Analysis of the Ty3 Retrotransposon in *Saccharomyces cerevisiae* Using Transposition Assays and High Throughput Sequencing Technologies

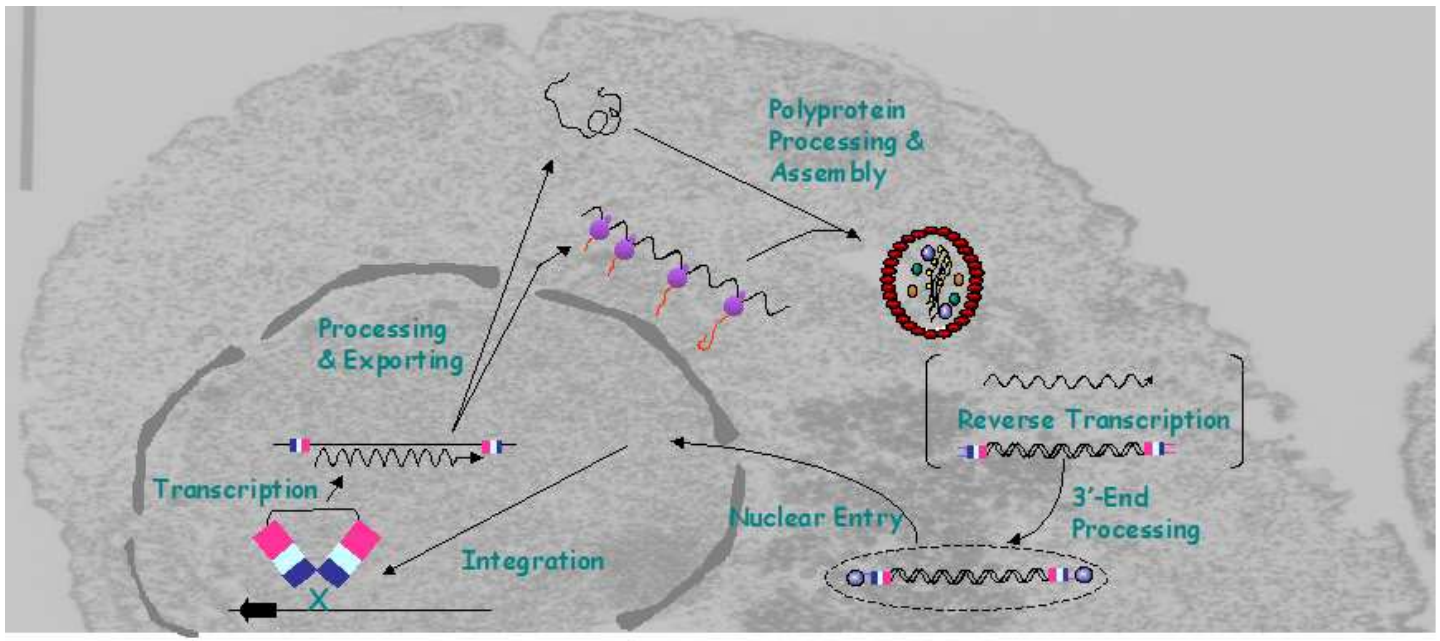
Kenny Daily, Paul Rigor, Sholeh Forouzan, Kim Nguyen, Pierre Baldi, Suzanne Sandmeyer

June 21, 2009

UCIrvine | University of California

Ty3 Retrotransposon Introduction

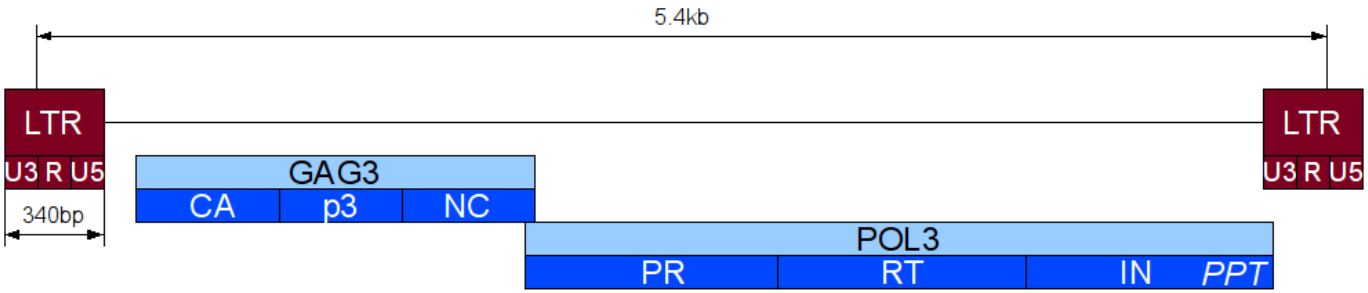
- Life Cycle in Budding Yeast
  - Entire life cycle happens intracellularly
  - Replicates through reverse transcription
  - Integrates full length DNA into the host, specifically to Pol III-transcription initiation sites
- Reasons to study
  - A model for retroviruses (MLV/HIV)
  - Targeting properties for gene therapy vectors
  - Discovery of new Pol III-transcribed genes



