Automated Homology-based Approach for the Identification of Transposable Elements: TESeeeker

Ryan C. Kennedy
November 2010
Transposable Elements (TEs)

- First found and analyzed by Barbara McClintock in 1948
  - Won Nobel Prize in 1983
- TEs are mobile pieces of DNA
- Typically divided into Class I and Class II elements
  - Class I elements are RNA-mediated
  - Class II elements are DNA-mediated
- Example *mariner* Class II TE:

<table>
<thead>
<tr>
<th>TA</th>
<th>ACGC...GTAA</th>
<th>GTATCAGCCA...CAAATTACG</th>
<th>TTAC...GCGT</th>
<th>TA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target Site Duplication</td>
<td>Inverted Repeat</td>
<td>Transposase</td>
<td>Inverted Repeat</td>
<td>Target Site Duplication</td>
</tr>
<tr>
<td>2 bp</td>
<td>20–30 bp</td>
<td>~900 bp</td>
<td>20–30 bp</td>
<td>2 bp</td>
</tr>
</tbody>
</table>
Motivation

Why study transposable elements?

- Have been found in all eukaryotic genomes
- Occupy large portions of genomes
  - 50% of human genome
  - 47% of *Aedes aegypti* mosquito genome
- Can influence genome evolution and gene expression

Mouse Genome Sequencing Consortium:

“The single most prevalent feature of mammalian genomes is their repetitive sequences, most of which are interspersed repeats representing ‘fossils’ of transposable elements. *Transposable elements are a principal force in reshaping the genome, and their fossils thus provide powerful reporters for measuring evolutionary forces acting on the genome.*”

TE Discovery Techniques

- Bergman and Quesneville categorize TE discovery into four categories:
  1) Comparative Genomic Methods
     - Perform multiple sequence alignment of related genomes and look for large changes amongst them
     - Good for finding new TE families, but relies on readily available, properly sequenced, related genomes
  2) De novo
     - Detect similar sequences found throughout the genome and cluster
     - Can discover new TE families, but often difficult to distinguish closely related TEs

TE Discovery Techniques

3) Structure-based
   - Use TE structural data, such as inverted repeats, to find TEs
   - Works well for characterized TEs, but does not locate degraded TEs or TEs with non-distinct structures

4) Homology-based (our approach)
   - Use known TEs as seeds to search in novel genomes
   - Can discover new TE families, but requires additional verification
Challenges in Locating TEs

- Although present in all eukaryotic genomes, difficult to annotate
  - Varying structural characteristics
  - Mobile nature often leads to copies within copies
  - TEs often are very degraded
Class II TE Evolution

- TE Insertion: past
- Transposon DNA
- Host DNA
- TE Multiplication
- TE Regulation by host: present
Manual Approach

- Developed and utilized during TE search on very different genome projects:
  - *Pediculus humanus humanus* (body louse)
    - Comprehensive search for all TE families
  - *Culex quinquefasciatus* (mosquito)
    - Search for non-LTR TEs
- Homology-based
  - Assembled representative TE library of high-quality TEs
    - Intact open reading frames
  - Results appear in TE sections of respective genome papers
Manual Approach

Representative transposases

tblastn

Combine, add flanks, extract

Assemble in DNASTAR SeqMan II

Consensus TE
### Results

<table>
<thead>
<tr>
<th>Class I</th>
<th>Family</th>
<th>Element</th>
<th>Length (bp)</th>
<th>Full-length Copies</th>
<th>Partial Hits</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-LTR</td>
<td>SART</td>
<td>Hope-like</td>
<td>4655</td>
<td>1</td>
<td>522</td>
<td>0.18%</td>
</tr>
<tr>
<td></td>
<td>R4</td>
<td>Dong-like</td>
<td>5266</td>
<td>4</td>
<td>1739</td>
<td>0.45%</td>
</tr>
<tr>
<td>LTR</td>
<td>Ty3/gypsy</td>
<td>Mdg1</td>
<td>5395</td>
<td>2</td>
<td>976</td>
<td>0.28%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class II</th>
<th>Family</th>
<th>Element</th>
<th>Length (bp)</th>
<th>Full-length Copies</th>
<th>Copies</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mariner/Tc1</td>
<td>mariner</td>
<td></td>
<td>1276</td>
<td>24</td>
<td>216</td>
<td>0.09%</td>
</tr>
</tbody>
</table>

