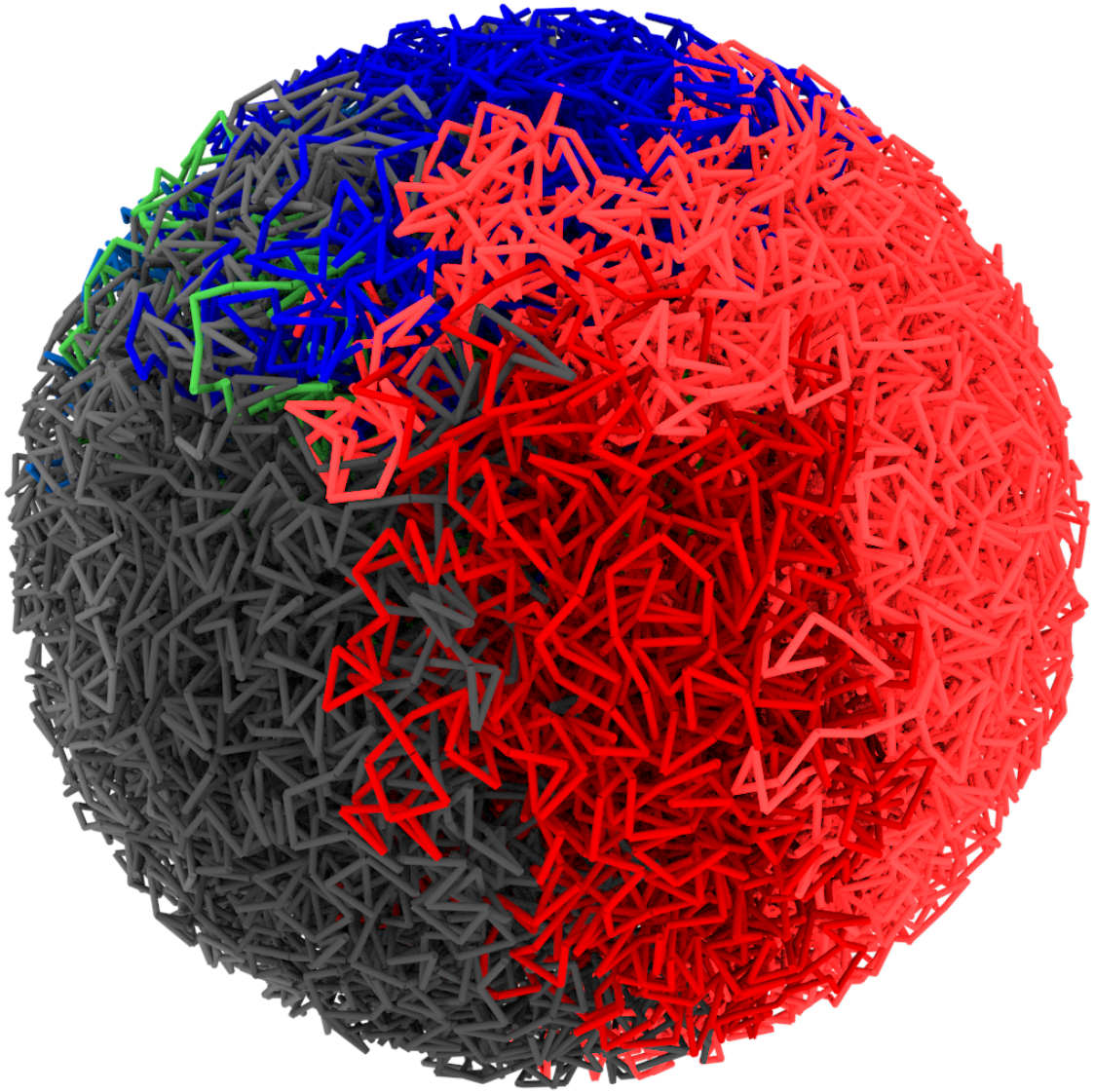
Capture numérique de contacts chromosomiques chez la mouche du vinaigre

P. Carrivain*, Y. Ogiyama†, C. Vaillant*, D. Jost‡, G. Cavalli†and R. Everaers*

*** Univ Lyon, Ens de Lyon, Univ Claude Bernard, CNRS, Laboratoire de Physique, F-69342 Lyon, France**
‡BCM Team, TIMC-IMAG Lab, CNRS, University Grenoble-Alpes
†Institute of Human of Genetics, CNRS UPR 1142

26 October 2018

Vision schématique d'un noyau

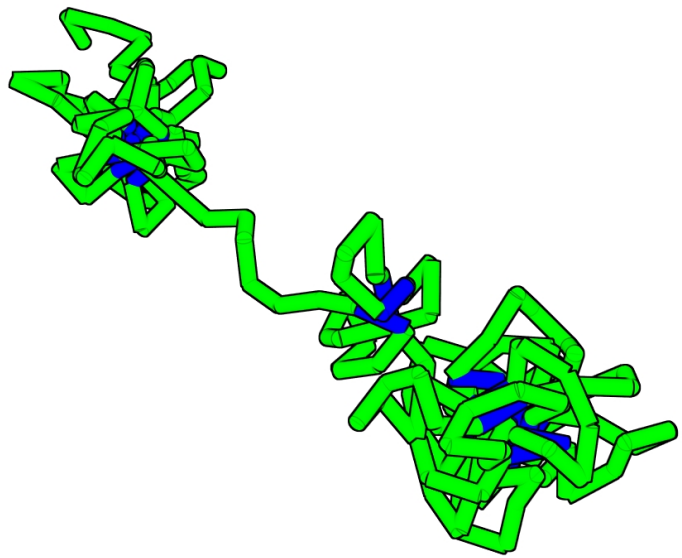


Capture de contacts chromosomiques

Capturing Chromosome Conformation

Job Dekker,[1]* Karsten Rippe,[2] Martijn Dekker,[3] Nancy Kleckner[1]

We describe an approach to detect the frequency of interaction between any two genomic loci. Generation of a matrix of interaction frequencies between sites on the same or different chromosomes reveals their relative spatial disposition and provides information about the physical properties of the chromatin fiber. This methodology can be applied to the spatial organization of entire genomes in organisms from bacteria to human. Using the yeast *Saccharomyces cerevisiae*, we could confirm known qualitative features of chromosome organization within the nucleus and dynamic changes in that organization during meiosis. We also analyzed yeast chromosome III at the G_1 stage of the cell cycle. We found that chromatin is highly flexible throughout. Furthermore, functionally distinct AT- and GC-rich domains were found to exhibit different conformations, and a population-average 3D model of chromosome III could be determined. Chromosome III emerges as a contorted ring.



