

Cours Cartographie génomique, version
préliminaire sans garantie de l'auteur!, ne pas
reproduire les documents insérés ne sont pas
libre de droit, merci!

C. Gautier

12 juin 2007

1 plan du cours

L'objectif de ce cours est de présenter les concepts et les principaux outils informatiques associés à la cartographie génomique. Les principales étapes seront :

- la cartographie génomique : pourquoi faire, comment ?
- les différents types de cartes
 - les cartes de restriction (l'ancêtre!)
 - les cartes cytogénétiques (vertébrés, drosophile)
 - l'hybridation in situ
 - les cartes de restriction
 - les cartes génétiques
 - les cartes d'hybrides d'irradiation
 - les cartes déduites des séquences génomiques
- cartographies et structures des génomes
- la cartographie comparée
- quelques applications à l'agronomie et à la médecine

2 la cartographie génomique : pourquoi faire, comment ?

Les génomes ont été longtemps considérés comme un ensemble de gènes n'entretenant entre eux que des relations fonctionnelles, en particulier en terme d'épistasie. Le séquençage complet des génomes a relancé l'étude des relations spatiales entre gènes (et autre éléments génétiques) le long des génomes. La prise en compte de l'arrangement linéaires des gènes a montré à la fois son intérêt dans la compréhension des mécanismes évolutifs et son efficacité dans l'exploration de mécanismes biologiques aussi divers que la recombinaison, la recherche de gènes orthologues ou encore la fluidité des génomes. Plus récemment l'accès à l'arrangement spatial des chromosomes au sein d'une cellule a mis en avant l'intérêt d'une cartographie en 3D. Il apparaît en effet de plus en plus clairement que la proximité spatiale de gènes distant dans le génome, à un sens biologique.

Ce cours ne doit être considéré que comme une introduction à un domaine qui est encore en plein développement et qui soulève de nombreux défis tant biologiques qu'informatiques.

Avant de faire un bilan des différents types de cartes utilisées sur les génomes, les concepts communs peuvent être rapidement cités.

chromosome : Le génome des organismes est (dans la plupart des cas chez les eucaryotes) constitué de plusieurs molécule d'ADN. Chacune est intégrée au sein d'une structure complexes impliquant de multiples protéines. Deux structures jouent un rôle important à la fois fonctionnel et dans la description des chromosomes ce sont le centromère et les télomères. Les chromosomes ont deux bras de part et d'autre du centromère un bras court (p) et un bras long (q). Si les deux bras ont des longueurs voisines le chromosome est dit **métacentrique**, si au contraire le bras p est réduit à une zone très courte principalement hétérochromatique le chromosome est **acrocentrique**. Enfin si le centromère est à l'extrémité du chromosome celui-ci est dit **télocentrique**. Il faut enfin signaler le cas d'organismes n'ayant pas de centromère clairement identifié, on parle alors de chromosomes **holocentriques** (c'est entre autre le cas des lépidoptères).

marqueur : n'importe quelle structure pouvant être localisée sur un génome (un gène, une séquence anonyme, une région colorée par un produit, ...). En pratique deux classes de marqueurs peuvent être définies en fonction de l'importance ou pas qu'il soit polymorphe. Les marqueurs peu polymorphes ont une localisation qui peut être facilement déterminée (par hybridation expérimentale ou in silico), ils peuvent ainsi servir de points d'ancrage pour la cartographie. Les marqueurs fortement polymorphes sont nécessaires à la fois pour les cartes génétiques (voir plus loin) et pour les études populationnelles (détermination de paternité, flux de gènes, ...). On distingue de plus parmi ces derniers marqueurs ceux qui sont dits neutres : on suppose que leur variabilité n'est pas à une variabilité de la qualité (fitness) des individus qui les porte.

méiose : Division cellulaire conduisant à la formation des gamètes ...

carte : c'est une succession de marqueurs ou d'ensembles de marqueurs. Certaines cartes sont associées à une distance entre marqueurs.

synténie un ensemble de paires de gènes orthologues dans deux espèces et appartenant dans chacune des deux espèces à un seul chromosome.

segment conservé : la notion la plus floue ! en théorie un ensemble de gènes orthologues synténiques et dont l'ordre dans les deux chromosomes est le même. les difficultés sont de deux sortes :

- un relâchement plus ou moins important de la contrainte d'ordre
- la prise en compte des gènes sans orthologues connus ou dupliqués différemment dans les deux génomes.

En pratique il est impossible actuellement d'utiliser ce terme sans donner la définition que l'on utilise.

réarrangements : les cartes génomiques évoluent sous l'effet de réarrangements du génome. Ces réarrangements génomiques sont de plusieurs natures et surviennent au cours de l'évolution avec des probabilités différentes dépendant à la fois de la nature du réarrangement et du taxon

en cause. Les données permettant de mesurer l'impact évolutif des différents réarrangements sont à la fois dispersées et de natures diverses. Les données médicales fournissent une vue relativement exhaustive des réarrangements associées à des dysfonctionnements (maladies génétiques) non léthaux mais donnant lieu à des symptômes. Le polymorphisme "caryotypique" est largement étudié dans de nombreux taxons en particulier celui des rongeurs. Dans ce contexte les réarrangements Robertsoniens sont clairement les plus fréquents, même se ce ne sont pas les seuls (voir FIG. 1).

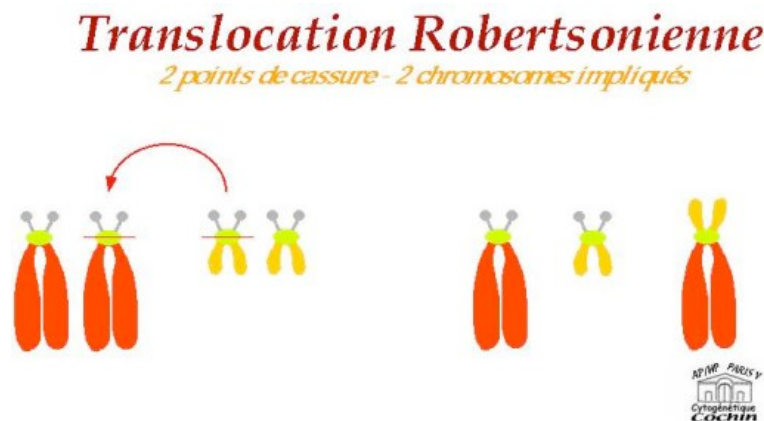


FIG. 1 – Les mutation par réarrangement Robertsonien réalise une fusion de deux chromosomes acrocentriques. Cela semble être le polymorphisme de réarrangement le plus fréquemment rencontré

Pour donner une idée un plus générale du type de réarrangement possible, la figure 2 extraite de l'Atlas of genetics and cytogenetics in oncology and haematology" ([http ://www.infobiogen.fr/services/chromcancer/](http://www.infobiogen.fr/services/chromcancer/)) donne les réarrangements observés dans des dysfonctinnements cellulaires.