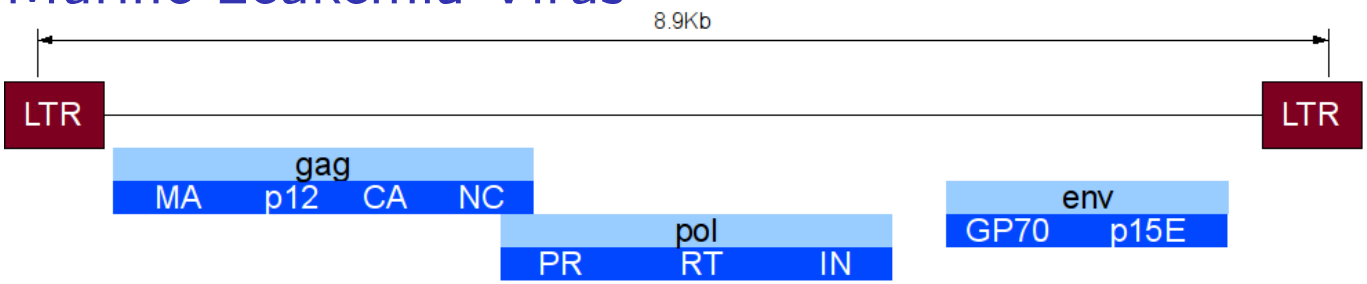
Ty3 Retrotransposon Genome Comparison

The Ty3 genome is very similar to retroviruses, such as MLV.

