Georgia Tech

BIOLOGY

Repetitive DNA and the logic of human gene regulation

King Jordan

School of Biology
Georgia Institute of Technology

http://esbg.gatech.edu

Overview

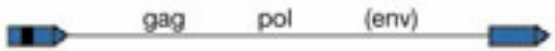
- Biological significance of transposable elements
- Epigenetic regulatory effects
  - o Promoter architecture & nucleosome binding
  - o TE-gene neighborhood and gene expression
- Genetic regulatory effects
  - o Transcription factor binding sites
  - o TE-genome evolution model & test
  - o microRNAs

Accelerated evolution

TE classification & abundance

publication of the human genome sequence underscores the abundance of TEs

Classes of interspersed repeat in the human genome

			Length	Copy number	Fraction of genome
LINEs	Autonomous		6–8 kb	850,000	21%
SINEs	Non-autonomous		100–300 bp	1,500,000	13%
Retrovirus-like elements	Autonomous		6–11 kb	450,000	8%
	Non-autonomous		1.5–3 kb		
DNA transposon fossils	Autonomous		2–3 kb	300,000	3%
	Non-autonomous		80–3,000 bp		

45% TEs, 3% SSR, 5% segmental duplications

more than ½ of human genome is repetitive DNA

(~1.5% protein coding sequence)

Lander ES et al. Nature (2001) 409: 860

Selfish DNA theory of TEs

- How can the abundance & ubiquity of TEs be explained?
- Adaptive explanations – role for host
- Phenotypic paradigm of molecular evolution
- Paradigm shift – led by Richard Dawkins ‘Selfish Gene’
- Selfish replicators – TEs as parasites with no role for the host
- Helps to avoid tautological thinking – ‘adaptive storytelling’
- Discourages research into TE biology

Doolittle & Sapienza Nature (1980) 284: 601
Orgel & Crick Nature (1980) 284: 604

Molecular domestication

- More nuanced view emerges – continuum parasitism to mutualism
- Evolution is opportunistic – novelty from materials on hand
- TEs ideal source of genetic building blocks
- Numerous examples of host functions from TEs
  - o Protein coding sequences from TEs
  - o Regulatory sequences from TEs

Brosius and Gould (1992) PNAS 89: 10706
Miller et al. PNAS (1992) 89: 4018
Kidwell & Lisch Evolution (2001) 55: 1

The TE domestication question

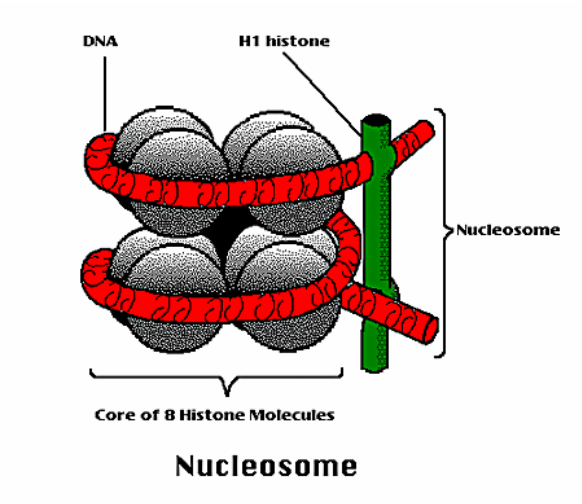
How do TEs – & repetitive DNA in general – contribute to the structure, function & evolution of eukaryotic genomes?

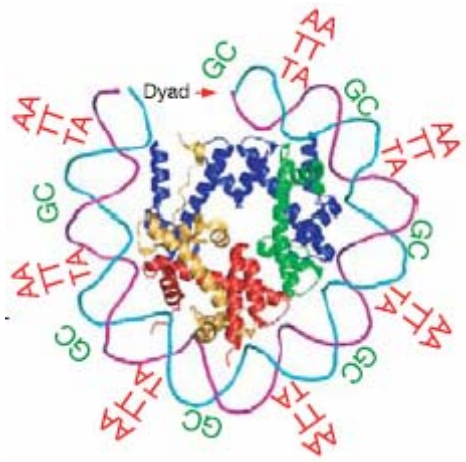
Epigenetic regulatory effects of repetitive DNA

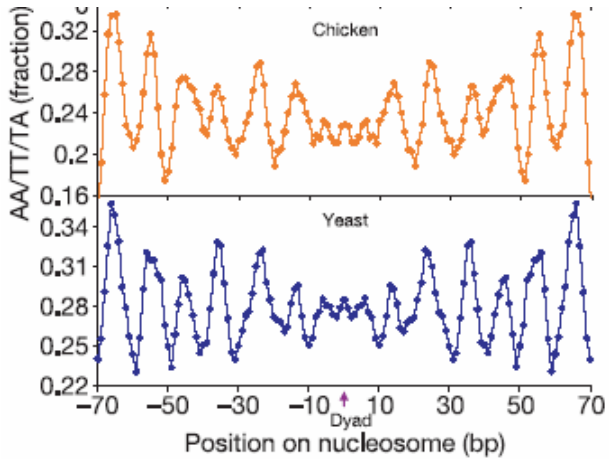
Transposable elements (TEs), Simple sequence repeats (SSRs), chromatin environment and modifications

Repetitive DNA & promoter architecture

- sequence specificity of nucleosome binding recently characterized
- can predict nucleosome binding locations and affinities accurately
- nucleosomes do not bind tightly near transcriptional start sites
- allows for DNA access by transcriptional machinery
- how does repetitive DNA relate to this phenomenon??

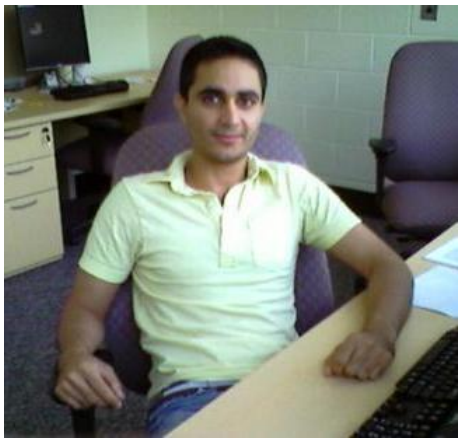






Segal et al. (2006) Nature 442: 772

Promoter sequence analysis



Ahsan Huda

1 – isolate proximal promoter regions, 1kb upstream TSS

2 – identify locations of repetitive DNA
TEs and low complexity/simple repeats

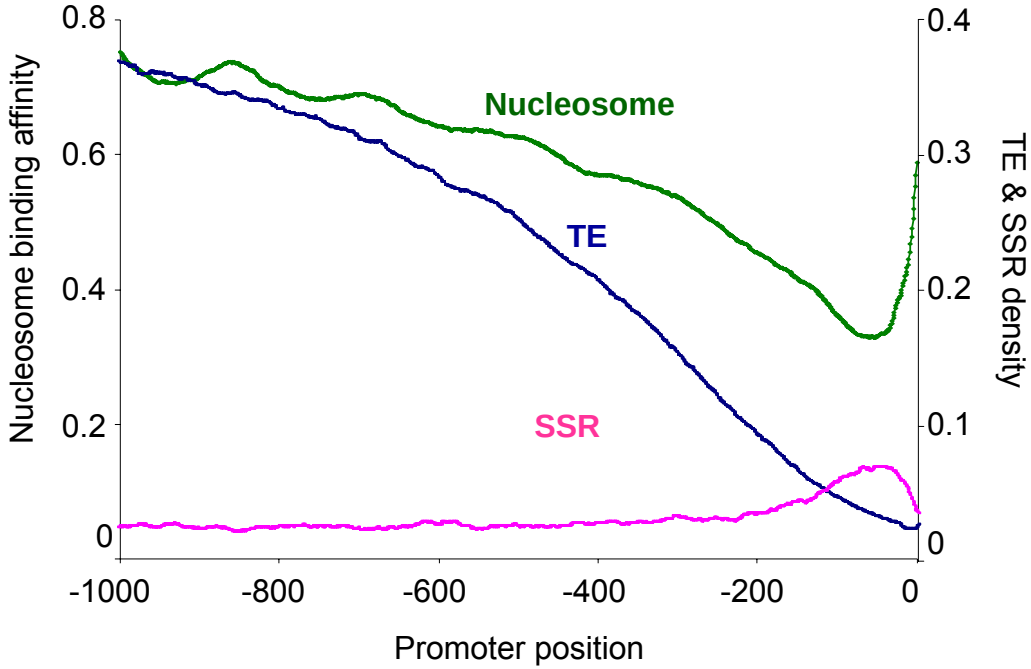
3 – predict nucleosome binding affinities using Segal model (Chicken)

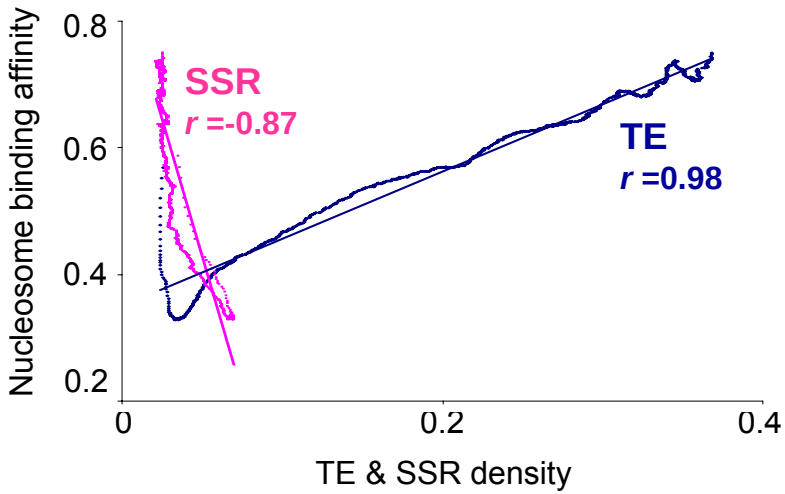
4 – compare nucleosome binding affinities to repetitive DNA

5 – cluster promoters wrt repetitive DNA content

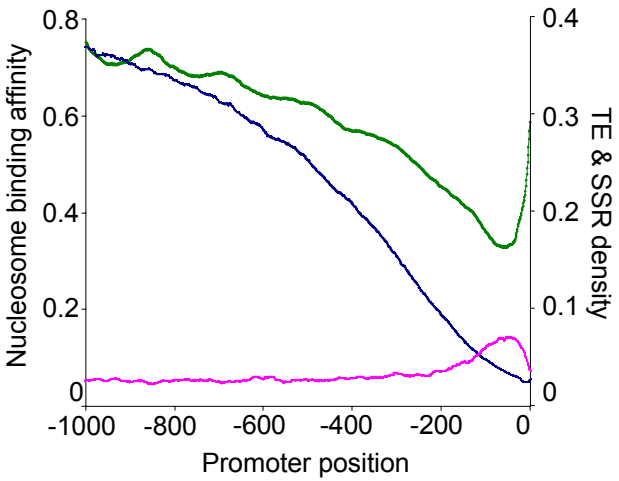
6 – assess regulatory properties (co-regulation?) for clusters

