Deciphering the function and evolution of the centromeric repeats in Primates

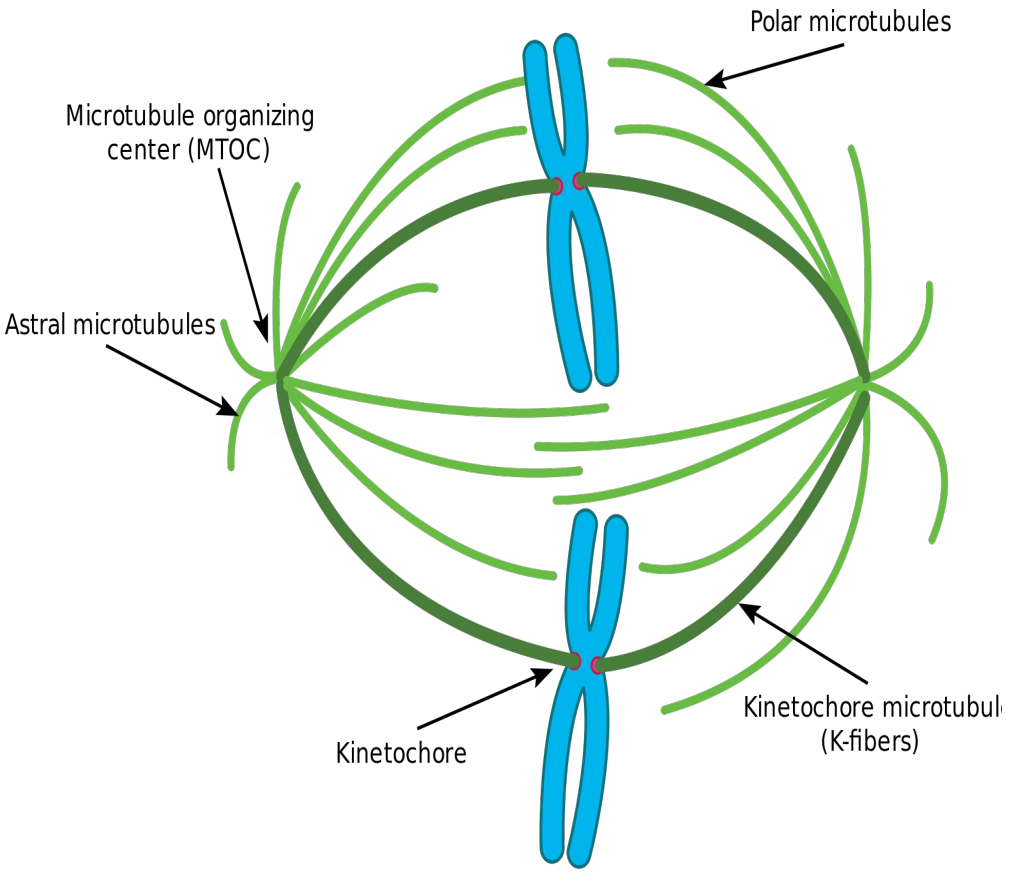
Sarah Kaddah
M2 BI

19/01/2018

Tuteur: Loic Ponger
Département: Régulation, Développement, Diversité Moléculaire
Unité: Structure et Instabilité des génomes
CNRS UMR 7196 / INSERM U1154 / MNHN