**TOTAL**                      |        |          |              |        |        | 1.0%    |

---

# C. quinquefasciatus Results

<table>
<thead>
<tr>
<th>Class I</th>
<th>Family</th>
<th>Full-length Copies</th>
<th>Partial Hits</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-LTR</td>
<td>CR1</td>
<td>31</td>
<td>973</td>
<td>0.28%</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>11</td>
<td>63</td>
<td>0.02%</td>
</tr>
<tr>
<td></td>
<td>Jockey</td>
<td>14</td>
<td>5028</td>
<td>1.77%</td>
</tr>
<tr>
<td></td>
<td>L1</td>
<td>57</td>
<td>662</td>
<td>0.15%</td>
</tr>
<tr>
<td></td>
<td>L2</td>
<td>9</td>
<td>1416</td>
<td>0.61%</td>
</tr>
<tr>
<td></td>
<td>Loa</td>
<td>9</td>
<td>184</td>
<td>0.09%</td>
</tr>
<tr>
<td></td>
<td>Loner</td>
<td>2</td>
<td>127</td>
<td>0.12%</td>
</tr>
<tr>
<td></td>
<td>Outcast</td>
<td>4</td>
<td>15</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>R1</td>
<td>32</td>
<td>250</td>
<td>0.14%</td>
</tr>
<tr>
<td></td>
<td>RTE</td>
<td>8</td>
<td>892</td>
<td>0.38%</td>
</tr>
<tr>
<td></td>
<td>Unclassified LINE</td>
<td>32</td>
<td>11,117</td>
<td>0.88%</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td></td>
<td></td>
<td><strong>4.44%</strong></td>
</tr>
</tbody>
</table>

Manual Approach

Representative transposases

Genome

tblastn

Combine, add flanks, extract

BLAST hits

major burden

Assemble in DNASTAR SeqMan II

Consensus TE
DNASTAR SeqMan II

- Manually trimming hits and additional processing is time consuming
- Can only assemble limited number of sequences at a time
Automated Approach

- Homology-based
- Replace DNASTAR SeqMan II and manual analysis with other tools
  - CAP3, Clustal, various scripts
- Iterative - repeat steps if necessary
Automated Approach Steps

1. Identify transposase in target genome
2. Find copies in target genome with flanks
3. Generate consensus from multiple sequence alignment of copies
4. Use consensus to identify TE

- Output: putative high-quality consensus TE which can in turn be used locate instances within the genome
- Runs in a matter of minutes/hours
  - Dependent on genome size, size of representative TEs, and richness of TEs in the genome
- Runs via web interface or via automated scripts
Automated Approach

1. Better Consensus TE
2. Density

For each contig...

CAP3 Assembly and trimming yields representative transposase from genome

Coding Region

1st

2nd

Consensus

3rd

MSA

1) Better Consensus TE
2) Density
Step 1

- Identify transposase(s) in target genome
  - `tblastn` representative transposases against genome
  - Parse BLAST file with the following parameters:
    - combine threshold: maximum distance sequences can be apart to join as a single hit
    - minimum length percentage: must be at least this percentage of query sequence to be considered
    - e-value cutoff: ignore everything worse than this value, typically 1E-20
    - flank size: amount of extra sequence to add to each end of hit (0)
  - Extract genomic sequences from above and iteratively assemble with CAP3
    - With CAP3, specify quality window size and threshold, as well as combine threshold

transposase(s) within genome

17
Step 2

- Find copies in target genome with flanks
  - `blastn` transposase(s) against genome
    - Parse BLAST file with the following parameters:
      - combine threshold: maximum distance sequences can be apart to join as a single hit
      - minimum length percentage: must be at least this percentage of query sequence to be considered
      - e-value cutoff: ignore everything worse than this value, typically 1E-20
      - flank size: amount of extra sequence to add to each end of hit
  - Extract genomic sequences from above

