BMC Genomics

BioMed Central

Database

RepPop: a database for repetitive elements in *Populus trichocarpa*

Fengfeng Zhou[1,2] and Ying Xu*[1,2]

Address: [1]Computational Systems Biology Laboratory, Department of Biochemistry and Molecular Biology, and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA and [2]BioEnergy Science Center, Tennessee, USA

Email: Fengfeng Zhou - ffzhou@csbl.bmb.uga.edu; Ying Xu* - xyn@bmb.uga.edu

* Corresponding author

Published: 9 January 2009

BMC Genomics 2009, **10**:14 doi:10.1186/1471-2164-10-14

Received: 4 June 2008

Accepted: 9 January 2009

This article is available from: http://www.biomedcentral.com/1471-2164/10/14

Abstract

Background: *Populus trichocarpa* is the first tree genome to be completed, and its whole genome is currently being assembled. No functional annotation about the repetitive elements in the *Populus trichocarpa* genome is currently available.

Results: We predicted 9,623 repetitive elements in the *Populus trichocarpa* genome, and assigned functions to 3,075 of them (31.95%). The 9,623 repetitive elements cover ~40% of the current (partially) assembled genome. Among the 9,623 repetitive elements, 668 have copies only in the contigs that have not been assigned to one of the 19 chromosome while the rest all have copies in the partially assembled chromosomes.

Conclusion: All the predicted data are organized into an easy-to-use web-browsable database, *RepPop*. Various search capabilities are provided against the *RepPop* database. A Wiki system has been set up to facilitate functional annotation and curation of the repetitive elements by a community rather than just the database developer. The database *RepPop* will facilitate the assembling and functional characterization of the *Populus trichocarpa* genome.

Background

The *Poplar* was selected to be the first tree genome to be sequenced, mainly because of its extraordinarily rapid growth rate and its relatively compact genome size (450–500 Mbps [1,2]). Biofuels are produced mainly through two sources, i.e. crops high in sugar or cellulose, e.g. sugar canes [3] and plants [4], and plants high in vegetable oils like soybean[5]. The *Populus trichocarpa* genome's rapid growth coupled with the high content of lignocelluloses has made it one of the model systems for the new generation of biofuels [4]. The current assembly of the *Poplar* genome was released in June 2004, and its total length is ~485 Mbps. The assembled 19 chromosomes with 7.66% gaps count for 63.41% of the whole genome. Further efforts are still needed to close the gaps in the sequenced chromosomes.

Repetitive elements represent a significant fraction of eukaryotic genomes and they could occupy as high as 80% of some land-plant genomes like wheat [6] and as low as 10–35% for *Arabidopsis thaliana* [7] and rice [8]. There are three main classes of repetitive elements, namely, local repeats (tandem and satellite repeats) [9], interspersed repeats (transposons) and segmental duplications (duplicated genomic segments). Among them, transposable elements are the most extensively studied repetitive elements, and they can be classified as retrotransposons or DNA transposons based on whether they

are transposed through the RNA or DNA intermediates [10]. Both interspersed repeats [11-16] and other duplicated elements [17] may induce homologous recombinations and insertions/deletions in the host genome, which may introduce great difficulties to the correct assembly of the repetitive regions in the host genome.

Typically repetitive elements have been identified in a genome using two approaches: (1) identification of homologous sequences to known repetitive elements [18], and (2) identification of repeats based on self-comparison a given genome and clustering them into families [19-21]. The first approach requires manually curated repetitive elements, which may not be feasible for newly sequenced genomes, though it can identify the precise boundaries of repetitive elements, even for the embedded partial copies. The second approach identifies repetitive elements in a *de novo* fashion, though it may require additional manual curations for the boundaries of the predicted elements.

Construction and content

Data resources

The current assembly of the *Populus trichocarpa* genome was released in June 2004 as version 1.1, which consists of 22,012 nucleotide sequences, covering large pieces of the 19 chromosomes and some unassembled short contigs, and the total length is 485,510,911 bps. This data was downloaded from the web site of *Populus trichocarpa* genome sequencing project [22].

We downloaded four of the most comprehensive databases of repetitive elements in eukaryotes, RepBase [23] version 12.05 (release of July 13, 2007), TREP [24] version 10 (release of July 2008), RetrOryza [25] and AtRepBase [26], for homology search. We also downloaded the databases RDP [27] and Rfam [28], and RNA genes in the rice RAP-DB database [29]. The NCBI database NT [30] containing all the non-redundant protein sequences was also downloaded for homology search.

Identification of repetitive elements

Due to the very large computer memory requirement by many repeat identification programs [19-21], we implemented our *RepPop* database and associated tools on a 64-bit Linux operating system with 32 GB memory. The repetitive elements with at least 2 copies in the *Poplar* genome were identified using *RepeatScout* [19]. We then removed any repetitive elements predicted to be low complexity regions using program *NSEG* [31] and tandem repeats using program *TRF* [32]. All the programs were run using the default parameters.

