BMC Bioinformatics

Software

Discovery and assembly of repeat family pseudomolecules from sparse genomic sequence data using the Assisted Automated Assembler of Repeat Families (AAARF) algorithm

Jeremy D DeBarry[1], Renyi Liu[1,2] and Jeffrey L Bennetzen*[1]

Address: [1]Department of Genetics, University of Georgia, Athens, GA 30602-7223, USA and [2]Current Address: Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA

Email: Jeremy D DeBarry - jdebarry@uga.edu; Renyi Liu - rliu.biocomp@gmail.com; Jeffrey L Bennetzen* - maize@uga.edu

* Corresponding author

Published: 13 May 2008

BMC Bioinformatics 2008, **9**:235 doi:10.1186/1471-2105-9-235

Received: 10 July 2007
Accepted: 13 May 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/235

Abstract

Background: Higher eukaryotic genomes are typically large, complex and filled with both genes and multiple classes of repetitive DNA. The repetitive DNAs, primarily transposable elements, are a rapidly evolving genome component that can provide the raw material for novel selected functions and also indicate the mechanisms and history of genome evolution in any ancestral lineage. Despite their abundance, universality and significance, studies of genomic repeat content have been largely limited to analyses of the repeats in fully sequenced genomes.

Results: In order to facilitate a broader range of repeat analyses, the Assisted Automated Assembler of Repeat Families algorithm has been developed. This program, written in PERL and with numerous adjustable parameters, identifies sequence overlaps in small shotgun sequence datasets and walks them out to create long pseudomolecules representing the most abundant repeats in any genome. Testing of this program in maize indicated that it found and assembled all of the major repeats in one or more pseudomolecules, including coverage of the major Long Terminal Repeat retrotransposon families. Both Sanger sequence and 454 datasets were appropriate.

Conclusion: These results now indicate that hundreds of higher eukaryotic genomes can be efficiently characterized for the nature, abundance and evolution of their major repetitive DNA components.

1 Background

All higher eukaryotic genomes are rich in multiple classes of repetitive DNA. Transposable elements (TEs) are particularly abundant, and are the most important factor responsible for genome size variation in both animals and plants [1]. Although TEs have been judged to be 'junk DNA', existing within host genomes as purely selfish denizens [2,3], it has been found that some repetitive elements perform important roles in their host genomes [4-7], for instance, as with the telomere-generating Het-A and TART retroelements of Drosophila [8]. In the last decade, the generation of whole genome sequence data from multiple species has provided the opportunity to investigate the relative contributions of repetitive elements to genomic organization and evolution [9-14].

TEs have been identified in nearly all organisms studied to date. They are reported to account for 3% of the 4 Mb yeast genome [6], 20% of the 140 Mb Arabidopsis genome [13], 22% of the 165 Mb Drosophila genome [15], 35% of the 390 Mb rice genome [14], 15% of the 1200 Mb chicken genome [16], > 60% of the 2400 Mb maize genome [17], 46% of the 3200 Mb human genome [18], > 70% of the 4800 Mb barley genome [19] and > 90% of the 16 Gb wheat genome [20].

One type of TE is the Long Terminal Repeat (LTR) retrotransposon (LRP). LRPs account for the great majority of the repetitive DNA in plant genomes [19]. LRPs are named for the LTRs that flank the coding regions of the element. LTRs contain regulatory sequences important for the proper expression of the LRP, such as the transcription start site and polyadenylation signals. Located between the LTRs of the element are the coding regions that provide the protein products necessary for the element's transposition. LRPs typically contain two open reading frames (ORFs). The first of these, gag, contains products necessary for the formation of a virus-like particle where reverse transcription of the RNA intermediate takes place. The second ORF, pol, contains the protease, reverse transcriptase, RNase-H and integrase regions necessary for element protein processing, reverse transcription, degradation of the RNA intermediate and integration into a new genome location [6,19].

Historically, LRPs, other TEs and other types of repetitive DNA have been identified in a genome by their presence in or near genes, or by the amplification of sequences with homology to TEs from other species [11,21]. Due in part to the wealth of sequence information provided by whole genome sequencing projects, the opportunities to detect TEs have greatly expanded in recent years. However, the high cost of completed whole genome sequencing makes this approach inappropriate to investigate the TE content of a large number of species.

As interest in TEs and other repetitive elements has grown, techniques have been developed to discover and investigate them directly. Without the availability of a large amount of assembled genome sequence, studies focused on the identification of TEs within a genome have been restricted to the use of hybridization and PCR techniques [6,22-25]. While these methods are useful for the identification of repeats that are highly homologous to already-discovered repeats, they lack the power necessary to discover or precisely quantitate new classes of repetitive DNA. Sample sequence analysis, wherein a small amount of DNA sequence is generated from randomly selected clones [26-28], can efficiently provide unbiased genomic information, that could potentially be analyzed for repetitive DNA content. The programs RECON [29] and ReAS [30] have been designed for the *de novo* discovery of repeats. RECON utilizes assembled genomic sequence as input, while ReAS was designed for highly redundant genomic coverage with Sanger sequence data sets [30]. ReAS was found to not be adaptable for use with 454 sequence data [31]. Thus, there is currently no method available that is designed to discover and describe genomic repeats using small quantities of unassembled Sanger or 454 sequence data.

