Molecular Phylogenetics and Evolution 86 (2015) 90–109

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Review

A survey of transposable element classification systems – A call for a fundamental update to meet the challenge of their diversity and complexity

Benoît Piégu [a], Solenne Bire [a,b], Peter Arensburger [a,c,*], Yves Bigot [a,*]

[a] UMR INRA-CNRS 7247, PRC, Centre INRA de Nouzilly, 37380 Nouzilly, France
[b] Institute of Biotechnology, University of Lausanne, Center for Biotechnology UNIL-EPFL, 1015 Lausanne, Switzerland
[c] Biological Sciences Department, California State Polytechnic University, Pomona, CA 91768, United States

ARTICLE INFO

Article history:
Received 24 December 2014
Revised 11 March 2015
Accepted 12 March 2015
Available online 20 March 2015

Keywords:
Transposon
Mobility
Host range
Ribozyme
Nuclease
Recombinase

ABSTRACT

The increase of publicly available sequencing data has allowed for rapid progress in our understanding of genome composition. As new information becomes available we should constantly be updating and reanalyzing existing and newly acquired data. In this report we focus on transposable elements (TEs) which make up a significant portion of nearly all sequenced genomes. Our ability to accurately identify and classify these sequences is critical to understanding their impact on host genomes. At the same time, as we demonstrate in this report, problems with existing classification schemes have led to significant misunderstandings of the evolution of both TE sequences and their host genomes. In a pioneering publication Finnegan (1989) proposed classifying all TE sequences into two classes based on transposition mechanisms and structural features: the retrotransposons (class I) and the DNA transposons (class II). We have retraced how ideas regarding TE classification and annotation in both prokaryotic and eukaryotic scientific communities have changed over time. This has led us to observe that: (1) a number of TEs have convergent structural features and/or transposition mechanisms that have led to misleading conclusions regarding their classification, (2) the evolution of TEs is similar to that of viruses by having several unrelated origins, (3) there might be at least 8 classes and 12 orders of TEs including 10 novel orders.

In an effort to address these classification issues we propose: (1) the outline of a universal TE classification, (2) a set of methods and classification rules that could be used by all scientific communities involved in the study of TEs, and (3) a 5-year schedule for the establishment of an International Committee for Taxonomy of Transposable Elements (ICTTE).

Contents

---

* Corresponding authors at: Biological Sciences Department, California State Polytechnic University, Pomona, CA 91768, United States (P. Arensburger). UMR INRA-CNRS 7247, PRC, Centre INRA de Nouzilly, 37380 Nouzilly, France (Y. Bigot).

E-mail addresses: parensburger@csupomona.edu (P. Arensburger), yves.bigot@tours.inra.fr (Y. Bigot).

http://dx.doi.org/10.1016/j.ympev.2015.03.009
1055-7903/

0. Introduction

Any collection of objects, including biological entities, may be classified in multiple ways in order to create groups based on phenotypic features (Mayr and Bock, 2002). Within scientific disciplines examples of such classifications range from the periodic table to the Enzyme Commission number system (Webb, 1992). Within the biological sciences the principle of shared common ancestry is so widely used as a classification criterion that many classifications in this field are assumed (sometimes incorrectly) to incorporate this criterion. Biological classification, a subfield of the study of systematics, is the grouping of species on the basis of evolutionary relationships (Daly et al., 2012). Mayr and Bock (2002) defined species classification as "The arrangement of entities in a hierarchical series of nested classes, in which similar or related classes at one hierarchical level are combined comprehensively into more inclusive classes at the next higher level". The classification of most living organisms has been codified by four international codes of nomenclature: one for animals (Ride et al., 2000); one for algae, fungi and plants (McNeill et al., 2012); one for prokaryotes (Lapage et al., 1992) and one for viruses (King et al., 2011). All four codes share several organizational levels including kingdom, phylum/division, class, order, family, genus, and species. The placement of individuals within these levels implies a series of evolutionary relationships that will often be used as a basis for subsequent research. For the working scientist a well organized biological classification provides the following four advantages: (1) it simplifies the identification of unknown organisms, (2) it reveals connections between groups of closely related organisms, (3) it indicates evolutionary relationship, and (4) it allows the integration of data from a few representatives from distinct groups into a web connection of all living organisms. Methods and criteria used to establish biological classifications have changed over time as evolutionary concepts and technical innovations have progressed. Most recently, phylogenetic analyses from protein and DNA sequences have had a significant impact on classification schemes. Over the last decade debates regarding the classification of some groups, such as viruses, have been the subject of passionate exchanges of views. Indeed, within the virus community discussions range from the definition of what a virus species is (van Regenmortel et al., 2013) to the possibility that certain viruses might represent a distinct fourth domain of life (Bandea, 2009; Williams et al., 2011; Philippe et al., 2013; Pennisi, 2013; Raoult, 2013). Finally, these discussions are complicated by the connection between viruses and a number of mobile genetic elements (more commonly referred to as transposable elements, TEs) that have been characterized in both prokaryotic and eukaryotic genomes (Weiss, 2006; Stoye et al., 2012; Desnues et al., 2012; Yutin et al., 2013). The classification of these TEs is the subject of this review and we begin by examining how to define a TE species.

TEs represent most of the interspersed repeats in the genomes of prokaryotes and eukaryotes. It is therefore striking that a simple but comprehensive definition of what constitutes a TE is not easily found in the literature. Haren et al. (1999) proposed that "TEs are discrete segments of DNA capable of moving from one locus to another in their host genome or between different genomes". Similarly, Kidwell and Lisch (2001) stated that "TEs are DNA sequences that have the capacity to change genomic locations". Since their publication, the above definitions have been widely used in the literature. Currently, as knowledge of the diversity of the TE and virus worlds has grown extensively we would suggest that these definitions could rephrased as (based in part on the evidence we present below) "TEs are discrete segments of DNA capable of moving within a host genome from one chromosome or plasmid location to another and which do not use a specific molecular machinery that they encode to infect the genome of new hosts by lateral transfer". An important aspect of any TE definition is that it includes mobile DNA sequences that are primarily maintained by vertical transmission as copies integrated into the chromosomes or plasmids of their hosts. Therefore, our amended definition considers that viruses, phages, and integrative conjugative elements (ICE) have similar features to TEs but they are not considered TEs since they are able to move between hosts independent of transmission vectors. To round out our proposed TE definition, it should also be understood that the state of TE copies within a host genome varies depending on the age and activity of the element. Autonomous TEs encode the enzymes required for their mobility while non-autonomous elements depend for their mobility on enzymes supplied by autonomous elements belonging to the same or a related element. TE sequences in a genome accumulate mutations over time which will most often inactivate the ability of these sequences to mobilize further. This ageing process has led to the presence, in most genomes, of many fossil TE sequences alongside a few active copies (Kidwell and Lisch, 2001).

In this report we review various TE classification systems, with particular attention to how each system has affected the development of TE biology. We also examine how these systems have held up in light of the exponential increase in genomic data.

This manuscript is organized into three sections. We begin by reviewing existing TE classifications and outline their respective strengths and weaknesses. Next we describe a number of TE sequences that have not been included in some TE classification systems. Finally, we outline a proposal for an international committee to help draft a unified TE classification.

1. History of existing TE classifications

1.1. TE classification pioneer

Finnegan (1989, 1992) launched the field of TE systematics based largely on what was previously proposed in this field from human, drosophila, and yeast models (Singer, 1982; Boeke, 1989; Finnegan and Fawcett, 1986). His proposal was that TEs could be