Chapter 6

Methods and Software in NGS for TE Analysis

Cristian Chaparro and Francois Sabot

Abstract

The recent development of next-generation sequencing (NGS) technologies allowed various authors to imagine, test, and validate new approaches for TE analysis, in their nature, type, activity, or quantity. In this chapter, we describe briefly the technologies used, then the various approaches and methods used already, and finally some potential new methods. In contrast to the more molecular chapters of the book, the approaches described here are purely bioinformatics, and have a set of NGS data as a starting point. Moreover, as these analyses are quite recent in the field, most of them were only performed once, and we cannot be sure that they could be reused in other species or context than the original one. However, there are a lot of interesting approaches and results that NGS can provide in the TE field.

Key words: NGS, Repeats, Interspersion, Transposon, Retrotransposon

Abbreviations

BAC	Bacterial artificial chromosome
LTR	Long terminal repeat
NGS	Next-generation sequencing
siRNA	Small interfering RNA
SNP	Single-nucleotide polymorphism
TE	Transposable element
VS	Structural variant

1. Introduction to NGS Technologies

Currently, most of the transposable element (TE) analyses using next-generation sequencing (NGS) are performed with 454 or Illumina data. The other NGS technology available at the time of

Yves Bigot (ed.), *Mobile Genetic Elements: Protocols and Genomic Applications*, Methods in Molecular Biology, vol. 859,
DOI 10.1007/978-1-61779-603-6_6,

writing, SOLiD, is mainly used in metagenomics, and is not used in TE analysis. These last years, the companies developing these technologies have concentrated at making them more efficient and in making the runs cheaper, thus lowering the cost per base. With the apparition of IonTorrent (now owned by Life Technologies) on the market with its bench sequencer, the race to produce small versions of the 454 (GS Junior) and Illumina (MySeq Personal Sequencing System) machines, affordable by more scientists, has begun. While these smaller versions have smaller throughputs, they present an undeniable interest to the mass of laboratories that cannot afford the bigger versions. One other technology is entering the field, the SMAR-T technology from Pacific Biosciences, which promises long reads in the order of several thousand base pairs.

1.1. The Pyrosequencing/454

The 454 technology is based on the detection of the released phosphate during DNA synthesis. The system will detect the A, then the T, then the G, and then the C, sequentially. If three A's are following, the machine will detect a threefold increase in signal and provide a three A information for peak calling. The current technology allows up to one million reads of 300–500 bases each (Titanium kit) per run. Thus, we obtain a quite large array of long reads and 400 Mb per run in less than 2 days after DNA (or RNA) extraction to results. The main limit with this technology is that for homopolymers longer than four bases (e.g., more than four As in a row) the detection system reaches the limits and cannot decipher the correct number of bases. The company released a smaller more affordable version of their machine (GS Junior), which still produces 400-bp reads and can produce 35 Mbases per run in 10 h. This technology, being the oldest of NGS, is fully mature and no more technical advances are expected.

1.2. The Solexa/Illumina

The second NGS technology used in TE studies is the Illumina system. This system relies on the detection of fluorescent dyes released after the incorporation of the tagged bases. It will detect one base at a time, thus avoiding the homopolymers' effect from 454. The current machines (HiSeq2000) provide a minimum of 50 millions of clusters per lane (8 lanes per flow cell), in which 100 high-quality bases are read from each border (pair-end sequencing). Thus, around 10 Gbases of sequences per lane are expected in around 8 days (the sequencing lasts for 6 days) with a total of 180–200 Gbytes of raw data per run. The main limit is the size of the reads that are produced, currently two times 100 bases and two times 150 bases expected for the end of 2011. The Illumina company continues the development of this technology, increasing read length, read quality, and cluster density, and as mentioned above, they also started recently to propose a smaller bench version of their machine, the MySeq, providing 3.2 million clusters per run, representing from 120 Mbases (for 35-bases reads) to 1 Gbases (for 150-bases paired-end reads).

