

## La région HLA de classe I : une organisation compliquée par la présence de nombreuses familles multigéniques

Laurent Pichon  
Thierry Giffon  
Bruno Chauvel  
Jean-Yves Le Gall  
Véronique David

Le complexe majeur d'histocompatibilité (CMH) est un segment génomique situé sur le chromosome 6 comportant un nombre considérable de gènes impliqués dans la fonction immunitaire dont la carte commence à être relativement détaillée. Les gènes *HLA* de classe I occupent environ 2 000 kb dans la moitié télomérique du CMH. Trois gènes extrêmement polymorphes codent pour les antigènes leucocytaires classiques *HLA-A*, *HLA-B* et *HLA-C*, glycoprotéines ancrées à la surface des cellules qui, d'une part, présentent aux récepteurs des lymphocytes T les peptides antigéniques issus de protéines endogènes et, d'autre part, sont reconnues par les récepteurs des cellules *natural killer*; trois gènes, *HLA-E*, *HLA-F* et *HLA-G*, à la structure très proche, ont une fonction biologique encore incertaine. Outre ces six gènes, le CMH contient de nombreux pseudogènes de classe I et des fragments de gènes issus d'un gène ancestral *HLA* de classe I, mais aussi d'autres gènes, copies uniques ou familles multigéniques, tout cela formant une région dont la densité génique est extrêmement élevée.

### ADRESSES

L. Pichon, T. Giffon : docteurs ès sciences. B. Chauvel : docteur en médecine, docteur ès sciences. Poste d'accueil Inserm U. 391. J.Y. Le Gall, V. David : professeurs de biochimie et de biologie moléculaire à l'université de Rennes I, praticiens hospitaliers au CHU de Rennes. UPR 41 Cnrs, Recombinaisons génétiques, Faculté de Médecine, 2, avenue du Pr-Léon Bernard, 35043 Rennes Cedex.

**L**e complexe majeur d'histocompatibilité humain (CMH) est localisé sur le bras court du chromosome 6 en 6p21.3, où il recouvre environ 4 000 kb. Cette région du génome est l'une des plus étudiées aujourd'hui; elle apparaît comme l'une des plus denses en gènes et d'une particulière complexité structurale. Le CMH est classiquement divisé en trois régions (figure 1): (1) la plus centromérique est la région *HLA* de classe II, d'une taille de 900 kb et dont les principaux

locus sont *HLA-DP*, *-DQ* et *-DR*. Des analyses systématiques de cette région ont révélé son extraordinaire richesse en séquences codantes, avec plus de 32 gènes déjà identifiés [1]; (2) la région intermédiaire dite de classe III, s'étend sur 1 100 kb et contient également un grand nombre de gènes, au moins 39, parmi lesquels les gènes codant pour les *tumor necrosis factors*  $TNF\alpha$  et  $\beta$ , les gènes de la 21 hydroxylase et les gènes du complément C2 et C4; (3) la plus télomérique, la région *HLA*

## RÉFÉRENCES

1. Campbell RD, Trowsdale J. Map of the human MHC. *Immunol Today* 1993; 14: 349-52.
  2. Geraghty DE, Pein J, Lipsky B, Hansen JA, Taillon-Miller P, Bronson SK, Chaplin DD. Cloning and physical mapping of the HLA class I region spanning the *HLA-E* to *HLA-F* interval by using yeast artificial chromosomes. *Proc Natl Acad Sci USA* 1992; 89: 2669-73.
  3. El Kahloun A, Vernet C, Jouanolle AM, Boretto J, Mauvieux V, Le Gall JY, David V, Pontarotti P. A continuous restriction map from *HLA-E* to *HLA-F*. Structural comparison between different *HLA-A* haplotypes. *Immunogenetics* 1992; 35: 183-9.
  4. Venditti CP, Chorney MJ. Class I gene contraction within the *HLA-A* subregion of the human MHC. *Genomics* 1992; 14: 1003-9.
  5. Geraghty DE, Koller BH, Hansen JA, Orr HT. The *HLA* class I gene family includes at least six genes and twelve pseudogenes and gene fragments. *J Immunol* 1992; 149: 1934-46.
  6. Carillo S, Pariat M, Jariel-Encontre I, Steff A, Lorca T, Piechaczyk M. Le catabolisme protéique intracellulaire: une fonction biologique majeure. Partie II: exemples de dégradation conditionnelle et genèse des peptides antigéniques. *médecine/sciences* 1995; 11: 845-52.
  7. Vély F, Vivier E. Mécanismes moléculaires de la cytotoxicité des cellules NK. *médecine/sciences* 1996; 12: 458-64.
  8. Bodmer JG, Marsh SGE, Albert ED, Bodmer WF, Bontrop RE, Charron D, Dupont B, Erlich HA, Mach B, Mayr WR, Parham P, Sasazuki T, Schreuder GMY, Strominger JL, Svejgaard A, Terasaki PI. Nomenclature for factors of the HLA system 1995. *Hum Immunol* 1995; 43: 149-64.
  9. Le Bouteiller P. HLA class I chromosomal region, genes, and products: facts and questions. *Crit Rev Immunol* 1994; 14: 89-129.
  10. Hughes AL. Origin and evolution of *HLA* class I pseudogenes. *Mol Biol Evol* 1995; 12: 247-58.
  11. Chimini G, Boretto J, Marguet D, Lanau F, Lauquin G, Pontarotti P. Molecular analysis of the human MHC class I region using yeast artificial chromosome clones. *Immunogenetics* 1990; 32: 419-26.
  12. Krishnan BR, Jamry I, Chaplin DD. Feature mapping of the HLA Class I region: localization of the *POU5F1* and *TCF19* genes. *Genomics* 1995; 30: 53-8.
  13. Takeda J, Seino S, Bell GI. Human *OCT 3* gene family: cDNA sequences alternative splicing gene organization, chromosomal location and expression at low levels in adult tissues. *Nucleic Acids Res* 1992; 20: 4613-20.
  14. Zhou Y, Chaplin DD. Identification in the *HLA* class I region of a gene expressed late in keratinocyte differentiation. *Proc Natl Acad Sci USA* 1993; 90: 9470-4.
- de classe I, est aussi la plus vaste puisqu'elle recouvre près de 2000 kb, le gène tronqué *HLA-17* et le locus *HLA-F* constituant, respectivement, ses limites théoriques centromérique et télomérique. Son analyse structurale n'a été entreprise que relativement tard par rapport à celles des régions de classe II et classe III. Il a fallu, en effet, attendre le début des années 1990 pour que son clonage dans des chromosomes artificiels de levure et les techniques d'électrophorèse en champ pulsé permettent de préciser les distances physiques réelles: *HLA-B* et *-C*, séparés de 85 kb, sont situés à 1200 kb de *HLA-A*, *HLA-E* est localisé en position 700 kb centromérique par rapport à *HLA-A* et *HLA-F* en position 250 kb télomérique [2, 3]. Ces analyses cartographiques ont également démontré l'existence d'un polymorphisme de taille entre différents haplotypes *HLA-A*: par rapport à *HLA-A3*, l'haplotype *HLA-A24* est délété d'une quarantaine de kilobases, entre *HLA-A* et *HLA-H* [4] alors que l'haplotype *HLA-A31* présente une insertion d'environ 50 kilobases, immédiatement centromérique par rapport à *HLA-A* [3] (figure 2).
- De la même façon, au début des années 1990, seuls les gènes du système HLA avaient été positionnés dans cette vaste région de 2000 kb. Depuis, l'identification, la localisation et la caractérisation de nouvelles séquences se poursuivent à un rythme rapide; on y retrouve schématiquement trois grandes catégories de gènes: (1) la famille des gènes *HLA* de classe I, (2) des gènes copies uniques non *HLA*, (3) des gènes appartenant à de nouvelles familles multigéniques, présentant un nombre plus ou moins important de séquences.

