



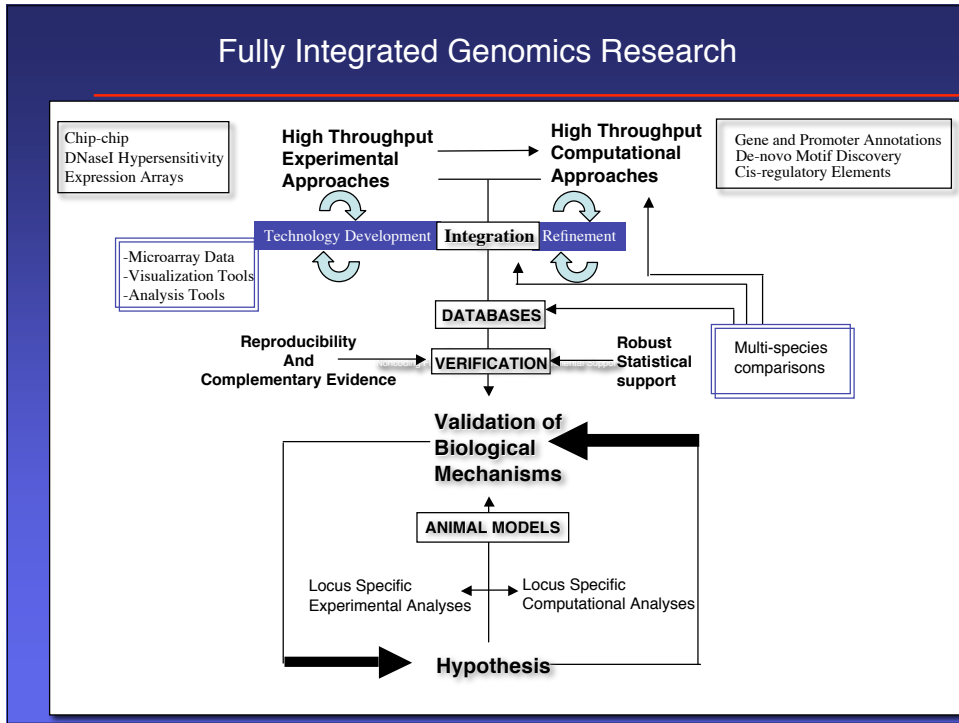
*Current Topics in Genome Analysis
Fall 2006*

*Week 5- Part I: Detection and Characterization of
Noncoding Functional Elements*

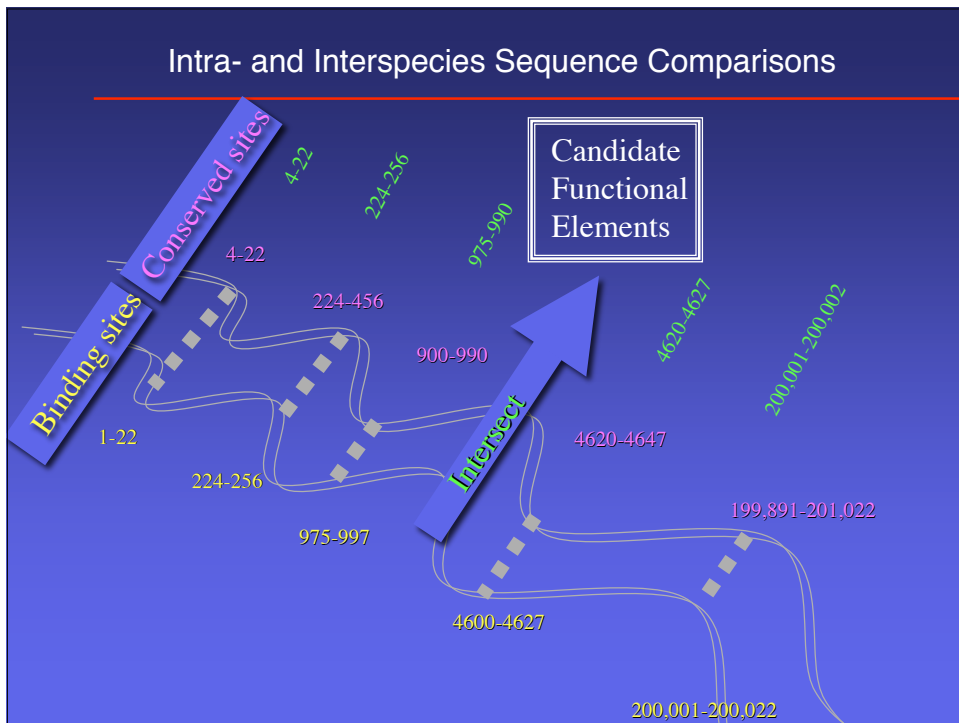
Laura Elnitski, Ph.D.



