Identification and Analysis of Transposable Elements in Genomic Sequences

9

Laurent Modolo and Emmanuelle Lerat

Abstract

Genome sequences are composed of different compartments, among which transposable elements (TEs) represent one of the most important. Not only do these elements correspond to a particularly large proportion of genomes, they are also involved in different mechanisms implicated in the evolution of genomes, such as chromosome rearrangement and gene innovation. Thus, the precise determination of TEs in genomes is of significant importance. This step is becoming more and more complex with the emergence of new types of sequence data coming from next-generation sequencing (NGS) technologies. In this chapter, we present the current status of bioinformatic developments made in the detection and analysis of TEs in genomic sequences. We first present the classic tools dedicated to the identification of TEs in classic genomic data, which originate from whole-genome sequences. Because these sequences are significantly different from the new types of sequences generated by NGS and because the problem of repeats in these data is not trivial, we then present how it is possible to handle TEs in NGS data. We also provide some examples of tools designed to answer particular questions about TEs using NGS data and how these types of data are particularly valuable for deepening our knowledge of the dynamics of TEs. Although this is a still a fast-growing field for which new developments are made every day, we hope to provide a broader view of what currently exists in this field and what allows for TE analyses in genomic sequences.

Introduction

Eukaryotic genomes are composed of different elements, which are classified according to their function in the organism. The protein-coding genes, the non-coding genes (specifically rRNA, tRNA, and small RNA genes), and the regulatory elements associated with these genes are typically considered the most important groups. The remaining elements, which include pseudogenes and repeated sequences, have long been considered trivial in terms of genome functioning. However, this limited vision has begun to change over the past several years. Measured genome size has been demonstrated to be highly variable. However, this metric has been found to have no relationship with the 'complexity' of the organism. When considering the expected number of genes, this paradox has had the implication that a given genome can contain more DNA than required. Early genome sequencing projects helped to answer this question with the discovery that the functional genome, which predominantly consists of protein-coding genes, represents only a small percentage of the genome, whereas the more variable regions of the genome, the 'non-coding regions', were demonstrated to represent a higher proportion of the complete genome. With regard to the human genome, this finding has been particularly striking. The first estimations of the gene number in our own genome were radically revised after the human genome sequencing project, which revealed a small number of genes (fewer than 25,000) and more than 98% 'non-functional' DNA (Lander *et al.*, 2001). The main source of size variation in genomes was then identified as the non-coding regions of genomes.

Transposable elements (TEs) are a part of these still underestimated genomic components, which can play important roles in the functioning and in the evolution of organisms.

TEs are dispersed repeat DNA sequences that have the ability to move from one position to another along chromosomes. These elements typically encode for all the proteins necessary for their movement and possess internal regulatory regions, allowing for their independent expression. Different categories of TEs have been identified, and several attempts were made to classify them (Wicker *et al.*, 2007; Kapitonov and Jurka, 2008). Globally, two main classes have been described according to their transposition intermediates (retrotransposons use an RNA intermediate and form class I, while transposons use a DNA intermediate and form class II). Within each class, subclasses have been created to group sequences with the same structural features (Fig. 9.1). According to the classification system of Wicker *et al.*, five orders containing 17 superfamilies were described as class I elements, and two subclasses, containing 12 superfamilies, were described as class II elements. As the number of the newly sequenced genomes increases, new elements and potential new types of TEs are being discovered, which increases and enriches the complexity of TE classification.

Since B. McClintock discovered and first described these elements in maize in the 1950s (McClintock, 1956), TEs have been searched for and discovered in almost all eukaryotic organisms. Depending on the organism, the proportion of TEs can be highly variable and at times large, for example, 3% in yeast (Kim *et al.*, 1998), 15%

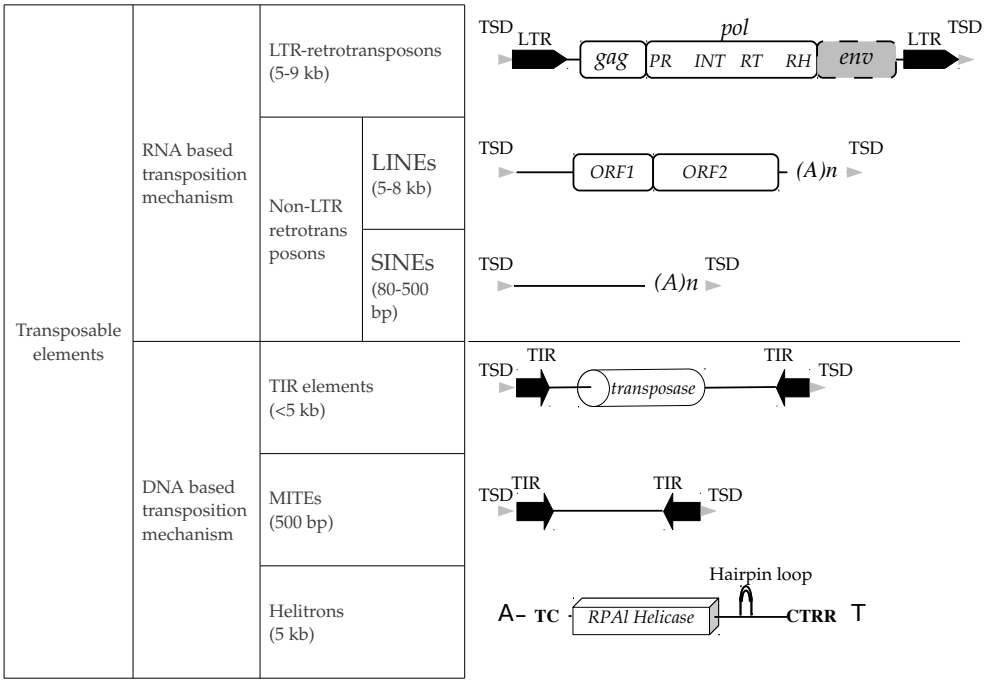

Figure 9.1 The different types of transposable elements (TEs) according to their transposition intermediates and their structural features. Almost all TEs possess two target site duplication (TSD) sequences at each extremity of the copy, corresponding to duplications of the insertion site. LTR-retrotransposons possess long terminal repeat (LTR) sequences at each extremity. Some LTR-retrotransposons encode for a third ORF specifying the ENV protein. The pol gene is formed by different domains, which encode for a protease (PR), an integrase (INT), a reverse transcriptase (RT) and an RNaseH (RH), respectively. The non-LTR retrotransposons contain a poly-A tail at their 3′ extremity. The LINEs possess two ORFs, whereas the SINEs have no coding capacity. Contrary to the MITEs, TIR elements contain an ORF encoding for a transposase. Both contain terminal inverted repeat (TIR) sequences at each extremity. The autonomous helitrons possess a coding capacity for a helicase and an RPA-like (RAPI) single-stranded DNA-binding protein.

in *Drosophila* (Dowsett and Young, 1982), 45% in human (Lander *et al.*, 2001) and more than 80% in maize (Schnable *et al.*, 2009). By proportion, TEs are directly linked to the host's genome size. Their proportion is typically very high in organisms with a very large genome, indicating their role in the expansion of the host's genome size. The variability of TE proportions in closely related species is linked to different parameters, such as the effective population size and reproduction mode of the host organism, the genetic drift, and the host's regulation of transposition activity. For example, in the sibling species *Drosophila melanogaster* and *D. simulans*, the proportion of TEs varies threefold (15% in *D. melanogaster* and 5% in *D. simulans*), although the two species have only been separated for 2.3 to 5.4 million years (Li *et al.*, 1999; Tamura *et al.*, 2004; Cutter, 2008). This discrepancy in TE content has been hypothesized to be linked to either a stronger selection against the deleterious effects of TE insertion in *D. simulans*, which could be due to a larger effective population size compared to *D. melanogaster*, or to a stronger resistance of *D. simulans* to an increase in TE copy number (Kimura and Kidwell, 1994; Vieira *et al.*, 1999). The self-incompatible plant *Arabidopsis lyrata* presents a larger genome (207 Mb) than does the self-compatible plant *A. thaliana* (125 Mb); however, these two species are only separated by 10 million years (Hu *et al.*, 2011). The size variation observed between these two plant species is predominantly due to TEs and appears, in part, to be related to differences in their mating systems (Lockton and Gaut, 2010).

