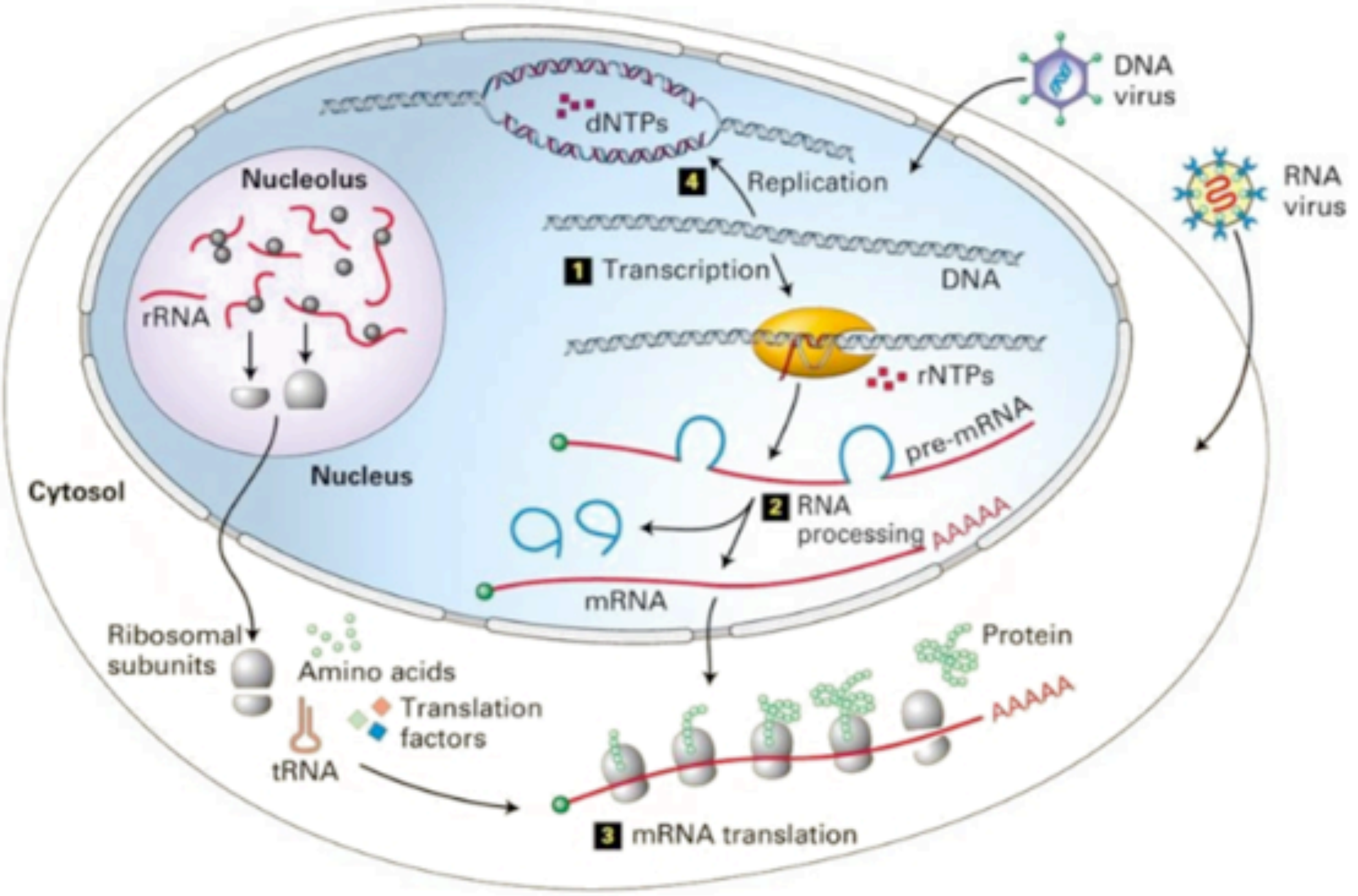
Discovering gene regulatory control using ChIP-chip and ChIP-seq

"An introduction to gene regulatory control, concepts and methodologies"

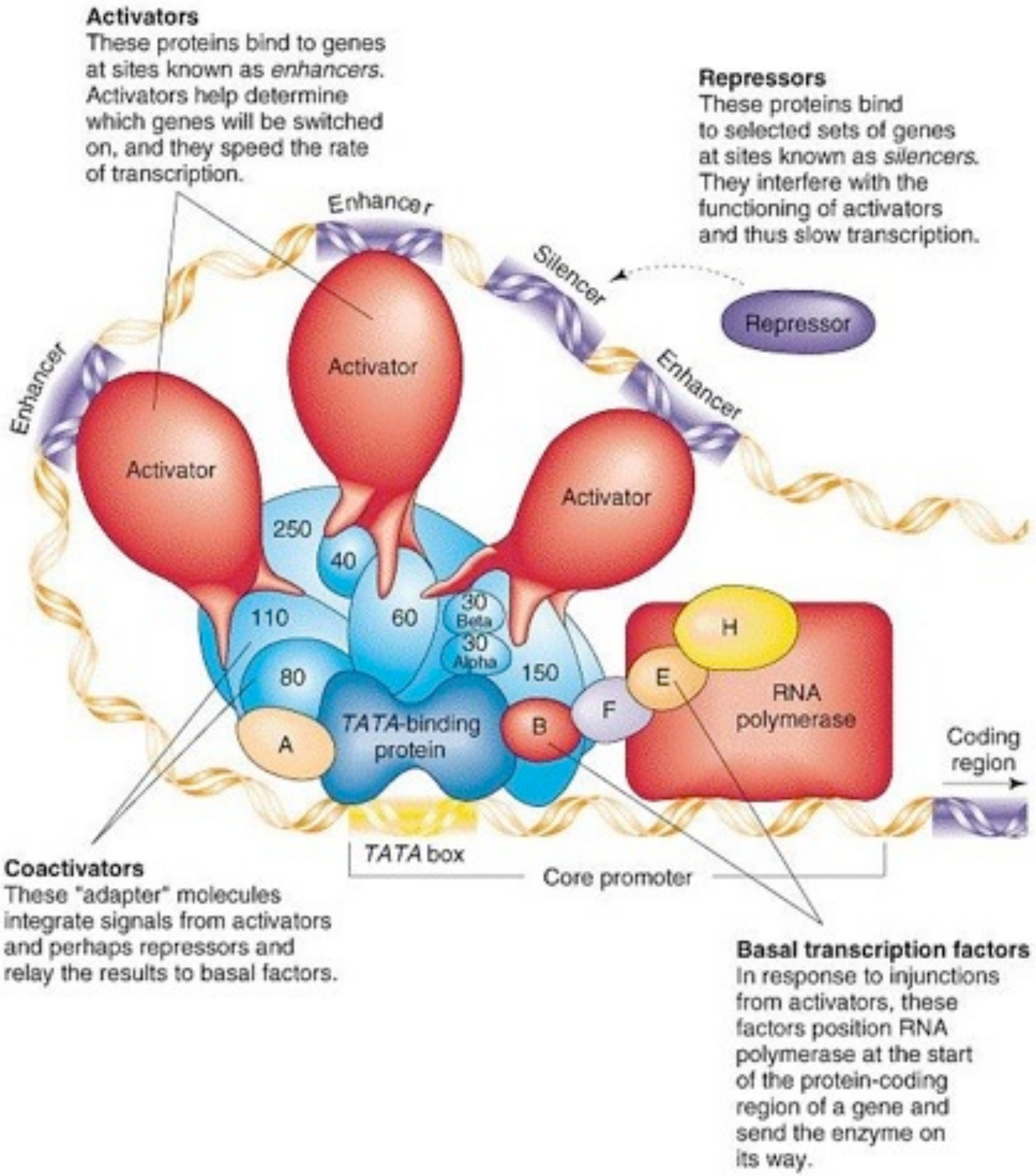
Ian Simpson

ian.simpson@.ed.ac.uk
bit.ly/bio2_2012

The Central Dogma of Molecular Biology



Transcription



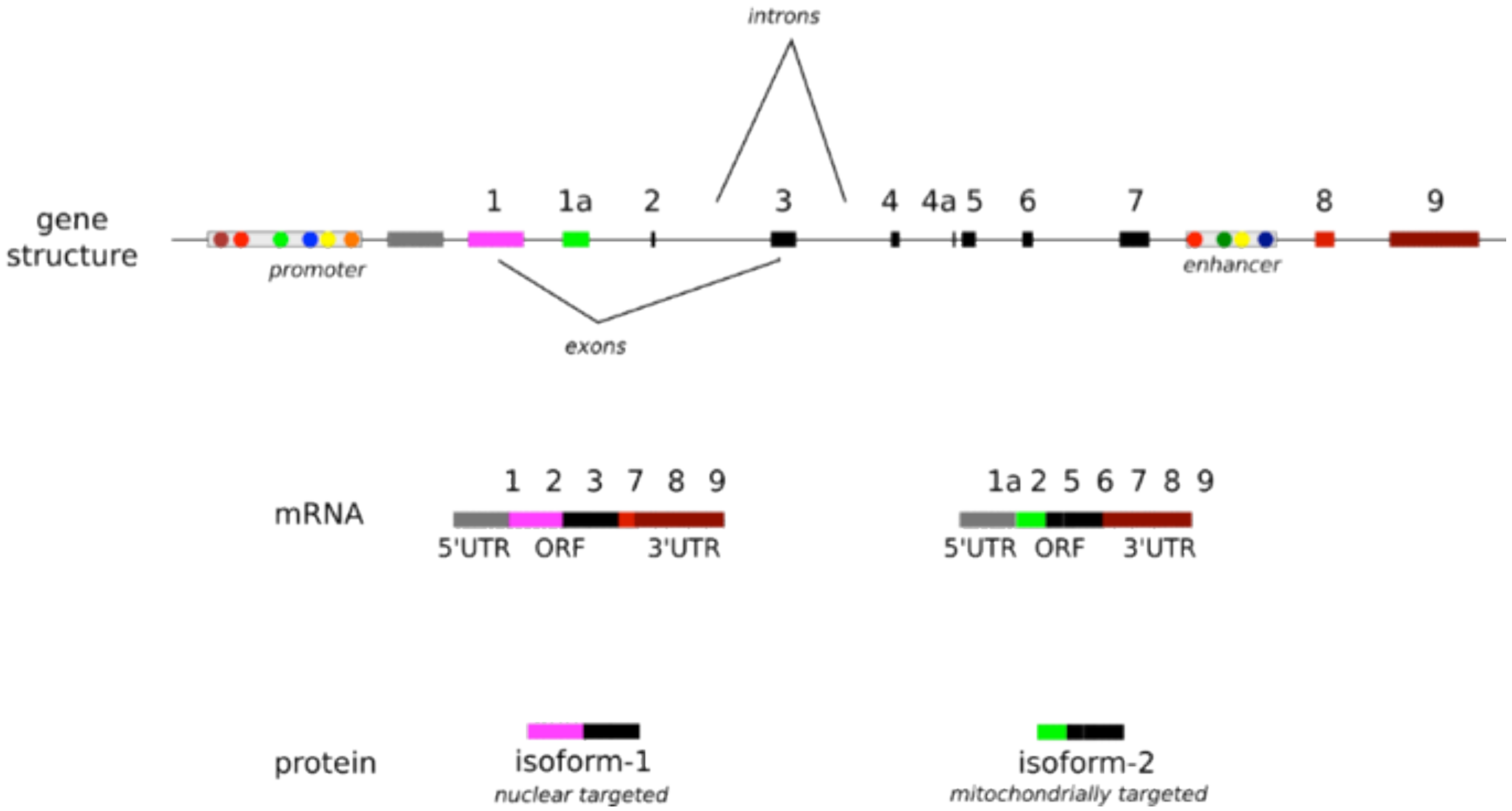
Activators
These proteins bind to genes at sites known as *enhancers*. Activators help determine which genes will be switched on, and they speed the rate of transcription.

Repressors
These proteins bind to selected sets of genes at sites known as *silencers*. They interfere with the functioning of activators and thus slow transcription.

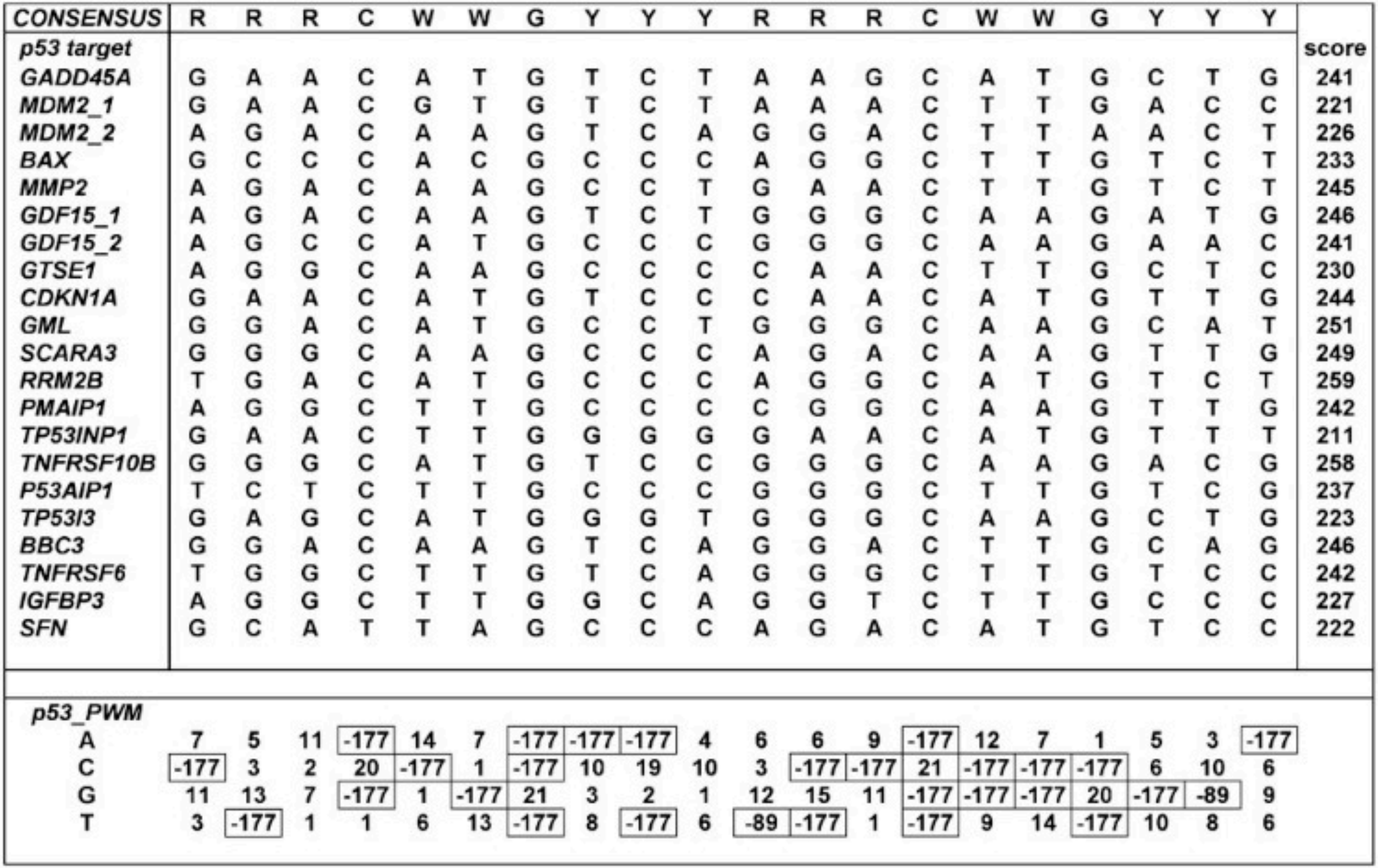
Coactivators
These "adapter" molecules integrate signals from activators and perhaps repressors and relay the results to basal factors.