### Les gènes HLA de classe I

Il s'agit d'une famille multigénique formée de dix-sept séquences apparentées et comprenant, outre les trois gènes *HLA-A*, *-B* et *-C* (encore appelés gènes *HLA* de classe I classiques ou de classe Ia), trois autres gènes fonctionnels *HLA-E*, *-F* et *-G* (encore appelés gènes *HLA* de classe I non classiques ou de classe Ib), ainsi que onze séquences non fonctionnelles [5] (figure 3).

Les gènes *HLA-A*, *-B* et *-C* codent pour les chaînes lourdes des molécules de classe I (44kDa) qui sont présentes à la surface de toutes les cellules, à l'exception des cellules germinales et de certaines cellules nerveuses. En association non covalente avec la  $\beta 2$  microglobuline ( $\beta 2m$ ), elles fixent les peptides antigéniques apprêtés puis les présentent aux récepteurs T des lymphocytes T CD8<sup>+</sup> cytotoxiques [6], d'une part, et préviennent la lyse par les cellules NK via leurs récepteurs de type inhibiteur, d'autre part [7].

Les gènes se composent de 8 exons séparés par 7 introns et engendrent un transcrit de 1,6 kb. Le premier exon correspond à la région 5' non traduite et au peptide signal, les exons 2, 3 et 4 codent pour les trois domaines extracellulaires  $\alpha 1$ ,  $\alpha 2$  et  $\alpha 3$  de 90 acides aminés chacun, l'exon 5 pour la région transmembranaire, les deux exons suivants pour le début de la partie intracytoplasmique, enfin, l'exon 8 correspond au reste de cette partie carboxyterminale ainsi qu'à la partie 3' non traduite. Ces gènes sont parmi les plus polymorphes décrits dans le génome humain; on dénombre en effet 59 allèles pour le locus *HLA-A*, 118 pour *HLA-B* et 36 pour *HLA-C* [8]. Le nombre de résidus polymorphes et le taux de variabilité sont essentiellement concentrés dans les domaines  $\alpha 1$  et  $\alpha 2$ . Sur les vingt positions hypervariables définies dans ces domaines  $\alpha 1$  et  $\alpha 2$ , 18 d'entre elles sont susceptibles d'affecter la spécificité de la reconnaissance par le récepteur T et le degré d'intensité de la réponse immunitaire.

Les gènes *HLA-E*, *HLA-F* et *HLA-G* sont également appelés gènes *HLA* de classe I non classiques. Bien qu'ils codent pour une molécule de structure très proche de celle des protéines de classe I classiques, leur distribution tissulaire est beaucoup plus restreinte et leur polymorphisme reste très faible. L'organisation structurale intron/exon de ces trois gènes est identique à celle des gènes *HLA* classiques et leur produit s'associe également de façon non covalente à la  $\beta 2m$ . Leur fonction n'est pas encore clairement définie et deux écoles s'opposent: pour les uns, l'absence d'équivalent chez la souris, le polymorphisme limité et le faible

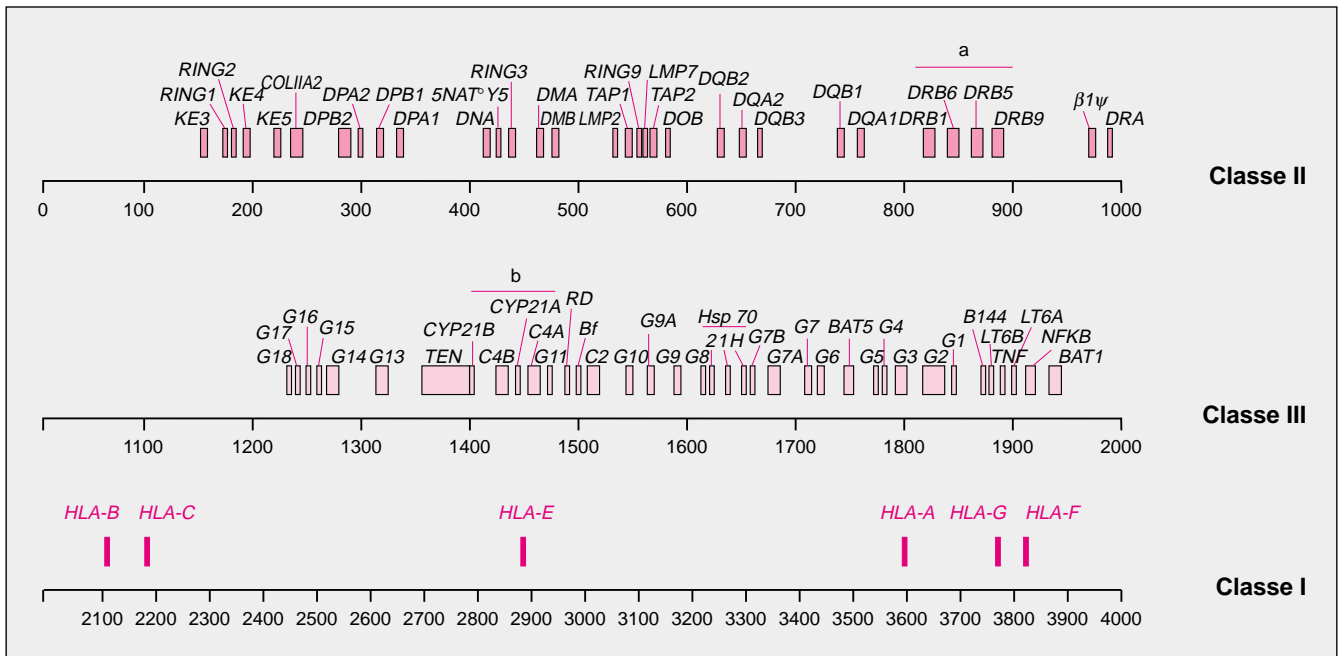


Figure 1. **Carte générale du complexe majeur d'histocompatibilité humain.** La région HLA de classe II s'étend sur les 1 000 kb centromériques. Elle comprend les gènes HLA-DP, -DQ et -DR, codant pour les chaînes  $\alpha$  et  $\beta$  des molécules de présentation de l'antigène, les gènes RING (really interesting new genes) dont les fonctions sont encore inconnues, les gènes DMA et DMB, dont les produits agissent comme molécules accessoires dans la présentation de l'antigène par les molécules de classe II, les gènes TAP (transporter associated with antigen processing) 1 et 2, ainsi que les gènes LMP (low molecular weight polypeptides) 2 et 7 impliqués dans la constitution des protéasomes nécessaires à la dégradation des antigènes cytosoliques en peptides [40]. La région de classe III s'étend sur les 1 000 kb suivants et comprend, entre autres, les gènes codant pour les molécules du complément, C2, C4A et C4B, le facteur Bf, les gènes TNFA et TNFB (tumor necrosis factor  $\alpha$  et  $\beta$ ), les gènes HSP70 (heat shock protein), les gènes BAT (HLA-B adjacent transcript) ainsi que d'autres gènes dont la fonction n'est pas déterminée. Enfin, la région de classe I s'étend sur les 2 000 kb télomériques. Des différences dans l'organisation des gènes DRB (a) et des gènes C4/CYP21 (b) ont été observées entre différents haplotypes [1].

niveau d'expression en surface les désignent comme des vestiges inactivés au cours du temps ; pour les autres, la capacité de présentation des peptides, au moins dans le cas de *HLA-E* et *HLA-G*, la conservation de *HLA-F*, le rôle de plus en plus probable de *HLA-G* dans la tolérance foeto-maternelle et l'augmentation progressive du nombre de tissus décrits où ces gènes sont transcrits en font des candidats potentiels pour un rôle, limité mais réel, de présentation d'antigènes [9].