Because of their presence in genomes, TEs have a significant impact on genome evolution and not only on the evolution of genome size (Lisch and Kidwell, 2000; Biémont and Vieira, 2006). TEs can promote mutations, which can be deleterious. For example, in humans, approximately 96 transposition events are directly linked to single-gene diseases (Hancks and Kazazian, 2012), and half of the spontaneous mutations observed in *Drosophila* are due to TEs (Eickbush and Furano, 2002). However, the effects of TEs can also increase the genetic diversity of an organism. The repetitive nature of TEs makes them responsible for chromosomal rearrangements via homologous recombination between copies, which, in some cases, can lead to the emergence of new species, such as has been hypothesized for *D. virilis*, for example (Evgen'ev *et al.*, 2000). When inserting in or near a gene, TEs can provide new regulatory elements, altering the expression of the gene, or they can contribute to gene innovation by providing a new coding region to a gene (Lisch and Kidwell, 2000). The implications of TEs in all epigenetic mechanisms have now been clearly established (Slotkin and Martienssen, 2007). All of these facts make TEs particularly important in the adaptation of organisms to environmental changes. During the years since their discovery, the status of TEs has moved from simple junk DNA to major players in genome evolution (for a historical review, see Biémont, 2010).

Thus, TEs are important components that cannot be neglected when analysing genome sequences. These elements can be quite numerous and, therefore, very important, and they should not be simply removed to ease gene annotations. The study of TEs is crucial for understanding their dynamics, which then allows us to better appreciate how genomes function and evolve. The question of identifying TEs in genomic sequences is a crucial point that has become more and more complex with the emergence of new types of sequencing data. In this chapter, we will first summarize the classic methods that exist to search and annotate TEs in assembled genome sequences. We will then identify the difficulties that can be encountered when using next-generation sequencing (NGS) data for this task and new methods that have been developed. Finally, we will provide examples concerning the specific analyses that can be performed on TEs using NGS data and how these new types of data constitute a significant advance in the field of TE dynamics.

Classic detection methods for TEs in genome sequences

Since the beginning of genome sequencing, significant efforts have been made to annotate the functional regions of genomes. The presence of transposable elements (TEs) and other repeats has made this task particularly difficult. Thus, to facilitate the annotation of genomes, methods to identify TEs and other repeats in genome

sequences were developed (Tang, 2007). Indeed, the ability to recognize these types of sequences has been a suitable starting point to allow for the assembly of a genome and also to ease the prediction of genes. This task is particularly important for genomes containing very high proportions of TEs. Moreover, given the importance of TEs in genome evolution, the identification of these sequences has been considered crucial to allow the access to entire populations of TE copies present in a given organism. Having access to all copies of a given TE family is particularly interesting for studying the evolution and dynamics of a TE family. For example, an analysis of the TE copies from the majority of families integrated into the *D. melanogaster* genome surprisingly demonstrated that the majority of these TEs had recently moved because these TE copies were almost identical within families and very few ancient copies were present (Bowen and McDonald, 2001; Lerat *et al.*, 2003). These observations are in favour of either the hypothesis of recurrent and numerous horizontal transfers of these elements or the hypothesis of a very high turnover in this genome to remove ancient and inactive copies. The identification of all the TE copies present in the human genome has allowed us to obtain information concerning the waves of amplification of the non-LTR retrotransposons LINE and SINE and the formation of retro-processed pseudogenes in this genome (Lander *et al.*, 2001; Ohshima *et al.*, 2003). Numerous methods dedicated to the identification and classification of TEs have been developed over the past 15 years. Several reviews have described these methods exhaustively and provide lists of available programs in each category (Bergman and Quesneville, 2007; Saha *et al.*, 2008a; Lerat, 2010; Janicki *et al.*, 2011). In this section, we will mainly summarize the different categories of existing programs and those, which are currently more used and more successful in performing their tasks.

Similarity- or library-based methods

The principle of these methods is to compare genome sequences to a library of TE reference sequences to search for the occurrence of the TEs in a genome. The library used can be defined by the user or can be a public database. The most widely used public database employed in this type of work is Repbase (Jurka *et al.*, 2005). This database contains the consensus sequences of different repeat sequences from a large set of eukaryotic organisms and is typically employed jointly with the program RepeatMasker (Smit *et al.*, 1996–2010), which performs a similarity search using the library as a reference. The main advantage of this type of method is that it is fast and accurate. Although, it obviously cannot discover new TE families, this method is still a good starting point to explore a new genome, particularly if TE sequences from closely related species are described.

Signature-based methods

This type of method uses particular structural features (such as nucleotide or protein motifs) of known TE classes to determine their occurrence in a genome sequence. Thus, this approach can locate new elements from a given class but will fail to discover new classes of elements. Another drawback of this method is that it will only discover nearly complete and potentially active copies and miss degraded ones. Thus, such an approach can be complemented using a library-based method once complete reference elements have been discovered based on their structure. Moreover, such an approach will depend on the level of knowledge available for a given class and if it is possible to determine fixed and shared features among several families of the same class. Signature-based programs typically concentrate on a particular type of TE.

For example, it is possible to specifically search for LTR-retrotransposons given several shared characteristics between families, such as the presence of an LTR (long terminal repeat) at each end of the sequence, the fact that the two LTRs are almost identical for complete and potentially active copies, a particular distance between them, or the presence of particular protein motifs in the ORFs contained inside the element. Different programs have been designed to detect LTR-retrotransposons, of which the most successful to date is LTRharvest (Ellinghaus *et al.*, 2008; Lerat, 2010). However, the user needs to determine the perfect parameters for the analysed genome to avoid the occurrence of numerous

false positives. Because of structural features such as the presence of a poly-A tail at the 3′ end of the sequence or target site duplications at each extremity of the copy, other programs have been designed to detect non-LTR retrotransposons (Szak *et al.*, 2002; Tu *et al.*, 2004; Lucier *et al.*, 2007). Particular DNA transposons known as MITEs have also been the subject of several programs because of their specific features, *i.e.* a short size (approximately 500 bp) and the presence of terminal inverted repeats (TIRs) at each end of a copy; and because the use of similarity-based methods to find these transposons is difficult due to their short size and a lack of coding capacity. Recently, the program MITE-Hunter (Han and Wessler, 2010) was developed to decrease the number of false positives typically obtained using other programs to locate MITEs.

De novo methods

With these types of approaches, it is possible to search for new types of elements because there is no *a priori* knowledge of the sequence itself. Indeed, these programs take advantage of the repetitive nature of TEs. These methods are particularly interesting when sequencing the genomes of species for which no close relatives are currently annotated and for which nothing is known about their repeat content. However, these methods are particularly sensitive to genomic coverage and to the quality of the sequence assembly. Another drawback is that these approaches will find any type of repeated sequences, even tandem repeats, satellites or segmental duplications, in addition to identifying TE sequences, which implies a classification step for the results to identify TEs. Moreover, TE families containing very few copies will not be detected.

There are two main approaches that are considered *de novo* methods. In the first approach, the genome sequence is first compared against itself to locate all the repeated sequences. Several programs use BLAST (Altschul *et al.*, 1990) to perform this step. The repeated sequences that are located are then grouped into clusters of similar sequence. A consensus sequence is then built for each cluster, and all the consensus sequences are used in a library-based approach to retrieve all occurrences within the genome. Among the most utilized programs that are currently used in the annotation of genomes, we can cite RECON (Bao and Eddy, 2002), PILER (Edgar and Myers, 2005), and BLASTER (Quesneville, unpublished).