A la fin du cours on reviendra sur ces concepts et sur quelques autres en les éclairant par les caractéristiques que nous aurons vu des différentes cartes.

3 les différents types de cartes

La diversité des cartes génomiques résulte à la fois des la diversité des approches expérimentales mais aussi de l'existence de plusieurs concepts biologiques de distances entres marqueurs génomiques. Si la distance "physique" comptée en nombre de bases séparant les deux marqueurs semble la plus naturelle, il faut être conscient que cette distance n'a pas de réel sens biologique. Le nombre moyen de "crossing over" est beaucoup plus signifiant à la fois en termes évolutifs que génétique.

3.1 les cartes de restriction (l'ancêtre!)

Le principe est la coupure du fragment d'ADN à cartographier par des enzymes de restriction ou des ensembles de 2 voire 3 enzymes. La migration des

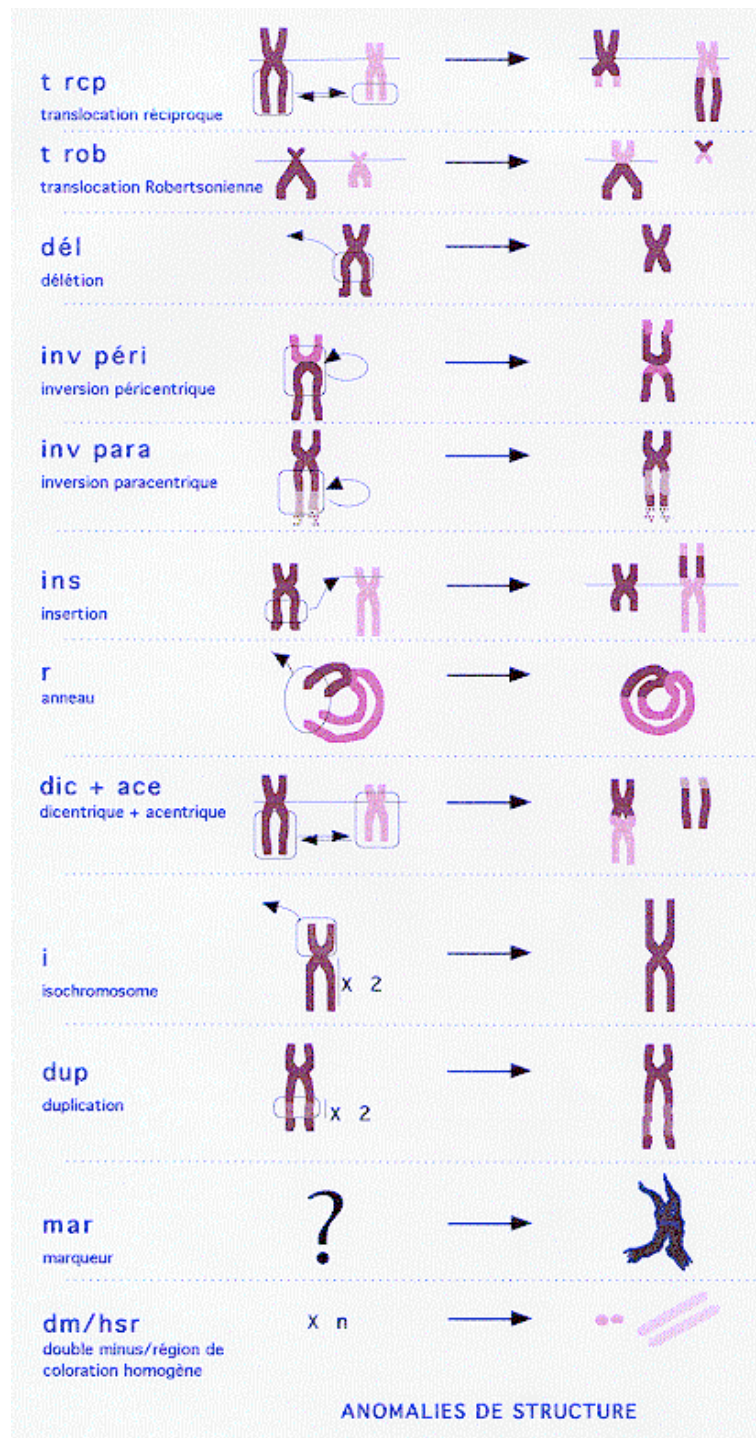


FIG. 2 – réarrangement décrits dans l'Atlas of genetics and cytogenetics in oncology and haematology"

fragments sur gels les classes par taille. On tente ensuite de reconstituer les positions de coupures cohérentes avec ces tailles de fragments. Un exemple sur un élément répété d'une drosophile est donnée figure 3