De la levure de boulanger ...

1. organisme unicellulaire $\sim 10^7$ bp
2. un ingrédient : brosse de polymères

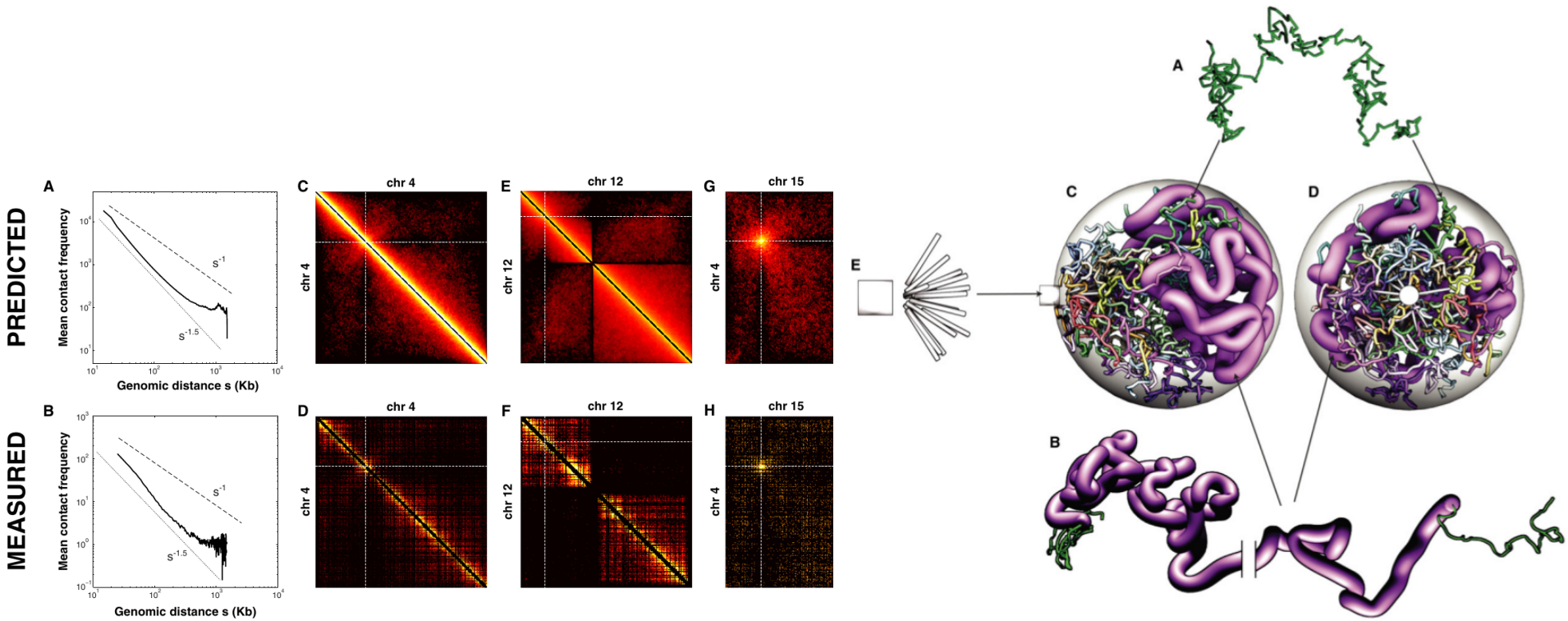


Figure: Courtesy of H. Wong et *al*, A Predictive Computational Model of the Dynamic 3D Interphase Yeast Nucleus. Current Biology 2012. Carrivain et *al*, *In silico* single-molecule manipulation of DNA with rigid body dynamics. Plos Computational Biology 2014.

... à la mouche du vinaigre

1. organisme pluricellulaire $\sim 10^8$ bp/noyau
2. + embryogénèse
3. + épigénétique
4. $\sim 10^6$ de cellules par expérience

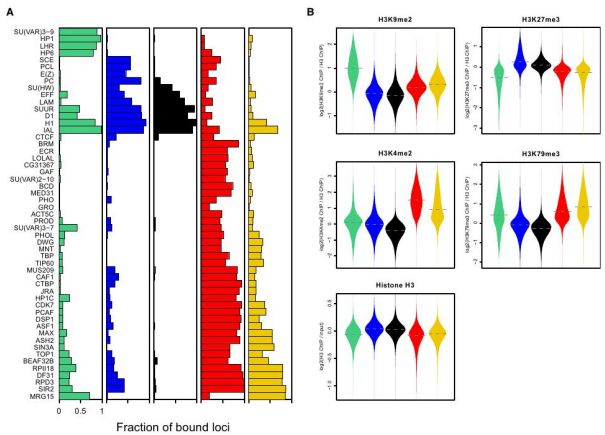
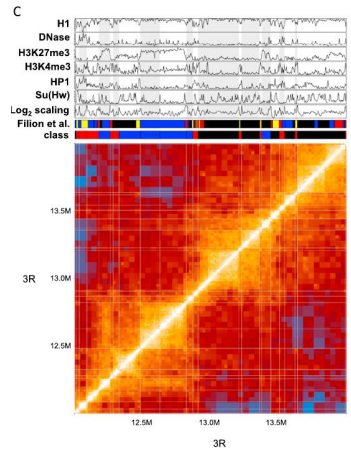
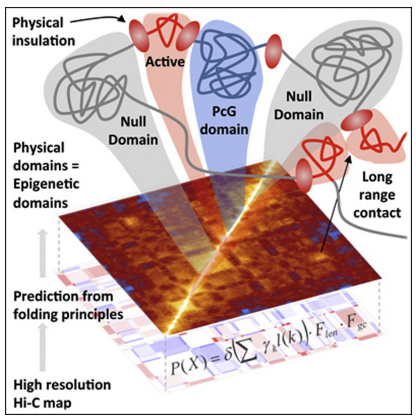
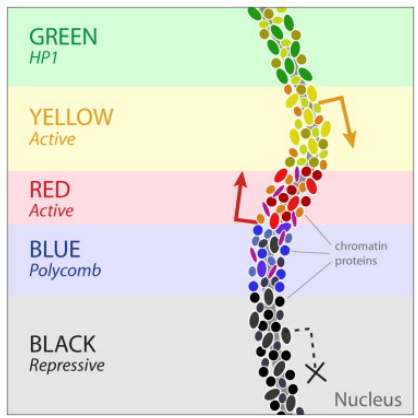