Le gène *HLA-E*, le plus divergent, présente une délétion de cinq nucléotides dans l'exon 7, ce qui entraîne un raccourcissement de la partie intracytoplasmique de la molécule. Il contient trois séquences *Alu*, situées, respectivement, dans la

région 5' flanquante en amont de la boîte CAT, dans l'intron 5 et dans la région 3' non traduite où elle crée un nouveau site de polyadénylation. Le gène *HLA-F* est le plus télomérique de tous les gènes de classe I. Sa séquence présente une mutation au niveau du signal d'épissage de l'intron 6 (G en A), entraînant l'absence de l'exon 7 dans l'ARNm mûr. La fin de sa séquence 3' non traduite, de la base 3826 à la base 4316, diverge complètement par rapport à la séquence de classe I classique. Cette zone divergente, très conservée au cours de l'évolution et contenant deux sites de polyadénylation supplémentaires, montre de très fortes analogies avec des séquences codant pour des sous-unités ribosomiques. Présent sur un fragment

Hind III de 6 kb, le gène *HLA-G* a été localisé à 100 kb télomérique par rapport à *HLA-A*. Il contient un codon de terminaison au début de l'exon 6, ce qui entraîne la perte de toute la partie 3' de la molécule codée par les exons 6, 7 et 8. Du fait que la protéine *HLA-G* soit la seule molécule de classe I synthétisée dans la barrière foeto-placentaire, il a été suggéré que cette molécule pourrait relayer l'interaction entre les cellules fœtales et les lymphocytes maternels et jouer ainsi un rôle dans la tolérance foeto-maternelle et dans le maintien de l'allogreffe foetale.

En dehors de ces gènes classiques *HLA-A*, *-B* et *-C*, et non classiques *HLA-E*, *-F* et *-G*, la famille *HLA* de classe I comprend onze séquences apparentées de nature diverse :

## RÉFÉRENCES

15. Volz A, Weiss E, Trowsdale J, Ziegler A. Presence of an expressed  $\beta$ -tubulin gene (*TUBB*) in the HLA class I region may provide the genetic basis for HLA-linked microtubule dysfunction. *Hum Genet* 1994; 93: 42-6.
16. Vernet C, Ribouchon MT, Chimini G, Pontarotti P. Structure and evolution of a member of a new subfamily of GTP-binding proteins mapping to the human MHC class I region. *Mamm Genome* 1994; 5: 100-5.
17. Vernet C, Boretto J, Mattei MG, Takahashi M, Jack LJW, Mather IH, Rouquier S, Pontarotti P. Evolutionary study of multigenic families mapping close to the human MHC class I region. *J Mol Evol* 1993; 37: 600-12.
18. Chu TW, Capossela A, Coleman R, Goei VL, Nallur G, Gruen JR. Cloning of a new finger protein gene (*ZNF173*) within the class I region of the human MHC. *Genomics* 1995; 29: 229-39.
19. El Kahloun A, Chauvel B, Mauvieux V, Dorval I, Jouanolle AM, Gicquel I, Le Gall JY, David V. Localization of seven new genes around the *HLA-A* locus. *Hum Mol Genet* 1993; 2: 55-60.
20. Giffon T, Lepourcelet M, Pichon L, Jezequel P, Bouric P, Carn G, Pontarotti P, Le Gall JY, David V. Cloning of a human homologue of the murine *Tctex-5* gene within the class I region of the CMH. *Immunogenetics* 1996 (sous presse).
21. Chauvel B, Dorval I, Fergelot P, Pichon L, Giffon T, Gicquel I, Le Gall JY, David V. A human homologue to the yeast omnipotent suppressor 45 maps 100 kb centromeric to *HLA-A*. *Immunogenetics* 1994; 41: 380-2.
22. Frovola L, Le Goff X, Rasmussen HH, Cheperegin S, Drugeon G, Kress M, Arman I, Haenni AL, Celis JE, Philippe M, Justesen J, Kisselev L. A highly conserved eukaryotic protein family possessing properties of polypeptide chain release factor. *Nature* 1994; 372: 701-3.
23. Pham-Dinh D, Della Gaspera B, Kerlero de Rosbo N, Dautigny A. Structure of the human myelin/oligodendrocyte glycoprotein gene and multiple alternative spliced isoforms. *Genomics* 1995; 29: 345-52.
24. Fan W, Liu Y, Parimoo S, Weissman SM. Olfactory receptor-like genes are located in the human major histocompatibility complex. *Genomics* 1995; 27: 119-23.

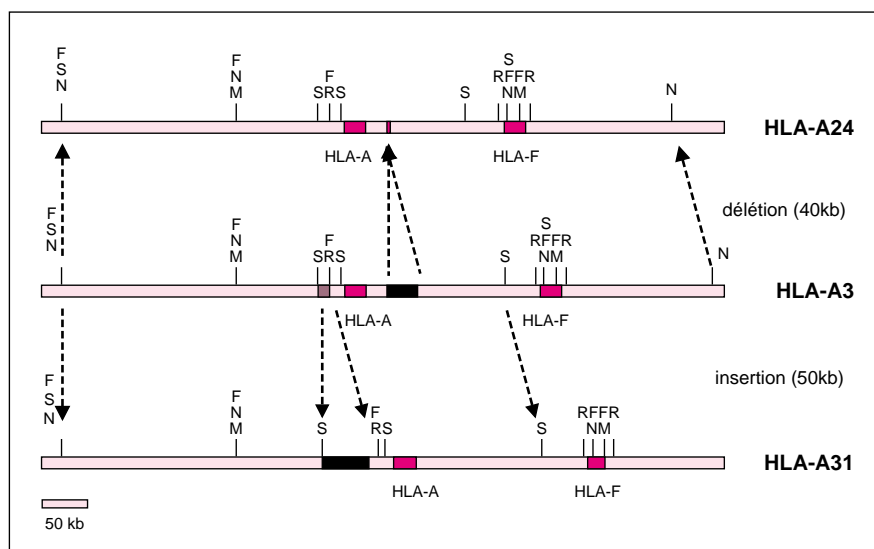


Figure 2. **Cartographie physique à grande échelle.** Elle démontra l'existence de polymorphismes de taille entre différents haplotypes HLA-A. Par rapport à HLA-A3, l'haplotype HLA-A24 présente une délétion d'une quarantaine de kilobases, entre HLA-A et HLA-F [2], alors que l'haplotype HLA-A31 présente une insertion d'environ 50 kilobases, immédiatement centromérique par rapport à HLA-A [3]. Les portions du génome impliquées dans ces polymorphismes de taille sont représentées par des rectangles noirs. Les sites enzymatiques sont indiqués par une lettre : F: Sfi I, S: Sal I, N: Not I, M: Mlu I, R: Nru I.

quatre pseudogènes entiers *HLA-H*, *-J*, *-K* et *-L*, quatre gènes tronqués *HLA-75*, *-80*, *-90* et *-16* et trois fragments de gènes *HLA-17*, *-21* et *-30*, dénommés d'après la taille du fragment Hind III dans lesquels ils sont inclus [5] (figure 3). Les pseudogènes sont retrouvés sur des fragments Hind III de 5,4 (*HLA-H*), 5,9 (*HLA-J*), 7,0 (*HLA-K*) et 9,2 kb (*HLA-L*). Ils présentent tous une structure intron/exon identique à celle des gènes de classe I mais diverses anomalies ponctuelles les rendent non fonctionnels. Ils sont tous localisés à proximité de *HLA-A*.