Ty3 Retrotransposon



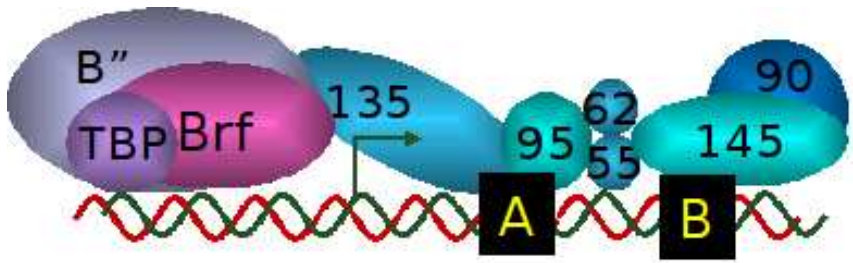
Murine Leukemia Virus



RNA Pol III Transcription Initiation Sites are Targets of Ty3 Integration

RNA Pol III Machinery

TFIIIB and TFIIIC

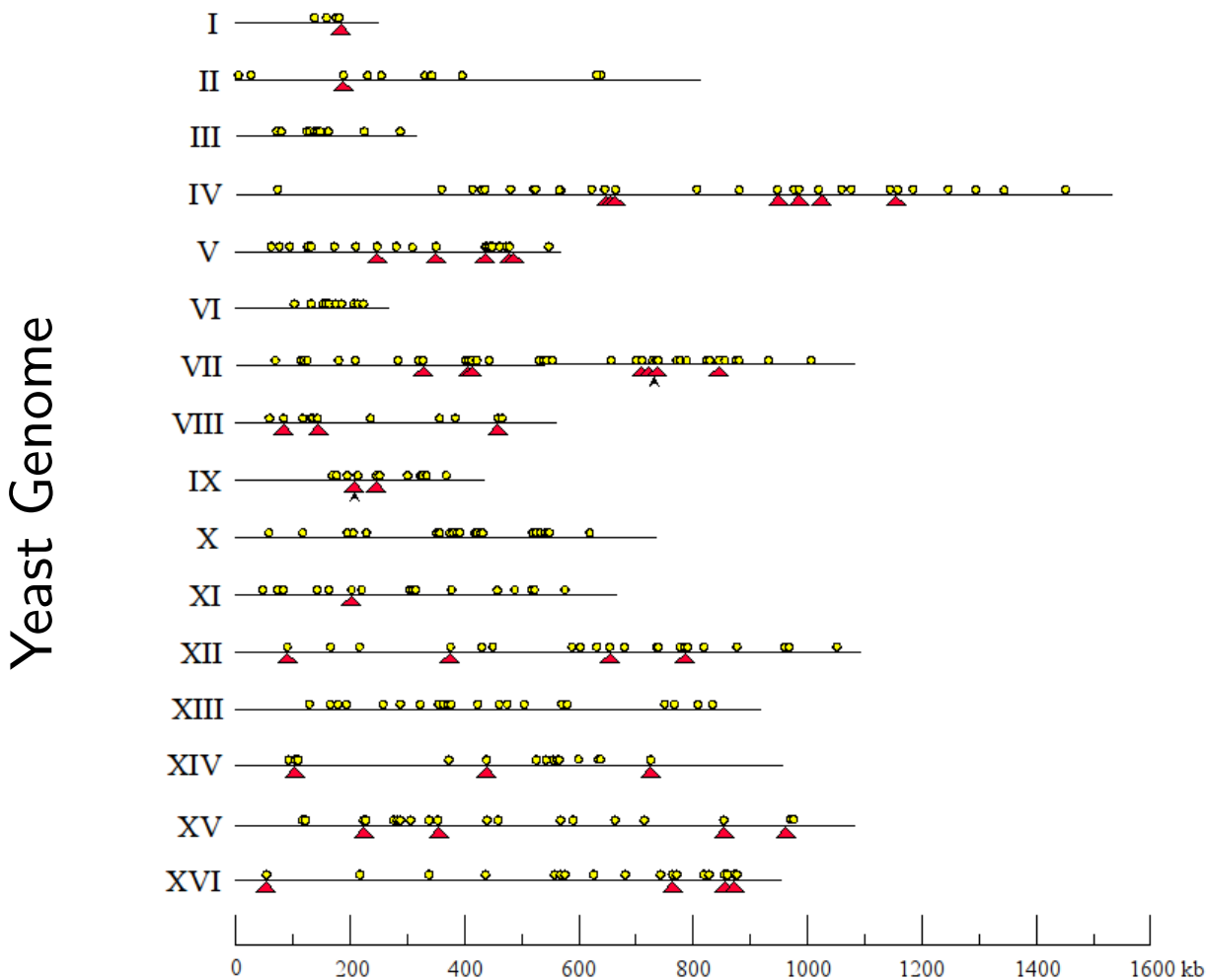


Plasmid Studies of Ty3 Integration

A link was found between insertion events and
Pol III-transcribed genes encoding small structural RNAs.

Genomic LTRs are Associated With tRNA Genes

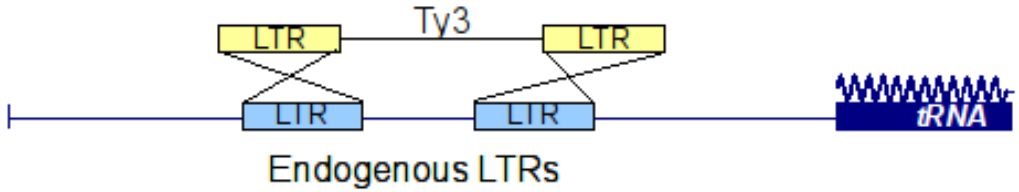
- Two full length Ty3 insertions in the yeast genome
- A total of 41 LTRs, 37 of these are solo