Figure 3. Chromatin Types Are Characterized by Distinctive Protein Combinations and Histone Modifications
(A) Fraction of all probed genomic loci within each chromatin type that is bound by each protein. Bound loci were determined separately for each protein as described in the text.
(B) Levels of histone H3 and four histone modifications as determined by genome-wide ChIP. The distribution of values is shown as "violin plots," which are symmetrized density plots of binding values per chromatin type: the wider the violin, the more data points are associated to that value. Dashed horizontal lines indicate the median binding value for each chromatin type. Histone modification ChIP data were normalized to H3 occupancy.
See also Figure S2.







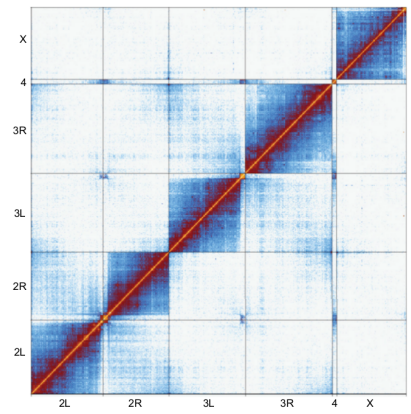
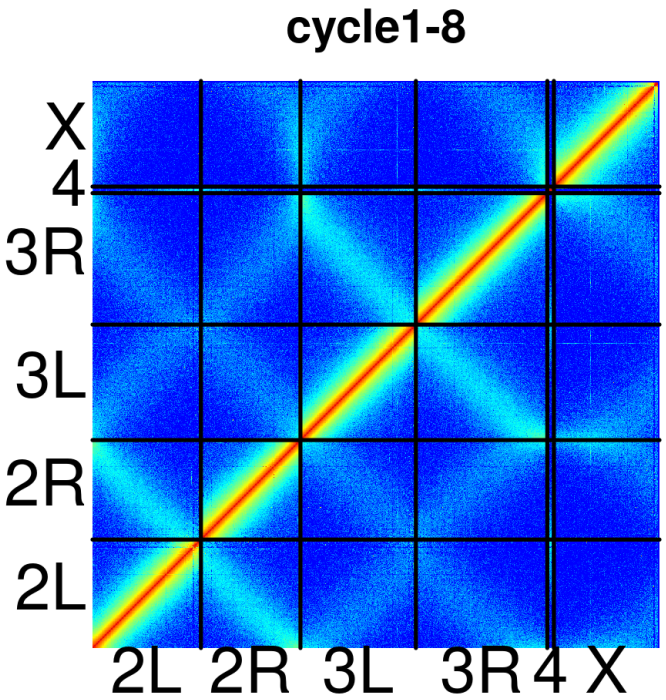
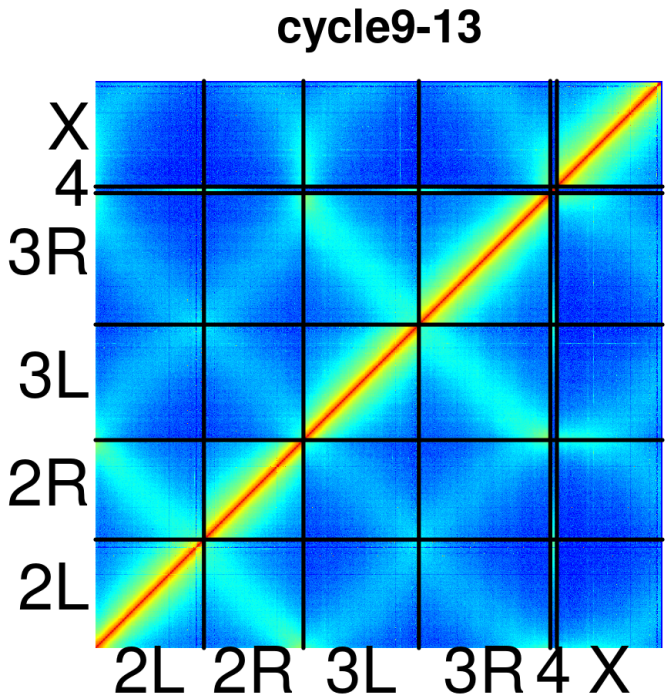
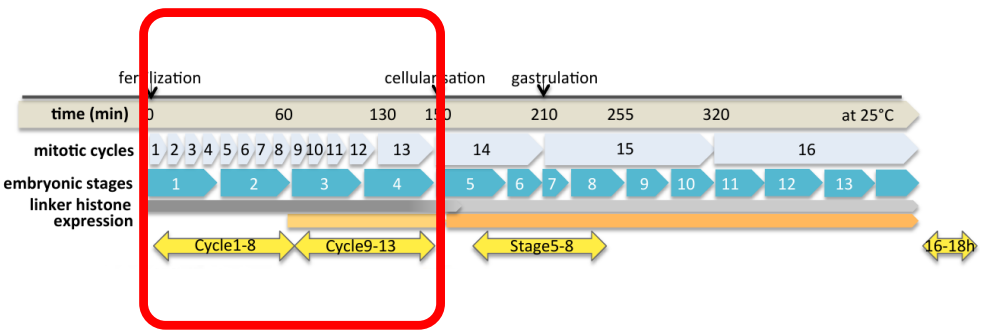


Figure: T. Sexton et *al*. Three Dimensional Folding and Functional Organization Principles of the *Drosophila* genome, Cell 2012 and G. Filion et *al*. Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in *Drosophila* Cells, Cell 2010.