Human promoter analysis

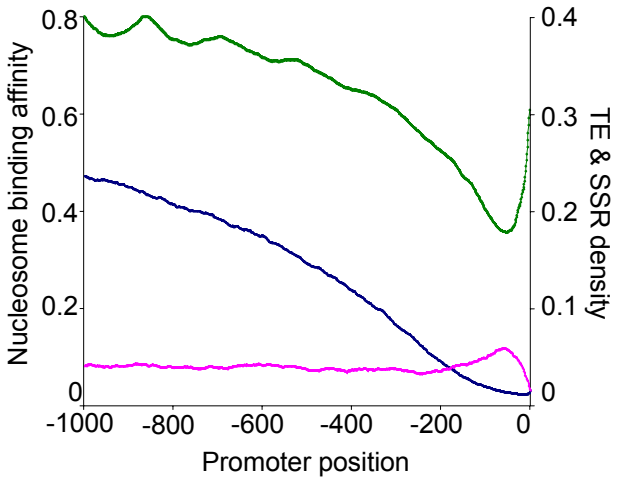




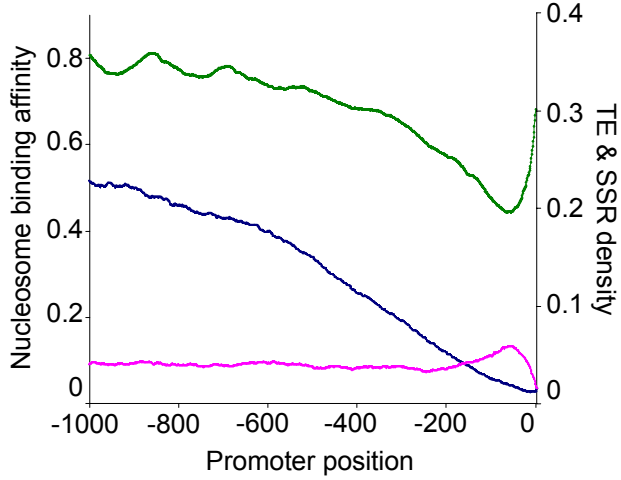
Multi-species comparison



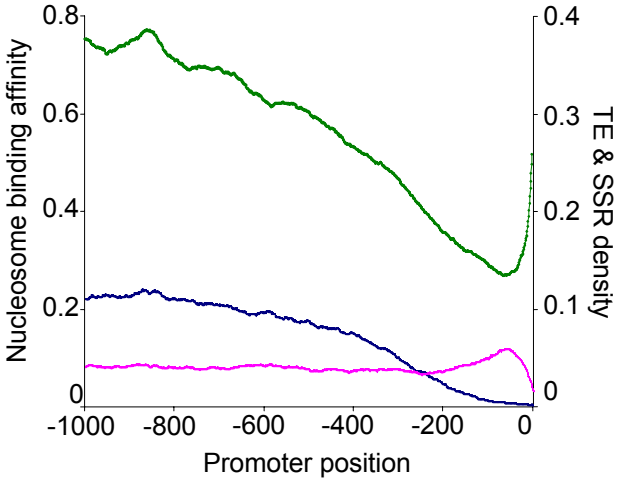
Chimpanzee



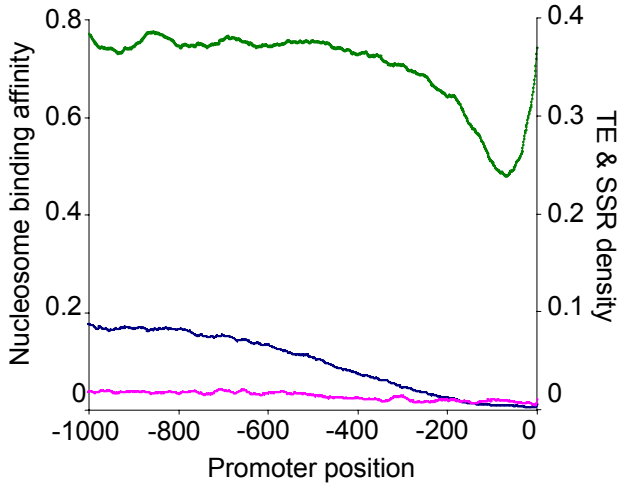
Mouse



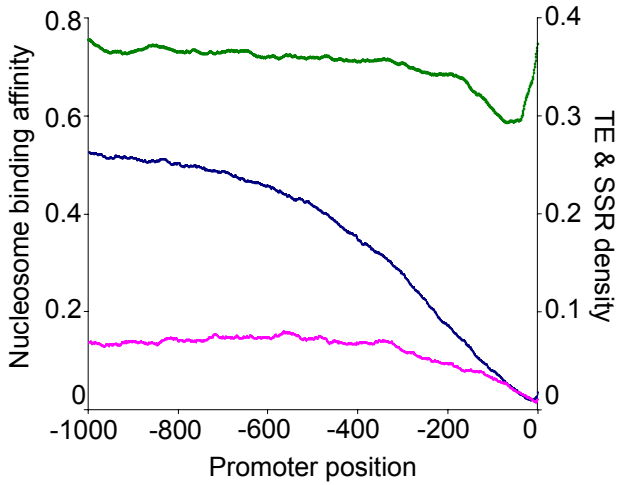
Rat



Chicken



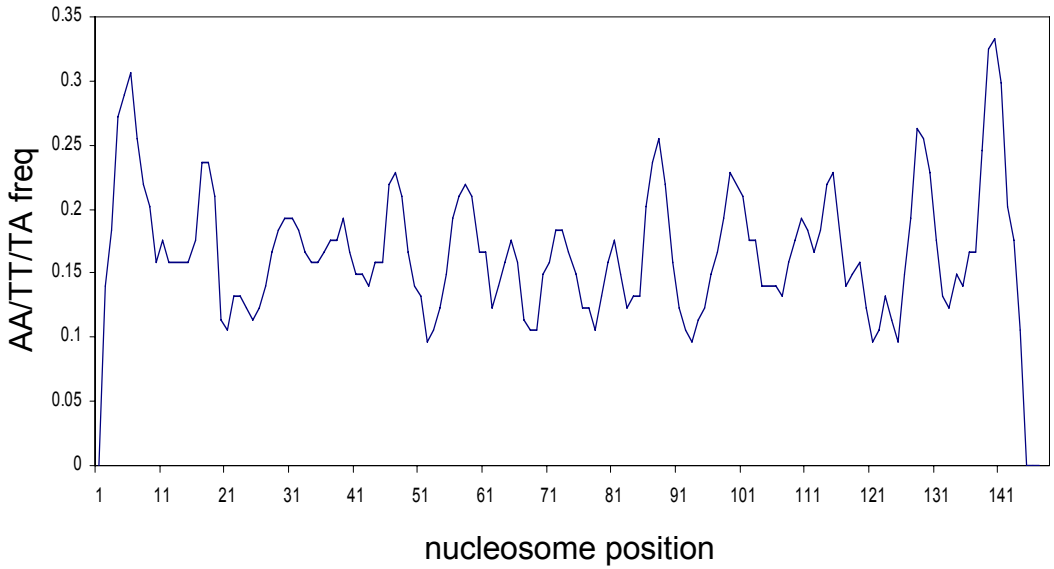
Xenopus

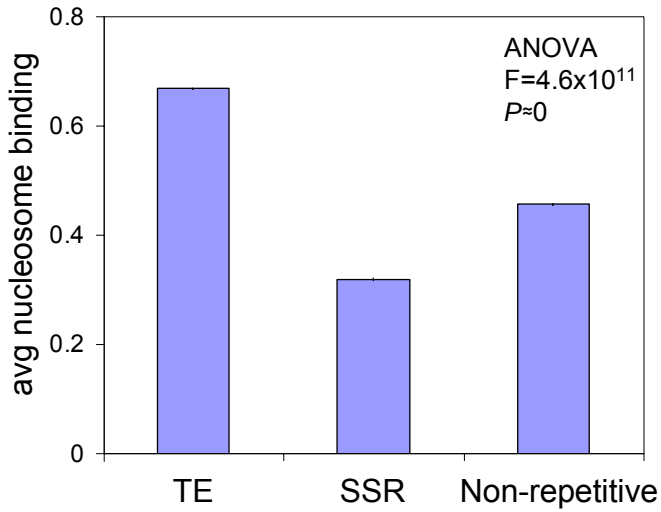


Zebrafish

Repetitive DNA nucleosome affinity

- experimentally characterized TE-derived nucleosome seqs show periodicity
- TEs have tighter nucleosome binding than non-repetitive DNA
- SSR sequences have lower nucleosome binding affinity





Promoter characterization with repetitive DNA

- promoter sequences can be characterized wrt the position and density of repetitive DNA sequences
- visually – any p_i can be color coded

  TE =  LC/SR =  Non-repetitive = 

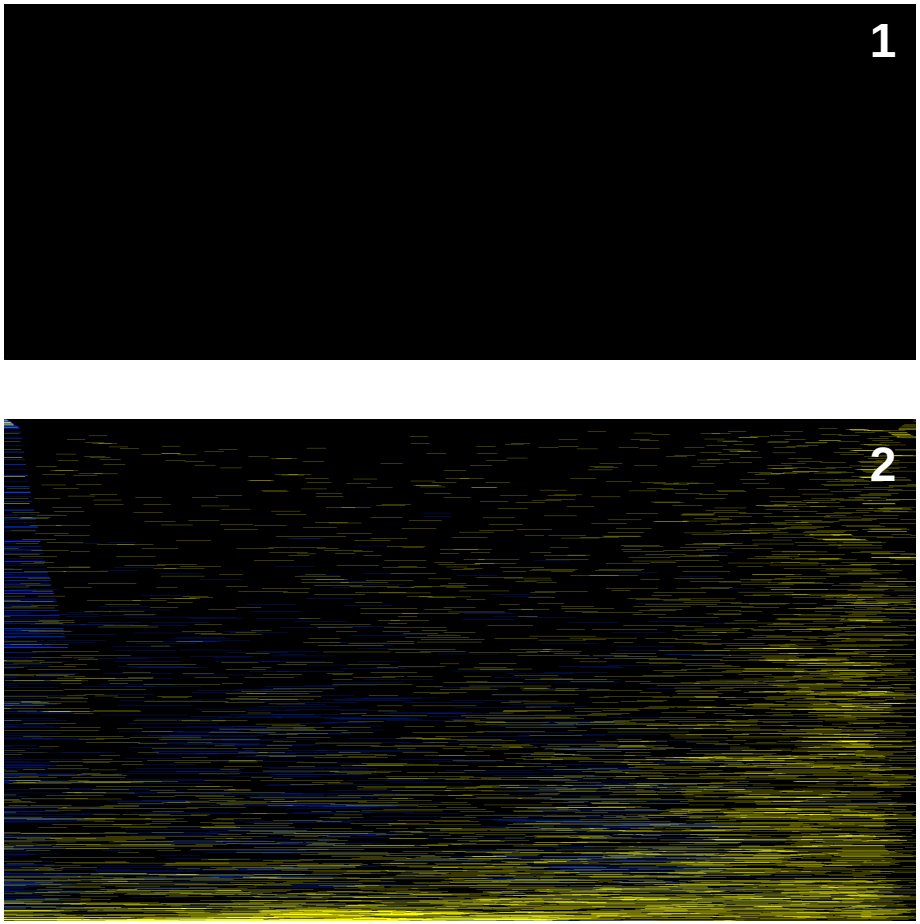
- numerically – any p_i can be assigned a value $p_i \in \{-1,0,1\}$