In order to provide an automated method for the efficient characterization of all of the high copy number repeats within a genome from sparse sample sequence data, the Assisted Automated Assembler of Repeat Families, or AAARF, algorithm is described in this article. Tests of AAARF on the *Zea mays* genome, using random shotgun sequence data from a Sanger sequencing output and from a simulated 454 sequence data set are presented. The *Z. mays* genome has been well studied in terms of repeat content and provides an excellent opportunity to test AAARF's effectiveness. For both data sets, the program constructed builds representing repeats necessary for genome structure and function (centromeric, ribosomal and knob repeats) and the seven most abundant LRPs in the *Z. mays* genome.

2 Implementation

2.1 The AAARF algorithm

AAARF works by comparing sample sequences from a genome to one another via BLAST [32] and then using a series of BLAST analyses and multiple alignments to "walk out" an *in silico* produced molecule, or "build", that represents a discreet family of repeats from the target organism. A schematic of the AAARF process is shown in Figure 1. Initially, AAARF accepts a fasta file of sample sequences as input. An unused sequence (the first in the input file in Figure 1) in the dataset is BLASTed against all other sample sequences. Next, a coverage matrix representing the detected similarities for the sequence is generated based on this BLAST output. The coverage matrix is a representation of the coverage depth for each nucleotide position in the sequence being considered. The coverage matrix is used to assess the repetitive nature of the sample sequence. The program calculates start and stop points for the sequence that represent the boundaries of a user defined minimum coverage threshold, based on a minimum depth of coverage requirement. This section of the sequence is known as the Minimally Covered Sequence, or MCS. If this sequence doesn't meet the minimum coverage requirement, or the MCS is too short, the sequence is rejected and the process starts again with the next sequence in the dataset. The portion of the sample sequence that corresponds to the MCS is extracted from the sample sequence. Next, the BLAST output is searched to locate sequences that overlap the MCS in the current search direction (right, in Figure 1). Sequences that meet

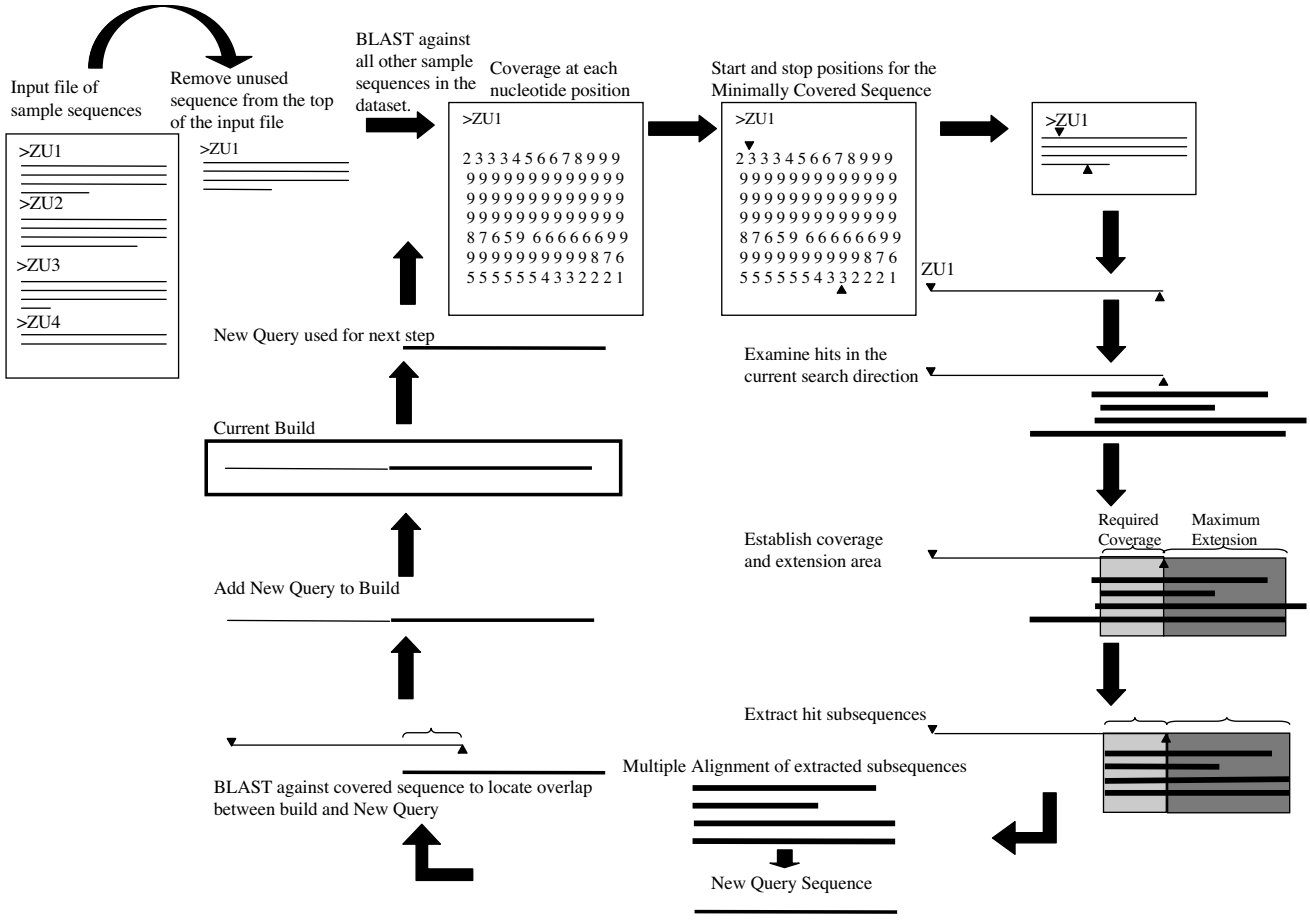

Figure 1
Schematic of the AAARF algorithm. Shown is an illustration of the AAARF process used to assemble builds representing high copy number repeat families from sample sequence data.