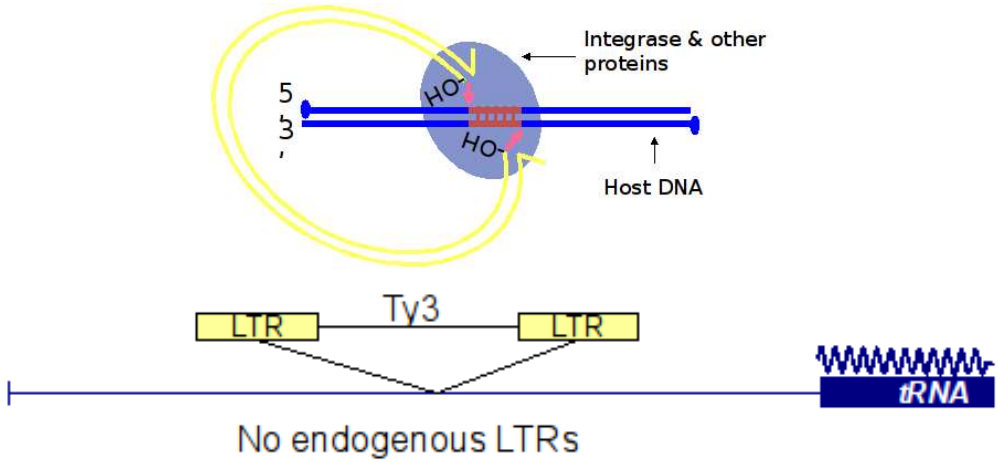
Yellow circle: tDNA
Red triangle: LTR
Black arrow: Full Ty3

Mechanisms of Ty3 Transposition

Homologous Recombination (Endogenous *RAD52*)



Integration (Ty3 Integrase)

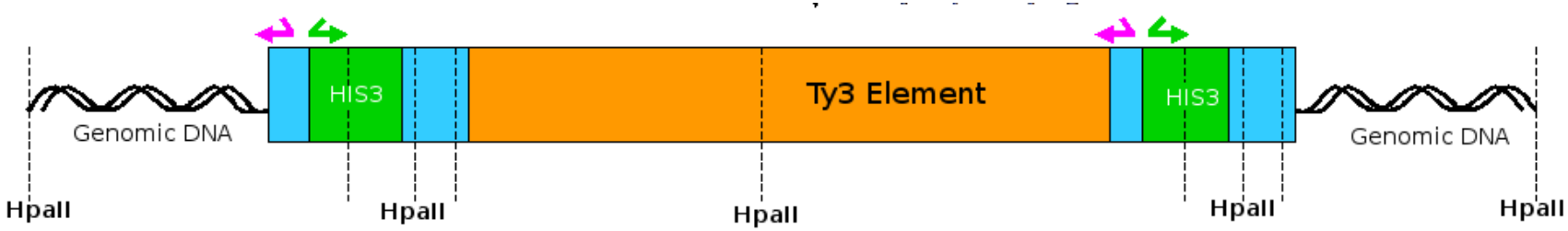


Data Generation and Collection

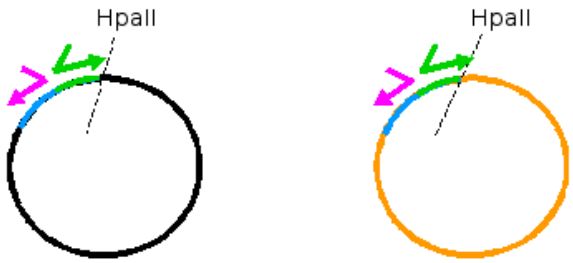
- First, locate the insertion sites

Amplification of Ty3 Insertion Sites for Sequence Analysis

- Digest DNA, Circularize, Inverse PCR
- Sequence PCR products (Solexa Illumina)

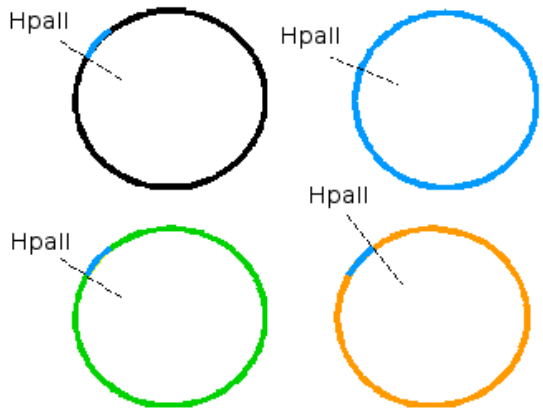


PCR products



CATTTTGAGATACAACATACGTACCGTACG

No PCR products



Mapping Sequence Data to the Yeast Genome

Results of sequencing

- millions of short "reads" (36bp)
- A read contains both LTR sequence from the end of Ty3 (17bp) and genomic sequence (19bp)