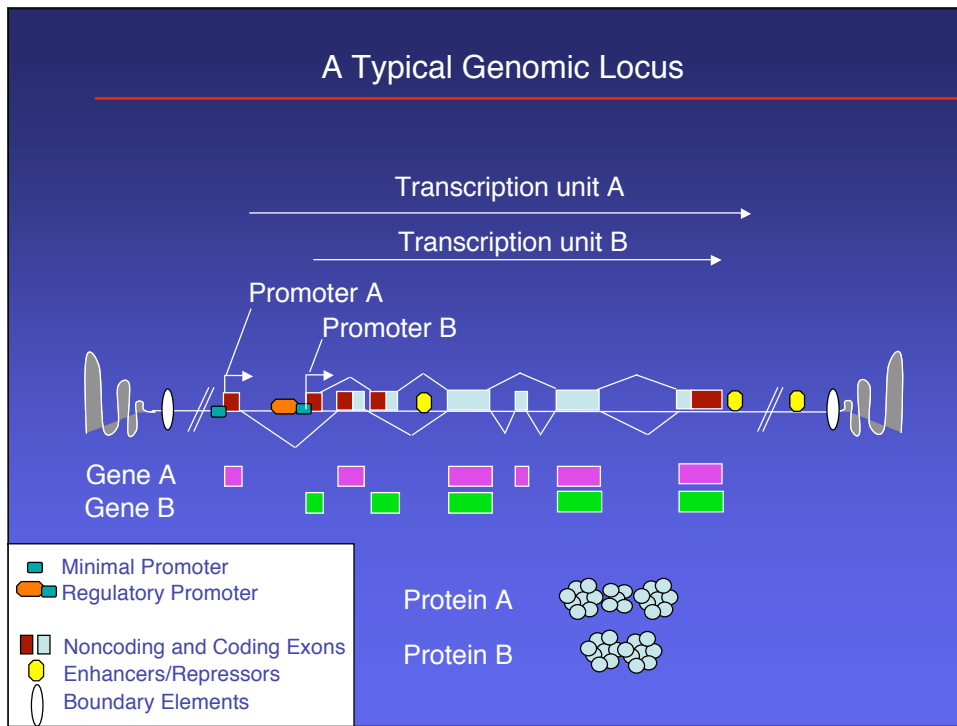
Fully Integrated Genomics Research



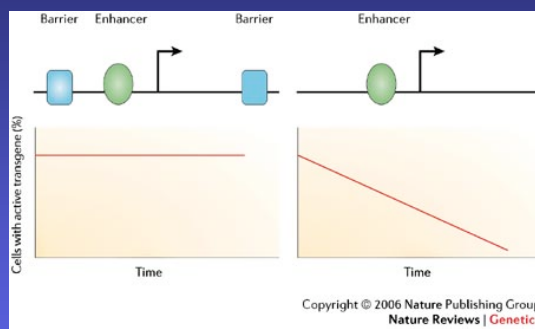
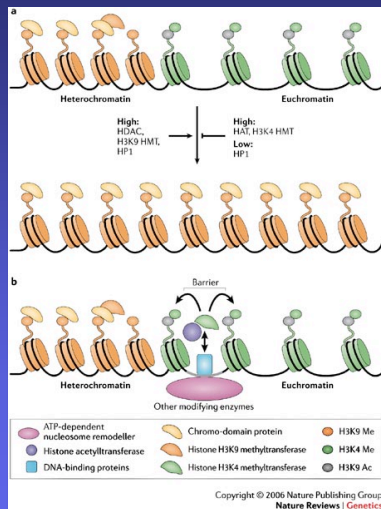
Intra- and Interspecies Sequence Comparisons



A Typical Genomic Locus



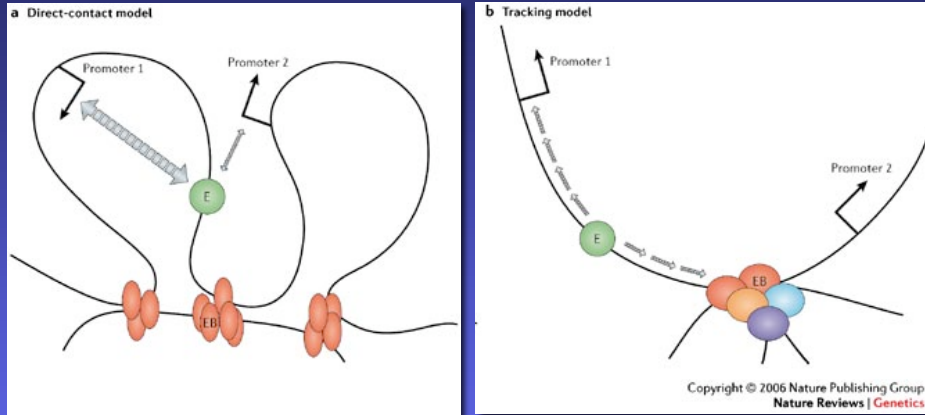
Boundary Elements



CTCF - 1st insulator
HS4 of the chicken beta-globin locus

Gaszner and Felsenfeld, 2006

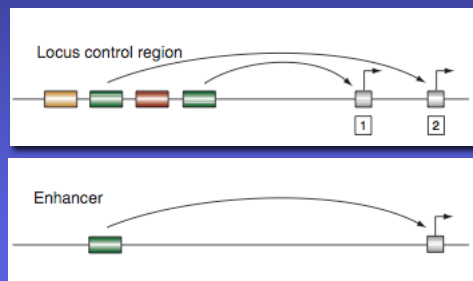
Boundary Elements / Enhancer Blockers



Can we find them through sequence specific searches?
Other indicators?
Chromatin structure
CTCF binding motifs
CTCF occupancy

Enhancers

Distally located regulatory elements, 1kb - 1 Mb
Locus control regions vs single gene regulators



Enhancers

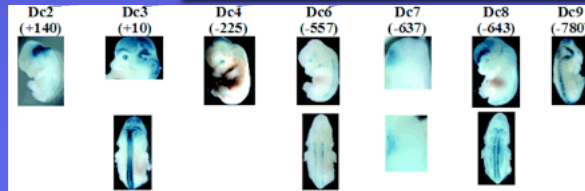
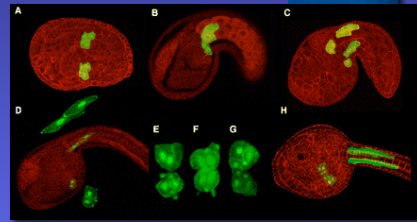
Many experimental systems using Luciferase or GFP assays:

Zebrafish
Drosophila
Chicken
Vertebrate cell lines

Ciona

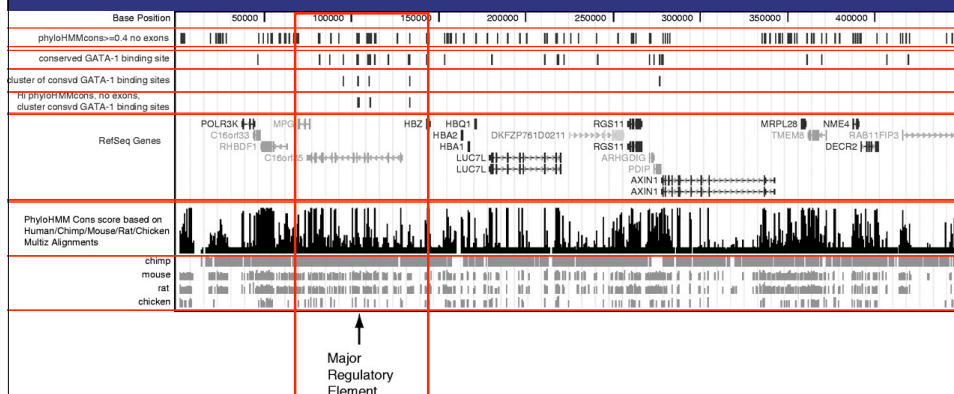


Xenopus



Transgenic mice

Computational Prediction of Enhancers



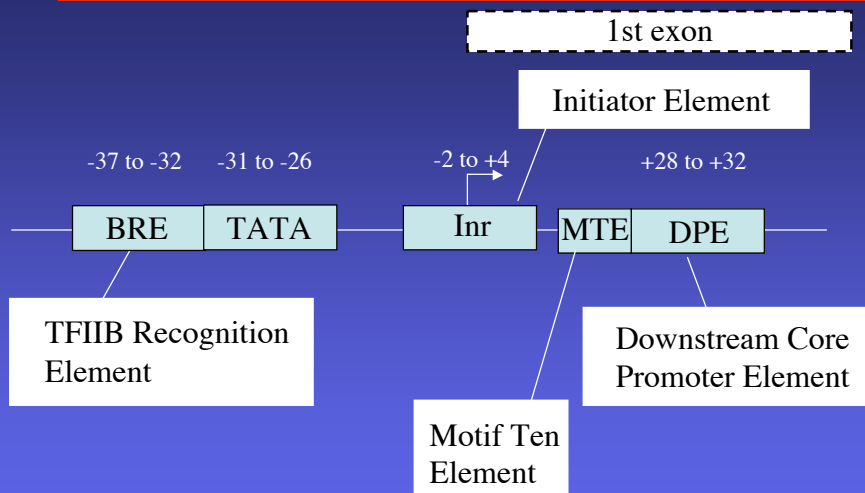
Core Promoters

- Proximal to transcription start site (+/- 35 bp)
- Determine timing of gene expression during development
- Interact with the pre-initiation complex

Riken CAGE Data Cap Analysis of Gene Expression
<http://fantom.gsc.riken.go.jp/>

dbTSS Database of Transcription Start Sites
<http://dbtss.hgc.jp/>

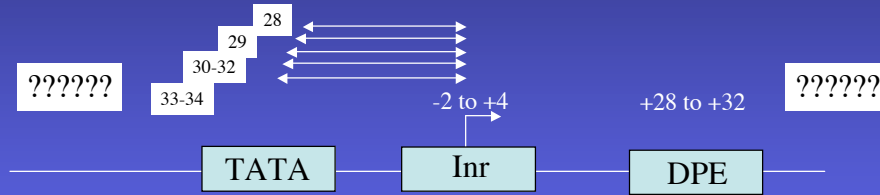
Core Promoters



Core Promoters

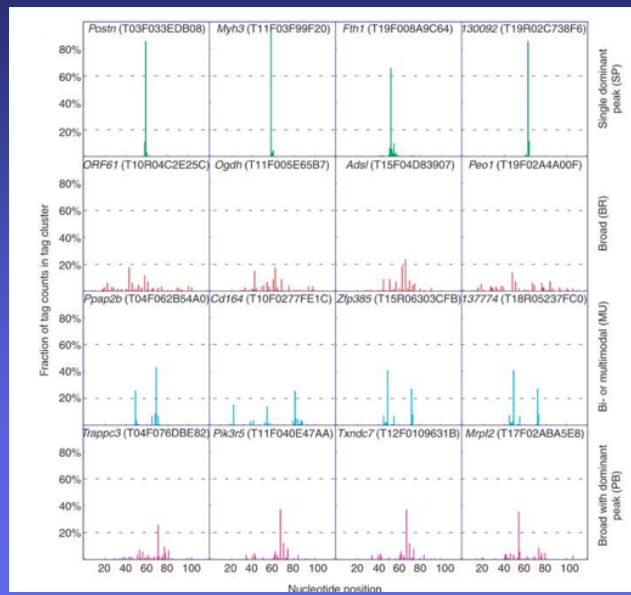
Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters

Ponjavic, et al. 2006



Sequence searching
Stratification of datasets

Core Promoters



Carninci et al. 2006

SP

BR

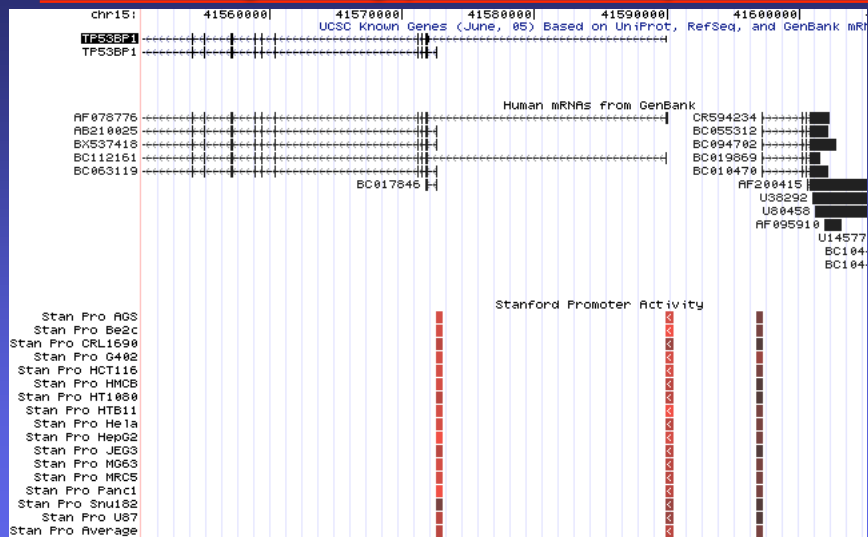
MU

PB

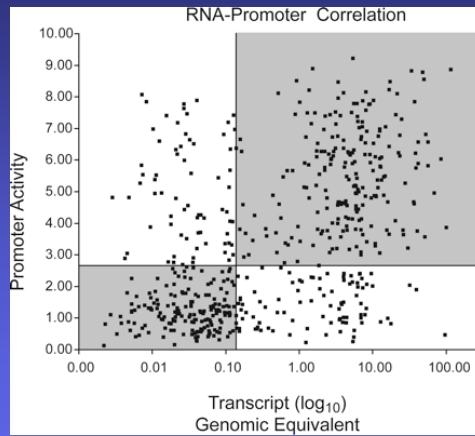
Core Promoters

- The transcriptional regulation and evolutionary plasticity of CpG island-associated promoters is also linked to epigenetic control of transcriptional activity. These promoters are rapidly evolving in mammals.
- The BR classes, most commonly based on CpG islands, represent the majority of mammalian promoters.
- Classical TATA-box promoter architecture represents a minority of the set of mammalian promoters in mouse and humans. This class is commonly associated with tissue-specific genes and high conservation across species

High Throughput Promoter Analyses



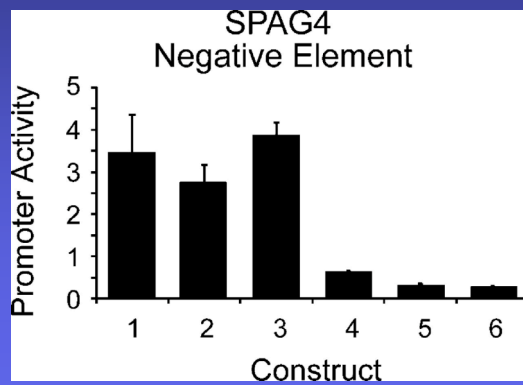
High Throughput Promoter Analyses



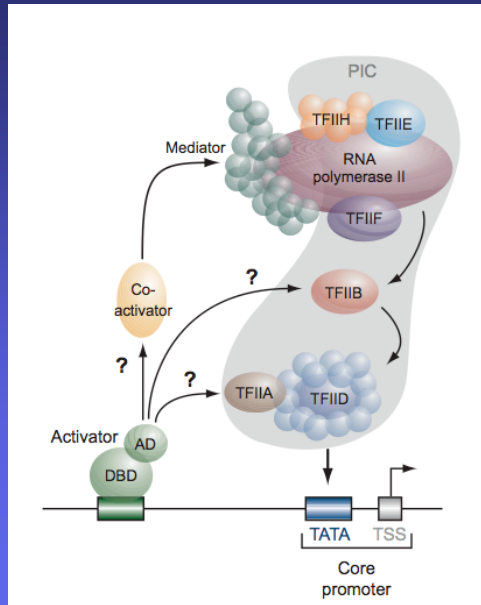
Cooper et al. Genome Res. 2006

Extended Promoters

- Upstream elements (~ 500-1 kb) that synergize with basal promoter
- Augment the level of expression or choice of tissue specificity
- Separation from enhancers? (negative elements- Cooper et al 2006)