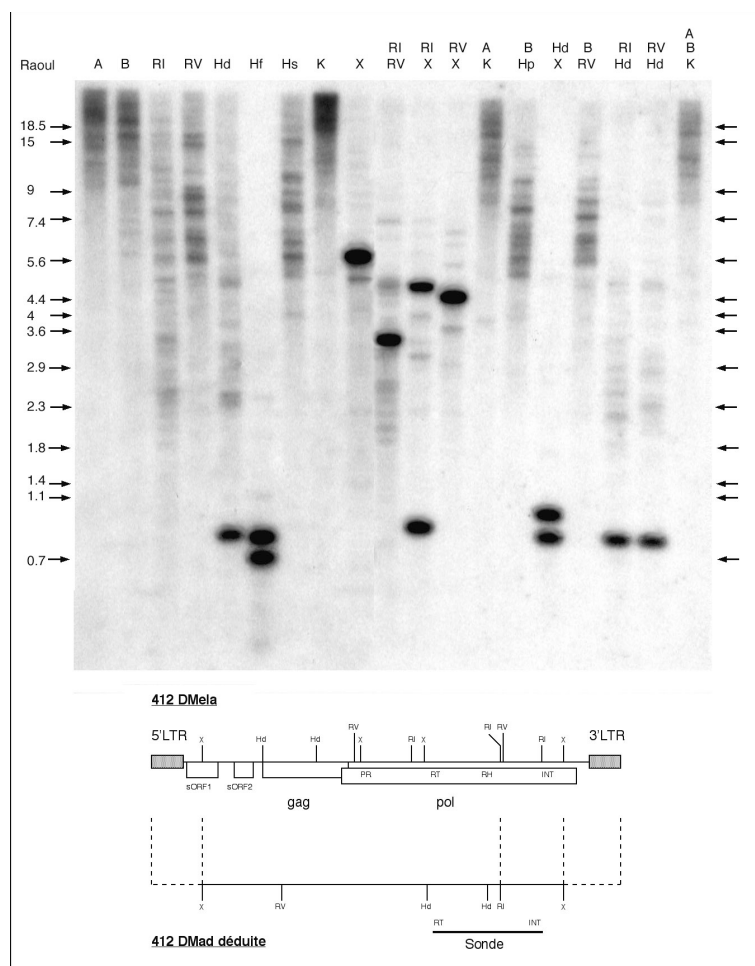


FIG. 3 – Carte de restriction établie par Southern-blots de l'élément 412 de *D. madeirensis*. A :ApaI, B :BamHI, RI :EcoRI, RV :EcoRV, Hd :HindIII, Hs :HpaI, K :KpnI, X :XhoI

3.2 les cartes cytogénétiques (vertébrés, drosophile)

les cartes cytogénétiques les plus utilisées proviennent de deux phénomènes sans rapport. Chez les vertébrés le traitement des chromosomes avec certains colorant (en particulier le giemsa) fait apparaître des bandes claires et sombres à l'examen au microscope des chromosomes. Ces bandes sont reproductible et constitue une "signature" de chaque chromosome et découpe chaque chrmosome en sous ensembles (les bandes) dans lesquels on peut placer des marqueurs. L'ordre des marqueurs au seins d'une bandes n'est pas connu, il en résulte des cartes constituées d'une suite d'ensemble de marqueurs, chaque ensemble

correspondant à une bande.



FIG. 4 – Caryotype humain à gauche la femme à droite l'homme

Chez la drosophile certaines cellules, localisées dans la glande salivaire, subissent une très forte multiplication de leur génomes. Ces différentes copies s'associent pour permettre l'examen microscopique des chromosomes. Sur ces chromosomes apparaissent des bandes (sans qu'une coloration chimique soit nécessaire) qui sont reproductibles et peuvent s'utiliser comme les bandes G des génomes vertébrés.

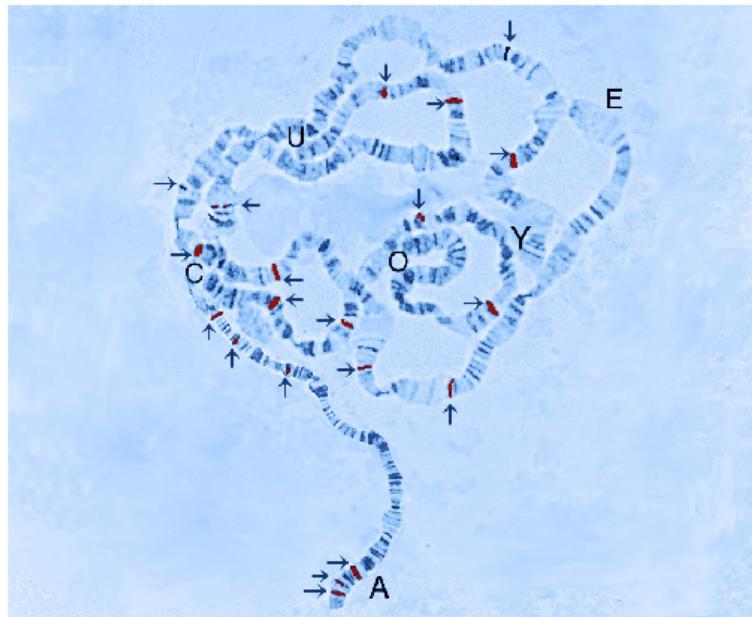


FIG. 5 – insertion de copies d'un élément transposable de la drosophile. Par hybridation avec une sonde fluorescente sur les chromosomes polytènes des glandes salivaires on peut repérer la position des éléments par rapport aux bandes

3.3 l'hybridation in situ

Pour positionner des marqueurs sur les cartes précédentes la technique la plus utilisée consiste à tirer partie des mécanismes d'hybridation. D'une manière très schématique on associe à un fragment de séquence du marqueur (une

sonde) une molécule permettant son repérage (molécule radioactive, molécule émettant une longueur d'onde visible par exemple sous excitation laser). Le marqueur apparaît ainsi visible soit directement soit par photographie avec des films adaptés au type d'émission de la molécule utilisée (FIG. 5).

L'utilisation de fluorochromes de plusieurs couleurs permet de visualiser des relations complexes entre génomes. La figure 3.3 montre une utilisation en génétique médicale.

ZOO-FISH analysis in a species of the order Chiroptera: *Glossophaga soricina* (Phyllostomidae)

Marianne Volleth^{1,4}, Christine Klett², Antje Kollak², Christa Dixkens^{2,5}, York Winter³, Walter Just², Walther Vogel² & Horst Hameister^{*2}
¹Institut für Humangenetik, Universität Erlangen-Nürnberg, Erlangen, Germany; ²Abteilung Medizinische Genetik, Universität Ulm, Ulm, Germany, Tel.: #-731-5023436, Fax.: #-731-5023438, email: horst.hameister@medizin.uni-ulm.de; ³Institut für Zoologie, Universität Erlangen-Nürnberg, Erlangen, Germany; Present address: ⁴M.V. Institut für Humangenetik, Otto-von-Guericke-Universität, Leipziger Straße 44, D-39120 Magdeburg; ⁵C.D. Institut für Humangenetik, Universität Göttingen, Göttinger Straße 12d, D-37073 Göttingen
^{*}Correspondence

Received 19 October 1998; accepted for publication by M. Schmid November 1998

Key words: Chiroptera, chromosomal evolution, comparative gene mapping, ZOO-FISH

Abstract

Glossophaga soricina is a flower-visiting bat which lives in the neotropics. The diploid chromosome number is $2n = 32$ with a fundamental number of autosomal arms, FN, of 60. *G. soricina* belongs to the Microchiroptera which have a lower diploid DNA content and a higher AT composition in their DNA compared with other mammals. By ZOO-FISH analysis with human chromosome-specific DNA probes, the human autosomes were found conserved in 41 segments. This is an arrangement similar to other mammals which have been analyzed. Several chromosomal associations already known from ZOO-FISH studies in other species were also present in *G. soricina*.

