



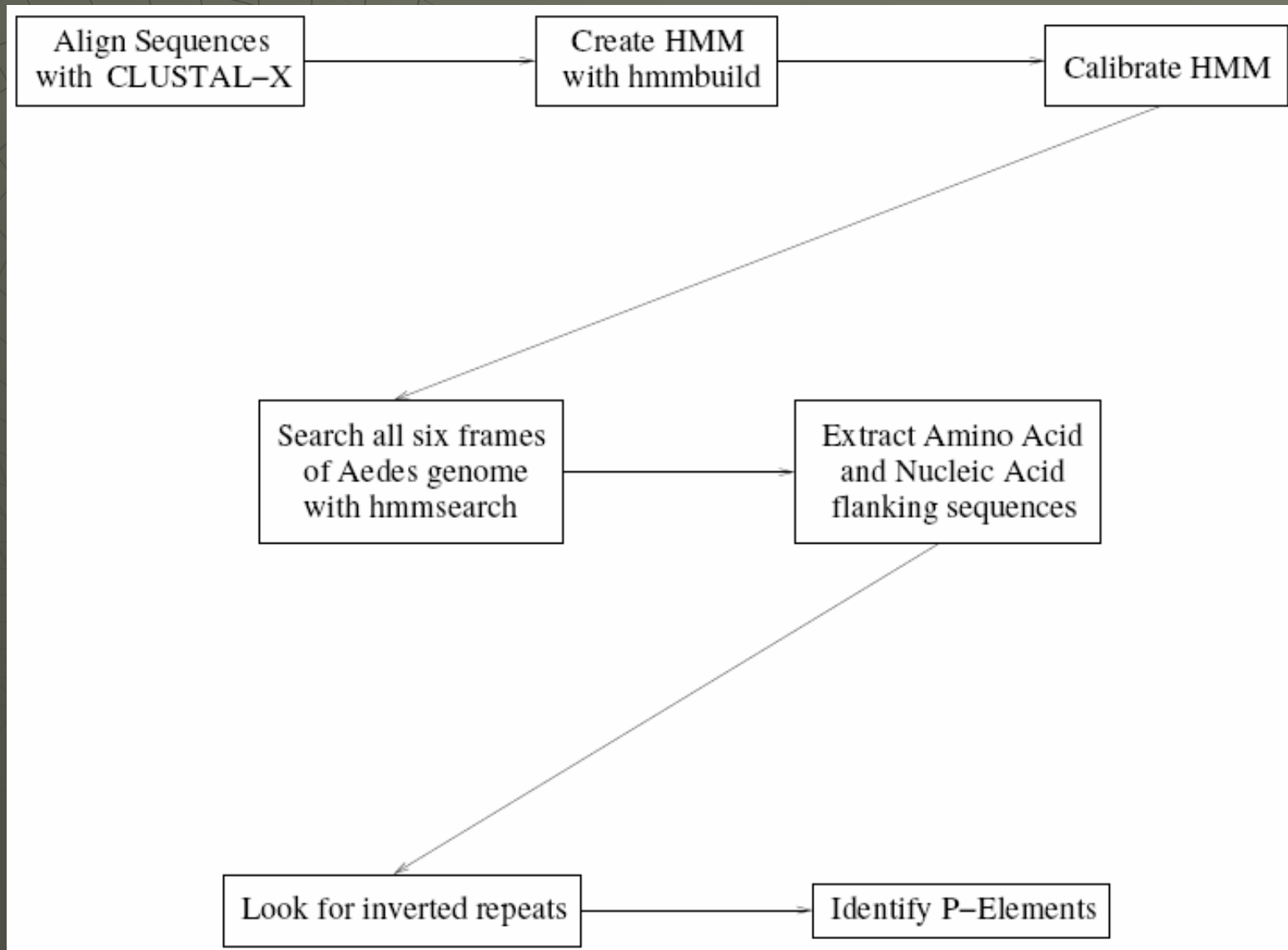
# Analysis of $P$ element sequences in *Aedes aegypti*