In the second approach, the occurrence of multiple small words known as k-mers is searched for within the genome sequence. The k-mer can then be extended to obtain longer sequences. Among the existing programs using this approach, some have been used to discover TEs in genome sequences, for example, REPuter (Kurtz and Schleiermacher, 1999), RepeatScout (Price *et al.*, 2005), and ReAS (Li *et al.*, 2005). This last program has the peculiarity of running not on an assembled genome but on the unassembled reads of a whole genome shotgun sequence to avoid the problems related to a bad assembly.

TEs in the next-generation sequencing data era

The availability of next-generation DNA sequencing (NGS) technologies has revolutionized our approach to genomics (Margulies *et al.*, 2005). These technologies allow us to obtain huge amounts of data at a relatively low cost and with less bias than older technologies (Wicker *et al.*, 2006), thus opening new avenues to the study of TEs. These new types of data also imply that the classic methods described previously will no longer be adapted. To describe how to deal with TEs in NGS data, it is first important to understand why these data are different from older sequencing data and what methodologies are currently available to handle them before performing the TE studies.

With NGS technologies, not only has the volume of data generated dramatically increased, but the range of applications has broadened from methylation pattern detection (MeDIP-Seq) and the study of DNA–protein interactions (ChIP-Seq) to quantifying and detecting gene expression (RNA-Seq). Whatever the application, three steps can always be highlighted in sequencing methodologies. First, the DNA of interest is randomly fragmented and amplified. Second, the ends of each of these fragments are sequenced into reads. Finally, the original sequence of the

DNA is reconstructed from the reads. Currently, the first two steps are highly automated, and the only concern of the researcher is to determine the appropriate sequencing cost to balance the read length and the depth of coverage necessary for the study. Even if the read size has increased with the development of NGS technologies, the reconstruction of the original DNA sequences (assembly) is currently still the most challenging and time-consuming step.

NGS data are subject to particular artefacts that need to be taken into account before performing analyses. These artefacts are predominantly adaptor sequences originating from failed or short DNA insertions during library preparation or near identical reads originating from PCR error. This step of trimming and filtering can be performed using several tools, such as SeqTrim or Quake (Falgueras *et al.*, 2010; Kelley *et al.*, 2010). Another problem is the presence of sequencing errors. Variations between reads can be caused by real sequencing errors or by single-nucleotide polymorphisms (SNPs), accounting for polyploidy and pooled samples. Thus, it is important to perform error corrections that will be linked to the NGS technologies used to generate the data because the different sequencing methods produce different types of sequencing error.

Sequencing and analysing DNA using NGS

Once the reads are sequenced, an important step is to perform their assembly in order to reconstruct the complete genome. This step is complex but can be eased if a reference genome is available. In such a case, it is possible to map the reads directly onto the reference genome. As the number of reads can be very large, classic mapping programs such as the one from the BLAST suite (Altschul *et al.*, 1990) have become too computationally demanding. Thus, a number of alternative approaches have been developed over the past three years to handle NGS data. Two strategies exist for mapping reads onto a genome. The first uses a hash table of the reads, and the other uses a Burrows-Wheeler (BW) transform of the genome (Schbath *et al.*, 2012). Generally, a hash table can better address mismatches, whereas the more complex BW transform approach can easily handle repeats. There are different alternatives to deal with reads occurring at multiple positions, which are known as multi-reads. The program can either ignore them, keep the best matches, keep a specific number of them, or ignore the ones mapping to more than a specific number of locations (Treangen and Salzberg, 2012). Taking into account all of these considerations, it appears that the programs BWA (Li *et al.*, 2010) and Bowtie (Langmead *et al.*, 2009) can outperform the other programs on many criteria, such as computational time, the correct positions of the reads, the number of unmapped reads, and the multi-reads with no more than three mismatches (Schbath *et al.*, 2012). Other parameters that need to be taken into account in addition to mismatches are indels. For this question, BWA is currently the only mapping program using the BW transform that is able to handle indels. Mapping programs using the BW transform appear to be the most appropriate for the study of TEs because they can better handle genomic repeats. Moreover, even if this class of algorithms relies on heuristics to address mismatches, we expect to have a better error correction and less mismatch on TE sequences because they are sequenced with a better coverage than the rest of the genome. For example, with a coverage of 10×, we expect an average coverage of 50× for a given TE that is present with five copies in the genome.

Most of the time there is no reference genome, and the original DNA sequences have to be reconstructed *de novo*. The mapping approach can still be used with the LAST program, which can take into account the divergence between species to relax the mapping parameters and perform *xeno-mapping* (Frith *et al.*, 2010). Otherwise, two different approaches exist for assembling without a reference genome, the first using a seed approach and the other using a de Bruijn graph. In the first approach, the algorithm tries to elongate short sequences of k nucleotides (k-mers) using overlapping reads by computing an overlapping graph where all the paths in the graph consist of overlapping reads. Developed for Sanger technologies, the construction of such a graph is often computationally intractable in the case of NGS data (Pevzner *et al.*, 2001). Since the publication of the EULER program, most assemblers use

the de Bruijn graph approach to assemble reads (Fig. 9.2). The first step of this approach is to build an index of all the possible sequences of size *k* (*k*-mers, often between 24 and 27 bp). The graph itself is built by adding the information from each read, and the sequence of a read is represented by a path between nodes. The nodes correspond to the *k*-mers and their reverse complements to handle more efficiently the two strands of DNA. The addition of each read will correspond to the addition of more edges between the nodes. The construction of a contig is a byproduct of the graph itself. Retrieving the original DNA sequence is a matter of linearizing the information contained in the graph by following the most supported edge, i.e. the one with the highest number of reads (Pevzner *et al.*, 2001; Zerbino and Birney, 2008).

In practice, even with a high coverage, there are always some parts of a DNA sequence that are more difficult to assemble, such as repetitive elements. If the repeat length is larger than the read length, which is often the case for TEs, then the coverage cannot help to reconstruct the original sequence. In this case, a read cannot be associated with one particular copy of the repeated element. Unassembled or uncovered regions are going to form gaps in the assembly, thus decreasing the connectivity between the resulting sequences. Paired-end sequencing technology can be very useful for solving some of the problems caused by repetitive content and short or very short reads (Treangen and Salzberg, 2012). This type of technology is a specific way of sequencing a DNA fragment at both ends. As the size of the fragments is known, the two resulting reads can be reliably positioned relative to each other. To obtain pairs of reads that are more than 500 bp apart (called the insert size), a specific library must be built (a mate-pair or long paired-end library) (Bentley *et al.*, 2008). With paired-end sequencing, a coverage of 10× is sufficient to have at least one mate-pair spanning every instance of each repeat in the genome and to be able to anchor this repeat if the paired read is uniquely mappable (Wetzel *et al.*, 2011). This mate pair information is very useful to position and order the contigs between each other, to fill the gaps, and to build scaffolds (Fig. 9.3). Contrary to the read length, longer inserts will not correspond to a better assembly (Wetzel *et al.*, 2011). It appears that the best strategy is to use different insert sizes to be able to resolve small repeats with short inserts and long repeats with large inserts. The size of the insert can be specifically tuned to optimally assemble the repeat content of the genome under study (Wetzel *et al.*, 2011). For example, it was possible to obtain the same assembly quality found using the Sanger-based approach by using insert sizes of 180 bp, 3 kb, 6 kb and 40 kb for the mouse and human genomes (Gnerre *et al.*, 2011). As with the classic Sanger sequencing methods, the mis-assembly of the repetitive parts of a genome can lead to numerous errors in the reconstructed sequences (Phillippy *et al.*, 2008). For example, a collapse is formed when the assembler incorrectly joins reads originating from distinct repeat copies. On the contrary, expansions are formed

