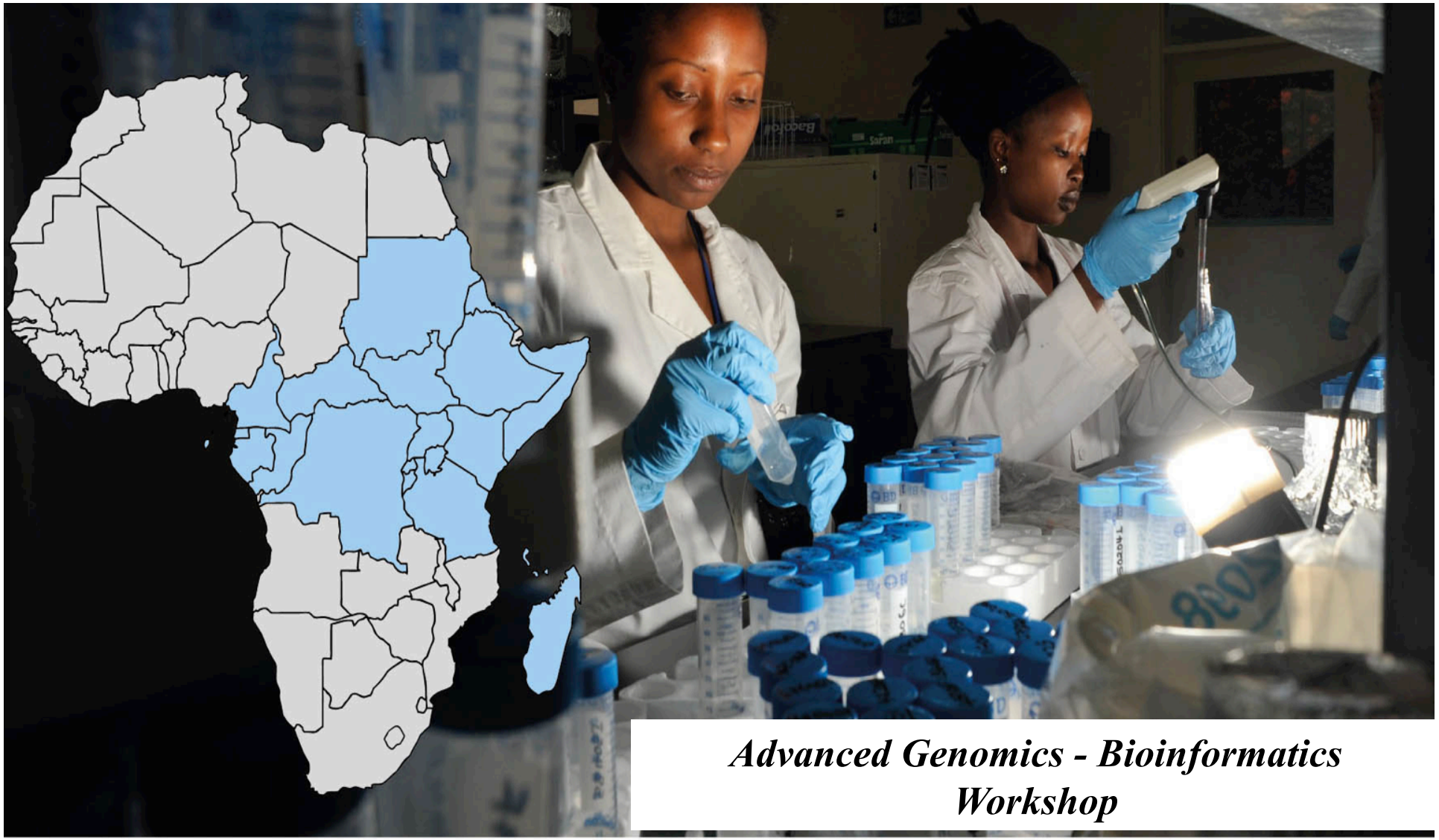


Advanced Genomics - Bioinformatics Workshop

Mark Wamalwa
BecA-ILRI Hub, Nairobi, Kenya
http://hub.africabiosciences.org/
m.wamalwa@cgiar.org

