs]) but instead insertions, deletions, duplications, and copy number variations (Feuk et al. 2006). However, contrary to the situation in human or inbred model species such as mouse or Drosophila, polymorphism has caused serious difficulties during the genome assembly

Table 1

Estimation of Nucleotide and Structural Polymorphism in Some Available Whole Genome Sequences

Species	Genome Size	π_{nt} (%)	Haplome-Specific DNA	Haplome Insertions[a]	References
Ciona intestinalis	160 Mb	4.5	16.6%	11.4%	Small et al. (2007a, 2007b)
Branchiostoma floridae	520 Mb	3.7%	16.4%	6.8%	Putnam et al. (2008)
Strongylocentrotus purpuratus	800	4–5%	n.d.	n.d.	Sodergren et al. (2006)
Drosophila simulans	150 Mb	1.8%	n.d.	n.d.	Begun et al. (2007)
Homo sapiens	3 Gb	0.09%	n.d.	3.5%	Venter et al. (2001); Levy et al. (2007); Xing et al. (2009)
Spadella cephaloptera	1.03 Gb	2.5%[b]	48.3%	30.9%	This study

NOTE.—n.d., not determined.

[a] This indicates polymorphic insertions that have been consistently identified during the annotation process as haplome-specific regions flanked by homologous sequences in each haplome and thus constitute clear insertions.

[b] This value is based on the alignment of conserved regions from the four BAC sequences (see fig. 1).

process of other organisms for which the starting material was collected in wild populations (Vinson et al. 2005). Marine species were especially found to display high allelic variation at both nucleotide and structural levels (table 1). For example, the sea urchin, the sea squirt and amphioxus all display an average of 5% heterozygosity for SNPs, based on comparison of haplomes from a single individual (Sodergren et al. 2006; Small et al. 2007a; Putnam et al. 2008), which is approximately 5-fold and 10-fold higher than what has been reported among diverse *Drosophila simulans* lines (Begun et al. 2007) and within the diploid human genome (Venter et al. 2001), respectively. Furthermore, structural polymorphism in amphioxus and sea squirt causes ~15% of the total DNA to be haplome specific, that is, present in one haplome only, with no homolog in the other one (Small et al. 2007a; Putnam et al. 2008). In *Ciona intestinalis*, there is preliminary evidence that these levels of structural variation could be related to polymorphic insertions of mobile elements because 12.7% of such insertions are reported to be specific to one haplome (Small et al. 2007a, 2007b). Very recently, the availability of complete haploid human genomes has confirmed the role of mobile genetic elements as key drivers of allelic structural variation because insertion events account for 10% of structural discrepancies in the human genome (Levy et al. 2007; Xing et al. 2009). However, these findings were based on the global analysis of allelic variation detected in the one or few individuals selected for whole genome sequencing, and they thus do not address the question of the distribution of these structural variations within populations nor do they attempt to examine in detail the relation between such polymorphic insertions and functional elements in the genome. Furthermore, the low number of organisms for which whole genome data are available limits our ability to discuss the general implications of genome variations in terms of interrelationships between population size, life history traits, and modalities of molecular evolution.

In the present study, we focus on the structural variation found in the genomic region harboring *Hox1* and several other coding genes of *Spadella cephaloptera*. This species belongs to the enigmatic phylum of chaetognaths or arrow worms, whose puzzling morphological characteristics and rapid evolutionary rates have made it very difficult to branch this lineage in the tree of metazoans (Marlétaz et al. 2006). A phylogenomic approach only recently succeeded in positioning this phylum within the protostomes (arthropods, annelids, and mollusks, etc), most likely in an early diverging position (Dunn et al. 2008; Marlétaz et al. 2008). Nevertheless, high levels of genetic variation have been reported within a single population of *S. cephaloptera*, and it is thus tempting to extend these preliminary observation to the structural level (Marlétaz et al. 2008).

The *Hox1* gene that belongs to the Hox class of transcription factors, which play a major role in anteroposterior patterning during bilaterian development (Pearson et al. 2005). Hox regulation involves a series of interaction between multiple *cis*-regulatory elements that are often distantly related. This complex interplay is thought to have imposed strong structural constraints on the Hox region, which have maintained the integrity of the Hox clusters among bilaterians (Kmita and Duboule 2003). Some exceptions to this highly constrained structure were, however, recently described in several squamate species. In the lizard *Anolis carolinensis*, for example, a massive bombardment of mobile elements drove the expansion of the Hox regions, which in turn led to alteration of Hox regulation and function and caused multiple morphological changes (Di-Poi et al. 2009, 2010).

Here, we analyze the sequence of four bacterial artificial chromosomes (BACs) clone inserts (40–184 kb) corresponding to four different alleles of the *Hox1* region of *S. cephaloptera*. Combining large-scale alignment, genotyping, recombination, and phylogenetic analyses, we show that the *Hox1* region is characterized by unprecedented levels of allelic structural variation in chaetognath populations. We further demonstrate that most of this variation is mediated by the activity of multiple families of retrotransposons and DNA transposons. The high level of polymorphism detected in this study suggests that chaetognaths have developed an unparalleled level of robustness in gene regulation.

Materials and Methods

BAC Screening and Sequencing

A genomic BAC library was built from a thousand adult *S. cephaloptera* individuals from the Sormiou population (Marseille) by Bio S&T Inc. in the pIndigoBAC-5 vector (Epicentre) (Marlétaz et al. 2008). The library represents 55,256 clones arrayed at high density on positively charged nylon filters (Protéigène). Average insert size is 135 kb, which provides a 7× coverage of the 1,050 Mb *S. cephaloptera* genome, as estimated by Feulgen densitometry (Gregory 2007). A hybridization screen was carried out on these filters using 40-bp oligonucleotide probes defined within the *Hox1* homeodomain. Positive clones were then checked for the presence of *Hox1* by polymerase chain reaction (PCR) using specific primers. Recovery of four positive clones for the *Hox1* gene is then consistent with library coverage, and these four inserts were found to be representative of the genomes of four individuals from the original pool employed for building the library. For positive clones, sequencing of BAC inserts was carried out using Sanger dye-terminator chemistry using a shotgun approach that yielded an average 12× coverage at Génoscope (Centre National de Séquençage). Assembly of each BAC was verified by digesting the clones using six different restriction enzymes (Bal I, Bgl II, Hind III, Nco I, Pvu II, and SCA I), and the obtained digestion pattern was compared with the pattern deduced from the assembled sequence.

Annotation and Comparison of Genomic Sequences

Genomic sequences corresponding to the four BAC inserts were annotated by Blast comparison with SwissProt and NR databases (Altschul et al. 1997). Putative coding genes were predicted using Augustus (Stanke and Morgenstern 2005). Whenever similarity with a mobile element component was detected (transposase or polyprotein), another Blast search was conducted against Repbase (Jurka 2000), and terminal repeats were examined at the extremities using pustell DNA matrix (dot plot) implemented in MacVector (MacVector Inc.). Alternatively, nonautonomous elements devoid of coding abilities were generally identified on the basis of dot plot inspection of terminal repeats, with special attention paid to polymorphic insertions. Generally, all occurrences of a newly identified element were investigated by similarity searches in the whole data set of previously recovered mobile elements of *S. cephaloptera* (table 1). Multiple BAC sequence alignment was performed using BlastZ with $K = 1{,}800$ and chainings (Schwartz et al. 2003). Alternative alignment was conducted using Mlagan (Brudno et al. 2003) in order to deal accurately with conserved blocks. The alignment and annotation plotting displayed in figure 1 and figure 3 was made possible by the Perl GD::SVG library. REPEATMASKER was used to detect simple repeats and low-complexity regions for each BAC sequence, as summarized in supplementary table S2 (Supplementary Material online) (Smit et al. 1996–2010).