1.3. The Upcoming Technologies

Currently, other technologies are being developed and perfected and will provide new possibilities for TE analysis. These are the IonTorrent/PGM (Personal Genome Machine) and PacificBiosciences/SMAR-T technologies. The first, based on a variant of pyrosequencing, aims at providing cheap systems with a limited amount of reads (100,000 of 100–300 bases, for $500). The second technology is geared toward de novo sequencing and the production of long reads (currently, 3 kb, up to 40 kb announced). While these machines do not provide enough throughput to analyze whole large eukaryotic genomes for the moment, an array of new analyses will become available for individual laboratories, from variant detection to sequencing of BAC pools. These next years will then see the emergence of new methods and ideas for TE analysis.

2. Approaches

2.1. Recurrence of Sequences and Recurrence of Insertions

One straightforward possibility of using NGS is to detect TEs using their "More-Than-One-Copy" property, as was performed by Wicker et al. (1). For that, authors used a 0.1× (10% of the genome, which represents a low coverage) sequencing of the barley genome. Using this set of Illumina data, they counted the number of occurrences of specific k-mers. Those k-mers were small sequences of 20 bases, and for each sequence they can report then a score representing its frequency in the genome. Plotting this frequency on the sequence of barley BACs, i.e., reporting the corresponding k-mers' frequency at each position (1–20, 2–21, etc.), they generated graphs representing the repeated level of the underlying sequence. The higher the plots, the more repeated the region is. The automatic identification was compared to the manually curated annotations available for those BACs, and more than 90% of the TE annotations were common between the two methods. Thus, even without any previous knowledge, it is possible to perform repeats annotation on a genome and even to detect LTR retrotransposons (their specific features of two quite identical LTR for each copy provide a "Batman ears" effect on the graph; Fig. 1). The software used here is *Tallymer* (2), on a currently standard bioinformatics personal machine. The counts were plotted logarithmically to allow better viewing. To perform such analysis, only short reads are needed, as no identification is possible with this method. The original paper was performed with 36-bases single-end reads from an Illumina GA I sequencer. Such lengths are still available, or more generally a 50-bases single-end set of sequences. The only prerequisite is to have clean data, i.e., without any adapters or primer or vector contamination. Other tools, such as *JellyFish* (3), are also able to perform fast and powerful k-mers counting, reducing the