Introduction

In the past it has been repeatedly shown that karyotype evolution in mammals is highly conservative (Ohno 1973; Wakefield & Graves 1996). It is therefore tempting to speculate on the structural composition of an ancestral mammalian founder karyotype which more than 100 million years ago gave rise to the different mammalian karyotypes of today's species. To develop ideas about the make-up of this hypothetical karyotype, it is necessary to have available karyotype and gene mapping data from at least one species out of each mammalian order. It would be most informative if the karyotype of the species

studied represents a common type in that order. For example, in the order of Carnivora, the cat (*Felis catus*) karyotype is representative for a phylogenetically old karyotype in that order (Dutrillaux & Couturier 1983). Therefore the comparison of the cat karyotype and cat gene arrangement with those from other species and from other orders has been highly informative. The second largest mammalian order, Chiroptera, with nearly 1000 extant species, has been scarcely studied in this respect. The common long-tongued bat, *Glossophaga soricina*, studied here, lives in the tropical and subtropical regions of the New World. *G. soricina* belongs to the suborder Microchiroptera, family Phyllostomidae, for which

3.4 les cartes génétiques

Le mécanisme biologique qui soutient la construction de ces cartes est la recombinaison. Au cours de la méiose les chromosomes paternel et maternel homologues s'apparient et des échanges ont lieu. Je ne reprendrais pas ici la description biologique du phénomène que vous avez vu dans un autre cours

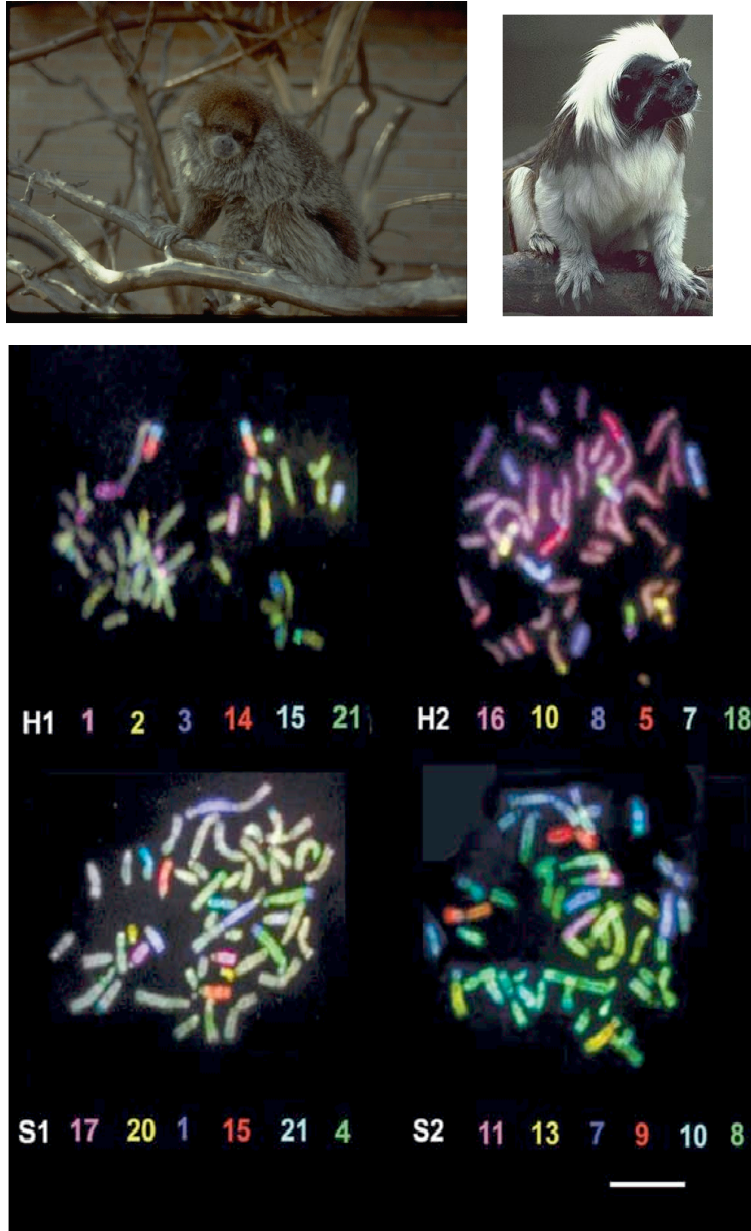


FIG. 6 – expérience FISH entre *Callicebus donacophilus* et *Homo sapiens* (H) d'une part et *Saguinus oedipus* (S) d'autre part ; Barros RMS (2003) Chromosome Res., 11 :327-334,

(relation entre la recombinaison et la structure en isochores). On se limitera ici à une description formelle : étant donnés deux marqueurs (A avec ces deux copies a_m et a_p et B avec ses copies b_m et b_p) il existe une probabilité P de formation de gamètes recombinés ($a_m - b_p$ et $a_p - b_m$) et P est reliée à la distance entre les deux locus par une fonction croissante (dites fonctions de cartographie (mapping functions)). La fonction la plus simple suppose que le nombre d'événements de recombinaison entre deux marqueurs suit une loi de Poisson dont le paramètre est proportionnel à la distance génétique entre les deux marqueurs. Ce modèle permet de gérer le fait que les recombinaisons observées entre deux marqueurs correspondent à un nombre impair de recombinaisons entre ces marqueurs. on reviendra sur ces aspects mathématiques dans la dernière partie du cours.

