Tableau 2 : Nombres de paralogues et taille des familles multigéniques dans plusieurs génomes entièrement séquencés.

régne	espèce	souche	nbr chromosomes°	nbr d'épisomes	taille du génome	nbr gènes codants	taille moyenne gènes	% d'ADN codant	GC %	nbr de paralogues	% paralogues	nbr de famille	nbr max de membres	références
archea	*Methanobacterium thermoautotrophicum*	ΔH	1c	—	1 751 377pb	1 855	—	92%	50%	409	22%	111	24	Douglas *et al.* , 1997
archea	*Archaeoglobus fulgidus*	VC-16	1c	—	2 178 400pb	2 436	822pb	92%	49%	719	30%	242	40	Klenk *et al.* , 1997
bacteria	*Mycoplasma pneumoniae*	M129	1?	—	816 394pb	677	1 011pb	89%	40%	298	44%	58	—	Himmelreich *et al.* , 1996
bacteria	*Chlamydia trachomatis*	—	1?	1	1 042 519pb	894	—	—	41%	256	29%	58	—	Stephens *et al.* , 1998
bacteria	*Ricettsia prowazekii*	Madrid E	1c	—	1 111 523pb	834	1 005pb	75%	29%	147	18%	54	—	Andersson *et al.* , 1998
bacteria	*Treponema pallidum*	Nichols	1c	—	1 138 006pb	1 041	1 023pb	93%	53%	129	12%	42	14	Fraser *et al.* , 1998
bacteria	*Borrelia burgdoferi*	B31	1ℓ		910 725pb	853	992pb	93%	29%	—	—	—	—	Fraser *et al.* , 1997
		B31		9ℓ+2c*	533 000pb	430	507pb	71%	23-32%	169	39%	47	12	
bacteria	*Helicobacter pylori*	J99	1c	—	1 643 831pb	1 495	998pb	91%	39%	337	23%	113	—	Alm *et al.* , 1999
bacteria	*Helicobacter pylori*	26695	1c	—	1 667 867pb	1 590	945pb	91%	39%	266	16%	95	32	Tomb *et al.* , 1997
bacteria	*Heamofilus influenze*	Rd KW20	1c	—	1 830 137pb	1 680	900pb	85%	38%	700	40%	208	43	Fleischmann *et al.* , 1995; Brenner *et al.* , 1995
bacteria	*Thermotoga maritima*	MSB8	1c	—	1 860 725pb	1 877	947pb	95%	46%	—	—	214	67	Nelson *et al.* , 1999
bacteria	*Neisseria meningitidis*	MC58	1c	—	2 272 351pb	2 158	874pb	83%	52%	678	32%	234	24	Tettelin *et al.* , 2000
bacteria	*Deinococcus radiodurans*	R1	2c	2c	3 284 156pb	3 187	937pb	91%	67%	1 665	52%	95	120	White *et al.* , 1999
bacteria	*Mycobacterium tuberculosis*	H37Rv	1c	—	4 411 529pb	3 924	—	91%	66%	—	51%	—	—	Cole *et al.* , 1998
bacteria	*Bacillus subtilis*	—	1c	—	4 214 810pb	4 100	890pb	87%	44%	1 927	47%	—	77	Kunst *et al.* , 1997
bacteria	*Escherichia coli*	K-12	1c	—	4 639 221pb	4 288	951pb	88%	51%	1 345	31%	—	80	Blattner *et al.* , 1997
eukaryote	*Leishmania major*	Friedlin	chr1	—	268 984pb	79	—	—	—	13	16%	54?	5	Myler *et al.* , 1999
eukaryote	*Saccharomyces cerevisiae*	—	16ℓ	—	12 068 000pb	6 241	—	70%	—	1 858	30%	—	—	Gofeau *et al.* , 1996
eukaryote	*Caenorhabditis elegans*	—	6 paires†	—	97 000 000pb	18 424	—	27%	36%	8 971	49%	—	—	*C. elegans* Sequencing Consortium, 1998
eukaryote	*Drosophila melanogaster*	—	4 paires†	—	120 000 000pb §	13 601 #	3 058pb	20%	—	5 536	41%	—	—	Adams *et al.* , 2000

° c : ADN circulaire, ℓ : ADN linéaire * 11 épisomes sur les 17 présents dans cette souche ont été séquencés § 120Mb d'euchromatine pour 180Mb d'ADN total # codant pour 14113 transcripts † dont une paires de chromosomes sexuels

b) Exemples :

Exempli gratia, suivant ces cinq lemmes, on qualifiera, ainsi, les relations phylogénétiques de la β_2-microglobuline humaine et du récepteur Fcγ I humain d'allologues et celles des récepteurs Fcγ I et II d'isologues (Figure 9) ; les trois gènes étant des paralogues. De manière symétrique, cette assertion est valable pour les équivalents murins des gènes respectivement précités. De manière plus communément admise, les relations phylogénétiques des β_2-microglobulines humaine et murines, des récepteurs Fcγ I humain et murin et des récepteurs Fcγ II humain et murin sont qualifiées d'orthologues. Le reste des relations établies par le réseau est, par défaut, qualifié d'homologue.