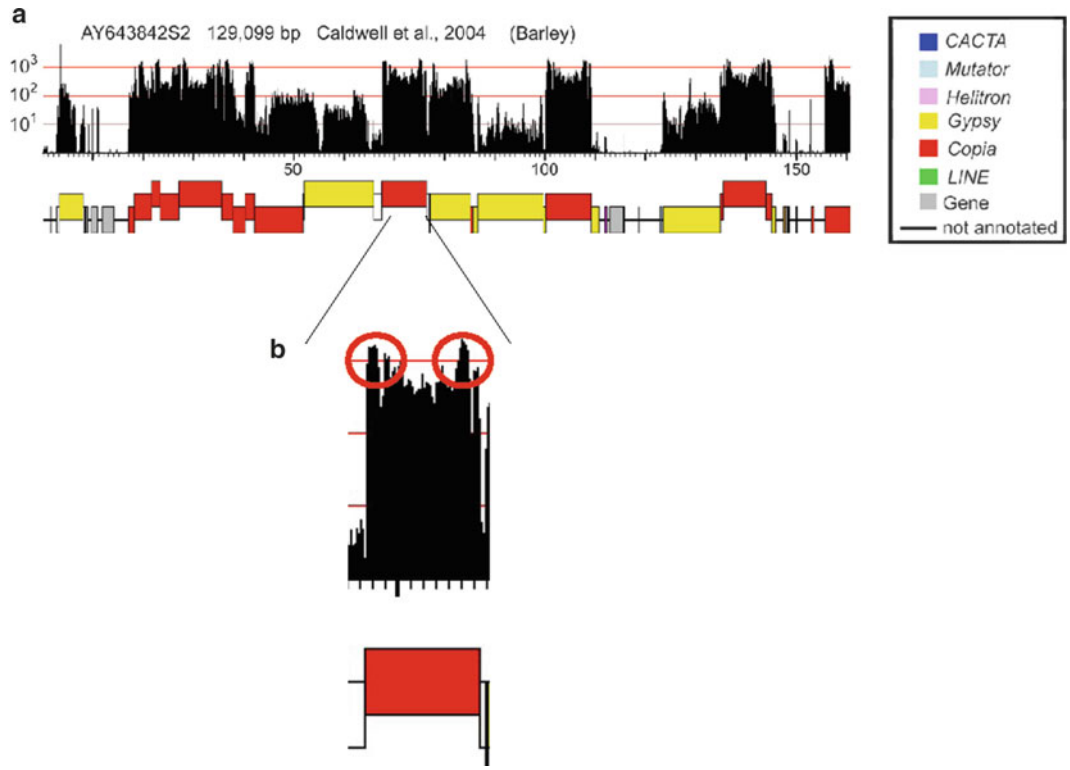

Fig. 1. Automatic TE detection through k-mers count. Images came from Wicker et al. (1). (**a**) Automatic annotation (k-mers logarithmic count) is compared to manual curated annotation. (**b**) "Batman ears" effect produced by the LTRs surrounding a *Copia* element.

time and memory necessary for doing these kind of analyses. However, this method is limited to the possibility of detecting repeated regions without being able to detect the precise boundary of elements or identify them.

2.2. Identification of New Elements Using the Nonassembled Reads in De Novo Sequencing

The amount of data produced by NGS technologies, which makes sequencing a genome a matter of days, led the community to the development of software that would assemble whole genomes. The most important problem that is hindering the assembly of genomes, to at least the same quality of the more expensive and time-consuming Sanger-sequenced genomes, is that due to the short reads the repeated regions cannot be spanned by the assembling algorithms. The repeated nature of these regions produces dead ends or many possible paths in DeBruijn graphs, commonly used in NGS assembly algorithms (Fig. 2). Gnerre et al. (4) have approached this issue in their *Allpaths-LG* software by producing libraries that span different lengths, and thus have shown that they can achieve the same finishing quality as Sanger-assembled genomes. They rely on very high coverage, and they have partly

Fig. 2. Simplified schema of a DeBruijn graph depicting the problem of repeated sequences. Repeat regions can be depicted as the crossing of several paths through the genome. When extending region 1, eventually the extended sequence will enter the repeat region. At the end of the repeat region, the graph will have multiple possible paths to follow, in this simplified case region 1 and region 2. As the correct path cannot be established, the sequence will be reverse complemented and the algorithm will start tracking back reaching another impasse at the other end of the repeat region.

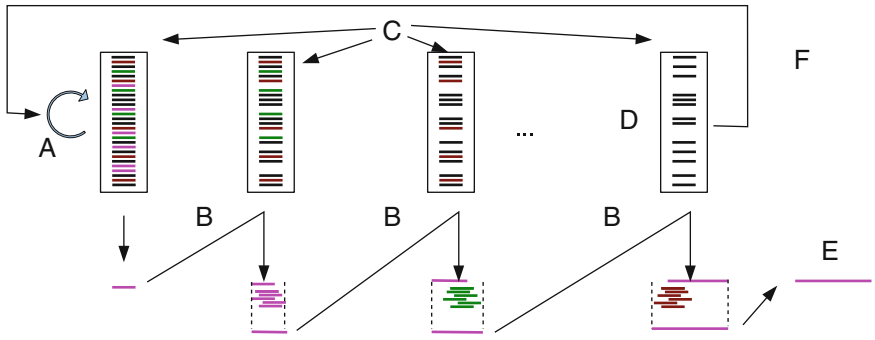

Fig. 3. Search for repeated elements in nonassembled reads. (**a**) Blast the reads database against itself in order to select the most abundant read. (**b**) Blast of the extended read to the reads database. (**c**) The reads that match the extension are extracted from the nonassembled reads database. (**d**) When the extension cannot continue, reverse complement the sequence and continue extending the other end. (**e**) The identified reference sequence for the repeat element is manually curated.

solved the repetitive sequence problem by collapsing read pairs if there is enough evidence that they overlap or span a gap. Still, over 60% of the unaligned reads pertains to transposable elements.