Totally 9,623 repetitive elements were identified, covering 194.00 Mbps (~40%) of the *Poplar* genome. The distributions of copy numbers and lengths of these repetitive elements are given in Figure 1. Most of the repetitive elements are short and of low copy numbers.

Annotation procedure

We first identified the homologous regions of the 9,623 repetitive elements in the databases *RepBase* [23], TREP [24], RetrOryza [25] and AtRepBase [26] using the NCBI Blast [30] with *E-value* cutoff e-5. One region might match two homologous elements in the database. We then removed the redundant annotations by keeping only the region with the lowest *E-value* for the overlapping regions. A total of 226 homologous regions were identified.

We then predicted 30 tRNA genes using the program *tRNAscan-SE* with default parameters [33]. 8 and 40 homologous regions to the RNA genes in databases *RDP* [27] and *RAP-DB* [29] were identified using the NCBI Blast [30] with *E-value* cutoff e-5 after removing the redundancy like above. No homologous regions were identified based on the RNA profiles of *Rfam* [28] using the program *infernal* [19] with default parameters.

2,720 homologous regions to sequences in the database NT [30] were identified using NCBI Blast [30] with *E-value* cutoff e-5, and annotated as having the functions of the best matched homologous proteins.

Utility and discussion

We organized the 9,623 predicted repetitive elements and their annotations into an easy-to-use web-browsable database system, *RepPop* [34] (Figure 2). *RepPop* is and will continue to be under continuing update on quarterly basis for annotation and curation of these repetitive elements. The composition of the *RepPop* database can be found in Table 1. We provided a Wiki interface for the whole community to help curate the annotations [35].

Data browsing

A user may browse all the 9,623 repetitive elements in the browsing interface of *RepPop*, as shown in Figure 3. The detailed annotation of each repetitive element can be retrieved using a popup window by clicking the corresponding entry under REName. Some repetitive elements have as high as 8,455 copies (RepPop694) in the *Populus trichocarpa* genome. We believe that it is not necessary to list the information of all the copies for such repetitive elements. So the browsing interface lists the information of at most 5 copies for each repetitive element as the default. The user can get the additional information, if needed, of all the copies by clicking the button "Get all". The user could also choose to browse only one of the following types of repetitive elements, namely, transposons, RNA genes, protein coding genes, and repetitive elements with

