Topics in Current Genetics 24

Marie-Angèle Grandbastien
Josep M. Casacuberta *Editors*

Plant Transposable Elements

Impact on Genome Structure and Function

Springer

Topics in Current Genetics **24**

Series Editor: *Stefan Hohmann*

For further volumes:
http://www.springer.com/series/4735

Marie-Angèle Grandbastien • Josep M. Casacuberta
Editors

Plant Transposable Elements

Impact on Genome Structure and Function

Springer

Editors
Marie-Angèle Grandbastien
Institut Jean Pierre Bourgin
UMR 1318 INRA/AgroParisTech
INRA-Versailles
78026 Versailles, France
e-mail: gbastien@versailles.inra.fr

Josep M. Casacuberta
Department of Molecular Genetics
Center for Research in Agricultural Genomics (CRAG)
CSIC-IRTA-UAB-UB
Campus UAB
Bellaterra - Cerdanyola del Vallés
08193 Barcelona, Spain
e-mail: josep.casacuberta@cragenomica.es

ISBN 978-3-642-31841-2 ISBN 978-3-642-31842-9 (eBook)
DOI 10.1007/978-3-642-31842-9
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012956160

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Series description

Topics in Current Genetics publishes review articles of wide interest in volumes that center around a specific topic in genetics; genomics; as well as cell, molecular, and developmental biology. Particular emphasis is placed on the comparison of several model organisms. Volume editors are invited by the series editor for special topics, but further suggestions for volume topics are highly welcomed. Each volume is edited by one or several acknowledged leaders in the field, who ensure the highest standard of content and presentation. All contributions are peer reviewed. All volumes of Topics in Current Genetics are part of the Springer eBook Collection. The collection includes online access to more than 3,500 newly released books, book series volumes, and reference works each year. In addition to the traditional print version, this new, state-of-the-art format of book publications gives every book a global readership and a better visibility.

Preface

Transposable elements (TEs) are ubiquitous mobile DNA sequences found in both prokaryotic and eukaryotic genomes. They are able to insert at different positions of the genome, either by excising from one position and reinserting into another or by replicating into daughter copies. TEs are particularly abundant in plant genomes, where they can represent over 80 % of the bulk of large cereal genomes. Their discovery by B. McClintock, and the subsequent introduction of the notion of genome fluidity, was a major shift in our concepts on heredity. TEs can dramatically modify the structure of host genomes, affect genome sizes and generate genetic variation, not only by transposition but also by providing the raw material for genomic rearrangements due to their repetitive nature. Until recently, and in spite of B. Mc Clintock's seminal concept of "Controlling Elements," the impact of TEs on host genome function was merely regarded as circumstantial. A rather different representation has been brought to light in the last decade, which strongly argues that TEs may also act as pivotal factors in generating genic variation and modulating cellular gene expression. This book is intended at presenting the latest advances on the importance of TEs and on their impact on plant genome dynamics and function.

The TE research scene has recently seen major advances, with new tools such as Next Generation Sequencing (NGS) technologies opening tremendous possibilities for rapid global analyses of genomes at reduced costs. This has led to an exponential increase in the amount of TE-related data and to a deeper knowledge of their impact on host genomes. As a consequence, all plant researchers engaged in genomic studies are more or less unwillingly bumping into this wealth of TEs and are now realizing that these TEs cannot be discarded as annoying junk sequences anymore. TEs are encountered in both genomic and transcriptomic data, and in a tremendous variety of elements, including highly defective and deleted versions sometimes mobilized at surprisingly high levels via related copies, making their classification a difficult task. There is therefore a need for researchers to find guidelines to recognize and classify TEs and better understand their importance and potential impact.

This book is intended both for scientists familiar with the field and for nonspecialists. It is organized in 14 chapters written by recognized researchers and is centered, on one hand on how to recognize and study plant TEs, notably using NGS technologies, and, on the other hand, on how TEs impact plant genome structure and genome function, with a few final examples of exciting TE-mediated phenotypic impacts. The first few chapters cover important aspects of what are TEs and how they can be identified and analyzed. Chapter 1 covers recent developments in TE classification and annotation and tackles the complex issue of defining consistent guidelines, while Chap. 2 summarizes and compares computational tools available for TE identification and provides a road map for efficient annotation. Chapter 3 then explores how NGS technologies can be used to study TE-mediated genome size variations and evolutionary patterns that shape the TE compartment, and Chap. 4 describes the recent development of NGS technologies to monitor TE mobility. The three following chapters provide further insights on some of the best known plant TEs. Chapter 5 describes the predominant type of TEs found in plant genomes, the LTR retrotransposons, and the subtle functional interplay between their autonomous and nonautonomous versions, while Chap. 6 explores the intriguing possibility of the existence of plant endogenous retroviruses, and Chap. 7 updates our knowledge on the highly abundant miniature elements, MITEs, and their impact on plant genomes. Chapter 8 summarizes the current state of affairs for epigenetic mechanisms developed by plant genomes to control TE mobility and highlights the plasticity of these mechanisms. The two following chapters address the important issue of TEs in polyploid contexts: Chap. 9 summarizes current knowledge on TE involvement in the drastic structural and functional changes resulting from allopolyploidy, a major speciation process in the plant kingdom, while Chap.10 compares the nature and evolution of TEs between polyploid sugarcane and other grass genomes. The four following chapters are dedicated to several striking mechanisms by which TEs have been exapted by host genomes to distil invaluable tools for modifying genome function. Chapter 11 describes how a fascinating type of TEs, Helitrons, can capture gene fragments and describes how such process can lead to new regulatory functions, and Chap. 12 reviews in detail how plant TE coding sequences have been frequently domesticated into functional cellular genes. Chapter 13 assesses current knowledge on the ubiquitous process of SINE exaptation for the production of regulatory RNAs, and Chap. 14 updates current data on plant LTR retrotransposon stress response and examines the possibility that LTRs could play a role in modulating host gene expression. Finally, the last two chapters present particularly striking examples of TE-associated phenotypic changes. Chapter 15 illustrates the role of the Rider LTR retrotransposon in several morphological and physiological changes in tomato, while Chap. 16 describes how small RNAs produced by a non-LTR retrotransposon are involved in the desiccation tolerance of resurrection plants.