Molecular Evolution Analyses

For recombination and molecular evolution analyses, a 30,222-nt gap-free alignment was generated from the global Mlagan alignment by excluding all areas displaying insertions of >50 nt in at least one sequence variant. Most likely recombination breakpoints were detected using the hidden Markov models approach, which modeled the probability of a recombination event along the sequence alignment (Husmeier and McGuire 2003). Probability of a portion of alignment to account for one of the three topological alternatives was calculated and plotted in figure 3. This model is well suited for alignments of a small number of sequences over large distances and is included in Topali v2 package (Milne et al. 2009). To extend visual inspection of discrete molecular signatures, occurrence of recombination in the *Hox1* anchor locus was statistically evaluated using the Phi test based on the compatibility principle (Bruen et al. 2006), which is implemented in SplitsTree (Huson and Bryant 2006). Nucleotide heterozygosity (π) was calculated along the 20,022 nt gap-free alignment using a sliding windows of 500 bp windows shifted every 250 bp using Variscan (Vilella et al. 2005). In parallel, indels <50 nt were counted in each 500 nt window using a Perl program, and indel levels were calculated as frequency of indel occurrence per base (number of indel events per window/window size). Molecular evolution parameters such as Tajima and Fu and Li statistics were calculated using DNAsp software (Librado and Rozas 2009). Phylogenetic analyses of marker regions (figs. 2 and 3) were conducted using PhyML 3.0 assuming the K80+F model, and node support values were computed using the likelihood-ratio test approach (Guindon and Gascuel 2003; Anisimova and Gascuel 2006).

Individual Anchoring and Genotyping

Individuals of *S. cephaloptera* were collected in Sormiou and Malmousque, two coastal areas near Marseille (France), which are 10 km apart. Genomic DNA extraction was performed using a Qiamp DNA mini kit (Qiagen) and PCR reactions using GoTaq Hot-Start polymerase (Promega). The following primers were used to amplify the *Hox1* anchor locus: forward, 5′-CATTCATTCATTCATTCGCCCTC-3′ and reverse, 5′-CGGTCTCGCCAGTTGTATCAAG-3′ (Tm = 58 °C). For the amplification of ribosomal proteins *L36a* and *L40a* introns as well as mitochondrial *Cytochrome Oxydase I* (COI) gene, the same primer set and conditions were employed as previously defined (Marlétaz et al. 2008). Cycling conditions

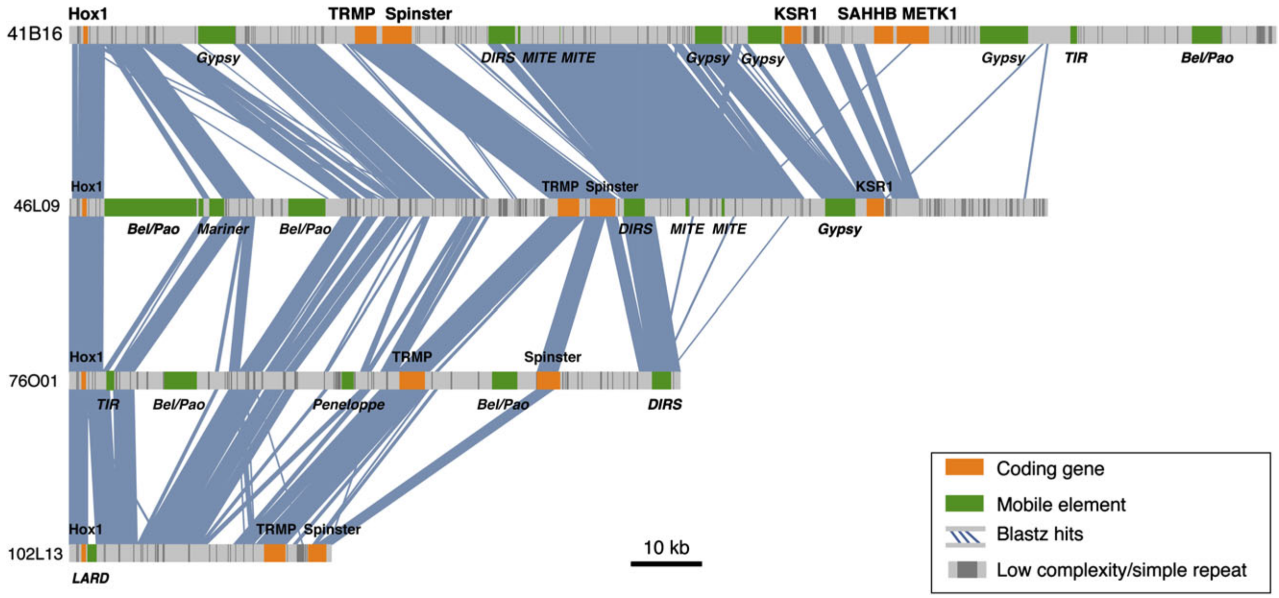

Fig. 1.—Structural variation and conserved blocks between variants of the *Hox1* region. This plotting shows how polymorphic insertions of mobile elements have generated structural shifts between conserved blocks in the four sequenced variants. Sequences obtained from four BAC inserts (light gray background) were annotated for coding genes (orange), mobile elements (green), and low-complexity regions (dark gray). Similarity blocks (blue lines) are drawn from BlastZ alignment between those sequences. Gene names are displayed above BAC sequences and mobile elements underneath (italics). Abbreviations used: *TRPM*, Transient Receptor Potential cation channel subfamily M; *KSR1*, kinase suppressor of ras 1; *SAHH*, S-adenosylhomocysteine hydrolase; *METK*, methionine adenosyltransferase.

were defined to match primer specificity and limit nonspecific amplification by using the touchdown approach, which features progressive reduction of annealing temperature during cycles. PCR products were gel-purified with the SV Purification kit (Promega) and cloned using pGEM-T vector (Promega). Inserts were sequenced using Sanger technology on an ABI 3730xl analyzer (Applied Biosystems). Sequences were recovered and edited using MacVector (MacVector Inc.). ClustalW alignment was refined by hand, which made it possible to consider insertion polymorphism as well as mobile element insertions.

Results

Variations of the *Hox1* Region Driven by Mobile Element Insertions

A BAC library of *S. cephaloptera* was screened, and four BACs positive for the *Hox1* gene were identified and further sequenced using the shotgun approach. A preliminary annotation of these BACs by Blast comparison against the SwissProt database indicated that they all include several coding genes: *Hox1*, *TRPM*, and *Spinster*-related genes (fig. 1 and supplementary table S1, Supplementary Material online). However, we found that intergenic distances vary markedly between the BAC sequences examined, with, for instance, the distance between *Hox1* and *TRPM* ranging from ~30 kb in BAC 102L13 to ~70 kb in BAC 46L09. To better characterize these variations, we computed a pairwise alignment and plotted conserved stretches between the four sequences, which shows extensive structural shifts occasioned by large insertions at several places, but we detected no inversions (fig. 1). This insertion pattern is also clearly outlined by dot plot comparisons of whole BAC sequences (supplementary fig. S1, Supplementary Material online). In contrast, we observed a set of tightly conserved blocks, which generally correspond to coding regions but not always. We attempted to further characterize the inserted regions responsible for these structural variations. Low-complexity regions and simple repeats were not necessarily found clustered within inserted areas (gray in fig. 1, see also supplementary table S2, Supplementary Material online). Conversely, in these areas, we found numerous conserved reverse transcriptase domains typical of retroviruses, which prompted us to suspect that structural variation could be explained by polymorphic insertion of mobile elements.