the minimum coverage requirements and have the potential to extend the process are selected. A sub-sequence corresponding to the coverage and extension criteria are extracted from the chosen sequences. A multiple alignment of the selected sequences is performed to generate a consensus sequence called the New Query (NQ). The NQ is BLASTed against the MCS to locate the overlap region. Such an overlap indicates that the multiple alignment has produced a sequence that is able to extend the build process correctly. The overlap region is trimmed from the MCS and the NQ is added to the growing build. To complete a single step, the NQ re-enters the loop and is BLASTed against the sample sequence dataset. Directional extension will continue until either there is insufficient evidence in the coverage matrix to indicate a repetitive sequence or the requirement for a minimum number of sequences to extend the process is not met. Extension then begins from the same starting sequence in the opposite direction. Sample sequences are only allowed to participate in a single build. This prevents the seeding of additional builds by sequences already used to construct a build. A sample sequence can participate in both directional extensions of a build in an effort to produce a build that most closely represents a full-length element family. The size of a single round of extension can be smaller than an individual sample sequence. To ensure that an entire sample sequence is used in build construction, the user can adjust the number of times that a sequence can be used to extend the build in a given direction. Using the program output along with the sample sequence dataset under investigation, it is possible to locate the biological ends of the elements represented by the builds.

2.2 Program construction input and output

AAARF consists of a single script written in the PERL programming language (see Additional files 1 and 2). The program makes use of a suite of freely available BioPerl modules [33]. BLAST and Clustalw [34] are used for

sequence comparisons and multiple alignments, respectively. A single file of fasta sequences and a BLAST database made from the same sequences, are used as input to the program. The program produces a fasta file of builds and a diagnostic log file. The log file is produced with the Log4Perl module [35], a customizable PERL logger. AAARF's activities during the build process are recorded in this file for later inspection. AAARF was run on a 2 GHz Macintosh with 1 Gb of RAM. Analysis of the Sanger sequence dataset test was completed in 302 minutes (10,000 sequences), and analysis of the simulated 454 dataset was completed in 1.5 days (50,419 sequences).

2.3 Build issues

There are a number of potential issues that arise when attempting to construct full-length repetitive elements from sample sequence data. It is important that these concerns are understood in order to properly implement the AAARF algorithm and interpret its results. The first of these deals with the internalization and single copy nature of LTRs within a build. LTRs are arranged at either end of a LRP in a direct repeat orientation. Because AAARF utilizes sequence similarity to construct builds, the two LTRs of a LRP will usually be combined into a single LTR composed of sample sequences from either end of multiple elements. Since it is unlikely that both LTRs of a single LRP will be present in a random set of sample sequences, a build's LTR region will likely be a combination of LTRs from multiple related LRPs. Also, because the construction of a build begins at a random point along a repeat (as dictated by the randomly chosen sample sequence that initiates the build), the LTR region will most likely not be present at the end of the build, but internalized within the build.

The next issue to consider is the size of the builds as compared to actual repeats. AAARF builds may be larger than the corresponding type of native repeat. It is possible that a full-length copy of the element may be produced for both search directions for a single build. The number of times that a sequence is allowed to participate in a build are restricted in a given search direction based on the sizes of the sample sequences use as input (Sanger or 454). The count is reset at the start of each directional search, allowing for multiple full-length representations to be present in a single build. This restriction on the number of times that a sequence is used in a search direction ensures that an entire sample sequence is only used once in each build direction. The parameter controlling the number of times that a sequence can be used to extend a build in a single direction is adjustable by the user.

Another way that builds may exceed the size of actual repeats reflects possible size differences within the repeat family. Some families will contain members that are larger than others due to insertions or deletions in certain family members. As long as the affected member is able to replicate, it will be maintained in the genome, and if it is present at a sufficient copy number, be incorporated into a build. Thus, it is possible that a build may represent the largest members of a family, while the smaller members are also contained within the build.

Alternatively, builds may be smaller than native repeats. In general, it is likely that a build will be shorter than a native LRP by the length of one LTR because of LTR internalization. A small build will also be produced in any situation where there are not enough sequences present for a particular type of repeat to be assembled by the program. For instance, LTRs by their nature should be present in numbers at least 2× the amount of sequence for any other region of the element. Since they are present more frequently in the dataset, it is possible that builds representing the LTRs of a family of elements will be produced in cases where construction of the internal regions of the elements is not possible. Also, a build may be broken up due to indels in some family members. Indels can be incorporated into a build as long as there are enough sequences in the dataset to cross the indel-generated build gap. However, if the indel is large enough that it causes the depth of coverage to drop below the requisite threshold, build construction will stop in that direction.

Repeat families may contain levels of sequence diversity that make it impossible for AAARF to assemble all members into a single build. In this case, it is possible that the family will be broken up into multiple builds. For nearly all repeat families that AAARF was able to construct in full-length or near full-length form, there were multiple builds for each family (Table 1). Because of the issues discussed above, it is possible for a build to be fragmented due to low coverage. If this is the case, then there will be no full-length build for a particular family. Rather, the builds for that family will be present in fragmented form. It is also possible that sample sequences from regions that differ between members of the same repeat family may initiate their own build. In this instance there may be a full-length build for a family and additional builds representing regions that were unable to collapse into the full-length build. In some cases it is possible to resolve build fragmentation issues and identify builds that belong to the same family. Shared sequence similarity among AAARF-produced builds can be used to infer relationships between builds that were not combined because of sequence divergence issues or positional effects of sequence used in the build process. For the three fragmented builds in the Sanger and 454 tests, comparison of the builds to one another was able to successfully resolve one of the fragmentation events.