Copies within genome with flanks
Step 3

- Obtain consensus from multiple sequence alignment (MSA) of copies
  - Perform MSA on sequences
  - Generate consensus from MSA
    - Can specify percentage of nucleotides that must be common amongst sequences to count in consensus

Putative Consensus
Step 4

- Use consensus to identify proper TE
  - *blastn* representative transposases against genome
    - Parse BLAST file with the following parameters:
      - combine threshold: maximum distance sequences can be apart to join as a single hit
      - minimum length percentage: must be at least this percentage of query sequence to be considered
      - e-value cutoff: ignore everything worse than this value, typically 1E-20
      - flank size: amount of extra sequence to add to each end of hit
  - Extract genomic sequences from above and iteratively assemble with CAP3
    - With CAP3, specify quality window size and threshold, as well as combine threshold

Consensus TE $\rightarrow$ Density
Automated Approach Schematic

Step 1
- TE Database
  - Representative transposases
  - BLAST
    - BLAST Results
      - PERL Scripts

Step 2
- CAP3
  - Extracted Sequences
  - BLAST
    - BLAST Results
      - PERL Scripts

Step 3
- BLAST
  - BLAST Results
  - PERL Scripts
  - Copies
  - MSA
    - Consensus
      - "Step 1"

Step 4
- TE
Validation Strategy

1. Initially evaluated automated approach on *P. humanus humanus* and *C. quinquefasciatus*
   - Validate against high-quality manually verified annotation
   - Identify default starting parameters

2. Check automated results versus published results

3. Genomes in General:
   - Translate consensus TE sequences
     - Identify open reading frame
       - blastp against *non-redundant protein (nr)* database at NCBI and check for conserved domains/hits
   - Can check for structural signatures
Validation (1): *P. humanus humanus mariner*

- Full *mariner* element identified following Step 4
- Validated against manual effort
  - TSDs; 14 bp terminal inverted repeats (TIRs); well-trimmed
Validation (2): *Anopheles gambiae* PEST

- **P** elements (Class II)
  - Sarkar et al. (2003) identified 6 distinct elements
  - Oliveira de Carvalho et al. (2004) identified 4 additional elements
  - Quesneville et al. (2006) identified 9 elements at least 30% divergent at nucleotide level
  - **Total:** 12 elements at least 30% divergent at nucleotide level

- Automated Approach
  - Identified 11/12 elements + 2 partial hits
  - Captured TIRs where previously described
Validation (3)

- Searched for *mariner* in a number of genomes
  - In agreement where previously reported
    - Human, frog, chicken
  - In agreement where not reported
    - Dog, cat, horse
  - Possible discovery
    - *Drosophila melanogaster* putative *mariner*
      - 1061 bp element has TIRs
      - 26 bp TIRs
      - No apparent TSDs
      - Single full-length copy, as well as several partial hits
      - Transposase is most similar to that of *Chymomyza amoena*, 77% identical at the amino acid level
      - Searches for this element in existing TE annotations for D. melanogaster produced no hits
Implementation

- Approach implemented as TESeeker
  - VirtualBox virtual appliance
    - Cross-platform
    - Completely configured, no need to install scripts
      - Provide only genome FASTA file
      - Optionally provide additional library files
  - Local web interface

- [http://www.nd.edu/~teseeker](http://www.nd.edu/~teseeker)
  - Virtual appliance
  - Documentation
  - TE Library
TESeeker Desktop
TESeeker Desktop