One can already suspect that most of the nonassembled sequences in such cases belong to TE sequences. So, new methods were developed to identify new TEs in the nonassembled portion of sequences. It consists on taking the most represented reads from the pool of nonassembled sequences, and then these sequences are extended by blasting one end against the remaining nonassembled reads. In this case, the best matching sequence that extends the repeated sequence is kept and all similar sequences are deleted from the database. This extension process is repeated until the sequence cannot be extended any more. The next step extends the other end of the sequence until no more sequences can be added or an overlap with the already assembled sequence is detected (Fig. 3). Afterward, the whole process is repeated and the remaining pool of sequences are searched for other repeated elements.

This method was used in Cocoa (5), where the *Gaucho* LTR retrotransposon was identified. Generally, homemade customized scripts are used for such purpose. The sequences identified with this method are considered as reference sequences and serve as a representative sequence of a family as it is almost impossible to identify a true genomic copy through this method.

2.3. Transcriptome and Expressed TEs

In a more classical way, the NGS can be used in transcriptomic analyses to detect expressed TEs. The method is more reliable and reproducible than microarray detection (6), and also provides more information, as all expressed sequences can be detected, even the currently unknown. The main limit in such analysis is the same as in microarray, i.e., identifying the stages or conditions where the TEs are expressed. However, the multiplexing capacities of NGS technologies provide an increased potential of identifying expressed TEs.

2.4. Improved Annotation of TEs Using siRNA

Small interfering RNAs (siRNAs) are supposed to be generated in order to block the expression of TEs (7). Thus, if a set of siRNAs are sequenced and mapped upon their reference genome, one can determine genomic loci that present an important production of siRNA. One could infer that such loci might pertain to repeated sequences. Thus, if a region is identified by an important set of siRNA but is not annotated yet as containing repeats, it will become a good candidate place to identify a new type of TE.

2.5. Identification of Active TEs Through Structural Variant Detections

The main interesting application of NGS technology is the detection of actively transposing elements by structural variant detection. While transcriptomics can be used to identify transcribed elements, it is not possible to know if these transcribed elements are capable of completing the transposition cycle. Using a whole-genome sequence as a template, it is possible to detect structural variants (SV) in this genome. The idea is to detect an anomaly in the mapping of a pair in two far away positions or in two chromosomes, or even with only one mate mapped (Fig. 4). The method, originally used on human genome, is extensively employed today in tumor cell studies, and recently in the TE field. In Asian rice, this method was used to show that in a single regeneration event at least 34 new insertions appeared in the close vicinity of genes (8). In *Arabidopsis*, this method was successfully used to detect the neoinsertions of the *Evade* elements (9). Recently, Fiston-Lavier et al. (10) released a program called *T-Lex* allowing the detection of specific copies in a sample compared to a reference. The software relies on the detection of NGS reads that map to the border between the repeated element and its surrounding genomic region. The strength of this software is that it will return an estimate of the population frequency for each annotated transposable element. Such method is very useful and powerful for population genomics of TEs and

