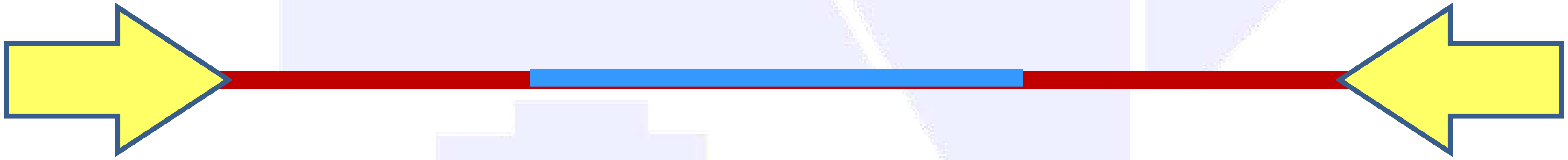
Automated Homology-based Approach for the Identification of Transposable Elements: TESeeker

Ryan C. Kennedy
November 2010

UNIVERSITY OF NOTRE DAME

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:



TA	ACGC…GTAA	GTATCAGCCA…CAAATTACG	TTAC…GCGT	TA
Target Site Duplication	Inverted Repeat	Transposase	Inverted Repeat	Target Site Duplication
2 bp	20-30 bp	~900 bp	20-30 bp	2 bp

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.*"

R.H. Waterston, et al. Initial sequencing and comparative analysis of the mouse genome. Nature, 420:520-562, December 2002.

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

C. M. Bergman and H. Quesneville. Discovering and detecting transposable elements in genome sequences. Briefings in Bioinformatics, 8(6):382-392, 2007.

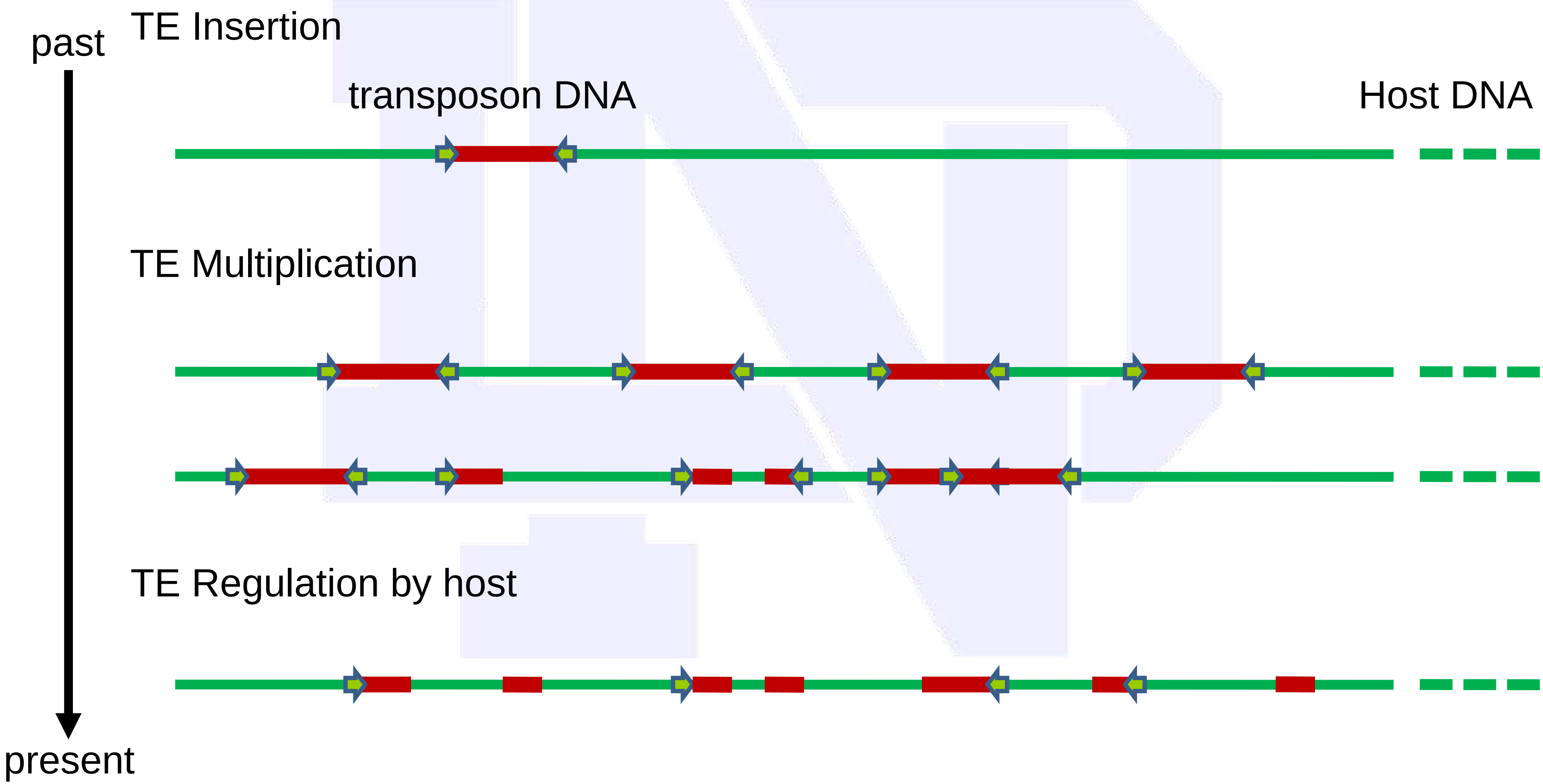
TE Discovery Techniques

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

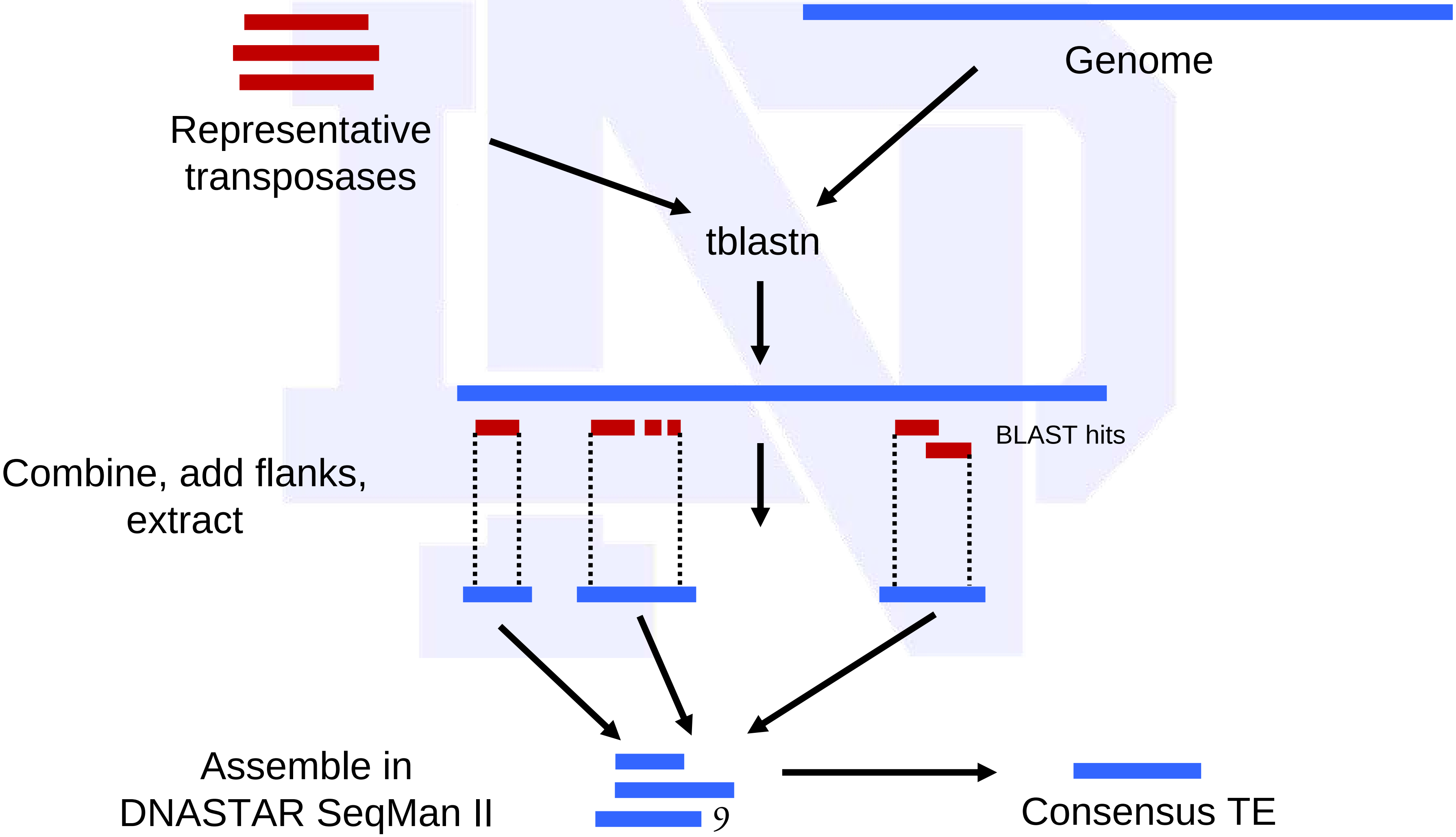
Class II TE Evolution



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

Genome

tblastn

BLAST hits

Combine, add flanks, extract

Assemble in DNASTAR SeqMan II

Consensus TE

P. humanus humanus Results

Class I	Family	Element	Length (bp)	Full-length Copies	Partial Hits	Density
Non-LTR	SART	*Hope-like*	4655	1	522	0.18%
	R4	*Dong-like*	5266	4	1739	0.45%
LTR	Ty3/gypsy	*Mdg1*	5395	2	976	0.28%
Class II	Family	Element	Length (bp)	Full-length Copies	Copies	Density
	Mariner/Tc1	*mariner*	1276	24	216	0.09%
TOTAL						1.0%

