Thirsting for Theoretical Biology
International Centre for Theoretical Sciences, Bangalore
June 3-7, 2019

Making sense of the threads of ACGT

Rakesh K Mishra

CCMB

How information is

- accumulated,

- stored/maintained &

- expressed

in DNA based information system of life?

Information content in the DNA sequence and its biological output

- The basis of all forms of life on earth -

- The substrate for the evolutionary process -

How information is encoded, regulated and expressed ?

If we can ever tell, looking at the DNA sequence,

- **the shape, size, behavior, etc., of its owner…**
- **response to biotic and abiotic factors, …**
- **the disease susceptibility, longevity, etc.,**

That would meant that we have answered this question!

Why it is getting more and more important

The new ways of understanding biology and potential applications

Availability of increasingly large data sets

Genomes, exomes, ESTs, …

ENCODE (ENCyclopedia **O**f **D**NA **E**lements) **modENCODE**

(**mod**el organism **ENC**yclopedia **O**f **D**NA **E**lements)

New HTP techniques

NGS and its multiple applications other than genome sequencing (cheaper/faster DNA sequencing)
4C, 5C, HiC, …
Bioinformatic/computational tools for DNA and Genome analysis

The genomic way of healthcare, lifestyle & agriculture

Personalized and precision medicine

Designer's plants and animals: genome editing technology
fast track breeding / screening

Large scale data generation programs

- **Earth BioGenome Project**
- **Population level human genome sequencing projects**
- **Microbiome**
- **eDNA**

How information is encoded, regulated and expressed ?

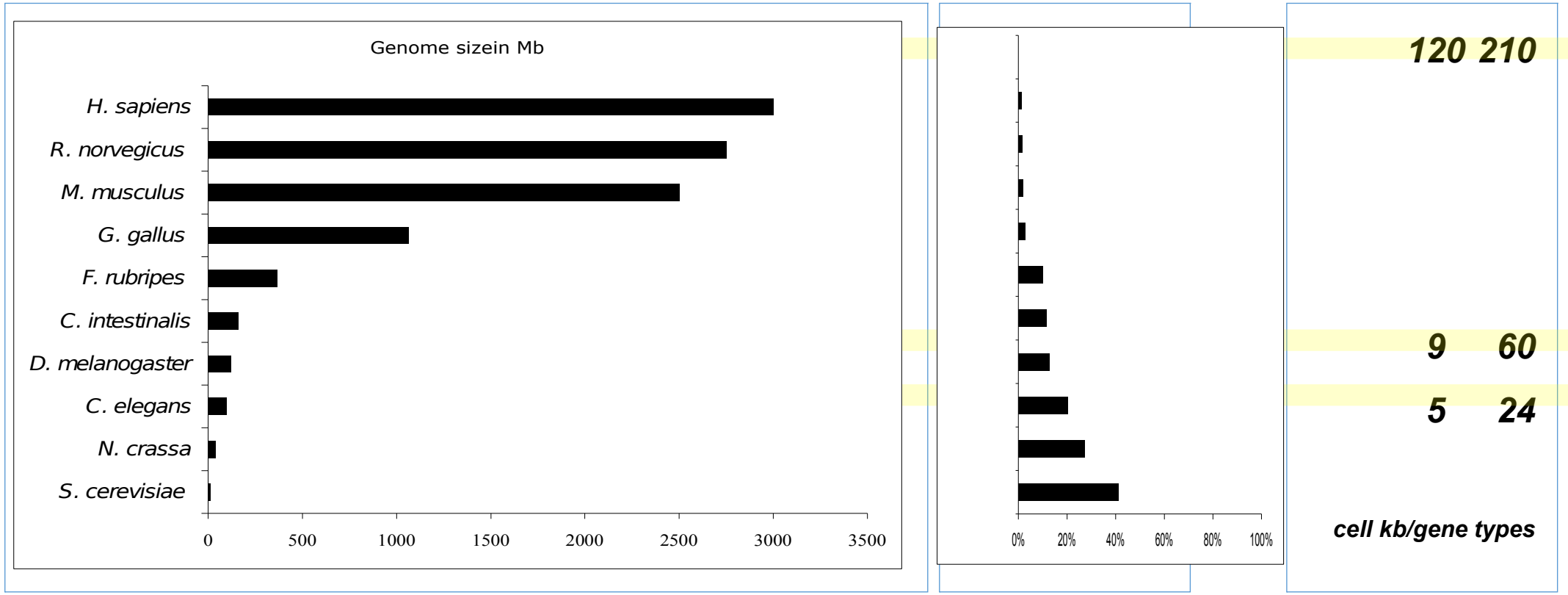
One approach may be to compare genomes of organisms of different level of complexity to decipher the information and DNA sequence relationship.

By this, we can ask:

What is it that adds to increase in complexity?

Genome size & number of genes

from simple to complex organisms



Outcome of evolutionary process:

Static number of genes but more non-coding DNA (more DNA/gene)

Parallel to the increasing complexity / epigenome potential

What is the function of this 'excess' DNA?

Kumar et al. BioEssays 2010

Reading the information content of the genome

[What is the function of the ‘excess’ DNA?]

Evolution of complexity

-not by more genes
-by more sophisticated regulation of genes!

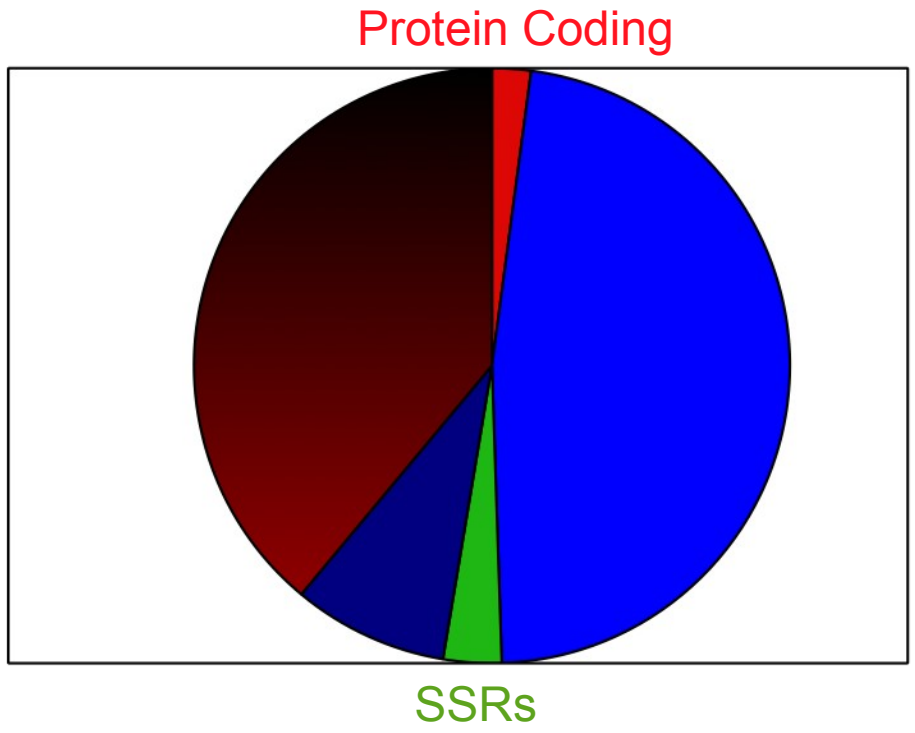
Multiple outputs from one genome

-each cell type has its epigenome
-multiple ways to package a complex genome!

Non-coding part of the genome

-contains the regulatory elements
-has the cell type specific packaging code

Composition of a genome
(human genome)



Many functional elements of genome are yet to be discovered

Why so much DNA?

How one genome acquires cell type specific distinct functional forms: ‘epigenomes’?

How to find novel functions of genomic elements?

Function in the non-coding part of genome

Providing grammar to the genomic language

common regulatory elements : Promoter, enhancer, repressor, …
[binding sites for transacting factors]

introns & inter genic regions : ? Variety of other regulatory elements

repetitive elements : ? Regulatory, selfish, mutagenic,
stress response, stability, dynamic, …

Function in the non-coding part of genome

Providing grammar to the genomic language

Differential packaging: genome packaging code
[cell type specific epigenome]

Chromatin level regulation: long-range interaction & sub-nuclear compartmentalization

ncRNA mediated effects: local / sequence specific effects
structural role

Regulatory elements in complex genomes

- *Repeats [SSRs] abundance & patterns*		*4 %*
- Motif cluster / patterns (boundary & PREs)		*7 %*
- Conservation across species – CNCS		*4 %*
- Context dependent search		*-*
- RNA version of repetitive DNA	*HTP*	*-*
- Biochemically defined regions [MAR/SAR]	*HTP*	*3 %*
- Epigenetic marking / patterns	*HTP*	*-*
- Accessibility	*HTP*	*-*
- Long range interactions [CCC…]	*HTP*	*-*
		~18 %

Bioinformatic and experimental approaches to identify novel functional elements in genomes

BMC Genomics 2019
Bioinformatics 2018
Genome Biol Evol 2018
Genome Biol Evol 2017
Gene 2017
BMC Genomics 2014
Gene 2014
Nucleic Acids Research 2013
Nucleic Acids Research 2012
Nucleic Acids Research 2011
J. Mol. Biol. 2010
Development 2010
BMC Bioinformatics 2010
BMC Genomics 2009
BMC Genomics 2004
Bioinformatics 2003a
Bioinformatics 2003b
Genome Biology 2003a
Genome Biology 2003b

Function in the non-coding part of genome

Providing grammar to the genomic language

Gopal Kushwah

Conservation across species – CNCS

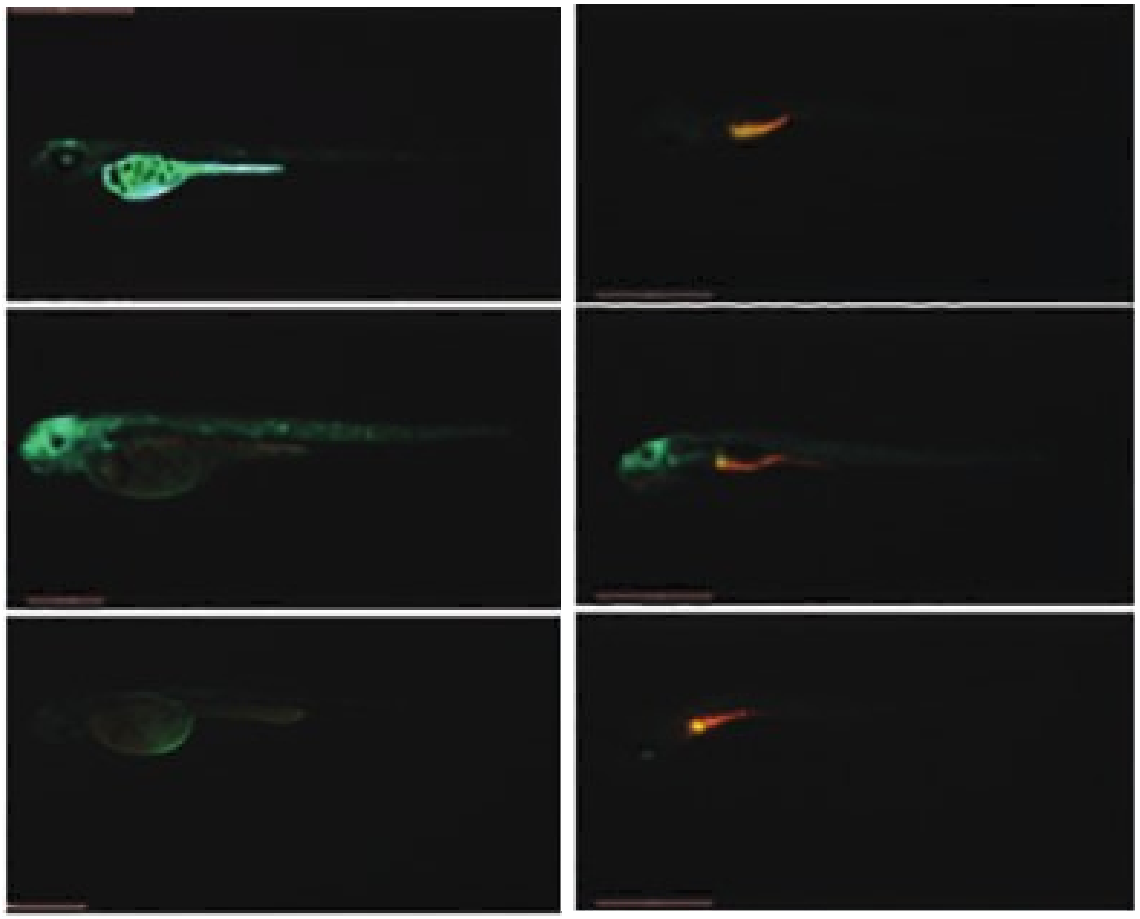
(ultra conserved sequences)

Vertebrate utility!