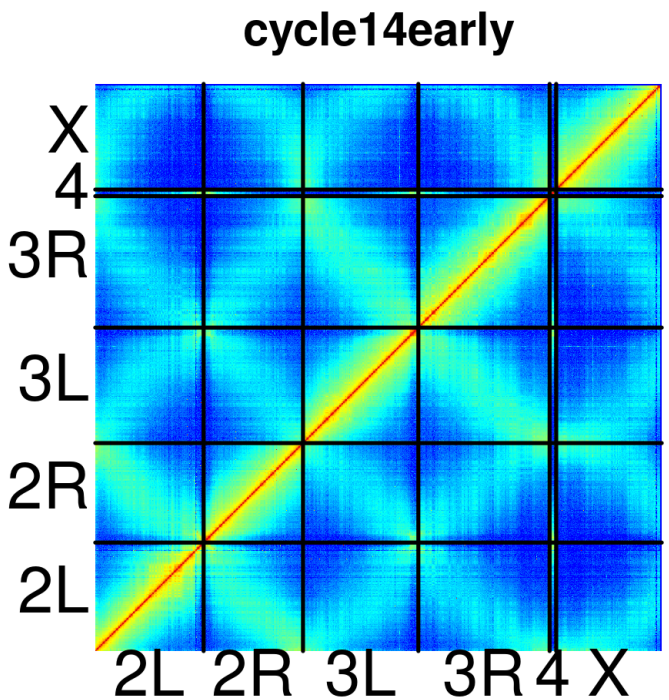
Les premiers cycles : des mitoses rapides.

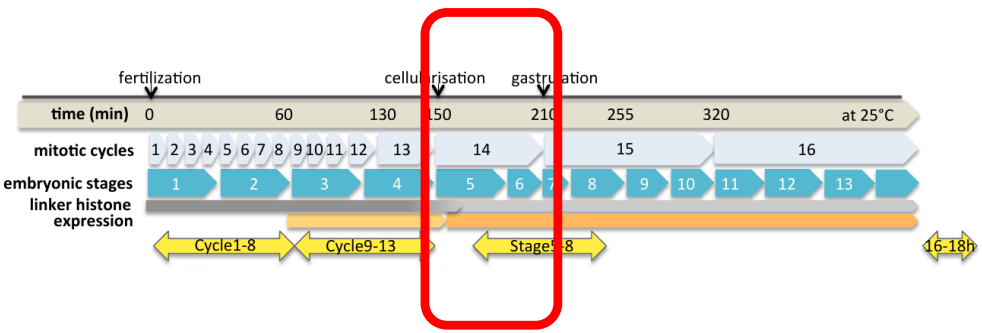






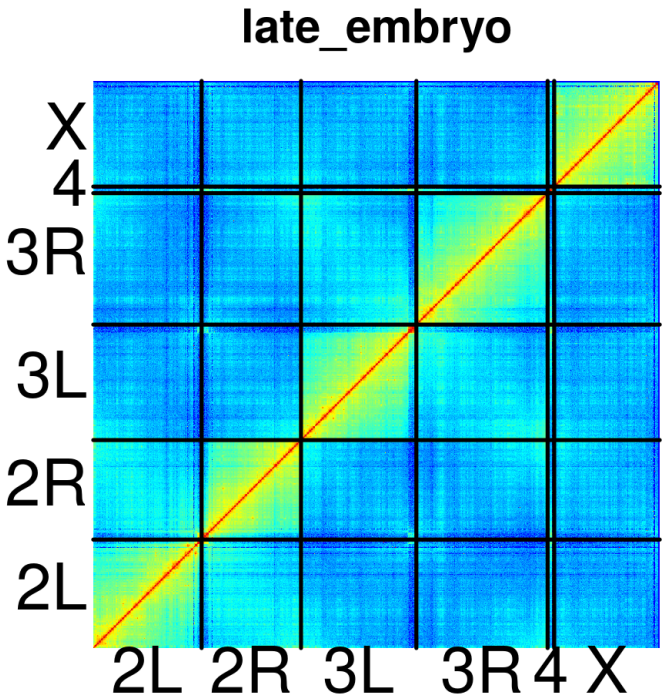
Le cycle 14 : apparition de la cellule.

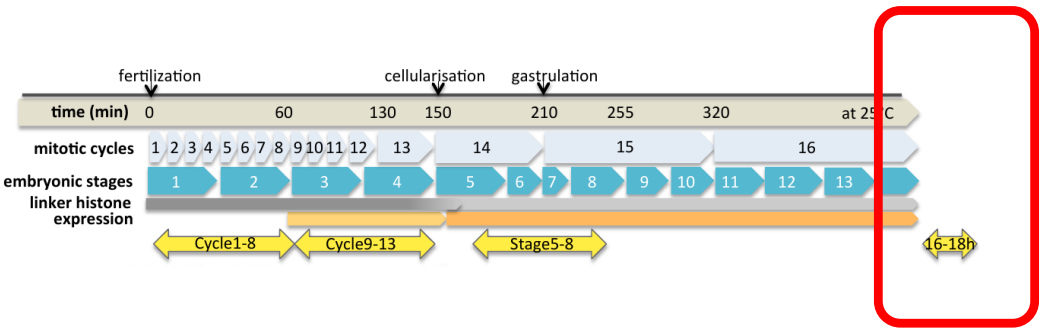




“apparition” de l’épigénétique.