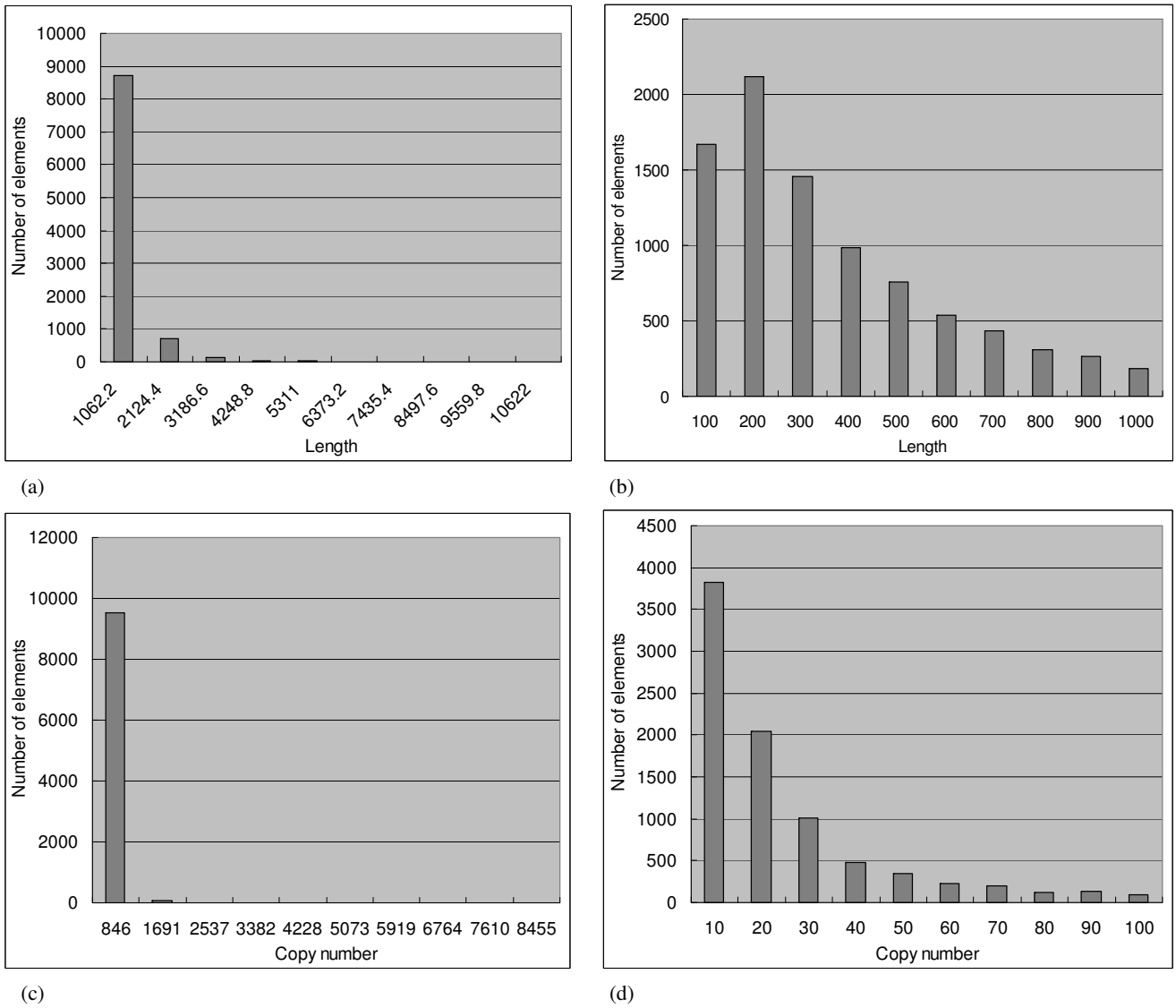

Figure 1
Basic information of the identified 9,623 repetitive elements. Distributions of (a) lengths and (c) copy numbers of repetitive elements in the *Populus trichocarpa* genome. The two distributions within a shorter range are in (b) and (d).

Table 1: Basic knowledge of the *RepPop* database

Annotations	Number	Number%	Length (bps)	Length%
Transposons	161	1.67	21,044,639	4.33
RNA genes	15	0.16	36,051	0.01
Protein-coding genes	2,983	31.00	157,586,923	32.46

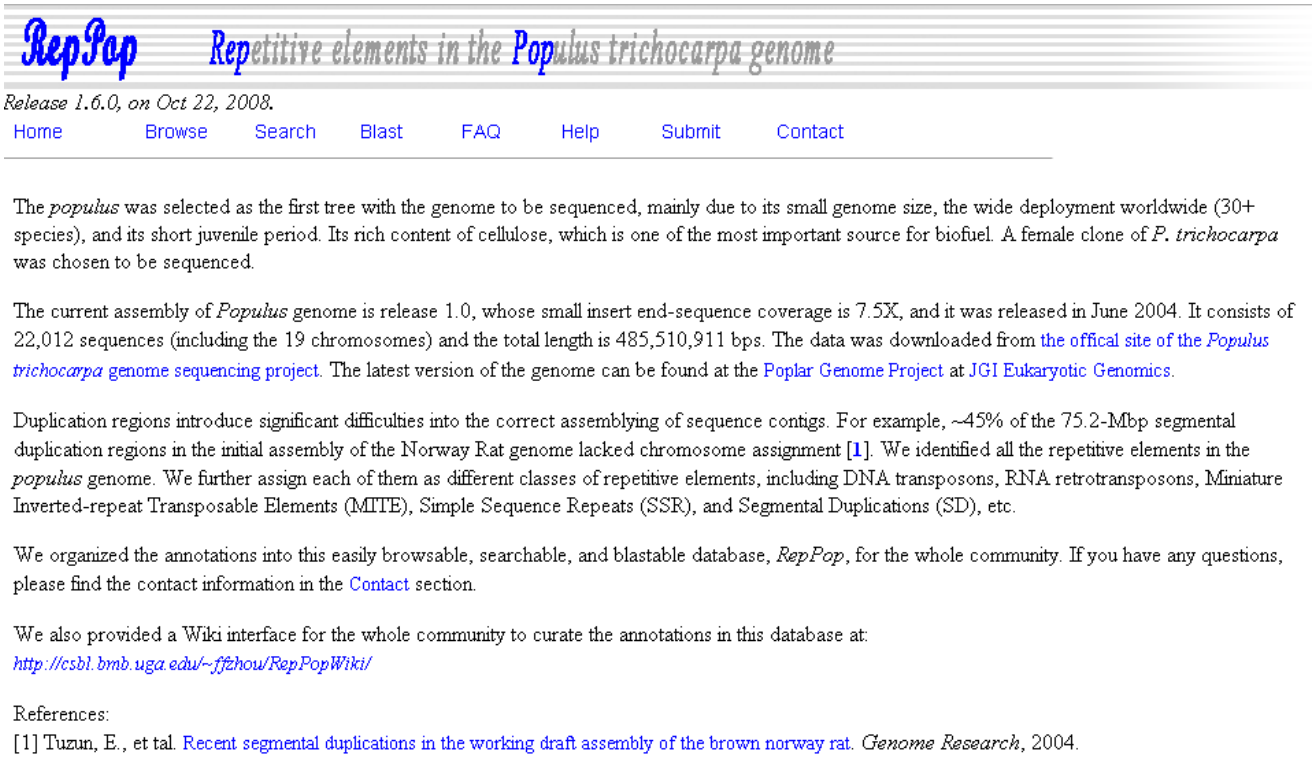

Figure 2
The main web page of database *RepPop*.

no annotations, by clicking the corresponding entry in the drop-down menu.