Near developmentally regulated genes

Increasing size from fish to mammal

3-5% of the genome



Sabarinadh et al. Genome Biol 2003
Sabarinadh et al. BMC Genomics 2004
Kushwah & Mishra Genome Biol & Evo. 2018

Function in the non-coding part of genome

Providing grammar to the genomic language

Motif cluster & pattern

chromatin domain Boundary Element Search Tool *[cdBEST]*

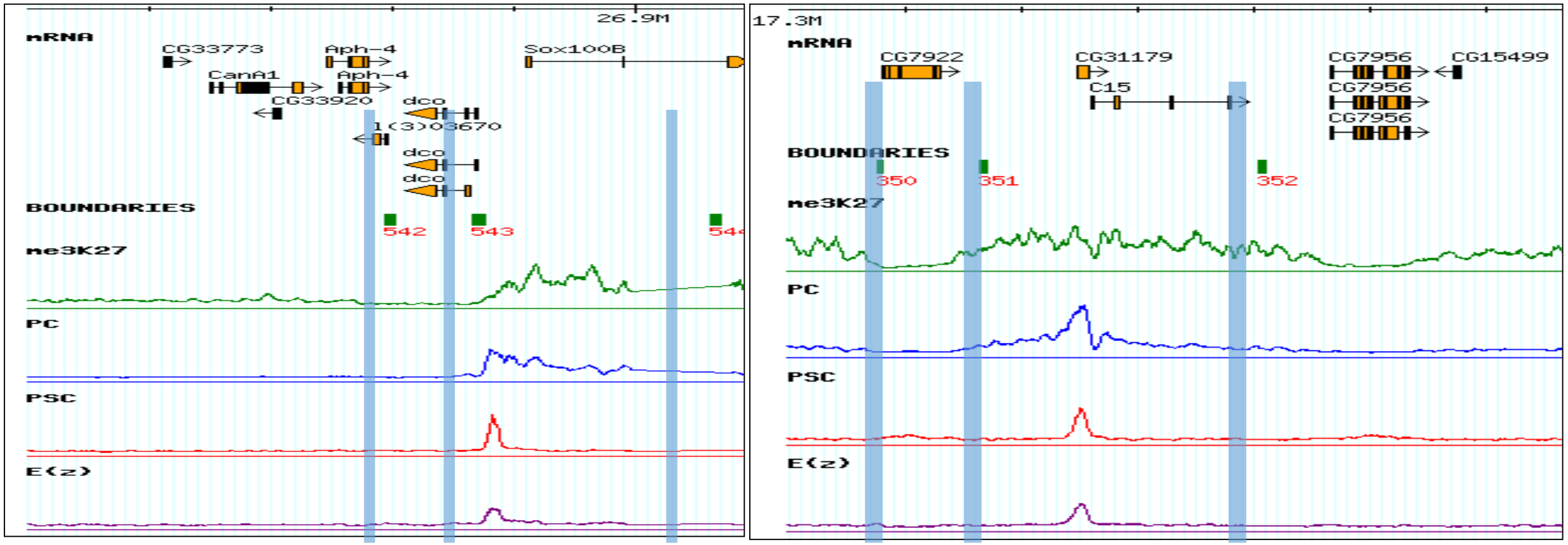
- **>4500 boundary elements predicted in *Drosophila melanogaster***
- **Great majority locating in the intergenic regions**
- **Genomes of 12 species of *Drosophila* analysed give similar results**
- **Transposable elements as boundaries is common feature in all drosophilids**
- **Accounts for ~3% of the genome**
- **>80% of cdBEST predictions function as boundaries *in vivo***
- **Applicable to other insects (e.g., malaria mosquito *An. gambiae*)**

A Srinivasan

Srinivasan & Mishra, Nucleic Acids Res. 2012

Ahanger et al Nucleic Acids Res.2013

Borders of genes and epiprofile



[separating functional domains of genome]

Polycomb Response Elements prediction tool (PREPT)

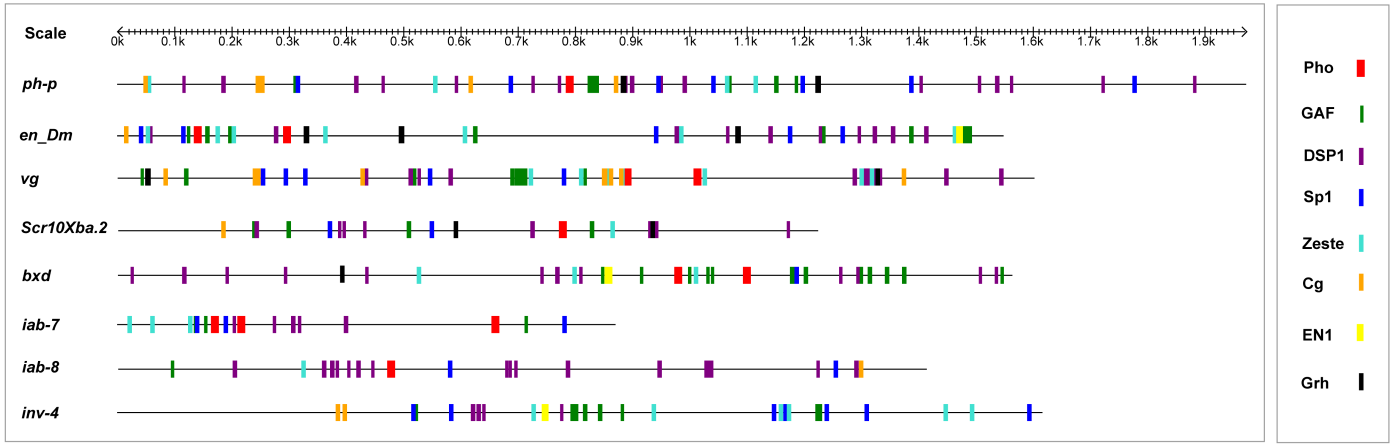
An improved PRE prediction tool that identified 8040 PREs in *Drosophila* genome with an average 6.7 PREs per 100 kb of DNA.

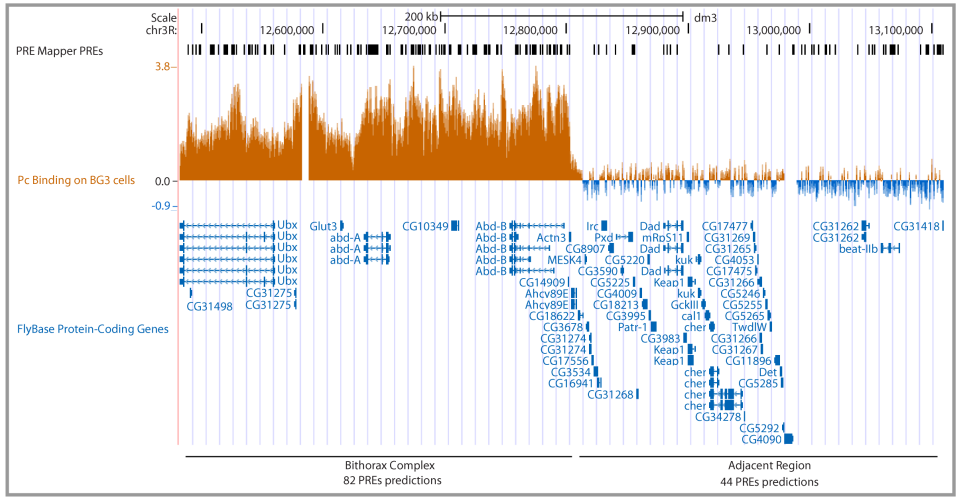
Identified PREs represent 3-4% of *Drosophila* genome.

PREs and boundaries often found co-habited in the genome, but bound by different proteins factors.

Relationship with TAD boundaries

Relationship with certain group of promoters





Function in the non-coding part of genome

Providing grammar to the genomic language

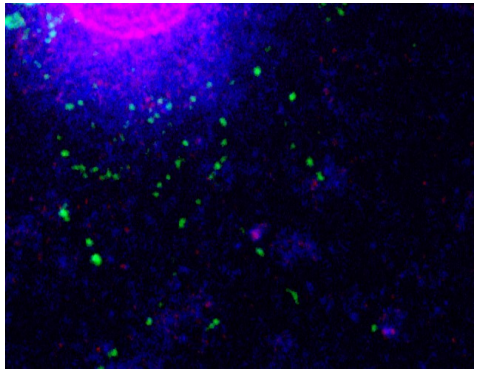
Srinivasan

Rashmi U Pathak

Matrix Associate Region (MARs)

MARs are the regions of genomic DNA that attaches with the nuclear matrix and thought to play an important role in higher order chromatin organization.

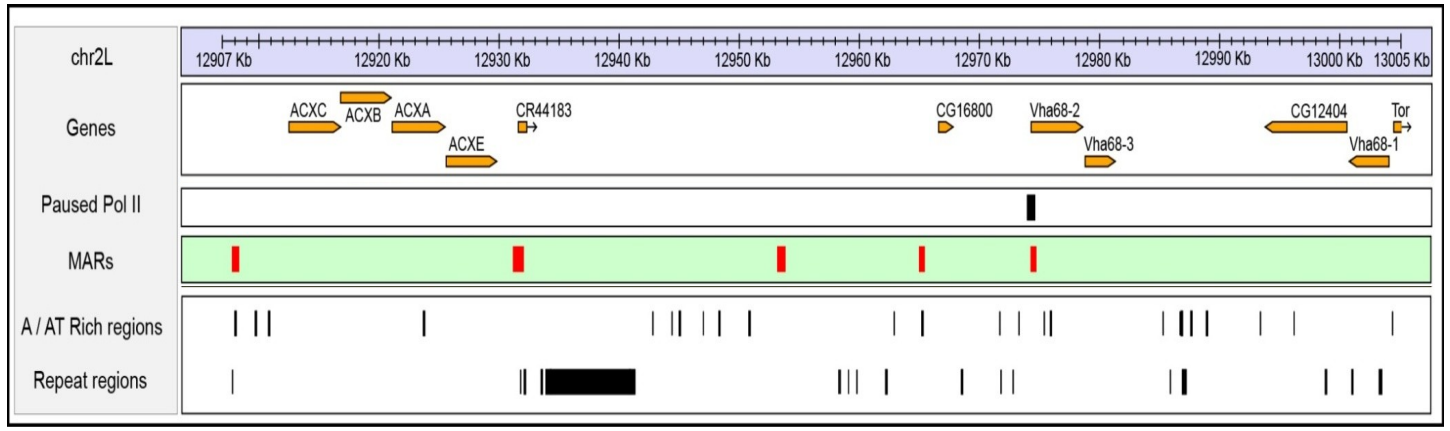
We used NGS approach and identified >7000 MARs across the euchromatic portion the *Drosophila* genome.



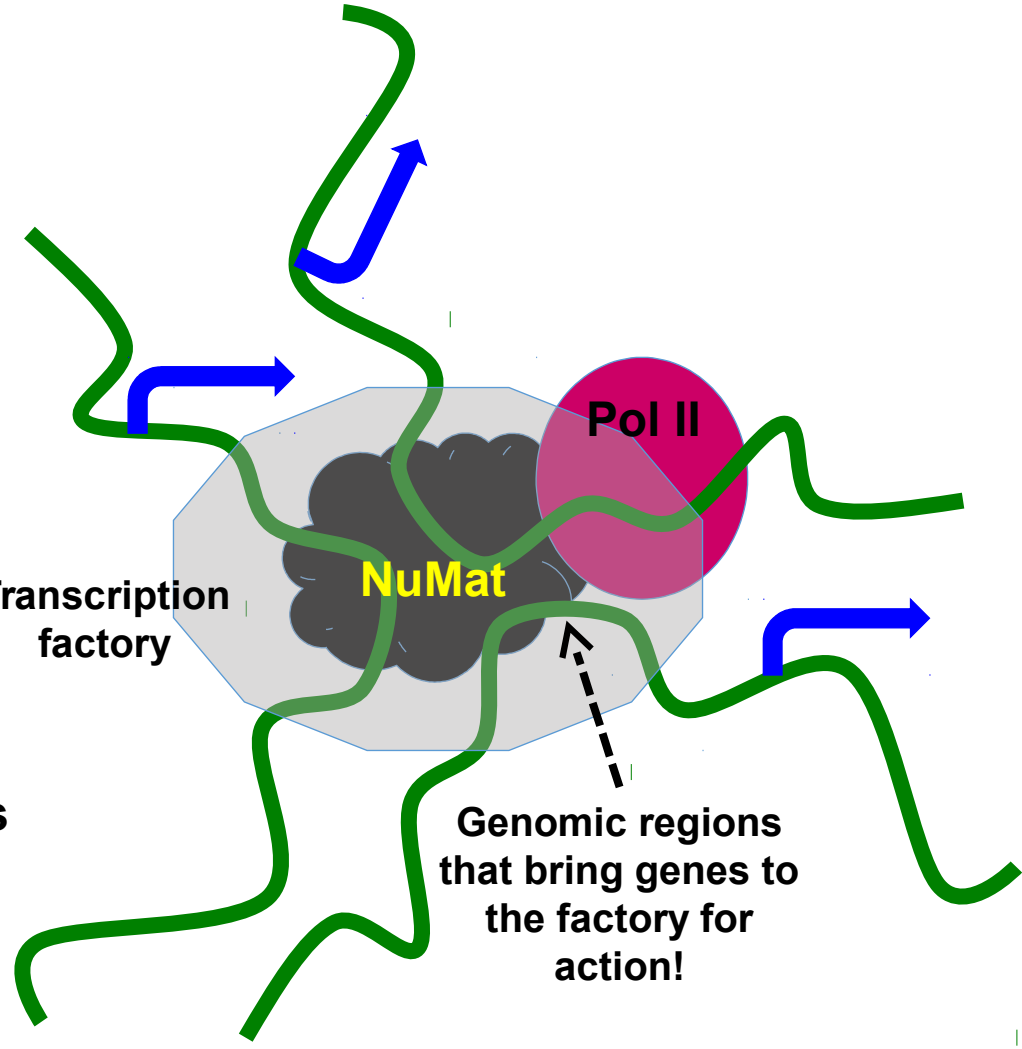
This accounts for about 2.5% of the *Drosophila* genome!

Pathak et al, BMC Genomics 2014

Anchoring chromatin regions to functional nuclear compartments?