Embryon tardif : après la gastrulation 16 – 18 h.

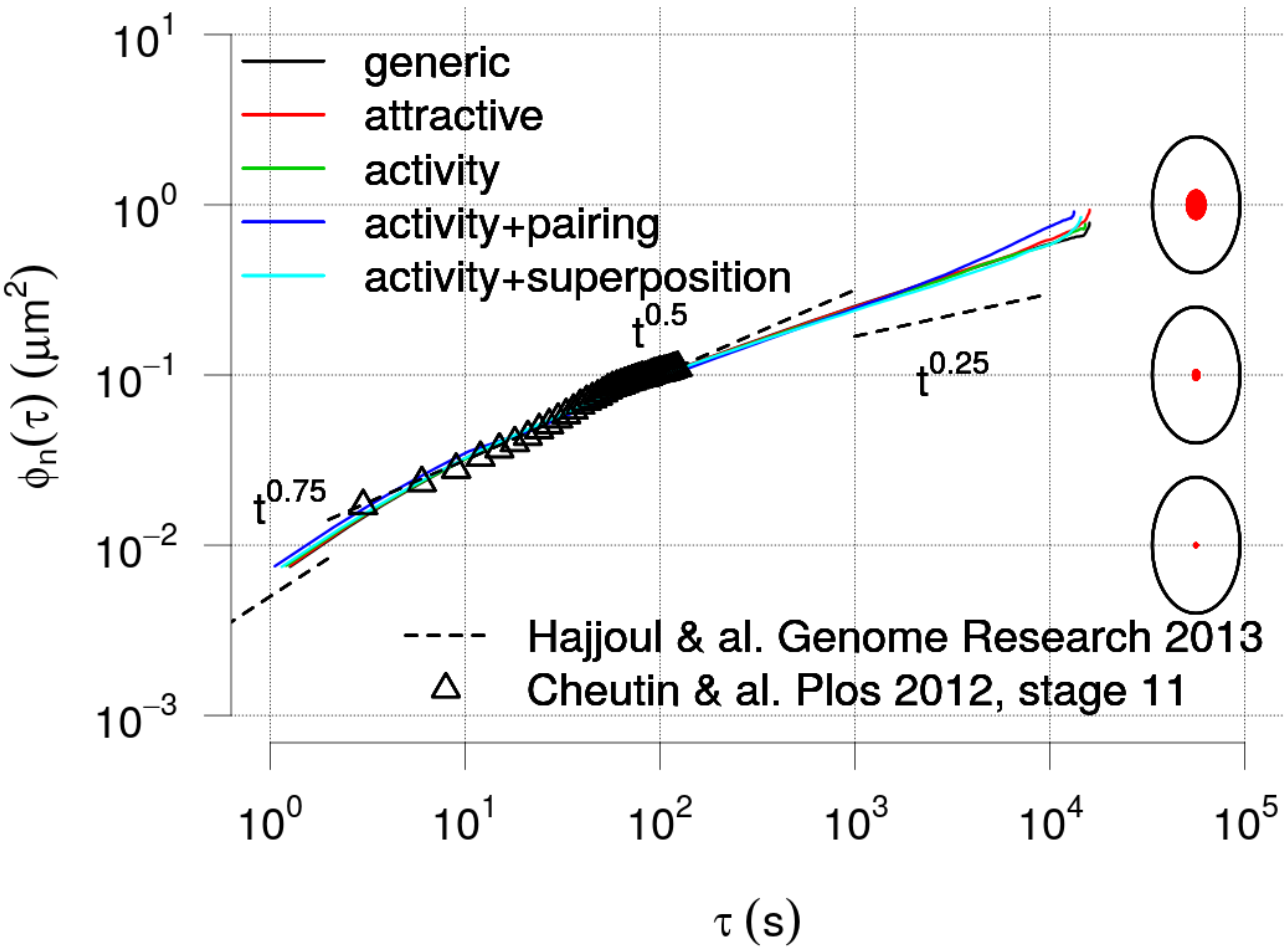




épigénétique différente de celle du cycle 14.

Comment trouver le temps "physique" de la simulation ?

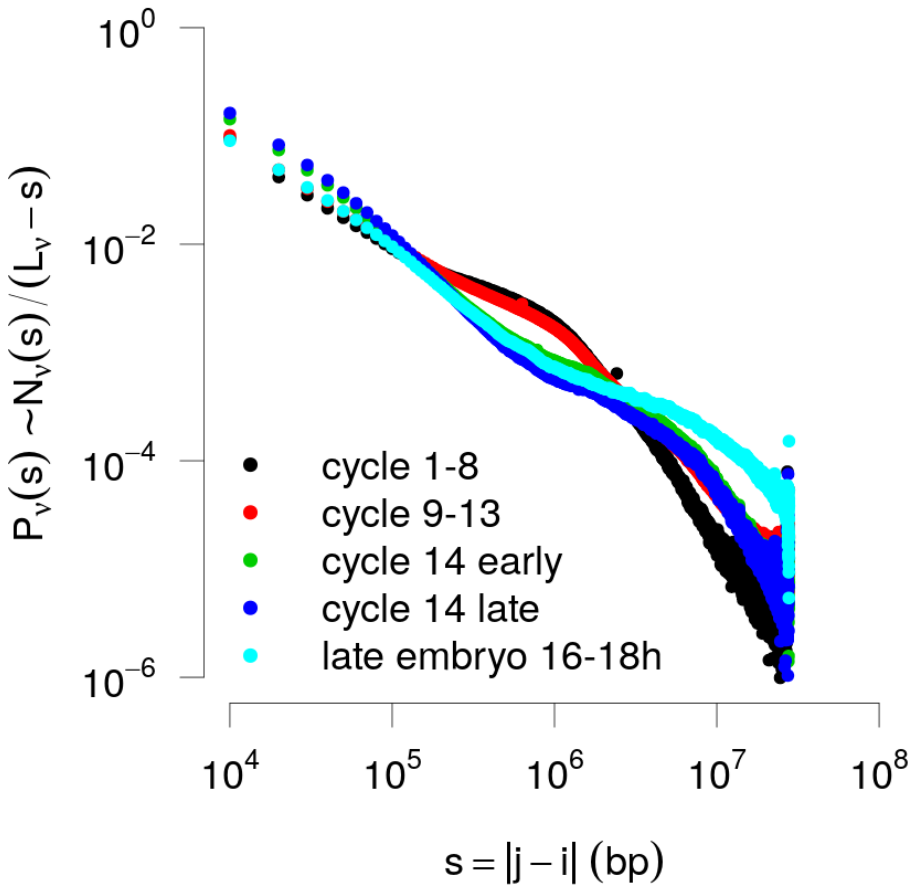
"Mean-Square-Displacement" : distance spatiale parcourue en X itérations, en Y secondes.