Data search with key words

The keyword search interface of *RepPop* makes it possible for a user to find items interesting to the user using a few keywords, as shown in Figure 4. Besides the typical keyword matching in the annotations, *RepPop* also provides the flexibility to support search for items within a specified range in terms of, say, elements with certain lengths or with certain copy numbers. For example, a pattern like "Length:Min-Max" as part of a keyword search can be used to find repetitive elements whose lengths are between the specified parameters Min and Max, whose default values are 0 and 10,622 if such range information is not specified. 10,622 is the maximum length of repetitive elements. Another example is including "Length>100" as part of a keyword search to find repetitive elements with at least 100 bps; and similarly "Length<500" is used to find elements with at most 500 bps. Similar range specifications are also available for specifying other quantities, say, "CopyNum:Min-Max" used for finding repetitive elements with copy numbers between specified parameters Min and Max. Tips for keyword searches are available in the Help box of *RepPop*.

Sequence homology search

An interface is provided to facilitate Blast search, using the NCBI Blast, against the 9,623 repetitive elements. Through this interface (Figure 5), a user can simultaneously specify up to 10 nucleotide or protein sequences of no more than 10 kbps for homology search. A query example is provided for using this interface and can be found and used by clicking the button "Example (RepPop25)".

User input to RepPop

We were able to assign functions for 3,075 of the 9,623 repetitive elements, based on homology search against the NT database, leaving 6,548 (~68.05%) repetitive elements functionally unassigned. We have designed the *RepPop* interface in such a way that a user can submit his/her own functional annotations of any repetitive elements in *RepPop* through the Submission interface, as shown in Figure 6. We have provided a Wiki web site for the general user community to directly annotate and curate the assigned functions of repetitive elements and to keep track of updates of annotations of each repetitive element, using PmWiki [36]. The developer of the *RepPop* database has the ultimate right to keep, revise or delete a particular contribution made by a user, which will be done on regular

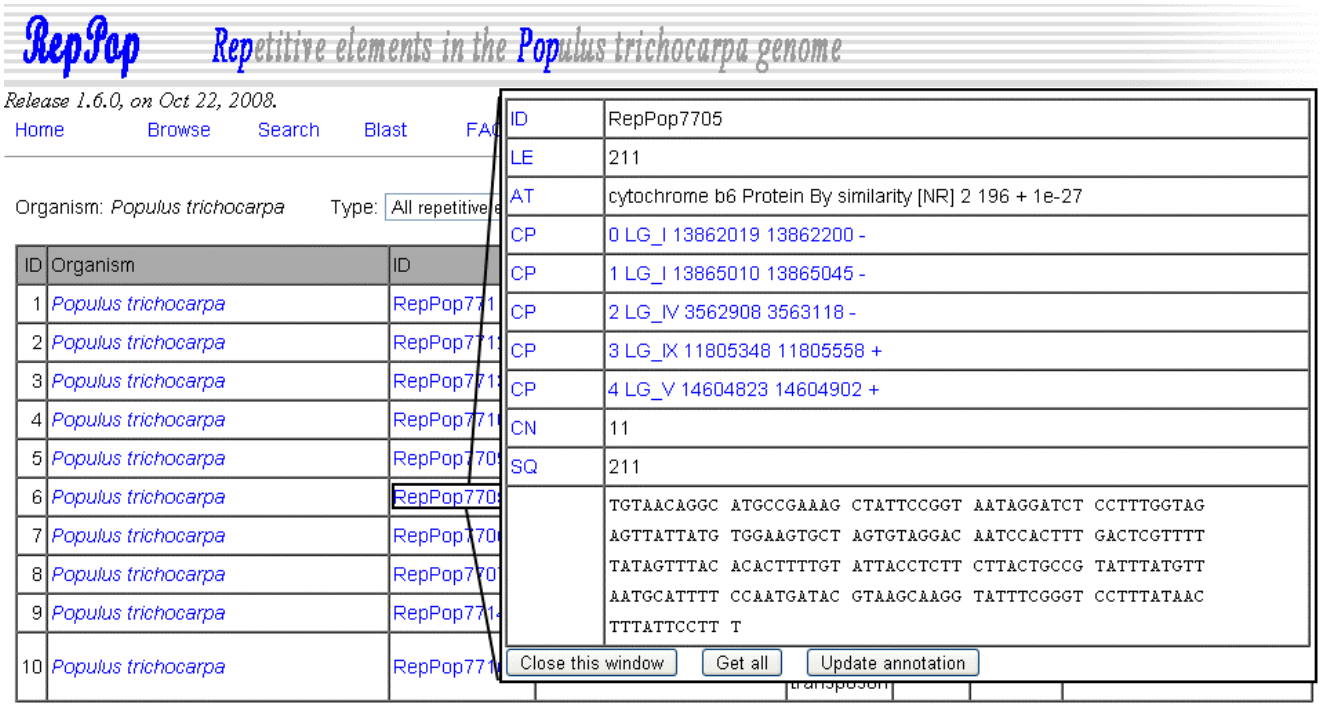

Figure 3
The browsing interface of database *RepPop*. All 9,623 predicted repetitive elements in the *Populus trichocarpa* genome. The user may click the entries in the column REName to retrieve detailed annotation of each repetitive element. A detailed description of the plant *Populus trichocarpa* could be found in a popup window by clicking on *Populus trichocarpa*.