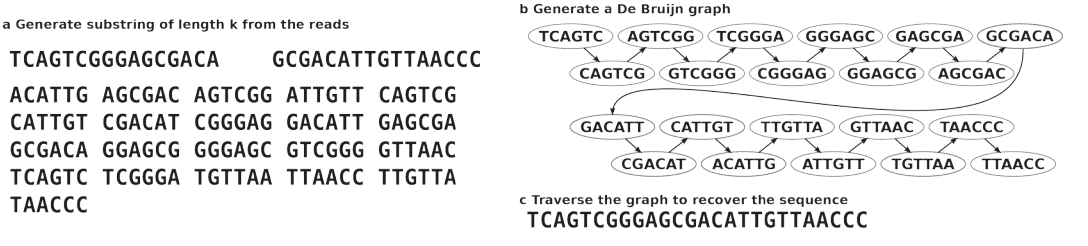

Figure 9.2 Overview of a sequence reconstruction using the de Bruijn graph-based strategy. (a) All substrings of length k (k-mers) were generated from the reads (here k=6). (b) Each unique k-mer represents a node in the de Bruijn graph. The read information is stored in the graph by connecting two nodes with an edge each time that two k-mers have a k – 1 overlap in the reads. Note that generally each node in the graph represents a k-mer and its complement. (c) Nodes from a chain of adjacent nodes link with each other by only one edge and are collapsed into a single node. The graph is then traversed to form contigs.

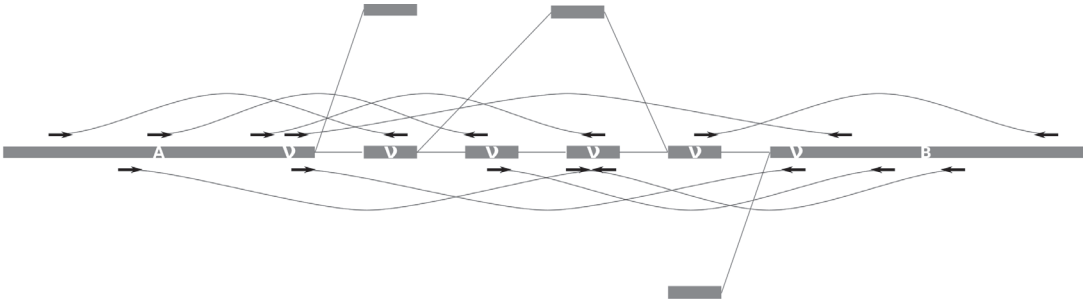

Figure 9.3 Utilization of paired-end information to resolve repeats and to fill gaps. A and B represent long contigs, and the small grey squares are other nodes in the graph. The dashed lines represent all the possible paths built from the read information. The reads are represented by black arrows, and the paired-end information is represented by grey curves linking the two arrowheads to each other. Finding the exact path in the graph from A to B is not a simple task. However, by using the paired-end information linking each pair of reads, we can explore a much simpler graph composed of the 'v' marked blocks.

when reads originating from different copies of a repeat are included in the assembly of one copy of this repeat. These cases are often associated with a greater or lesser density of reads than expected over the repeat(s), and the mate-pairs are stretched out or compressed. Sequence rearrangements are another type of assembly artefact and appear when blocks of DNA are separated by repeats. In this case, the order of these blocks can be wrongly recovered because they are anchored by similar repeats. A special case of rearrangements is inversion, where the two repeat copies are in opposite directions. Sometimes all the above cases of mis-assembly can be possible without violating any constrains on the paired-end information. The amosvalidate package is a collection of tools aimed at detecting mis-assemblies in an automated pipeline (Phillippy *et al.*, 2008).

All the DNA reconstruction problems that we have just described will appear in the de Bruijn graph by forming three types of structures (Zerbino and Birney, 2008). The first corresponds to 'tips', where a chain of nodes is disconnected from the rest of the graph at one end. The second corresponds to 'bubbles', where two paths start from the same node. The third corresponds to erroneous connections, which have no identifiable structure and correspond to chimeric reads. The 'bubble' structures of the graph can correspond to perfect circles, in the case of SNPs or can form densely connected 'tangles' in the case of repetitive sequences (Pevzner *et al.*, 2001). The shape of these 'tangles' can be used for the characterization of the TE, as described in Macas *et al.* (2007). There are many programs that can handle whole genome assemblies using the de Bruijn graph and that are capable of using paired-end information, such as the programs Velvet (Zerbino and Birney, 2008) and ABySS (Simpson *et al.*, 2009). More recently, SOAPdenovo (Li *et al.*, 2010) was used for the *de novo* assembly of the giant panda genome with an average of 20× coverage using 52 bp reads and 37 paired-end libraries with insert sizes ranging from 150 bp to 10 kb. This genome contains 36.1% TEs. The most recent program, ALLPATHS-LG (Gnerre *et al.*, 2011), was created to account for paired-end information with mixed insert size libraries. This program produces a better assembly quality than SOAPdenovo but it is slower (for mammalian-sized genomes, it requires three weeks when SOAPdenovo requires only three days (Gnerre *et al.*, 2011)). Even if the development of assemblers allows for an ever improving assembly of genomes, including their repeat content, there will always remain some unassembled fractions in the data. The study of these leftover contigs or reads can be a source of information about the TE content. These analyses can be conducted using mapping approaches of the data to known TE databases of DNA or protein sequences (Sun *et al.*, 2012).

Sequencing and analysing RNAs using NGS

New sequencing technologies have opened up new vistas of knowledge at the various levels of an organism's biology. Having access to an entire genome is particularly valuable, but having access to the landscape of gene expression allows us to go deeper into the analysis of genomes. Thus, RNA-Seq technology has been developed, which predominantly consists of adding a reverse transcription step to transform RNA into DNA before performing the regular NGS steps (Marioni *et al.*, 2008). The goal of this approach is to obtain the sequence and the abundance of all the transcripts of a given organism in a given condition. Using these types of sequencing methods, it is important to handle particular artefacts. In addition to the errors generated by the reverse transcription step, the reconstruction of the RNA sequences accounts for the different levels of expression for the different transcripts and for the mechanism of alternative splicing (Martin and Wang, 2011). Because of the presence of shared exons in different transcripts, expression counting or quantification is not trivial for RNA-Seq data (Trapnell *et al.*, 2010). Moreover, there is a competition between the most abundant transcripts, which can be over-sequenced, and the less abundant ones, which can be missed. To sequence and quantify less abundant transcripts, hybridization-based depletion methods to remove the more abundant transcripts can be utilized (He *et al.*, 2010). However, these depletion methods induce biases for the quantification and for the assembly of the most expressed transcripts.

Once the main artefacts are corrected, there are two strategies to analyse the RNA-Seq data. The first approach, known as 'map first', consists of mapping the reads onto a reference genome. This approach can be confronted with many problems, ranging from the correctness of the alignment, the possibility of losing the splicing information and the completeness of the reference genome. There are many programs for mapping RNA-Seq data to a reference genome, but it appears that the most commonly utilized is a combination of the program TopHat, which is able to discover splice junctions, and the program Cufflinks, which can be used for the quantification of transcripts (Trapnell *et al.*, 2010). The second approach, 'assembly first' (*de novo* method), consists of directly assembling reads to reconstruct the transcripts. Using this approach, the resultant transcripts can subsequently be mapped to a reference genome. A number of transcriptome assembly programs have been developed, and most use the de Bruijn graph approach. This data structure naturally handles the high redundancy of the data because each repeat or transcript is only present once in the graph. The most commonly used transcriptome assemblers are Velvet (Zerbino and Birney, 2008) and Trans-ABySS (Robertson *et al.*, 2010). More recently, the Trinity program was proposed (Grabherr *et al.*, 2011). Other approaches, such as KISSPLICE (Sacomoto *et al.*, 2012), have been developed to identify and quantify *de novo* polymorphisms such as alternative splicing, SNP and tandem repeats in RNA-Seq data.