Therefore, we carried out a careful annotation of mobile elements by comparing genomic sequences with SwissProt and Repbase but also by searching potential long terminal repeats (LTR) and terminal inverted repeats (TIR) using similarity dot plot matrices (supplementary fig. S1, Supplementary Material online). Annotation information is summarized in supplementary table S1, Supplementary Material online. In this way, we identified 22 mobile elements of diverse classes in the *S. cephaloptera* genomic sequences examined (table 2). As detected by pairwise alignment (fig. 1), among the 19 mobile element insertions located in the aligned

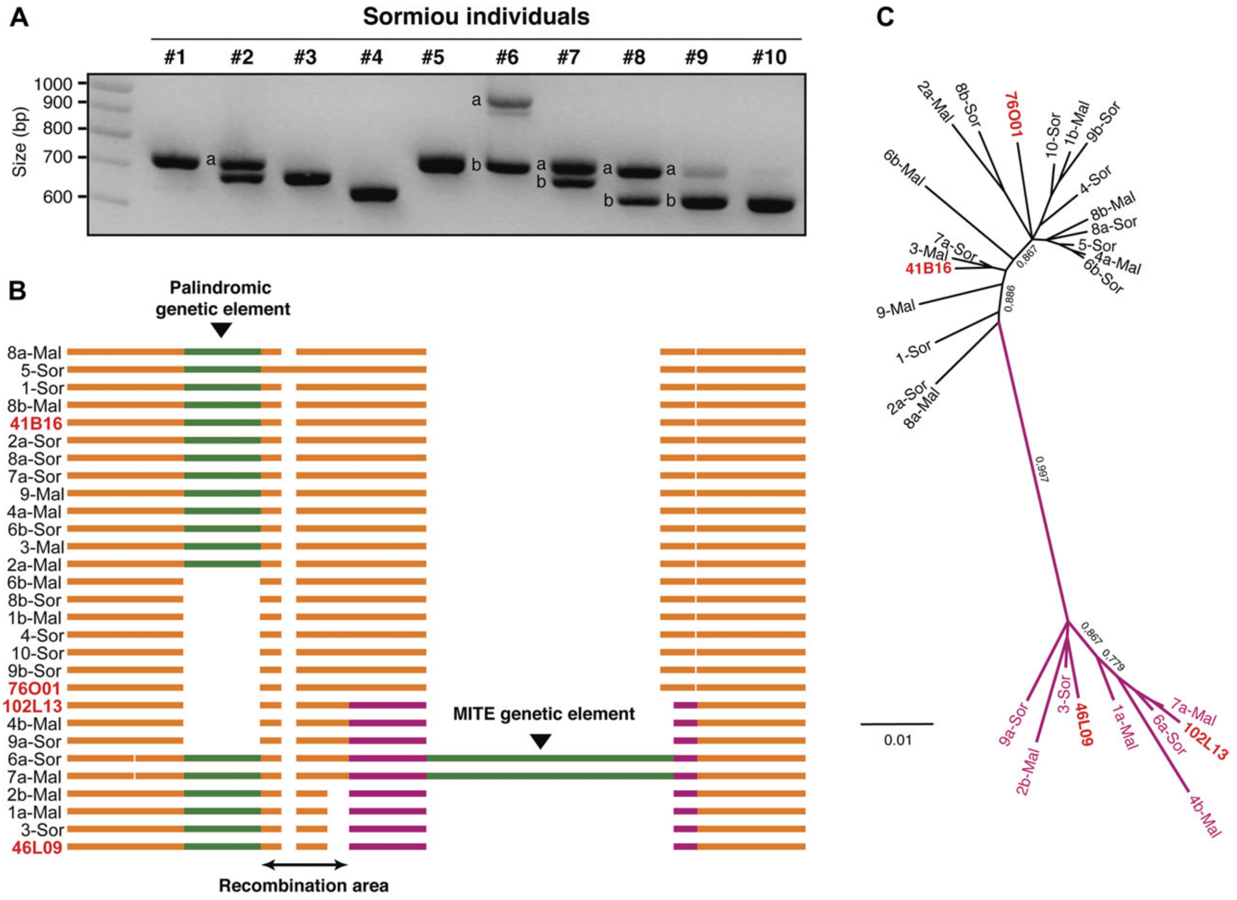

Fig. 2.—Individual anchoring of structural variation. The diagnostic locus was selected on the 5′ side of *Hox1* gene for its ability to discriminate between structural variants through PCR amplification. (*A*) Distinct size of the DNA fragment obtained from amplification of the marker locus for individuals from the Sormiou population. (*B*) Detailed insertion pattern in the sequence variants from BACs and individuals. Size differences observed in panel (*A*) correspond to mobile element insertions (in green, palindromic element and MITE). The region in purple has a very strong nucleotide divergence responsible for the split of alleles into two clades in the phylogenetic analysis (PhyML, K80 with LTR node supports) as presented in (*C*). Notably, sharing of palindromic insertions and divergent regions (purple) between some individuals and one BAC sequence shows evidence of recombination.

regions, ten insertions are clearly polymorphic between sequenced BAC inserts and, altogether, those insertions represent ~70 kb of DNA insertions. The span of insertions generally coincides very well with the range of mobile elements, as exemplified by Bel/Pao-related element in position 5396 of BAC 46L09 and Gypsy elements in BAC 41B16 (fig. 1). Polymorphic insertions are all single occurrences, and none of them are shared between variants. However, some other insertions are present in all BAC inserts and thus clearly predate divergence between them, such as, for example, the DIRS-like element shared between sequences of three BACs (fig. 1).

Diversity and Dynamics of Mobile Elements in Chaetognath

In *S. cephaloptera*, we uncovered a large diversity of mobile elements that encompasses members of the two superclasses of mobile elements, DNA transposons and retrotransposons (table 2). Retrotransposons are particularly abundant in the surveyed region of *S. cephaloptera* with representatives of the Gypsy and Bel/Pao families, which are widespread among metazoans and could constitute a large part of their genomes (Jurka et al. 2007). For instance, the Bel/Pao class of retrotransposons exhibits a peculiar distribution because they are absent from mammals but present in nearly all other groups of metazoans, from platyhelminthes to fishes (Copeland et al. 2005). In *S. cephaloptera*, LTRs were impossible to identify in several of these Bel/Pao elements, whereas those elements include a homologous Bel/Pao polyprotein (table 2). This lack of LTRs could probably be explained by loss due to strong divergence after transposition, which is corroborated by the observed strong divergence of polyprotein genes. A similar situation is observed around the Mariner transposon whose typical TIRs are missing. Possible loss of LTRs is also