(say, monthly) basis, based on the input provided by the users through PmWiki. Figure 7 shows a screen shot of using this feature. The user needs to register to open an account through the provided link in the right sidebar of main interface of RepPopWiki [35] before being able to add and revise the annotations for selected elements.

Manual input and other useful information

A Help interface is provided to help the users to get familiar with how to use *RepPop*. A detailed description of using various interfaces of *RepPop* can be found on this page. A collection of comprehensive databases of repetitive elements and computational programs for identifying such elements is provided in this Help interface, a user of which may be interested in identifying repetitive elements in other genomes. A list of Frequently Asked Questions (FAQs) is included in the FAQ interface.

Comparison with other databases of plant repetitive elements

There are quite a few databases focusing on the repetitive elements in plants. RetrOryza [25] collects 242 families of LTR retrotransposons in the rice genome, AtRepBase [26] provides the browsing and blasting interfaces for the 63 well annotated repetitive elements in the *Arabidopsis* genome and TREP [24] represents a community joint effort to collect and annotate the repetitive elements in the *Triticeae* genomes. All above three databases collect a limited number of repetitive elements with well curated annotations in one or a few closely related organisms. Our database, *RepPop*, computationally identified all the families of repetitive elements and tried to annotate them using sequence mapping. We have classified them as RNA, transposon and unknown genes, which is similar to the classification system of TREP.

Conclusion

RepPop is a database currently consisting of all the 9,623 predicted repetitive elements in the *Populus trichocarpa* genome along with functional annotations for some of them. Various search capabilities are provided in support of using this database by a large community of users. One unique feature of the database is that it allows users to add their annotations and curations to selected repetitive elements in a fashion similar to Wikipedia, which should help to rapidly increase the amount of information stored in this database.

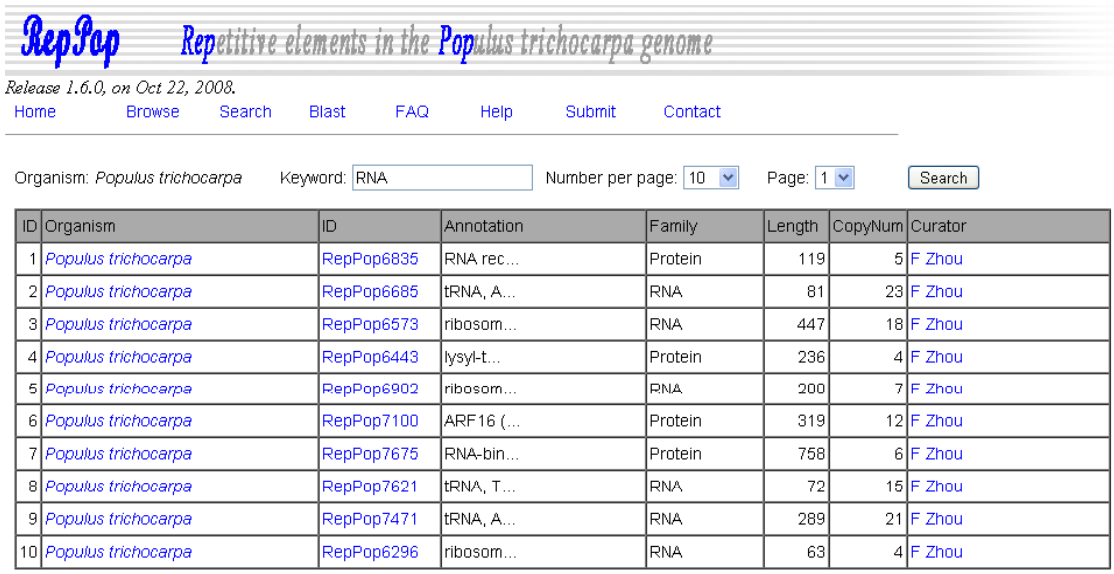

Figure 4
The searching interface of database *RepPop*. A user can search for repetitive elements with keywords through the keyword search interface.