Even with a reference genome, the best results are obtained using mixed strategies of the two approaches (Surget-Groba and Montoya-Burgos, 2010). Based on the confidence we have in the reference genome, there are two possibilities. When the reference genome is of a very good quality, it is possible to first align the reads onto it and then to assemble the reads using the mapped reads as long contigs. The assembling step allows for resolving of the reads coming from the expressed regions not present in the reference genome. In a case where the reference genome is not of a very good quality, another strategy is used, consisting of first making the assembly of reads before mapping them onto the reference genome. In this case, the errors present in the reference genome have little impact because they are not present in the assembled contigs. The mapping step is then used to resolve scaffolds from the more fragmented contigs obtained using the *de novo* approaches. This last approach was used to successfully assemble the transcriptome of the mosquito *Anopheles funestus* (Crawford *et al.*, 2010)

What do NGS data bring to TE analyses?

With the new technologies of DNA and RNA sequencing, new opportunities to study TEs have

appeared, once the difficulty for handling these repeated sequences taken into account when processing the data. In this section, we will exemplify different analyses that have allowed us to improve on our understanding of TEs.

TE identification in a genome survey

Even if the cost of NGS sequencing has continuously diminished it is still of interest to perform genomic surveys to characterize large genomes. The historic approach of using the end sequences of bacterial artificial chromosome (BAC) vectors is biased towards sequences that can be successfully loaded into BACs. NGS technologies allow us to perform more representative surveys of large genomes by sequencing at a low depth using a whole genome shotgun (WGS) approach. This approach was first used to further characterize the soybean genome, with particular attention paid to its repeat content (Swaminathan *et al.*, 2007). The first step of the data analysis is to characterize a maximum number of reads using databases of annotated sequences. For this step, classical programs such as BlastX and BlastN may be used. For example, in the survey analysis of the barley genome using 454 sequencing at 0.1× coverage, it was possible to determine 7.4% of the barley genes using BlastX on a database of predicted rice proteins and to characterize the presence of many TEs using BlastN on a TE plant database (Wicker *et al.*, 2009).

Another characteristic resulting from genome sampling with WGS and NGS technologies is to expect an increased coverage of the repetitive content. For example, with a genomic coverage of 0.01×, we can expect a coverage of 10× for each repeat occurring in the genome with 1000 copies (Macas *et al.*, 2007). With this characteristic, the identification of repeat content using a NGS survey is not limited to a homology search. The first program developed to use the expected increase in coverage for TE sequences is ReAS, which has already been referenced in section 1 of this chapter and allows us to assemble consensus TE sequences (Li *et al.*, 2005). The TEs must exist at a sufficient copy number to be recognized by their read number and must not be too degraded to have sufficient sequence similarities between their copies to be able to build a consensus sequence. The ReAS program starts by building a k-mer index of the reads. The high copy number k-mers are then picked out to retrieve the reads containing them. The reads are then assembled and expanded to recover the consensus sequences of the different elements. When possible, the contigs formed are linked using the paired-end information. The main drawback of this method is the fact that it is not designed for short or very short reads and cannot process reads less than 104 bp (Macas *et al.*, 2007). The AAARF (Assisted Automated Assembler of Repeat Families) program was designed to overcome this problem and can process short or very short reads (DeBarry *et al.*, 2008). This program uses one read as a query sequence to obtain its nucleotide coverage against the rest of the dataset using BlastN. This nucleotide coverage is then used to select the overlapping reads, which are then aligned using ClustalW to build a new query sequence. This program iteratively elongates each query sequence and assembles a set of TE contigs.

To recover repetitive sequences, it is also possible to cluster overlapping unannotated DNA sequences. This method was used to assemble 41% of the reads of the soybean genome into contigs using the Phrap program (Swaminathan *et al.*, 2007). A similar approach was used to assemble 31.6% of the unannotated reads into contigs for the barley genome survey (Wicker *et al.*, 2009). However, using this approach and because of the low depth of the data, some links are absent in the overlapping graphs, leading to contig fragmentation (Novák *et al.*, 2010). A slightly different approach was used to reconstruct the repeat sequences from a survey of the pea genome (Macas *et al.*, 2007). In this study, the program tclust from the TGICL package was used to cluster the reads based on a mutual similarity and to assemble each cluster into contigs. With this method, each cluster contains related repeat sequences, which can be used to better characterize the variability in the TE content of a genome. This method was successfully used to characterize the repeat content of the banana genome (Hribová *et al.*, 2010). One drawback of this more sensible approach is the formation of chimeric clusters, which are caused by the presence of reads spanning two TE sequences and form bridges

between two clusters (Macas *et al.*, 2007). More recently, a novel cluster-based approach was proposed with the program SeqGraphR (Novák *et al.*, 2010). This program uses a hierarchical agglomeration algorithm to cluster the reads and to characterize the TE sequences. This graph–based clustering allows for a better segregation of groups of unrelated sequences than tclust. Moreover, it allows for a better characterization of the cluster structure by computing various graph metrics to discriminate between different types of repeats. The assembly of TE sequences from a cluster results in the formation of consensus sequences. With this type of sequence representation, a significant amount of information about TE sequence variability is lost. SeqGraphR can provide an alternative representation of TEs by direct graph visualization. This approach can be very useful for deciphering contig assembly or for distinguishing between two closely related TEs. SeqGraphR was successfully used to characterize the TE content of three species of *Nicotiana tabacum* in a genome survey using 454 sequencing at 0.1× (Renny-Byfield *et al.*, 2011).

All previous methods work for genome survey data with relatively small dataset sizes, ranging from 33 Mb to 90 Mb, but may not be suitable for larger datasets. Moreover, approaches using read overlap information will only work for sparse genome survey data, where the only sequences that can be assembled come from repeat sequences. If the coverage reaches or exceeds 1×, most of the genome can be assembled, and the described approaches will lose their specificity for TE sequences. However, a genome survey dataset can always be generated from a deep genome sequencing output by randomly selecting a subset of reads to easily study the most abundant repeats. LTR retrotransposons can also be *de novo* identified using mapping approaches. For this class of TEs, the reads will pileup on the two LTRs, and a 'batman ears'-like structure will appear when using programs such as Tallymer or JellyFish if we allow for multi-read mapping (Kurtz *et al.*, 2008; Marçais and Kingsford, 2011).

Structural variant detection

One of the most interesting characteristics of TEs is their ability to replicate and to colonize a genome. This transposition activity can be studied between species or populations. NGS technologies allow us to study copy number variation (CNV) by sequencing pooled DNA from different individuals (or pool-Seq). The main advantages of these approaches are that the approaches are fast and not copy specific and offer a higher sensitivity compared to other technologies (Alkan *et al.*, 2011). To perform these types of studies, one needs an assembled reference genome, a database of the TE sequences (which can be built from the sequencing data) and, optionally, paired-end technology to better resolve the TE information.

For this purpose, the T-lex program was developed to compute the population frequencies of individual TE insertions (Fiston-Lavier *et al.*, 2011). This program is a pipeline using four modules. The first module uses RepeatMasker to identify TEs and their flanking regions in the reference genome. The second module uses MAQ (Li *et al.*, 2008) to determine the presence of TEs by mapping reads across the sequences formed by an identified TE and its flanking regions in the reference genome. The third module uses SHRiMP (Rumble *et al.*, 2009), which can align sequences with long gaps, to identify the absence of a TE insertion in the analysed populations by mapping reads spanning only the two flanking regions of the TE sequence in the reference genome. Finally, the last module combines the information of the previous modules to obtain the frequencies of each TE family in the populations. The second version of the T-lex program is able to automatically use paired-end information to detect novel TE insertions. By using a similar approach to T-lex, it was possible to successfully analyse the activity of TEs using pooled DNA samples from 114 isofemale lines of *D. melanogaster* (Kofler *et al.*, 2012). For this study, 80 million paired-end fragments were produced with the Illumina Genome Analyser IIx. These reads were mapped onto the reference genome, where all the repeats were first masked using RepeatMasker. The mapping step was then performed using the BWA-SW program (Li *et al.*, 2010). The authors were then able to identify novel insertions with at least three paired couples of reads, with, for each couple, one read mapping to a genomic locus and the others mapping to a TE sequence. In this study, novel TE insertions

were detected if they were present in at least 7% of the populations.