Table 1: Sanger and 454 results compared to the seven most abundant LTR retrotransposon families in maize

Most Abundant LRPs in Maize (a)	Percent Genome Composition	Element Size (kb)	LTR Size (kb)	Sanger: Best Build Size (bp)	Sanger: Number of Builds	454: Best Build Size (bp)	454: Number of Builds
Huck	10.7	11–14	1.6	23706 (FL)	8	11269 (F)	10
Ji	9.4	8.5–10	1.3	10041 (FL)	4	9432 (FL)	3
Opie	7.1	6.5–9	1.3	8150 (FL)	3	9599 (FL)	2
Zeon	4.8	7.3	0.6	7412 (FL)	7	1225 (I)	4
Grande	3.9	10.5–13.5	0.6	8469 (F)	4	6459 (I)	3
Cinful	3.5	8.5	0.6	7264 (F)	5	1600 (I)	4
Xilon **(b)**	3.1	11.7	2.7	8971 (I)	1	2107 (I)	5

(a) Percent genome composition data from Meyers et al. 2001.
(b) *Xilon* percent genome composition from Meyers, pers. comm.
FL = Builds representing full-length copies of repeat families, F = Builds representing fragmented copies of repeat families, I = Incomplete builds

3 Results

3.1 Testing the AAARF algorithm

A wealth of information exists regarding the repeats found in the maize genome (Table 1), particularly the LRP content [17,26,36-39]. The maize genome is approximately 2400 Mb in size [17], and > 60% of the genome is composed of repetitive DNA [40], primarily LRPs [6]. In order to ascertain AAARF's effectiveness, a database of known maize genomic repeats was assembled. This database is a combination of TIGR's maize repeat database [41] and a maize repeat database developed by P. San Miguel at Purdue University (P. San Miguel, pers. comm.). Each of these databases contains the sequences of many different repeat families and individual family members found in maize. Builds produced by AAARF were BLASTed against this known repeat database to investigate how accurately the builds represent actual genomic repeats and to examine how well the program constructs builds representing distinct families of repeats.

To classify a build as representing a type of known repeat, several criteria were used. A build was required to be at least 1 kb in size, and to have at least one hit with a minimum BLAST score of 100 when compared to the known repeat database. A score of less than 100 was taken as evidence that no useful sequence similarity existed between the build and the known repeat database. Builds were also inspected to ensure that they did not improperly fuse two repeat families. Finally, each build was examined to ensure that it showed similarity to a single family of known repeats over at least 90% of its length. In this regard, all builds generated by AAARF from the maize data analyzed (below) were found to be homologous to an already-known maize repeat family, indicating both the quality and comprehensiveness of the TIGR and San Miguel databases.

The ultimate goal of the AAARF algorithm is to construct the best build possible for a given family of repeats. How well AAARF is able to accomplish this is dependent on the amount of sequence in the sample data set for a given repeat family. There are many issues regarding the build process that were considered in the examination of the builds (discussed in Implementation). A build was classified as full-length only if it showed similarity to the entire length of multiple members of a single discreet repeat family.

3.2 AAARF analysis of Sanger sample sequence data from maize

Random unfiltered shotgun sequence reads produced by Sanger sequencing for maize are available from TIGR [42]. Sequences were obtained from TIGR in December of 2005. We selected the first 10,000 available sequences from this database for input into AAARF. This sample sequence dataset totaled 7,821,671 bp (average read size 782 bp), representing 0.33% of the maize genome (Table 2). Input sequences were screened for vector content using NCBI's UNIVEC database[43]. AAARF produced 180 builds from the Sanger sample sequence dataset described above (Table 2) and the parameter set described in Table 3. Of these, 57 were chosen for further analysis (Table 2). As expected, AAARF assembled builds for non-TE repeats, including centromeric repeats, ribosomal repeats and knob repeats. In addition, builds representing all 7 of the most abundant LRP families in the maize genome were constructed (Table 1). Full-length builds were constructed for the four most abundant families. Builds representing the *Grande* and *Cinful* families were fragmented, such that there was a region missing from each build. For both *Grande* and *Cinful*, the missing region was assembled intact in additional builds for each family. For the *Xilon* build, a 700 bp portion of the LTR region found in native *Xilon* elements was missing. Thus, AAARF constructed full-length or near full-length builds representing the seven most abundant LRP families in the maize genome. In all cases, AAARF was able to assemble a build for each of these families that readily identified the repeat family. For

Table 2: Overall results of AAARF tests of Sanger and simulated 454 data sets

	Number of Sample Sequences	Sequence Amount (bp) (%Genome)	Total Builds	Builds < 1 kb (a)	Fused Builds (a)	Build Coverage < 90% (a)	No Hits/ Score Too Low (a)	Analyzed Builds
Sanger Sequence Build Results	10000	7821671 (0.33%)	180	46	5	49	23	57
454 Sequence Build Results	50419	5045000 (0.21%)	63	2	2	12	1	46

(a) Builds not further analyzed

the *Grande* family (the build with the largest missing fragment) the largest build covered 80% of the expected family size.

3.3 454 sequence analysis

454 sequence analysis [44] is an emerging high throughput technology that greatly lowers the cost of data generation. This will facilitate the generation of sample sequence datasets for a wide array of species. The initial 454 sequences had an average length of ~100 bp [44]. In order to test the ability of AAARF to utilize this type of data, a simulated dataset of 454 sequences for the maize genome was generated (see below). Sequences were screened for vector content using NCBI's UNIVEC database [43]. A total of 50,419 sequences representing 5,045,000 bp (average read size 100 bp) were used as input. This dataset represents ~0.21% of the maize genome (Table 2).