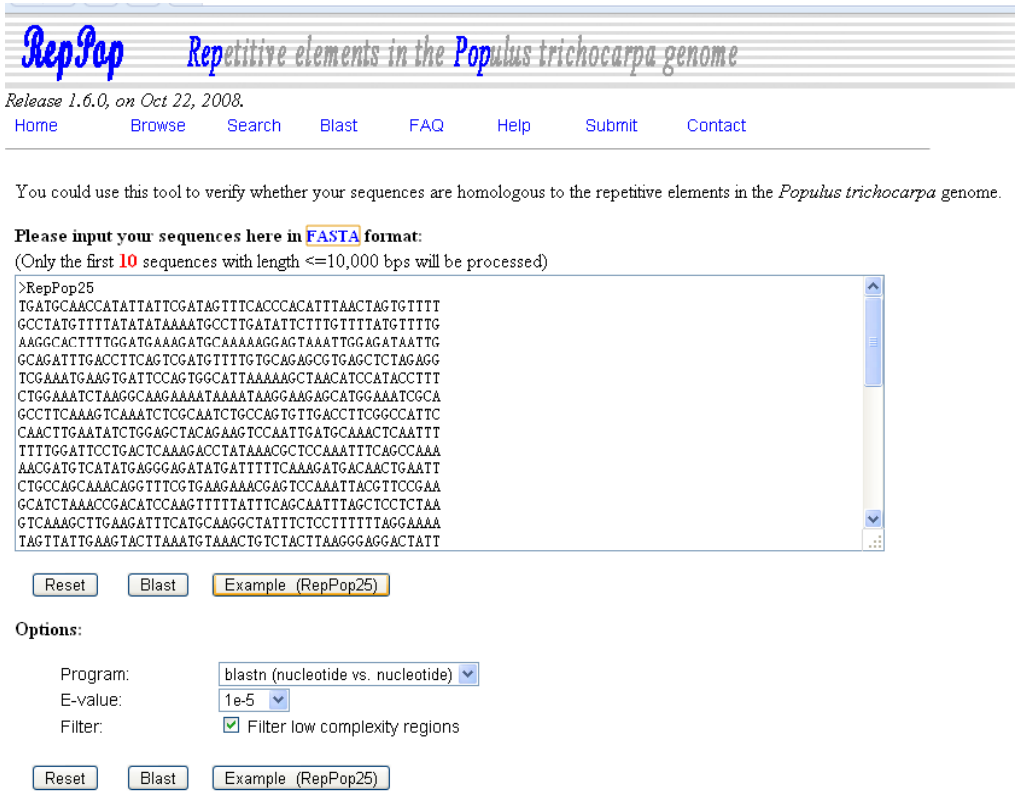

Figure 5
The blasting interface of database *RepPop*. Searches for the homologous regions for user-specified DNA or protein sequences in *RepPop*.

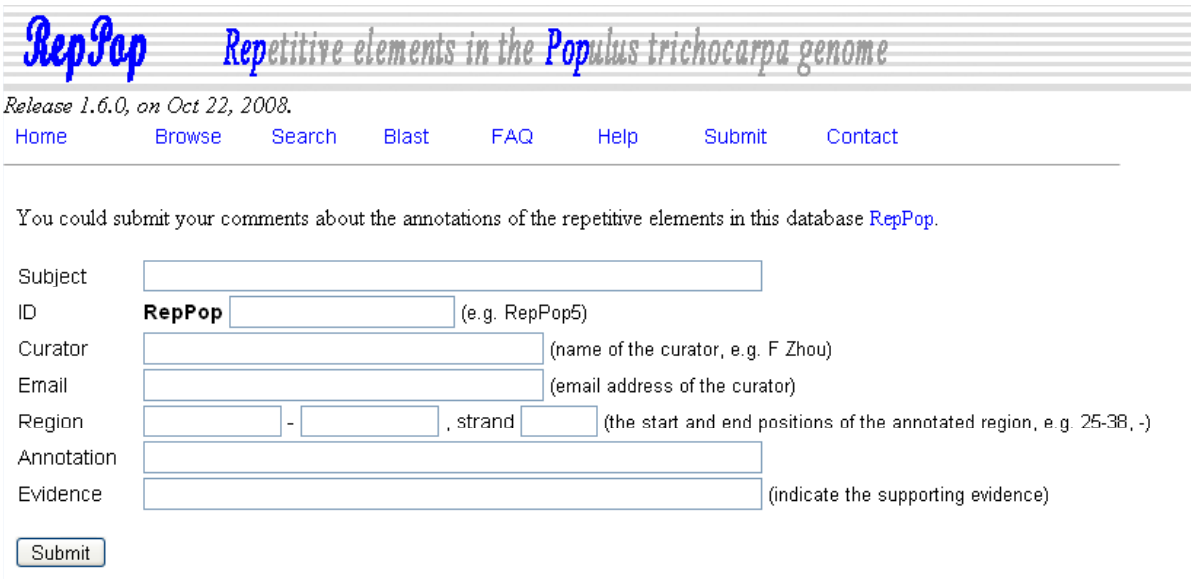

Figure 6
The submision interface of database *RepPop*. A user of *RepPop* may submit his/her annotations on a specific repetitive element with supporting evidence through this interface.