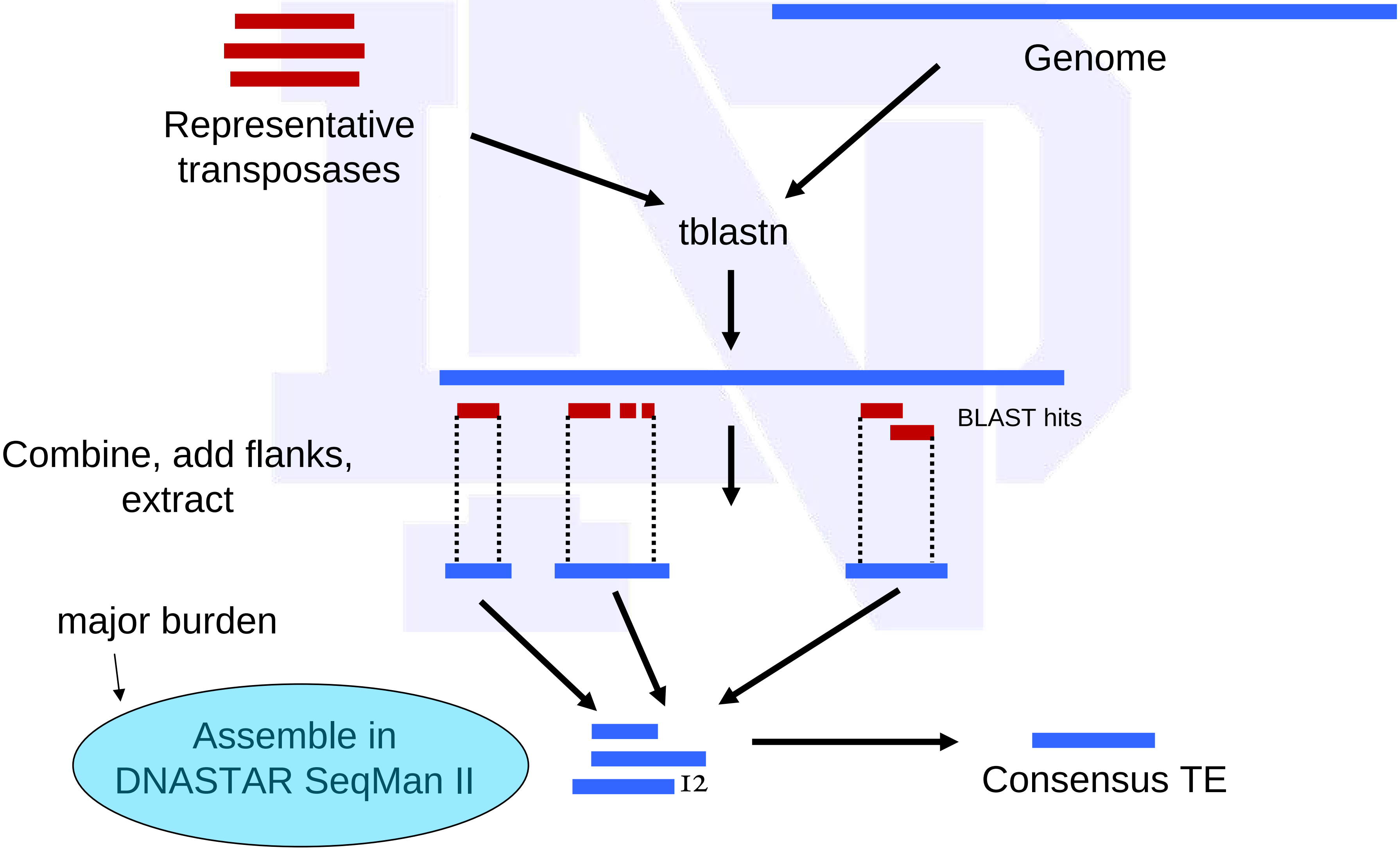
E.F. Kirkness et al., “Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle.” *Proceedings of the National Academy of Sciences*, 107(27):12168-12173, July 2010.

C. quinquefasciatus Results

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

P. Arensburger et al., "Sequence of *Culex quinquefasciatus* Establishes a Platform for vector Mosquito Comparative Genomics." *Science*, 330(6000):86-88, October 2010.

Manual Approach



Representative transposases

Genome

tblastn

BLAST hits

Combine, add flanks, extract

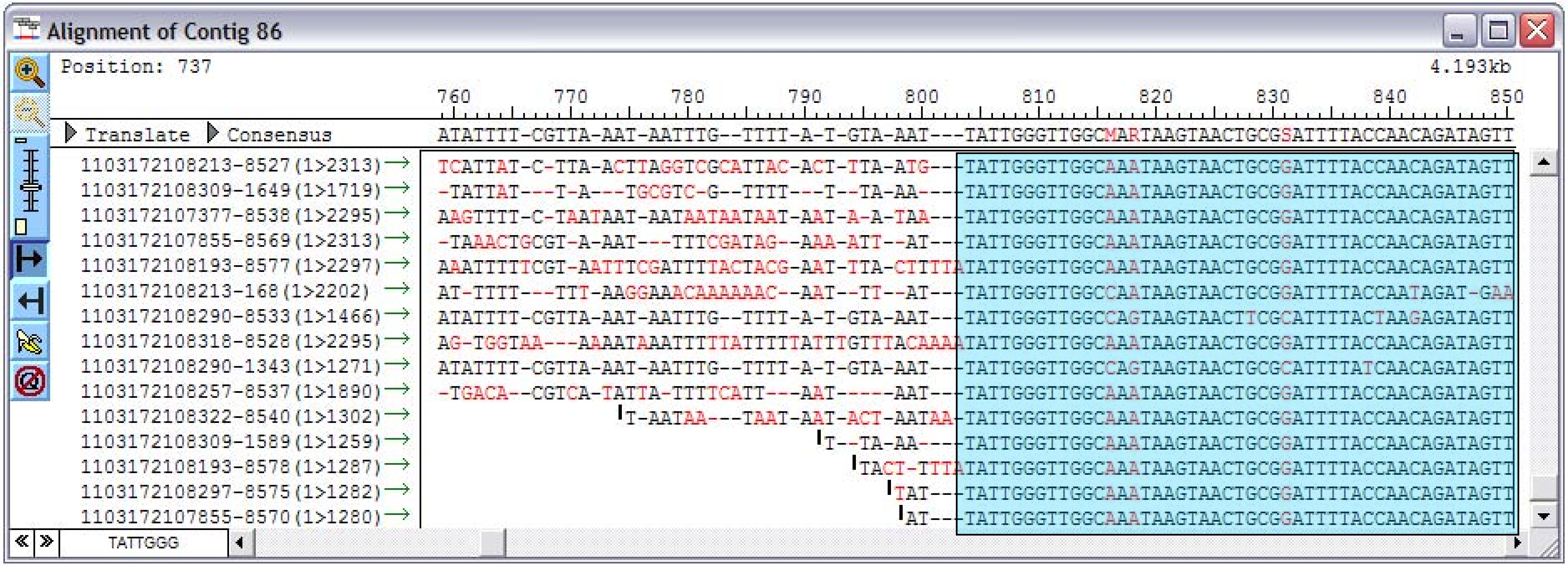
major burden

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



```
Position: 737                                                                                        4.193kb
                                  760       770       780       790       800       810       820       830       840       850
Translate  Consensus              ATATTTT-CGTTA-AAT-AATTTG--TTTT-A-T-GTA-AAT---TATTGGGTTGGCMARTAAGTAACTGCGSATTTTACCAACAGATAGTT
  1103172108213-8527(1>2313)→    TCATTAT-C-TTA-ACTTAGGTCGCATTAC-ACT-TTA-ATG---TATTGGGTTGGCAAATAAGTAACTGCGGATTTTACCAACAGATAGTT
  1103172108309-1649(1>1719)→    -TATTAT---T-A---TGCGTC-G--TTTT---T--TA-AA----TATTGGGTTGGCAAATAAGTAACTGCGGATTTTACCAACAGATAGTT
  1103172107377-8538(1>2295)→    AAGTTTT-C-TAATAAT-AATAATAATAAT-AAT-A-A-TAA---TATTGGGTTGGCAAATAAGTAACTGCGGATTTTACCAACAGATAGTT
  1103172107855-8569(1>2313)→    -TAAACTGCGT-A-AAT---TTTCGATAG--AAA-ATT--AT---TATTGGGTTGGCAAATAAGTAACTGCGGATTTTACCAACAGATAGTT
  1103172108193-8577(1>2297)→    AAATTTTTCGT-AATTTCGATTTTACTACG-AAT-TTA-CTTTTATATTGGGTTGGCAAATAAGTAACTGCGGATTTTACCAACAGATAGTT
  1103172108213-168(1>2202)  →   AT-TTTT---TTT-AAGGAAACAAAAAAC--AAT--TT--AT---TATTGGGTTGGCCAATAAGTAACTGCGGATTTTACCAATAGAT-GAA
  1103172108290-8533(1>1466)→    ATATTTT-CGTTA-AAT-AATTTG--TTTT-A-T-GTA-AAT---TATTGGGTTGGCCAGTAAGTAACTTCGCATTTTACTAAGAGATAGTT
  1103172108318-8528(1>2295)→    AG-TGGTAA---AAAATAAATTTTTATTTTTATTTGTTTACAAAATATTGGGTTGGCAAATAAGTAACTGCGGATTTTACCAACAGATAGTT
  1103172108290-1343(1>1271)→    ATATTTT-CGTTA-AAT-AATTTG--TTTT-A-T-GTA-AAT---TATTGGGTTGGCCAGTAAGTAACTGCGCATTTTATCAACAGATAGTT
  1103172108257-8537(1>1890)→    -TGACA--CGTCA-TATTA-TTTTCATT---AAT-----AAT---TATTGGGTTGGCAAATAAGTAACTGCGGATTTTACCAACAGATAGTT
  1103172108322-8540(1>1302)→                    T-AATAA---TAAT-AAT-ACT-AATAA-TATTGGGTTGGCAAATAAGTAACTGCGGATTTTACCAACAGATAGTT
  1103172108309-1589(1>1259)→                                     T--TA-AA----TATTGGGTTGGCAAATAAGTAACTGCGGATTTTACCAACAGATAGTT
  1103172108193-8578(1>1287)→                                        TACT-TTTATATTGGGTTGGCAAATAAGTAACTGCGGATTTTACCAACAGATAGTT
  1103172108297-8575(1>1282)→                                           TAT---TATTGGGTTGGCAAATAAGTAACTGCGGATTTTACCAACAGATAGTT
  1103172107855-8570(1>1280)→                                            AT---TATTGGGTTGGCAAATAAGTAACTGCGGATTTTACCAACAGATAGTT
```

Automated Approach

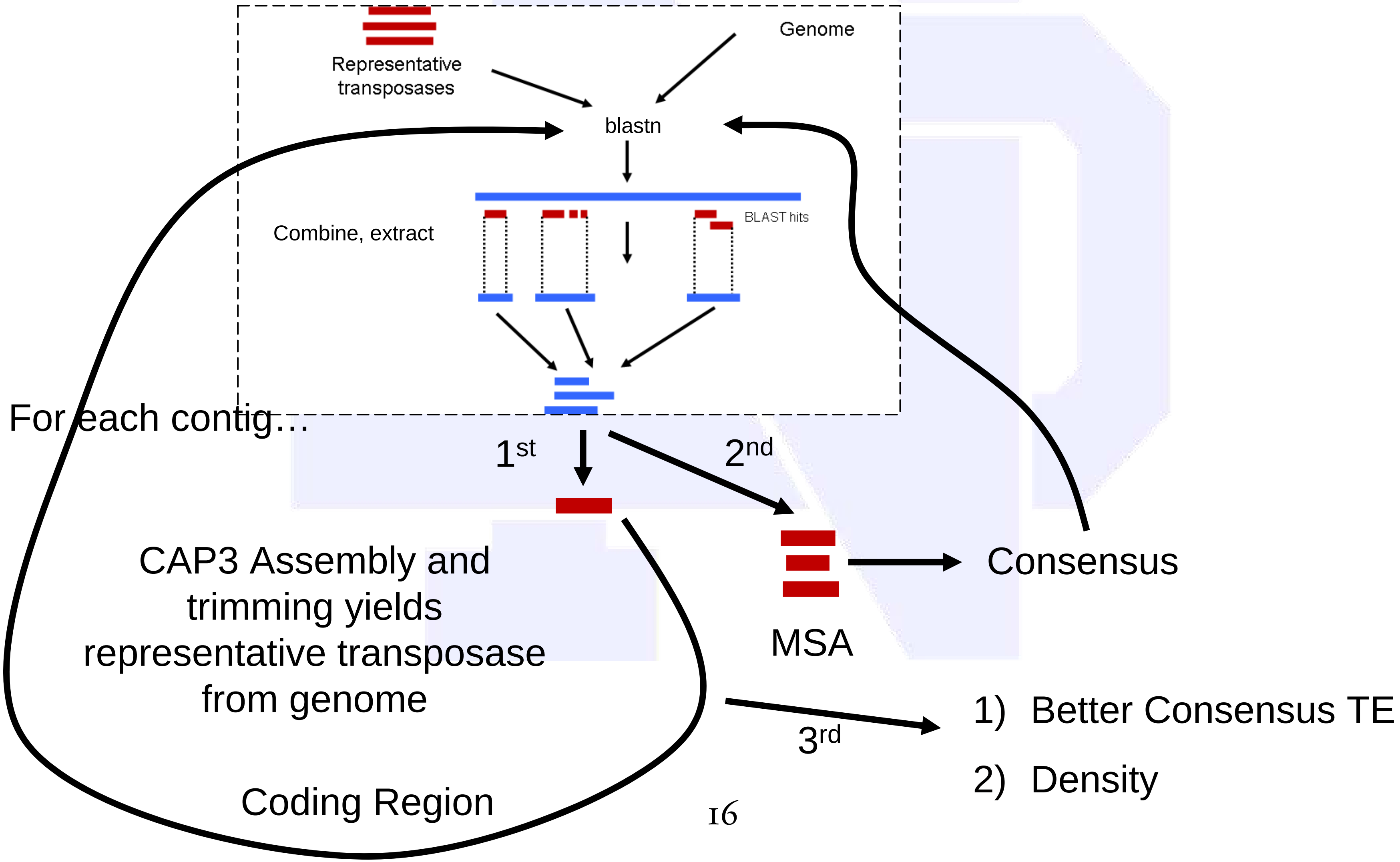
- Homology-based
- Replace DNASTAR SeqMan II and manual analysis with other tools
  - CAP3, Clustal, various scripts
- Iterative- repeat steps if necessary