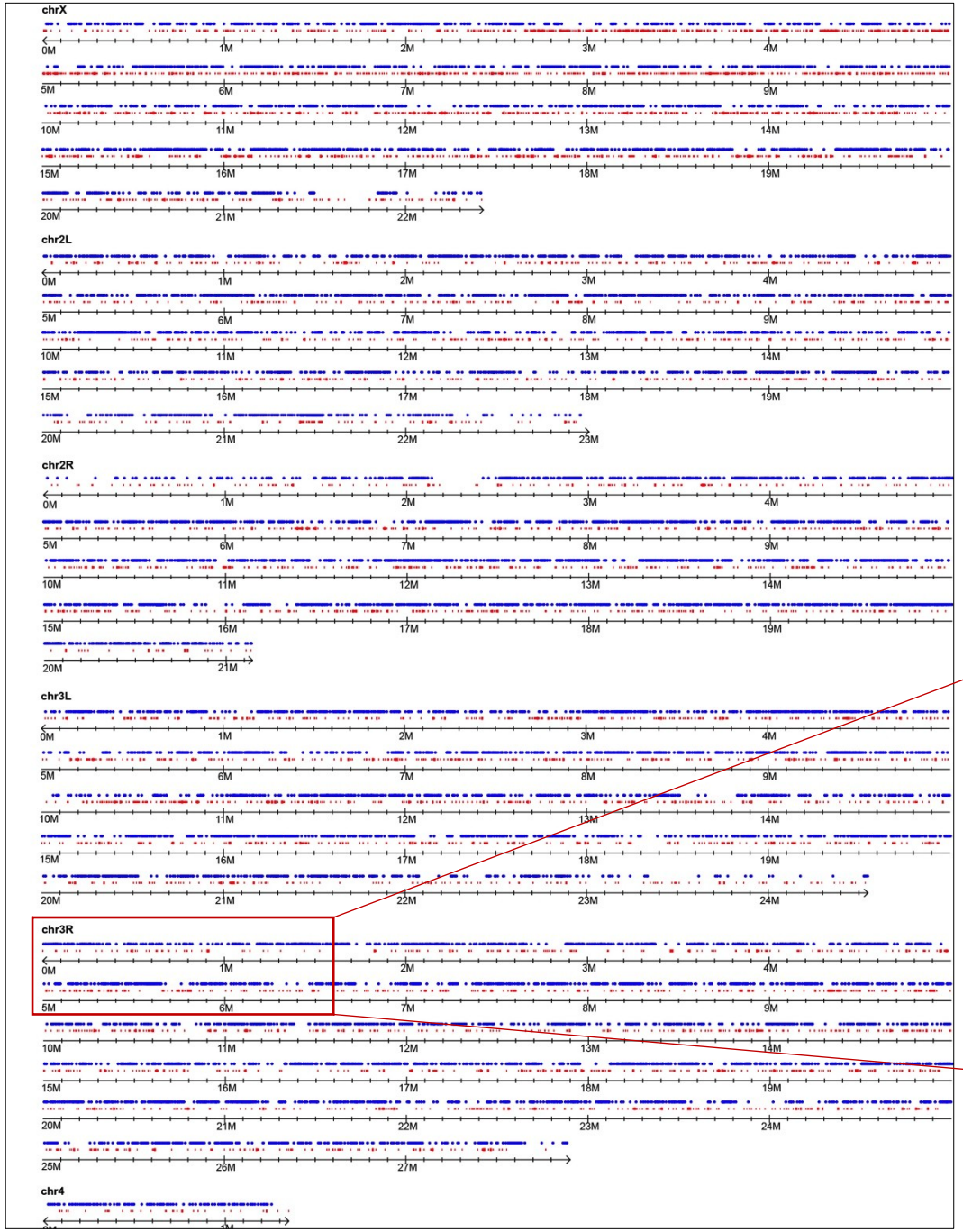


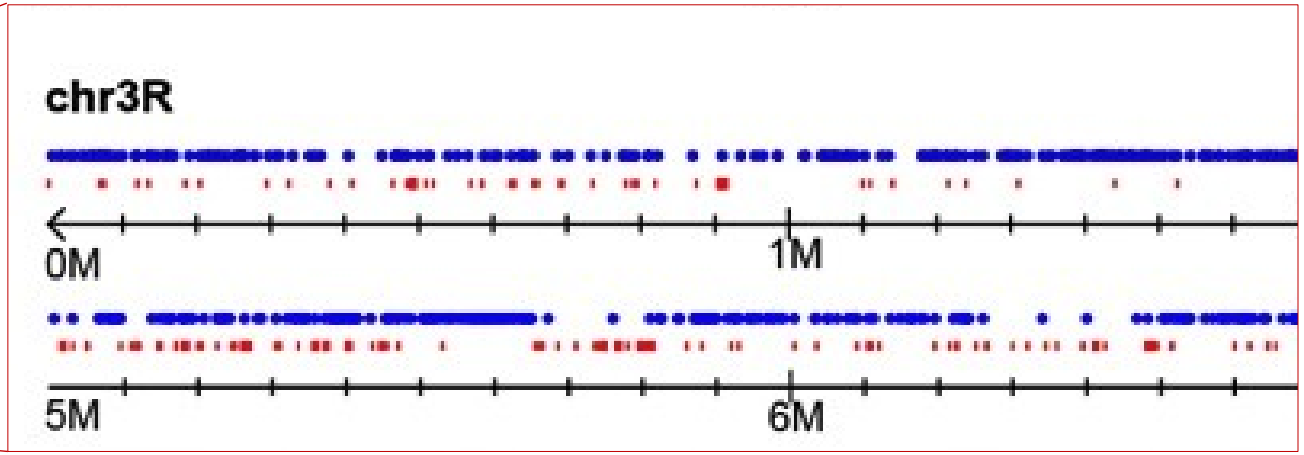
75% of the MARs are associated with genes
Higher density of MARs on X-chromosome
TSS/stalled Pol II on MAR – anchoring to transcription factories
Origin of replication overlap – anchoring to replication factories
MAR association with repeats: TEs / SSR / New motifs



MAR map of *Drosophila* genome

[Gene/MAR]





Function in the non-coding part of genome

Providing grammar to the genomic language

How to look for more functional elements?

Context dependent search

RNA version of repetitive DNA

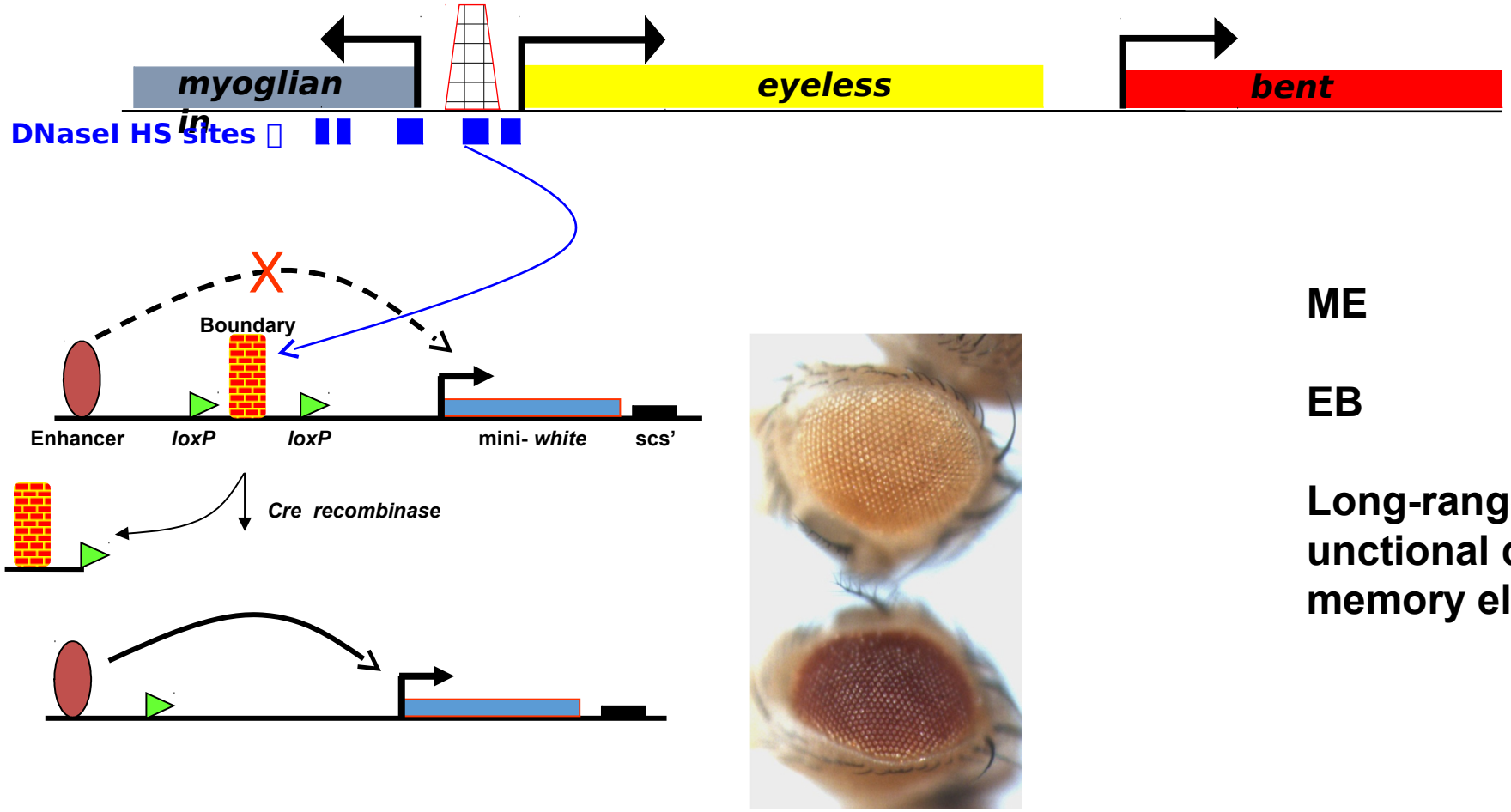
Epigenetic marking / patterns

Accessibility & protein bound regions

Long range interaction regions

Context dependent logic to search for boundary elements

Hina Sultana



ME

EB

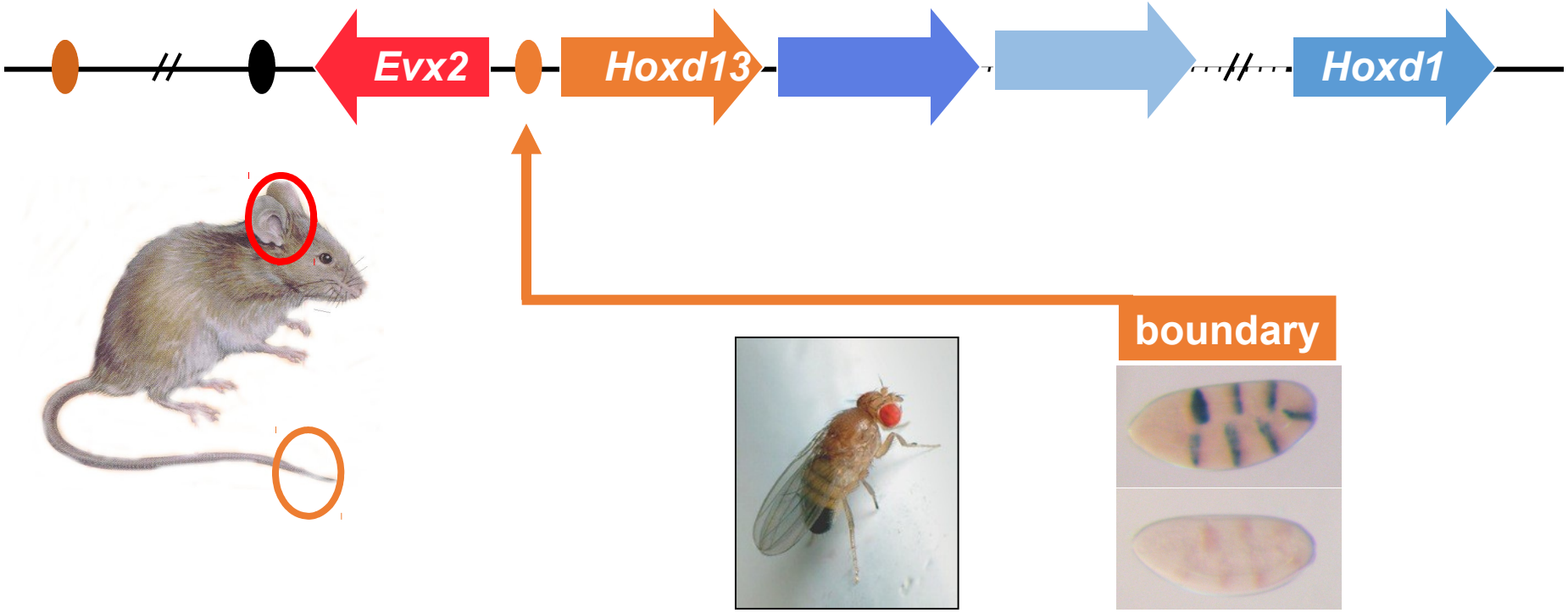
Long-range interaction, unctional domain and memory elements