About the centromere

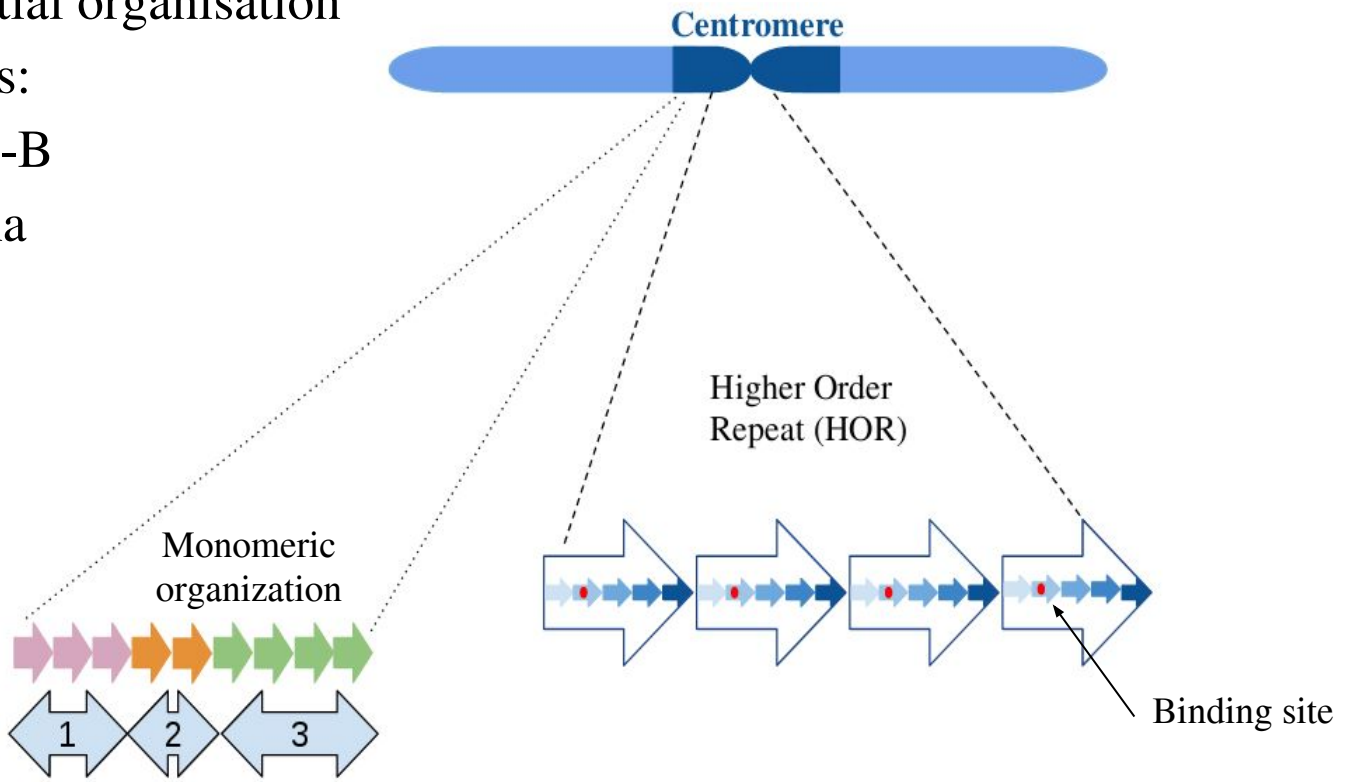
- Chromatin structure
- Cell division
- DNA diverse
- Protein involved conserved
- **Satellite DNA:**
  Tandemly repeated sequence
  (~100 to +10 000)
- Function ?
  (fixation of proteins …)



Albert et al, 2002

Satellite DNA in Primates

- Tandemly repeated sequence: α-satellite
- Repeat length ~170 pb
- Similarity: more than 70%
- Hundreds of thousands repeats
- Composed by several families
- Specific spatial organisation
- Binding sites:
  - -CENP-B
  - -pJalpha