The same database of known repeats and the same build classification criteria used for testing the output of the Sanger sequence test were used for analyzing the 454 sequences. Smaller input sequences cause a variety of build issues stemming from the required BLAST parameter settings used by AAARF. Because of the reduced size of the input sequences compared to the previous dataset, the number of builds with a total size of less than 1 kb was greatly increased. Since AAARF only allows each sample sequence to participate in a single build, builds of less than 1 kb in length were rejected. This ensured that as many sequences as possible were available to the program for each build, instead of being utilized in smaller, ultimately uninformative builds.

Despite the inherent difficulties posed by shorter sequences, and the overall reduction in sample sequence dataset size, AAARF generated 46 builds that were identified as belonging to known repeat families (Table 2). All *Huck* family builds were fragmented in the output, with an approximately 1.4 kb fragment missing from the largest build. This fragment was present intact in an additional *Huck* build. *Ji* and *Opie* were constructed in full-length form. For the remaining 4 families in Table 1, only *Grande* was constructed in a large build. It is possible that the inability of the program to construct the other four elements in a full-length size is due to the 454 sample sequence data being only ~64% the size of the Sanger

Table 3: Parameter settings for Sanger and 454 tests

	Minimum Hit Length	Minimum Hit Identity	Maximum e-value	Required Length of MCS	Required MCS Coverage Depth	Minimum Number of Hits for Extension
Sanger	150	89	1.00E-25	150	3	2
454	30	88	1.00E-10	30	3	2

	Required Coverage Length	Maximum Extend Length	Minimum BL2SEQ Hit Size	Maximum BL2SEQ e-value	Maximum Number of Times a Sequence Can Be Used in One Direction	Other
Sanger	150	50	90	1.00E-10	13	
454	30	40	15	10	4	BL2SEQ Word Size: 7

sequence dataset. It is also possible that further parameter optimization for 454 data will yield superior results. As 454 sequencing technology continues to develop, the average size of the reads is increasing. Such an increase in sequence size will facilitate their use as input sequences for the AAARF approach.

Apart from sequence size issues, 454 technology brings with it a new set of issues with regard to reliability statistics and error rates [45]. In particular homopolymers pose specific problems due to the nature of the pyrosequencing technology that 454 sequencing employs [46]. Precise parameter adjustments to account for these issues can be made with the use of actual 454 data.

3.4 454 dataset construction

In order to simulate a 454 sequence dataset for maize, all available unfiltered shotgun sequences for the maize genome were downloaded from TIGR [42]. At the time of this analysis, there were 50,877 shotgun sequences in this dataset. The data were divided into three subsets of 16,959 sequences each. To simulate an average read size of 100 bp, a subsequence of each read was extracted. Positions 100–150, 100–200 and 100–250 were extracted from all sequences in each set respectively. The extraction was initiated from the 100 bp position for each read to avoid any possible sequencing errors at the end of the read. Only one sequence was extracted from each shotgun read to provide a random sampling. This produced three datasets composed of 50, 100 and 150 bp sequences. For each of the three subsets, there were sequences that were not large enough for the extraction process, resulting in a final count of 50,419 sequences.

3.5 Parameters

AAARF parameters for the Sanger sequencing and 454 datasets were determined by trial and error in order to examine how changes in the program parameters affected program output. Adjustable parameters for both tests are found in Table 3. Parameters affecting the required length of BLAST hits, coverage, extension length and maximum number of times that a sequence is allowed to participate in a search direction were chosen based on the sizes of the sample sequences used for each test. Identity and e-value requirements for both tests were determined by trial and error.

Required depth of coverage for a sequence to be classified as repetitive, and the required minimum number of sequences for extension were chosen based on an interpretation of what was necessary to recognize a repetitive sequence. The presence of a particular sequence in the sample sequence dataset at least 3 times was seen as evidence of its repetitive nature. Because of slight positional variation of the coordinates of sequences participating in the AAARF process, it is possible that a sequence that belongs in a build may be rejected due to a difference with required positional parameters generated during MCS construction. In order to account for this phenomenon, only 2 sequences were required for extension. This did not affect the accuracy of the builds when compared to a test requiring a 3 sequence minimum for extension.

For the 454 test, the word-size BLAST parameter for the BL2SEQ was lowered to 7 from the standard 11. During testing it became apparent that the small size of the 454 sequences presented problems with the detection of overlap between the New Query Sequence and the MCS (Figure 1). Reduction in the required word size facilitated overlap detection.

3.6 End finding

As the name of the program indicates, there is a hands-on component to the AAARF process. The program assembles builds representing discreet families of genomic repeats while maintaining the correct order and orientation of the elements. However, the element components are unlikely to be placed in the same end-to-end fashion as a typical element. This is due to the random starting point for a build. To alleviate this issue, a method for the identification of the biological element endpoints of LRPs for AAARF-produced builds has been developed.