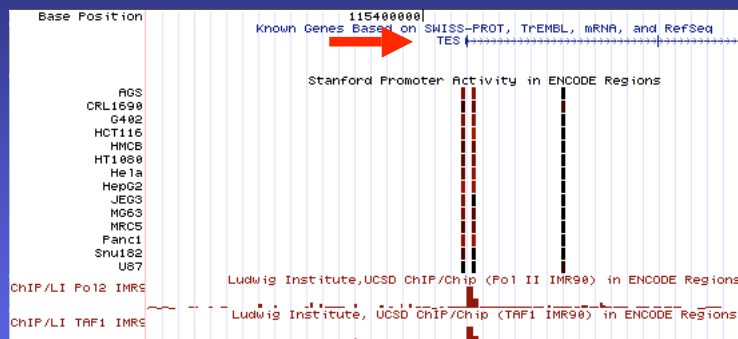


Promoter Classes



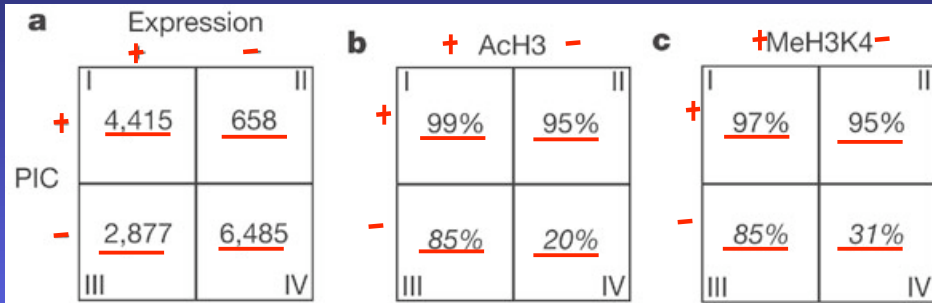
PIC =
Pre-initiation
complex

Promoter Features



UCSC Genome Browser

Promoter Classes



Kim et al., Nature 2005

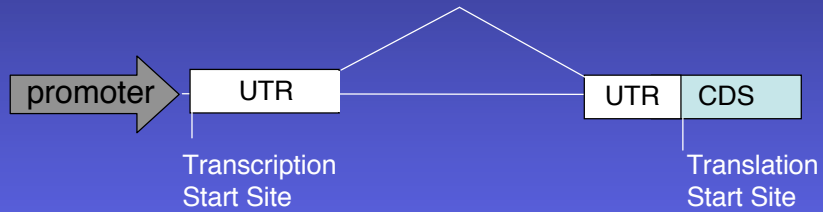
Promoter Classes

As expected, formation of the PIC on promoters leads to transcription

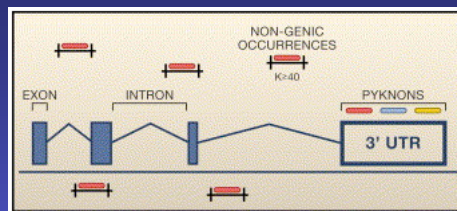
- When PIC does not correlate with expression
 - (a) PIC formation is not sufficient to initiate transcription or mRNAs are post-transcriptionally degraded.
 - (b) Transcription occurs, but the PIC assembles transiently, weakly or only during the early stage of fibroblast differentiation.
- Acetylation and methylation marks can correspond to both active and poised genes, defining the transcriptome capacity of the cell, requiring additional transcription regulators and machinery to collaborate with these epigenetic markers

5' UTRs

- Regulatory binding sites
- Small ORFs
- Spliced introns
- Secondary structure



3' UTRs- pyknons



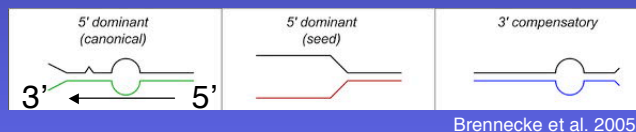
- from the Greek word meaning "dense" Rigoutsos et al, Meynert and Birney 2006
- typically 60-80 nucleotides long
- present in coding and noncoding DNA
- spacing of 18 to 22 nucleotides between copies
- 1/3 predicted to form double-stranded, energetically stable, hairpin-shaped RNA secondary structures with reverse complement copies.
- subsume approximately 40% of the known microRNA sequences, thus suggesting a possible link with posttranscriptional gene silencing and RNA interference.

3' UTRs- microRNA target sites

- MicroRNAs (miRNAs) are small non-coding RNAs that serve as post-transcriptional regulators of gene expression in plants and animals.

MicroRNA target sites:

- 8-base length
- end with the nucleotide 'A'
- match through Watson–Crick pairing

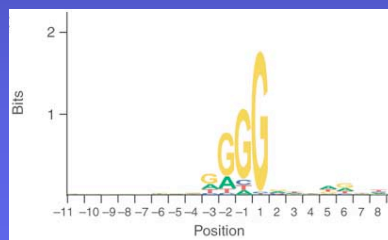


<http://mirna.imbb.forth.gr/microinspector/>

<http://microrna.sanger.ac.uk/targets/v3/>

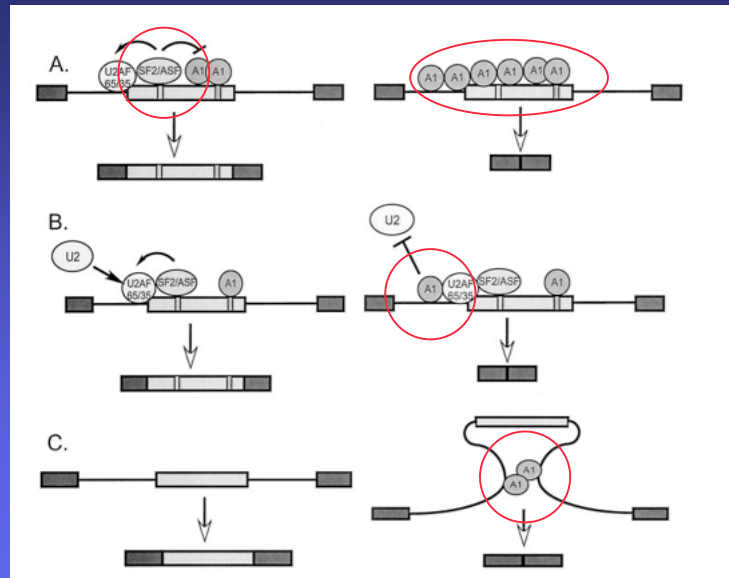
3' UTRs

- Polyadenylation signals
- Various (A + T)-rich elements involved in controlling mRNA stability and degradation
- Binding sites for the Puf family of RNA-binding proteins (yeast) with possible human homologues
- Promoters