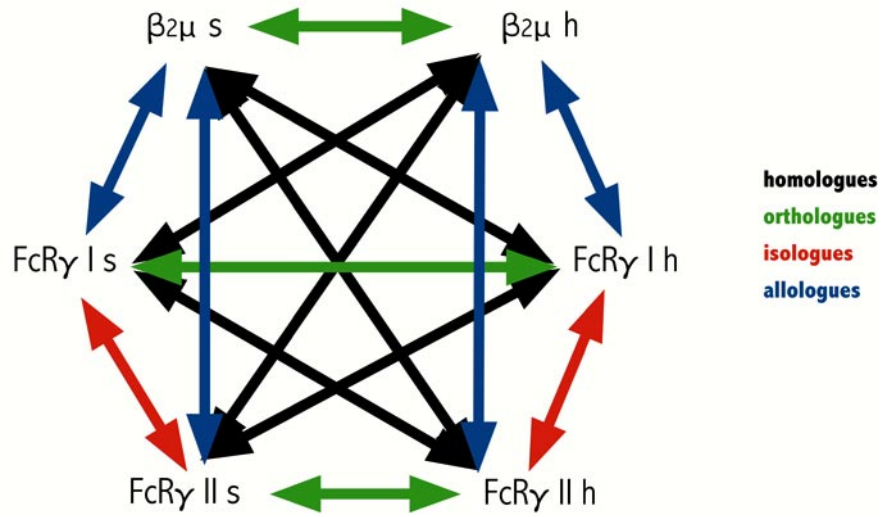

Figure 9 : Réseau de relations phylogénétiques au sein de la super famille des Ig.
β_2-microglobuline humaine et murine ($\beta_2\mu$ h $\beta_2\mu$ s, respectivement) ; Récepteurs Fcγ I et II humain et murins (FcRγ I h, FcRγ II h, FcRγ I s, FcRγ II s, respectivement).

c) Remarques importantes :

Dans un premier temps, l'établissement d'une relation d'homologie entre deux gènes n'est pas une observation mais une hypothèse sur leur évolution qui, à ce titre, doit être évaluée dans un contexte phylogénétique (Thornton & DeSalle, 2000). D'autre part, l'homologie est un caractère qualitatif et, en tant que tel, il ne peut être divisé. Il est, par conséquent, absurde de parler de pourcentage d'homologie : c'est la similitude qui quantifie le résultat de la comparaison de deux séquences (Reeck *et alii*, 1987).

Par ailleurs, il est, nonobstant, extrêmement important de noter que les définitions d'« allologues » et d'« isologues » sont contingentes de la stringence de la définition que l'on donne du terme « fonction » ou de la fonction à laquelle on fait référence dans le cas de famille de protéines dont les membres sont multifonctionnels et dont une seule fonction leur est commune. Tout cela dépend de la problématique posée et du type d'inférences qui en découle (*exempli gratia confer* § Pfemp1, allologie et isologie : une question de point de vue. page 40).

2) Notions de familles :

Ces collections de gènes sont réunies en famille et leurs relations sont hiérarchisées en sous-famille et superfamille ; créant ainsi une taxonomie moléculaire. Il est important de noter, dès maintenant, que ces notions de familles s'appliquent autant aux homologues qu'aux paralogues et aux orthologues, de même qu'aux allologues et aux isologues. Ainsi, selon le type d'inférence envisagée (relations d'une collection de gènes au sein d'un organisme, entre organismes ou recherche d'un ancêtre commun), les notions de famille prennent en compte des homologues, des orthologues, ou des paralogues (allologues ou isologues).

La locution « famille de gène », dans la litérature, n'est que trop souvent improprement usitée. On prendra, donc, garde à parler de **famille de gènes** lorsqu'on fait référence à des homologues ou à des orthologues (collections de gènes d'espèces différentes) et à parler de **famille multigénique** lorsqu'on fera référence

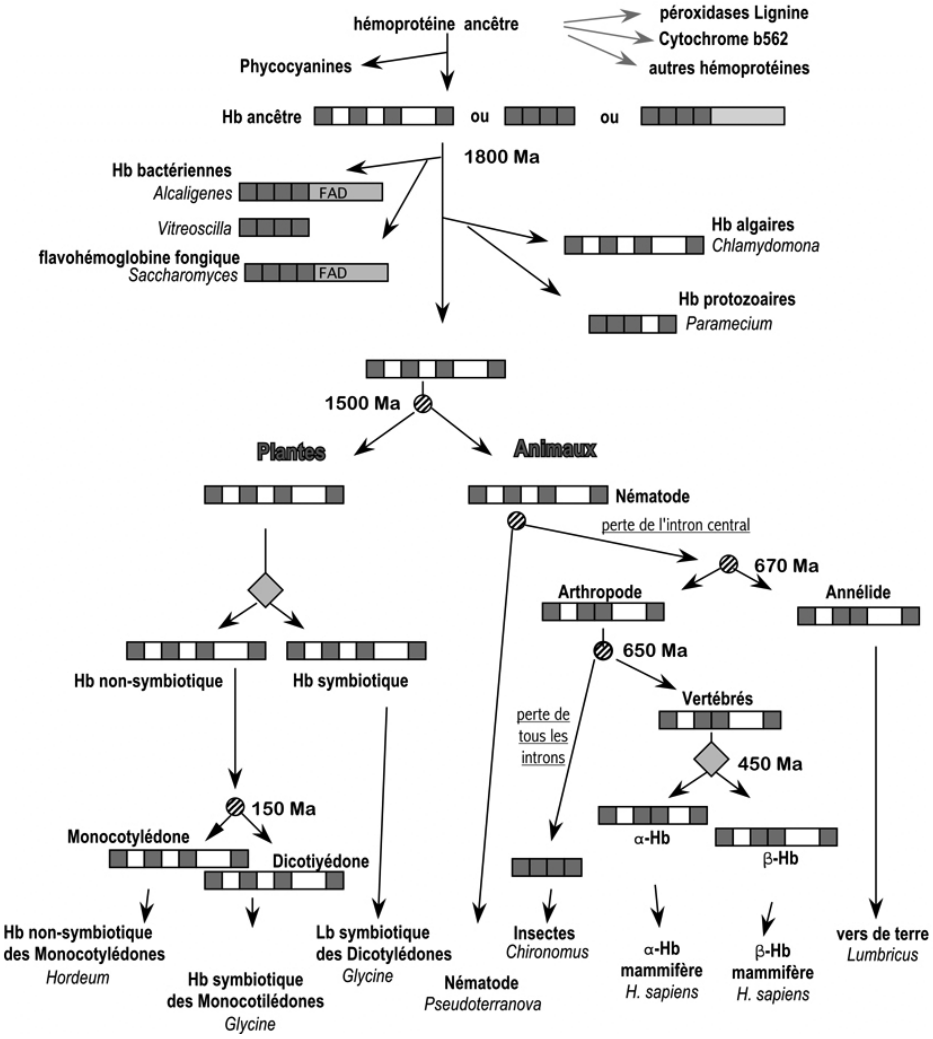

Figure 10 :Représentation schématique de l'évolution du gène hémoglobine de la bactérie à l'Homme (d'après Hardison, 1998).
Les exons sont schématisés en gris et les introns en blanc. Les événements de spéciation sont symbolisés par des cercles et les événements de duplication par des losanges. Ma : Million d'années.