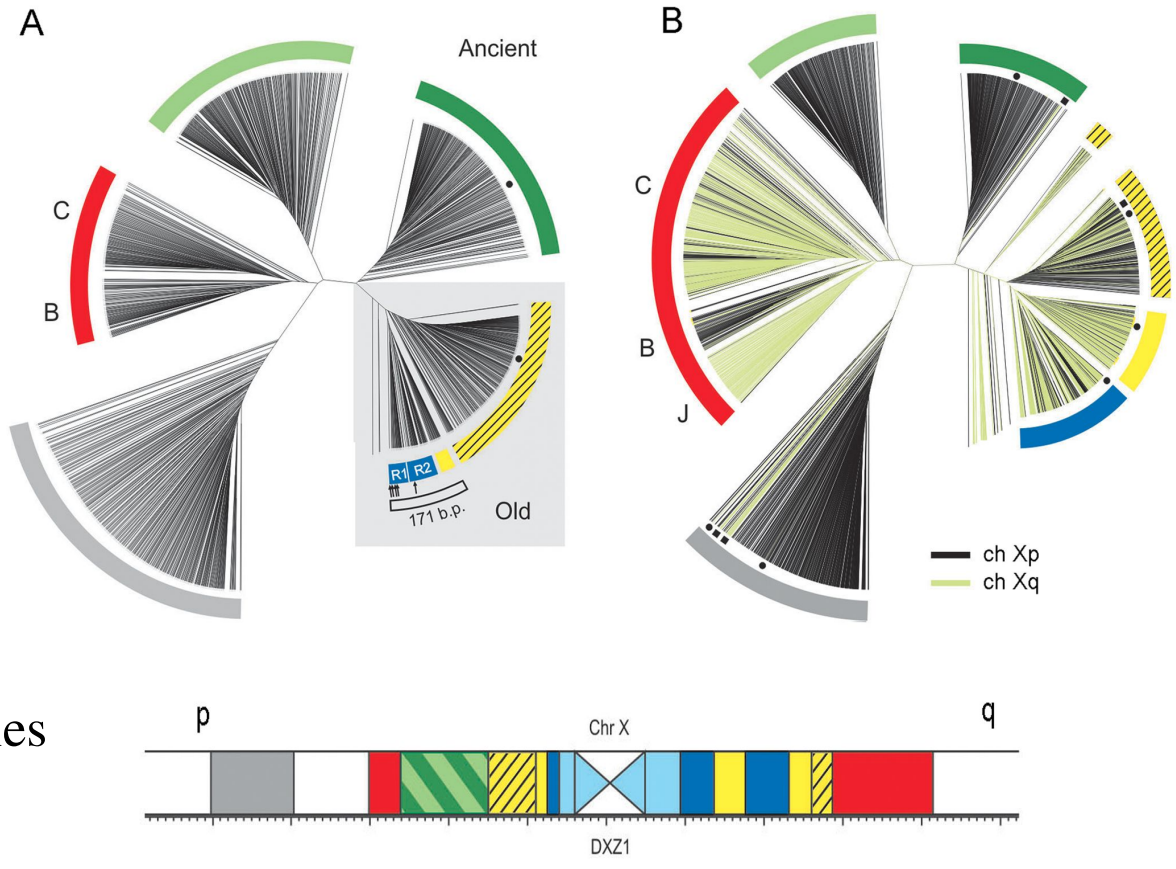


α-satellites DNA

Studies in human:

Phylogenetic analysis of pericentromeric monomers,
Shepelev et al., 2009

- Age-gradient hypothesis: youngest families on centromeric regions
- Identification of 20 families