3.5 les cartes d'hybrides d'irradiation (largement inspirée de la thèse de R. Guyon)

Les cartes d'hybrides d'irradiation repose sur la probabilité d'appartenance de deux marqueurs à un même segment génomique après cassure du génome par irradiation. Expérimentalement on tire parti du comportement (non encore expliqué) de cellules hybrides obtenue par fusion de cellules d'un rongeur (hamster par exemple) avec des cellules de l'organismes d'intérêt. Le succès de la fusion étant relativement faible, un enrichissement en cellules fusionnées est obtenu par l'utilisation d'un gène de sélection (par exemple TK, thymidine kinase). Dans un milieu particulier ce gène est indispensable à la survie des cellules. Les cellules de rongeurs utilisées sont dépourvues de ce gène et leur culture sur le milieu sélectif ne conserve que les lignées dans lesquelles le gène orthologue de l'organisme étudié complète le génome déficient du rongeur.

Chacune des lignées sélectionnées ne va retenir qu'une partie des segments génomiques. Par hybridation avec des marqueurs spécifiques on retient des lignées ayant un taux de rétention satisfaisant (par exemple autour de 10%). Une centaine de telles lignées est en générale suffisant pour une première cartographie d'un génome. La précision de la cartographie d'épend de la taille des segments elle même déterminé par l'intensité de l'irradiation.

La construction de la carte repose sur l'analyse expérimentale de la présence de marqueurs dans les différentes lignées. La présence simultanée de deux marqueurs dans la même lignée est un argument pour leur proximité génomique (appartenance à un même segment). Il est cependant clair que ceci nécessite une analyse mathématique et informatique plus fine qui sera présenté dans la partie "Informatique et cartographie".

3.6 les cartes déduites des séquences génomiques

3.7 comparaison entre cartes construites sur un même génome

On se limitera ici à la comparaison entre les cartes physiques et génétiques. Si on peut supposer que l'ordre des gènes est le même sur ces deux type de cartes, la relation entre les distance n'est pas linéaire. En prenant comme référence la distance physique (nombre de nucléotides séparant deux marqueur) les différentes régions du génome montrent des taux de recombinaison différent. Le

graphe de la distance génétique en fonction de la distance physique est appelé cartes de Marey. La figure 7 en donne un exemple.

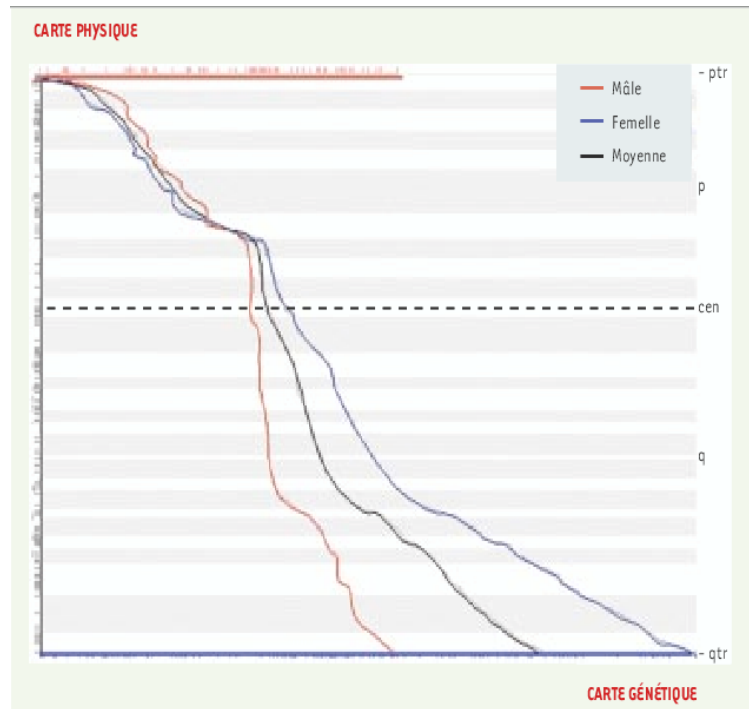


FIG. 7 –

4 cartographies et structures des génomes

5 la cartographie comparée

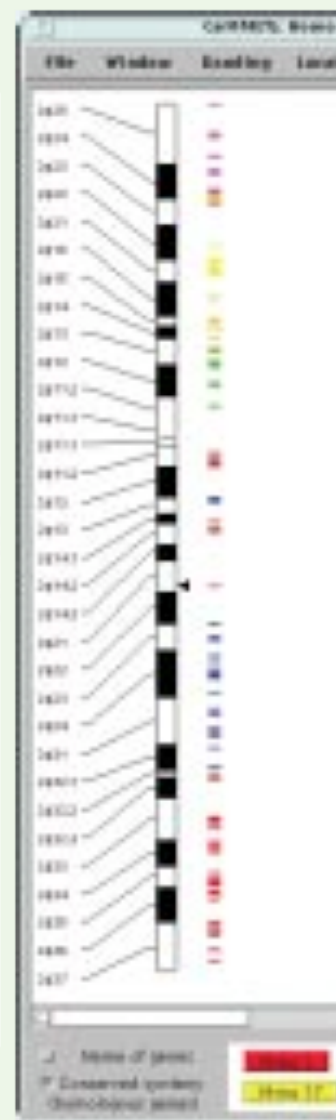
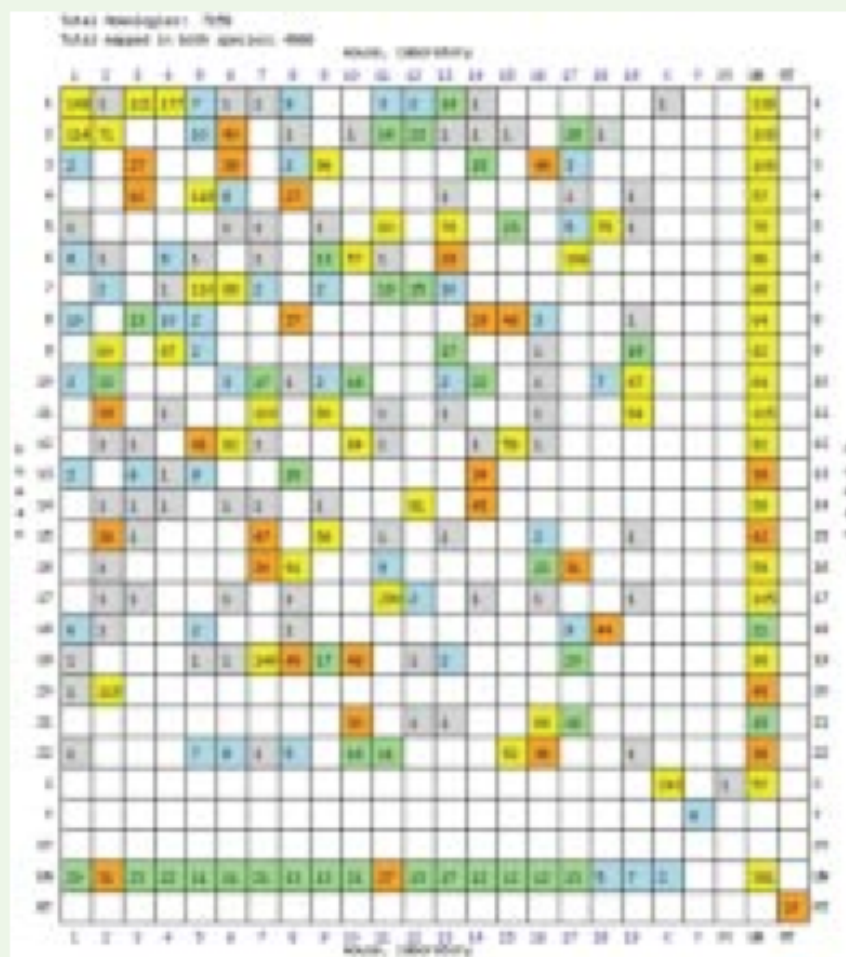
6 quelques applications à l'agronomie et à la médecine