à des paralogues, tant allologues qu'isologues (collection de gènes au sein d'une même espèce). Il est important de noter qu'une famille multigénique peut elle-même appartenir à une famille de gène homologue, comme la famille multigénique des globines humaines fait partie de la famille des gènes des globines.

Le fait de regrouper des gènes en famille implique toujours une notion d'évolution sous-jacente ; l'analyse d'homologues peut potentiellement faire remonter l'ancêtre commun au fameux LUCA, la comparaison d'orthologues permet d'analyser des événements de spéciation, celle de paralogues d'analyser la dynamique du génome après un événement de spéciation (Figure 10 : comparer les relations allologiques de l'α-Hb et de la β-Hb humaines à la relation d'homologie de α-Hb humaine et des flavohémoglobines fongiques) et celle d'isologues d'analyser la dynamique du génome sur quelques générations. Ainsi, comme en astronomie, plus on « regarde loin » (*id est* plus c'est « différent »), plus on observe des événements éloignés dans le temps.

3) Pseudogènes et familles multigéniques:

Un pseudogène est un gène non-exprimé:

- soit parce qu'il est non-transcript ; il est alors soit :
    a. défectif dans sa région régulatrice.
    b. défectif dans sa région promotrice.
        a. Pas d'ARNm.
- soit parce qu'il est non-traduit ; il est alors soit :
    c. défectif en signaux d'initiation de la traduction.
    d. soit sa traduction est stoppée prématurément (comme le gène resa2 de P. falciparum, Cappai *et alii*, 1992).
        b. Pas de protéine.

Cependant, ils possèdent, en général, un équivalent "actif" dans le génome, dont ils dérivent par un ancêtre commun. Ils peuvent, donc, entretenir des relations paralogiques avec d'autres gènes. En conséquence de quoi, on prendra garde à ne pas oublier que les pseudogènes sont des membres à part entière des familles multigéniques.

4) Unités de séquence des familles de protéines :

Le critère utilisé pour l'établissement des relations de parenté entre ces gènes est la similitude. Celui-ci peut être appliqué à tout ou partie de la séquence considérée — nucléique ou protéique — et permet, en ce qui concerne les gènes codant pour des protéines, de définir différents niveaux de structuration des protéines : les motifs , les modules et les domaines. Nécessairement, le fait que le critère de similitude ne puisse concerner qu'une portion des séquences considérées implique la non-transitivité d'une relation phylogénétique (pour exemple, voir Figure 45 : Réseau de relations d'homologie au sein des familles multigéniques DLC et DBL de différents *Plasmodium*. page 97).

a) Les motifs :

Le motif est la plus petite unité de séquence des familles de protéines ; il est défini comme une région de forte similitude dans des alignements de segments de protéines. Il peut être très simple, comme l'hexamère répété qui forme l'hélice β parallèle gauche de l'Uridine 5'-diphosphate-*N*-acetylglucosamine acyltransférase (Raetz & Roderick, 1995). Les motifs sont très utilisés pour identifier des régions fonctionnelles dans les séquence protéiques et, quand ils partagent un ancêtre commun, ils sont très utiles pour classifier les familles.

b) Les modules et les domaines:

Le motif « *C_2H_2 zinc finger* » qui lie l'ADN, définit la plus grande famille de protéines connue (Henikoff *et alii*, 1997). Comme il forme une structure contiguë indépendamment repliée, il est lui-même un module, dont la petite taille de 21 à 26 acides aminée est attribuable au cation Zn^{2+}, qui grâce à ces liaisons de coordination avec 2 résidus cystéine et 2 résidus histidine, confère sa structure au module. De manière plus générale, les modules consistent en de multiples motifs, qui forment le corps de la structure de la protéine. Les motifs contribuant à un module structurel peuvent être distants dans la séquence primaire, comme les motifs « HIGH » et « KMSKS » des ARNt aminoacyl synthase de classe I, où ils sont séparés par une centaine de résidus (Schimmel, 1991).

Les modules sont obligatoirement contigus dans la séquence, alors que les domaines structuraux sont des unités fonctionnelles indépendantes, de conformations spatiales fonctionnelles, qui n'ont pas besoin d'être contigus (Patthy, 1985; Henikoff *et alii*, 1990; Green *et alii*, 1993; Doolittle, 1995). Les modules sont constitués d'un ou plusieurs motifs. Puisqu'ils sont les unités structurelles et fonctionnelles fondamentales, ils sont plus pertinents dans la classification des protéines.