Shepelev et al., 2009

α-satellites DNA

Studies on Gorillas, Catacchio et al., 2015

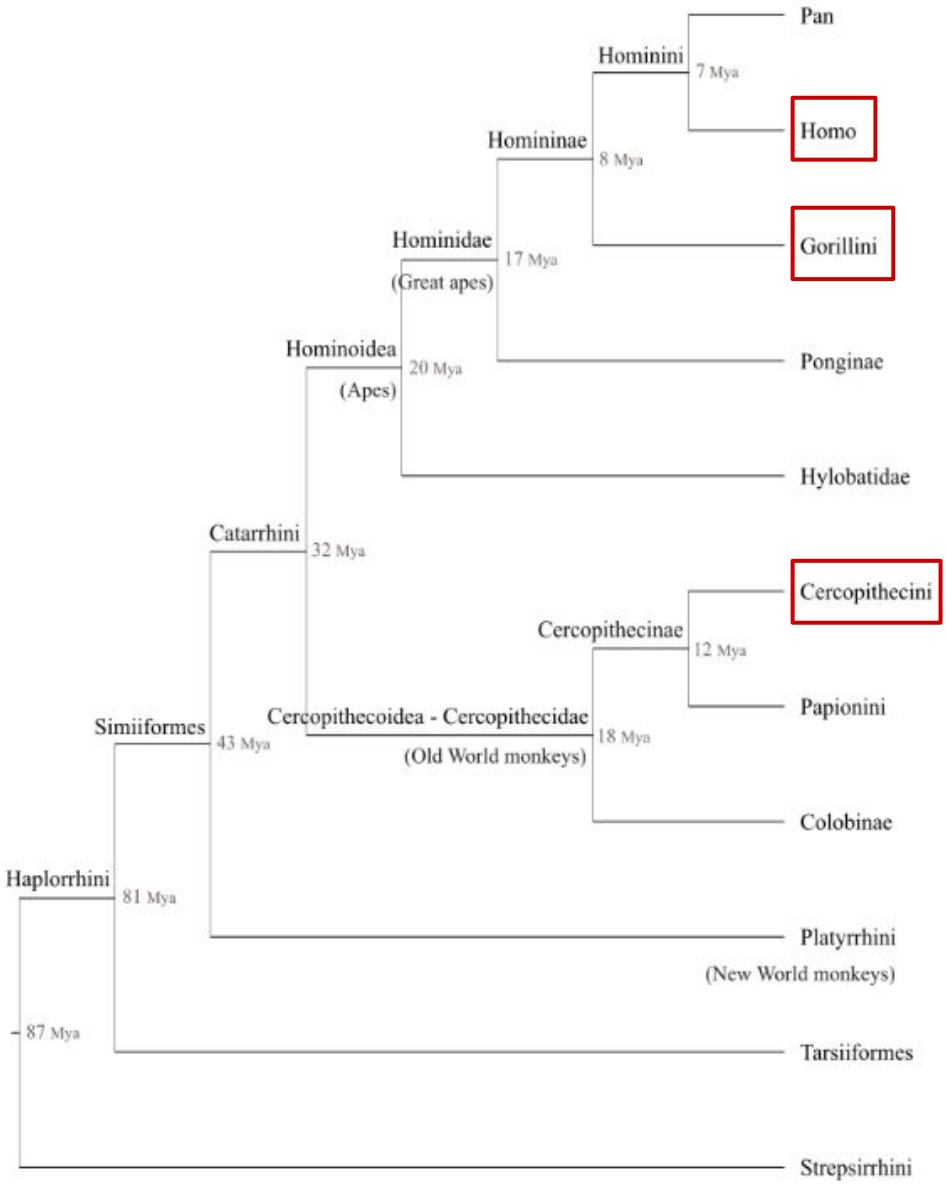
- Medium throughput-sequencing
- Identification of 5 families
- Complex HOR organization
- Binding sites for CENP-B

Studies on Cercopithecini,

Cacheux et al., 2016

- High throughput-sequencing
- Identification of 6 families
- Binding sites for pJalpha

Few interspecific comparison



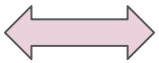
Goals:

Understand the function and evolutionary mechanisms of α-satellite DNA

Workplan

1 - Choose species for analysis

1



Workplan

1 - Choose species for analysis: several (~3) out of 16

Species	Alpha- Satellite Number
Cercocebus atys (SRA)	80 884
Chlorocebus sabaeus	29 842
Gorilla gorilla	120 864
Homo sapiens (HGSC, HuRef)	37 204 ; 63 167
Macaca fascicularis (genome and SRA)	195642 ; 39893
Macaca mulatta (SRA)	7 365
Macaca nemestrina (SRA)	462 063
...	...