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



For each contig…

CAP3 Assembly and trimming yields representative transposase from genome

Coding Region

Consensus

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

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

↓

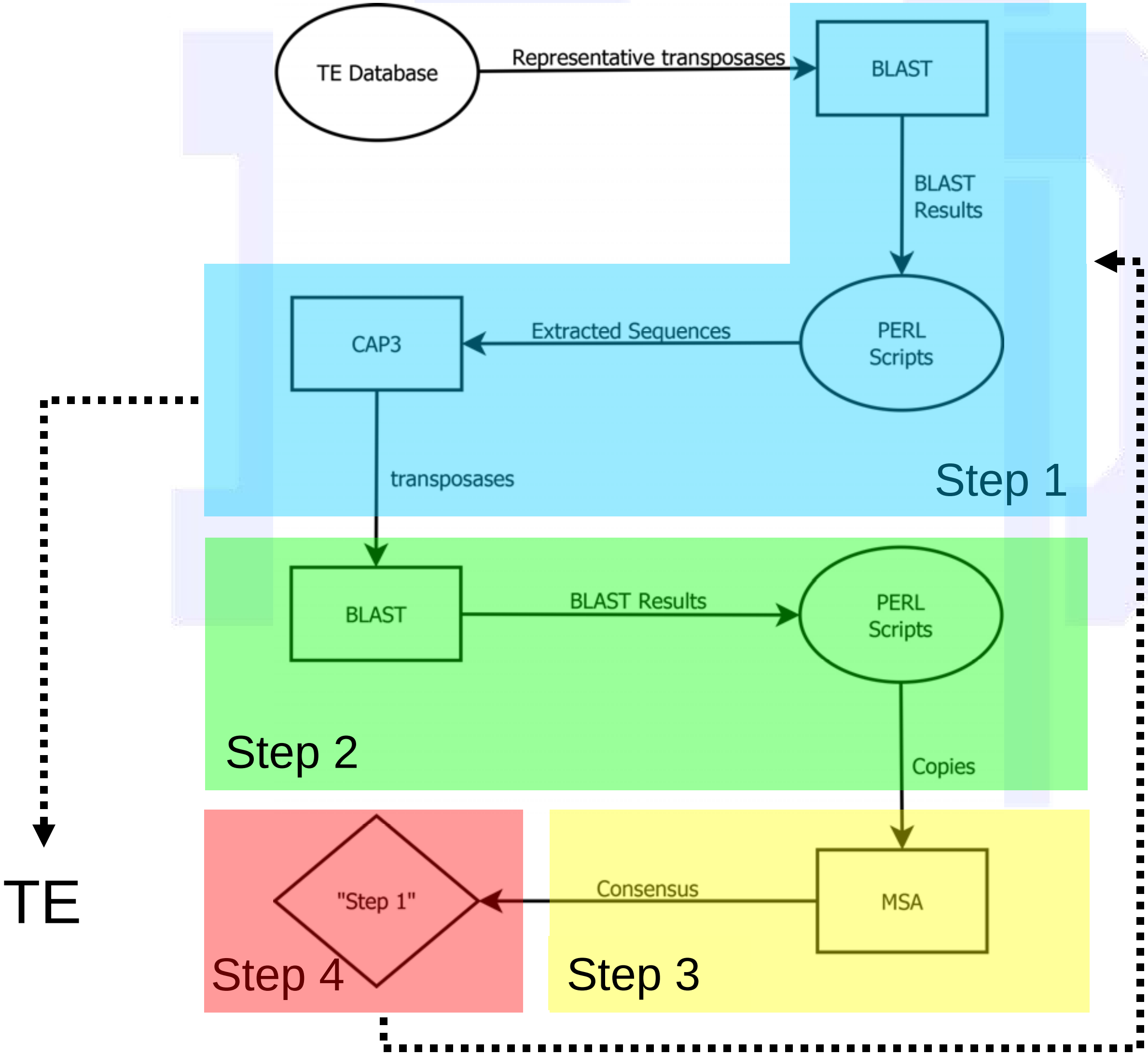
Putative Consensus

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 → Density

Automated Approach Schematic



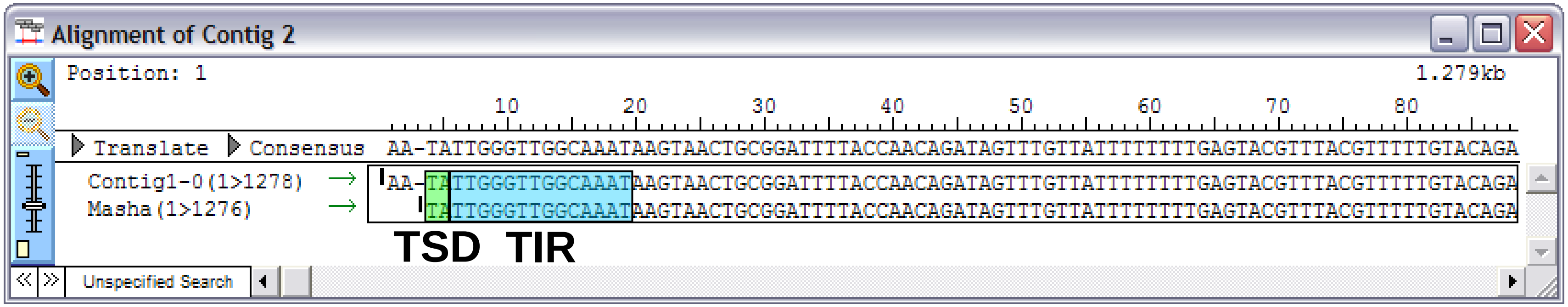
UNIVERSITY OF NOTRE DAME

Validation Strategy

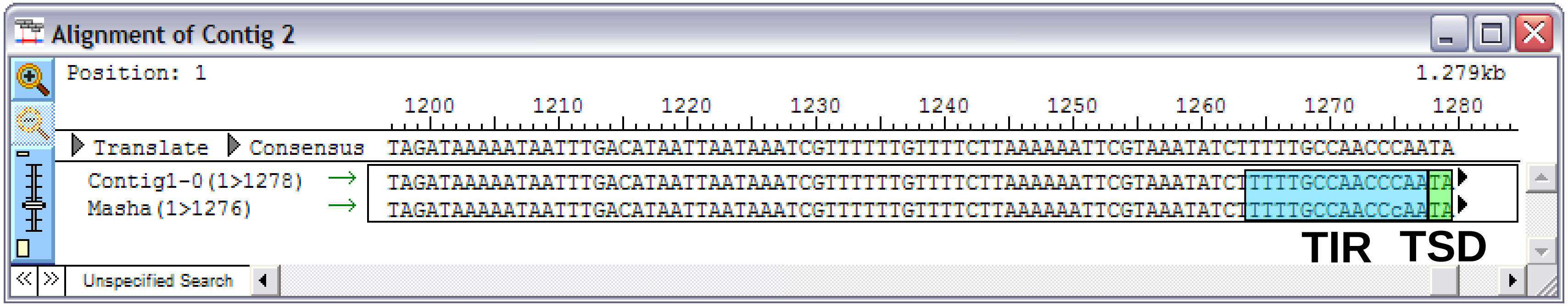
1. Initially evaluated automated approach on *P. humanus humanus* and *C. quinquefasciatus*
   - Validate against high-quality manually verified annotation
   - Identify default starting parameters
2. Check automated results versus published results
3. Genomes in General:
   - Translate consensus TE 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



```
Position: 1                                                                                  1.279kb
                                 10        20        30        40        50        60        70        80
Translate  Consensus   AA-TATTGGGTTGGCAAATAAGTAACTGCGGATTTTACCAACAGATAGTTTGTTATTTTTTTTGAGTACGTTTACGTTTTTGTACAGA
Contig1-0(1>1278)  →  AA-TATTGGGTTGGCAAATAAGTAACTGCGGATTTTACCAACAGATAGTTTGTTATTTTTTTTGAGTACGTTTACGTTTTTGTACAGA
Masha(1>1276)      →     TATTGGGTTGGCAAATAAGTAACTGCGGATTTTACCAACAGATAGTTTGTTATTTTTTTTGAGTACGTTTACGTTTTTGTACAGA
```



```
Position: 1                                                                                  1.279kb
                       1200      1210      1220      1230      1240      1250      1260      1270      1280
Translate  Consensus   TAGATAAAAATAATTTGACATAATTAATAAATCGTTTTTTGTTTTCTTAAAAAATTCGTAAATATCTTTTTGCCAACCCAATA
Contig1-0(1>1278)  →  TAGATAAAAATAATTTGACATAATTAATAAATCGTTTTTTGTTTTCTTAAAAAATTCGTAAATATCTTTTTGCCAACCCAATA
Masha(1>1276)      →  TAGATAAAAATAATTTGACATAATTAATAAATCGTTTTTTGTTTTCTTAAAAAATTCGTAAATATCTTTTTGCCAACCcAATA
```