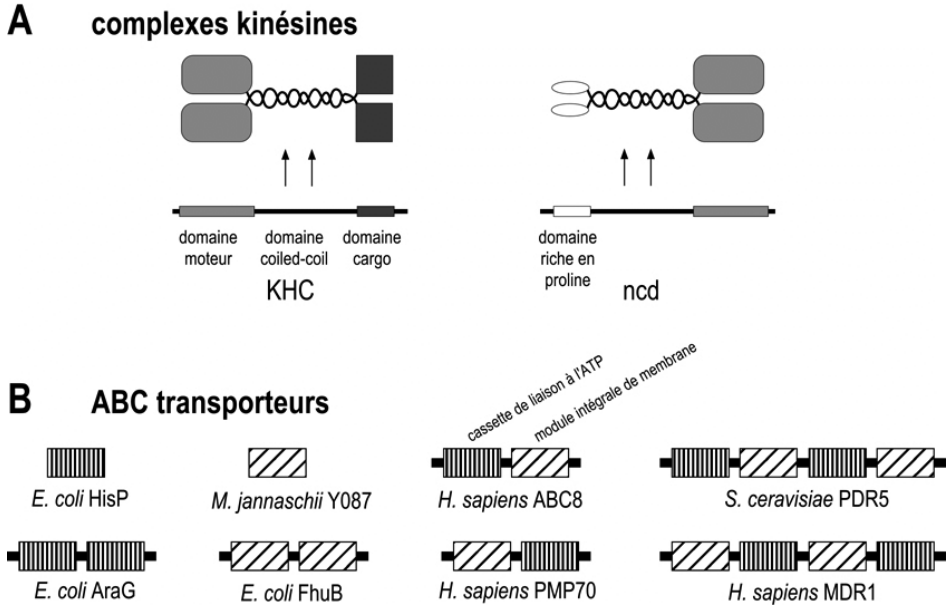

Figure 11 : Représentations schématiques des différents modules et domaines des complexes kinésines (A) et des ABC transporteurs (B).

Souvent, les domaines définissant une même famille présentent différents types de connectivité, comme les kinésines et les ABC transporteurs. Le domaine moteur des kinésines peut, ainsi, être placé à l'une où l'autre des deux extrémités de la chaîne polypeptidique qui est constituée d'un domaine « *coiled-coil* » et d'un domaine globulaire (Figure 11 A) ; la place du domaine moteur, en N- ou C-terminal, déterminant le sens de déplacement sur le microtubule (Sablin, 2000). Les ABC transporteurs sont, quant à eux, constitués de 2 types de modules : la cassette de liaison de l'ATP et le module intégral de membrane qui peuvent être connectés de différentes manières (Figure 11 B ; (Dean & Allikmets, 1995))

c) Chimèrisme :

Il n'est pas rare de trouver deux modules fonctionnels, ordinairement indépendants, fusionnés, aboutissant à la création d'une protéine chimère. De nombreuses enzymes biosynthétiques eucaryotes sont sous forme d'un seul polypeptide, alors que leurs équivalents bactériens sont codés séparément par des orthologues. Cela est très bien illustré par les enzymes de la voie de biosynthèse des purines et des pyrimidines : notamment la Glycinamide ribonucléotide synthase (GARS), l'Aminoimidazole ribonucléotide synthase (AIRS) et la Glycinamide ribonucléotide transformylase (GART).

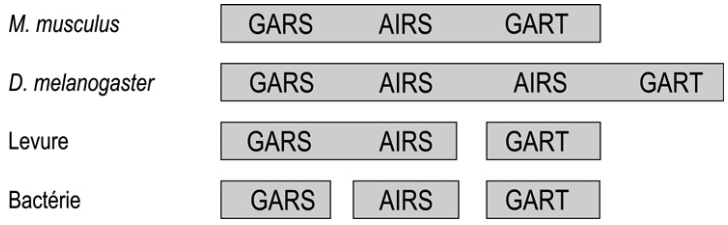

Figure 12 : Organisation des gènes et des domaines de GARS, AIRS et GART chez différentes espèces (d'après Davidson & Peterson, 1997).

Chez *E. coli*, AIRS et GART sont codées par des cistrons différents (*PurM* et *PurN* respectivement) du même opéron (Smith & Daum, 1986; Smith & Daum, 1987); et GARS (*Pur D*) est codé par un gène à un autre locus (Aiba & Mizobuchi, 1989; Cheng *et alii*, 1990). Chez *S. cerevisiae* (Henikoff, 1986) et *S. pompe* (Fluri *et alii*, 1976) GARS et AIRS font partie d'une enzyme bifonctionnelle codée par le gène *Ade5,7* et GART fait partie d'une protéine monoenzymatique codée par le gène *Ade8*. Chez *D. melanogaster* (Henikoff *et alii*, 1986; Henikoff & Eghtedarzadeh, 1987), le poulet (Aimi *et alii*, 1990), *Mus musculus* (Kan *et alii*, 1993) et l'espèce humaine (Aimi *et alii*, 1990; Schild *et alii*, 1990), les trois fonctions sont des domaines d'une protéine trienzymatique, quand un transcrit complet est synthétisé à partir du gène *gart* . Le gène de la drosophile, à la différence des autres, possède une duplication interne du domaine AIRS (Figure 12). Chez les parasites, tel que *T. gondii* et *P. falciparum*, la dihydrofolate réductase est fusionnée avec la thymidine synthase, afin de former une enzyme bifonctionnelle, contrairement aux mamifères.

Certains membres de familles de paralogues sont eux-mêmes des chimères, à l'exemple des ABC transporteurs, composés d'un domaine de liaison de l'ATP et d'un

module intégral de membrane (Figure 11B) ; complexifiant, ainsi, les relations paralogiques entre les différentes familles de gènes impliquées.

C.Mécanismes de genèse des familles multigéniques :

Deux types de mécanismes concourent, de manière séquentielle, à la mise en place de famille multigéniques: la duplication de gènes et, ensuite, les processus évolutifs. La duplication de gène est l'événement primordial et essentiel à partir duquel un gène ancestral donne naissance à deux copies, dont l'évolution individuelle et l'expansion ultérieure et successive dans le génome concourent à la formation d'une famille multigénique.

Les mécanismes de duplications impliqués sont divers et multiples ; de plus, la séquence des événements est, la plupart du temps, impossible à déterminer. Tous les mécanismes qui contribuent à l'évolution des génomes peuvent être impliqués, comme la recombinaison homologue ectopique entre séquences répétées, la translocation, la rétrotranscription et l'intégration d'ARNm et les événements, rares mais avérés, de tétraploïdisation.

