Computational analysis of transposable element evolution in *Drosophila* genomes.

Casey M. Bergman

Faculty of Life Sciences
University of Manchester

casey.bergman@manchester.ac.uk
http://bioinf.manchester.ac.uk/bergman
Overview of Talk

• Noncoding DNA (ncDNA) & transposable elements (TEs)

• Discovery and detection of TEs in genomes

• Abundance and distribution of TEs in *Drosophila* genomes

• Demographic history of TEs in *D. melanogaster*

• TE population genomics using 454 sequencing
Higher organisms have a higher proportion of ncDNA

- Bacteria: 15%
- Yeast: 30%
Higher organisms have a higher proportion of ncDNA

- Fly: 75%
- Human: 98%
- Bacteria: 15%
- Yeast: 30%
TEs are significant & variable part of ncDNA
3 major types of transposable element (TE)

DNA transposons (cut+paste)

RNA retrotransposons (copy+paste)
Why is the discovery and detection of TE sequences in genomes important?

- Genome assembly
- Genome alignment
- Genome organization
- Genome evolution
- Population genomics
Discovery of new TE families

• Homology to TE proteins (e.g. transposase)
  => HMMer, tBLASTx

• dispersed repeats (all-by-all, k-mers)
  => RECON, PILER, RepeatScout, ReAS

• structural motifs (e.g. LTRs)
  => LTRstruc, LTRharvest

• comparative genomics
  => compTE

Detection of individual TE copies

Quesneville, Bergman et al. (2005) PLoS Comp. Biol. 1:e22
Genomic TE distribution in *D. melanogaster*

~3%  ~20%

Genome-wide average ~5.5%

Bergman, Quesneville et al. (2006) Genome Biology 7:R112
Duplication generates regions of high TE density

Bergman, Quesneville et al. (2006) Genome Biology 7:R112
Chimeric gene-TE transcripts create nested genes

TE abundance in 12 *Drosophila* genomes

- **Proportion TE/Repeat in Scaffolds > 200 Kb**

**Legend**
- BLASTER-tx+Repbase-NoDros
- BLASTER-tx+BDGP
- BLASTER-tx+PILER
- RepeatMasker+ReAS
- RepeatRunner+PILER
- CompTE

**TE abundance in 12 *Drosophila* genomes**

- **D.mel**
  - BLASTER-tx+Repbase-NoDros: 5.3%
  - BLASTER-tx+BDGP: 2.7%
  - BLASTER-tx+PILER: 3.7%
  - RepeatMasker+ReAS: 6.9%
  - RepeatRunner+PILER: 12.0%
  - CompTE: 24.9%

- **D.sim**
  - BLASTER-tx+Repbase-NoDros: 2.7%
  - BLASTER-tx+BDGP: 2.7%
  - BLASTER-tx+PILER: 8.5%
  - RepeatMasker+ReAS: 8.5%
  - RepeatRunner+PILER: 15.6%
  - CompTE: 13.9%

- **D.sec**
  - BLASTER-tx+Repbase-NoDros: 8.9%
  - BLASTER-tx+BDGP: 8.9%
  - BLASTER-tx+PILER: 13.9%
  - RepeatMasker+ReAS: 13.9%
  - RepeatRunner+PILER: 24.9%
  - CompTE: 24.9%

Abundance of major TE types is conserved across genus *Drosophila*

Is the genomic distribution of TEs in *D. melanogaster* affected by historical activity?
non-LTR retrotransposons exhibit a 'pseudogene-like' mode of sequence evolution

Alignment of paralogous TEs

non-LTR retrotransposons exhibit a 'pseudogene-like' mode of sequence evolution

Retrotransposon demographics in *D. melanogaster*

Summary of retrotransposon demographics inferred from intra-genomic comparisons

• LTR elements systematically younger than non-LTR elements

• non-LTR families inserted in waves since speciation

• most LTR families inserted since colonization of non-African habitats

• LTR insertions not at transposition-selection equilibrium

• Low frequency of LTR insertions may not be due to selection

From evolutionary statics to dynamics: population genomics of TEs using 454 sequencing
Population genomics of TEs using 454 sequencing

Strain X

454 Reads

Hybrid TE-unique reads
“Unique Flank Tags”
Population genomics of TEs using 454 sequencing

Strain X

454 Reads

Hybrid TE-unique reads
“Unique Flank Tags”

Reference

TEs

KNOWN ✓

NEW!
Population genomics of TEs using 454 sequencing

<table>
<thead>
<tr>
<th>chrX:</th>
<th>8170000</th>
<th>8180000</th>
<th>8190000</th>
<th>8200000</th>
<th>8210000</th>
<th>8220000</th>
<th>8230000</th>
<th>8240000</th>
<th>8250000</th>
<th>8260000</th>
<th>8270000</th>
<th>8280000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H(+)64</td>
<td>INE-1(-)4505</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AF28-5</td>
<td>unique flanks</td>
<td>AF46-2</td>
<td>unique flanks</td>
<td>jockey_189203_0529_0962</td>
<td>jockey_173763_1355_3963</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AF56-4</td>
<td>unique flanks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AF50-5</td>
<td>unique flanks</td>
<td>NC301 unique flanks</td>
<td>NC303 unique flanks</td>
<td>NC306 unique flanks</td>
<td>NC358 unique flanks</td>
<td>NC375 unique flanks</td>
<td>NC732 unique flanks</td>
<td>Gap Locations</td>
<td>FlyBase Noncoding Genes</td>
<td>FlyBase Noncoding Genes</td>
<td></td>
</tr>
</tbody>
</table>

**TEs in reference sequence**
Population genomics of TEs using 454 sequencing

- **Known** *INE-1* insertion present in NC and AF strains
- **Novel** jockey insertion present in >1 strain

**TEs in reference sequence**
Preliminary findings using 454 sequencing

• 10 strains of *D. melanogaster*: 6 USA, 4 Malawi

• ~24% of annotated TEs found in >=1 wild strain (1300/5400)

• ~72% found in nature in low recomb. regions (950/1300)

• DNA transposons (~30%) found in nature more often than LTR/non-LTR retrotransposons (~10%)

• ~1/3 of INE-1 found in nature, so estimate ~3900 annotated TEs present in >=1 wild strain

• consistent with all TEs fixed in low recomb. regions plus few hundred segregating in high recomb. regions
Summary

• Mature methods exist for analysis of TEs in genomes
  http://www.bioinf.manchester.ac.uk/bergman/te-tools.html

• Structural classes of TEs have different genome dynamics

• Recent LTR insertion has implications for transposition-selection balance paradigm

• Population genomics using next generation sequencing will help resolve forces controlling TE evolution
Hadi Quesneville
Ruqiang Li
Douda Bensasson
Andy Clark

2 Post-doctoral positions (available ~Sep 2008)