File Browser:
- Genomes
- TELibrary
- Documentation
- TESeeker
- BLAST
- Extract
- license.txt

Files:
- gambel_ac.fa
- hAT_ac.fa
- mariner_ac.fa
- p_ac.fa
- piggybac_ac.fa
- pogo_ac.fa
- TCI_ac.fa

File size: 7 items, Free space: 32.5 GB
Transposable elements (TEs) are a type of repetitive sequence that have been found in nearly all eukaryotic genomes. First discovered and analyzed by McClintock in the 1950s, TEs have the ability to move about and replicate within a genome. Due to their mobile and replicative nature, TEs often occupy large portions of genomes. This prevalence of TEs poses a major difficulty in sequence assembly, as repeat regions are prone to misassembly. TEs can impact host genomes in a number of ways. They are believed to play a major role in genome evolution, as they can insert themselves into, mutate, and move genes, thereby influencing gene expression, causing gene variation, and transferring genetic material.

With the number of sequenced genomes rapidly rising, the need to identify TEs within them also grows. The ability to do this automatically and effectively in a manner similar to the methods used for genes is of increasing importance. This document describes how to use the implementation of our approach, TESeeeker to identify high-quality consensus TEs.
TESeeker Desktop
TESeeker Desktop
TESeeker Desktop
TESeeker Walkthrough

Identify *mariner* element in *P. humanus humanus*:
- Start with default parameters
- Make sure genome file and library file are present
- Start search
TESeeker Desktop
# TESSeeker Local Web Interface

![TESSeeker](http://localhost/TESSeeker.png)

## TESSeeker

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLAST Query Library</td>
<td>mariner_ac.fa</td>
</tr>
<tr>
<td>Closeness to Combine BLAST Hits</td>
<td>50 bp</td>
</tr>
<tr>
<td>BLAST type</td>
<td>tblastn</td>
</tr>
<tr>
<td>BLAST Database</td>
<td>phumanus.SUPERCONTIGS-USDA.PhumU1.fa</td>
</tr>
<tr>
<td>Closeness to combine CAP3 Hits</td>
<td>50 bp</td>
</tr>
<tr>
<td>CAP3 Window Size</td>
<td>20 bp</td>
</tr>
<tr>
<td>CAP3 Quality Threshold Multiplier</td>
<td>.9</td>
</tr>
<tr>
<td>CAP3 Quality Threshold</td>
<td>18</td>
</tr>
<tr>
<td>Desktop Output Folder Name</td>
<td>louseOut</td>
</tr>
<tr>
<td>Find Consensus?</td>
<td>checked</td>
</tr>
<tr>
<td>Flank Size</td>
<td>300 bp</td>
</tr>
</tbody>
</table>

Submit Reset

Done
TESeeker Status

Job 5 is Running
TESeeker Status

Job Finished! See results in desktop folder IouseOut.
# TESSeeker Results

![Index of file](file:///home/teseeker/Desktop/louseOut)

## Index of file:///home/teseeker/Desktop/louseOut

- **Name**
  - cap2
  - codingRegion_files
  - consen_files
    - consensus_contigs.fas
    - consensus_iter1_singlets.fas
    - consensus_singlets.fas
  - output