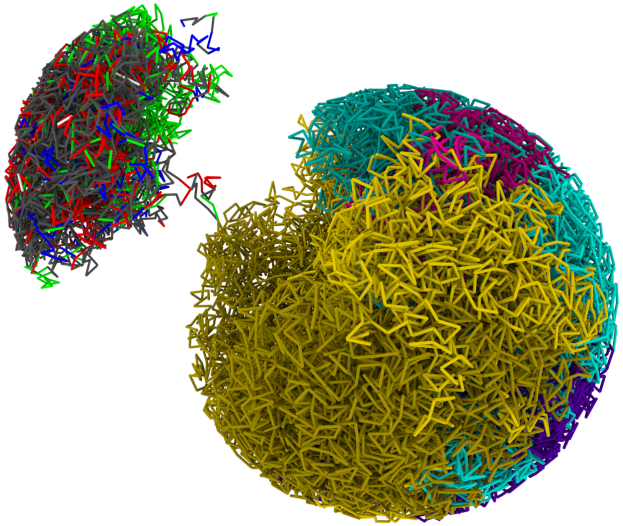


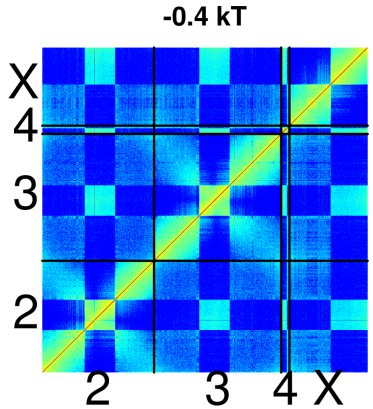
La statistique : combien de simulations par expérience ?

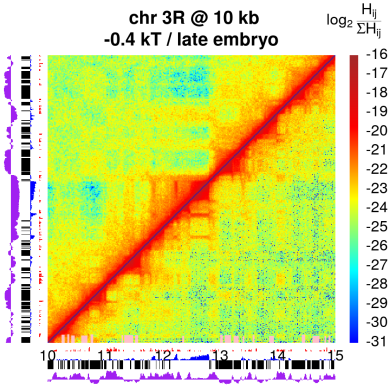
La probabilité de contact : une représentation gros-grain d'une carte de contacts. Quel que soit le stage de l'embryogénèse, la probabilité de contacts couvre quatre ordres de grandeurs.

environ 10000 **simulations par expérience (embryogénèse, mutant et autre espèce de mouche) !**







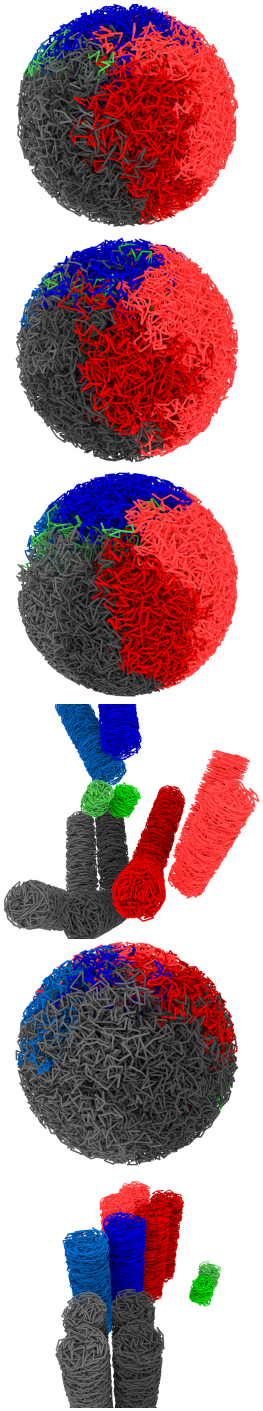


> 10 Mbp
memory effects & topology

< 10 Mbp
co-polymer & epigenetics

< 1 Mbp
transcription, replication ...

GENOMIC SCALE



Volume et traitement des données biologiques

En utilisant l'infrasctructure mise à disposition au **CBP** par Emmanuel Quemener :

- cat /proc/cpuinfo : Intel(R) Xeon(R) CPU E5-2637 v4 @ 3.50GHz
- cat /proc/meminfo : 131960340 KBytes
- Téléchargement et désarchivage de $\sim$ 400 G de données brutes.
- Assemblage des données brutes en une carte de contacts chromosomiques : $\sim$ 12 h par cycle.
- Après assemblage il ne reste plus que $\sim$ 7 G de données.
- Utilisation en local (**CBP**) des outils de bio-informatique (bowtie2, samtools, bedtools et logiciel fait "maison").