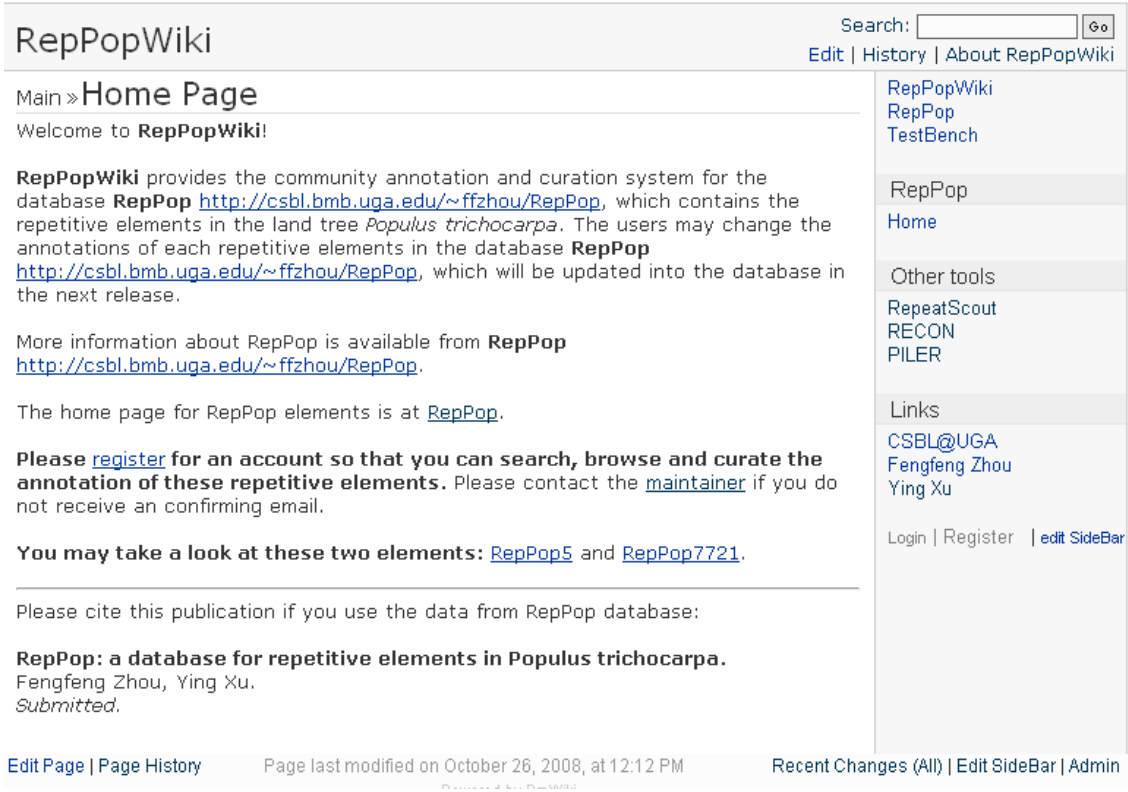

Figure 7
The Wiki interface of database *RepPop*. A user can revise the annotation of a specific repetitive element through this interface.

Future perspectives

More efforts are being put into manual curations to provide more accurate annotations of the predicted repetitive elements, especially for the chimeric ones. Curations from other researchers, including users, are encouraged, as discussed above, through the web site of *RepPop*.

Availability and requirement

Project name: The repetitive elements in *Populus trichocarpa* genome.

Project home page: http://csbl.bmb.uga.edu/~ffzhou/RepPop/.

Operating system(s): Platform independent.

Programming languages: PHP.

License: Not required.

Any restrictions to use by non-academics: None.

Abbreviations

RepPop: Repetitive elements in the *Populus trichocarpa* genome; IS: Insertion Sequence; LTR: Long Terminal Repeat.

Authors' contributions

FZ conceived the project, performed the identification and annotation of the data, and wrote the manuscript. YX wrote and polished the manuscript, and served as the principle investigator of the project. All authors have read and approved the final submitted version of this manuscript.

Acknowledgements

This work is supported in part by the National Science Foundation (DBI-0354771, ITR-IIS-0407204, DBI-0542119, CCF0621700), also National Institutes of Health (1R01GM075331 and 1R01GM081682) and a Distinguished Scholar grant from the Georgia Cancer Coalition, and the grant for the BioEnergy Science Center http://genomicsgtl.energy.gov/centers/center_ORNL.shtml, which is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. We thank the colleagues in the Biofuel group of UGA CSBL for their comments on this work. We would also like to thank the two anonymous reviewers for helpful and constructive comments on our work.