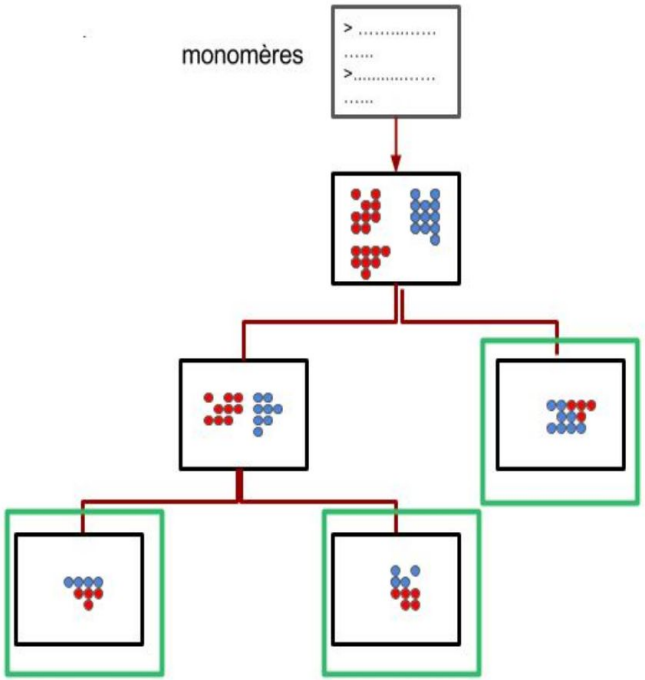
1

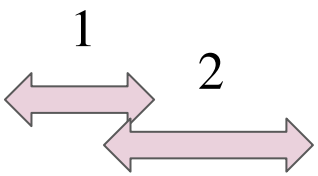
Workplan

1 - Choose species for analysis

2 - Identify families for each species : objective and reproducible method

- Process big amount of short sequences
- Classification into families
- Binary classification



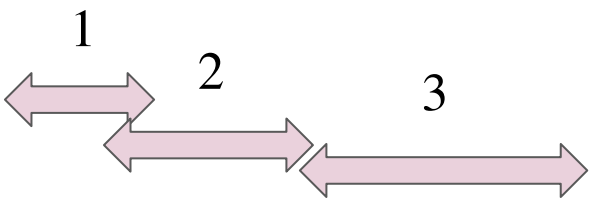


Workplan

1 - Choose species for analysis

2 - Identify families for each species

3 - Characterize families into each species:
-percentage of similarity
-binding sites...