The chapters were conceived and written autonomously, so that they can be read independently, even though this may have resulted in a few redundancies. Many other topics could have been covered, and many other beautiful examples of TE impact on plant genomes could have been exposed, however it was impossible to assemble all of the chapters that we would have liked to have in this volume, due to

lack of space. Nevertheless, we feel that the 14 chapters presented in this book provide altogether a global overview of the most interesting current advances in the field of plant TE studies, while providing a useful reference vademecum volume for all (highly welcomed!) newcomers to the field. We hope that they will feel the urge to better understand what are these repetitive sequences that compose more than half of their data and that, after consulting this book, they will become convinced that Transposable Elements are certainly not "junk," but may actually be by far the most interesting and fun part of their data!

Finally, we wish to heartily thank all authors of this volume, that all have made substantial efforts to share our common passion with you and to provide excellent contributions. We also thank Stefan Hohmann for providing us the opportunity to compile this volume, the staff at Springer Verlag for their continuous help and support to make this book possible, and Tom Bureau for correcting this text.

September 2012
Versailles, France — Marie-Angèle Grandbastien
Barcelona, Spain — Josep M. Casacuberta

Contents

Chapter 1
So Many Repeats and So Little Time: How to Classify Transposable Elements

Thomas Wicker

Abstract Transposable elements (TEs) are present in all genomes. Often there are hundreds to thousands of different TE families contributing the majority of the genomic DNA. Although probably only a very small portion of TEs actually contributes to the function and thereby to the survival of an organism, they still have to be analysed, annotated and classified. To filter out the scarce meaningful signals from the deluge of data produced by modern sequencing technologies, researchers need to be able to efficiently and reliably characterise TE sequences. This process requires three things: First, clear guidelines how to classify and characterise TEs. Second, high-quality databases that contain well-characterised reference sequences, and third, computational tools for efficient TE searches and annotations. This article is intended as a summary of recent developments in TE classification as well as a "little helper" for researchers burdened with the epic task of TE annotation in genomic sequences.

Keywords Transposable element • Retrotransposon • DNA transposon • Superfamily • Family • Classification

1.1 Introduction

1.1.1 Early Findings on Genome Sizes and Sequence Complexity

Even before DNA could be sequenced, researchers realised that eukaryotic genomes show an extreme variation in size (Bennett and Smith 1976). Some studies reported an over 200,000-fold variation in genome size, namely between the amoeba *Amoeba dubia* that has an estimated genome size of 670,000 Mbp (Gregory

T. Wicker (✉)
Institute of Plant Biology, University of Zurich, Ollikerstrasse 107, CH-8008 Zurich, Switzerland
e-mail: wicker@botinst.uzh.ch

M.-A. Grandbastien and J.M. Casacuberta (eds.), *Plant Transposable Elements*, Topics in Current Genetics 24, DOI 10.1007/978-3-642-31842-9_1,

2001) and the 2.9 Mbp genome of the microsporidium *Encephalitozoon cuniculi* (Biderre et al. 1995; Katinka et al. 2001). In the absence of DNA sequence information, genome sizes were measured by estimating nuclear DNA amounts through densitometric measurements (e.g. Bennett and Smith 1976). The "sequence complexity" of genomes was assessed by DNA re-association kinetics. These experiments showed that the vast differences in genome sizes are due to the presence of different amounts of "repeating DNA sequences" (Britten et al. 1974), although their nature was completely unknown at that time. Nevertheless, it was clear early on that the repetitive fraction of a genome is relatively complex and consists of many different types of repeats. Genomes could even be fractionated into highly and moderately repetitive sequences by DNA re-association kinetics (Peterson et al. 2002).

1.1.2 Definition of "Gene Space" and the "C-Value Paradox"

Only when technological advances allowed near-complete sequencing of eukaryotic genomes, actual gene numbers could finally be estimated. Here, it needs to be noted that the definition of what actually constitutes the "gene space" of a genome is still a topic of debate. It certainly includes all "typical" protein-coding genes. Additionally, many components of the gene space do not encode proteins, such as the highly repetitive ribosomal DNA clusters, tRNAs and small nucleolar and small interfering RNAs. Probably, gene space should also include conserved non-coding sequences (Freeling and Subramaniam 2009) and ultraconserved elements (Bejerano et al. 2004), although their functions are barely understood. In the following discussion of gene numbers, I will only refer to protein-coding genes.

1.1.3 The Number of Genes is Similar in All Genomes

As Table 1.1 shows, the estimates of gene numbers differ from species to species, but for all sequenced eukaryotic genomes they are in a range from 5,000 to 50,000. Thus, at a first glance, gene numbers vary only by a factor of 10 while genomes sizes, as described above, vary more than 200,000-fold. The recently finished genome of *Brachypodium distachyon* probably has the most stringent gene annotation so far and possesses 25,554 genes. This gene number is very similar to that of the most recent version of the *Arabidopsis thaliana* genome (version 9) that has 26,173 annotated genes. Even the large maize genome is estimated to contain only about 30,000 genes (Schnable et al. 2009). Interestingly, these numbers are very similar to those for vertebrate genomes, because for all sequenced vertebrate genomes, such as human, mouse, or chicken, genes numbers are now estimated in the range of 25,000–30,000 (Table 1.1). Only fungi and invertebrate animals have clearly fewer genes. Yeast, with its compact 12 Mbp genome has less than 6,000 genes while insects such as *Anopheles gambiae* or *Drosophila melanogaster* have approximately 12,000 genes