Sulatana et al., Nucleic Acids Res. 2011

Context dependent logic to search for boundary elements

conserved across the species

D Vasanthi



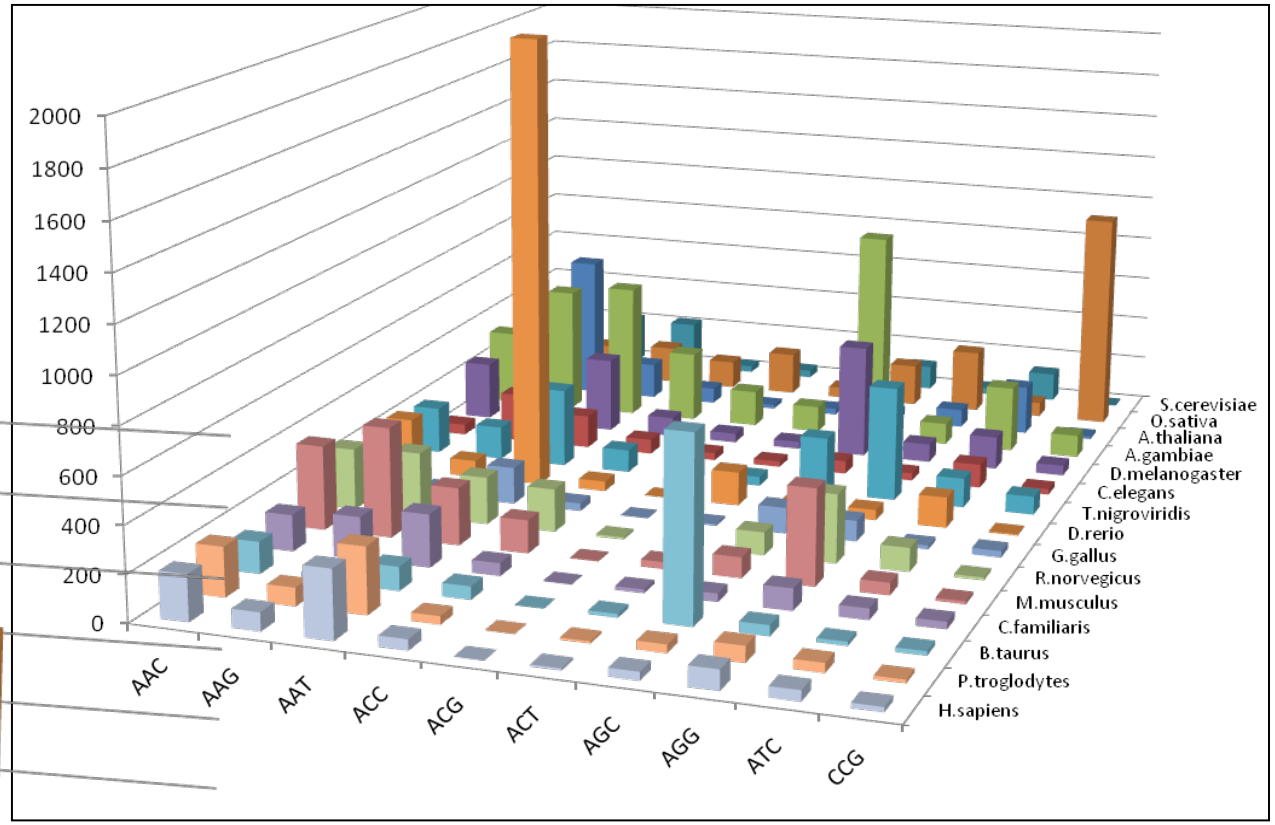
Vasanthi et al. Development 2010

Function in non-coding part of genome

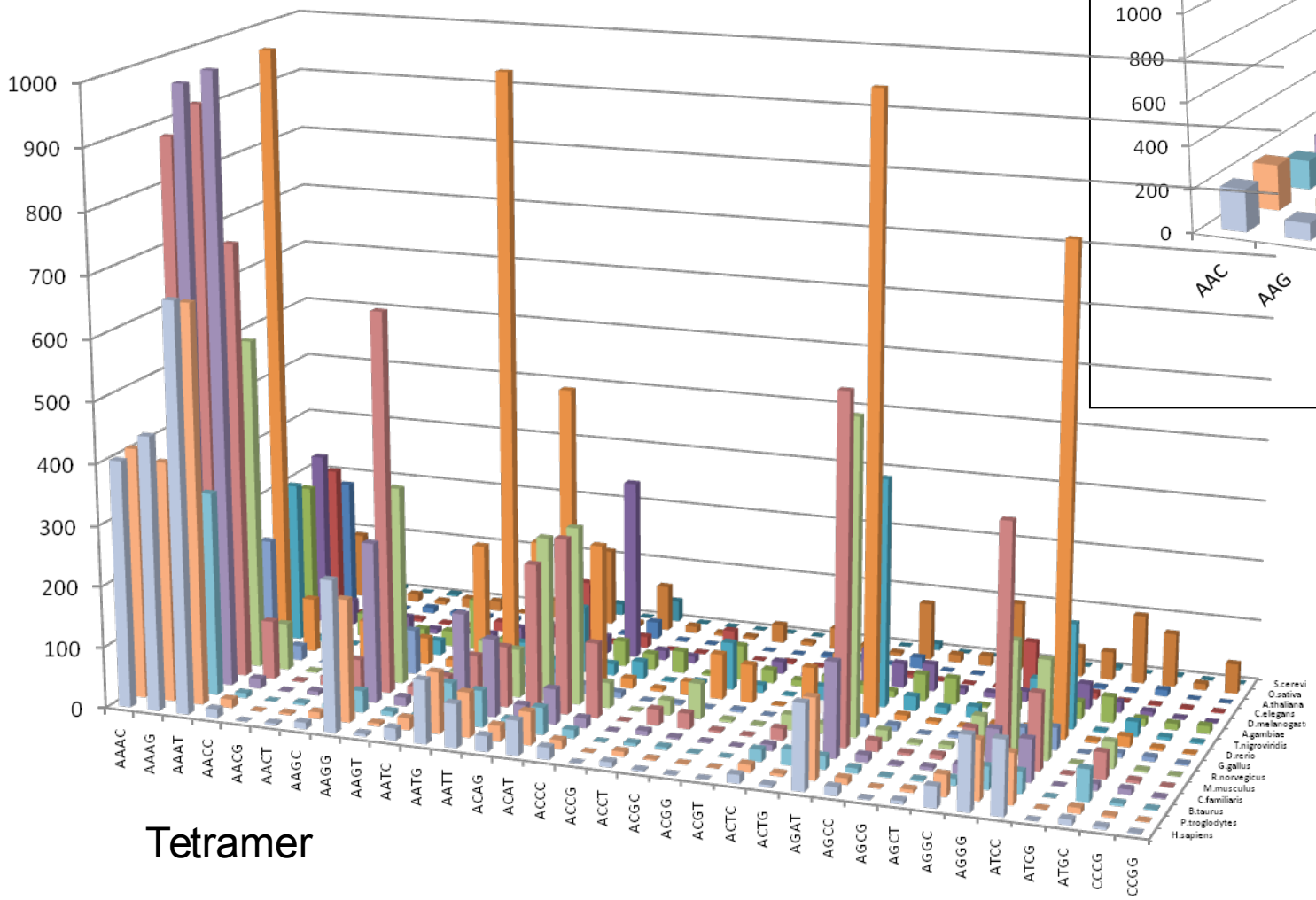
Simple Sequence Repeats [SSR] have functional significance

~3% of the human genome

SSRs are selectively enriched in different eukaryotes

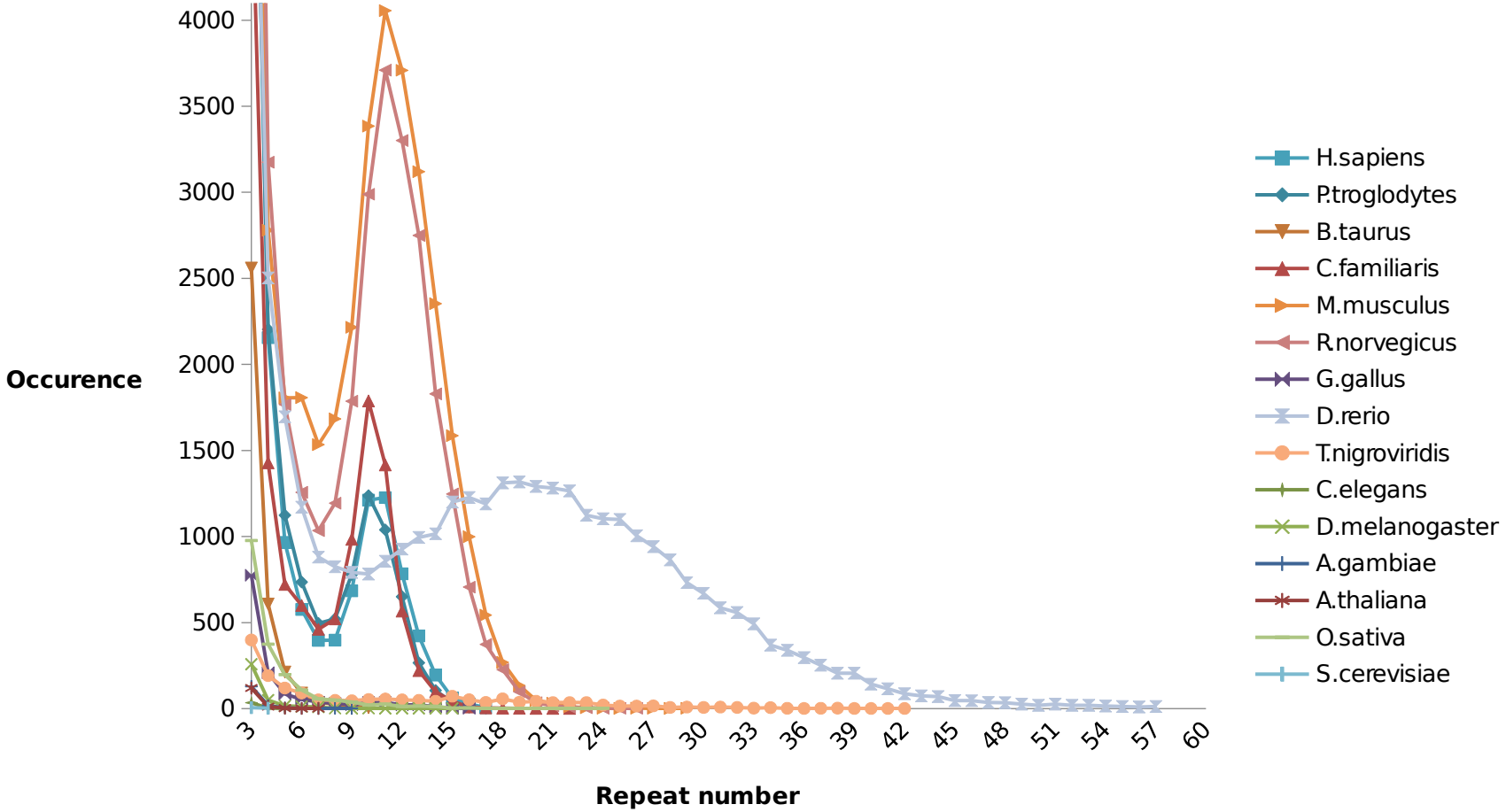


Trimer



Tetramer

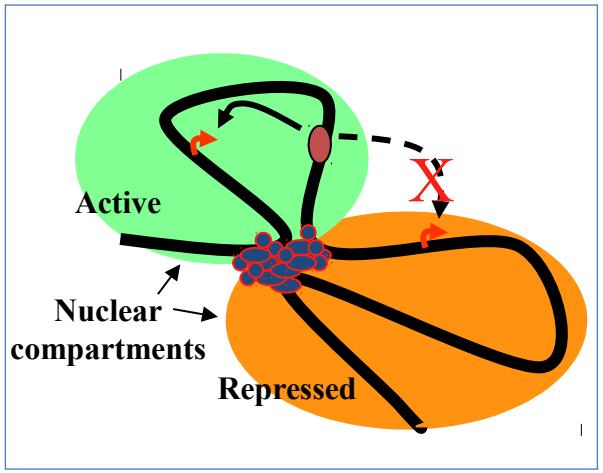
Some more unusual features of SSRs

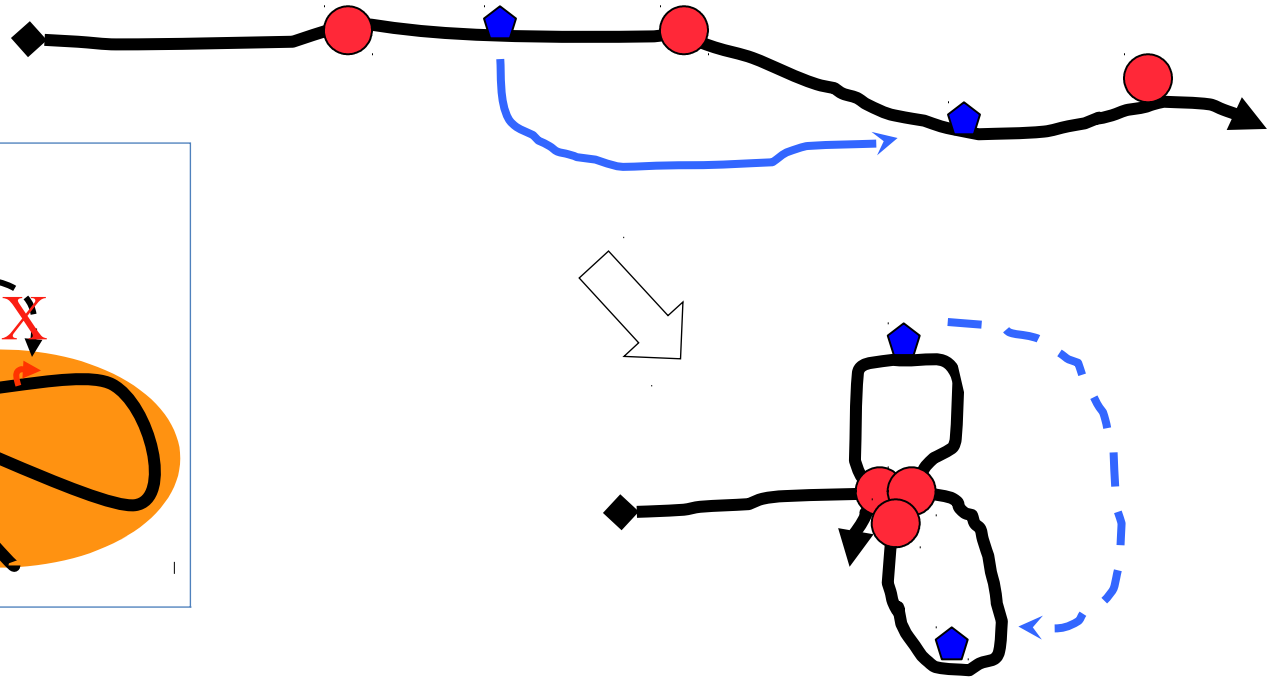


Preferred size and sequence
Selection pressure to maintain it!