For LRPs, the ultimate goal is to locate the LTR region of the build as this region contains both element ends. Initially, BLASTx is used to locate possible protein coding regions within the build, to narrow the area of the build where the LTR may be found. Next, the build is used in a BLAST analysis against the sample sequence data set that was used as input for the AAARF program. This BLAST provides coverage information for the build. In addition to facilitating the location of the biological endpoints of the build, this information can be compared to the AAARF-generated diagnostic log file to ensure that all suitable sequences were used to construct a given build. If sequences are found in this comparison that were not used to construct a build, program parameters can be altered to incorporate these sequences. The Apollo program [47] was used to visualize this comparison (Figure 2). Using this information, it is possible to examine the build for regions that are represented at a greater depth of coverage in the sample sequence dataset (Figure 2). The region of the build that contains the LTR should be covered by sample sequences at least twice as deeply as the rest of the build. This is due to the presence of LTRs at either end of a native full-length element and the solo-LTRs present in the genome as a result of partial element removal by unequal recombination. Around this region of increased coverage, the individual sample sequences that cover the region are inspected to locate sequences that are

Figure 2: Comparison of an AAARF-Produced Build to the Sample Sequence Dataset

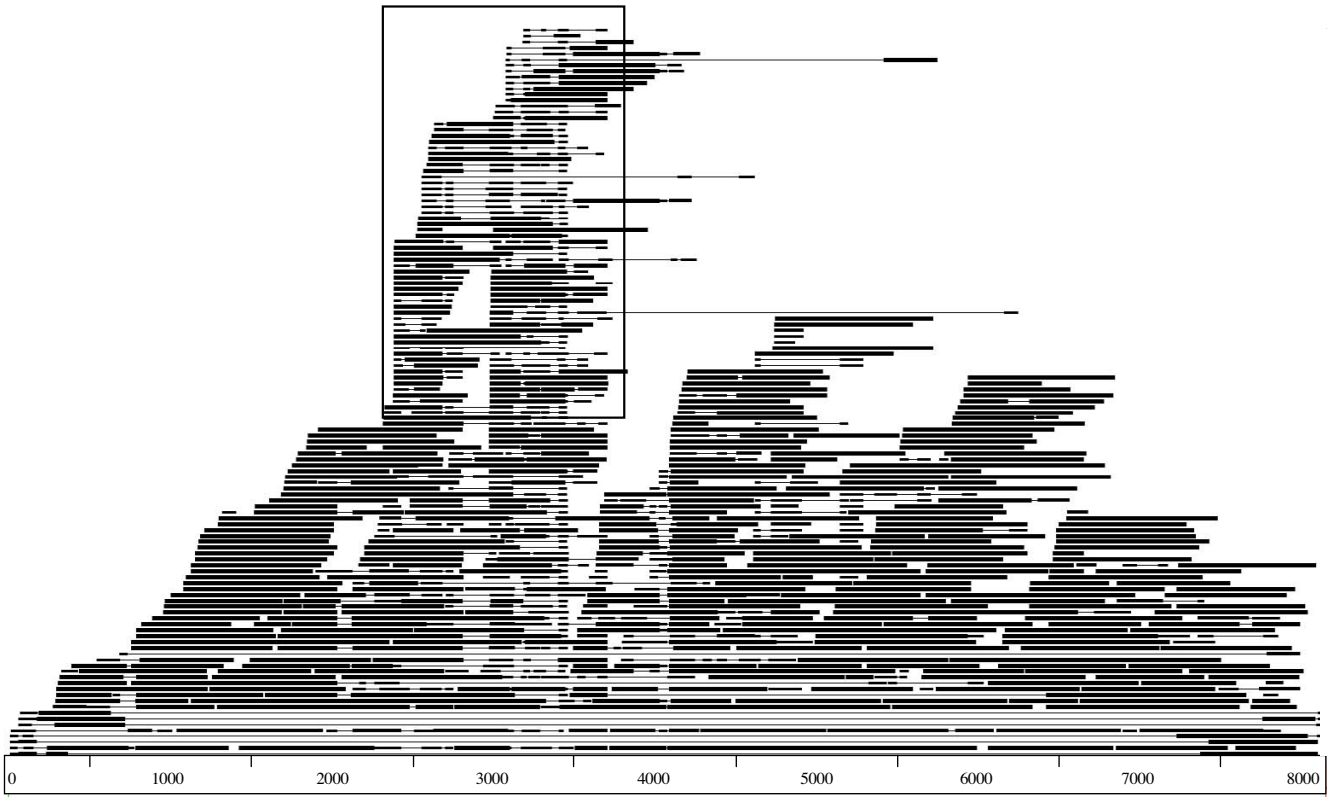

Figure 2

Comparison of an AAARF-produced build to the sample sequence dataset. Shown is the BLAST result of an AAARF-generated build compared to the sample sequence dataset used to create it. The bottom, metered line represents the full-length *Opie* build (8,150 bp) from the Sanger sequence test. Smaller lines above represent sample sequences. Regions of shared similarity between the build and the sample sequences are indicated by the position of the sample sequences relative to the build. A region of increased coverage on the build, combined with sample sequences whose similarity to the build stops at the same position (boxed area), indicates the likely presence of an LTR.

truncated at approximately the same position on the build (Figure 2). Because the LTRs are present on either side of a native full-length element, sample sequences that include the LTR boundaries may include either sequence from the interior of the element or sequence from the genome surrounding the element.

Once these possible boundaries are identified, this region is extracted from the build for further inspection. There are four possible orientations for the LTR region, depending on the orientation and strandedness of the sequence that initiated the build. The extracted section is then inspected for conserved LTR terminal dinucleotide motifs and the Primer Binding Site and Polypurine Tract for the element family that has been constructed. These structures are localized to the LTR region of native LRPs and can be used to indicate the presence of an LTR within the build. For a build representing a full-length element, this information can then be used to manually reconstruct a build with LTRs at either end of the build. For builds that are less than full-length, this approach will be useful in identifying the LTR region if it is contained in the build. As a proof of principle, this method was used to identify biological element endpoints for the full-length *Opie* build in the Sanger sequence test. The location of the LTR within the build was verified using actual elements from the known repeat database.