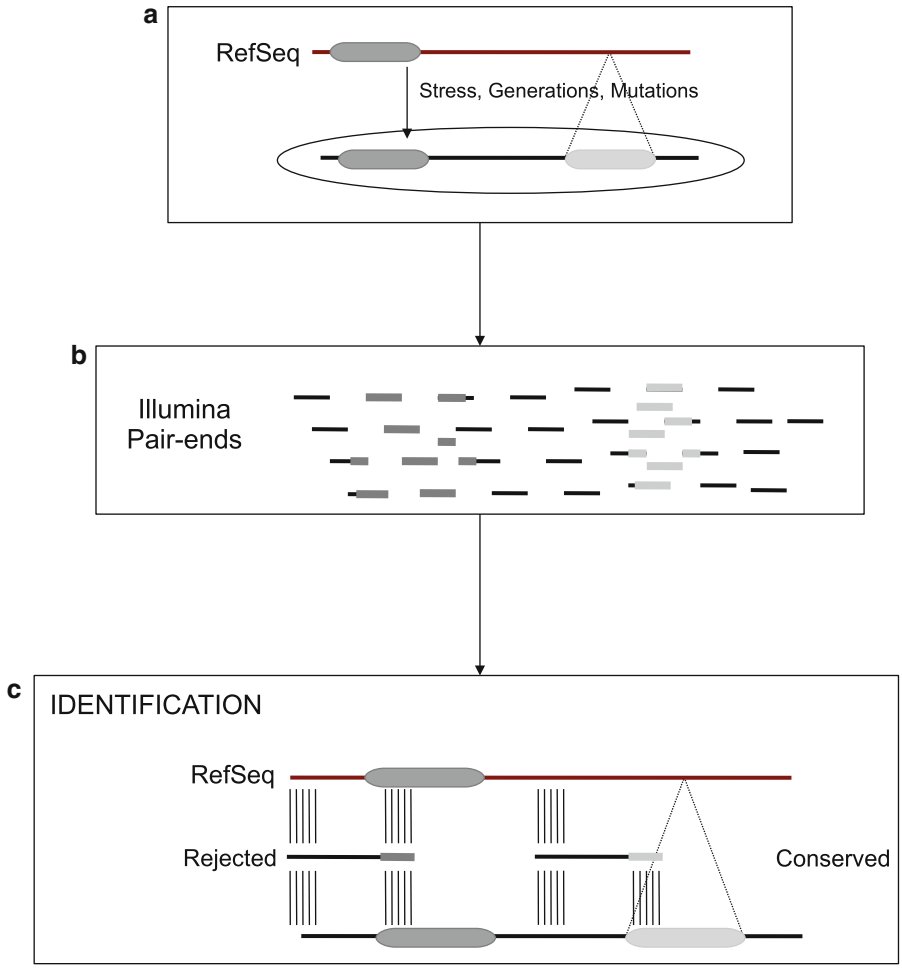

Fig. 4. Detection of structural variants. (**a**) A genome for which a reference (RefSeq) is available is resequenced after stress/generations/mutations. (**b**) The sequencing is performed using (Illumina or 454) pair-ends fragments. (**c**) Pairs are mapped on the RefSeq, and mapping anomalies are detected.

for studying the potential impact of specific insertions upon a specific phenotype.

This method relies on the initial knowledge of annotated TE insertions and on the existence of the reference genome. Only a good coverage of the genomic sequence is needed. One drawback is that this approach is not efficient in identifying insertions of transposable elements into other repeated or duplicated sequences as the flanking regions are not unique.

In the same way, markers based on TEs can be quickly generated using this method, with a bulk of DNAs from various individuals, and in detecting all the SV in the mix. Those SV positions can then be transformed in retrotransposon-based insertion polymorphism (RBIP)-like markers, suitable for genetic screening or positional cloning.

This is generally performed using pair-ended reads, sometimes using mate paired. This can also in theory be performed using longer single reads provided that a very stringent analysis is performed, as anomalies in mapping will probably result in a "non-mapped" state of the overlapping read.

3. Methods

3.1. Mapping with a Direct Reference

If one possesses the corresponding genome (or transcriptome) from the studied organism, the main used method (and the easiest) is the mapping. It consists in positioning the reads at the most probable position upon the reference. The software generally first calculates all the potential occurrences of each read, and then provides the best one in an outfile. If the data are pair-ended reads, the output will take into account the best reciprocal position to map the pair. There are various algorithms, methods, and programs to map short reads. The most used so far (and the most robust) are *Eland* (Illumina software), *BWA* (11), *Bowtie* (12), *MAQ* ((13); less used by now), and *SOAP2* (14). The classical tools, such as *BLAST*, *BLAT*, and *MegaBLAST*, are far too slow to ensure a good analysis. The parameters depend on the heterogeneity and type of organism: Is it homozygous or heterozygous? Is the species really variant in terms of SNP? Generally, one can assume that having a mapping allowing three mismatches in the short sequence (up to 75 bases, 4 for longer reads) will ensure a good mapping in terms of sensibility, even at the detriment of sensitivity (less false positive, but more false negative). The main limits of mappers are that, in the case of TEs, they cannot decipher between the various possibilities for each pair and provide a random location over all the possible ones. However, mapping is highly efficient and, coupled with sufficient posttreatment, can effectively be used to detect and identify SV.