Métrologie

code Cpp "maison" lancé au :

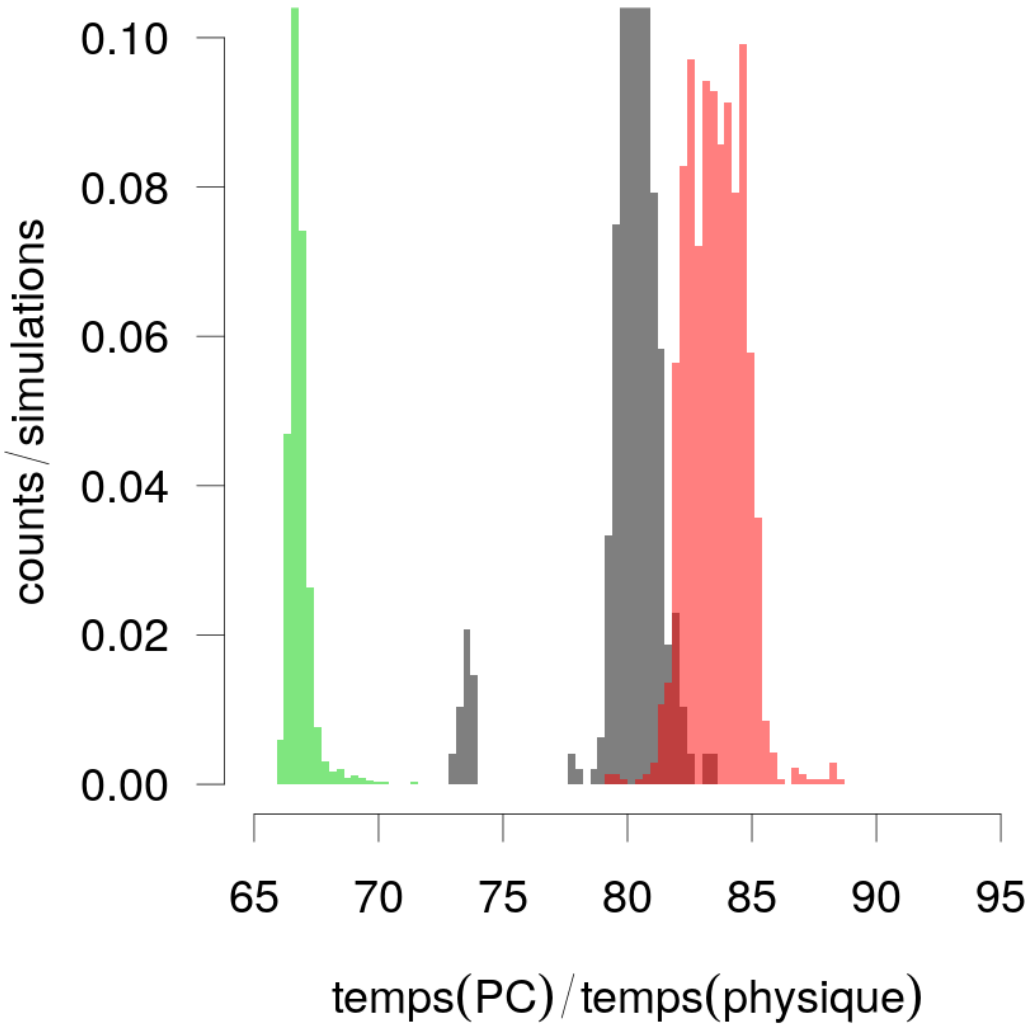
CBP : Emmanuel Quemener, Micaël Calvas, Cerasela Calugaru et Ralf Everaers
plateforme d'expérimentation avec 500 coeurs de calculs et 42 GPUs

PSMN : Lois Taulelle, Coraline Petit, Micaël Calvas, Cerasela Calugaru et Hervé Gilquin
plateforme de production avec environ 8000 coeurs de calculs et quelques GPUs

GENOTOUL : INRIA Toulouse
plateforme de production limitée à 100000 h de calcul par an

rapport temps réel temps processeur & mémoire

Métrologie : temps



Métrologie : temps, détail pour une simulation

Une expérience numérique, c'est 1 mois (temps d'attente dans la queue de soumission compris).

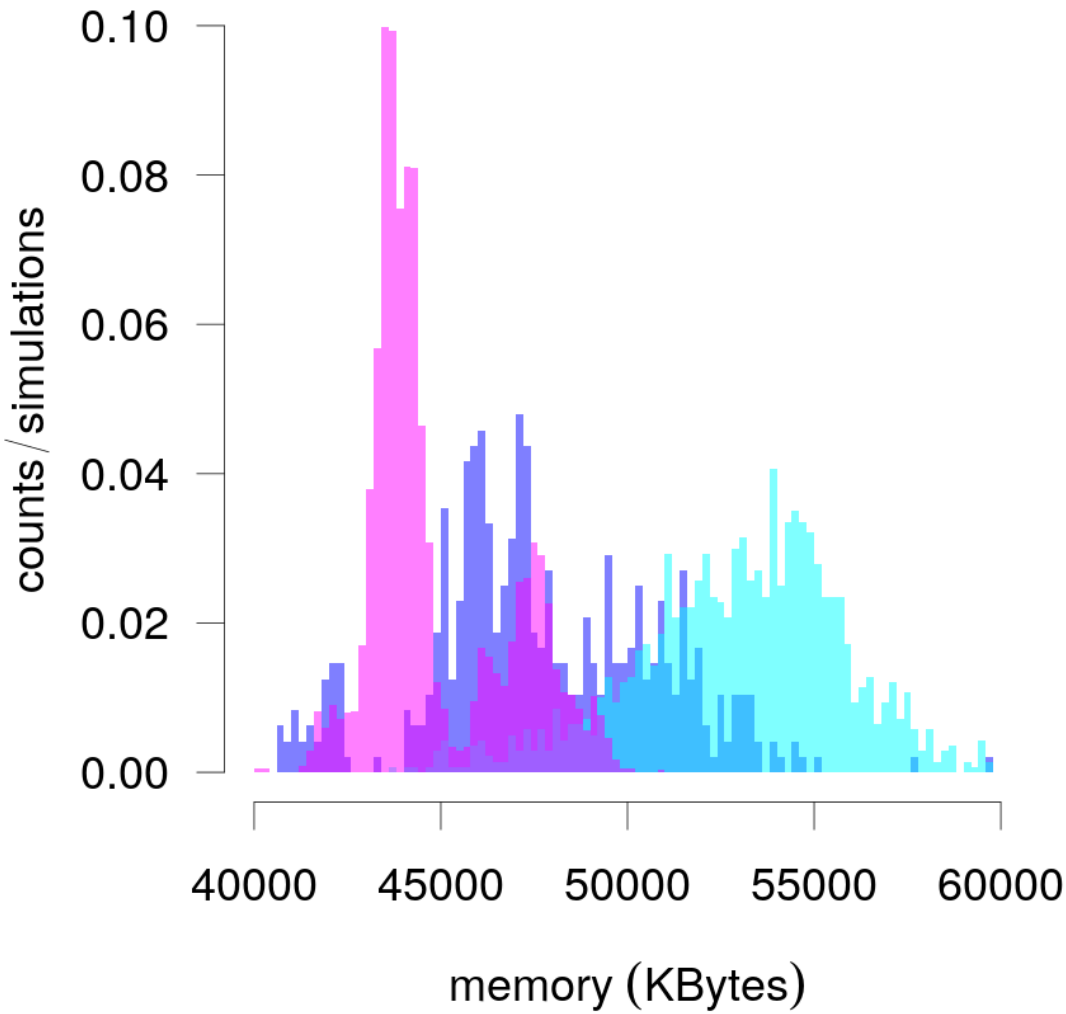
Détail :