- promoter-specific vectors can be clustered – *e.g.* using SOM – to yield groups with similar repetitive DNA profiles

Human promoter clusters

TE-

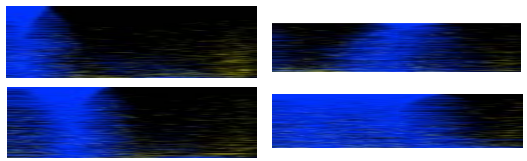
TE+

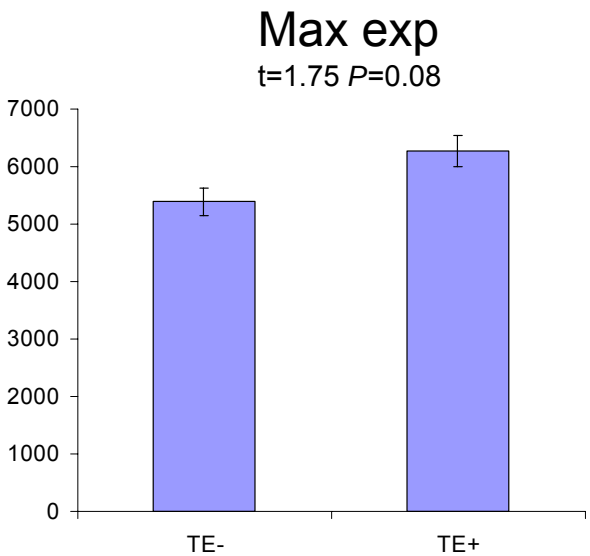


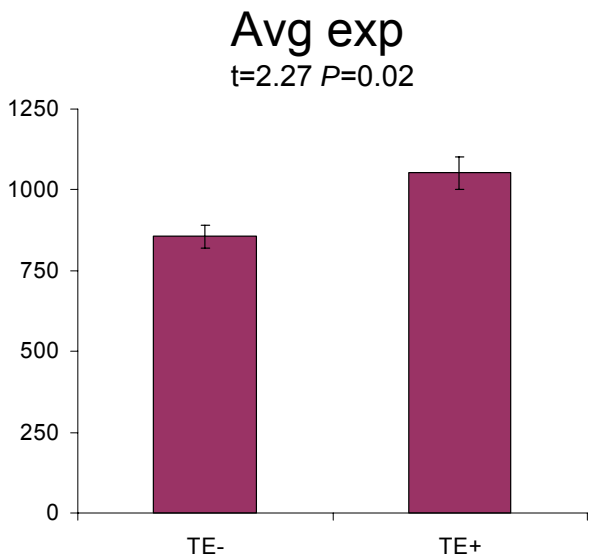
Promoter cluster characteristics

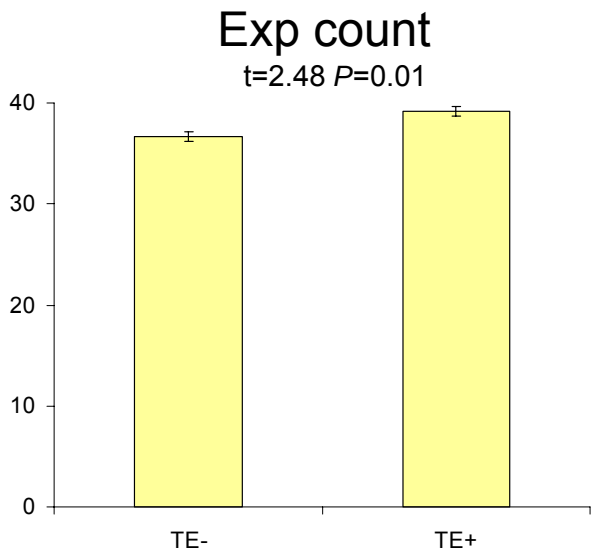
- promoter clusters can be characterized according to the expression patterns of their genes
- TE- clusters 

  have significantly lower expression (max, avg, count) than

  TE+ clusters 



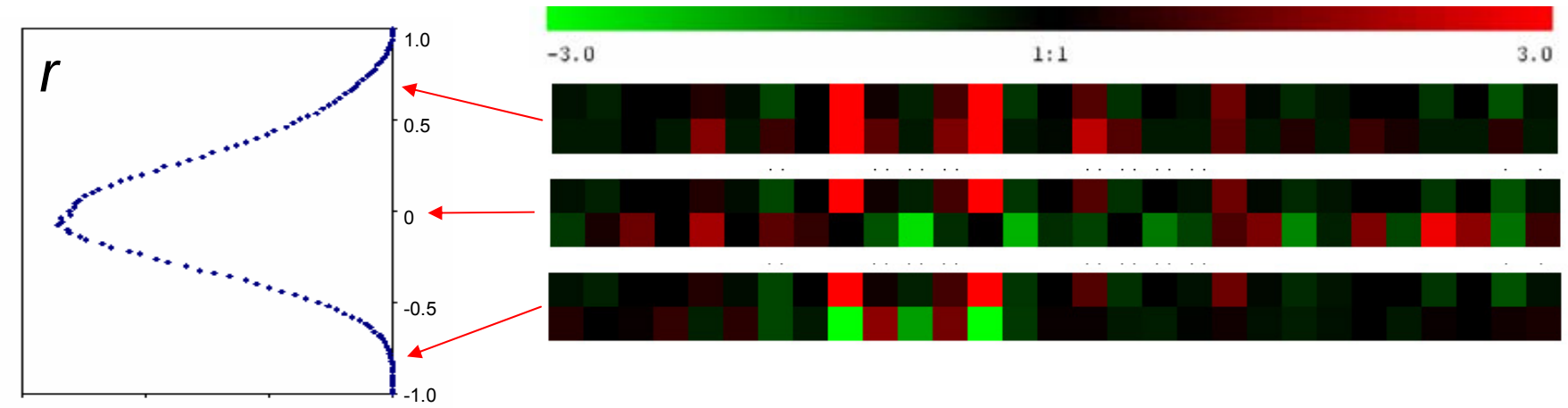




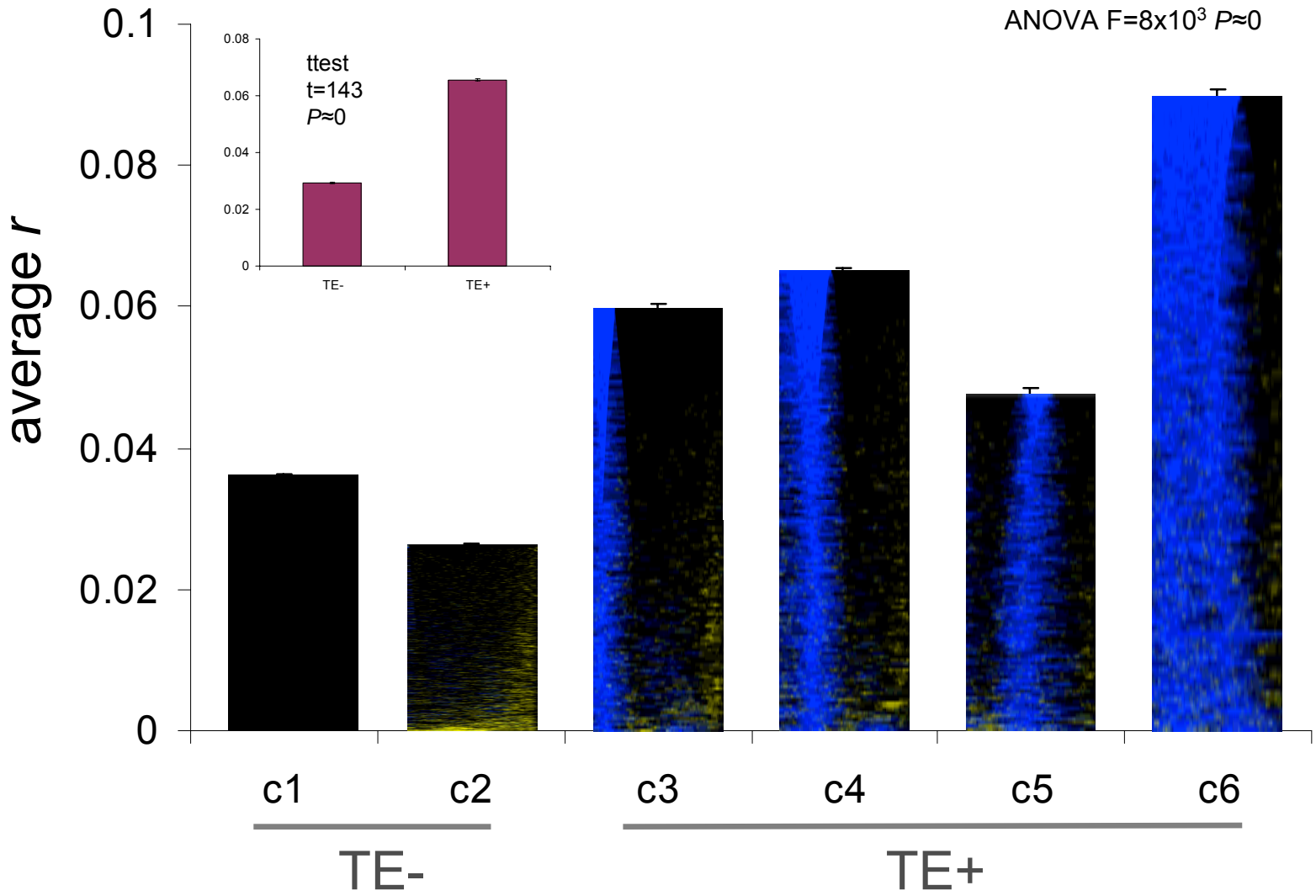
Co-regulation of related promoters

- differences between pairs of genes for various expression parameter values can be used to search for evidence of cluster-specific co-regulation