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

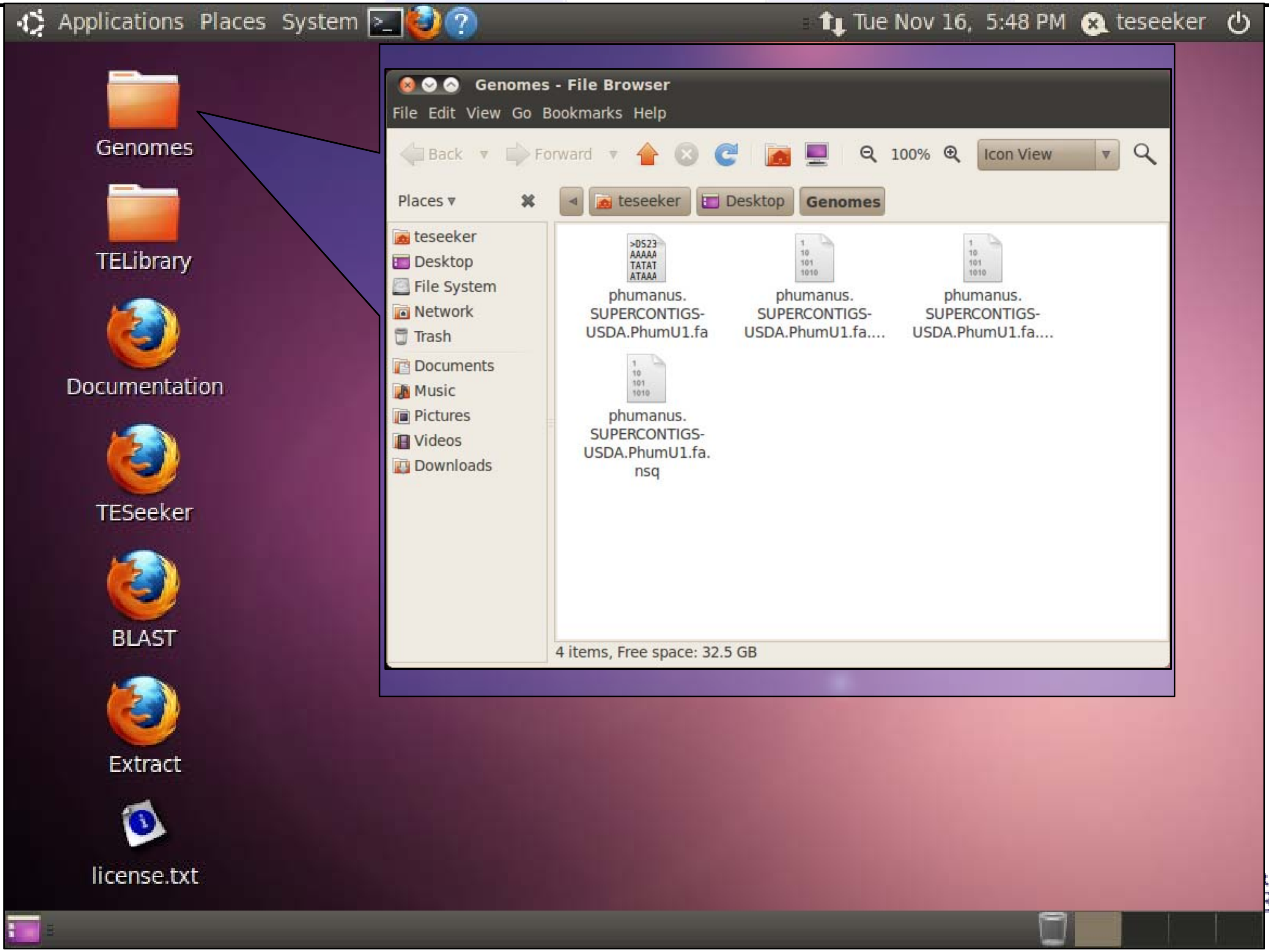
Validation (3)

- 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

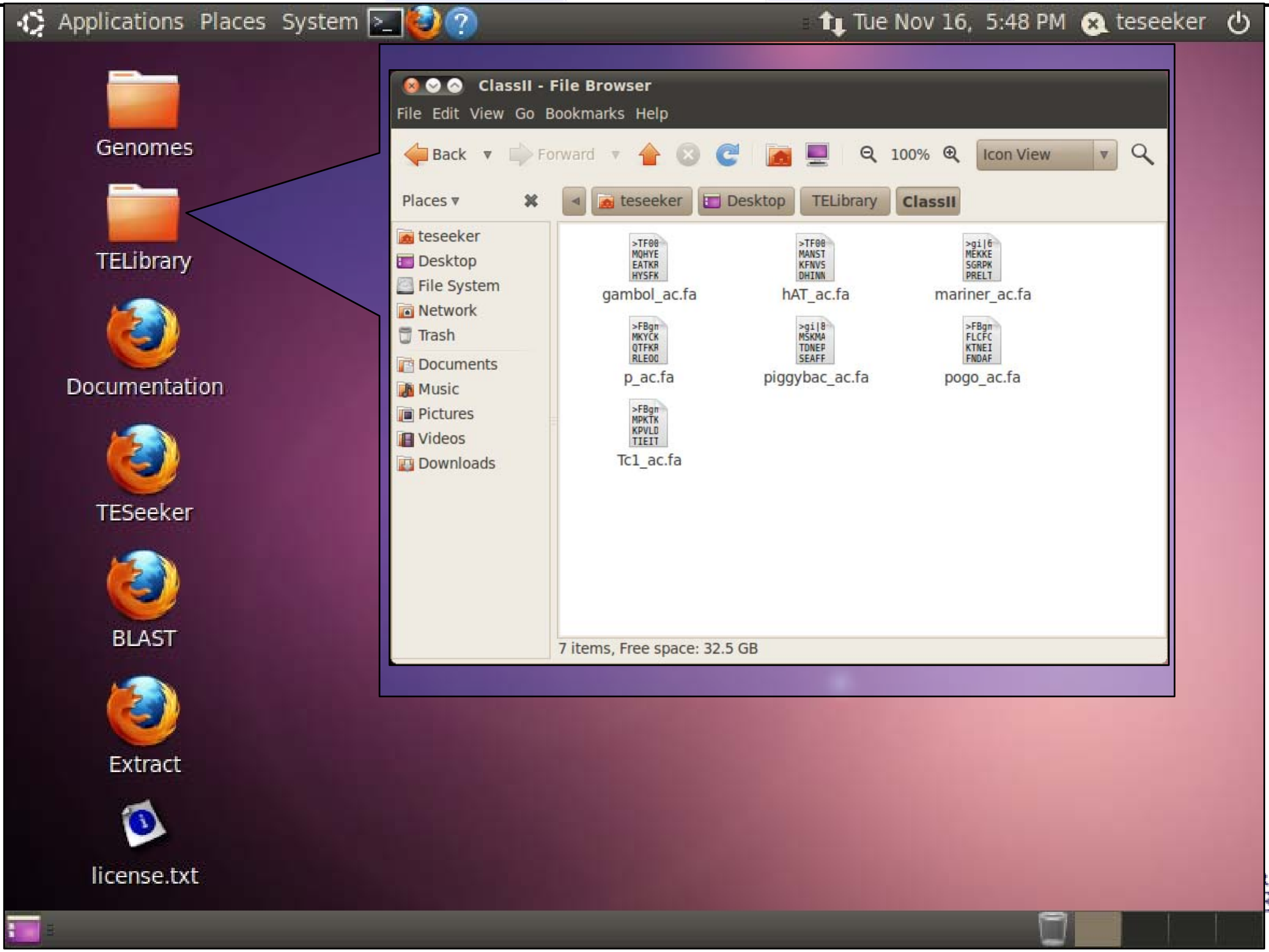
- Approach implemented as TESeeker
  - VirtualBox virtual appliance
    - Cross-platform
    - Completely configured, no need to install scripts
      - Provide only genome FASTA file
      - Optionally provide additional library files
    - Local web interface
- http://www.nd.edu/~teseeker
  - Virtual appliance
  - Documentation
  - TE Library

TESeeker Desktop

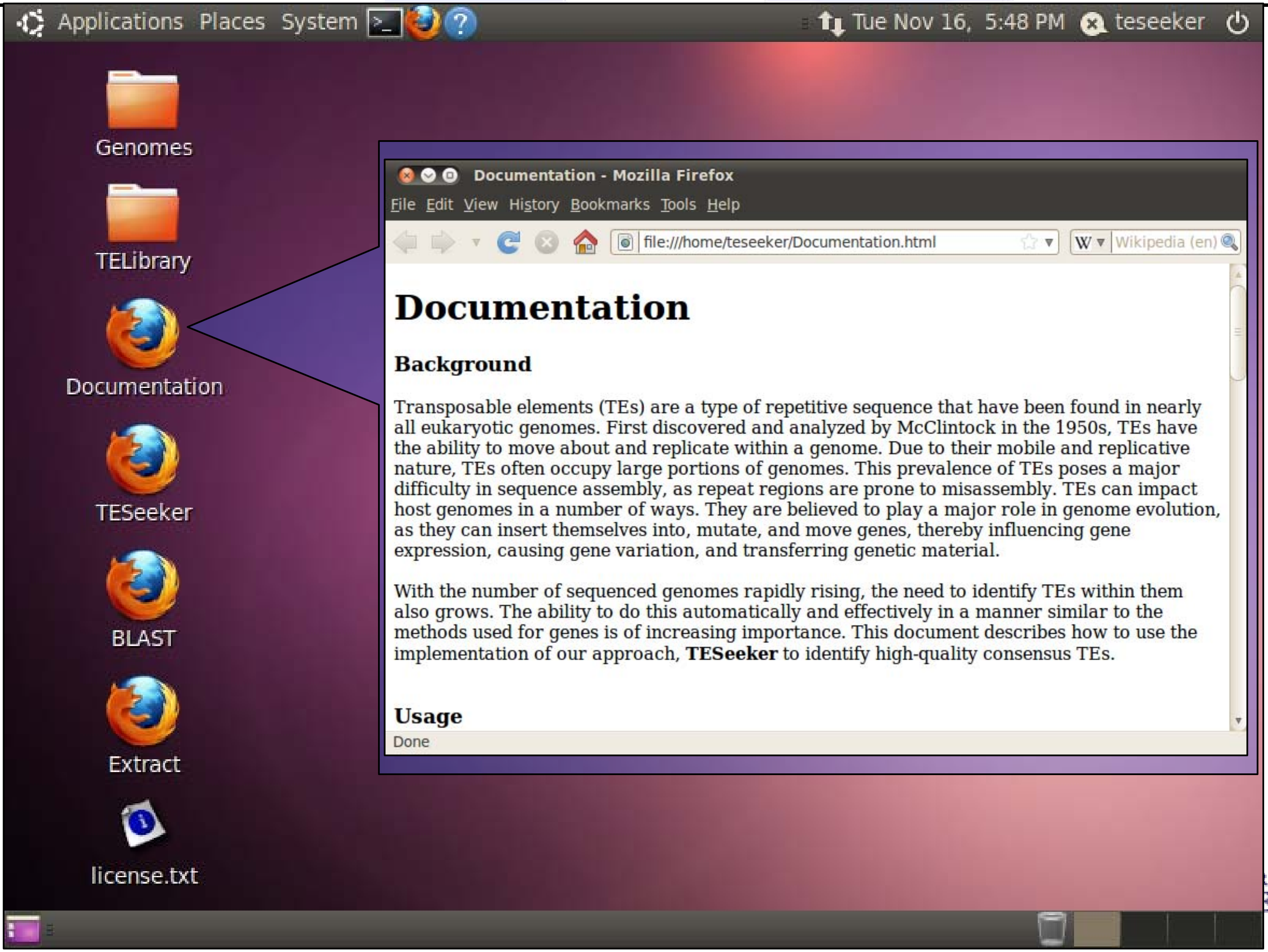


Applications Places System
Tue Nov 16, 5:48 PM
teseeker
Genomes
TELibrary
Documentation
TESeeker
BLAST
Extract
license.txt
Genomes - File Browser
File Edit View Go Bookmarks Help
Back
Forward
100%
Icon View
Places
teseeker
Desktop
Genomes
teseeker
Desktop
File System
Network
Trash
Documents
Music
Pictures
Videos
Downloads
phumanus.SUPERCONTIGS-USDA.PhumU1.fa
phumanus.SUPERCONTIGS-USDA.PhumU1.fa....
phumanus.SUPERCONTIGS-USDA.PhumU1.fa....
phumanus.SUPERCONTIGS-USDA.PhumU1.fa.nsq
4 items, Free space: 32.5 GB