Trois des gènes tronqués, *HLA-80*, *-90* et *-16*, semblent avoir un ancêtre commun proche dans l'évolution du fait de la grande analogie retrouvée entre leurs séquences [10]. *HLA-80* et *90* ont une structure identique, constituée d'un fragment génomique allant de l'exon 2 à la fin de la région 3' non traduite. Ces deux séquences sont localisées à 30 kb en position centromérique par rapport à *HLA-A* pour la première, et à 50 kb centromérique par rapport au gène *HLA-F* pour la seconde. *HLA-16*, localisé

juste à côté du pseudogène *HLA-H*, est constitué de la même portion de séquence classe I amputée de l'exon 2 [5]. Le gène tronqué *HLA-75* est constitué d'une portion de la séquence classique allant de la région promotrice jusqu'à l'exon 3 et est adjacent à *HLA-90*. La séquence de ce locus présente 85 % d'identité avec les séquences de *HLA-A* et de *HLA-G*, suggérant une origine commune [10].

Trois fragments de gènes appartiennent également à cette famille multigénique. *HLA-17*, la plus centromérique de toutes les séquences de classe I, est apparenté aux locus *HLA-B* et *-C* mais ne contient que 600 pb de séquence classique, de la fin de l'exon 5 au début de l'exon 8. *HLA-21* présente des analogies avec l'exon 4 de *HLA-A* et est localisé entre ce gène et le pseudogène *HLA-70*. *HLA-30* semble plus apparenté à *HLA-E* et au pseudogène *HLA-L* et est situé à environ 50 kb en position centromérique par rapport à ce dernier. Enfin, trois séquences supplémentaires ont été identifiées mais non clonées, la séquence *HLA-11*, spéci-

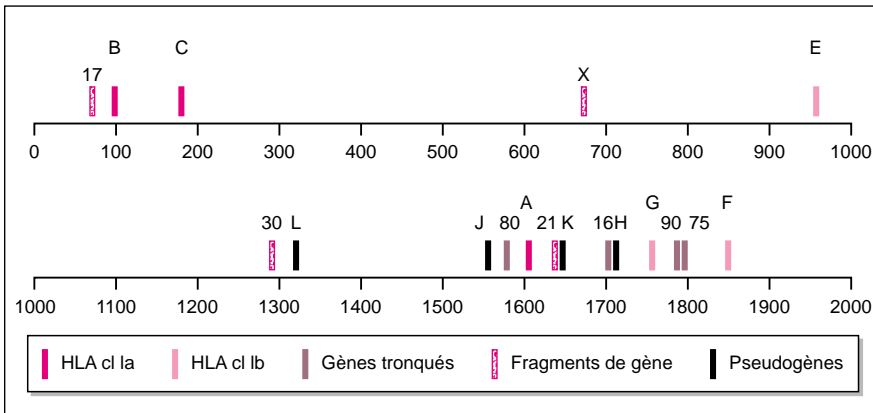


Figure 3. Localisation des membres de la famille multigénique HLA de classe I, dans les 2000 kb séparant HLA-17, la séquence HLA la plus centromérique, de HLA-F, la plus télomérique. Elle comprend les gènes fonctionnels classiques HLA-A, -B, -C, et non classiques HLA-E, -F et -G, les pseudogènes HLA-F, -J, -K et -L, les gènes tronqués HLA-75, -80, -90, -16 et -X et enfin les fragments de gènes HLA-17, -21 et -30.

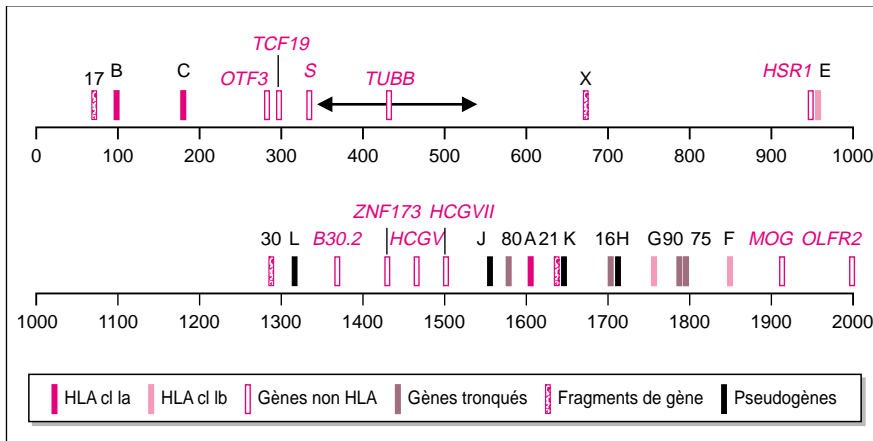


Figure 4. Localisation physique des 11 nouveaux gènes de type copie unique récemment identifiés dans la région HLA de classe I du CMH. Les gènes OTF3, encore appelés POU5F1, TCF19, S, TUBB, HSR1, ZNF173, HCGV, HCGVII et l'exon B30.2 sont localisés entre HLA-C et HLA-J, les gènes MOG et OLFR2 sont télomériques par rapport à HLA-F.

fique de l'haplotype *HLA-A11-B60*, la séquence *HLA-14*, identifiée par une sonde spécifique de *HLA-E* et la séquence *HLA-X*, isolée dans la lignée lymphoblastoïde HHKB et localisée en position centromérique par rapport à *HLA-E* [5, 11].

### Les gènes non HLA « copie unique »

Outre cette famille *HLA* de classe I, un certain nombre de nouveaux gènes a été isolé et caractérisé dans cette région (figure 4).

Également connu sous le nom de *OTF3* (*octamer-binding transcription factor n°3*) ou *OCT3*, le gène *POU5F1* a été physiquement localisé à 100 kb puis à 130 kb en position télomérique par rapport à *HLA-C* [12]. Ce gène est formé de 5 exons contenus dans un fragment génomique de 7 kb. Deux transcrits de 1,37 et 1,15 kb, engendrés par épissage alternatif et appelés *OCT3A* et *3B*, ont été détectés dans tous les tissus adultes analysés. Ce gène appartient à une famille codant pour des facteurs de transcription, la famille *POU* (ce nom dérive des protéines Pit 1, Oct 1 et 2

et Unc 86, qui présentent un homéodomaine polypeptidique appelé *POU*). Son analogue murin, *Otf-3*, localisé dans le complexe H2, est exprimé au début de l'embryogenèse [13].