Basal transcription factors
In response to injunctions from activators, these factors position RNA polymerase at the start of the protein-coding region of a gene and send the enzyme on its way.

Gene structure



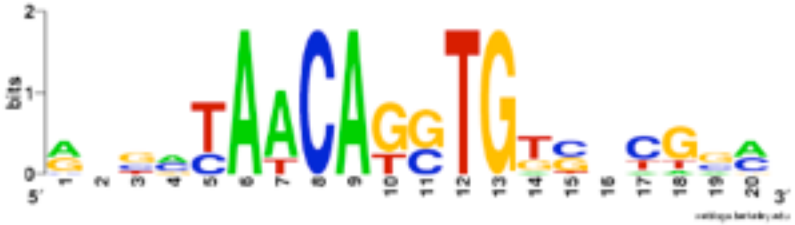
Example PWM for the human P53 protein



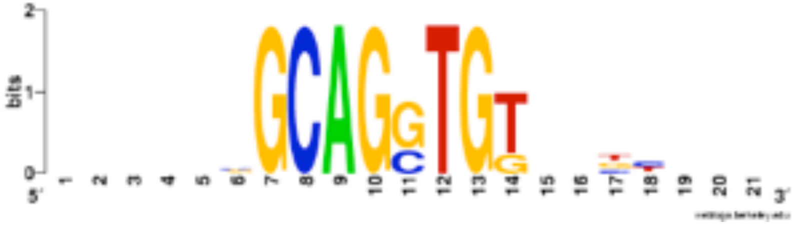
CONSENSUS	R	R	R	C	W	W	G	Y	Y	Y	R	R	R	C	W	W	G	Y	Y	Y	
p53 target																					score
GADD45A	G	A	A	C	A	T	G	T	C	T	A	A	G	C	A	T	G	C	T	G	241
MDM2_1	G	A	A	C	G	T	G	T	C	T	A	A	A	C	T	T	G	A	C	C	221
MDM2_2	A	G	A	C	A	A	G	T	C	A	G	G	A	C	T	T	A	A	C	T	226
BAX	G	C	C	C	A	C	G	C	C	C	A	G	G	C	T	T	G	T	C	T	233
MMP2	A	G	A	C	A	A	G	C	C	T	G	A	A	C	T	T	G	T	C	T	245
GDF15_1	A	G	A	C	A	A	G	T	C	T	G	G	G	C	A	A	G	A	T	G	246
GDF15_2	A	G	C	C	A	T	G	C	C	C	G	G	G	C	A	A	G	A	A	C	241
GTSE1	A	G	G	C	A	A	G	C	C	C	C	A	A	C	T	T	G	C	T	C	230
CDKN1A	G	A	A	C	A	T	G	T	C	C	C	A	A	C	A	T	G	T	T	G	244
GML	G	G	A	C	A	T	G	C	C	T	G	G	G	C	A	A	G	C	A	T	251
SCARA3	G	G	G	C	A	A	G	C	C	C	A	G	A	C	A	A	G	T	T	G	249
RRM2B	T	G	A	C	A	T	G	C	C	C	A	G	G	C	A	T	G	T	C	T	259
PMAIP1	A	G	G	C	T	T	G	C	C	C	C	G	G	C	A	A	G	T	T	G	242
TP53INP1	G	A	A	C	T	T	G	G	G	G	G	A	A	C	A	T	G	T	T	T	211
TNFRSF10B	G	G	G	C	A	T	G	T	C	C	G	G	G	C	A	A	G	A	C	G	258
P53AIP1	T	C	T	C	T	T	G	C	C	C	G	G	G	C	T	T	G	T	C	G	237
TP53I3	G	A	G	C	A	T	G	G	G	T	G	G	G	C	A	A	G	C	T	G	223
BBC3	G	G	A	C	A	A	G	T	C	A	G	G	A	C	T	T	G	C	A	G	246
TNFRSF6	T	G	G	C	T	T	G	T	C	A	G	G	G	C	T	T	G	T	C	C	242
IGFBP3	A	G	G	C	T	T	G	G	C	A	G	G	T	C	T	T	G	C	C	C	227
SFN	G	C	A	T	T	A	G	C	C	C	A	G	A	C	A	T	G	T	C	C	222
p53_PWM																					
A	7	5	11	-177	14	7	-177	-177	-177	4	6	6	9	-177	12	7	1	5	3	-177	
C	-177	3	2	20	-177	1	-177	10	19	10	3	-177	-177	21	-177	-177	-177	6	10	6	
G	11	13	7	-177	1	-177	21	3	2	1	12	15	11	-177	-177	-177	20	-177	-89	9	
T	3	-177	1	1	6	13	-177	8	-177	6	-89	-177	1	-177	9	14	-177	10	8	6	

The classic footprinting method

Atonal E-box

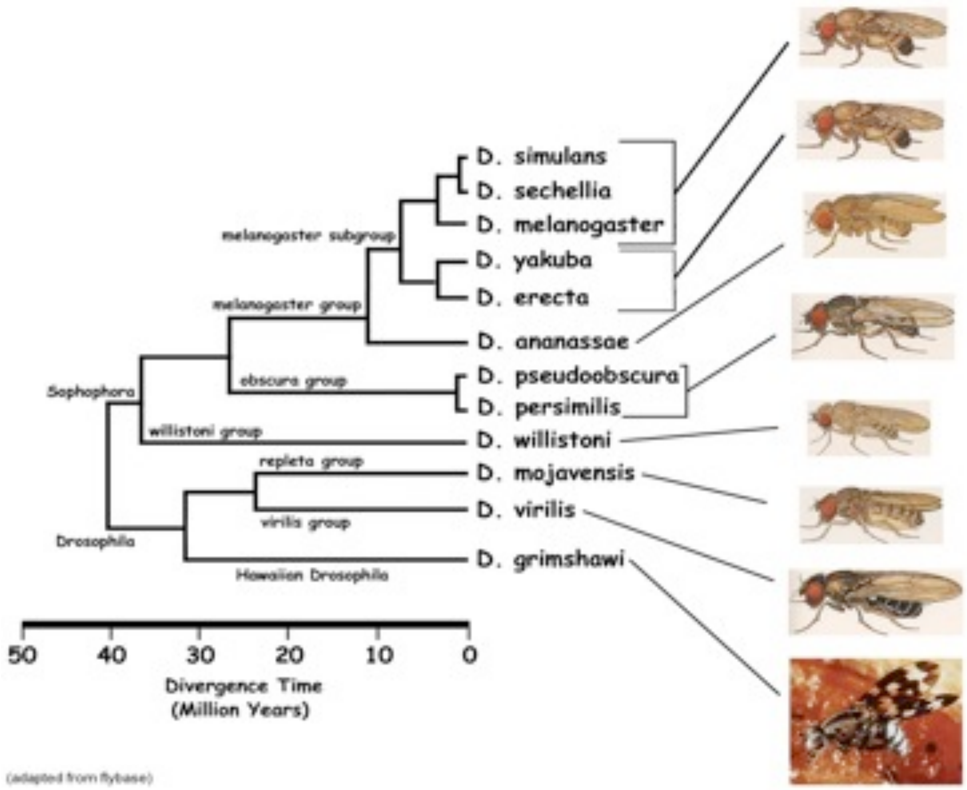


Scute E-box



TF binding site screening

- PWM GibbsSampler
- MOODS fast forward



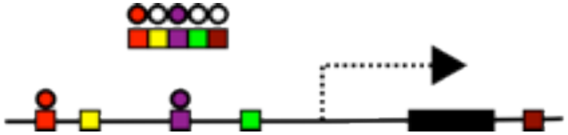
(adapted from flybase)

phylogenetic conservation

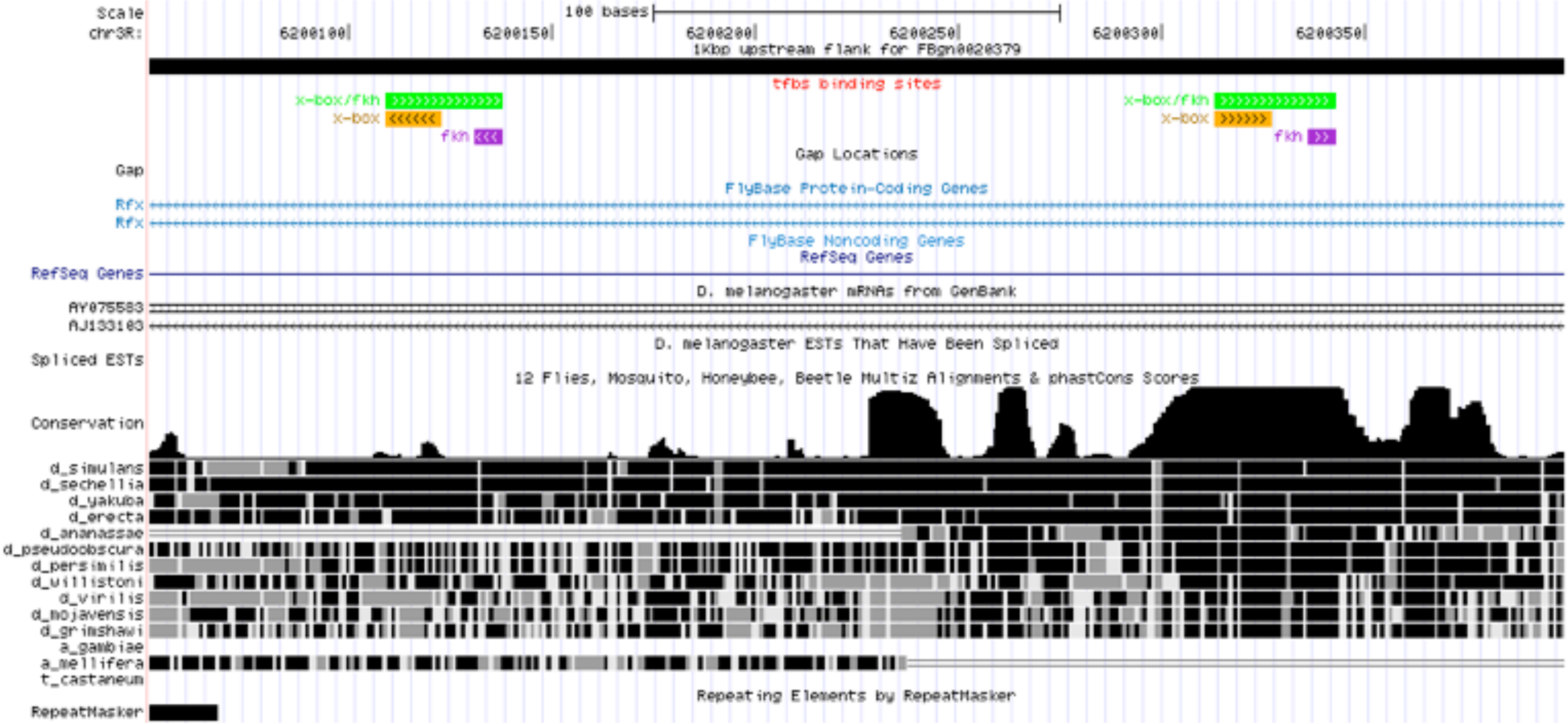
- PhastCons, UCSCMultiz
- BioProspector

promoter/enhancer detection

- HMM/cis-module
- cluster-buster
- BioTIFFIN
- RedFly
- ModENCODE



Classic phylogenetic footprinting approach



Limitations of the classical approach to finding TFBSs

•The number and quality of binding site sequences is low

•There is no explicit relation between conservation and function
i.e. sites are often conserved, but conserved sites do not necessarily function

•Assumptions have to be made about where to look and how to score

•Extremely biased information, low number of experiments to determine sites

•Non-physiological conditions used during assessment

•Measurements made only in specific tissue or cells at specific times
local solutions to the PWM problem, may be wrong for other conditions

Problems with the available data sources

- Main source of site specific data remains pattern or PWM (or HMM)

Common name	Binary nomenclature	Number of PWMs
human	Homo sapiens	476
mouse	Mus musculus	423
rat	Rattus norvegicus	253
chick	Gallus gallus	133
clawed frog	Xenopus laevis	84
fruit fly	Drosophila melanogaster	68
thale cress	Arabidopsis thaliana	45
yeast	Saccharomyces cerevisiae	39
monkey	Cercopithecus aethiops	29
gibbon ape	Hylobates lar	24
cattle	Bos taurus	23
domestic pig	Sus scrofa	20
zebra fish	Brachydanio rerio	19