7th – 18th September 2015

biosciences eastern and central africa

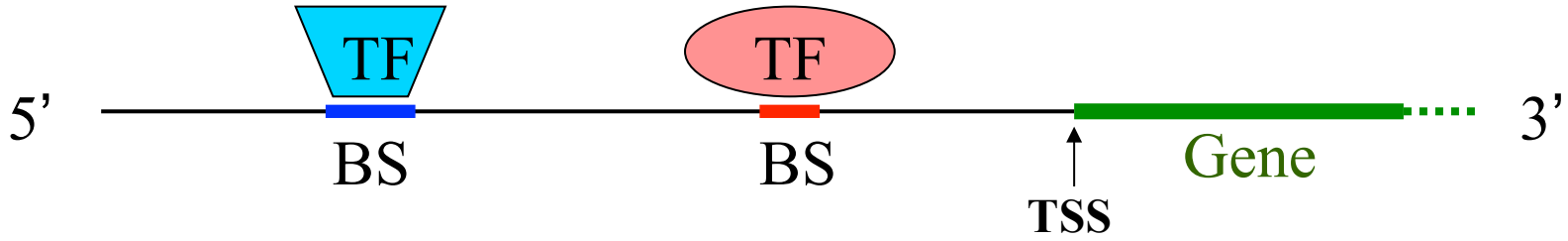
REGULATORY SEQUENCE ANALYSIS TOOLS (RSAT)

Motif Discovery Platform

Promoter Analysis:

Exteremely brief intro

- **Transcription is regulated primarily by transcription factors (TFs) – proteins that bind to DNA subsequences, called binding sites (BSs)**
- **TFBSs are located mainly (not always!) in the gene's promoter – the DNA sequence upstream the gene's transcription start site (TSS)**
- **TFs can promote or repress transcription**



Promoter Analysis (cont.)
TFBS models

- **The BSs of a particular TF share a common pattern, or motif, which is often modeled using:**
  - **Consensus string**