Without a reference genome, comparative studies of the TE content of different species can also be achieved using genome surveys. For example, an analysis of genomic gigantism in plethodontid salamanders was performed using a genome survey of six species sequenced with a coverage of 0.1× using 454 technology to obtain reads of a maximum length of 400 bp (Sun *et al.*, 2012). In this study, the RepeatModeler program (Smit and Hubley 2008–2010) was used to identify those repeats covered by a minimum of four reads. The REPCLASS program (Feschotte *et al.*, 2009) was then used to further classify the unknown repeats. The results demonstrated that the analysed salamander species accumulate large amounts of LTR-retrotransposons compared to other vertebrates.

Analysis of TE regulation by the host

With the prevalence of TEs and their capacity to invade a genome, it is crucial for the host to be able to regulate their activity to avoid too many deleterious effects. During the past few years, the links between TEs and the epigenetic systems of regulation, such as DNA methylation, histone modifications, and RNA interference, have been shown to be linked to the repeat content of a genome (Siomi and Siomi, 2008; Rebollo *et al.*, 2010). In particular, diverse RNA-mediated defences have been discovered in different eukaryotic organisms (Slotkin and Martienssen, 2007; Blumenstiel, 2011). These discoveries have allowed for the development of new models for TE dynamics in natural populations in which four phases have been described: an initial phase of TE invasion; a second phase of TE proliferation, leading to the appearance of TE insertion alleles initiating the production of small RNAs; and finally a quiescent state, leading to the stabilization of TE copy number (see (Blumenstiel, 2011) for a review).

With the development of NGS technologies, it has become easier to analyse the epigenetic control of TEs. Particular modifications can regulate TE activity, such as DNA methylation. This type of modification can be determined using a MeDIP-seq approach and has already been used in several organisms. For example, in black cottonwood, the TEs possessed variable methylation according to their family, with LTR retrotransposons being globally more methylated than other classes (Vining *et al.*, 2012).

RNA-Seq is also a reliable way to study active TEs because the complete RNA sequences of the TEs can be found in the data output. The control of TEs can also be studied when considering small RNAs such as piRNAs or siRNAs, which are expected to be copy specific to the TEs that they control. This type of analysis can also be used to validate *de novo* annotations of TEs. The results of these approaches are naturally linked with the condition or the stage where the different TEs are expressed. Due to piRNA regulation, the quantification aspect of RNA-Seq is of a lesser interest for the study of TEs because the number of transcripts is not directly correlated with the activity of a TE (Brennecke *et al.*, 2007). However, these studies allow us to obtain information on the potentially complete and active copies that are inserted into the genome. The analysis of small RNAs is typically performed by mapping the reads on to a reference genome to help identify clusters of small RNAs and to determine which copy in the genome is associated with a particular small RNA. This approach has been used in the analysis of the control of particular TEs in *D. melanogaster* (Brennecke *et al.*, 2007; Brennecke *et al.*, 2008; Grentzinger *et al.*, 2012) and in plant species (Hollister *et al.*, 2011). Of course, these genome-mapping approaches have limitations due to the differences that exist between individuals. For example, in the case of an analysis of *P*-elements, it was not possible to map them onto the *D. melanogaster* reference genome because this particular element is absent in the sequenced strain (Brennecke *et al.*, 2008). Thus, it was necessary to find another strategy, in this case, using those reads that did not map onto the reference genome.

The regulation of TEs can also be linked to the histone modifications of the DNA. Using ChIP-seq data, it is possible to determine what types of modifications are associated with TEs or with genes given their TE neighbourhood. For example, in mouse embryonic stem cells, an analysis of ChIP-seq data revealed that the

majority of gene promoters surrounded by numerous TEs were depleted of the bivalent marks H3K27me3 + H3K4me3 compared to genes surrounded by few or no TEs. This bivalent mark has been demonstrated to be specific to a 'poised state' of developmental genes that are temporarily repressed in embryonic stem cells but that will be activated later during development (Zhang and Mager, 2012). This previous analysis indirectly observed modifications associated with TEs. Generally, during the mapping step of reads sequenced using ChIP-seq analysis, the reads cannot be associated with TE copies because only uniquely mappable reads are conserved (those having a unique location on the genome). Thus, an alternative mapping approach was proposed by Huda *et al.* to allow for the direct examination of ChIP-seq reads associated with TEs (Huda *et al.* 2010). In that analysis, the authors took advantage of several ChIP-seq experiments, which allowed them access to a genome-wide map of 38 histone modifications in human $CD4^+$ T cells (Barski *et al.*, 2007; Wang *et al.*, 2008). They used the MAQ program to align the reads, allowing for redundant genomic locations. It was then possible to characterize what type of TE was present in the mapped reads. Their results demonstrated a high variation in TE histone modifications according to the TE family, with the older TE families and the TEs close to genes carrying more modifications than the younger TE families and those TEs distal to genes.

These different examples demonstrate how valuable NGS data are in the analysis of TEs, and also identify the need for specific tools to handle these data for studying TEs.

Conclusions

Transposable elements are important components of genomes that cannot simply be put aside when analysing genomes. It is important to understand how TEs function and evolve to better understand all of the impacts of TEs on genome evolution. Given the large amount of genomic data that has been continuously generated, this task becomes more and more difficult. However, these data allow us an access to new information that we did not have several years ago.

Since the first sequencing projects were undertaken, efforts have been made to develop programs that allow for the detection and analysis of TE sequences. Various programs have been developed that use different or complementary approaches to detect TE sequences. These tools have very different performances, but even the best ones cannot discover all the TE sequences in a genome because each has its own drawback(s) that prevents it from finding each and every TE (Saha *et al.*, 2008b; Lerat, 2010). Thus, the best approach to exhaustively describe the landscape of TEs in a genome is to use several of these different programs and to cross-reference the results. Similarly, the best approach to locate TE sequences in complete genomes appears to reside in the use of pipelines of programs. For example, the RepeatModeler pipeline includes different programs to build, refine and classify consensus sequences of putative interspersed repeats (Smit and Hubley, 2008–2010). The REPET pipeline has been built to integrate the findings of similarity- and *de novo*-based programs (Quesneville *et al.*, 2005). This pipeline was recently updated to retain those programs that provide the best results after the authors tested different *de novo* programs (Flutre *et al.*, 2011). Other pipelines have generally been developed to answer specific questions (see Lerat, 2010).

In all cases, another important step after the identification of putative TE sequences is the classification of the repeats into families. This is a difficult step because it must take into account the biological aspect of TEs, such as the fact that some copies can be fragmented and thus not only full-length elements exist in a genome and that TEs often insert inside each other, producing what are known as nested TEs. Some programs have been developed to integrate the classification step, such as the TEclass program, which tries to determine the main classes of unknown elements using machine-learning algorithms (Abrusán *et al.*, 2009), or the REPCLASS program, which uses different approaches to annotate TEs (Feschotte *et al.*, 2009).

With the new type of genomics data generated by next-generation sequencing technologies, it is necessary to develop new approaches to detect and analyse TEs. Indeed, most of the programs

that have been designed for classic genomic data cannot handle these new types of data. This is predominantly because NGSs produce small sequence fragments, which increase the difficulties involved in assembling the repeat content of a genome, but also because the amounts of these data are too large to be handled with the existing tools. However, new programs have been designed to take these problems into account. Even if all these programs are not specific to TEs, specific programs are now available to answer particular questions on TEs with regards to population genomics, and we can hope for new developments in more specific areas. Questions about the epigenetic regulation of TEs are particularly important at different levels, such as the impact of TEs in cancer development in humans. NGS data now offer the possibility of having access to this information, which will necessitate the development of particular tools specific for TEs.