4 Discussion

AAARF provides an excellent resource for the initial characterization of high copy number repeats in a genome that has been subjected to very limited shotgun sequence anal-

ysis. Because of the nature of the AAARF process, the pseudomolecules it produces are "patchwork" representations of native repeat elements. A multiple alignment of selected overlapping and extending sequences is used for each directional extension. Thus, each step represents sections from actual repeats found in the target genome. Sequence divergence information for actual elements is available via the sample sequences used to construct the pseudomolecule. By comparing these pseudomolecules to the input sample sequence dataset, information about the evolutionary history of the assembled repeat family and the percent of the target genome composed of that repeat can be determined.

While these tests have focused on AAARF's ability to construct builds representing LRP sequences, the utility of the program extends to the construction of builds representing any high copy number repeat in a genome. As long as there is sufficient sequence in the sample sequence dataset to represent the repeat, AAARF will construct a build for it.

The parameters used here were developed for use in the maize genome. Depending on the type of sample sequence being used or the species being investigated, it will be necessary to alter the program parameters to produce the most accurate builds possible. This is the primary reason that the diagnostic test log is produced as a part of the AAARF process. Using this log file, it will be possible to optimize the parameter set for any species under investigation.

5 Conclusion

As understanding of the prevalence and effects of LRPs has increased, it has become apparent that understanding the evolutionary dynamics of LRPs both within individual genomes and among different species is necessary for a complete understanding of genome structure and history. The true utility of the AAARF approach is its ability to facilitate such an understanding. Because AAARF is designed to function on sample sequence data, important information about the TE content of a genome can be investigated with a small amount of sequence, making this type of analysis feasible for studies that involve hundreds of species.

6 Availability and requirements

• **Project Name:** Assisted Automated Assembler of Repeat Families

• **Project Home Page:** https://sourceforge.net/projects/aaarf

• **Operating System:** Mac OS X

• **Programming Language:** Perl

• **Other Requirements:** Bioperl 1.2.3 or higher, Bioperl Run Package 1.4 or higher, Log4perl 1.01 or higher, NCBI BLAST 2.2.9 or higher, Clustalw 1.8.3 or higher

• **License:** GNU Lesser General Public License

• **Restrictions:** none

7 Authors' contributions

JDD co-designed the algorithm, conducted the testing and implementation of the algorithm and drafted the manuscript. RL co-designed the algorithm and reviewed the manuscript. JLB conceived the approach, participated in its design, provided valuable guidance and critically edited the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

AAARF Algorithm Perl Script. This is the AAARF algorithm. The file can be viewed in a text editor. Examples are BBedit for Macintosh and Wordpad for Windows. This file is also available at the project homepage https://sourceforge.net/projects/aaarf.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-9-235-S1.pl]

Additional file 2

AAARF Program Documentation. This is the AAARF algorithm documentation. The file is intended to aid in the use of the program and provides instructions for its use. The file can be viewed in a text editor. Examples are BBedit for Macintosh and Wordpad for Windows. This file is also available at the project homepage https://sourceforge.net/projects/aaarf.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-9-235-S2.txt]

References

1. Kidwell MG: **Transposable elements and the evolution of genome size in eukaryotes.** *Genetica* 2002, **115:**49-63.
2. Orgel LE, Crick FH, Sapienza C: **Selfish DNA.** *Nature* 1980, **288:**645-646.
3. Doolittle WF, Sapienza C: **Selfish genes, the phenotype paradigm and genome evolution.** *Nature* 1980, **284:**601-603.
4. Makalowski W: **Genomics. Not junk after all.** *Science* 2003, **300:**1246-1247.
5. Kidwell MG, Lisch DR: **Perspective: transposable elements, parasitic DNA, and genome evolution.** *Evolution Int J Org Evolution* 2001, **55:**1-24.
6. Kumar A, Bennetzen JL: **Plant retrotransposons.** *Annu Rev Genet* 1999, **33:**479-532.
7. Bennetzen JL: **Transposable element contributions to plant gene and genome evolution.** *Plant Mol Biol* 2000, **42:**251-269.
8. Pardue ML, Danilevskaya ON, Lowenhaupt K, Slot F, Traverse KL: **Drosophila telomeres: new views on chromosome evolution.** *Trends Genet* 1996, **12:**48-52.
9. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J,

Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.

10. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420:**520-562.

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

12. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, Coulson A, D'Eustachio P, Fitch DH, Fulton LA, Fulton RE, Griffiths-Jones S, Harris TW, Hillier LW, Kamath R, Kuwabara PE, Mardis ER, Marra MA, Miner TL, Minx P, Mullikin JC, Plumb RW, Rogers J, Schein JE, Sohrmann M, Spieth J, Stajich JE, Wei C, Willey D, Wilson RK, Durbin R, Waterston RH: **The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics.** *PLoS Biol* 2003, **1:**E45.

13. Kellogg EA, Bennetzen JL: **The evolution of nuclear genome structure in seed plants.** *Am J Bot Am J Bot* 2004, **91:**1709-1725.

14. Vij S, Gupta V, Kumar D, Vydianathan R, Raghuvanshi S, Khurana P, Khurana JP, Tyagi AK: **Decoding the rice genome.** *Bioessays* 2006, **28:**421-432.

15. Kapitonov VV, Jurka J: **Molecular paleontology of transposable elements in the Drosophila melanogaster genome.** *Proc Natl Acad Sci U S A* 2003, **100:**6569-6574.