Align Reads to Genome

Use short read aligner to find the point of insertion in the yeast genome, giving us genomic coordinates of insertion event (**Bowtie**, MAQ, SOAP)

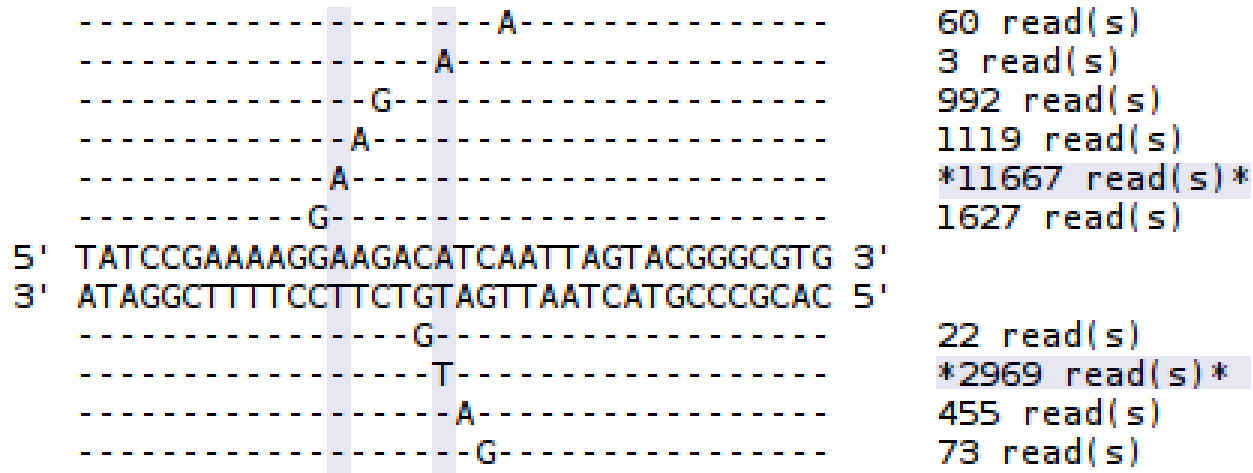
Reads are Mapped to the Genome

The **reads** are mapped to the reference genome. The stars above and below the reference sequence indicate positions where **insertions** occurred. This is a truncated example; a region may have thousands of reads.

Determine Insertion Sites

Reads are collapsed to insertion sites

Number of reads at one position on one strand = **density** of that insertion site



Clustering of insertion sites

- Assume co-localized (distance = 10bp) sites arise from a common target.
- We group these sites into a **cluster** and summarize the cluster c for dataset d by:

$$c_d = \frac{\sum_{i=c_{start}}^{c_{stop}} \text{reads}_i}{\sum_{d} \text{reads}_i}$$

- giving us a matrix of C clusters by D datasets where each entry indicates the percentage of the total reads for cluster c in dataset d.
- Confident that the number of reads is proportional to the insertion usage (supported by qPCR, reproducible patterns with different PCR strategies)

Questions

Does Ty3 target only known Pol III-transcribed genes?

Are there preferentially targeted Pol III-transcribed genes?

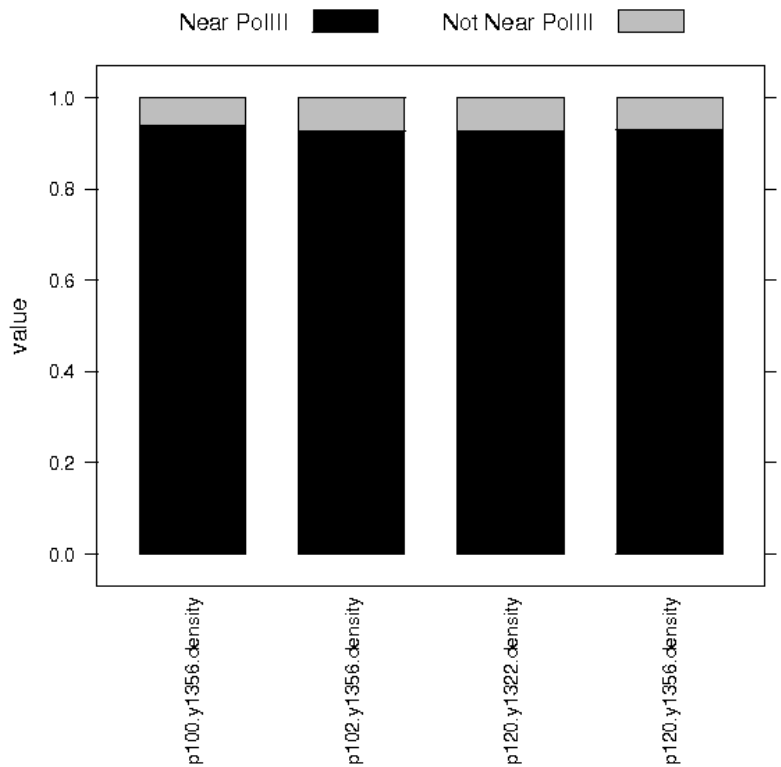
Ty3 inserting via recombination or integration?