Subramanian et al. Genome Biology 2003
Subramanian et al. Bioinformatics 2003
Ramamoorthy et al. Gene 2014

What is the selection pressure?

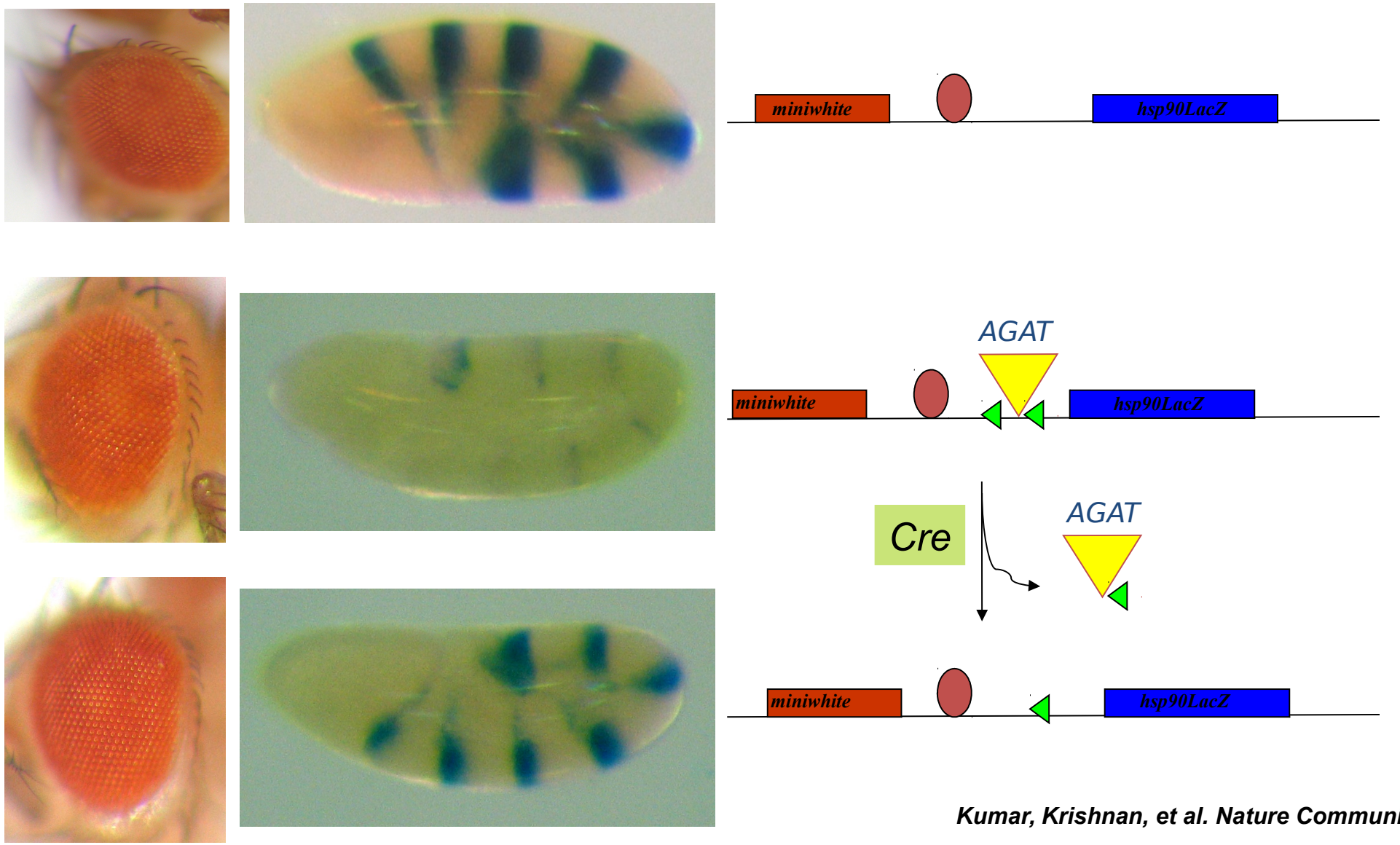




Packaging the genome with the help of boundary elements

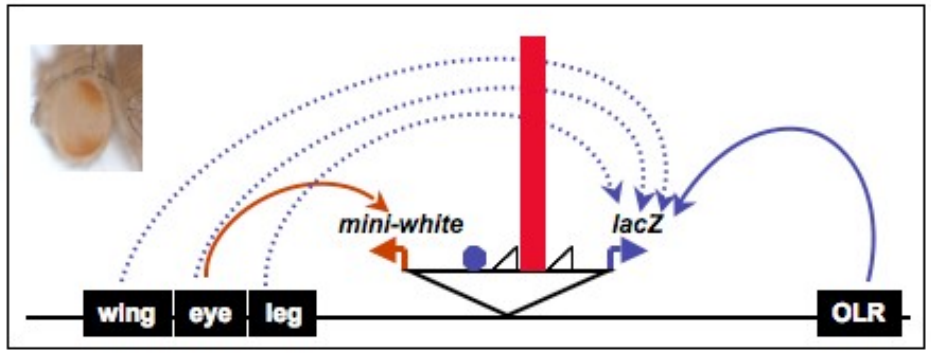
Kumar et al., BioEssays 2010

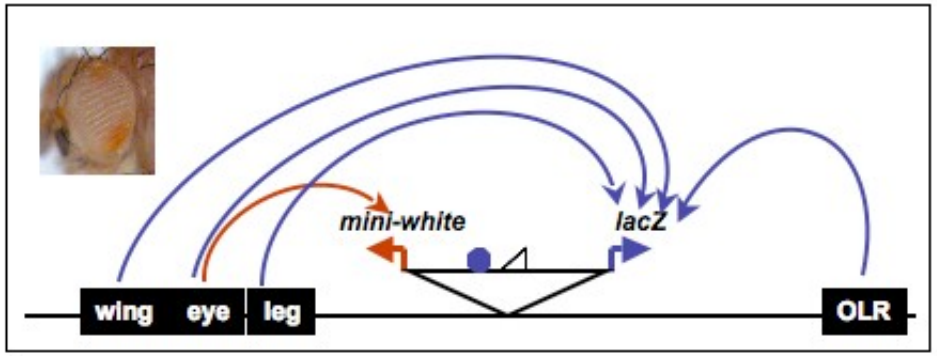
AGAT functions as enhancer-blocker boundary element

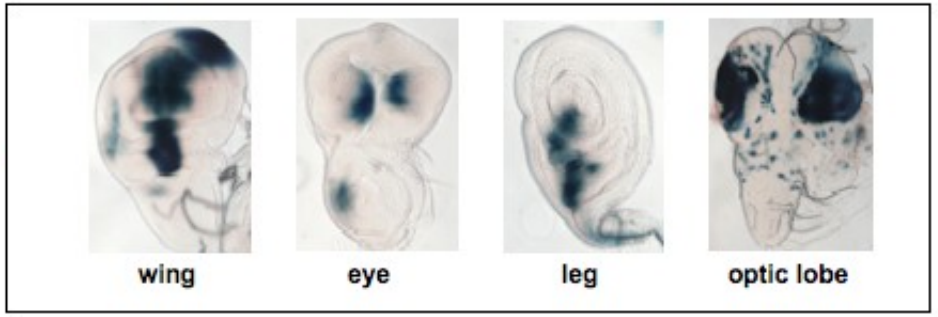


Kumar, Krishnan, et al. Nature Communications, 2013

Enhancer blocker in native context





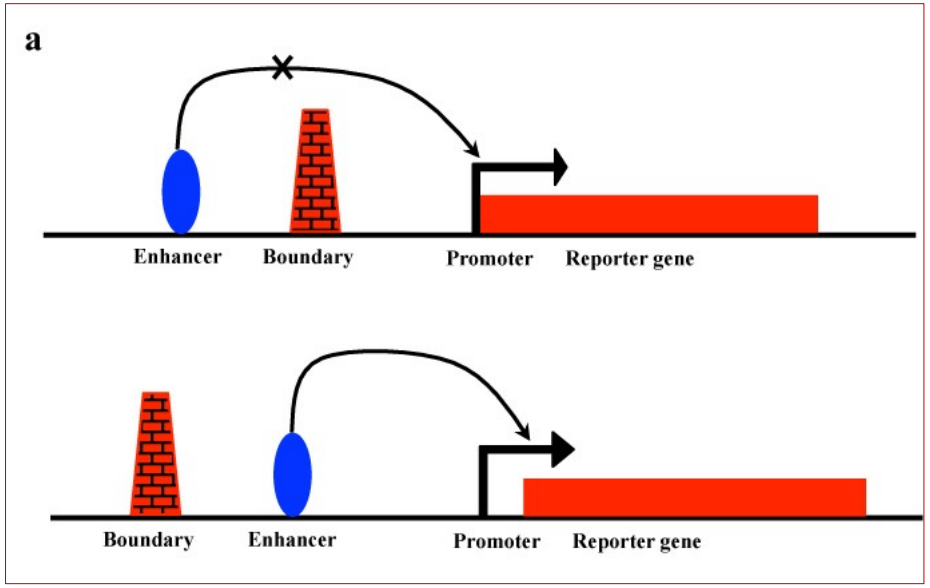


SSR repeats from human Y chromosome function as boundary element in fly

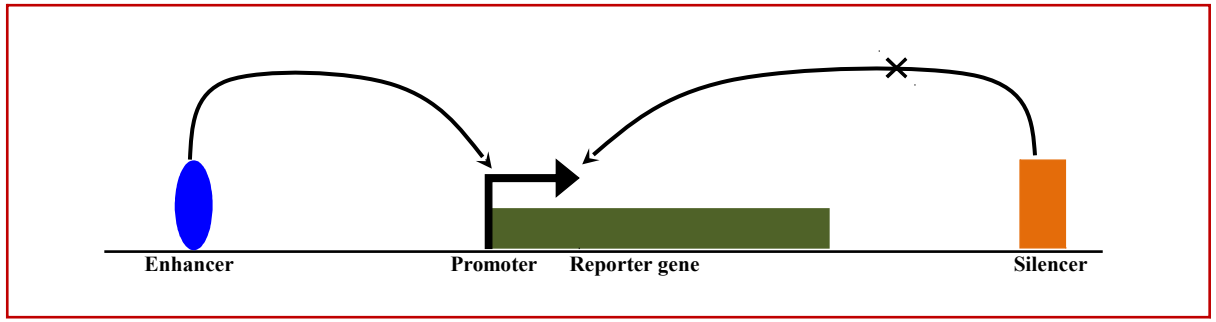
Kumar, Krishnan, et al. Nature Communications, 2013

SSRs for functions

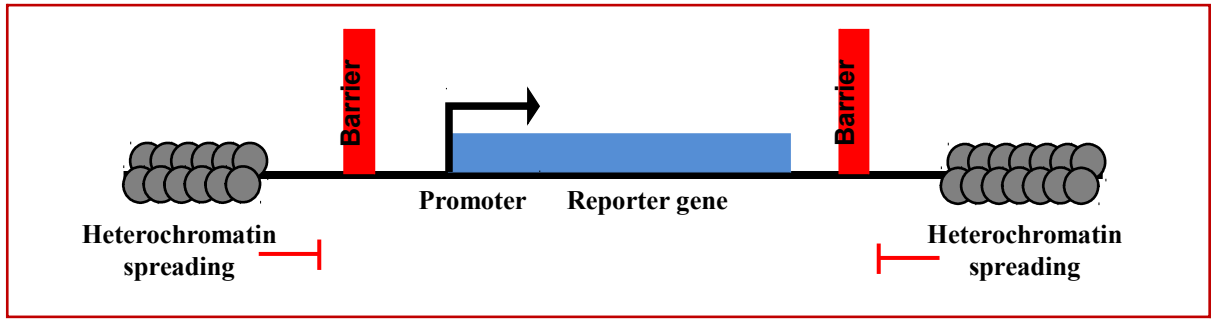
enhancer/repressor, boundary/barrier activity in cell based assays



Enhancer blocker



Enhancer/Repressor



Barrier assay

Survey of 23 SSRs for promoter modulation (enhancer/repressor, boundary/barrier) activity in cell based assays

Fathima Athar

Summary of assay in 5 different cell lines

		Promoter modulation assay					Boundary assay	Barrier assay
S.no.	SSR	IMR-32	MCF7	HeLa	HEK293T	K562	(K562)	(K562)
1	A	-	↑↑	↑↑↑	-	-	-	NA
2	AT	↓↓	↓↓	-	↓	↓↓	√	-
3	AAG	↑	-	↑	↑↑	-	-	NA
4	AAT	-	-	↑↑↑	↑↑	↑↑	-	NA
5	ATC	↓↓	-	-	↑	-	√	-
6	AGAT	↓	-	↓↓↓	-	-	√	-
7	AAAG	↓↓↓	↓↓	↑	-	-	-	NA
8	AAAT	↑↑	-	↑	↑↑	↑↑↑	√	-
9	AAGG	↑	-	↓	-	↑↑↑	-	NA
10	ACAT	-	-	-	-	↑	-	NA
11	ATCC	-	↓	↓↓↓	↑	-	-	NA
12	AAAAG	↑	-	↑	↑	↑↑↑	-	NA
13	AAAAT	-	-	↑↑↑	-	↑↑	√	-
14	AAAGG	↑	↑	↑↑↑	↑	↑↑↑	-	NA
15	AACAT	↑	-	↑↑↑	↑	-	√	w
16	AAGAG	-	-	↓	-	↑↑	√	-
17	AAGGG	↑	-	↓↓	-	↑	-	-
18	AATAC	↓	-	↑↑	↑↑	↑↑	√	-
19	AATAG	↓↓	-	↓↓↓	↓↓	↓↓	-	NA
20	AATAT	-	-	↓↓	-	↑↑	-	NA
21	AATGG	-	-	↑↑	-	↑	-	-
22	ACATAT	↓↓	-	↓↓	-	↑	-	NA
23	AGATAT	-	↓	↓↓↓	↓	-	-	NA
Resp. Positive controls *		-	↓↓	↑	↑	↑	√	

↑1.5-2, ↑↑2-2.5, ↑↑↑ >2.5, ↓ 0.8-0.6, ↓↓ 0.6-0.4, ↓↓↓<0.4-fold promoter activity

'-' no change in promoter/boundary/barrier activity when compared to respective vector controls.