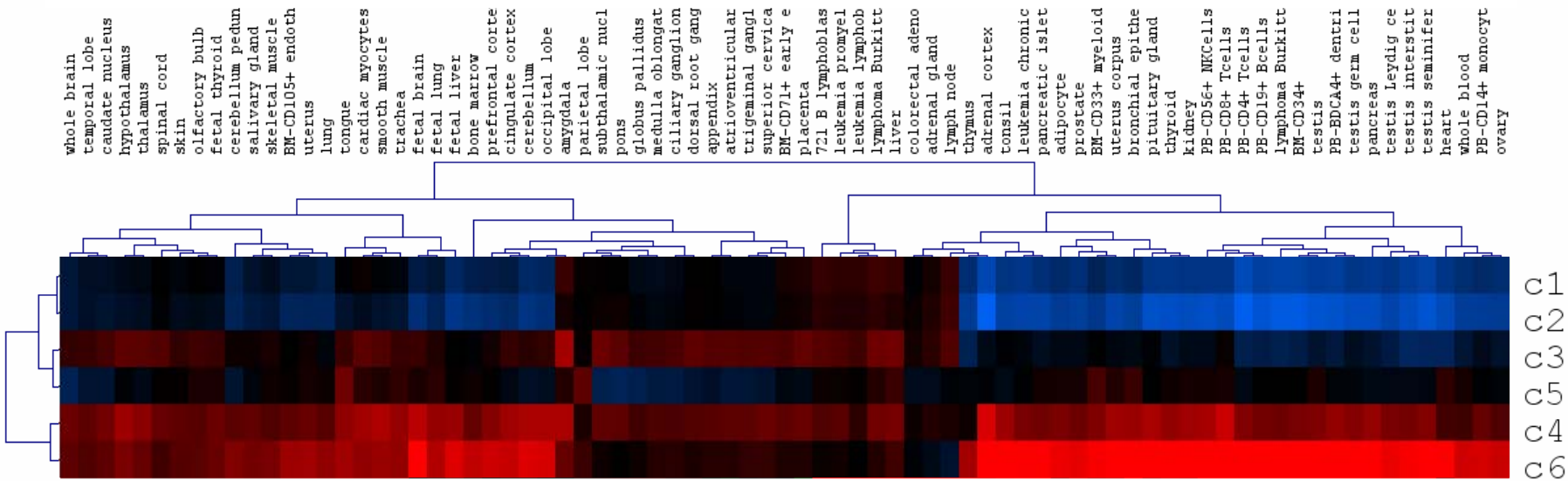
- e.g. compare expression profiles with Pearson correlation coefficient (*r*)

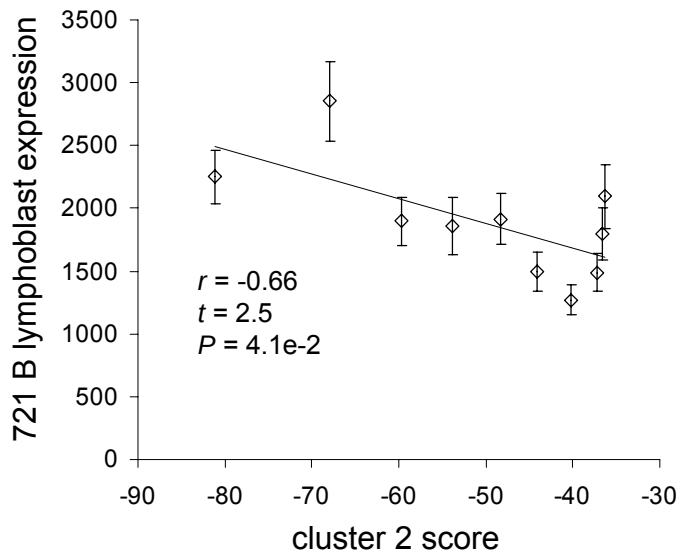


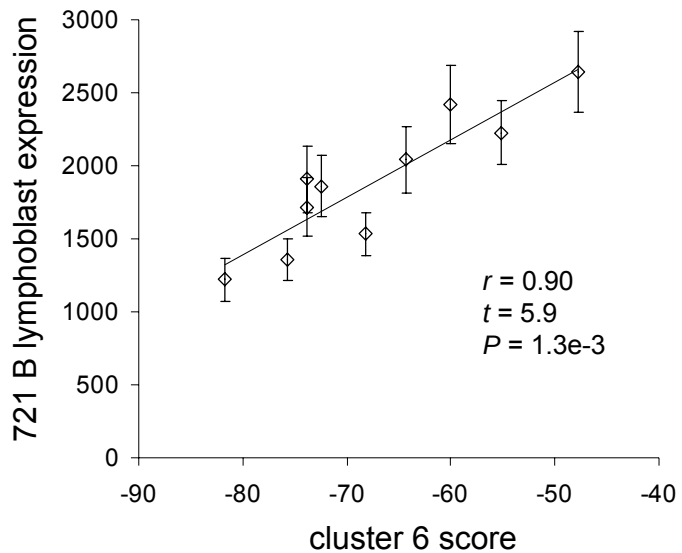
TE+ promoters are coregulated



TE promoter clusters and tissue-specific gene expression patterns

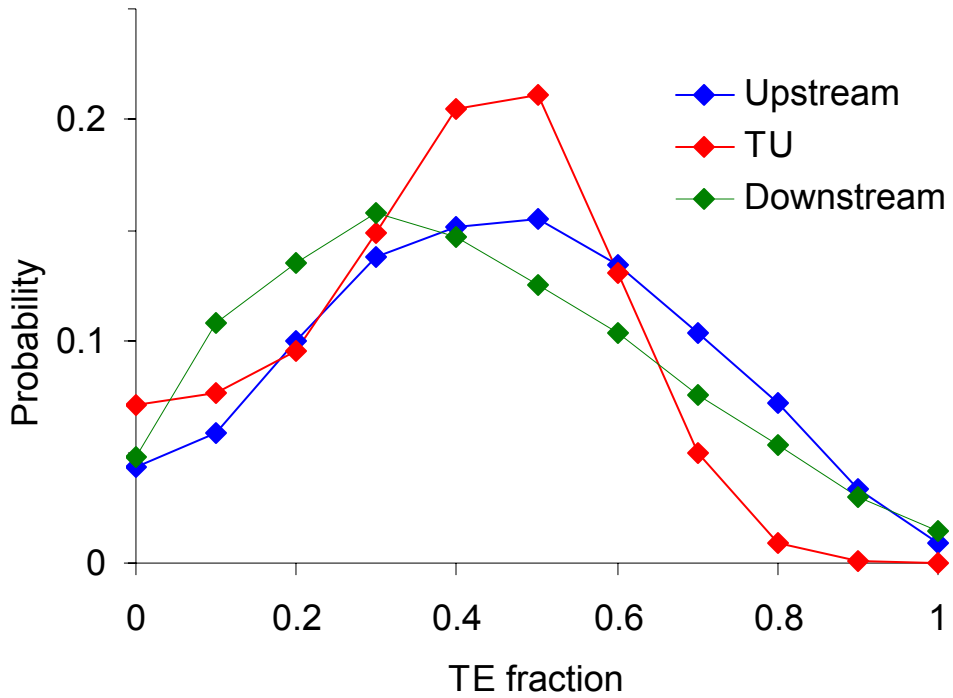






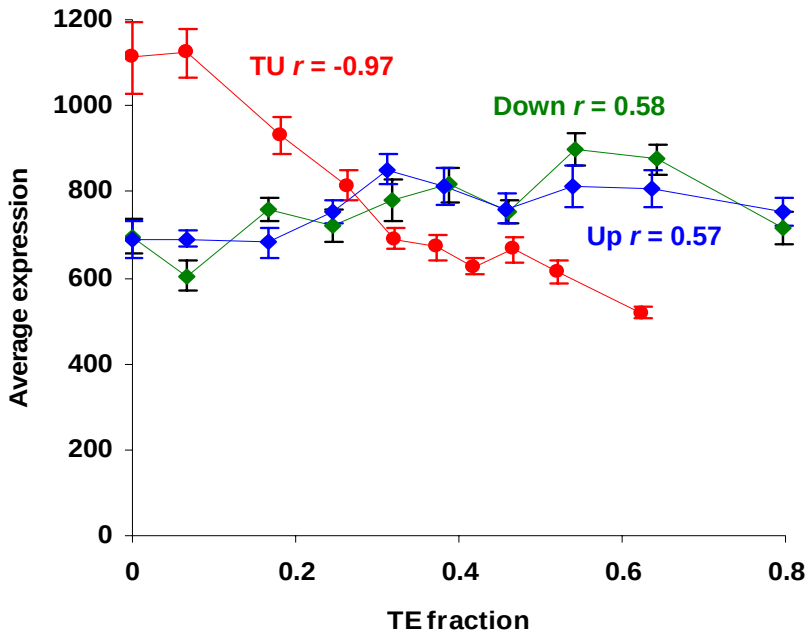
TE insertion profiles & gene expression

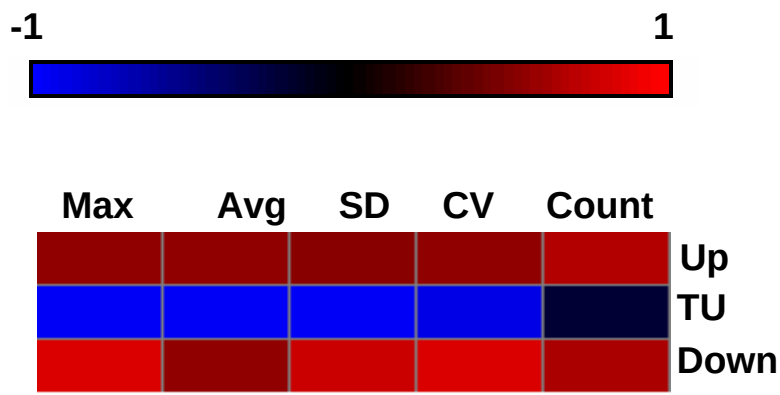
- TE insertion profiles characterized for all human gene loci
- transcriptional unit (TU) taken as upstream most TSS and downstream most TTS
- fraction of TE residues computed for TU and 5k up/down



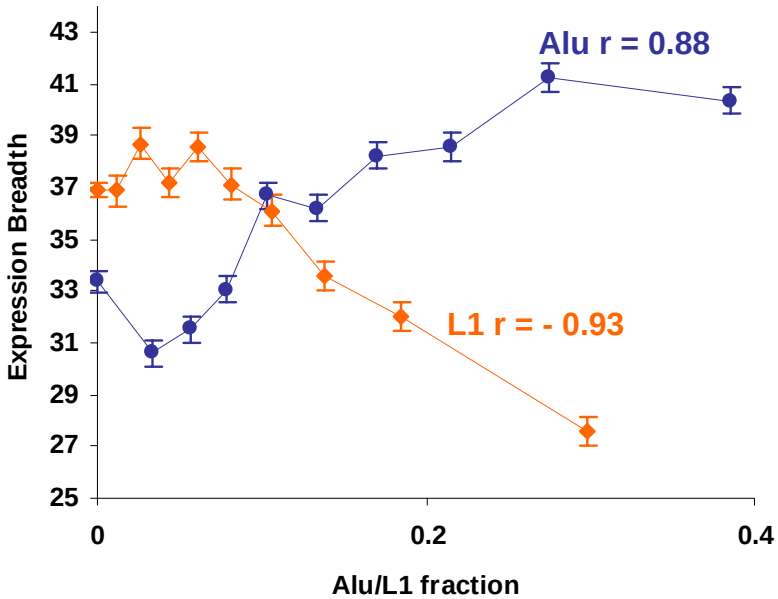
Effect of TE neighborhood on gene expression