*TCF19*, gène de régulation de la croissance cellulaire, a été localisé à environ 130 kb de *HLA-C* et à seulement 600 pb du gène *POU5F1*. Il est composé de trois exons, avec une séquence de type *Alu* dans le premier intron et semble bien conservé chez les mammifères. Il se présente dans un sens de transcription inverse par rapport à celui du gène *HLA-C* [12]. Un autre gène, appelé *S* (*skin*), exprimé spécifiquement dans les kératinocytes en différenciation sous forme de deux ARNm de 2,2 et 2,6 kb, a été localisé à 160 kb en position télomérique par rapport à *HLA-C*, dans une orientation centromère/télocère [14]. Ce gène présente deux exons séparés par un intron de 2,9 kb et l'ARNm séquencé, de 2547 pb, contient deux signaux de polyadénylation distincts.

Le gène *TUBB*, codant pour la chaîne  $\beta$  de la tubuline, protéine de structure des microtubules, a été localisé approximativement, sur des hybrides d'irradiation contenant des fragments du chromosome 6, dans un segment génomique allant de 170 à 370 kb en position télomérique par rapport à *HLA-C* [15].

Le gène *HSR1* est situé en position centromérique par rapport à *HLA-E*. Un transcrit majoritaire de 2 kb a été mis en évidence dans tous les tissus analysés, et un autre, de plus grande taille, dans le cerveau. L'ARNm de 2 kb présente un cadre de lecture ouvert de 1290 pb. La séquence protéique déduite, de 430 acides aminés pour une masse moléculaire de 47,3 kDa, présente une très forte analogie avec une protéine de fixation du GTP. Une séquence analogue de ce gène, appelée *MMRI*, a également été retrouvée dans le complexe H2, sur le chromosome 17 de la souris [16].

En utilisant une stratégie fondée sur la conservation de séquence entre l'homme et le porc, un nouvel exon, appelé *B30.2*, a été isolé et localisé à 200 kb centromérique par rapport à *HLA-A* [17]. Cet exon présente un cadre de lecture ouvert, de 166 acides aminés, qui correspond à un

## RÉFÉRENCES

25. Amadou C, Ribouchon MT, Mattei MG, Jenkins NA, Gilbert DJ, Copeland NG, Avoustin P, Pontarotti P. Localization of new genes and markers to the distal part of the human major histocompatibility complex (MHC) region and comparison with the mouse: new insights into the evolution of mammalian genomes. *Genomics* 1995; 26: 9-20.
26. Leelayuwat C, Townend DC, Degli-Esposti MA, Abraham LJ, Dawkins RL. A new polymorphic and multicopy MHC gene family related to nonmammalian class I. *Immunogenetics* 1994; 40: 339-51.
27. Bahram S, Bresnahan M, Geraghty DE, Spies T. A second lineage of mammalian major histocompatibility complex class I genes. *Proc Natl Acad Sci USA* 1994; 91: 6259-63.
28. Avoustin P, Ribouchon MT, Vernet C, N'Guyen B, Crouau-Roy B, Pontarotti P. Non-homologous recombination within the major histocompatibility complex creates a transcribed hybrid sequence. *Mamm Genome* 1994; 5: 771-6.
29. Venditti CP, Harris JM, Geraghty DE, Chorney MJ. Mapping and characterization of non *HLA* multigene assemblages in the human MHC class I region. *Genomics* 1994; 22: 257-66.
30. Leelayuwat C, Pinelli M, Dawkins L. Clustering of diverse replicated sequences in the MHC. Evidence for en bloc duplication. *J Immunol* 1995; 155: 692-8.
31. Pichon L, Carn G, Bouric P, Giffon T, Chauvel B, Lepourcelet M, Mosser J, Le Gall JY, David V. Structural analysis of the *HLA-A/HLA-F* subregion: precise localization of two new multigene families closely associated with the *HLA* class I sequences. *Genomics* 1996; 32: 236-44.
- domaine protéique très fortement conservé dans plusieurs molécules telles que la butyrophiline bovine, le facteur nucléaire du xénope XNF7, la protéine humaine RFP (*ret finger protein*) ou encore un antigène nucléaire humain de 52 kDa appelé SS-A/Ro. Le gène complet n'a pas encore été isolé.
- Le gène *ZNF173* [18], également identifié sous le nom de *HCGVI* (*hemochromatosis candidate gene n°VI*) [19], est situé à 170 kb en position centromérique par rapport au locus *HLA-A*. Il est constitué de 7 exons répartis sur 14 kb et est transcrit en un ARNm de 3,6 kb, avec un cadre de lecture ouvert de 1 620 pb codant pour une protéine de 539 acides aminés, de type AFP (*acid finger protein*). Cette séquence protéique présente une analogie importante avec celle traduite du gène *RFP* humain. Ces deux protéines contiennent de multiples résidus histidine et cystéine, qui forment des structures en «doigt» susceptibles de représenter des domaines de fixation dépendante d'un métal. Le gène *ZNF173* est conservé chez les mammifères et exprimé de façon ubiquiste.
- Le gène *HCGV* a été localisé à 150 kb en position centromérique par rapport à *HLA-A*. Il est composé de trois exons distincts, code pour un transcrit ubiquiste de 1,8 kb, et est traduit en une protéine potentielle de 126 acides aminés, d'une masse moléculaire évaluée à 14 kDa. Ce gène est l'analogue humain du gène murin *Tctex 5* (*T complex testis expressed*) [20].
- Le gène *HCGVII* est l'un des quatre gènes humains analogues du gène *omnipotent suppressor SUP45* de la levure *S. cerevisiae*. Une séquence est située sur le chromosome 6, à 100 kb en position centromérique par rapport à *HLA-A* [21] tandis que les autres copies sont réparties sur les chromosomes 5, 7 et X. L'ADNc montre également une analogie forte avec l'ARNm à effet maternel *Cl1*, isolé dans l'œuf de xénope. Ces gènes, bien conservés chez les vertébrés supérieurs, codent vraisemblablement pour des facteurs de terminaison de la traduction [22].
- Le gène *MOG* (*myelin oligodendrocyte glycoprotein*), est localisé à 60 kb de *HLA-F*, en position télomérique. Il appartient à la superfamille des immunoglobulines et la protéine correspondante, exprimée à la surface des oligodendrocytes en culture ou de la lamelle périphérique du manchon de myéline du système nerveux central, ne semble être retrouvée que chez les mammifères. Ce gène, dont la taille est d'environ 17 kb, comprend 9 exons et est transcrit en 6 ARNm différents engendrés par épissage alternatif [23].
- Le gène *OLFR2* (*olfactory receptor 2*), d'abord isolé sous le nom de *FAT 11* (*HLA-F adjacent transcript n°11*), est situé à 150-200 kb en position télomérique par rapport à *HLA-F*, et code pour un transcrit ubiquiste de 2 kb. La séquence de la protéine potentielle présente une identité de 60 % avec la séquence déduite de l'ADNc spécifique du récepteur olfactif murin OR-3 [24]. Comme dans le cas d'autres gènes de récepteurs olfactifs, la région codante de *OLFR2* ne présente pas d'intron.
- Le gène *RFP* a été localisé physiquement chez la souris, dans le complexe H2 puis à environ 700 kb télomérique par rapport au gène *HLA-F* humain, au cours d'une étude de syntenie entre les chromosomes 6 humain et 17 murin [25].