7 mathématiques et cartographie génomique

7.1 représentation des connaissances

7.1.1 le projet GeM

Ce projet a été initié au laboratoire par B. Spataro et G. Bronner et se poursuit actuellement par la thèse de V. Navratil. L'objectif est l'analyse des relation entre le fonctionnement et l'évolution des génomes de vertébrés d'une part et l'organisation spatiales le long des chromosomes de l'information génomique d'autre part.



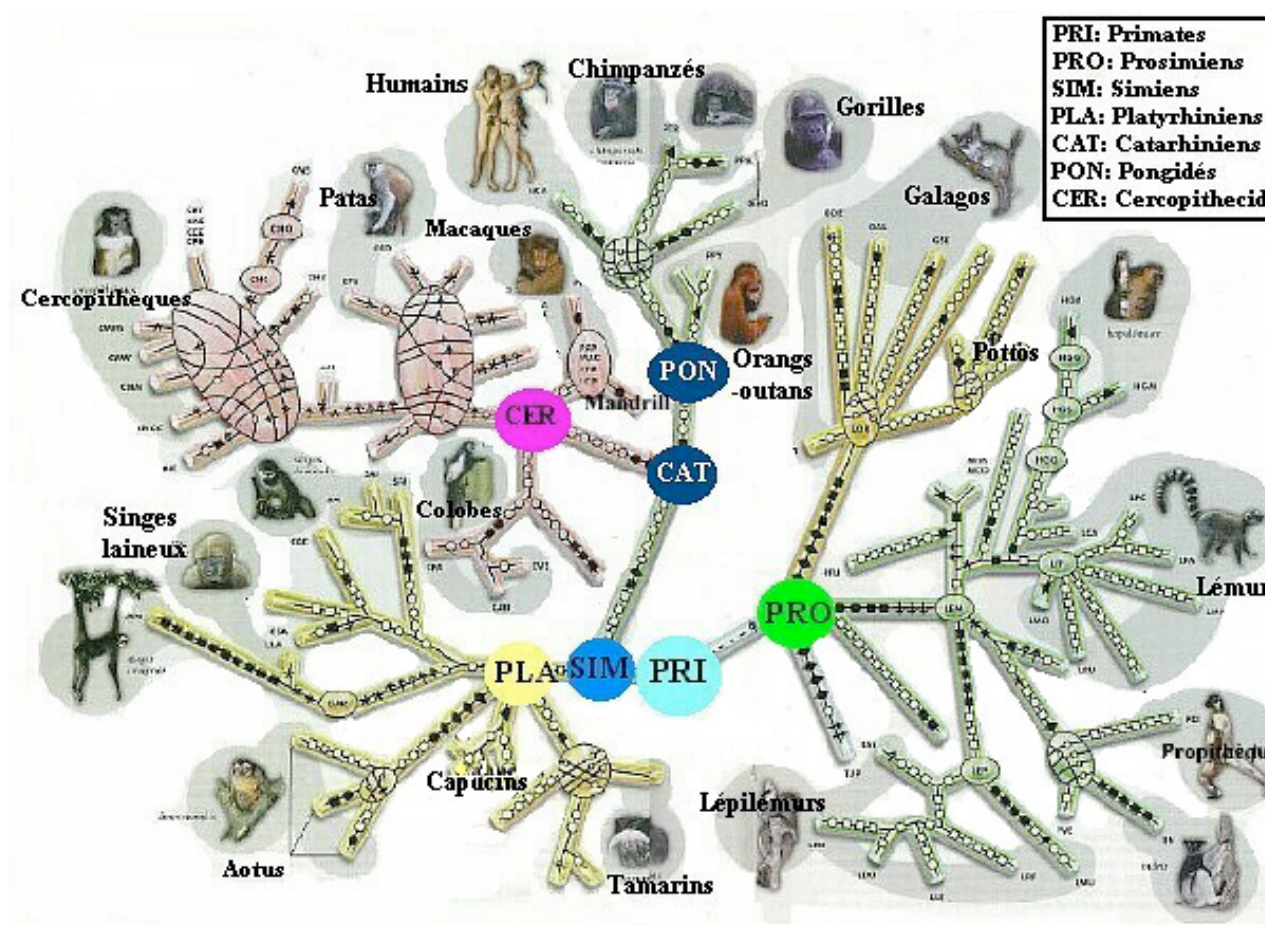


FIG. 9 –

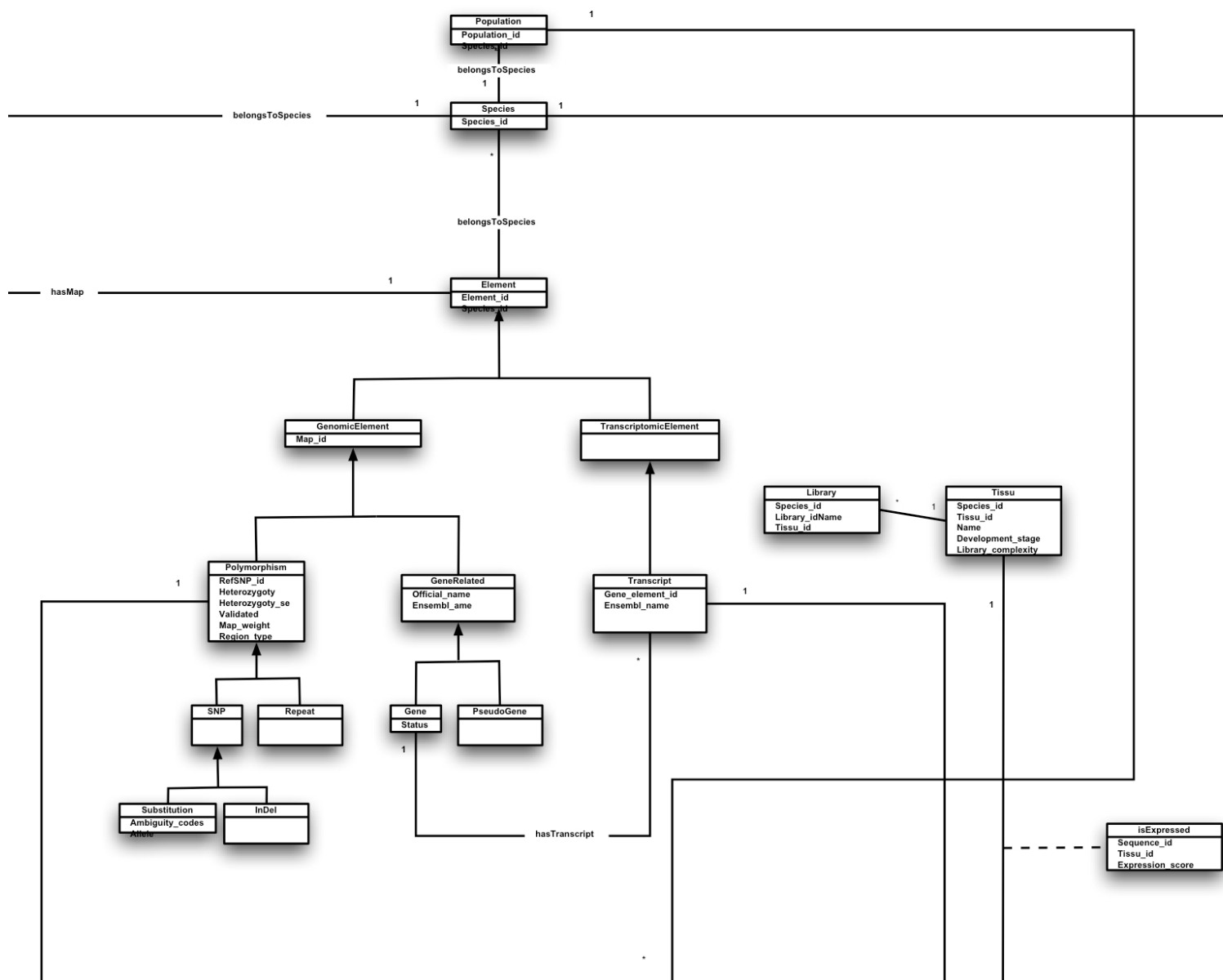


FIG. 11 –

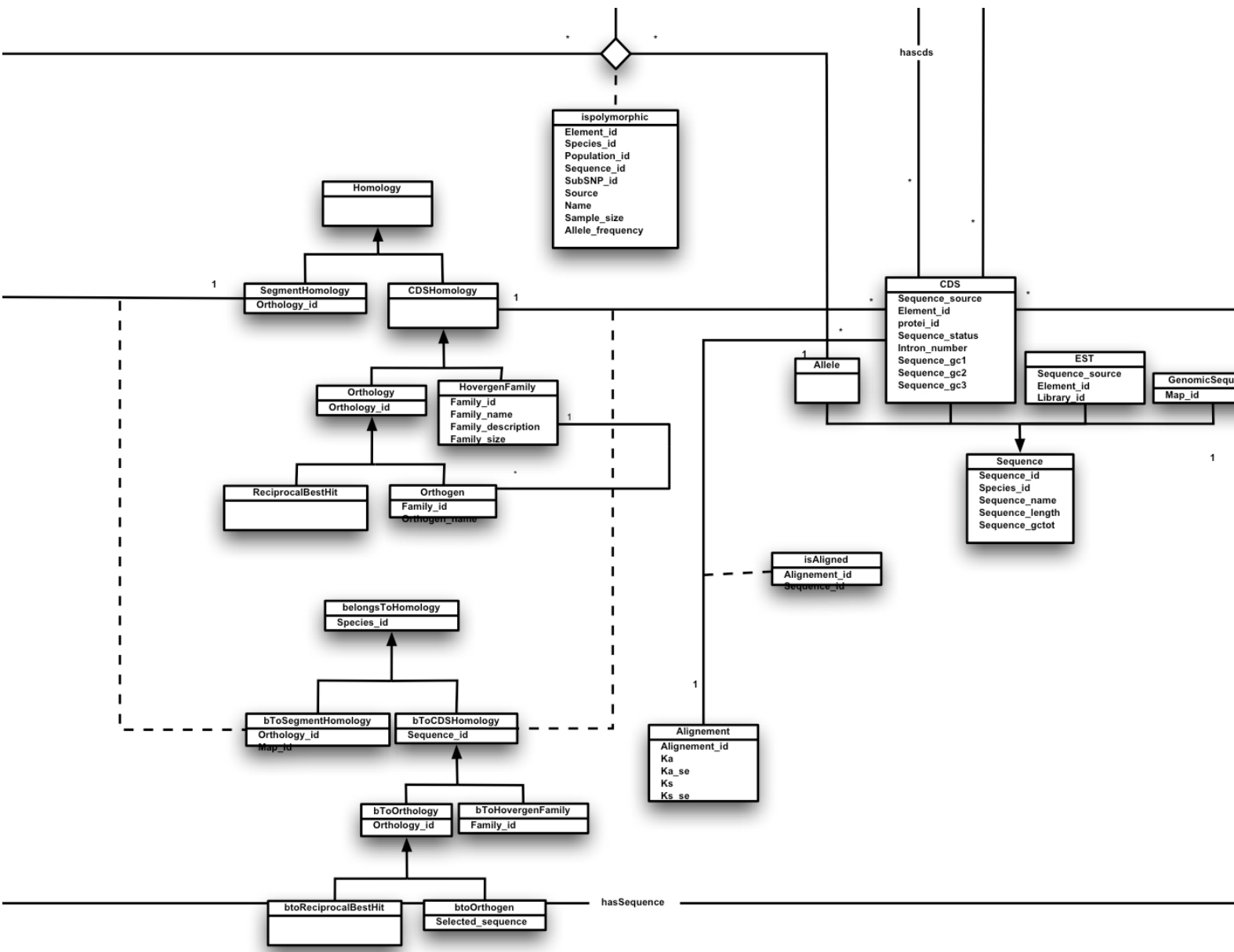


FIG. 12 –

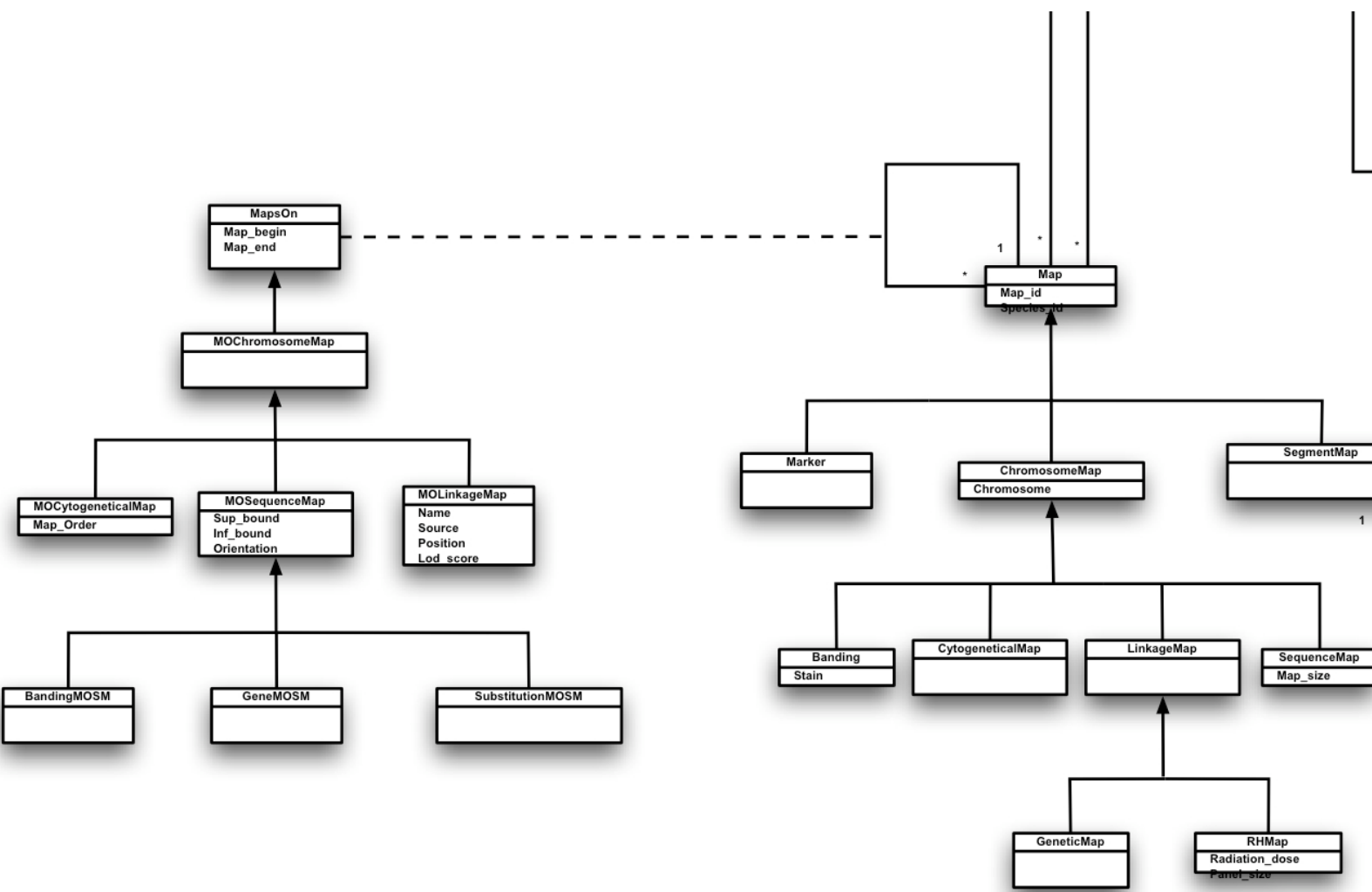


FIG. 13 –

7.2 hybride d'irradiation