TESeeker Desktop



TESeeker Desktop



Documentation

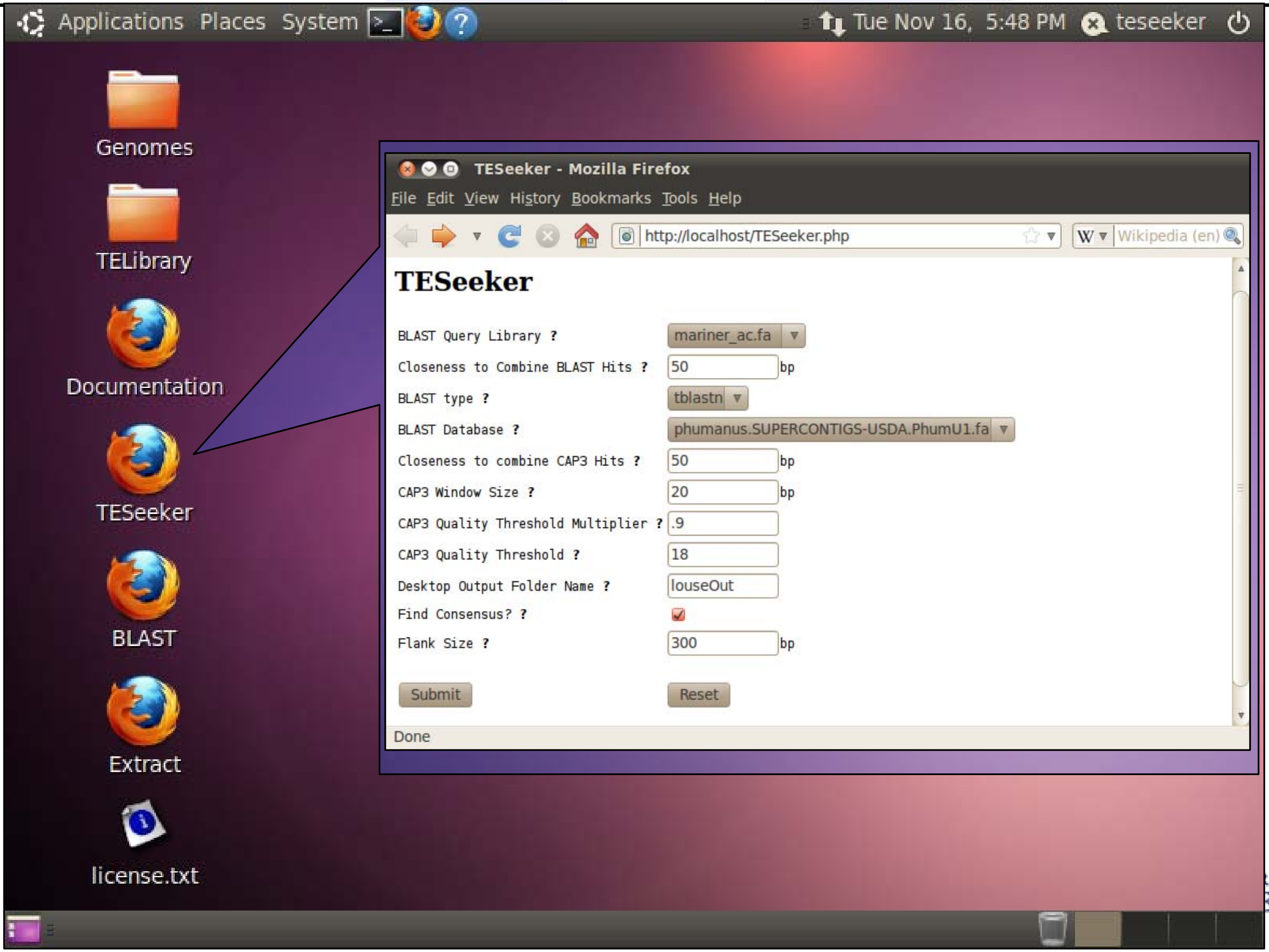
Background

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, **TESeeker** to identify high-quality consensus TEs.

Usage

TESeeker Desktop



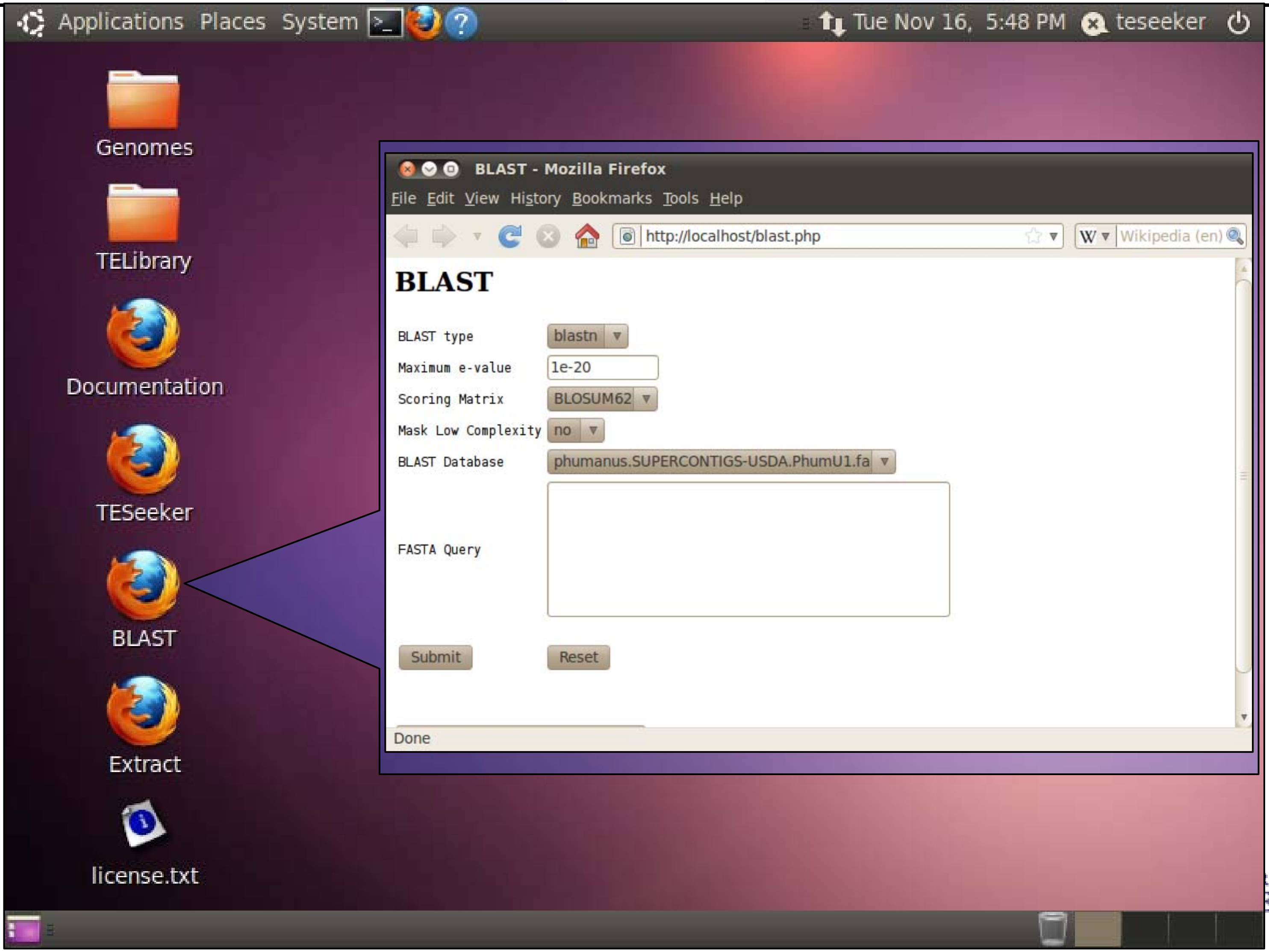
Applications Places System
Tue Nov 16, 5:48 PM
teseeker
Genomes
TELibrary
Documentation
TESeeker
BLAST
Extract
license.txt
TESeeker - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost/TESeeker.php
Wikipedia (en)
Done

TESeeker

BLAST Query Library ?	mariner_ac.fa
Closeness to Combine BLAST Hits ?	50 bp
BLAST type ?	tblastn
BLAST Database ?	phumanus.SUPERCONTIGS-USDA.PhumU1.fa
Closeness to combine CAP3 Hits ?	50 bp
CAP3 Window Size ?	20 bp
CAP3 Quality Threshold Multiplier ?	.9
CAP3 Quality Threshold ?	18
Desktop Output Folder Name ?	louseOut
Find Consensus? ?	☑
Flank Size ?	300 bp

Submit Reset

TESeeker Desktop



Applications Places System
Tue Nov 16, 5:48 PM teseeker
Genomes
TELibrary
Documentation
TESeeker
BLAST
Extract
license.txt
BLAST - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost/blast.php
Wikipedia (en)
Done

BLAST

BLAST type: blastn

Maximum e-value: 1e-20

Scoring Matrix: BLOSUM62

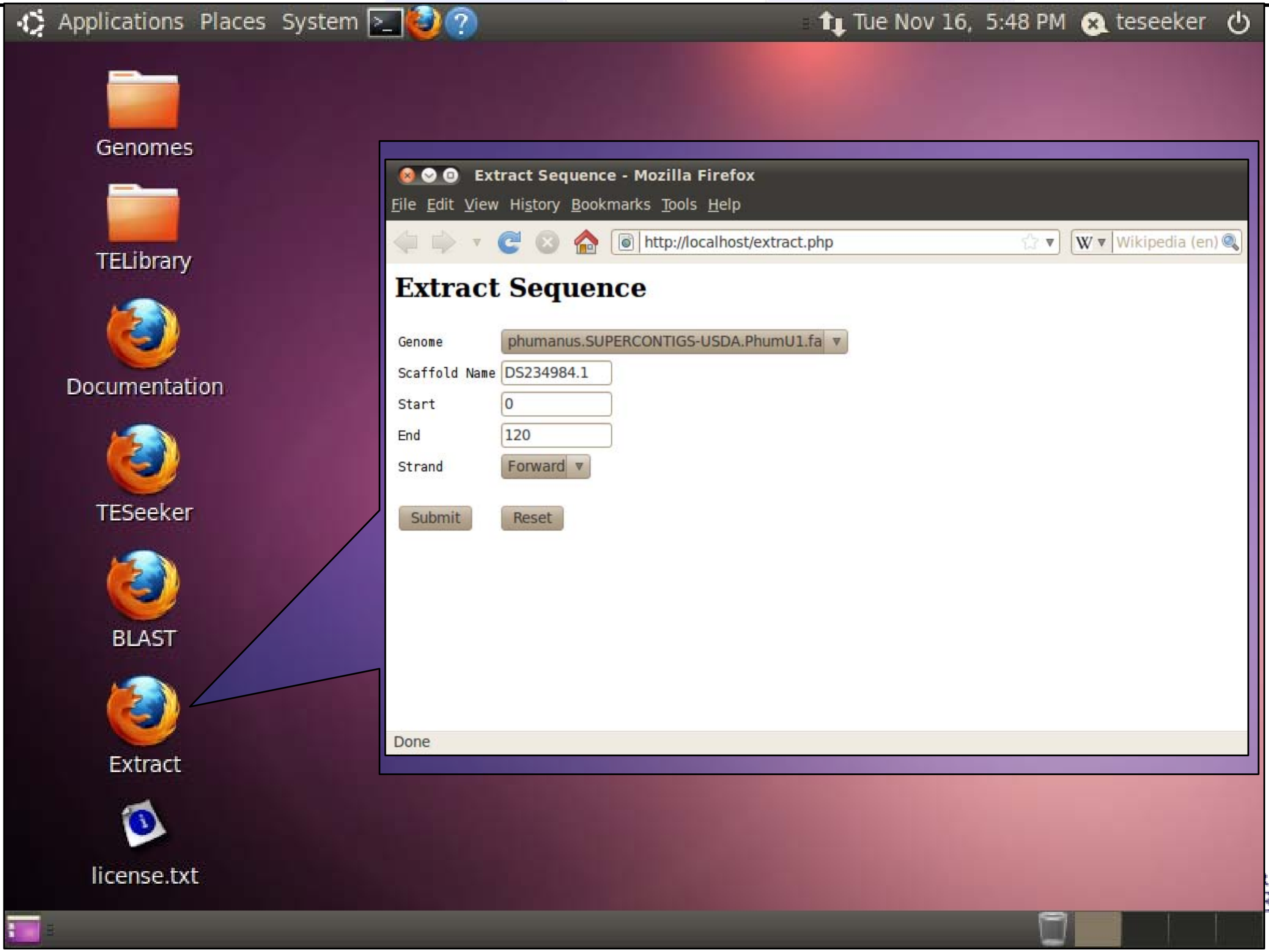
Mask Low Complexity: no

BLAST Database: phumanus.SUPERCONTIGS-USDA.PhumU1.fa

FASTA Query:

Submit Reset

TESeeker Desktop



Extract Sequence

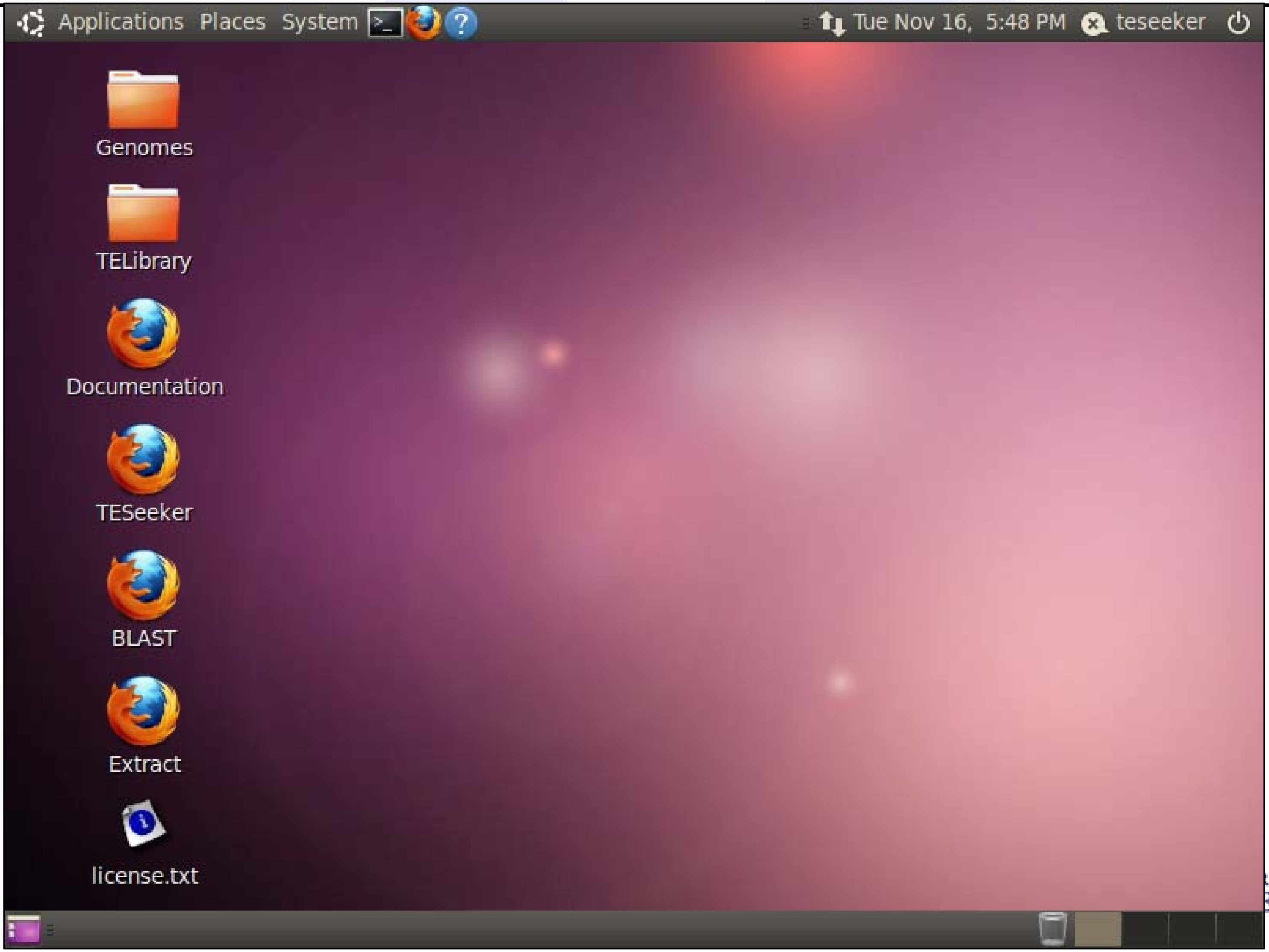
Genome	phumanus.SUPERCONTIGS-USDA.PhumU1.fa
Scaffold Name	DS234984.1
Start	0
End	120
Strand	Forward

SITY OF
DAME

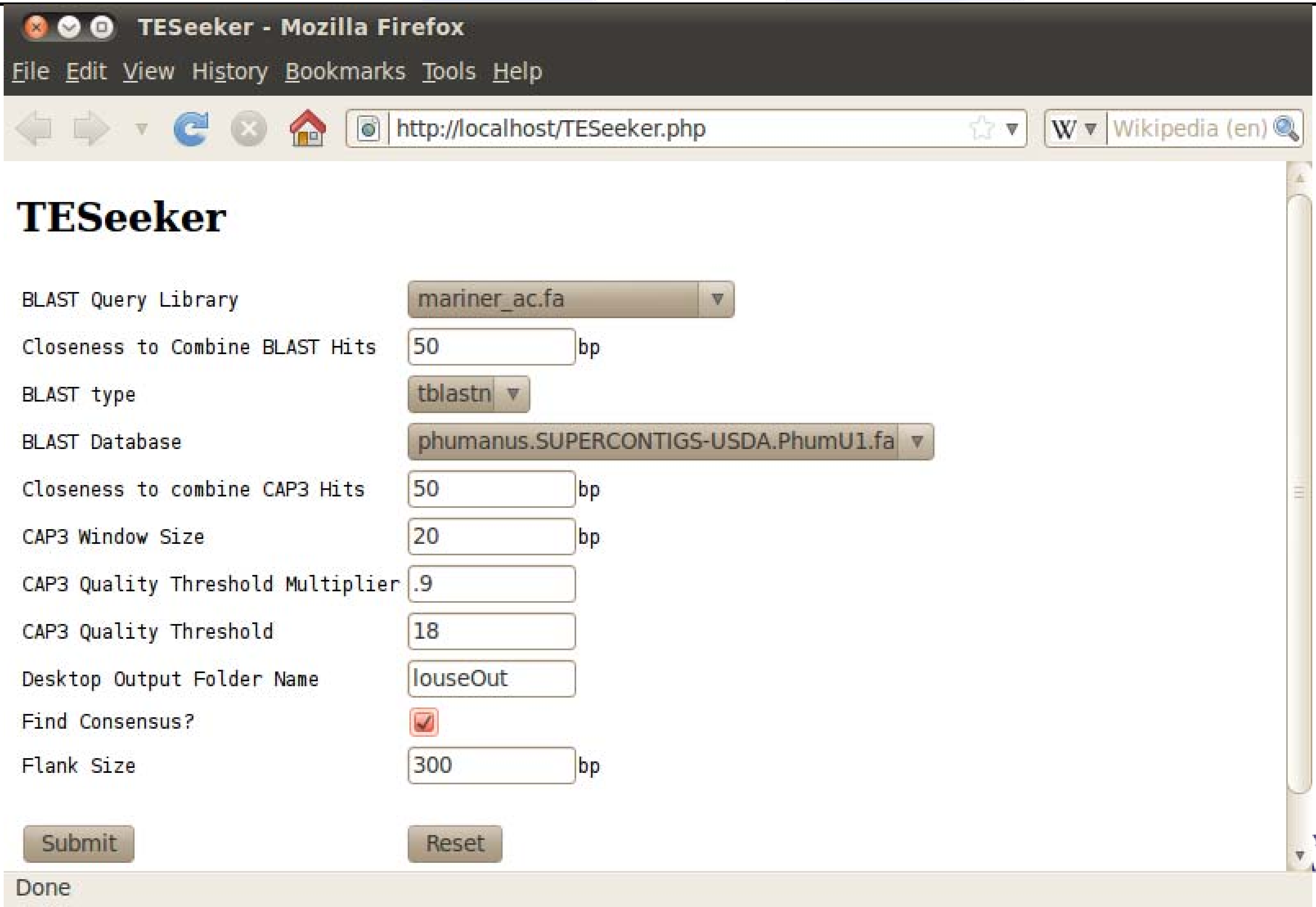
TESeeker Walkthrough

- Identify *mariner* element in *P. humanus humanus:*
  - Start with default parameters
  - Make sure genome file and library file are present
  - Start search