√ - positive boundary activity;

'w' - weak barrier activity;

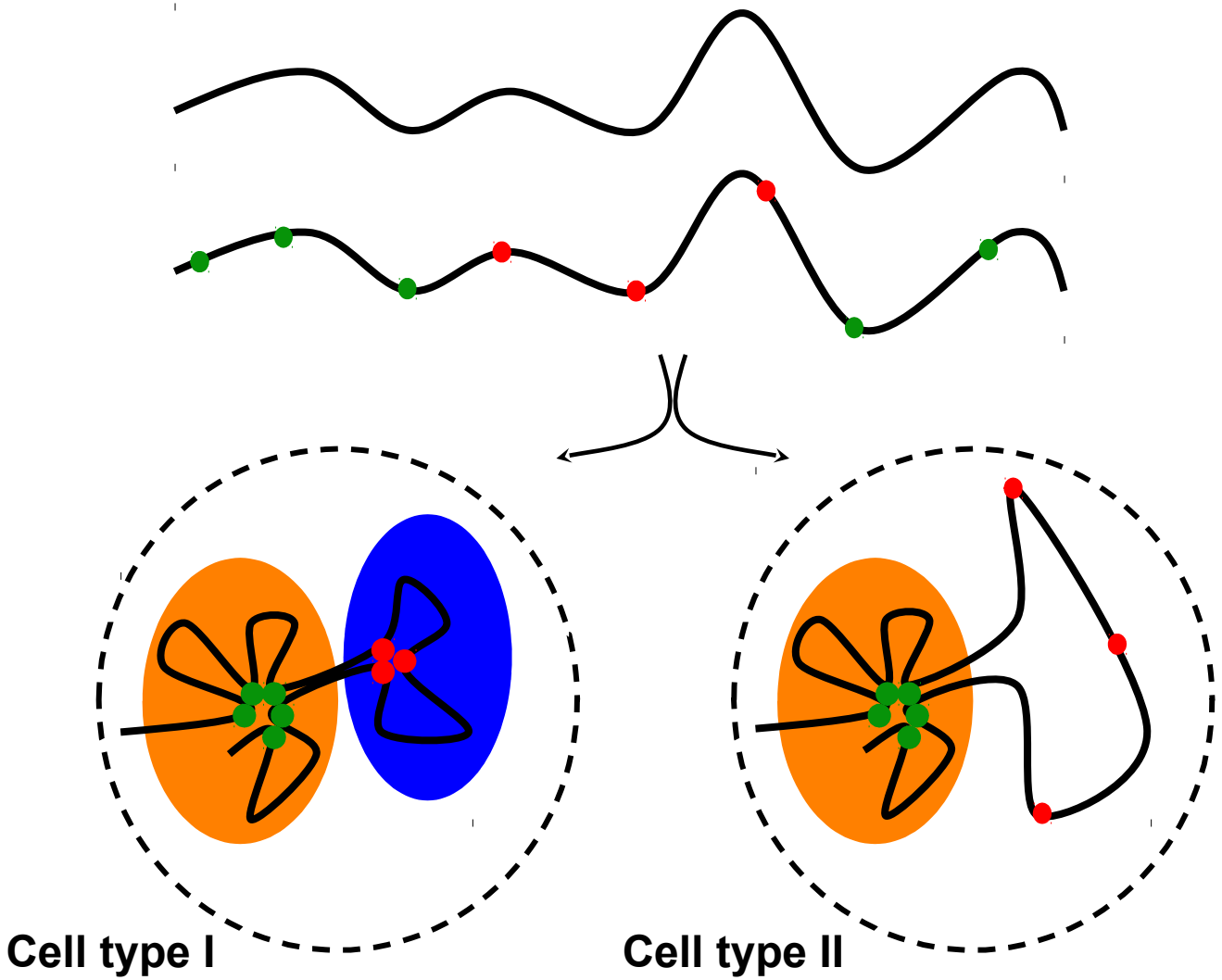
* mHoxPRE-FI for promoter modulation assay and β-globin boundary element for boundary assay were used as positive controls;

NA – not analysed.

Krishnan et al. Gene 2017

Repeat mediated interaction of distant loci

functional clustering / spatial anchoring



SSR: part of genome packaging code

- Multiple loci can be marked with fewer landmarks for coordinated regulation

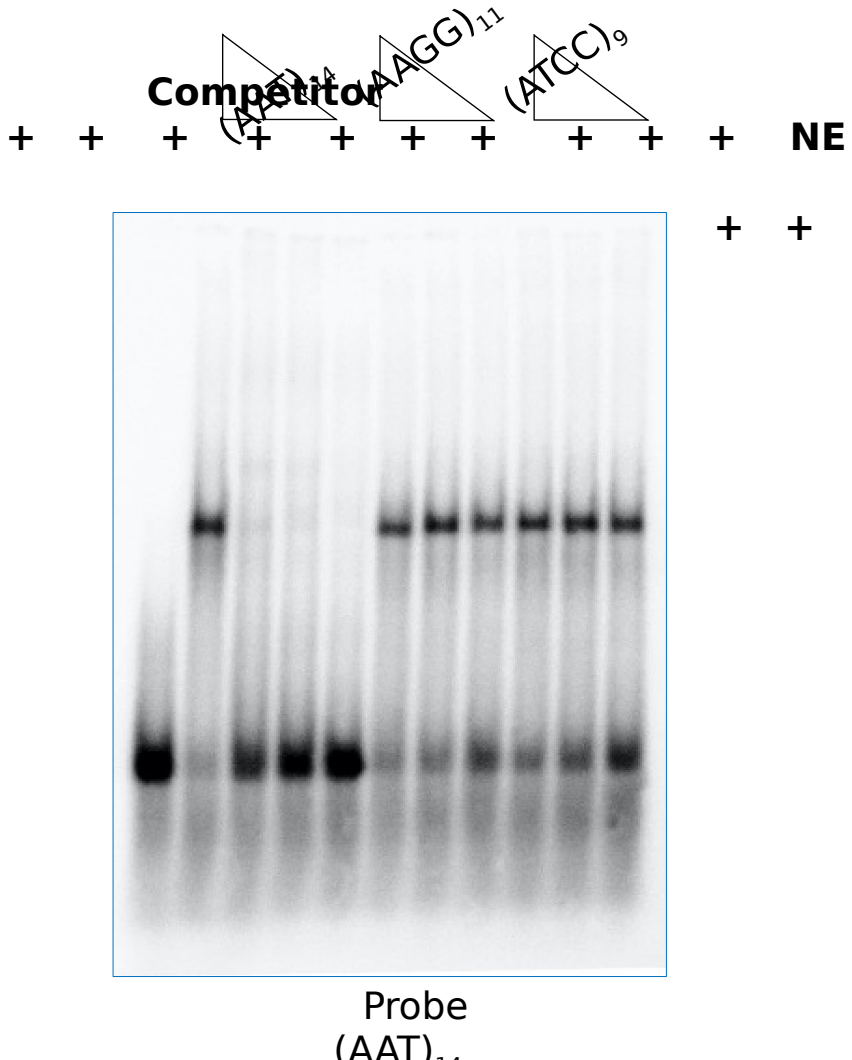
- Few factors can control large number of genes

- Targeting genomic loci marked by repeats via guide molecules:

~proteins
~transcripts

Kumar et al., BioEssays 2010

Sequence specific SSR binding proteins



S. No	Simple Sequence Repeat	Specific DNA binding activity
1	A_{36}	+
2	AT_{21}	+
3	AAG_{19}	+
4	AAT_{14}	+
5	$AAAG_{13}$	+
6	$AAAT_{10}$	+
7	$AGAT_{15}$	+
8	$AAAAG_{11}$	+
9	$AAGAG_{12}$	+
10	$AATAC_{12}$	+
11	$AAAAT_8$	+
12	$AATAG_{11}$	+
13	$AATAT_9$	+
14	$AACAT_{10}$	+
15	$ACATAT_8$	+
16	AATGG	+
17	$AGATAT_7$	+
18	$AAAGG_{12}$	+
19	ATC_{12}	To be confirmed
20	$ACAT_{10}$	To be confirmed
21	$ATCC_9$	To be confirmed
22	$AAGG_{11}$	To be confirmed
23	$AAGGG_{11}$	To be confirmed

K. Phanindhar

Summary of EMSAs done with Nuclear extract from *Drosophila* embryos (0-16h)

SSRs are: Selectively enriched and conserved repeat elements

Distributed non-randomly

Have functional significance

Most SSRs are transcribed [cell type dependent, Nuclear]

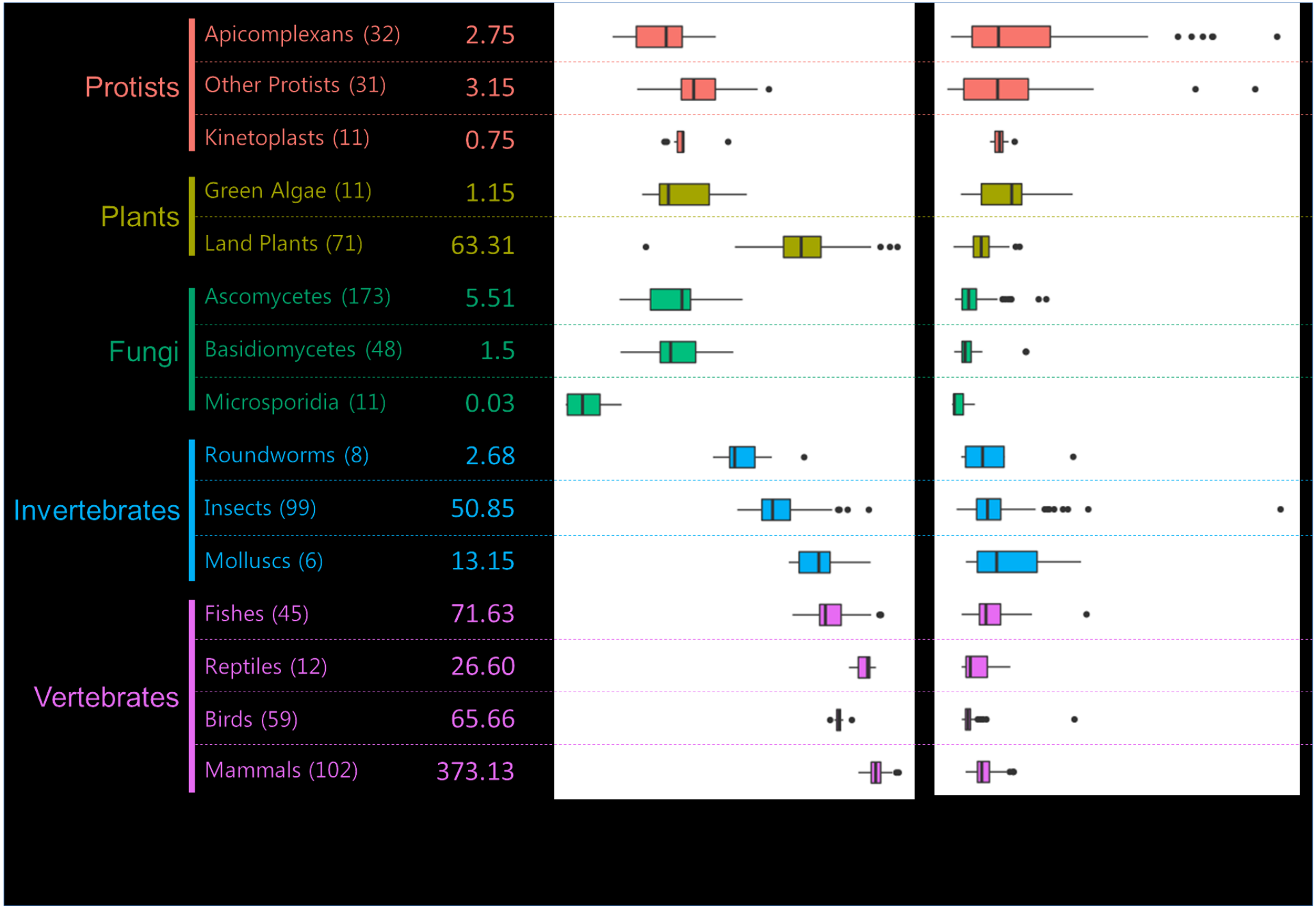
Their evolution, however, remains poorly understood.