Future trends

With the development of NGS data, access to individuals' sequence data should provide us with valuable information on the specific content of TEs, allowing us to be more precise in terms of the insertion profiles of TEs. Currently, the only available possibility is to compare data to reference genomes. However, this shortcut is not precise enough when delving deeper into our understanding of the mechanisms of TE dynamics.

Web resources

A list of existing TE detection tools is available at:

- http://bergmanlab.smith.man.ac.uk/?page_id=295
- Quesneville. *BLASTER suite* <http://urgi.versailles.inra.fr/Tools/Blaster>.
- Smit, AFA, Hubley, R and Green, P. *RepeatMasker Open-3.0*. 1996–2010 <http://www.repeatmasker.org>.
- Smit, AFA, Hubley, R. *RepeatModeler Open-1.0*. 2008–2010 <http://www.repeatmasker.org>.

References

Abrusán, G., Grundmann, N., DeMester, L., and Makalowski, W. (2009). Teclass – a tool for automated classification of unknown eukaryotic transposable elements. Bioinformatics *25*, 1329–1330.

Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nat. Rev. Genet. *12*, 363–376.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Bao, Z., and Eddy, S.R. (2002). Automated *de novo* identification of repeat sequence families in sequenced genomes. Genome Res. *12*, 1269–1276.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., *et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53–59.

Bergman, C.M., and Quesneville, H. (2007). Discovering and detecting transposable elements in genome sequences. Brief. Bioinform. *8*, 382–392.

Biémont, C. (2010). A brief history of the status of transposable elements: from junk DNA to major players in evolution. Genetics *186*, 1085–1093.

Biémont, C., and Vieira, C. (2006). Genetics: junk DNA as an evolutionary force. Nature *443*, 521–524.

Blumenstiel, J.P. (2011). Evolutionary dynamics of transposable elements in a small RNA world. Trends Genet. *27*, 23–31.

Bowen, N.J., and McDonald, J.F. (2001). Drosophila euchromatic LTR retrotransposons are much younger than the host species in which they reside. Genome Res. *11*, 1527–1540.

Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G.J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell *128*, 1089–1103.

Brennecke, J., Malone, C.D., Aravin, A.A., Sachidanandam, R., Stark, A., and Hannon, G.J. (2008). An epigenetic role for maternally inherited piRNAs in transposon silencing. Science *322*, 1387–1392.

Crawford, J.E., Guelbeogo, W.M., Sanou, A., Traoré, A., Vernick, K.D., Sagnon, N., and Lazzaro, B.P. (2010). *De novo* transcriptome sequencing in *Anopheles funestus* using Illumina RNA-seq technology. PLoS One *5*, e14202.

Cutter, A.D. (2008). Divergence times in Caenorhabditis and Drosophila inferred from direct estimates of the neutral mutation rate. Mol. Biol. Evol. *25*, 778–786.

DeBarry, J.D., Liu, R., and Bennetzen, J.L. (2008). Discovery and assembly of repeat family pseudomolecules from sparse genomic sequence data using the assisted automated assembler of repeat families (AAARF) algorithm. BMC Bioinformatics *9*, 235.

Dowsett, A.P., and Young, M.W. (1982). Differing levels of dispersed repetitive DNA among closely related species of Drosophila. Proc. Natl. Acad. Sci. U.S.A. *79*, 4570–4574.

Edgar, R.C., and Myers, E.W. (2005). PILER: identification and classification of genomic repeats. Bioinformatics *21* (Suppl 1), i152–i158.

Eickbush, T.H., and Furano, A.V. (2002). Fruit flies and humans respond differently to retrotransposons. Curr. Opin. Genet. Dev. *12*, 669–674.

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. BMC Bioinformatics *9*, 18.

Evgen'ev, M.B., Zelentsova, H., Poluectova, H., Lyozin, G.T., Veleikodvorskaja, V., Pyatkov, K.I., Zhivotovsky, L.A., and Kidwell, M.G. (2000). Mobile elements and chromosomal evolution in the virilis group of Drosophila. Proc. Natl. Acad. Sci. U.S.A. *97*, 11337–11342.

Falgueras, J., Lara, A.J., Fernández-Pozo, N., Cantón, F.R., Pérez-Trabado, G., and Claros, M.G. (2010). SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. BMC Bioinformatics *11*, 38.

Feschotte, C., Keswani, U., Ranganathan, N., Guibotsy, M.L., and Levine, D. (2009). Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. Genome Biol. Evol. *1*, 205–220.

Fiston-Lavier, A.-S., Carrigan, M., Petrov, D.A., and González, J. (2011). T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. Nucleic Acids Res. *39*, e36.

Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. (2011). Considering transposable element diversification in *de novo* annotation approaches. PLoS One *6*, e16526.

Frith, M.C., Wan, R., and Horton, P. (2010). Incorporating sequence quality data into alignment improves DNA read mapping. Nucleic Acids Res. *38*, e100.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., *et al.* (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. U.S.A. *108*, 1513–1518.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., *et al.* (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.

Grentzinger, T., Armenise, C., Brun, C., Mugat, B., Serrano, V., Pelisson, A., and Chambeyron, S. (2012). piRNA-mediated transgenerational inheritance of an acquired trait. Genome Res. *22*, 1877–1888.

Han, Y., and Wessler, S.R. (2010). MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res. *38*, e199.

Hancks, D.C., and Kazazian, H.H. (2012). Active human retrotransposons: variation and disease. Curr. Opin. Genet. Dev. *22*, 191–203.

He, S., Wurtzel, O., Singh, K., Froula, J.L., Yilmaz, S., Tringe, S.G., Wang, Z., Chen, F., Lindquist, E.A., Sorek, R., *et al.* (2010). Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. Nat. Methods *7*, 807–812.

Hollister, J.D., Smith, L.M., Guo, Y.-L., Ott, F., Weigel, D., and Gaut, B.S. (2011). Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. Proc. Natl. Acad. Sci. U.S.A. *108*, 2322–2327.

Hribová, E., Neumann, P., Matsumoto, T., Roux, N., Macas, J., and Dolezel, J. (2010). Repetitive part of the banana (*Musa acuminata*) genome investigated by low-depth 454 sequencing. BMC Plant Biol. *10*, 204.

Hu, T.T., Pattyn, P., Bakker, E.G., Cao, J., Cheng, J.-F., Clark, R.M., Fahlgren, N., Fawcett, J.A., Grimwood, J., Gundlach, H., *et al.* (2011). The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat. Genet. *43*, 476–481.

Huda, A., Mariño-Ramírez, L., and Jordan, I.K. (2010). Epigenetic histone modifications of human transposable elements: genome defense versus exaptation. Mob. DNA *1*, 2.

Janicki, M., Rooke, R., and Yang, G. (2011). Bioinformatics and genomic analysis of transposable elements in eukaryotic genomes. Chromosome Res. *19*, 787–808.

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

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

Kelley, D.R., Schatz, M.C., and Salzberg, S.L. (2010). Quake: quality-aware detection and correction of sequencing errors. Genome Biol. *11*, R116.

Kim, J.M., Vanguri, S., Boeke, J.D., Gabriel, A., and Voytas, D.F. (1998). Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. Genome Res. *8*, 464–478.

Kimura, K., and Kidwell, M.G. (1994). Differences in *P* element population dynamics between the sibling species *Drosophila melanogaster* and *Drosophila simulans*. Genet. Res. *63*, 27–38.

Kofler, R., Betancourt, A.J., and Schlötterer, C. (2012). Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. PLoS Genet. *8*, e1002487.

Kurtz, S., and Schleiermacher, C. (1999). REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics *15*, 426–427.