Table 1.1 Genome sizes and gene numbers in publicly available genomes

	Size [Mbp]	Genes	Reference
Animal genomes			
Anopheles gambiae	278	14,000	Holt et al. (2002)
Caenorhabditis elegans	97	19,000	CSC (1998)
Drosophila melanogaster	120	15,200	Adams et al. (2000)
Gallus gallus	1,200	20,000–23,000	ICGSC (2004)
Homo sapiens	2,850	24,000	IHGSC (2004)
Mus musculus	2,500	30,000	MGSC (2002)
Plant genomes			
Arabidopsis thaliana	120	26,200	AGI (2000)
Brachypodium distachyon	273	25,500	IBI (2010)
Fritillaria uva-vulpis	87,400	unknown	Leitch et al. (2007)
Hordeum vulgare	5,700	38,000–48,000	Mayer et al. (2009)
Oryza sativa	372	40,600	IRGSC (2005)
Physcomitrella patens	462	35,900	Rensing et al. (2008)
Populus trichocarpa	410	45,500	Tuskan et al. (2006)
Sorghum bicolor	659	34,500	Paterson et al. (2009)
Triticum aestivum	16,000	50,000	Choulet et al. (2010)
Vitis vinifera	342	30,400	Jaillon et al. (2007)
Zea mays	2,061	30,000	Schnable et al. (2009)
Fungal genomes			
Aspergillus nidulans	30	10,600	http://www.broadinstitute.org
Aspergillus flavus	36.8	12,600	http://www.broadinstitute.org
Fusarium verticilloides	41.8	14,200	http://www.broadinstitute.org
Magnaporthe grisea	42	11,100	Dean et al. (2005)
Saccharomyces cerevisiae	11.7	5,700	http://www.broadinstitute.org
Stagonospora nodurum	37	16,600	http://www.broadinstitute.org
Tuber melanosporum	125	7,500	http://www.broadinstitute.org
Botrytis cinerea	42.6	16,400	http://www.broadinstitute.org
Other genomes			
Encephalitozoon cuniculi	2.9	1,997	Katinka et al. (2001)
Amoeba dubia	670,000	unknown	Gregory et al. (2001)

AGI Arabidopsis genome initiative, *CSC C. elegans* sequencing consortium. *IBI* International Brachypodium initiative, *ICGSC* International chicken genome sequencing consortium, *IHGSC* International human genome sequencing consortium, *IRGSP* International rice genome sequencing consortium, *MGSC* Mouse genome sequencing consortium

(Table 1.1). Thus, a consensus transpires that most eukaryotes possess between 5,000 and 30,000 genes, making it obvious that only a relatively small fraction of the genomes sequenced to date actually encode functional genes.

1.1.4 The C-Value Paradox

The fact that gene numbers are very similar while genome sizes vary extremely came to be known as the "C-value Paradox". Moreover, depending on which taxonomic group is analysed, there may be little or no correlation between genome

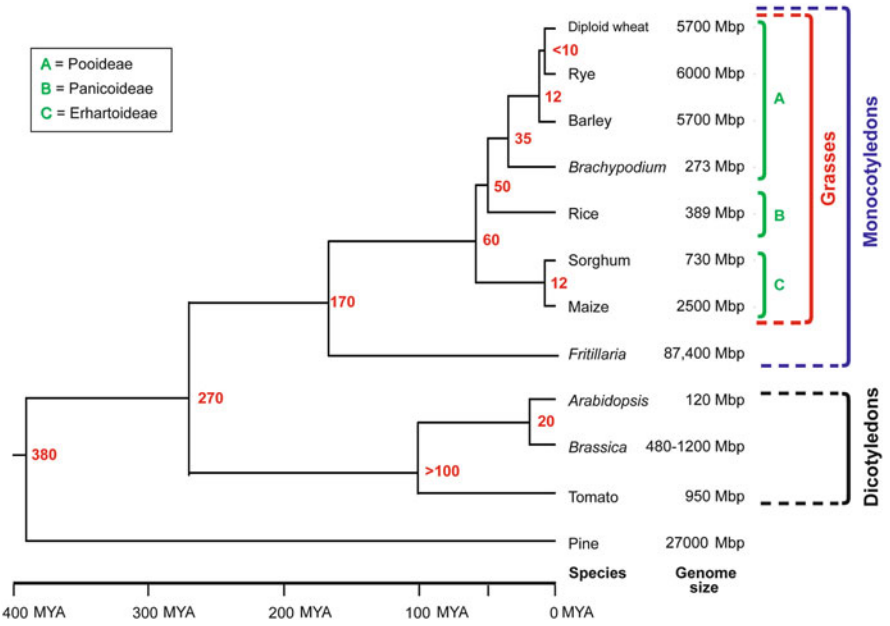

Fig. 1.1 Phylogenetic relationships and genome sizes in selected plant species. Divergence times of specific clades are indicated in red numbers next to the corresponding branching. These numbers are averages of the published values provided in Table 1.1. The scale at the bottom indicates divergence times in million years ago (MYA). Major taxonomic groups that are discussed in the text are indicated at the *left*

size and phylogenetic relationships. This effect is particularly strong on plants where even very closely related species can have very different genome sizes (Fig. 1.1). Among the dicotyledonous plants, there is *Arabidopsis thaliana*, the first plant which had its genome completely sequenced. With a size of about 120 Mbp (*Arabidopsis Genome Initiative* 2000), it is one of the smallest plant genomes known. In contrast, closely related *Brassica* species that diverged from *Arabidopsis* only 15–20 MYA (Yang et al. 1999) have five to ten times larger genomes. In monocotyledonous plants, variation is even more extreme: The grasses *Brachypodium dystachion*, rice and sorghum have genome sizes of 273 Mbp, 389 Mbp and 690 Mbp, respectively, considerably larger than the *Arabidopsis* genome but roughly an order of magnitude smaller than the genomes of some agriculturally important grass species such as wheat and maize, with haploid genome sizes of 5,700 and 2,500 Mbp, respectively. And even they are still dwarfed by the genomes of some lilies, among them *Fritillaria uva-vulpis* which has a genome size of more than 87,000 Mbp, over 700 times the size of the *Arabidopsis* genome (Leitch et al. 2007). Also among *Dicotyledons*, closely related species often differ dramatically in their genome sizes. Maize and sorghum, for example diverged only about 12 MYA (Swigonova et al. 2004), but the maize genome is more than four times the size of the sorghum genome (Table 1.1, Fig. 1.1).

