## Overview

Purpose

This document details the steps needed to install and execute TransposonScout for personal use. TransposonScout is a bioinformatics tool for researchers interested in determining possible horizontal gene transfers of an organism based on primary DNA sequence.

Scope

TransposonScout is a Perl-based program that pipes together the functions of RepeatScout (determines repetitive regions of a DNA sequence which correspond to transposable elements) and BLAST search (searches NCBI sequence database for DNA matches in non-query species using RepeatScout transposable elements for queries). BLAST results are parsed, organized by species and displayed in a GUI interface.

System Requirements

* Linux OS
* 1GB or more of RAM
* 1MB of hard disk space for TransposonScout
* Internet connection

NOTE: TransposonScout may produce a large number of files during operation and more than 1MB of hard disk space may be required.

## Installation

Prerequisites

* + 1. BioPerl distribution containing RemoteBlast module.
    2. Tk distribution.

Installation

* + 1. Download TransposonScout-2.0-linux.tar.gz.
    2. ‘cd’ to directory containing TransposonScout-2.0-linux.tar.gz.
    3. ‘tar –zxvf TransposonScout-2.0-linux.tar.gz –C {desired directory}’ in the directory containing the archive file. A directory called “TransposonScout-2.0-linux” will be created in the desired directory.
    4. ‘cd’ to “TransposonScout-2.0-linux” directory.
    5. ‘perl TransposonScout.pl’ to execute TransposonScout.

## Usage

Executing full pipeline

* + 1. Click C:\Users\Prasath\Desktop\Bioinformatics\TransposonScout\v2.0\folder_icon.gif and open FASTA formatted genomic sequence. Full path name will appear within “Select sequence (FASTA format):” entry field.   
         
       CAUTION: TransposonScout does **not** validate whether the file is FASTA formatted. This may cause an error to be thrown during execution.
    2. Enter the binomial species name of the genomic sequence selected into the “Binomial species name:” entry field.   
         
       CAUTION: TransposonScout does **not** validate whether the species name is correct and may **not** throw an error during execution. This may cause final outputs to be not what the user desired (i.e., blast results returning matches to the species of the input genome). Ensure that species name is correct before proceeding.
    3. Click “Submit” button.

If either of the above fields is left empty a dialog box will appear informing the user to ensure they are filled.

* + 1. A progress window will appear detailing the operation currently being executed. Files from the operation are placed into a folder within “TransposonScout-2.0-linux/results” in a directory based on the filename of input sequence. For example, if the input sequence was “human3M.fa”, a directory called “human3M-Wed-May-22-18-48-12-2013” will be created where the date corresponds to the date and time the file was executed.
    2. Wait until execution of the pipeline is completed. This may take several hours depending on the number of repetitive regions identified within the input sequence. Please refer to the “Troubleshooting” section of this document if the progress window displays any errors.
    3. Once execution is completed, a dialog box will appear asking the user to proceed to the output window. Press “OK”.
    4. Transposable elements identified within the input sequence are queried to BLAST and matches of these elements in species other than the input species are returned. Results are organized by the matching species. Click on a species from the list on the right-side.
    5. Transposable element consensus sequences (e.g., R=1, R=2, R=3, etc…) that match to the selected species will appear below in the list. Click on a transposable element consensus sequence.
    6. Information about the match (Hit information) and High-scoring Segment Pair (HSP) information for a particular transposable element will appear in the “Output” box on the left.

Executing RemoteBlast  
  
Execution of RepeatScout results in the creation of a file containing transposable element consensus sequences found in the input sequence. This file is located under the directory created for input sequence (see 3.1.4). This file can be blasted and parsed without having to execute RepeatScout again.

* + 1. From the starting window, click “Run” from the menu bar and choose “Run RemoteBlast”.
    2. Click C:\Users\Prasath\Desktop\Bioinformatics\TransposonScout\v2.0\folder_icon.gif and locate file containing transposable element consensus sequences.
    3. Enter the name of the species that these sequences originate from. Ensure that this correct (see 3.1.2).
    4. Click the “Submit” button.
    5. A progress window will appear displaying the progress of the fragmentation, BLAST and parsing process. Please refer to the “Troubleshooting” section of this document if any errors occur.
    6. Once execution is completed, a dialog box will appear asking the user to proceed to the output window. Press “OK”.
    7. “.unfrag” file is fragmented into multiple “.filtered\_1” files each containing one FASTA formatted transposable element consensus sequence. These “.filtered\_1” files are queried to BLAST and matches of these elements in species other than the input species are returned. Results are organized by the matching species. Click on a species from the list on the right-side.
    8. Transposable element consensus sequences (e.g., R=1, R=2, R=3, etc…) that match to the selected species will appear below in the list. Click on a transposable element consensus sequence.
    9. Information about the match (Hit information) and High-scoring Segment Pair (HSP) information for a particular transposable element will appear in the “Output” box on the left.