## Les familles multigéniques non HLA

Dix autres familles multigéniques, dont probablement huit différentes, ont été décrites dans cette région. La fréquence et l'intrication de ces séquences multiples rend compte, pour une large part, de la complexité structurale de cette région (*figure 5*). Parmi celles-ci, deux au moins présentent des caractéristiques susceptibles de les rapprocher de la famille *HLA*, avec un rôle éventuel dans la fonction immunitaire, les familles *PERB11* et *P5*. Une analyse informatique de séquence de la région située entre *TNF* et *HLA-B* a identifié le premier gène de la famille *PERB11* (*PER* pour Perth, B pour bloc  $\beta$ , n°11). Quatre autres locus de cette famille ont ensuite été reconnus, dont trois dans la région *HLA* de classe I: *PERB11.2* situé en position centromérique par rapport à *HLA-B*; *PERB11.3* à proximité du gène *HLA-E* et *PERB11.4* à proximité de *HLA-A*. Le dernier locus, *PERB11.5*, n'avait pas été localisé physiquement [26].

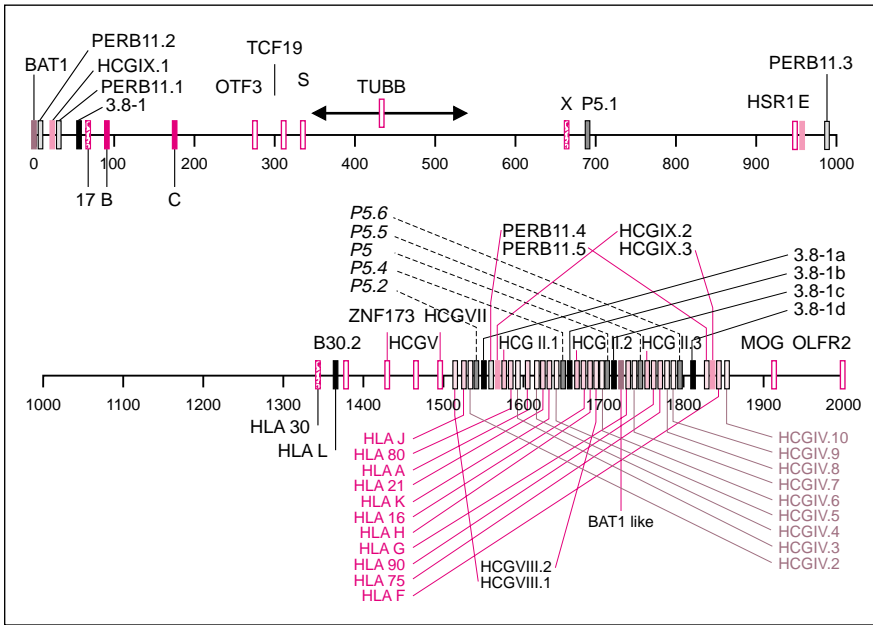


Figure 5. **Cartographie physique récapitulative de l'ensemble de la région HLA de classe I du CMH faisant apparaître la localisation physique précise des différents locus appartenant aux huit familles multigéniques P5.1, 3.8-1, PERB11, BAT1, HCGII, HCGIV, HCGVIII et HCGIX identifiées dans cette région.** Il est intéressant de noter que l'essentiel de ces locus redondants est retrouvé dans deux zones distinctes séparées d'environ 1000 kb, autour de HLA-B d'une part, et entre HLA-J et HLA-F d'autre part. Seules deux séquences appartenant aux familles multigéniques (PERB11.3 et P5.1) ont été retrouvées en dehors de ces deux zones, respectivement à proximité de HLA-E et HLA-X. Cette figure met également en évidence le contraste important existant entre la portion centrale de cette région (environ 1000 kb), caractérisée par la présence en un seul exemplaire de différents gènes, et les deux portions flanquantes.

Le gène *PERB11.1* présente un transcrit de 3 kb dans le muscle squelettique, un autre de 15 kb au moins dans le poumon, et deux de 8 et 10 kb dans le foie, le thymus, la prostate, les testicules, le duodénum et les leucocytes [26]. L'analyse informatique de l'ARNm indique que la protéine déduite présente d'importantes analogies avec les molécules de classe I du CMH de plusieurs espèces animales, telles que le lézard, la grenouille ou le poulet, et laisse envisager un rôle éventuel dans la réponse immunitaire. La famille *PERB11* est, en fait, identique à une autre famille multigénique décrite sous le nom de *MIC* et comprenant cinq séquences distinctes, appelées *MIC-A* à *MIC-E* [27]. Le cinquième locus, *MIC-E*, correspond au locus *PERB11.5* qui n'avait pu être localisé et est situé à proximité de *HLA-F*.

La famille *P5* a été caractérisée en utilisant comme sonde un sous-clone d'un chromosome artificiel de levure, le YAC 6F6, localisé entre *HLA-C* et *HLA-A* et contenant *HLA-X*. Ces séquences, appelées *P5-1* à *P5-7*, sont toutes situées à proximité de séquences *HLA* de classe I. La transcription de ces gènes, effectuée sous le contrôle d'un promoteur de classe I, est restreinte aux tissus lymphoïdes suggérant, une fois encore, une fonction immunologique. La séquence *P5.1* est localisée à proximité du gène *HLA-X* alors que les autres séquences sont rassemblées dans la région entourant *HLA-A* [28]. Deux autres familles multigéniques ont été isolées lors d'une étude de la zone délétée dans les haplotypes *HLA-A9* (*HLA-A23* et *A24*) [29]. La famille multigénique *3.8-1* comprend cinq séquences dont quatre sont

localisées à proximité de *HLA-A* et la dernière entre les gènes *HLA-B* et *TNF*. Cette dernière séquence, non conservée au cours de l'évolution, code pour un ARNm de 9,5 kb. La seconde famille multigénique, décrite sous le nom de *10.0-1*, comprend huit fragments dont cinq sont situés autour de *HLA-A*, un à proximité du gène *HLA-B* et un autre à proximité de *HLA-E*, le dernier n'étant pas localisé. Malgré quelques divergences dans le nombre et la localisation des séquences, l'analyse de l'ADNc *10.0-1* montre qu'il y a vraisemblablement identité entre les familles *P5.1* et *10.0-1* [29].

La famille *BAT1* (*HLA-B adjacent transcript n°1*) a été caractérisée après étude de cosmides recouvrant 400 kb de la région centromérique par rapport à *HLA-B*. Cette famille comprend, outre le gène *BAT1* localisé à 30 kb centromérique par rapport au gène *TNFA*, un second locus, appelé *PERB2* ou *BAT1 like*, situé à proximité de *HLA-A* [30].

Les gènes *HCGII* et *HCGIV* ont été isolés par sélection d'ADNc à partir d'un YAC contenant *HLA-A*, le YAC B30H3. Ces deux gènes appartiennent à des familles multigéniques comprenant un grand nombre de séquences redondantes (11 locus pour la famille *HCGIV* et 4 pour *HCGII*). Ces séquences sont étroitement associées aux gènes et aux pseudogènes de la famille *HLA-A* ainsi qu'aux membres des autres familles multigéniques de la région, en particulier *P5* et *3.8-1*, formant ainsi des groupes de gènes qui pourraient être le résultat de remaniements chromosomiques successifs au cours de l'évolution [31].