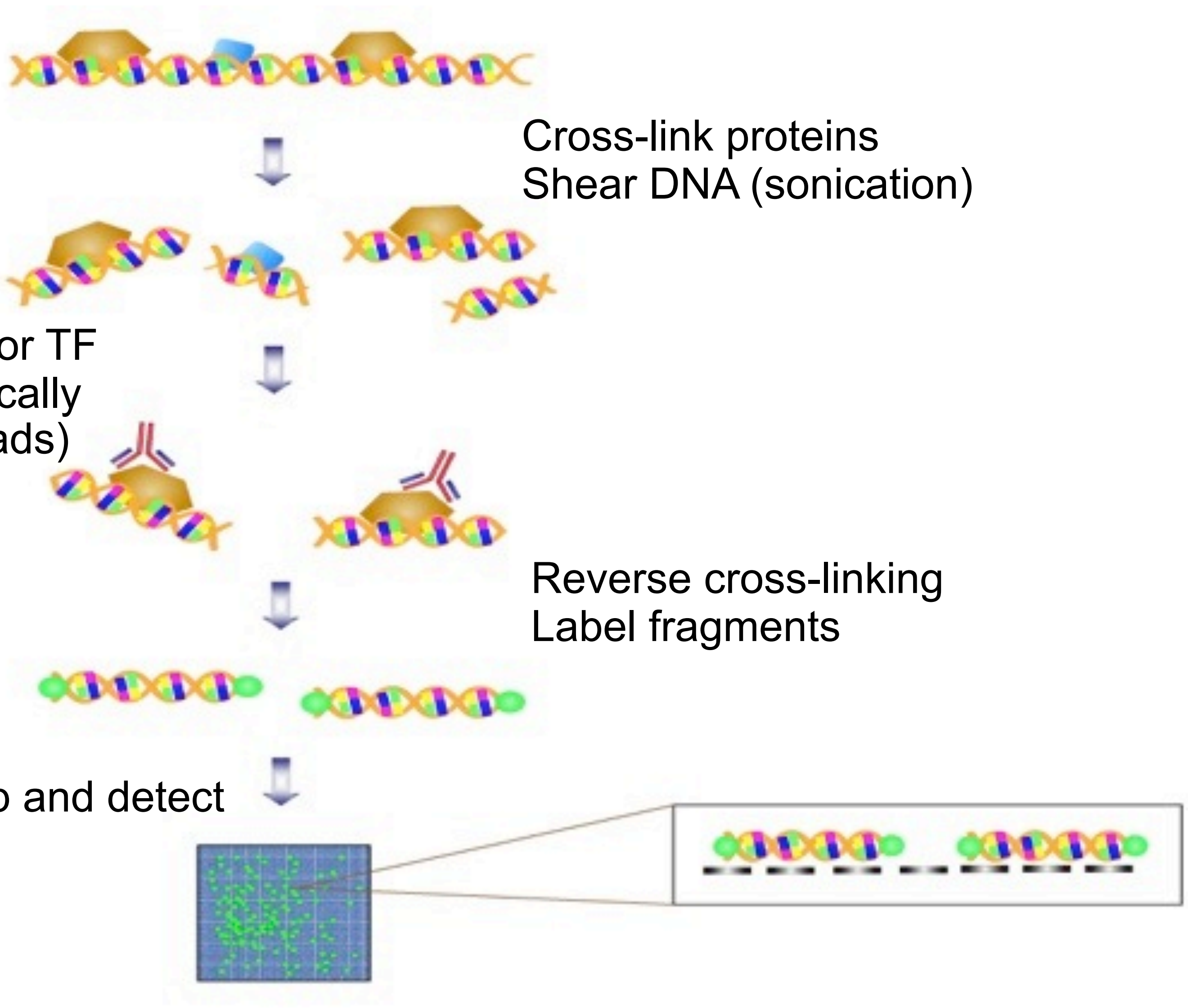
TransfacPro2009.1

Replacing classical prediction with direct localisation

What do we need

- Assays that cover the whole genome (aren't biased)
- Applicable to all transcription factors (good coverage)
- Can be measured in lots of different conditions (condition specific, biologically relevant)
- Can be mapped onto precise (and small) genome locations (high resolution)
- Cost effective, accurate and reliable

Chromatin immuno-precipitation (ChIP)



Cross-link proteins
Shear DNA (sonication)

Recover fragments for TF
using antibody (typically
pulled down on beads)

Reverse cross-linking
Label fragments

Hybridise to chip and detect

How do we get from populations of DNA fragments to positions on chromosomes ?

Currently there are two main choices

ChIP-chip	Hybridisation onto a genomic tiling array
Chip-seq	Direct sequencing of the bound (now released) fragments

ChIP-chip

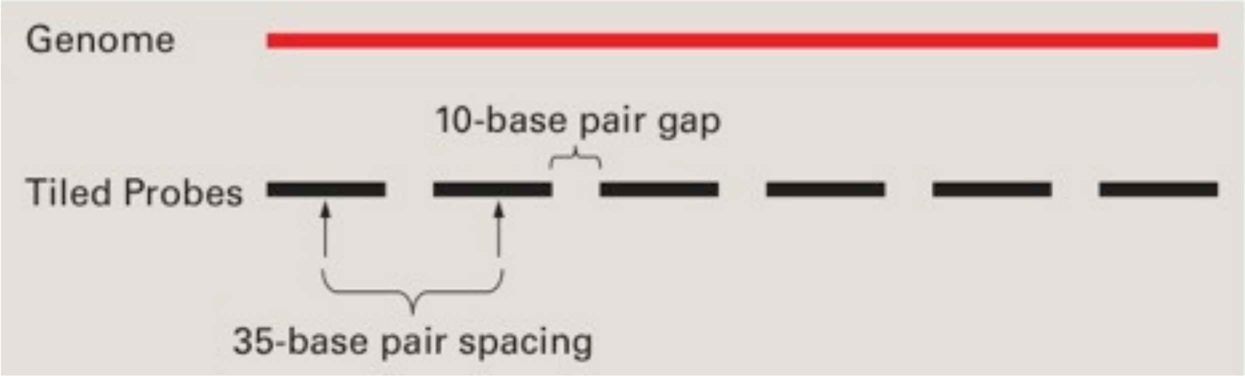
Here a manufactured slide is used in which fragments spanning the genome have been synthesised and attached to the slide surface in a geometric Arrangement. We label our TF retrieved fragments, hybridise them to the slide and then read fluorescence from the features.

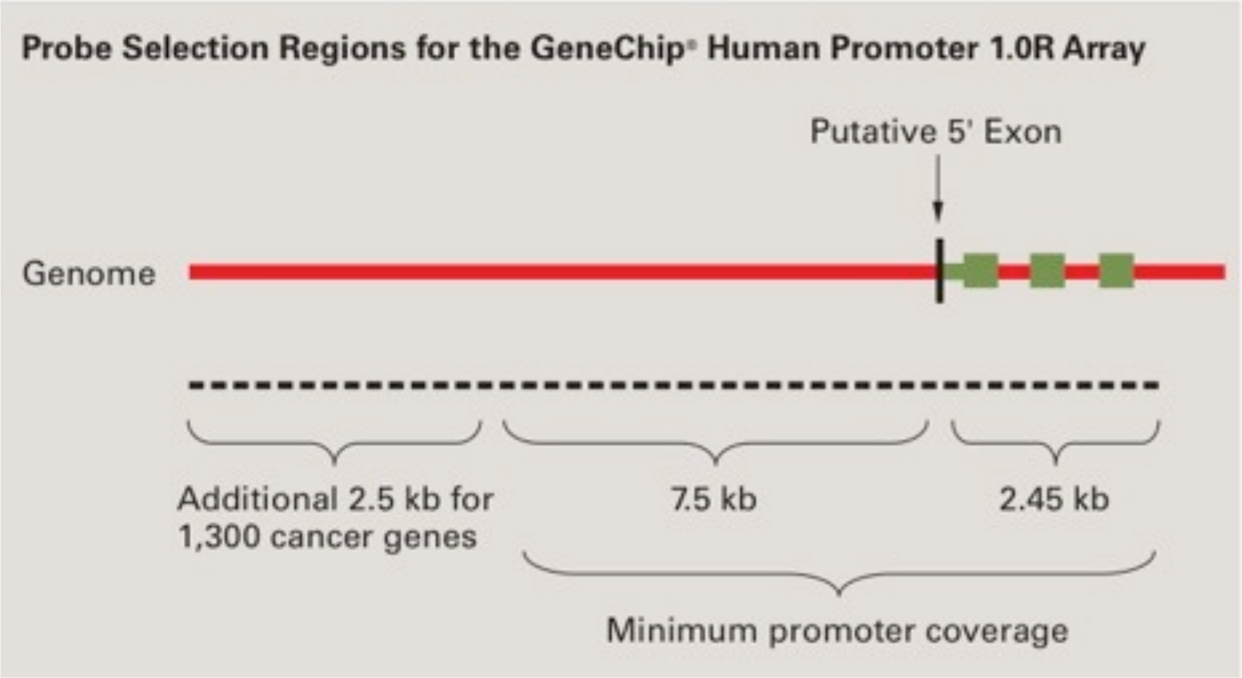
ChIP-seq

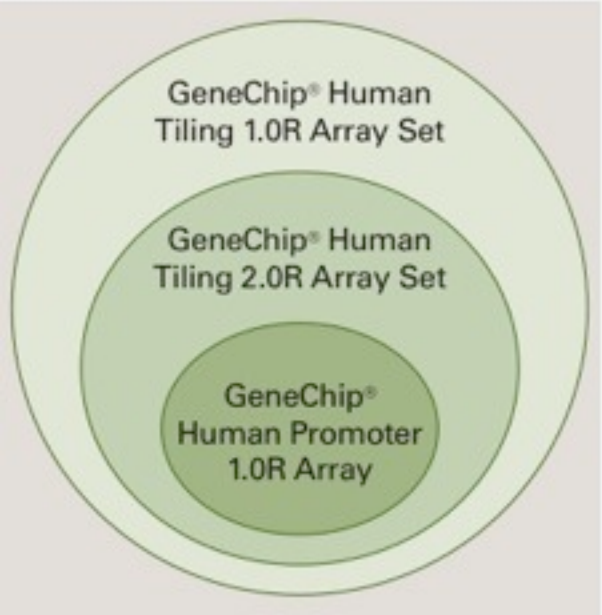
Taking advantage of high throughput sequencing technology (so called next-gen) we attempt to sequence all the fragments. This is quantitative.

In both cases we have issues with mapping, signal processing (noise) and significance testing

Detection method 1 - Genome tiling arrays (ChIP-chip)




Probe Selection Regions for the GeneChip® Human Promoter 1.0R Array
Putative 5' Exon
Genome
Additional 2.5 kb for 1,300 cancer genes
7.5 kb
2.45 kb
Minimum promoter coverage


GeneChip® Human Tiling 1.0R Array Set
GeneChip® Human Tiling 2.0R Array Set
GeneChip® Human Promoter 1.0R Array

Features of genome tiling arrays

•Generally resolution can be as low as ~3kb, Tfs bind to on average 6-8bp

•How do we know which gene to map to ? (meta-data)
microarray, gene proximity, functional annotation, in-vivo expression
comparison to true positive

•Redundancy probes map to more than one location

•Coverage, cannot cover the genome. This introduces bias.
even in Drosophila commonly only 50% of genome possible
2 human chromosomes at 35bp resolution → 1 million features

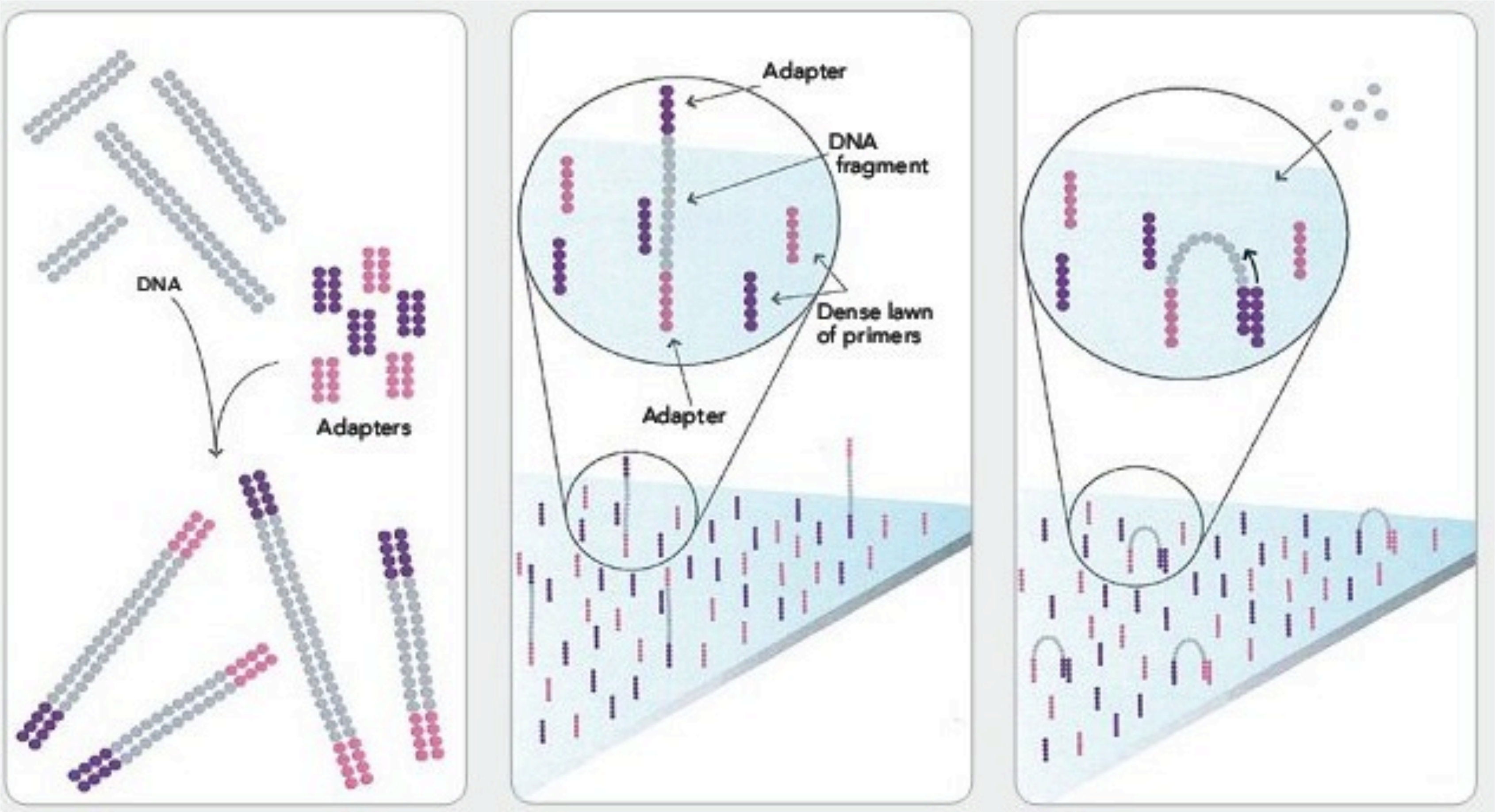
•Can estimate site occupancy frequency

•Cross-hybridisation can be big problem with repetitive DNA (~5% human genome)

•Processed just like a gene expresison microarray
SAM, limma (modelled error, tight control of FDR)

Detection method 2 – direct sequencing (ChIP-seq)