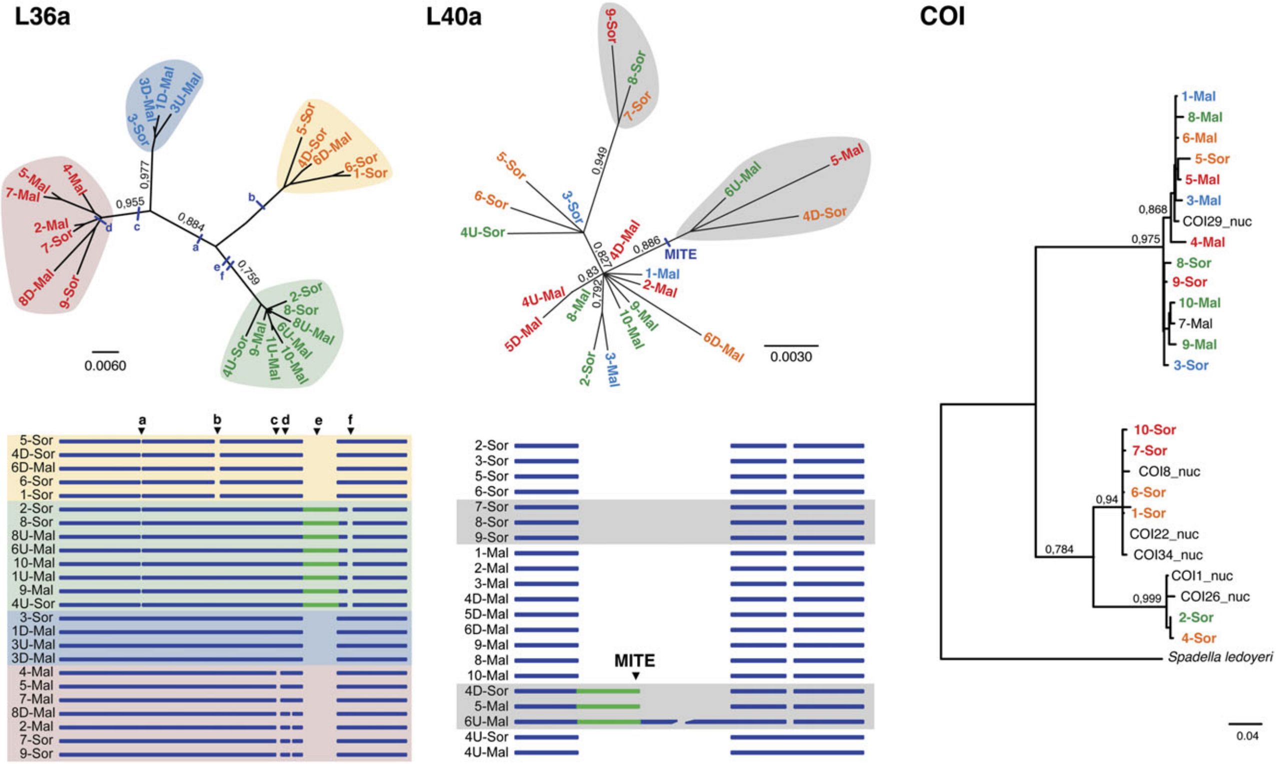

Fig. 3.—Nucleotide and insertion polymorphisms in introns of L36a and L40a and cytochrome oxydase I (COI) mitochondrial gene. Nucleotide divergence between individuals was considered in both cases using a phylogenetic tree (PhyML, K80 with LTR node support). Structural changes such as mobile element insertions and microdeletions are presented as a schematic alignment. In both cases, structural events concur remarkably with nucleotide divergence for splitting individual alleles into several clades, as noted by annotation of branches with discrete events (a, b, c, etc.). To illustrate the lack of congruence between the two marker loci, individuals which belong to a clade defined from L36a were colored accordingly in the L40a and COI tree, which shows they are not clustered according to this other marker. The MITE sequence recovered in the L40a intron from individuals 4,5, and 6 is the same as that described elsewhere in this paper. Some other haplotypes derived from expressed sequence tag data were also included in COI tree (Marlétaz et al. 2008).

supported by the constant ~10 kb size of insertions around polyproteins of these LTR-devoid elements, a size similar to that of the intact LTR Bel/Pao element with ~1 kb-long LTRs (start of 46L09, see table 2). Such losses of LTRs were not observed in Gypsy elements, which do not reach these levels of nucleotide divergence in polyprotein genes. These differences suggest that successive waves of retrotransposition took place within the *S. cephaloptera* genome, with Bel/Pao elements first and Gypsy elements next. On the other hand, another element recovered in BAC 102L13 has lost its coding region but its LTRs have been maintained, a situation associated with so-called large retrotransposon derivative elements (Kalendar et al. 2004). Therefore, the *S. cephaloptera* genome has likely undergone a complex transposition history with several waves of transposition involving diverse elements, from the well-known retrotransposons to the recently discovered DIRS and Penelope class elements (Wicker et al. 2007).

Beyond large elements whose large coding regions are relatively easy to annotate, mobile element diversity also includes small elements like short interspersed nuclear element (SINE) or miniature inverted repeat transposable elements (MITEs) that have played important roles in genome evolution of multiple species (Feschotte et al. 2002). In *S. cephaloptera*, we found a large number of ~300 bp elements that could be considered as MITE elements of their flanking short TIR motifs and lack of any coding domain (table 2). MITE elements were proposed to originate from DNA transposons that have lost their reading frame and were subsequently disseminated after simplification (Feschotte and Mouches 2000). In *S. cephaloptera*, multiple insertions of MITEs were detected; they are generally highly polymorphic, they occur in the vicinity of genes, and are highly polymorphic. We found them not only in the BACs (supplementary table S1, Supplementary Material online) but also in the anchor marker locus of one of the genotyped individuals (see below and fig. 2) and in the intronic region of *L40a* (fig. 3). These multiple occurrences nevertheless indicate broad genomic distribution and high insertion polymorphism for these MITE elements, which makes them insightful for population studies. Moreover, the short length of these MITEs allows detection of their insertion

Table 2
Diversity and classification of mobile elements retrived in BAC sequences from *S. cephaloptera*

Class	Repeat Type	Repeat Size (bp)	Composition	Size Range (bp)	Occurrence
Retrotransposons (class I)					
Gypsy related	LTR	200–250	GAG?-AP-RT-RH-INT	4,052–7,273	41B16 (4) 46L09 (1)
Bel/Pao related	Lost	—	GAG-AP-RT-RH-INT	3,798–5,585	41B16 (1) 46L09 (1) 76O01 (2)
	LTR	914	GAG-AP-RT-RH-INT	13,941	46L09 (1)
DIRS-related	TIR	166	?-YR-?	2,840–3,930	41B16 (1) 46L09 (1) 76O01 (1)
Penelope	TIR	226	RT-?	1,763	76O01 (1)
DNA transposons (class II)					
Tc1/Mariner	Lost	—	Transposase	662	46L09 (1)
Other					
LARD	LTR		—		102L13 (1)
MITE related	TIR	28	—	280–400	41B16 (2) 46L09 (2)

NOTE.—Repeat type refers to type of terminal repeats flanking internal sequences of transposons: repeat size (bp) provides the sequence length of the identified mobile elements among the multiple copies recovered in the region. If applicable, composition of internal sequence is indicated with AP, aspartic protease; RT, reverse transcriptase; RH, RNase H; GAG, capsid protein; INT, integrase; YR, tyrosine recombinase; LARD, large retrotransposon derivative.

site by simple PCR amplification and gel electrophoresis (fig. 2).

Some other short motifs exhibit insertion polymorphism and could be related to some kind of mobile elements of enigmatic nature. This is the case of a ~90-bp palindromic motif, representing the reverse complement of itself, it was found in the sequence of two BACs out of 4, in a very conserved region located just after the end of the *Hox1* gene (fig. 1). Similarly, the size and insertion pattern of this motif provided the opportunity to discriminate between the variants represented by the four BAC sequences at the population scale.

By comparing and annotating four distinct copies of the *Hox1* region, we characterized a strong structural polymorphism, which is mediated by polymorphic insertions of diverse mobile elements. To accurately extend the interpretation of this pattern, we need to assign these structural variations to individuals in a population context.

Population Anchoring of Structural Variants

In order to stress the allelic nature of structural variation uncovered in BAC sequencing, we selected a marker locus on the 5′ side of the *Hox1* gene for its ability to discriminate between the different forms sequenced. In this locus, both a diverging nucleotide stretch and the insertion site of a palindromic element (see previous section) may be used as diagnostic characters to discriminate between the four structural variants (fig. 2*B*). Mobile element insertions notably constitute discrete events with low probability of homoplasy and thus represent valuable markers for population analyses (Ray 2007). Therefore, we performed PCR amplification for this locus on a set of 20 individuals collected in Sormiou and Malmousque, two locations near Marseille (10 individuals each). This yielded, for each variant, a product that had a specific size and could then be distinguished from others by gel electrophoresis (fig. 2*A*). Sormiou is the Calanque area locality in which the individuals used for building the BAC library were collected (Sormiou, SOR in fig. 2). Malmousque is a nearby locality (10 km) within the bay of Marseille (Malmousque, MAL in fig. 2). A previous study, which focused on mysid crustaceans, suggested that these locations are separated by a barrier to gene flow, possibly caused by sea current patterns (Lejeusne and Chevaldonne 2006). We recovered a maximum of two variants for each individual with a high proportion of heterozygous individuals (e.g., individuals #2,6,8, or 9 in fig. 1*A*), which was a first indication of the allelic nature of the observed structural variation.