1) Mécanismes de duplication de gène:

a) Recombinaison homologue ectopique :

La recombinaison homologue ectopique entre des éléments répétés de part et d'autre d'un gène est très certainement l'événement génétique intervenant le plus fréquemment dans la duplication de gènes. C'est ainsi qu'on explique la duplication du gène γ du regroupement de loci β-Hb dans la lignée simienne (Figure 13). Les deux gènes γ-Hb des catarrhiniens[14] et des platyrrhiniens[15] sont encadrés par trois éléments

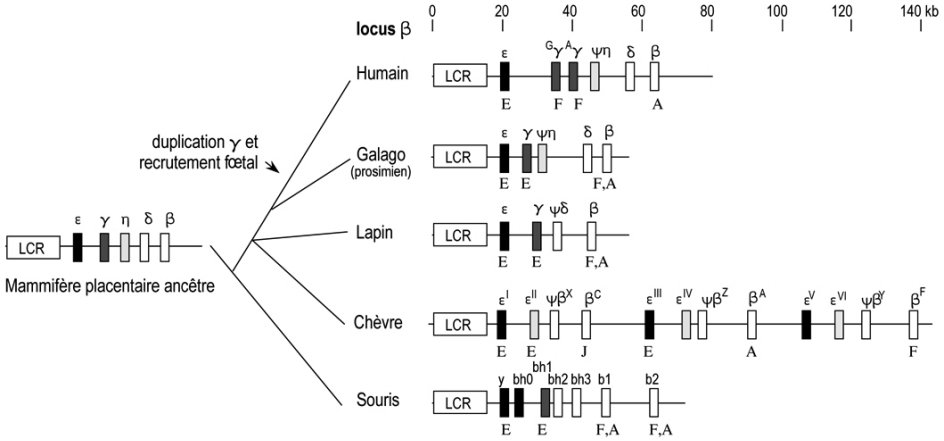

Figure 13 : Représentation schématique de l'évolution du regroupement de gène des β-Hémoglobines chez les euthériens.
Les gènes sont encadrés, leur nom indiqué au dessus et le profil d'expression temporel en dessous (E : embryonnaire ; F : fœtal ; A : adulte ; J : juvénile). Les gènes orthologues sont représentés de la même couleur. Galago : *confer* Annexe B, page 117.

[14] Singes de l'Ancien Monde.
[15] Singes du Nouveau Monde.

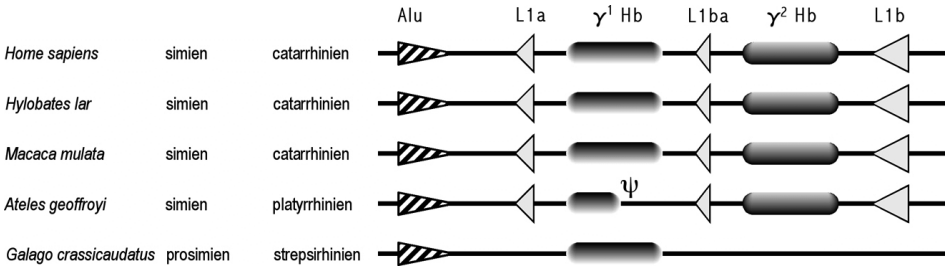

Figure 14 : Représentation schématique des loci γ-Hb et de leur contexte génétique chez différents primates.
Hyalobates lar : gibbon ; *Macaca mulata* : rhésus ; *Ateles geoffroyi* : singe araigné ; ψ : pseudogène, Galago : *confer* Annexe B, page 117.

L1 (de type « long interspersed repetive element », Fitch *et alii*, 1991), Figure 14.

L'élément L1ba, séparant les loci γ^1 et γ^2, est en partie homologue à L1b et en partie à L1a (Fitch *et alii*, 1991); suggérant fortement que : ① la duplication γ est due a une recombinaison homologue ectopique entre un élément L1b et un élément L1a (Figure 15) ; et ② l'ancêtre commun des simiens possédait un gène γ-Hb encadré par deux éléments L1 dont l'insertion est postérieure à la divergence des simiens des prosimiens (≈ 50 millions d'années).

La duplication de gène est, en général, suivie, par des événements de recombinaison additionnels. A partir d'un certain nombre de copies, les membres de familles multigéniques constituent des éléments répétés du génome et peuvent promouvoir, par eux-mêmes, des événements de duplication/délétion par ce mécanisme de recombinaison homologue ectopique. C'est ainsi qu'on peut expliquer l'amplification des membres du regroupement en tandem des 8 gènes paralogues *sera* (PfB0325c à PfB0355c) identifié sur le chromosome 2 (Gardner *et alii*, 1998).

Il est important de noter que la recombinaison homologue ectopique est à la fois un moyen de duplication et, à la fois, un mécanisme de délétion; suivant la chromatide

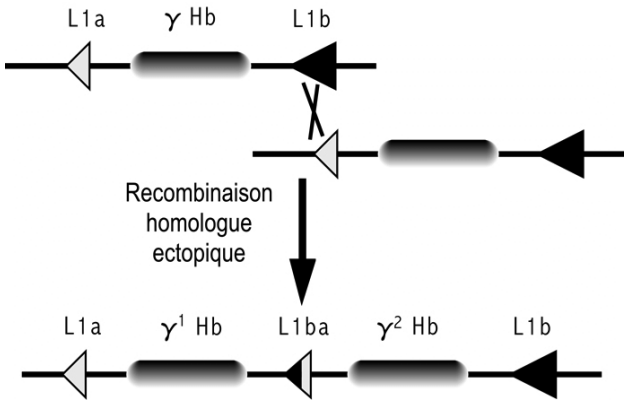

Figure 15 : Représentation schématique de l'événement de recombinaison non-homologue entre les éléments L1a et L1b ancestraux, à l'origine de la duplication du locus γ-Hb chez l'ancêtre des simiens.

retenue par les processus évolutifs, c'est soit la duplication soit la délétion qui sera fixée dans l'espèce considérée.

b) Conversion génique :