Illumina/Solexa SBS sequencing system

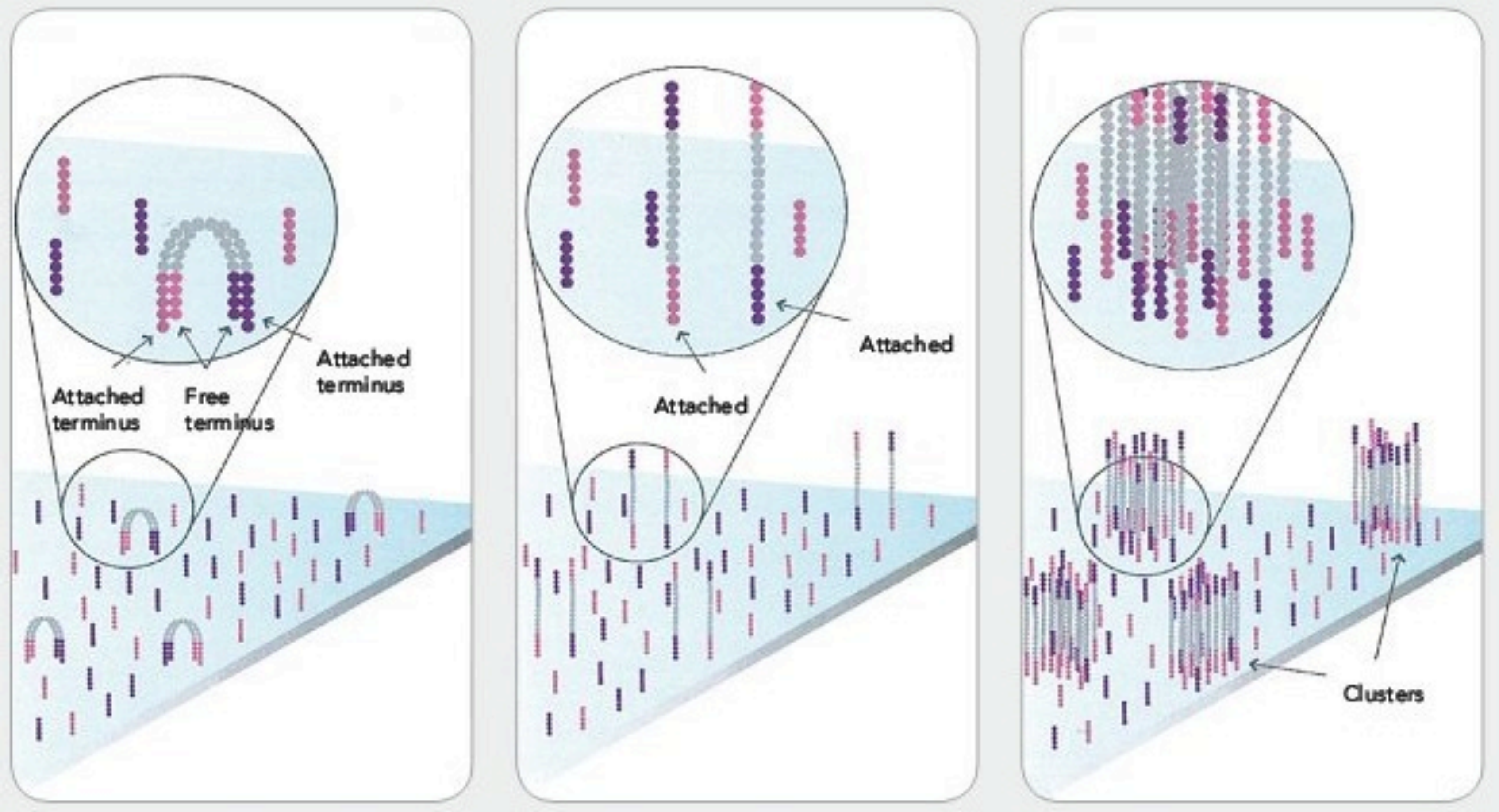


Ligate adaptors onto DNA fragments

Denature and attach to substrate

Anneal and extend bridge

Detection method 2 – direct sequencing (ChIP-seq)

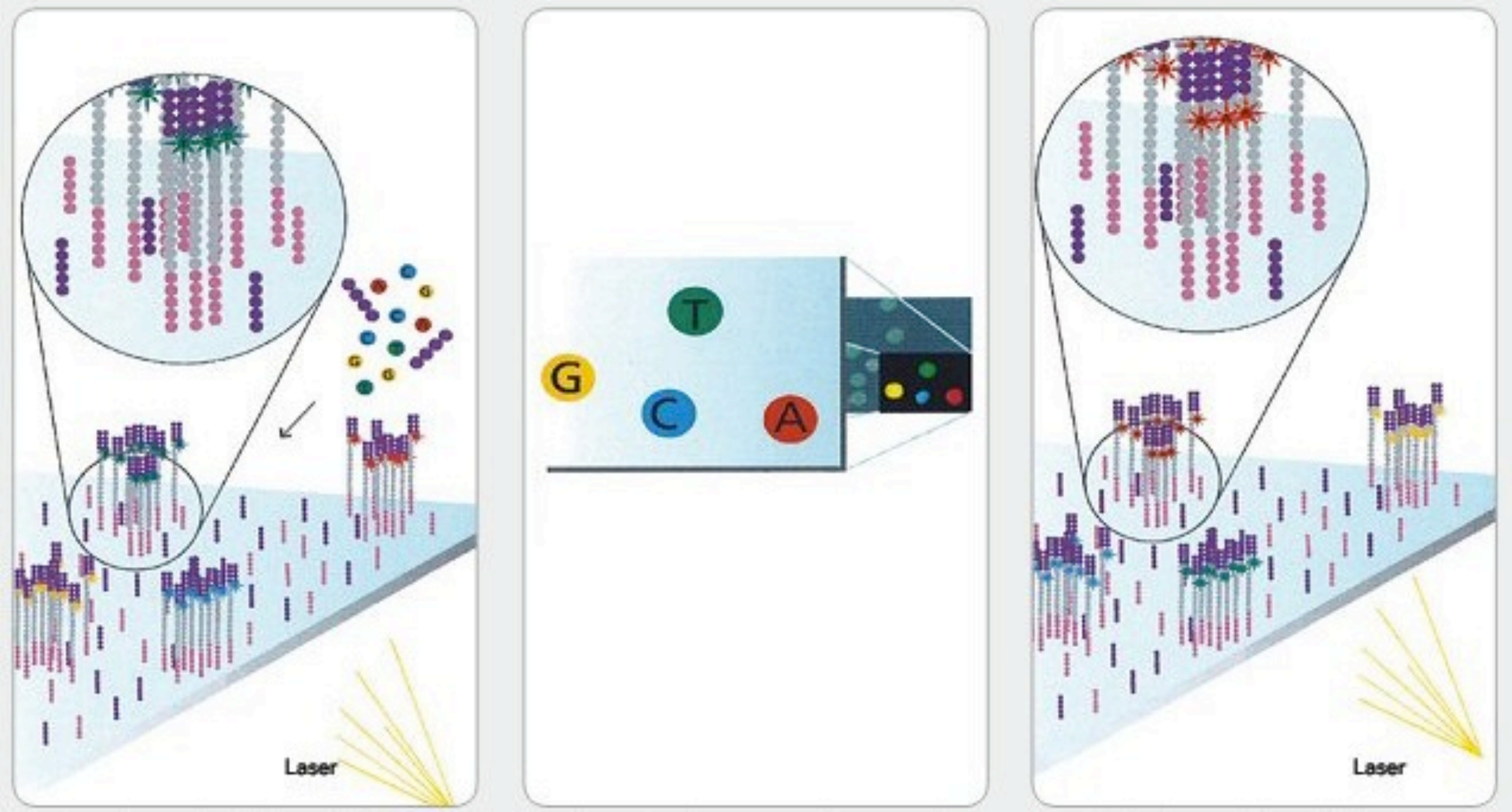


Complete extension

Denature ready for next round

Repeat to build cluster

Detection method 2 – direct sequencing (ChIP-seq)

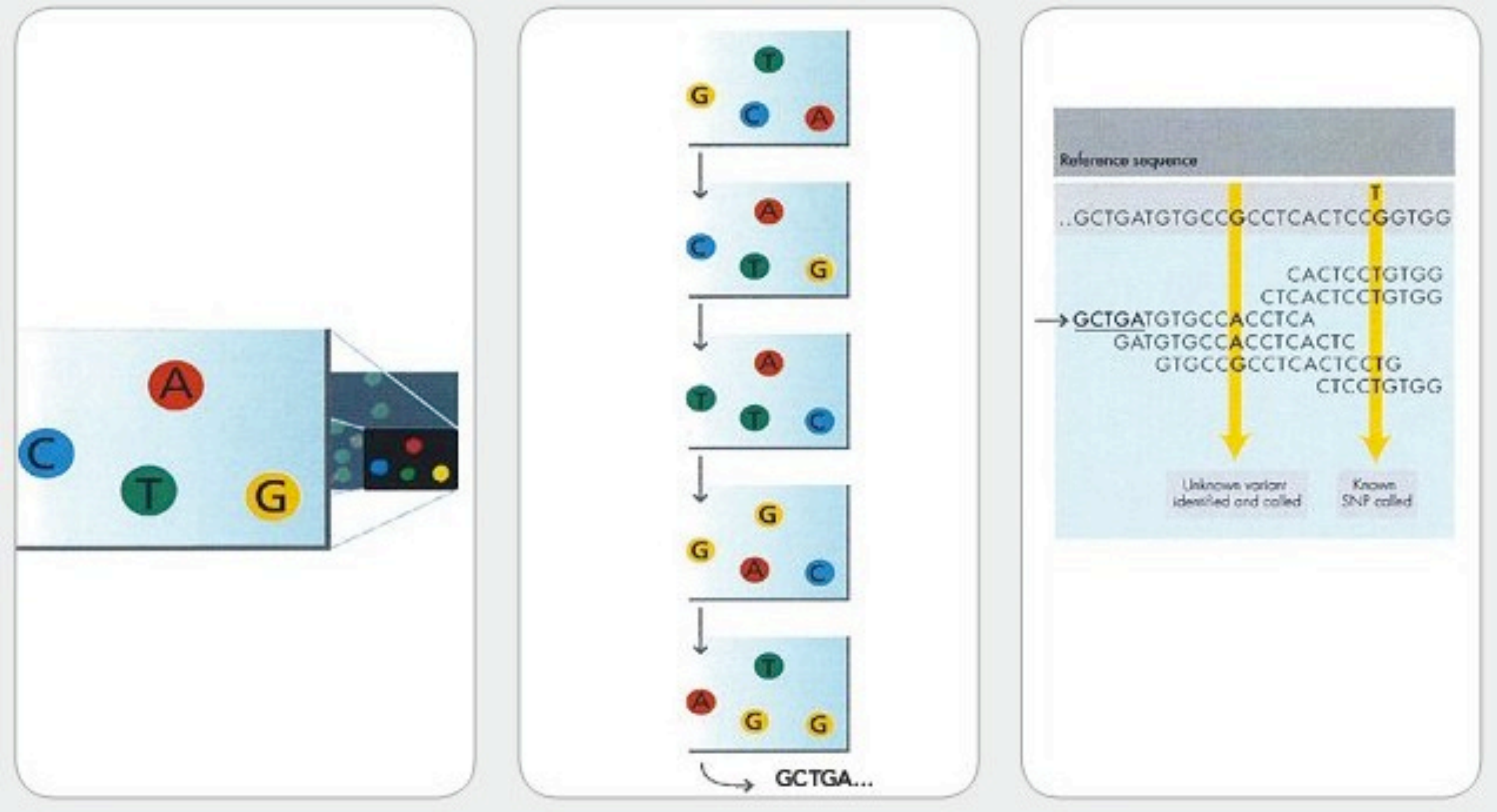


Add fluorescent nucleotides and primer

Scan chip for first base

Enzymatically release block and repeat addition of fluorescent base

Detection method 2 – direct sequencing (ChIP-seq)



Read next base

Repeat until complete

Assemble, align and map sequences

Features of high-throughput sequence data

•Very high resolution, typically 25-mers with mid-spacing ~35bp

•Huge datasets, many Gb of sequence, assembly non-trivial

•Complete genome coverage, no assumption, no bias

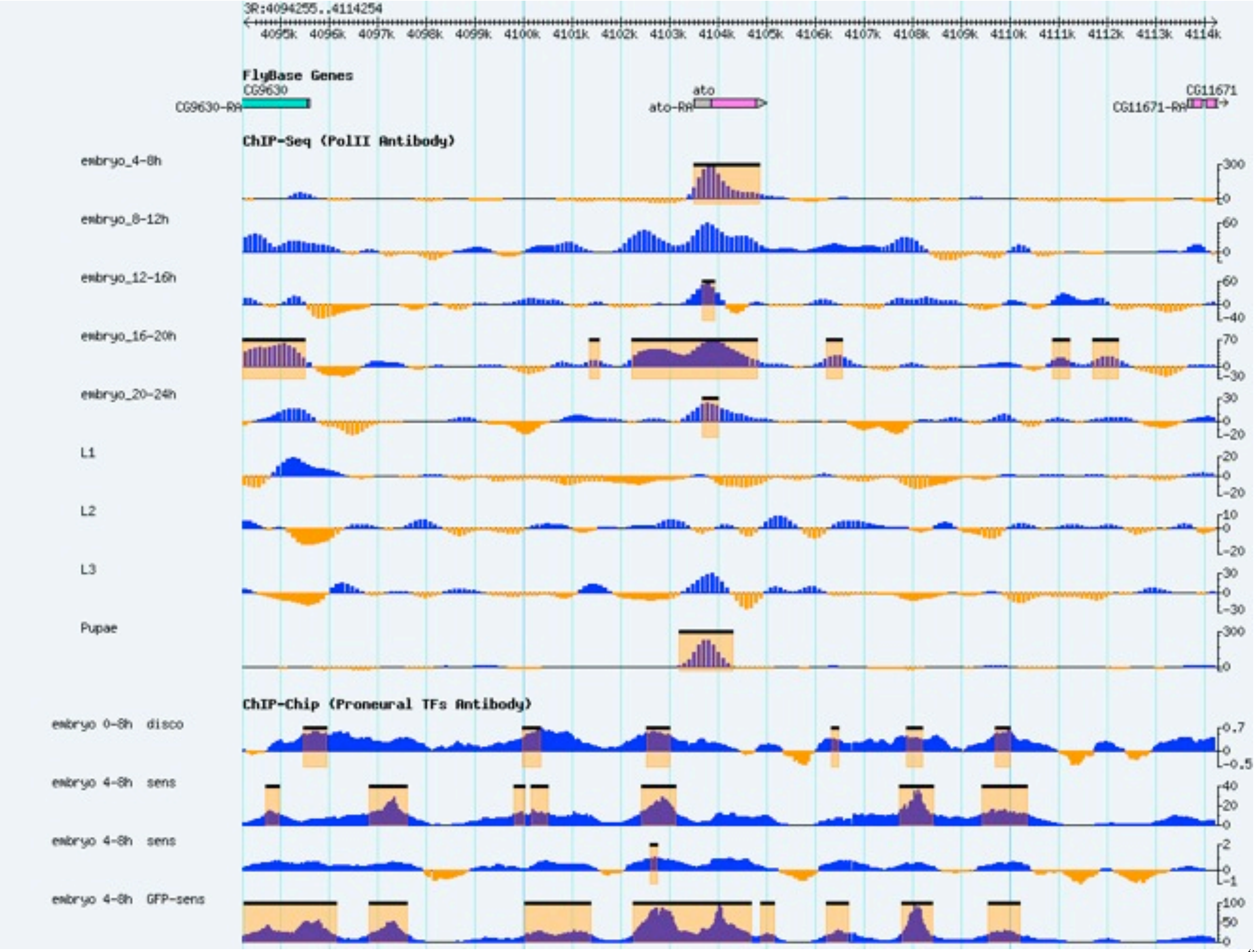
•Generally superior at identifying bound sites beyond expectation
(this is related to a more accurate ability to discriminate signal from noise)

•Sequences are counted to determine the frequency of site occupancy
(better than chip, here seq num is proportional to bound sites)

•Sequences are mapped and converted into signal peaks
(typical sizes of bound peaks can range from 50bp-1kb)

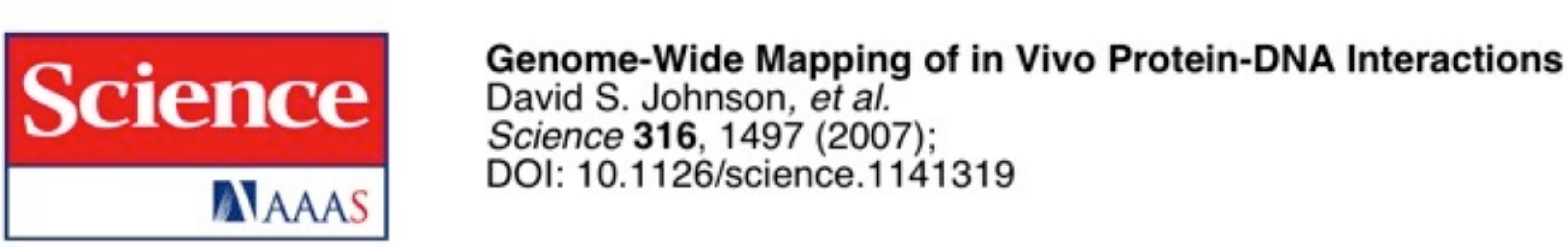
•Strong correlation between statistical significance of peak and presence of binding motif (might seem obvious!)