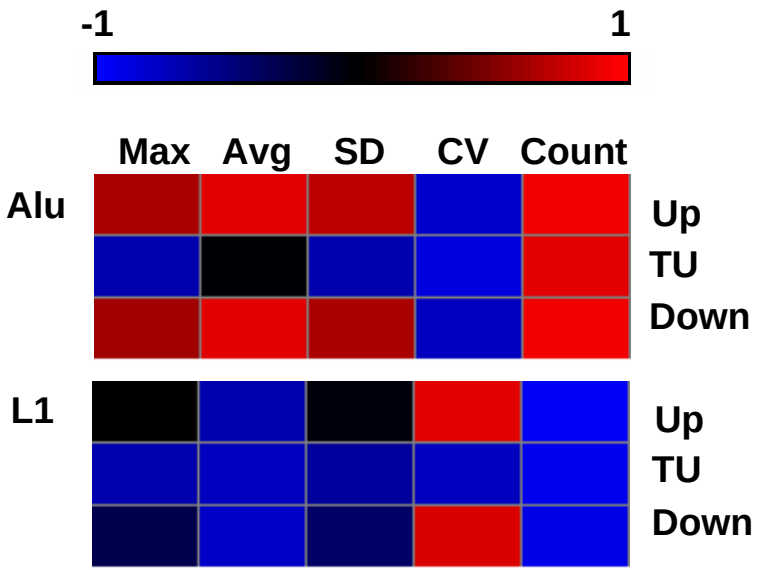
- Novartis Gene Expression Atlas - Affymetrix microarray expression data for 44,775 probes across 79 human tissues
- gene-specific expression parameters calculated: average, maximum, breadth (count), standard deviation and coefficient of variation
- gene-specific TE fractions compared to expression paramter values



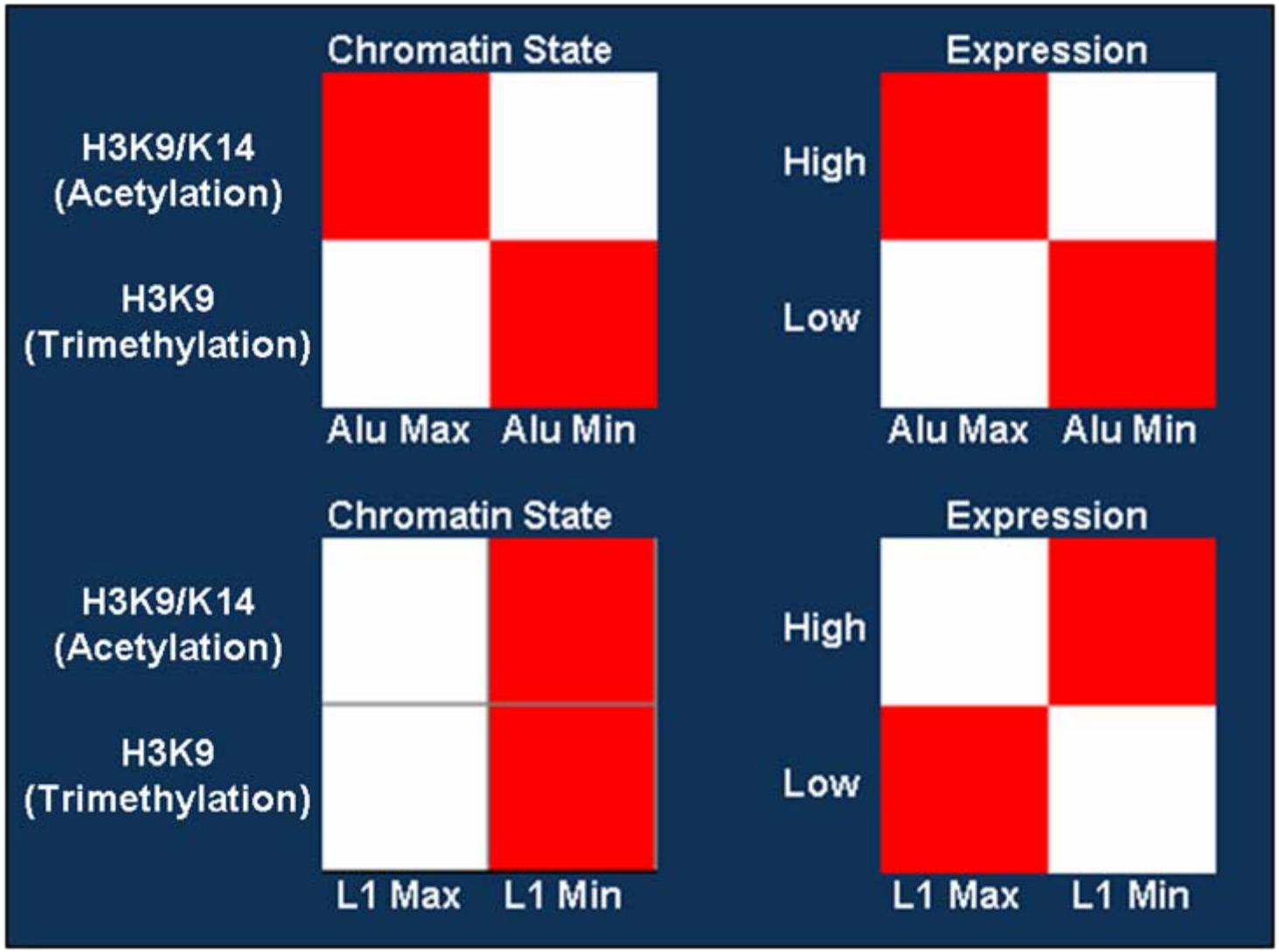


Divergent effects of Alu and L1 on gene expression





TE neighborhood and chromatin state (histone tail modifications)



Genetic regulatory effects of repetitive DNA

Sequence-specific TE-derived regulatory elements

TE-derived human regulatory sequences

- Analyzed several different classes of regulatory sequence for TE origins
- Proximal promoter regions – 25% TE+, 8% TE positions
- 5' & 3' UTRs – 3% TE+, 2.5% TE positions & 14.5% TE+, 7% TE positions
- Lower than for genome (45%) but higher than for CDS (2.5%, <1%)

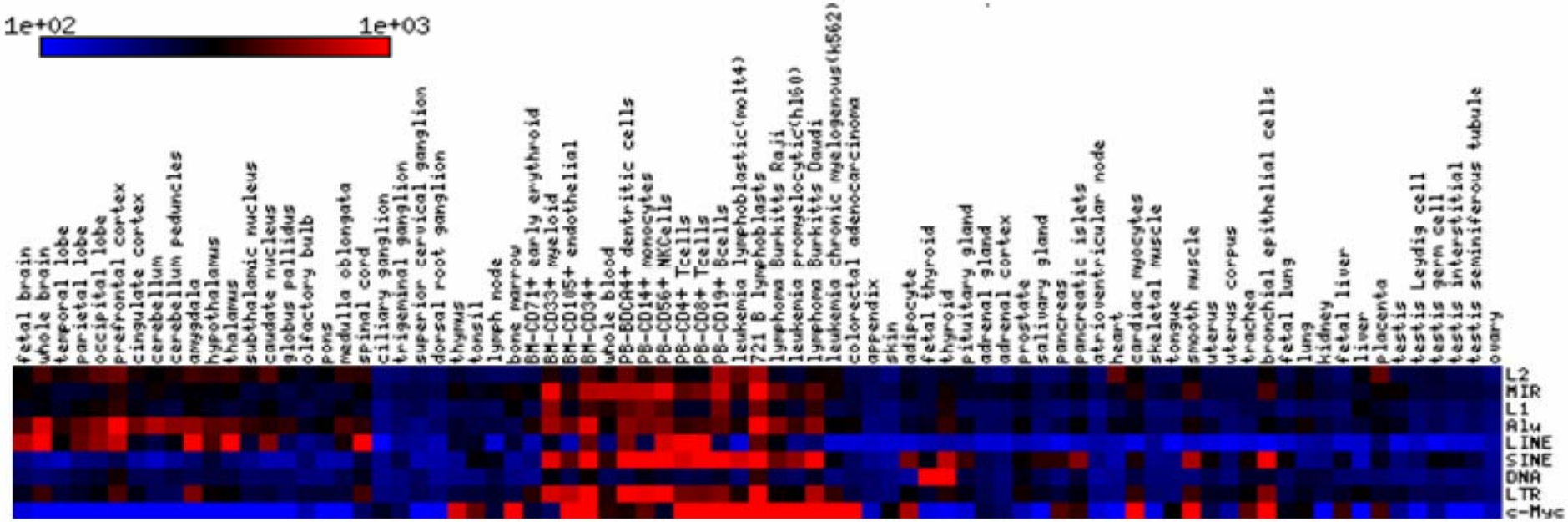
<u>Experimentally characterized sites</u>

- Cis-regulatory binding sites
  - o TRANSFAC - 846 sites from 288 genes
  - o 21 TE-derived (2.5%) from 13 genes (4.5%)
  - o Extrapolate to >1,000 human genes with TE-derived cis-sites
- Global regulatory sites
  - o LCRs (β-globin locus) derived from TEs
  - o Scaffold/matrix attachment regions (56% TE, 40% LINE >genome)

Jordan et al. (2003) Trends Genet 15: 95

High-throughput identification of TE-derived TFBS

- genome-scale identification of c-Myc using chromatin immunoprecipitation (ChIp-seq)
- co-locate with TEs and identify statistically significant hits to c-Myc binding site motif (PWM)
- thousands of TE-derived TFBS identified in this way (4,564)
- map to genes and evaluate expression patterns of genes with TE-derived c-Myc sites
- demonstrates regulatory activity (cancer-related) of TE-derived TFBS