- génération de nombres pseudo-aléatoires : ~ 10 %.
  - pour chaque simulation indépendante $\sim 10^{12}$ nombres pseudo-aléatoires à générer.
  - stockage dans un fichier de $\sim 10^{13}$ bytes ?
  - peut être parallélisée.
- calcul des interactions épigénétiques "deux-à-deux" : ~ 32 %.
- résolution de systèmes matriciels $\boldsymbol{\mathcal{A}x} = \boldsymbol{b}$ sous-contraintes $\boldsymbol{x} \geq 0$: ~ 58 %.

Limitation des queues de calcul à un certain nombre d'heures :

- sauvegarde de la configuration du système et de l'état du générateur pour une reprise à "chaud".
- une simulation de T itérations + une simulation de T itérations == une simulation de $2T$ itérations.

Métrologie : mémoire



Métrologie : entrée-sortie

- En entrée : un fichier (une fois) de 628 KBytes.
- En sortie : deux fichiers de structure (environ toutes les 3 - 4 heures) de $\sim$ 1.7 MBytes (écriture des valeurs en binaire pour ne rien perdre de la configuration du système dans la reprise à "chaud").
- Si toutes les simulations écrivent en même temps c'est $\sim$ 2 GBytes d'écriture dans le /scratch.
- Un run complet (temps et statistiques) c'est 65 fois $\sim$ 2 GBytes d'écriture dans le /scratch.
- Transfert du /scratch vers le /local d'une station de travail (at the edge) pour analyse.
  - plusieurs "rsync" au cours de l'expérience numérique.
  - analyse des dernières avancées de l'expérience numérique ($2 - 3$ h de temps PC pour extraire les observables).
- au total : volume d'environ 2T de données produites pour une exploration non-exhaustive d'interactions épigénétiques.

Métrologie : simulation à un niveau de détail plus important

d'une simulation au détail de 10000 bp à une simulation au détail de 1000 bp :

- algorithme en $N^{1.2}$: au moins un facteur 16 dans le temps PC.
- volume des données produites par un run : au moins un facteur 10.
- volume des données produites par l'assemblage des simulations : au moins un facteur 100.

c'est équivalent à passer de la simulation de la mouche du vinaigre à celle de l'humain à un niveau de détail constant.

D'un centre de calcul vers un GPU ?

Une simulation d'un noyau ne contient pas de parallélisme.

- jobs séquentiels sur des queues mono-cœur
- réservation de 16 et 32 coœurs pour un nœud entièrement dédié

Le "parallélisme" se trouve dans la réalisation d'un grand nombre de simulations indépendantes.

Portage sur GPU : un worker travaille sur une simulation (travail sur la mémoire utilisée par le programme).

D'un centre de calcul vers une ferme de GPUs ?

Algorithme hautement parallèle de chromosomes sur réseau (calcul en entier, implémentation OpenCL).

Tests sur l'infrasctructure GPUs développée et mise en place au **CBP** par Emmanuel Quemener.

simulation de taille N sur $2N$ itérations.

GPU	$N = 2^{11}$	$N = 2^{12}$	$N = 2^{13}$	$N = 2^{14}$	$N = 2^{15}$	$N = 2^{16}$	$N = 2^{17}$
Tesla P100-PCIE-16GB	0.09	0.18	0.37	0.87	2.68	8.99	31.53
Tesla K80	0.14	0.32	0.85	2.23	7.04	24.27	86.48
Tesla K40m	0.12	0.26	0.75	2.29	7.24	25.42	90.26
GTX 1080	0.11	0.23	0.55	1.78	6.31	21.72	75.88
GTX 980 Ti	0.13	0.26	0.56	1.36	5.23	18.97	65.65

GPU or CPU	total (en secondes)
Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz (one core)	1683.688345
GTX 780	51.62
GTX 980	40.25
GTX 980 Ti	36.34
GTX 1080	42.23
GTX TITAN	47.74
Tesla K40c	59.00
Tesla P100-PCIE-16GB	24.06

1. Avec une Tesla P100-PCIE-16GB on est plus 60× plus rapide qu'un Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz : **un noyau en direct !**
2. 30 Tesla P100-PCIE-16GB pendant un mois pour une expérience !

Conclusions

- “parallélisme” statistique
- reproductibilité des calculs
- exploration des interactions épigénétiques chez la mouche du vinaigre

- deux années “humaines” de calcul au **CBP** et au **PSMN** avec un service de qualité et une très grande réactivité
- quelques centaines d’années “processeurs” ($\sim 10^7$ h) de calcul au **CBP** et au **PSMN**

- portage sur carte graphique, en cours ...