Example ChIP-Chip and ChIP-seq data spanning the atonal locus



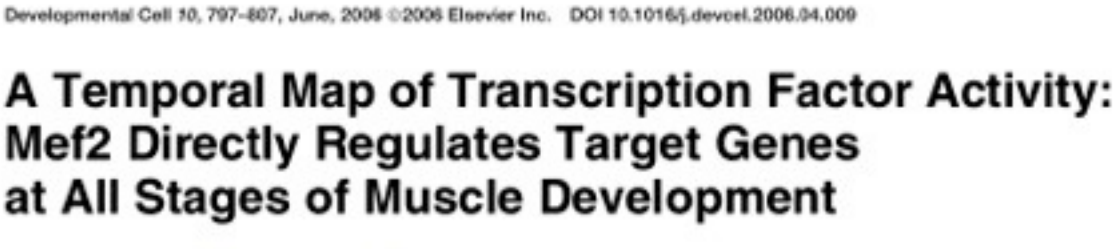
Ian Simpson, Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh

Real world examples of ChIP-chip and ChIP-seq in use



Science
AAAS

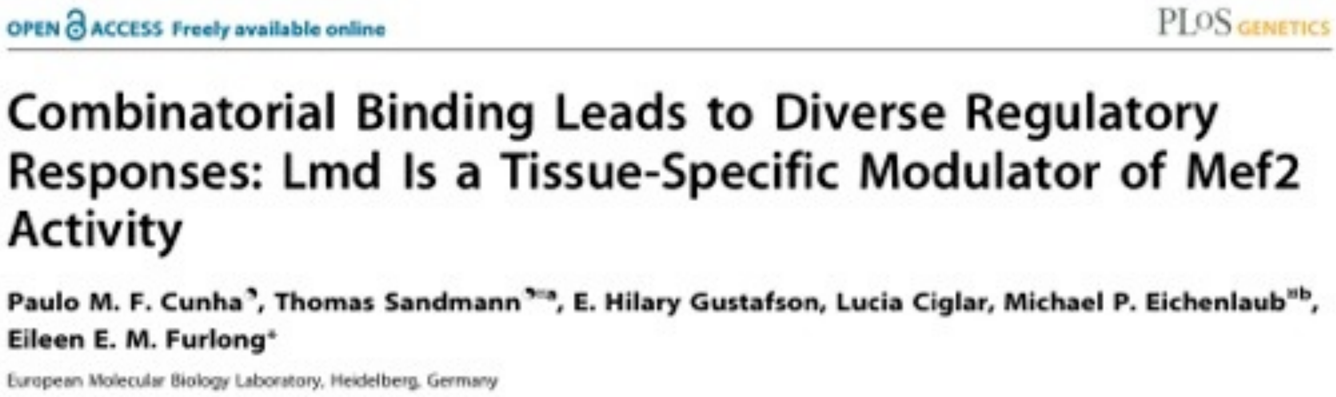
Genome-Wide Mapping of in Vivo Protein-DNA Interactions
David S. Johnson, *et al.*
Science **316**, 1497 (2007);
DOI: 10.1126/science.1141319



Developmental Cell *10*, 797–807, June, 2006 ©2006 Elsevier Inc. DOI 10.1016/j.devcel.2006.04.009

A Temporal Map of Transcription Factor Activity: Mef2 Directly Regulates Target Genes at All Stages of Muscle Development

Thomas Sandmann,[1] Lars J. Jensen,[1]
Janus S. Jakobsen,[1] Michal M. Karzynski,[1]
Michael P. Eichenlaub,[1] Peer Bork,[1]
and Eileen E.M. Furlong[1,*]
[1] European Molecular Biology Laboratory
D-69117 Heidelberg
Germany



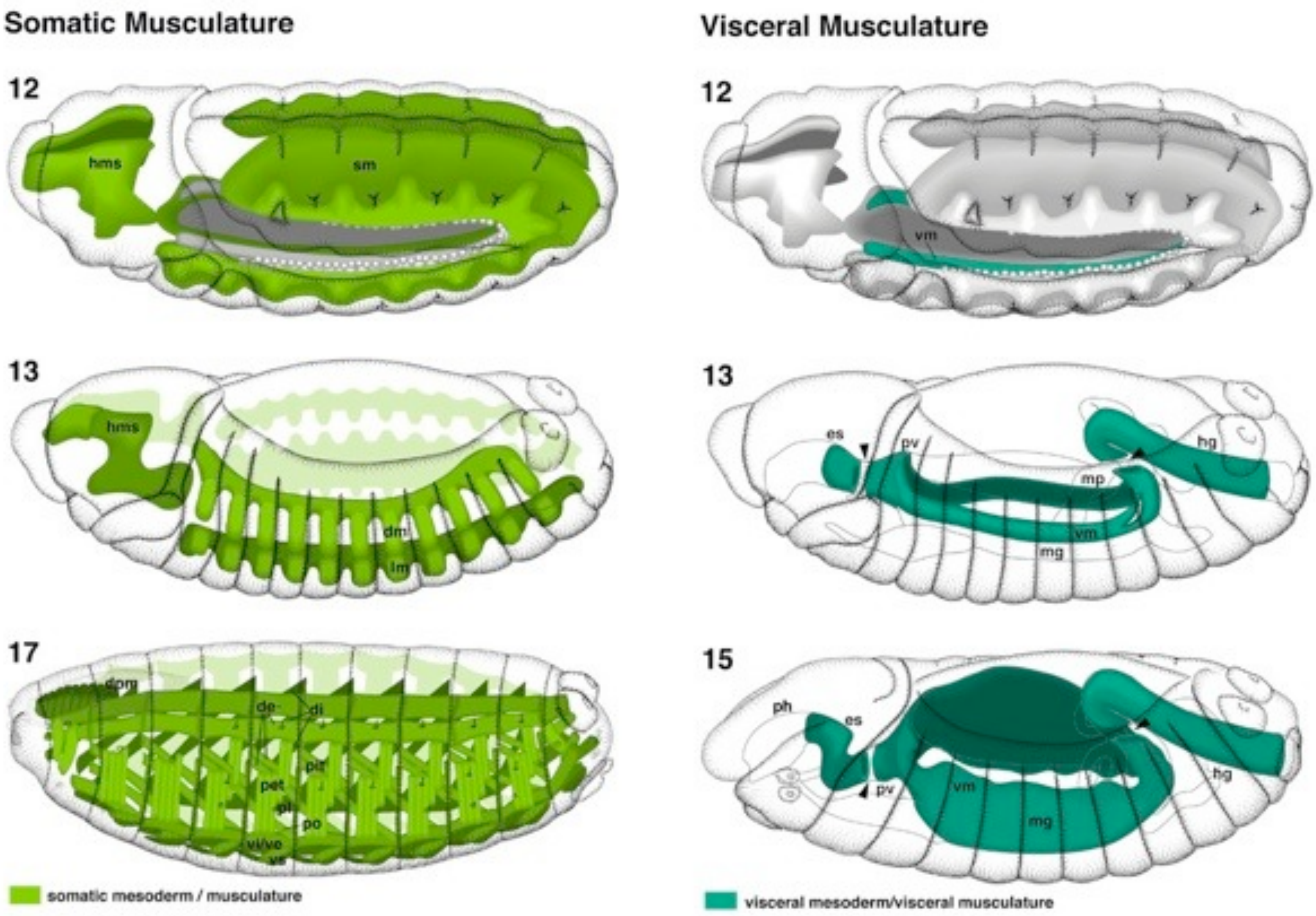
OPEN ACCESS Freely available online PLoS GENETICS

Combinatorial Binding Leads to Diverse Regulatory Responses: Lmd Is a Tissue-Specific Modulator of Mef2 Activity

Paulo M. F. Cunha[¶], Thomas Sandmann[¶¤a], E. Hilary Gustafson, Lucia Ciglar, Michael P. Eichenlaub[¤b], Eileen E. M. Furlong[*]

European Molecular Biology Laboratory, Heidelberg, Germany

Studying Drosophila musculature development using ChIP-chip

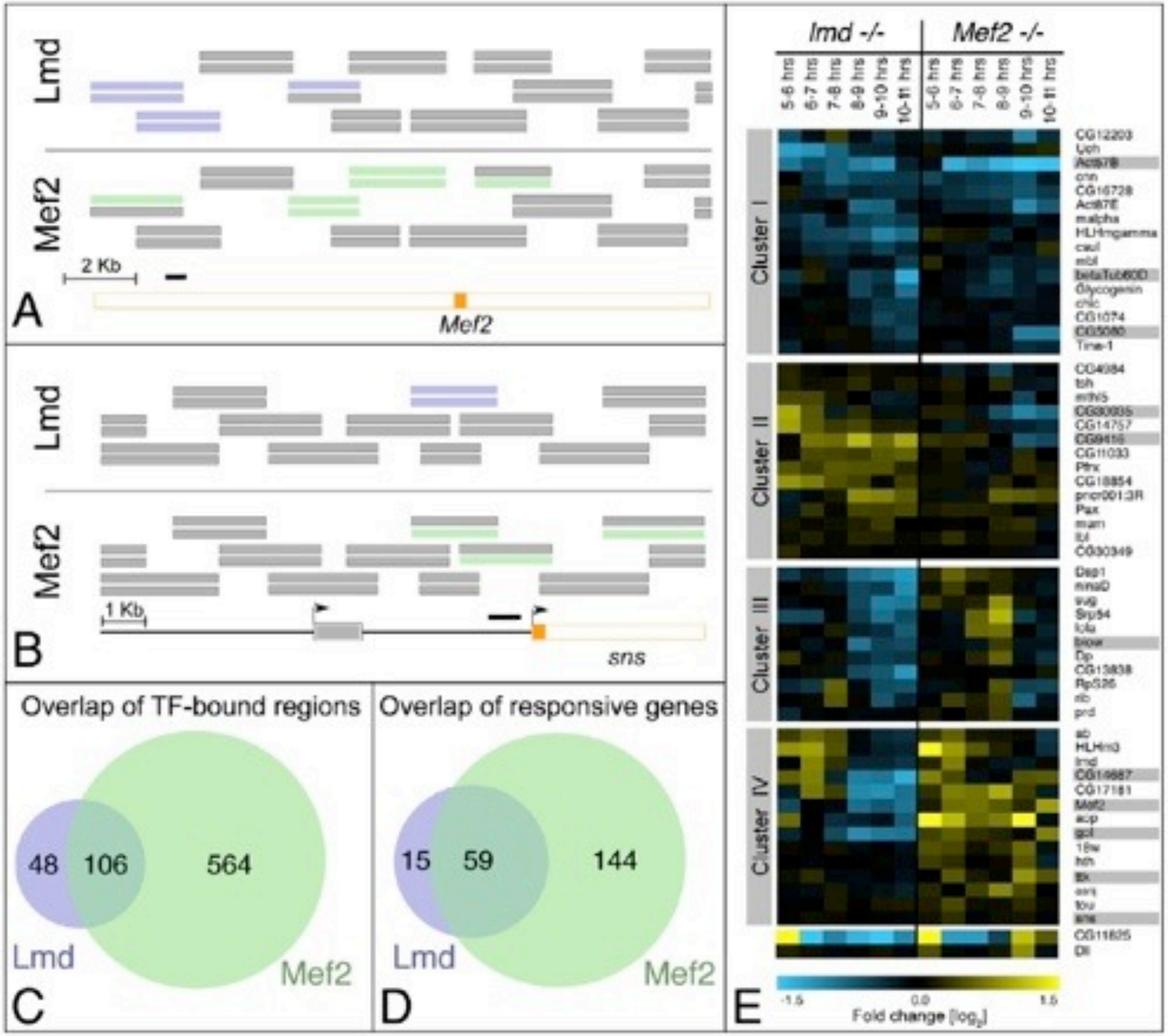

Somatic Musculature

Visceral Musculature

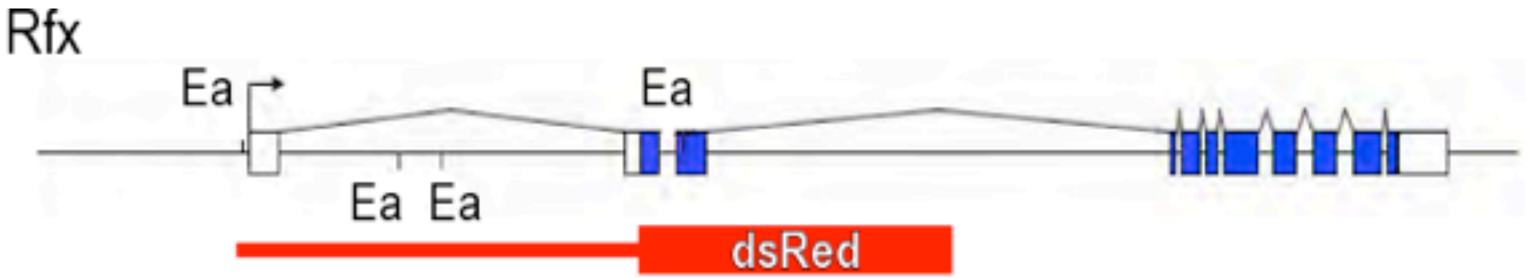
🟩 somatic mesoderm / musculature

🟩 visceral mesoderm/visceral musculature

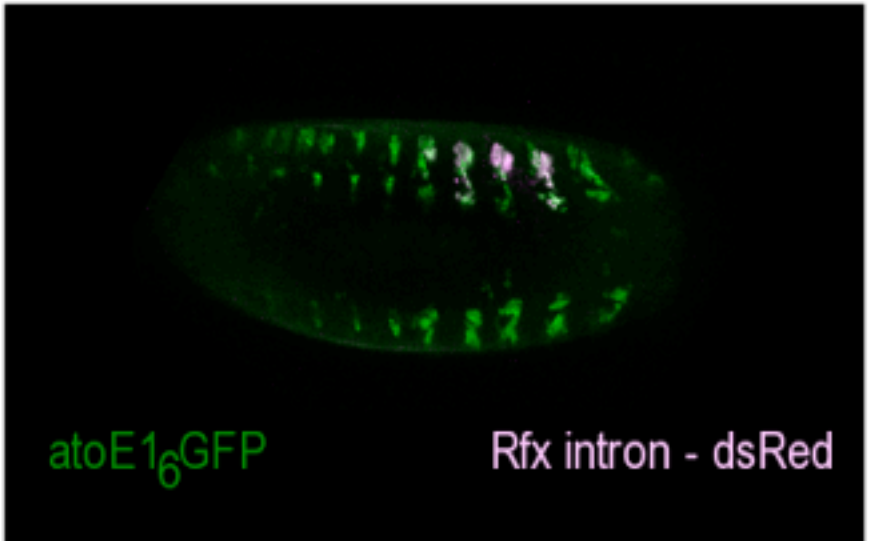
ChIP-chip blocks integrated with gene expression data for Mef2 and Lmd