Les familles *HCGVIII* et *IX*, mises en évidence par des sondes issues du sous-clonage du YAC 168E5 (recouvrant la région *HLA-A/HLA-F*), sont moins complexes. *HCGVIII* comprend simplement deux séquences, situées respectivement à 100 kb centromériques et 110 kb télomériques par rapport à *HLA-A* [32]. Cette famille révèle sur *Northern blot*, un transcrit ubiquiste de 9,3 kb et un autre de 2,7 kb retrouvé spécifiquement dans le testicule et l'ovaire adultes ainsi que dans le cerveau et le rein d'origine fœtale. La famille *HCGIX* présente trois séquences situées à proximité et centromé-

## RÉFÉRENCES

32. Pichon L, Giffon T, Chauvel B, Carn G, Bouric P, El Kahloun A, Le Gall JY, David V. Physical map of the *HLA-A/HLA-F* subregion and identification of two new coding sequences. *Immunogenetics* 1996; 43: 175-81.
33. Pichon L, Hampe A, Giffon T, Carn G, Le Gall JY, David V. A new non-HLA multi-gene family associated with the *PERB11* family within the the MHC class I region. *Immunogenetics* 1996 (sous presse).
34. Wei H, Fan WF, Xu H, Parimoo S, Shukla H, Chaplin DD, Weissman SM. Genes in one megabase of the *HLA* class I region. *Proc Natl Acad Sci USA* 1993; 90: 11870-4.
35. Fan WF, Wei X, Shukla H, Parimoo S, Xu H, Sankhavaram P, Li Z, Weissman SM. Application of cDNA selection techniques to regions of the human MHC. *Genomics* 1993; 17: 575-81.
36. Marshall B, Leelayuwat C, Degli-Esposti MA, Pinelli M, Abraham LJ, Dawkins RL. New major histocompatibility complex genes. *Hum Immunol* 1993; 38: 24-9.
37. Mizuki N, Ando H, Kimira M, Ohno S, Miyata S, Goto K, Ishihara M, Yamazaki M, Watanabe K, Ono A, Taguchi S, Nakamura S, Sato M, Kikuti YY, Ando A, Shigenari A, Okumura T, Inoko H. Nucleotide sequence determination of the *HLA* class I region spanning the 240 kb segment around the *HLA-B* and *-C* genes. HUGO's human genome meeting Heidelberg 22-23 March 1996.
38. Goei VL, Parimoo S, Capossela A, Chu TW, Gruen JR. Isolation of novel non-*HLA* gene fragments from the hemochromatosis region (6p21.3) by cDNA hybridization selection. *Am J Hum Genet* 1994; 54: 244-51.
39. Totaro A, Rommens JM, Grifa A, Lunardi C, Carella M, Huizenga JJ, Roetto A, Camaschella C, De Sandre G, Gasparini P. Hereditary hemochromatosis: generation of a transcription map within a refined and extended map of the *HLA* class I region. *Genomics* 1996; 31: 319-26.
40. Bahram S. Transporteurs de peptides et présentation de l'antigène. *médecine/sciences* 1993; 9: 1204-13.
- riques, respectivement, par rapport aux locus *HLA-B*, *HLA-80* et *HLA-F*; elle code pour un petit messager de 0,7 kb [33].
- Outre ces gènes et familles de gènes assez bien caractérisés que nous venons de passer en revue, un grand nombre d'autres séquences codantes a également été identifié dans cette région par des techniques de sélection d'ADNc et de séquençage systématique. A défaut d'une analyse poussée, ils n'ont pour l'instant fait l'objet que de publications préliminaires: *HCGI* et *III*, situés respectivement à 250 et 170 kb et en position centromérique par rapport à *HLA-A* [17], 32 locus *CAT* et 11 transcrits *FAT* (pour respectivement *HLA-C* et *-F adjacent transcript*) [34, 35], 9 autres séquences *PERB* [36], 4 gènes *NOB* (new organization associated with *HLA-B*) [37] ainsi que différents ADNc sans dénomination particulière [38, 39]. Il n'est pas impossible que certaines de ces séquences codantes soient identiques, mais cela ne pourra être vérifié que lorsque la totalité des informations sera disponible dans les banques de données internationales.
- Enfin, plusieurs transcrits ont été assignés en 6p21.3 par FISH (*fluorescence in situ hybridization*) sans qu'une recherche de localisation précise ait été effectuée à ce jour, tels que les gènes *ZNF76* et *165*, le gène *RAGE* (*receptor for advanced glycosylation end*), le gène *ID4*, réglant négativement l'expression tissulaire spécifique de certains gènes humains, le gène *H1.1* de la famille des histones ou encore le gène *DDR*, l'analogue humain du gène *NEP/RTK*, localisé dans le complexe H2 murin. Il est donc possible qu'une partie d'entre eux appartienne également à la région *HLA* de classe I.
- La complexité structurale de la région *HLA* de classe I résulte de l'intrication de nombreuses séquences appartenant à huit familles multigéniques au moins (*figure 5*). A l'instar du travail effectué pour les séquences *HLA-A*, l'inventaire des gènes fonctionnels, pseudo-gènes ou gènes tronqués de ces familles reste à entreprendre. En fait, la structure de la région *HLA* de classe I n'est pas uniforme et un contraste apparaît nettement entre deux types d'organisation: d'une part, une intrication de séquences multiples, c'est le cas pour l'environnement immédiat des gènes *HLA-B* et *-C* d'un côté, et *HLA-A* et *-F* de l'autre; et, d'autre part, une succession de gènes présents à l'état d'une seule copie; c'est le cas pour la vaste région qui s'étend entre ces deux pôles. Différents schémas évolutifs ont été proposés pour rendre compte de la complexité actuellement observée; aucun, cependant, ne permet d'expliquer pourquoi les parties de la région *HLA* de classe I ayant fait l'objet du plus grand nombre de remaniements sont celles qui contiennent les gènes *HLA-A*, *-B* et *-C* dont on sait, par ailleurs, qu'ils sont fonctionnellement très conservés.
- Les associations différentes de chacune des familles multigéniques avec les locus *HLA* peuvent fournir des informations quant à l'évolution de la région. En résumé, les familles *P5.1* et *PERB11* présentent, en plus des locus localisés autour de *HLA-B* et *HLA-A*, au moins un locus associé à *HLA-E*. Les locus des familles *3.8-1* et *HCGIX* ne sont retrouvés associés qu'aux locus *HLA-A* et *-B*. Enfin, les membres des familles *HCGII*, *IV* et *VIII* sont tous localisés autour de *HLA-A*. L'association de *P5.1* et de *PERB11* avec le locus *HLA-E*, décrit comme étant le locus *HLA* ayant divergé le premier, il y a environ 40 millions d'années, aurait donc précédé l'apparition des locus ancestraux des familles *3.8-1* et *HCGIX*. Les familles *HCGII* et *HCGIV* seraient apparues plus tardivement. Les séquences de la famille *HCGII* n'étant retrouvées qu'à proximité des locus *HLA-80*, *-90* et *-16*, tous les trois décrits comme dérivant d'un ancêtre commun récent [10], il est probable que cette famille soit apparue à l'occasion des derniers phénomènes de duplication. Enfin, le nombre restreint de locus de la famille *HCGVIII* suggère également une apparition tardive. Les études en *zooblots*, semblent confirmer cette chronologie car seule la famille *PERB11* est conservée.
- Enfin, la question d'une éventuelle unité fonctionnelle de cette région se pose également. Quelques renseignements vont dans ce sens: les gènes de la famille *P5.1* ont une expression réglée par un promoteur de type