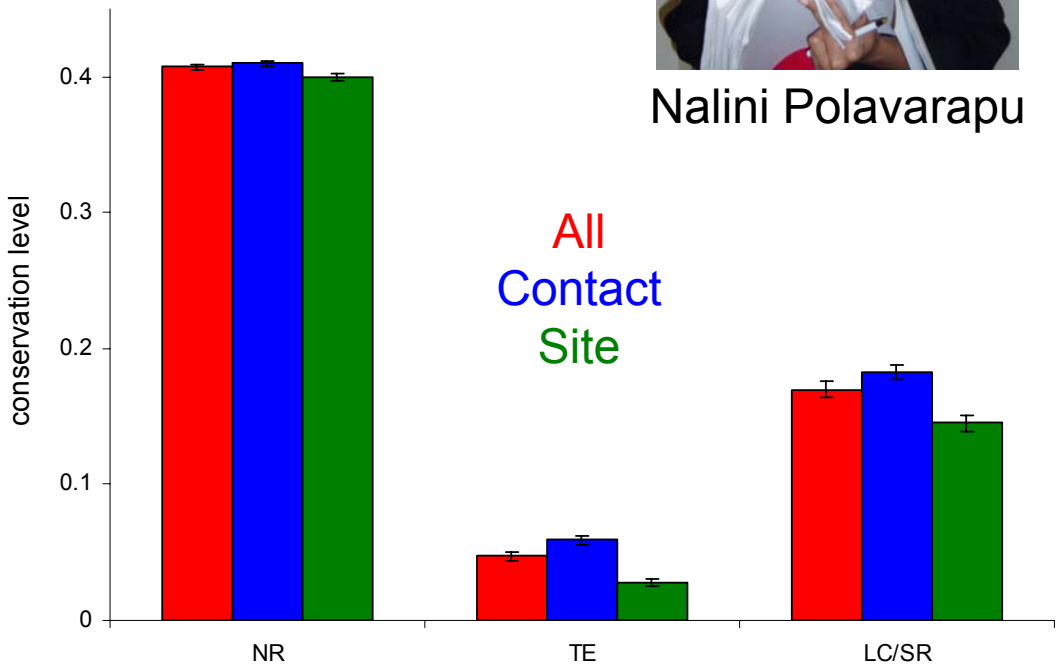
Accelerated evolution of repetitive DNA derived regulatory sites

- TEs are the most lineage-specific elements in eukaryotic genomes
- TEs are often found to be rapidly evolving
- TE-derived regulatory sites should be rapidly evolving
- TEs may provide a mechanism for driving regulatory divergence
- Phylogenetic footprinting methods will overlook TE-derived regulatory sites
- Evaluate for: cis-sites, HS-sites & miRNAs

Accelerated evolution of cis-sites derived from repetitive DNA

- 1,799 experimentally characterized cis-regulatory sites mapped to the human genome
- 182 co-located with repetitive DNA sequences: 79 TE & 103 LC/SR
- relative evolutionary rates (conservation levels) computed using whole genome alignments of 17 vertebrate species
- TE & LC/SR derived cis-sites are less conserved than non-repetitive sites
- residues in physical contact with trans factors are more conserved for all 3 classes of cis-sites
- suggests functional relevance of repetitive DNA derived sites

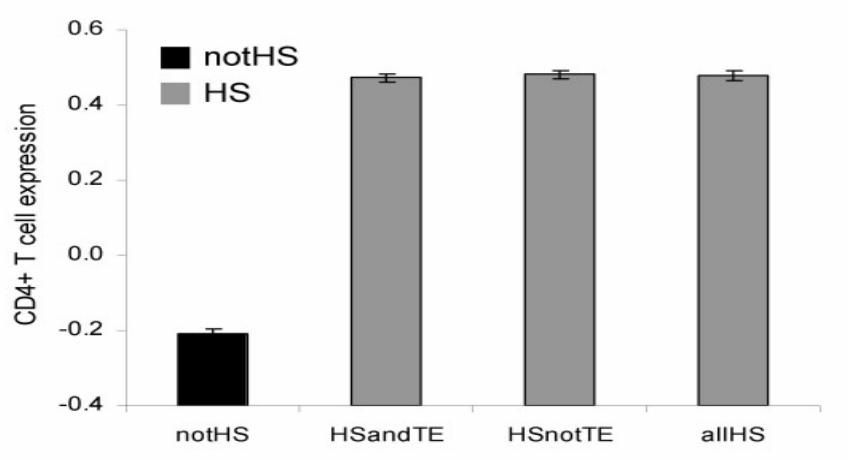
Nalini Polavarapu

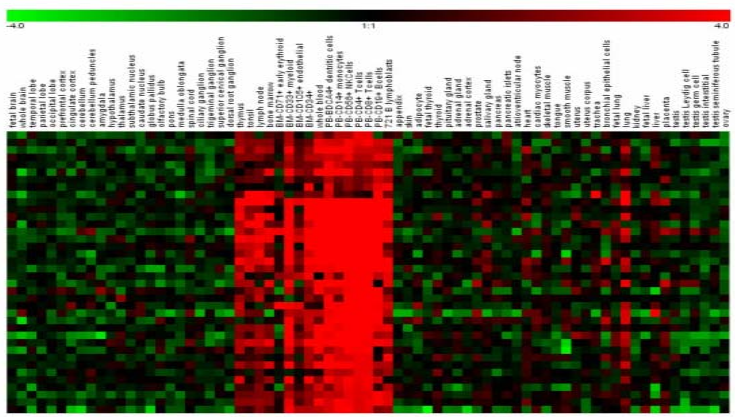


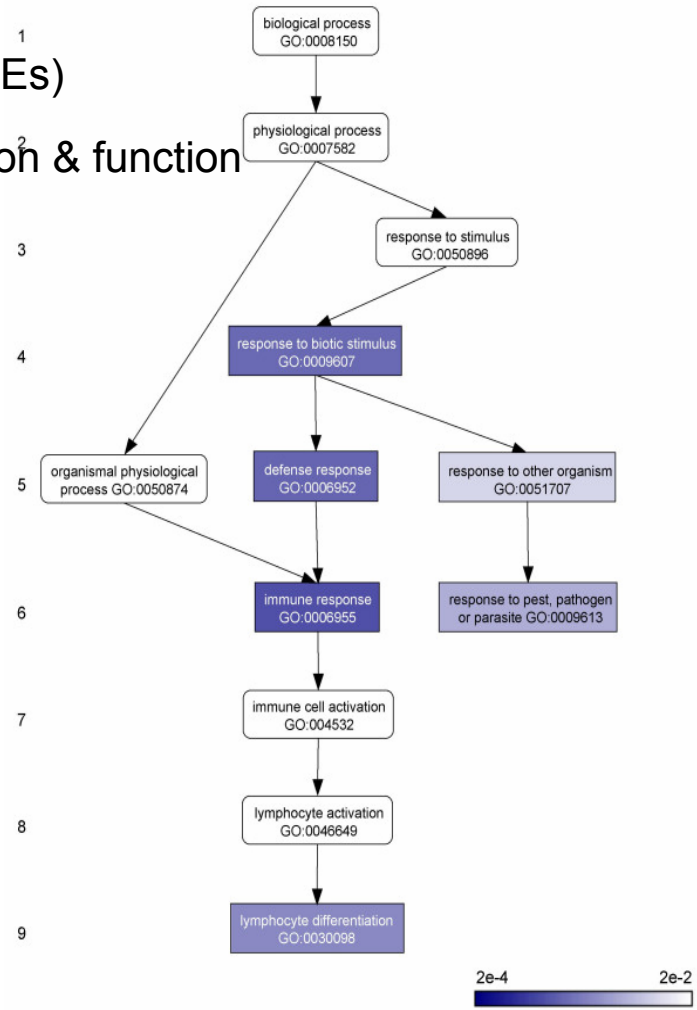
Polavarapu et al. (2008) BMC Genomics 9: 226

TE-derived DNaseI-hypersensitive sites

- DNaseI-hypersensitive (HS) sites identify regulatory regions
- 14,216 HS sites mapped to human genome
- 3,229 HS sites are TE-derived (11% of all positions from TEs)
- TE-derived HS sites are relevant wrt CD4+ T-cell expression & function



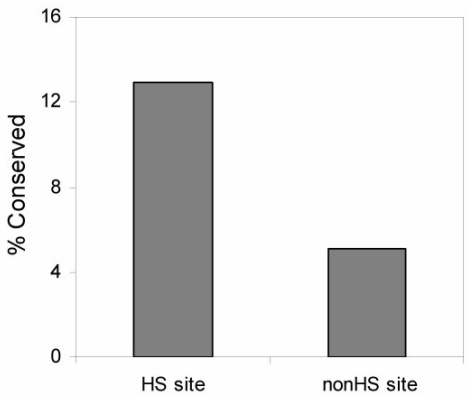


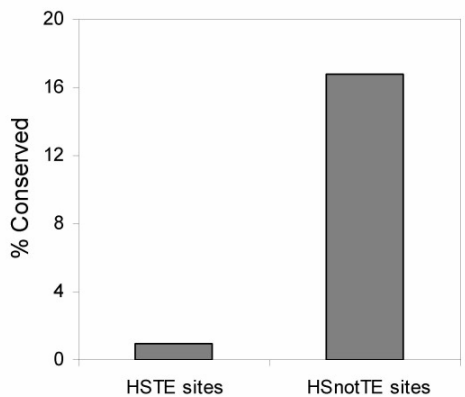


Mariño-Ramírez & Jordan (2006) Biol Direct 1: 20

Accelerated evolution of TE-derived HS sites

- Human HS sites are conserved [consistent with functional relevance]
- Human TE-derived HS sites are rapidly evolving
- Phylogenetic footprinting will overlook these





	HS	nonHS
PC	539,401	146,083,741
notPC	3,638,723	2,701,069,048

$\chi^2=5.2\times10^6$ $P=0$

	HSTE	HSnotTE
PC	4,281	535,120
notPC	450,283	3,188,440

$\chi^2=6.5\times10^5$ $P=0$

- Genes with TE-derived HS sites have higher levels of human-mouse ortholog expression divergence