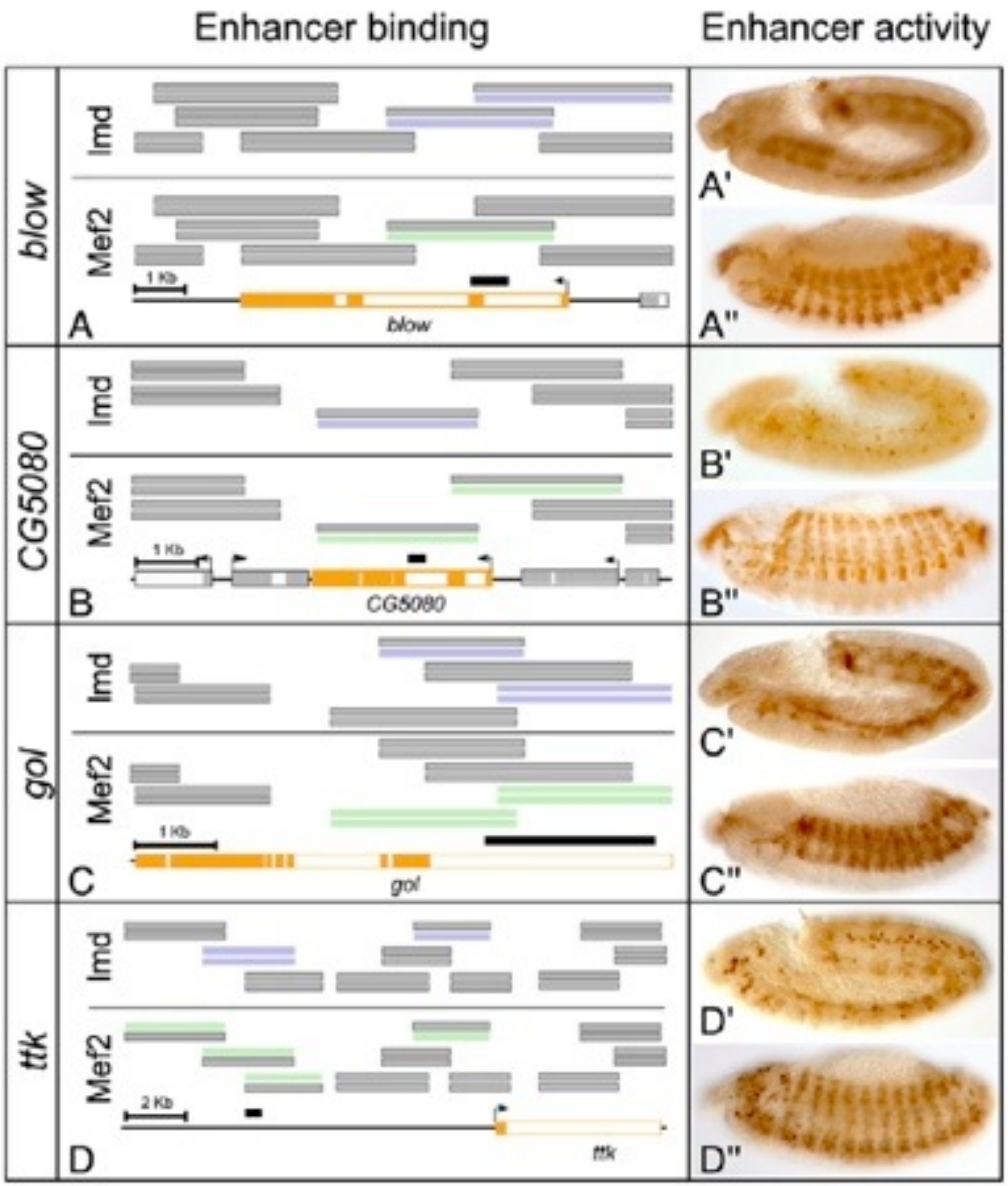
Validation of enhancers and TF binding sites



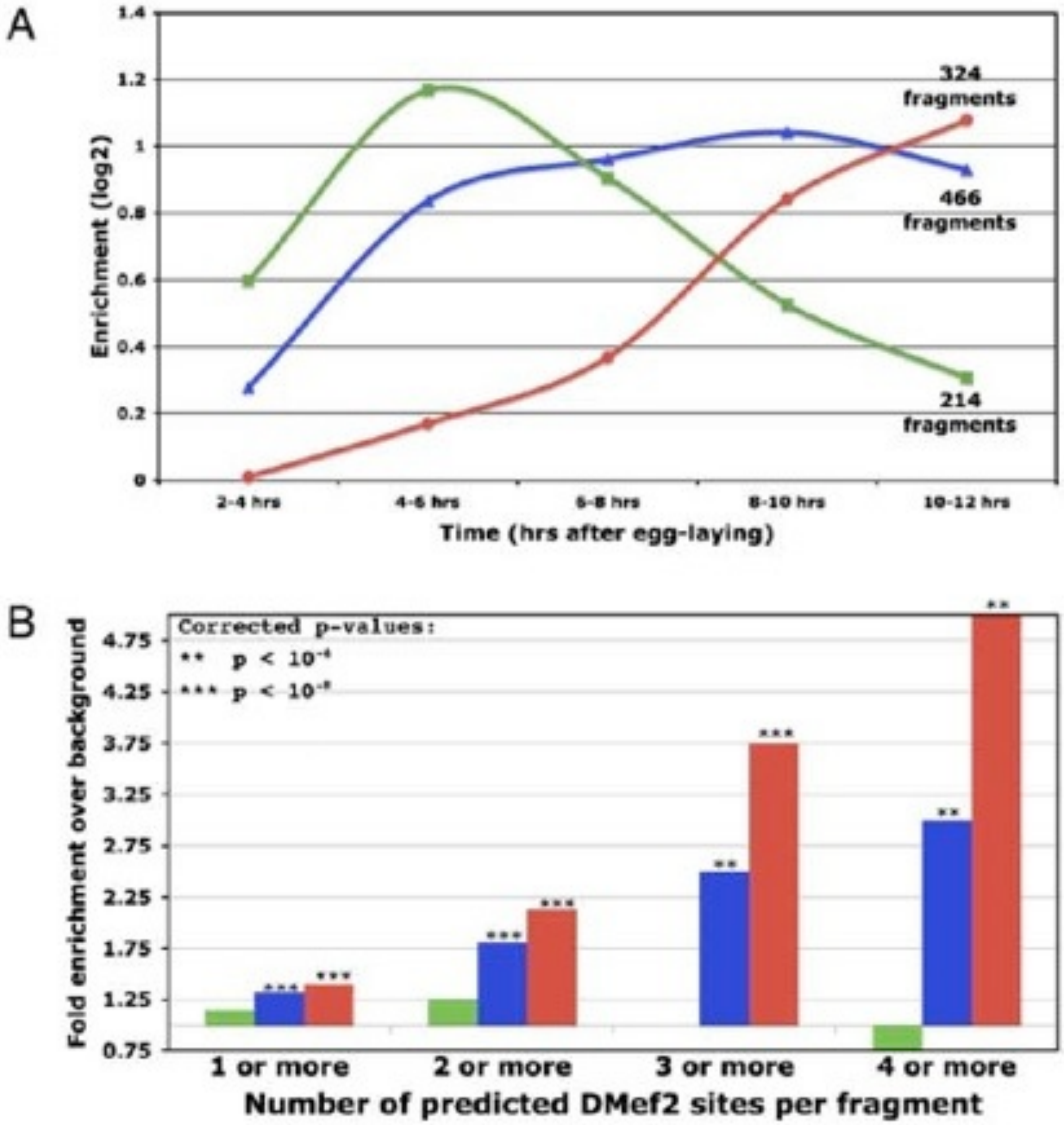
Dmel ATGCTTACCACCACTATTCGAA**CAGCTG**TGAGCGTTGCCACTTGTCTTGAGGATTAACCAA
Dsec ATGCTTACCAGCACTGTTCGAA**CAGCTG**TGAGCGTTGCCACTTGTCTTGGGGGTTAACCAG
Dere -TGCTAGCCAACACTGTTCGAA**CAGCTG**GGAGCGTTGCCACTTGTGCTCGCGGGCTAACCAA
Dyak ATGTTAACCAACACTGTTCGCA**CAGCTG**TAAGCGTTGCCACTTGTGCTTGCGAATTAGCCAG



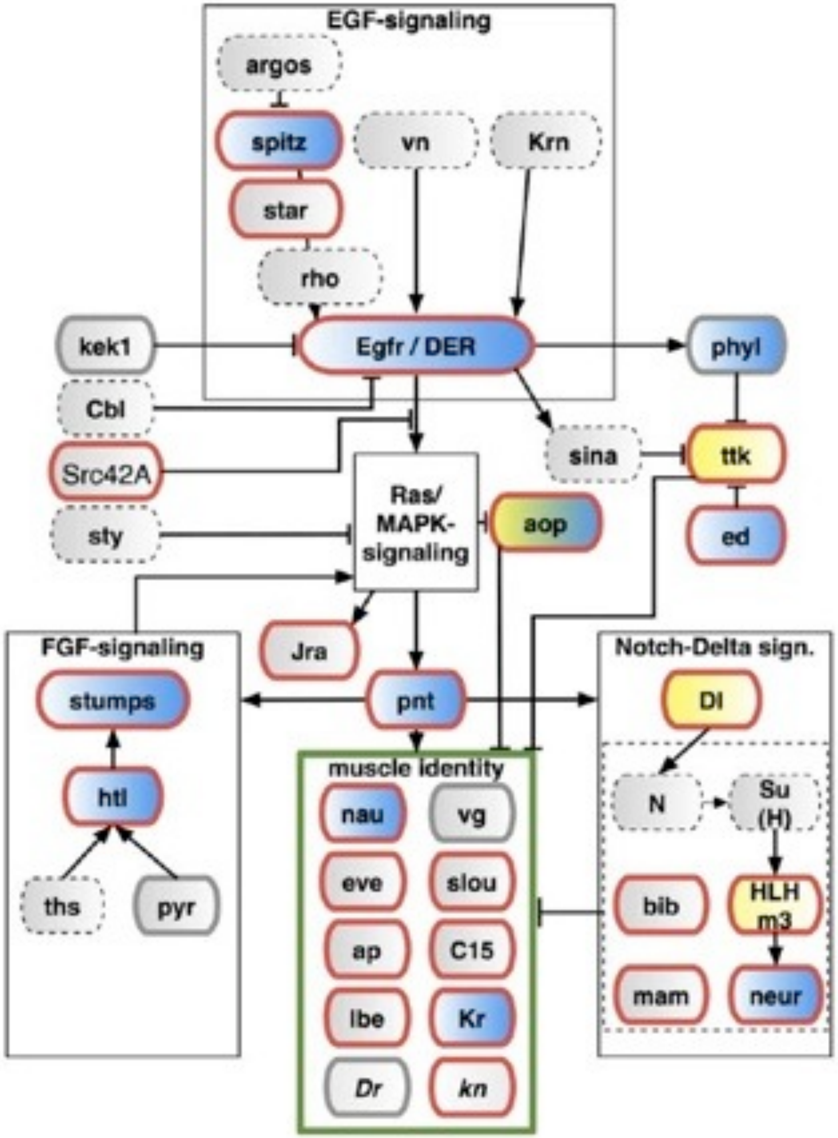
Validation of enhancer activity for Mef2/Lmd candidate target genes



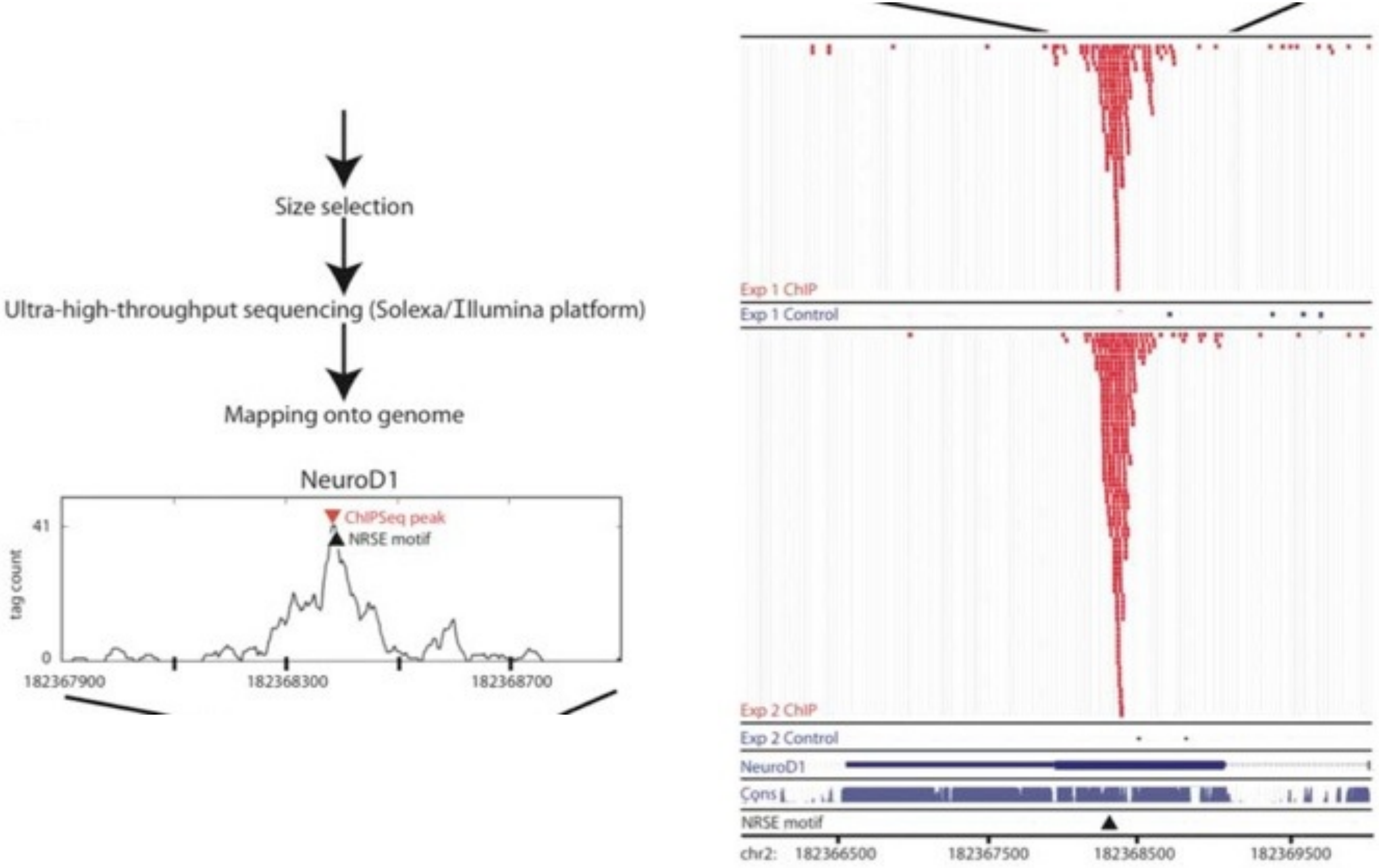
Temporal binding profiles of over-represented Mef2 bound blocks