La conversion génique est définie comme un transfert non-réciproque d'information génétique entre deux gènes possédant un important degré de similitude (Radding, 1978). Elle permet, au sein d'une famille, la duplication de petites portions d'ADN, de quelques paires de bases à quelques centaines de paires de bases. C'est un élément important dans la dynamique des familles multigéniques. Des exemples seront développés ci-dessous, dans le cadre de l'évolution concertée (*confere* § 3ème alternative: (non-)divergence des copies. page 27) et de l'établissement de manière pérenne de pseudogènes (*confere* § 2ème alternative: pseudogène / gène fonctionnel. page 27).

c) Transcription inverse d'ARNm et intégration:

De nombreux pseudogènes sont caractérisés par l'absence de promoteur et d'introns et la présence d'une séquence poly-A en 3', ils sont appelés retropseudogènes. Ils résultent de la rétrotranscription d'ARNm puis de son intégration dans le génome (Mighell *et alii*, 2000). L'interprétation initiale des données de séquence du chromosome 22 humain indique que 19% des séquences codantes sont des pseudogènes dont 82% sont des rétropseudogènes (Dunham *et alii*, 1999). En général, ce processus n'aboutit qu'à une copie dupliquée sous forme non-fonctionnelle. En effet, de par la nature même du mécanisme, seule la séquence codante est dupliquée. Cependant, dans certains cas, l'intégration peut se produire, dans le sens adéquat, en amont d'un promoteur ou en aval d'un promoteur bidirectionnel.

Ainsi, on dénombre, chez la souris, trois gènes codant pour la S-adénosylméthionine décarboxylase. Le gène *Amd-1* possède des introns et code une protéine de 334 acides aminés, alors que *Amd-2* présente toutes les caractéristiques d'un rétropseudogène, mais permet l'expression d'une protéine qui diffère par 2 acides aminés de AMD-1 (Nishimura *et alii*, 1998). Cependant leur promoteur respectif ne possède aucune similitude. Le gène *Amd-3*, quant à lui, est un « véritable» rétropseudogène sans promoteur. Chez *P. falciparum*, le gène *resa2*, dont la phase codante est interompue par un codon stop (Cappai *et alii*, 1992), est très certainement issu d'une intégration d'un messager : en effet, contrairement à *resa*, il ne possède pas d'intron. Il est cependant transcrit (Vazeux *et alii*, 1993).

d) Tétraploïdisation :

Ohno, en 1970, avait proposé que la duplication de génome entier, provenant d'une erreur à la méiose, puisse être un mécanisme important de l'évolution. L'analyse de l'arrangement des gènes dupliqués du génome de *S. cerevisiae* a mis en évidence la présence de 84 paires de régions contenant un minimum de trois gènes dupliqués, représentant 905 gènes, soit 16% du protéome (Wolfe & Shields, 1997; Seoighe &

Wolfe, 1999). La levure serait donc un tétraploïde dégénéré, dont seuls 16% des gènes dupliqués ont été retenus au cours de l'évolution.

Les vertébrés auraient, quant à eux, subi deux cycles de tétraploïdisation. Leur protéome est en moyenne 4 fois plus important que celui des invertébrés. De plus, de nombreux gènes de vertébrés sont présents sous forme de 2, 3 ou 4 paralogues tandis que leurs orthologues invertébrés sont uniques ; cette constatation est connue sous le nom de « *one-to-four rule* » (Ohno, 1970; Ohno, 1999).

2) Processus évolutifs:

Après duplication d'un gène, les processus évolutifs et la sélection naturelle, qui agissent sur les deux copies, peuvent conduirent à plusieurs *scenarii* que l'ont peut formaliser de manière conceptuelle et consécutive en 4 étapes dichotomiques:

- (In-)stabilité de la duplication.
- Fonctionnalité de la copie: pseudogène/gène fonctionnel.
- (Non-)divergence des copies.
- Type de séquence évoluant: région régulatrice non-codante/séquence codante.

Il faut, cependant, ne pas perdre de vue que cette description, tout en étant formelle, est aussi — afin de clarifier le raisonnement — volontairement formaliste et statique. Il apparaîtra, donc, bien évident au lecteur, qu'il faut la replacer dans le contexte dynamique des processus évolutifs et qu'à chaque instant ces quatre alternatives se posent simultanément et ne sont pas, nécessairement, irréversibles, ni exclusives.

a) 1ère alternative: (in-)stabilité de la duplication.

La duplication peut avoir plusieurs conséquences, qui en terme d'évolution conduisent soit à la non-transmission du génotype, soit à l'éviction du génotype de la population des génotypes de l'espèce, soit à la délétion d'une des copies, soit au maintien des 2 copies. La résultante de cette première étape est, donc, soit la perte d'une des copies, soit le maintien des deux copies.

Si la duplication confère un phénotype létal (*fitness* = 0), alors la mort de l'organisme conduit à la non-transmission du génotype. Si elle est délétère ou entraine une baisse de la *fitness*, il y aura soit non-transmission du génotype, soit éviction rapide (en quelques générations) du génotype de la population, soit délétion pure et simple afin de retrouver un équilibre dynamique de la *fitness* par rapport à la *fitness* moyenne des génotypes de la population considérée. Dans le cas d'un gain de *fitness*, les deux copies seront maintenues par la sélection naturelle. Enfin, quant au cas où il n'y a ni gain ni diminution de *fitness*, en l'absence de pression (positive ou négative) de la part de la sélection naturelle, seul le hasard déterminera la perte d'une des copies ou le maintien des 2 copies. Si deux duplicates sont évolutivement stables, alors se pose la question de la fonctionnalité des deux copies.

b) 2ème alternative: pseudogène / gène fonctionnel.

D'aucuns pourront s'étonner de la dichotomie introduite par cette seconde alternative : celle-ci pouvant paraître redondante par rapport à la première. Cependant, il y a une différence entre ne pas maintenir une seconde copie dupliquée et maintenir une deuxième copie dupliquée sous forme non-fonctionnelle. Par ailleurs, cette seconde alternative permet de rendre compte de la création des nombreux pseudogènes rencontrés dans certaines familles multigéniques (comme Pf60/*var*) et d'évoquer les "rôles biologiques" possibles des pseudogènes. En effet, il n'est que trop souvent admis que les pseudogènes sont **uniquement** un "cul-de-sac" ou la "poubelle" de l'évolution.