Executing Blast Parser  
  
Execution of Remote BLAST will result in the creation of multiple “.blast\_xml” files. Each file contains BLAST results for the transposable element consensus sequences. The files are located in the “blast reports” directory within the directory created for the input sequence (see 3.1.4). These files can be parsed without execution the entire pipeline.

* + 1. From the starting window, click “Run” from the menu bar and choose “Run Blast Parser”.
    2. Click C:\Users\Prasath\Desktop\Bioinformatics\TransposonScout\v2.0\folder_icon.gif and locate the directory containing the “.blast\_xml” files.
    3. Enter the name of the species that these sequences originate from. Ensure that this correct (see 3.1.2).
    4. Click the “Submit” button.
    5. A progress window will appear displaying the progress of the parsing process. Please refer to the “Troubleshooting” section of this document if any errors occur.
    6. Once execution is completed, a dialog box will appear asking the user to proceed to the output window. Press “OK”.
    7. “.blast\_xml” files are parsed and results are organized by the matching species. Click on a species from the list on the right-side.
    8. Transposable element consensus sequences (e.g., R=1, R=2, R=3, etc…) that match to the selected species will appear below in the list. Click on a transposable element consensus sequence.
    9. Information about the match (Hit information) and High-scoring Segment Pair (HSP) information for a particular transposable element will appear in the “Output” box on the left.

## Troubleshooting

Errors

* + 1. Dialog box saying “Please ensure both fields are filled.”  
         
       One or both entry fields were left empty upon clicking the “Submit” button. Ensure the input sequence (or directory) field and “Binomial species name” field are filled.
    2. “Error: Cannot create results directory this file.”  
         
       There was an error when creating the directory to store the files created during the execution of the pipeline. Ensure that TranposonScout.pl has “write” privileges.
    3. “Error: Cannot build seeds.”  
         
       An error occurred during execution of build\_lmer\_table script of RepeatScout. Ensure that input sequence is a FASTA-formatted DNA sequence.
    4. “Error: Unable to complete consensus building.”  
         
       An error occurred during execution of the RepeatScout script of RepeatScout. This may be due to any number of reasons. One solution may be to bypass the RepeatScout step in this pipeline and execute RemoteBlast step after running RepeatScout manually.
       1. Execute RepeatScout via command line.
       2. Execute RemoteBlast (see 3.2) on the file generated by the filter-stage-1 script of RepeatScout.
    5. “Error: Unable to complete filtering process.”  
         
       An error occurred during the execution of filter-stage-1 script of RepeatScout. This may have occurred because the file generated from RepeatScout script did not contain any sequences (it may be possible that the sequence does not contain any transposable elements). See 4.1.4 for a possible solution.
    6. “Error: Unable to locate repeat library.”  
         
       An error occurred during library fragmentation step of TransposonScout pipeline. TransposonScout was unable to locate the “output\_repeats.unfrag” file generated from completion of the RepeatScout portion of the pipeline. See 4.1.4 for a possible solution.
    7. “Error: Unable to locate BLAST input files.”  
         
       An error occurred during Remote BLAST execution. The fragmentation step of the pipeline failed to produce the “.filtered\_1” files for BLAST querying. Ensure that the file containing transposable element consensus sequences are in FASTA format.
    8. “Error: Unable to locate BLAST XML reports.”  
         
       An error occurred during parsing of BLAST reports. Ensure that the selected directory contains XML formatted BLAST reports with the “.blast\_xml” file extensions. It is also possible that RemoteBlast was not executed due to the lack of network connectivity. Ensure that there is a stable Internet connection and execute RemoteBlast (see 3.2).

## Frequently Asked Questions

Can TransposonScout run on Windows?  
  
It is possible to execute TranposonScout.pl (via Cygwin) assuming that BioPerl and Tk libraries are installed. Unfortunately the RepeatScout program that is part of TransposonScout’s pipeline fails to execute on Windows. If you are content with simply using the “RemoteBlast” and “Blast Parser” features of TransposonScout then it is possible to run TransposonScout on Windows.

## Maintenance

Please contact [ppararaj@gmail.com](mailto:ppararaj@gmail.com) to report any bugs associated with this program.