Trevor Cickovski

Jim Hogan

Ryan Kennedy

# Approach



# Align Sequences with *Clustal-X*

CLUSTAL X (1.81) multiple sequence alignment

```
gi|58386693|ref|XP_314978.2| -----
gi|55239599|gb|EAA10370.3| -----
gi|58381516|ref|XP_311297.2| -----
gi|55243097|gb|EAA06886.2| -----
gi|58388308|ref|XP_316192.2| -----
gi|55238944|gb|EAA44169.2| -----
...
gi|28630150|gb|AAM94946.1| -----
gi|23664272|gb|AAN39288.1| -----
gi|38564327|ref|NP_078948.2| -----
gi|10439964|dbj|BAB15609.1| -----
gi|57109260|ref|XP_544956.1| MTHNSEIQIQTKNHKLNKVNQSDPEARVLP SASGPAPLCAQGCECEGLLT
gi|50746565|ref|XP_420555.1| -----
```

Alignment of amino acid sequences

# HMM Process

- ◆ Used 6 genome frames from course directory
- ◆ Built 21 different HMM Models (*hmmbuild*)
  - One containing all sequences
  - The rest containing subsets of sequences
- ◆ *hmmcalibrate* – scores sequences
- ◆ Ran *hmmsearch* on input genome files in FASTA format
- ◆ Extracted relevant sequences from results, including 5kb flanking regions

# *hmmsearch* Output

Sequence	Description	Score	E-value	N
Aedes-0F	aegypti supercontig 1.152	192.0	7.5e-55	1
Aedes-0F	aegypti supercontig 1.309	166.2	4.4e-47	1
Aedes-0F	aegypti supercontig 1.723	148.1	1.2e-41	1
Aedes-0F	aegypti supercontig 1.1056	145.4	7.8e-41	1
Aedes-0F	aegypti supercontig 1.421	53.4	4.1e-13	1
Aedes-0F	aegypti supercontig 1.98	49.7	5.1e-12	1

Used this output to extract specific regions of the sequence

# Inverted Repeats

- ◆ *Blast2seq*
  - *Blast* alignment of a sequence against self
- ◆ Designate the boundaries of a given  $P$  element
- ◆ Important component for  $P$  elements
  - Transposase recognizes it, enabling it to transpose

# Hits with Inverted Repeats

Supercontig	Evalue	Genomic Hits	Inverted Repeats	Hit
1.1056	e-41	6	cagcgacattcgctctctatagttattacctctatt	PREDICTED: similar to THAP domain containing 9 [Danio rerio]
1.604	e-24	137	gacgtaacaagtgaaaaaacgtaaaatttaggtgatttacacatttttaccacaactt (etc)	ENSANGP0000005847 [Anopheles gambiae str. PEST]
1.893	e-18	119	catgaaaacgacttgtttaattcacgcaaggccctt aacaataaaatcagcaacaactg	hypothetical protein LOC54875 [Homo sapiens]
1.327	e-15	126	cacgtggtatatggacggtcct	No significant similarity found
1.421	e-13	9	aagtggtagagcccgcgctcacagcaaagc catattaaagggtgctggt (etc)	PREDICTED: similar to THAP domain containing 9 [Strongylocentrotus purpuratus]
1.639	e-09	218	tcttcgcgcaactctccattcatagactctg	PREDICTED: similar to THAP domain containing 4 [Strongylocentrotus purpuratus]
1.458	e-09	2	ccgcggtacaaagcaaagccatgctgaagtgtctgggttcgagtcceggctgctccag (etc)	PREDICTED: similar to THAP domain containing 4 [Strongylocentrotus purpuratus]
1.5	e-06	386	acatacagcttctctctcttcgggtgattgatgctctgaaaata (etc)	reverse transcriptase-like protein
1.927	e-06	150	gagtcggtcattcttttcttctaaccatcttg atgcat	transposase [Drosophila melanogaster]
1.266	e-05	4	gtccacaaattacgtaacgctttaagggaagggggtaggctcaaacattacgactcata	PREDICTED: similar to THAP domain containing 9 [Strongylocentrotus purpuratus]