1.2 Transposable Elements

1.2.1 Basics of Selfishness and Junk

As the number of genes is similar in all organisms, it became clear early on that the factor which mainly determines genome size is the amount of repetitive sequences. Nowadays we know that the vast majority of these repetitive sequences are in fact transposable elements (TEs). These elements contain no genes with apparent importance for the immediate survival of the organism. Instead they contain just enough genetic information to produce copies of themselves and/or move around in the genome. For this reason, such sequences are often referred to as "selfish" DNA (Orgel and Crick 1980). To some degree that disparaging view is justified, because TEs are small genetic units, actual "minimal genomes", which contain exactly enough information to be able to replicate, move around in the genome or both. They use the DNA replication and translation machinery of their "host" and thrive within the environment of the genome. For this reason, the term "junk DNA", is often used almost synonymously with TE sequences, reflecting the view of TEs being largely a parasitic burden to the organism.

1.2.2 TE Taxonomy and Classification

Pioneering work in TE classification was done by Hull and Covey (1986), Finnegan (1989) and Capy et al. (1996). The first publicly available database for TEs was RepBase (girinst.org/repbase/) by Jerzy Jurka and colleagues who also proposed a classification system for all TEs (Jurka et al. 2005). In 2007, a group of TE experts met at the Plant and Animal Genome Conference in San Diego (CA, USA) with the goal to define a broad consensus for the classification of all eukaryotic transposable elements. This included the definition of consistent criteria in the characterisation of the main superfamilies and families and a proposal for a naming system (Wicker et al. 2007). The proposed system is a consensus of previous TE classification systems and groups all TEs into 2 major classes, 9 orders and 29 superfamilies (Fig. 1.2). A practical aspect of the classification system is that the TE family name should be preceded by a three-letter code for class, order and superfamily (Fig. 1.2). This was intended to make working with large sets of diverse TEs easier as it enables simple text-based sorting and allows the immediate recognition of the classification when seeing the name of a TE. The proposed classification system is open to expansion as new types of TEs might still be identified in the future. A system that attempts to cover such a vast and complex biological field is by its nature reductionist and tends to oversimplify matters. Thus, there is still an ongoing scientific debate about various aspects of the system (Kapitonov and Jurka 2008; Seberg and Petersen 2009), some of which will be discussed in more detail below.

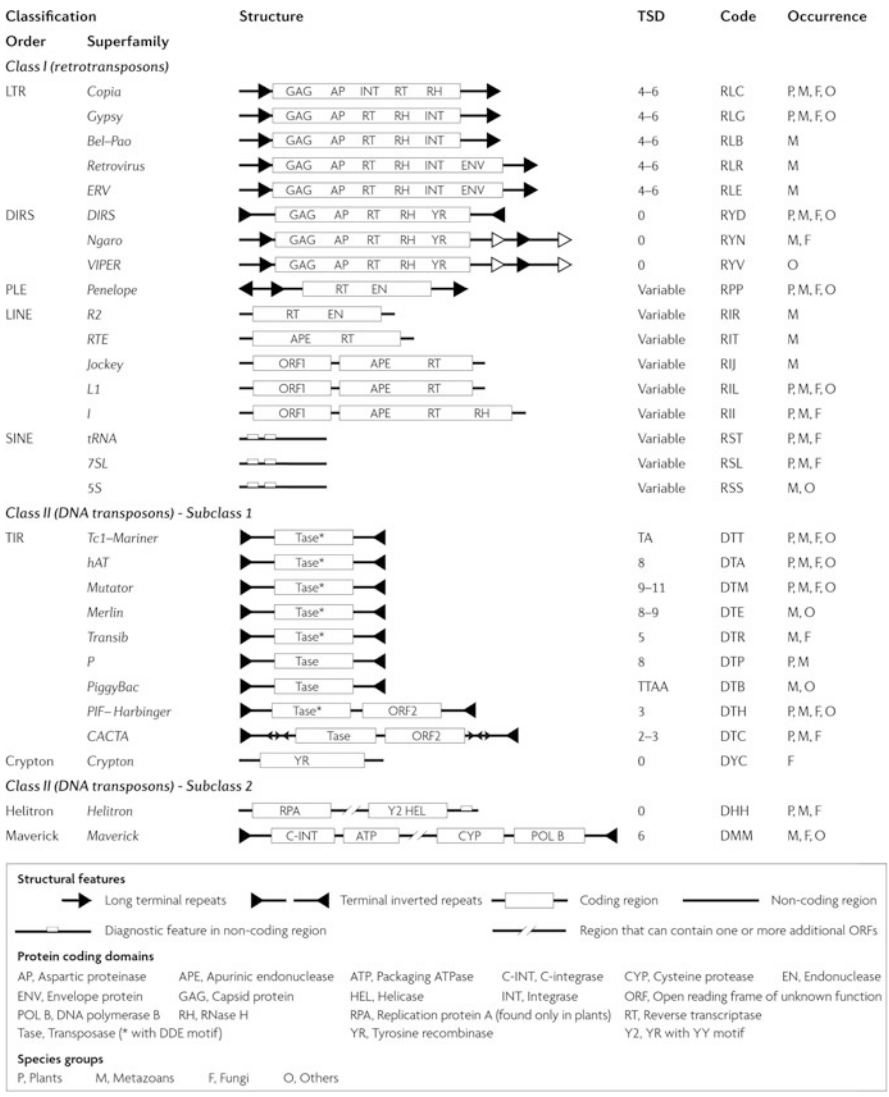

Fig. 1.2 Classification system for transposable elements (Wicker et al. 2007a). The classification divides TEs into two main classes on the basis of the presence or absence of RNA as a transposition intermediate. They are further subdivided into subclasses, orders and superfamilies. The size of the target site duplication (TSD), which is characteristic for most superfamilies, can be used as a diagnostic feature. A three-letter code describes all major groups and is added to the family name of each TE