What features explain insertion preferences?

- Differences in Pol III Box A/B sites
- Chromosomal context

De novo Insertions Target Pol III-transcribed Genes

- Majority of the reads fall $\leq$ 50bp from 5' end of a Pol III-transcribed gene
- Including all known non-tRNA Pol III-transcribed genes
- Other 10% could be unkown Pol III-transcribed genes, or novel targets of insertion.
  - Two that are found are *UFO1* and *TIM21*, which are nearby degenerate tRNAs.



Questions

Does Ty3 target only known Pol III-transcribed genes?

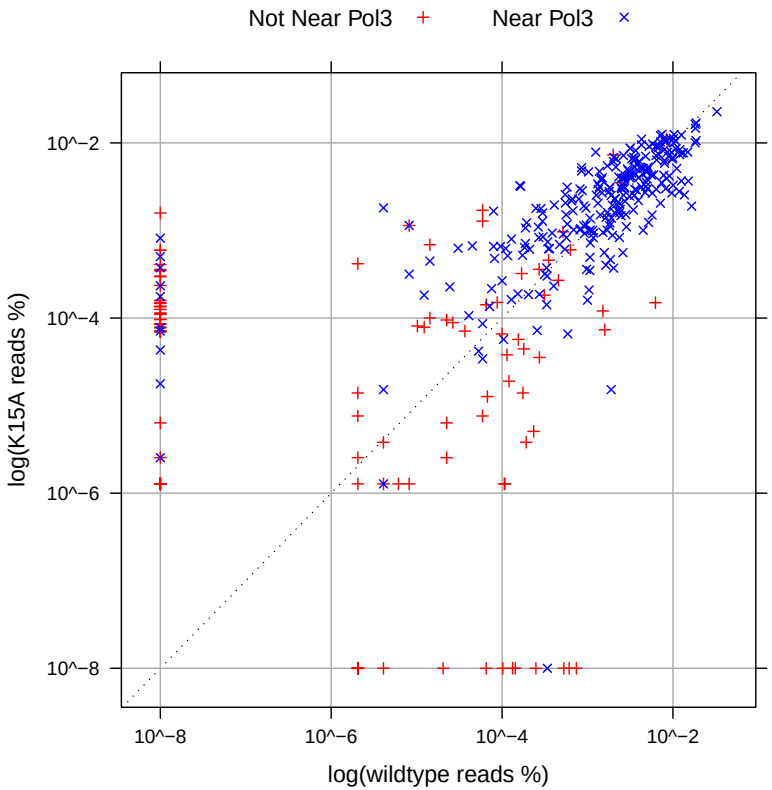
Are there preferentially targeted Pol III-transcribed genes?