Mariño-Ramírez & Jordan (2006) Biol Direct 1: 20

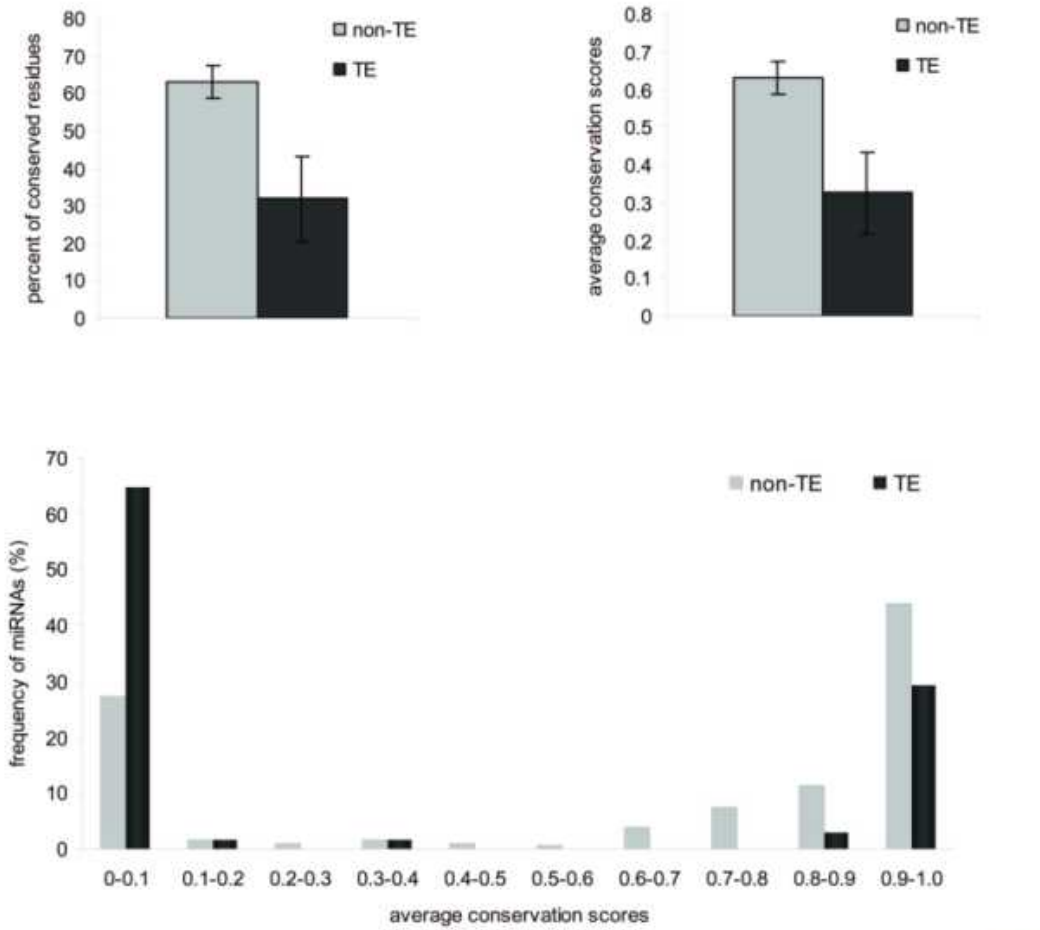
Human microRNAs from TEs

Jittima ‘Jing’ Piriyapongsa

- 462 human miRNAs from miRBase database (80% exp char, 20% orthologous)
- co-locate miRNA genes with TEs
- 55 TE-miRNA associations (12% of miRBase w/ 90% exp char)
- 49 intronic & 19 intergenic
- 50 >50% TE-derived
- several nested insertions
- *ab initio* prediction using conservation of secondary structure
- 85 novel TE-derived miRNA genes predicted

Piriyapongsa & Jordan (2007) PLoS ONE 2:e203
Piriyapongsa et al. (2007) Genetics 176: 1323

Accelerated evolution of TE-derived miRNAs

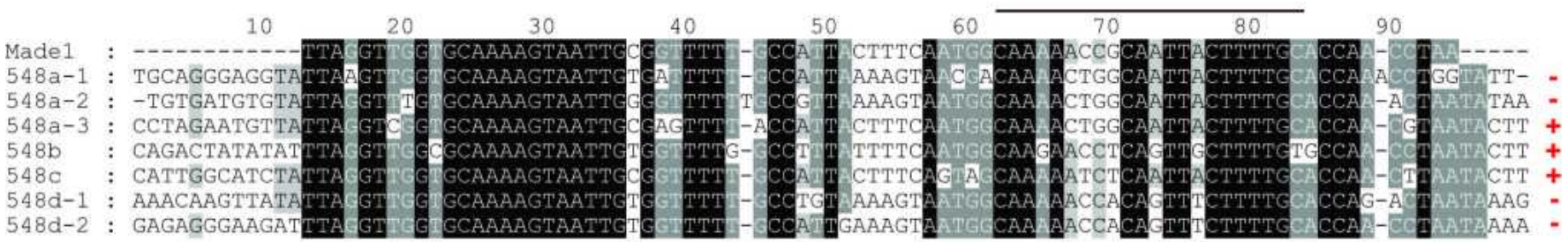


Piriyapongsa et al. (2007) Genetics 176: 1323

Paralogous family of human miRNAs from TEs

hsa-mir-548

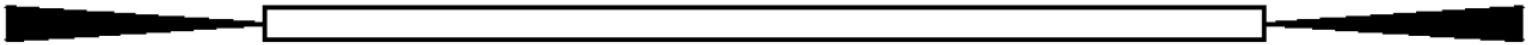
- family of 7 closely related miRNAs from the Made1 family
- recently experimentally characterized by SAGE related technique
- Made1 is a miniature-inverted repeat TE (MITE) family
- pri-mRNAs derived from elements inserted in both directions?
- Made1 elements are nearly perfect palindromes
  - o 37bp terminal inverted repeats (TIRs) with 6bp intervening



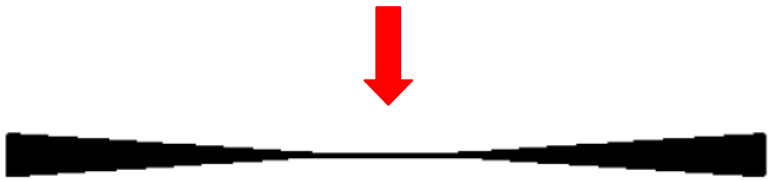
Cummins et al. (2006) PNAS 103: 3687
Piriyapongsa & Jordan (2007) PLoS ONE 2:e203

Derivative MITE genomic structure

Made1 genomic structure suggests mechanism for pri-miRNA formation



Autonomous DNA-type transposon (Hsmar1)

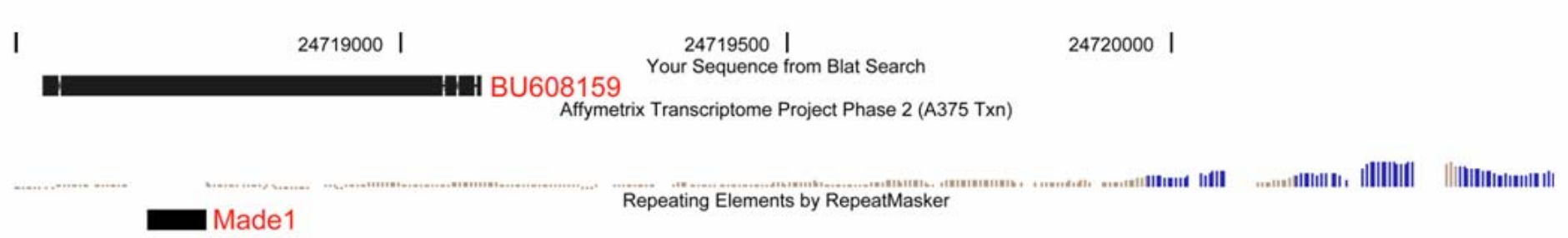


Non-autonomous MITE (MADE1)