Moreover, some fragments did not match the sizes expected for the anchor locus but were found to be significantly larger (fig. 2*A*). We further sequenced all distinguishable bands in order to characterize the extent of nucleotide and structural variation found in genotyped individuals. We determined that these larger fragments correspond to the insertion of a MITE element, a pattern recovered in individuals sampled in the two locations (individual 6 of Sormiou and individual 7 of Malmousque, fig. 2). Phylogenetic analysis indicates a unique origin of a strongly divergent nucleotide stretch (purple, fig. 2*B* and *C*) and the insertion of a palindromic genetic element, which should be considered as a discrete event in the ancestry of the population. The presence of both these discrete characters in individuals suggests occurrence of recombination between original alleles, which was further confirmed by the Phi test for recombination (*P* value 4.46 × 10^{-5}, fig. 2*B*). Therefore, by anchoring

Table 3.
Patterns of Nucleotide Substitution in Various Genomic Loci of *Spadella cephaloptera*

Region/Locus	Length	*n*	Sites	*S*	π	θ	Tajima's *D*	Fu and Li's *D**
Hox1 region All	30,222	4	28,292	1,318	0.0247	0.0258	−0.4151	−0.3671
Hox1 region coding	4,398	4	4,386	67	0.0083	0.0085	−0.1663	−0.1365
Hox1 regionNoncoding	25,824	4	23,051	1,154	0.0264	0.0275	−0.4033	−0.3563
*Hox*1 anchor locus	432–796	20	262	60	0.0456	0.0660	−1.1721	−1.8732
L36a intron	720–820	20	657	139	0.0488	0.0575	−0.5818	−1.2990
L40a intron	352–1,315	20	336	29	0.0118	0.0239	−1.9535[a]	−2.6524[a]

Note.—n, number of sampled individuals; S, number of variable sites; π and θ correspond to nucleotide diversity and nucleotide polymorphism, respectively, as estimators of heterozygosity.

[a] Significant *P* values (<0.05).

variants within the wild population, we showed not only that they are part of the natural allelic variation of the population but also that they have been involved in recombination processes at the population level.

To ascertain these conclusions, we further carried out the genotyping of the same individuals for several other independent loci. For each individual, the intronic region of ribosomal protein *L36a* and *L40a* as well as the *COI* mitochondrial gene were sequenced, including multiple alleles from the same individual, when corresponding bands could be distinguished on gel eletrophoresis. This set of loci has previously proven its ability to refute the hypothesis of cryptic speciation, thereby promoting *S. cephaloptera* as a case study for population genetics (Marlétaz et al. 2008). Intronic loci exhibit a strong differentiation at both the nucleotide (table 3) and structural levels, with the insertion of palindromic (fig. 3, *L36a*) and MITE elements (fig. 3, *L40a*).

We observed incongruent genealogies between both nuclear and mitochondrial loci, which is indicative of genetic shuffling and recombination between individuals. For each locus, individuals were distributed within two to four distinct well-defined clades, on the basis of phylogenetic reconstruction, indel events, and mobile element insertions, that are remarkably congruent in terms of nucleotide divergence (fig. 3). The *Hox1* anchor locus, *L36a* intron and the *L40a* intron, all predict a markedly different clustering of individuals, as does the mitochondrial *COI* gene that underwent a different evolutionary history from that of nuclear loci (color code in fig. 3). This incongruence indicates a lack of genetic differentiation between haplotypes of the sampled individuals, which are thus part of the allelic diversity of a single population.

Recombination at the Region Scale

Since we found evidence of recombination within the *Hox1* anchor locus, we attempted to further extend these investigations. We estimated most likely recombination breakpoints along the alignment of *Hox1* region variants using a Markovian approach, and we detected a total of 75 putative recombination breakpoints that correspond to most likely topological changes between blocks, as inferred from nucleotide and indel variations (fig. 4). Coding regions are not excluded from recombination, despite their low heterozygosity that reduces our ability to accurately predict recombination in some coding regions (e.g., *TRPM* gene, fig. 4). The majority of mobile element insertions occurred within recombination-free blocks, which excludes a trend toward an association between recombination breakpoints and insertion sites. Alternatively, no correlation was observed between recombination pattern and the distribution of simple repeats or low-complexity regions (supplementary fig. S2, Supplementary Material online). For example, in the sequence of BAC 46L09, both a LTR retrotransposon and a transposon are clustered within a single block of common origin (alignment position 5,000–6,000 in fig. 4). This reasoning could also be illustrated by the MITE insertion in the anchor marker locus, which probably happened after recombination (between the signature region and palindromic genetic element in fig. 2). This pattern suggests that the major portion of insertions took place after recombination events documented here. This tentative dating suggests that the recombination events that took place in this region could be quite ancient because they would predate mobile element insertions that are likely quite old, as suggested by significant divergence between copies and loss of LTRs (e.g., Bel/Pao in fig. 1). This observation could be related to the inverse correlation between transposition and recombination previously observed in Drosophila, which could be explained by selection against transposable element insertions (Rizzon et al. 2002).

Selection and Functional Elements

To consider how structural polymorphism and recombination affect functional elements harbored by the genomic region under investigation, we carried out a survey of nucleotide substitution patterns in order to detect clues for selection. We recovered a nucleotide diversity of 2.47% (table 3) in the genomic alignment, which likely constitutes a lower estimate at the genome scale because markers that have undergone a greater sampling exhibit stronger values, such as those observed for *L36a* and *L40a* introns as well as the *Hox1* anchor locus, (n = 20, maximum 4.88%, table 3). However, this average high level of

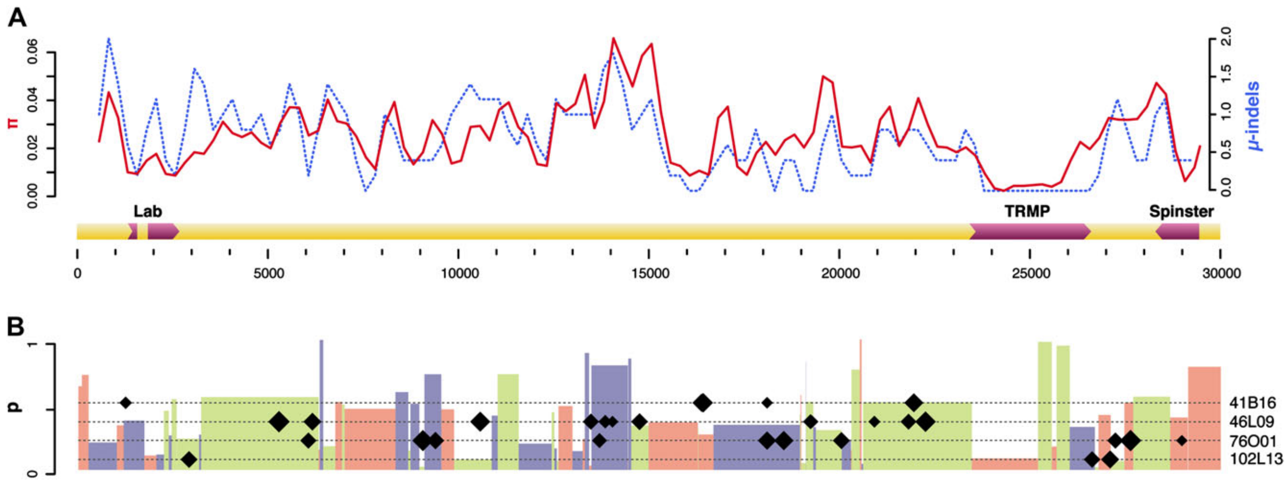

Fig. 4.—Heterozygosity and recombination in conserved block alignment along the *Hox1* region. (*A*) Nucleotide heterozygosity π (blue in *A*) and μ indels (<50 bp, orange in A) were computed in 500 bp windows, every 250 bp and plotted along the alignment. These parameters could be considered in the light of the alignment annotation displayed underneath (*A*). (*B*) Each color represents a block of similar topology between genomic variants in the alignment, and its height is proportional to its probability (related to the amount substitutions and indels supporting this topology). Recombination breakpoints thus correspond to switches from one colored box to another. The positions of large insertions (>50 bp) that were removed from the alignment and generally correspond to transposition events are represented as diamond squares whose size is proportional to the logarithm of insertion length. This plot allowed verification that some large insertions occurred within blocks devoid of recombination points, suggesting that detectable recombination predates insertions in many cases.