Sans revenir précisément sur les aspects expérimentaux nous allons considérer en premier le cas où un chromosome (haploïde) humain est fusionné avec un génome de rongeur. Les cassures générées par l'irradiation seront supposées suivre un processus Poissonnien de paramètre λ .

Si on suppose un taux d'irradiation correspondant en moyenne à 4 points de cassure sur le chromosome 21 (données obtenues par Cox avec une irradiation de 8000 rads [*Cox DR et al (1990), Science 250 :245-250*]) dont la longueur est de l'ordre de $4 \cdot 10^4$, on obtient $\lambda = \frac{4}{4 \cdot 10^4} = 10^{-4}$.

Soit θ la probabilité d'au moins un point de cassure entre deux loci séparés d'une distance (physique) de δ on a évidemment :

$$1 - \theta = e^{-\lambda\delta}$$

8 TP

L'objectif est de relier le taux de recombinaison au banding et au taux de CG chez l'homme. Le plan général est :

1. charger et installer le package R MareyMap à partir du site : <http://pbil.univ-lyon1.fr/rezvoy/mareymap/>
2. regarder la documentation pour avoir une vue intuitive des méthodes d'estimation du taux de recombinaison.
3. choisir un chromosome et chercher des régions à fort ou faible taux de recombinaison
4. aller sur le site d'Ensembl pour repérer la zone choisie, y trouver des gènes, extraire la séquences nucléotidique de leur partie codante
5. calculer CG_3