TESeeker Desktop



TESeeker Local Web Interface

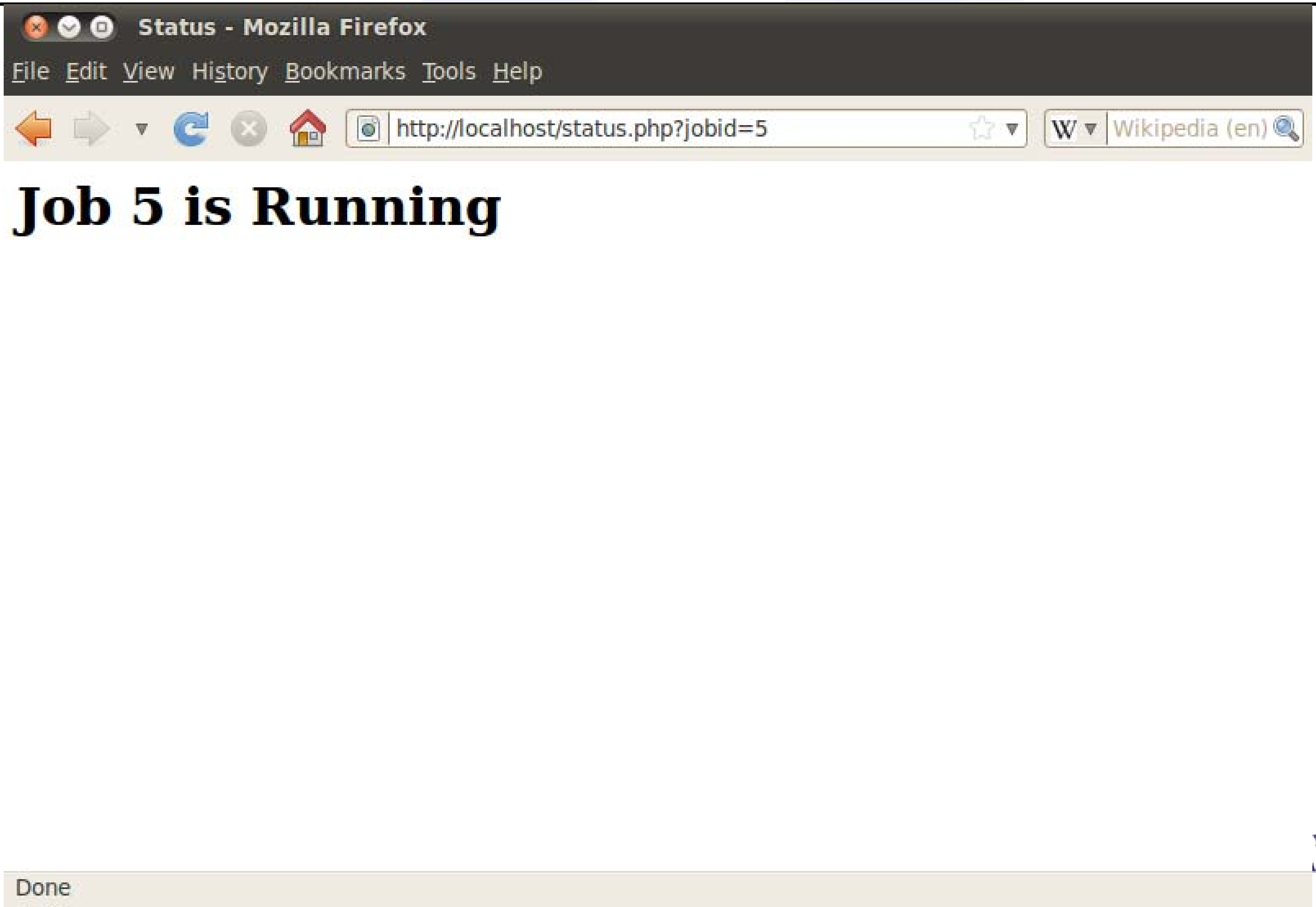


TESeeker - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost/TESeeker.php
Wikipedia (en)
Submit
Reset
Done

TESeeker

BLAST Query Library	mariner_ac.fa ▾	
Closeness to Combine BLAST Hits	50	bp
BLAST type	tblastn ▾	
BLAST Database	phumanus.SUPERCONTIGS-USDA.PhumU1.fa ▾	
Closeness to combine CAP3 Hits	50	bp
CAP3 Window Size	20	bp
CAP3 Quality Threshold Multiplier	.9	
CAP3 Quality Threshold	18	
Desktop Output Folder Name	louseOut	
Find Consensus?	☑	
Flank Size	300	bp

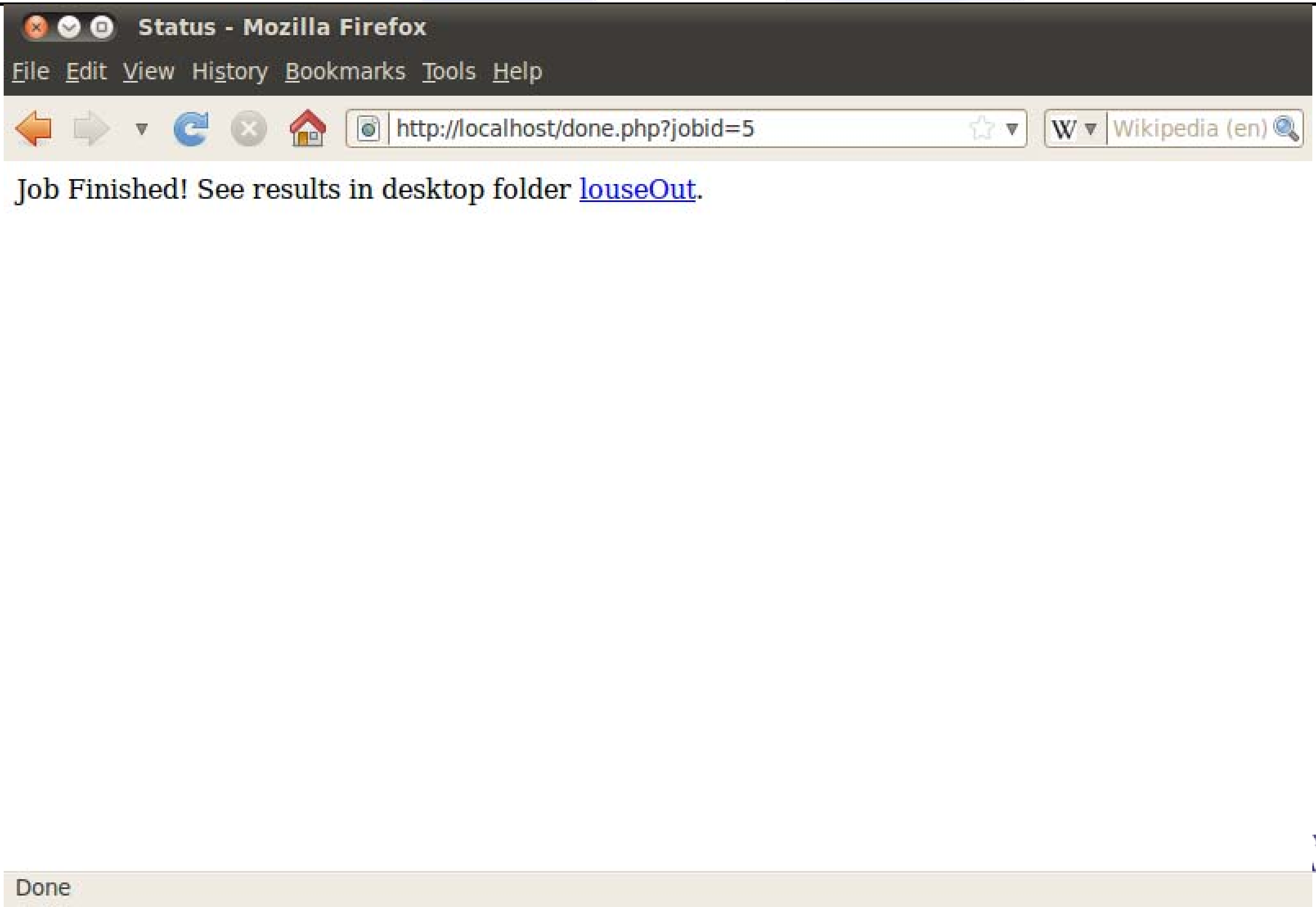
TESeeker Status



Status - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost/status.php?jobid=5
Wikipedia (en)

Job 5 is Running

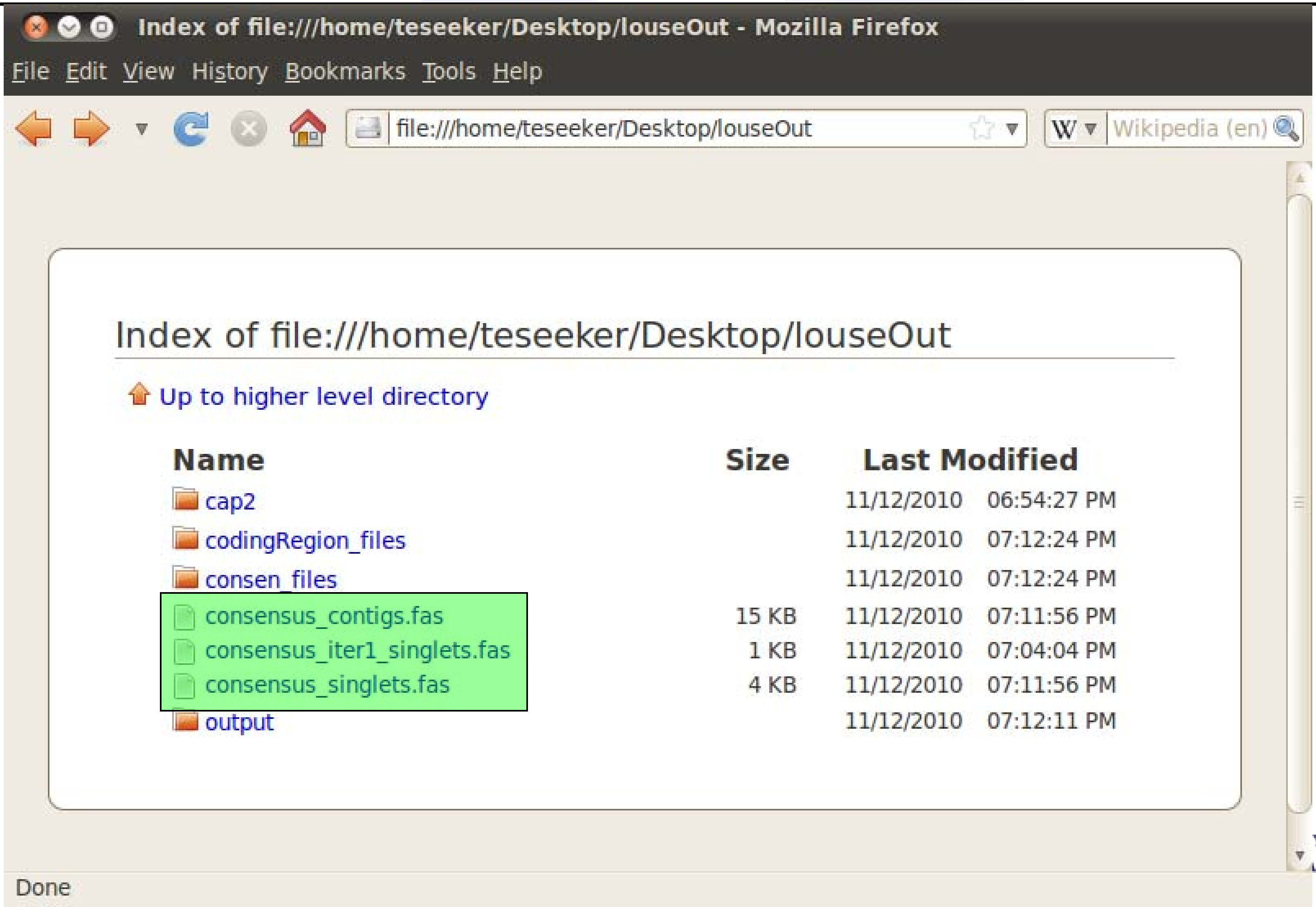
TESeeker Status



Status - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost/done.php?jobid=5
Wikipedia (en)
Done

Job Finished! See results in desktop folder louseOut.