1.2.3 Class and Subclass: The Highest Levels of TE Classification

At the highest taxonomic level, TEs are divided into two classes. Class 1 contains all TEs that replicate via an RNA intermediate in a "copy-and-paste" process. This class includes both LTR as well as non-LTR retrotransposons. In Class 2 elements,

the DNA itself is moved analogous to a "cut-and-paste" process. Class 2 elements are further subdivided into subclass 1 and 2. Subclass 1 are the classic cut-and-paste elements where the DNA is moved with the help of a transposase enzyme. Subclass 2 includes TEs whose transposition process entails replication without double-stranded cleavage and the displacement of only one strand. The Order Helitron from Subclass 2 seems to replicate via a rolling-circle mechanism (Kapitonov and Jurka 2001). Their placement within class 2 reflects the common lack of an RNA intermediate, but not necessarily common ancestry.

1.2.4 TE Superfamilies Represent Ancient Evolutionary Lineages

The most commonly used level of classification is the assignment of a TE to a particular superfamily. Superfamilies are ancient evolutionary lineages that arose during the very early evolution of eukaryotes, some even before the divergence of prokaryotes and eukaryotes. Superfamilies are mainly defined by homology at the protein level. That means that two TEs belong to the same superfamily if their predicted protein sequences show clear homology and can be aligned over most of their length. Terms like "clear homology" and "most of their length" reflect a plea to common sense and should not be tightly bound to arbitrary cut-offs based on E-Values or percent sequence similarity. The fact is that TEs belonging to the same superfamily (even if they come from very distantly related species) usually share many conserved amino acid motifs along the length of their predicted proteins which, importantly for practical work, is usually picked up in a blastx or blastp search. In contrast, TEs from different superfamilies usually show hardly any sequence similarity in their encoded proteins. Protein similarity between members of different superfamilies is reduced to very ancient sequence motifs such as the DDE or Zn-finger motifs (Capy et al. 1997). Here it has to be noted that sequence similarity within the same superfamily can only be expected in the "core" enzymes of the TE elements such as the transposase, reverse transcriptase or integrase, while fast-evolving proteins such as gag (in LTR retrotransposon) and ORF2 (in many DNA transposons) often cannot be aligned between members of the same superfamily. The superfamily of SINEs (small interspersed nuclear elements) has a special status. These small elements do not encode any proteins but are derived from RNA Polymerase promoters and can therefore only be classified based on specific DNA motifs.

1.2.5 TEs Show Most Diversity at the Family Level

It is at the family level is where things get really complicated. While the 29 superfamilies are relatively clearly defined, the exact definition of a TE family is still topic of debate (Kapitonov and Jurka 2008; Seberg and Petersen 2009).

It is clear that within superfamilies TEs have diverged in to an almost incomprehensibly large number of sub-groups and clades. Here, researchers usually introduce the family as the next lower level (after Superfamily). Early on, it became clear that there must be hundreds or even thousands of different types of TEs populating genomes (SanMiguel et al. 1998; Wicker et al. 2001). However, the challenge has been to define criteria for a family that, on one hand, make at least some biological sense and on the other hand are reasonably simple to apply. Of course, the most biologically meaningful TE classification would be based on phylogenetic analysis (Seberg and Petersen 2009). Construction of phylogenetic trees deduced from DNA or predicted protein sequences allows the identification of specific clades, and is therefore a classification scheme based on biological criteria. Such analyses are essential for our understanding of how TEs and genomes evolve. However, phylogenetic analyses are complex and very labour intensive and require a thorough knowledge of TEs, but they are relatively irrelevant when it comes to the initial task of TE identification and annotation, especially in large-scale genome projects.

1.2.6 The 80–80–80 Rule Revisited

In 2007, several colleagues and I proposed the "80–80–80" rule (Wicker et al. 2007) which became both famous and infamous among researchers working on TE annotation. The rule says that two TEs belong to the same family if they share at least 80 % sequence identity at the DNA level over at least 80 % of their total size. The third criterion simply refers to the minimal size of a putative TE sequence that should be analysed in order to avoid that unspecific signals are over-interpreted. The rule was mainly based on practical criteria. We assumed that most researchers on task to annotate TE sequences would need a simple guideline to classify TE sequences. In most cases, blastn (DNA against DNA) searches would be performed as a first step for TE identification. The BLAST algorithm is not able to align DNAs which are significantly less than 80 % identical. Thus, a given TE sequence will produce no strong BLASTN alignments if its sequence is significantly less than 80 % identical to sequences in the reference database. The second criterion (80 % of the entire length of the TE) was introduced to address the problem that different parts show different levels of sequence conservation within the same TE family. Most TEs are comprised of protein-coding sequences and regulatory regions. Good examples illustrating that problem are the long terminal repeat (LTR) retrotransposon superfamilies. The two LTRs contain promoter and downstream regions while the internal domain contains mainly protein-coding regions. Comparisons between many different TE families shows that the regulatory regions evolve much faster than the coding sequences. Thus, often the DNA sequences of the coding region might be alignable while up- and downstream regions (e.g. LTRs) are completely diverged and cannot be aligned. The second criterion of the 80–80–80 rule requires that at least some of the regulatory sequences can be aligned at the

DNA level. There is at least some biological justification for the 80/80 rule, as elements which are similar at the DNA level must have originated from a common "mother" copy in evolutionary recent times.