References

1. Taylor G: **Populus: arabidopsis for forestry. Do we need a model tree?** *Ann Bot (Lond)* 2002, **90(6):**681-689.
2. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, *et al.*: **The genome of black cottonwood, Populus trichocarpa (Torr. & Gray).** *Science* 2006, **313(5793):**1596-1604.
3. Basso LC, de Amorim HV, de Oliveira AJ, Lopes ML: **Yeast selection for fuel ethanol production in Brazil.** *FEMS yeast research* 2008.
4. Li X, Weng JK, Chapple C: **Improvement of biomass through lignin modification.** *Plant J* 2008, **54(4):**569-581.
5. Schirmer-Michel AC, Flores SH, Hertz PF, Matos GS, Ayub MA: **Production of ethanol from soybean hull hydrolysate by osmotolerant Candida guilliermondii NRRL Y-2075.** *Bioresource technology* 2008, **99(8):**2898-2904.
6. Smith DB, Flavell RB: **Characterisation of the wheat genome by renaturation kinetics.** *Chromosoma* 1975, **50(3):**223-242.
7. **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.** *Nature* 2000, **408(6814):**796-815.
8. **The map-based sequence of the rice genome.** *Nature* 2005, **436(7052):**793-800.
9. Tuskan GA, Gunter LE, Yang ZK, Yin T, Sewell MM, DiFazio SP: **Characterization of microsatellites revealed by genomic sequencing of *Populus trichocarpa*.** *Can J For Res* 2004, **34(1):**85-93.
10. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, *et al.*: **A unified classification system for eukaryotic transposable elements.** *Nature reviews* 2007, **8(12):**973-982.
11. Ciampi MS, Schmid MB, Roth JR: **Transposon Tn10 provides a promoter for transcription of adjacent sequences.** *Proc Natl Acad Sci USA* 1982, **79(16):**5016-5020.
12. Reynolds AE, Felton J, Wright A: **Insertion of DNA activates the cryptic bgl operon in E. coli K12.** *Nature* 1981, **293(5834):**625-629.
13. Saedler H, Reif HJ, Hu S, Davidson N: **IS2, a genetic element for turn-off and turn-on of gene activity in E. coli.** *Mol Gen Genet* 1974, **132(4):**265-289.
14. Louarn JM, Bouchâe JP, Legendre F, Louarn J, Patte J: **Characterization and properties of very large inversions of the E. coli chromosome along the origin-to-terminus axis.** *Molecular & general genetics: MGG* 1985, **201(3):**467-476.
15. Reif HJ, Saedler H: **IS1 is involved in deletion formation in the gal region of E. coli K12.** *Molecular & general genetics: MGG* 1975, **137(1):**17-28.
16. Schneider D, Duperchy E, Coursange E, Lenski RE, Blot M: **Long-term experimental evolution in Escherichia coli. IX. Characterization of insertion sequence-mediated mutations and rearrangements.** *Genetics* 2000, **156(2):**477-488.
17. Li X, Heyer WD: **Homologous recombination in DNA repair and DNA damage tolerance.** *Cell Res* 2008, **18(1):**99-113.
18. Smit A, Hubley R, Green P: **RepeatMasker Open-3.0.** 2004 [http://www.repeatmasker.org].
19. Price AL, Jones NC, Pevzner PA: **De novo identification of repeat families in large genomes.** *Bioinformatics* 2005, **21(Suppl 1):**i351-358.
20. Edgar RC, Myers EW: **PILER: identification and classification of genomic repeats.** *Bioinformatics* 2005, **21(Suppl 1):**i152-158.
21. Levitsky VG: **RECON: a program for prediction of nucleosome formation potential.** *Nucleic Acids Res* 2004:W346-349.
22. **JGI_Populus_trichocarpa_v1.1** 2006 [http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html].
23. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements.** *Cytogenet Genome Res* 2005, **110(1–4):**462-467.
24. **TREP** 2008 [http://wheat.pw.usda.gov/ITMI/Repeats/].
25. Chaparro C, Guyot R, Zuccolo A, Piegu B, Panaud O: **RetrOryza: a database of the rice LTR-retrotransposons.** *Nucleic Acids Res* 2007:D66-70.
26. **AtRepBase** 1999 [http://nucleus.cshl.org/protarab/AtRepBase.htm].
27. Cole JR, Chai B, Farris RJ, Wang Q, Kulam-Syed-Mohideen AS, McGarrell DM, Bandela AM, Cardenas E, Garrity GM, Tiedje JM: **The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data.** *Nucleic Acids Res* 2007:D169-172.
28. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A: **Rfam: annotating non-coding RNAs in complete genomes.** *Nucleic Acids Res* 2005:D121-124.
29. Tanaka T, Antonio BA, Kikuchi S, Matsumoto T, Nagamura Y, Numa H, Sakai H, Wu J, Itoh T, Sasaki T, *et al.*: **The Rice Annotation Project Database (RAP-DB): 2008 update.** *Nucleic Acids Res* 2008:D1028-1033.
30. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, *et al.*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2008:D13-21.

31. Wootton JC, Federhen S: **Analysis of compositionally biased regions in sequence databases.** *Methods Enzymol* 1996, **266:**554-571.
32. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27(2):**573-580.
33. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res* 1997, **25(5):**955-964.
34. **RepPop** 2008 [http://csbl.bmb.uga.edu/~ffzhou/RepPop/].
35. **RepPopWiki** 2008 [http://csbl.bmb.uga.edu/~ffzhou/RepPopWiki/].
36. **PmWiki** 2008 [http://www.pmwiki.org/].