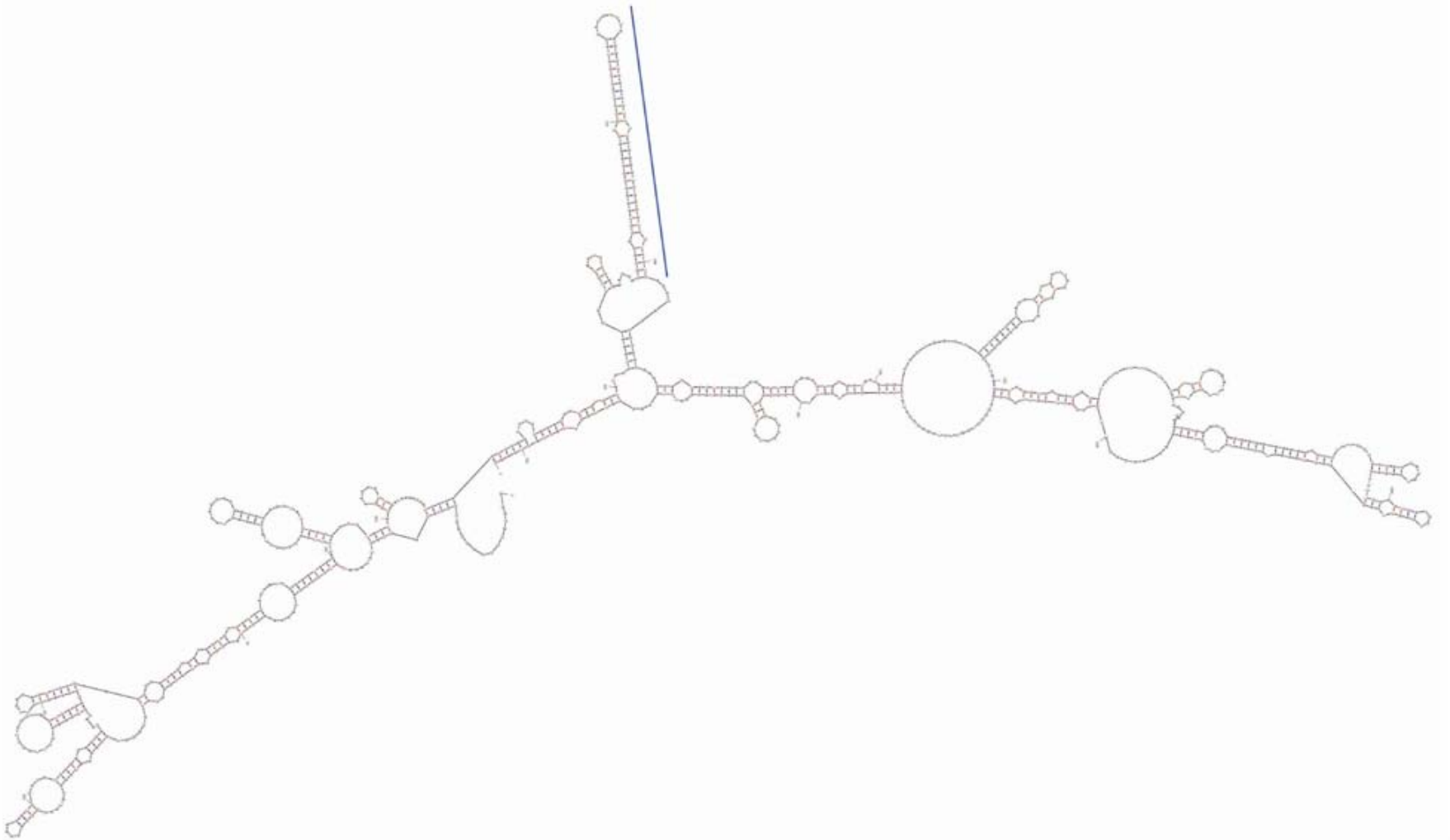
Predicted hairpin structure

Read-through expression of Made1

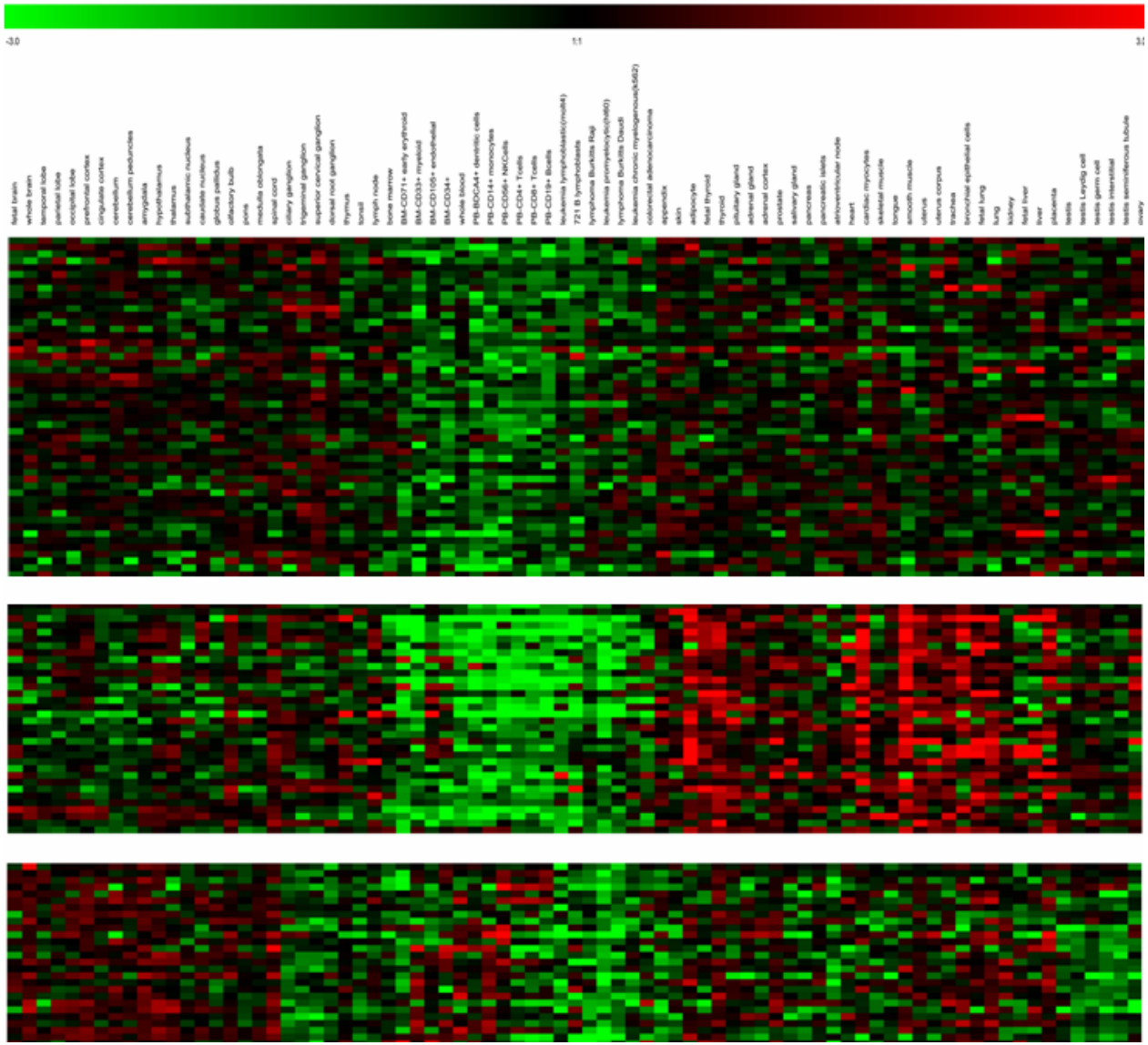
- Hundreds of Made1 ESTs can be found
- Transcription initiated from adjacent genomic positions



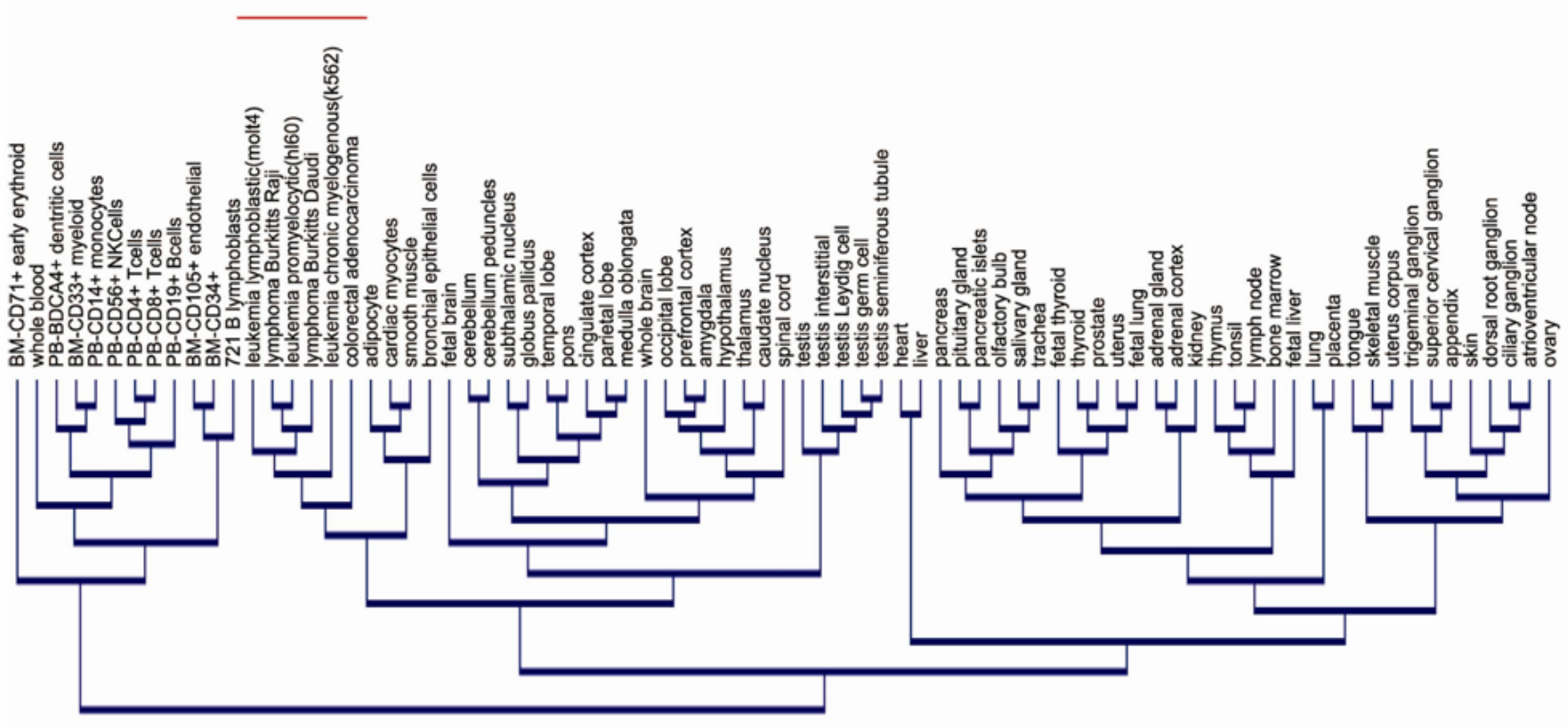
Secondary structure of Made1 containing transcript



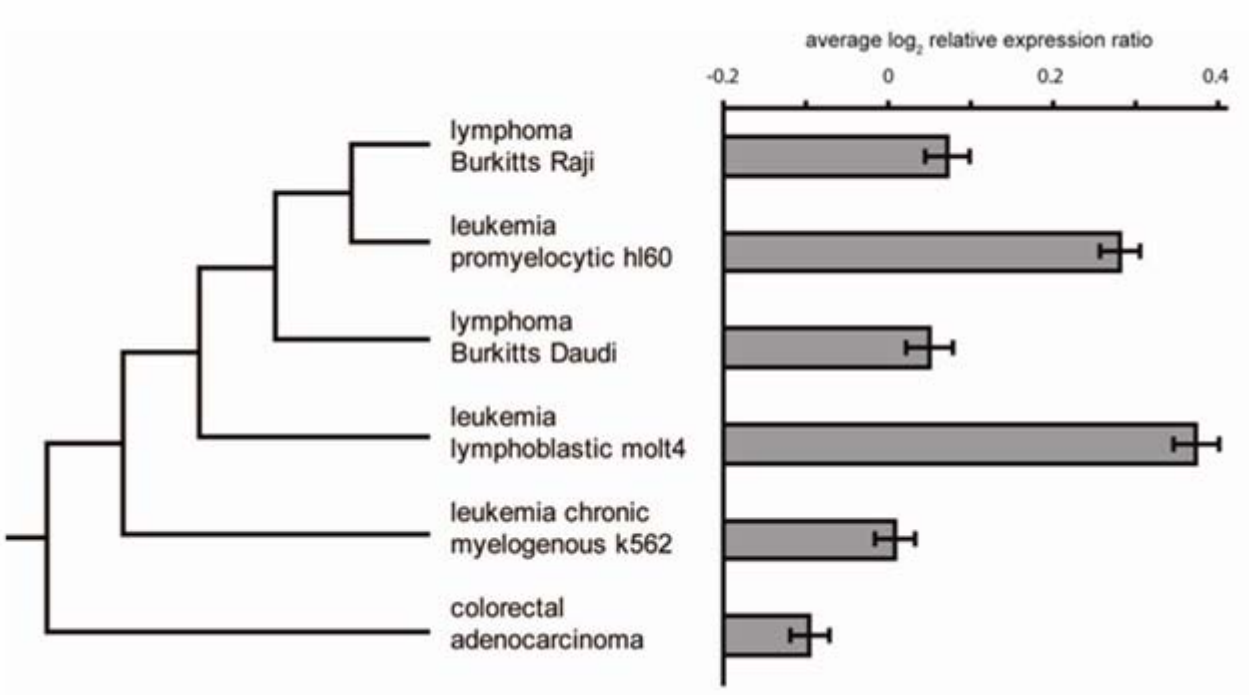
Potential cancer-related regulatory effects of hsa-mir-548



Cancer samples cluster together



Low expression for colorectal sample



Conclusions

- Repetitive DNA – TEs in particular – contribute many regulatory elements (epigenetic & genetic) to mammalian genomes, e.g. TFBS, promoter seqs, microRNAs

- Growing awareness of the connection between repetitive DNA and chromatin structure – along with regulatory implications

- TE-related regulatory sequences are functionally relevant but diverge rapidly between evolutionary lineages

- As such, they may play a role in driving regulatory divergence between evolutionary lineages and/or between normal and cancerous cells

The Selfish DNA theory is dead !

Long live the Selfish DNA theory !

Kreitman Bioessays (1996) 18: 678

Georgia Tech

BIOLOGY

Acknowledgements

Jittima Piriyapongsa

Ahsan Huda

ESBG members

Nalini Polavarapu

Daniel Gonzalez

NCBI

Leonardo Mariño-Ramírez

David Landsman

Igor B. Rogozin

Eugene V. Koonin

http://esbg.gatech.edu