heterozygosity masks the strong fluctuations that underlie the heterogeneity of the genomic landscape in our alignment. First, comparison of average levels of heterozygosity between coding and noncoding regions revealed a lower nucleotide variation in coding regions, consistent with purifying selection (table 3). This is confirmed by comparison of synonymous and nonsynonymous rates of substitutions inferred for the three coding genes found in the region explored (table 4). Low values of nonsynonymous to synonymous ratio, so-called K_a/K_s, were found, similarly indicating selection on coding regions (average $K_a/K_s = 0.129$ in coding regions). These observations are in agreement with a previous report of low nonsynonymous divergence in duplicated genes as measured by K_a/K_s (Marlétaz et al. 2008). Then, Tajima's *D* statistics were calculated for both genomic and marker alignments and were found negative for all alignments. Limited sampling ($n = 4$) may limit the significance of this test for genomic alignment because lower values were obtained for markers tested in a larger number of individuals ($n = 20$). Nevertheless, these negative values indicate an excess of low frequency alleles, generally considered as a sign of purifying selection or, alternatively, as an indication of a recent demographic shift such as population expansion.

In order to examine relationships between nucleotide polymorphisms and functional domains in detail, several parameters were computed along the alignment of genomic regions using a sliding-window approach (fig. 4*A*). First, nucleotide diversity (π) was found to correlate with the rate of microinsertions (μ-indel, size <50 bp), a trend previously reported for other species (Small et al. 2007a). Observations of reduced polymorphism in coding regions were confirmed by the correlation of heterozygosity with gene intronic structure, a strong decrease being observed, for example, in the two *Hox1* exons. Remarkably, both π and μ indel rates reach a nearly null value in a large portion of the *TRPM* gene (fig. 4*A*). However, notable fluctuations are observed along the alignment, and low heterozygosity regions are not restricted to coding genes because polymorphism parameters fall to particularly low values in some noncoding regions. For instance, a sudden decrease of heterozygosity around position 15000 could be related to the presence of a putative functional element, such as gene regulatory region or functional noncoding RNA (fig. 4*A*). These parameters generally confirm the strong levels of heterozygosity within the chaetognath population but also the occurrence of selection limiting nucleotide variation in both coding and noncoding regions.

Discussion

Origin and Maintenance of Polymorphism

In this study, we report strong structural and nucleotide polymorphism in a genomic region of the chaetognath *S. cephaloptera* harboring several coding genes including the *Hox1* gene. Previous studies have surveyed genome polymorphism from raw genome sequencing data (Small et al. 2007a; Putnam et al. 2008), whereas we have characterized four alleles of a 100-kb genomic region that presents significant structural variations, and we have subsequently anchored these variations at the individual scale

Table 4

Selection Profiles in Selected Coding Genes as Revealed by Synonymous Versus Nonsynonymous Ratios

Gene	Sites	π	K_s	K_a	K_a/K_s
Hox1	618	0.0223	0.0346 (0.0177)	0.0050 (0.0021)	0.1942 (0.1332)
TRPM	3195	0.0278	0.0253 (0.0052)	0.0030 (0.0008)	0.1182 (0.0251)
Spinster	579	0.0172	0.0168 (0.0032)	0.0012 (0.0014)	0.0756 (0.0853)
Total	4,398	0.0083	0.0255 (0.0126)	0.0030 (0.0021)	0.1294 (0.1005)

NOTE.—π, nucleotide diversity. Standard deviations of pairwise comparisons are indicated in parentheses.

within a reference population. Several lines of evidence clearly support the allelic nature of this variation in regard to alternative hypotheses, such as intraspecific locus duplication. Indeed, partial genome duplication was reported in a previous study that was estimated to have affected ~30% of genes (Marlétaz et al. 2008). However, this duplication is very ancient, as evidenced by high synonymous substitution rates in duplicated genes despite purifying selection (mean $K_s \approx 2$ and $K_a/K_s \approx 0.01$). Recovery of some duplicated genes, such as 18S rRNA, in chaetognath species dispersed throughout the tree of the phylum have also suggested that this duplication may be traced back to the origin of the phylum (Telford and Holland 1997; Papillon et al. 2006). Conversely, a haploid number of nine chromosome was found in three Sagittidae species suggesting a stable number of chromosomes (Bone et al. 1991). Furthermore, genotyping in our study revealed a typical allelic pattern characterized by high heterozygosity and recombination, at both a local and global scale (figs. 2 and 4, respectively). Occurrence of recombination is fully consistent with the sexual reproductive behavior of chaetognaths, for which morphological adaptations are said to prevent self-fertilization (Bone et al. 1991). Moreover, allelic variation is still probably underestimated in the surveyed genomic data, as stressed by new SNPs and structural variants that were characterized in genotyped individuals (fig. 2). Generally, the levels of observed nucleotide heterozygosity are of the same order of magnitude than those observed in genome data of amphioxus (Putnam et al. 2008), tunicate (Small et al. 2007a), or sea urchin (Sodergren et al. 2006) with a typical 2–5% heterozygosity (table 1). This bulk of evidence confirms unprecedented levels of structural variation occurring in a single population that we report here.

Several criteria could be put forward to explain the origin of the nucleotide and structural diversity that we found for *S. cephaloptera*. First, a very large effective population size is said to be a prerequisite for the maintenance of extensive genetic variation in the populations (Kimura 1983). Both the planktonic lifestyle of chaetognaths (Bone et al. 1991) and the current view of their population genetics (Peijnenburg et al. 2004, 2006) are in agreement with a large population size, which is also consistent with the strong purifying selection affecting their duplicated genes (Marlétaz et al. 2008). However, fast molecular rates as well as intrinsic dynamics of mobile element activity should have played a prominent role in shaping the genomic landscape uncovered in this study. This prompts questions about the way polymorphic insertions have spread among the genomes of individuals within the *S. cephaloptera* repartition area. This goal would require extensive sequencing in diverse populations and could become reachable in the next few years with the rise of next generation sequencing methods (Harismendy et al. 2009; Rokas and Abbot 2009).

Mobile Elements

One of the most striking aspects of the variation reported here is the major role played by mobile elements in shaping structural diversity of the genomic region (fig. 1). Interest has been paid only very recently to the impact of transposition dynamics on structural variation (Lynch 2007). Careful examination of newly sequenced diploid human genomes (Levy et al. 2007) identified 846 insertion and deletion events mediated by mobile elements in the human genome, which represent a total 431 kb of structural variation. Experimental assessment of these events in five human individuals sampled so far (Xing et al. 2009) indicated that they are polymorphic in most cases (70%). Other studies have focused on Drosophila in which transposable elements are finely annotated, and they pointed out the low fixation rate of insertions (Perez-Gonzalez and Eickbush 2002; Petrov et al. 2003), despite some recently discovered cases of adaptation induced by mobile elements (Gonzalez et al. 2008). Alternatively, several examples of genome structural rearrangements driven by transposable elements have been described (Caceres et al. 2001; Hughes and Coffin 2001) but the extent of such phenomena in natural populations is not documented to date.