De facto, la similitude entre un pseudogène et son paralogue fonctionnel peut promouvoir, *per se*, des réarrangements génomiques et la conversion génique ; et ce d'autant plus facilement et fréquemment qu'ils sont nombreux et/ou proches de leur pendants fonctionnels. D'une part, ils concourent, donc, à la dynamique du génome et, d'autre part, ils peuvent être une réserve de diversité génétique. Le rôle des pseudogènes dans la génération de diversité du domaine variable de la chaîne lourde (V_H) et légère (V_L) des immunoglobulines des aviaires, au cours de l'ontogénie B, en est l'illustration la plus frappante. Contrairement aux mammifères, qui possèdent et réarrangent différents gènes V_H et V_L, les oiseaux ne possèdent qu'une seule copie fonctionnelle de chaque gène V_H et V_L. La diversité des domaines variables est, alors, assurée par conversion génique du gène fonctionnel par un des nombreux et différents pseudogènes paralogues aux gènes V_H et V_L (McCormack *et alii*, 1993).

c) 3ème alternative: (non-)divergence des copies.

Dès lors que les deux gènes dupliqués sont maintenus sous forme fonctionnelle, les deux séquences peuvent diverger ou subir une **évolution concertée**. La conversion génique et les « *crossing over* » inégaux sont les deux principaux mécanismes générant une évolution concertée (pour revue Li, 1997). La conversion génique est impliquée dans l'homogénéisation de petites portions d'ADN, de quelques paires de bases à quelques centaines de paires de bases ; tandis que l'homogénéisation de très grandes régions d'ADN impliquerait, quant à elle, les « *crossing over* » inégaux.

Dans le cas du locus γ-Hb des catyrrhiniens, précédemment évoqué (Figure 14, page 24, et Figure 15, page 24), une régison d'environ 1,5 kb, comprenant la région régulatrice et la partie 5' de la région codante du gène de la Gγ-globine, est quasiment identique à la région correspondante du gène de la Aγ-globine (Shen *et alii*, 1981) ; indiquant qu'un événement de conversion génique eu lieu, il y a environ 1Ma, probablement induit par l'élément répété γ-$(TG)_n$ (Kilpatrick *et alii*, 1984). L'inactivation du gène γ1-globine des platyrrhiniens laisse supposer l'importance d'un mécanisme de correction spécifique de séquence dans la fixation de gènes dupliqués.

Tableau 3 : La famille multigénique des filaments intermédiaires.

Membre	Classe de similitude de séquence	Masse Moléculaire (kDa)	Localisation chez les mammifères
Groupe d'assemblage 1			***Cytoplasme***
Cytokératines acides (CK9-20)	I	40-64	tous les épithéliums
Cytokératines basiques (CK1-8)	II	52-68	tous les épithéliums
Groupe d'assemblage 2			***Cytoplasme***
Vimentine	III	55	cellules mésenchymateuses
Desmine	III	53	cellules musculaires
Glial fibrillary acidic protein (GFAP)	III	50-52	cellules gliales, astrocytes, cellules stellaires du foie
Périphérine	III	54	différentes cellules neuronales
Synémine	III / IV	182	cellules musculaires
Paranécine	IV / I	178	cellules musculaires
Nestin	III / IV	240	cellules souches neuroépithéliales, muscles
α-Internexine	IV	56	neurones
Neurofilaments triplets proteins	IV	—	neurones
NF-L	—	68	neurones
NF-M	—	110	neurones
NF-H	—	130	neurones
Groupe d'assemblage 3			***Noyau***
Lamines types A/C	V	62-72	la plupart des cellules différenciées
Lamines types B	V	65-68	tous les types cellulaires

De manière plus globale, sur l'ensemble des gènes de *C. elegans* examinés, 40% sont des paralogues, dont 2% des paires de copies paraissent avoir subi un événement de conversion génique (Semple & Wolfe, 1999).

d) 4ème alternative: divergence des séquences régulatrices non-codante/séquence codante.

Enfin, dans le cas où la duplication est stable, où le gène dupliqué est fonctionnel et où les deux copies ne subissent pas de processus d'évolution concertée, les deux gènes dupliqués vont diverger. Or, un gène est constitué de deux entités dont la nature informative est distincte : une région régulatrice et une région codante. La divergence peut avoir lieu dans l'une ou l'autre de ces entités : les conséquences en étant différentes. Dans un cas, l'une des copies pourra avoir l'occasion d'acquérir une nouvelle fonction ; dans l'autre cas, le gène concerné pourra acquérir une nouvelle régulation. Bien évidemment les deux régions peuvent diverger, combinant, ainsi, leurs effets.

i. Acquisition d'une nouvelle fonction :

De nombreux exemples, connus de tous, peuvent illustrer ce cas. On se contentera, donc, de revenir sur l'exemple de la super-famille des Ig dont font partie la β_2-microglobuline, qui possède un rôle structural, et le récepteur Fcγ I, dont la fonction est de lier les IgG.

ii. Acquisition d'une nouvelle régulation :

La régulation d'un gène peut être de deux ordres : topologique [spécifique de tissu(s)] ou chronologique (spécifique d'une phase de développement). Dans le premier cas, l'un des deux gènes dupliqués sera exprimé en même temps que l'autre,

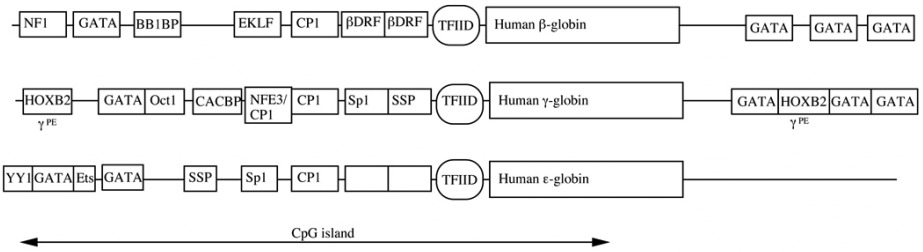

Figure 16 : représentation schématique des sites de liaisons de protéines des promoteurs des gènes du locus β-Hb humain.

mais dans un autre type cellulaire. Dans le second, ils ne seront pas exprimés simultanément, soit par la même cellule, soit par un autre tissu.