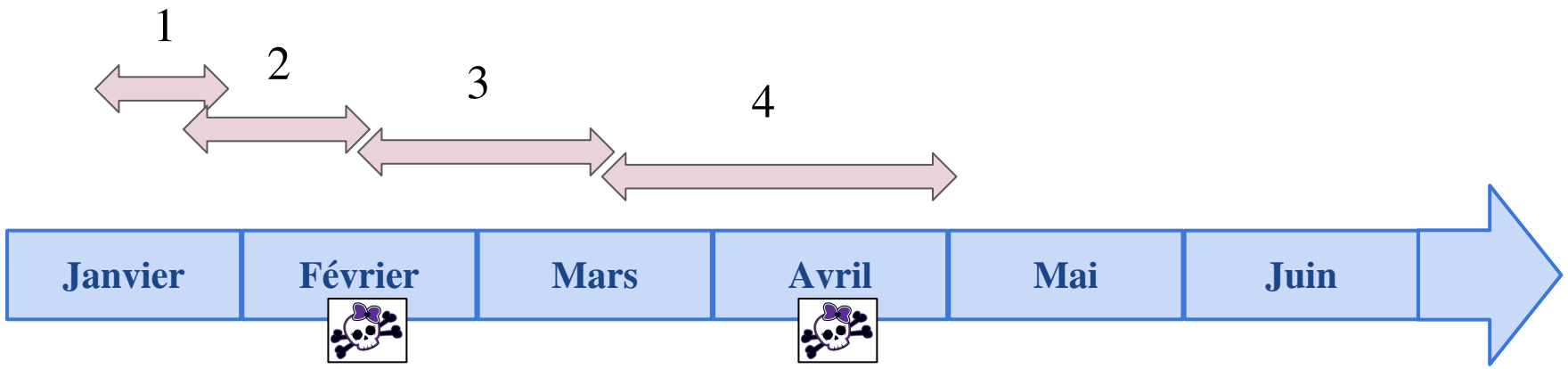
Workplan

1 - Choose species for analysis

2 - Identify families for each species

3 - Characterize families into each species

4 -1. Interspecific comparison
4 -2. Spatial organization analysis (HOR)



Workplan

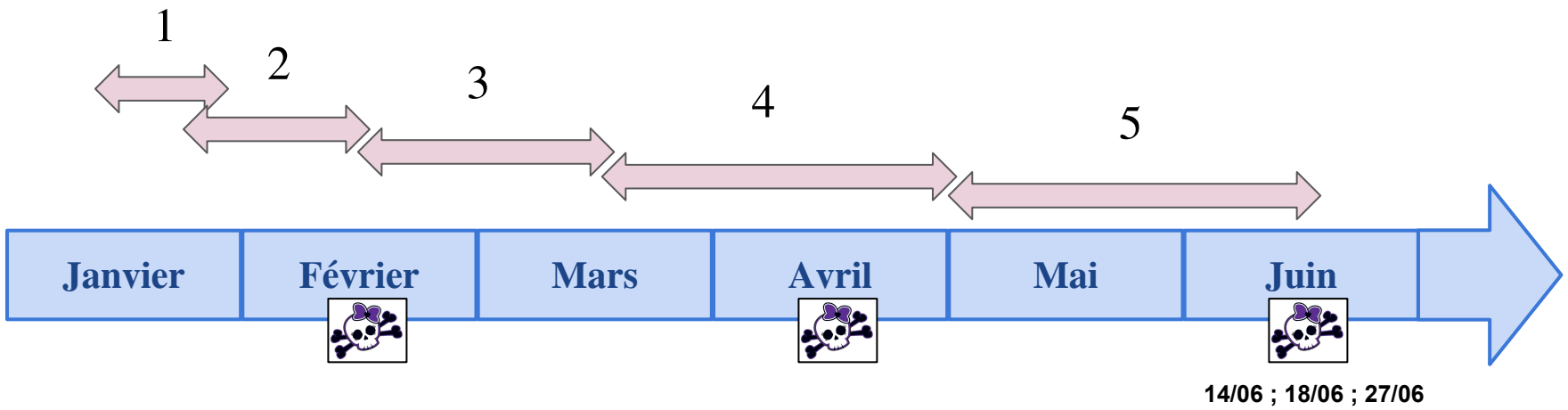
1 - Choose species for analysis

2 - Identify families for each species

3 - Characterize families into each species

4 -1. Interspecific comparison
4 -2. Spatial organization analysis (HOR)

5 - Writing of the report



Thank you for your attention

Data

Espèces	Nombre de séquences
Cercocebus atys (SRA)	80 884
Chlorocebus sabaeus	29 842
Gorilla gorilla	120 864
Homo sapiens (HGSC, HuRef)	37 204 ; 63 167
Macaca fascicularis (genome, SRA)	195642 ; 39893
Macaca mulatta (SRA)	7 365
Macaca nemestrina (SRA)	462 063
Mandrillus leucophaeus (WGS)	34 140
Nasalis larvatus	21 399
Nomascus leucogenys	392 948
Pan paniscus (SRA)	272 661
Pan troglodytes	114 104
Papio anubis	515 969
Papio hamadryas (Contigs)	70188
Pongo_abelii (genome, NCBI)	173 428 ; 175 240
Rhinopithecus_roxellana	109 811

Algorithm