Kurtz, S., Narechania, A., Stein, J.C., and Ware, D. (2008). A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics *9*, 517.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Lerat, E. (2010). Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity *104*, 520–533.

Lerat, E., Rizzon, C., and Biémont, C. (2003). Sequence divergence within transposable element families in the *Drosophila melanogaster* genome. Genome Res. *13*, 1889–1896.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. *18*, 1851–1858.

Li, R., Ye, J., Li, S., Wang, J., Han, Y., Ye, C., Wang, J., Yang, H., Yu, J., Wong, G.K.-S., *et al.* (2005). ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. PLoS Comput. Biol. *1*, e43.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., *et al.* (2010). The sequence and *de novo* assembly of the giant panda genome. Nature *463*, 311–317.

Li, Y.J., Satta, Y., and Takahata, N. (1999). Paleo-demography of the *Drosophila melanogaster* subgroup: application of the maximum likelihood method. Genes Genet. Syst. *74*, 117–127.

Lisch, D.R., and Kidwell, M.G. (2000). Transposable elements and host genome evolution. Trends Ecol. Evol. *15*, 95–99.

Lockton, S., and Gaut, B.S. (2010). The evolution of transposable elements in natural populations of self-fertilizing *Arabidopsis thaliana* and its outcrossing relative *Arabidopsis lyrata*. BMC Evol. Biol. *10*, 10.

Lucier, J.-F., Perreault, J., Noël, J.-F., Boire, G., and Perreault, J.-P. (2007). RTAnalyzer: a web application for finding new retrotransposons and detecting *L1* retrotransposition signatures. Nucleic Acids Res. *35*, W269–W274.

Macas, J., Neumann, P., and Navrátilová, A. (2007). Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. BMC Genomics *8*, 427.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376–380.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. *18*, 1509–1517.

Martin, J.A., and Wang, Z. (2011). Next-generation transcriptome assembly. Nat. Rev. Genet. *12*, 671–682.

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics *27*, 764–770.

McClintock, B. (1956). Controlling elements and the gene. Cold Spring Harb. Symp. Quant. Biol. *21*, 197–216.

Novák, P., Neumann, P., and Macas, J. (2010). Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics *11*, 378.

Ohshima, K., Hattori, M., Yada, T., Gojobori, T., Sakaki, Y., and Okada, N. (2003). Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. Genome Biol. *4*, R74.

Pevzner, P.A., Tang, H., and Waterman, M.S. (2001). An Eulerian path approach to DNA fragment assembly. Proc. Natl. Acad. Sci. U.S.A. *98*, 9748–9753.

Phillippy, A.M., Schatz, M.C., and Pop, M. (2008). Genome assembly forensics: finding the elusive mis-assembly. Genome Biol. *9*, R55.

Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). *De novo* identification of repeat families in large genomes. Bioinformatics *21* (Suppl 1), i351–i358.

Quesneville, H., Bergman, C.M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M., and Anxolabehere, D. (2005). Combined evidence annotation of transposable elements in genome sequences. PLoS Comput. Biol. *1*, 166–175.

Rebollo, R., Horard, B., Hubert, B., and Vieira, C. (2010). Jumping genes and epigenetics: towards new species. Gene *454*, 1–7.

Renny-Byfield, S., Chester, M., Kovařík, A., Comber, S.C.L., Grandbastien, M.-A., Deloger, M., Nichols, R.A., Macas, J., Novák, P., Chase, M.W., *et al.* (2011). Next generation sequencing reveals genome downsizing in allotetraploid *Nicotiana tabacum*, predominantly through the elimination of paternally derived repetitive DNAs. Mol. Biol. Evol. *28*, 2843–2854.

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S.D., Mungall, K., Lee, S., Okada, H.M., Qian, J.Q., *et al.* (2010). De novo assembly and analysis of RNA-seq data. Nat. Methods *7*, 909–912.

Rumble, S.M., Lacroute, P., Dalca, A.V., Fiume, M., Sidow, A., and Brudno, M. (2009). SHRiMP: accurate mapping of short color-space reads. PLoS Comput. Biol. *5*, e1000386.

Sacomoto, G.A.T., Kielbassa, J., Chikhi, R., Uricaru, R., Antoniou, P., Sagot, M.-F., Peterlongo, P., and Lacroix, V. (2012). KISSPLICE: *de-novo* calling alternative splicing events from RNA-seq data. BMC Bioinformatics *13* (Suppl 6), S5.

Saha, S., Bridges, S., Magbanua, Z., and Peterson, D. (2008a) Computational approaches and tools used in identification of dispersed repetitive DNA sequences. Trop. Plant Biol. *1*, 85–96.

Saha, S., Bridges, S., Magbanua, Z.V., and Peterson, D.G. (2008b) Empirical comparison of *ab initio* repeat finding programs. Nucleic Acids Res. *36*, 2284–2294.

Schbath, S., Martin, V., Zytnicki, M., Fayolle, J., Loux, V., and Gibrat, J.-F. (2012). Mapping reads on a genomic sequence: an algorithmic overwiew and a practical comparative analysis. J. Comput. Biol. *19*, 796–813.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., *et al.* (2009). The B73 maize genome: complexity, diversity, and dynamics. Science *326*, 1112–1115.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M., and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. Genome Res. *19*, 1117–1123.

Siomi, H., and Siomi, M.C. (2008). Interactions between transposable elements and Argonautes have (probably) been shaping the Drosophila genome throughout evolution. Curr. Opin. Genet. Dev. *18*, 181–187.

Slotkin, R.K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. Nat. Rev. Genet. *8*, 272–285.

Sun, C., Shepard, D.B., Chong, R.A., López Arriaza, J., Hall, K., Castoe, T.A., Feschotte, C., Pollock, D.D., and Mueller, R.L. (2012). LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. Genome Biol. Evol. *4*, 168–183.

Surget-Groba, Y., and Montoya-Burgos, J.I. (2010). Optimization of *de novo* transcriptome assembly from next-generation sequencing data. Genome Res. *20*, 1432–1440.

Swaminathan, K., Varala, K., and Hudson, M.E. (2007). Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey. BMC Genomics *8*, 132.

Szak, S.T., Pickeral, O.K., Makalowski, W., Boguski, M.S., Landsman, D., and Boeke, J.D. (2002). Molecular archeology of L1 insertions in the human genome. Genome Biol. *3*, research0052.

Tamura, K., Subramanian, S., and Kumar, S. (2004). Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. Mol. Biol. Evol. *21*, 36–44.

Tang, H. (2007). Genome assembly, rearrangement, and repeats. Chem. Rev. *107*, 3391–3406.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Treangen, T.J., and Salzberg, S.L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. *13*, 36–46.

Tu, Z., Li, S., and Mao, C. (2004). The changing tails of a novel short interspersed element in *Aedes aegypti*: genomic evidence for slippage retrotransposition and the relationship between 3′ tandem repeats and the poly(dA) tail. Genetics *168*, 2037–2047.

Vieira, C., Lepetit, D., Dumont, S., and Biémont, C. (1999). Wake up of transposable elements following *Drosophila simulans* worldwide colonization. Mol. Biol. Evol. *16*, 1251–1255.

Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Peng, W., Zhang, M.Q., *et al.* (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. Nat. Genet. *40*, 897–903.

Wetzel, J., Kingsford, C., and Pop, M. (2011). Assessing the benefits of using mate-pairs to resolve repeats in *de novo* short-read prokaryotic assemblies. BMC Bioinformatics *12*, 95.

Wicker, T., Schlagenhauf, E., Graner, A., Close, T.J., Keller, B., and Stein, N. (2006). 454 sequencing put to the test using the complex genome of barley. BMC Genomics *7*, 275.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., *et al.* (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., and Stein, N. (2009). A whole-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.

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. *18*, 821–829.

Zhang, Y., and Mager, D.L. (2012). Gene properties and chromatin state influence the accumulation of transposable elements in genes. PLoS One *7*, e30158.