1.2.7 Biological Meaning vs. Pragmatism in TE Classification

It is clear that a classification rule based simply on the fact that DNA sequences can be aligned is arbitrary, and it was justifiably criticised (Kapitonov and Jurka 2008; Seberg and Petersen 2009). Indeed, TE families (we shall stick to the term "family" for this discussion) sometimes form a continuum, where a sequence from one end of the spectrum might not be properly alignable with one from the other end. But within the continuum, it is possible to move from one end to the other by continuously aligning the most similar sequences. Thus, the simple criterion of whether the DNA sequence of two TEs can be aligned over most of their length can lead to unclear situations. Nevertheless, in most cases, the criterion works quite well. Indeed, usually it is not possible to cross the boundary from one TE family to the other simply by continuously aligning the most similar sequences. For example the *Copia* families *BARE1* and *Maximus* from barley show practically no DNA sequence identity, not even in the most conserved parts of the CDS (Wicker and Keller 2007). It is, therefore, not possible to cross the boundary from one family to the other based on alignments of the DNA sequences. If nothing else, the strategy of defining TE families based on sequence homology is at least pragmatic and allows classification without complex phylogenetic analyses. Nevertheless, it does not replace phylogenetic analyses when it comes to the study of evolution.

1.2.8 How Many Different TE Families Are There?

Recently, the classification system of Wicker et al. (2007) was put to the test in the framework of the International Brachypodium Initiative (2010). The stated goal was to obtain a TE annotation that is comparable in quality to gene annotation. Thus, Brachypodium became the first plant genome where a special group, the Brachypodium repeat annotation consortium (BRAC), was responsible solely for TE annotation. Great care was taken to isolate and characterise as many TE families as possible. As shown in Table 1.2, a total of 499 TE families were characterised. The largest variety was found in LTR retrotransposons which contribute over two-thirds of all families. They are also the class of elements that contributes most to the total genome sequence due to their large size. Most abundant in numbers of copies were small Miniature Inverted-Repeat Transposable Elements (MITEs; Bureau and Wessler 1994), small non-autonomous DNA transposons. Over 20,000 Stowaway MITEs of 23 different families were identified. Despite the large effort invested in TE annotation in the Brachypodium genome, TE annotation is still not complete.

Table 1.2 Numbers of TE families in the genome of the model grass *Brachypodium distachyon*

Superfamily	Code	Families
Gypsy	RLG	147
Copia	RLC	133
LTR unknown	RLX	56
Non-LTR	RIX	3
CACTA	DTC	13
Harbinger	DTH	44
Mariner	DTT	36
Mutator	DTM	62
Helitron	DHH	5
Total		499

TE are categorised into superfamilies. These numbers refer to TE families that were characterised in detail in the framework of the Brachypodium repeat annotation consortium. The actual number of TE families is known to be higher

When sequences were annotated carefully in comparative analyses, dozens of additional TE families could be identified (Jan Buchmann, pers. comm). Many of them are low-copy elements which have weak or no homology to previously described TE families. Thus, the 499 TE families identified in the framework of the genome project are certainly a minimal number. The Brachypodium genome is relatively small compared to other plant genomes. However, there is evidence that the size of larger genomes is mainly due to the excessive expansion of relatively few TE families, rather than the diversification of countless small families. Especially in plants, single or a few LTR retrotransposon families can contribute large parts to the genome (Paterson et al. 2009; Schnable et al. 2009; Wicker et al. 2009). In fungi, the situation is similar: in the very repetitive genome of barley powdery mildew, a few dozen TEs completely dominate the repetitive fraction (Spanu et al. 2010). In summary, in most genomes one has to expect hundreds of different TE families, in some probably thousands. However, fears that there might more TE families in a single genome than words in the English language (SanMiguel et al. 2002), and thus naming of all individual families would be impossible, seem to be unfounded.

1.2.9 The Necessity of TE Databases

For the researcher confronted with the epic task to annotate TEs in a genome, it is essential to have a good reference database of TE sequences. In the best case, this is a dataset of well-characterised TE sequences. In the worst case, it is a collection of sequences that are simply known to be repetitive and which were assembled automatically into contigs. Often the reality lies somewhere between the two. The most abundant TEs are usually well characterised with respect to their precise termini and proteins they encode. But for many sequences, one only knows that

they are repetitive, but the exact size or classification is not known. Repeat classification and characterisation is still done very much on a species by species. This is mainly because TEs from different species (if they diverged more than a dozen million years ago) share very little sequence identity at the DNA level. Thus, only protein-coding TEs can usually be identified across species boundaries. If one also wants to precisely annotate non-coding regions and non-autonomous TEs, one usually needs to generate a TE database for the respective species. There are too many TE databases for different species available to describe here. The most inclusive product available today is probably RepBase (girinst.org/repbase/), which includes TE sequences from many different species. However, the task of compiling an all-inclusive TE database which adheres to consistent rules is a monumental one, and it is growing literally by the day.

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Sidén-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, WoodageT WKC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC (2000) The genome sequence of *Drosophila melanogaster*. Science 287:2185–2195

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D (2004) Ultraconserved elements in the human genome. Science 304:1321–1325

Bennett MD, Smith JB (1976) Nuclear DNA amounts in angiosperms. Philos Trans R Soc Lond B Biol Sci 274:227–274

Biderre C, Pages M, Metenier G, Canning EU, Vivaras CP (1995) Evidence for the smallest nuclear genome (2.9 Mb) in the microsporidium *Encephalitozoon cuniculi*. Mol Biochem Parasitol 74:229–231

Britten RJ, Graham DE, Neufeld BR (1974) Analysis of repeating DNA sequences by reassociation methods. Enzymology 29:363–418

Bureau TE, Wessler SR (1994) Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Plant Cell 6:907–916

C. elegans Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282:2012–2018

Capy P, Vitalis R, Langin T, Higuet D, Bazin C (1996) Relationships between transposable elements based upon the integrase-transposase domains: is there a common ancestor? J Mol Evol 42:359–368

Capy P, Langin T, Higuet D, Maurer P, Bazin C (1997) Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor? Genetica 100:63–72

Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, Leroy P, Schlub S, Le Paslier MC, Magdelenat G, Gonthier C, Couloux A, Budak H, Breen J, Pumphrey M, Liu S, Kong X, Jia J, Gut M, Brunel D, Anderson JA, Gill BS, Appels R, Keller B, Feuillet C (2010) Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. Plant Cell 22:1686–1701