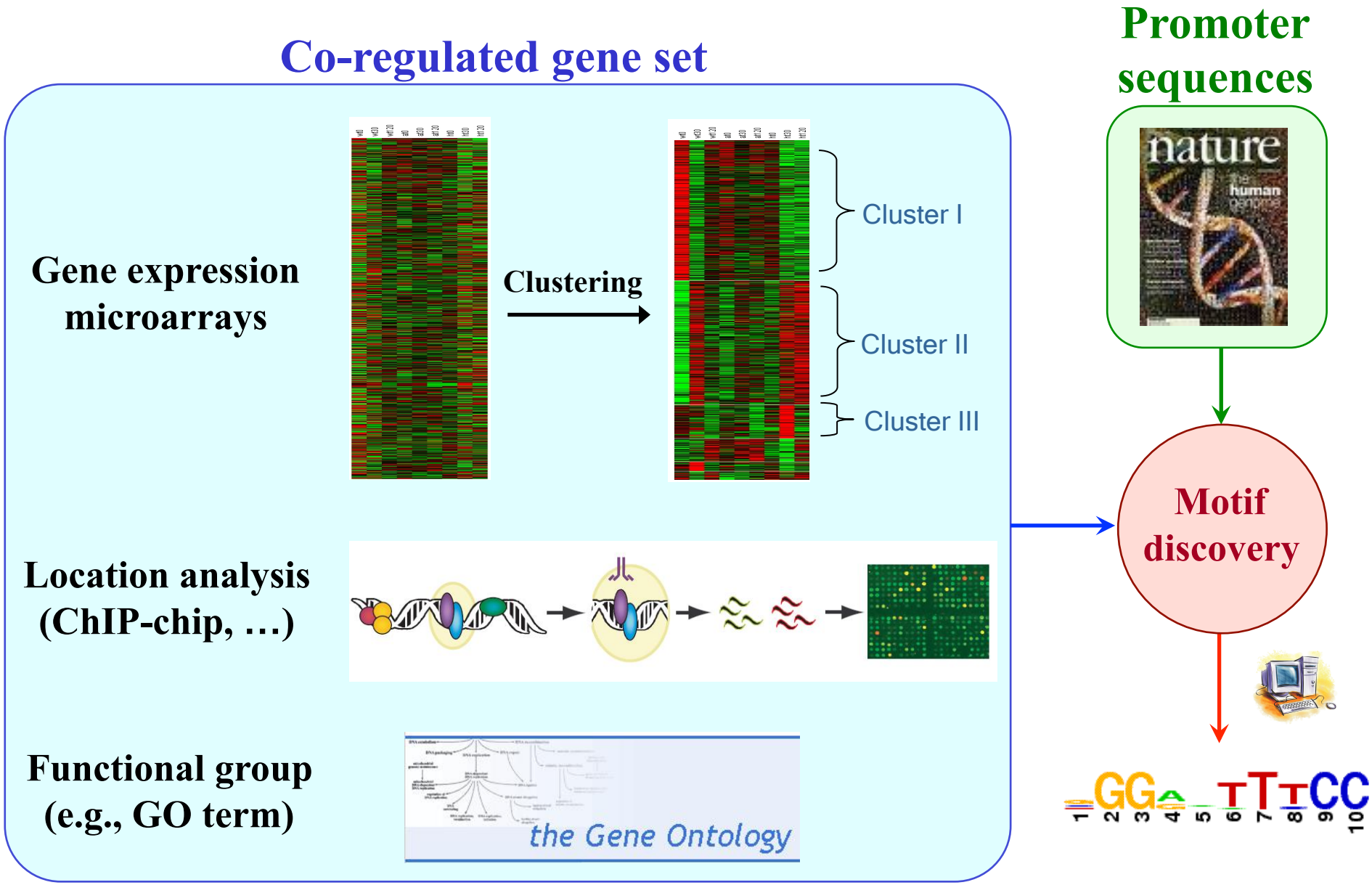
    `TASDAC (S={C,G} D={A,G,T})`
  - **Position weight matrix (PWM / PSSM)**

A	0.1	0.8	0	0.7	0.2	0
C	0	0.1	0.5	0.1	0.4	0.6
G	0	0	0.5	0.1	0.4	0.1
T	0.9	0.1	0	0.1	0	0.3

> Threshold = 0.01:

TACACC (0.06)
TAGAGC (0.06)
TACAAT (0.015)
…

Promoter Analysis (cont.): Typical pipeline



Promoter Analysis (cont.): Goals

Reverse-engineer the transcriptional regulatory network = find the TFs (and their BSs) that regulate the studied biological process

Input: A set of co-expressed genes

Output: "Interesting" motif(s):

1. **Known motifs:**
   PRIMA, ROVER, …
2. **Novel motifs:**
   MEME, AlignACE, …
3. A **group** of co-occurring motifs = cis-regulatory module (**CRM**):
   MITRA, CREME, …

RSAT

Promoter Analysis (cont.): Challenges

Why is it so difficult?