With this study in *S. cephaloptera*, we add a perspective from a wild marine population to the recent account of structural variations mediated by mobile elements in the human genome (Xing et al. 2009). If our observations are to be extended to the whole genome of *S. cephaloptera*, the level of structural variation and polymorphism would be an order of magnitude higher than that in the human population with ~30% of a genomic locus involved. Moreover, whereas in the human genome, mainly SINEs (Alu) and LINEs (L1) are involved, diverse mobile elements were found to account for polymorphic insertions in *S. cephaloptera*, with several known classes of retrotransposons (Gypsy, Bel/Pao, and Penelope) and DNA transposon (Tc1/mariner). These elements are described here for the first time in the chaetognath phylum, which further stresses their widespread distribution among metazoans (table 2). Furthermore, this study confirms the description of new classes like Penelope, Helitrons, or Politrons permitted by whole genome sequencing of several organisms and emphasizes the phylogenetic extent of this diversity (Jurka et al. 2007). Retrotransposons seem particularly abundant in chaetognaths but current resources make it difficult to trace back their respective relationships and ancestry. Although remnant similarity in their coding regions and LTR sequences are clues of a common origin, their high nucleotide divergence could be either accounted by accumulation of substitutions after insertion or alternatively by diversification within families that could for instance have yielded several distinct but related members of the Bel/Pao class.

Regulatory Robustness

One of the most puzzling aspects of the genomic variation reported here is how gene regulation takes place in such a variable genomic landscape. The region studied contains a high density of genes with 3 to 6 coding genes (in the longest available BAC insert) and at least two of these genes could be presumed to be expressed in a highly regulated fashion: *Hox1* is characterized by an anteroposterior domain shared between all bilaterians (Pearson et al. 2005) and *Spinster* expression in Drosophila is limited to a subset of glial and ovary cells (Nakano et al. 2001). Therefore, both coding and regulatory sequences are expected to be under strong functional constraints for these genes and, particularly, for the *Hox1* genes whose appropriate regulation has been shown to be critical for development (Kmita and Duboule 2003). For instance, mobile element insertion in the *Hox* gene cluster has caused alteration of the expression of posterior *Hox* genes in squamates (Di-Poi et al. 2010). The lack of redundancy of *Hox* gene clusters in invertebrates as well as the key role of *Hox1* in regard to posterior *Hox* genes increasingly stress the importance of its regulation, calling for collection of further experimental evidence about Hox gene expression and genomic location in *S. cephaloptera*. Nevertheless, alterations of expression reported in squamates may not necessarily affect chaetognaths. First, contrary to squamates that exhibit prominent morphological divergence from the tetrapod body plan (e.g., loss of limbs in snakes), the body plan of chaetognaths is remarkably conserved and the phylum displays very little morphological variation that could possibly be related to a Hox expression shift. Second, it has been previously suggested that scattered expression of *Hox* genes along the anteroposterior axis could be maintained in the absence of a conserved cluster organization and collinear expression (Seo et al. 2004) and that multiple mechanisms of *Hox* gene regulation could take place in distinct bilaterian lineages (Duboule 2007). Third, in contrast to mobile insertion events taking place in the *Hox* cluster of squamates, those reported here are not fixed and still belong to the allelic diversity of the reference population, which suggest that the *Hox* regulation mechanism has to cope with the variety of genomic landscapes in diverse individuals. Deciphering how this regulation could be triggered would provide a tremendous insight into the evolution of gene regulation mechanisms and thereby the evolution of body organization.

Several kinds of functional constraints could affect intergenic spaces: regulatory regions such as transcription factor binding sites, regulatory noncoding RNA but also structural constraints related to chromatin structure or transcription initiation (Castillo-Davis 2005). Intergenic distances have been hypothesized to evolve through rapid DNA turnover involving stepwise large insertions and persistent DNA loss through small deletions, but it remains unclear whether these processes are mainly driven by selection or dynamic equilibrium (Singh and Petrov 2004). Clear evidence of purifying selection in coding sequences indicates that selective processes are at play in this genomic region and population, which suggests that observed structural variations have been spared by selective processes and do not significantly affect fitness. Furthermore, it should also be noted that none of the polymorphic insertions is related to inversion or modification of the gene order, which has been remarkably conserved among the four BACs studied. This conservation could be consistent with the elimination of most divergent structural variants by selection, indicating that observed structural variation is below a selection threshold. In any case, although modalities of selection on coding nucleotide sequences are well characterized, the way selection impacts noncoding sequences and structure of genomes deserves further investigation.

Moreover, the presence of functional elements in this genomic region underlines the robustness of gene regulatory processes controlling physiology and development functions, ensured for instance by *Spinster* and *Hox1*, respectively. This stresses the apparent paradox of the neutral theory of molecular evolution in which large population size favors accumulation of an increasing genetic diversity despite improved selection efficiency. Interestingly, mathematical evaluation of the potential link between robustness and population parameters suggests that maximal robustness is most likely to evolve in a polymorphic large population, such as that reported here for *S. cephaloptera* (Wagner 2005). This hypothesis constitutes an interesting opportunity to bridge population genomics with system biology by connecting the evolution of genetic network properties to population parameters.

"Intraspecific Footprinting" in Nonmodel Organisms

High levels of variation within genomes of a single species have profound implications from a comparative genomics perspective by promoting a possible practical way to identify functional elements through an intraspecific footprinting approach. This approach is based on the high conservation of functional elements between genomes of related organisms, a particularly relevant feature for the identification of noncoding functional elements, such as transcription factor binding sites (Wasserman and Sandelin 2004). However, one of the major caveats of footprinting is the need to select the appropriate phylogenetic range between compared genomes: if the species are too close, divergence is insufficient to be conclusive, whereas if they are too far, alignment could be difficult, and moreover, some elements could have undergone major turnover (Hare et al. 2008). These principles have recently been extended to the intraspecific level in *C. intestinalis*, where they have allowed the recovery of regulatory sites for several genes (Boffelli et al. 2004). This has been made possible by evaluating genetic diversity in sampled individuals and subsequently using reporter gene constructs (Harafuji et al. 2002). This use of polymorphism data for annotation purpose could be applied in the future to new biological models, as *S. cephaloptera* or other relevant marine species. Furthermore, species with big genomes and abundant mobile elements could be particularly well suited for this purpose because their functional elements are more easily identified because they are widespread and flanked by large stretches of poorly conserved DNA (Peterson et al. 2009). Nevertheless, relationships between robustness of gene regulatory networks, genome polymorphism, and population size are underexplored paths which could help to understand major evolutionary events, such as the radiation of metazoans at the beginning of the Cambrian period (Knoll and Carroll 1999).

Supplementary Material

Supplementary figures S1–S2 and tables S1–S2 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We thank Clément Gilbert and David Enard for their insightful comments on the manuscript. We are also grateful to Susan Cure for correcting the manuscript. This work was funded by the Centre National de la Recherche Scientifique (CNRS) and F.M. is supported by a CNRS BDI PhD fellowship.

Literature Cited

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Anisimova M, Gascuel O. 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst Biol. 55:539–552.

Begun DJ, et al. 2007. Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. PLoS Biol. 5:e310.

Boffelli D, et al. 2004. Intraspecies sequence comparisons for annotating genomes. Genome Res. 14:2406–2411.

Bone Q, Kapp H, Pierrot-Bults AC, editors. 1991. The biology of chaetognaths. Oxford: Oxford University Press.

Britten RJ, Cetta A, Davidson EH. 1978. The single-copy DNA sequence polymorphism of the sea urchin Strongylocentrotus purpuratus. Cell. 15:1175–1186.

Brudno M, et al. 2003. LAGAN and multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res. 13:721–731.

Bruen TC, Philippe H, Bryant D. 2006. A simple and robust statistical test for detecting the presence of recombination. Genetics. 172: 2665–2681.

Caceres M, Puig M, Ruiz A. 2001. Molecular characterization of two natural hotspots in the Drosophila buzzatii genome induced by transposon insertions. Genome Res. 11:1353–1364.

Carthew RW, Sontheimer EJ. 2009. Origins and mechanisms of miRNAs and siRNAs. Cell. 136:642–655.

Castillo-Davis CI. 2005. The evolution of noncoding DNA: how much junk, how much func? Trends Genet. 21:533–536.