α. Nouvelle régulation topologique:

Les filaments intermédiaires forment une famille multigénique et sont des constituants du cytosquelette cellulaire, exprimés dans tous les types cellulaires. Cependant, ils ne sont pas tous ubiquitaires, loin s'en faut, comme le montre le Tableau 3.

β. Nouvelle régulation chronologique:

Un bel exemple de régulation chronologique, nous est, encore, fourni par les paralogues du locus β-Hb. Les gènes ε, γ et β sont des isologues , le premier est exprimé au stade embryonnaire dans le sac vitellin chez l'embryon, le second au stade fœtal par le foie et la rate et le troisième après la naissance par la moelle osseuse (*confere* Figure 13, page 23). Leurs séquences régulatrices sont, en effet, très différentes (Figure 16).

3) Conclusions:

La création d'une famille multigénique confère 3 grands avantages sélectifs et adaptatifs. C'est, tout d'abord, un mécanisme rapide (à l'échelle de l'évolution) et peu «coûteux » (toujours en termes d'évolution) de création de nouvelles fonctions. Deuxièmement, c'est un moyen de mettre en place des profils d'expression temporels et/ou topologiques différents, conférant une meilleure adaptation à différents stades de développement et/ou une régulation plus adaptée. D'autre part, des isologues peuvent constituer un moyen d'adaptation ultrarapide (sur une échelle de temps inférieure à la génération cellulaire ou à celle de l'organisme).

Enfin, il est important de remarquer que les propriétés des familles multigéniques, qui rendent compte de leur utilisation extensive dans les génomes (Tableau 2, page 17), découlent directement de leurs mécanismes de genèse. Ainsi, c'est en expliquant le comment qu'on trouve le pourquoi.

III. PROTEINES A DOMAINES DBL : UNE SUPERFAMILLE DE P. FALCIPARUM:

De nombreuses familles de gènes et de familles multigéniques ont été décrites chez les parasites du genre *Plasmodium*, comme la famille multigénique des gènes *SICAvar* codant pour l'antigène variant de *P. knowlesi* (al-Khedery *et alii*, 1999) ou la famille multigénique codant pour la protéine de rhoptrie p235 de *P. yoelii* (soumise à une variation phénotypique clonale, Preiser *et alii*, 1999). L'objet de la présente introduction n'est pas d'en faire un catalogue exhaustif ; nonobstant, on s'attardera sur une superfamille des *Plasmodium*, connexe à la famille multigénique Pf60/*var* de *P. falciparum* : la famille des protéines à domaines DBL.

Tableau 4 : Les membres de la super famille des protéines à « Duffy Binding Like » Dommaine.

Nom	espéce	nbr de copies	MM (kDa)	Nbr Dom.	localisation	fonction	références
DABP	*P. knowlesi*	*3*		*1*	*micronème*	*invasion: liaison au récepteur Duffy*	*Adams et alii , 1990*
DABP	*P. vivax*	*1*		*1*	*micronème*	*invasion: liaison au récepteur Duffy*	*Adams et alii , 1992*
EBA-175	*P. falciparum*	*1*	*175*	*2*	*micronème*	*invasion: liaison à la Glycophorine A*	*Camus et alii , 1985; Sim et alii , 1992*
EBL-1	*P. falciparum*	*1*		*2*	*—*	*invasion?*	*Peterson & Wellems, 2000*
PfEMP1	*P. falciparum*	*≈50*	*200-400*	*2-7*	*surface GRp*	*variation antigénique, adhésion, immunorégulation*	

Cette superfamille est définie par la présence de domaines riches en cystéine homologues au domaine fonctionnel des protéines DABP de *P. knowlesi* et *P. vivax* (Figure 17). Ces domaines sont impliqués dans des interactions de type récepteur/ligand entre le parasite et des cellules de son hôte. L'ensemble de ces protéines homologues peut être subdivisé en 2 familles : les protéines DABP, EBA-175 et EBL-1 d'un côté et les membres de la famille *var* (PfEMP1) de l'autre. En effet, les premières sont impliquées dans l'invasion de l'érythrocyte par le mérozoïte et se caractérisent par la présence d'une séquence signal, une localisation dans les micronèmes, un à deux domaine par molécule et des gènes uniques ou très peu redondants (Tableau 4). Quant aux molécules PfEMP1, elles sont impliquées dans la variation antigénique et la séquestration ; elles ne possèdent pas de séquence signal, présentent 2 à 7 domaines DBL, ainsi qu'un domaines CIDR qui leur est propre. De plus, elles se caractérisent par leur chimérisme ; en effet, elles appartiennent aussi à la famille Pf60/*var*.

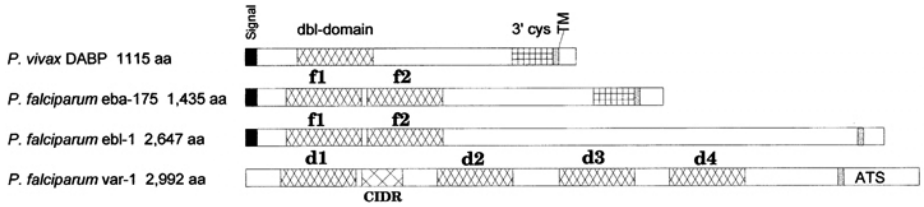

Figure 17 : Représentation schématique des molécules de la superfamille DBL.