- **BSs** are short and degenerate (non-specific)
- **Promoters** are long + complex (hard to model):
  - Multiple BSs of several TFs
  - Old (non-functional) BSs
  - Other genetic/structural signals (e.g., GC content)
- **Search space** is huge:
  - 15^{10} (500 billion) consensus strings of length 10
  - 1Kbp promoter × 20K genes in human = 20 Mbps
- Which **score** to use - what makes a motif "interesting"?
  - Enrichment: over-representation w.r.t. BG model
  - Location and/or strand bias
  - Conservation across related species

Promoter Analysis (cont.): Challenges (II)

- **Additional complications: alternative promoters, wrong TSS annotations, paralogs (→ dependencies), …**
- **Many TFs have BSs in distant upstream locations, as well as in introns, UTRs, …**

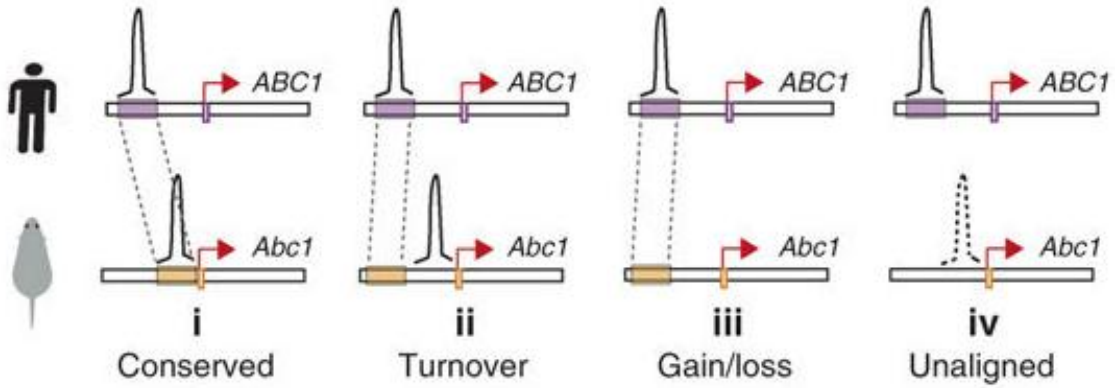
[Lin et al. '07]: Used ChIP-PET to identify BSs of ER-α in breast cancer cells.

- **Only 5% of BSs are within 5kb upstream of TSS!**
- **Only 23% of the BSs are conserved among vertebrates, "which suggests limited conservation of functional binding sites".**

Promoter Analysis (cont.): Challenges (III)

[Odom et al. '07]: Used ChIP-chip to map BSs of 4 TFs in human+mouse liver.

- Function and binding motifs are conserved
- 41-89% of BSs are species specific
- When a pair of orthologous genes contain a BS of the same TF, the BSs are aligned only in 1/3 of the cases



Promoter Analysis:

Status of motif discovery tools

- **Extant tools perform reasonably well for:**
  - **Finding known/novel motifs in organisms with short, simple promoters, e.g., yeast**
  - **Identifying some of the known motifs in complex species, e.g., TFs whose BSs are usually close to the TSS**
- **… but often fail in other cases!**
- **Each tool is custom-built for a *specific* target score, often *parametric* (i.e., assumes a BG model) or uses a *small* part of the genome as BG reference;**

  Majority of tools can efficiently handle only *dozens* of genes
- **Comparison of tools: [Tompa et al. '05]**

RSAT - TOOLS

- **Research platform:**
  - **Extensible: add new algs, scores, motif models**
  - **Flexible: control params, algs, scores of execution**
- **Experimental tool:**
  - **Sensitive: find subtle signals**
  - **Efficient: analyze many long sequences**
  - **Informative: show lots of info on motifs**
  - **User-friendly: nice GUI**

Main features: I/O

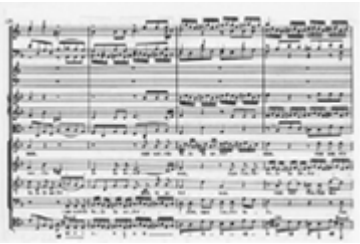
Input:

- **Type: target set / expression data**
- **Multiple species / target-sets**
- **Sequence region (promoter, 1st intron, 3' UTR, …)**

Output:

- **Non-redundant set of motifs**
- **Rich info per output motif:**
  1. **Graphical motif logo**
  2. **Multiple scores & combined *p*-value**
  3. **Similarity to known TFBS models**
  4. **List of target genes**
  5. **BS localization graph**
  6. **Targets mean expression graph**

Main features: scores



Motif scores:

- **User selects scores to use, a subset of:**
  - **<u>Target-set:</u> Over/under-representation:**
    1. **Hypergeometric**
    2. **GC-content+length binned binomial**
  - **<u>Expression:</u>**
    1. **Enrichment of ranked expression (multiple conditions) (Not yet in the public version)**
  - **<u>Global/spatial:</u>**
    1. **Localization**
    2. **Strand-bias**
    3. **Chromosomal preference**
- **Scores are combined into a single *p*-value**
- **Doesn't assume specific models for distribution of BSs and/or expression values**

Main features: misc.

GUI:

- Control all parameters
- Save/load parameters from file
- Save textual+graphical output to file
- TFBS viewer

Other:

- Ignore redundant sequences (with identical subsequence)
- Applicable to multiple genome-scale promoter sequences
- Bootstrapping: Empirical *p*-value estimation using random target sets / shuffled data
- Execution modes: GUI , batch
- Interoperability: Java application

Combining *p*-values

Each motif receives *p*-values from various sources (several scores, multiple species): $p_1, p_2, \ldots, p_n$

We combine them into a single *p*-value *p*:

$$p = \text{Prob}\{ \varphi_1 \cdot \varphi_2 \cdot \ldots \cdot \varphi_n \le p_1 \cdot p_2 \cdot \ldots \cdot p_n \mid \varphi_i \sim U[0,1] \}$$

Denote: $\phi = p_1 \cdot p_2 \cdot \ldots \cdot p_n$

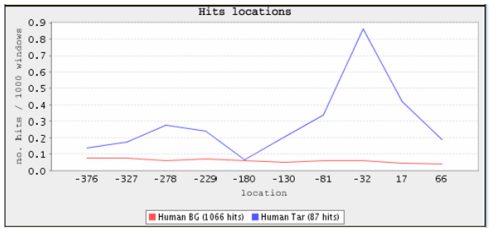
$$\rightarrow p = 1 - \phi \cdot \Sigma(\ln 1/\phi)^i / i! \quad , i=0,\ldots,n-1$$

Also developed a weighted version when each *p*-value has a different weight

Case study
Global Analysis I:
Localized human+mouse motifs

Input:

- All human & mouse promoters (2 x ~20,000)
- Region: -500…100 (w.r.t. TSS)
- Total sequence length: ~26 Mbps
- [No target-set / expression data]
- Score: localization



Results:

- Recovered known TFs:

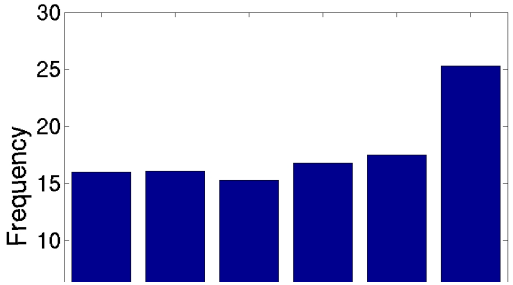
  Sp1, NF-Y, GABP, TATA, Nrf-1, ATF/CREB, Myc, RFX1

- Recovered the splice donor site
- Identified several novel motifs

Global Analysis II:
Chromosomal preference

Input:

- **All fly promoters (~14,000)**
- **Region: -1000…200 (w.r.t. TSS)**
- **Total sequence length: ~11 Mbps**
- **[No target-set / expression data]**
- **Score: chromosomal preference**