Synthesis of the target gene network and known myogensis pathway



Chip-seq analysis of the neuron restrictive silencing factor (NRSF)



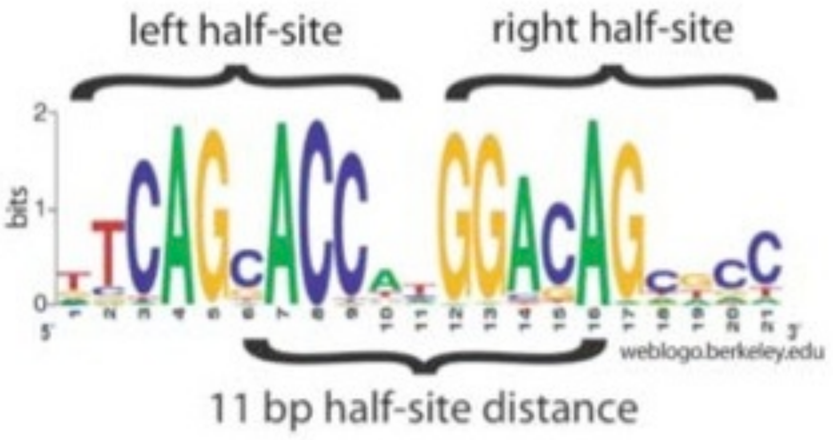
Size selection

Ultra-high-throughput sequencing (Solexa/Illumina platform)

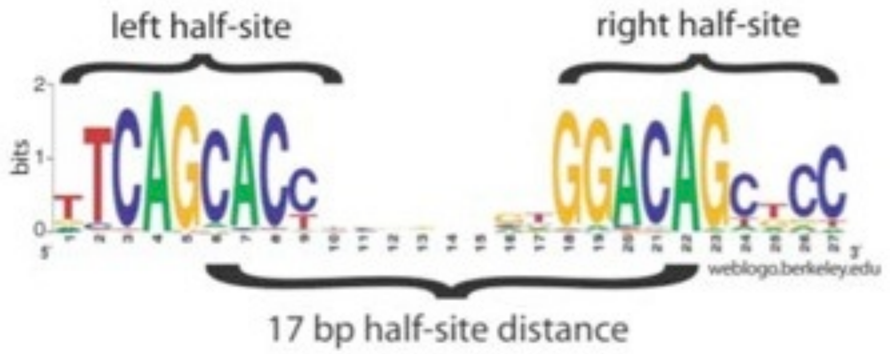
Mapping onto genome

ChIP-seq reveals new binding motif flexibility for NRSF

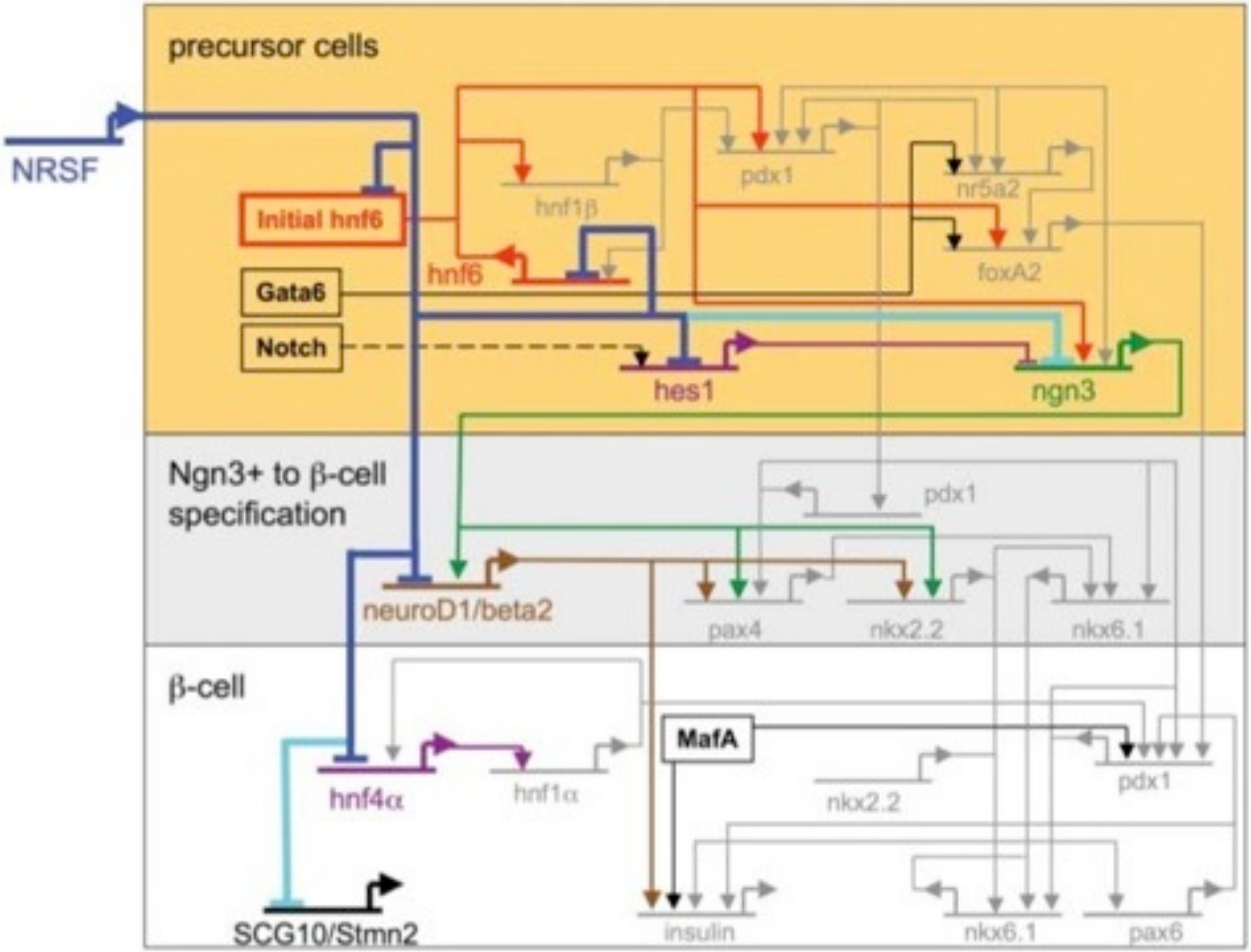
Canonical NRSF PWM logo



Novel NRSF PWM logo



The gene regulatory network downstream of NRSF constructed from ChIP-seq data



Summary for ChIP based target prediction methods

•ChIP-chip and ChIP-seq allow for the first time physical identification of bound regions on the genomic scale

•ChIP-seq presents higher resolution and is replacing ChIP-chip

•Both methods require large data-processing and analysis

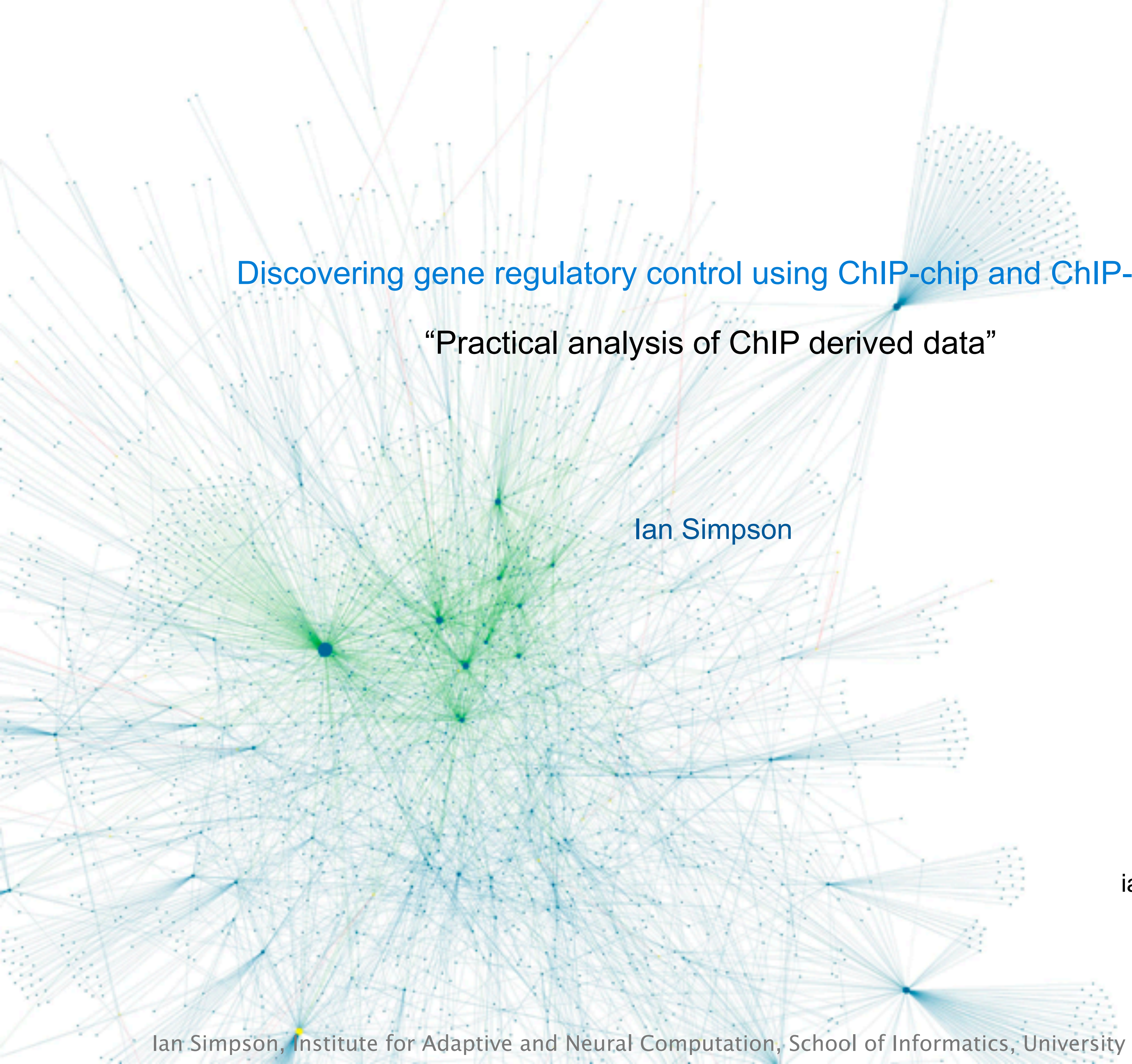
•Novel methods have been developed to call bound regions from these data

they are predominantly based on hidden markov models (HMM) and are naturally normally 2-state models (peak, non-peak)

•The resulting regions can be used with classical methods to refine the nature of the regulatory element (PWM Gibbs/HMM profiling, motif detection, conservation)

•Can also be refined by more precise experiments on the ChIP DNA such as targeted PCR

•Revolutionises the analysis of gene regulatory networks by integration with gene expression data

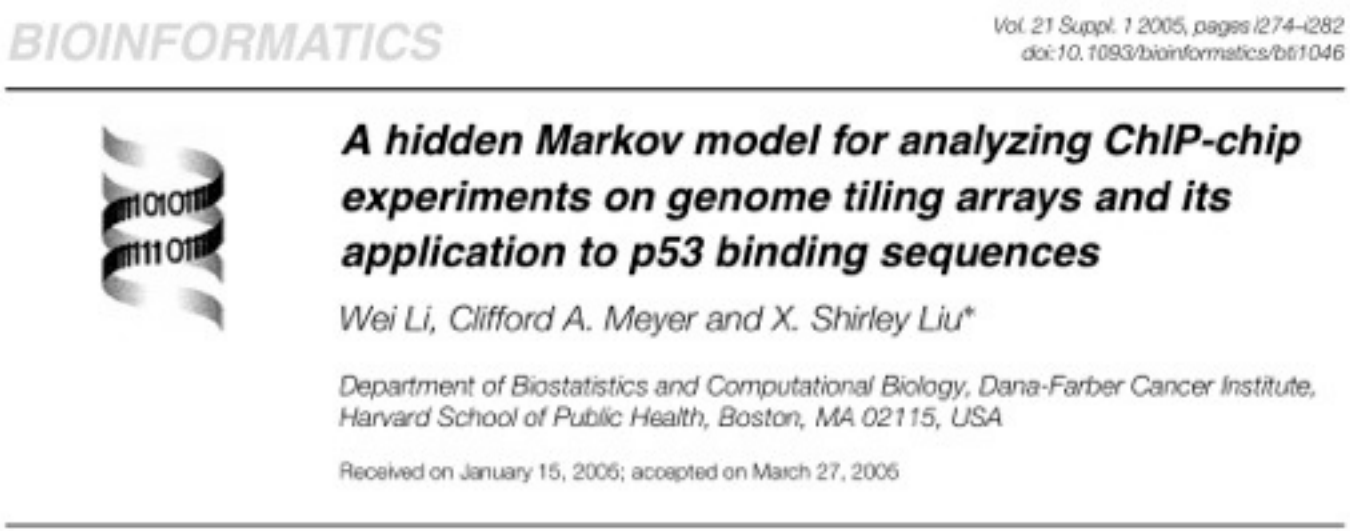


Discovering gene regulatory control using ChIP-chip and ChIP-seq

“Practical analysis of ChIP derived data”

Ian Simpson

ian.simpson@.ed.ac.uk
http://bit.ly/bio2links



BIOINFORMATICS

Vol. 21 Suppl. 1 2005, pages i274–i282
doi:10.1093/bioinformatics/bti1046

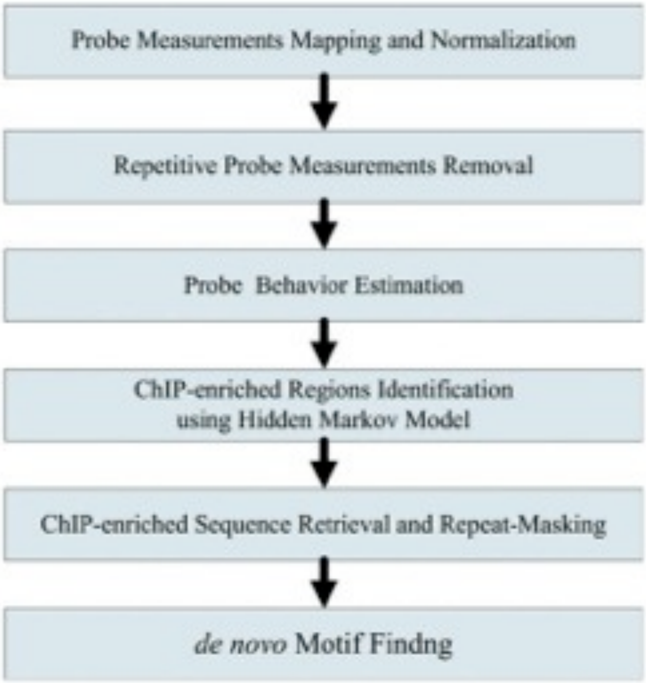
A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences

Wei Li, Clifford A. Meyer and X. Shirley Liu*

Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute,
Harvard School of Public Health, Boston, MA 02115, USA

Received on January 15, 2005; accepted on March 27, 2005

HMMtiling



Probe Measurements Mapping and Normalization

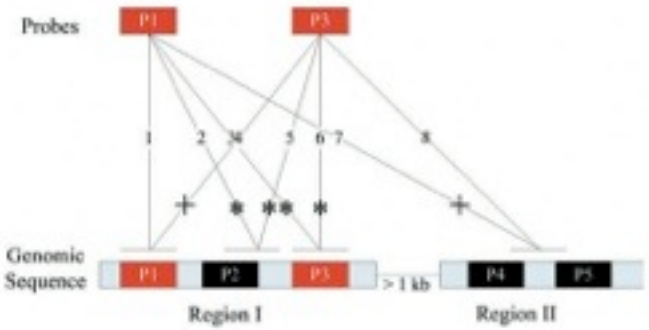
Repetitive Probe Measurements Removal

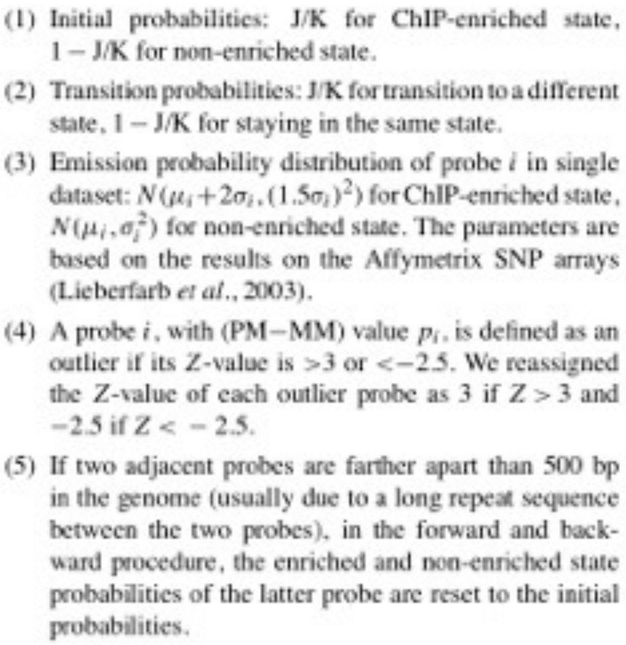
Probe Behavior Estimation

ChIP-enriched Regions Identification using Hidden Markov Model

ChIP-enriched Sequence Retrieval and Repeat-Masking

de novo Motif Findng





(1) Initial probabilities: J/K for ChIP-enriched state, 1 − J/K for non-enriched state.

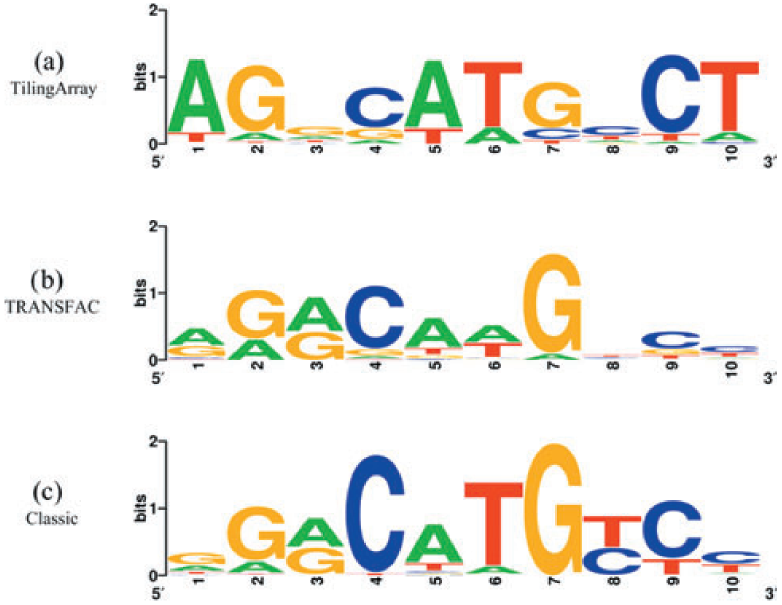
(2) Transition probabilities: J/K for transition to a different state, 1 − J/K for staying in the same state.

(3) Emission probability distribution of probe i in single dataset: $N(\mu_i + 2\sigma_i, (1.5\sigma_i)^2)$ for ChIP-enriched state, $N(\mu_i, \sigma_i^2)$ for non-enriched state. The parameters are based on the results on the Affymetrix SNP arrays (Lieberfarb *et al.*, 2003).

(4) A probe i, with (PM−MM) value p_i, is defined as an outlier if its Z-value is >3 or <-2.5. We reassigned the Z-value of each outlier probe as 3 if $Z > 3$ and -2.5 if $Z < -2.5$.

(5) If two adjacent probes are farther apart than 500 bp in the genome (usually due to a long repeat sequence between the two probes), in the forward and backward procedure, the enriched and non-enriched state probabilities of the latter probe are reset to the initial probabilities.

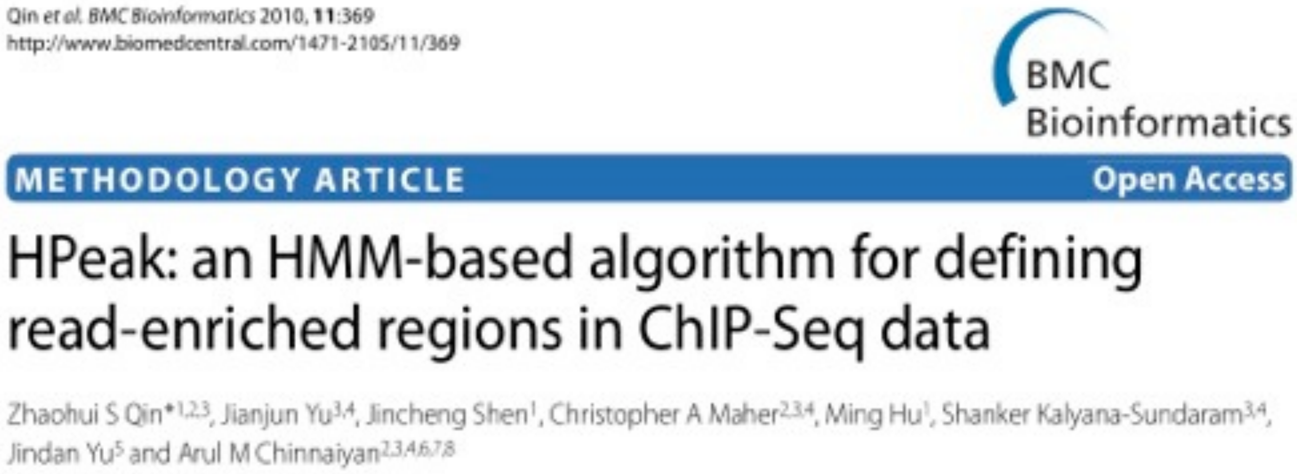
Comparison to known p53 binding sites



Tiling Array MDScan profile

Published PWM in Transfac

The canonical consensus



Qin *et al. BMC Bioinformatics* 2010, **11**:369
http://www.biomedcentral.com/1471-2105/11/369

BMC Bioinformatics

METHODOLOGY ARTICLE

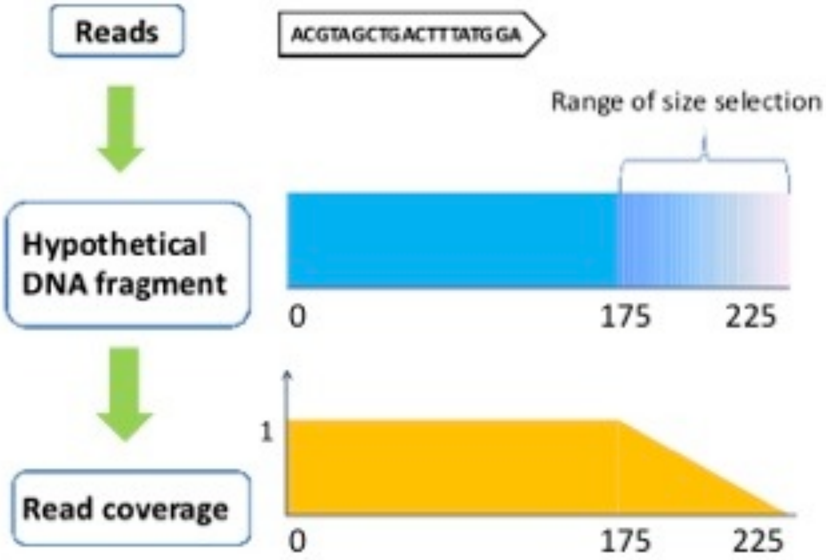
Open Access

HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data

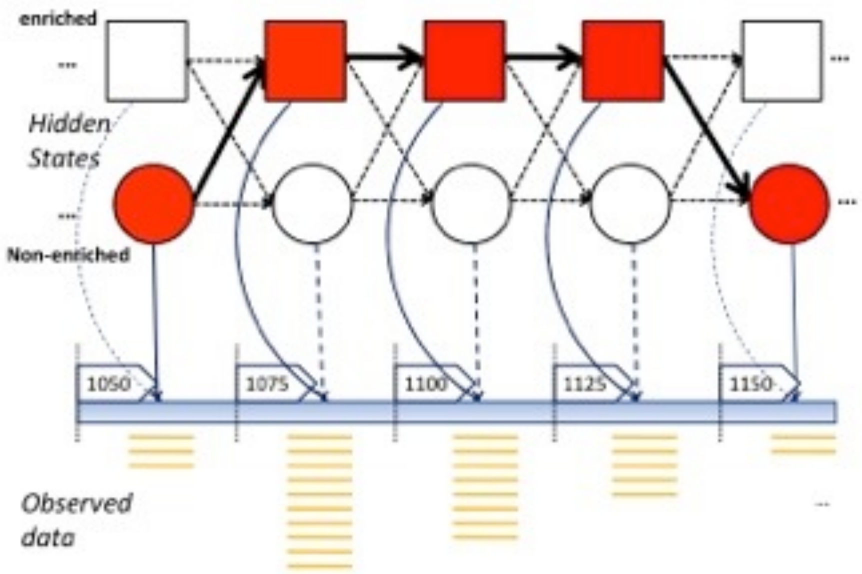
Zhaohui S Qin*[1,2,3], Jianjun Yu[3,4], Jincheng Shen[1], Christopher A Maher[2,3,4], Ming Hu[1], Shanker Kalyana-Sundaram[3,4], Jindan Yu[5] and Arul M Chinnaiyan[2,3,4,6,7,8]

HPeak

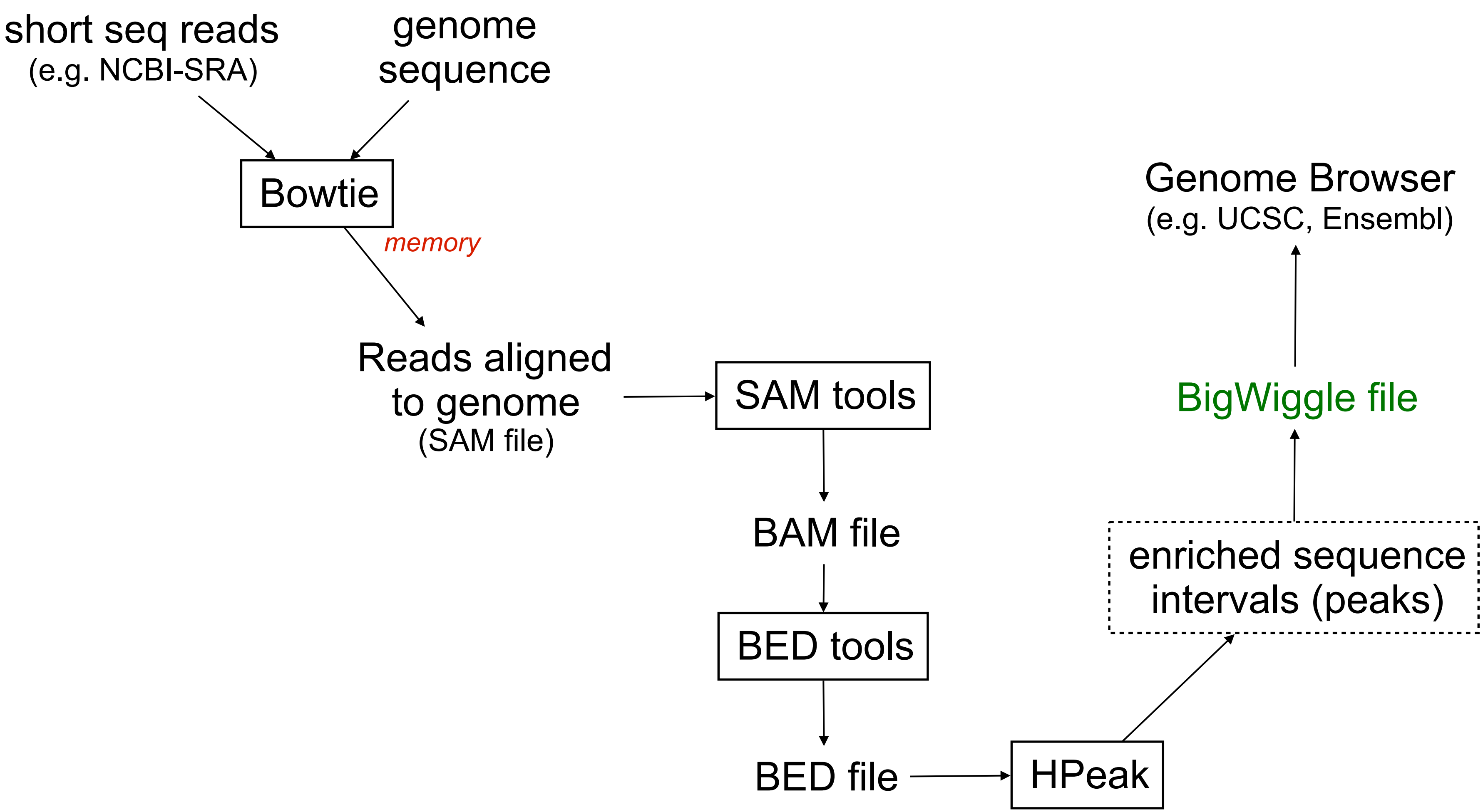
Hypothetical DNA fragment



Model architecture



ChIP-seq, from short sequence reads to enriched intervals analysis pipeline using HPeak



Finding and using resources

Where to find the data

How to visualise genome scale data