Dean RA, Talbot NJ, Ebbole DJ, Farman ML, Mitchell TK, Orbach MJ, Thon M, Kulkarni R, Xu JR, Pan H, Read ND, Lee YH, Carbone I, Brown D, Oh YY, Donofrio N, Jeong JS, Soanes DM, Djonovic S, Kolomiets E, Rehmeyer C, Li W, Harding M, Kim S, Lebrun MH, Bohnert H, Coughlan S, Butler J, Calvo S, Ma LJ, Nicol R, Purcell S, Nusbaum C, Galagan JE, Birren BW (2005) The genome sequence of the rice blast fungus *Magnaporthe grisea*. Nature 434:980–986

Finnegan DJ (1989) Eukaryotic transposable elements and genome evolution. Trends Genet 5:103–107

Freeling M, Subramaniam S (2009) Conserved noncoding sequences (CNSs) in higher plants. Curr Opin Plant Biol 12:126–132

Gregory TR (2001) Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol Rev Camb Philos Soc 76:65–101

Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JM, Wides R, Salzberg SL, Loftus B, Yandell M, Majoros WH, Rusch DB, Lai Z, Kraft CL, Abril JF, Anthouard V, Arensburger P, Atkinson PW, Baden H, de Berardinis V, Baldwin D, Benes V, Biedler J, Blass C, Bolanos R, Boscus D, Barnstead M, Cai S, Center A, Chaturverdi K, Christophides GK, Chrystal MA, Clamp M, Cravchik A, Curwen V, Dana A, Delcher A, Dew I, Evans CA, Flanigan M, Grundschober-Freimoser A, Friedli L, Gu Z, Guan P, Guigo R, Hillenmeyer ME, Hladun SL, Hogan JR, Hong YS, Hoover J, Jaillon O, Ke Z, Kodira C, Kokoza E, Koutsos A, Letunic I, Levitsky A, Liang Y, Lin JJ, Lobo NF, Lopez JR, Malek JA, McIntosh TC, Meister S, Miller J, Mobarry C, Mongin E, Murphy SD, O'Brochta DA, Pfannkoch C, Qi R, Regier MA, Remington K, Shao H, Sharakhova MV, Sitter CD, Shetty J, Smith TJ, Strong R, Sun J, Thomasova D, Ton LQ, Topalis P, Tu Z, Unger MF, Walenz B, Wang A, Wang J, Wang M, Wang X, Woodford KJ, Wortman JR, Wu M, Yao A, Zdobnov EM, Zhang H, Zhao Q, Zhao S, Zhu SC, Zhimulev I, Coluzzi M, della Torre A, Roth CW, Louis C, Kalush F, Mural RJ, Myers EW, Adams MD, Smith HO, Broder S, Gardner MJ, Fraser CM, Birney E, Bork P, Brey PT, Venter JC, Weissenbach J, Kafatos FC, Collins FH, Hoffman SL (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 298:129–149

Hull R, Covey SN (1986) Genome organization and expression of reverse transcribing elements: variations and a theme. J Gen Virol 67:1751–1758

International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463:763–768

International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695–716

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyère C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pè ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quétier F, Wincker P (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467

Kapitonov V, Jurka J (2001) Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci USA 98:8714–8719

Kapitonov V, Jurka J (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet 9:411–412

Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, Prensier G, Barbe V, Peyretaillade E, Brottier P, Wincker P, Delbac F, El Alaoui H, Peyret P, Saurin W, Gouy M, Weissenbach J, Vivares CP (2001) Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature 414:450–453

Leitch IJ, Beaulieu JM, Cheung K, Hanson L, Lysak MA, Fay MF (2007) Punctuated genome size evolution in *Liliaceae*. J Evol Biol 20:2296–2308

Mayer KF, Taudien S, Martis M, Simková H, Suchánková P, Gundlach H, Wicker T, Petzold A, Felder M, Steuernagel B, Scholz U, Graner A, Platzer M, Dolezel J, Stein N (2009) Gene content and virtual gene order of barley chromosome 1H. Plant Physiol 151:496–505

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562

Orgel LE, Crick FHC (1980) Selfish DNA: the ultimate parasite. Nature 284:604–607

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob-ur-Rahman WD, Westhoff P, Mayer KF, Messing J, Rokhsar DS (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551–556

Peterson DG, Schulze SR, Sciara EB, Lee SA, Nagel A, Jiang N, Tibbetts DC, Wessler SR, Paterson AH (2002) Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. Genome Res 12:795–807

Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud PF, Lindquist EA, Kamisugi Y, Tanahashi T, Sakakibara K, Fujita T, Oishi K, Shin-I T, Kuroki Y, Toyoda A, Suzuki Y, Hashimoto S, Yamaguchi K, Sugano S, Kohara Y, Fujiyama A, Anterola A, Aoki S, Ashton N, Barbazuk WB, Barker E, Bennetzen JL, Blankenship R, Cho SH, Dutcher SK, Estelle M, Fawcett JA, Gundlach H, Hanada K, Heyl A, Hicks KA, Hughes J, Lohr M, Mayer K, Melkozernov A, Murata T, Nelson DR, Pils B, Prigge M, Reiss B, Renner T, Rombauts S, Rushton PJ, Sanderfoot A, Schween G, Shiu SH, Stueber K, Theodoulou FL, Tu H, Van de Peer Y, Verrier PJ, Waters E, Wood A, Yang L, Cove D, Cuming AC, Hasebe M, Lucas S, Mishler BD, Reski R, Grigoriev IV, Quatrano RS, Boore JL (2008) The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. Science 319:64–69

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20:43–45

SanMiguel PJ, Ramakrishna W, Bennetzen JL, Busso CS, Dubcovsky J (2002) Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome 5A(m). Funct Integr Genomics 2:70–80

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC, Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

Seberg O, Petersen G (2009) A unified classification system for eukaryotic transposable elements should reflect their phylogeny. Nat Rev Genet 10:276