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Last Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>cap2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>codingRegion_files</td>
<td></td>
<td></td>
</tr>
<tr>
<td>consen_files</td>
<td></td>
<td></td>
</tr>
<tr>
<td>consensus_contigs.fas</td>
<td>15 KB</td>
<td>11/12/2010 06:54:27 PM</td>
</tr>
<tr>
<td>consensus_iter1_singlets.fas</td>
<td>1 KB</td>
<td>11/12/2010 07:11:56 PM</td>
</tr>
<tr>
<td>consensus_singlets.fas</td>
<td>4 KB</td>
<td>11/12/2010 07:12:11 PM</td>
</tr>
<tr>
<td>output</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
TESeeker Results

```plaintext
>Contig1-0  0 1309 f
ATAATACCTAATATGGTGGCCAAAATAGTAACTGCGGATTTTACCAACAGATT
TTTGTATTTTGTGAGTACGTATGGTGGGCTAACAGACATGAACCTTTTGATATGA
TTTACCTGTGTTCTCTGTACATCCCTGCTAAAAAAATATGCTGAACTCTTTAAATC
AACGCAAAGATATATTAAAATCTGAGGTGATCCAGGCCCAGGCGAACATTTCTT
TTTTTTTTGCAAAAGGTGTATATGTCTCTCCAGGCCCAGGTAAAAGGATGTGAGG
GGGATGAACGCTATATAGAAACGCAATGCTCAAAACTGTGTTGCGAAATTCGTTCTGAG
ATTTTCTTCTCGAAAATGAGAAAGCTGCTCCGGCGCCTACATGGAGTTTAAAGATGACACAGA
TAAGGCCCCCTATTAGTTATATGTCCGCAAGTATTTGCAACTAAGGAACCTAG
ATGATTCACATCTGGCCTCAAACACCGCTGCTGCCGCTCTTGGTGCCAAAAGAAGATCGT
ATACGTTATTTTTTGGGAACTGTAGTTAAACAGAGCCGAATCTGCTTTCGCAATGTCTCT
AAAACGCAAATGGCAATGACCATTTTGGAAAAAGATGCCGCTTAAATGGTCCAAGGCAACCCATACAC
TCTATGATGACTTTTTTGGAAAAAGATGCCTGTTTACGCAAAGGAAACGACCCACCAACAC
TCTAAGGCTCAGATTACCCAAAACAGAGATTTTGTATATTTTGTGGGATTACAAAGG
CATAGTCCACCTTGAGCTGCTGACGATCTGCGACCCATATAATCAAGATGTTTACCTCG
ACATTGACAAAAATTTAATGATACCATCAGAAAGAACGACCCGGAATACGCATAGCAA
AGAAAATGTCCTTACACCCACATATGCGGACCACCCTCCCACCCTTTAGGCCACTGGAACAAAA
ACTACTGCGGCTAGGCTGAAATGTCTTTCGTCGCCACCTCCCTATATGCGCAAAACATCGTCC
AAATAATTATACATTCTTTCCGGATCTCTAAAATTTTTTTAAAAAGGACAAAAATTCACAAAA
CGGCTAATGCGTCGAAAAACGCTGATGAGGACAGTATTGCTCTCAAAAAAATAGAGTTGCTG
TGAAAAAAAGGAAATGGGAGACTGACTACCCGAAAATGTCGAAAGATCTGACTAATATATATACAAAA
TCAATAATAGTAAAAATATTTGTGACATATATATTAAATGCGTTTCTTTTTCTTTA
AAACAGTGAAATTTTCTTGGCCACCCCAATAAATATGATGATGAA
```

Done
**TESeeker Results**

- **Alignment**
  - TESeeker top result with default parameters
  - 99% identity with manually annotated *mariner*
Developed and automated a homology-based approach to identify TEs
- Tedious and time-consuming task now automated
  - From months to hours or days
- Output: high-quality consensus TEs
  - Can be used to determine instances in genome (density)

Implemented as TESeeker
- Distributed as a virtual appliance
  - All tools and scripts
- Web interface
- Distributed with high-quality library of representative coding regions from major TE families

Approach contributed to multiple genome annotation projects
- Sequences available in TEfam database
- Most rigorously tested in arthropod genomes
Acknowledgments

This research was supported in part by:

- NIAID/NIH contracts HHSN272200900039C and HHSN266200400039C for “VectorBase: A Bioinformatics Resource Center for Invertebrate Vectors of Human Pathogens.”

- Eck Institute for Global Health
- Notre Dame Center for Research Computing
- Biocomplexity Cluster, supported in part by NSF MRI Grant No. DBI-0420980
Questions or Comments?