16. Wicker T, Robertson JS, Schulze SR, Feltus FA, Magrini V, Morrison JA, Mardis ER, Wilson RK, Peterson DG, Paterson AH, Ivarie R: **The repetitive landscape of the chicken genome.** *Genome Res* 2005, **15:**126-136.

17. SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Bennetzen JL: **Nested retrotransposons in the intergenic regions of the maize genome.** *Science* 1996, **274:**765-768.

18. Wong LH, Choo KH: **Evolutionary dynamics of transposable elements at the centromere.** *Trends Genet* 2004, **20:**611-616.

19. Feschotte C, Jiang N, Wessler SR: **Plant transposable elements: where genetics meets genomics.** *Nat Rev Genet* 2002, **3:**329-341.

20. Flavell RB: **Repetitive DNA and chromosome evolution in plants.** *Philos Trans R Soc Lond B Biol Sci* 1986, **312:**227-242.

21. Turcotte K, Srinivasan S, Bureau T: **Survey of transposable elements from rice genomic sequences.** *Plant J* 2001, **25:**169-179.

22. Goldberg ML, Sheen JY, Gehring WJ, Green MM: **Unequal Crossing-over Associated with Asymmetrical Synapsis between Nomadic Elements in the Drosophila-Melanogaster Genome.** *Proc Natl Acad Sci U S A* 1983, **80:**5017-5021.

23. Miller K, Lynch C, Martin J, Herniou E, Tristem M: **Identification of multiple Gypsy LTR-retrotransposon lineages in vertebrate genomes.** *J Mol Evol* 1999, **49:**358-366.

24. Judelson HS: **Sequence variation and genomic amplification of a family of Gypsy-like elements in the oomycete genus Phytophthora.** *Mol Biol Evol* 2002, **19:**1313-1322.
25. Voytas DF, Cummings MP, Koniczny A, Ausubel FM, Rodermel SR: **copia-like retrotransposons are ubiquitous among plants.** *Proc Natl Acad Sci U S A* 1992, **89:**7124-7128.
26. Meyers BC, Tingey SV, Morgante M: **Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome.** *Genome Res* 2001, **11:**1660-1676.
27. Elgar G, Clark MS, Meek S, Smith S, Warner S, Edwards YJ, Bouchireb N, Cottage A, Yeo GS, Umrania Y, Williams G, Brenner S: **Generation and analysis of 25 Mb of genomic DNA from the pufferfish Fugu rubripes by sequence scanning.** *Genome Res* 1999, **9:**960-971.
28. Li WL, Zhang P, Fellers JP, Friebe B, Gill BS: **Sequence composition, organization, and evolution of the core Triticeae genome.** *Plant Journal* 2004, **40:**500-511.
29. Bao Z, Eddy SR: **Automated de novo identification of repeat sequence families in sequenced genomes.** *Genome Res* 2002, **12:**1269-1276.
30. Li RQ, Ye J, Li SG, Wang J, Han YJ, Ye C, Wang J, Yang HM, Yu J, Wong GKS, Wang J: **ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun.** *Plos Computational Biology* 2005, **1:**313-321.
31. Macas J, Neumann P, Navratilova A: **Repetitive DNA in the pea (Pisum sativum L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and Medicago truncatula.** *BMC Genomics* 2007, **8:**427.
32. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
33. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E: **The bioperl toolkit: Perl modules for the life sciences.** *Genome Research* 2002, **12:**1611-1618.
34. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680.
35. **The Log4Perl Project** [http://log4perl.sourceforge.net/]
36. Bennetzen JL: **The contributions of retroelements to plant genome organization, function and evolution.** *Trends Microbiol* 1996, **4:**347-353.
37. Sanz-Alferez S, SanMiguel P, Jin YK, Springer PS, Bennetzen JL: **Structure and evolution of the Cinful retrotransposon family of maize.** *Genome* 2003, **46:**745-752.
38. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL: **The paleontology of intergene retrotransposons of maize.** *Nat Genet* 1998, **20:**43-45.
39. Sanmiguel P, Bennetzen JL: **Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons.** *Annals of Botany* 1998, **82:**37-44.
40. Bennetzen JL, SanMiguel P, Chen M, Tikhonov A, Francki M, Avramova Z: **Grass genomes.** *Proc Natl Acad Sci U S A* 1998, **95:**1975-1978.
41. Ouyang S, Buell CR: **The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants.** *Nucleic Acids Res* 2004, **32:**D360-3.
42. **The TIGR Maize Database** [http://maize.tigr.org/]
43. **Univec Database** [ftp://ftp.ncbi.nih.gov/pub/UniVec]
44. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen ZT, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu PG, Begley RF, Rothberg JM: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437:**376-380.
45. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM: **Accuracy and quality of massively parallel DNA pyrosequencing.** *Genome Biol* 2007, **8(7):**R143.
46. Goldberg SMD, Johnson J, Busam D, Feldblyum T, Ferriera S, Friedman R, Halpern A, Khouri H, Kravitz SA, Lauro FM, Li K, Rogers YH, Strausberg R, Sutton G, Tallon L, Thomas T, Venter E, Frazier M, Venter JC: **A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes (vol 103, pg 11240, 2006).** *P Natl Acad Sci USA P Natl Acad Sci USA* 2006, **103:**16057-16057.
47. Lewis SE, Searle SM, Harris N, Gibson M, Lyer V, Richter J, Wiel C, Bayraktaroglir L, Birney E, Crosby MA, Kaminker JS, Matthews BB, Prochnik SE, Smithy CD, Tupy JL, Rubin GM, Misra S, Mungall CJ, Clamp ME: **Apollo: a sequence annotation editor.** *Genome Biol* 2002, **3:**RESEARCH0082.