Carninci et al, 2006

Splicing Elements



Douglas L. Black 2003

Splicing Elements

- The same ESR sequence can function as an enhancer in one exon and a silencer in the other.
- Orthologous exons that are alternatively spliced in both human and mouse are more conserved than constitutively spliced ones
- This may be due to the presence of ESRs and is consistent with previous studies showing selective pressure against synonymous mutations in alternatively spliced exons

Goren et al. 2006

Splicing Enhancers

ESE FINDER

<http://rulai.cshl.edu/tools/ESE/>

RESCUE-ESE

<http://genes.mit.edu/burgelab/rescue-ese/>

Cis-acting elements bound by *trans*-acting factors

Sequence specific patterns recognized and bound by proteins

May mediate secondary interactions over long distances

May utilize proteins that

- do not directly bind DNA
- alter the DNA temporarily

Are impacted by the spatial positioning of nucleosomes

Predict them genome wide
Find mutations that affect them
Understand the biological role of proteins that bind there

Predicting transcription factor binding sites

Pre-mapped

- False positives
- False negatives

ab initio tools

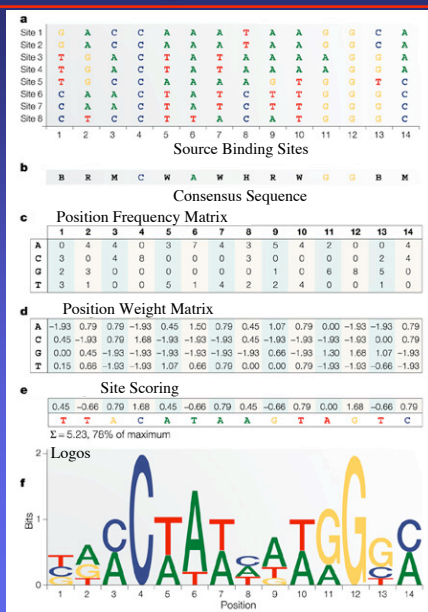
- See PDF table

Additional considerations

- Sequence conservation
- Evolutionary constraint
- Regulatory potential
- Novel Motifs

Clusters of interacting factors
Tissue specific factors

Binding Motif Notation



Resources for detecting regulatory sites

Service	Site
Whole genome binding site predictions	GALA Browser
	UCSC Genome Browser
	rVista
Position-weight matrix repositories	JASPAR
	SELEXdb
	TRANSFAC
Experimental data repositories	CAGE Basic/Analysis Databases
	COMPEL
	EPD
	GenBank
	MPromDb
	OMGProm
	TRRD

Resources for detecting binding sites

Pattern matching	AliBaba2
	AlignACE
	ANN-Spec
	Bayesian Phylogenetic Footprint
	cisRED
	CONreal
	ConSite
	CompelPatternSearch
	DoOP
	Dragon ERE Finder
	ECR browser
	Mapper
	MatInspector
	MDscan -
	MotifViz
	P-Match
	PROMO
	PromoterPlot
	RSAT
	SeSiMCMC
	SiteSeer
	TESS
	TFbind
	TFSEARCH
Toucan	
TRED	

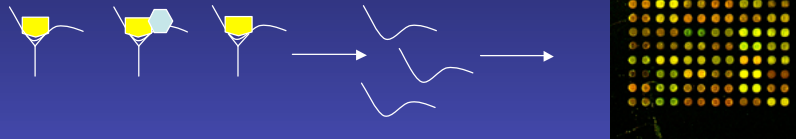
Resources for detecting binding sites

Pattern discovery	Consensus
	Dcode.org
	Ensembl.org
	Evoprinter
	FootPrinter
	Gibbs Motif Sampler
	GLAM
	MITRA
	motifSampler
	MSCAN
	Oligo/Dyad-analysis
	oPOSSUM
	Weeder Web
	YMF
Target Explorer	

Resources for detecting binding sites

Cluster detection	Cluster Buster
	CREME
	Improbizer
	ModuleSearcher
	MEME
	Over-represented Transcription Factor Binding Site Prediction Tool (OTFBS)
	TraFaC
	TFBind
	TRANSCompel

ChIP-chip



- Detects Protein-DNA Interactions - Direct
- Protein-Protein Interactions - Indirect

ENCODE Consortium

31 ChIP datasets of 18 sequence specific factors were examined for enrichment in motifs.

Resources for detecting binding sites

ChIP-chip	Chip-chip antibodies
	Chip-chip protocols
	MPEAK
	PeakFinder
Transcription factor information	ooTFD
PWM comparisons	T-Reg Comparator

Resources for ChIP-chip data

Resource	URL
ArrayExpress	www.ebi.ac.uk/arrayexpress
GALAXY2	www.bx.psu.edu
GEO	www.ncbi.nlm.nih.gov/geo
ENCODEdb	http://research.nhgri.nih.gov/ENCODEdb
Ensembl	www.ensembl.org
UCSC	www.genome.ucsc.edu

ChIP-chip Results

- The factors whose binding sites are most enriched at the 5' ends of genes include histone modifications, TAF1, RNA PolIII with hypo-phosphorylated C terminal domain
- E2F2, a sequence-specific factor that regulates the expression of many genes at the G1 to S transition, is often tightly associated with TSSs.
- Repressive elements CTCF and SIRT1, and PU1 are underrepresented near the TSSs.
- Several factors had no enriched TRANSFAC motif nor could a motif be discovered using *ab initio* methods.