Spanu PD, Abbott JC, Amselem J, Burgis TA, Soanes DM, Stüber K, Loren V, van Themaat E, Brown JK, Butcher SA, Gurr SJ, Lebrun MH, Ridout CJ, Schulze-Lefert P, Talbot NJ, Ahmadinejad N, Ametz C, Barton GR, Benjdia M, Bidzinski P, Bindschedler LV, Both M, Brewer MT, Cadle-Davidson L, Cadle-Davidson MM, Collemare J, Cramer R, Frenkel O, Godfrey D, Harriman J, Hoede C, King BC, Klages S, Kleemann J, Knoll D, Koti PS, Kreplak J, López-Ruiz FJ, Lu X, Maekawa T, Mahanil S, Micali C, Milgroom MG, Montana G, Noir S, O'Connell RJ, Oberhaensli S, Parlange F, Pedersen C, Quesneville H, Reinhardt R, Rott M, Sacristán S, Schmidt SM, Schön M, Skamnioti P, Sommer H, Stephens A, Takahara H, Thordal-Christensen H, Vigouroux M, Wessling R, Wicker T, Panstruga R (2010) Genome expansion and gene loss in powdery mildew fungi reveal functional tradeoffs in extreme parasitism. Science 330:1543–1546

Swigonova Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL, Messing J (2004) On the tetraploid origin of the maize genome. Comp Funct Genomics 5:281–284

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Déjardin A, Depamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjärvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leplé JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouzé P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y, Rokhsar D (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596–1604

Wicker T, Keller B (2007) Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. Genome Res 17:1072–1081

Wicker T, Stein N, Albar L, Feuillet C, Schlagenhauf E, Keller B (2001) Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. Plant J 26:307–316

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982

Wicker T, Taudien S, Houben A, Keller B, Graner A, Platzer M, Stein N (2009) A hole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. Plant J 59:712–722

Yang YW, Lai KN, Tai PY, Li WH (1999) Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between Brassica and other angiosperm lineages. J Mol Evol 48:597–604

Chapter 2
Transposable Element Annotation in Completely Sequenced Eukaryote Genomes

Timothée Flutre, Emmanuelle Permal, and Hadi Quesneville

Abstract With the development of new sequencing techniques, the number of sequenced plant genomes is increasing. However, accurate annotation of these sequences remains a major challenge, in particular with regard to transposable elements (TEs). The aim of this chapter is to provide a roadmap for researchers involved in genome projects to address this issue. We list several widely used tools for each step of the TE annotation process, from the identification of TE families to the annotation of TE copies. We assess the complementarities of these tools and suggest that combined approaches, using both *de novo* and knowledge-based TE detection methods, are likely to produce reasonably comprehensive and sensitive results. Nevertheless, existing approaches still need to be supplemented by expert manual curation. Hence, we describe good practice required for manual curation of TE consensus sequences.

Keywords Annotation • Bioinformatics • Classification • Curation • Identification • Pipeline

2.1 Introduction

Transposable elements (TEs) are mobile genetic elements that shape the eukaryotic genomes in which they are present. They are virtually ubiquitous and make up, for instance, 20% of a typical *D. melanogaster* genome (Bergman et al. 2006), 50% of a *H. sapiens* genome (Lander et al. 2001), and 85% of a *Z. mays* genome (Schnable et al. 2009). They are classified into two classes depending on their transposition mode: via RNA for class I retrotransposons and via DNA for class II transposons

T. Flutre • E. Permal • H. Quesneville (✉)
INRA, UR 1164, URGI, Unité de Recherche en Génomique-Info,
78026 Versailles cedex, France
e-mail: hadi.quesneville@versailles.inra.fr

M.-A. Grandbastien and J.M. Casacuberta (eds.), *Plant Transposable Elements*,
Topics in Current Genetics 24, DOI 10.1007/978-3-642-31842-9_2,

(Finnegan 1989). Each class is also subdivided into several orders, superfamilies, and families (Wicker et al. 2007). Due to their unique ability to transpose and because they frequently amplify, TEs are major determinants of genome size (Petrov 2001; Piegu et al. 2006) and cause genome rearrangements (Gray 2000; Fiston-Lavier et al. 2007). Once described as the "ultimate parasites" (Orgel and Crick 1980), TEs are commonly found to regulate the expression of neighboring genes (Feschotte 2008; Bourque 2009) or even to have been domesticated so as to provide a specific host function (Zhou et al. 2004; Bundock and Hooykaas 2005; Santangelo et al. 2007; Kapitonov and Jurka 2005).

As a consequence of the development of new rapid sequencing techniques, the number of available sequenced eukaryotic genomes is constantly increasing. However, the first step of the analysis, i.e., accurate annotation, remains a major challenge, particularly concerning TEs. Correct genome annotation of genes and TEs is an indispensable part of thorough genome-wide studies. Consequently, efficient computational methods have been proposed for TE annotation (Bergman and Quesneville 2007; Lerat 2010; Janicki et al. 2011). Given that the pace at which genomes are sequenced is unlikely to decrease in the coming years; the process of TE annotation needs to be made widely accessible.

This chapter lays down a clear road map detailing the order in which computational tools (or combinations of such tools) should be used to annotate TEs in a whole genome. We distinguish three steps (1) identifying TEs by searching for reference sequences (e.g., full-length TE sequences) and building consensuses from similar sequences, (2) manual curation to define and classify TE families, and (3) annotation of every TE copy. We also provide some hints on manual curation, a step that is still necessary.

2.2 *De Novo* Detection of Transposable Elements

Various efficient computational methods are available to identify unknown TEs in genomic sequences. Each method is based on specific assumptions that have to be understood to optimize selection and combination of the methods to ensure they are appropriate for any particular analytic goal.

2.2.1 Computing Highly-Repeated Words

TEs, due to their capacity to transpose, are often present in a large number of copies within the same genome. Although TE sequences degenerate with time, words (i.e., short subsequences of few nucleotides) that compose them are consequently repeated throughout the genome. Software, such as the TALLYMER (Kurtz et al. 2008) and P-CLOUDS (Gu et al. 2008), has been designed to find repeats rapidly in genome sequences by counting highly frequent words of a given length k, called k-mers. These programs are very useful for quickly providing a view of the repeated