In the case of a transcriptome as reference or as data set, either the same tools mentioned above or more specialized software, like *Exonerate* (15), *CuffLinks* (16), or *TopHat* (17), can be used. Those last will take into account the exonic nature of data to try to predict intron splicing sites in order to improve the mapping.

Mapping is not a highly time-consuming method these days, and it can be performed by the previously cited tools on normal desktop/laptop computers in a decent period of time (less than 1 h for a 10× coverage of a 400 Mbases genome, such as rice).

3.2. Mapping Without a Direct Reference

If no reference from the same species is available, the mapping can be performed however upon a more distant one (but not too far). The same tools can be used, relaxing even more the mapping tolerance (e.g., allowing more SNP/indels or larger gaps in the

mapping). Another tool, *LAST* (18), originally dedicated to perform alignments on large sequences (as genomes), was enhanced to allow the so-called "xeno-mapping", i.e., mapping upon a reference which is not the one from which the data originated. Tested on *Drosophila* genomes (*melanogaster* data upon *simulans* genome), this approach ensures a better mapping. The program in fact increases the mapping using the provided information of theoretical genetic distance between the two organisms.

3.3. Assembly of TEs

Classical assembly softwares can be used, such as *Newbler* (454 software), *Velvet* (19), *Abyss* (20), or *Trans-Abyss* (for transcriptomics data (21)). However, these softwares were not designed to reconstruct repeated regions. It is possible to use part of the program to be able to gather all related reads, and then use homemade scripts to "reassemble" a generic element.

This method is particularly time- and processor-consuming, and cannot be achieved for a classical eukaryotic genome upon a normal computer, but must be run on clusters or grid.

A specific tool for that purpose was developed by DeBarry et al. (22), *AAARF*, and efficiently tested on maize genome.

3.4. Count of TEs

Using NGS data and a set of clean sequences (representative true copies or genome-wide consensus), it is possible to calculate the copy number variation of TEs between individuals or closely related species. For that, a mapping and a coverage calculation upon this reference will theoretically be enough. However, technical biases from DNA extractions, library preparation, and cluster sequencing may provide biased results. Thus, such analysis must be performed using a reiterative experimental procedure to ensure strong and statistically valid results.

References

1. Wicker T, et al. (2008) Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. BMC Genomics 9:518
2. Kurtz S, et al. (2008) A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics 9:517
3. Marçais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27: 764–770
4. Gnerre S, et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci USA 108:1513–1518
5. Argout X, et al. (2010) The genome of Theobroma cacao. Nat Genet 43:101–108
6. Marioni JC, et al. (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18:1509–1517
7. Blumenstiel JP (2011) Evolutionary dynamics of transposable elements in a small RNA world. Trends in Genetics 27:23–31
8. Sabot F, et al. (2011) Transpositional landscape of rice genome revealed by Paired-End Mapping of high-throughput resequencing data. Plant J doi: 10.1111/j.1365-313X.2011.04492.x. [Epub ahead of print]
9. Mirouze M, et al. (2009) Selective epigenetic control of retrotransposition in Arabidopsis. Nature 461:427–430

10. Fiston-Lavier A, et al. (2010) T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. Nucleic Acids Res 39:e36
11. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760
12. Langmead B, et al. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25
13. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18:1851–1858
14. Li R, et al. (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25:1966–1967
15. Slater G, Birney E (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31
16. Trapnell C, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotech 28: 511–515
17. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111
18. Frith MC, Wan R, Horton P (2010) Incorporating sequence quality data into alignment improves DNA read mapping. Nucl. Acids Res 38:e100
19. Zerbino, D.R., Birney, E., 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18, 821–829
20. Simpson JT, et al. (2009) ABySS: A parallel assembler for short read sequence data. Genome Res 19:1117–1123
21. Robertson G, et al. (2010) De novo assembly and analysis of RNA-seq data. Nat Meth 7:909–912
22. DeBarry J, Liu R, Bennetzen J (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