ENCODE 2006

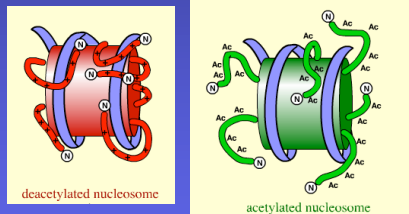
Detectable Histone Modifications

Gross rearrangements : movement or removal

Detected by DNase I sensitivity

Reversible alterations: methylation, acetylation, phosphorylation

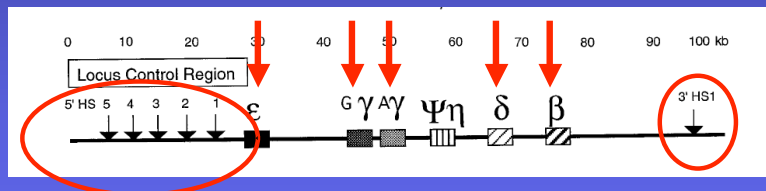
Detected by ChIP assays



http://www.broad.harvard.edu/chembio/lab_schreiber/anim/animations/hdac.html

Chromatin and Gene Expression

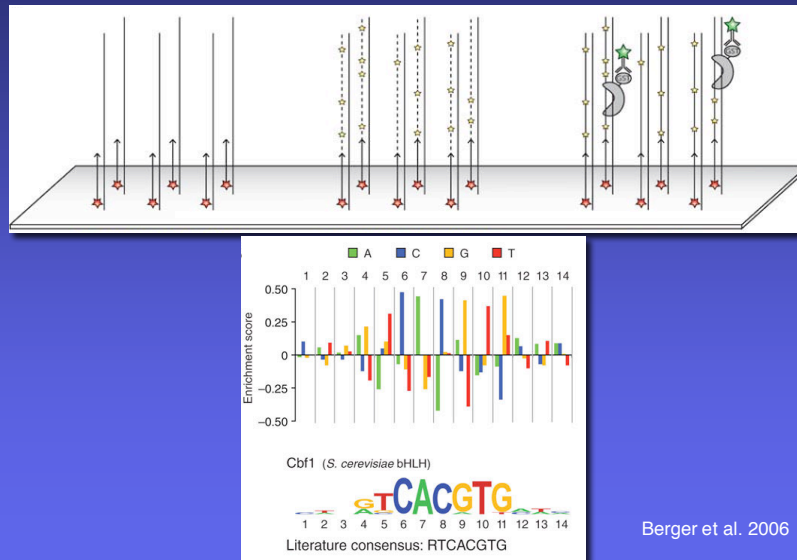
- Changes in chromatin conformation can be measured by sensitivity to DNase I
- DNase I *hypersensitivity* is an indicator of relaxed chromatin
- Often seen prior to detectable gene expression



Novel Motifs

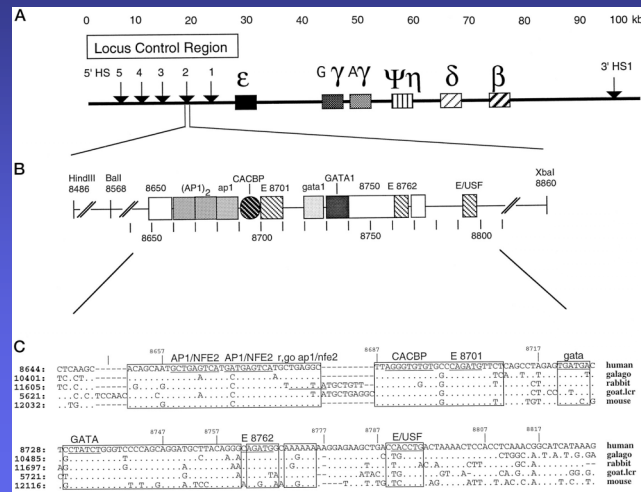
- Certain TFs, such as Sp1 and C/EBP, bind to target sequences that vary widely.
- Evolutionary pressure to retain such binding sites is minimal owing to the high likelihood that alternative sites will be available within a regulatory region.
- This indicates that there might be two subtypes of TFBS: highly selected sites that rarely occur by chance and auxiliary sites that are available by convenience.
- Phylogenetic footprinting methods will be well suited for binding sites of TFs with greater binding specificity.

Protein Binding Arrays



Novel Motifs

Phylogenetic Footprinting Shadowing



Characteristics Implicating Function

- Location in genome
- Conservation in a multiple sequence alignment
- Evolutionary Constraint
- Predictive tracks
- Clusters of transcription factor binding sites
- Experimental evidence
- Chromatin structure

Functional multiplicity

GALAXY2 Server

Enables Integrative Genomic Analysis

Extracting Orthologous Sequences
Phylogenetic Tree Construction
Basic Statistics
Operations
 Complement
 Restrict
 Merge overlapping regions
 Intersect
 Union
 Join Lists
 Cluster
 Proximity
 Subtract