TESeeker Results



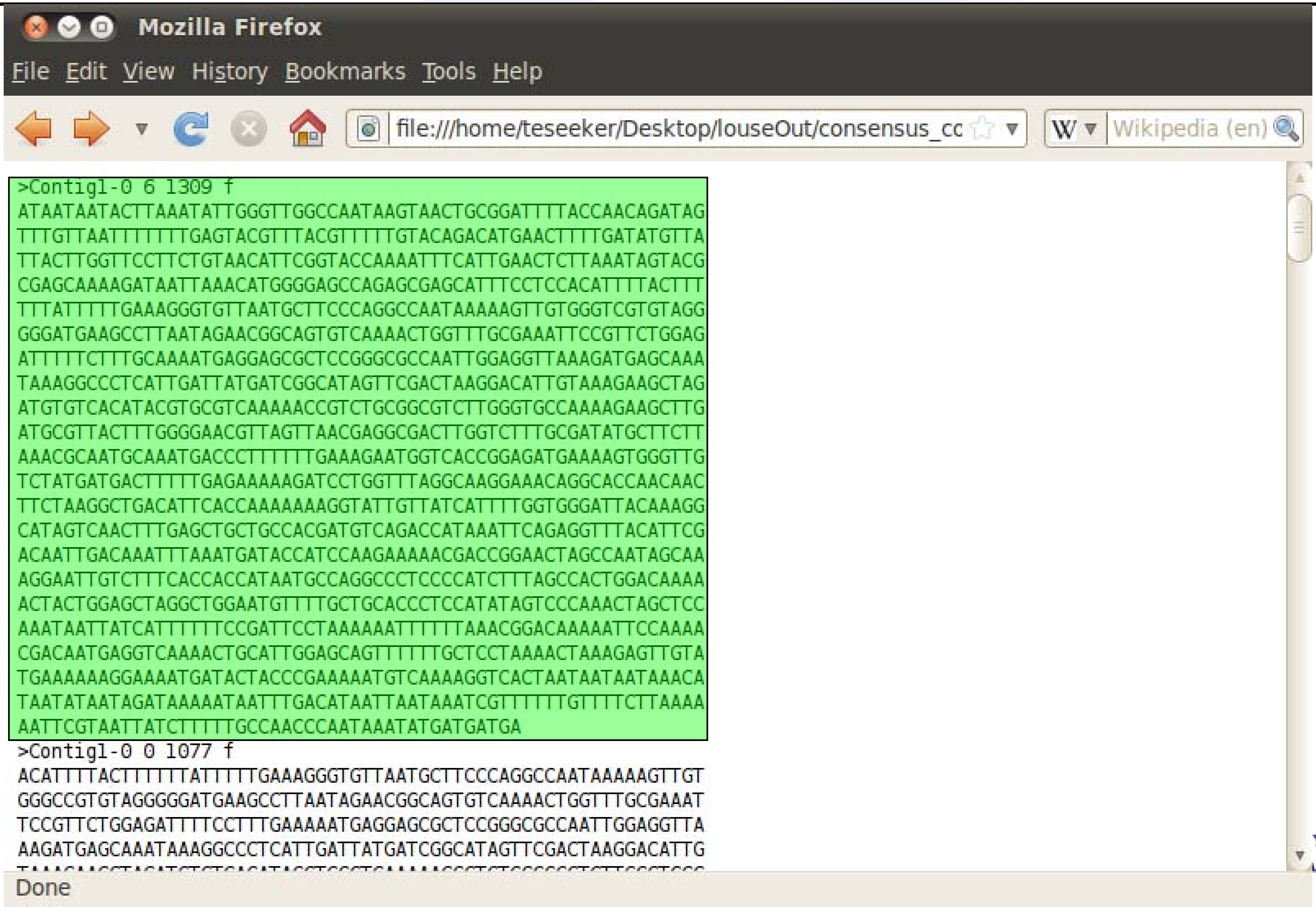
Index of file:///home/teseeker/Desktop/louseOut - Mozilla Firefox
File Edit View History Bookmarks Tools Help
file:///home/teseeker/Desktop/louseOut
Wikipedia (en)
Done

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

Up to higher level directory

Name	Size	Last Modified	
cap2		11/12/2010	06:54:27 PM
codingRegion_files		11/12/2010	07:12:24 PM
consen_files		11/12/2010	07:12:24 PM
consensus_contigs.fas	15 KB	11/12/2010	07:11:56 PM
consensus_iter1_singlets.fas	1 KB	11/12/2010	07:04:04 PM
consensus_singlets.fas	4 KB	11/12/2010	07:11:56 PM
output		11/12/2010	07:12:11 PM

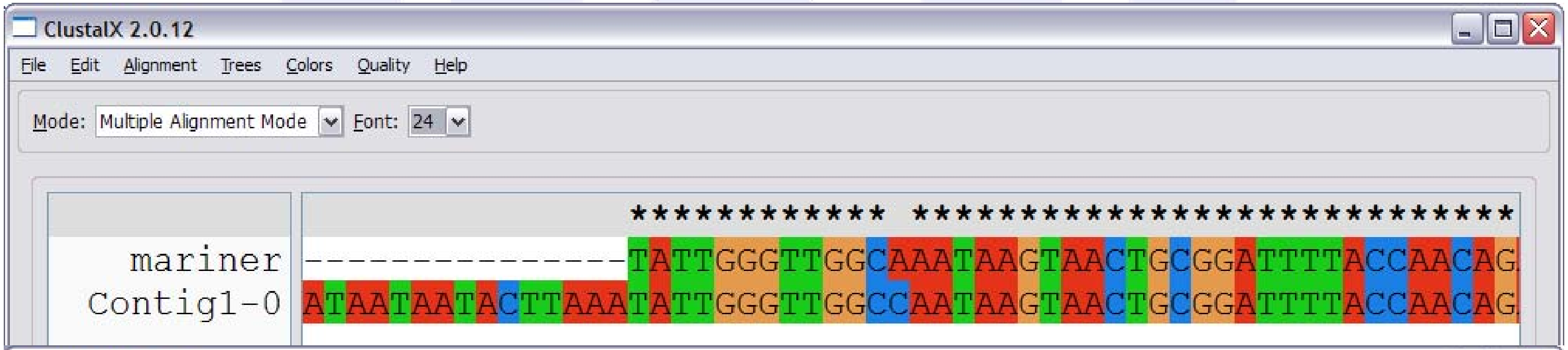
TESeeker Results


Mozilla Firefox
File Edit View History Bookmarks Tools Help
file:///home/teseeker/Desktop/louseOut/consensus_cc
Wikipedia (en)
Done

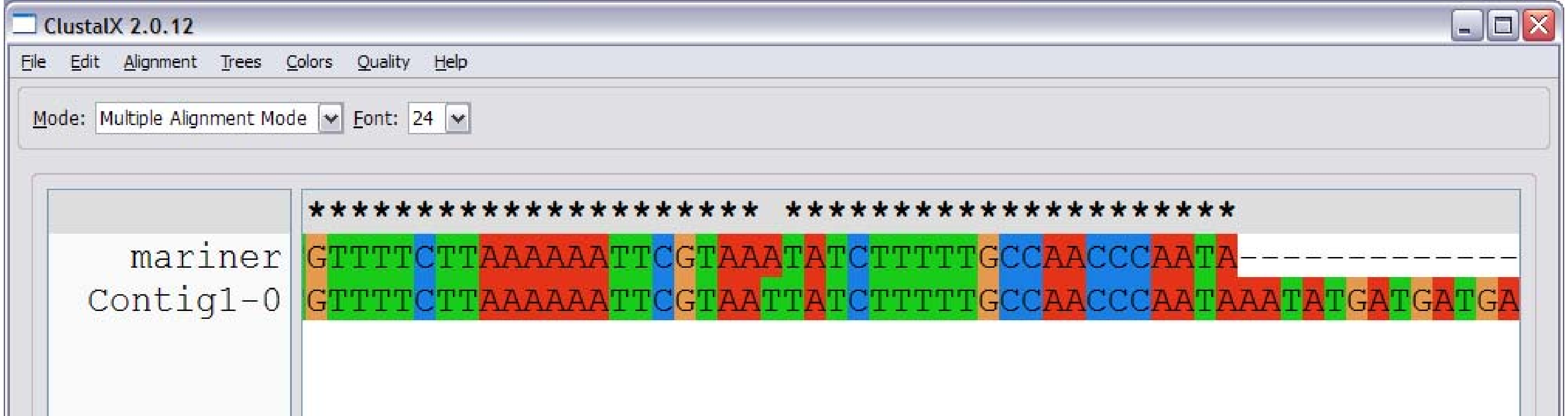
```
>Contig1-0 6 1309 f
ATAATAATACTTAAATATTGGGTTGGCCAATAAGTAACTGCGGATTTTACCAACAGATAG
TTTGTTAATTTTTTTGAGTACGTTTACGTTTTTGTACAGACATGAACTTTTGATATGTTA
TTACTTGGTTCCTTCTGTAACATTCGGTACCAAAATTTCATTGAACTCTTAAATAGTACG
CGAGCAAAAGATAATTAAACATGGGGAGCCAGAGCGAGCATTTCCTCCACATTTTACTTT
TTTATTTTTGAAAGGGTGTTAATGCTTCCCAGGCCAATAAAAAGTTGTGGGTCGTGTAGG
GGGATGAAGCCTTAATAGAACGGCAGTGTCAAAACTGGTTTGCGAAATTCCGTTCTGGAG
ATTTTTCTTTGCAAAATGAGGAGCGCTCCGGGCGCCAATTGGAGGTTAAAGATGAGCAAA
TAAAGGCCCTCATTGATTATGATCGGCATAGTTCGACTAAGGACATTGTAAAGAAGCTAG
ATGTGTCACATACGTGCGTCAAAAACCGTCTGCGGCGTCTTGGGTGCCAAAAGAAGCTTG
ATGCGTTACTTTGGGGAACGTTAGTTAACGAGGCGACTTGGTCTTTGCGATATGCTTCTT
AAACGCAATGCAAATGACCCTTTTTTGAAAGAATGGTCACCGGAGATGAAAAGTGGGTTG
TCTATGATGACTTTTTGAGAAAAAGATCCTGGTTTAGGCAAGGAAACAGGCACCAACAAC
TTCTAAGGCTGACATTCACCAAAAAAAGGTATTGTTATCATTTTGGTGGGATTACAAAGG
CATAGTCAACTTTGAGCTGCTGCCACGATGTCAGACCATAAATTCAGAGGTTTACATTCG
ACAATTGACAAATTTAAATGATACCATCCAAGAAAAACGACCGGAACTAGCCAATAGCAA
AGGAATTGTCTTTCACCACCATAATGCCAGGCCCTCCCCATCTTTAGCCACTGGACAAAA
ACTACTGGAGCTAGGCTGGAATGTTTTGCTGCACCCTCCATATAGTCCCAAACTAGCTCC
AAATAATTATCATTTTTTCCGATTCCTAAAAAATTTTTTAAACGGACAAAAATTCCAAAA
CGACAATGAGGTCAAAACTGCATTGGAGCAGTTTTTTGCTCCTAAAACTAAAGAGTTGTA
TGAAAAAAGGAAAATGATACTACCCGAAAAATGTCAAAAGGTCACTAATAATAATAAACA
TAATATAATAGATAAAAATAATTTGACATAATTAATAAATCGTTTTTTGTTTTCTTAAAA
AATTCGTAATTATCTTTTTGCCAACCCAATAAATATGATGATGA
>Contig1-0 0 1077 f
ACATTTTACTTTTTTATTTTTGAAAGGGTGTTAATGCTTCCCAGGCCAATAAAAAGTTGT
GGGCCGTGTAGGGGGATGAAGCCTTAATAGAACGGCAGTGTCAAAACTGGTTTGCGAAAT
TCCGTTCTGGAGATTTTCCTTTGAAAAATGAGGAGCGCTCCGGGCGCCAATTGGAGGTTA
AAGATGAGCAAATAAAGGCCCTCATTGATTATGATCGGCATAGTTCGACTAAGGACATTG
```

TESeeker Results

- Alignment
  - TESeeker top result with default parameters
  - 99% identity with manually annotated *mariner*



```
                ************ ****************************
   mariner      ---------------TATTGGGTTGGCAAATAAGTAACTGCGGATTTTACCAACAG
 Contig1-0      ATAATAATACTTAAATATTGGGTTGGCCAATAAGTAACTGCGGATTTTACCAACAG
```



```
                ********************* *********************
   mariner      GTTTTCTTAAAAAATTCGTAAATATCTTTTTGCCAACCCAATA-------------
 Contig1-0      GTTTTCTTAAAAAATTCGTAATTATCTTTTTGCCAACCCAATAAATATGATGATGA
```

TE Identification Summary

- 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?

UNIVERSITY OF
NOTRE DAME