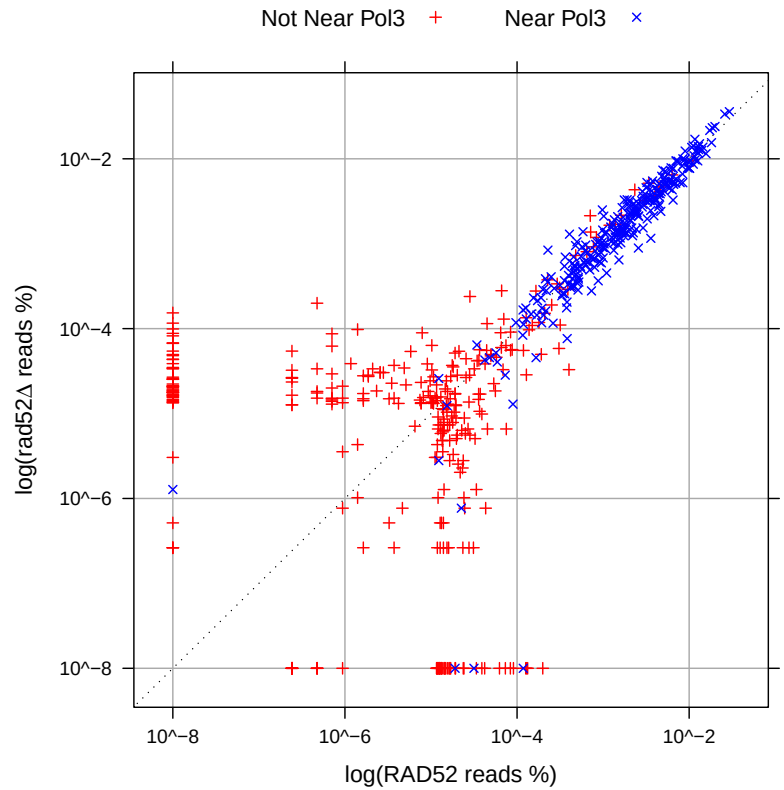
Ty3 inserting via recombination or integration?

What features explain insertion preferences?

- Differences in Pol III Box A/B sites
- Chromosomal context

Insertion Densities Vary Across the Genome





Questions

Does Ty3 target only known Pol III-transcribed genes?

Are there preferentially targeted Pol III-transcribed genes?

Ty3 inserting via recombination or integration?

What features explain insertion preferences?

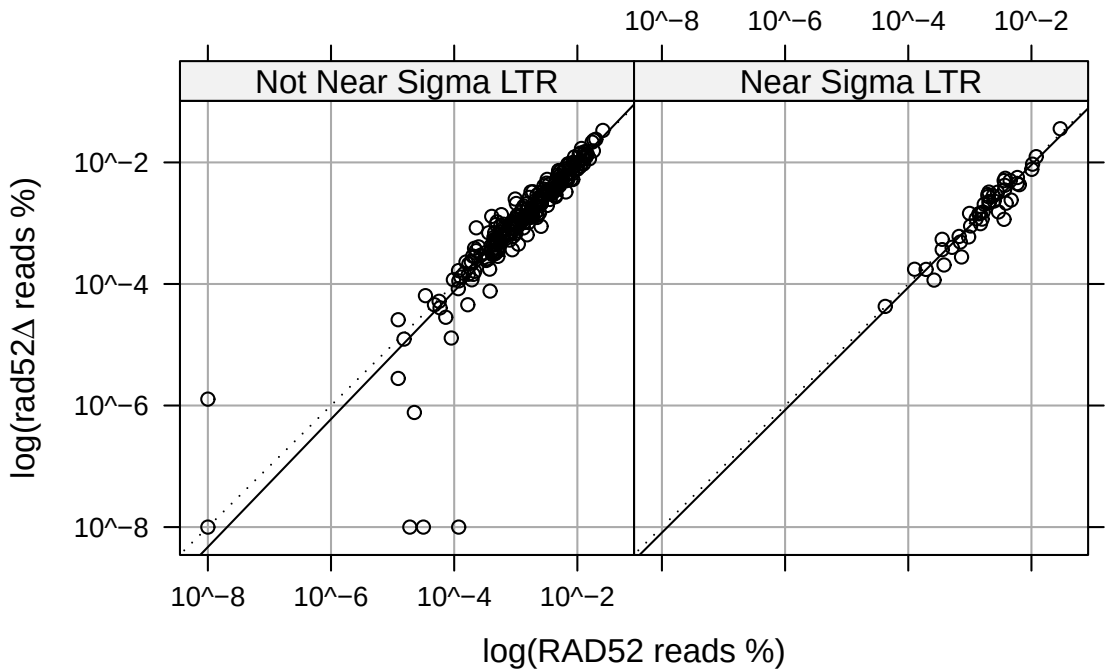
- Differences in Pol III Box A/B sites
- Chromosomal context

Is Ty3 Targeting to tRNA Genes Dependent on Recombination with Existing LTRs?

- Compare Ty3 transposition patterns in the presence (*RAD52*) and absence (*rad52*Δ) of recombination
  - Strain without *RAD52* cannot integrate Ty3 via homologous recombination, only via the integrase product of the Ty3 genome.