Here is the most comprehensive analysis of SSRs we have carried out:

>680 million microsatellites

719 eukaryotes

to explore the evolutionary trends from protists to mammals



Overview of SSRs

~685 million SSRs

719 eukaryotes

Major findings and propositions

1. SSR density: taxon-specific variations in exonic, intronic and intergenic densities

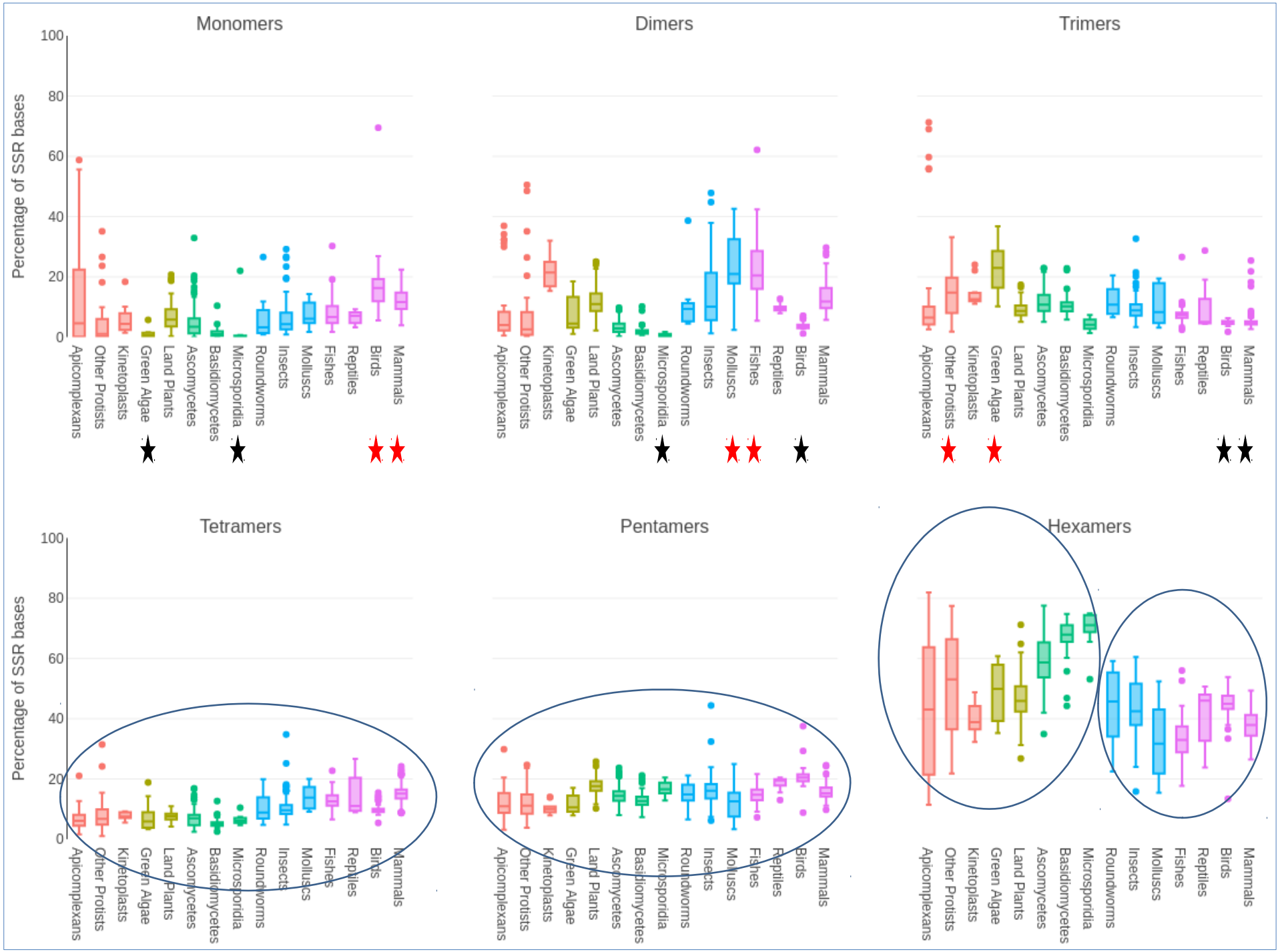
2. Composition: i) highly constrained in organisms with heterogeneous cell types
ii) greater diversity in motif abundance, density and GC content in simpler organisms such as protists, green algae and fungi

3. SSR lengths: increased in complex organisms
(indicative of an evolutionary selection pressure)

We propose:

SSRs are integral components in speciation and the evolution of organismal complexity, as they bring coordinated genome regulatory features

further supported by the fact that there are species-specific repeat signatures that mirror phylogenetic relationships, which brings up utility of SSRs in phylogenomic studies.

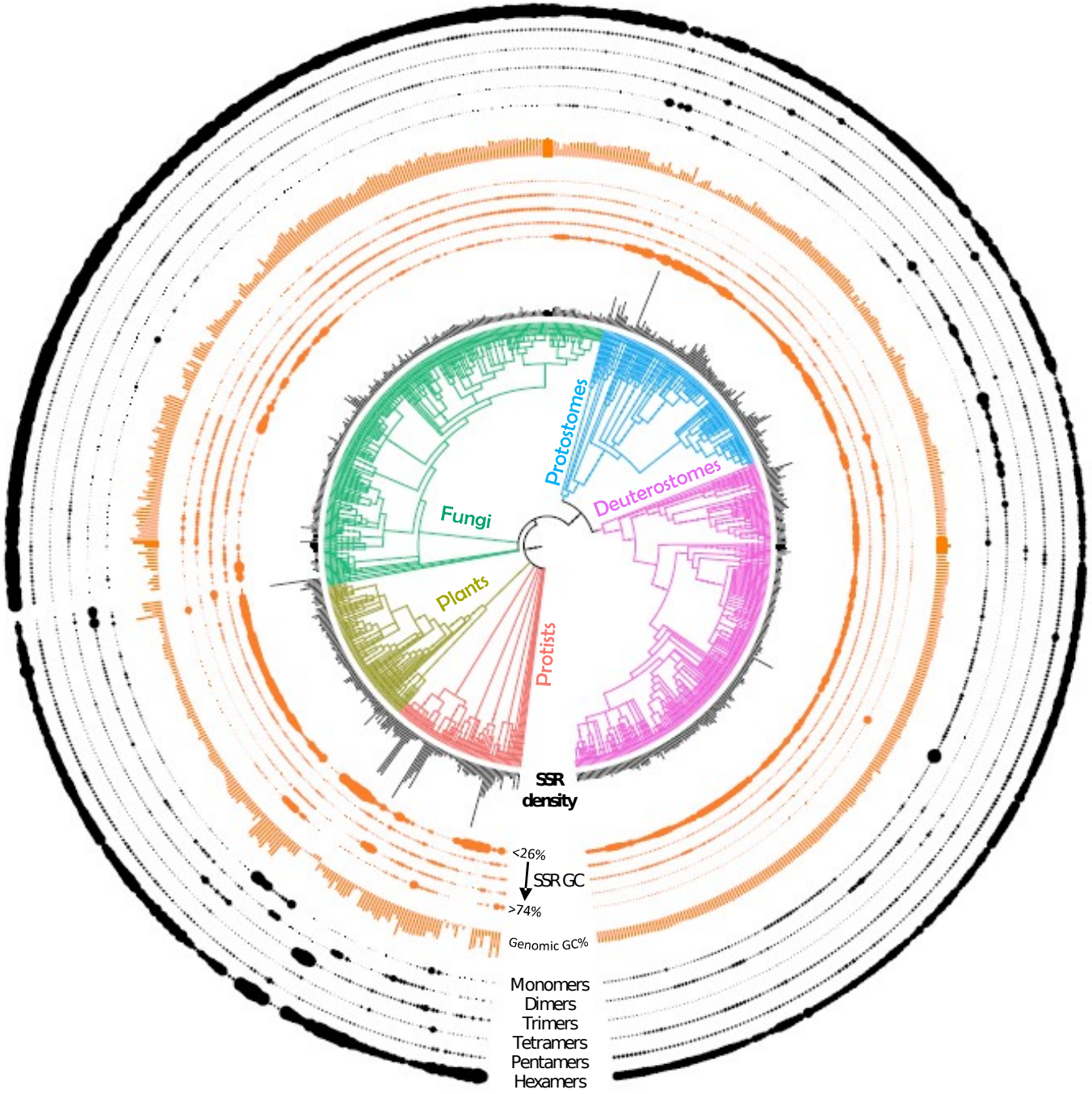


Composition of SSRs by their motif sizes

Box plots

Y-axis:
% of k-mer base coverage

X-axis:
subgroup



Attributes of all SSRs analyzed

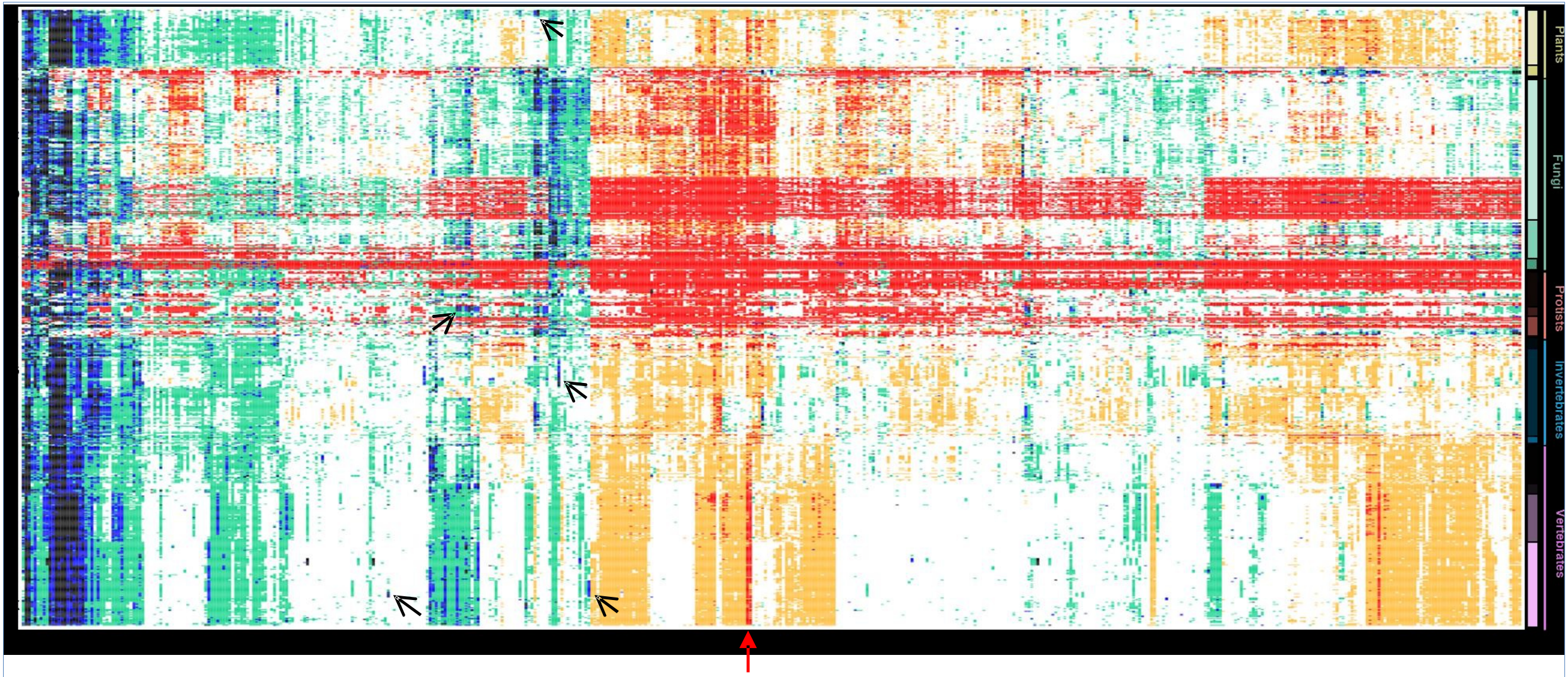
The tree was constructed using iTOL (interactive Tree Of Life) webserver.

Black bars (the innermost track) around the organisms represent the SSR density.

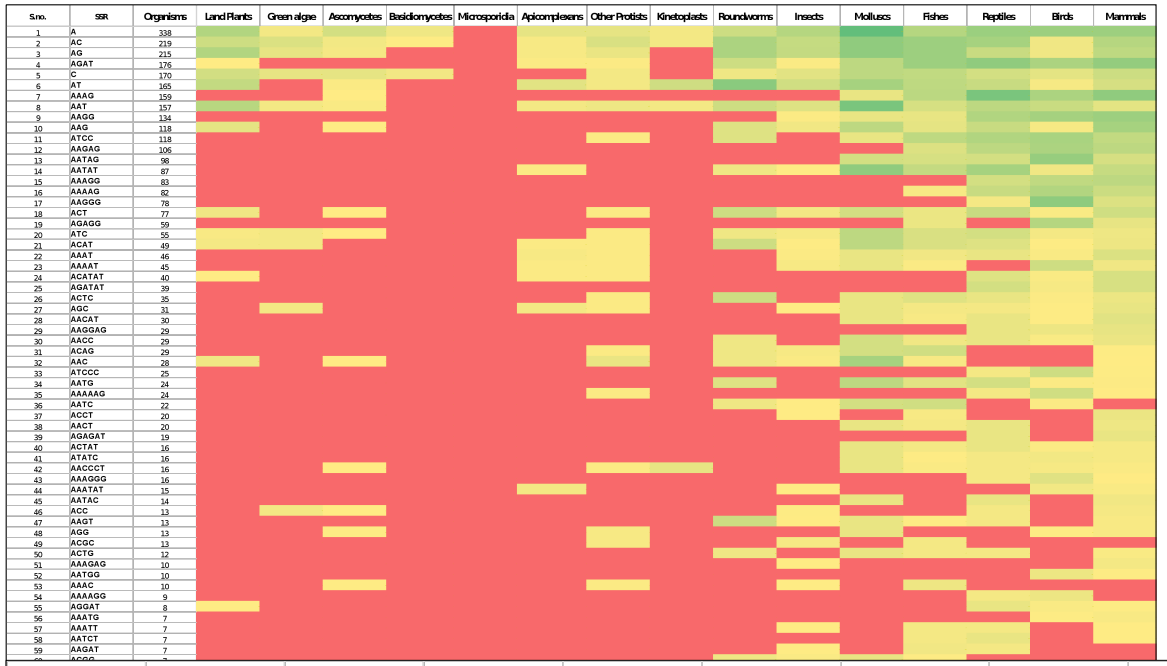
The orange tracks show the SSR GC %.

The orange bars represent the genomic GC content.

The black tracks: distribution based on the motif size.