Copeland CS, et al. 2005. The Sinbad retrotransposon from the genome of the human blood fluke, Schistosoma mansoni, and the distribution of related Pao-like elements. BMC Evol Biol. 5:20.

Di-Poi N, Montoya-Burgos JI, Duboule D. 2009. Atypical relaxation of structural constraints in Hox gene clusters of the green anole lizard. Genome Res. 19:602–610.

Di-Poi N, et al. 2010. Changes in Hox genes' structure and function during the evolution of the squamate body plan. Nature. 464:99–103.

Duboule D. 2007. The rise and fall of Hox gene clusters. Development. 134:2549–2560.

Dunn CW, et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature. 452:745–749.

Feschotte C, Jiang N, Wessler SR. 2002. Plant transposable elements: where genetics meets genomics. Nat Rev Genet. 3:329–341.

Feschotte C, Mouches C. 2000. Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the Arabidopsis thaliana genome has arisen from a pogo-like DNA transposon. Mol Biol Evol. 17:730–737.

Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. Nat Rev Genet. 7:85–97.

Gonzalez J, et al. 2008. High rate of recent transposable element-induced adaptation in Drosophila melanogaster. PLoS Biol. 6:e251.

Gregory TR, et al. 2007. Nucleic Acids Res. 35 (Database issue):D332–8.

Gregory TR. 2010. Animal Genome Size Database [Internet]. [cited 2010 August 25]. Available from: http://www.genomesize.com.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.

Hahn MW. 2008. Toward a selection theory of molecular evolution. Evolution. 62:255–265.

Harafuji N, Keys DN, Levine M. 2002. Genome-wide identification of tissue-specific enhancers in the Ciona tadpole. Proc Natl Acad Sci U S A. 99:6802–6805.

Hare EE, et al. 2008. Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLoS Genet. 4:e1000106.

Harismendy O, et al. 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. 10:R32.

Hughes JF, Coffin JM. 2001. Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. Nat Genet. 29:487–489.

Husmeier D, McGuire G. 2003. Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. Mol Biol Evol. 20:315–337.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23:254–267.

Jurka J. 2000. Repbase update: a database and an electronic journal of repetitive elements. Trends Genet. 16:418–420.

Jurka J, Kapitonov VV, Kohany O, Jurka MV. 2007. Repetitive sequences in complex genomes: structure and evolution. Annu Rev Genomics Hum Genet. 8:241–259.

Kalendar R, et al. 2004. Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. Genetics. 166:1437–1450.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Kmita M, Duboule D. 2003. Organizing axes in time and space; 25 years of colinear tinkering. Science. 301:331–333.

Knoll AH, Carroll SB. 1999. Early animal evolution: emerging views from comparative biology and geology. Science. 284:2129–2137.

Lejeusne C, Chevaldonne P. 2006. Brooding crustaceans in a highly fragmented habitat: the genetic structure of Mediterranean marine cave-dwelling mysid populations. Mol Ecol. 15:4123–4140.

Levy S, et al. 2007. The diploid genome sequence of an individual human. PLoS Biol. 5:e254.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 25:1451–1452.

Luikart G, et al. 2003. The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet. 4: 981–994.

Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates.

Marlétaz F, et al. 2008. Chaetognath transcriptome reveals ancestral and unique features among bilaterians. Genome Biol. 9:R94.

Marlétaz F, et al. 2006. Chaetognath phylogenomics: a protostome with deuterostome-like development. Curr Biol. 16:R577–578.

Milne I, et al. 2009. TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops. Bioinformatics. 25:126–127.

Nakano Y, et al. 2001. Mutations in the novel membrane protein spinster interfere with programmed cell death and cause neural degeneration in Drosophila melanogaster. Mol Cell Biol. 21: 3775–3788.

Papillon D, Perez Y, Caubit X, Le Parco Y. 2006. Systematics of Chaetognatha under the light of molecular data, using duplicated ribosomal 18S DNA sequences. Mol Phylogenet Evol. 38:621–634.

Pearson JC, Lemons D, McGinnis W. 2005. Modulating Hox gene functions during animal body patterning. Nat Rev Genet. 6: 893–904.

Peijnenburg KT, Breeuwer JA, Pierrot-Bults AC, Menken SB. 2004. Phylogeography of the planktonic chaetognath Sagitta setosa reveals isolation in European seas. Evolution. 58: 1472–1487.

Peijnenburg KT, Fauvelot C, Breeuwer JA, Menken SB. 2006. Spatial and temporal genetic structure of the planktonic Sagitta setosa (Chaetognatha) in European seas as revealed by mitochondrial and nuclear DNA markers. Mol Ecol. 15: 3319–3338.

Perez-Gonzalez CE, Eickbush TH. 2002. Rates of R1 and R2 retrotransposition and elimination from the rDNA locus of Drosophila melanogaster. Genetics. 162:799–811.

Peterson BK, et al. 2009. Big genomes facilitate the comparative identification of regulatory elements. PLoS One. 4:e4688.

Petrov DA, et al. 2003. Size matters: non-LTR retrotransposable elements and ectopic recombination in Drosophila. Mol Biol Evol. 20:880–892.

Ponting CP, Oliver PL, Reik W. 2009. Evolution and functions of long noncoding RNAs. Cell. 136:629–641.

Prud'homme B, Gompel N, Carroll SB. 2007. Emerging principles of regulatory evolution. Proc Natl Acad Sci U S A. 104(Suppl 1): 8605–8612.

Putnam NH, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. Nature. 453:1064–1071.

Ray DA. 2007. SINEs of progress: mobile element applications to molecular ecology. Mol Ecol. 16:19–33.

Rizzon C, Marais G, Gouy M, Biemont C. 2002. Recombination rate and the distribution of transposable elements in the Drosophila melanogaster genome. Genome Res. 12:400–407.

Rokas A, Abbot P. 2009. Harnessing genomics for evolutionary insights. Trends Ecol Evol. 24:192–200.

Schwartz S, et al. 2003. Human-mouse alignments with BLASTZ. Genome Res. 13:103–107.

Seo HC, et al. 2004. Hox cluster disintegration with persistent antero-posterior order of expression in Oikopleura dioica. Nature. 431:67–71.

Singh ND, Petrov DA. 2004. Rapid sequence turnover at an intergenic locus in Drosophila. Mol Biol Evol. 21:670–680.

Small KS, Brudno M, Hill MM, Sidow A. 2007a. Extreme genomic variation in a natural population. Proc Natl Acad Sci U S A. 104:5698–5703.

Small KS, Brudno M, Hill MM, Sidow A. 2007b. A haplome alignment and reference sequence of the highly polymorphic Ciona savignyi genome. Genome Biol. 8:R41.

Smit AFA, Hubley R, Green P. 1996-2010. RepeatMasker. Version 3.2.8. Seattle (WA): Distributed by the authors, Institute for Systems Biology, Available from: http://repeatmasker.org.

Sodergren E, et al. 2006. The genome of the sea urchin Strongylocentrotus purpuratus. Science. 314:941–952.

Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 33:W465–W467.

Telford MJ, Holland PW. 1997. Evolution of 28S ribosomal DNA in chaetognaths: duplicate genes and molecular phylogeny. J Mol Evol. 44:135–144.

Venter JC, et al. 2001. The sequence of the human genome. Science. 291:1304–1351.

Vilella AJ, Blanco-Garcia A, Hutter S, Rozas J. 2005. VariScan: analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. Bioinformatics. 21:2791–2793.

Vinson JP, et al. 2005. Assembly of polymorphic genomes: algorithms and application to Ciona savignyi. Genome Res. 15:1127–1135.

Wagner A. 2005. Robustness and evolvability in living systems. Princeton (NJ): Princeton University Press.

Wasserman WW, Sandelin A. 2004. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet. 5:276–287.

Wicker T, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8:973–982.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet. 8:206–216.

Xing J, et al. 2009. Mobile elements create structural variation: analysis of a complete human genome. Genome Res. 19:1516–1526.

Associate editor: Michael Purugganan