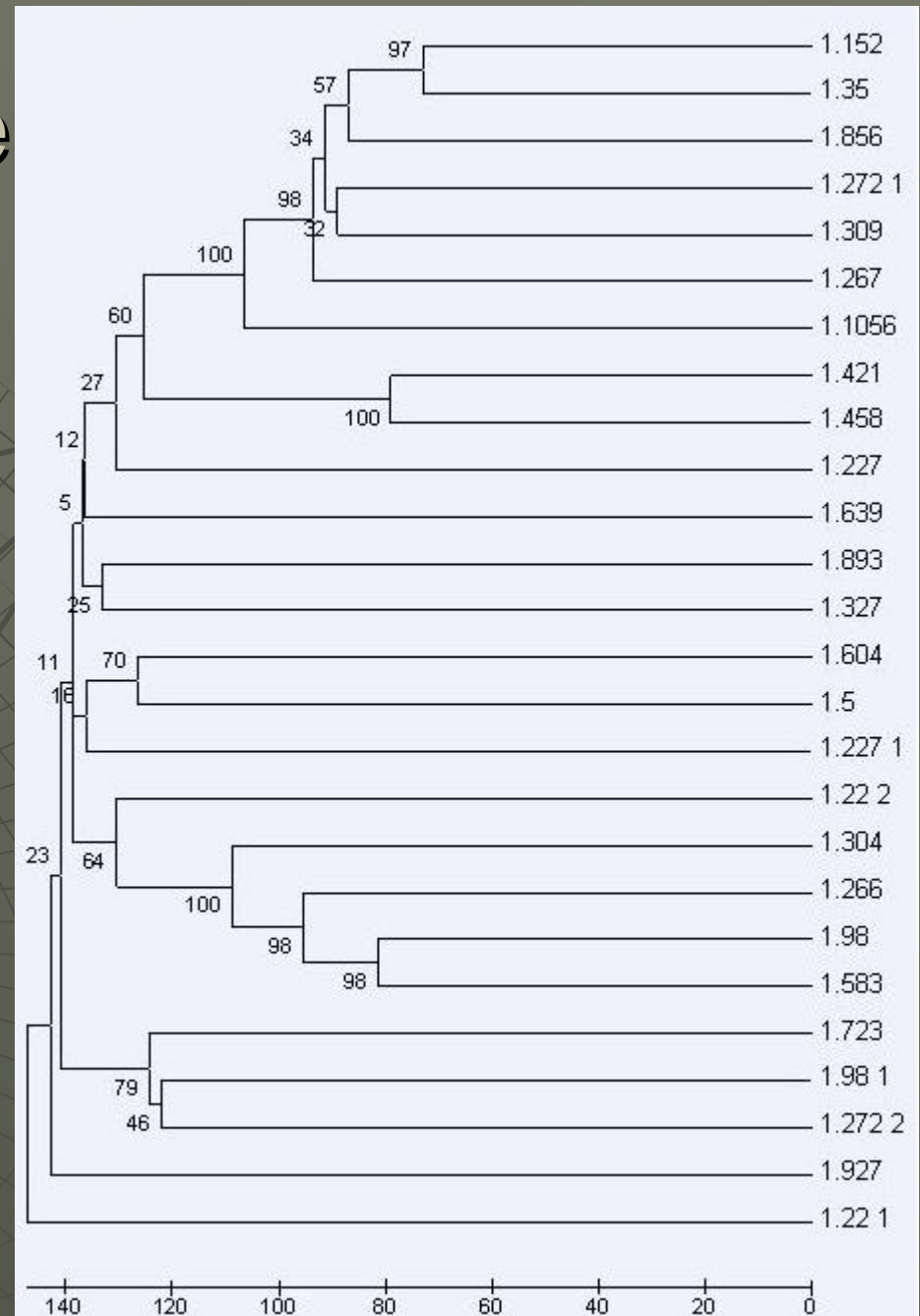
# Aligned Sequences with *Clustal-X*



Multiple alignment of 26 transposon sequences

# Phylogenetic Tree

- ◆ 3 clades of sequences clustered together supported by *Blast* results



# Results

- ◆ Sequences seem to fall into at least 3 clades, based on *Clustal-X* alignment and phylogenetic tree
- ◆ These clades are also supported by *Blast* results against NCBI database nr
- ◆ None of the sequences show exact homology to each other
  - Indicates several insertions or a long time since first insertion

# Conclusion

- ◆ P elements in *Aedes aegypti* seem to come from ancient lineages
  - Don't resemble one another on nucleotide level
  - Still seems to be some similarity on the amino acid level, as seen through the *Blast* results (THAP protein)
- ◆ Successful in finding most intact P elements from the genome
  - From blasting P elements against the genome, resulting in more than 100 hits for some sequences
- ◆ *hmmsearch* doesn't account for frame shifts, so we may be missing some P elements
  - Would have used *genewise*, but it is computationally much more expensive