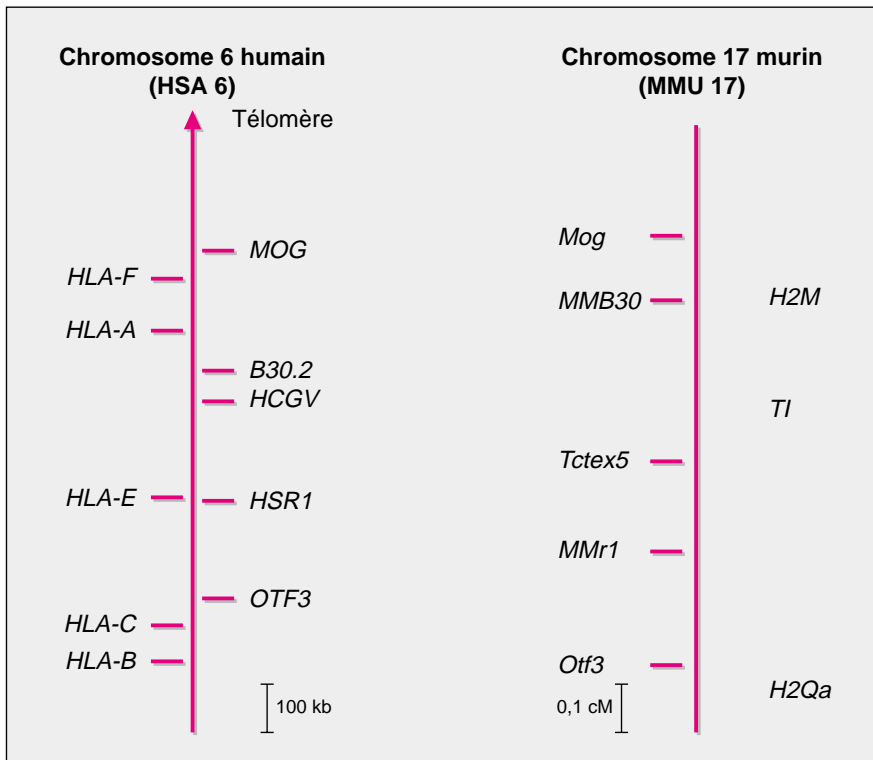


Figure 6. **Carte comparative de la partie distale des CMH humain (HLA) et murin (H2M, TI et H2Qa), sur les chromosomes HSA 6 et MMU 17, avec localisation respective des gènes analogues OTF-3 et Otf-3, HSR1 et MMr1, HCGV et Tctex-5, B30.2 et MMB30, MOG et Mog [25].**

HLA de classe I et limitée à des tissus lymphoïdes; l'analyse informatique des séquences *PERB11* suggère une capacité des protéines déduites de ces séquences de lier des peptides; enfin, au cours de réarrangements évolutifs successifs, les gènes sont restés étroitement associés, ce qui indiquerait également une coopération de leur fonction biologique.

Quoi qu'il en soit, en quelques années, plusieurs gènes autres que les gènes *HLA* de classe I, ont été identifiés dans cette région. Cette liste n'est, bien entendu, que provisoire. En effet, tous les résultats préliminaires obtenus par FISH ou sélection d'ADNc concordent pour estimer que la densité en gènes de cette région est apparemment identique à celle des deux autres régions du CMH, les régions *HLA* de classe II et de classe III, avec un gène différent tous les 10 à 20 kb. L'identification de ces séquences codantes sera sans doute accélérée par les différents programmes de séquençage systématique à grande échelle du CMH

actuellement mis en place dans différents pays. Au-delà d'un inventaire exhaustif, il restera bien entendu à comprendre précisément les fonctions biologiques de ces gènes. D'ores et déjà, quelques renseignements se dégagent de l'analyse des séquences géniques du fait d'analogies plus ou moins grandes des séquences protéiques déduites avec certains groupes fonctionnels: facteurs de transcription (OTF3), récepteurs olfactifs (OLFR2), fonctions immunologiques potentielles (P5 ou *PERB11*). Par ailleurs, la mise en évidence d'un bloc synténique entre cette région du chromosome 6 et une partie du chromosome 17 de la souris (figure 6) a été fructueuse sur le plan structural en conduisant à l'identification, chez l'homme, de gènes préalablement connus chez la souris (*Otf3*, *Mog* et *RFP*) et en montrant que, dans les deux espèces, l'ordre de ces gènes était rigoureusement identique. Elle l'a été également sur le plan biologique en démontrant que cette région du

chromosome 17 de la souris contenait des gènes intervenant dans le développement embryonnaire et dans la spermatogenèse; il est probable que c'est également le cas chez l'homme comme tend à le prouver l'identification du gène *HCGV*, l'analogue humain du gène murin *Tctex-5*. Il restera également à expliquer la signification de cette très forte concentration en gènes conservés dans cette région du génome et à déterminer si elle est ou non en relation avec les fonctions immunologiques des gènes *HLA* ■

TIRÉS À PART

V. David.



**MOSAQUES**  
Association des « X fragile »

77, rue Raspail  
92270 Bois-Colombes, France  
Tél./Fax : 01 47 60 24 99

le samedi 22 mars 1997  
au CNIT-EXPO  
(amphithéâtre GOETHE)  
2, place de la Défense  
92503 PARIS LA DÉFENSE

sur le thème :

LE SYNDROME DE L'X FRA-  
GILE, aspects génétiques,  
cliniques et thérapeutiques

avec la participation de :

Dr Eric Fombonne,  
Président du Conseil Scientifique  
Dr Christophe-Loïc Gérard (Paris)  
Pr Randi J. Hagerman (Denver  
USA)  
Pr Jean-Louis Mandel (Stras-  
bourg)  
Pr Arnold Munnich (Paris)  
Pr Gérard Ponsot (Paris)  
Pr Allan Reiss (Baltimore USA)

## Summary

**The class I region of the MHC genes is one of the most complex in the whole human genome**

The human major histocompatibility complex (MHC) is a 4-Mb genomic segment located on chromosome 6, in the distal portion of the 6p21.3 band. The MHC is a well described region of the human genome and represents a cluster of genes intensively studied because of the important functions of the gene products in the immune system. The total number of genes located within the MHC is now approximately 100, but more can be predicted. The density of genes in the Class II and Class III regions has been known for several years, but the high density of genes in the Class I region did only recently become apparent. This class I region spans 2,000 kb on the telomeric half of the MHC, and contains six class I genes, three of which, the classical human leukocyte antigen genes *HLA-A*, *HLA-B* and *HLA-C*, encode membrane-anchored cell surface glycoproteins that present the endogenous antigenic peptides to the T cell receptors and are recognized by NK cell receptors. These genes are characterized by a high polymorphism

and a nearly ubiquitous expression. The biological function of the three nonclassical class I genes *HLA-E*, *HLA-F*, and *HLA-G* still remains uncertain. In addition to these six genes, the MHC contains a number of class I pseudogenes and gene fragments whose close relationship indicates a common origin from a typical *HLA* class I ancestor gene. This region appears also to harbour various non-*HLA* single copy genes such as *OTF3*, *TCF19*, *S gene*, *TUBB*, *HSR1*, *B30.2*, *ZNF173*, *HCG V*, *HCG-VII*, *MOG* or *OLFR2* as well as eight recently described multigene families named *P5.1*, *3.8-1*, *PERB11*, *BAT1*, *HCGII*, *HCGIV*, *HCGVIII* and *HCGIX* containing a number of sequences displayed throughout the *HLA* Class I region. The presence of all these coding sequences and multigene families suggests that the gene density of the *HLA* Class I region is comparable to that of the class II and class III regions, *i.e.* one gene per 20 or 30 kb. We present here a new map compiling the update knowledge of the structure of this Class I region.

(Tel.: