Molecular Phylogenetics and Evolution 67 (2013) 266–276

Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

# Convergent intron gains in hymenopteran elongation factor-1α

Seraina Klopfstein *, Fredrik Ronquist

*Swedish Museum of Natural History, Box 50007, SE-104 05 Stockholm, Sweden*

ARTICLE INFO

*Article history:*
Received 4 September 2012
Revised 17 January 2013
Accepted 29 January 2013
Available online 7 February 2013

*Keywords:*
eEF1A
Elongation factor 1-α
Hymenoptera
Intron evolution
Ancestral state reconstruction

ABSTRACT

The eukaryotic translation elongation factor-1α gene (eEF1A) has been used extensively in higher level phylogenetics of insects and other groups, despite being present in two or more copies in several taxa. Orthology assessment has relied heavily on the position of introns, but the basic assumption of low rates of intron loss and absence of convergent intron gains has not been tested thoroughly. Here, we study the evolution of eEF1A based on a broad sample of taxa in the insect order Hymenoptera. The gene is universally present in two copies – F1 and F2 – both of which apparently originated before the emergence of the order. An elevated ratio of non-synonymous versus synonymous substitutions and differences in rates of amino acid replacements between the copies suggest that they evolve independently, and phylogenetic methods clearly cluster the copies separately. The F2 copy appears to be ancient; it is orthologous with the copy known as F1 in Diptera, and is likely present in most insect orders. The hymenopteran F1 copy, which may or may not be unique to this order, apparently originated through retroposition and was originally intron free. During the evolution of the Hymenoptera, it has successively accumulated introns, at least three of which have appeared at the same position as introns in the F2 copy or in eEF1A copies in other insects. The sites of convergent intron gain are characterized by highly conserved nucleotides that strongly resemble specific intron-associated sequence motifs, so-called proto-splice sites. The significant rate of convergent intron gain renders intron–exon structure unreliable as an indicator of orthology in eEF1A, and probably also in other protein-coding genes.



## 1. Introduction

Mistaking paralogs for orthologs can cause drastic errors in phylogenetic inference. Although methods are available to account for the duplication history of a gene and even use the associated information for phylogenetic reconstruction (Boussau and Daubin, 2010), such methods require data on a genomic scale, and are not applicable to non-model organisms for which very little molecular data are available. Traditional phylogenetic studies, especially among non-model organisms, thus relied on either single-copy genes, mitochondrial genes or genes with a very high level of concerted evolution between the copies (e.g. rRNA).

A notable exception is the elongation factor 1α (eEF1A, also referred to as EF-1α), a protein responsible for delivering aminoacylated tRNAs to the ribosome during translation (Andersen et al., 2003). Primers for eEF1A were available early on during the history of molecular phylogenetics and it has been used extensively in phylogenetic studies of insects and other arthropods, at a range of taxonomic levels (Danforth et al., 2006b; Giribet et al., 2001; Kjer et al., 2001; Klopfstein et al., 2011; Pilgrim et al., 2008; Simon et al., 2010). In an attempt to unify the efforts in insect phylogenetics, Caterino et al. (2000) proposed to include eEF1A in a canon of genes for insect systematics, and it is to date one of very few nuclear protein-coding gene for which sequences are available from most arthropod groups. However, multiple copies of eEF1A were soon found in various taxa, including Hemiptera (Downie and Gullan, 2004), Neuropterida (Haring and Aspöck, 2004), Thysanoptera (Morris et al., 2002), and the three largest insect orders Coleoptera (Jordal, 2002; Ruiz et al., 2009), Diptera (Hovemann et al., 1988), and Hymenoptera. In the latter, a second copy was found in various members of the superfamily Apoidea and in other Aculeata (Brady and Danforth, 2004; Brady et al., 2011; Danforth et al., 2006a,b; Danforth and Ji, 1998), and the duplication event was even used in an attempt to root the bee tree (Brady et al., 2011). More recently, it was also reported from *Nasonia*, a parasitic wasp whose genome has been sequenced (Niehuis et al., 2007).

Orthology determination within Hymenoptera was perceived as straightforward, as the two copies – usually called F1 and F2 – are clearly divergent in their coding sequences, and each was believed to show a consistent intron–exon structure. Outside Hymenoptera, orthology is still not satisfactorily resolved. The hymenopteran F2 copy is generally assumed to be the ancestral copy and orthologous

* Corresponding author. Address: Naturhistoriska Riksmuseet, Department of Biodiversity Informatics, Box 50007, SE 104 05 Stockholm, Sweden. Fax: +46 (0)8 519 540 99.

*E-mail addresses:* klopfstein@nmbe.ch (S. Klopfstein), fredrik.ronquist@nrm.se (F. Ronquist).

1055-7903 
http://dx.doi.org/10.1016/j.ympev.2013.01.015

to the single copies found in many insect orders, as well as to the C2 copy in Coleoptera and to the so-called F1 copy in Diptera (the naming of the latter being rather unfortunate). The relationships of the hymenopteran F1 copy with other paralogs, however, remain unclear. Some authors assume that duplication events were independent in each insect order, implying that the hymenopteran F1 copy is unique to Hymenoptera (Simon et al., 2010). Mainly based on the presence of an intron not found in any other paralog, the hymenopteran F1 copy has also been associated with the dipteran F2 copy, but the latter also shares an intron position with the hymenopteran F2, and previous phylogenetic analyses of the coding sequence remained inconclusive (Danforth and Ji, 1998; Jordal, 2002). Given that F1 was reported from relatively few hymenopterans previously, it was also quite possible that it originated within a subclade of the order.

Relying upon intron–exon structures to ascertain orthology is based on the assumption that intron–exon patterns in eEF1A in arthropods are sufficiently stable; intron losses must be rather rare, and, more importantly, parallel intron gains at identical positions in the coding region must not occur (Djernæs and Damgaard, 2006). Indeed, introns were long believed to mostly originate from a period very early in eukaryotic history. This so-called "intron-early" hypothesis assumed that most of the differences in intron–exon structure observed among living organisms stems from differential intron loss in different groups (Logsdon, 1998). While it is not easy to prove that introns have originated recently, convincing evidence is today available from population genomic analyses (Li et al., 2009; Torriani et al., 2011) and from near intron pairs, i.e. introns found in related organisms which cannot possibly have co-occurred in their ancestor simply because the exon between them would have been too short for accurate splicing (Lehmann et al., 2010). To prove convergent appearance of two introns at identical positions is even more difficult, and evidence here stems mainly from introns with a very disparate taxonomic distribution, e.g. an intron occurring in plants and in *Drosophila* but not in any other animals (Tarrío et al., 2003). However, convergent intron gains at identical positions actually seem plausible since it is known that the exonic positions flanking introns throughout the studied genomes are not random, but instead represent highly conserved motifs, the so-called proto-splice sites (Dibb and Newman, 1989), which might be both sites of preferential intron insertion and retention (e.g. Babenko et al., 2010; Cohen et al., 2012; Lehmann et al., 2010; Lim and Burge, 2001; Ruvinsky and Ward, 2008).

Given the prevalent use of eEF1A in arthropod and especially insect phylogenetics and the lack of knowledge concerning the origin of the multiple copies and their introns, we studied the evolutionary history of this gene in one of the largest insect orders, the Hymenoptera. We take an explicit statistical approach based on phylogenetic methods and a broad taxonomic sample. After establishing that both the F1 and the F2 copies occur not only in the derived groups from which they were previously reported but throughout the order, we demonstrate that the intron–exon structure is not as conserved as previously thought. In particular, we show that the F1 copy, which apparently originated through retroposition, was originally intron free. It has successively accumulated introns in the Hymenoptera, three of which have appeared at the precise positions occupied by introns in the F2 copy or in eEF1A copies in other insects. The sites of convergent intron gain belong to a small set of eEF1A sites characterized by conserved nucleotides that strongly resemble motifs known to facilitate intron gain or retention. We conclude that molecular phylogeneticists need to be aware that the significant rate of convergent intron gain renders intron–exon structure unreliable as an indicator of orthology in elongation factor-1α, and likely also in many other protein-coding nuclear genes.

## 2. Materials and methods

### 2.1. Taxon sampling and molecular methods

We included 108 species from 68 families and all the 22 superfamilies of Hymenoptera, and 23 outgroup sequences from a variety of insect orders (Supplementary file 1). The taxon sampling spanned the whole hymenopteran radiation and specifically included the most basal families. Outgroup sampling covered two paralogs each in the orders Coleoptera and Diptera. We aimed to amplify a fragment of more than 1 kb, spanning bases 219–1261 of the coding region of the Drosophila F1 copy (GenBank accession X06869). We first tried to amplify this genomic region in two parts by standard PCR approaches, using a variety of primers (Supplementary file 2). PCR protocols in general followed Heraty et al. (2011), or we used PuReTaq PCR beads (GE healthcare) according to the protocol of the manufacturer, with 1.5 μL of 10 mM primer added to a final volume of 20 μL. A typical PCR cycle started with 4 min of denaturation at 94 °C, followed by 35–38 cycles of 1 min at 94 °C, 1 min at the primers' annealing temperature, and 1.5 min of elongation at 72 °C, and ended with a terminal elongation period of 5 min. Touch-down PCRs were chosen for some of the more degenerate primers, and started with 2 cycles at annealing temperature + 4 °C, 2 cycles at annealing temperature + 2 °C, and 18–22 cycles at the primers' annealing temperature. For many taxa, direct PCR was unsuccessful, as the two copies were often superimposed and some of the highly degenerate primers also amplified unrelated fragments. In these cases we used semi-nested or nested PCR approaches, using primers HAF2for1 and Cho10mod in a first amplification round of 30 cycles. 0.5 μL of the PCR products were then used as template for a second PCR with primers for shorter regions (25–30 cycles). Even with this approach, specific amplification of either copy was often difficult, and we separated the two copies on an agarose gel. The difficulties to specifically amplify each copy persisted because much of the variation between the eEF1A copies in Hymenoptera is restricted to silent third codon positions, as most of the amino acid sequence is highly conserved even between the copies. Where differences in amino acids could be observed, they often involved an invariably conserved position in one copy, and a larger range of amino acids in the other copy (see results section). Copy-specific primers can thus only be developed within subgroups of Hymenoptera. Supplementary file 2 lists all primers used in this study, commenting on which were successful at amplifying single copies in certain groups. In the end, we obtained 87 sequences of F1 (80 of which newly for this study), and 96 (92 new) of F2.

### 2.2. Characteristics of the two eEF1A copies

To denote copy identity in Hymenoptera in comparison with copies found in outgroups, we use three letters from the name of the respective insect order as a prefix, e.g. "dip-F2" for the F2 copy found in Diptera, or "col-C1" for the C1 copy from Coleoptera. The naming of the copies in Diptera and Hymenoptera is unfortunate, as the dip-F1 copy is probably orthologous to hym-F2 and vice versa. If no prefix is used, we refer to a hymenopteran copy.

Initially, sequences were assigned to either hym-F1 or hym-F2 based on the presence or absence of two introns previously thought to be diagnostic for either copy. As this approach proved uninformative or even contradictory in numerous cases (see results), we assigned sequences by calculating pairwise *p*-distances of amino acid sequences from the coding sequences of F1 and F2 in *Apis mellifera*, which proved to be a conclusive approach. Phylogenetic methods were then applied to test whether this assignment resulted in monophyly of at least one of the copies. To