Comparison of *RAD52* and *rad52Δ*

Ty3 is not only using homologous recombination with existing genomic LTRs, but is integrating at *de novo* sites using integrase.



Questions

Does Ty3 target only known Pol III-transcribed genes?

Are there preferentially targeted Pol III-transcribed genes?

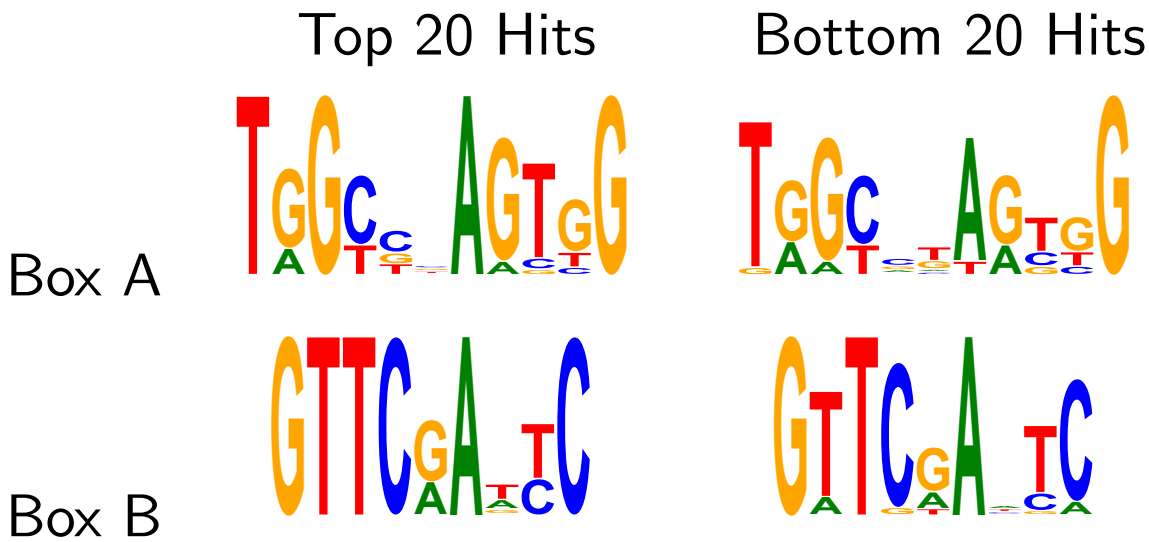
Ty3 inserting via recombination or integration?

What features explain insertion preferences?

- Differences in Pol III Box A/B sites
- Chromosomal context

Differences in Box A and B sites of Top/Bottom 20 Insertion Clusters

- Rank insertion clusters by percentage of reads
- Examine Box A and B sites in Pol III genes with many insertions versus those with few insertions
- Degenerate (hence possibly weaker) Box A and B motifs in those with few insertions!



Conclusions

- High-throughput sequencing can be used to study Ty3 transposition targeting

Conclusions

- High-throughput sequencing can be used to study Ty3 transposition targeting
- Ty3 targets all known Pol III-transcribed genes, not only tRNA genes

Conclusions

- High-throughput sequencing can be used to study Ty3 transposition targeting
- Ty3 targets all known Pol III-transcribed genes, not only tRNA genes
- Findings suggest that:
  - Either unkown Pol III-transcribed genes exist in the genome,
  - Ty3 can insert without targeting factors at low frequencies, or
  - Ty3 targeting factors bind other genomic sites.

Conclusions

- High-throughput sequencing can be used to study Ty3 transposition targeting
- Ty3 targets all known Pol III-transcribed genes, not only tRNA genes
- Findings suggest that:
  - Either unkown Pol III-transcribed genes exist in the genome,
  - Ty3 can insert without targeting factors at low frequencies, or
  - Ty3 targeting factors bind other genomic sites.
- Differences may be explained by variations in promoter elements, affecting the binding of Pol III machinery

Thank You!

Advisors and Collaborators

- Pierre Baldi
- Suzanne Sandmeyer
- Baldi Lab (Paul Rigor, Sholeh Forouzan)
- Sandmeyer Lab (Kim Nguyen)
- WUSTL Sequencing Lab (Haoyi Wang, Rob Mitra, Mark Johnston, David Myers)

Funding

- NIH Biomedical Informatics Training Grant