<http://www.bx.psu.edu/>

Questions / Follow-up

elnitski@mail.nih.gov Re: Current Topics

Supplementary Tables

Table 1(A). Web Servers for *in silico* predictions

Service	Site	URL
Whole genome binding site predictions	GALA Browser	www.bx.psu.edu
	UCSC Genome Browser	genome.ucsc.edu
	rVista	rvista.dcode.org
Position-weight matrix repositories	JASPAR	jaspar.cgb.ki.se
	SELEXdb	www.mgs.bionet.nsc.ru/mgs/systems/selex/
	TRANFAC	www.gene-regulation.com
Experimental data repositories	CAGE Basic/Analysis Databases	fantom3.gsc.riken.jp
	COMPEL	compel.bionet.nsc.ru/new/compel/compel.html
	EPD	www.epd.isb-sib.ch
	GenBank	www.ncbi.nlm.nih.gov
	MPromDb	bioinformatics.med.ohio-state.edu/MPromD
	OMGProm	bioinformatics.med.ohio-state.edu/OMGProm
	TRRD	www.mgs.bionet.nsc.ru/mgs/gnw/trrd/
Pattern matching	AliBaba2	www.gene-regulation.com/pub/programs/alibaba2/index.html
	AlignACE	Atlas.med.harvard.edu
	ANN-Spec	www.cbs.dtu.dk/services/DNAarray/ann-spec.php
	Bayesian Phylogenetic Footprint	bayesweb.wadsworth.org/cgi-bin/bayes_align12.pl
	cisRED	www.cisred.org
	CONreal	conreal.niob.knaw.nl
	ConSite	mordor.cgb.ki.se/cgi-bin/CONSITe/consite
	CompelPatternSearch	compel.bionet.nsc.ru/FunSite/CompelPatternSearch.html
	DoOP	doop.abc.hu
	Dragon ERE Finder	sdmc.lit.org.sg/ERE-V2/index
	ECR browser	ecrbrowser.dcode.org
	Mapper	bio.chip.org/mapper
	MatInspector	www.genomatix.de/products/MatInspector
	MDscan -	ai.stanford.edu/~xslui/MDscan/
	MotifViz	biowulf.bu.edu/MotifViz
	P-Match	www.gene-regulation.com/cgi-bin/pub/programs/pmatch/bin/p-match.cgi
	PROMO	www.lsi.upc.es/~alggen
	PromoterPlot	promoterplot.fmi.ch
	RSAT	rsat.ulb.ac.be/rsat/
	SeSiMCMC	favorov.hole.ru/gibbslfm
	SiteSeer	rocky.bms.umist.ac.uk/SiteSeer/
	TESS	www.cbil.upenn.edu/tess
	TFbind	tfbind.ims.u-tokyo.ac.jp
	TFSEARCH	www.cbrc.jp/research/db/TFSEARCH.html
	Toucan	homes.esat.kuleuven.be/~saerts/software/toucan.php
	TRED	rulai.cshl.edu/TRED

Table 1(B). Web Servers for *in silico* predictions

Pattern discovery	Consensus	bifrost.wustl.edu/consensus
	Dcode.org	www.dcode.org
	Ensembl.org	www.ensembl.org
	Evoprinter	evoprinter.ninds.nih.gov
	FootPrinter	bio.cs.washington.edu/software.html
	Gibbs Motif Sampler	bayesweb.wadsworth.org/gibbs/gibbs.html
	GLAM	Zlab.bu.edu/glam
	MITRA	www1.cs.columbia.edu/compbio/mitra/
	motifSampler	www.esat.kuleuven.ac.be/~dna/Biol/Software.html
	MSCAN	mscan.cgb.ki.se/cgi-bin/MSCAN
	Oligo/Dyad-analysis	rsat.scmbb.ulb.ac.be/rsat/
	oPOSSUM	www.cisreg.ca
	Weeder Web	www.pesolelab.it/
	YMF	bio.cs.washington.edu/software.html
	Target Explorer	trantor.bioc.columbia.edu/Target_Explorer
Cluster detection	Cluster Buster	zlab.bu.edu/cluster-buster
	CRÈME	creme.dcode.org
	Improbizer	www.cse.ucsc.edu/~kent/improbizer/improbizer.html
	ModuleSearcher	homes.esat.kuleuven.be/~saerts/software/modulesearcher.html
	MEME	meme.sdsc.edu/meme/
	Over-represented Transcription Factor Binding Site Prediction Tool (OTFBS)	www.bioinfo.tsinghua.edu.cn/%7Ezhengjsh/OTFBS/index.html
	TraFaC	trafac.cchmc.org/trafac/index.jsp
	TFBind	tfbind.ims.u-tokyo.ac.jp
	TRANSCompel	www.gene-regulation.com/pub/databases.html#transcompel
ChIP-chip	Chip-chip antibodies	www.chiponchip.org/antibody.html
	Chip-chip protocols	genomecenter.ucdavis.edu/farnham/farnham/protocol.html
	MPEAK	www.stat.ucla.edu/~zmdl/mpeak/
	PeakFinder	research.stowers-institute.org/jeg/2004/cohesin/peakfinder
Transcripton factor information	ooTFD	www.ifti.org/oofdf
PWM comparisons	T-Reg Comparator	treg.molgen.mpg.de

* Note – tools or servers may belong to more than category in the table. Additional tools are frequently provided at each site. Web sites may occasionally be down due to server errors or discontinued altogether.

Table 2. Resources for Experimental protocols

Assay	Reference (Chapter)	
	Molecular Cloning (Sambrook and Russell 2001)	Current Protocols in Mol. Biol. (Ausubel et al. 2005)
DNase I Hypersensitivity	17.18	21
EMSA	17.13	12
DNase I Protection	17.4	21
ChIP*	N/A	21
Transfection	16	9

*Note: Additional information is available in (Allis and Wu, 2003).

Table 3. Web servers and repositories for publicly available high-throughput ChIP-chip binding data

Resource	URL
ArrayExpress	www.ebi.ac.uk/arrayexpress
GALAXY2	www.bx.psu.edu
GEO	www.ncbi.nlm.nih.gov/geo
ENCODEdb	http://research.nhgri.nih.gov/ENCODEdb
Ensembl	www.ensembl.org
UCSC	www.genome.ucsc.edu

Allis C.D. and C. Wu. 2003. *Chromatin and Chromatin Remodeling Enzymes, Part B*. Methods in Enzymology. Elsevier Inc.

Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, K.A., Struhl, K. 2005. *Current Protocols In Molecular Biology*. John Wiley & Sons, Inc.

Sambrook, J. and Russell, D.W. 2001. *Molecular Cloning: A Laboratory Manual (3-Volume Set)*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.