R

- In *Drosophila*, dosage compensation is achieved by a twofold up-regulation of the male X-linked genes and requires the association of the male-specific lethal complex (MSL) on the X chromosome. How the MSL complex is targeted to X-linked genes and whether its recruitment at a local level is necessary and sufficient to ensure dosage compensation remain poorly understood. Here we report the MSL-1-binding profile along the male X chromosome in embryos and male salivary glands isolated from third instar larvae using chromatin immunoprecipitation (ChIP) coupled with DNA microarray (ChIP–chip). This analysis has revealed that majority of the MSL-1 targets are primarily expressed during early embryogenesis and many target genes possess DNA replication element factor (DREF)-binding sites in their promoters. In addition, we show that MSL-1 distribution remains stable across development and that binding of MSL-1 on X-chromosomal genes does not correlate with transcription in male salivary glands. These results show that transcription per se on the X chromosome cannot be the sole signal for MSL-1 recruitment. Furthermore, genome-wide analysis of the dosage-compensated status of X-linked genes in male and female shows that most of the X chromosome remains compensated without direct MSL-1 binding near the gene. Our results, therefore, provide a comprehensive overview of MSL-1 binding and dosage-compensated status of X-linked genes and suggest a more global effect of MSL complex on X-chromosome regulation.

Global Analysis II:
Chromosomal preference (cont.)

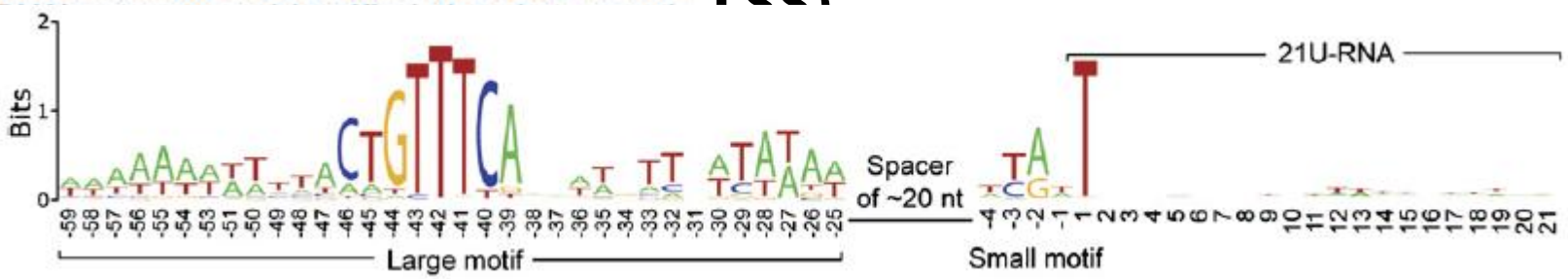
Input:

SUMMARY

We sequenced ~400,000 small RNAs from *Caenorhabditis elegans*. Another 18 microRNA

000)

rss)



a third class of nematode small RNAs, called 21U-RNAs, was discovered. 21U-RNAs are precisely 21 nucleotides long, begin with a uridine 5′-monophosphate but are diverse in their remaining 20 nucleotides, and appear modified at their 3′-terminal ribose. 21U-RNAs originate from more than 5700 genomic loci dispersed in two broad regions of chromosome IV—primarily between protein-coding genes or within their introns. These loci share a large upstream motif that enables accurate prediction of additional 21U-RNAs. The motif is conserved in other nematodes, presumably because of its importance for producing these diverse, autonomously expressed, small RNAs (dasRNAs).

207

ng Reveals
nal MicroRNAs
As in *C. elegans*

ael J. Axtell,[1,4] William Lee,[3] Chad Nusbaum,[3]

Summary

- **RSAT Tools:**
  - **Easy to use**
  - **Feature-rich, informative**
  - **Sensitive & efficient**
- **Constructed a large, real-life, heterogeneous benchmark for testing motif finding tools**
- **Demonstrated various applications of motif discovery**
- **http://41.204.190.30/rsat/**
- **http://acgt.cs.tau.ac.il/amadeus**

Acknowledgements

Chaim Linhart
Yonit Halperin
Ron Shamir