Enrichment trend of the 501 motifs (X axis) in 685 million SSRs from 719 genomes (Y axis)



S.no.	SSR	Organisms	Land Plants	Green algae	Ascomycetes	Basidiomycetes	Microsporidia	Apicomplexans	Other Protists	Kinetoplasts	Roundworms	Insects	Molluscs	Fishes	Reptiles	Birds	Mammals
1	A	338															
2	AC	219															
3	AG	215															
4	AGAT	176															
5	C	170															
6	AT	165															
7	AAAG	159															
8	AAT	157															
9	AAGG	134															
10	AAG	118															
11	ATCC	118															
12	AAGAG	106															
13	AATAG	98															
14	AATAT	87															
15	AAAGG	83															
16	AAAAG	82															
17	AAGGG	78															
18	ACT	77															
19	AGAGG	59															
20	ATC	55															
21	ACAT	49															
22	AAAT	46															
23	AAAAT	45															
24	ACATAT	40															
25	AGATAT	39															
26	ACTC	35															
27	AGC	31															
28	AACAT	30															
29	AAGGAG	29															
30	AACC	29															
31	ACAG	29															
32	AAC	28															
33	ATCGC	25															
34	AATG	24															
35	AAAAAG	24															
36	AATC	22															
37	ACCT	20															
38	AAGT	20															
39	AGAGAT	19															
40	ACTAT	16															
41	ATATC	16															
42	AACCCT	16															
43	AAAGGG	16															
44	AAATAT	15															
45	AATAC	14															
46	ACG	13															
47	AAGT	13															
48	AGG	13															
49	ACGC	13															
50	ACTG	12															
51	AAAGAG	10															
52	AATGG	10															
53	AAAC	10															
54	AAAAGG	9															
55	AGGAT	8															
56	AAATG	7															
57	AAATT	7															
58	AATCT	7															
59	AAGAT	7															

SSR length preference

The 131 SSRs that show a length preference in any organism are listed in the 2nd column.

The number of organisms (from all subgroups) showing length preference for a SSR is in the 3rd column.

The heatmap shows % of organisms in a subgroup (maximum 83%) that show length preference for a SSR.

Uniquely abundant SSRs showing species-specific enrichment

#	Species	Uniquely abundant SSRs	Divergence from LCA
1	*Leishmania* sp.	AGGG, AGGGG, AGGGGG, ACACGC	1660 Ma
2	Green algae	CG, ACGCG, CCCCG ACGGCG, ACGCCG, AGCGCG, ACGCGG, ACGTCG	1160 Ma
3	Cereals	CCGGCG, CCCGCG, ACGGCC	104 Ma
4	*Drosophila* sp.	AACAGC	127 Ma
5	Birds	AAACC, AAAGG, AAAACC, AAAAGG	111 Ma
6	Ruminants	AACTG, AAAGTG, AAGCTG	56 Ma
7	Primates	AATGG, ACCTCC	67 Ma

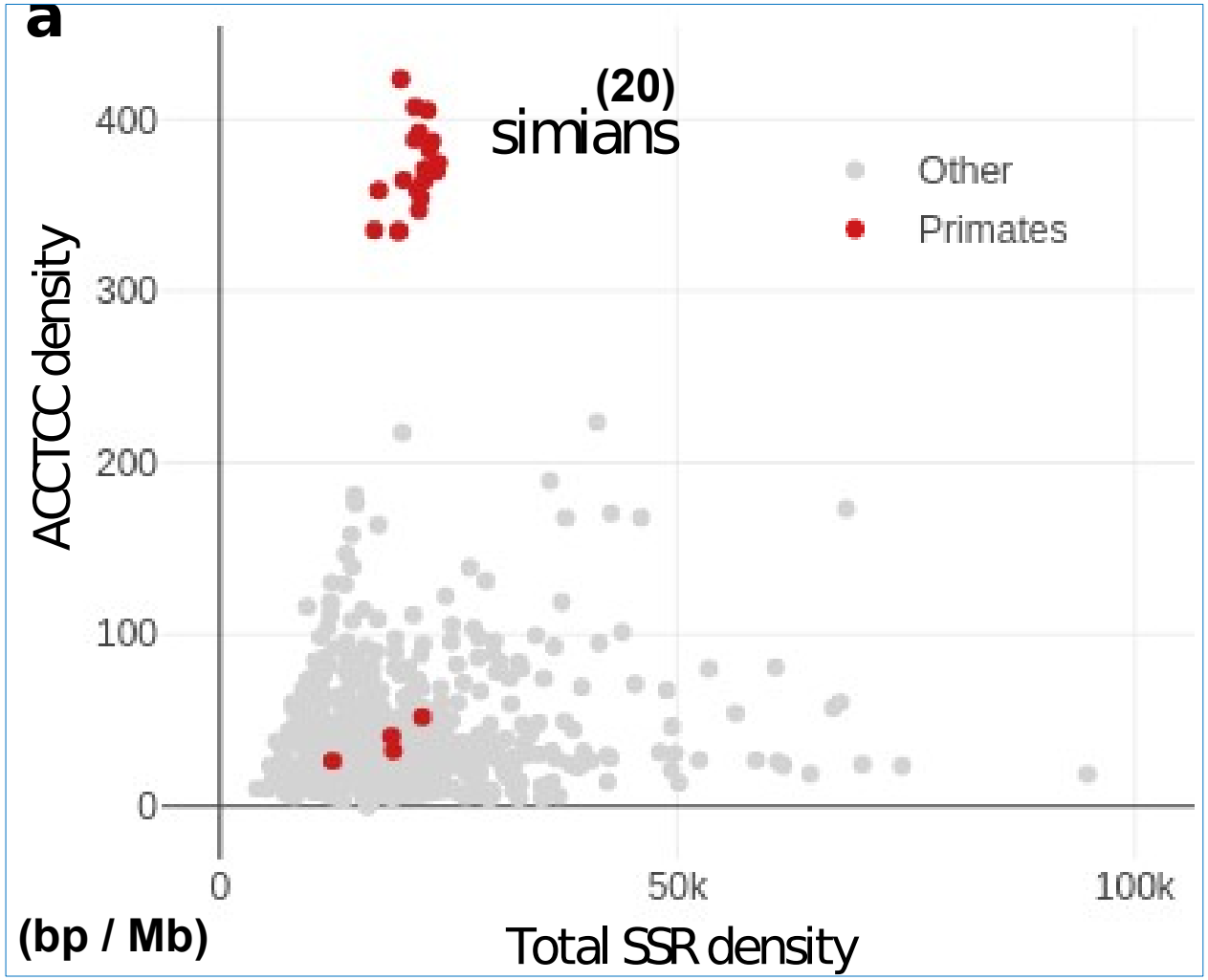
uniquely abundant in that clade/taxon

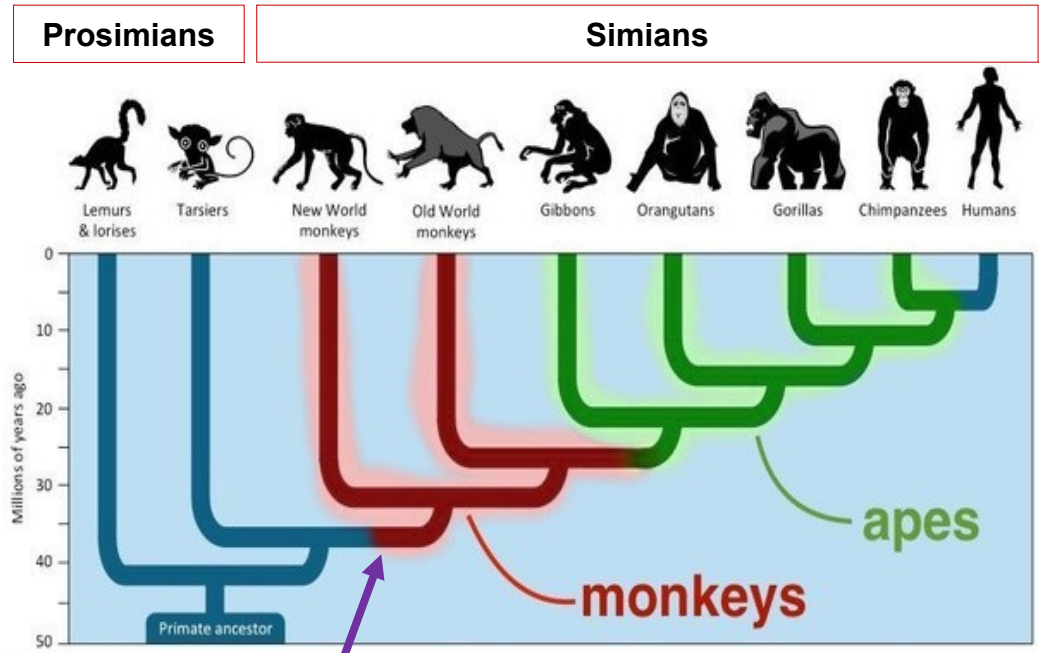
mostly unique for the clade, with a couple of other species showing enrichment

uniquely abundant in the clade, but are also enriched in several unrelated species

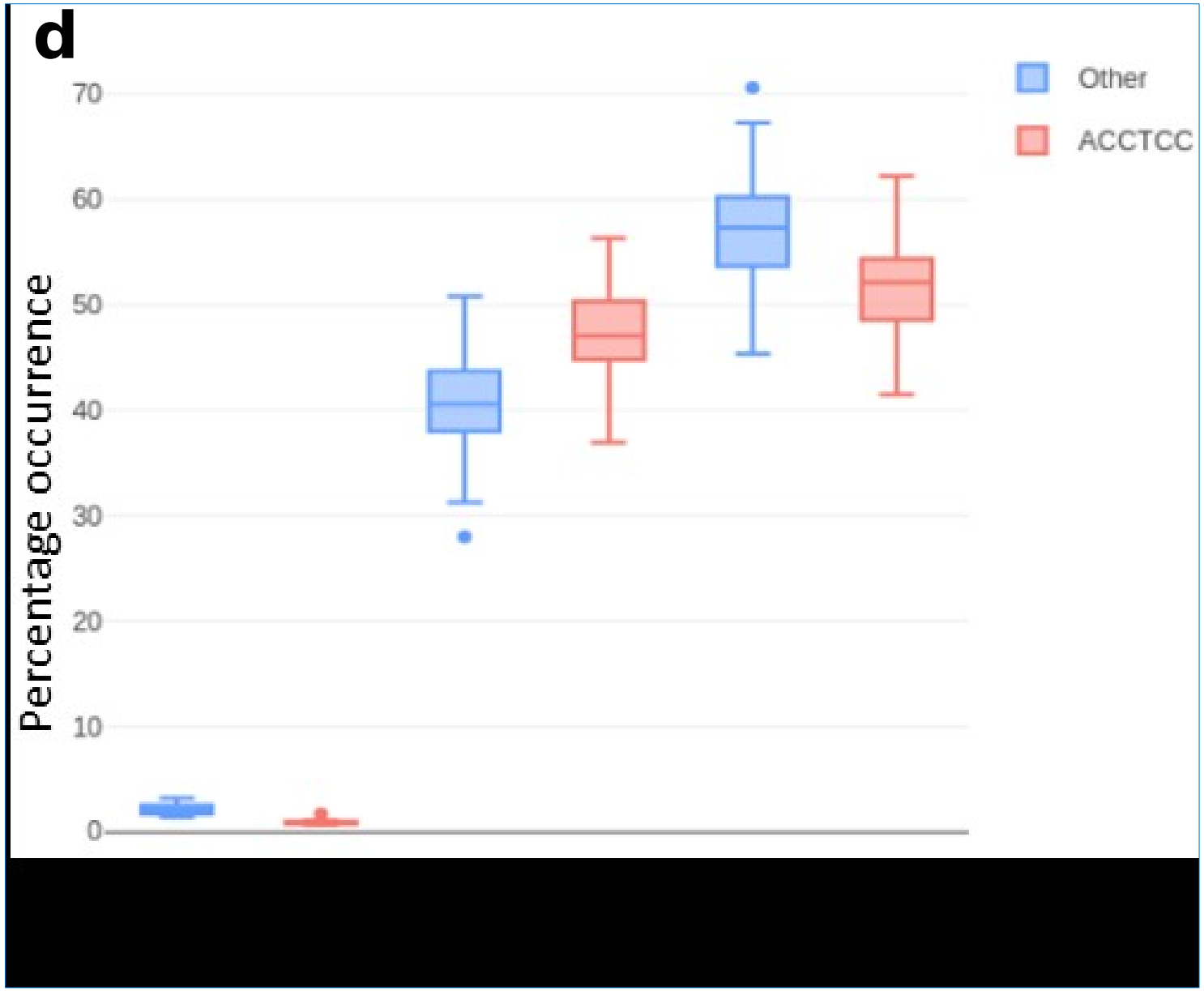
ACCTCC

Density of ACCTCC compared to the total SSR density for all organisms





ACCTCC emergence: 35 million years ago (coinciding with that of simians) and persisting since then



ACCTCC is significantly under-represented in exons and over-represented in intron compared to other repeat classes in simians

Indicative of:

- the mechanism of expansion

- gene regulatory role

- splicing regulation role

Conclusions:

Evolution of complexity is largely due to emergence of novel and complex regulatory mechanisms

Much of non-coding DNA is reflection of this process

SSRs have been selectively enriched by active process and retained due to positive selection pressure

Among the possible roles: genomic packaging, boundary function, coordinated regulation, activator, repressor, …

SSRs are likely to function with help of sequence specific DNA binding proteins and the corresponding strand specific ncRNA

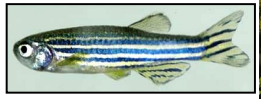
There are species-specific SSR signatures that mirror phylogenetic relationships, indicating specific roles of such elements

CSIR
Council of Scientific and Industrial Research

Thank Y

Department of Science & Technology
Ministry of Science & Technology
विज्ञान और प्रौद्योगिकी विभाग

Department of Biotechnology
Ministry of Science and Technology



our_lab@CCMB

evo-devo of Hox	A Srinivasan Nikhil Hajirnis
chromatin & epigenome	Divya Tej Sowpati Shreekant Verma Shagufta Khan K Phanindhar Fathima Athar Ravina Saini Runa Hamid Avvaru Akshay M S Soujanya Sonu Yadav
nuclear Architecture	Rashmi U Pathak Rahul Sureka Ashish Bihani